nature biotechnology

Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips

Sriram Kosuri[1,2,6], Nikolai Eroshenko[1,3,6], Emily M LeProust[4], Michael Super[1], Jeffrey Way[1], Jin Billy Li[2,5] & George M Church[1,2]

Development of cheap, high-throughput and reliable gene synthesis methods will broadly stimulate progress in biology and biotechnology[1]. Currently, the reliance on column-synthesized oligonucleotides as a source of DNA limits further cost reductions in gene synthesis[2]. Oligonucleotides from DNA microchips can reduce costs by at least an order of magnitude[3–5], yet efforts to scale their use have been largely unsuccessful owing to the high error rates and complexity of the oligonucleotide mixtures. Here we use high-fidelity DNA microchips, selective oligonucleotide pool amplification, optimized gene assembly protocols and enzymatic error correction to develop a method for highly parallel gene synthesis. We tested our approach by assembling 47 genes, including 42 challenging therapeutic antibody sequences, encoding a total of ~35 kilobase pairs of DNA. These assemblies were performed from a complex background containing 13,000 oligonucleotides encoding ~2.5 megabases of DNA, which is at least 50 times larger than in previously published attempts.

The synthesis of DNA encoding regulatory elements, genes, pathways and entire genomes provides powerful ways to both test biological hypotheses and harness biology for our use. For example, from the use of oligonucleotides in deciphering the genetic code[6,7] to the recent complete synthesis of a viable bacterial genome[8], DNA synthesis has engendered tremendous progress in biology. Currently, almost all DNA synthesis relies on the use of phosphoramidite chemistry on controlled-pore glass (CPG) substrates. The synthesis of gene-sized fragments (500–5,000 base pairs (bp)) relies on assembling many CPG oligonucleotides together using a variety of gene synthesis techniques[2]. Technologies to assemble verified gene-sized fragments into much larger synthetic constructs are now fairly mature[8–12].

The price of gene synthesis has fallen drastically over the last decade. However, the current commercial price of gene synthesis, ~$0.40–1.00/bp, has begun to approach the relatively stable cost of the CPG oligonucleotide precursors (~$0.10–0.20/bp)[1], suggesting that oligonucleotide cost is limiting. At these prices, the construction of large gene libraries and synthetic genomes is out of reach to most. There are many ongoing efforts to lower the cost of gene synthesis that focus on reducing the cost of the oligonucleotide precursors. For example, microfluidic oligonucleotide synthesis can reduce reagent cost by an order of magnitude and has been used for proof-of-concept gene synthesis[13].

Another promising route is to harness existing DNA microchips, which can produce up to a million different oligonucleotides on a single chip, as a source of DNA. Previous efforts have demonstrated that genes can be synthesized from DNA microchips[3–5,14]. Thus far it has not been possible to scale up these approaches for at least three reasons. First, the error rates of oligonucleotides from DNA microchips are higher than traditional column-synthesized oligonucleotides. Second, the assembly of gene fragments becomes increasingly difficult as the diversity of the oligonucleotide mixture becomes larger. Finally, the potential for cross-hybridization between individual assemblies imposes strong constraints on the sequences that can be constructed on an individual microchip.

Recently, the quality of microchip-synthesized oligonucleotides was improved by controlling depurination during the synthesis process[15]. These arrays produce up to 55,000 200-mer oligonucleotides on a single chip and are sold as a ~1–10 picomole pools of oligonucleotides, termed OLS pools (oligo library synthesis). Several groups have used OLS pools in DNA capture technologies, promoter analysis and DNA barcode development[16–20]. We have previously shown that individual oligonucleotides in a 55,000 150-mer OLS pool were evenly distributed[18]. We reanalyzed this data set to provide an estimate of the frequency of transitions, transversions, insertions and deletions in this OLS pool (Online Methods) and found the overall error rate to be ~1/500 bp both before and after PCR amplification, suggesting that OLS pools can be used for accurate large-scale gene synthesis (**Supplementary Table 1**).

To test whether OLS pools could be used for DNA microchip-based gene synthesis, we designed two pools (OLS pools 1 and 2) of different lengths, each containing ~13,000 130-mer or 200-mer oligonucleotides, respectively. **Figure 1** is a general schematic of our methods for using OLS pools to perform gene synthesis. Briefly, we designed oligonucleotides that were then printed on DNA microchips and recovered as a mixed pool of oligonucleotides (OLS pool). Next, we took advantage of the long oligonucleotide lengths to

[1]Wyss Institute for Biologically Inspired Engineering, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Harvard School of Engineering and Applied Sciences, Cambridge, Massachusetts, USA. [4]Agilent Technologies, Santa Clara, California, USA. [5]Present address: Department of Genetics, Stanford University, Stanford, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to S.K. (sri.kosuri@wyss.harvard.edu) or N.E. (eroshenk@wyss.harvard.edu).

Received 3 August; accepted 25 October; published online 28 November 2010; doi:10.1038/nbt.1716

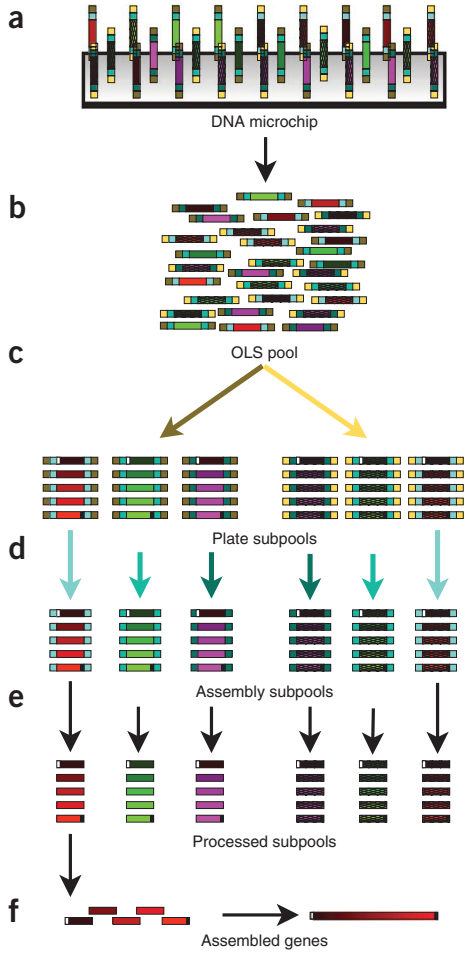

Figure 1 Schematic for scalable gene synthesis from OLS pool 2. (**a**,**b**) Pre-designed oligonucleotides (no distinction is made between dsDNA and ssDNA in the figure) are synthesized on a DNA microchip (**a**) and then cleaved to make a pool of oligonucleotides (**b**). (**c**) Plate-specific primer sequences (yellow or brown) are used to amplify separate plate subpools (only two are shown), which contain DNA to assemble different genes (only three are shown for each plate subpool). (**d**) Assembly-specific sequences (shades of blue) are used to amplify assembly subpools that contain only the DNA required to make a single gene. (**e**) The primer sequences are cleaved using either type IIS restriction enzymes (resulting in dsDNA) or by DpnII/USER/λ exonuclease processing (producing ssDNA). (**f**) Construction primers (shown as white and black sites flanking the full assembly) are then used in an assembly PCR reaction to build a gene from each assembly subpool. Depending on the downstream application, the assembled products are then cloned either before or after an enzymatic error correction step.

independently PCR amplify and process only those oligonucleotides required for a given gene assembly. For the 200-mer OLS pool 2, we first amplified a 'plate subpool' that contained DNA to construct up to 96 genes, and then amplified individual 'assembly subpools' to separate the oligonucleotides for an individual gene. For the 130-mer OLS pool 1, we directly amplified assembly subpools, foregoing the plate subpool step. Next, the primers used for these amplification steps were removed by either type IIS restriction endonucleases to form double-stranded DNA (dsDNA) fragments (OLS pool 2), or a combination of enzymatic steps to form single-stranded DNA (ssDNA) fragments (OLS pool 1). Finally, we constructed full-length genes using PCR assembly, performed enzymatic error correction to improve error rates if necessary, and, finally, cloned and characterized the constructs.

Obtaining subpools of only those DNA fragments required for any particular assembly is crucial for robust gene synthesis in very complex DNA backgrounds. In addition, isolating subpools relieves constraints on sequence similarity inherent in past approaches. To facilitate the partitioning of OLS pools into smaller subpools, we designed 20-mer PCR primer sets with low potential cross-hybridization ('orthogonal' primers) derived from a set of 244,000 25-mer orthogonal sequences developed for barcoding purposes[21]. Two separate orthogonal primer sets were constructed for the different OLS pools because of their varying requirements for downstream processing. Both sets were screened for potential cross-hybridization, low secondary structure and matched melting temperatures to construct large sets of orthogonal PCR primer pairs.

To construct genes from the OLS pools, we developed algorithms to split the sequence into overlapping segments with matching melting temperatures such that they could be later assembled by PCR. Genes on OLS pool 1 and 2 were designed differently to test the effect of different overlap lengths. We designed genes on OLS pool 1 such that the processed ssDNA pools fully overlapped to form a complete dsDNA sequence. In OLS pool 2, the processed dsDNA fragments partially overlapped by ~20 bp and could be assembled into a contiguous gene sequence using PCR. We initially constructed a set of fluorescent proteins to test the efficacy of the gene synthesis methods on both OLS pools (**Fig. 2**).

For OLS pool 1, we designed two independent 'assembly subpools' that encoded GFPmut3b plus flanking orthogonal primer sequences that were later used for PCR assembly (construction primers). The two assembly subpools, GFP43 and GFP35, differed in the average overlap length (43 and 35 bp, respectively), total length (82–90 and 64–78 bases, respectively) and number of oligonucleotides (18 and 22, respectively). We also designed two subpools, control subpools 1 and 2, containing ten and five 130-mer oligonucleotides, respectively, to test amplification efficacy. The other eight subpools, containing a total of 12,945 130-mer sequences, were constructed on the same chip but were not used in this study except to provide potential sources of cross-hybridization. Each of these 12 subpools was flanked with independent orthogonal primer pairs (assembly-specific primers). As a control, we used these same algorithms to design a set of shorter column-synthesized oligonucleotides (20 bp average overlap; 35–45 bases in length; and 39 total oligonucleotides) encoding GFPmut3b and obtained them from a commercial provider (IDT). These oligonucleotides were combined to form a third pool (GFP20) that was also tested. (All synthesized oligonucleotides used in the study can be found in **Supplementary Sequences**).

Each of the four subpools (GFP43, GFP35, control 1 and control 2) were PCR amplified from the synthesized OLS pool using modified primers that facilitated downstream processing (**Supplementary Figs. 1** and **2a**)[18]. The oligonucleotides were then processed to remove primer sequences (**Supplementary Figs. 2b** and **3**). Briefly, lambda exonuclease was used to make the PCR products single stranded, and then uracil DNA glycosylase, endonuclease VIII and DpnII restriction endonuclease were used to cleave off the assembly-specific primers. The resultant gel shows that although the reaction was efficient, unprocessed oligonucleotide still remained. In addition, we observed spurious cleavage by DpnII that was likely due to the extensive overlap within the subpool that is inherent in the gene synthesis process. We assembled the GFP43, GFP35 and GFP20 subpools using PCR, which resulted in GFP-sized products as well as many incorrect low molecular weight products (**Fig. 2a**).

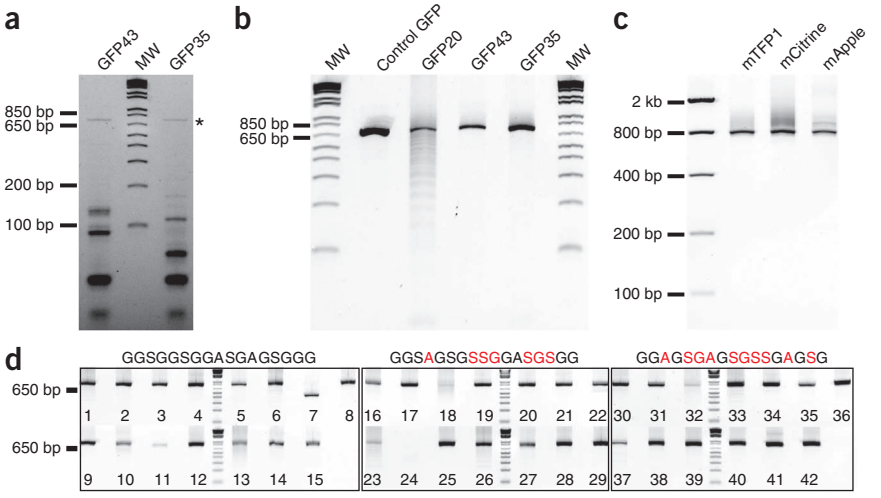

Figure 2 Gene synthesis products. (**a**) Results of PCR assembly of GFPmut3 from two different assembly subpools (GFP43 and GFP35) that were amplified from OLS pool 1. Full-length GFPmut3 is expected to be 779 bp and is indicated with an asterisk (*). Other bands show lower molecular weight misassembled products. (**b**) Gel purification and re-amplification of the full-length assembled GFPmut3. (**c**) Results of assembling three fluorescent proteins using the longer oligonucleotides in OLS pool 2 and a PCR assembly protocol that did not require gel isolation. (**d**) Results of assembling 42 variable regions of single-chain antibody fragments that contained challenging GC-rich linkers. Of the 42 assemblies, all but two (7 and 24) resulted in strong bands of the correct size. We gel isolated and re-amplified these two, resulting in bands of the correct size (**Supplementary Fig. 10b**). The antibody that corresponds to each number is given in **Supplementary Table 3** and the amino acid sequence of the linker region used is given above each gel with differing amino acids in red.

We gel isolated, digested and then cloned the assembly products into an expression vector (**Fig. 2b** and **Supplementary Fig. 4**). We used flow cytometry tests, manual colony counts and sequencing of individual clones to measure the error rates (**Supplementary Fig. 5a**,**b**). All three of the assays correlated well, and the error rates determined through sequencing were 1/1,500 bp, 1/1,130 bp and 1/1,350 bp for the GFP43, GFP35 and GFP20 synthesis reactions, respectively (**Fig. 3** and **Supplementary Table 2**).

These results illustrate a number of important points. First, our subpool assembly primers were sufficiently well-designed to provide stringent subpool amplification of as few as 5 oligonucleotides out of a 12,995 oligonucleotide background. Second, the relative quantities of the oligonucleotides in the assembly subpools were sufficient to allow PCR assembly. Third, the error rates from 130-mer OLS pools were sufficiently low to construct gene-sized fragments (717 bp) such that >50% of the sequences were perfect. In fact, the error rates from both the GFP43 and GFP35 assemblies were indistinguishable from the column-synthesized GFP20 assemblies. Fourth, our data show that the level of fluorescence of our gene assemblies correlated with the number of constructs with perfect sequence, providing a useful screen to test fluorescent gene assemblies in OLS pool 2 (**Supplementary Fig. 6**). Finally, although PCR assembly was able to generate full-length product, many smaller misassembled products were also formed, requiring the use of difficult-to-automate gel isolation steps.

In OLS pool 2, we designed 836 assembly subpools split into 11 plate subpools, encoding 2,456,706 bases of oligonucleotides that could potentially result in 869,125 bp of final assembled sequence. We first constructed three fluorescent proteins to test assembly protocols in OLS pool 2: mTFP1, mCitrine and mApple. We found that the PCR assembly protocols developed for ssDNA subpools in OLS pool 1 only produced short (<200 bp) misassemblies when applied to the dsDNA subpools in OLS pool 2. We tested over 1,000 assembly PCR conditions by varying parameters such as DNA concentration, annealing temperatures, cycle numbers, polymerase choice and buffer conditions. Using the best protocol (**Supplementary Note**), we assembled the genes encoding the three proteins with no detectable misassemblies, thereby removing the need for gel

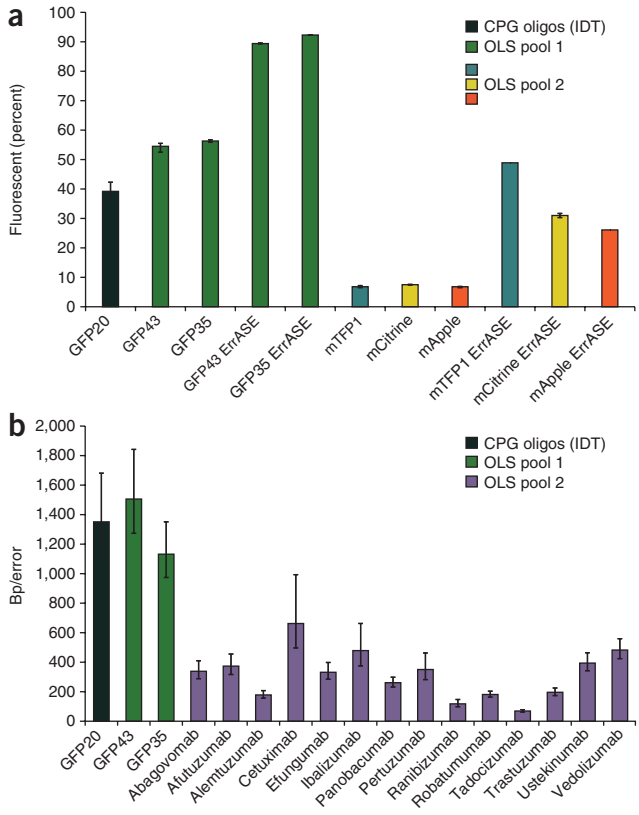

Figure 3 Characterization of products from OLS pools 1 and 2. (**a**) Percentage of fluorescent cells resulting from synthesis products derived from column-synthesized oligonucleotides (black), OLS Chip 1 subpools GFP43 and GFP35 (green) and the three fluorescent proteins produced on OLS Chip 2 with and without ErrASE treatment (blue, yellow and orange). Error bars correspond to the range of replicates from 3 (GFP20, GFP43, GFP35), 2 (GFP43 ErrASE, GFP35 ErrASE), 4 (mTFP1, mCitrine, mApple, mCitrine ErrASE) and 1 (mTFP1 ErrASE, mApple ErrASE) separate electroporations. (**b**) Error rates (average bp of correct sequence per error) from various synthesis products. Error bars show the expected Poisson error based on the number of errors found (±√n). Deletions of more than two consecutive bases are counted as a single error (no such errors were found in OLS pool 1).

isolation (**Fig. 2c** and **Supplementary Fig. 7a**,**b**). Cloning followed by flow cytometry screening showed that 6.8%, 7.5% and 6.8% of the cells were fluorescent for mTFP1, mCitrine and mApple assemblies, respectively (**Fig. 3a**).

Assuming 6% correct sequence per construct and no selection against errors in the assembly process, the error rate was ~1/250 bp for 200-mer OLS pool 2, significantly above that of the estimates for 130-mer OLS pool 1 (~1/1,000 bp) and the sequenced 55,000 150-mer OLS pool (~1/500 bp). This is not completely unexpected, as the amount of depurination is dependent upon the number of deprotection steps during synthesis and thus the oligonucleotide length. Despite the higher error rate, there were several advantages to the 200-mer OLS pool 2. First, the extensive overlaps designed in OLS pool 1 caused spurious processing of the primers from the assembly subpools. The use of type IIs restriction endonucleases to process primers to form dsDNA resulted in more robust processing. Second, the use of two amplification steps conserves chip-eluted DNA to allow for future scaling of the gene synthesis process (**Supplementary Note**). Third, the assemblies of OLS pool 1 produced many smaller bands and required lower-throughput gel isolation procedures. This could be due to mispriming during PCR assembly because of the long overlap lengths used in the design process. The assemblies in OLS pool 2 used much shorter overlap lengths and resulted in no smaller molecular weight misassembled products.

To improve the error rates of the genes assembled from OLS pool 2, we used ErrASE, a commercially available enzyme cocktail that detects and corrects mismatched base pairs, to remove errors in the assembled fluorescent proteins. For each gene, we applied ErrASE at six different stringencies, reamplified the constructs, cloned the PCR products and rescreened the cloned genes using flow cytometry. Improvement of the level of fluorescence progressively increased with greater ErrASE stringency. At the highest levels of error correction, the fluorescence levels were 31%, 49% and 26% for mTFP1, mCitrine and mApple respectively (**Fig. 3a** and **Supplementary Fig. 8**). We also performed the ErrASE procedure on our GFP43 and GFP35 pools from OLS pool 1, resulting in fluorescence levels of 89% and 92%, respectively (**Fig. 3a** and **Supplementary Fig. 5c**). We sequenced clones of GFP43 and GFP35 and found three errors in 21,510 (1/7,170 bp) and four errors in 20,076 (1/5,019 bp) sequenced bases, respectively.

As a more challenging test for our DNA synthesis technology, we designed and synthesized oligonucleotides in OLS pool 2 for 42 genes encoding the variable regions of single-chain antibody fragments (scFv) regions corresponding to a number of well-known antibodies. We have previously had trouble synthesizing these genes using commercial gene synthesis companies. This might be partly due to the prototype $(Gly_4Ser)_3$ linker, which is designed to maximize flexibility and allow the heavy and light V regions to assemble[22]. The repetitive nature and high GC content of the linker-encoding sequences often represents a challenge for accurate DNA synthesis. We therefore tested three different linker sequences that varied in GC content and repetitive character of the linker encoding sequence. In addition, the presence of high sequence homology in the antibody backbones and linkers represented a potential source of cross-hybridization that could interfere with assembly (61% average sequence identity).

As expected, the antibody sequences did not assemble as robustly as the fluorescent proteins, and thus we further optimized the conditions during pre- and post-assembly (**Supplementary Figs. 7c**, **9** and **10a**). Under the best protocol, 40 of the 42 constructs assembled to the correct size (**Fig. 2d** and **Supplementary Table 3**). The two misassembled genes displayed faint bands at the correct size, which were gel isolated and reamplified to produce strong bands of the correct size. We sequenced 15 antibodies including representatives from all three linker types. We performed enzymatic error correction using ErrASE, gel isolated the product and finally cloned the constructs into an expression vector. One of the 15 antibodies did not clone, and another had a deleted linker region in all 21 sequenced clones. Both of these antibodies were encoded with the highest GC content linker. The average error rate of the 14 antibodies that did clone was 1/315 bp (**Fig. 3b** and **Supplementary Table 2**); this was considerably higher than the GFP assemblies, but still sufficient for construction of genes of this size (~10% of clones should be perfect, on average). In addition, the high levels of sequence similarity between the antibodies, combined with the successful assembly and sequencing (which showed no instances of cross-contamination) further validates that the selective amplification is at least stringent enough to make highly related protein sequences.

Our results show the assembly of gene-sized DNA fragments totaling ~35,000 bp from oligonucleotide pools of more than 50 kilobases. A number of key features are important to make the process work, including the use of low-error starting material, well-chosen orthogonal primers, subpool amplification of individual assemblies, optimized assembly methods and enzymatic error correction. Together, these features enabled gene assembly from oligonucleotide pools containing at least 50 times more sequences than previously reported (**Supplementary Note**). We describe two separate OLS pool lengths and assembly methods, which have their own advantages and disadvantages (**Supplementary Fig. 1**). The shorter, 130-mer OLS pool 1 assemblies have lower error rates, but because there are no plate amplifications, will be harder to scale as we begin to utilize larger OLS pools. The longer 200-mer OLS pool 2 is easier to scale, but contained higher error rates. The costs of oligonucleotides in both processes are <$0.01/bp of final synthesized sequence, and thus the dominant costs are enzymatic processing, cloning and sequence verification. Future work on reducing the cost of perfect sequence will focus on the ability to lower sequencing costs by using cheaper next-generation sequencing technologies, or by incorporating other error-correction techniques such as PAGE selection of oligonucleotide pools or mutS-based error filtration[3,23].

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was supported by the US Office of Naval Research (N000141010144), National Human Genome Research Institute Center for Excellence in Genomics Science (P50 HG003170), Department of Energy Genomes to Life (DE-FG02-02ER63445), Defense Advanced Research Projects Agency (W911NF-08-1-0254) and the Wyss Institute for Biologically Inspired Engineering (all to G.M.C.). We thank H. Padgett for providing ErrASE and expertise during optimization and J. Boeke for advice on gene assembly protocols. We also thank S. Raman, F. Vigneault and F. Zhang for critical readings of the manuscript, G. Dantas for pZE21 (Washington University), F. Isaacs (Yale University) for pZE21G and J.S. Workman (Wyss Institute) for pSecTag2A.

AUTHOR CONTRIBUTIONS

S.K. and N.E. wrote the paper with contributions from all authors; S.K. and G.M.C. conceived the study; S.K. wrote all algorithms and designed all sequences; S.K. and N.E. designed and performed all experiments; E.L. provided the oligonucleotides libraries; M.S. and J.F. designed the single-chained versions of commercial antibodies; J.B.L. performed the OLS high-throughput sequencing experiment and provided critical advice on the processing of subpools.

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Carr, P.A. & Church, G.M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
2. Tian, J., Ma, K. & Saaem, I. Advancing high-throughput gene synthesis technology. *Mol. Biosyst.* **5**, 714–722 (2009).
3. Tian, J. *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**, 1050–1054 (2004).
4. Richmond, K.E. *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* **32**, 5011–5018 (2004).
5. Zhou, X. *et al.* Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.* **32**, 5409–5417 (2004).
6. Nirenberg, M.W. & Matthaei, J.H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* **47**, 1588–1602 (1961).
7. Söll, D. *et al.* Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc. Natl. Acad. Sci. USA* **54**, 1378–1385 (1965).
8. Gibson, D.G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
9. Gibson, D.G. Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* **37**, 6984–6990 (2009).
10. Li, M.Z. & Elledge, S.J. Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. *Nat. Methods* **4**, 251–256 (2007).
11. Bang, D. & Church, G.M. Gene synthesis by circular assembly amplification. *Nat. Methods* **5**, 37–39 (2008).
12. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **37**, e16 (2009).
13. Lee, C.-C., Snyder, T.M. & Quake, S.R. A microfluidic oligonucleotide synthesizer. *Nucleic Acids Res.* **38**, 2514–2521 (2010).
14. Kim, C. *et al.* Progress in gene assembly from a MAS-driven DNA microarray. *Microelectron. Eng.* **83**, 1613–1616 (2006).
15. LeProust, E.M. *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
16. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
17. Schlabach, M.R. *et al.* Synthetic design of strong promoters. *Proc. Natl. Acad. Sci. USA* **107**, 2538–2543 (2010).
18. Li, J.B. *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res.* **19**, 1606–1615 (2009).
19. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
20. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
21. Xu, Q. *et al.* Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. USA* **106**, 2289–2294 (2009).
22. Huston, J.S. *et al.* Medical applications of single-chain antibodies. *Int. Rev. Immunol.* **10**, 195–217 (1993).
23. Carr, P.A. *et al.* Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.* **32**, e162 (2004).

ONLINE METHODS

Reanalysis of OLS pool error rates. We reanalyzed a previously published data set for determining sequencing error rates[18]. Briefly, the data set was derived from high-throughput sequencing using the Illumina Genome Analyzer platform of a 53,777 150-mer OLS pool. Two sequencing runs were performed, the first before any amplification, and the second after two rounds of ten cycles of PCR (20 cycles total). As our previous analyses were mostly looking for distribution effects, we reanalyzed these existing data to get an estimate of error rates before and after PCR amplification. We realigned the data set using Exonerate to allow for gapped alignments and analysis of indels[24]. Specifically, we used an affine local alignment model that is equivalent to the classic Smith-Waterman-Gotoh alignment, a gap extension penalty of −5, and used the full refine option to allow for dynamic programming–based optimization of the alignment. These reads were solely mapped on base calls by the Illumina platform. We used these alignments to count mismatches, deletions and insertions as compared to the designed sequences. However, as base-calling can be more error prone on next-generation platforms than traditional Sanger-based approaches, we filtered the results based only on high-quality base-calls (Phred scores of ≥30 or >99.9% accuracy). This was accomplished by converting Illumina quality scores to Phred values using the Maq utility sol2sanger[25] and only using statistics from base calls of Phred 30 or higher. All error rate analysis scripts were implemented in Python and are available upon request. Although this method provides an estimate for error rates, unmapped reads may have higher error rates, thus underestimating the total average error rate. In addition, base-calling errors might still overestimate the error rate. Finally, using only high-quality base calls, which usually occur only in the first ten bases of a read, might only reflect error rates on the 5′ end of the synthesized oligonucleotide.

Design and synthesis of OLS pools. The 13,000 oligos in the first OLS library (OLS pool 1) were broken up into 12 separately amplifiable subpools (assembly subpools). Each assembly subpool was defined by unique 20 bp priming sites that flanked each of the oligos in the pool. The priming sites were designed to minimize amplification of oligos not in the particular assembly subpool. This was done by designing set of orthogonal 20-mers (assembly-specific primers) using a set of 240,000 orthogonal 25-mers[21] as a seed. From these sequences we selected 20-mers with 3′ sequence ending in thymidine or GATC for the forward and reverse primers, respectively. We screened for melting temperatures of 62–64 °C and low primer secondary structure. After the additional filtering, 12 pairs of forward and reverse primers were chosen to be the assembly-specific primers. The 13,000 oligos in the second OLS library (OLS pool 2) were broken up into 11 subpools corresponding to 11 sets of up to 96 assemblies (plate subpools), which were further divided into a total of 836 assembly subpools. A new set of orthogonal primers were designed similarly to the previous set (without the GATC and thymidine constraints) but further filtered to remove type IIS restriction sites, secondary structure, primer dimers and self-dimers. The final set of primer pairs was distributed among the plate-specific primers, assembly-specific primers and construction primers. See **Supplementary Methods** for more detailed design information and primer sequences.

OLS pools were synthesized by Agilent Technologies and are available upon signing a Collaborative Technology Development agreement with Agilent. Costs of OLS pools are a function of the number of unique oligos synthesized and of the length of the oligos (<$0.01 per final assembled base-pair for all scales used in this study). OLS pools 1 and 2 were independently synthesized, cleaved and delivered as lyophilized ~1–10 picomole pools.

Amplification and processing of OLS subpools. Lyophilized DNA from OLS pools 1 and 2 were resuspended in 500 µl TE. Assembly subpools were amplified from 1 µl of OLS pool 1 in a 50 µl qPCR reaction using the KAPA SYBR FAST qPCR kit (Kapa Biosystems). A secondary 20 ml PCR amplification using Taq polymerase was performed from the primary amplification product. The barcode primer sites were removed using a technique previously described[20]. In brief, the forward primers contained a phosphorothioate bond at the 5′ end and the last nucleotide on the 3′ end was a deoxyuridine; the reverse primers contained a DpnII recognition site (GATC) at the 3′ end and a phosphorylated 5′ end. PCR amplification was followed by λ-exonuclease digestion of 5′ phosphorylated strands, hybridization of the 3′ primer site to its complement, and cleavage of the 5′ and 3′ primer sites using USER enzyme mix and DpnII (New England Biolabs), respectively. Plate subpools were amplified from 1 µl of OLS pool 2 in 50 µl Phusion polymerase PCR reactions. Assembly subpools were amplified from the plate subpools by 100 µl Phusion polymerase PCR reactions. A BtsI digest removed the forward and reverse primer sites. See the **Supplementary Methods** for more detailed protocols.

Assembly of fluorescent proteins. GFPmut3 (ref. 26) was assembled from OLS pool 1 assembly subpools by PCR. The GFP43 and GFP35 subpools were designed such that there was full overlap between neighboring oligos during assembly, with average overlaps of 43 bp and 35 bp for GFP43 and GFP35, respectively. For the first set of assemblies, 330 pg of the GF43 subpool or 40 pg of the GFP35 subpool was used per 20 µl Phusion polymerase PCR assembly. The full-length product was gel-isolated, amplified using Phusion polymerase and cloned into pZE21 after a HindIII/KpnI digest. The second set of assemblies was built using a similar procedure, except that the assembly PCR used 170 pg or 190 pg of GFP43 and GFP35 subpools, respectively; and the gel-isolated product was not re-amplified before cloning.

Oligonucleotides for mTFP1, mCitrine and mApple were designed such that there was on average a 20 bp overlap between adjacent oligonucleotides. The proteins were built from OLS pool 2 assembly subpools by first performing a KOD polymerase pre-assembly reaction that was done in the absence of construction primers followed by a KOD polymerase assembly PCR in which the construction primers were included. ErrASE error correction was then performed on aliquots of the synthesis products following the manufacturer's instructions. The assembled product was digested with HindIII and KpnI and cloned into pZE21. Sequencing of clones was performed by Beckman Coulter Genomics. See the **Supplementary Methods** for more detailed protocols.

ErrASE. ErrASE is an enzyme cocktail designed to remove errors in synthetically assembled genes (Novici Biotech). Assembled genes are denatured and re-annealed to allow for the formation of hetero-duplexes. A resolvase enzyme in ErrASE then recognizes and cuts at mismatched positions. Other enzymes in the cocktail remove these cut mismatched positions. The products could then be reamplified by PCR to reassemble the full-length gene.

Specifically, six aliquots of 10–50 ng of each assembled gene was added to 10 µl of PCR buffer (we have also tested the effects of including betaine in the buffer; see **Supplementary Fig. 11**). Hetero-duplexes were formed by denaturing at 95 °C and slowly cooling to 0 °C. Each aliquot was then used to resuspend six different lyophilized ErrASE mixtures of increasing stringency provided by the manufacturer. After a 1–2 h at incubation 25 °C, the assemblies were re-amplified and visualized on an agarose gel. Of the reactions that resulted in a correctly sized band, the one that used the most stringent ErrASE protocol was selected for cloning.

Flow cytometry. Fluorescent cell fractions of the cloned libraries of assembly products were quantified using a BD LSR Fortessa flow cytometer either a 488 nm laser with a 530 nm filter (30 nm bandpass) or a 561 nm laser with a 610 nm filter (20 nm bandpass).

Synthesis of antibodies. 125 ng of each antibody assembly pool was pre-assembled in 20 µl KOD pre-assembly reactions. We then tested nine amplification protocols for the ability to amplify the 42 antibody pre-assemblies into full-length genes. We attempted to clone eight constructs from the best assembly protocol (afutuzumab, efungumab, ibalizumab, oportuzumab, panobacumab, robatumumab, ustekinumab and vedolizumab; see **Supplementary Fig. 10a** and **Supplementary Table 3**). The eight assemblies were error-corrected using ErrASE, gel-isolated, re-amplified using Phusion polymerase, gel-isolated again, and cloned into pSecTag2A after an ApaI/SfiI digest. Sequencing was performed by Genewiz. All but oportuzumab cloned successfully. We then repeated the experiment, increasing the amount of assembly-pool DNA in the pre-assembly reaction to 400 ng. We selected a different set of eight constructs from this second set of assemblies for cloning (abagovomab,

doi:10.1038/nbt.1716

alemtuzumab, ranibizumab, cetuximab, efungumab, pertuzumab, tadocizumab and trastuzumab; see **Fig. 2d** and **Supplementary Table 3**). Using the same methods as with the first set of cloned antibodies, this second set was error-corrected, gel-isolated, cloned and sequenced. See the **Supplementary Methods** for more detailed protocols.

24. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
25. Li, H. Maq: mapping and assembly with qualities. *Welcome Trust Sanger Institute* (2010). Available at: <http://maq.sourceforge.net>.
26. Cormack, B.P., Valdivia, R.H. & Falkow, S. FACS-optimized mutants of the green fluorescent protein (GFP). *Gene* **173**, 33–38 (1996).