Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 617469, 8 pages
doi:10.1155/2010/617469

Methodology Report

Semi-Automated Library Preparation for High-Throughput DNA Sequencing Platforms

Eveline Farias-Hesson,[1] Jonathan Erikson,[1] Alexander Atkins,[1] Peidong Shen,[2] Ronald W. Davis,[2] Curt Scharfe,[2] and Nader Pourmand[1]

[1] *Department of Biomolecular Engineering, University of California, 1156 High Street, Santa Cruz, CA 95064, USA*
[2] *Stanford GenomeTechnology Center, Stanford University, 855 S. California Avenue, Palo Alto, CA 94304, USA*

Correspondence should be addressed to Nader Pourmand, pourmand@soe.ucsc.edu

Received 26 January 2010; Accepted 5 April 2010

Academic Editor: Lori Snyder

Next-generation sequencing platforms are powerful technologies, providing gigabases of genetic information in a single run. An important prerequisite for high-throughput DNA sequencing is the development of robust and cost-effective preprocessing protocols for DNA sample library construction. Here we report the development of a semi-automated sample preparation protocol to produce adaptor-ligated fragment libraries. Using a liquid-handling robot in conjunction with Carboxy Terminated Magnetic Beads, we labeled each library sample using a unique 6 bp DNA barcode, which allowed multiplex sample processing and sequencing of 32 libraries in a single run using Applied Biosystems' SOLiD sequencer. We applied our semi-automated pipeline to targeted medical resequencing of nuclear candidate genes in individuals affected by mitochondrial disorders. This novel method is capable of preparing as much as 32 DNA libraries in 2.01 days (8-hour workday) for emulsion PCR/high throughput DNA sequencing, increasing sample preparation production by 8-fold.

1. Introduction

Next-generation sequencing technologies such as Applied Biosystems/SOLiD, Roche/454, and Illumina/Solexa Genome Analyzer have revolutionized genomic research. Furthermore, the applications of these technologies have greatly expanded the ability of researchers to use genomic information to solve biological and clinical questions by generating gigabases of data in single sequencing runs [1–4]. Next-generation sequencing has spawned diverse applications such as the 1000 Genomes Project [5], RNA sequencing of a single cell [6], epigenetic [7], and transcriptome profiling [8]. Sample preparation is an essential step in the DNA sequencing process; however, it has been a bottleneck to more cost-effective and time-efficient applications of these technologies.

The steps involved in using a next-generation sequencing platform are: (1) DNA library preparation (including shearing the DNA to desired size, end-polishing, adaptor ligation, nick-translation- amplification, and gel purification of libraries); (2) quantification of the product from step one; (3) emulsion PCR or bridge amplification (Solexa); (4) depositing templated beads onto the instrument for sequencing. Steps (2)–(4) can be performed quickly and efficiently in pooled fashion. However, DNA library preparation is time-consuming and requires highly trained and qualified personnel to perform nearly 60 substep operations to prepare usable samples. For example, an average laboratory technician may take as much as 12 hours to prepare just one sample or up to 4 samples in parallel without increasing the risk of making mistakes. Streamlining this step could dramatically expedite the sequencing pipeline.

One means of expediting sample preparation is to enable automated multiplexing and pooling of several small genomes or samples for a single sequencing run. This would allow for studying hundreds of target sequences in hundreds of individuals [9]. Currently available sample preparation protocols process one sample at a time and rely heavily on spin column purification technologies for isolating DNA. This labor-intensive system is not suitable

for automation because it requires multiple centrifugation steps. Furthermore, the purification processes as currently performed can result in significant reductions as well as variability in DNA yield, limiting preparation of samples in which generally relatively low DNA amounts are available. To overcome this problem we substituted ~2.8 μM carboxylated magnetic beads in the purification steps where spin column purification technologies are used. This simple and inexpensive reagent allows rapid automated purification, requiring minimal labor, and reagent input.

Automated sample preparation has significant advantages over the manual preparation of samples for next-generation sequencing. Through automation, human error can be reduced and experimental costs can be lowered at the same time. Additionally, automation may significantly eliminate the variability found in the manual processing, providing identical conditions to create a more reproducible process [10].

In this paper, we describe how we devised and tested an approach to overcome limitations in DNA sample preparation for high-throughput DNA sequencing platforms. First, we developed a parallel semi-automated library maker (PSALM) to replace the manual sample preparation protocols of sample processing for the Applied Biosystems' SOLiD next-generation sequencer by using carboxylated magnetic bead technology in conjunction with a liquid-handling robot. We then applied PSALM to study nuclear-encoded mitochondrial genes associated with hereditary disorders that we prioritized from a recent study [11]. We show that our approach can be used for high-throughput DNA library construction and sequencing of hundreds of target sequences. Our automated sample preparation provides reproducibility and high coverage for most targeted sequences. Achievement of this high quality and throughput is a prerequisite for cost-effective medical re-sequencing of disease candidate genes in phenotyped populations.

2. Methods

2.1. Selection of Mitochondrial Candidate Genes and DNA Sample Preparation. We selected thirty-nine candidate genes based on subcellular localizations of their gene products to human mitochondria, and the association of these genes with mitochondrial disorders [11] (Tables S2, and S3). We identified a total of 438 exons (211,841 bp) for these genes that were PCR-amplified with Oligonucleotide primer pairs (designed using Primer3: http://primer3.sourceforge.net/) and AmpliTaq Gold DNA Polymerase (Applied Biosystems/Foster City/CA/USA) using established PCR protocols. Shorter exons separated by short introns were paired and amplified together as one amplicon, and longer exons (>600 bp) were amplified using multiple overlapping amplicons.

Our study included 32 samples that represented 25 individuals with mitochondrial disorders and 7 healthy controls ($211,841\,\text{bp} \times 32\ \text{samples} = 6.77\,\text{Mb}$). In order to obtain sufficient amounts of starting material from small amounts of genomic DNA (~100 ng), we performed whole genome amplification (WGA) using the REPLI-g Mini Kit (Qiagen, Inc., CA) following Qiagen's protocol. DNA amounts were quantified using PicoGreen reagent (Invitrogen, Inc., CA). The PCR amplicons for each sample were inspected by 1.2% Agarose gel electrophoresis, pooled together in equimolar amounts, and then purified using the QIAquick PCR Purification Kit (Qiagen, Inc., CA).

2.2. Semi-Automated Barcoded Fragment Library for SOLiD Sequencing

2.2.1. Parallel Semi-Automated Library Maker (PSALM). Thirty two samples (25 patients and 7 controls), each containing 438 gene exons, were used to prepare 32 barcoded fragment libraries. Sixteen samples were simultaneously processed in parallel in a 96-well plate using a Magnatrix 8000 plus Liquid-handling Robot (NorDiag, Oslo, Norway). This liquid-handling robot has two peltier heating/cooling units for incubations and a retractable pipette tip magnet system for magnetic separations.

The parallel semi-automated library maker (PSALM) employs a combination of automated and manual approaches for library preparation. The end-repair, ligation of molecular barcoded adaptors and purification reactions required for ligating SOLiD adaptors to the processed DNA were done in an automated fashion, while gel purification steps and library amplification were performed manually (Figure 1).

A total of 250 ng of each sample was sheared (50–300 bp) using the following program: 20% duty cycle, 5 intensity, and 200 cycles per burst for 8 minutes at 5°C in the Covaris S2 system (Covaris, Inc. Woburn, MA). Shearing was done in a 50 μL reaction in micro tubes containing nuclease-free water and 10 mM TE.

Automated Steps of PSALM. Each sheared DNA was end-repaired by adding 7 μL 10X End-it Buffer, 7 μL 10 mM ATP, 7 μL 2.5 mM dNTPs, and 1 μL enzyme mix using the End-It DNA End-Repair Kit (Epicentre, Madison, WI). The enzymatic reaction was transferred from the Peltier unit and mixed with samples in the 96-well plate containing 2 rows with 8 samples per row. After 30 minutes incubation at room temperature, with 2 mixes in the middle of the incubation step, a purification reaction was performed using a Carboxy Clean up kit according to the manufacturer's instructions (NorDiag, Oslo, Norway).

The purification step consisted of adding to the end-repair reaction, three volumes (210 μL) of a binding solution containing 15% (v/v) of 2XBinding Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl) and 85% (v/v) ethanol. 100 μG of carboxy beads were mixed with this solution and incubated for 15 minutes at room temperature, with 2 mixes in the middle of the incubation step. Beads containing the attached DNA were then magnetically collected. The collected beads were washed twice in 45 μL 70% (v/v) ethanol. The DNA was eluted by mixing the beads in 50 μL 10 mM Tris (pH 8.0) for 1 minute (Ambion, Foster City, CA).

The end-repaired DNA of each library was ligated to both SOLiD P1 and one of the 16 multiplex adaptors P2 (IDT, Coralville, IA) using the Quick Ligase Kit (NEB, Ipswich,

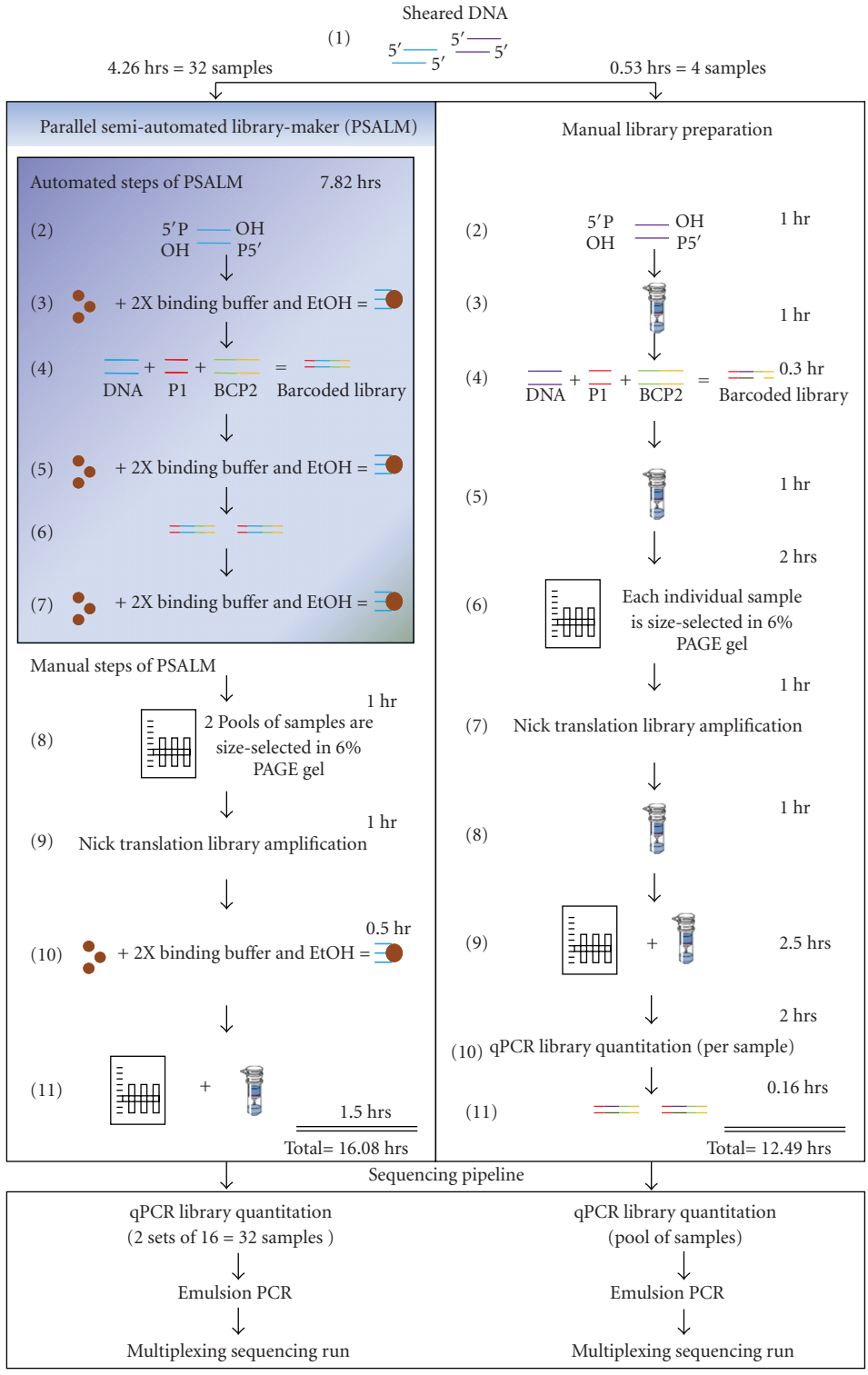

Figure 1: The Parallel Semi-Automated Library Maker (PSALM) uses a NorDiag Magnatrix 8000 plus to prepare 32 samples simultaneously. DNA of 16 samples is sheared (1) and submitted to the automated steps of PSALM (Highlighted in Blue). After end-repair (2), and ligation of molecular barcoded adaptors (4), the 16 samples are pooled (6) and stored at 4°C or −20°C. A new automated cycle is started to prepare additional 16 samples. After 12.08 hours, the 2 pools containing 32 samples enter the manual steps of PSALM for size selection in 6% PAGE gel (8), nick-translation, and PCR amplification (9). A final gel purification step is done in 4% Agarose gel for removing self-ligated adaptors. Next, the library is eluted from the gel using spin columns (11). Carboxy beads substitute for columns used in the manual library preparation to clean up enzymatic reactions (3, 5, and 10), and in the steps where DNA concentration is required (7). Thirty-two libraries are prepared to enter the sequencing pipeline in 16.08 hours. In contrast, only 4 samples are processed in parallel in the manual preparation. Spin columns are used in all purification steps (3, 5, 8 and 9), and a pool with only 4 barcoded libraries is obtained after 12.49 hours (11).

MA). (see Table 1 for supplementary information.) The ligation was achieved by adding 100 μL 2X Quick Ligase Buffer, 5 μL Quick Ligase Enzyme, 0.3 μL 50 pmol/μL of each adaptor P1 and P2, and 45.56 μL nuclease free water. The ligation reagents were transferred from the Peltier unit and mixed with the samples. After incubating for 10 minutes at room temperature, with 2 mixes in the middle of the incubation step, a purification step was performed by adding three volumes (600 μL) of binding solution. Carboxy beads were mixed with this solution and incubated for 15 minutes. at room temperature, with 2 mixes in the middle of the incubation step. Beads containing the attached DNA were magnetically collected and washed twice in 45 μL 70% (v/v) ethanol. The DNA was eluted by mixing the beads in 50 μL 10 mM Tris (pH 8.0) for 1 minute.

After completing the first automated cycle, in which 16 samples were processed, the libraries were organized so that samples containing barcodes 1–16 could be pooled in a set. After purification samples were stored at 4°C or −20°C, and a new plate was used to process the 16 additional samples. Two sets (Sets A, and B), each with 16 barcoded libraries, were obtained after pooling 32 libraries. Each pool of libraries was individually purified using carboxy beads and entered into the manual steps of the PSALM.

Manual Steps of PSALM. The two pools of libraries were resolved in a 6% polyacrylamide gel (Invitrogen, Inc., CA), and a150–200 bp fraction was excised from the gel, shredded, and subjected to 10 cycles of PCR amplification. The number of cycles was determined by both the ability to visualize the amplified product in a 2.2% FlashGel (Lonza) and the ability to yield enough products for performing a gel purification of the barcoded fragment libraries. The amplified products were purified and resolved in 4% Agarose gel for removing self-ligated adaptors. The excised fragment, 150–200 bp, was purified with the QIAquick Gel Extraction kit according to manufacturer instructions (Qiagen, Valencia, CA).

2.2.2. Robot Accuracy and Analysis of the Yield of DNA Recovered by Carboxy Beads. Samples containing DNA were purified using both MinElute Reaction Cleanup kit (Qiagen, Valencia, CA) and Carboxy Cleanup kit (NorDiag, Oslo, Norway). The MinElute Reaction Cleanup kit was used for manually preparing libraries, while the Carboxy Cleanup kit was used in the automated system. Bead purification was done in the liquid-handling robot. The amount of DNA in the sample before and after the purification process was measured using the DNA 1000 assay for the 2100 Bioanalyzer (Agilent, Foster City, CA).

A TaqManGene Expression Assay was performed to determine the accuracy of the library preparation process by PSALM. SOLiD libraries prepared by the eight channels of our liquid-handling robot were diluted to 50 pg/μL and used in 20 μL real-time PCR reactions. All reactions were performed in triplicate. The SOLiD TaqMan Assay (Ac00010015_a1) was done following the manufacturer recommendations for quantitating SOLiD libraries (Applied Biosystems, Foster City, CA) in a MX3005 (Stratagene, La Jolla, CA).

2.3. Sequencing Pipeline. Once libraries were available, emulsion PCR reactions were performed by mixing 280 pg of each library pool with 1.6 billion P1 beads with primers covalently attached to their surfaces. The barcoded and 50-base sequences were obtained using SOLiD3 multiplexing sequencing (Applied Biosystems, Foster City, CA). The barcoded samples were sequenced using a 4-well partitioned slide.

2.4. Bioinformatics Analysis. SOLiD reads were mapped against the human genome and the targeted 438 exon sequences using the BLAST-Like Alignment Tool (BLAT) strict matching parameters according to Kent, 2002 [12].

3. Results

3.1. Parallel Semi-Automated Library Maker (PSALM). In this study we processed 32 samples from 25 individuals affected by mitochondrial disorders and 7 healthy controls.

The time presented in Figure 1 for both our parallel semi-automated library maker (PSALM) and the manual approach was calculated considering all steps required in the library preparation, including the hands-on time for setting up plates containing samples and reagents for the automated steps of PSALM. Once all plates used in the automated steps of PSALM were set up, the automated steps were performed unsupervised up to the point where plates containing the 70% ethanol required in the purification steps were placed in the robot (Figure 1: steps 3, 5 and 7 of PSALM). This procedure reassures us that there was no modification in the concentration of the ethanol used for washing the beads due to evaporation while plates were sitting in the instrument. The hands-on time for placing the 70% ethanol plates in the instrument is 15 minutes for all 3 purification steps.

We simultaneously prepared 32 different samples in 16.08 hours using PSALM, increasing library production in 8-fold (Figure 1).

The automated steps of PSALM allow us to prepare 16 samples simultaneously in 6.04 hours (2.13 hours for shearing the DNA and 3.91 hours in the automated steps of PSALM). Once an automated cycle is completed and a set of 16 samples are ready for the gel size-selection step, the samples are pooled, purified and stored at 4°C or −20°C. Because the robot requires very little hands-on time, 16 additional samples had their DNA sheared and prepared for the next automated cycle of PSALM. Two pools (A and B) containing 32 samples were ready for gel size selection in 12.08 hours or 1.51 days (8-hour workday). In addition to reducing the amount of reagents used by reducing the number of samples for the steps that follow in the library preparation, one great advantage of working with two pools of samples instead of individual samples is that less human errors occur in this laborious step of the sample preparation. During the gel size selection, gel slices in the size range corresponding to the adaptor ligated DNA were cut, and the excised pieces were shredded and used in the library amplification step. Working with a higher number of samples increases the chances of the technician mistakenly mixing

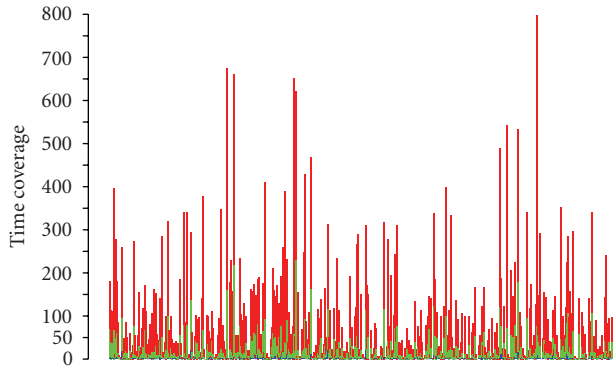

Figure 2: Coverage values for each exon across all groups. Each bar has three colors. Blue shows the minimum coverage value found across all data sets. Green shows the average of the coverage values over all the data sets, and red shows the maximum coverage value from all the data sets.

samples. By using PSALM, samples are primed to enter the sequencing pipeline in 2.01 days (8-hour workday).

In our experience, a trained technician using the manual sample preparation takes over 12.49 hours or 1.56 workdays (8-hour workday) to prepare 4 samples simultaneously (Figure 1). A trained technician would take nearly 6.2 days to prepare 16 libraries, and 12.49 workdays to prepare 32 libraries (data not shown). Based on these estimates, 11 workdays are saved for working on the sequencing pipeline or other projects.

To automate the library preparation, most purification steps that use spin column technology in the manual library preparation were replaced by a carboxy-terminated bead technology. The optimisation of our pipeline included an analysis of the DNA yield recovered using the bead-based purification approach, as well as an analysis for the accuracy of PSALM for preparing libraries.

While the highest DNA yield obtained using the column approach was 81% each time they were used in the manual library preparation, the bead-based purification approach recovers an average of 93% of the starting material (Figure 3(a)).

In addition, the semi-automated system has the advantage of accurately reproducing the library preparation process. Seventy-five percent of the libraries prepared using PSALM were amplified with the same number of PCR cycles (Average CT value = 15.5), indicating a similar concentration of libraries in the tested samples (Figure 3(b)).

The DNA concentration of the each pool set (A and B) containing 16 samples obtained using the PSALM approach was determined in a qPCR reaction. Set A had 19.58 ng/μL, while pool B had 20.16 ng/μL.

3.2. Sequencing of Samples from Sets A and B. Two sets of samples were sequenced in a SOLiD3 Multiplexing run. The reads obtained for samples present in sets A and B were used for mapping against the 438 targeted exons.

Examining the color space quality values (Figure S1), a slight decline can be seen in the quality value (QV) after base 34. These color space QVs are still good, and looking at the base space QVs we can see the base calling quality of all the reads. Low points at the beginning and end of the base space quality are expected because SOLiD analyzes emissions from transitions between the bases. Since there is no transition before base1 or after base 50, these values are lower.

The sequencing from the two different sample sets (A and B) produced a total of 137,594,672 reads, each 50 base pairs long. We then took these approximately 137 million reads and removed any reads containing a poly-N (unknown nucleotide) sequence. This removed 32,331,834 (23%) of reads, leaving 105,262,838 valid reads.

Valid reads were mapped against the human genome and the exon sequences using BLAT. Of the 105 million reads, 39,963,500 mapped uniquely to the 438 exons, 8,151,977 mapped to the exons as well as to other locations in the genome, and 2,245,206 reads mapped to the human genome, but not to one of the 438 exons. Considering all groups, approximately 40% of reads mapped uniquely to the desired exons and out of all mapping reads.

Consistent with other studies, we observed extremely high coverage at the 5′ and 3′ ends on the exons (Figure S4). This bias reflects over representation of the amplicon ends in the DNA samples after fragmentantion prior to library preparation, and primer bias over expressing the 3′ and 5′ ends The overrepresentation of amplicon end sequences is not only wasteful for the sequencing yield but also decreases the expected average coverage depth across the targeted intervals [13]. Because this phenomenon could have a larger effect on the sequence coverage, we defined the coverage for an amplicon as the minimum base coverage across every base in this amplicon.

In addition, there is often a bias against GC-rich areas during PCR amplification [14], which leads to sequencing bias. In order to predict the expected GC content of our samples, we plotted the GC content for the reference sequences used to match the 438 exons.While we did observe that 21% of the amplicons had GC content higher than 55%, the number of reads obtained per amplicon was high enough to disregard GC bias (Figure S2).

Of the 2 sets of pooled libraries (sets A and B), set B was sequenced in duplicate (B1 and B2). This set was sequenced in two partitions of the sequencing slide. Comparing the two duplicates of set B against each other, there was a bias in a number of reads for set B2. Each sample in set B2 consistently had ~75% of the total number of reads than its counterpart had in set B1 (Figure S3).

Every read was mapped against the exon library and every unique read was used to provide coverage for each position in the exon, with coverage being the minimum previously defined. This was calculated for each exon over all 32 data sets. Figure 2 shows the minimum, average, and maximum coverage across all samples for each exon. The average minimum coverage for each exon was 19x coverage, but 28% of the exons had at least one sample with 0x coverage. However, 375 of the 438 amplicons had an average coverage higher than 10x across all samples. The maximum coverage across all samples provides excellent coverage; 424 amplicons had >10x coverage and 406 of these have >20x

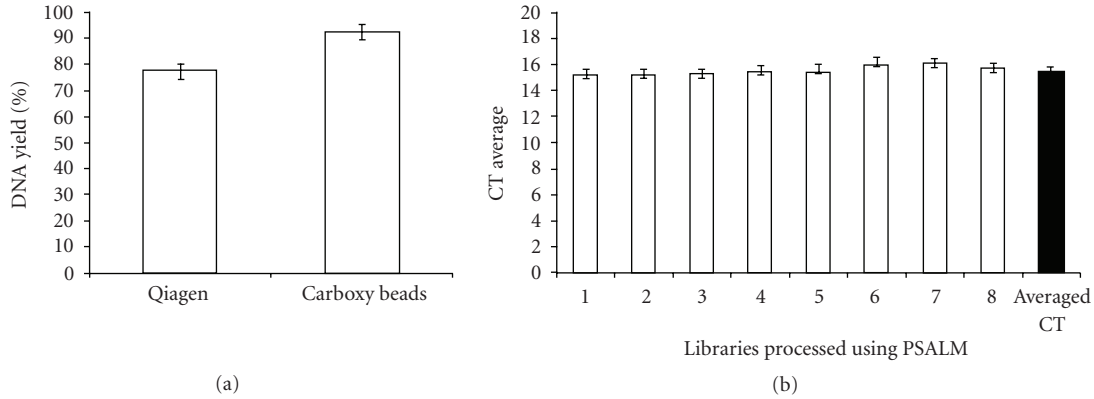

Figure 3: Comparison between the purification methods used in PSALM (CarboxyBeads) and traditional manual spin column preparation (Qiagen) is presented in the histogram (a), while histogram (b) plots the average CT values obtained for quantification of barcoded libraries prepared according to PSALM. The error bars represent a 95% confidence interval. The carboxy bead method recovers an average of 93% of the initial DNA (a), while libraries prepared using the semi-automated pipeline are amplified with same number of PCR cycles (Average CT value = 15.5), which indicates similar concentration of libraries in the tested samples.

coverage (Tables S4 and S5). The minimum values can be low, but on the average each exon had good coverage. When analyzing coverage per sample present in sets A and B, samples in set B had overall higher coverage than samples from set A. Samples p21 and p24 from set A have the lowest average coverage of 8x while samples 13, p23, and c9 from set B had the lowest average coverage of 40x. All samples from set B had average coverage higher than 40x, while the highest average coverage for set A was 43x (Figure S5, Tables S4 and S5). Harismendy et al. studied the fold difference in average coverage of sequences using three commercially available next-generation sequencing platforms. While error in pooling equimolar amounts of samples or amplicon specific bias (sequence, length) explains only a small fraction of the observed coverage variability, unique sequences present in equimolar amounts in the library generation step end up being covered at vastly different read depths. ABI SOLiD demonstrated a strong bias against coverage of repetitive elements. In addition, despite having considerably higher average sequence coverage compared to other next-gen platforms, the ABI SOLiD data had the largest number of no and low coverage intervals, the majority of which were AT-rich repetitive sequences [13]. While it is important to understand the reason for variation in the sequence coverage depth, our results provide enough coverage across all samples to perform many analyses such as SNP detection.

4. Discussion

In this study, we developed procedures to automate DNA library preparation for next-generation DNA sequencing platforms.

Next-generation sequencing technologies have been generating gigabases of data, greatly expanding the ability of researchers to use genomic information to solve biological and clinical questions. On the other hand, the application of these technologies may not be cost-effective when applied to studies involving multiple individual samples such as the targeted re-sequencing of disease populations, small genome sequencing, and ChIPSeq. These studies would greatly benefit from parallel sample processing that would require barcoding and pooling of individual samples for tracking through emulsion PCR and high-throughput sequencing. However, established protocols that involve manual barcoding of individual samples require the preparation of large sample amounts that are often not readily available and involve multiple and time-consuming laboratory steps that are prone to error. To overcome these problems, we have developed a novel semi-automated barcoding strategy and DNA sample processing protocols that increase throughput and accuracy in high-throughput DNA sequencing studies.

The use of liquid-handling robotics improved two critical components of sample preparation. First, by replacing the spin column purification with carboxy magnetic beads, the time-consuming process of multiple centrifugation steps was removed, while significantly increasing the DNA yields. Secondly, barcoding allows multiple types of samples to be prepared and sequenced together, removing the need for separate sequencing runs. Although an actual cost evaluation has not been conducted, these benefits should in effect greatly reduce cost in next-generation sequencing studies. Notably, 32 libraries were processed in 2.01 workdays in this study, whereas manual preparation would have required at least 12.49 workdays. When using the full capacity of the automated system, 96 samples can be processed in 6 days, thus reducing preparation time by nearly 30-fold.

One of the major advantages of using automation is the fact that sample preparation can be easily scaled to prepare up to 96 samples simultaneously, reducing labor and preparation time. In this study, we used a Magnatrix 8000 plus robot, but any liquid-handling robot can be programmed to incorporate the automated steps of PSALM. On the other hand, even though there is no data available estimating the time spent to manually perform the automated steps of PSALM, a manual approach using automatic multichannel

pipettes is an alternative for smaller labs that may not have the ability to incorporate custom robotics. However, a well-trained technician would be performing repetitive laborious tasks that are boring and tend to cause human error. In addition, by eliminating automation, operator-introduced variability might increases [9].

We did not present in this paper a bioinformatics analysis where we can evaluate the level of contamination between samples. All samples used in this study, whether they were part of the patients or control group, were positive for the 39 mitochondrial genes and for the 438 exons. A major bioinformatics analysis is underway to identify the SNPs associated with individual samples as well as to compare the SOLiD sequencing results with other methods previously used to characterize these samples. Once this data is available, we should be able to evaluate and estimate contamination between samples.

However, since contamination is a major concern of sample preparation, automation is one of the measures we recommend for increasing production, reproducibility and removing the major source of contamination which is human manipulation and error. A series of measures were taken to standardize the automation steps of our sample preparation. First, our automated system, reagents, equipment, and supplies are stored and used in a clean-room designated as an amplicon-free area. Second, while programming the robot, we investigated and readjusted accordingly the pipetting system of our robot. One of our major concerns was whether micro droplets were visible at the end of tips during the pipetting cycles. The droplets could be a source of contamination if they were to drip down into different wells of the plate as the pipetting system sweeps over plates to dispense liquid or tips in the trash. Even though there were no visible droplets on the tips, we placed a tray device under them while moving the tips across the platform. We also adjusted the speed of the pipetting system for mixing samples and used 1.2 mL round-well deep-well plates to reduce risk of solutions splashing out of the wells and contaminating adjacent wells. Finally and more importantly, we used a molecular barcode-based strategy for sample preparation. A unique 6bp molecular barcode (also known as a tag or multiplex identifier) was ligated to the 3′ of each individual sample in the adaptor ligation step. The SOLiD System barcodes contain unique sequences designed for optimal multiplexing. Sixteen barcodes were selected based on uniform melting temperature (Tm), low error rate, and unique color space profile. Rounds of ligation-based sequence were performed using primer sets complimentary to the barcode set following sequencing of the target DNA. A data analysis in color space was performed at the end of the sequence run and the sequence data of each specific sample was traced back and sorted using its unique identifier. Data files containing reads in which zero mismatches were allowed for the barcodes were used for the coverage data analysis.

While there was an observed GC bias in number of reads and a bias towards mapping near the primer sites, other studies (where samples were prepared manually) have shown similar biases [13]. Apparently, the semi-automation of the library-making stage does not produce any new anomalies, but does reproduce those biases inherent to other steps in the process.

While automation renders sample preparation more efficient and less laborious, some problems remain to be solved. One such limitation is the gel extraction step, which is currently not suitable for automation and is a source of inconsistency in the sequencing process (data not shown). To remedy this problem, we are working to develop an automated size selection method which, together with the removal of the amplification step, will result in a fully automated sample preparation process. Currently available enzymatic fragmentation processes without bias used in a manual preparation process may be an alternative to mechanical shearing [15].

In addition, while library preparation has been automated, the sequencing pipeline would greatly benefit from having the emulsion PCR process automated. Manual preparation of samples is time consuming and labor intensive, and this step could be performed simultaneously for all samples, which might mitigate potential bottlenecks in the workflow at this step. However, efforts remain underway to solve this problem through the use of automation.

We propose that robotic automation of all preliminary steps can significantly reduce sample preparation time for DNA sequencing using next generation technologies. Here we demonstrate that even semi-automation of just one step can save time and increase yield. Ultimately, the transition to automated preparation will render next-generation sequencing a more usable tool, and we anticipate that this will increase the scope of its use by researchers addressing epidemiological and other problems that benefit from the examination of large DNA or RNA sample sizes.

Acknowledgments

Eveline Farias-Hesson and Jonathan Erikson are equal contributors to this paper. The authors wish to thank Jason Smith and Martin Storm from Applied Biosystems for providing technical support and binning of the barcoded libraries. This work was supported in part by a National Eye Institute grant (to R. Davis and C. Scharfe), National Aeronautics and Space Administration Cooperative Agreements NCC9-165 and NNX08BA47A, and National Institutes of Health [P01-HG000205].

References

[1] P. K. Gupta, "Single-molecule DNA sequencing technologies for future genomics research," *Trends in Biotechnology*, vol. 26, no. 11, pp. 602–611, 2008.

[2] R. A. Holt and S. J. M. Jones, "The new paradigm of flow cell sequencing," *Genome Research*, vol. 18, no. 6, pp. 839–846, 2008.

[3] J. R. ten Bosch and W. W. Grody, "Keeping up with the next generation: massively parallel sequencing in clinical diagnostics," *Journal of Molecular Diagnostics*, vol. 10, no. 6, pp. 484–492, 2008.

[4] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[5] N. Siva, "1000 Genomes project," *Nature Biotechnology*, vol. 26, no. 3, article 256, 2008.
[6] F. Tang, C. Barbacioru, Y. Wang, et al., "mRNA-Seq whole-transcriptome analysis of a single cell," *Nature Methods*, vol. 6, no. 5, pp. 377–382, 2009.
[7] E. Goossens, M. de Rycke, P. Haentjens, and H. Tournaye, "DNA methylation patterns of spermatozoa and two generations of offspring obtained after murine spermatogonial stem cell transplantation," *Human Reproduction*, vol. 24, no. 9, pp. 2255–2263, 2009.
[8] N. Cloonan, A. R. R. Forrest, G. Kolle, et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.
[9] N. J. Lennon, R. E. Lintner, S. Anderson, et al., "A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454," *Genome Biology*, vol. 11, no. 2, article R15, 2010.
[10] J. Lundeberg, D. Klevebring, M. Gry, J. Lindberg, and A. Eidefors, "Automation of cDNA synthesis and labelling improves reproducibility," *Journal of Biomedicine and Biotechnology*, vol. 2009, Article ID 396808, 7 pages, 2009.
[11] C. Scharfe, H. H.-S. Lu, J. K. Neuenburg, et al., "Mapping gene associations in human mitochondria using clinical disease phenotypes," *PLoS Computational Biology*, vol. 5, no. 4, Article ID e1000374, 2009.
[12] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.
[13] O. Harismendy, P. C. Ng, R. L. Strausberg, et al., "Evaluation of next generation sequencing platforms for population targeted sequencing studies," *Genome Biology*, vol. 10, no. 3, article R32, 2009.
[14] R. J. Henry and K. Oono, "Amplification of a GC-rich sequence from barley by a two-step polymerase chain reaction in glycerol," *Plant Molecular Biology Reporter*, vol. 9, no. 2, pp. 139–144, 1991.
[15] N. Caruccio, H. Grunenwald, and F. Syed, "Nextera™ technology for NGS DNA library preparation: simultaneous fragmentation and tagging by in vitro transposition," *Nature Methods*, vol. 16, no. 3, 2009.