Synthesis and Error Correction Methods in Gene Fabrication

by

Jason Sun-hyung Park

Submitted to the Biological Engineering Division
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering

at the

Massachusetts Institute of Technology

May 2006
[June 2006]

Signature of Author: _____
Biological Engineering Division
May 25, 2006

Certified by: _____
Joseph Jacobson
Associate Professor of Media Arts and Sciences and Mechanical Engineering
Thesis Supervisor

Certified by: _____
Bevin P. Engelward
Associate Professor of Biological Engineering
Thesis Reader in Biological Engineering

Accepted by: _____
Ioannis Yannas
Professor of Polymer Science & Engineering
Chair, Biological Engineering Graduate Program Committee

Synthesis and Error Correction Methods in Gene Fabrication

by

Jason Sun-hyung Park

Submitted to the Biological Engineering Division
on May 25, 2006 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in
Biological Engineering as Recommended by the Biological Engineering Division

ABSTRACT

Gene Fabrication technology involves the development and optimization of methods relevant to the in vitro synthesis of any given target gene sequence(s) in the absence of template. The driving purpose of this field of research is to bring about the capability for on-demand fabrication of a DNA construct of arbitrary length and sequence quickly, efficiently, and cost-effectively.

The first part of this document describes many of the important considerations in performing successful de novo gene synthesis from a survey of the literature as well as from our own work. Recommendations are made for a universally effective, robust, and simple protocol for potential users of gene synthesis, discussing important factors such as choice of protocol, source of commercial oligonucleotides, and polymerase choice.

The second part of this document focuses on error correction. Reducing error rates is one of the main challenges in gene fabrication because high error rates preclude the possibility of fabricating long gene targets in a practical and economical manner. Improvements in error rates are essential for continued progress in the development of gene fabrication technology. I discuss the importance of error rate in gene synthesis from a practical standpoint and show results in the development of novel methods for the removal of errors from a pool of synthesized DNA.

Thesis Supervisor: Joseph Jacobson
Title: Associate Professor of Media Arts and Sciences and Mechanical Engineering

Table of Contents

List of Figures

List of Tables

Introduction

Background / Motivation

Gene fabrication involves the development and optimization of methods relevant to the in vitro synthesis of any given target gene sequence(s) in the absence of template. The driving purpose of this field of research is to bring about the capability for on-demand fabrication of a DNA construct of arbitrary length and sequence quickly, efficiently, and cost-effectively. Current gene synthesis technology is effective for making single genes, but can be slow and costly, especially for long sequences of DNA or when multiple genes are desired. Through developments in gene fabrication technology, we hope to remove these boundaries and provide an invaluable tool in biological and biomedical research and engineering.

The isolation and manipulation of genes and other large DNA molecules is crucial to many areas of molecular biology research. While powerful techniques in traditional molecular biology have been developed over the years for all sorts of DNA manipulations, the time and labor expended on such work is substantial. An alternative would be to directly synthesize desired genes, but synthetic genes are expensive with typical costs (commercially ordered) of $1 to $2 per base (ie. $1000-$2000 for a 1 kb gene) and turn-around times of 2-4 weeks. Inexpensive, reliable, and fast gene snthesis would greatly increase the number and types of experiments which could be carried out in such areas as the study of gene pathways and de novo protein design, and facilitate the construction of large gene libraries at reasonable cost.

There are many uses for synthetic DNA, especially with the continued growth of DNA sequence and protein structure databases available online. For example, the ability to directly synthesize a gene and several variants can often be useful in the study of a gene in biological research, especially with codon optimization for expression in bacterial or other systems. Also, synthesizing a gene *de novo* eliminates the need to obtain an organism from its natural habitat in order to study one of its genes. Gene fabrication technology also allows for the synthesis of genes with novel functionalities that do not even exist in nature. The ability to synthesize multiple genes at once also gives researchers the ability to engineer and analyze complete biochemical pathways, genetic networks, and entire genomes.

One can imagine that with the continued development and optimization of gene synthesis technology, ordering synthesized genes will one day be as quick and affordable as ordering single-stranded DNA oligonucleotides today. When one wants to obtain an oligonucleotide today, the process is as simple as filling out an order form, waiting a couple days, and paying $10-$20. While this was not true as recently as 5-10 years ago, costs have come down dramatically for oligo synthesis through new developments. Once gene synthesis technology reaches the maturity that oligonucleotide synthesis technology has achieved, many basic DNA manipulations now performed in the laboratory (for example mutagenesis, cloning, purifications) could be replaced by simply ordering the exact DNA species desired.

The goals for gene fabrication technology do not stop at the single-gene level, however. Whereas even robust single-gene synthesis would be an enabling technology for many research applications, the ability to synthesize large sets of genes at minimal cost would enable certain fields of research that are currently prohibitively expensive. These include studying interactions between every member of a particular functional class (Newman et al, 2003), synthesizing many mutants of a single gene (for example, alanine scanning mutagenesis (Cunningham et al, 1989)), protein design, or labeling all the genes from a single genome with an antigenic peptide tag or fluorescent protein (Ehrhardt, 2003) for detection.

New kinds of research and design projects will become possible with the development of improved gene synthesis technology. In the field of Synthetic Biology, researchers are developing increasingly large and more complex artificial genetic systems, built mostly from modified genes found in nature (Elowitz et al, 2000; Basu et al, 2005; Levskaya et al, 2005; Voigt et al, 2005). Notable examples of completely non-biological DNA designs are beginning to emerge as well (Shih et al, 2004; Park et al 2006). The DNA constructs required for some proposed projects are so large that conventional gene synthesis becomes prohibitively expensive. These include complex in vitro genetics systems such as that proposed by Church and colleagues (Tian et al 2004). Other groups are pursuing the direct synthesis of entire simple genomes (Smith et al, 2003). Such projects aim both to test hypotheses on the fundamental requirements for life, and to generate organisms that have been dramatically re-engineered for new purposes, such as waste processing, energy production, and complex syntheses of useful compounds.

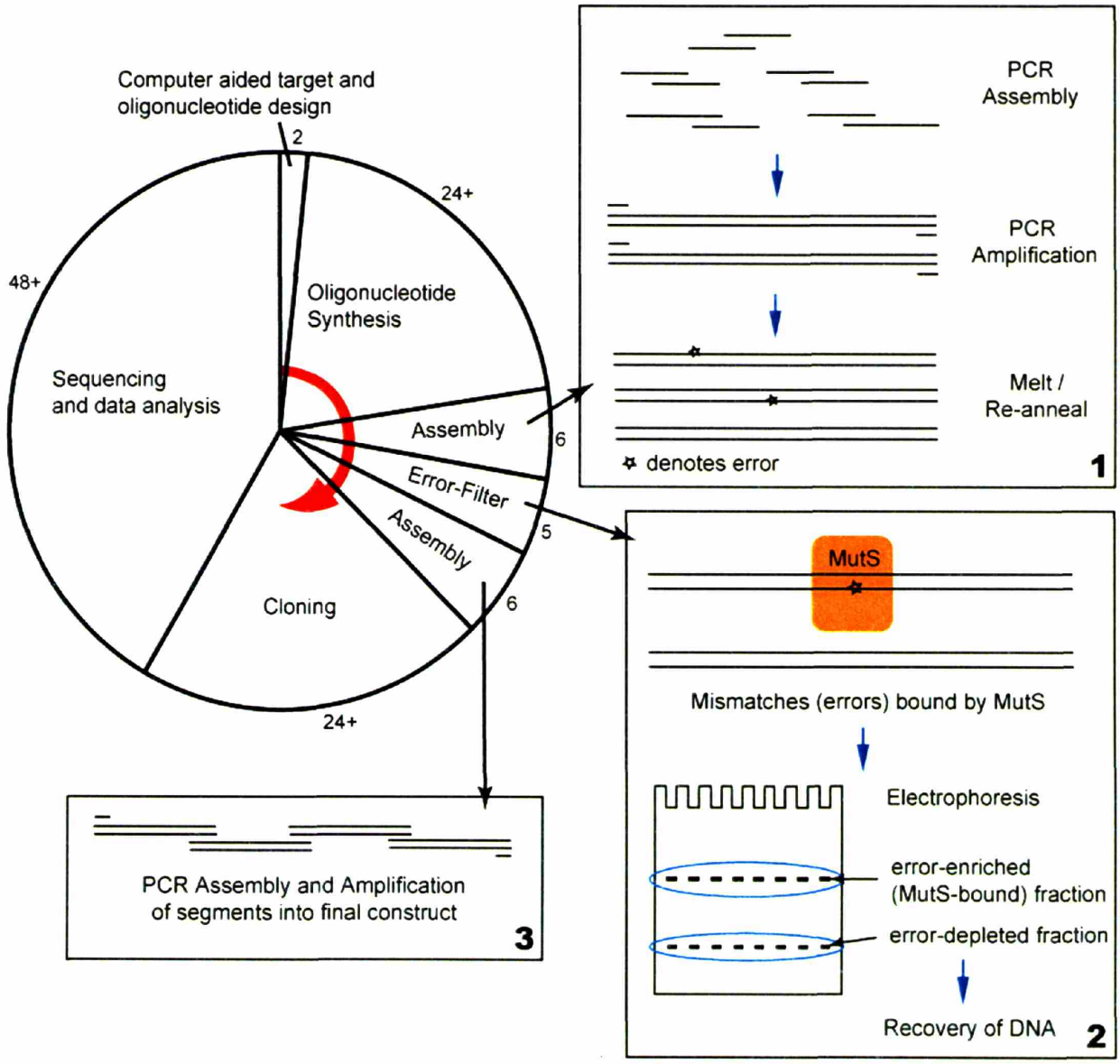

Figure 1 Principle steps in the construction of synthetic genes employing MutS for error-reduction. The pie chart indicates the approximate amount of time consumed by each step (in hours), with a red arrow indicating the order of operations. The most time-consuming steps in this process are often oligonucleotide 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 DNAheteroduplexes (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. (From Carr et al. 2004)

Gene Synthesis Methods

While commercial sources of synthetic DNA are widely available and becoming increasingly affordable (Carlson, 2003), researchers may benefit from the ability to make one's own DNA targets, if they can do so without too much experience or effort.

The first technique for designing synthetic genes in the laboratory was published over 35 years ago (Khorana, 1968). Since then, many variations of protocols on gene synthesis have arisen in the literature. However, there has been relatively little work comparing variables and protocols in gene synthesis methods. Some of the parameters that vary most often between protocols have been the source of oligonucleotides, the method used to parse the gene targets, choice of ligase and/or polymerase-based approaches, assembly protocol, and more. We set out to address the need to give potential users of gene synthesis a clear picture and set of recommendations regarding the most important factors to optimize and address in doing robust and high-quality (low error rate) gene synthesis, combining an analysis of the gene synthesis literature and empirical measurements.

Recent Landmarks in Gene Synthesis and the Need for Error Reduction

There have been a number of recent landmarks in gene synthesis from a number of separate teams led by Venter, Cello, Church, and Santi (Smith et al, 2003; Cello et al, 2002; Tian et al, 2004; Kodumal et al; 2004). These groups have demonstrated the ability to synthesize large molecules of DNA from oligos, with products 5.4, 7.5, 15 and 32 kb in length, respectively. It is notable that error reduction was of critical importance to each of these groups' accomplishments. Both the Venter and Church teams found it necessary to purify their oligonucleotides prior to assembly. The Venter team used a gel-based size separation prior to ligation while the Church team used a series of selective hybridization selections with selection oligonucletides (Smith et al, 2003; Tian et al, 2004). In one of the reports, the target was a bacteriophage genome and natural selection reminiscent of work by Stemmer et al in 1995 allowed for automatic selection of functional clones (though silent mutations could not be detected by the functional assay). While natural selection is a useful tool, the technique was not generalizable as most gene targets are not selective. In the other reports, an interrmediate step of cloning and sequencing of intermediate segments was necessary to get correct full length product.

Significance of Error Rate and Error Rate Reduction for Gene Fabrication

Error Correction: Theory and Consequences

The theory behind error correction in gene fabrication lies at the intersection of computer science and biology: error-correcting codes. The central idea behind error-correcting codes is that multiple noisy or unreliable inputs can be combined to give high-fidelity output. In "consensus voting," copies of a signal are compared against one another and the consensus is chosen as the output. In this way, each input signal "votes" on the output signal.

Von Neumann formulated the requirements to simulate a given number of perfect inputs (Winograd and Cohen, 1967): "A circuit containing N (error-free) gates can be simulated with probability of error at most ε using N log (N/ε) faulty gates, which fail with probability p, so long as $p < p_{th}$." The key feature here is the logarithmic relationship. In terms of gene fabrication, in order to improve the reliability of the system by a factor of x, the number of DNA molecules required for error correction scales by only log x.

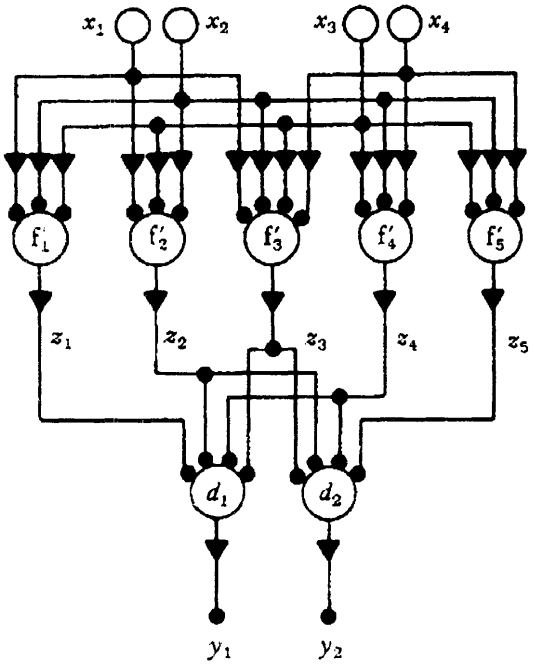

Figure 2 Example of a circuit employing error correcting codes
Multiple noisy signals x_i are routed to components f_i for "consensus voting." The result proceeds to a second round of "consensus voting" by similar components d_i. (Winograd and Cohen, 1967).

In either an electronic or a biological setting, one needs a means of detecting errors in order to apply the principles of error-correcting codes. In nature, biological systems maintain low error rates in DNA replication by comparing newly synthesized DNA to its original template. For example, methylation of a strand of DNA at a recognition site marks it as an original – and presumably error-free – template. In the case of an error, a number of repair mechanisms are set in motion. The mismatch repair system of many organisms detects an error where a misincorporated base fails to form a Watson-Crick base pair with the base on its complementary strand (Modrich, 1991).

The concept of an original, presumably error-free copy of DNA acting as a template for detection of errors is key component of error correction *in vivo*. However, gene fabrication, which is *de novo* DNA synthesis, is quite different. There is no distinguishable original, perfect copy of DNA. All DNA molecules produced have a discrete probability of containing fairly randomly distributed errors in their sequence. The standard of quality also differs between gene fabrication and natural processes. Biological systems, because of silent mutations, mutations with insignificant effect, or compensatory mutations, are more fault-tolerant than gene fabrication processes. In contrast, a perfect copy of DNA must be made from scratch in gene fabrication. Error filtration, correction, and prevention methods are therefore of critical importance.

The error rate of a gene fabrication process directly affects its usability to synthesize DNA for its various applications. What error rate is required to economically and realistically make DNA of a given sequence length? In terms of molecular biology, this translates to: "How many clones must be sequenced to have adequate confidence of obtaining at least one perfect clone?" The calculation goes as follows: For a per-base error rate of P and a DNA target of length N, the probability that a given clone contains no errors at any position is $(1-P)^N$. For X clones sequenced, the confidence (C) of obtaining at least one perfect clone is $C = 1-(1-(1-P)^N)^X$. Conversely, for a given confidence value, the number of clones one expects to sequence is $X = \ln(1-S) / \ln(1-(1-P)^N)$. Figure 3 indicates how much more difficult it is to synthesize, for example, a 6000mer than a 600mer. At a standard error rate of 1 error in 600 bp (Stemmer et al, 1995, Withers-Martinez, et al., 1999, Hoover and Lubkowski, 2002, as well as our own observed rates) a 600mer can be produced with high confidence and only a few clones sequenced. On the other hand, without a better error rate, one can see that synthesizing a 6000mer is unrealistic and

nearly impossible. One way to get around this problem would be two synthesize ten 600mer pieces and clone and sequence them, and then assemble these together into a 6000mer and clone and sequence them again. On the other hand, an improvement in error rate to 1 in 6000 would allow one to generate a DNA target about 6000 bp in length with little trouble.

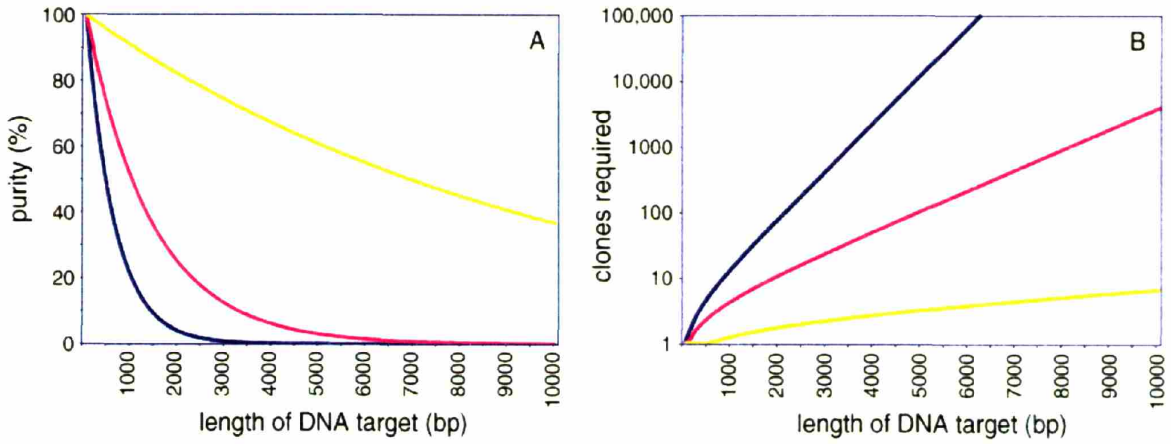

Figure 3 Gene fabrication of long targets requires improved error rates
A. The purity (defined as the % error-free clones) of gene synthesis products decreases exponentially with the length of the product synthesized. Error rates shown are 1 in 600 bp (blue) typical of conventional gene synthesis approaches, 1 in 1400 bp (red), and 1 in 10,000 bp (yellow). B. The number of clones that must be sequenced to obtain at least one that is error-free (95% Confidence Interval. The same three error rates as in (A) are indicated. (From Carr et al., 2004)

Error correction is also a vital aspect of gene fabrication because it cuts down on sequencing and cloning. Cloning, preparing samples, and sequencing take up almost two-thirds of the total process time in gene fabrication. Figure 1 shows the process of gene fabrication as a time pie chart, illustrating the point that by more accurately synthesizing gene targets, one can not only cut down on the number of clones necessary to get one perfect copy, but also avoid the need for any time-consuming additional steps such as site-directed mutagenesis.

In order to employ the concept of error-correcting codes in gene fabrication, one makes the assumption that errors are far outnumbered by correct DNA. Therefore, with many copies of DNA, a consensus vote between the DNA strands allows for determination of the correct sequence and allows for errors to be found and fixed. To accomplish this, an error correction system must be able to accomplish error detection, error removal, and error repair. The overarching functional requirements in the development of error correction protocols are: a) utilization of a molecule with affinity for errors, b) a method for removal of the errors which are

bound by aforementioned molecules, and c) a process that can be cycled for improvement of error rate by multiple rounds of consensus voting. The ideal result is a cheap, robust, and hopefully machine-automatable procedure with short cycle time that results in large improvements of error rate.

Recognizing errors with MutS: in vivo vs. in vitro

The error-binding molecule of choice in much of our current and future work in the development of error correction protocols employs the protein MutS. As part of a common natural mismatch repair mechanism involving MutL, MutH, and MutS, MutS is a protein with affinity for binding to DNA duplexes in places where they deviate from Waston-Crick base pairings. MutS, which has homologs in many organisms, exhibits different sensitivity to different types of mismatches (Brown et al. 2001). The sensitivity and specificity of the mismatch binding also varies across species.

Affinity is highest for single base deletions, which we have determined to be the most prominent error in gene synthesis. The different types of mismatches are: AA, CC, GG, TT, AC, AG, TC, TG, and the bulges caused by short insertions and deletions. E. coli MutS is known to have poor affinity for CC mismatches (Brown et al., 2001). In eukaryotes, multiple MutS homologs with distinct functions together perform the task of mismatch recognition (Eisen, 1998). MutS proteins from thermophilic bacteria, seem to have affinity for all of the types of mismatches (Biswas and Hsieh, 1995, Whitehouse et al. 1997).

Figure 4 details the mechanism of action of MutS, MutL, and MutH. As mentioned earlier, methylation of one of the strands of DNA identifies it as the original copy. In the absence of strand methylation, both strands are cut by MutH (Smith and Modrich, 1997). This feature has been utilized in the literature for the removal of errors from synthesized DNA. The cleaved, error-containing products are separated from non-error-containing products by size separation in gel electrophoresis.

In de novo gene fabrication, we do not have the luxury of knowing which DNA strands are "original" and "correct." Further, because of the nature of PCR-based gene synthesis, errors in DNA are copied into the comlementary strands of DNA as well. Thus there is the necessity for dissociating and reassociating DNA duplexes in order to re-assort DNA strands and create error heteroduplexes in which errors are matched with their corresponding correct bases. This

mismatch can then be recognized by a protein such as MutS. For lengths of DNA <10-20 kbp, this re-assortment can be accomplished by thermal denaturation and re-annealing. For larger targets, one might utilize proteins such as RecA to effect strand transfer between duplexes.

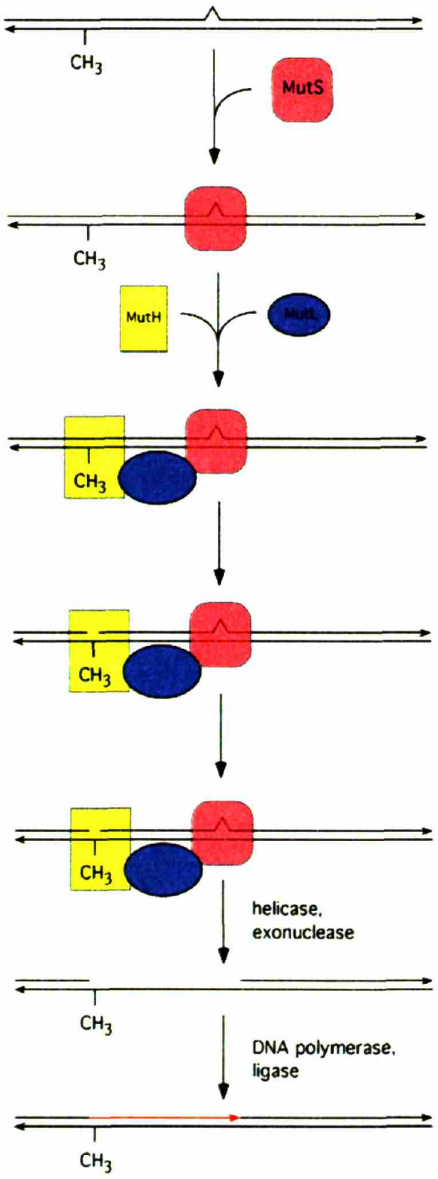

Figure 4 The mismatch repair system of E. coli.
A DNA heteroduplex is shown. Arrowheads indicate the 3' end of each strand. The top strand contains a single error (indicated by bulge). The bottom (template) strand contains a methylated GATC sequence. MutS protein (a dimer, shown as one unit for simplicity) binds to the site of the mismatch. MutL and MutH proteins bind to the MutS-DNA complex, and MutH scans the nearby DNA for a GATC (potentially looping out intervening sequence, not shown). When MutH finds the site, it nicks the unmethylated strand. Actions of a helicase and an exonuclease digest part of the top strand, until the error is degraded. A DNA polymerase and ligase then fill in the gap, resulting in a corrected strand (Modrich 1991).

One of the main issues with MutS for use in error correction in gene fabrication is nonspecific binding. Especially with longer DNA constructs, nonspecific binding may be a problem. Depending on the application, our data shows that MutS can be used effectively to remove errors from DNA of length about 1000 bp or less. In the synthesis of various gene products including GFP (~1000 bp) and Thermus aquaticus MutS (~2500 bp), we circumvented this issue parsed the sequences into chunks of ~300 bp, error-correcting these pieces, and then assembling them into the full gene targets through a PCR-based method. However, a method of synthesis and error correction that does not require this additional step would be preferable.

There exist a number of other molecules used for binding to DNA mismatches, including T7 endonuclease I (Babon et al., 2000), T4 endonuclease VII (Youil, 1996), and CEL I endonuclease (Oleykowski et al, 1998). These may serve as ready alternatives to the use of MutS and its homologs.

Scope of this document

This document focuses on two aspects of gene fabrication: gene synthesis and error correction. Other issues in Gene Fabrication e.g. automation/parallelization by microfluidics and use of DNA microarrays as a source of oligonucleotides, while important, are beyond the scope of this work.

Methods

General protocols

Listed here are a number of general protocols and reagents used in our work. These are not specific to Gene Fabrication but are listed here for reference. For information on proprietary reagents and protocols, please refer to manufacturers' documentation.

Cloning

We use the plasmid vector pDONR221 with the Clonase II (Invitrogen) recombination system for convenient, low-background cloning of some gene targets, such as GFP.

We also use the T7lac-promoter based pET system (Novagen) of vectors for protein expression. We use different variants of the vector depending on additional features we may want included in our protein expression. For example, we could use pET-44 to have a NusTag fused to our protein of interest for enhanced protein solubility.

Restriction enzymes are obtained from New England Biolabs.

Chemically competent cells are obtained from Invitrogen: some common types we use include DH5α MAX Efficiency, DH5α Library Efficiency, and BL21(DE3).

Protein expression/purification

As mentioned above, we use isopropyl-beta-D-thiogalactopyranoside (IPTG)-inducible T7lac systems for protein expression. Our standard procedure for protein expression involves an overnight culture growth at 37°C and 300pm from a colony pick followed by 1:100 dilution into fresh LB with antibiotic and re-growth at 37°C and 300pm to mid-log phase (~0.6 OD_{600}). The culture is induced with 1 mM IPTG and incubated at 37°C and 300rpm for 2 hours before harvesting the cells in a centrifuge. We use either sonication or BugBuster reagent with Benzonase nuclease (Novagen) to lyse cells.

We use an AKTApurifier (GE Healthcare) system to automate much of our protein purification work. We use high-flow columns for affinity, ion exchange, and other types of column protein purification.

Commercial polymerases for PCR

We have used the following commercially available polymerases in our work:
Taq (NEB), PfuTurbo (Stratagene), PfuUltra (Stratagene), PfuUltraII (Stratagene), Phusion (Finnzymes), Pfx50 (Invitrogen), KOD HiFi (Novagen), and BD Advantage 2 (Clontech). We use the manufacturer-supplied buffers and manufacturer-recommended thermal cycling parameters for our PCRs.

Agarose gel electrophoresis

1% agarose gel electrophoresis with 0.5 μg/mL ethidium bromide or 1X SYBR Safe (Molecular Probes) is used to analyze PCR products. Qiagen Gel Extraction Kits (Qiagen) are used to extract DNA from agarose gel.

Polyacrylamide gel electrophoresis – protein and DNA gels

Polyacrylamide gel electrophoresis (PAGE) is used to analyze protein and DNA samples. Pre-cast TBE gradient gels (4%-12%) (Invitrogen) are used for DNA analysis and stained with SYBR Gold (Molecular Probes). The crush and soak procedure is used to recover DNA samples from PAGE. Precast Bis-Tris gels (6%) (Invitrogen) are used for protein analysis and stained with Simply Blue SafeStain (Invitrogen).

Gene synthesis: PCR-based methods

As discussed earlier, there are a variety of methods employed for gene synthesis, each with their unique advantages and disadvantages. The most prominent among these – and the protocol that we favor – involves the use of the Polymerase Chain Reaction (PCR) to assemble a pool of oligonucleotides that together make up the target sequence of interest. We favor the nomenclature Polymerase Construction and Amplification (PCA) for the process – this was the term coined by Mullis in the first report to employ a thermocycled polymerase-based method (Mullis et al, 1986).

Sequence parsing

A target DNA sequence for gene fabrication is parsed into a set of overlapping oligonucleotides (~40-70 bp in length) in computer software. The choice of parsing algorithm depends on the particular needs of a project.

In the simplest parsing scheme, the target sequence can be divided up into equal chunks of n base pairs. In more complicated algorithms, the sequence can be parsed in ways so as to optimize for certain parameters including consistent melting temperature, no hairpin formation, no self-annealing, no primer-dimerization, codon frequency in host organisms, and more. To allow an extra degree of freedom in oligonucleotide sequence choice, some software allows gaps and overlaps between adjacent oligonucleotides of the same strand. An example of this type of software, hosted at the NIH, is DNAWorks (http://molbio.info.nih.gov/dnaworks/).

The EGFP and Tma MutS sequences used for the optimization and comparison of One-Step PCA and Two-Step PCA described in this document were parsed using DNAWorks.

Purchasing oligonucleotides commercially

The parsed oligonucleotides are purchased from a commercial vendor. Choice of vendor is important, as different vendors provide DNA with different error rates. We have demonstrated that the errors in a starting oligonucleotide pool can make up a high proportion of the errors present in its resultant synthesized gene product. We compared oligonucleotides from a number of vendors including Integrated DNA Technologies (IDT), Sigma, Operon, and MWG Biotech.

The EGFP oligonucleotides used for the optimization and comparison of One-Step PCA and Two-Step PCA described in this document were purchased commercially from Operon with no additional purification. The Tma MutS olignonucleotides were purchased commercially from IDT with no additional purification.

One-Step Gene Synthesis: Polymerase Construction / Assembly (One-Step PCA)

While a number of variations on polymerase-based methods for gene synthesis are in the literature (Gao et al, 2003), we favor simple approaches we call One-Step PCA and Two-Step PCA. In the next two protocols, the term "outer amplification oligonucleotides" refers to the first oligonucleotides (5') of the coding and non-coding strands from each pool.

In One-Step PCA, 300 nM of each of the outer amplification oligonucleotides and an amount of the entire pool of oligonucleotides ranging in concentration between 0-50 nM per oligo are combined (See Figure 5). dNTPs are added to the reaction to a concentration of 1 μM total (250 nM each). Reactions are carried out in polymerase manufacturer-provided 1X reaction buffer(s) with a manufacturer-recommended amount of polymerase.

Reaction mixtures are thermocycled for 45 cycles of denaturation, annealing, and extension. Temperatures and times for all steps, as well as those for the initial denaturation and final extension steps, are those recommended by each polymerase's manufacturer.

Products are analyzed and purified by agarose gel electrophoresis.

We tested a number of important variables in the One-Step PCA protocol to determine its limitations and compare it with the Two-Step PCA protocol described next.

Two-Step Gene Synthesis: Polymerase Construction / Assembly (Two-Step PCA)

Two-Step PCA, our preferred protocol for gene synthesis, consists of an Assembly PCR followed by an Amplification PCR (See Figure 6).

Assembly PCR reactions (in a total volumes of 20 μL) are set up with an oligo pool concentration of about 15 nM (each). dNTPs are added to the reaction to a concentration of 0.8 μM total (200 nM each). Reactions are carried out in polymerase manufacturer-provided 1X reaction buffer(s) with manufacturer-recommended amounts of polymerase units used in each reaction. Amplification PCR reactions (in total volumes of 50 μL) are set up with a 1:20 dilution of the assembly PCR material, 300 nM of each of the outer amplification oligonucleotides, and 0.8 μM total dNTP (200 nM each). Again, reactions were carried out in polymerase manufacturer-provided 1X reaction buffer(s) with manufacturer-recommended amounts of polymerase units used in each reaction.

Both steps of the PCA are carried out for 30 cycles, with initial denaturation, final extension, denaturation, annealing, and extension steps' cycling parameters as recommended by each polymerase's manufacturer.

Products are analyzed and purified by agarose gel electrophoresis.

A number of important variables in Two-Step PCA were examined and their impact on an important readout of quality, error rate, was determined. Two-Step PCA was also compared side-by-side with One-Step PCA in terms of robustness of assembly.

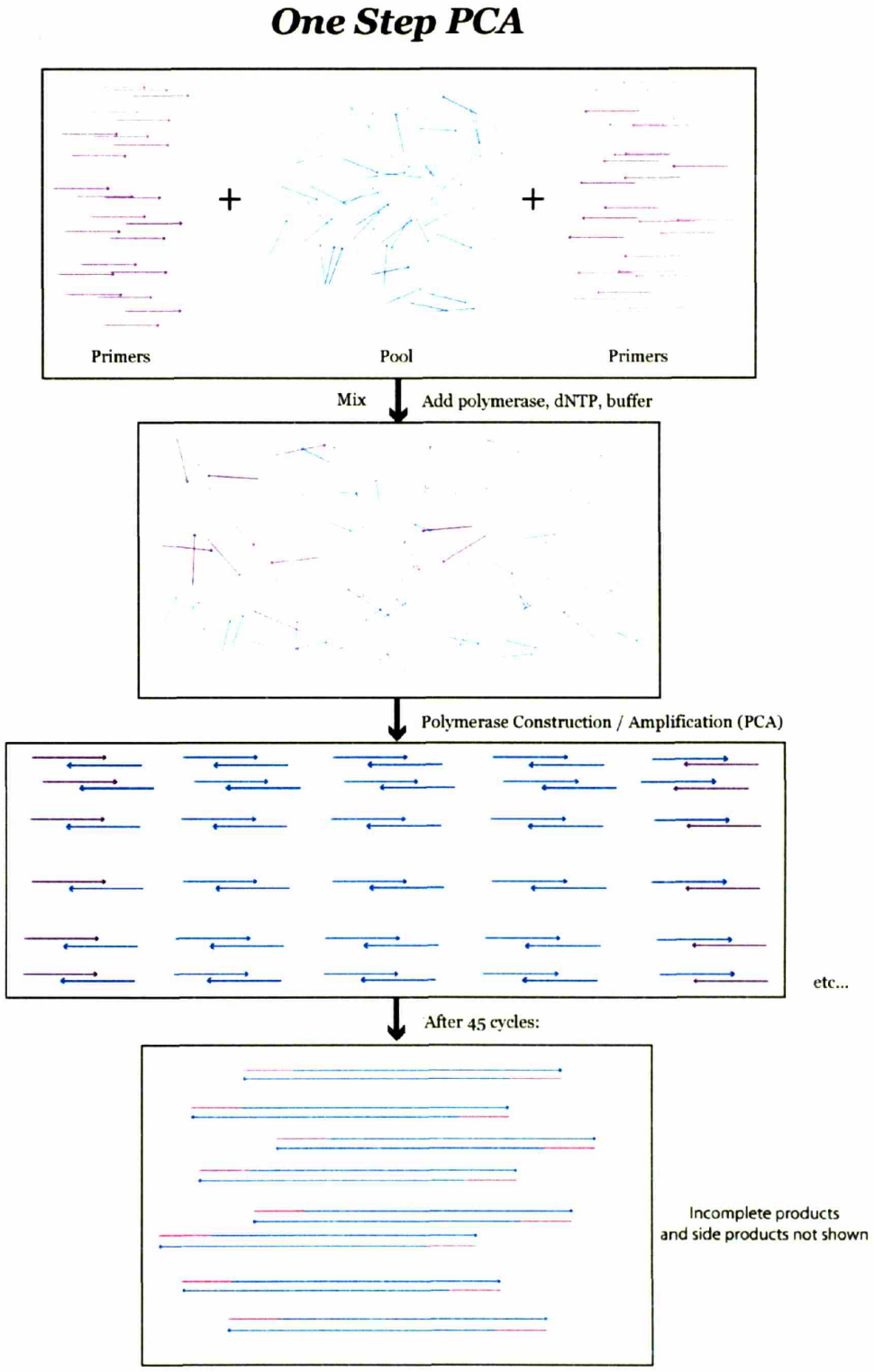

Figure 5 One-Step Polymerase Construction / Assembly (PCA)
In One-Step PCA, a pool of construction oligonucleotides and a much higher concentration of the outermost (5' end) oligos of the coding and non-coding strands are combined in a Polymerase Chain Reaction. Reaction mixtures are thermocycled for 45 cycles of denaturation, annealing, and extension. The product is a mixture of full-length gene target and shorter incomplete products and side-products. Thermocycling parameters used are those recommended by each polymerase's manufacturer.

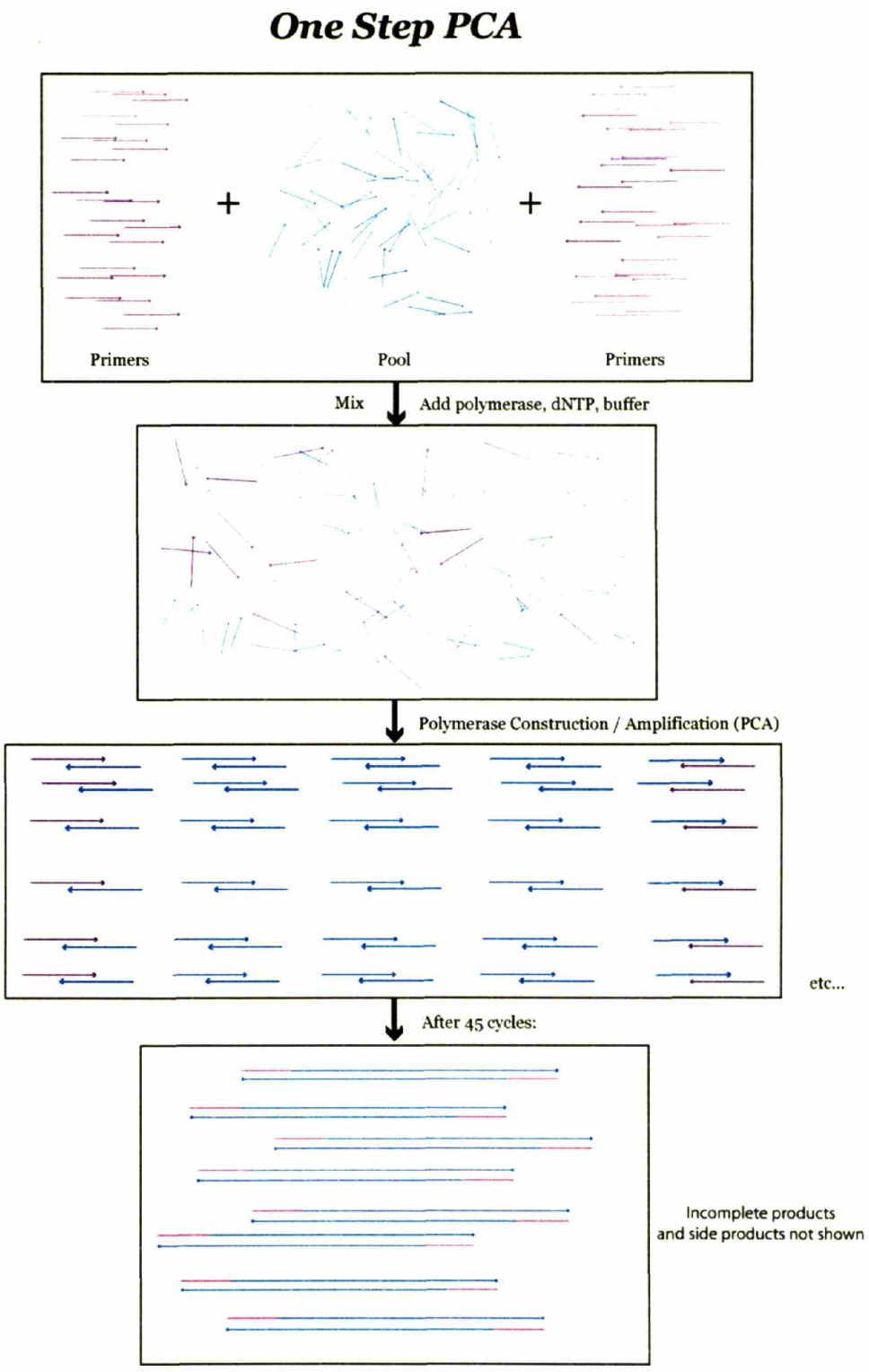

Figure 5 One-Step Polymerase Construction / Assembly (PCA)
In One-Step PCA, a pool of construction oligonucleotides and a much higher concentration of the outermost (5' end) oligos of the coding and non-coding strands are combined in a Polymerase Chain Reaction. Reaction mixtures are thermocycled for 45 cycles of denaturation, annealing, and extension. The product is a mixture of full-length gene target and shorter incomplete products and side-products. Thermocycling parameters used are those recommended by each polymerase's manufacturer.

Assays to determine the error rate in a synthesized DNA construct

As discussed earlier, one of the primary measures of quality and a limiting factor in gene synthesis is the rate at which errors occur in the synthetic DNA. Therefore, it is of paramount importance to optimize gene synthesis and error reduction protocols to yield error rates that are as low as possible. This implies the need for quick, accurate, and economical assays for error rate. Below are the assays we employ most frequently in our work.

Colony Count: Green Fluorescent Protein (GFP) in E. coli

By using Enhanced Green Fluorescent Protein (EGFP) (pEGFP, BD Biosciences) as the desired target for gene synthesis, one can roughly assay the error rate of a given gene fabrication process by assessing the percentage of colonies on a plate of selective agar-containing media that are glowing green under UV illumination.

Full-length EGFP gene constructs are synthesized as described earlier and inserted into the pDONR 221 plasmid using the BP Clonase II recombination reaction (Invitrogen), with overnight incubation for maximum transformation efficiency. Competent DH5alpha cells (Invitrogen) are transformed with these reaction products and cells are plated with kanamycin selection on LB agar plates and incubated at 37°C. Colonies are grown to maturity in 16–18 h. The plates are imaged on a Fluorchem8900 gel documentation system (Alpha Innotech) under UV illumination (365 nm) with a SYBR Green filter. Adobe Photoshop is used for image enhancement (gamma setting ~3, white/black threshold adjusted for optimal contrast, false color green).

From a green/white colony count, one can calculate an approximate error rate per base pair using the following equation:

$$(\text{average \% silent mutations})^{-1} * (\%\ \text{green}) = (1\text{-error rate per bp})^{\text{DNA length}}$$

$$1 - ((\text{average \% silent mutations})^{-1} * (\%\ \text{green}))^{1/(\text{DNA length})} = \text{error rate per bp}$$

Note that we can do a similar assay and calculation synthesizing the LacZ-alpha gene and doing a blue/white screen.

Flow cytometry of E. coli expressing Green Fluorescent Protein from synthesized DNA

This assay can roughly be thought of as a scaling up of the GFP colony count to analyze thousands of single *E. coli* cells in a flow cytometer instead of single colonies on a plate.

After transformation as described in the Colony Count procedure, the mixture of transformed cells are diluted in 5 mL LB with 15 μg/mL kanamycin and grown at 37°C in a 300 r.p.m. shaking incubator (Lab-Line). Cultures are grown ~18-20 hours and diluted 400-fold into LB just prior to analysis. Cultures are analyzed using either a FACScan or FACSCalibur system (BD Biosciences) with a 488nm argon laser in the MIT Center for Cancer Research Flow Cytometry Core facility. The Low flow setting is used for the FACScan system and the Medium flow setting is used for the FACSCalibur system. Live cells are differentiated from dead cells and debris by analysis of the forward and side scattering properties of cells in the sample. The live cells are analyzed for green fluorescence. The ratio of green fluorescence (530 nm) to yellow fluorescence (585 nm) is measured so that any cells exhibiting autofluorescence (green:yellow ratio ~1:1) can be excluded from the green fluorescent cell count. CellQuest Pro (BD Biosciences) and FlowJo (Tree Star) software are used to analyze flow cytometry data.

Use of MutS in a gel mobility shift assay

Binding varying amounts of MutS to a known amount of synthesized DNA and analyzing by polyacrylamide gel electrophoresis can serve as a rough assay for error rate. The DNA is thermally denatured and re-annealed before this process. MutS binds more readily to a pool of DNA that contains many errors than to error-free DNA. This effect can be observed by staining the gel for DNA, i.e. with SYBR Gold Stain (Molecular Probes). MutS-bound DNA migrates through the gel slower than unbound DNA.

This process is also used to purify error-depleted DNA from error-containing DNA. Refer to the "Gel-based error-filtration via MutS-binding" protocol in the next section for more information.

Sequencing

Sequencing is the gold standard for error rate determination. The downside of this tool though, is that it is costly both in terms of time as well as money. This limits the sample size of

data sets collected – whereas we can easily collect tens of thousands of data points for a set using flow cytometry, we can only realistically send off tens to hundreds of clones for sequencing.

Regardless, sequencing gives us information that is simply not available using any of the other techniques. We can precisely count what types of errors exist in our synthesized DNA and we can note their location on the DNA strand. Also, GFP flow cytometry and colony counts do not take into account silent mutations or mismatch mutations that only change the intensity of fluorescence.

We use MIT Biopolymers and Genaissance Pharmaceuticals for sequencing. Both accept 96-well plates – Genaissance only accepts 96-well plates. Samples are provided to MIT Biopolymers at appropriate concentrations / volumes with primers premixed. Genaissance accepts bacterial glycerol stocks (10% glycerol in selective LB media) (> 10 μL) shipped overnight on dry ice, from which they produce material for sequencing using a TempliPhi reaction (GE Healthcare). Sequencing primers are sent for the entire plate in 1.5 mL tubes or in 96-well plate format.

One sequencing reaction is performed per sample per ~500bp in length using appropriate primers (internal to the target sequence, sequences in the vector flanking the insert, etc as necessary). Errors in the sequenced products are analyzed by sequence alignment using VectorNTI Advance 10: AlignX (Invitrogen) and verified by direct visual examination of electropherogram output files with Chromas Lite (Technylesium).

Error Removal methods

Reassorting errors into heteroduplexes

As discussed in the Introduction, DNA duplexes in a pool of DNA duplexes in a pool of synthesized DNA need to be dissociated and reassociated after mixing in order to assort DNA strands to create error heteroduplexes where errors are matched with their corresponding correct bases. For lengths of DNA <10-20 kbp, this re-assortment can be accomplished by thermal denaturation and re-annealing. (For larger targets, one might utilize proteins such as RecA to effect strand transfer between duplexes). In our experience, heating the DNA mixture in a thermocycler to 95°C for 2 minutes and then gradually ramping down the temperature to room

temperature over the course of 20-30 minutes has been sufficient for strand reassortment. The presence of salt is important to prevent DNA backbone repulsion during this process.

We have observed that this relatively fast anneal protocol seems to create higher-order structures, some of which are presumably similar in structure to Holliday junctions in recombination. This is evidenced by high-molecular mass aggregates of DNA that do not migrate well through non-denaturing TBE polyacrylamide gel electrophoresis. These aggregates disappear in denaturing TBE polyacrylamide gel electrophoresis. In addition, we have observed that application of a Holliday junction resolvase such as Thermus T7 Endonuclease I (NEB) causes the DNA to migrate in gel electrophoresis as normal DNA duplexes of the expected length.

Gel-based error-filtration via MutS-binding

Newly synthesized DNA product is thermally denatured and re-annealed to re-assort errors as described above. The DNA is then bound to recombinant MutS by incubation at 60°C for 10 minutes in a binding buffer (50 mM NaCl, 10 mM Tris pH 8, and 8 mM $MgCl_2$).

The mixture is then analyzed by TBE polyacrylamide gel electrophoresis (PAGE) and the gel is stained for DNA by SYBR Gold (Molecular Probes). Error-containing DNA is bound by MutS, which shifts its mobility in the gel. DNA that is not shifted In mobility is error-depleted. This error-depleted DNA is cut out and extracted from the gel by the crush and soak method at 37°C in an elution buffer (10 mM Tris pH 7.5, 1 mM EDTA, 50 mM NaCl). This minute amount of DNA can be amplified by PCR. The procedure can be repeated or the DNA can be used for cloning.

The precise ratio of MutS to DNA used is important: too much MutS causes nonspecific binding to dominate and there is effectively no error-free DNA not bound by the MutS, while too little MutS does not bind enough error-containing DNA to make a substantial improvement in error rate.

The optimal MutS:DNA molar ratio varies depending on the origin of the MutS (what species), the length of the DNA, and the error rate of the DNA. However, as a starting point, we note that the optimal ratio for *Thermus aquaticus* (Taq) MutS seems to be about 275:1 (MutS monomer to double-stranded DNA). This corresponds to a 40:1 mass ratio for a 1000bp DNA construct.

Removal of error-containing DNA via immobilized MutS

In order to improve the cycle time and ease of implementation of error-filtration using MutS, we worked to extend our results in MutS-mediated gel-based error-filtration by developing a protocol to perform error-filtration with MutS immobilized to solid supports. We made use of the His tags we put on our proteins (originally for protein purification) to bind the MutS to Ni-NTA resins (Pierce) and Ni-IDA HisMag magnetic agarose beads (Novagen).

In this protocol, we bind an excess of our His-tagged MutS to the resin or magnetic beads and incubate at room temperature for several minutes. After two successive washes in binding buffer (50 mM NaCl, 10 mM Tris pH 8, and 8 mM $MgCl_2$) (sediment by centrifugation or applying magnet), the synthesized DNA – melted and re-annealed to re-assort errors as described earlier – is added. After incubation at room temperature for 30 minutes, the MutS-solid support complex is sedimented and the supernatant (error-depleted DNA) is set aside for analysis. The DNA can be roughly quantitated by taking an Abs_{220nm} spectrophotometer reading or running on a polyacrylamide gel and staining with SYBR Gold stain (Molecular Probes). The material can also be amplified by PCR and the procedure can be repeated. Sequencing, GFP colony count, or flow cytometry (in the case of GFP construct) can be used to determine the error rate of the DNA.

Synthesis and expression of several variants of MutS

Using the methods discussed in this document, we have synthesized and expressed variants of MutS from the following organisms: *Escherichia coli* (Eco), *Thermus aquaticus* (Taq), *Thermatoga maritima* (Tma), and *Aquifex aeolicus* (Aae). Because MutS is a relatively large protein (~100 kDa), we employed MutS-mediated gel-based error-filtration to facilitate the successful construction / cloning of these genes (first using commercially available Taq MutS from Epicentre, later using our own recombinant Taq MutS).

Results

Critical Variables and Considerations in Practical Gene Synthesis

Polymerase-based methods vs. Ligase-based methods

We set out to fill a need in the literature for a comprehensive set of practical recommendations for users of gene synthesis. We began by surveying the literature and examining the many different approaches in existence for gene synthesis.

First, we decided on polymerase-based approaches to gene synthesis – in which a pool of oligos prime off of each other to make progressively longer DNA species until they reach full-length and are amplified exponentially as in PCR – over ligase-based approaches. In ligation-based methods, complementary groups of oligos are annealed together and covalently joined by DNA ligase. One of the disadvantages of ligation-based approaches include the requirement that oligos be phosphorylated at their 5' ends in order for the ligation reaction to occur. Phosphorylating oligos during oligo synthesis is expensive, and phosphorylating oligos with polynucleotide kinase post-synthesis is not difficult but adds an additional biochemical processing step to the gene synthesis process.

In addition, no gaps or overlaps are allowed in parsing genes for ligase-based approaches. In polymerase-based methods, gaps and overlaps are allowed because the DNA polymerase can extend the DNA chain from the end of an oligo according to the complementary DNA strand, but ligase simply covalently links adjacent DNA molecules. Allowing gaps and overlaps adds a degree of freedom in software parsing in terms of avoiding "hotspots" for mispriming, adjusting the melting temperature of an oligo, and more. Also, PCR-based gene synthesis is tolerant of a good portion of the errors resulting from olignucleotide synthesis in which (capping groups..) – provided that no errors were introduced into the truncated oligo – whereas incorporation of such an oligonucleotide by ligation-based gene synthesis could possibly result in an error in the final product.

We favor the nomenclature Polymerase Construction and Amplification (PCA) for the aforementioned family of polymerase-based approaches to gene synthesis, after the term coined by Mullis in the first report to employ a thermocycled polymerase-based method (Mullis et al, 1986). Variations of this general protocol in the literature include approaches with several

sequential assembly/amplification reactions (Stemmer et al, 1995; Smith et al, 2003) and methods where the entire gene synthesis process is condensed into a single reaction (Tian et al, 2004; Wu et al, 2006). The recommended arrangement and concentration of construction olignucleotides also varies in the literature. Most of the time, the oligonucleotides are parsed to arrange themselves linearly (refer to Figure 5 or Figure 6). However, alternatives presented in the literature include a circular arrangement to generate a concatamer (Stemmer et al, 1995) and oligonucleotides arranged to assemble from the center of the construct on out (Gao et al, 2003). In terms of concentration, some recommend the same concentration of all olignucleotides in the pool, some a high concentration of the "outside" primer oligos, and others a gradient.

One-Step PCA vs. Two-Step PCA

Given a straightforward parse and a linear arrangement of construction oligonucleotides, one needs to make the choice between using one or two steps of PCR in synthesizing the gene construct. In Two-Step PCA, the first PCR is an assembly PCR – all oligos are present at the same concentration and successively longer gene products are formed as the oligos prime and extend off one another. The second PCR is an amplification PCR, where a high concentration of the "outside" primers and a dilution of the assembly PCR are put in the reaction to efficiently and selectively amplify the full-length gene product of interest. In One-Step PCA, the assembly and amplification steps are condensed into a single reaction with both a high concentration of "outside" primers and an otherwise uniform concentration of construction oligos in the pool. One-Step PCA is presented in (Tian et al, 2004) and studied further in (Freeland et al, 2006), though under different nomenclature in each case.

While One-Step PCA has the advantage of reduced sample handling and somewhat reduced reaction time, our experience had been that One-Step PCA is very effective for short gene products (<500bp) but that the approach gave less specific product and was less robust than Two-Step PCA and required more optimization for longer products.

We systematically evaluated One-Step PCA, taking into account four key parameters: polymerase/buffer choice, length of construct, number of construction oligos in the pool, and initial concentration of construction oligos in the reaction.

DNAWorks-parsed oligonucleotides derived from GFP target sequence were separated into four pools with expected sizes of correctly assembled products from each being 264 bp (12

oligonucleotides), 475 bp (22 oligonucleotides), 682 bp (32 oligonucleotides), and 993 bp (42 oligonucleotides). A 2406 bp target sequence (Tma MutS gene) was separated into five pools in a similar manner. The sizes of the correctly assembled products from each of the five pools were expected to be 545 bp (20 oligonucleotides), 1075 bp (38 oligonucleotides), 1621 bp (56 oligonucleotides), 2163 bp (74 oligonucleotides), and 2406bp (90 oligonucleotides).

The polymerases / buffers used for this experiment were PfuTurbo, BD Advantage 2, and Phusion. The range of oligo concentrations used was 0-50 nM (each).

Looking at Figure 7 and Figure 8, one can see some trends important in choosing these key parameters in One-Step PCA. BD Advantage 2 (Clontech) outperforms the other polymerases in terms of maximum gene length, yielding products as lage as 1621bp (and possibly 2163bp). However, this polymerase exhibits an unacceptably high error rate and is not recommended for use in PCA. There is clearly an optimal oligo pool concentration in the vicinity of 10-20 nM (each). There is length-dependence – longer products with more oligos tend to be more difficult to build by One-Step PCA than shorter products with fewer oligos.

Next, we compared non-optimized Two-Step PCA builds with optimized One-Step PCA builds. For simplicity and to test sensitivity to construction oligo construction, we left the total oligo pool concentration of the first assembly PCR at 500 nM total for all reactions. Thus, per-oligo concentration varied depending on construct, and for the 250 bp assembly each oligonucleotide was present at 35 nM while for the 2500 bp product each oligo was present at only 5 nM. Despite this, strong product bands were generated in all Two-Step PCA's. See Figure 9 for a direct comparison of optimized One-Step PCA builds and non-optimized Two-Step PCA builds for constructs of a variety of lengths with Phusion polymerase.

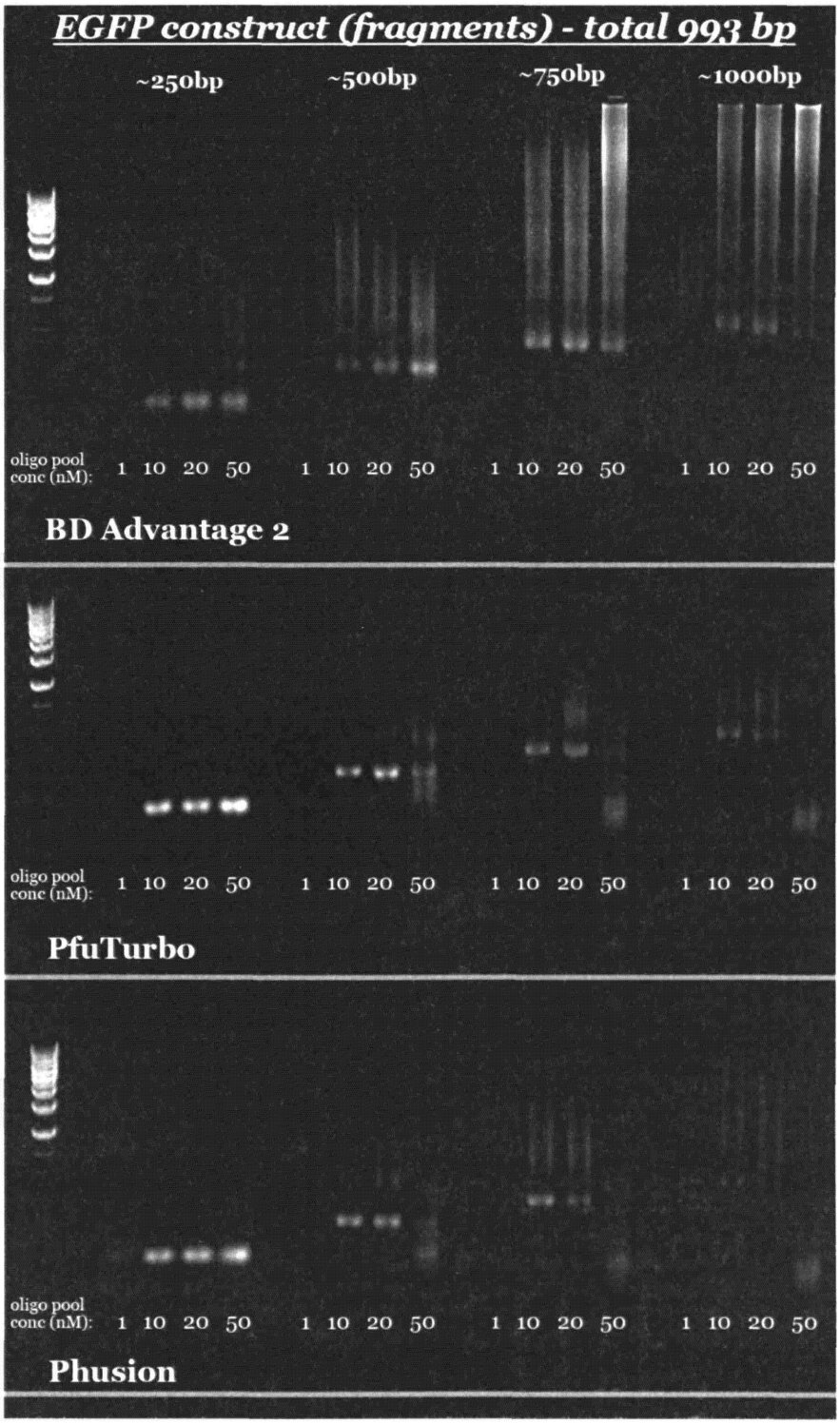

Figure 7 One-Step PCA of EGFP fragments with various polymerases at varied oligo pool concentrations. Target DNA constructs of length 264 bp (12 oligos), 475 bp (22 oligos), 682 bp (32 oligos), and 993 bp (42 oligos) from the EGFP gene were assembled using the One-Step PCA protocol. Robustness of assembly for varying concentrations of oligo pool (1-50 nM each) was assessed. 4 uL of each PCA product was run on the 1% agarose gel alongside 2 uL kb ladder (Stratagene). All images were enhanced for contrast with the same parameters.

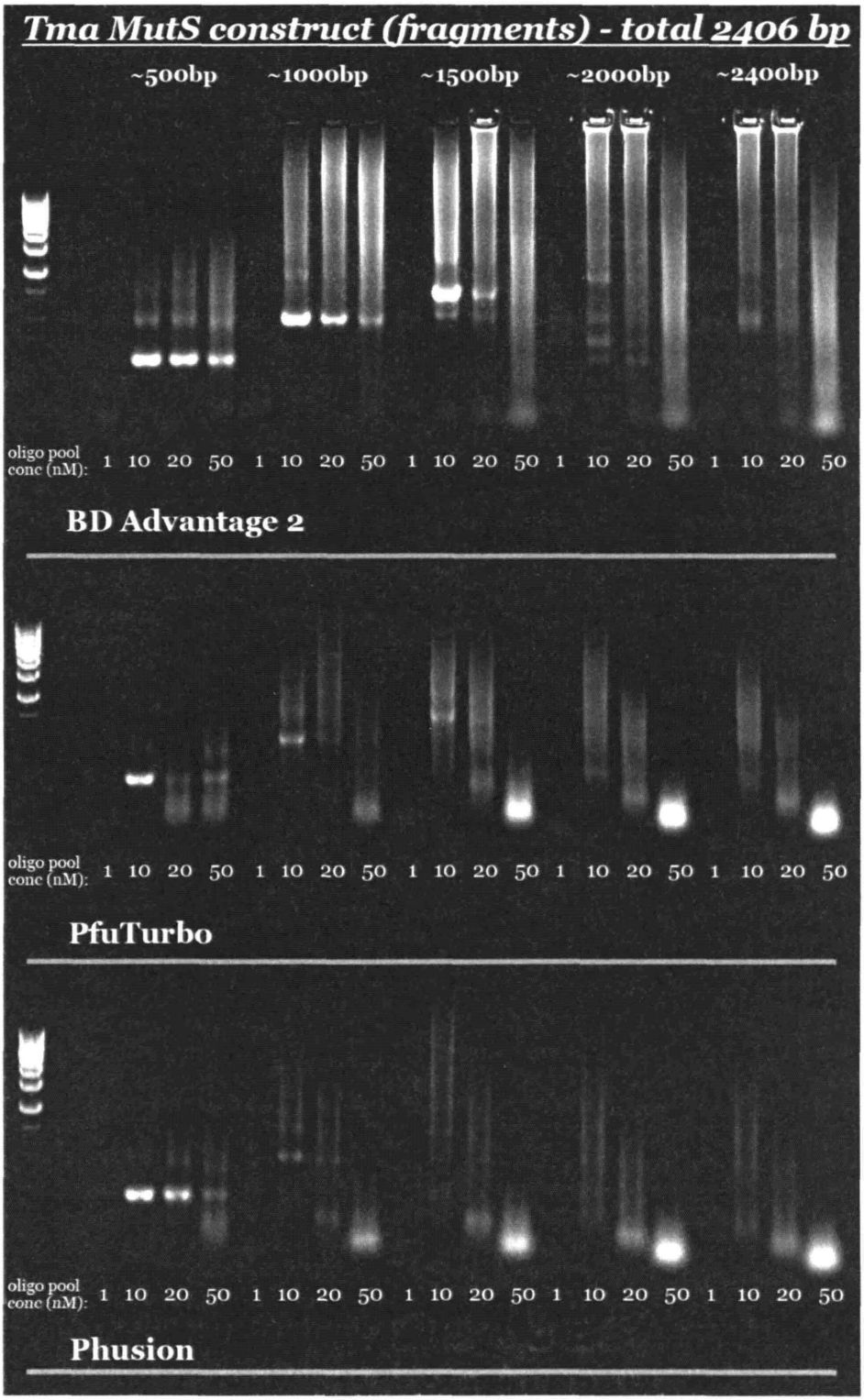

Figure 8 One-Step PCA of EGFP fragments with various polymerases at varied oligo pool concentrations. Target DNA constructs of length 545 bp (20 oligos), 1075 bp (38 oligos), 1621 bp (56 oligos), 2163 bp (74 oligos), and 2406 bp (90 oligos) from the Tma MutS gene were assembled using the One-Step PCA protocol. Robustness of assembly for varying concentrations of oligo pool (1-50 nM each) was assessed. 4 uL of each PCA product was run on the 1% agarose gel alongside 2 uL kb ladder (Stratagene). All images were enhanced for contrast with the same parameters.

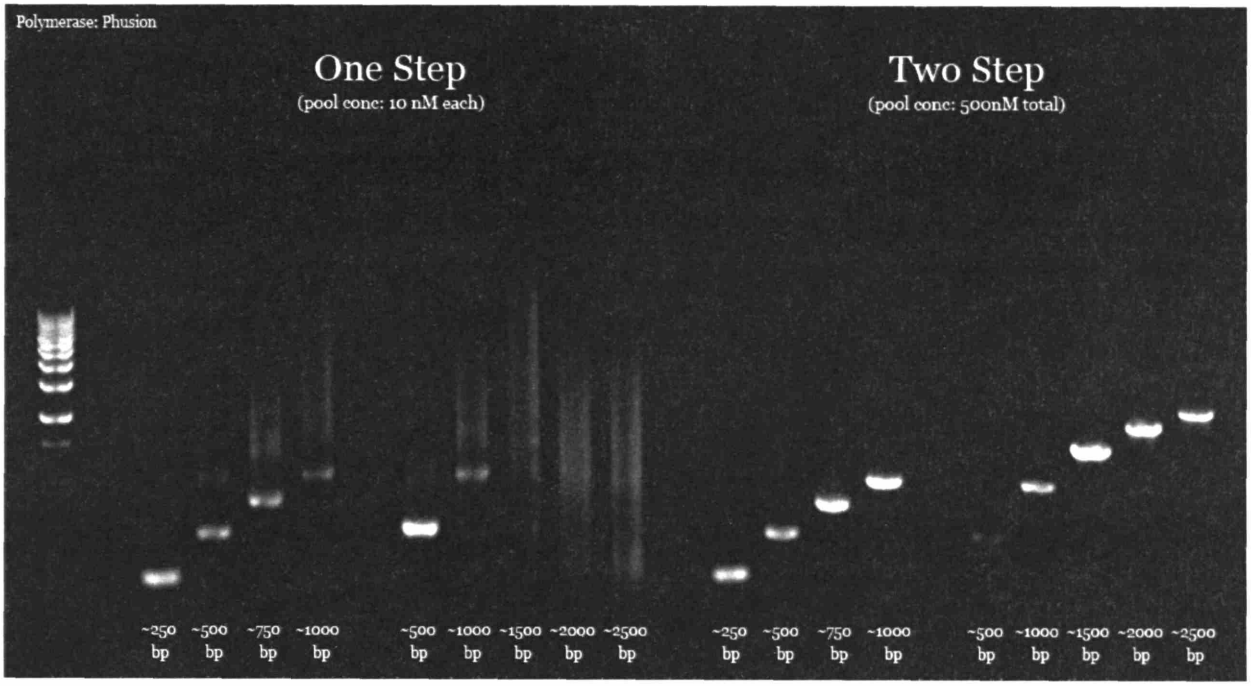

Figure 9 One-Step PCA vs. Two-Step PCA for constructs of various sizes with Phusion polymerase. Target DNA constructs of length 264 bp (12 oligos), 475 bp (22 oligos), 682 bp (32 oligos), and 993 bp (42 oligos) from the EGFP gene and of constructs length 545 bp (20 oligos), 1075 bp (38 oligos), 1621 bp (56 oligos), 2163 bp (74 oligos), and 2406 bp (90 oligos) from the Tma MutS gene were assembled using either One-Step PCA (with 10 nM each oligo pool concentration) or Two-Step PCA using Phusion polymerase (Finnzymes). Robustness of assembly was assessed. 4 uL of each PCA product was run on the 1% agarose gel alongside 2 uL kb ladder (Stratagene). All images were enhanced for contrast with the same parameters.

Parsing the sequence

We recommend the use of a software parsing tool such as DNAWorks (http://molbio.info.nih.gov/dnaworks/) for gene synthesis in order to optimize for consistent melting temperature, hairpin formation, no self-annealing, no primer-dimerization, codon frequency in host organisms, and more. In our experience, this software has been very easy to use and has given us reliable results. Oligonucleotides that are parsed with DNAWorks seem to build with greater robustness in One-Step PCA.

Error rate from the gene synthesis process: Vendor comparison

We found that the error rate of synthesized DNA depended greatly on the quality of the oligonucleotides ordered commercially. Vendors were seen to vary dramatically in the suitability of their oligonucleotides for gene synthesis.

The primary error in chemical oligonucleotide synthesis is a failure to add a new nucleotide monomer to the growing chain (synthesized 3' to 5') – this causes a deletion. Oftentimes, this error-product will be capped with an acetyl group, preventing further additions to the chain. This results in a shortened oligonucleotide, but there is no mutation actually incorporated into the DNA, so these truncated products do not add any errors to the final product in gene synthesis. This is evidenced by the fact that while average stepwise yields for oligo synthesis are roughly 99% (1% error), use of these oligos in gene synthesis results in an error rate of only about 0.2% - many of the dominant errors in oligo synthesis are masked by the PCA process. However, sometimes the acetylation step fails, and the growing nucleotide chain ends up with a deletion. The result of this is the incorporation of a deletion into the synthetic DNA in gene synthesis. Single base deletions are the most common type of error found in our sequencing data.

We compared synthetic EGFP built from oligonucleotides ordered from many different vendors via flow cytometry and sequencing analysis. (See Table 1 and Figure 10).

SUMMARY OF ERRORS

	Deletions			Insertions			Substitution						Other		Total errors	Bases Sequenced	Reciprocal error rate (per bp)
	Single deletion			Single insertion			Transition		Transversion								
Sample name	-G:C	-A:T	Multiple deletion	-G:C	-A:T	Multiple insertion	G:C to A:T	A:T to G:C	G:C to C:G	G:C to T:A	A:T to C:G	A:T to T:A	=	Listing of Other Errors			
Operon - PfuTurbo	1	1		1		2			1			1			7	1058?	1512
Operon (2nd set resent) - PfuTurbo	2	1		2			1	1	2	1		1	1	GAC->A--	12	11473	956
Operon (2nd set resent) - Phusion	2	2		2		2	2		3						13	10480	806
Sigma - PfuTurbo	10	6		3	2		2	1	1		3	2			30	21846	728
IDT - Phusion	7	1	1		1			1	1				2	TT->GA, TT->GA	14	9930	709
IDT - PfuTurbo, safestain	15	4	5	1	1	3	1		1				1	GTC -> TCA	33	21617	655
IDT - Pfx50	1	3	3			3	2		3	1	1		1	CTCGCCGGA -> CAGGGTGGT	18	10587	588
IDT - PfuTurbo	10	2	3	2					2	2					21	10923	520
IDT - PfuUltra II	8	7	1	2	2	2	2		1						25	10037	401
Operon (2nd set) PfuTurbo	1		1			2	6	9	5	3	9	7			43	9151	213
IDT - BD Advantage 2	12	2	3			1	53	6	3	11	1	11	7	TC->AT, GC->AG, CC->AG, TGCT->___A, TG->C_, GT->TG, AT->CA	110	9174	83
IDT - Taq (NEB)	22	1	1	1	1		55	19	8	9	3	8	2	GG->C_, CA->TG	130	10037	77

Table 1 Sequencing analysis - Polymerase and Vendor comparison

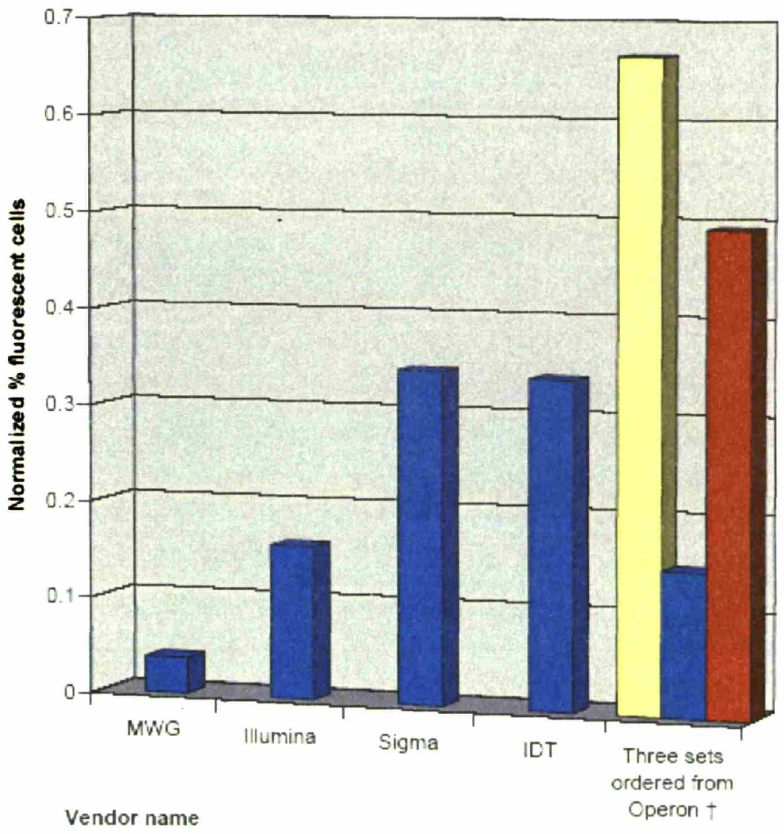

Figure 10 Flow cytometry data: Oligo vendor comparison
EGFP was synthesized using Two-Step PCA with oligos from five different vendors. The products were cloned with Clonase II (Invitrogen), transformed, cultured, and analyzed by flow cytometry as described earlier. All % fluorescent cells measurements were normalized to the % fluorescent cells measurement of cells expressing a perfect, sequenced copy of the pDONR221-EGFP plasmid to account for run-to-run variation. Oligos from MWG, Illumina, Sigma, IDT, and the first set of Operon oligos were parsed "naively" into 50mers. The second and third sets of Operon oligos were parsed using DNAWorks 3.0.

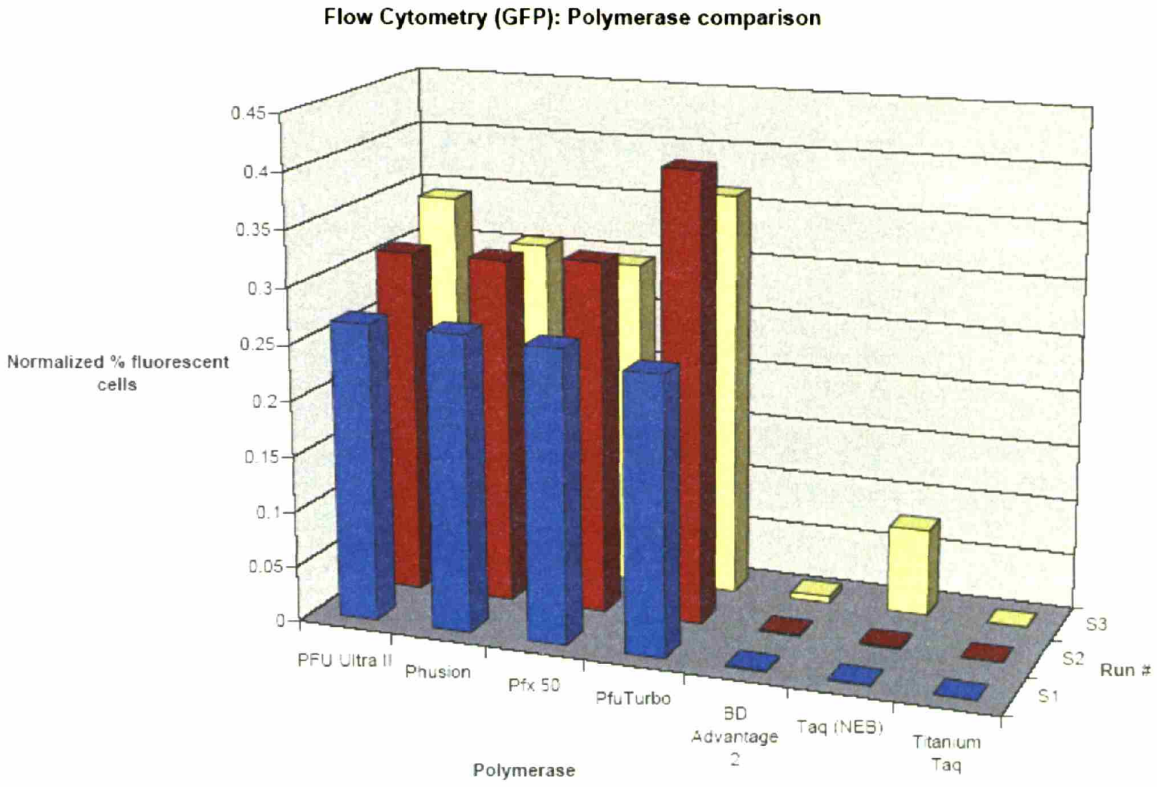

Figure 11 Flow cytometry data: Polymerase Comparison
EGFP was synthesized using Two-Step PCA with oligos from IDT and a panel of different polymerases: PfuUltra II (Stratagene), Phusion (Finnzymes), Pfx50 (Invitrogen), PfuTurbo (Stratagene), BD Advantage 2 (Clontech), Taq (NEB), and Titanium Taq (Clontech). The products were cloned with Clonase II (Invitrogen), transformed, cultured, and analyzed by flow cytometry as described earlier. All % fluorescent cells measurements were normalized to the % fluorescent cells measurement of cells expressing a perfect, sequenced copy of the pDONR221-EGFP plasmid to account for run-to-run variation.

While an exhaustive survey of batch-to-batch variation from each of these vendors was not within the scope of this work, our results from IDT are consistent with that which we have seen in the past – we have consistently seen low error rates in the range of 1 in 500bp to 1 in 700bp with excellent reproducibility from IDT. Oligos purchased from Sigma also seem promising for gene synthesis with an error rate comparable or slightly better than IDT, but we did not choose to assess batch-to-batch variability at this time and would recommend IDT over Sigma for most users for the time being. Operon seems to have excellent potential for producing oligonucleotides of high quality (the first batch we got from them gave us an error rate of 1 in 1500 with matching FACS data, prompting us to order another set from them), but they seem to have a problem with variability in quality. After speaking with Operon representatives about the

relatively low quality of the second set of oligos ordered from them, they agreed to send us another set on a machine they use primarily to produce oligos for gene synthesis applications. We refer to this third set as "Operon 2 New" or "Operon 2nd resent". (Note: the second and third sets of oligonucleotides from Operon were parsed using DNAWorks, all other oligos were parsed identically, into 50mers without any optimization). The third set of oligos yielded product with an intermediate error rate on the order of 1 in 1000.

Thus, for the average user, we recommend ordering oligos from IDT for most common gene synthesis applications.

Error rate from the gene synthesis process: Polymerase comparison

We also studied the effect of using various polymerases – some high-fidelity such as PfuTurbo (Stratagene) and some non-high-fidelity such as Taq – and found that most high fidelity (proofreading) PCR polymerases make negligible contributions to the overall error rate of gene synthesis. On the other hand, the use of Taq polymerase or Taq-based blends (often used for long templates in PCR) contributed substantially to the error rate of gene synthesis. We again used flow cytometry and sequencing as our main readouts for error rate.

Figure 11 shows the performance of various polymerases for gene synthesis in our GFP flow cytometry assay. It is important to note that this functional assay by flow cytometry is not a good readout for substitution errors and silent mutations because these types of errors may not abolish function. On the other hand, a functional assay is very good for detecting the single base deletions that are so prevalent in gene synthesis.

Within the limits of our assay, all non-Taq polymerases tested gave similar performance. We confirmed our results by sequencing analysis. Sequencing confirmed disastrous error rates for Taq (NEB) and BD Advantage 2 (Clontech), a Taq-containing blend, of 1 in <100bp. PfuTurbo Hotstart (Stratagene), Phusion (Finnzymes), and Pfx50 (Invitrogen) yielded similar error rates (~1 in 600 with IDT oligos), while Pfu Ultra II Fusion sequencing yielded a roughly 50% higher error rate (about 1 in 400 with IDT oligos).

Protein-mediated error correction for synthetic DNA by MutS gel-shift

Fragments of EGFP were synthesized and then thermally denatured and re-annealed to re-assort errors and create error heterodimers. MutS was observed to selectively bind error-containing DNA and cause a gel-mobility shift in polyacrylamide gel. The DNA was visualized by staining with SYBR Gold. As can be seen in Figure 12, the error-containing DNA can be seen as a shifted band of reduced mobility. The MutS-filtered DNA, the MutS-bound DNA, twice-MutS-filtered DNA, and appropriate controls were amplified and analyzed by flow cytometry, green/white colony count and sequencing.

FACS analysis (see Figure 13) showed substantial improvement in fluorescence in those samples that were error-filtered, both in terms of percent green cells as well as in mean intensity of green fluorescence.

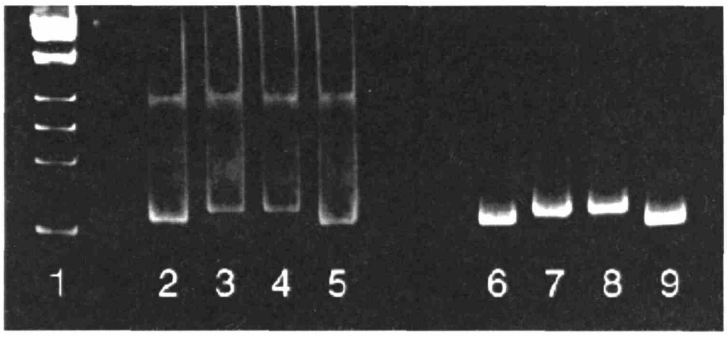

Figure 12 MutS pull-down filter.
Lane 1: kb ladder. Lanes 2,3,4,5: ~300mer pieces of GFP (993bp), treated with MutS. Lanes 6,7,8,9: Same as lanes 2,3,4,5, except without MutS treatment. (From Carr et al., 2004)

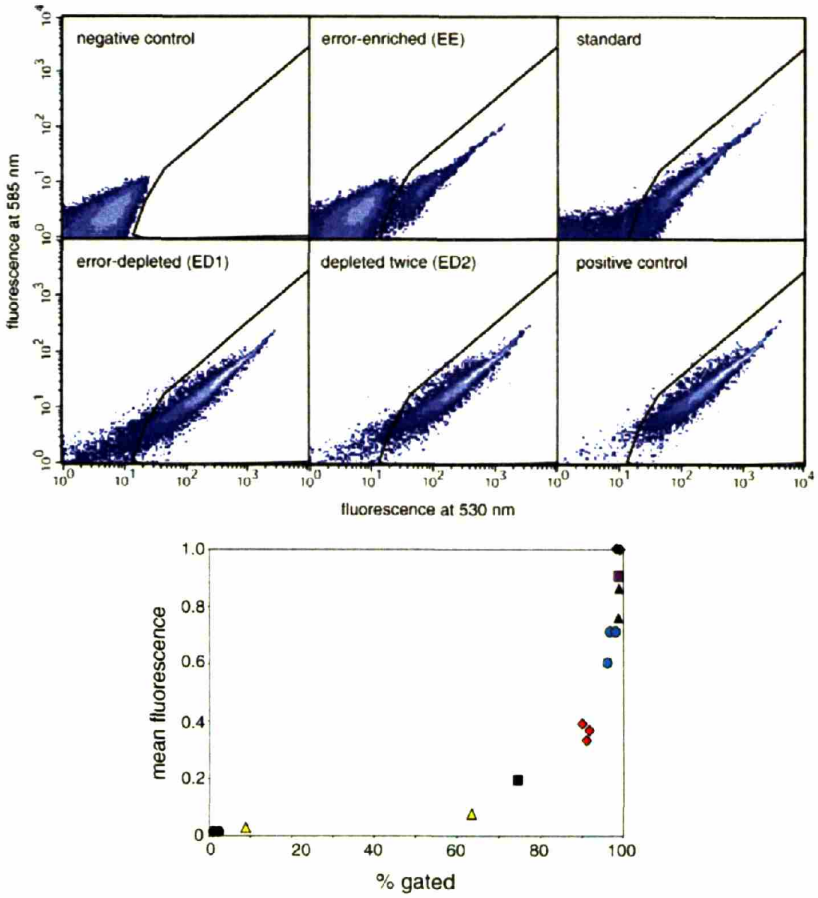

Figure 13 FACS data for GFP synthesized with MutS pull-down error filter (from Carr et al. 2004)

(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 nonspecific 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 14.

Sequencing data was used to quantitate errors and characterize the types and locations of errors surviving the error-filtration procedure. Error rate was decreased ~15-fold in two cycles of the filter protocol from a standard error rate of about 1 in 600 to about 1 in 10000.

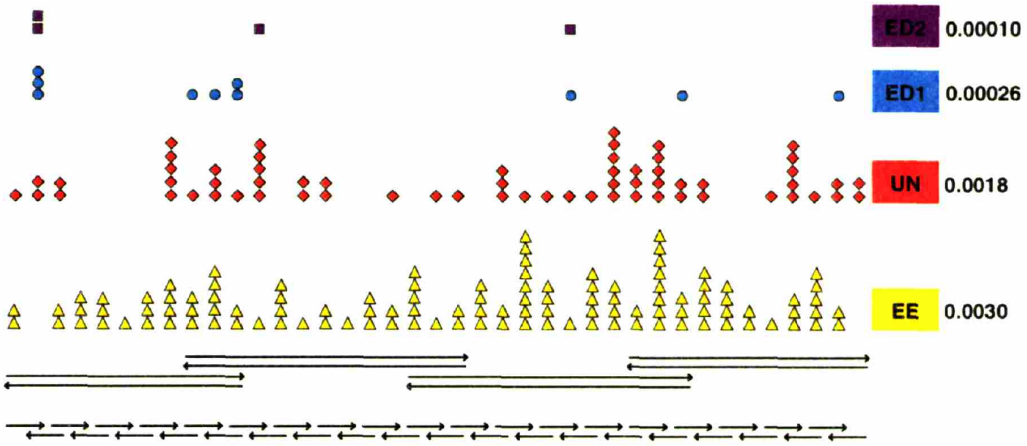

Figure 14 Locations of errors within the GFP DNA synthesis product - MutS pull-down error filter (from Carr et al. 2004). From bottom to top: the overlapping set of 38 oligonucleotides (thirty-six 50-mers and two 5'-terminal 59-mers) used to build the GFP gene and flanking sequences (arrowheads indicate the 3'-terminus); 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 are indicated.

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

Table 2 Summary of errors in GFP gene synthesis - MutS pull-down error filter (from Carr et al. 2004).

Discussion

Critical Variables and Considerations in Practical Gene Synthesis

The results of our work in developing a universally effective, robust, and simple gene synthesis protocol allow us to make some broad recommendations for the average gene synthesis user for most applications. We have shown that three of the most important considerations in gene synthesis are the choice of protocol, the commercial source of oligonucleotides, and the use of a high-fidelity polymerase.

First, Two-Step PCA is an easy-to-use and straightforward protocol that is effective and robust (it is forgiving of wide variations in oligo pool concentration, for example). On the other hand, while One-Step PCA has its advantages, such as reduced sample handling and shorter overall assembly time, in general situations, the shortcut is not worth it because optimizing and troubleshooting a failed procedure is usually significantly more time-consuming than running the protocol itself. Also, as one can see highlighted in yellow in the Gene Fabrication time pie (see Figure 15), the time spent doing gene assembly in the overall process of Gene Fabrication is so small that the extra time and labor spent performing the Two-Step PCA vs. the One-Step PCA is negligible.

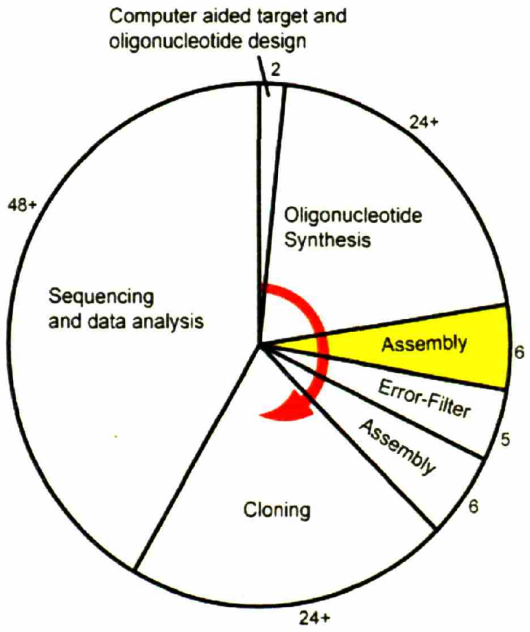

Figure 15 Gene Fabrication: Process Time Pie

Second, we have shown that making a good choice for the commercial source of oligonucleotides for gene synthesis is of great importance, as the quality of the oligo can be the dominating source of error in gene synthesis (with use of a high-fidelity polymerase, discussed next). Sigma and Operon are both promising, but the latter exhibits great variability and the former has not yet been thoroughly studied for batch-to-batch variation. For most applications (target length ~1000bp), we recommend the tried-and-true oligo vendor IDT for an overall error rate in the range of 1 in 500bp to 1 in 700bp. If the user can discuss one's needs with the company and get high quality oligos from Operon with less variability, that could be a useful option. Given batch-to-batch variability of Operon oligos at present though, for building longer constructs or in other applications requiring lower error rates, we would recommend building the target in overlapping ~1000bp pieces with the more reliable and reasonably high-quality IDT oligos and following up the procedure with an error correction protocol such as a MutS-mediated gel-based error-filter. We have demonstrated an error rate improvement to 1 in 10000bp starting from 1 in 600bp in two cycles of this procedure (see next section).

Finally, choosing a high-fidelity polymerase that is robust in gene assembly PCA is absolutely critical. Taq and BD Advantage 2 (a Taq blend) gave error rates that were unacceptable for gene synthesis applications (see Figure 11 and Table 1). Of the commercially available polymerase we used, we recommend the use of Phusion (Finnzymes). It performs quite well in both One-Step PCA and Two-Step PCA, has reduced cycle times (less than half the time), requires half as many units per reaction, and demonstrates low error rates for gene synthesis. It is also in the "new generation" of polymerases that fuse the traditional thermophilic proofreading polymerase to a domain for improved processivity.

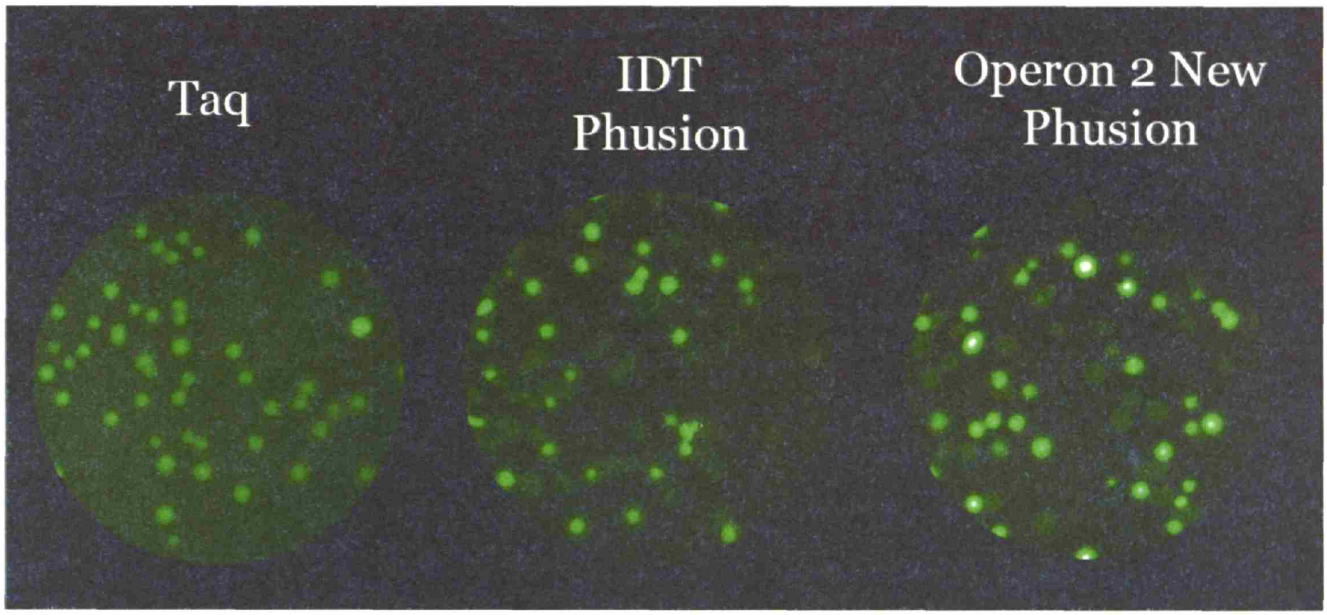

Figure 16 GFP green/white colony count comparing three data points demonstrating the importance of choosing optimal parameters for successful gene synthesis.
GFP green/white colony count was performed with cultures from cells transformed with gel-purified product from Two-Step PCA performed with the following oligos and polymerases (from left to right): IDT oligos / Taq (NEB) polymerase; IDT oligos / Phusion (Finnzymes) polymerase; Operon (3rd set) / Phusion (Finnzymes) polymerase. Plates were imaged on a Fluorchem8900 gel documentation system (Alpha Innotech) under UV illumination (365 nm) with a SYBR Green filter. Adobe Photoshop was used for image enhancement (gamma setting ~3, white/black threshold adjusted for optimal contrast, false color green). The same image enhancement settings and camera exposure settings were used for all samples.

As one of our primary assays, flow cytometry is a useful qualitative and semi-quantitative tool that allows us to sample large populations of cells at once. However, in our experience, we have often seen substantial sample-to-sample variation on the order several percent (% fluorescent cells). Normalizing against a pEGFP positive control is useful for data analysis and comparing data sets. Machine run-to-run variability with the same sample, however, is very low (less than 0.5%). We hope to continue to improve our knowledge of GFP flow cytometry as a useful tool for assessing error rates in gene synthesis. Some considerations include staining dead cells with propidium iodide or another marker to more effectively analyze only live cells (we currently rely on forward and side light-scattering distributions), filtering our media for particulate matter, and using higher quality culture media.

Protein-mediated error correction for synthetic DNA by MutS gel-shift

Error rates are 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 (Cello et al 2002). By contrast, a 2703 bp plasmid synthesis (Stemmer et al 1995) and a 5386 bp bacteriophage phiX174 synthesis (Smith et al, 2003) 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.

The method described here and reported in *Nucleic Acids Research* in 2004 demonstrates a 15-fold reduction in error rates to 1 in 10,000bp. Referring back to Figure 3, one can see that this improvement in error rate allows for the synthesis of targets on the order of 10kb. Or, more dramatically, it should be possible to synthesize a 5000bp target, error correct it, clone it, and sequence only three clones to have very high probability that it will be error-free. A case in point: we generated the Taq MutS gene using this error correction protocol, sending off only one clone for sequencing, which came back perfect. When we had previously tried to synthesize correct DNA products of this length, the process required two cycles: we generated smaller pieces of the overall construct, cloned and chose perfect constructs by sequencing, and then assembled the full-length target from the perfect smaller pieces and cloned and sequenced again. One can see that production of even a moderately large gene target without error correction can take substantial labor, time, and expense (especially for additional sequencing reactions necessary for the higher error rate).

Analysis of the sequencing data gave some interesting information for follow-up study. Based on what is known about MutS protein affinity for different types of DNA mismatches (Whitehouse et al, 1997; Brown et al, 2001), one may expect some errors to be removed by this protocol better (or worse) than others. However, reductions in all categories of errors were observed in our results (See 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 (Whitehouse et al,

1997). 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.

The distribution of error locations within the sequence also changes as a result of MutS treatment (see Figure 14). 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 14). 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 (Figure 14).

Possible directions for improving on this protocol in terms of improved error rate, decreased sample handling, amenability to automation, and more are discussed in the following section.

Follow-up material to MutS gel-shift error correction: Initial Results

To date, our results show the best error rate published so far in the literature. Figure 17 shows a side-by-side comparison of three methods of DNA error reduction published (ours and two published subsequently by other groups). The Belshaw group demonstrated a protocol for error correction called consensus shuffling in which synthetic DNA is fragmented with a variety of restriction endonucleases, error-filtered via binding to immobilized MutS, and re-assembled via PCR (Binkowski et al 2005). Two iterations of the process decreased errors 3.5- to 4.3-fold to final values of ~1 in 3500. The Hegemann group demonstrated a novel strategy for error

reduction in cloned synthetic DNA by applying enzymatic mismatch cleavage with the resolvases T4 endonuclease VII or E. coli endonuclease V at the site of errors and using an 3'-5' exonuclease (or the 3'-5' exonuclease activity of proofreading polymerases) to remove the error (Fuhrmann et al, 2005).

Further improvements in error rate will be helpful in enabling the use of gene synthesis technology in applications requiring large DNA constructs as described earlier.

At present, preliminary experiments have been completed to make improvements on the MutS gel-shift error correction described above and reported in *Nucleic Acids Research* in 2004. Some of the improvements we have been working on, many of which are noted in the 2004 report, include: use of PCR primers (containing cloning sequences) that lie outside of the coding sequence of the gene target sequence to prevent errors from final PCR amplification steps after error reduction protocols, optimizing the parameters of the MutS error-filtration protocol to work with longer pieces of DNA so as to mitigate end-effects (the footprint of MutS, determined by DNase I footprinting by Biswas and Hsieh (1997), is ~24-28bp), further characterizing the optimal stoichiometry for MutS binding to DNA, and the development of separation techniques other than polyacrylamide gel (time-intensive, not automatable) such as attaching MutS to solid supports such as resins or magnetic agarose beads and sedimenting by centrifugation or applying a magnet. We have also synthesized the genes for two hyperthermophilic variants of MutS from *Thermatoga maritima* and *Aquifex aeolicus* and successfully expressed the proteins in *E. coli*. We have been characterizing the properties of these proteins, especially in regards to their potential utility in error reduction protocols for gene synthesis.

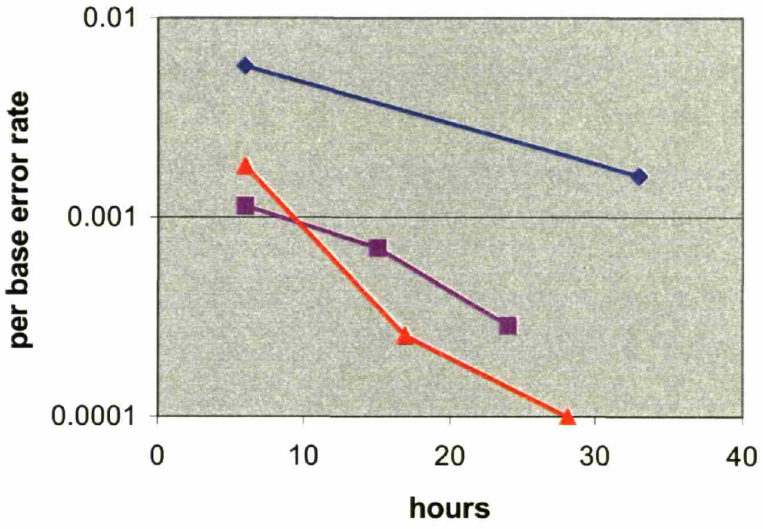

Figure 17 Comparison of three methods of DNA error reduction.
MutS error-filtration described in this document (Carr et al 2004) (red triangles), consensus shuffling by Binkowski et al 2005 (purple squares), and enzymatic mismatch cleavage by Fuhrmann et al 2005 (blue diamonds). Times are estimated.

Characterization of new MutS variants: Preliminary results/discussion

As mentioned above, we have constructed the genes for MutS genes derived from other species (from *Thermotoga maritima*—"Tma MutS," and *Aquifex aeolicus*—"Aae MutS"). We have assessed the thermostability of these proteins from hyperthermophilic organisms using a temperature scan with circular dichroism. The observed melting temperature of Tma MutS and Aae MutS are about 82°C and >95°C, respectively. This property of the two proteins makes them promising candidates for high-temperature reactions such as for error-prevention reagents in PCR.

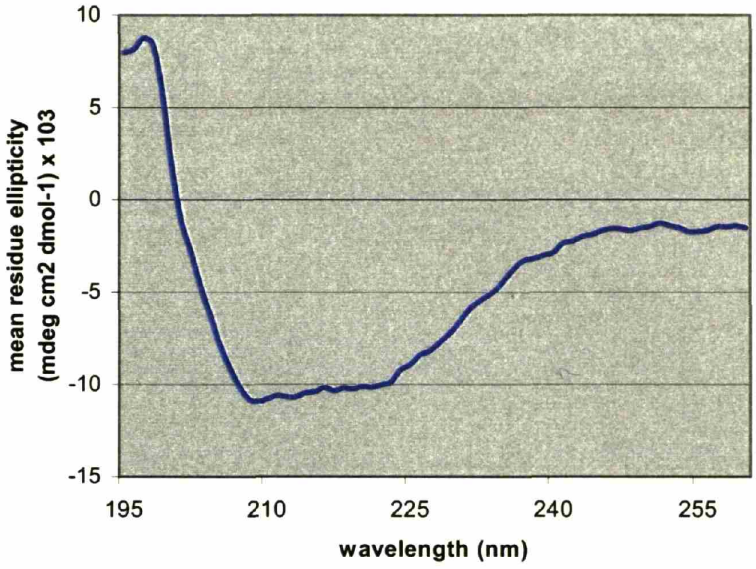

Figure 18 Circular dichroism spectrum of Tma MutS compares favorably to that of Takamatsu et al. 1996 for T. thermophilus MutS

We plan to use the new MutS proteins for a number of efforts in error correction. First off, we plan to compare them to Taq MutS for error-binding characteristics and general error removal post-gene synthesis. Both proteins have shown function in gel mobility shifts of mismatched DNA. Initial GFP flow cytometry results indicate that Aae MutS may improve error rate as much as 15-fold in one cycle. It is possible that the new variants of MutS have improved, or at least different, performance characteristics (affinity, specificity) for error-binding – as a case in point, MutS from *Thermus thermophilus*, HB8 was found to have broader specificity for mismatches than E. coli MutS (Whitehouse, 1997). Multiple species of MutS might complement each others' strengths and weaknesses for optimal error removal.

We are also in the process of further characterizing the error-binding properties of these MutS proteins with two different tools: surface plasmon resonance (SPR) and Fluorescence Correlation Spectroscopy (FCS) with the Olympus MF20 – an instrument that makes ingle molecule fluorescence measurements within the volume of a confocal field. In the latter approach, fluorescently labeled oligos can be observed in the presence of varying amounts of MutS protein to get a measure of binding affinity for different types of mismatches (AA, CC, GG, TT, AC, AG, CT, GT, single base deletions/insertions).

MutS attached to solid supports: Preliminary results/discussion

We now have some preliminary data on error-filtration using MutS attached to solid supports via affinity tags. We expect such methods to outperform the current approach using gel electrophoresis in terms of both speed and ease of implementation. One of the advantages of a quick protocol is that it can be iterated multiple times to yield better and better error rates (up to a certain threshold).

Our recombinant MutS proteins have N-terminal polyhistidine tags used for protein purification following expression. Our preliminary results suggest that MutS can simultaneously bind to both Ni-NTA (or Ni-NDA) conjugated beads or resins and DNA mismatches. We have noted better performance with DNA of shorter lengths, however, with 70mers binding better than 400mers binding better than 1000mers. Our immobilized Taq MutS gave no evidence of preferentially binding to 1000bp error-containing DNA vs. error-free DNA.

These results are similar to those of Geschwind and co-workers (Geschwind et al, 1996), who fused E. coli MutS to an N-terminal biotinylated peptide domain for mutation detection with streptavidin-coated beads. They observed ~4-fold discrimination of mismatches for 400 bp DNA, and saw no discrimination at 1400 bp. Similarly, Binkowski et al. later showed a modest ~2-fold improvement of a pool of synthetic GFP DNA using a MBP-MutS (Taq) fusion protein for error reduction in synthetic genes (Binkowski et al, 2005).

While non-specific binding of MutS to DNA is definitely a concern and is a likely a contributing factor in the decreased efficiency of differential binding by the immobilized MutS by our group and others, the fact that our MutS-mediated gel-based error-filtration protocol yields substantial error-rate improvement gives us reason to believe that an effective immobilized-MutS error-filtration protocol can be developed. Recently, using one of our new MutS variants (from *Aquifex aeolicus*), we have observed as much as 15-fold improvement in error rate in one cycle of the gel-based error-filter. In recent work, Erie and coworkers studied nonspecific MutS binding from an imaging / statistical approach by observing DNA binding on MutS by atomic force microscopy and found that a high fraction of non-specific binding occurs at DNA ends (Yang et al, 2005). This is encouraging for us because the number of DNA ends of a linear molecule (two) does not increase with the length of the construct. One way to block non-specific end-binding of MutS to DNA is to add bulky end-modifications such as streptavidin on end biotinylated DNA.

We hope to design an improved immobilized-MutS error-filtration protocol in a number of ways. First of all, we have noticed that the two new MutS variants we have recently synthesized and expressed appear to have improved binding characteristics for larger DNA while bound to Ni-NTA resin. This could have to do with the characteristics of the proteins themselves (we are in the process of characterizing the proteins with techniques such as SPR, circular dichroism, and Fluorescence Correlation Spectroscopy), or it could have to do with steric considerations: the new MutS proteins have linkers between themselves and their His-tags of 10 amino acids (pET-15b vector) whereas the Taq MutS protein has only a linker of 2 amino acids (Gly-Gly). DNA-binding occurs relatively close to the N-terminus, where the His-tags are located. As a first experiment, we are planning to subclone the Taq MutS protein into the pET-15b vector to give it the same linker as the other recombinant MutS proteins. By experimenting with this construct, we hope to determine the relative importance of linker length on the binding characteristics of immobilized MutS.

Other possibilities for immobilizing MutS on solid supports include utilization of any of a variety of affinity tags and domains (CBD, GST, strep-tag, S-tag) or chemical cross-linkers (via thiol groups on cysteine residues engineered into the protein sequence at appropriate locations via gene synthesis or mutagenesis techniques). These affinity tags/domains or crosslinking groups could be placed on the C-terminus or the N-terminus of the MutS protein (or in the case of a cysteine residue for crosslinking, anywhere on the protein).

We are also interested in the effect of using our different variants of MutS in developing this approach. In preliminary results, we have noticed that *A. aeolicus* MutS seems more effective than T. *maritima* MutS at separating mismatched 1000mer DNA while bound to Ni-NTA resins. Estimates from monitoring flow-through by UV absorbance (OD_{220nm}) and PAGE show approximately four-fold preferential binding of mismatched 1000mercompared to its error-free counterpart. We will follow through with these results by flow cytometry and sequencing.

As a sidenote, we have had some concerns about improperly annealed DNA (e.g. Holliday junction-like higher-order structures) being poorly bound by MutS in the immobilized-MutS error-filter protocol. In the gel-based assay, such higher-order structures have such poor mobility that they effectively remove themselves from the error-filtering by migrating only a small distance (or not at all) through the gel. In an immobilized-MutS error-filter system, such DNA – which would never have been treated for error-reduction – could presumably make it

through the filter and reduce the overall effectiveness of the protocol. One might consider filter membranes to exclude higher-order structures, the use of resolvases to resolve those structures, or other methods, if it turns out that this is an important problem to deal with in optimizing this protocol.

This also brings up the issue of the dependence of error reduction efficiency on DNA target sequence (apart from the usual discussions of differential binding specificity of MutS to different types of mismatches). As a consequence of the possibility that MutS does not bind well to higher-order structures of DNA, sequences containing repeats, palindromes, or other sequences conducive to DNA secondary structure-formation may be somewhat poor targets for error reduction. In these cases especially, it may be helpful to develop and utilize methods for excluding, reducing, or resolving higher-order DNA structures prior to error correction. A number of simple initial experiments investigating sequence dependence of error reduction efficiency are advisable.

Optimizing MutS stoichiometry and binding conditions: Preliminary results/discussion

In the MutS gel-based filter, we observe differing amounts DNA in the "unbound" error-depleted band in the gel depending on the relative amounts of MutS and DNA used in the procedure. Preliminary data via GFP flow cytometry suggests that using increasing amounts of MutS to get fainter and fainter bands improves the error rate of the surviving pool. This issue, and others concerning the optimal stoichiometry for MutS-DNA interactions in error-reduction protocols, deserves consideration in future work.

Conclusions / Recommendations

Gene Synthesis

Gene synthesis holds promise as a remarkably powerful tool to not only make current research efforts requiring DNA manipulation more productive and efficient, but also as a technology that will enable entire new fields of research that require long constructs, custom-designed DNA molecules, or large sets of designed constructs.

This document gives a set of recommendations for a robust, quick, and easy gene synthesis protocol for the typical user. Among the most important considerations were protocol choice, oligo vendor, and polymerase choice. We demonstrate a Two-Step PCA protocol that gives reasonably low error rates (assayed for 1000bp construct) and robust building for gene products up to at least 2.5kb in length (90 oligonucleotides).

There appears to be a ceiling for the best error-rate achievable (without error correction) around 1 in ~1500bp (possibly better, given the uncertainty of our error rate measurement with the number of bases sequenced). With the use of a high-fidelity polymerase, it seems that oligo synthesis errors dominate this overall error rate. There may be room for newly optimized chemistry development for oligo vendors, though much optimization has already been done. If so, we can expect to be able to push the overall error rate of the gene synthesis process even lower.

Future work should explore the optimization of gene synthesis protocols for use in specialized circumstances, such as in microfluidic devices coupled to DNA microarrays (a remarkably economical source of oligonucleotides that could help push down the price of gene synthesis several orders of magnitude). Because reduced sample-handling may be important, use of One-Step PCA or other protocols may be better in the case of these devices, even though we recommend Two-Step PCA for normal, in vitro benchtop gene synthesis.

Error Correction

It was noted above that at present, ~1 in 1500bp may be the best gene synthesis error rate we can achieve, given errors due to the oligo synthesis process. However, achieving better error rates is of paramount importance in opening up the use of gene synthesis for some of the more

ambitious objectives laid out in the Introduction section. This points to the necessity of developing improved post-synthesis error correction protocols. While our demonstrated MutS-mediated gel-based error rate reduction (~15-fold) from 1 in 600bp to 1 in 10000bp is significant, there is still much room for improvement. Future work will continue to push the limits of error rate reduction using strategies involving MutS as well as a number of other candidate proteins.

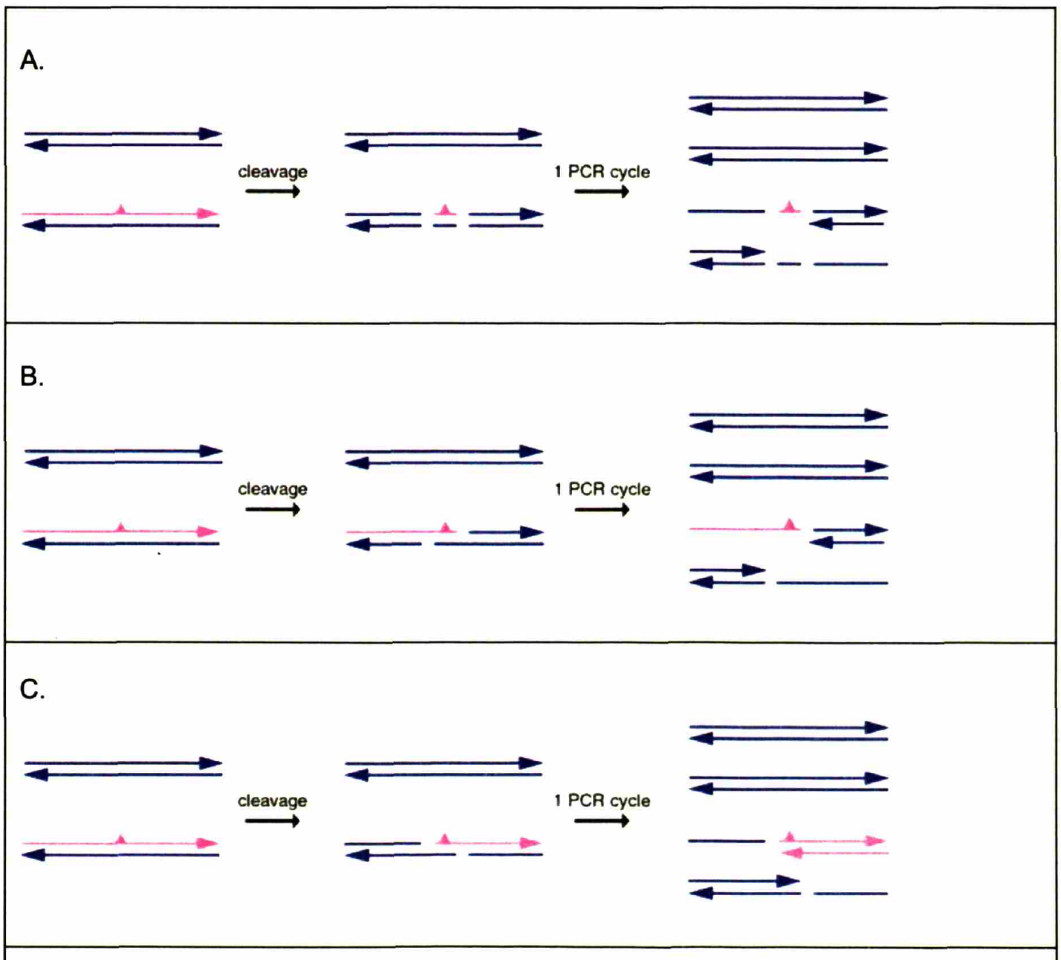

Figure 19 The effect of mismatch cleavage followed by amplification.
The fate of a given error (pink triangle, strands containing errors shown in pink) is followed through cleavage of (A) both strands on each side of the mismatch; (B) both strands 3' to the mismatch; and (C) both strands 5' to the mismatch. In each case, propagation of the error into full length strands is retarded relative to amplification of uncleaved material.

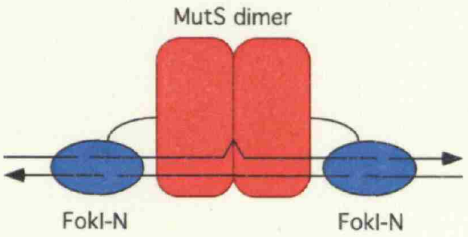

Figure 20 Schematic of a mismatch endonuclease designed for gene synthesis error correction.
It is composed of domains from MutS and FokI proteins. MutS binds as a dimer to a DNA mismatch, positioning the FokI nuclease domain to cleave both strands of DNA on each side

Some of our future directions are obvious extensions of our MutS gel-based error-filter protocol. These are described in some detail and some preliminary data is given in the Discussions section. Other future directions include the creation of a mismatch endonuclease by fusing MutS to the nuclease domain of FokI, the use of said fusion mismatch endonuclease for iterations of error reduction in PCR by cleavage followed by amplification (see Figure 19), the use of various thermophilic resolvases for enzymatic mismatch cleavage methods (as in Hegemann 2005), the use of MutS as a PCR-additive to block extension of error-containing duplexes, and more.

The ideal error correction protocol will improve error rate substantially, be quick, generalizable and easy to implement, require minimal sample-handling, and be an automatable and iterative process. In terms of error rate improvement, the sky is the limit – *in vivo* systems regularly maintain DNA fidelity with error rates as low as 1 in 10^8 bp to 1 in 10^{10} bp.

Acknowledgments

The author would like to thank Peter Carr, Joseph Jacobson, and Shuguang Zhang for teaching and advising him throughout the years of research represented in this document, as well as to Bevin Engelward for providing useful feedback and for agreeing to be his BE Thesis Reader on such short notice. Thanks also to members of the Zhang and Jacobson groups for helpful discussions, and to Glenn Paradis (MIT CCRR Flow Cytometry Core Facility) and Ky Lowenhaupt (Rich Lab, MIT) for assistance and training with flow cytometry and circular dichroism instrumentation.

References

Babon JJ, McKenzie M, Cotton RG. (2000) The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. Methods Mol Biol. 152:187-99.

Basu, S., Gerchman, Y., Collins, C.H., Arnold, F.H., and Weiss, R., (2005) A synthetic multicellular system for programmed pattern formation. Nature, 434(7037): p. 1130-4.

Biswas, I. and Hsieh, P. (1996) Identification and Characterization of a Thermostable MutS Homolog from Thermus aquaticus. J. Biol. Chem 271:5040–5048.

Biswas I, Hsieh P. (1997) Interaction of MutS protein with the major and minor grooves of a heteroduplex DNA. J Biol Chem. 1997 May 16;272(20):13355-64.

Binkowski, BF, Richmond, KE, Kaysen, K, Sussman MR, Belshaw PJ (2005) Correcting errors in synthetic DNA through consensus shuffling. Nucleic Acids Research, Vol. 33, No. 6, e55.

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

Campbell, Virginia W. and David A. Jackson. (1980) The Effect of Divalent Cations on the Mode of Action of DNase I. The Journal of Biological Chemistry. Vol. 255, No. 8, April 25, pp. 3726-3735.

Carlson, R., (2003) The pace and proliferation of biological technologies. Biosecur Bioterror, 1(3): p. 203-14.

Carr, P.A., Park J.S., Lee Y.J., Yu T., Zhang, S., Jacobson J.M. (2004) Protein-mediated error correction for *de novo* DNA synthesis. Nucleic Acids Research. Vol. 32 No. 20 e162

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(5583): p. 1016-8.

Chen, Junghuei and Seeman, Nadrian C. (1991) Synthesis from DNA of a molecule with the connectivity of a cube. Nature Apr 18;350(6319):631-3.

Cunningham, B.C. and Wells, J.A. (1989) High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science. 244(4908): p. 1081-5.

Ehrhardt, D. (2003) GFP technology for live cell imaging. Curr Opin Plant Biol, 6(6): p. 622-8.

Eisen, J.A. (1998) A phylogenomic study of the MutS family of proteins. Nucleic Acids Research 26:4291–4300

Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. (2000) Nature 403:335-8.

Fuhrmann M, Oertel W, Berthold P, Hegemann P (2005) Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. Nucleic Acids Res. 2005 Mar 30;33(6):e58.

Gao, X., Yo, P., Keith, A., Ragan, T.J., and Harris, T.K. (2003) Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. Nucleic Acids Res, 31(22): p. e143.

Geschwind, D.H., Rhee, R., and Nelson, S.F. (1996) A biotinylated MutS fusion protein and its use in a rapid mutation screening technique. Genet Anal, 13(4): p. 105-11.

Goho, A.M. (2003) Life Made to Order. Technology Review (April)

Harrison, R.G. (2000) Expression of soluble heterologous proteins via fusion with NusA. inNovations 11:4-7.

Hoover D.M., and Lubkowski J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res. 30:43-

Khorana, H.G., Buchi, H., Caruthers, M.H., Chang, S.H., Gupta, N.K., Kumar, A., Ohtsuka, E., Sgaramella, V., Weber, H. (1968) Progress in the total synthesis of the gene for ala-tRNA. Cold Spring Harb Symp Quant Biol, 33: p. 35-44.

Kim Y.G., Chandrasegaran S. Chimeric restriction endonuclease. Proc Natl Acad Sci U S A. 1994 Feb 1;91(3):883-7.

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 USA. 101(44): p. 15573-8.

Levskaya, A., Chevalier, A.A., Tabor, J.J., Simpson, Z.B., Lavery, L.A., Levy, M., Davidson, E.A., Scouras, A., Ellington, A.D., Marcotte, E.M., and Voigt, C.A., (2005). Synthetic biology: engineering Escherichia coli to see light. Nature. 438(7067): p. 441-2.

Modrich, P. Mechanisms and biological effects of mismatch repair. (1991) *Annu. Rev. Genet.* 25:229-53.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol, 51 Pt 1: p. 263-73.

Newman, J.R. and Keating, A.E. (2003) Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science. 300(5628): p. 2097-101.

Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT. (1998) Mutation detection using a novel plant endonuclease. Nucleic Acids Res, 26(20):4597-4602.

Park, S.H., Pistol, C., Ahn, S.J., Reif, J.H., Lebeck, A.R., Dwyer, C., and Labean, T.H., (2006) Finite-size, fully addressable DNA tile lattices formed by hierarchical assembly procedures. Angew Chem Int Ed Engl, 45(5): p. 735-9.

Porteus M.H. and Baltimore, D. (2003) Chimeric nucleases stimulate gene targeting in human cells. Science 300:763.

Sambrook, and Russell, (2001) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Shih, W.M., Quispe, J.D., and Joyce, G.F. (2004) A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. Nature. 427(6975): p. 618-21.

Sixma, T.K (2001) DNA mismatch repair: MutS structures bound to mismatches. Curr. Op. Struct. Biol. 11:47–52

Smith, H.O., Hutchison, C.A., 3rd, 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(26): p. 15440-5.

Smith J, Modrich P. (1997) Removal of polymerase-produced mutant sequences from PCR products. PNAS 94:6847-50.

Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene. 164:49-53.

Takamatsu, S., Kato, R., and Kuramitsu, S., Mismatch DNA recognition protein from an extremely thermophilic bacterium, Thermus thermophilus HB8. Nucleic Acids Res, 1996. **24**(4): p. 640-7.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. Nature. 432(7020): p. 1050-4.

Voigt, C.A. and Keasling, J.D. (2005) Programming cellular function. Nat Chem Biol, 1(6): p. 304-7.

Wah, D.A. et al. (1997) Structure of themultimodular endonuclease FokI bound to DNA Nature 388:97.

Wang, L., Brock, A., Herberich, B., and Schultz, P.G. (2001) Expanding the Genetic Code of Escherichia coli. Science 292:498-500.

Whitehouse, A. et al. (1997) Analysis of the Mismatch and Insertion/Deletion Binding Properties of Thermus thermophilus, HB8, MutS. Biochem Biophs Res. Commun. 233:834–837

Winograd, S., and Cowan, J.D. (1967) Reliable computation in the presence of noise. MIT Press, Cambridge, MA.

Withers-Martinez, C. et al. (1999) PCR-based gene synthesis as an efficient approach for expression of the A+T-rich malaria genome. Protein Engineering 12:1113-1120.

Wu G, Wolf JB, Ibrahim AF, Vadasz S, Gunasinghe M, Freeland SJ. (2006) Simplified gene synthesis: A one-step approach to PCR-based gene construction. J Biotechnol. 2006 Mar 1; [Epub ahead of print] .

Yang, Y., Sass, L.E., Du, C., Hsieh, P., and Erie, D.A. (2005) Determination of protein-DNA binding constants and specificities from statistical analyses of single molecules: MutS-DNA interactions. Nucleic Acids Res, 33(13): p. 4322-34.

Youil, R., Kemper,R. and Cotton, R.H.G (1996) Detection of 81 of 81 Known Mouse b-Globin Promoter Mutations with T4 Endonuclease VII—The EMC Method. Genomics 32:431–435. Genomics 32:431-435.

Appendix

PCR recipe / thermocycling charts

One-Step PCA recipe

For PfuTurbo, BD Adv 2

uL	Component
0.5	dNTP (10 mM each, stock)
0.4	pol (PfuTurbo or BDAdv2)
2	10X Buffer
2	Pool
2.4	Primers (2.5 uM each)
12.7	ddH2O

For Phusion

uL	Component
0.5	dNTP (10 mM each, stock)
0.2	pol (PfuTurbo or BDAdv2)
4	5X Buffer
2	Pool
2.4	Primers (2.5 uM each)
10.9	ddH2O

Thermocycling parameters for EGFP constructs and ~500bp and ~1000bp Tma MutS construct build

PfuTurbo

95°C / 2 minutes
    Repeat 30X:
    95°C / 30 seconds
    55°C / 30 seconds
    72°C / 1 minute
72°C / 10 minutes
4°C / Hold

BD Advantage 2

95°C / 1 minute
    Repeat 30X:
    95°C / 30 seconds
    55°C / 30 seconds
    68°C / 1 minute
68°C / 1 minutes
4°C / Hold

Phusion

98°C / 30 seconds
    Repeat 30X:

98°C / 10 seconds
55°C / 30 seconds
72°C / 15 seconds
72°C / 5 minutes
4°C/Hold

Thermocycling parameters for Tma MutS construct build (~1500bp, ~2000bp, 2406bp)

PfuTurbo
95°C / 2 minutes
Repeat 30X:
95°C / 30 seconds
55°C / 30 seconds
72°C / 3 minute
72°C / 10 minutes
4°C / Hold

BD Advantage 2
95°C / 1 minute
Repeat 30X:
95°C / 30 seconds
55°C / 30 seconds
68°C / 3 minute
68°C / 3 minutes
4°C / Hold

Phusion
98°C / 30 seconds
Repeat 30X:
98°C / 10 seconds
55°C / 30 seconds
72°C / 40 seconds
72°C / 5 minutes
4°C/Hold

Two-Step PCA recipes and thermocycling (PCR 1 = Assembly PCR, PCR 2 = Amplification PCR)

GFP PCR 1 **20 uL total**

	REACTION MIXTURE									PCR PROGRAM (30 cycles)						
NAME	OLIGOS	uL	dNTP	uL (10 mM each)	POLYMERASE		BUFFER	uL	WATER	DENATURE		ANNEAL		EXTEND		Initial denaturation and final extension
40mer	MWG40	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
50mer	MWG50	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
60mer	MWG60	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Sigma oligos	Sigma	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon Oligos	Operon	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Illumina Oligos (invitrogen)	Illumina	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
normal	IDT50	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
old dNTP	IDT50	2.0	old	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
freeze-thaw dNTP	IDT50	2.0	freeze-thaw	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
KOD HiFi	IDT50	2.0	normal	0.4	KOD	0.4	KOD #1	2.0 + 2.0 MgCl2	13.2	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Pfx 50	IDT50	2.0	normal	0.4	Pfx 50	0.4	Pfx 50	2.0	15.2	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Taq (NEB)	IDT50	2.0	normal	0.4	Taq	0.4	Thermo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/2min
Phusion	IDT50	2.0	normal	0.4	Phusion	0.2	Phusion HF	4.0	13.4	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Titanium Taq	IDT50	2.0	normal	0.4	Tit Taq	0.4	Tit Taq	2.0	15.2	95°	0:30	55°	0:30	68°	1:05	95/1min, 68/2min
BD Advantage 2	IDT50	2.0	normal	0.4	BD Adv	0.4	PCR buffer	2.0	15.2	95°	0:30	55°	0:30	68°	1:05	95/1min, 68/2min
PFU Ultra	IDT50	2.0	normal	0.4	PFU Ultra	0.4	PFU Ultra	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon 2nd, PfuTurbo	Operon 2nd	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon 2nd, Pfx50	Operon 2nd	2.0	normal	0.4	Pfx 50	0.4	Pfx 50	2.0	15.2	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Operon 2nd, Phusion	Operon 2nd	2.0	normal	0.4	Phusion	0.2	Phusion HF	4.0	13.4	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Operon 2nd, PfuUltraII	Operon 2nd	2.0	normal	0.4	PFU Ultra	0.4	PFU Ultra	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
dNTP abuse (4x)	IDT50	2.0	normal 4x	1.6	PFU turbo	0.4	PFU turbo	2.0	14.0	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 50	IDT50	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 55	IDT50	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 60	IDT50	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon2New, PfuTurbo	Operon 2nd	2.0	normal	0.4	PFU turbo	0.4	PFU turbo	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon2New, Pfx50	Operon 2nd	2.0	normal	0.4	Pfx 50	0.4	Pfx 50	2.0	15.2	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Operon2New, Phusion	Operon 2nd	2.0	normal	0.4	Phusion	0.2	Phusion HF	4.0	13.4	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Operon2New, PfuUltraII	Operon 2nd	2.0	normal	0.4	PFU Ultra	0.4	PFU Ultra	2.0	15.2	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min

GFP PCR 2 50 uL total

	REACTION MIXTURE												PCR PROGRAM (30 cycles)						initial denaturation and final extension
NAME	PCR 1	uL	dNTP	uL	POLYMERASE		BUFFER	uL	FWD PRIMER		RVS PRIMER		WATER	DENATURE		ANNEAL		EXTEND	
40mer	MWG40	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	MWG40	2.5	MWG40	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
50mer	MWG50	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	MWG50	2.5	MWG50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
60mer	MWG60	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	MWG60	2.5	MWG60	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Sigma oligos	Sigma	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	Sigma	2.5	Sigma	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon Oligos	Operon	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	Operon	2.5	Operon	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Illumina Oligos (Invitrogen)	Illumina	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	Illumina	2.5	Illumina	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
normal	IDT50	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
old dNTP	IDT50	2.5	old	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
freeze-thaw dNTP	IDT50	2.5	freeze-thaw	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
KOD HF!	IDT50	2.5	normal	1.0	KOD	1.0	KOD #1	5.0 + 5.0 MgCl2	IDT50	2.5	IDT50	2.5	30.5	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Pfx 50	IDT50	2.5	normal	1.0	Pfx 50	1.0	Pfx 50	5.0	IDT50	2.5	IDT50	2.5	35.5	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Taq (Strategene)	IDT50	2.5	normal	1.0	Taq	1.0	Thermo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/2min
Phusion	IDT50	2.5	normal	1.0	Phusion	0.5	Phusion HF	10.0	IDT50	2.5	IDT50	2.5	31.0	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Titanium Taq	IDT50	2.5	normal	1.0	Tit Taq	1.0	Tit Taq	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	68°	1:05	95/1min, 68/2min
BD Advantage 2	IDT50	2.5	normal	1.0	BD Adv	1.0	PCR buffer	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	68°	1:05	95/1min, 68/2min
PFU Ultra	IDT50	2.5	normal	1.0	PFU Ultra	1.0	PFU Ultra	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon 2nd. PfuTurbo	Operon 2nd	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon 2nd. Pfx50	Operon 2nd	2.5	normal	1.0	Pfx 50	1.0	Pfx 50	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Operon 2nd. Phusion	Operon 2nd	2.5	normal	1.0	Phusion	0.5	Phusion HF	10.0	Operon 2nd	2.5	Operon 2nd	2.5	31.0	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Operon 2nd. PfuUltraII	Operon 2nd	2.5	normal	1.0	PFU Ultra	1.0	PFU Ultra	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
dNTP abuse (4X)	IDT50	2.5	normal 4x	4	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	32.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 45	IDT50	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 50	IDT50	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Anneal temp 55	IDT50	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	IDT50	2.5	IDT50	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon2New PfuTurbo	Operon2Ne	2.5	normal	1.0	PFU turbo	1.0	PFU turbo	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min
Operon2New Pfx50	Operon2Ne	2.5	normal	1.0	Pfx 50	1.0	Pfx 50	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	94°	0:15	55°	0:30	68°	1:05	94/2min, 68/2min
Operon2New Phusion	Operon2Ne	2.5	normal	1.0	Phusion	0.5	Phusion HF	10.0	Operon 2nd	2.5	Operon 2nd	2.5	31.0	98°	0:10	55°	0:30	72°	0:35	98/30sec, 72/5min
Operon2New PfuUltraII	Operon2Ne	2.5	normal	1.0	PFU Ultra	1.0	PFU Ultra	5.0	Operon 2nd	2.5	Operon 2nd	2.5	35.5	95°	0:30	55°	0:30	72°	1:05	95/2min, 72/10min

Sequences

Taq MutS protein sequence

Taq MutS Construct DNA sequence (with flanking sequences included)

```
   1    CATGCCATGGGCCATCACCATCATCACCACGGGGGTATGGAAGGCATGTTAAAAGGGGAA   60
  61    GGACCAGGCCCTCTGCCCCCCCTGCTTCAGCAATACGTGGAATTGCGGGATCAATACCCG  120
 121    GATTATTTGCTCCTCTTCCAGGTCGGTGATTTCTACGAATGCTTTGGGGAAGATGCAGAG  180
 181    CGGCTGGCTCGGGCATTAGGACTCGTATTAACTCACAAAACCTCGAAAGACTTCACAACT  240
 241    CCGATGGCGGGCATACCTTTGCGTGCGTTCGAGGCGTATGCTGAACGTTTATTGAAAATG  300
 301    GGATTTCGGCTCGCTGTAGCCGATCAAGTGGAACCCGCAGAGGAAGCAGAAGGTCTGGTT  360
 361    CGCCGGGAAGTTACCCAGCTCCTTACCCCGGGAACGTTGTTACAGGAAAGCCTGTTGCCC  420
 421    CGCGAGGCTAACTATTTAGCAGCCATAGCCACAGGTGATGGGTGGGGATTAGCCTTTTTG  480
 481    GACGTTAGCACAGGTGAATTCAAGGGAACTGTGCTTAAATCTAAGTCGGCTCTGTATGAC  540
 541    GAGCTCTTCCGCCACCGTCCTGCGGAGGTCCTTCTCGCGCCTGAGTTGTTGGAGAACGGG  600
 601    GCTTTCTTGGACGAGTTTCGGAAACGGTTCCCAGTCATGTTGAGTGAAGCACCTTTTGAA  660
 661    CCAGAAGGTGAGGGTCCATTGGCTCTTCGTCGTGCGCGTGGCGCTTTACTGGCGTACGCC  720
 721    CAACGGACTCAAGGAGGGGCGTTAAGCCTTCAACCTTTTCGGTTCTACGATCCAGGCGCT  780
 781    TTTATGCGGTTACCGGAGGCCACTCTTCGTGCGTTAGAGGTGTTCGAACCCCTCCGCGGC  840
 841    CAGGATACGTTGTTCTCAGTTCTGGACGAAACACGTACGGCTCCTGGCCGTCGTCTCCTT  900
 901    CAAAGCTGGCTGCGCCATCCGCTCCTGGACCGCGGTCCCCTGGAGGCCCGTCTGGATCGC  960
 961    GTCGAGGGATTCGTACGCGAAGGCGCTTTGCGTGAAGGAGTGCGGCGTTTATTATACCGC 1020
1021    TTAGCAGACCTCGAACGGCTCGCGACACGGCTGGAGCTCGGACGGGCTTCCCCGAAGGAC 1080
1081    CTGGGAGCACTGCGTCGCTCGTTACAAATTTTACCGGAACTTCGGGCCCTCTTAGGGGAG 1140
1141    GAAGTTGGGTTGCCAGATCTCTCACCTTTGAAGGAGGAATTGGAAGCTGCTCTGGTTGAG 1200
1201    GATCCTCCCCTCAAGGTTTCGGAAGGAGGCCTTATACGTGAGGGCTACGATCCTGATCTT 1260
1261    GATGCCCTGCGCGCGGCTCACCGTGAGGGTGTAGCGTACTTCCTGGAACTCGAAGAACGC 1320
1321    GAGCGGGAGCGGACTGGCATACCCACTTTGAAGGTCGGTTACAACGCTGTATTTGGTTAC 1380
1381    TACCTCGAAGTTACTCGGCCATACTATGAGCGTGTCCCGAAAGAATACCGCCCTGTACAA 1440
1441    ACACTTAAGGATCGCCAGCGCTATACGCTCCCAGAAATGAAGGAGAAAGAGCGGGAAGTT 1500
```

```
1501      TATCGGCTGGAAGCTCTCATCCGCCGTCGTGAGGAGGAGGTGTTCTTGGAGGTTCGTGAG     1560
1561      CGTGCTAAGCGGCAGGCCGAAGCTCTCCGTGAGGCGGCTCGGATACTCGCAGAACTTGAC     1620
1621      GTTTATGCGGCCCTGGCGGAGGTTGCGGTACGGTATGGATACGTACGCCCACGTTTCGGT     1680
1681      GATCGTCTTCAGATCCGTGCAGGGCGGCACCCAGTAGTCGAGCGTCGTACTGAATTCGTC     1740
1741      CCCAATGACCTGGAGATGGCGCATGAGCTGGTCCTTATTACAGGCCCGAACATGGCAGGT     1800
1801      AAGTCCACTTTCTTACGTCAGACAGCGTTGATCGCGCTTTTGGCGCAAGTAGGTTCTTTC     1860
1861      GTGCCTGCTGAGGAAGCGCACCTGCCGCTCTTCGATGGCATTTATACTCGTATAGGGGCC     1920
1921      TCGGATGATTTAGCTGGCGGAAAGTCAACATTTATGGTCGAGATGGAGGAAGTCGCCTTG     1980
1981      ATCCTGAAAGAAGCTACCGAGAACAGTCTCGTACTCCTTGATGAAGTCGGGCGCGGGACG     2040
2041      AGTTCACTCGACGGGGTAGCTATAGCCACTGCTGTTGCTGAGGCCCTTCACGAGCGCCGC     2100
2101      GCGTATACCTTGTTTGCCACACACTACTTCGAACTTACTGCGCTCGGATTACCGCGGTTG     2160
2161      AAAAATTTACACGTCGCCGCTCGCGAGGAAGCTGGGGGCCTGGTCTTTTATCACCAAGTT     2220
2221      CTGCCTGGACCAGCCTCTAAATCCTACGGAGTAGAAGTTGCTGCGATGGCCGGATTGCCC     2280
2281      AAGGAAGTCGTAGCACGTGCCCGTGCCCTGCTCCAGGCGATGGCGGCGCGGCGGGAAGGG     2340
2341      GCACTCGATGCAGTGTTGGAACGGTTGCTCGCGTTAGATCCCGACCGGCTGACTCCGCTG     2400
2401      GAAGCACTCCGGCTTCTGCAGGAATTAAAAGCCTTGGCGCTCGGGGCCCCCCTCGACACG     2460
2461      ATGAAGTGACTCGAGCGGAC                                             2480
```

Tma MutS protein sequence

```
   1 MKVTPLMEQYLRIKEQYKDSILLFRLGDFYEAFFEDAKIVSKVLNIVLTRRQDAPMAGIP
  61 YHALNTYLKKLVEAGYKVAICDQMEEPSKSKKLIRREVTRVVTPGSIVEDEFLSETNNYM
 121 AVVSEEKGRYCTVFCDVSTGEVLVHESSDEQETLDLLKNYSISQIICPEHLKSSLKERFP
 181 GVYTETISEWYFSDLEEVEKAYNLKDIHHFELSPLALKALAALIKYVKYTMIAEDLNLKP
 241 PLLISQRDYMILDSATVENLSLIPGDRGKNLFDVLNNTETPMGARLLKKWILHPLVDRKQ
 301 IEERLKAVERLVNDRVSLEEMRNLLSNVRDVERIVSRVEYNRSVPRDLVALRETLEIIPK
 361 LNEVLSTFGVFKKLAFPEGLVDLLRKAIEDDPVGSPGEGKVIKRGFSSELDEYRDLLEHA
 421 EERLKEFEEKERERTGIQKLRVGYNQVFGYYIEVTKANLDKIPDDYERKQTLVNSERFIT
 481 PELKEFETKIMAAKERIEELEKELFKSVCEEVKKHKEVLLEISEDLAKIDALSTLAYDAI
 541 MYNYTKPVFSEDRLEIKGGRHPVVERFTQNFVENDIYMDNEKRFVVITGPNMSGKSTFIR
 601 QVGLISLMAQIGSFVPAQKAILPVFDRIFTRMGARDDLAGGRSTFLVEMNEMALILLKST
 661 NKSLVLLDEVGRGTSTQDGVSIAWAISEELIKRGCKVLFATHFTELTELEKHFPQVQNKT
 721 ILVKEEGKNVIFTHKVVDGVADRSYGIEVAKIAGIPDRVINRAYEILERNFKNNTKKNGK
 781 SNRFSQQIPLFPVX
```

Tma MutS Construct DNA sequence (with flanking sequences included)

```
   1 CAGGGAATTCCATATGAAGGTAACCCCCCTTATGGAACAGTACCTGCGGATAAAAGAGCA
  61 ATACAAGGATTCTATTCTTCTCTTTCGCCTCGGAGACTTTTATGAAGCCTTCTTCGAGGA
 121 TGCGAAGATCGTCAGCAAAGTCCTTAACATTGTCTTAACTCGCCGTCAGGATGCTCCAAT
 181 GGCTGGGATACCTTACCATGCACTTAATACCTACCTGAAAAAACTCGTGGAAGCGGGATA
 241 CAAAGTGGCAATATGCGATCAGATGGAAGAGCCGAGCAAGTCCAAGAAGTTAATCCGTCG
 301 TGAAGTAACCCGCGTTGTGACGCCTGGCTCCATAGTAGAGGACGAGTTCTTATCGGAAAC
 361 TAACAATTACATGGCGGTGGTTTCGGAAGAAAAGGGTCGCTACTGCACAGTGTTTTGCGA
 421 CGTATCTACTGGTGAGGTTTTAGTTCATGAGTCTTCGGATGAGCAGGAGACGTTGGATCT
 481 TCTCAAGAACTACAGTATTTCTCAGATTATCTGCCCGGAACACCTGAAGTCTAGTCTCAA
 541 AGAGCGTTTCCCCGGAGTCTATACAGAGACAATCAGCGAATGGTACTTTAGTGATCTTGA
 601 GGAAGTAGAGAAGGCGTACAATCTTAAAGACATCCATCACTTCGAATTAAGCCCCTTAGC
 661 CCTTAAAGCCCTTGCCGCATTGATTAAGTACGTCAAATATACTATGATCGCCGAGGATCT
 721 CAACTTAAAACCGCCGTTGTTAATTAGCCAGCGTGACTATATGATATTGGACTCTGCAAC
 781 CGTGGAGAACTTGAGTCTGATTCCAGGTGACCGGGGTAAAAATCTGTTTGATGTGCTTAA
 841 TAACACTGAAACACCTATGGGGGCTCGTTTGTTGAAGAAATGGATATTACATCCGCTGGT
 901 AGACCGTAAACAAATAGAGGAGCGCCTCAAAGCAGTCGAACGCCTTGTGAATGATCGTGT
 961 TTCGTTGGAGGAAATGCGTAATCTGCTCAGTAACGTTCGGGATGTTGAACGCATAGTAAG
1021 TCGTGTCGAGTACAACCGCAGCGTTCCACGGGATTTGGTCGCCCTTCGGGAAACCTTGGA
1081 AATCATTCCCAAGTTGAATGAGGTGCTTTCTACTTTTGGTGTATTCAAGAAACTCGCATT
1141 CCCGGAAGGTTTAGTAGATTTACTCCGGAAAGCAATTGAAGACGACCCCGTAGGGTCCCC
```

```
1201 TGGGGAGGGAAAGGTCATCAAGCGCGGATTTTCCAGTGAGCTCGATGAATACCGCGATCT
1261 CTTAGAACATGCCGAAGAACGGCTTAAGGAGTTTGAGGAGAAGGAACGCGAACGTACGGG
1321 CATACAGAAGCTGCGGGTGGGTTACAACCAGGTCTTTGGATACTACATTGAGGTGACAAA
1381 GGCCAACTTAGATAAGATACCGGACGACTACGAACGGAAGCAGACACTCGTGAACTCCGA
1441 GCGCTTCATCACGCCGGAGCTGAAAGAGTTCGAGACAAAAATCATGGCTGCCAAAGAACG
1501 TATCGAGGAGTTGGAGAAAGAACTGTTCAAATCAGTTTGTGAGGAGGTAAAGAAACACAA
1561 GGAGGTGTTGCTCGAAATCTCAGAGGACCTGGCAAAGATAGACGCCCTGTCAACACTTGC
1621 ATACGATGCAATCATGTACAACTACACTAAGCCGGTGTTTTCGGAGGATCGGCTGGAAAT
1681 TAAGGGTGGCCGGCATCCAGTCGTGGAGCGGTTTACCCAAAATTTCGTAGAAAACGACAT
1741 CTACATGGACAACGAGAAACGGTTCGTGGTTATTACTGGTCCAAACATGTCGGGTAAGTC
1801 TACGTTTATACGTCAGGTTGGGCTCATTAGCTTAATGGCACAAATTGGGTCTTTTGTACC
1861 AGCTCAGAAAGCGATTCTCCCCGTATTCGATCGGATCTTTACGCGCATGGGTGCCCGTGA
1921 CGATTTAGCTGGAGGTCGGTCAACTTTCCTCGTAGAGATGAACGAAATGGCCCTCATTTT
1981 GTTAAAATCCACGAATAAGTCGCTCGTGCTGTTAGACGAGGTAGGGCGTGGCACGTCTAC
2041 ACAAGACGGCGTCTCTATTGCGTGGGCCATCTCTGAGGAACTGATAAAGCGTGGATGCAA
2101 GGTCCTGTTCGCGACACATTTTACGGAGTTAACCGAGCTTGAAAAGCACTTCCCTCAAGT
2161 ACAGAACAAGACCATCTTGGTGAAGGAGGAGGGGAAGAATGTCATCTTCACTCATAAAGT
2221 CGTAGATGGAGTTGCCGACCGCTCTTATGGGATAGAGGTTGCTAAAATTGCAGGTATTCC
2281 AGACCGTGTTATCAATCGGGCCTACGAGATTTTGGAGCGCAATTTCAAAAACAATACCAA
2341 AAAGAATGGCAAGAGTAACCGCTTCTCACAGCAGATCCCACTTTTCCCGGTTTGACTCGA
2401 GCGGAC
```

Sequencing primers: Tma (2406 bp total, shoot for –500, 360, 860, 1360, 1860):

```
TmaMutSseq1: GGTGTTCCGGGCAGATAATCTGAG (24 bases, 54%GC, Tm~58, 3' end @501)
TmaMutSseq2: CCATAGTAGAGGACGAGTTCTTATCG (26 bases, 46%GC, Tm~58, ends @355)
TmaMutSseq3: GAAGAAATGGATATTACATCCGCTGG (26 bases, 42%GC, Tm~56, ends @899)
TmaMutSseq4: GGGTGGGTTACAACCAGGTCTTTGG (25 bases, 56%GC, Tm~57, ends @1359)
TmaMutSseq5: ATGGCACAAATTGGGTCTTTTGTACC (26 bases, 42%GC, Tm~56, ends @1860)
```

Aae MutS protein sequence

```
   1 MEKSEKELTPMLSQYHYFKNQYPDCLLLFRLGDFYELFYEDAYIGSKELGLVLTSRPAGK
  61 GKERIPMCGVPYHSANSYIAKLVNKGYKVAICEQVEDPSKAKGIVKREVVRVITPGTFFE
 121 RDTGGLASLYKKGNHYYVGYLNLAVGEFLGAKVKIEELLDLLSKLNIKEILVKKGEKLPE
 181 ELEKVLKVYVSELEEEFFEEGSEEILKDFGVLSLQAFGFEEDTYSLPLGAVYKYAKTTQK
 241 GYTPLIPRPKPYRDEGFVRLDIKAIKGLEILESLEGRKDISLFKVIDRTLTGMGRRRLKF
 301 RLLSPFRSREKIERIQEGVQELKENREALLKIRQILEGMADLERLVSKISSNMATPRELV
 361 YLKNSLKKVEELRLLLLELKAPIFKEILQNFEDTKKIINDIEKTLVEDPPLHVKEGGLIR
 421 EGVNAYLDELRFIRDNAETYLREYEKKLRQETGIQSLKIGYNKVMGYYIEVTKPNLKYVP
 481 SYFRRRQTLSNSERFTTEELQRLEEKILSAQTRINDLEYELYKELRERVVKELDKVGNNA
 541 SAVAEVDFIQSLAQIAYEKDWAKPQIHEGYELIIEEGRHPVIEEFVENYVPNDTKLDRDS
 601 FIHVITGPNMAGKSSYIRQVGVLTLLSHIGSFIPARRAKIPVVDALFTRIGSGDVLALGV
 661 STFMNEMLEVSNILNNATEKSLVILDEVGRGTSTYDGIAISKAIVKYISEKLKAKTLLAT
 721 HFLEITELEGKIEGVKNYHMEVEKTPEGIRFLYILKEGKAEGSFGIEVAKLAGLPEEVVE
 781 EARKILRELEEKENKKEDIVPLLEETFKKSEEAQRLEEYEEIIKKIEEIDIGNTTPLQAL
 841 LILAELKKKCSFSKKESGAX
```

Aae MutS Construct DNA sequence (with flanking sequences included)

```
   1 CAGGGAATTCCATATGGAGAAGTCAGAAAAGGAATTAACCCCTATGTTAAGCCAATACCA
  61 TTATTTCAAGAACCAGTACCCGGACTGTTTACTCTTGTTCCGGCTCGGGGATTTTTACGA
 121 GTTGTTCTATGAAGATGCGTATATAGGGTCTAAAGAGCTCGGGTTGGTGCTGACATCCCG
 181 CCCAGCGGGTAAGGGGAAAGAACGCATTCCCATGTGCGGAGTGCCATATCATTCTGCCAA
```

```
 241 CTCCTACATTGCGAAACTGGTAAACAAAGGATACAAAGTTGCTATTTGTGAGCAGGTCGA
 301 GGATCCGTCTAAGGCAAAGGGAATAGTAAAACGGGAGGTTGTACGGGTGATTACCCCGGG
 361 AACCTTTTTCGAGCGCGACACAGGGGGGCTTGCTAGTCTCTATAAGAAAGGCAACCATTA
 421 CTATGTGGGATATCTCAATCTCGCTGTAGGCGAATTCTTAGGAGCGAAAGTTAAGATTGA
 481 GGAACTGCTTGACCTGCTTAGCAAACTCAATATCAAGGAAATCTTGGTTAAGAAGGGCGA
 541 GAAACTTCCAGAGGAGCTGGAAAAGGTCCTGAAGGTGTACGTCAGCGAACTTGAGGAGGA
 601 GTTTTTTGAGGAAGGTTCAGAAGAAATTCTCAAGGACTTCGGCGTTTTGTCGCTGCAAGC
 661 ATTCGGGTTTGAAGAAGATACCTATAGCTTACCACTTGGTGCAGTCTACAAATACGCCAA
 721 AACTACTCAGAAAGGGTACACGCCTTTGATACCACGCCCTAAACCATACCGCGATGAGGG
 781 GTTCGTTCGCCTCGACATAAAAGCGATAAAGGGTCTCGAAATATTGGAGTCACTTGAAGG
 841 CCGGAAGGACATTAGCCTGTTTAAGGTAATTGACCGTACGTTAACTGGGATGGGACGGCG
 901 TCGTTTGAAATTCCGTCTCTTATCGCCATTCCGCAGTCGGGAGAAAATAGAACGCATTCA
 961 GGAAGGCGTACAGGAATTGAAAGAGAACCGCGAAGCCTTGCTGAAAATCCGTCAGATACT
1021 CGAAGGGATGGCAGATCTCGAACGCCTCGTGTCGAAAATTTCGTCGAACATGGCGACTCC
1081 GCGCGAACTCGTTTATTTGAAAAATAGTTTAAAGAAGGTGGAAGAACTGCGGTTGTTGTT
1141 GCTTGAGTTAAAGGCGCCGATATTCAAGGAGATACTGCAGAACTTCGAGGACACGAAGAA
1201 AATTATCAATGACATCGAGAAGACTTTGGTAGAGGATCCACCGTTACATGTTAAAGAGGG
1261 GGGGTTGATACGTGAGGGGGTGAATGCATATCTTGATGAGCTTCGTTTTATTCGGGACAA
1321 CGCGGAGACCTATTTACGTGAATACGAAAAGAAGCTGCGCCAAGAGACTGGCATCCAGAG
1381 CTTGAAGATAGGATATAACAAGGTCATGGGTTACTATATCGAGGTGACGAAACCAAACTT
1441 AAAGTACGTTCCATCTTACTTCCGCCGTCGTCAGACTCTTAGTAATAGCGAACGGTTCAC
1501 TACAGAGGAACTTCAGCGCCTGGAGGAAAAGATTCTGAGCGCCCAGACACGGATTAACGA
1561 TCTTGAGTATGAACTTTACAAGGAGTTACGCGAGCGGGTCGTCAAAGAGTTGGACAAGGT
1621 AGGGAACAACGCTTCGGCCGTTGCCGAGGTCGACTTTATACAATCTCTCGCTCAAATTGC
1681 TTATGAAAAAGACTGGGCAAAACCCCAGATCCATGAGGGCTACGAGCTTATAATCGAGGA
1741 GGGACGTCATCCTGTGATCGAGGAATTTGTTGAAAACTATGTCCCCAACGATACGAAGTT
1801 AGACCGTGACTCGTTCATCCATGTTATTACCGGACCCAATATGGCCGGTAAGTCTAGTTA
1861 TATTCGGCAAGTCGGCGTTTTGACTTTGCTCAGCCACATTGGATCTTTTATACCTGCACG
1921 TCGGGCGAAGATCCCTGTCGTAGACGCTCTCTTTACGCGGATTGGTTCAGGCGACGTTTT
1981 AGCTTTAGGCGTTTCTACATTTATGAATGAGATGCTGGAGGTGTCTAACATCCTTAACAA
2041 TGCGACGGAAAAGTCGCTCGTGATCCTGGACGAAGTGGGTCGTGGTACAAGCACTTATGA
2101 TGGAATCGCAATTAGTAAGGCTATCGTGAAATACATATCCGAAAAGCTGAAAGCCAAGAC
2161 TCTGTTGGCAACCCATTTCTTAGAGATCACCGAATTAGAGGGGAAGATCGAGGGAGTCAA
2221 GAACTATCACATGGAAGTCGAGAAAACACCTGAGGGTATTCGCTTTCTGTATATTCTGAA
2281 GGAGGGCAAAGCCGAAGGCTCGTTTGGCATAGAAGTAGCAAAATTAGCTGGCCTGCCGGA
2341 GGAGGTGGTCGAAGAAGCACGTAAGATCCTCCGCGAGTTAGAGGAGAAAGAAAACAAGAA
2401 AGAGGACATCGTGCCATTGTTGGAAGAGACATTCAAAAAGTCCGAGGAAGCCCAACGTTT
2461 GGAGGAGTACGAGGAAATAATCAAAAAAATCGAAGAGATCGACATCGGGAATACCACTCC
2521 CTTGCAGGCTCTGCTCATACTTGCGGAACTTAAGAAAAAATGTTCATTCTCCAAGAAGGA
2581 ATCCGGTGCCTGACTCGAGCGGAC
```

Sequencing primers

Aae (2604 bp total, shoot for -580, 440, 970, 1500, 2030):
AaeMutSseq1: TCCTCAAGTTCGCTGACGTACACC (24 bases, 54%GC, Tm~59, 3' ends @574)
AaeMutSseq2: AGGCAACCATTACTATGTGGGATATCTC (28 bases, 43%GC, Tm~58, ends @436)
AaeMutSseq3: GGAAGGCGTACAGGAATTGAAAGAG (25 bases, 48%GC, Tm~58, ends @985)
AaeMutSseq4: CGTCGTCAGACTCTTAGTAATAGCG (25 bases, 48%GC, Tm~59, ends @1490)
AaeMutSseq5: GAATGAGATGCTGGAGGTGTCTAAC (25 bases, 48%GC, Tm~57, ends @2029)

EGFP sequence:

```
GGGGACCACTTTGTACAAGAAAGCTGGGTCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCT
TTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGA
CCATGATTACGCCTAGCTTGCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACCGGTCGCCACCATGGTGAGC
AAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAG
CGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGC
```

CCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAG
CAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAA
CTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCA
AGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAG
CAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTA
CCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGA
GCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATG
GACGAGCTGTACAAGTAAAGCGGCCGCGACTCTAGAATTCAGCCTGCTTTTTTGTACAAACTTGTGGGG

EGFP oligos: Operon 2nd set, parsed with DNAWorks 3.0 (including flanking sequences for Clonase (Invitrogen))

EGFP-N50t1	GGGGACCACTTTGTACAAGAAAGCTGGGT
EGFP-N50b2	ACTCACATTAATTGCGTTGCGACCCAGCTTTCTTGTACAAAG
EGFP-N50t3	GCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCC
EGFP-N50b4	CGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAG
EGFP-N50t5	GGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAAT
EGFP-N50b6	GTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGC
EGFP-N50t7	TGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCA
EGFP-N50b8	GCATGCAAGCTAGGCGTAATCATGGTCATAGCTGTTTCCTGT
EGFP-N50t9	ACGCCTAGCTTGCATGCCTGCAGGTCGACTCTAGAGGATCCC
EGFP-N50b10	CCATGGTGGCGACCGGTACCCGGGGATCCTCTAGAGTCGACC
EGFP-N50t11	CCGGTCGCCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACC
EGFP-N50b12	GCTCGACCAGGATGGGCACCACCCCGGTGAACAGCTCCTCGC
EGFP-N50t13	CCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAG
EGFP-N50b14	CTCGCCCTCGCCGGACACGCTGAACTTGTGGCCGTTTACGTC
EGFP-N50t15	CCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCC
EGFP-N50b16	TGCCGGTGGTGCAGATGAACTTCAGGGTCAGCTTGCCGTAGG
EGFP-N50t17	ATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTC
EGFP-N50b18	CTGCACGCCGTAGGTCAGGGTGGTCACGAGGGTGGGCCAGGG
EGFP-N50t19	GACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACAT
EGFP-N50b20	GGACTTGAAGAAGTCGTGCTGCTTCATGTGGTCGGGGTAGCG
EGFP-N50t21	AGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCC
EGFP-N50b22	GTCCTTGAAGAAGATGGTGCGCTCCTGGACGTAGCCTTCGGG
EGFP-N50t23	GCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCG
EGFP-N50b24	TGTCGCCCTCGAACTTCACCTCGGCGCGGGTCTTGTAGTTGC
EGFP-N50t25	TGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGA
EGFP-N50b26	CGTCCTCCTTGAAGTCGATGCCCTTCAGCTCGATGCGGTTCA
EGFP-N50t27	CATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCT
EGFP-N50b28	GTTGTGGCTGTTGTAGTTGTACTCCAGCTTGTGCCCCAGGAT
EGFP-N50t29	TACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAG
EGFP-N50b30	CACCTTGATGCCGTTCTTCTGCTTGTCGGCCATGATATAGAC
EGFP-N50t31	GAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACAT
EGFP-N50b32	CGAGCTGCACGCTGCCGTCCTCGATGTTGTGGCGGATCTTGA
EGFP-N50t33	GCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCA
EGFP-N50b34	GGCAGCAGCACGGGGCCGTCGCCGATGGGGGTGTTCTGCTGG
EGFP-N50t35	CCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTC

EGFP-N50b36	CGTTGGGGTCTTTGCTCAGGGCGGACTGGGTGCTCAGGTAGT
EGFP-N50t37	TGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGC
EGFP-N50b38	TCCCGGCGGCGGTCACGAACTCCAGCAGGACCATGTGATCGC
EGFP-N50t39	ACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG
EGFP-N50b40	TTCTAGAGTCGCGGCCGCTTTACTTGTACAGCTCGTCCATGC
EGFP-N50t41	CGGCCGCGACTCTAGAATTCAGCCTGCTTTTTTGTACAAACT
EGFP-N50b42	CCCCACAAGTTTGTACAAAAAAGCAGGCT

GFP Sequencing primers

GFP-F: CCTCGTGACCACCCTGAC (faces forward, internal to sequence)
GFP-R: CACCAGGGTGTCGCCCTC (faces backward, internal to sequence)