Published online November 23, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 20 **e162**
doi:10.1093/nar/gnh160

Protein-mediated error correction for *de novo* DNA synthesis

Peter A. Carr[1,2], Jason S. Park[3], Yoon-Jae Lee[4], Tiffany Yu[5], Shuguang Zhang[6] and Joseph M. Jacobson[1,2,*]

[1]Center for Bits and Atoms, [2]Media Laboratory, [3]Department of Mechanical Engineering, [4]Department of Biology, [5]Department of Chemical Engineering and [6]Center for Biomedical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

Received October 8, 2004; Revised and Accepted October 30, 2004

ABSTRACT

The availability of inexpensive, on demand synthetic DNA has enabled numerous powerful applications in biotechnology, in turn driving considerable present interest in the *de novo* synthesis of increasingly longer DNA constructs. The synthesis of DNA from oligonucleotides into products even as large as small viral genomes has been accomplished. Despite such achievements, the costs and time required to generate such long constructs has, to date, precluded gene-length (and longer) DNA synthesis from being an everyday research tool in the same manner as PCR and DNA sequencing. A critical barrier to low-cost, high-throughput *de novo* DNA synthesis is the frequency at which errors pervade the final product. Here, we employ a DNA mismatch-binding protein, MutS (from *Thermus aquaticus*) to remove failure products from synthetic genes. This method reduced errors by >15-fold relative to conventional gene synthesis techniques, yielding DNA with one error per 10 000 base pairs. The approach is general, scalable and can be iterated multiple times for greater fidelity. Reductions in both costs and time required are demonstrated for the synthesis of a 2.5 kb gene.

INTRODUCTION

Major advances in DNA synthesis have been central to progress in biotechnology and basic biomedical research. Powerful examples of this progress include elucidation of the genetic code (1), production of the first synthetic gene (2), sequencing of the human genome (3,4) and the widespread uses of PCR (5,6). Throughout these applications and many others, the ability to synthesize oligonucleotides (7) typically single strands of DNA 10–80 bases in length, has been an essential enabling technology. This synthetic capacity in turn has bred strong interest in the fabrication of larger constructs, genes and gene circuits, from such synthetic oligonucleotide precursors.

Unfortunately, regardless of the approach, the costs and time involved to create genes and longer DNA constructs with high fidelity [currently ∼\$2 per base (8)] still prevent this technology from being an everyday resource in the same manner as oligonucleotide synthesis, PCR or DNA sequencing. In addition, there is strong interest in applications requiring synthesis of far more than a single gene. These include the design of genetic circuitry (9), engineering of entire biochemical pathways (10,11) and even the construction of small genomes (12). Thus, of great appeal would be the availability of DNA synthesized rapidly, with costs of \$0.10 per base or less (i.e. below even current prices for oligonucleotides) at length scales up to or beyond 1 Mb.

In order to enable this vision, new core capabilities are required. The first need is to dramatically decrease the cost of the oligonucleotide 'building blocks'. Important steps in this direction have recently been achieved, building large numbers of genes by harnessing the massively parallel form of oligonucleotide synthesis used to produce oligonucleotide microarrays (13,14).

The second need is to drastically reduce errors. The pervasiveness of flaws in the DNA product forms a substantial obstacle to fast, ultra-low cost gene synthesis. Effort and resources consumed by steps for clonal selection and sequencing are unnecessarily high for short targets, and prohibitive for long ones. Figure 1 illustrates the impact of errors on gene synthesis. For a gene synthesis with a typical error rate of one per 600 bp synthesized (15–18), a 1 kb gene can be obtained by sequencing ∼10 clones. But at this same error rate, sequencing the 100 clones required for even a 2 kb product becomes impractical. Thus, for a large target, multiple rounds of assembly, cloning and sequencing are typically required, as the long product is extremely difficult to synthesize without errors (19). Other strategies include choosing a synthesis target amenable to natural selection (20,21) (useful only for special cases), or performing site-directed mutagenesis to fix mistakes (15) (which still requires at least two rounds of cloning and sequencing, as well as additional oligonucleotide synthesis). Instead, improving the error rate to one per 10 kb synthesized would allow one to sequence a single clone for the 1 kb product, and two clones for 2 kb.

Here, we report an alternative approach to DNA error reduction, demonstrating an error rate 15-fold lower than is typical for *de novo* gene synthesis. This approach is based on the MutS protein, a part of the DNA mismatch repair pathway in a wide

[*]To whom correspondence should be addressed. Tel: +1 617 253 7209; Fax: +1 617 258 6264; Email: jacobson@media.mit.edu

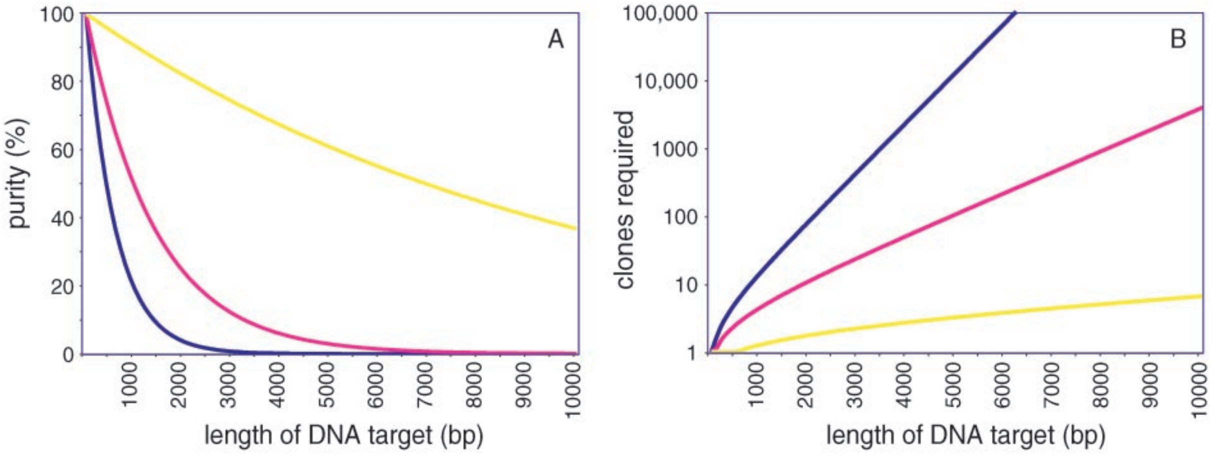

Figure 1. Influence of error rates on *de novo* DNA synthesis. (**A**) The purity of gene synthesis products (yield of error-free clones) decreases exponentially with the length of the product synthesized. Error rates shown are 1 in 600 bp (blue) typical of conventional gene synthesis approaches, 1 in 1400 bp (red) (13), and 1 in 10 000 bp (yellow) as reported here. (**B**) The number of clones which must be sequenced to have a high (95%) probability of obtaining at least one which is error-free. The same three error rates as in (A) are indicated. Calculations are described in Supplementary Table A.

variety of organisms, which binds to many different kinds of DNA mismatches (22). Though affinity for these different types of errors varies, MutS proteins have been shown to bind to all simple one base mismatches, as well as short deletions or insertions of one to four bases (23). In the approach demonstrated here, this affinity for mismatches is used to separate flawed DNA molecules from the desired products in *de novo* DNA synthesis. The method is first illustrated with a simple reporter construct for ease of assaying success, followed by a larger gene (2.5 kb).

As the ease and reliability of such techniques advance, *de novo* DNA synthesis will probably replace all other production methods for which a desired DNA sequence is known. Examples include basic cloning of genes (with sequences known from a database, or designed), expression optimization (already a common use for gene synthesis), site-directed mutagenesis (single changes, or many in tandem) and construction of complex genetic systems. For those engaged in the design of proteins, gene circuits and larger systems, the decreased time and cost associated with producing the molecules of each 'draft' will result in a rapid redesign cycle previously unattainable. We expect that these new DNA synthesis approaches will serve as a key enabling tool and essential foundry for the formative field of synthetic biology.

METHODS

Parsing of the GFP target sequence

A 993 bp target sequence for gene assembly was designed by combining the coding and promoter sequences for enhanced green fluorescent protein (pEGFP, BD Biosciences) with flanking sequences for BP Clonase recombination using the Gateway cloning system (Invitrogen). In addition, a silent mutation (A169 to T, removing a HindIII restriction site) was included near the beginning of the GFP coding sequence to easily differentiate between DNA built *de novo* and possible contamination from other sources.

The sequence was parsed simply into 50mer oligonucleotides (plus two 59mers, one at the 5′ termini of each strand), which were purchased commercially (Integrated DNA Technologies, Inc.) with no additional purification. The oligonucleotides were chosen to represent both the sense and antisense strands of DNA, and were offset by 25 bp to allow maximum overlap between complementary pairs of 50mers. Complete sequences for these oligonucleotides are given in the Supplementary Table B.

PCR assembly/amplification of GFP gene pools

The general scheme for assembly, amplification and error removal is shown in Figure 2A. Oligonucleotides were grouped into four pools, representing four overlapping subsets of the target sequence. The oligonucleotides of each pool were combined and then diluted in deionized water to a stock concentration of 5 μM total (130 nM each oligonucleotide).

Assembly PCR was carried out on each of the four pools of DNA under the following conditions: 1 mM dNTP (250 μM each), 1 U PfuTurbo HotStart Polymerase (Stratagene), pooled oligonucleotides (500 nM total oligonculeotide concentration) in 1× cloned Pfu buffer (Stratagene, 20 mM Tris–HCl pH 8.8, 2 mM $MgSO_4$, 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 0.1% Triton X-100, 0.1 mg/ml BSA) in a total volume of 20 μl. HotStart was performed at 94°C for 2 min prior to thermal cycling of the reaction. Thirty cycles of PCR were performed: melting at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 1 min, with a final 2 min extension at 72°C.

Amplification PCR was subsequently performed using the products of the assembly PCRs as templates: 250 μM each dNTP, 1 U PfuTurbo HotStart Polymerase, 300 nM each primer, 1 μl PCR assembly product (from the previous step) in 1× cloned Pfu buffer, in a total volume of 20 μl. The first (5′) oligonucleotide of the coding and non-coding strands from each pool were used as the PCR primers. The same thermocycling program was used. PCR products were purified by agarose gel electrophoresis.

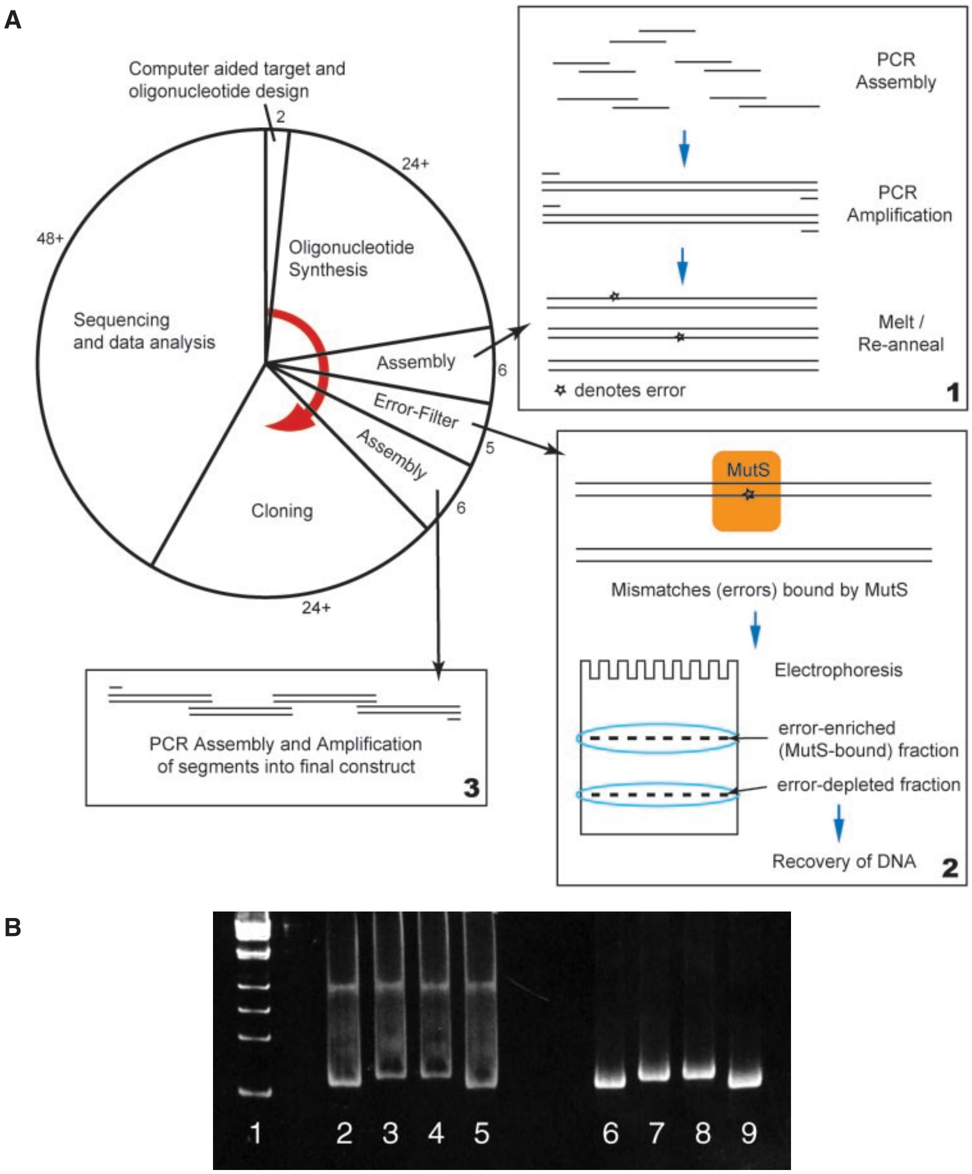
A
Computer aided target and oligonucleotide design
2
24+
48+
Oligonucleotide Synthesis
Sequencing and data analysis
Assembly
6
Error-Filter
5
Assembly
6
Cloning
24+
PCR Assembly
PCR Amplification
Melt / Re-anneal
✫ denotes error
1
MutS
Mismatches (errors) bound by MutS
Electrophoresis
error-enriched (MutS-bound) fraction
error-depleted fraction
Recovery of DNA
2
PCR Assembly and Amplification of segments into final construct
3
B
1
2
3
4
5
6
7
8
9

Figure 2. (**A**) Principal steps performed in the construction of synthetic genes employing MutS protein for error-reduction. The pie chart indicates the approximate amount of time consumed by each step (in hours), with a red arrow indicating the order of operations. The most time-consuming steps in this process are often oligonucleotide synthesis and DNA sequencing (including plasmid production). The 24+ and 48+ hours indicated for each of these represent lower bounds on these processes, possible if performed with immediate access to the appropriate equipment. If these steps are performed by outside providers, 3–5 days are typical of each step. Box 1: gene segments are synthesized and amplified using conventional PCR protocols. The resulting products are dissociated and re-annealed so that errors are present as DNA heteroduplexes (mismatches). Box 2: MutS protein is mixed with this pool of molecules and binds to mismatches. The error-enriched (MutS-bound) fraction is resolved from the error-depleted fraction by electrophoresis. Box 3: The error-depleted segments are assembled into the desired gene and amplified by PCR prior to cloning. (**B**) Polyacrylamide gel electrophoresis of DNA segments used to assemble the GFP gene construct (contrast enhanced). Lane 1: size standard (Kb DNA Ladder, Stratagene; from bottom, sizes are 250, 500, 750 and 1000 bp). Lanes 2–5: the four segments, each complexed with MutS. Lower bands are the error-depleted fractions; upper bands are the error-enriched (MutS-bound) fractions. Lanes 6–9: the same four segments, with no MutS present. Some smearing of the DNA is consistently observed between the two bands in all lanes containing MutS, probably representing protein–DNA complexes which have dissociated.

PAGE-based error-filtration of GFP gene constructs with MutS

Gene constructs with errors were induced to form heteroduplexes by melting the DNA at 94°C and slowly lowering the temperature to 50°C over the course of 20 min. Thus, most errors become part of a heteroduplex in which each error is very likely to be paired with a non-complementary base on the opposite strand. MutS from *Thermus aquaticus* (Epicentre) was mixed with the amplified gene fragments: 2.5 µg MutS with ~50 ng DNA in a 5 µl solution of 8 mM $MgCl_2$, 50 mM NaCl and 10 mM Tris, pH 8. This mixture was incubated at 60°C for 20 min to allow formation of MutS-heteroduplex complexes.

The MutS/DNA mixtures were electrophoresed through a 4–12% gradient TBE PAGE gel (Invitrogen) at 108 V for 30 min and visualized using SYBR-Gold stain (Molecular Probes). The crush and soak method (24) was used to recover DNA from the gel in an elution buffer (10 mM

Tris–HCl pH 7.5, 50 mM NaCl, 1 mM EDTA) for a minimum of 2 h at 37°C, and concentrated by ethanol precipitation using PelletPaint-NF (Novagen) as a co-precipitant. DNA was resuspended in 20 μl of 10 mM Tris–HCl (pH 8). The DNA recovered from the gel was taken from the areas of the gel corresponding to error-enriched, error-filtered and untreated DNA (see Figure 2B). The above error-removal experiment was later repeated to gauge consistency of the technique; in this repeat experiment, error-depleted material from the first experiment was also subjected to error-removal a second time (this time as the full-length 1 kb construct), amplified by PCR, and cloned.

PCR assembly/amplification of final GFP gene construct

Two steps of PCR were used to assemble and then amplify DNA of the four gene fragments into the full-length final gene construct, in PCR reactions identical to those used for assembling and amplifying the original fragments above, with the following modifications. 10 μl (2.5 μl of each) of the resuspended fragments was used in each assembly PCR instead of pooled oligos, and samples were thermocycled for 35 cycles. The first (5′) oligonucleotide of the coding and non-coding strands from the full-length gene were used as the PCR primers.

Cloning of synthesized GFP genes

The full-length EFGP gene constructs were inserted into the pDONR 221 plasmid using the BP Clonase recombination reaction (Invitrogen), with overnight incubation for maximum transformation efficiency. Library Efficiency DH5α cells (Invitrogen) were transformed with these reaction products and grown with kanamycin selection on LB agar plates. Colonies were grown to maturity in 16–18 h and then chosen at random for DNA sequencing (20 or more colonies picked per set), without regard to GFP expression, and grown in LB media in the presence of 15 μg/ml kanamycin. Plasmid DNA was isolated by alkaline lysis (QIAGEN).

Flow cytometry analysis

Cultures were grown at 37°C in a 300 r.p.m. shaking incubator (Lab-Line) from each transformation for use in flow cytometry analysis. Cultures were grown ~18 h and then diluted to 0.6 OD_{600}. The diluted cultures were allowed to grow for an additional hour before being analyzed by flow cytometry using a FACSCalibur system (BD Biosciences) with an argon laser. Live cells were differentiated from dead cells, and debris by analysis of the forward and side scattering properties of cells in the sample. The live cells were analyzed for green fluorescence. The ratio of green fluorescence (530 nm) to yellow fluorescence (585 nm) was measured so that any cells exhibiting autofluorescence (green:yellow ratio ~1:1) could be excluded from the green fluorescent cell count.

Clone selection, sequencing and analysis of errors

Plasmid DNA samples were sequenced for each of four categories of GFP gene synthesis: error-enriched, untreated, error-depleted and twice-depleted (see Figure 4). The number of bases sequenced was similar for each category, with a minimum of 35 kb each. Two sequencing reactions were performed per sample to cover the 993 bp target with two 500+ bp sequence reads. The sequencing primers used (GFP-F: CCTCGTGACCACCCTGAC and GFP-R: CACCAGGGTG-TCGCCCTC) were designed to bind internally to the target sequence, so as to be vector-independent. Errors in the sequenced products were analyzed by sequence alignment using ClustalX (25). Each error was verified by direct visual examination of electropherogram output files with Chromas (Technylesium).

Building the *T.aquaticus* MutS gene

A 2480 bp target sequence was designed by combining the coding sequence for *T.aquaticus* MutS protein with flanking sequences containing an N-terminal His_6 tag and unique restriction sites for cloning. The sequence was parsed using the software DNAWorks (17) into oligonucleotides that were purchased commercially (Integrated DNA Technologies) with no additional purification. The oligonucleotides were separated into eight pools, corresponding to a series of overlapping fragments of ~350 bp each. These sequences are given in Supplementary Table C.

Gene fragments were constructed by the two-step PCR assembly/amplification method, and error-filtered using MutS (Epicentre) and PAGE as described earlier for the GFP gene. The error-depleted fragments were then also assembled and amplified in a two-step PCR procedure as with GFP. dNTP concentrations were increased to 400 μM each for this final PCR amplification step to reflect the increased length of the gene target.

Cloning of the MutS gene

The assembled MutS gene was digested with both NcoI and XhoI restriction enzymes (New England BioLabs) and ligated into the pET-15b vector (Novagen) using T4 DNA ligase (New England BioLabs). The resulting plasmids were transformed into chemically competent MAX Efficiency DH5α cells (Invitrogen). Transformants were grown 16–18 h on LB agar plates containing 50 μg/ml ampicillin. Colonies were picked from these plates, and cultures were grown overnight in LB medium containing 50 μg/ml ampicillin. Plasmid DNA was isolated by alkaline lysis (QIAGEN). DNA samples were screened for the correct size insert by digestion with NcoI and XhoI. One of these samples was sequenced (MIT Biopolymers Laboratory), using five sequencing primers chosen to bind internally to the MutS gene. Analysis of sequences and electropherograms were performed using ClustalX and Chromas as above.

RESULTS

Error rates in DNA synthesis were assessed using the synthesis of a 993 bp DNA target containing a gene for green fluorescent protein, including an upstream promoter and flanking recombination sites for efficient cloning. When correctly synthesized, cloned and transformed into *E.coli*, this product gave rise to fluorescent cells. Thus, error rates could be estimated using colony counts on agar plates, or by flow cytometry, measuring the fluorescence of each cell (Figure 3). Such measurements allow us to quickly determine the effectiveness of a given error correction procedure.

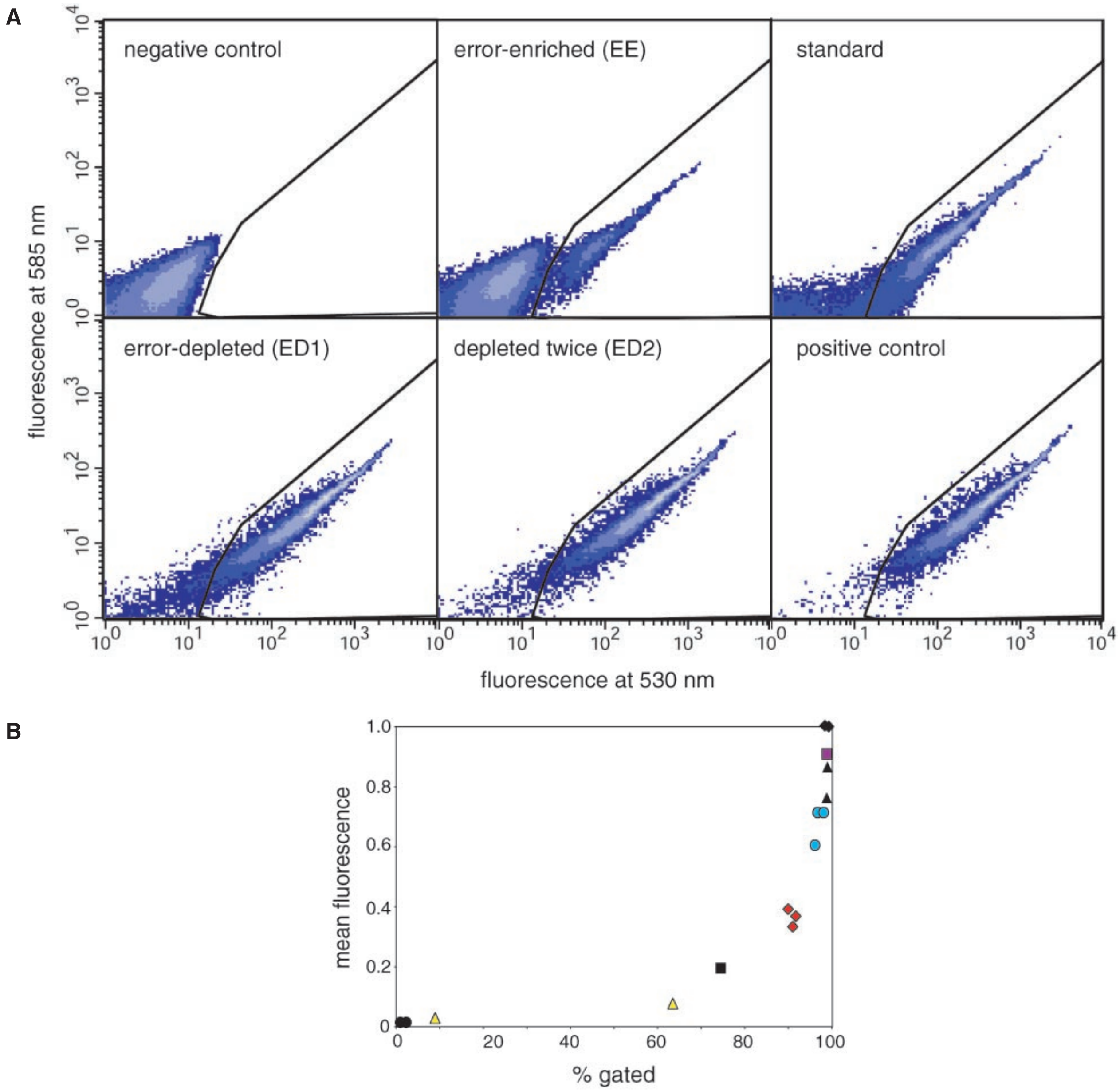

Figure 3. (**A**) Effect of error removal on GFP gene synthesis. Flow cytometry measurements of cells expressing GFP from synthetic genes. Error removal as shown in Figure 2 has been used to improve the quality of the synthesis products. Horizontal axes indicate fluorescence intensity specific to this gene, while vertical axes indicate non-specific fluorescence at a different frequency. Thus, cells which contain successfully synthesized GFP genes are expected to display a minimum level of fluorescence at 530 nm, and substantially less fluorescence at 585 nm (the bounded region in the lower right of each graph). Higher contours (lighter plot color) indicate greater density of cells at a given coordinate. Negative control: expressing a non-fluorescent gene (Tet) in the same vector; Error-enriched: GFP genes produced from MutS-bound DNA fragments; Standard: GFP genes produced by conventional gene synthesis, with no additional processing to remove errors; Error-depleted: GFP genes which have undergone one cycle of error removal; Depleted twice: after two cycles of error removal; Positive control: a correct copy of the same GFP gene, in the same vector. (**B**) Mean fluorescence intensity of each population of cells (50 000 per experiment) as a function of the proportion of fluorescent cells (those in the cut-off region indicated in panel A). Each application of the error-removal process yields an improvement in the quality of the synthetic genes. (black circles): negative control; (yellow triangles): error-enriched; (black square): standard; (red diamonds): 'untreated' DNA subjected to the same manipulations shown in Figure 2, but without the application of MutS protein; (blue circle): DNA error-depleted once using MutS protein; (black triangles): the same GFP DNA employed for the positive control, but amplified by PCR and re-cloned; (purple square): depleted twice; (black diamonds): positive control. Values have been normalized to the mean intensity of the positive control (set at 1). Color symbols indicate sets which were subjected to DNA sequencing and correspond to the symbols shown in Figure 4.

Removal of errors from the pool of synthetic DNA was accomplished with the scheme shown in Figure 2. Initial studies indicated that with the full-length gene, the observable population of DNA was in the MutS-bound fraction, but that smaller pieces gave better separation, probably because of the lower incidence of errors per DNA duplex (P.A. Carr and T. Yu, unpublished data). Thus, the desired DNA product was synthesized in overlapping segments, using conventional

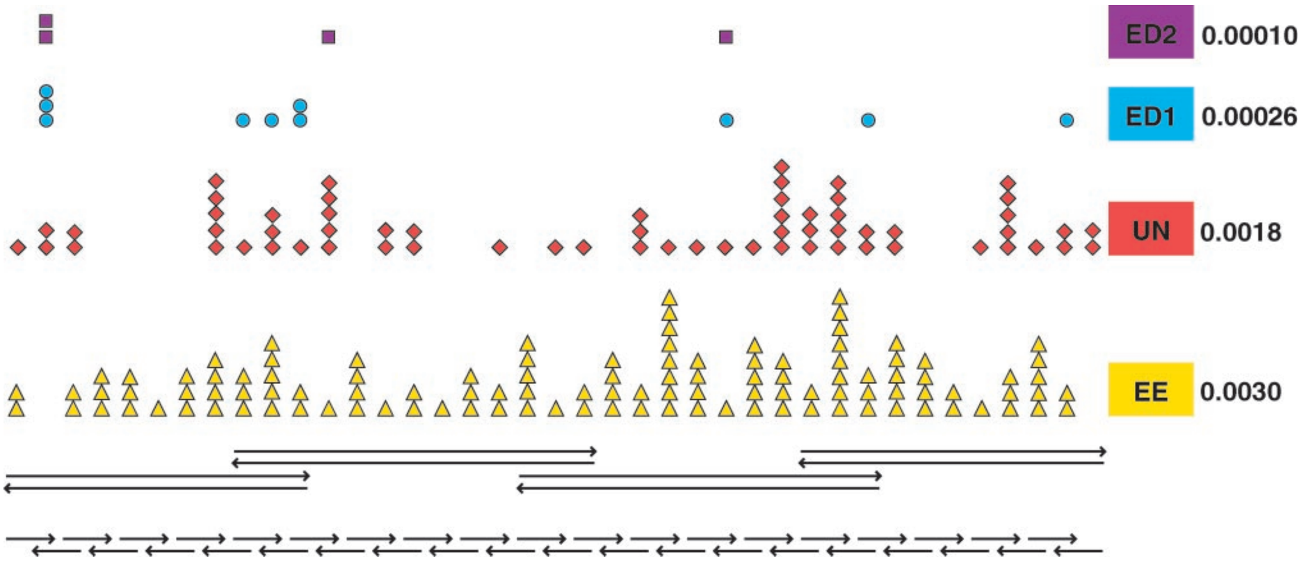

Figure 4. Positions of errors within the GFP DNA synthesis product. From bottom to top: the overlapping set of 38 oligonucleotides (thirty-six 50mers and two 5′-terminal 59mers) used to build the GFP gene and flanking sequences (arrowheads indicate the 3′-terminus of each molecule); the four intermediate assembly products used for the first round of error depletion; positions of errors present in the error-enriched (EE, yellow triangles), untreated (UN, red diamonds), error-depleted (ED1, blue circles), and twice depleted (ED2, purple squares) gene synthesis products. Per-base error rates for each of these sets are also indicated.

PCR assembly techniques (17). After a melting and re-annealing step, errors are present in these mixtures as DNA heteroduplexes—at the site of the error, the DNA duplex is mismatched. This mismatch is then a target for MutS binding. After complexing the DNA mixture with MutS, segments with and without errors were resolved from each other by electrophoresis. (MutS binding changes the electrophoretic mobility of the bound DNA.) The error-depleted fraction is then recovered and used to assemble the full-length DNA product. After the initial reduction in error rate, this process was iterated with the full-length DNA product to further minimize any surviving errors.

To demonstrate the effectiveness of this method, we analyzed the GFP gene products of the error-depleted fraction, as well as untreated samples, and the fraction which was presumed error-enriched. Figure 3 shows flow cytometry data for these sets, with a dramatic reduction in non-fluorescent cells, i.e. the failure products. The error-enriched sample likewise shows a strong increase in non-fluorescent cells. The error-depleted sample was also subjected to a second round of error depletion, with an additional improvement more clearly visible in Figure 3B. The apparent error rate in this twice-depleted pool was indistinguishable from that of one of the controls, a clone known to contain the correct sequence which was amplified using the same PCR primers and cloned back into the same vector (see Figure 3B).

The fluorescence of this gene product is a useful, simple measure for assessing the success of *de novo* DNA synthesis. That being said, some errors will likely be invisible to this procedure, such as those which do not change the amino acid sequence of the encoded protein. In order to reveal silent errors as well as to acquire more detail about which types of errors survived the MutS treatment, sequencing was carried out on the various DNA products (roughly 38 kb sequenced per set). Figure 4 shows the observed error frequencies for samples which were error-enriched, untreated, error-depleted and twice-depleted. Significant improvement is seen after one cycle (1 error per 3.8 kb synthesized) and after two cycles (1 error per 10 kb) of error removal. Table 1 details the number of each type of error observed in the different samples, and errors are tabulated in detail in Supplementary Table D.

Table 1. Summary of errors in GFP gene syntheses

Error type	Error-enriched	Untreated	Error-depleted	Depleted twice
Deletion				
Single deletion				
−G/C	47	28	1	0
−A/T	18	9	0	1
Multiple deletion	25	4	6	0
Insertion				
Single insertion				
+G/C	4	0	0	0
+A/T	3	3	0	0
Multiple insertion	0	0	0	0
Substitution				
Transition				
G/C to A/T	10	9	1	1
A/T to G/C	1	1	0	0
Transversion				
G/C to C/G	4	6	0	1
G/C to T/A	0	2	2	1
A/T to C/G	0	0	0	0
A/T to T/A	0	1	0	0
Other				
GA to T	1	0	0	0
Total errors	113	63	10	4
Bases sequenced	37 440	35 977	38 103	39 080
Error rate (per base)	0.0030	0.0018	0.00026	0.00010

To demonstrate applicability to other, larger DNA synthesis targets, we used the method shown in Figure 2A to construct a 2480 bp gene, a variant of MutS. After extensive characterization of error rates with the GFP gene syntheses, the goal in this instance was to achieve the desired gene with a minimum of effort. Following selection for the correct length insert, a single clone was sequenced, and found to contain the correct sequence.

DISCUSSION

Error rates have been a significant barrier to the construction of large DNA targets. For example, the 7501 bp poliovirus synthesis was achieved at great cost and required many months, largely due to the multiple iterations of assembly and sequencing needed to yield the correct product (19). By contrast, a 2703 bp plasmid synthesis (20) and a 5386 bp bacteriophage ϕX174 synthesis (21) were relatively rapid and inexpensive, but required targets which were easily selected for function (such as antibiotic resistance, or a viable genome) and thus are not general to most DNA synthesis goals.

Errors in synthetic DNA can come from many sources. The dominant source is the oligonucleotides themselves, i.e. errors arise during oligonucleotide synthesis. These can be of different types. Oligonucleotide synthesis can have average stepwise yields (ASWY) of 99%, i.e. roughly one error introduced per 100 bases. Paradoxically, most PCR-based DNA assembly approaches use these oligonucleotides to build larger products with much better error rates, typically one in 600 bases. This is possible because the principal error in olionucleotide synthesis occurs when a single base monomer is not added successfully to the growing polymer chain. This flawed product is acetylated to terminate further chain growth, and interferes little with gene assembly. Instead, the dominant error observed after gene assembly is a short deletion, most often a single base. This can be traced back to a failure to both couple and then acetylate during oligonucleotide synthesis, or to a failure to deprotect during a given addition cycle. The assembly process also has the potential to introduce errors, such as mistakes made by the polymerases used to amplify the products. Errors can occur once the desired product is replicating within a biological host, such as a plasmid within a bacterium, though these errors are expected at much lower frequency.

This method yields a 15-fold reduction in error rates relative to conventional *de novo* DNA synthesis, one error per 10 kb produced (a yield of 0.9999 per base). With this improvement, larger DNA targets can be conveniently synthesized, without resorting to additional cloning steps or excessive sequencing. For example, following the curves shown in Figure 1, it should now be possible to synthesize a 5000 bp target, clone only once, and expect one of three clones to yield the desired product. In the case of our synthesis of the 2.5 kb MutS gene, one clone was sufficient. Our previous efforts to synthesize DNA products of this length required two stages of cloning and sequencing, first to generate smaller segments which were confirmed for the correct content, and then amplified from these separate clones and assembled into the full-length product, at considerably greater time and expense (P.A.C. and J.S.P., unpublished data). These earlier syntheses typically required several more days (for the extra cycles of cloning, transforming and growing bacteria, and for sequencing purified plasmids). That effort also required ten times as many sequencing reactions in order to yield clones with the correct sequence.

Based on the known affinity range of MutS proteins for different types of mismatches (23,26), one might expect some errors to be preferentially removed by the above procedure. Within the limits of the samples sequenced, reductions in all categories of errors are observed (Table 1). Deletions are the dominant form of error in untreated samples (59%), especially single base deletions (44%). The reduction in this category is the most dramatic: for the twice-depleted products only 1 out of 4 errors was a deletion. Of greater surprise was the absence of longer deletions in the twice-depleted products. MutS proteins are not known to bind well to heteroduplexes with deletions longer than 4 bases (23). In contrast, longer deletions were observed in the error-enriched (MutS-bound) fraction—several deletions 5 or 6 bases in length, and one 54 bp deletion (Supplementary Table D).

The distribution of error locations within the sequence also changes as a result of MutS treatment (Figure 4). For the errors which survived the removal procedure, a bias is seen towards the ends of the DNA product, i.e. where PCR primers were used for final amplification after error removal. These account for 43% of the errors in these groups, in a region which is only 12% of the total sequence. (In untreated samples, 11% of the errors fall in this region—there is no apparent bias.) Thus it seems likely that the final amplification is introducing some low level of errors through the PCR primers. In addition, the DNA polymerase used for PCR has the potential to introduce errors. Data shown in Figure 3B is consistent with this hypothesis. Two cycles of error correction brought the overall error rate to roughly the same level as one of the control experiments, a correct copy of the gene which was simply PCR-amplified and re-cloned into the same vector.

Some bias in error location is also observed in favor of errors at the ends of the four gene fragments which underwent the first round of error-depletion, followed by final assembly and amplification (Figure 4, ED1 errors). Many of these errors are within 15 bp of the edge of the DNA duplex, implying MutS binding at these edges may be less effective. These errors are not observed after the second round of error-depletion, performed on the full-length product (ED2 errors).

The current method could be implemented in high-throughput synthetic processes by coupling to capillary electrophoresis for separation of MutS-DNA complexes. However, to make this approach even more amenable to rapid iteration and automation, electrophoretic gel-based separation of MutS-mismatch complexes should be replaced with a simple affinity-based approach. MutS with an additional affinity label can be attached to small amounts of resin, and used to separate mismatched DNA in a simple spin filter in microcentrifuge tubes. Toward this end, we utilized our current error reduction approach to synthesize a version of the *T.aquaticus* MutS gene, adding an affinity motif (a six-histidine peptide). The gene product was shown to be correct after sequencing only one clone, further demonstrating the effectiveness of this technique.

For some applications of synthetic DNA, *in vitro* transcription and translation provide an attractive option for making protein. But in order to produce the correct product, one typically needs to select individual clones and verify their DNA sequences. Now, with sufficiently low error rates, synthetic DNA can be used without any clonal selection step. For example, as shown in Figure 1A, the results of a conventional

1000 bp gene synthesis would typically be 19% pure. But with this form of error correction, that figure improves to 90%. An additional 10-fold improvement in error rate would raise this value to 99%.

How might one achieve the next order of magnitude of improvement? The above data suggest some initial steps. First, errors introduced by late-stage PCR amplification can be addressed by performing a final error removal step immediately before cloning. As an alternative, errors introduced by PCR primers can be made irrelevant by including additional short regions at the termini of the DNA product exclusively for amplification. In combination with these steps, further cycles of the error removal process are likely to yield additional improvement.

Other means for removing errors have been employed for *de novo* DNA synthesis, with varying degrees of success. Initial purification of the oligonucleotides has been performed based on length (21,27), affinity for hydrophobic media and the presence of mismatches (13). However, after such an early purification step, errors can still accumulate in the final DNA product, such as from PCR amplification. Thus, late-stage error processing (i.e. immediately before cloning) is highly desired.

Some error reduction approaches are more amenable to processing larger-sized DNA products. Smith and Modrich applied the mismatch error correction system of *E.coli* (MutS, MutL and MutH) to the challenge of errors generated during PCR. In this method, these components together act to cleave flawed DNA products at a four-base GATC site remote from the errors, and the smaller cleavage products were resolved from the desired DNA by electrophoresis (28). The requirement for a specific site is of some inconvenience for the sequence designer, as targets lacking the site would have to be engineered to include it. In addition, the method was demonstrated for only a small DNA reporter (308 bp), was applied to error rates substantially lower than those encountered in *de novo* DNA synthesis, and only phenotypically observable errors were analyzed by sequencing.

Denaturing high performance liquid chromatography (dHPLC) also has the potential to separate mismatch-containing DNA on the basis of mobility under partial denaturation, even up to 1.5 kb (29). However, the discrimination of dHPLC is most effective in assays designed for mutation detection, i.e. when protocols have been designed around specific known variations. In such cases the choice of DNA fragments, melting temperature and other experimental parameters are optimized around these specifics. In the more general case of *de novo* synthesized DNA, a method must be robust for all possible errors which may be encountered.

Use of functional selection has also proven a powerful method for obtaining higher quality clones in some examples, such as synthesis of a gene for antibiotic resistance (20) or for a DNA product encoding a replicative bacteriophage (21). However, only occasionally will such an approach be feasible for a desired target—most *de novo* DNA synthesis targets will not be easily selectable. In addition, such approaches do not prevent silent mutations, or even some conservative mutations within coding regions, and the selective pressure is typically even weaker outside coding regions.

With all other approaches employed to act on the full-length product, error reduction is global with respect to the DNA product, i.e. an entire molecule is discarded based on an error somewhere in the sequence, even though >99% of that molecule may be error-free. This is also the case with the method introduced here. However, the binding of MutS to an error-containing duplex is an intrinsically local event. Thus, we expect it is possible to achieve local removal of errors, in essence salvaging the correct regions. This capability would be especially valuable in the synthesis of long (>1 kb) products, where the initial assembly becomes very unlikely to provide many of the correct molecules to select from. Such an approach need only be limited in size by the length of DNA products which can be effectively assembled and amplified by PCR. PCR has been used to successfully amplify targets of at least 42 kb in length (30).

Furthermore, no other general method has demonstrated error rates commensurate with those reported here. While it is possible that the combination of two or more methods may provide even further improvement in the quality of synthesized DNA, it seems likely that multiple rounds of MutS-mediated error reduction will render these other approaches unnecessary.

The method described here yields DNA error reduction which is effective, robust and general with respect to the choice of synthetic target. It also raises the possibility of error correction applied directly to large targets, independent of size, without the need to synthesize smaller intermediate products. Options exist for the automation of such MutS-mediated error reduction strategies, which will enable high-throughput production of high quality synthetic DNA.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank George Church, Jingdong Tian, Brian Chow, Christopher Emig, David Kong and members of the Zhang and Jacobson groups for helpful discussions. This work has been supported by NSF grants EIA-0323439 and CCR-0122419, and by DARPA grants DAAD16-01-C-0048 and F30602-01-2-0596.

REFERENCES

1. Khorana,H.G. (1968) Nucleic acid synthesis in the study of the genetic code, in Nobel Lectures: Physiology or Medicine (1963–1970). Elsevier Science Ltd, Amsterdam, pp. 341–369.
2. Agarwal,K.L., Buchi,H., Caruthers,M.H., Gupta,N., Khorana,H.G., Kleppe,K., Kumar,A., Ohtsuka,E., Rajbhandary,U.L., Van de Sande,J.H. *et al*. (1974) Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. *Nature*, **227**, 27–34.
3. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
5. Kleppe,K., Ohtsuka,E., Kleppe,R., Molineux,I. and Khorana,H.G. (1971) Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *J. Mol. Biol.*, **56**, 341–361.

6. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
7. Caruthers,M.H. (1985) Gene synthesis machines: DNA chemistry and its uses. *Science*, **230**, 281–285.
8. Carlson,R. (2003) The pace and proliferation of biological technologies. *Biosecur. Bioterror.*, **1**, 203–214.
9. Elowitz,M.B. and Leibler,S. (2000) A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**, 335–338.
10. Martin,V.J., Pitera,D.J., Withers,S.T., Newman,J.D. and Keasling,J.D. (2003) Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nat. Biotechnol.*, **21**, 796–802.
11. Mehl,R.A., Anderson,J.C., Santoro,S.W., Wang,L., Martin,A.B., King,D.S., Horn,D.M. and Schultz,P.G. (2003) Generation of a bacterium with a 21 amino acid genetic code. *J. Am. Chem. Soc.*, **125**, 935–939.
12. Hutchison,C.A., Peterson,S.N., Gill,S.R., Cline,R.T., White,O., Fraser,C.M., Smith,H.O. and Venter,J.C. (1999) Global transposon mutagenesis and a minimal Mycoplasma genome. *Science*, **286**, 2165–2169.
13. Tian,J., Gong,H., Sheng,N., Zhou,X., Gulari,E., Gao,X. and Church,G.M. Accurate multiplex gene syntheses from programmable DNA chips. *Nature*, in press.
14. 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.
15. Baedeker,M. and Schulz,G.E. (1999) Overexpression of a designed 2.2 kb gene of eukaryotic phenylalanine ammonia-lyase in *Escherichia coli*. *FEBS Lett.*, **457**, 57–60.
16. Withers-Martinez,C., Carpenter,E.P., Hackett,F., Ely,B., Sajid,M., Grainger,M. and Blackman,M.J. (1999) PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome. *Protein Eng.*, **12**, 1113–1120.
17. Hoover,D.M. and Lubkowski,J. (2002) DNA works: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.*, **30**, e43.
18. Chalmers,F.M. and Curnow,K.M. (2001) Scaling up the ligase chain reaction-based approach to gene synthesis. *Biotechniques*, **30**, 249–252.
19. 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.
20. Stemmer,W.P., 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.
21. Smith,H.O., Hutchison,C.A.,III, Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl Acad. Sci. USA*, **100**, 15440–15445.
22. Modrich,P. (1991) Mechanisms and biological effects of mismatch repair. *Annu. Rev. Genet.*, **25**, 229–253.
23. 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.
24. Sambrook,J. and Russell,D.W. (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
25. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, **25**, 4876–4882.
26. 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.
27. Au,L.C., Yang,F.Y., Yang,W.J., Lo,S.H. and Kao,C.F. (1998) Gene synthesis by a LCR-based approach: high-level production of leptin-L54 using synthetic gene in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **248**, 200–203.
28. Smith,J. and Modrich,P. (1997) Removal of polymerase-produced mutant sequences from PCR products. *Proc. Natl Acad. Sci. USA*, **94**, 6847–6850.
29. Xiao,W. and Oefner,P.J. (2001) Denaturing high-performance liquid chromatography: a review. *Hum. Mutat.*, **17**, 439–474.
30. Cheng,S., Fockler,C., Barnes,W.M. and Higuchi,R. (1994) Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl Acad. Sci. USA*, **91**, 5695–5699.