Applied and Environmental Microbiology

Comparison of Normalization Methods for Construction of Large, Multiplex Amplicon Pools for Next-Generation Sequencing

J. Kirk Harris, Jason W. Sahl, Todd A. Castoe, et al. 2010. Comparison of Normalization Methods for Construction of Large, Multiplex Amplicon Pools for Next-Generation Sequencing. Appl. Environ. Microbiol. 76(12):3863-3868. doi:10.1128/AEM.02585-09.

Applied and Environmental Microbiology, June 2010, p. 3863–3868 Vol. 76, No. 12
0099-2240/10/$12.00 doi:10.1128/AEM.02585-09

Comparison of Normalization Methods for Construction of Large, Multiplex Amplicon Pools for Next-Generation Sequencing▽†

J. Kirk Harris,[1,5]* Jason W. Sahl,[2] Todd A. Castoe,[3,5] Brandie D. Wagner,[4] David D. Pollock,[3,5] and John R. Spear[2]

Department of Pediatrics, University of Colorado School of Medicine, Aurora, Colorado 80045[1]; Department of Environmental Science and Engineering, Colorado School of Mines, Golden, Colorado 80401[2]; Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado 80045[3]; Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, University of Colorado Denver, Aurora, Colorado 80045[4]; and Consortium for Comparative Genomics, University of Colorado Denver, Aurora, Colorado 80045[5]

Received 23 October 2009/Accepted 18 April 2010

Constructing mixtures of tagged or bar-coded DNAs for sequencing is an important requirement for the efficient use of next-generation sequencers in applications where limited sequence data are required per sample. There are many applications in which next-generation sequencing can be used effectively to sequence large mixed samples; an example is the characterization of microbial communities where ≤1,000 sequences per samples are adequate to address research questions. Thus, it is possible to examine hundreds to thousands of samples per run on massively parallel next-generation sequencers. However, the cost savings for efficient utilization of sequence capacity is realized only if the production and management costs associated with construction of multiplex pools are also scalable. One critical step in multiplex pool construction is the normalization process, whereby equimolar amounts of each amplicon are mixed. Here we compare three approaches (spectroscopy, size-restricted spectroscopy, and quantitative binding) for normalization of large, multiplex amplicon pools for performance and efficiency. We found that the quantitative binding approach was superior and represents an efficient scalable process for construction of very large, multiplex pools with hundreds and perhaps thousands of individual amplicons included. We demonstrate the increased sequence diversity identified with higher throughput. Massively parallel sequencing can dramatically accelerate microbial ecology studies by allowing appropriate replication of sequence acquisition to account for temporal and spatial variations. Further, population studies to examine genetic variation, which require even lower levels of sequencing, should be possible where thousands of individual bar-coded amplicons are examined in parallel.

Emergent technologies that generate DNA sequence data are designed primarily to perform resequencing projects at reasonable cost. The result is a substantial decrease in per base costs from traditional methods. However, these next-generation platforms do not readily accommodate projects that require obtaining moderate amounts of sequence from large numbers of samples. These platforms also have per run costs that are significant and generally preclude large numbers of single-sample, nonmultiplexed runs. One example of research that is not readily supported is rRNA-directed metagenomics study of some human clinical samples or environmental rRNA analysis of samples from communities with low community diversity that require only thousands of sequences. Thus, strategies to utilize next-generation DNA sequencers efficiently for applications that require lower throughput are critical to capitalize on the efficiency and cost benefits of next-generation sequencing platforms.

Directed metagenomics based on amplification of rRNA genes is an important tool to characterize microbial communities in various environmental and clinical settings. In diverse environmental samples, large numbers of sequences are required to fully characterize the microbial communities (15). However, a lower number of sequences is generally adequate to answer specific research questions. In addition, the levels of diversity in human clinical samples are usually lower than what is observed in environmental samples (for example, see reference 7).

The Roche 454 genome sequencer system FLX pyrosequencer (which we will refer to as 454 FLX hereafter) is the most useful platform for rRNA-directed metagenomics because it currently provides the longest read lengths of any next-generation sequencing platform (1, 14). Computational analysis has shown that the 250-nucleotide read length (available from the 454 FLX-LR chemistry) is adequate for identification of bacteria if the amplified region is properly positioned within variable regions of the small-subunit rRNA (SSU-rRNA) gene (9, 10).

In this study, we used the 454 FLX-LR genome sequencing platform and chemistry, which provides >400,000 sequences of ~250 bp per run. After we conducted this study, a new reagent set (454 FLX-XLR titanium chemistry) was released, which further increases reads to >1,000,000 and read lengths to >400 bp (Roche). The 454 FLX platform dramatically reduces per base costs of obtaining sequence, and physical separation into between 2 and 16 lanes is available; this physical separation on the plate reduces sequencing output overall, up to 40% com-

* Corresponding author. Mailing address: Department of Pediatrics, University of Colorado Denver, 13123 E. 16th Ave., B395, Aurora, CO 80045. Phone: (720) 777-4943. Fax: (720) 777-7284. E-mail: jonathan.harris@ucdenver.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 23 April 2010.

paring 2 lanes versus 16 lanes. For applications where modest sequencing depth (~1,000 sequences per sample) is adequate to address research questions, physical separation does not allow adequate sample multiplexing because even a 1/16 454 FLX-LR plate run is expected to produce ~15,000 reads. Further, the utility of the platform as a screening tool at 16-plex is limited by cost per run.

A solution to make next-generation sequencing economical for projects such as rRNA-directed metagenomics is to use bar-coded primers to multiplex amplicon pools so they can be sequenced together and computationally separated afterward (6). To successfully accomplish this strategy, precise normalization of the DNA concentrations of the individual amplicons in the multiplex pools is essential for effective multiplex sequencing when large numbers of pooled samples are sequenced in parallel. There are several potential methods available for normalizing concentrations of amplicons included in multiplex pools, but the relative and absolute performance of each approach has not been compared.

In this study, we present a direct quantitative comparison of three available methods for amplicon pool normalization for downstream next-generation sequencing. The central goal of the study was to identify the most effective method for normalizing multiplex pools containing >100 individual amplicons. We evaluated each pooling approach by 454 sequencing and compared the observed frequencies of sequences from different pooled bar-coded amplicons. From these data, we determined the efficacy of each method based on the following factors: (i) how well normalized the sequences within the pool were, (ii) the proportion of samples failing to meet a minimum threshold of sequences per sample, and (iii) the overall efficiency (speed and labor required) of the process to multiplex samples.

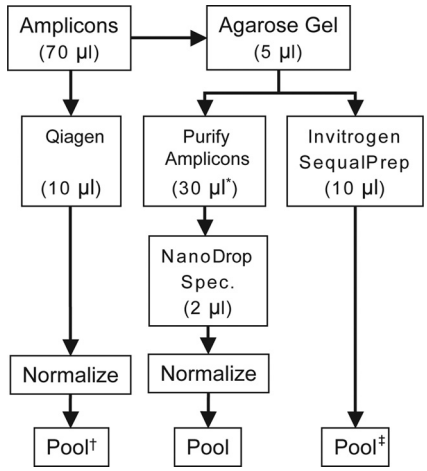

FIG. 1. Schematic of normalization methods. PCR amplicons were generated for 144 different bar codes and processed for normalization and pooling by the three methods shown schematically. The three methods were direct quantification (NanoDrop), size-restricted DNA quantification (Qiagen), and (iii) quantitative DNA binding (SequalPrep kit; Invitrogen). A single pool of PCR product was split and processed by the three normalization methods. Individual amplicons were purified with the Ampure magnetic bead purification kit prior to determination of DNA concentration. †, the Qiagen pool was purified with the same kit prior to sequencing. ‡, the Invitrogen pool was concentrated using the Zymo DNA Clean & Concentrator kit prior to sequencing. Spec., specimen.

MATERIALS AND METHODS

DNA extraction and PCR. A total of 53 DNAs, from a wide diversity of samples, were extracted with the Mobio Powersoil kit (Mobio). Representative samples were from deep phreatic sinkholes (5), deep biosphere borehole fluids (12), and groundwater monitoring wells. PCR was performed with Promega mastermix (Promega) using bar-coded primers targeting bacterial SSU-rRNA genes previously designed for use with the Roche 454 FLX pyrosequencer (6). Of the total 53 DNAs, 44 were amplified with three different bar-coded primers, three samples were amplified with two different bar-coded primers, and six samples were amplified with a single bar-coded primer for a total of 144 amplicon products that were combined in the sequencing pools. The PCR program consisted of an initial denaturation step of 2 min at 94°C and 25 cycles of PCR, with 1 cycle consisting of 30 s at 94°C, 20 s at 52°C, and 60 s at 65°C.

Conventional Sanger sequences were also obtained for many of the DNAs sequenced with pyrosequencing. Methods for PCR, cloning, and sequencing have been published previously (13). Only Sanger sequences obtained using bacterium-specific primers were used for comparison with pyrosequencing results.

Normalization approaches. Three methods were used to construct the amplicon pools: (i) direct quantification (NanoDrop 1000; NanoDrop), (ii) size-restricted DNA quantification (QIAxcel; Qiagen), and (iii) quantitative DNA binding (SequalPrep kit; Invitrogen). Amplicon pool construction requires multiple steps, which include removal of unincorporated primers, quantification of the amplicon products, normalization of the amplicon concentration, and combining an equal amount of each amplicon into a common pool of DNA. Figure 1 shows a schematic comparison of each method tested. All PCR amplicons were checked by electrophoresis on a 1% agarose gel to confirm amplification before amplicon pool construction was initiated.

Direct DNA quantification pool. To construct the direct DNA quantification pool, each individual PCR product (144 in total) was processed to remove unincorporated primers and nucleotides using the Ampure magnetic bead purification kit in 96-well format (Agencourt). The DNA concentration was determined for each purified product by measuring its 260-nm absorbance with a Nanodrop ND-1000 spectrophotometer (Nanodrop, Wilmington, DE). All amplicons were then manually diluted to match the lowest concentration (7.1 ng/μl). From the normalized amplicons, an equal amount of volume was taken from each and added to a single tube to form the final multiplexed pool. The DNA concentration of this pool was determined by spectroscopy to be 10.7 ng/μl with an A_{260}/A_{280} ratio of 2.05.

Size-restricted DNA quantification pool. To construct the size-restricted DNA quantification pool, we used the QIAxcel capillary fragment analysis platform to directly measure the DNA concentration of the amplicons in the size range of 350 to 450 bp (expected amplicon size of ~400 bp). The instrument was run in the low mode (lowest dynamic range, 1 to 10 ng/μl) with the QIAxcel DNA screening kit (2400) cartridge. Each amplicon was then manually diluted to the lowest measured concentration (2.8 ng/μl), and equal volumes of amplicons were combined in a single tube to construct the pool. Unincorporated primers and nucleotides were removed from an aliquot of the Qiagen pool with Ampure magnetic beads (Agencourt). Due to the low concentration expected (~3 ng/μl), 100 μl of the pool was concentrated to 10 μl using the DNA Clean & Concentrator kit (Zymo Research). The DNA concentration of this concentrated pool was checked by spectroscopy and was determined to be ~42 ng/μl with an A_{260}/A_{280} ratio of 1.95.

Quantitative DNA binding pool. We used the recently available SequalPrep kit (Invitrogen) to construct the quantitative DNA binding pool. This product binds approximately the same amount of DNA in each well (25 ng) when DNA is present in excess (≥250 ng recommended). The DNA was normalized per the manufacturer's instructions. The pool was constructed by adding an equal volume of each amplicon (after normalization) to a single tube. Due to the low concentration expected (1 to 2 ng/μl) of the quantitative DNA binding pool, 100 μl of the pool was concentrated to 10 μl, as described above for "Size-restricted DNA quantification pool." The DNA concentration of this concentrated pool was checked by spectroscopy and determined to be 8.2 ng/μl with an A_{260}/A_{280} ratio of 1.92.

Sequence determination. Each amplicon pool was independently sequenced using the Roche genome sequencer system 454 FLX instrument, using the FLX-LR chemistry and reagents (Roche). Emulsion PCR (emPCR) and emPCR cleanup were conducted independently for each pool, using FLX-LR emPCR kit II (Roche). Each pool was sequenced on 1/8 of a subdivided 70- by 75-cm 454 FLX-LR picotiter plate. For a 1/8 plate run, Roche provides an expected number

TABLE 1. Counting statistics for each amplicon pool

Counting statistic	Value for pool constructed by method[a]:		
	Invitrogen	NanoDrop	Qiagen
No. of sequences			
Total	52,896	29,353	39,372
Minimum	195	0	21
Maximum	629	553	1,584
Avg	367.3	203.8	273.4
SD	79.8	107.9	206.7
Maximum/minimum ratio	3.2	>553[b]	75.4
Failure rate (%)	2.1	43.8	62.5

[a] Three methods were used to construct the amplicon pools: (i) direct quantification (NanoDrop 1000; NanoDrop), (ii) size-restricted DNA quantification (QIAxcel; Qiagen), and (iii) quantitative DNA binding (SequalPrep kit; Invitrogen).
[b] The minimum was set at 1 for the purpose of this calculation.

TABLE 2. Processing time for each step of pool construction

Processing step	Time required (min) by method:		
	Nanodrop	Qiagen	Invitrogen
Primer removal	90	(40[a])	
Concentrating DNA	120	(500[b])	90
Adjust DNA concn	(90)	90	
Mixing	10	10	10
Total	310	640	90

[a] Primer removal was performed on a single aliquot of the pooled amplicons by the same procedure used for the individual amplicons in the NanoDrop pool.
[b] This represents manual adjustment of the output for each sample. Software modification would shorten this time significantly.

of reads of ≥30,000. On the basis of this expected number of sequences per 1/8 plate, we calculated the average expected number of sequences per bar-coded amplicon at 209 (30,000/144) under the assumption that the normalization process was perfect.

Community diversity statistics. Sanger sequences and pyrosequences were aligned with the Infernal aligner (11) and clustered at a distance of 0.03 by the furthest-neighbor algorithm implemented in the RDP pyrosequencing pipeline (4). Nonparametric species richness estimates were calculated with the Chao1 estimator (2), which is integrated into the EstimateS software program (version 8; R. K. Colwell, University of Connecticut [http://viceroy.eeb.uconn.edu/EstimateS]). Coverage was calculated by dividing the observed number of operational taxonomic units (OTUs) by the Chao1 estimate and converting this value to a percentage. The coverage of Sanger sequences in the context of pyrosequencing results was calculated by dividing the number of observed OTUs from Sanger sequencing by the pyrosequencing Chao1 estimate.

RESULTS

In total, 121,700 sequences were determined from 144 amplicon products to contain legitimate recognized bar codes and were assigned to the associated amplicon sample. An additional 2,800 sequences (2.25%) did not contain correct bar codes and were excluded from the analysis; the data for all bar codes are provided in Table S1 in the supplemental material. The number of sequences determined for each multiplexed sample ranged nearly 2-fold in the three experiments, from ~30,000 (the manufacturer's expectation) to ~54,000 (Table 1). Given the variation in actual sequences obtained per experiment, we recalculated the appropriate expectations of reads per sample for each pool based on the actual reads obtained for each sample (Table 1).

The quantitative DNA binding method (Invitrogen SequalPrep) yielded the best-normalized pool, with the tightest range of numbers of sequences per bar-coded amplicon across the multiplexed pool (Table 1). The range of sequences observed was approximately 3-fold, which is consistent with the product description from Invitrogen. This pool also had the highest number of sequences of the three pools, but the significance of this result is not immediately clear; speculatively, it could mean that the pool was higher quality due to greater removal of PCR contaminants, such as primer dimers, oligonucleotides, and nucleotides that often lead to lower sequence yield on the 454 FLX. The Invitrogen SequalPrep plate kit was also the most rapid and efficient means of constructing the amplicon pool. This method accomplishes most of the steps required for amplicon pool construction in a single operation (DNA binding), which requires limited hands-on time (Fig. 1 and Table 2). The time required to construct the 144-member pool was approximately 1.5 h, which is significantly shorter than for the other pools (Table 2). Furthermore, additional amplicons would not dramatically extend the time required for normalization; thus, the technique scales well with the number of samples to be multiplexed.

Comparison of the triplicate sample libraries (samples included multiple times in the pool with different bar codes) demonstrated variability in the methods, providing another perspective on the precision of normalization accomplished by each method. In this analysis, the lower the coefficient of variation (CV), the better the triplicate samples were normalized. Thus, a lower CV represents a greater precision of normalization. The quantitative DNA binding pool method (SequalPrep kit; Invitrogen) performed the best of the three methods with >40% of the triplicate samples having a CV of less than 10% (Fig. 2). In contrast, only about 12% performed this well in the direct DNA quantification pool (NanoDrop 1000; NanoDrop), and none were under 10% CV in the size-restricted DNA quantification pool (QIAxcel; Qiagen). In further support of the performance of the quantitative DNA binding pool method, 78% of the samples had a CV under 20%; this is in strong contrast to the 35% for the direct DNA quantification pool method and 2.4% for the size-restricted DNA quantification pool method with CV values less than 20%.

Microbial community analysis. One of the applications of sample multiplexing is the characterization of microbial communities in complex and diverse environments. The ability to cheaply obtain thousands of sequences per sample using next-generation sequencing methods enables substantial insight into microbial diversity in systems where a limited number of Sanger sequences cannot yield insight. In this study, pyrosequences and Sanger sequences were obtained for 18 DNAs included in the normalization experiment described here (12). Table 3 contains a summary of bacterial diversity estimates observed for these 18 environments. Diversity estimates using Chao1 (3) suggest that there is additional diversity present in all samples, but the level of diversity suggested by the pyrosequencing libraries is much higher. On average, Sanger sequencing identified 30.5% (range, 10.8 to 45.5%) of the estimated diversity, and pyrosequencing identified 42.6% (range,

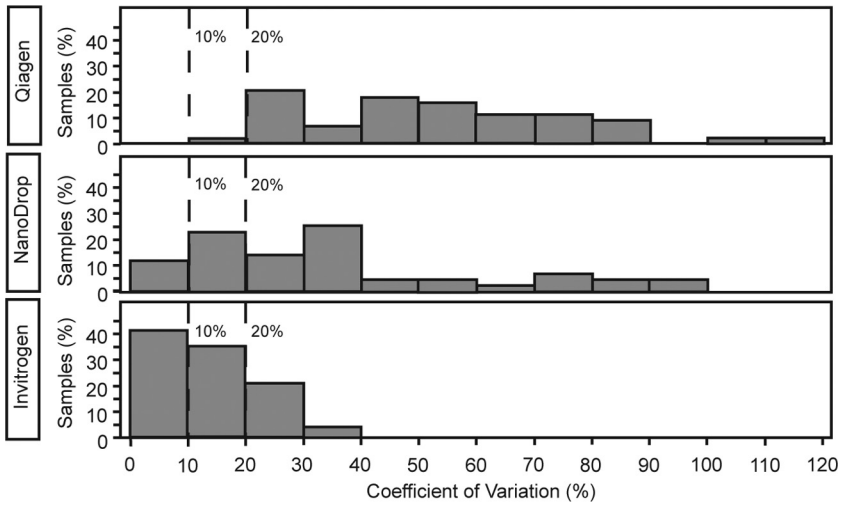

FIG. 2. Coefficient of variation from triplicate samples by the three methods. The three methods were direct quantification (NanoDrop), size-restricted DNA quantification (Qiagen), and (iii) quantitative DNA binding (Invitrogen). The coefficient of variation calculated from sequence counts for samples amplified with three independent bar codes ($n = 44$) were calculated to show how well the sequence counts agreed. A lower CV represents tighter clustering of the counts.

28.9 to 59.3%). The Sanger estimates when weighed in the context of the higher numbers identified by pyrosequencing results in observation of an average of 4.3% (range, 0.6 to 22.9%) of the apparent diversity.

Taxonomic composition of replicate libraries. To examine the taxonomic diversity estimates between triplicate libraries, we determined the percentage of sequences that were in OTUs from all three libraries for each sample. An average of 37% (range, 15 to 90%) of sequences belonged to OTUs shared across the triplicate libraries. As expected, diversity of each environment correlated with the observed agreement. We also observed a similar trend for OTUs when only a single sequence was observed in each library.

DISCUSSION

A substantial hindrance to the effective utilization of next-generation sequencing technology has been the lack of methods for accomplishing projects that require a smaller number of sequences from a large number of samples. It is expected, however, that large collections of samples will become commonplace as high-throughput DNA sequencing approaches

TABLE 3. Microbial richness estimation for each DNA sample

DNA sample	Sanger sequencing				% Chao estimate observed (Sanger/pyrosequencing)[b]	Pyrosequencing			
	No. of sequences	No. of OTUs	Diversity estimate using Chao1	% Chao estimate observed[a]		No. of sequences	No. of OTUs	Diversity estimate using Chao1	% Chao estimate observed[c]
1	75	38	88	43.18	2.30	1,761	655	1,654	39.60
2	96	69	289	23.88	3.27	1,818	873	2,107	41.43
3	94	17	45	37.78	0.64	2,338	1,185	2,674	44.32
4	60	64	180	35.56	4.77	1,303	616	1,342	45.90
5	37	36	333.5	10.79	1.38	1,956	992	2,613	37.96
6	74	53	139.7	37.94	3.09	2,120	775	1,714	45.22
7	92	11	39	28.21	2.59	2,109	192	424	45.28
8	89	60	273	21.98	2.26	3,680	1,209	2,659	45.47
9	93	55	213	25.82	1.82	1,898	1,129	3,025	37.32
10	80	58	271	21.40	2.53	2,119	1,056	2,295	46.01
11	85	53	140	37.86	1.28	3,948	1,753	4,149	42.25
12	62	36	148	24.32	1.58	2,705	901	2,280	39.52
13	73	66	361	18.28	2.21	1,459	923	2,991	30.86
14	90	25	55	45.45	1.39	2,375	519	1,795	28.91
15	68	54	178	30.34	2.34	2,689	960	2,312	41.52
16	81	55	228	24.12	7.71	3,150	324	713	45.44
17	92	21	51	41.18	22.93	540	46	91.6	50.22
18	85	20	48	41.67	13.33	505	89	150	59.33

[a] Number of OTUs in Sanger sequencing divided by Sanger Chao1 (percent).
[b] Number of OTUs in Sanger sequencing divided by pyrosequencing Chao1 (percent).
[c] Number of OTUs in pyrosequencing divided by pyrosequencing Chao1 (percent).

become more frequently used in clinical and environmental research. Combined with innovations in massively multiplexed bar coding, efficiently accomplishing the precise normalization of pooled samples is of critical importance to enable a wide range of sequencing experiments. For example, the efficiency and relative precision of normalized amplicon pool construction can significantly affect the tractability, in terms of labor and cost, of a wide array of next-generation sequencing projects. Ultimately, if this process can be accomplished rapidly, cheaply, and precisely, the maximum benefit of the extreme low costs of next-generation sequencing can be fully realized.

The experiment described in this study is a model for the application of the 454 FLX platform as a screening tool for microbial ecology, although the methods could be easily transferred to any other platform by changing the platform-specific primers. It is now possible to economically obtain >1,000 sequences from hundreds of samples in parallel, which provides the ability for more robust experimental design for ecological studies to incorporate additional samples including temporal and spatial variation and replicates. When this study was conducted, it was possible to obtain >1,000 sequences from 300 to 400 samples in a single experiment (on a full 454 FLX-LR plate run). The recent release of the 454 FLX-XLR titanium chemistry has increased this to around 1,000 samples (each with 1,000 sequences) per 454 FLX-XLR run. There is no inherent limit to the number of amplicons that can be multiplexed, but the overhead to track the multiplexing will affect the efficiency of projects at large numbers.

The bacterial diversity detected in the 18 samples where triplicate bar codes were used demonstrated the expected need for larger numbers of sequences to describe the communities more completely. Richness increased on average by approximately 20-fold (Table 3). We also found examples that suggested that using ~1,000 sequences per sample was fairly adequate; in three samples, the increase in richness was 5-fold or less when richness estimates based on Sanger sequences versus 454 pyrosequences were compared. Further, there is evidence that in some cases the Sanger data, based on longer amplicons, did greatly underestimate the richness of the sample (Table 3, DNA sample 3); this has also been observed by others (8). These data imply that different questions regarding microbial community diversity and different types of samples will require different levels of sequence coverage to adequately address experimental hypotheses. The abilities to massively multiplex and to normalize a large pool of samples are thus important for achieving appropriate numbers of sequences per sample, without strict dependence on the physical sample separation abilities of the instrument.

In this comparative study of amplicon pool normalization methods, the quantitative DNA binding approach based on the Invitrogen SequalPrep plate kit outperformed the other two methods in all areas evaluated. It provided the best normalized pool, producing an extremely precise range of normalized amplicons with approximately 3-fold range in difference in sequence counts (Table 1). The Invitrogen SequalPrep method outperformed the other two methods, including having the lowest failure rate, lowest standard deviation of sequence counts, lowest CV for triplicate samples, highest efficiency, and limited hands-on time (also amenable to automation). The procedure also seems quite robust. For example, the amount of PCR product added likely fell below the recommended minimum (250 ng) for a subset of the amplicons in this study, which suggests that the binding step remains sufficiently quantitative at amounts below the recommended minimum amount of starting amplicon material.

TABLE 4. Normalization performance of additional multiplex amplicon pools

Pool	Sequencing effort	Plex	Maximum/minimum[a]
1	Full plate	288	16.4
2	1/8 plate	96	7.0
3	1/8 plate	89	6.1
4	1/8 plate	49	2.6

[a] Maximum number of sequences divided by minimum number of sequences.

One potential issue that was not addressed by this experiment is the impact of variable DNA template concentrations over a wide range, which could interfere with normalization due to variation in template DNA binding. This potential interference is an important factor in some clinical applications where the ratio of human DNA to bacterial DNA is high, which limits the ability to normalize input DNA template concentrations during PCR. Initial experiments suggest that the human DNA increases the normalization range about 10-fold, which is consistent with the range of template concentrations used in the amplification step. We constructed multiple pools (49- to 288-plex) from clinical samples using the quantitative DNA binding method and found that the range of sequences within a pool is generally less than 20-fold (Table 4).

The direct DNA quantification pool method is representative of prior published normalization approaches (6). This approach should perform well, but there are several issues with this approach that are difficult to overcome. First, there are a large number of pipetting steps, which will introduce error at each step. Robotic liquid handling may reduce the pipetting error to some extent, but it will not completely eliminate error due to multiple pipetting steps. Further, the DNA concentrations determined include any double-stranded DNA present, which does not accurately reflect the amount of target amplicon (e.g., template DNA).

The size-restricted DNA quantification pool method was the poorest performing normalization method in this study. Before the normalization step, we were particularly enthusiastic about this method as a means to overcome problems caused by highly variable initial DNA concentrations that focus normalization on the target amplicon. This occurs because the method calculates the concentration of a discrete window of amplicon size around the expected amplicon product size and ignores the quantity of DNA or template DNA outside this amplicon size window. This would allow for the quantification of DNA prior to cleaning up, which would lower the cost of pool construction by eliminating the need to remove primers and other nontarget amplification products (e.g., primer dimers) from each individual sample. The automation of the quantification step is another attractive feature of this approach. Multiple issues are likely involved in the poor performance. One issue is the need for multiple pipetting steps to normalize the DNA concentration manually. Another problem is the dynamic range setting

of the QIAxcel capillary fragment analysis platform, which was not optimal for all the amplicons included in the pool and therefore likely introduced error. Finally, the software used by the instrument was not intended for quantifying DNA over broad peaks (as opposed to distinct peaks) and required manual adjustments to all lanes, which could have introduced error. This manual adjustment accounted for the majority of time required to construct this pool. Software modification would dramatically reduce the time required for this step. With these adjustments, the QIAxcel platform would provide a good method for product confirmation and automated initial quantification of only the targeted amplicon.

The use of multiplex bar-coded amplicon pools substantially extends the utility and cost-effectiveness of the Roche 454 FLX pyrosequencer, as it would any high-throughput platform. This approach allows multiplex data acquisition at higher and more-flexible multiples than physical separation supported by the manufacturer's gasket-based sample separation options. Bar codes therefore provide the ability to obtain the maximal sequence reads in a run by not occluding any of the usable space on the sequencing plate by the overlying gasket. Another advantage is that the bar code also allows for internal quality control for sample tracking and contamination, including detecting instances where sequencing templates (bound to beads) are accidentally transferred between physically separated regions of the 454 picotiter plate due to gasket misplacement or poor sealing.

The multiplex design has several attractive features. First, the initial sequence data provides adequate information to assess the quality of the amplicon pools and to estimate the required level of sequencing needed for adequate coverage from each sample. Further, secondary pools can be constructed from the remaining normalized amplicons that are mixed at nonequimolar ratios to obtain different target numbers of sequences from each sample. This approach allows for optimal acquisition of predetermined levels of sequence data from each sample in the amplicon pool. The abilities to track coverage statistics and to collect adequate data to justify comparisons between samples are critical for comparisons between samples with ecological statistics. Thus, the bar-coded amplicon approach allows rational experimental design of microbial ecology experiments, whether in environmental samples or clinical samples, that greatly increases a researcher's ability to focus resources more effectively and efficiently.

ACKNOWLEDGMENTS

We acknowledge the support of the National Institutes of Health (NIH; R01 GM083127) to D.D.P. and an NIH training grant (LM009451) to T.A.C. Additional support was provided by NIH grant U01HL081335-01 and Cystic Fibrosis Foundation grant Harris08A0. J.R.S. is supported by a National Science Foundation grant (068282) and a U.S. Air Force Office of Scientific Research grant (R-8196-G1).

REFERENCES

1. **Ansorge, W. J.** 2009. Next-generation DNA sequencing techniques. N. Biotechnol. **25:**195–203.
2. **Chao, A.** 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics **43:**783–791.
3. **Chao, A.** 1984. Non-parametric estimation of the number of classes in a population. Scand. J. Stat. **11:**265–270.
4. **Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37:**D141–D145.
5. **Gary, M. O., and J. M. Sharp, Jr.** 2006. Volcanogenic karstification of Sistema Zacatón, Mexico, p. 79–89. *In* R. S. Harmon and C. M. Wicks (ed.), Perspectives on karst geomorphology, hydrology, and geochemistry: a tribute volume to Derek C. Ford and William B. White. GSA Special Paper 404. The Geological Society of America, Boulder, CO.
6. **Hamady, M., J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight.** 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods **5:**235–237.
7. **Harris, J. K., M. A. De Groote, S. D. Sagel, E. T. Zemanick, R. Kapsner, C. Penvari, H. Kaess, R. R. Deterding, F. J. Accurso, and N. R. Pace.** 2007. Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis. Proc. Natl. Acad. Sci. U. S. A. **104:**20529–20533.
8. **Huber, J. A., H. G. Morrison, S. M. Huse, P. R. Neal, M. L. Sogin, and D. B. M. Welch.** 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ. Microbiol. **11:**1292–1302.
9. **Liu, Z., T. Z. DeSantis, G. L. Andersen, and R. Knight.** 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. **36:**e120.
10. **Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight.** 2007. Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. **35:**e120.
11. **Nawrocki, E. P., D. L. Kolbe, and S. R. Eddy.** 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics **25:**1335–1337.
12. **Sahl, J. W., N. Fairfield, J. K. Harris, D. Wettergreen, W. C. Stone, and J. R. Spear.** 2010. Novel microbial diversity retrieved by autonomous robotic exploration of the world's deepest vertical phreatic sinkhole. Astrobiology **10:**201–213.
13. **Sahl, J. W., R. Schmidt, E. D. Swanner, K. W. Mandernack, A. S. Templeton, T. L. Kieft, R. L. Smith, W. E. Sanford, R. L. Callaghan, J. B. Mitton, and J. R. Spear.** 2008. Subsurface microbial diversity in deep-granitic-fracture water in Colorado. Appl. Environ. Microbiol. **74:**143–152.
14. **Shendure, J., and H. Ji.** 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26:**1135–1145.
15. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. **103:**12115–12120.