ACS Synthetic Biology

Technical Note

pubs.acs.org/synthbio

Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process

Elaine B. Shapland,[†,§] Victor Holmes,[†,§] Christopher D. Reeves,[†,§] Elena Sorokin,[†] Maxime Durot,[†,‡] Darren Platt,[†] Christopher Allen,[†] Jed Dean,[†] Zach Serber,[†] Jack Newman,[†] and Sunil Chandran*,[†]

[†]Amyris, Inc., 5885 Hollis, Suite 100, Emeryville, California 94608, United States

[‡]TOTAL New Energies USA, Inc., 5858 Horton Street, Suite 253, Emeryville, California 94608, United States

S *Supporting Information*

ABSTRACT: In recent years, next-generation sequencing (NGS) technology has greatly reduced the cost of sequencing whole genomes, whereas the cost of sequence verification of plasmids via Sanger sequencing has remained high. Consequently, industrial-scale strain engineers either limit the number of designs or take short cuts in quality control. Here, we show that over 4000 plasmids can be completely sequenced in one Illumina MiSeq run for less than $3 each (15× coverage), which is a 20-fold reduction over using Sanger sequencing (2× coverage). We reduced the volume of the Nextera tagmentation reaction by 100-fold and developed an automated workflow to prepare thousands of samples for sequencing. We also developed software to track the samples and associated sequence data and to rapidly identify correctly assembled constructs having the fewest defects. As DNA synthesis and assembly become a centralized commodity, this NGS quality control (QC) process will be essential to groups operating high-throughput pipelines for DNA construction.

Custom software
Low sample volume automation, high plexity
4000 plasmids sequenced at $2.68 each

KEYWORDS: *synthetic biology, next-generation sequencing, NGS, high throughput*

Synthetic biologists routinely assemble well-characterized DNA parts into larger constructs and introduce those DNA assemblies into host organisms to achieve desired phenotypes.[1−4] This is often a trial-and-error process that requires building and testing tens to thousands of DNA assemblies. For example, a comprehensive combinatorial exploration of five genes each expressed at five levels would require 3125 DNA assemblies. At a company like Amyris, it is common to build many constructs to test diverse hypotheses or to optimize a multigene pathway using iterative design−build−test−learn cycles similar to strategies described previously.[5−7] On this scale, quality control (QC) of large numbers of DNA assemblies creates logistical and economic challenges.

High-throughput strain engineering facilities routinely use automated workflows to assemble thousands of DNA constructs ranging in size from 3 to 30 kb and containing 2−12 parts. The DNA assemblies must hence undergo rigorous QC to avoid building and testing incorrectly engineered strains, which could lead to erroneous conclusions regarding genotype−phenotype relationships. Because no assembly method is perfect, finding a correct assembly requires QC analysis to be performed on multiple clones. Until recently, this involved comparing the observed restriction endonuclease fragment sizes to those computationally predicted[8] for four colonies, followed by Sanger sequencing of the chosen clone. To achieve 2× coverage across a 10 kb assembly using Sanger sequencing requires at least 24 reads spaced appropriately across the assembly and costs at least $72. This is too expensive and logistically onerous for a high-throughput operation.

By combining an Illumina MiSeq platform[9] with a LabCyte Echo acoustic liquid dispensing system, we have developed a rigorous, low-cost QC method that enables complete sequencing of almost every DNA assembly built by a high-throughput operation. The MiSeq can provide about 5 gigabases (GB) of data in a 24 h run using the 300-cycle v2 kit,[10,11] theoretically allowing 25 000 plasmids of 10 kb average size to be sequenced. However, there were several obstacles to be overcome before we could achieve even a fraction of this high level of multiplexing.

The Illumina Nextera method for preparing sequencing libraries is convenient and robust.[12] However, cost-effective sequencing of plasmids in the 3−30 kb range requires hundreds of barcode primers and a significant reduction in the use of the expensive Nextera reagents. A recent report described a Nextera workflow in which reaction volumes were reduced 8-fold relative to the Illumina protocol.[13] Here, in addition to showing that the volume of the tagmentation reaction can be reduced 100-fold using acoustic droplet ejection, we also demonstrate that thousands of uniquely barcoded samples can be handled with the appropriate automation infrastructure. We

Received: December 11, 2014
Published: April 25, 2015

DOI: 10.1021/sb500362n

demonstrate the simultaneous sequencing of over 4000 plasmids with an average size of 8 kb (largest ~20 kb) at a consumables cost of less than $3 per plasmid. Here, we focus on a description of the method and briefly discuss how such data is being used to optimize the DNA assembly process. A comprehensive analysis of the data vis-à-vis different assembly methods will be the topic of a separate publication.

■ RESULTS AND DISCUSSION

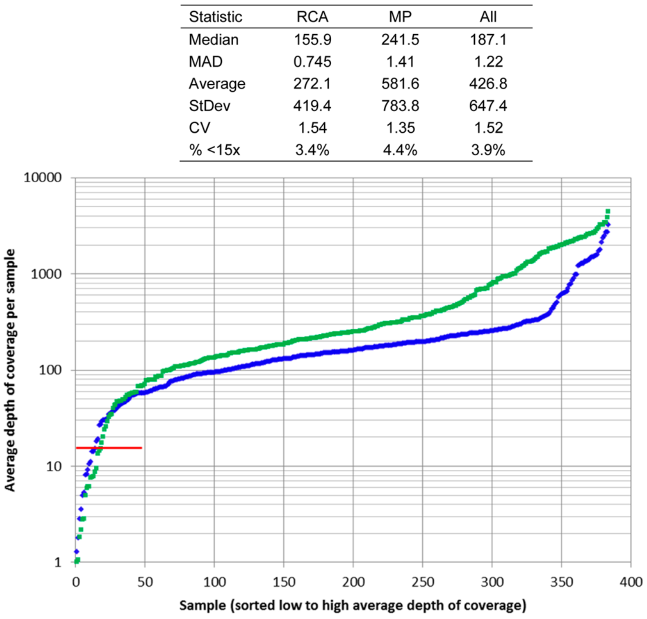
Reducing Tagmentation Reaction Volume. Tagmentation is like transposon insertion,[14] except the transposome cuts the target DNA and appends tags (transposon terminal sequences) to the resulting fragments. It is a stoichiometric, Poisson process, and the size distribution of the fragments is determined by the ratio of transposome to DNA. An Illumina Nextera kit for preparation of 96 samples costs $7000, so plasmid sequencing with these kits is very expensive and impractical. To reduce cost and establish a manageable workflow, we focused on reducing the volume of the tagmentation reaction in stepwise fashion, modifying other steps as necessary to adjust for reduced sample volume or total DNA mass. Early experiments showed that the tagmentation reagents could be used as a master mix and that 5 μL reactions gave sequence data quality equivalent to that obtained using the Nextera kit according to Illumina's protocol (50 μL tagmentation). This remained true upon further reduction of the reaction volume to 0.5 μL using the Echo acoustic liquid dispensing system.

After tagmentation, the transposase remains tightly bound to the DNA[14] and can inhibit the initial strand-displacing extension required for the PCR. In the Illumina protocol, the tagmented DNA is purified away from the transposase using Zymo Clean and Concentrate columns, but this is impractical for high throughput, and we explored other methods of removal. Tagmented DNA fragments or a control reagent (PCR products with ends identical to tagmented fragments after end repair) were subjected to various treatments, and the efficiency of PCR amplification was compared to that using Zymo column purification. Treatments tested included heat, pH, chaotropic agents, and detergents. After several experiments, it was determined that addition of SDS to a final concentration of 0.1% was most effective at removing the transposase without interfering with the subsequent PCR (Figure S1). This eliminated the cost-prohibitive column purification step.

Barcoding PCR. Barcoded adapters are attached to the ends of Nextera library fragments using a nonstandard PCR protocol requiring initial end repair with a strand-displacing polymerase. The volume of this PCR cannot be reduced too much or the subsequent size-selection by solid-phase reversible immobilization cannot be operationalized. By reducing the tagmentation reaction volume, the PCR reagents in the Nextera kit become limiting. As a potential replacement reagent to carry out this PCR, we chose Vent polymerase from New England Biolabs, which is reported to have strand-displacement activity and relatively high fidelity.[15] Figure S2 shows that Vent polymerase can replace the NPM reagent in the Illumina Nextera kit with only a slight decrease in PCR efficiency, which could be remedied by a compensatory increase in the number of PCR cycles.

To enable the required level of multiplexing, we designed a set of barcode adapter primers using previously described algorithms.[16,17] From all possible 8-base sequences, those with mononucleotide runs longer than two bases or GC content outside the range of 35−65% were removed. Sequences differing by at least three bases from all other barcodes in the set or from sequences complementary to all 8-base sequences present within the constant regions of the i5 and i7 adapter primers (see Illumina Customer Sequence Letter[18]) were then selected. These ~800 sequences were placed into the context of the full-length Illumina adapter primer, and the resulting adapter primers were analyzed using DINAMelt[19] to find those with the lowest predicted tendency to form inter- or intramolecular duplexes. Table S1 lists the set of barcodes used for the experiments in this work. Figure S2 shows that primers from Integrated DNA Technologies or from Illumina gave equivalent PCR efficiencies. At least 192 forward and 192 reverse barcode sequences (providing 36 864 unique barcode combinations) pass the filtering process described above.

Source of DNA for Nextera Library Preparation. For preparing plasmid DNA, rolling circle amplification (RCA) takes less than one-third of the hands-on time and produces more consistent final DNA concentrations compared to plasmid minipreps.[20] Data in the literature suggests that RCA DNA should be a good source for preparing Nextera libraries. For example, it gives good Sanger sequence data[20] and good restriction digest banding,[8] and whole genome-amplified DNA provides good Illumina sequence data.[22] A set of 384 DNA assemblies ranging in size from 4 to 20 kb was used to prepare both RCA DNA and plasmid DNA, and the 768 DNA samples were used to prepare a pool of 768 Nextera libraries for the MiSeq. Although the average depth of coverage for the 768 samples spanned over 3 orders of magnitude and displayed wide statistical variation (Figure 1), only 4% of the samples had an average coverage below 15×, an empirically determined point below which the sequence data is generally unreliable (see Illumina Technical Note: Estimating Sequencing Coverage[21]). Since the total yield of data in a MiSeq run is divided between the samples in the pool, it is most significant that the



Statistic	RCA	MP	All
Median	155.9	241.5	187.1
MAD	0.745	1.41	1.22
Average	272.1	581.6	426.8
StDev	419.4	783.8	647.4
CV	1.54	1.35	1.52
% <15x	3.4%	4.4%	3.9%

Figure 1. Distribution and statistics of read coverage for 768 samples prepared from DNA of 384 plasmids prepared by RCA (blue diamonds) or miniprep (MP; green squares). The red line indicates the 15× coverage threshold. MAD is the median absolute deviation.

plasmid DNA samples had about twice the coverage variation compared to the RCA DNA samples. This implies that a greater percentage of samples will have reliable data if the pool contains only RCA DNA samples instead of plasmid DNA samples. The sequence data for each DNA assembly was identical whether prepared by RCA or plasmid miniprep, with three exceptions where the samples prepared from plasmid DNA apparently lost the insert, perhaps because cells containing empty plasmid swept the population. We conclude that plasmid DNA prepared by RCA is superior to that prepared by alkaline lysis for highly multiplexed plasmid sequencing on the MiSeq

We observed that solutions of phage λ DNA at concentrations over ~20 ng/μL were not transferred by the Echo, apparently because long polymers can prevent ejection of emerging droplets. Since RCA DNA, like phage λ DNA, is high molecular weight (≥50 kb), we investigated how accurately RCA DNA was transferred by the Echo. A 384-well source plate was filled with precise concentrations of DNA generated from pure plasmid DNA using an Illustra Templiphi kit. As shown in Figure S3, the Echo accurately (>90%) and reliably transferred this DNA at concentrations up to 10 ng/μL.

Increasing the Number of Samples Receiving Sufficient Sequence Data. For a robust QC process, the samples should receive similar average read coverage and few should have less than 15× coverage. To achieve this, each sample in the pool should have a similar molar concentration of sequenceable fragments such that each forms a similar number of clusters on the MiSeq flow cell. When the same pool of Nextera libraries derived from the same set of plasmid constructs was sequenced in separate MiSeq runs, coverage was highly correlated between the runs (Figure S4), indicating that coverage variation arises during preparation and pooling of the libraries, not during the Illumina sequencing process. The sequence of each sample obtained from the two runs was identical, verifying the reliability of the sequence data itself (data not shown).

The large deviation in average coverage across the sample population in Figure 1 was observed early in the development of this method. Subsequently, the protocol was optimized, as described below, and the number of samples sequenced per run was steadily increased. To pool according to molar concentration, the average fragment size of thousands of samples must be determined in a reliable manner, which is time-consuming and labor-intensive. Therefore, we sought ways to minimize the variation in average fragment size across the libraries so that pooling could be based on mass concentration. The effect of input DNA concentration on coverage variability was studied using a plate of precise concentrations of RCA DNA to generate Nextera libraries. This revealed that input DNA concentrations of 3−10 ng/μL gave relatively consistent coverage, whereas coverage variation, and coverage itself, increased significantly as input DNA concentration fell below 2.5 ng/μL (Figure 2). Thus, coverage variation could be reduced by using RCA DNA at 3−10 ng/μL for tagmentation. In addition, the workflow could be streamlined because all samples could be diluted by a standard factor, instead of diluting each sample individually.

Samples at the edges of a plate sometimes had low concentrations, which we thought might be due to droplets veering to the sides such that reagents were not completely mixed at the bottom of wells. To mitigate this, plates were centrifuged at 1000*g* immediately after dispensing on the Echo.

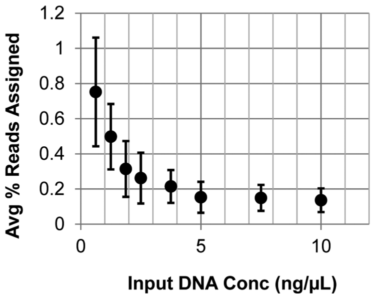


Figure 2. Effect of RCA DNA concentration in the tagmentation reactions on the percentage of reads assigned based on the barcodes. Each point represents the average of 48 samples; error bars are standard deviation. The expected average for the 384 samples is 0.26%.

We also decided to add the entire volume of any sample with a low concentration to the pool because such samples then had a chance of receiving coverage without significantly affecting the coverage of other samples.

The protocol changes discussed above were implemented for the parallel sequencing of 4078 plasmids. Figure 3 shows that the coverage variation and statistics for this MiSeq run were significantly improved over the run shown in Figure 1, with 98.4% receiving over 15× average coverage. Of the 1.6% samples with low coverage, most were found to be empty wells that had failed at the RCA step and would fail any QC method. We hypothesize that the slightly higher ratio of DNA to transposome during tagmentation reduced variation because the subsequent PCR to append the barcode adapter sequences uses a 30 s extension time that will not amplify fragments too large to form clusters. In other words, the higher DNA-to-protein ratio during tagmentation and the short PCR extension time may act to hold the variation within limits.

Statistic	Value
Median depth of coverage	82.3
Average depth of coverage	85.1
Standard deviation of coverage	37.1
Coefficient of coverage variation	0.44
Samples with less than 15x average coverage	1.6%

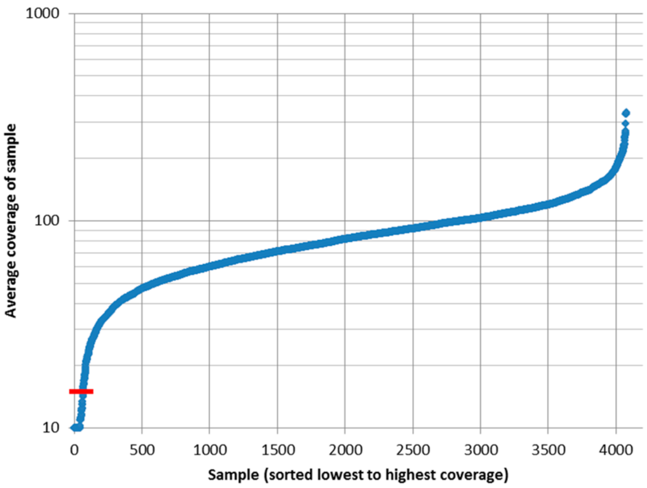


Figure 3. Distribution of read coverage for a run containing 4078 plasmid samples.

In the above QC of 4078 plasmids, the consumables cost was $2.68 per MiSeq sample, which breaks down as shown in Table S2. Although this is almost $11 per assembly (because four

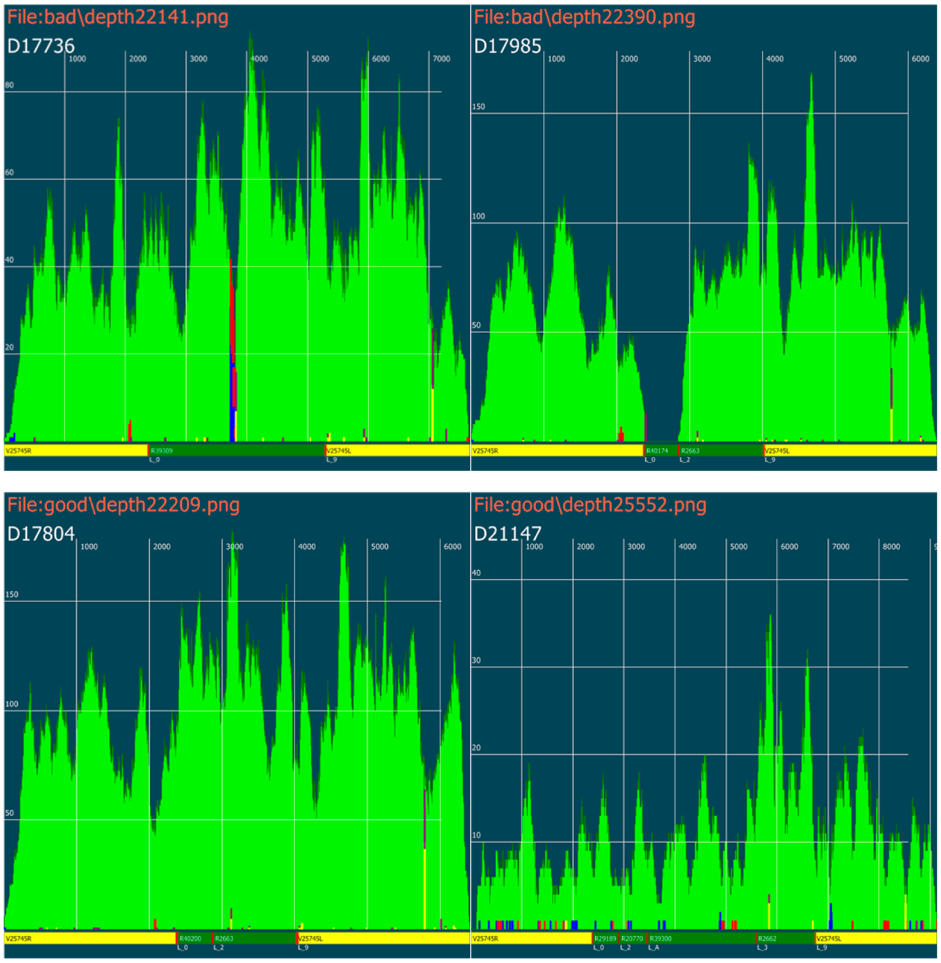


Figure 4. Example sequence data plots for samples from the run of 4078 samples described here. The top two show samples with differences between the reads and the reference, while the bottom two show samples that match the reference perfectly (not counting the vector). The green region shows the depth of coverage. Red and blue bars indicate a SNP in the forward and reverse reads, respectively. Purple and yellow bars indicate an indel in the forward and reverse reads, respectively. Note that even with less than 15× average coverage (bottom right) it is sometimes possible to obtain reliable QC data. At the bottom of each plot are the DNA parts in green and the vector in yellow.

replicates of each are sequenced), achieving only 2× coverage by Sanger sequencing of this same set of DNA assemblies would be about 20-fold more expensive and would include the need to order and track many primers to distribute the reads across the assemblies appropriately.

Analyzing the NGS QC Data. Aligning reads to a digital reference and choosing the best replicate of an assembly is conceptually simple but requires rapid, parallel analysis of many data sets. We initially tested SAMTOOLS and BCFTOOLS[29,30] to identify features (SNPs and indels), but we were unable to find appropriate settings to reliably call all mutations found in the plasmids. A possible cause for this could be the high read coverage seen in some samples (approaching 1000×), which may hinder some part of the mutation calling algorithm. We were reluctant to subsample the sequencing data in these cases, as this reduces resolution of SNP frequency and complicates base calling in regions of low coverage. Another possible cause is that our DNA samples may be mixed populations that do not resemble the diploid genomic samples against which these algorithms and tool sets were developed. For example, a SNP at 10% frequency does not match a heterozygous or homozygous situation. Interestingly, we found that the features were identified correctly at the level of read alignment but sometimes missed by the calling algorithms. Given the small size of the plasmids that we were sequencing (compared to genomes), we decided to implement our own simple feature detection method based on the pileup file. Software was written in F# (http://fsharp.org) to call mutations and assign severity scores to features based on their sequence context (e.g., part type and the probability that they could impair function). The software ranks the replicates of each assembly based on the number of mutations and their severity and reports which replicate best matches the digital template. In addition, the software stores all sequence variants found, along with other relevant information, in a postgreSQL database. Finally, the software generates a graphic for each sample (Figure 4) showing coverage and variant calls, which facilitates the investigation of specific cases when the algorithmic decision is in question. The uneven coverage in these examples is mostly due to Poisson sampling during the sequencing process. Some of the uneven coverage might also be due to bias for or against certain sequence motifs by either the transposome[23] or the polymerase used for the PCR.[24] On the other hand, it might also be an indication of sequence discrepancies that should be more closely investigated.

In the run with 4078 samples described here, 4056 were four replicates of 1014 constructs assembled by yeast homologous recombination. The remaining 22 samples were internal process controls, which were not used for data analysis. Table 1 shows the statistics for the sequence differences between the samples and the digital reference sequences. The importance of replicates is highlighted by the fact that, although 5.8% of the samples were misassembled, only 1% of the constructs had no correctly assembled replicate.

Table 1. Sequence Difference Statistics for the Four Replicates of 1014 Assemblies Assembled by Yeast Homologous Recombination

statistic	percent of 4056 samples or 1014 constructs
samples exactly matching the reference	54%
samples with only one SNP or one indel	23%
samples with more than one SNP or indel	16%
samples misassembled (zero coverage for >200 bp)	5.8%
constructs having at least one replicate matching reference	73%
constructs having at least one replicate correctly assembled	99%

When a SNP or indel is present in only one replicate of a construct, this is likely due to errors in the primers or errors by the polymerase during PCR amplification of parts. Alternatively, errors may arise during RCA for MiSeq sample prep. The frequency of this type of mutation appears to be consistent with the known fidelity of the polymerases[25] or with the reported frequency of errors in oligonucleotide primers.[26] Many indels were located at homopolymers, which are known to be susceptible to contraction during replication and are also prone to sequencing artifacts even on the Illumina platform. When the same SNPs or indels are present in all four replicates, or in the same part in different constructs, they are most likely due to errors in either the digital reference sequence (i.e., data entry) or the template used for PCR amplification of the part. Several errors were due to the use of a physical part for the PCR template that was not the same as the part specified in the digital request. The frequency of this type of mutation was higher than anticipated, and we can reduce it. Since the run with 4078 samples described here, we have used this NGS QC process in more than 10 assembly cycles, thus accumulating a large amount of NGS QC data. We intend to publish a comprehensive analysis of this data in the future and to identify how the assembly process generates the different types of mutations.

Applications of This Process. Plasmid DNA has been the workhorse of molecular biology for 4 decades. NGS protocols, however, have not yet enabled the cost-effective sequencing of these small bits of DNA. The method described herein bridges the power of NGS to the plasmid libraries used in gene synthesis, DNA assembly, enzyme engineering, amplicon sequencing, and library deconvolution, to name a few. These are common research areas in synthetic biology, and we expect the community to benefit from adoption of the NGS QC process described here.

■ METHODS

Instrumentation. Liquid transfers were carried out on Biomek FX or NX robots (Beckman Coulter, Brea, CA) for volumes greater than 2 μL or on an Echo 550 plus Access robot (Labcyte, Sunnyvale, CA) for volumes less than 2 μL. Sequencing was done on a MiSeq (Illumina, Inc., San Diego, CA). Fluorescence was read on an M5 plate reader (Molecular Devices, LLC, Sunnyvale, CA). DNA fragment size profiles were determined using either a Bioanalyzer 2100 (Agilent Technologies, Inc., Santa Clara, CA) or a Fragment Analyzer (Advanced Analytical Technologies, Inc., Ames, IA).

DNA Assembly and Quantitation. DNA parts with specific linker sequences at each end were assembled in a shuttle vector using yeast homologous recombination, followed by shuttling into *Escherichia coli* for isolation of DNA, as previously described.[8] DNA assemblies built using the ligase cycling reaction (LCR)[27] were also used in some experiments. Plasmid DNA was prepared by alkaline lysis and silica gel binding[8] or was amplified using an Illustra Templiphi kit (GE Healthcare Life Sciences, Piscataway, NJ). DNA concentration was measured using Quant-iT PicoGreen reagent (Life Technologies, Foster City, CA) in Costar 3658 or 3677 black 384-well plates (Corning, Inc., Corning, NY). The PicoGreen reagent was diluted with TE (10 mM Tris-HCl, pH 8, 0.5 mM EDTA) containing 0.05% Tween 20.

Preparing Nextera Libraries for Sequencing. Figure 5 depicts the chronological workflow for the highly multiplexed plasmid sequencing protocol described here. Using the reagents in an Illumina Nextera kit (FC-121-1031), the tagmentation reaction volume was reduced from 50 μL, as specified in the kit protocol, to 5 μL for the Biomek robots (2 μL of DNA solution and 3 μL of tagmentation master mix containing 0.5 μL tagmentation enzyme and 25 μL tagmentation buffer) or 0.5 μL (200 nL DNA and 300 nL of tagmentation master mix) for the Echo. Rolling circle amplified (RCA) DNA or plasmid DNA prepared by alkaline lysis was diluted with TE to achieve the desired concentration (2.5–10 ng/μL; see Results and Discussion). The transposase was dissociated from the tagmented DNA by adding SDS (sodium dodecyl sulfate) to a final concentration of 0.1% (e.g., 125 nL of 0.5% SDS added to 0.5 μL tagmented DNA).

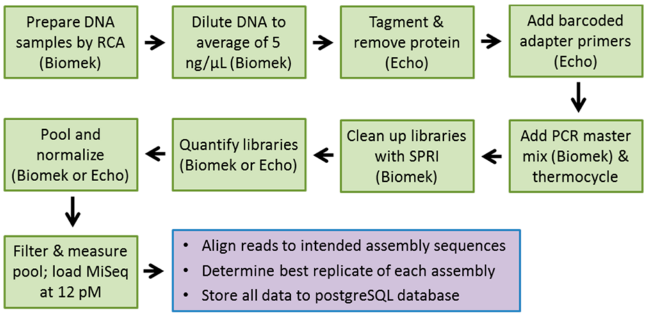

Prepare DNA samples by RCA (Biomek)
Dilute DNA to average of 5 ng/µL (Biomek)
Tagment & remove protein (Echo)
Add barcoded adapter primers (Echo)
Add PCR master mix (Biomek) & thermocycle
Clean up libraries with SPRI (Biomek)
Quantify libraries (Biomek or Echo)
Pool and normalize (Biomek or Echo)
Filter & measure pool; load MiSeq at 12 pM
• Align reads to intended assembly sequences
• Determine best replicate of each assembly
• Store all data to postgreSQL database

Figure 5. Diagram of the workflow for the plasmid QC process. The type of robot used at each step is indicated in parentheses.

Adapters for the Illumina sequencing process, including 8-base barcodes, were attached to each tagmented DNA sample using 12 cycles of PCR. All primers were obtained from IDT (Integrated DNA Technologies, Inc., Coralville, IA) with standard desalting. The barcodes inserted into the Illumina i5 and i7 adapter primer sequences (see Illumina Customer Sequence Letter[18]) are listed in Table S1. Using the Echo, each sample well received 125 nL of a forward barcode primer and 125 nL of a reverse barcode primer (each at 100 μM). A PCR master mix (24.5 μL) was then added using a Biomek robot. The master mix contained 0.2 units/μL of Vent DNA polymerase (New England Biolabs, Ipswich, MA), 1× Thermopol buffer (NEB), 2 mM $MgSO_4$, 200 μM of each deoxynucleotide triphosphate, and 200 nM of each terminal primer (to mitigate the fact that long oligonucleotides have 5′-end truncations). The thermocycler program was 3 min at 72 °C and then 12 cycles of 10 s at 98 °C, 30 s at 63 °C, and 60 s at 72 °C. Small fragments and unincorporated primers were removed from the resulting PCR products using 0.6 volumes of Ampure XP paramagnetic bead suspension (A63880, Beckman Coulter, Indianapolis, IN) per volume of PCR reaction according to the manufacturer's instructions.

Libraries were pooled and normalized based on DNA concentration and the size of the DNA assembly from which the library was generated. The goal of normalization is to achieve equal molar amounts of the DNA representing each plasmid (see Results and Discussion). The pool was filtered and concentrated using a Microcon fast-flow filter unit (EMD Millipore, Billerica, MA). The DNA concentration and average fragment size of the pool were determined by Picogreen fluorescence and a high-sensitivity DNA chip on a Bioanalyzer 2100, respectively. After diluting the filtered pool to 1.11 nM with water, 18 μL was denatured by adding 2 μL of 1 N NaOH. After 5 min at room temperature, 980 μL of ice-cold Illumina hybridization buffer was added, followed by 2 μL of 1 N HCl. The denatured pool was loaded on the MiSeq at 12 pM, which was empirically determined to give the optimum cluster density when following this protocol.

Sequence Data Processing. A web-based sequencing tracking system was created to manage the many samples and the large amounts of data generated. It facilitates the creation of runs, generation of sample sheets required by the MiSeq, and analysis of multiple data types, including the NGS QC data described here. Reads were demultiplexed using the embedded MiSeq Reporter software. For large numbers of multiplexed samples (>1000), it was necessary to increase the File Copy Timeout setting to avoid premature interruption of the demultiplexing process, which takes several extra hours after a highly multiplexed run appears to have completed. When a sequencing run completes, the system automatically retrieves the FASTQ files from the MiSeqOutput folder. Read mapping to the intended assembly sequences uses BWA v0.6.232 and the sampe method with default settings.[28] Alignments are stored in BAM file format using SAMTOOLS v0.1.19.[29,30] Mapping statistics are obtained using the SAMTOOLS flagstat utility. A pileup file is generated using SAMTOOLS mpileup with default options to obtain read coverage along the reference sequence.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figure S1: Identifying the optimum SDS and Triton X-100 concentrations for removal of the transposase after tagmentation. Figure S2: PCR efficiency using Vent polymerase and primers ordered from IDT or the Nextera kit reagents NPM and PPC. Figure S3: Transfer of RCA DNA by the Echo. Figure S4: Correlation of read coverage comparing two separate MiSeq runs of the same plasmids prepared for sequencing by the protocol described here. Table S1: Barcodes used in this work. Table S2: Consumables costs per sample when 4000 samples are sequenced in parallel. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/sb500362n.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: (510) 597 4765; Fax: (510) 225 2645; E-mail: chandran@amyris.com.

Author Contributions

§E.B.S., V.H., and C.D.R. contributed equally to this work. E.B.S., V.H., C.D.R., D.P., Z.S., J.D., J.N., and S.S.C. designed the research; E.B.S., V.H., C.D.R., E.S., M.D., D.P., and C.A. conducted the research; S.S.C. supervised the research; C.D.R., E.B.S., and S.S.C. wrote the paper.

Notes

The authors declare the following competing financial interest(s): E.B.S., V.H., C.D.R., D.P., M.D., C.A., Z.S., J.D., J.N., and S.S.C. possess stock or stock options in Amyris Inc. or TOTAL New Energies USA, Inc.

■ ACKNOWLEDGMENTS

This work was funded by Defense Advanced Research Projects Agency (DARPA) Living Foundries grant HR001-12-3-0006. We wish to thank Don Misumi for maintaining the Biomek robots and modifying the scripts for our needs. We also thank Daniel Hollis, Nalien Vongtharangsy, and other members of the Amyris Automated Strain Engineering (ASE) team for helping with automation and getting the NGS QC process into the ASE workflow.

■ REFERENCES

(1) Weenink, T., and Ellis, T. (2013) Creation and characterization of component libraries for synthetic biology. *Methods Mol. Biol. 1073*, 51−60.

(2) Polizzi, K. M. (2013) What is synthetic biology? *Methods Mol. Biol. 1073*, 3−6.

(3) Munnelly, K. (2013) Engineering for the 21st century: synthetic biology. *ACS Synth. Biol. 2*, 213−215.

(4) Stephanopoulos, G. (2012) Synthetic biology and metabolic engineering. *ACS Synth. Biol. 1*, 514−525.

(5) Gardner, T. S., Hawkins, K. M., Meadows, A. L., Tsong, A. E., and Tsegaye, Y. (2014) *Production of Acetyl-Coenzyme a Derived Isoprenoids*, Patent US 8415136 B1.

(6) Du, Y. L., Williams, D. E., Patrick, B. O., Andersen, R. J., and Ryan, K. S. (2014) Reconstruction of cladoniamide biosynthesis reveals nonenzymatic routes to bisindole diversity. *ACS Chem. Biol. 9*, 2748−2754.

(7) Ajikumar, P. K., Xiao, W. H., Tyo, K. E., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T. H., Pfeifer, B., and Stephanopoulos, G. (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science 330*, 70−74.

(8) Dharmadi, Y., Patel, K., Shapland, E., Hollis, D., Slaby, T., Klinkner, N., Dean, J., and Chandran, S. S. (2014) High-throughput, cost-effective verification of structural DNA assembly. *Nucleic Acids Res. 42*, e22.

(9) Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 456*, 53−59.

(10) Perkins, T. T., Tay, C. Y., Thirriot, F., and Marshall, B. (2013) Choosing a benchtop sequencing machine to characterise *Helicobacter pylori* genomes. *PLoS One 8*, e67539.

(11) Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., and Pallen, M. J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol. 30*, 434−439.

(12) Caruccio, N. (2011) Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation

and adaptor tagging by in vitro transposition. *Methods Mol. Biol. 733*, 241−255.

(13) Lamble, S., Batty, E., Attar, M., Buck, D., Bowden, R., Lunter, G., Crook, D., El-Fahmawi, B., and Piazza, P. (2013) Improved workflows for high throughput library preparation using the transposome-based nextera system. *BMC Biotechnol. 13*, 104.

(14) Reznikoff, W. S. (2008) Transposon Tn5. *Annu. Rev. Genet 42*, 269−286.

(15) Kong, H., Kucera, R. B., and Jack, W. E. (1993) Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis*. Vent DNA polymerase, steady state kinetics, thermal stability, processivity, strand displacement, and exonuclease activities. *J. Biol. Chem. 268*, 1965−1975.

(16) Bystrykh, L. V. (2012) Generalized DNA barcode design based on Hamming codes. *PLoS One 7*, e36852.

(17) Frank, D. N. (2009) BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinf. 10*, 362.

(18) (2014) *Illumina Customer Sequence Letter*, Illumina, Inc., San Diego, CA, http://supportres.illumina.com/documents/documentation/chemistry_documentation/experiment-design/illumina-customer-sequence-letter.pdf.

(19) Markham, N. R., and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res. 33*, W577−581.

(20) Dean, F. B., Nelson, J. R., Giesler, T. L., and Lasken, R. S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res. 11*, 1095−1099.

(21) *Technical Note: Estimating Sequence Coverage*, Illumina, Inc., San Diego, CA, http://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf.

(22) Indap, A. R., Cole, R., Runge, C. L., Marth, G. T., and Olivier, M. (2013) Variant discovery in targeted resequencing using whole genome amplified DNA. *BMC Genomics 14*, 468.

(23) Ason, B., and Reznikoff, W. S. (2004) DNA sequence bias during Tn5 transposition. *J. Mol. Biol. 335*, 1213−1225.

(24) Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol. 12*, R18.

(25) McInerney, P., Adams, P., and Hadi, M. Z. (2014) Error rate comparison during polymerase chain reaction by DNA polymerase. *Mol. Biol. Int. 2014*, 287430.

(26) Hecker, K. H., and Rill, R. L. (1998) Error analysis of chemically synthesized polynucleotides. *Biotechniques 24*, 256−260.

(27) de Kok, S., Stanton, L. H., Slaby, T., Durot, M., Holmes, V. F., Patel, K. G., Platt, D., Shapland, E. B., Serber, Z., Dean, J., Newman, J. D., and Chandran, S. S. (2014) Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol. 3*, 97−106.

(28) Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows−Wheeler transform. *Bioinformatics 25*, 1754−1760.

(29) Ramirez-Gonzalez, R. H., Bonnal, R., Caccamo, M., and Maclean, D. (2012) Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments. *Source Code Biol. Med. 7*, 6.

(30) Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics 25*, 2078−2079.