Published online 6 August 2010

Nucleic Acids Research, 2010, Vol. 38, No. 19 ***e180***
doi:10.1093/nar/gkq677

High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides

Alex Y. Borovkov[1,*], Andrey V. Loskutov[1], Mark D. Robida[1], Kristen M. Day[1], Jose A. Cano[1], Tien Le Olson[1], Hetal Patel[1], Kevin Brown[1], Preston D. Hunter[1] and Kathryn F. Sykes[1,2,*]

[1]Center for Innovations in Medicine of the Biodesign Institute at the Arizona State University, Tempe, AZ 85287 and [2]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA

Received March 23, 2010; Revised July 19, 2010; Accepted July 21, 2010

ABSTRACT

To meet the growing demand for synthetic genes more robust, scalable and inexpensive gene assembly technologies must be developed. Here, we present a protocol for high-quality gene assembly directly from low-cost marginal-quality microarray-synthesized oligonucleotides. Significantly, we eliminated the time- and money-consuming oligonucleotide purification steps through the use of hybridization-based selection embedded in the assembly process. The protocol was tested on mixtures of up to 2000 oligonucleotides eluted directly from microarrays obtained from three different chip manufacturers. These mixtures containing <5% perfect oligos, and were used directly for assembly of 27 test genes of different sizes. Gene quality was assessed by sequencing, and their activity was tested in coupled *in vitro* transcription/translation reactions. Genes assembled from the microarray-eluted material using the new protocol matched the quality of the genes assembled from >95% pure column-synthesized oligonucleotides by the standard protocol. Both averaged only 2.7 errors/kb, and genes assembled from microarray-eluted material without clonal selection produced only 30% less protein than sequence-confirmed clones. This report represents the first demonstration of cost-efficient gene assembly from microarray-synthesized oligonucleotides. The overall cost of assembly by this method approaches 5¢ per base, making gene synthesis more affordable than traditional cloning.

INTRODUCTION

Synthetic gene production is a relatively new and fast growing sector of the multi-billion dollar market of biotechnological reagents. Fueled by the advances of post-genomic technologies, the market has been doubling annually for the past 5 years and is expected to grow at this or higher rate for the next 5–10 years (1). Existing gene assembly technologies use individually synthesized oligonucleotides (oligos) as a starting material and require clonal validation of the assembled products. These methods will be unable to support the current pace of expansion of the synthetic gene market. Development of more robust, easily scalable, time- and cost-efficient gene assembly technologies will be needed (2). Significant progress has already been made toward this goal. Assembly of genes up to 2.5 kb is now a routine process that can take <2 weeks to complete. However, assembly of large DNA molecules remains a challenge. Only a few synthetic molecules over 10 kb have been reported to date (3–6), but those can be used for pasting together much larger DNA fragments. Gibson *et al.* (7,8), for instance used them for the assembly of *Mycoplasma genitalium* (580 kb) and *M. mycoides* (1.1 Mb) genomes.

The major drawback of current gene assembly technologies is the lack of internal quality control mechanisms. This leads to their dependency on the quality of the starting oligos. To date, commercial gene assemblies are performed using only the best available column-synthesized oligos; however, even this does not guarantee flawless assembly. For instance, only one in ten 1 kb genes assembled from 50-mer oligos synthesized with 99.9% coupling efficiency would be correct. The other nine would carry sequences that were divergent from the intended sequence. The number of mistakes

*To whom correspondence should be addressed. Tel: +1 469 774 2263; Email: aborovkov@synbuild.com
Correspondence may also be addressed to Kathryn F. Sykes. Tel: +1 480 727 0859; Fax: +1 480 727 0756; Email: kathryn.sykes@asu.edu
Present addresses:
Alex Y. Borovkov, Synbuild LLC, 1833 W. Main Street, Suite 129, Mesa, AZ 85201, USA.
Jose A. Cano and Tien Le Olson, Departmento de Biotecnologia, Psicofarma, Mexico City, Mexico 14050, USA.
Hetal Patel, Kevin Brown and Preston D. Hunter, Covance Laboratories Inc., Chandler, AZ 85249, USA.

grows exponentially with the length of the product, making direct assembly of large molecules impractical. More detailed evaluation of the current status of gene assembly technologies can be found in recent reviews (2,9–11).

Heterogeneity of assembly products is commonly addressed by the clonal selection, which includes cloning, isolation, preparation and sequencing of a large number of individual assemblies in order to identify those without mistakes (12). This significantly prolongs the gene production process and contributes greatly to the cost of the final product. One way to increase frequency of flawlessly assembled products is to minimize the number of faulty oligos available for the assembly. To some degree, this can be achieved by the additional purification of synthetic oligos prior to their use. Chromatographic purification is an option available from most oligo vendors, but the process is expensive and the separation is incomplete. Much more efficient separation can be achieved by polyacrylamide gel purification, but this option is even more expensive and is not available from all vendors. This option was used by the Craig Venter Institute for assembly of the first synthetic genomes (4,7,8). To reduce the oligo purification cost, a uniform length oligo design was used to enable bulk purification of synthesized oligos. Although still expensive, the approach was shown to be very effective. When used to assemble a 5.4 kb $\phi \times 174$ genome, ~10% of the generated molecules were found to be infectious. Sequence analysis of several infectious clones revealed the presence of ~5 single nucleotide substitutions per genome, most likely generated by DNA polymerase (4).

The success of this oligo purification scheme suggests that the majority of mutations observed in synthetic gene products are caused by the incorporation of poorly synthesized oligos into the assemblies. Once prevented, it led to dramatic improvement in the resulting gene quality. Tian *et al*. (13) showed that a similar effect could be achieved by a hybridization-based oligo purification process, but the high cost and complexity of the developed protocol complicated its commercial adaptation. However, a hybridization-based selection can be employed not as a separate step, but integrated into the gene assembly process. To enable such integration some modifications of the assembly process would be required. These would include: (i) a special oligo design enabling normalization of the thermodynamic parameters of the oligo pairings; and (ii) a way to maintain these parameters unchanged throughout the assembly process to ensure its efficiency. This can be accomplished by the implementation of a T_m-normalized oligo design and by avoidance of polymerization-based steps that would alter Tm of these oligos by their extension. Under the name of ligase chain reaction (LCR), such a process was introduced over 10 years ago (14), but remained largely unnoted for ~5 years until it was successfully used for the assembly of a 714 bp green fluorescence protein (GFP) gene from T_m-normalized microarray-synthesized oligos (15). The gene assembled by LCR directly from unamplified and unpurified low-quality oligos contained only 1–6 errors/kb in contrast to 15–60 errors/kb in the same gene assembled from similarly designed and synthesized oligos by a PCR-based method (16). The LCR-based assembly demonstrated the power of hybridization-based selection, but failed to address the low-yield issue unavoidably associated with the use of microarray-synthesized oligos. This drawback limited its use to low complexity, high redundancy oligo pools.

The need to use column-synthesized oligos has become the key factor limiting throughput and cost reduction potentials for commercial gene production. The cost of these oligos has remained constant for the past 4 years and any future reduction is unlikely (11). Despite advancements in modern column (or resin)-based oligo synthesizers, which can now produce over 1500 oligos in a single run, these machines are no match to the hundreds of thousands of oligos inexpensively synthesized on a single microarray chip (see (2) for review). Today a number of companies offer custom microarray synthesis using variety of technical platforms, but only three vendors will elute the oligos from the glass surface. These are LC Sciences, which adopted the microfluidic technology developed by Zhou *et al*. (15); Combimatrix, which uses the electrochemical platform developed by Egeland and Southern (17); and Agilent Technologies, which uses the inkjet technology pioneered by Lasky and Hood (18).

In addition to the poor product quality and high pool complexity, the use of microarray-synthesized oligos for gene assembly presents an additional challenge. Yields of microarray-synthesized oligos are six orders of magnitude lower than those of column-synthesized ones and are not nearly enough to accommodate assembly of all genes designed in the pools without their prior amplification. Tian *et al*. were first to overcome these problems by developing protocols for parallel amplification and purification of complex oligo mixtures. Using these protocols, they were able to assemble all 21 genes of the *E. coli* 30S ribosomal subunit cluster from a single pool of microarray-synthesized oligos (13). The quality of the genes assembled by their protocol was 2- to 3-fold higher than those built from microarray oligos that had been purified by polyacrylamide gel electrophoresis, and ~10-fold higher than those assembled without purification.

Although other examples of successful gene assembly from microarray-synthesized oligos exist (13,15,16,19) the broad adoption of the protocols is tempered by their added cost and technical complexity. The goal of this study was not to add to the list of examples, but to develop a less expensive and less complex alternative that would make microarray-synthesized oligos feasible for commercial gene production.

MATERIALS AND METHODS

Oligonucleotides

All T_m-normalized oligos were designed by the oligo design algorithm described here with T_m 55 ± 1°C calculated by the Meinkoth and Wahl formula (20) adjusted for a typical ligase/polymerase buffer ($T_m = 61 + 0.41 \times (\%G + \%C) - 675/n$). All column-made

oligos were ordered desalted from Invitrogen. The microarray-synthesized oligos were ordered from LC Sciences, Agilent Technologies or Combimatrix. All microarray-synthesized oligos in the amount ~10 pmol/pool were desalted on Biospin P6 mini-columns (BioRad) by the protocol recommended by the manufacturer. The desalted mixture was divided into two equal ~5 pM aliquots. One was stored at −80°C as a backup, and another used for block assembly. The aliquots of 5′ phosphorylated oligos from Invitrogen and LC Sciences were used without any pre-treatment. The aliquots of oligos from Agilent Technologies were 5′ phosphorylated in 100µl of 1× OptiKinase buffer supplemented with 1 mM ATP and 10 µ of OptiKinase (USB) by 30 min incubation at 37°C followed by 15 min incubation at 65°C for heat inactivation of the enzyme. The aliquots of 3′ phosphorylated oligos from Combimatrix were simultaneously 3′ dephosphorylated and 5′ phosphorylated in 100 µl of 1× T4 polynucleotide kinase buffer supplemented with 1 mM ATP and 10 µ of T4 polynucleotide kinase (NEB) by 30 min incubation at 37°C followed by 15 min incubation at 65°C for heat inactivation of the enzyme. The generated pools of 5′ phosphorylated oligos were precipitated by adjusting the total volume to 100 µl and adding 1 µl of glycogen (20 mg/ml), 10 µl of 3 M Na-acetate, pH 5.2, and 200 µl of acetone. The mixtures were incubated at −80°C for at least an hour, and then centrifuged at 16 000*g* for 10 min. The supernatants were discarded and the pellets washed several times with 70% ethanol, dried and re-dissolved in 5 µl of water.

Block assembly and amplification

The 5 µl aliquots of the acetone-precipitated oligos were mixed with 5 µl of 2× *Taq* DNA ligase buffer supplemented with 40% PEG6000 and 20 µ of *Taq* DNA ligase (NEB). The mixture was heated to 95°C, incubated for 2 min, then slowly cooled (30 min) down to 58°C and incubated at this temperature for 12 h. Then 90 µl of 1.1× iProof buffer supplemented with 1 mM dNTP and 2 µ iProof HF DNA polymerase (BioRad) was added directly to the ligation reaction and the resulting mixture was subjected to 10 cycles of polymerase cycling assembly (PCA, also known as recursive PCR), with incubations at 98°C for 10 s, 58'C for 30 s and 72°C for 15 s. Next, the PCA products were used for block amplification. The reactions were performed in 30 µl of 1× iProof buffer supplemented with 1 mM dNTP, 500 nM of block-specific primer, 0.6 µ iProof HF DNA polymerase and a 1 µl PCA aliquot used as a template. A and H blocks were amplified independently by 35 cycles of incubations at 98°C for 10 s, 58°C for 30 s, 72°C for 15 s. A blocks were generated with the A-specific primer (5′-GCAGCG TCTGGAGTCTCCTC-3′) and H blocks with the H-specific primer (5′-AGCAGCTCTCAGAGTCTT TTC-3′). The amplified A and H blocks were agarose gel purified (Qiagen), quantified and mixed together at a 1 : 1 ratio. The resulting block mixture (600 ng) was digested in 50 µl of 1× NEBuffer 4 with 10 µ *Mly*I (NEB) at 37°C for 3 h and purified from the cleaved flanking regions by a DNA purification kit (Qiagen).

Assembly of the gene amplification template and individual gene amplification

The mixture of flank-trimmed blocks was assembled into a gene amplification template by an overlapping PCR. The reaction was set in 30 µl of 1× iProof buffer with 1 mM dNTP, 2.5 ng of *Mly*I digested blocks and 0.6 µ iProof HF DNA polymerase, and performed by 8 cycles of incubations at 98°C for 15 s, 55°C for 28 min, 72°C for 1 min. The resulting gene template was used for individual amplification of each of the genes designed in the pool. The individual reactions were set up in 50 µl of 1× iProof buffer supplemented with 500 nM of gene-specific primers, 1 µ iProof HF DNA polymerase and 1 µl aliquots of the generated gene template. Genes were amplified by 35 cycles of 98°C for 10 s, 55°C for 30 s, 72°C and the amplified products were either used directly or cloned into pCR-Blunt II-TOPO (Invitrogen) vector for further evaluation.

In vitro translation and protein detection

Linear expression elements (LEEs) for expression in coupled *in vitro* transcription/translation (IVT) system (Invitrogen) were assembled by extension of the gene assembly products with T7 promoter and terminator regions taken from pEXP5 vector (Invitrogen) by overlapping PCR (21). One micrograms of the PCR-purified LEE was expressed in a 50 µl IVT reaction supplemented with 1 µl of S^{35} labeled methionine (PerkinElmer) and assembled according the manufacturer's recommendation. The reaction was performed for 4 h at 37°C with continuous agitation. Five microlitres of the IVT reaction was loaded per well on a pre-cast 4–12% polyacrylamide gradient gel (BioRad), electrophoretically separated and autoradiographed.

Special precautions

The high sensitivity of the developed protocols requires extra attention to potential contamination and cross contamination of samples when working with microarray-synthesized oligos. Setup of the PCR workstations should be in compliance with the forensic science good laboratory practice (GLP) (22) and only barrier tips that prevent DNA cross contamination should be used (www.mbpinc.com/html/pdf/techreport/TechReport_313.pdf).

RESULTS AND DISCUSSION

The synthetic gene assembly approach presented here is an adaptation of previously developed techniques that enable high-quality gene assembly from microarray-synthesized oligos. We were able to combine the ideas of hybridization-based oligo selection (13,15) and parallel amplification (13,19) into a single process that allows simple and cost-efficient production of unlimited amounts of high-quality building materials directly from unpurified pools of microarray-synthesized oligos. For them to work together, we reversed the typical order in

which the quality and yield issues are addressed: the oligo selection occurs first and the template amplification is performed second (schematic shown in Figure 1). This enabled the coupling of the oligo selection process with the assembly of intermediate double-stranded units, which we call blocks. Once assembled, blocks are co-amplified as a single pool, which is then used as a template for assembly of individual genes.

To enable gene assembly from the pools of amplified blocks, adjoining blocks were designed with short unique regions of overlap, which thereby defined their relative positions and orientations. Due to these overlapping regions, parallel amplification of all intended blocks in a single reaction is not possible, as amplification would result in the production of unconnected, internally primed, partial blocks. To avoid this complication, adjacent blocks are designed with alternatively distinct flanking sequences. We numbered the blocks as odd and even based on their flanking sequence, and amplified them in separate reactions with appropriate pairs of block amplification primers.

T_m-normalized oligo design and gene recoding

T_m-normalized oligos were designed by an algorithm developed in-house and based on the Meinkoth and Wahl formula adjusted for a typical ligase/polymerase buffer ($T_m = 61 + 0.41 \times (\%G + \%C) - 675/n$) (20). Oligo design can be performed on individual genes or on a set of genes concatenated into a virtual sequence to simplify uniformed block design. The algorithm tracks the sequence along its length and breaks it into a set of adjoining fragments having identical T_m values. In the experiments described here the T_m value was set at $55 \pm 1°C$. The identified fragments served as halves of the predicted oligos. Oligos for the upper strand were designed by combining sequences of the T_m-normalized fragments 1 with 2, 3 with 4, 5 with 6, etc., and designed for the lower strand by combining the sequences complimentary to the fragments 2 with 3, 4 with 5, 6 with 7, etc. Together they represented a thermodynamically uniformed set of gene assembling oligos (GAOs).

Although this algorithm is suitable for designing T_m-normalized oligos for any type of DNA molecule, we took advantage of the fact that to date protein-encoding sequences are in the highest demand. This enabled us to use the redundancy of the amino acid translation table to improve the accuracy and robustness of the assembling process by disrupting GC- and AT-rich islands, direct and inverted repeats, microsatellites and other undesirable DNA features. Using the codon frequency database, maintained by Kazusa DNA Research Institute, we identified a set of 12 codons rarely used by either the *E. coli* or mammalian translation machineries. To evaluate the accuracy of our selection, we took six rarely used by mammalian codons and used those to generate 'choke' points in a firefly luciferase gene (LUC) optimized for mammalian expression. This commercially available

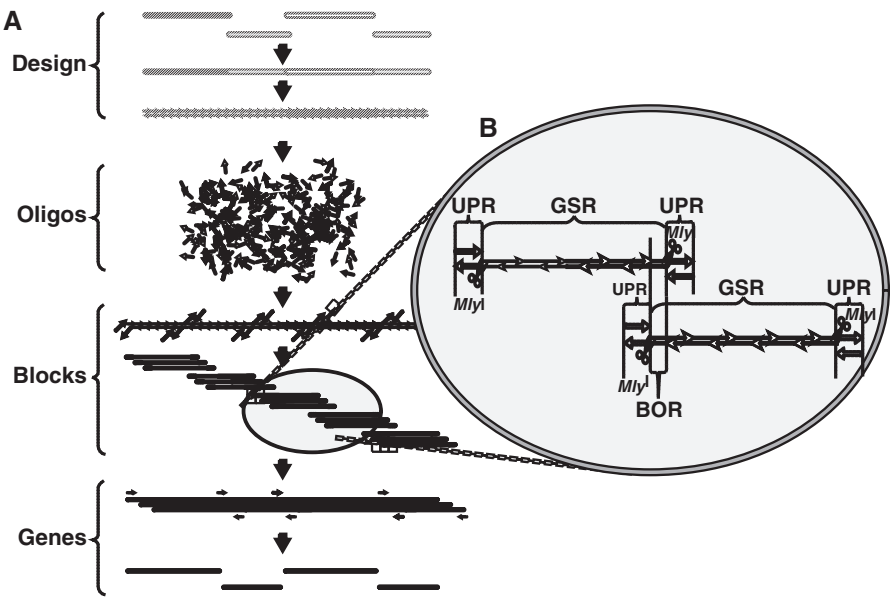

Figure 1. Block-based gene assembly. (**A**) Schematics of the process. Design: All genes in the set are combined into a virtual string and the resulting sequence is T_m-normalized. For simplicity, a set of four color-coded genes is shown. Oligos: T_m-normalized oligos are designed, synthesized and pooled into a single mixture. Blocks: a pool of oligos is converted into a mixture of partially overlapping blocks of uniform size. Blocks are assembled under tightly controlled, highly stringent conditions that prevent incorporation of improperly synthesized oligos. Assembled blocks are amplified to generate sufficient amount of material for the amplification of all individual genes designed to be built from the oligo pool. Genes: the amplified blocks are trimmed of their flanking regions and combined into a single sequence by several rounds of overlapping PCR. The assembled sequence is used as a template for amplification of all genes in the set, each with its own set of gene-specific primers. (**B**) Detailed diagram of block structure. Blocks are composed of unique gene-specific regions (GSR) built from GAOs and common flanking regions (UPR), which enables their co-amplification. UPRs carry *Mly*I recognition sites, which are used for UPR cleavage. Two distinct UPR sequences are used in alternation from block to block. Adjacent blocks share regions of overlap (BOR) enabling their covalent joining by overlapping PCR.

LUC+ gene (23) was further modified to carry only the most frequent mammalian codons (Max LUC). The 'choke' variants were generated from Max LUC by swapping the ~500 bp region between unique *Bst*BI and *Eco*RV restriction sites with modified 500 bp fragments. Ala, Arg, Leu, Pro, Ser and Thr variants were generated by replacing the corresponding codons in the region with their rare alternatives: GCG, CGT, CTA, CCG, TCG or ACG, respectively. Evaluation of luciferase activity in murine cells transfected with these variants revealed that replacement of Leu, Ser or Thr codons caused a statistically significant ~4-fold reduction in gene expression, whereas replacement of Ala, Arg, Pro codons had only marginal effects (Supplementary Figure S1). Based on this evaluation, three of the above codons (GCG for Ala, CGT for Arg and CCG for Pro) were removed from the original list, leaving the final list of codons to be avoided for an efficient gene expression in *E. coli* and mammalian systems with nine codons: four for Arg (AGG, AGA, CGG and CGA), two for Leu (TTA and CTA), and one for each of Ile (ATA), Ser (TCG) and Thr (ACG). These were excluded from usage to ensure adequate expression of the recoded genes in either system. All the remaining synonymous codons were used interchangeably to modify gene sequences as desired to facilitate assembly, without altering the amino acids of the encoded proteins.

Gene recoding was completed by performing several re-iterative rounds of intermediate recoding using an in-house program followed by output analysis. During the first round, the program replaced all codons specified as rare with randomly selected synonymous codons. In the second, it removed restriction sites for the list of enzyme specified by the operator. In the third, it disrupted homopolymeric regions longer than 3 nt and repeat units over 6 nt. And finally, the sequence was normalized for purine/pyrimidine content within a 21-bp-long sliding window. The resulting sequence was checked for all specified criteria; if any of them was not met, the sequence was sent back for another recoding attempt. If criteria were met, the sequence was fed into the oligo design program and the predicted oligo set was checked for unique pairings. If uniqueness was not confirmed, then the sequence was sent for recoding once again. Each of the output sequences carried *E. coli* and mammalian codon adaptation indexes (CAIs) (24) above 0.8, sufficient for efficient protein expression in either system (25). The source codes of the developed recoding and oligo design programs are provided in Supplementary Material. Online versions of are available at http://www.innovationsinmedicine.org/software/Tm-normalized-oligo-design/.

We tested our recoding algorithm on the *Plasmodium falciparum* merozoite surface protein (MSP) gene, which is known to be poorly expressed in murine cells (26). In addition to the wild type (WT) gene (mammalian CAI 0.25), two recoded versions were constructed. One by recoding the gene using a conventional codon optimization protocol (OPT) and another using the sequence normalization algorithm described here (NRM). The calculated mammalian CAI's of OPT and NRM variants were 0.99 and 0.85, respectively. The relative gene activities of the three constructed versions were assessed by (i) determination of MSP produced in transfected NIH 3T3 cell cultures by immunoblot and (ii) measurement of antigen-specific titers in sera of immunized mice by ELISA. The WT gene produced no detectable protein in transfected murine cells and stimulated very low antibody titers in mice. In contrast, both recoded versions expressed similarly robust levels of protein in cell culture and stimulated similarly strong immune responses in mice (Figure 2). Based on this study, we concluded that robust gene expression can be achieved without maximizing high-frequency codon usage. We were able to achieve comparable improvements in protein expression by avoiding only a small number of rare codons.

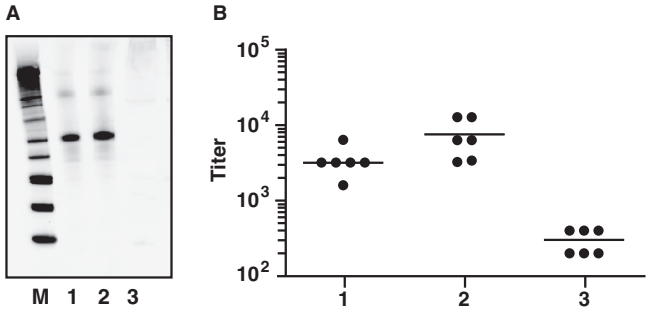

Figure 2. Effect of sequence normalization on gene expression. Activity of the native (WT) and two recoded versions of the malaria MSP gene (NRM and OPT, see text for details) were compared by the levels of their expression in NIH 3T3 cell cultures (**A**) and by the strength of the immune respond in genetically immunized mice (**B**). (A) Nearly confluent NIH 3T3 cell cultures were transfected with the test constructs using lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. After transfection, cells were incubated for 12 h at 37°C in a CO_2 incubator, and then harvested and lysed. The SDS soluble fractions were analyzed by immunoblot using in-house generated a MSP mouse sera (1 : 1000) and goat HRP-conjugated anti-mouse IgG + IgM (H + L) (1 : 10 000) purchased from KPL. Lane M: MagicMark XP chemiluminescent molecular weight marker (Invitrogen); Lanes 1–3: lysate of cells transfected with OPT, NRM, and WT MSP variants, respectively. (B) Three groups (6 mice/group) of 2-month old BALB/c mice were genetically immunized with a plasmid expressing one of the MSP versions (12 μg) delivered into the ear pinna at Weeks 0, 2 and 4. Two weeks after the last immunization, the animals were bled and the sera assayed for MSP antigen reactivity by ELISA using MSP-coated plates and goat HRP-conjugated anti-mouse IgG + IgM (H + L) (10 000) purchased from KPL. Individual mouse titers are plotted on a log scale as dots and group averages are represented by lines. 1: MSP OPT; 2: MSP NRM; and 3: MSP WT-immunized mice.

Enzyme and oligo pool size selections

Since microarray-synthesis provides only femtomolar amounts of oligos, the ability to use them without amplification relies on the substrate sensitivity of the enzyme used. Five commercial enzymes [*Taq* DNA polymerase (Promega), iProof (Bio-Rad), *Pfu* Ultra II (Stratagene), HiFi (Roche) and *Pf* × 50 (Invitrogen)] were tested for their ability to utilize oligo substrates at sub-nanomole concentrations. *Taq* DNA polymerase showed the highest robustness and specificity, followed by iProof, *Pfu* Ultra II and HiFi. *Pf* × 50 failed to produce any

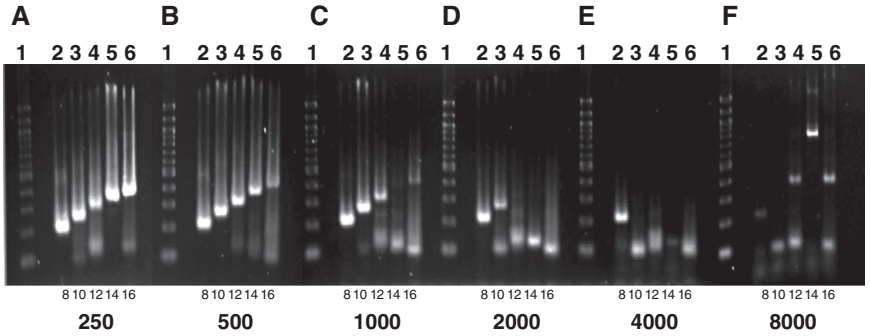

Figure 3. Effect of oligo pool complexity on the robustness of gene assembly. Five sets of GAOs designed for the assembling of five different test genes comprised of 8, 10, 12, 14 or 16 oligos were mixed with increasing numbers of irrelevant but similarly designed GAOs to create oligo pools with 250, 500, 1000, 2000, 4000 and 8000 complexities. The pools were diluted to 1 nM/oligo and used for assembling the test genes by 10 cycles of PCA followed by 35 cycles of a standard PCR as described in 'Materials and Methods' section. The expected 220, 270, 320, 370 and 420 bp products were separated by agarose gel electrophoresis and visualized by ethidium bromide staining. Products assembled from the 250, 500, 1000, 2000, 4000 and 8000 GAO complex pools are presented in (**A**), (**B**), (**C**), (**D**), (**E**) and (**F**), respectively. Lane 1—100 bp DNA ladder (NEB); lane 2—gene 1 (220 bp); line 3—gene 2 (270 bp); lane 4—gene 3 (320 bp); lane 5—gene 4 (370 bp); and lane 6—gene 5 (420 bp).

detectable product even at the highest concentration of substrate tested (100 pM) (Supplementary Figure S2). Despite the robustness of *Taq* DNA polymerase, we selected iProof for use in the subsequent experiments, since it was the only high-fidelity enzyme capable of utilizing substrate at picomole concentrations.

Selection of the enzyme for the ligation step was limited to *Taq* ligase, the only commercially available thermostable DNA ligase, which unfortunately is highly sensitive to substrate concentration (27). However, by supplementing the reaction with 20% PEG6000 and increasing its duration to over 6 h, we were able to use it with nanomole concentrations of oligos (Supplementary Figures S3 and S4).

To define the limits of oligo pool complexity for successful gene assembly, we selected five test genes of 220, 270, 320, 370 and 420 bp, comprised of 8, 10, 12, 14 or 16 GAOs, respectively. Six pools with complexities increasing from 250 to 8000 oligos were made by mixing column-synthesized GAOs designed for these test genes with those corresponding to other, irrelevant genes. Using a standard gene building protocol (28), all five test genes were successfully assembled from the pools comprised of 1000 or fewer oligos. The two smaller genes were assembled from the pool of 2000 oligos and the smallest gene was generated from the pools of 4000 and 8000 oligos. However, the 8000-complex pool also gave rise to the formation of unintended products, suggesting loss of pairing specificity at this level of pool complexity (Figure 3). Based on these results, we concluded that to ensure the accuracy and robustness of gene assembly pool complexities should be limited to 4000 oligos, and the block sizes limited to ~270 bp.

Block structure

The amount of material synthesized on a microarray chip is not adequate for the assembly of all genes designed on it. Amplification of this material at some point is unavoidable, and for the reasons discussed below we choose to do it on the block rather than on the oligo level. A structural diagram of the block-based gene assembly process and detailed block structure are shown in Figure 1. Each block is comprised of a unique core built of GAOs and common flanking regions enabling co-amplification of blocks sharing common flanking sequences. To enable removal of these regions after block amplification, they contained embedded recognition sites for a Type IIS restriction enzyme (*Mly*I) that accommodated their cleavage. The adjacent blocks were designed with short areas of overlap so that they could be combined into longer fragments by overlapping PCR (29) or some other covalent DNA attachment method. To enable parallel assembly and amplification of the entire plurality of blocks from a single oligo pool, adjoining blocks were designed with alternating flanking sequences. By alternating these common end sequences, the number of different block-flank categories was limited to two. Use of different flanking regions on adjacent blocks is essential for amplification of all blocks from a single oligo mixture. Without alternation of common flanks, an attempt to co-amplify two adjacent blocks would result in a partial block amplification of only their overlapping regions.

To evaluate the effect of flanking sequences on the specificity and robustness of block amplification and efficiency of *Mly*I cleavage, we compared the performance of 12 randomly designed flanking sequences and did not find any significant differences between them by these criteria (Supplementary Table S1). However, for the purpose of consistency all experiments in this study were performed using adapter sequences GCAGCGTCTGGA GTCTCCTC and AGCAGCTCTCAGAGTCTTTTC designated as A and H, respectively.

Block assembly and amplification

The initial DNA ligase step of block assembly requires oligos with 5′ phosphate and 3′ hydroxyl groups. Depending on the vendor, oligos were delivered in one of three formats: 5′-P/3′-OH, 5′-OH/3′-OH or 5′-OH/3′-P. Oligo mixtures in the 5′-P/3′-OH format were used without modifications; those delivered in the 5′-OH/3′-OH format were 5′ phosphorylated with OptiKinase; and

those delivered in the 5′-OH/3′-P format were 3′ dephosphorylated and 5′ phosphorylated with T4 polynucleotide kinase. Once the ends of the oligos were properly modified they were used in the DNA ligase-driven assembling reaction coupled with a highly stringent oligo selection process. We calculated that at the concentrations that are achievable with microarray-synthesized oligos proper pairing in 1× *Taq* DNA ligase buffer would take over 140 h (Supplementary Material). This rendered LCR, with its constantly interrupted short annealing steps, impractical and argued in favor of a reaction performed at a constant temperature that would not disrupt the oligo pairings. To shorten the reaction, time ligation buffer was supplemented with PEG6000. This resulted in 20-fold acceleration of the oligo pairing process, reducing the reaction time to 7 h. The reaction was performed at a temperature 3°C over the T_m of the oligos since an analysis of single nucleotide mismatch association kinetics (30) suggested that this temperature is optimal for achieving maximal stringency of the oligo selection process (Supplementary Figure S5). Our own data showed that raising the reaction temperature from 55°C to 58°C did not have a noticeable effect on the product yield (Supplementary Figure S6).

Next, the relatively short fragments generated by the ligase reaction were extended by 10 cycles of PCA. This step, although not essential, contributed significantly to the size of the fragments that could be assembled from the generated template and to the robustness of block amplification (Supplementary Figure S7). The PCA-generated mixture was split into two aliquots. One aliquot was used for amplification of the A-flanked blocks using the A-block specific primer, whereas the other was used for amplification of the H-flanked blocks using the H-block specific primer. Although parallel unbiased amplification of far more complex mixtures of much larger DNA fragments has been shown earlier (31) we performed our own evaluation. To test amplification bias, we followed 80 randomly selected blocks within a 252 block mixture generated from a 2000-complex oligo pool. We found all 80 blocks equally represented in the generated mixture (Supplementary Figure S8a). To assess the possibility of using the generated block mixture as a renewable source of the gene building material, the block mixture was re-evaluated after two rounds of re-amplification. Seventy eight blocks were still represented evenly in the mixture, but the presence of the other two blocks was significantly diminished (Supplementary Figure S8b). This suggests that block re-amplification should be limited to 1–2 rounds to avoid reduction of pool complexity.

Assembling blocks into a gene amplification template

In preparation for their use in gene assembly, the co-amplified blocks were treated with *Mly*I to remove the flanking regions and then assembled into long continuous DNA fragments by several cycles of PCA. The resulting mixture was used as a template for amplification of all individual genes designed in the pool. Optimization of PCA conditions was performed on the 252-complex block mixture described above assembled from microarray oligos synthesized by Agilent. Success of gene assembly by 35 cycles of a standard PCR was used as a criterion for evaluating the quality of the generated template. We found that block concentration, length of the annealing step, and the number of cycles were the most critical parameters defining assembly of an optimal gene amplification template (Figure 4). The reaction should contain at least 20 fmol of each individual block. The quality of the generated template gradually increased with extension of the annealing time and was found to be highest at 28 min. Assembly of the template required only 6 cycles of amplification. Additional cycling was not useful, and on the contrary, appeared to reduce the quality of the generated template.

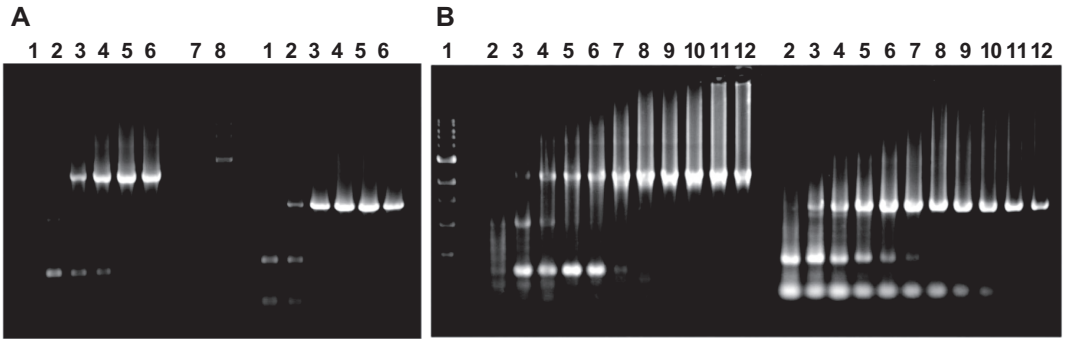

Figure 4. Optimizing assembly of blocks into template for gene production. (**A**) Evaluating effect of annealing time. A 252-element mixture of *Mly*I cleaved blocks (20 fmol/block) assembled from a 2000-complex pool of microarray-synthesized oligos, was subjected to five cycles of PCA with increasing annealing times from 0 to 35 min. To evaluate robustness of the PCA-generated template 1 µl aliquots of the reactions were used for amplification of 2.1 and 1.3 kb genes with pairs of gene-specific primers. The amplified products were separated by agarose gel electrophoresis and visualized by ethidium bromide staining. Lanes 1–6—products amplified from the templates generated by the reactions that carried, respectively, 0, 2, 9, 18, 28 and 35-min long annealing steps; lane 7—100 bp DNA ladder (NEB); and lane 8–1 kb DNA ladder (NEB). (**B**) Evaluating effect of PCR cycle number. The block mixture described in (**A**) was subjected to an increasing number of PCA cycles performed with a 25-min annealing step. Robustness of the generated templates was evaluated as described in the (**A**). Lane 1—1 kb DNA ladder (NEB); and lanes 2–12—products amplified from the templates assembled using 0 to 10 PCA cycles, respectively.

Table 1. Error types and frequencies in the genes assembled by the regular and the block-based protocols from the column- and microarray-synthesized oligos

Protocol	Oligos		Type of mutations per kb			Total
	Design	Source	Insertions	Deletions	Substitutions	
Regular[a]	Length normalized	Various (>100 Mb)	0.1	2.5	0.2	2.8
Block based[b]	T_m-normalized	Invitrogen (~100 kb)	0.1 (13[c])	2.2 (227)	0.3 (25)	2.7 (265)
		LC Sciences (~100 kb)	0.2 (17)	2.3 (232)	0.3 (32)	2.8 (281)
		Agilent (~50 kb)	0.1 (3)	2.4 (117)	0.3 (14)	2.7 (134)
		Combimatrix (~10 kb)	0.1 (1)	2.2 (22)	0.6 (6)	2.9 (29)

[a]Represent a summary of data collected by authors over years from analysis of 600–800-bp-long assemblies representing entire proteomes of several bacterial (*Bacillus anthracis*, *Chlamydia pneumoniae*) and viral (CPV, HSV) virus genomes built from 60-mer oligos synthesized by Invitrogen, Sigma and Macrogenics by the most commonly used protocol (28).
[b]Represent a summary of data collected during this study from the genes assembled by the described protocol from several independent T_m-normalizedmicroarray oligo pools (three from LC Sciences, two from Agilent and one from Combimatrix) and one artificial pool of column-made oligos (Invitrogen).
[c]Numbers in the parentheses represent the actual numbers of observed mutations.

Evaluation of gene quality

To evaluate efficiency of the introduced hybridization-based oligo selection process and the quality of the resulting products, we cloned and sequenced genes generated by this protocol from the pools of T_m-normalized microarray oligos synthesized by three independent vendors (LC Sciences, Agilent, Combimatrix) and from an identical mixture of the adequately diluted T_m-normalized column-synthesized oligos (Invitrogen). The results of this analysis and accumulated data on the quality of products assembled by the standard protocol from length-normalized (60-mer) column-made oligos (Invitrogen, Sigma, Qiagen, Macrogenics) are summarized in Table 1. All evaluated samples showed similar types and frequencies of errors regardless of the source of the oligos used for their assembly. We interpret these results as an indication that the coupled selection/assembly process presented here is effective. Incorporation of an oligo into an assembling molecule is no longer based on its availability in the mixture, but rather on the homology of its sequence to the target. This relaxes dependency of product quality on the quality of the starting oligos, and enables high-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligos. Regardless of the oligo source the genes assembled by the protocol described here contained 2.7–2.9 errors/kb, mostly represented by single-nucleotide deletions, compared to 2.8 errors/kb observed in the genes assembled from high-quality length-normalized column-synthesized oligos by standard protocols.

For 6 of the 27 genes assembled from one of the microarray oligo pools, we identified flawless clonal equivalents. These were used as reference standards for the evaluation of translation activity of the gene assemblies used directly without clonal selection. Each gene was assembled in two forms: one directly from the PCR-generated product of gene assembly, and another using the identified clonal equivalent as a template. All were flanked with T7 promoter and terminator and expressed in a coupled bacterial *in vitro* transcription/translation system (Invitrogen). Each pair expressed similar levels of polypeptide of the predicted molecular weight (Figure 5). In a separate experiment performed similarly, but on a different set of genes, we estimated the quantites of the generated polypeptides by measuring radioactivity of the TCA-insoluble fractions (Supplementary Table S2). Genes assembled directly from the PCR-generated templates averaged polypeptide yields of 30% less than those assembled from the sequence-confirmed templates, consistent with the observed error frequencies.

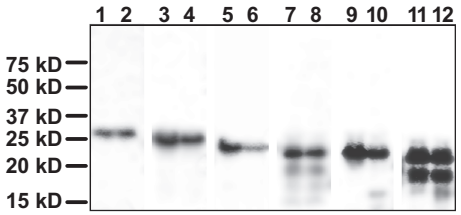

Figure 5. *In vitro* expression of the synthetic genes. PCR-generated gene assemblies and their cloned, sequence-confirmed equivalents were assembled into T7 promoter-driven expression cassettes and expressed in a coupled *in vitro* transcription/translation system. The ^{35}S–methionine labeled products were separated by polyacrylamide gel electrophoresis and detected by autoradiography. All genes were evaluated in pairs, although not all pairs were fractionated on the same gel. The picture is a composite of gels normalized relative to a common precision plus WesternC standard (Bio-Rad). Odd lanes—products generated by the cassettes assembled from the sequence confirmed clones; even lanes—products generated by the cassettes assembled directly from PCR-generated gene assemblies. Lanes 1 and 2—gene CA1; lanes 3 and 4—gene ENO1L1; lanes 5 and 6—gene LIN7A; lanes 7 and 8—gene PY00523; lanes 9 and 10—gene PY07543; and lanes 11 and 12—gene PY04421.

CONCLUSIONS

The protocol presented here eliminates the expensive and technically challenging oligo purification steps and makes the use of microarray-synthesized oligos feasible, offering a 100-fold material cost reduction. However, microarray-synthesized oligos present other challenges besides the low quality. They are available only in the

form of highly complex pools and in the amounts that are too low for use with the standard protocols. To date, several attempts have been made to adapt this source of oligos for gene assembly (13,15,16,19), but complexity of the protocols outweighed the practical benefits.

Our protocol is free from any oligo amplification or purification steps. Instead of physically removing poorly made oligos from the mixture, we used a highly stringent hybridization approach based on a T_m-normalized oligo design that prevented their incorporation into the assemblies. By avoiding the oligo amplification step, commonly used to increase the amount of gene building material, but unavoidably converting them into a double-stranded format, we were able to couple the oligo selection and gene assembly into a single process. The issue of the low amounts of microarray-synthesized oligos was addressed latter in the process by co-amplification of uniformly sized DNA fragments, rather than the oligos themselves. Block amplification had the added benefit of reducing the risk of biased amplification by reducing the complexity of the amplified mixture 10-fold. Furthermore, the generated blocks can be evaluated for perfection and used as a renewable source of error-free gene assembly material (12).

The oligo design algorithm we developed is applicable to all types of DNA sequences. However, when used to design protein-encoding genes it can be combined with the removal of rare codons, direct and inverted repeats, homopolymeric regions, GC or AT islands and other sequence abnormalities that may complicate the gene assembly process or reduce its efficiency. This method of recoding allows control of the specificity of oligo pairing and reduces the length of the oligos that simplifies their synthesis and reduces their cost.

However, one feature of this protocol still needs attention. This is the risk of cross-contamination upon reuse of the block amplification primers. So far, we were able to avoid it by strictly following GLP regulations similar to those used in forensic labs. To eliminate this risk entirely, we are currently modifying our oligo design program to design a new set of flanking regions for every newly designed oligo pool. We have shown that the yields and specificity of the block amplification step is not limited to a specific set of primers, suggesting an endless number of sequences are possible for them.

The genes constructed from microarray-synthesized oligos using our block-based gene assembly protocol were found to be indistinguishable from genes assembled by the standard protocol from the high-quality column-synthesized oligos. Sampling oligo pools from three different vendors, we demonstrated the robustness and cost-efficiency of the new method. It allows parallel assembly of several dozens of average size genes from a single oligo mixture purchased for less than a thousand dollars. Depending on the pool complexity, this reduces the material cost to a range from 1.5¢ to 0.5¢ per base and the total cost of gene assembly to from 5¢ to 3¢/bp. The entire process takes only a few days and is easily scalable. Even without the implementation of robotics, a single operator can produce several hundred genes per week. The high fidelity of the process makes the assembled genes usable without cloning. If accepted by the customer, the non-clonal gene format would contribute significantly to further cost reduction, and fast and cost-efficient production of large numbers of genes for a wide variety of research programs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Dr. Stephen Albert Johnston for providing support and suggestions, and to the management and scientists of Agilent Technologies, Combimatrix, Invitrogen and LC Sciences for helpful discussions and providing us with the synthetic material used in these experiments. We would also like to thank Elizabeth Lambert, Derek Stadie, Corina Prieto, Hiro Kitahara, Jenny Sanchez, Sachi Miyajima and Haroon Saleem for technical assistance.

FUNDING

National Institute of Health (grant number P01AI056295); Arizona Technology Research Initiative Fund. Funding for the open access charge: Arizona State University.

Conflict of interest statement. None declared.

REFERENCES

1. Newcomb,J. (2007) The new age of biological engineering: Implications for the US economy. *Ind. Biotechnol.*, **3**, 325–332.
2. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. BioSyst.*, **5**, 714–722.
3. Cello,J., Paul,A.V. and Wimmer,E. (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, **297**, 1016–1018.
4. Smith,H.O., Hutchison,C.A. III, Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: ϕX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl Acad. Sci. USA*, **100**, 15440–15445.
5. 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**, 15573–15578.
6. Chan,L.Y., Kosuri,S. and Endy,D. (2005) Refactoring bacteriophage T7. *Mol. Syst. Biol.*, **1**, 1–10.
7. Gibson,D.G., Benders,G.A., Andrews-Pfannkoch,C., Denisova,E.A., Baden-Tillson,H., Zaveri,J., Stockwell,T.B., Brownley,A., Thomas,D.W. and Algire,M.A. (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*, **319**, 1215–1220.
8. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.-Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M. *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329**, 52–56.
9. Xiong,A.S., Peng,R.H., Zhuang,J., Liu,J.G., Gao,F., Chen,J.M., Cheng,Z.M. and Yao,Q.H. (2007) Non-polymerase-cycling-assembly-based chemical gene synthesis: strategies, methods, and progress. *Biotechnol. Adv.*, **26**, 121–134.
10. Czar,M.J., Anderson,J.C., Bader,J.S. and Peccoud,J. (2009) Gene synthesis demystified. *Trends Biotechnol.*, **27**, 63–72.

11. Mueller,S., Coleman,J.R. and Wimmer,E. (2009) Putting synthesis into biology: a viral view of genetic engineering through de novo gene and genome synthesis. *Chem. Biol.*, **16**, 337–347.
12. Larsen,L.S.Z., Wassman,C.D., Hatfield,G.W. and Lathrop,R.H. (2008) Computationally optimised DNA assembly of synthetic genes. *Int. J. Bioinform. Res. Appl.*, **4**, 324–336.
13. 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**, 1050–1054.
14. Au,L.C., Yang,F.Y., Yang,W.J., Lo,S.H. and Kao,C.F. (1998) Gene synthesis by a LCR-based approach: high-level production of Leptin-L54 using synthetic gene in Escherichia coli. *Biochem. Biophys. Res. Commun.*, **248**, 200–203.
15. Zhou,X., Cai,S., Hong,A., You,Q., Yu,P., Sheng,N., Srivannavit,O., Muranjan,S., Rouillard,J.M., Xia,Y. *et al.* (2004) Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.*, **32**, 5409–5417.
16. Kim,C., Kaysen,J., Richmond,K., Rodesch,M., Binkowski,B., Chu,L., Li,M., Heinrich,K., Blair,S. and Belshaw,P. (2006) Progress in gene assembly from a MAS-driven DNA microarray. *Microelectronic Eng.*, **83**, 1613–1616.
17. Egeland,R.D. and Southern,E.M. (2005) Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. *Nucleic Acids Res.*, **33**, e125.
18. Lausted,C., Dahl,T., Warren,C., King,K., Smith,K., Johnson,M., Saleem,R., Aitchison,J., Hood,L. and Lasky,S. (2004) POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. *Genome Biol.*, **5**, R58.
19. Richmond,K.E., Li,M.H., Rodesch,M.J., Patel,M., Lowe,A.M., Kim,C., Chu,L.L., Venkataramaian,N., Flickinger,S.F. and Kaysen,J. (2004) Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.*, **32**, 5011–5018.
20. Meinkoth,J. and Wahl,G. (1984) Hybridization of nucleic acids immobilized on solid supports. *Anal. Biochem.*, **138**, 267–284.
21. Liang,X., Teng,A., Braun,D.M., Felgner,J., Wang,Y., Baker,S.I., Chen,S., Zelphati,O. and Felgner,P.L. (2002) Transcriptionally active polymerase chain reaction (TAP) high throughput gene expression using genome sequence data. *J. Biol. Chem.*, **277**, 3593–3598.
22. Mifflin,T.E. (2003) Setting up a PCR laboratory. In *PCR primer: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 5–14.
23. Sherf,B.A. and Wood,K.V. (1994) Firefly luciferase engineered for improved genetic reporting. *Promega Notes*, **49**, 14–21.
24. Sharp,P.M. and Li,W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.
25. Henry,I. and Sharp,P.M. (2007) Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.*, **24**, 10–12.
26. Narum,D.L., Kumar,S., Rogers,W.O., Fuhrmann,S.R., Liang,H., Oakley,M., Taye,A., Sim,B.K. and Hoffman,S.L. (2001) Codon optimization of gene fragments encoding Plasmodium falciparum merzoite proteins enhances DNA vaccine protein expression and immunogenicity in mice. *Infect. Immun.*, **69**, 7250–7253.
27. Housby,J.N., Thorbjarnardottir,S.H., Jonsson,Z.O. and Southern,E.M. (2000) Optimised ligation of oligonucleotides by thermal ligases: comparison of Thermus scotoductus and Rhodothermus marinus DNA ligases to other thermophilic ligases. *Nucleic Acids Res.*, **28**, e10.
28. Stemmer,W.P., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
29. Ho,S.N., Hunt,H.D., Horton,R.M., Pullen,J.K. and Pease,L.R. (1989) Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene*, **77**, 51–59.
30. Wallace,R.B., Shaffer,J., Murphy,R.F., Bonner,J., Hirose,T. and Itakura,K. (1979) Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res.*, **6**, 6353–6357.
31. Krishnakumar,S., Zheng,J., Wilhelmy,J., Faham,M., Mindrinos,M. and Davis,R. (2008) A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl Acad. Sci.*, **105**, 9296–9301.