

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Error correction in gene synthesis technology

Siying Ma¹, Ishtiaq Saaem^{1,2} and Jingdong Tian^{1,2}

¹ Department of Biomedical Engineering and the Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA

² Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

Accurate, economical and high-throughput gene and genome synthesis is essential to the development of synthetic biology and biotechnology. New large-scale gene synthesis methods harnessing the power of DNA microchips have recently been demonstrated. Yet, the technology is still compromised by a high occurrence of errors in the synthesized products. These errors still require substantial effort to correct. To solve this bottleneck, novel approaches based on new chemistry, enzymology or next generation sequencing have emerged. This review discusses these new trends and promising strategies of error filtration, correction and prevention in *de novo* gene and genome synthesis. Continued innovation in error correction technologies will enable affordable and large-scale gene and genome synthesis in the near future.

Gene synthesis

With the rise of synthetic biology, the era of creating new functional genes, genetic networks and whole genomes is upon us. Heralding the dawn of this new era are the rapid technological breakthroughs allowing for on-demand synthesis of DNA of any sequence or length. Recent breakthroughs have resulted in synthesis and assembly of an entire bacterial genome and creation of a new cell controlled by this transplanted synthetic genome [1]. The demand for synthetic genes and genomes will probably continue to increase as the scope of their applications expands.

Gene synthesis is typically accomplished by enzymatic assembly of chemically synthesized overlapping oligonucleotides that span the entire length of the gene construct (reviewed in [2,3]). The resulting products unavoidably contain errors such as deletions, insertions, or base substitutions, due largely to mistakes in chemical oligonucleotide synthesis and to a lesser extent, to the subsequent enzymatic gene assembly processes. Cloning and sequencing of multiple clones is normally required to identify the clones with the correct sequence. Improvement in fidelity of gene synthesis is essential to continued progress in the development of this technology, because a substantial fraction of the overall cost of gene synthesis goes to cloning and sequencing, which is required for selecting and confirming a correct sequence.

Nature has evolved sophisticated error-correction mechanisms to ensure that DNA replication proceeds with high

fidelity [4]. The error rates in prokaryotic and eukaryotic replication machineries range from 10^{-7} to 10^{-8} owing to various proofreading and mismatch repair mechanisms [5,6]. By contrast, the current gene synthesis process has a typical error rate of 10^{-2} to 10^{-3} , or 1–10 errors/kbp synthesized [7–10].

Given an error rate (R), the probability of a synthetic DNA sequence being error-free, $(1 - R)^N$, decreases exponentially as its length (N) increases. The number of clones that need to be sequenced in order to have 95% confidence of obtaining one perfect clone, $\ln(1 - 0.95)/\ln[1 - (1 - R)^N]$, can be dramatically reduced with a 10-fold reduction in error rate [8]. The presence of a variety of error types and error sources from the series of DNA synthesis and gene assembly steps allows establishment of error-control procedures at multiple levels. This article reviews the current methods for error filtration, correction and prevention. We also discuss new developments and new directions that may potentially enable error-free synthesis for oligonucleotides, genes, and genomes in the near future.

Error removal from synthetic oligonucleotides

The dominant source of errors in synthetic DNA comes from chemical synthesis of oligonucleotides. Standard solid-phase oligonucleotide synthesis uses the classical phosphoramidite chemistry [11,12], which adds each nucleotide monomers to the 5'-end of the elongating DNA chain in a four-step cycle. (i) Deprotection: an acid is used to remove the protecting demethoxytrityl (DMT) group from the 5'-end of the growing oligonucleotide chain and generate a reactive 5'-OH group. (ii) Coupling: the 5'-OH group generated from the deprotection step reacts with an activated monomer created by adding the desired phosphoramidite and an appropriate activator (i.e. tetrazole) simultaneously. Tetrazole, a weak acid, protonates the trivalent phosphorus on the 3'-end of the monomer. This results in slow displacement of the secondary amine and formation of a highly reactive tetrazolide that then immediately couples with the OH group. (iii) Capping: uncoupled 5'-OH groups are blocked by an acylating capping reagent, which is delivered along with a nucleophilic catalyst, to minimize deletion products. (iv) Oxidation: the unstable phosphite triester internucleotide linkages are oxidized to a more stable pentavalent phosphotriester [13].

The most frequent type of errors in oligonucleotide synthesis happens when a new phosphoramidite monomer fails to couple to the elongating chain, which results in a

Corresponding author: Tian, J. (jtian@duke.edu).

typical step-wise coupling efficiency of 98.5–99.5% [14–16]. Uncoupled chains will be terminated from growth by acetylation and result in truncated oligonucleotides. However, failures in acetylation or deprotection do happen with a frequency as high as 0.5% per position, which leads to deletion errors in the final synthetic DNA. Insertions also occur when DMT is cleaved by excess activator and can reach 0.4% per base [15].

Post-synthesis, the purity of the synthesized oligonucleotide pool can be improved by size exclusion purification using HPLC [17] or PAGE [15]. Hydrophobic purification cartridges can also be used for purification of the oligonucleotide pool before the hydrophobic DMT blocking group is removed (trityl-on). Full-length oligonucleotides with a hydrophobic DMT terminus can be readily separated from prematurely terminated sequences lacking the blocking group. With these methods, >90% of the impurities (mostly insertions/deletions and truncations) can be eliminated before assembly and, as a result, the error rate in the final product can be reduced by several fold [18–20]. Although such methods are relatively laborious, the effort can be justified if gene-construction oligos of the same length are pooled and purified together [19]. The drawback of size exclusion purification methods is that they are generally ineffective against base substitutions or single-base insertions/deletions, especially for long oligos. In fact, single-base deletions are the most frequently observed error type in assembled DNA constructs [8] and they cannot be removed effectively by size exclusion purification methods.

Besides post-synthesis purification, a fundamental approach to increase the accuracy of chemical DNA synthesis is to develop more efficient synthesis chemistry. Along this line, an alternative two-step DNA synthesis method utilizes a peroxy anion as the nucleophile to remove simultaneously a 5'-carbonate and oxidize the internucleotide phosphite trimer [21]. The removal of the 5'-protecting group with peroxy anion under mildly basic conditions is considered essentially irreversible and quantitative, and therefore has the potential to eliminate depurination and reduce mutation frequencies completely in cloned, synthetic DNA. To date, there has been no report on the widespread use of the new two-step DNA synthesis method. Whether it will become the next new standard remains unclear.

Error removal from chip-synthesized oligo pools

Oligonucleotide pools synthesized from microarrays have recently been used as an economical source for large-scale gene synthesis. However, oligos synthesized on planar surfaces tend to be more prone to errors. Depurination of purine bases seems to be of major concern [22]. As a result of prolonged exposure to deprotecting/detritylation agents, adenine and guanine bases often undergo degradation by hydrolysis, leaving behind only the ribose sugar backbone. The presence of the 5'-OH-presenting sugar allows the oligonucleotide chain to elongate further; however, during the final side-group removal step (typically using ammonium hydroxide), these apurinic bases are cleaved and thereby result in truncated products.

DNA synthesis on planar surfaces often uses a modified version of the four-step phosphoramidite chemistry where a certain step is gated to provide spatial control of the

individual oligonucleotides being synthesized. This is typically the coupling step for inkjet printing-based synthesis (Agilent, Protogene), or the deblocking step for light-based (LC Sciences, Invitrogen, Affymetrix) and redox reaction-based platforms (Combimatrix, Oxamer). In any case, truncated products could arise from misalignment of printed droplets or from partial deblocking as a result of poor light-source registration and improper sequestering of redox ions. Erroneous products caused by such 'edge effects' could be mitigated by using patterned substrates for synthesis [23,24]. Studies using the inkjet chip synthesis platform have demonstrated that the error rate can be reduced from one error in ~200 bases to one in ~600 bases by using patterned silica features on a plastic chip [25].

In addition to size exclusion purifications, strategies using the hybridization–selection principle have been used to reduce errors in microarray synthesized oligo pools. Error-containing oligos can be removed by stringent hybridization selections using short complementary oligos (selection oligos) immobilized on beads [7]. Gene construction oligos with errors form imperfect matches with the selection oligos and can be washed away under stringent washing conditions, whereas those without errors can be retained and enriched. This strategy may be useful for cleaning up large pools of microarray-derived oligos but may not be convenient or economical for purifying small numbers of oligos due to the burden of synthesizing complementary selection oligos for all gene-construction oligos.

Without a separate prepurification step, simply increasing hybridization stringency during gene assembly reaction helps prevent incorporation of erroneous oligos into the final assembly products [26]. This is because perfect hybridization among error-free sequences creates better templates for the polymerase or the ligase used in gene assembly. Ligation-based chain assembly methods tend to benefit more from this effect because no gap is allowed between oligos. Polymerase-based cyclic assembly (PCA) methods allow the use of longer oligos and, as a result, the middle portion of the oligos are not subject to hybridization selection and tend to carry over more errors into assembled gene products [7,26–28]. It takes fewer long oligos to assemble a gene, thus, there is an advantage of using long oligos provided that the sequence quality is satisfactory. High quality long oligos are available from commercial sources such as Integrated DNA Technologies ('Ultramers', as long as 200 bp; Coralville, IA, USA). Agilent Technologies (Santa Clara, CA, USA) has also recently started producing oligonucleotide libraries composed of >100 bp sequences synthesized by their SurePrint microarray platform [22].

The rapid development of next-generation sequencing (NGS) technology has made it possible to sequence large pools of oligo sequences at affordable costs. This has triggered the temptation of selecting sequence-verified oligos as input for gene assembly. A proof-of-concept experiment has been performed that has demonstrated that the so-called 'megacloning' method can reduce error rates by a factor of 500 compared to the starting oligonucleotide pool generated by microarray [29]. In principle, with future development in platform automation, millions of oligos can be sequenced and sorted in a single megacloner run, which will potentially enable gene construction up to megabases in length [29].

Error removal from synthetic genes

Despite exhaustive purification, errors that remain in synthetic oligos will be carried over during the assembly process and accumulate in downstream gene constructs. Additional errors may also be introduced by polymerase elongation or mis-hybridization among oligos [30]. Selection of an error-free synthetic gene sequence often requires expensive and time consuming cloning and sequencing steps and some good luck. Besides cloning and sequencing, expression assays or functional screens can also be used to select desired products against errors that can cause reading frame shift and/or loss of protein functions [19,31–34]. Nevertheless, this method is only selective to protein-coding regions or sequences encoding functional elements and is not effective for spotting silent or conservative mutations.

Longer DNA constructs (5–50 kb) are usually assembled step-wise to achieve desired quality and efficiency. Short segments (<1 kb) are first synthesized as building blocks, and their sequences verified individually by cloning and sequencing before being joined together into longer segments. Considering the error frequencies in starting synthetic oligonucleotides and the efficiency of cloning and sequencing, many current error-reducing strategies target this intermediate assembly stage to achieve maximum error-elimination efficiency through simple, robust and ideally automatable procedures [35,36].

Most of the current error-removal techniques make use of DNA mismatch recognition agents. In contrast to mutation detection and correction *in vivo* [6,37], synthetic DNA does not have chemical labels to distinguish between the 'correct' and the 'mutant' strands. Moreover, the PCR-mediated assembly process will copy the existing errors into the complementary strand. Therefore, the likelihood of errors existing as mismatches in the assembled polynucleotide constructs is very small. It is then necessary to reassociate randomly the polynucleotides through a denaturation and rehybridization process during which erroneous bases form mismatches with the corresponding correct

bases in the reverse-complementary strand. The bulging mismatch sites are then recognized and removed by using mismatch-binding proteins or mismatch-cleavage enzymes (Figure 1).

Error removal using MutS mismatch-binding protein

The MutS protein is part of the bacterial MutHLS mismatch repair machinery [38]. It detects and binds to a variety of mispaired bases and small single-strand loops *in vivo*. Methods have been developed that use MutS as an error-removal agent for gene synthesis. In one strategy, MutS protein from the thermophilic bacteria *Thermus aquaticus* was used to filter out directly full-length heteroduplexes [8]. After denaturation and reannealing, mismatch containing heteroduplexes are recognized and bound by the *Taq* MutS protein, which can be separated from the unbound homoduplexes (mostly error-free) by a gel-mobility shift assay (Figure 2a). This method has been reported to reduce error rate to one error/3.8 kb with one round of filtration, which represents several folds of improvement from raw synthesis products. Repeating the procedure one more time reportedly reduces the error rate to one error/10 kb, which is good enough to ensure finding a correct sequence with a single round of cloning and sequencing. A major limitation with removing full-length heteroduplexes is that it requires a substantial fraction of the sequences to be error-free to survive the MutS binding filtration.

For error-rich sequences, an alternative MutS-based error-correction strategy is called 'consensus shuffling' [39] (Figure 2b). Reannealed polynucleotide products are first cleaved by a mixture of restriction endonucleases and the resulting overlapping short fragments are then subjected to MutS column filtration. Short, mismatch-containing pieces are captured by immobilized *Taq* MutS, whereas error-free fragments are eluted and reassembled into full-length constructs by assembly PCR. Two iterations of consensus shuffling improve the error rate 3.5–4.3-fold to one error/3.5 kb [39].

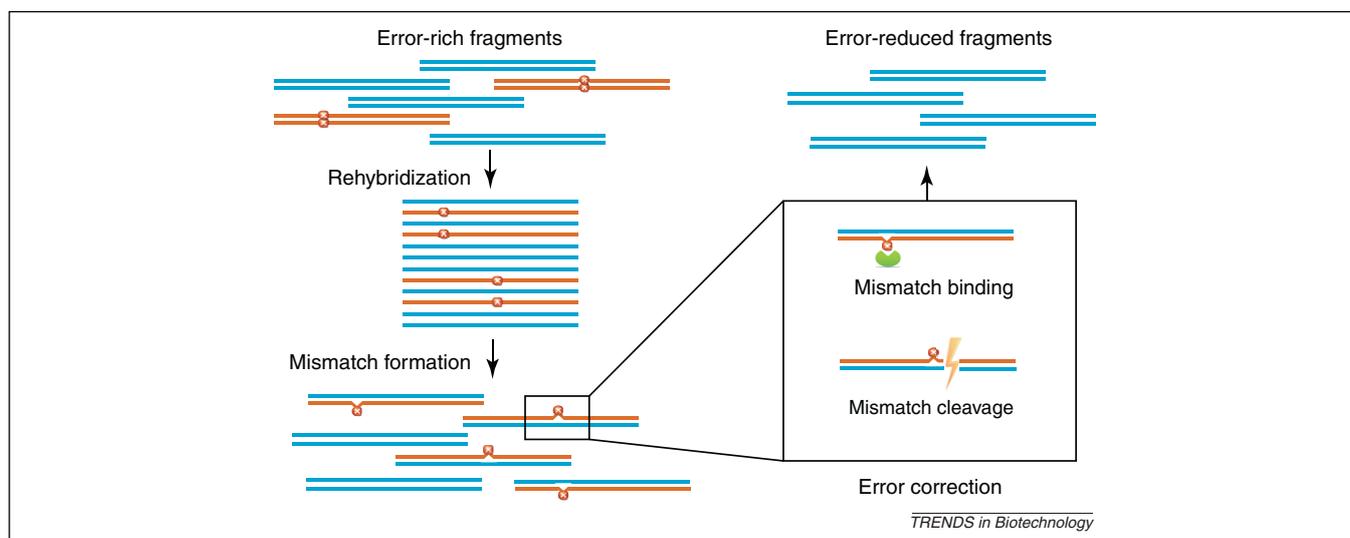


Figure 1. Summary of general schemes of mismatch-based error correction in synthetic DNA constructs. Assembly products are heat-denatured and then reannealed to allow correct (blue lines) and mutant (red lines) strands to rehybridize and form mismatches randomly. The mismatches are then removed by various error correction methods using either mismatch-binding proteins or mismatch-cleaving enzymes to yield products with correct sequences.

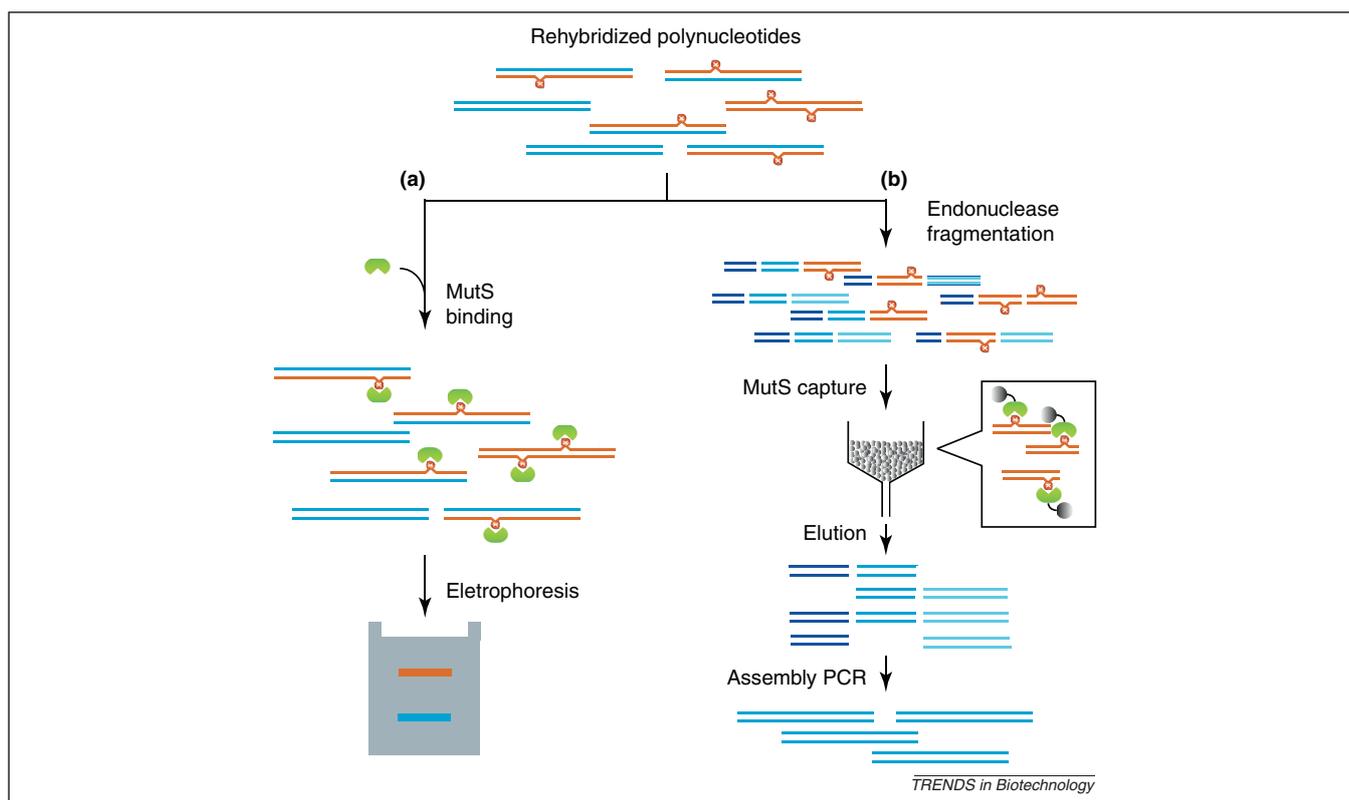


Figure 2. Schematic of error removal strategies using MutS mismatch-binding protein. Assembled gene segments are denatured and reannealed to form heteroduplexes where errors are revealed as mismatches between correct (blue) and mutant (red) sequences. (a) MutS proteins recognize and bind to mismatch structures. Full-length heteroduplexes with MutS bound can be separated from homodimers without MutS by gel-shift assay. (b) Consensus shuffling. Reannealed constructs are fragmented by type IIS restriction endonucleases. Fragments containing mismatches are filtered away by MutS binding column. Error-free fragments that pass through the column are collected and reassembled into full-length genes by assembly PCR.

Consensus shuffling is a mechanism that resembles DNA shuffling in which several iterations can be made until the consensus sequences are concentrated sufficiently to outnumber other species. Consensus shuffling provides several advantages over direct MutS filtration of full-length sequences: (i) it is more effective on longer DNA constructs in which perfect hybridization after reassortment is rare; (ii) it tolerates more errors in the starting material; and (iii) the removal of the entire DNA heteroduplex is not necessary. Importantly, fragmentation is ideally performed with type IIS restriction endonucleases that cut outside of the recognition sites. In this way, distinct ends can be generated by the same restriction endonuclease to minimize nonspecific hybridization of cut fragments after error filtration. However, despite the merit of the idea, consensus shuffling requires extraneous use of endonucleases and the reaction conditions may still need to be improved to achieve better outcome [39]

Error correction using mismatch-cleaving enzymes

Mismatch-cleaving enzymes here refer to a group of mismatch-specific endonucleases that recognize and cleave at or near mismatch sites in DNA heteroduplexes. The family embraces a variety of members including but not limited to: (i) resolvases, such as phage T4 endonuclease VII [40], T7 endonuclease I [41] and *Escherichia coli* endonuclease V [42]; (ii) mismatch repair endonuclease MutH [38]; and (iii) single-strand specific nucleases, such as S1 nuclease from *Aspergillus oryzae*, P1 nuclease from *Penicillium citrinum*,

mung bean nuclease, and CEL nuclease from celery [43,44]. The ability of these endonucleases to cleave heteroduplex DNA at the mismatch sites has led to their extensive applications as probes for mutation and polymorphism detection [43,45–49]. The mismatch-cleaving activity also makes them very useful in error removal for gene synthesis and several different assays have been developed utilizing this activity.

Resolvases

T7 endonuclease I was used in an early study: after assembly and amplification of a target DNA, the purified product was denatured, reannealed and treated with T7 endonuclease I. Afterwards, the pool was run on a gel and the remaining intact band was excised and extracted, while the cleaved fragments were discarded [50]. This way of enriching error-free sequences seems to work for high-quality synthetic products where correct sequences outnumber mutants.

The efficacy of various resolvases has been systematically explored to recognize and cleave single-base mismatches using a different error-removal assay, which is applicable for low-quality sequences. Three resolvases, T4 endonuclease VII, T7 endonuclease I and *E. coli* endonuclease V were tested respectively in the process of ligation-based gene synthesis and functional cloning of bacterial chloramphenicol-acetyltransferase (cat) gene. After treatment with T4 endonuclease VII and *E. coli* endonuclease V, the fraction of 'functional clones' increased over 10-fold, and sequence

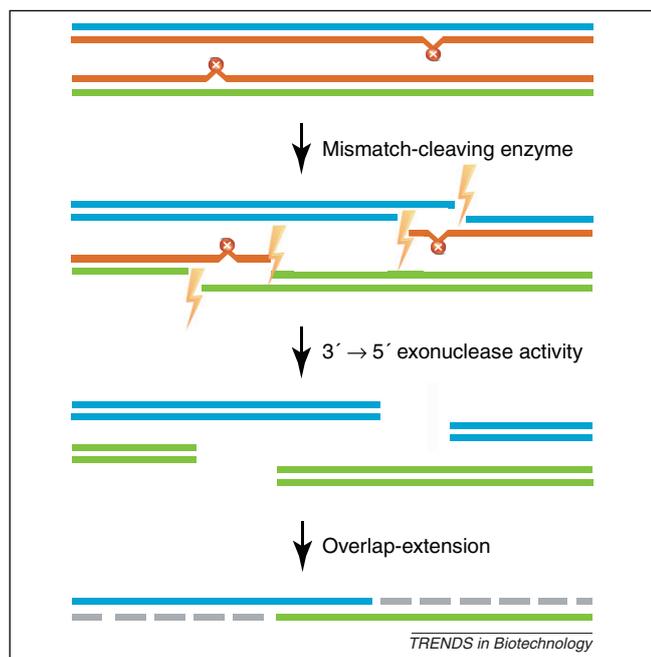


Figure 3. Schematic diagram showing the principle of error removal using mismatch-cleaving enzymes. The reannealed heteroduplexes are cleaved in both strands by mismatch-specific enzymes 2–5 bp downstream of the mismatches, generating stagger ends that are immediately dissociated at the reaction temperature. The single-stranded overhangs that contain mismatch bases are degraded by an exonuclease in the reaction mixture. Full-length genes with fewer errors are recovered by overlap assembly PCR [32].

analysis revealed an over 4-fold reduction of error rate. T7 endonuclease I was not fully characterized in this setting [32]. Instead of using gel purification, it was demonstrated that following mismatch-cleavage, an additional 3' to 5'-exonuclease treatment to remove single-stranded overhangs was essential for error removal, either by adding *E. coli* exonuclease I or using the intrinsic 3' to 5'-exonuclease activity of proofreading DNA polymerases (Figure 3). The processed DNA fragments were reassembled into full-length products by assembly PCR.

Another method, called circular assembly amplification, combines exonuclease and endonuclease treatments to improve gene synthesis quality [51]. In this method, the overlapping and complementary oligonucleotides are first ligated into circular molecules under stringent conditions. An exonuclease is added to degrade any uncircularized polynucleotides. The circular DNA constructs are then treated with *E. coli* endonuclease V, which linearizes and degrades the mismatch-containing circles in concert with the exonuclease. By using the above three tiers of selections, the method is able to reduce error rate by at least 7-fold compared to the conventional PCA protocol [51].

A problem with phage resolvase-based mismatch cleavage is preferential cleavage of certain types of mismatches. For instance, T7 endonuclease I has been reported to fail to achieve quantitative cleavage for selected mispairs [32]. In addition, the degree of specificity of T4-endonuclease VII has been found to be highly dependent on the length of the substrate and the sequence surrounding the mismatch sites [52,53]. This enzymology limitation is intrinsic to

these resolvases because they naturally function to recognize holiday junctions rather than mismatches in a heteroduplex [54]. Their potency and efficacy of mismatch cleavage are therefore compromised by both low sensitivity and high undesired background. Future work to optimize their substrate specificity or use enzyme combinations may potentially produce better results.

MutHLS

E. coli proteins MutH, MutL and MutS constitute a bacterial mismatch repair mechanism: MutS detects and binds to mispaired bases and small single-strand loops; MutL couples the MutH endonuclease to the MutS bound mutant sites, leading to incision of the unmethylated strand of a hemimethylated dsDNA. The error-containing staggered ends are trimmed by a helicase and an exonuclease, and the resulting gap is filled by a DNA polymerase and a ligase [5]. MutHLS can be used to remove polymerase-produced mutant sequences in PCR products [38]. Treating reannealed PCR products with MutHLS allows MutS to bind to the mismatches and create incisions on both unmethylated strands at d(GATC) in the vicinity of the mismatch sites. The cleaved heteroduplexes are then removed by gel electrophoresis. Mismatch removal using *E. coli* MutHLS reduces the error rate by one order of magnitude and that is effective to G–T, A–C, G–G, A–A and small insertion/deletion mispairs, but less effective for other types of mismatches [38].

Single-strand specific nucleases

Single-strand specific nucleases are a family of multifunctional enzymes that are present ubiquitously and display a wide variety of characteristics. Although more than 30 single-strand nucleases have been isolated from various sources, only a few have been thoroughly characterized. The most widely used single-strand nucleases are almost exclusively extracellular glycoproteins, including S1 and P1 nucleases from fungi, the mung bean nuclease and the CEL nuclease from plants. These nucleases vary drastically in the degree of specificity towards different types of mismatches. For instance, S1 and P1 nucleases appear to be strongly specific to AT-rich regions [54]; S1 nuclease seems incapable of recognizing single base mismatches [55]; and quantitative single nucleotide mismatch cleavage by mung bean nuclease occurs only at pH 6–6.5 [56]. Moreover, contradictory results have been reported regarding the use of S1, P1 and mung bean nucleases in detecting different types of single nucleotide polymorphism in heteroduplex DNA [56]. The incomplete capability of these single-strand nucleases to identify all possible types of mismatches greatly discourages their general use in synthetic gene error correction.

CEL endonuclease is an ortholog of the S1 nucleases isolated from celery [44,57]. The uniqueness of CEL endonuclease is that it is the first eukaryotic endonuclease known to cleave DNA with high specificity towards all types of base mismatches and DNA distortions at neutral pH. Following the discovery of CEL nuclease, the techniques for high-throughput mutation detection have significantly improved, and CEL has become the most commonly used enzyme in TILLING (targeting-induced local lesions

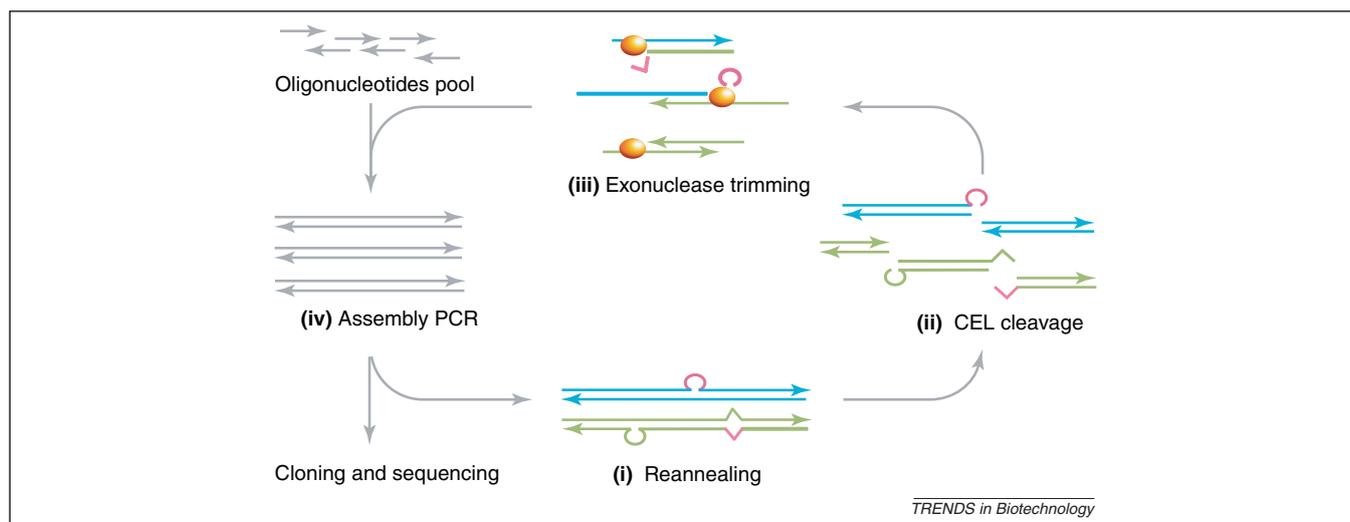


Figure 4. Schematic diagram of error correction strategy using CEL mismatch-specific endonuclease. Multiple CEL error-correction cycles may be integrated into a gene synthesis process. Each cycle consists of four steps: (i) reannealing of assembled gene constructs to present erroneous bases as mismatches; (ii) CEL nuclease cleavage on both strands at the 3' side of the mismatches; (iii) exonuclease trimming of single-stranded mismatch overhangs by added exonuclease or the 3'→5' exonuclease activity of the proofreading PCR enzyme; and (iv) reassembly and amplification of the processed fragments by assembly PCR. The final products are used for downstream applications such as cloning and sequencing.

in genomes) [58–61], polymorphism analysis [62,63] and disease diagnosis [57,62–67].

CEL endonuclease is a mannosyl glycoprotein that nicks a DNA strand at the 3'-end of the base-substitution mismatch and DNA distortion. Prolonged incubation and a high enzyme to substrate ratio facilitates double-stranded cleavage by CEL at opposite phosphodiester bonds 3' to the mismatch, generating: (i) two single-base 3' overhangs in the case of base substitutions; or (ii) a short stagger end in the case of insertion/deletion [44,57]. The commercialization of CEL endonuclease as SurveyorTM nuclease by Transgenomic, Inc (Omaha, NE, USA) has led to its extensive applications in mutation detection as a simple yet cost-effective method, resulting in the discovery of many novel mutations associated with human genetic disorders, including BRCA1 [44,57,68], EGFR [64], HCDC4 [65], p53 [66], mitochondrial [62,63] and kidney-related [67] genes. The broad substrate specificity and low undesired activity of CEL nucleases make them the most promising candidates for error correction in synthetic gene assemblies.

Recent studies employing on-chip gene synthesis technology have explored the effectiveness of SurveyorTM nuclease in error correction of microchip-synthesized genes [25,69] (Figure 4). The incorporation of a 60-min digestion step with SurveyorTM nuclease followed by PCR assembly with Phusion DNA polymerase was able to reduce the error rate of the synthetic gene products from one error/526 bp to one error/3883 bp. Two iterations of enzymatic error cleavage reactions further reduced the error rate to 1 error/8701 bp, representing an over 16-fold error reduction [69]. Selective sub-pool amplification of chip-synthesized oligonucleotides achieved an error rate of one error/7017 bp after enzymatic error correction using ErrASE kit, a commercially available CEL-based enzyme cocktail that corrects mistakes in a similar fashion [70].

The combined actions of mismatch cleavage by CEL nuclease and proofreading by Phusion DNA polymerase

makes the procedure particularly robust and suitable for correcting error-rich synthetic sequences. Reannealing of low-quality synthetic sequences often yields little error-free duplexes that can survive the mismatch-binding filtration or mismatch cleavage. Rather than discarding all error-containing heteroduplexes, CEL nuclease makes double-stranded cuts next to the mismatches. The mismatch bases are then chewed away by the exonuclease activity of the proofreading DNA polymerase. The resulting error-free fragments are reassembled into full-length gene constructs by assembly PCR (Figure 4). This method enables error elimination at the single-base level, while salvaging a majority of the correct portions of the sequences. More importantly, the 'inspection–correction–reassemble' cycle can be iterated several times until products with desired purity are obtained.

Minimizing errors in genome synthesis

Identifying and correcting errors in synthetic genomes appears to be costly and painful due to the length of the sequences and difficulty of construct assembly. The key then is to use error-free starting materials (i.e. sequence-confirmed intermediate sequences) and avoid the introduction of errors during the genome assembly process. Single-step assembly using yeast homologous recombination could produce long, high-quality DNA constructs with a very low error rate (0.054%), comparable to many *in vitro* error-removal protocols as discussed previously. A microbial genome >1 Mb has successfully been assembled using this strategy [1,71]. This technique has also been applied to rapid assembly of metabolic pathways with multiple parts [72]. In addition to using overlapping dsDNA fragments as starting materials, a yeast recombination mechanism has recently been successfully applied to direct assembly of single-strand oligonucleotides into double-stranded gene constructs [73].

Concluding remarks and perspectives

Over the past half century, *de novo* chemical DNA synthesis and enzymatic gene and genome assembly techniques have been continuously improved. The longest published synthetic DNA has extended from <1 kb to >1 Mb, which encoded functional genes, genetic pathways, and viral or bacterial genomes [1,2,74,75]. Nevertheless, the gap between the expanding demand and the capability to deliver synthetic genes and genomes in an accurate, economical and high-throughput fashion is still wide. To close the gap, current efforts in automation and miniaturization of gene synthesis need seamlessly to incorporate convenient and effective error-prevention and error-removal strategies. The methods discussed in this review reflect current thinking and practice in error correction. There is still plenty of room for improvements and innovations. For example, the new high-fidelity DNA synthesis chemistry has yet to be put into general use, which can reduce error generation from the beginning. The error-correction enzymes can be further engineered to be more effective in correcting single-base mutations. The cloning and sequencing process used to select correct sequences should be eliminated or simplified and automated using advanced NGS technology. Ultimately, what the field is looking for is a simple, effective and automatable method that ensures maximum quality of DNA synthesis at minimum cost. Concerted efforts combining more efficient DNA synthesis chemistry, more effective error-correction enzymes, and more powerful NGS sequencing technology seems to be the way to go. With continued innovation, this goal is expected to be accomplished in the near future, which will significantly benefit the synthetic biology and biotechnology fields.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by NIH (R01HG005862) and by the National Basic Research Program of China (2011CBA00800). J.T. was a Beckman Young Investigator and a recipient of The Hartwell Foundation Individual Biomedical Research Award.

References

- Gibson, D.G. *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56
- Tian, J. *et al.* (2009) Advancing high-throughput gene synthesis technology. *Mol. Biosyst.* 5, 714–722
- Czar, M.J. *et al.* (2009) Gene synthesis demystified. *Trends Biotechnol.* 27, 63–72
- Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. *Annu. Rev. Biochem.* 74, 681–710
- Schofield, M.J. and Hsieh, P. (2003) DNA mismatch repair: molecular mechanisms and biological function. *Annu. Rev. Microbiol.* 57, 579–608
- Li, G.M. (2008) Mechanisms and functions of DNA mismatch repair. *Cell Res.* 18, 85–98
- Tian, J. *et al.* (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432, 1050–1054
- Carr, P.A. *et al.* (2004) Protein-mediated error correction for *de novo* DNA synthesis. *Nucleic Acids Res.* 32, e162
- Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* 30, e43
- Xiong, A.S. *et al.* (2004) A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. *Nucleic Acids Res.* 32, e98
- Caruthers, M.H. (1985) Gene synthesis machines: DNA chemistry and its uses. *Science* 230, 281–285
- Beaucage, S.L. and Caruthers, M.H. (1981) Deoxynucleoside phosphoramidites – a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* 22, 1859–1862
- Caruthers, M.H. *et al.* (1987) Chemical synthesis of deoxypolynucleotides by the phosphoramidite method. *Methods Enzymol.* 154, 287–313
- Eckstein, F. (1991) *Oligonucleotides and Analogues: A Practical Approach*, IRL Press
- Ellington, A. and Pollard, J.D., Jr (2001) Introduction to the synthesis and purification of oligonucleotides. *Curr. Protoc. Nucleic Acid Chem.* Appendix 3, A.3C
- Zhou, X. *et al.* (2004) Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.* 32, 5409–5417
- Andrus, A. and Kuimelis, R.G. (2001) Analysis and purification of synthetic nucleic acids using HPLC. *Curr. Protoc. Nucleic Acid Chem.* Chapter 10, Unit 10.15
- Cello, J. *et al.* (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297, 1016–1018
- Smith, H.O. *et al.* (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15440–15445
- Xiong, A.S. *et al.* (2006) PCR-based accurate synthesis of long DNA sequences. *Nat. Protoc.* 1, 791–797
- Sierzchala, A.B. *et al.* (2003) Solid-phase oligodeoxynucleotide synthesis: a two-step cycle using peroxy anion deprotection. *J. Am. Chem. Soc.* 125, 13427–13441
- LeProut, E.M. *et al.* (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* 38, 2522–2540
- Saaem, I. *et al.* (2010) *In situ* synthesis of DNA microarray on functionalized cyclic olefin copolymer substrate. *ACS Appl. Mater. Inter.* 2, 491–497
- Chow, B.Y. *et al.* (2009) Photoelectrochemical synthesis of DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* 106, 15219–15224
- Quan, J. *et al.* (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.* 29, 449–452
- Borovkov, A.Y. *et al.* (2010) High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides. *Nucleic Acids Res.* 38, e180
- Smith, H.O. *et al.* (2003) Generating a synthetic genome by whole genome assembly: phi X174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15440–15445
- Gao, X. *et al.* (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, e143
- Matzas, M. *et al.* (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.* 28, 1291–1294
- Cline, J. *et al.* (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* 24, 3546–3551
- Stemmer, W.P. *et al.* (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* 164, 49–53
- Fuhrmann, M. *et al.* (2005) Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.* 33, e58
- Cox, J.C. *et al.* (2007) Protein fabrication automation. *Protein Sci.* 16, 379–390
- Kim, H. *et al.* (2010) A fluorescence selection method for accurate large-gene synthesis. *Chembiochem* 11, 2448–2452
- Kodumal, S.J. *et al.* (2004) Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15573–15578
- Reisinger, S.J. *et al.* (2007) Total synthesis of multi-kilobase DNA sequences from oligonucleotides. *Nat. Protoc.* 1, 2596–2603
- Modrich, P. (1991) Mechanisms and biological effects of mismatch repair. *Annu. Rev. Genet.* 25, 229–253

- 38 Smith, J. and Modrich, P. (1997) Removal of polymerase-produced mutant sequences from PCR products. *Proc. Natl. Acad. Sci. U.S.A.* 94, 6847–6850
- 39 Binkowski, B.F. *et al.* (2005) Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.* 33, e55
- 40 Kemper, B. and Garabett, M. (1981) Studies on T4-head maturation. 1. Purification and characterization of gene-49-controlled endonuclease. *Eur. J. Biochem. FEBS* 115, 123–131
- 41 West, S.C. (1992) Enzymes and molecular mechanisms of genetic recombination. *Annu. Rev. Biochem.* 61, 603–640
- 42 Yao, M. and Kow, Y.W. (1996) Cleavage of insertion/deletion mismatches, flap and pseudo-Y DNA structures by deoxyinosine 3'-endonuclease from *Escherichia coli*. *J. Biol. Chem.* 271, 30672–30676
- 43 Desai, N.A. and Shankar, V. (2003) Single-strand-specific nucleases. *FEMS Microbiol. Rev.* 26, 457–491
- 44 Yang, B. *et al.* (2000) Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry* 39, 3533–3541
- 45 Youil, R. *et al.* (2000) Enzymatic mutation detection (EMD) of novel mutations (R565X and R1523X) in the FBN1 gene of patients with Marfan syndrome using T4 endonuclease VII. *Hum. Mutat.* 16, 92–93
- 46 Babon, J.J. *et al.* (2003) The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Mol. Biotechnol.* 23, 73–81
- 47 Qiu, P. *et al.* (2004) Mutation detection using Surveyor nuclease. *Biotechniques* 36, 702–707
- 48 Pimkin, M. *et al.* (2007) Recombinant nucleases CEL I from celery and SP I from spinach for mutation detection. *BMC Biotechnol.* 7, 29
- 49 Triques, K. *et al.* (2008) Mutation detection using ENDO1: application to disease diagnostics in humans and TILLING and Eco-TILLING in plants. *BMC Mol. Biol.* 9, 42
- 50 Young, L. and Dong, Q. (2004) Two-step total gene synthesis method. *Nucleic Acids Res.* 32, e59
- 51 Bang, D. and Church, G.M. (2008) Gene synthesis by circular assembly amplification. *Nat. Methods* 5, 37–39
- 52 Babon, J.J. *et al.* (1999) Mutation detection using fluorescent enzyme mismatch cleavage with T4 endonuclease VII. *Electrophoresis* 20, 1162–1170
- 53 Norberg, T. *et al.* (2001) Enzymatic mutation detection method evaluated for detection of p53 mutations in cDNA from breast cancers. *Clin. Chem.* 47, 821–828
- 54 Yeung, A.T. *et al.* (2005) Enzymatic mutation detection technologies. *Biotechniques* 38, 749–758
- 55 Silber, J.R. and Loeb, L.A. (1981) S1 nuclease does not cleave DNA at single-base mis-matches. *Biochim. Biophys. Acta* 656, 256–264
- 56 Till, B.J. *et al.* (2004) Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Res.* 32, 2632–2641
- 57 Oleykowski, C.A. *et al.* (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Res.* 26, 4597–4602
- 58 Till, B.J. *et al.* (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Res.* 13, 524–530
- 59 Colbert, T. *et al.* (2001) High-throughput screening for induced point mutations. *Plant Physiol.* 126, 480–484
- 60 Comai, L. *et al.* (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant J.* 37, 778–786
- 61 Perry, J.A. *et al.* (2003) A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. *Plant Physiol.* 131, 866–871
- 62 Bannwarth, S. *et al.* (2006) Rapid identification of unknown heteroplasmic mutations across the entire human mitochondrial genome with mismatch-specific Surveyor Nuclease. *Nat. Protoc.* 1, 2037–2047
- 63 Bannwarth, S. *et al.* (2008) Rapid identification of mitochondrial DNA (mtDNA) mutations in neuromuscular disorders by using surveyor strategy. *Mitochondrion* 8, 136–145
- 64 Janne, P.A. *et al.* (2006) A rapid and sensitive enzymatic method for epidermal growth factor receptor mutation screening. *Clin. Cancer Res.* 12, 751–758
- 65 Nowak, D. *et al.* (2006) Mutation analysis of hCDC4 in AML cells identifies a new intronic polymorphism. *Int. J. Med. Sci.* 3, 148–151
- 66 Mitani, N. *et al.* (2007) Surveyor nuclease-based detection of p53 gene mutations in haematological malignancy. *Ann. Clin. Biochem.* 44, 557–559
- 67 Voskarides, K. and Deltas, C. (2009) Screening for mutations in kidney-related genes using SURVEYOR nuclease for cleavage at heteroduplex mismatches. *J. Mol. Diagn.* 11, 311–318
- 68 Kulinski, J. *et al.* (2000) CEL I enzymatic mutation detection assay. *Biotechniques* 29, 44–46 48
- 69 Saaem, I. *et al.* (2011) Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Res.* DOI: 10.1093/nar/gkr887
- 70 Kosuri, S. *et al.* (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.* 28, 1295–1299
- 71 Gibson, D.G. *et al.* (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215–1220
- 72 Shao, Z. and Zhao, H. (2009) DNA assembler, an *in vivo* genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* 37, e16
- 73 Gibson, D.G. (2009) Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* 37, 6984–6990
- 74 Wimmer, E. *et al.* (2009) Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nat. Biotechnol.* 27, 1163–1172
- 75 Carr, P.A. and Church, G.M. (2009) Genome engineering. *Nat. Biotechnol.* 27, 1151–1162