Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules

Jerrod J Schwartz, Choli Lee & Jay Shendure

We present dial-out PCR, a highly parallel method for retrieving accurate DNA molecules for gene synthesis. A complex library of DNA molecules is modified with unique flanking tags before massively parallel sequencing. Tag-directed primers then enable the retrieval of molecules with desired sequences by PCR. Dial-out PCR enables multiplex *in vitro* clone screening and is a compelling alternative to *in vivo* cloning and Sanger sequencing for accurate gene synthesis.

The prospect of generating synthetic genes and genomes has only been realized because of the availability of programmable building blocks in the form of synthetic oligonucleotides. However, the speed, cost and accuracy of gene synthesis have been limited by the high error rates found in these chemically synthesized precursors. We have developed a method that takes advantage of massively parallel sequencing to identify and retrieve accurate building blocks, thereby greatly improving the efficiency of gene synthesis.

The traditional approach of using controlled-pore glass (CPG) media for oligonucleotide synthesis is expensive and results in high error rates: 1 in every 100–1,000 bases[1]. This is a key bottleneck for high-throughput and inexpensive synthetic gene and genome construction[2], which generally rely on sequence-verified precursors derived from CPG-synthesized oligonucleotides. Isolating accurate precursors typically requires laborious cloning and Sanger sequencing to identify correct molecules for downstream processing.

Significant effort has also been directed at exploiting programmable microarrays as a source of oligonucleotides for inexpensive, multiplex gene synthesis[3–5]. However, multiplex assembly from microarray-derived precursors typically results in many inaccurate constructs because error rates are often higher than for CPG oligonucleotides, and the abundance of individual products can vary by orders of magnitude. Consequently, the retrieval of error-free sequences by cloning and Sanger sequencing remains a rate-limiting step regardless of whether CPG methods or microarrays are used for DNA synthesis.

Here we present a general strategy for massively multiplex verification followed by retrieval of specific molecules with accurate or desired sequences from a complex mixture of nucleic acids. Unlike sequence enrichment or error correction approaches requiring *in vivo* cloning[6], specialized instrumentation[7], enzymatic processing[8–11] or a specific next-generation sequencing (NGS) platform[7], our method is entirely *in vitro* and compatible with any NGS platform. Multiplex verification and tag-directed retrieval can be performed on mixtures of oligonucleotides, assembled DNA constructs or mutagenized DNA constructs. Here we focus on a complex mixture of microarray-synthesized oligonucleotides. We also demonstrate rapid assembly of accurate, full-length genes from retrieved precursors.

First, oligonucleotides corresponding to sequences of interest ('gene fragments') are designed, synthesized and released from a DNA microarray. This results in a mixture of accurate and inaccurate fragments whose relative abundances may vary considerably. Second, adaptors containing unique tags are added to the ends of each fragment with PCR (**Fig. 1a**). The tags are embedded as variable subsequences within the primers, and PCR conditions are chosen to impose a complexity bottleneck such that each molecule has a high probability of receiving a unique tag pair. Third, the tagged library is deeply sequenced using any NGS platform. In cases in which read lengths supported by the platform are shorter than the gene fragments or the platform has a high error rate, tag-directed 'subassembly'[12] can be used to generate long, accurate reads. For retrieval, primers complementary to the tags of a

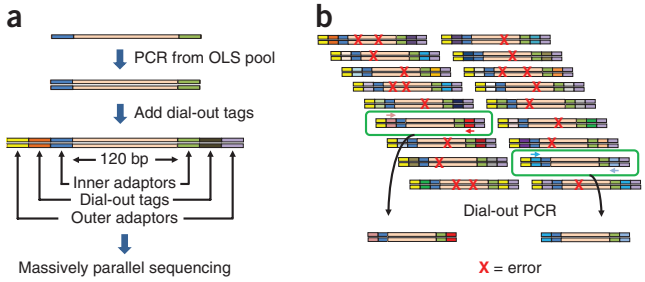

Figure 1 | Dial-out PCR for retrieving accurate sequences from a nonuniform, error-rich library of synthetic DNA molecules. (**a**) Groups of single-stranded oligonucleotides are synthesized on either controlled-pore glass columns or microarrays and are PCR amplified in parallel. The library is then modified with two unique, flanking 'dial-out' tags and PCR amplified using a set of common outer adaptors before being subjected to massively parallel sequencing. Paired-end reads match the tags with the internal sequence of the fragment. (**b**) Dial-out PCR primers are designed against tag pairs associated with accurate sequences and are used to selectively amplify and retrieve them from the original complex library pool at any time.

Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to J.J.S. (jschwar@uw.edu) or J.S. (shendure@uw.edu).

RECEIVED 30 APRIL; ACCEPTED 2 JULY; PUBLISHED ONLINE 12 AUGUST 2012; DOI:10.1038/NMETH.2137

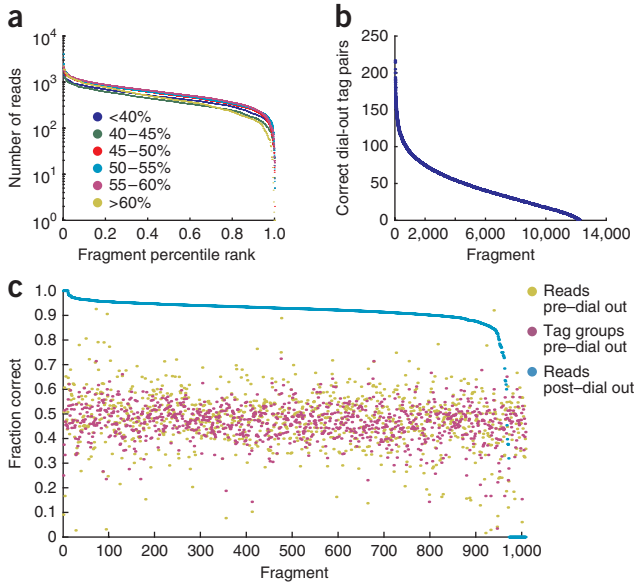

Figure 2 | Assessment of designed *E. coli* sequence fragments before and after dial-out PCR. (**a**) Uniformity of the six GC-content groups after initial amplification and tagging. The total number of reads mapping to each of the 12,472 fragments is plotted in percentile rank order for each group. (**b**) The number of unique dial-out tag pairs corresponding to accurate sequence is plotted in rank order for the 12,148 designed fragments that had at least one accurate sequence. (**c**) The fraction of correct molecules and tag groups for the 1,008 fragments before and after dial-out PCR, plotted in rank order.

desired error-free gene fragment are used for 'dial-out' PCR to produce high-purity DNA for downstream synthesis (**Fig. 1b**).

To demonstrate the method, we designed 12,472 synthetic fragments corresponding to a contiguous 1.25 Mb (27%) of the *Escherichia coli* K12 genome (bases 1–1,247,200). The fragments were 120 nucleotides (nt) long and tiled with 20-nt overlaps on both sides. We appended one of six pairs of 20-nt adaptor sequences that were selected based on the GC content of each fragment (total length = 160 nt). After microarray-based oligonucleotide library synthesis (OLS)[13], we amplified each of the six groups from the OLS pool separately. We added adaptors containing a tag region of 16 degenerate bases by further PCR. The tagged library was then subjected to massively parallel paired-end sequencing (Illumina MiSeq; 2 × 151 bp), which yielded 5.5 million mapping read pairs. Over 90% of the 12,472 fragments were present at abundances within a tenfold range, and the group with GC content ≥60% exhibited the highest dropout rate (28 of 1,048, or 2.7%) (**Fig. 2a**). We observed one or more read pairs for 12,422 (99.6%) fragments and one or more unique tag pairs corresponding to an accurate molecule for 12,148 (97.4%) fragments (**Fig. 2b** and **Supplementary Fig. 1**). Overall, the error rate in the tagged pool was ~1 in 230 bp (**Supplementary Fig. 2**).

To evaluate the robustness of tag-directed retrieval, we randomly selected 1,022 of 12,472 (8%) fragments for dial-out PCR corresponding to 122,640 bp of synthetic DNA. Within this subset, one fragment was not observed in the sequenced pool and 11 of 1,022 (1%) fragments were observed, but no accurate molecules were identified. For the remaining fragments, for which accurate molecules were identified, we designed PCR primers complementary to the tags flanking accurate molecules. Two fragments had tag pairs containing ≥5-base stretches of guanine bases, and retrieval was not attempted for these. We ordered dial-out primers for the remaining 1,008 fragments from a commercial vendor and used them with no additional purification. Dial-out PCR reactions for 984 of 1,022 (96%) selected fragments amplified as expected, either becoming detectable by real-time PCR between cycles 25 and 35 or yielding an appropriately sized gel band (**Supplementary Fig. 3**). We found that primers that failed to amplify appeared likely to form hairpins or dimers (**Supplementary Note 1**). We pooled aliquots from all 1,008 PCR reactions and sequenced the pool for validation (Illumina MiSeq; 2 × 151 bp).

We focused our analysis on the 120-nt region of interest and included only read pairs for which all base calls at overlapping positions agreed, which provided sequencing data for 971 of 1,022 (95%) selected fragments. Three fragments were only seen once in the tagged library, thus demonstrating successful retrieval of rare species at levels as low as ~1 molecule per 100,000 before dial-out PCR. Without quality score filtering, the top 949 fragments had a postretrieval error rate of 1 in 1,808 bp and a 93% median accuracy (**Fig. 2c**). When we imposed a Q30 minimum quality score requirement (see Online Methods) on all bases, the error rate dropped to 1 in 2,176 bp and the median accuracy improved to 96% (**Supplementary Fig. 4**), suggesting that observed errors are primarily sequencing errors rather than errors introduced during dial-out PCR. The remaining retrieved fragments (22 of 1,022, or 2%) had a much higher error rate (1 in 398 bp) and a lower median accuracy of 71%. Most of the errors in this subgroup were recurrent and likely due to polymerase errors in GC-rich regions during an early cycle of the dial-out PCR.

We were surprised at the robustness of dial-out PCR given our lack of strict criteria for primer selection and the potential for cross-hybridization. Nine reactions did not produce a visible gel band but still gave accurate reads, which suggests that retrieval was possible even with problematic primers. For the retrieved fragments in which errors were observed, one could perform a second dial-out PCR, either with the same set of primers or with primers directed at a different tag pair for the same fragment.

To demonstrate that fragments retrieved by dial-out PCR can be used for rapid and accurate gene synthesis, we randomly selected 27 *E. coli* genes between 900 and 1,100 bp in length that had at least one accurate copy of each fragment required for assembly. Dial-out PCR successfully retrieved 277 of 289 (96%) of the targeted fragments; the 12 (4%) dropouts appeared to fail because of primer issues. For all 15 genes not missing any fragments, we rapidly assembled full-length genes from 10 or 11 constituent fragments in a single step using *in vitro* recombination[14] (**Supplementary Fig. 5**). Cloning and Sanger sequencing confirmed that all 15 genes, spanning a total of 15,800 bp, were synthesized with no errors. This indicates that the actual postretrieval error rate is less than 1 in 15,800 bp and may be limited by only the 1 in 10^7 bp error rate of the polymerase used for retrieval. We did not attempt assembly of the 12 genes for which a single fragment failed dial-out PCR. We note, however, that each of these fragments had at least ten additional tag pairs that could be used in a second round of retrieval.

To be viable as a fully practical alternative to cloning and Sanger sequencing, dial-out PCR should enable validation and

retrieval for fragments over a range of sizes (100–1,000 bp) and should not be limited by current NGS read lengths and error rates. To this end, we evaluated an alternative approach for the preparation of longer gene fragments involving (i) multiplex polymerase cycling assembly (PCA) of microarray-derived oligonucleotides, (ii) tag-directed subassembly[12] to overcome sequencer read length limits and identify accurately synthesized and assembled molecules, and (iii) tag-directed retrieval via dial-out PCR (**Supplementary Note 2**).

Although we successfully demonstrated subassembly-based verification, retrieval of ~300-bp gene fragments and accurate assembly of three Neanderthal genes, the overall results were mixed. For example, with 64-plex PCA, only 42 of 192 (22%) designed fragments had at least one accurate copy available for retrieval, whereas with 4-plex or 7-plex PCA, all 19 of 19 (100%) fragments had at least one accurate copy. We surmise that our success was limited because of a more error-prone 200-mer OLS pool and a substantial nonuniformity in the abundance of individual fragments after multiplexed PCA that resulted in a high dropout rate.

While this manuscript was in review, a method conceptually similar to dial-out PCR was reported[15] that involves the multiplex assembly of ~300-bp gene fragments before 454 sequencing and tag-directed retrieval of accurate molecules. Although ultimately successful in generating accurate precursors for subsequent assembly, the authors encountered challenges analogous to ours. Future efforts with this alternative strategy (synthesis → multiplex assembly → verification → retrieval) may be directed at building longer intermediates using more accurate OLS pools and assembling with multiplex *in vitro* recombination instead of PCA. If the key challenges can be overcome, the benefits of the alternative strategy include fewer retrieval reactions per gene as well as fewer postretrieval assembly steps per gene.

With increasing scales of oligonucleotide synthesis comes a concomitant need to rapidly screen complex synthetic libraries and selectively retrieve specific error-free sequences. This is currently a significant bottleneck for the field of synthetic biology. The screening and retrieval steps are typically accomplished by cloning and serial colony picking, and they are among the few places where conventional Sanger sequencing is still essential. The first report using NGS as a preparative tool to retrieve desired sequences[7] required highly specialized instrumentation and was compatible only with the 454 platform. In contrast, dial-out PCR consists of broadly adoptable protocols and is compatible with any NGS platform.

As described here, the synthesis and retrieval of sequence-verified 120-bp gene fragments can be completed in ~5 d (**Supplementary Note 3**). Retrieval expenses are dominated by the cost of the dial-out PCR primers, although we believe that this could be greatly reduced by using a static library of tags. For example, a standardized adaptor library containing 10^4 forward and 10^4 reverse tags gives 10^8 unique possible forward-reverse tag combinations, which is more than sufficient for dial-out PCR of accurate molecules. We also show that retrieved fragments can be quickly assembled into accurate full-length genes. The efficiency and cost of gene assembly might be improved by increasing the gene fragment size, eliminating the adaptor removal step during gene assembly, performing multiple dial-out retrievals in a single reaction, or further multiplexing *in vitro* recombination–based assembly.

Many laboratories may not require DNA synthesis at a scale that utilizes the full capacity of a microarray OLS pool. However, there are ongoing efforts to develop 'synthetic DNA foundries' that might effectively consolidate the needs of the community to achieve maximal cost-effectiveness. We also note that tagged, validated material derived from an OLS pool can be stored, and the retrieval of accurate versions of specific targets can be performed as needed.

Dial-out PCR allows for the normalization of target sequence abundance before multiplex assembly steps, and it has the potential to decrease production costs for high-quality, sequence-verified synthetic DNA by over an order of magnitude. The dial-out concept also supports clone retrieval from other complex nucleic acid libraries, such as in the screening and recovery of specific mutants from a complex mutagenesis library.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.B. Hiatt, J.O. Kitzman, R.P. Patwardhan, J. Cooper, L. Pennacchio, E. Rubin and S. Deutsch for helpful discussions. We also thank Y. Zhang (Albert Einstein College of Medicine) for providing the SLiCE strain and R. Qiu for preparing the SLiCE extract. J.J.S. was funded by a Helen Hay Whitney Foundation postdoctoral fellowship. Our work was supported in part by a grant from the US National Institutes of Health–National Cancer Institute (1R21CA160080 to J.S.).

AUTHOR CONTRIBUTIONS

J.J.S. and J.S. designed the research, J.J.S. performed the research, C.L. sequenced the libraries, and J.J.S. and J.S. analyzed the data and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/nmeth.2137. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ma, S., Saaem, I. & Tian, J. *Trends Biotechnol.* **30**, 147–154 (2012).
2. Gibson, D.G. *et al. Science* **329**, 52–56 (2010).
3. Borovkov, A.Y. *et al. Nucleic Acids Res.* **38**, e180 (2010).
4. Kosuri, S. *et al. Nat. Biotechnol.* **28**, 1295–1299 (2010).
5. Tian, J. *et al. Nature* **432**, 1050–1054 (2004).
6. Gibson, D.G. *et al. Nat. Methods* **6**, 343–345 (2009).
7. Matzas, M. *et al. Nat. Biotechnol.* **28**, 1291–1294 (2010).
8. Bang, D. & Church, G.M. *Nat. Methods* **5**, 37–39 (2008).
9. Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman, M.R. & Belshaw, P.J. *Nucleic Acids Res.* **33**, e55 (2005).
10. Carr, P.A. *et al. Nucleic Acids Res.* **32**, e162 (2004).
11. Smith, J. & Modrich, P. *Proc. Natl. Acad. Sci. USA* **94**, 6847–6850 (1997).
12. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. *Nat. Methods* **7**, 119–122 (2010).
13. Maurer, K. *et al. PLoS ONE* **1**, e34 (2006).
14. Zhang, Y., Werling, U. & Edelmann, W. *Nucleic Acids Res.* **40**, e55 (2012).
15. Kim, H. *et al. Nucleic Acids Res.* published online, doi:10.1093/nar/gks546 (16 June 2012).

ONLINE METHODS

Oligonucleotide synthesis and design strategy. We used the *E. coli* K12 substrain DH10B (GenBank CP000948.1) genome sequence as the basis for the OLS pool synthesis design. The first 1,247,200 bp of the genome were partitioned into 120-nt individual fragments (12,472 total), and each fragment shared 20 nt of overlapping sequence with its adjacent neighbors to facilitate downstream assembly. No attempt was made to optimize the overlapping regions with regard to length or melting temperature (T_m). We binned the fragments into six groups according to GC content (<40%, 40–45%, 45–50%, 50–55%, 55–60% and >60% GC) and added one of six pairs of 20-nt adaptor sequences to the 5′ and 3′ ends to facilitate amplification (total length = 160 nt, **Supplementary Table 1**).

We obtained conventional oligonucleotides (adaptors, PCR primers and sequencing primers) from a commercial vendor (Integrated DNA Technologies). The OLS pool was synthesized on a programmable microarray using a semiconductor electrochemical process (CustomArray)[13].

Amplification and tagging. The raw oligonucleotide pool was initially size selected by loading 10 µL (53 ng/µL) on a 6% denaturing polyacrylamide gel (Invitrogen). A band corresponding to 140–180 nt was excised from the gel and purified. The six GC group fragment pools were then amplified separately from the size-selected OLS pool with Kapa HiFi Hotstart Ready Mix (Kapa Biosystems) using real-time PCR on a MiniOpticon (Bio-Rad) and group-specific primers (**Supplementary Table 1**; primers are named according to the GC content of the group they amplify—for example, ca_40-45_f and ca_40-45_r are for the 40–45% group). The cycling conditions were (i) 95 °C for 2 min, (ii) 98 °C for 20 s, (iii) 65 °C for 15 s, (iv) 72 °C for 15 s, (v) looping through (ii)–(iv) 35 times and (vi) 72 °C for 5 min. Reactions were pulled from the cycler just before plateauing, cleaned up using AMPure (Agencourt) and eluted in 30 µL water. Each pool was then quantified using a Qubit (Invitrogen).

For the tagging reaction, we did a second PCR with Kapa HiFi Hotstart Ready Mix containing 0.5 ng of the group-amplified template, 1 fmol of each tagging primer (for example, 40-45_f_tag and 40-45_r_tag for the 40–45% group, **Supplementary Table 2**) and 25 pmol of the outer flow cell primers (ill_tag_amp_f and ill_tag_amp_r). The cycling conditions used were the same as described above. The outer primers were added after five cycles to allow for appropriate bottlenecking during the initial extension. Following the PCR, the six reactions were run on a 6% polyacrylamide gel (Invitrogen) and the product band was size selected (~310 bp).

Illumina sequencing. We pooled the six groups in proportion to the number of designed sequences within each group and sequenced the combined pool on an Illumina MiSeq (2 × 151 bp). Custom read 1 and read 2 sequencing primers were added to the sequencing cartridge before starting the run (illum_read1 and illum_read2, **Supplementary Table 3**).

Analysis of tag-defined read groups. The paired-end 151-bp reads available on the Illumina MiSeq enabled the library to be sequenced with 110 bp of overlap between reads. We trimmed the first 36 bp from both read 1 and read 2 and placed this tag-plus-adaptor sequence in the read header for reference. Taken together, each tag pair established read-pair membership in a 'tag-defined read group' that formed the basis for subsequent analysis. We mapped the reads to the 12,472 designed target sequences using the Burrows-Wheeler Aligner (BWA)[16]. No base quality score filtering was done at this stage, and only reads that agreed within the overlap region were carried forward. Within a tag-defined read group, if all of the grouped reads had the same reference sequence and were accurate, we flagged the tag pair as being accurate. If there were any agreed-upon errors within the overlap region, or if there was an error within either of the 5-bp non-overlapping regions, we flagged the tag pair as being inaccurate. Tag pairs that mapped to more than one species in the same GC group were discarded (~2% of all unique tag pairs).

Dial-out PCR. We randomly selected 1,022 of the 12,472 designed fragments for retrieval. The vast majority of these targets had multiple unique dial-out tag pair candidates to choose from. To maximize retrieval success while keeping the tag selection process as simple as possible, we chose the tag pair that was the most abundant and that did not contain a stretch of five or more guanine bases (i.e., GGGGG). The reasoning behind this is that poly(G) nucleic acids have a tendency to form guanine tetraplex structures[17] and can be difficult to synthesize. The T_m of each tag of length n was calculated using the formula

$$T_m = 81.5 + 16.6 \times \log_{10}[Na^+] + 41 \times (\%GC) - \frac{600}{n}$$

If a tag had a T_m ≥60 °C, we selected it as a dial-out primer as is. If a tag had a T_m <60 °C, we added 3 nt to the 5′ end corresponding to the constant bases in the adaptor sequence. This process was repeated until the T_m of the tag's dial-out primer was ≥60 °C. We ordered premixed 96-well plates of dial-out primers for 1,008 of the 1,022 selected fragments with no additional purification (Integrated DNA Technologies, standard desalting).

Each dial-out PCR reaction used Kapa HiFi Hotstart Ready Mix and included 0.1 ng of the tagged template library and 25 pmol of each dial-out primer. The cycling conditions were (i) 95 °C for 2 min, (ii) 98 °C for 20 s (iii) 65 °C for 15 s, (iv) 72 °C for 15 s, (v) looping through (ii)–(iv) 35 times and (vi) 72 °C for 5 min. Reactions were pulled from the cycler just before plateauing, purified with AMPure (Agencourt) and eluted in 30 µL water. A random selection of 37 of the dial-out PCR products that amplified well was run on a 6% polyacrylamide gel (Invitrogen) to check the product size (**Supplementary Fig. 3**). All of the real-time PCR reactions that appeared to fail (came up early, late or not at all) were also checked on a gel for the presence of a product of the expected size.

Sequence verification of dial-out PCR products. To verify the accuracy of the retrieved fragments, for each 96-well plate we pooled 5 µL from each reaction (well) into a 'plate' pool and purified each plate pool with AMPure. Next, the ends of the dial-out products were end-repaired and 5′-phosphorylated (End-It DNA End-Repair, Epicentre). The molecules were then A-tailed (NEBNext, NEB) and ligated to Y-tailed sequencing adaptors (ill_yad_1 and ill_yad_2, **Supplementary Table 4**) using Ultrapure T4 DNA ligase (Enzymatics). Finally, we performed a light PCR with Kapa HiFi Hotstart Ready Mix to add Illumina-compatible

doi:10.1038/nmeth.2137

flow cell adaptors (ill_flow_1 and ill_flow_2). Following AMPure purification, all of the plate libraries were pooled and loaded on an Illumina MiSeq for sequencing (2 × 151 bp). The standard Illumina sequencing primers were replaced with ill_val_r1 and ill_val_r2.

We first aligned post–dial-out sequencing reads to the 12,472 fragment reference sequences to identify the adaptor-insert junction position. This step was performed because the molecules had a variable-length (16–25 nt) dial-out primer sequence at the start of every read. We then trimmed adaptor and tag sequences from each read and realigned the trimmed reads to the reference sequences. Next, we screened the reads for pairs that agreed in the overlap region. In parallel analyses, we imposed a minimum quality score requirement ($Q \geq 2$, $Q \geq 5$, $Q \geq 10$ and so on, up to $Q \geq 32$) across all insert bases such that only read pairs with all bases greater than or equal to the threshold were kept. For each quality-threshold–level data set, we evaluated read pairs for accuracy with respect to the reference and determined the median accuracy of all the passing read pairs (**Supplementary Fig. 4**).

Gene assembly with dial-out PCR products. We randomly selected 27 genes 900–1,100 bp in length that had at least one accurate copy of each fragment required for assembly. Dial-out PCR primers for the corresponding 289 fragments were designed and ordered from IDT; fragments were retrieved using dial-out PCR as described above. Gene assembly with *in vitro* recombination was performed for the subset of 15 genes for which dial-out PCR was successful for all constituent fragments (**Supplementary Table 5a**). We note that the 12 fragments that failed dial-out PCR each had at least ten other tag pairs corresponding to accurate molecules that could be used in a second round of retrieval (**Supplementary Table 5b**).

Although SLiCE is capable of recombination activity even with heterologous flanking sequences, we opted to remove the adaptors by performing a light PCR with 20-nt internal primers to maximize recombination efficiency. Internal primers were designed and synthesized by simply taking the first and last 20 bp of each fragment sequence and ordering the corresponding oligonucleotide primer. If the primer corresponded with the beginning or end of the gene, a 26-nt tail sequence was appended to facilitate cloning into the BamHI-EcoRI sites of pUC19 (primer tails of 5′-GTT GTA AAA CGA CGG CCA GTG AAT TC-3′ or 5′-GCC TGC AGG TCG ACT CTA GAG GAT CC-3′). Following adaptor removal, fragments were purified with AMPure and eluted in 30 µL water.

SLiCE buffer and extract were prepared as described in ref. 14. For each gene, we pooled 1.6 µL of each adaptor-removed fragment (10 or 11 fragments per gene), 5 µL of 5× SLiCE buffer (250 mM Tris-HCl (pH 7.5 at 25 °C), 50 mM $MgCl_2$, 5 mM ATP and 5 mM DTT) and 2 µL SLiCE extract, and we incubated the reaction at 37 °C for 1 h. Full-length genes were PCR amplified with Kapa HiFi Hotstart Ready Mix using the terminal primers of the first and last fragment in the assembly. The cycling conditions were: (i) 95 °C for 2 min, (ii) 98 °C for 20 s, (iii) 62 °C for 15 s, (iv) 72 °C for 60 s, (v) looping through (ii)–(iv) 35 times and (vi) 72 °C for 5 min. All 15 amplified full-length genes were checked for the correct size on a 6% polyacrylamide gel (**Supplementary Fig. 5**). At this stage, we size selected any genes that had multiple bands to improve cloning efficiency. We then cloned each of the genes into pUC19 with InFusion HD (Clontech) and transformed the plasmids into Fusion-Blue Competent Cells (Clontech). After plating on LB + Amp + IPTG + X-Gal plates and growing overnight, five or more white colonies for each gene were selected for plasmid amplification (TempliPhi, GE Healthcare) and forward and reverse Sanger sequencing (GENEWIZ).

16. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
17. Poon, K. & Macgregor, R.B. *Biopolymers* **45**, 427–434 (1998).