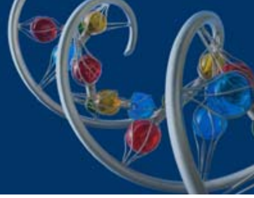
# SCALING UP THE 454 TITANIUM LIBRARY CONSTRUCTION AND POOLING OF BARCODED LIBRARIES

Wilson Phung[1], Christopher Hack[1], Harris Shapiro[1], Susan Lucas[2], and Jan-Fang Cheng[1]
[1]Lawrence Berkeley National Laboratory, [2]Lawrence Livermore National Laboratory

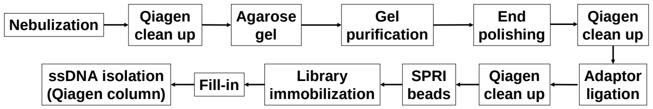


## Abstract

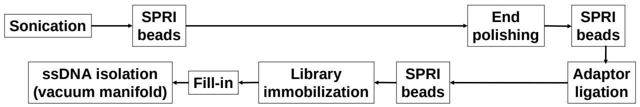
We have been developing a high throughput 454 library construction process at the Joint Genome Institute to meet the needs of de novo sequencing a large number of microbial and eukaryote genomes, EST, and metagenome projects. We have been focusing efforts in three areas: (1) modifying the current process to allow the construction of 454 standard libraries on a 96-well format; (2) developing a robotic platform to perform the 454 library construction; and (3) designing molecular barcodes to allow pooling and sorting of many different samples. In the development of a high throughput process to scale up the number of libraries by adapting the process to a 96-well plate format, the key process change involves the replacement of gel electrophoresis for size selection with Solid Phase Reversible Immobilization (SPRI) beads. Although the standard deviation of the insert sizes increases, the overall quality sequence and distribution of the reads in the genome has not changed. The manual process of constructing 454 shotgun libraries on 96-well plates is a time-consuming, labor-intensive, and ergonomically hazardous process; we have been experimenting to program a BioMek robot to perform the library construction. This will not only enable library construction to be completed in a single day, but will also minimize any ergonomic risk. In addition, we have implemented a set of molecular barcodes (AKA Multiple Identifiers or MID) and a pooling process that allows us to sequence many targets simultaneously. Here we will present the testing of pooling a set of selected fosmids derived from the endomycorrhizal fungus *Glomus intraradices*. By combining the robotic library construction process and the use of molecular barcodes, it is now possible to sequence hundreds of fosmids that represent a minimal tiling path of this genome. Here we present the progress and the challenges of developing these scaled-up processes.

## Scaling Up of 454 Titanium Std Library Construction

### Single tube library construction



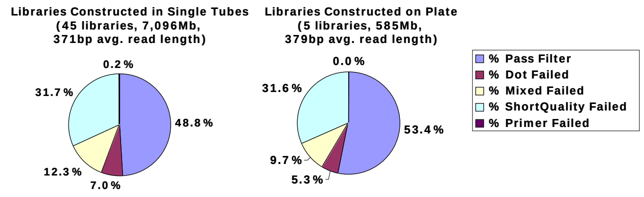
### 96-well plate library construction



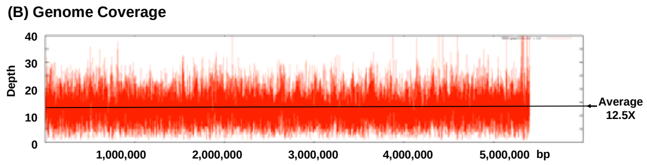
All the modifications were made so that the entire process can be adapted to the 96-well plate format. One major change is replacing the gel size selection step with a SPRI bead size selection. We thought that it might result in an increase of short quality read failure.

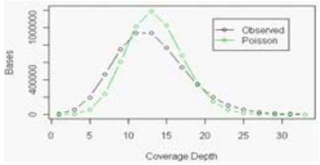
## Quality Assessment of Libraries Constructed on 96-well Plate

**(A) Read Quality**



**(B) Genome Coverage**



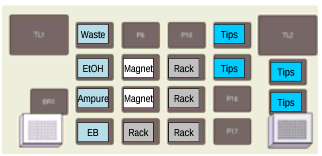


We assessed the quality of the libraries constructed using a plate format by (A) comparing the read quality with libraries constructed in individual tubes, and (B) examining the randomness of the sequence coverage over the entire genome (*Klebsiella variicola*, 5.4Mb). All the parameters examined show that the quality of the Titanium std libraries constructed on 96-well plates appears to be equivalent to those constructed in individual tubes.

## Programming the BioMek Robot

Currently, the construction of 12 shotgun libraries in parallel requires a single operator 2 days to complete, and the numerous pipetting steps require caution to ensure that safety thresholds against repetitive stress injuries are not exceeded. A method to construct 96 454 shotgun libraries in parallel using a Beckman-Coulter BioMek FX robot to automate the repetitive pipetting steps is in development at the JGI. The goal of this project is to enable a single operator to be able to construct 96 454 shotgun libraries in a single day with minimal ergonomic risk.

**Deck Layout of the BioMek Robot**



We were able to successfully develop and implement a dry test method without reagents and a wet test run with reagents but without samples. However, due to technical issues with the BioMek, we are currently working on the instrument to test the program with samples. Some of the remaining challenges to be solved include incorporating a vacuum for use with the final column purification step of the ssDNA library, and potential contamination issues from using a robot that is used by multiple groups for multiple procedures.

## Pooling of Titanium Libraries with Molecular Barcodes

Rationale: The fosmids were selected from a library derived from the endomycorrhizal fungus *Glomus intraradices*. The whole genome shotgun sequences generated by the Sanger reads form thousands of contigs. One possibility is that the assembly is hindered by the occurrence of multiple copies of many nuclear genes, somewhat diverged in sequence. This 454 library pooling approach, if it works, would provide a way to complete this genome.

**BARCODE SEQUENCE DESIGN REQUIREMENTS**

- Oligo length: 10 nucleotides (1,048,576 possible sequences)
- No consecutive same bases (78,732 sequences)
- 40-60% GC content (64,472 sequences)
- No more than 2 di-nt or tri-nt repeats (62,072 sequences)
- Adapter sequences differ by at least 4 bases (270 sequences)

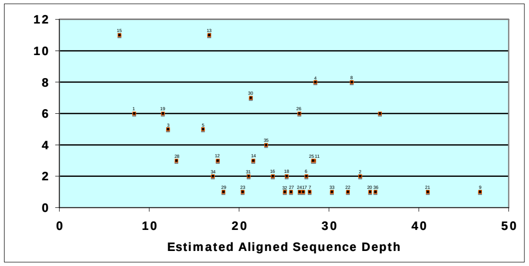
We have designed the molecular barcode sequences for the Titanium libraries based on the above requirements. We have selected 96 MIDs to be used in creating libraries in a 96-well plate format. To test the ability of pooling projects and Newbler's ability in resolving these MIDs, we constructed 36 fosmid subclone libraries, each containing a unique MID. These libraries were pooled with an equal mass according to the Bioanalyzer readings.

The pooled 36 barcoded fosmid subclone libraries were sequenced in a quarter of a Titanium run. This run yielded 229,356 reads, containing 89.7 MB of sequence. If divided evenly between the fosmids, there would be 6,371 reads each; the actual read numbers sorted into these projects ranged from a low of 790 (12% of the mean) to a high of 22,083 (~350% of the mean). Most of the projects have a less than 50% of read number deviation from the mean. We have also used the "number of allowed errors" parameter in the MID configuration file to test the ability of sorting these MIDs.

We have also examined the ability of assigning these MIDs accurately by Newbler:

a) Allowing no error: 97.02% of the initial reads were assigned to a MID, the missing 3% could be due to a combination of error rate and contamination.
b) Allowing up to one error: 99.61% of the initial reads were assigned to a MID, all uniquely.
c) Allowed up to two errors: 99.80% of the initial reads were assigned to a MID, and 99.63% were uniquely assigned.
d) Allowing up to three errors: 99.92% of the initial reads were assigned to a MID, but only 24.35% were assigned uniquely.

## The Assembly of Pooled Fosmids



The assembly of these Titanium reads formed 13 complete fosmids (36.1%), 6 with 2 large contigs (16.7%), 12 with 3-6 large contigs (33.3%), and 5 with greater than 6 large contigs (13.9%). More than half of the fosmids (11 of 20) that received greater than 25-fold sequence depth formed single complete contigs. Local repeats within each fosmids do not seem to affect the assembly. So this dataset suggests that the pooled barcoded fosmid libraries with 454 sequencing seems to be a valid approach to sequence the *Glomus intraradices* genome.

## Conclusions

1. **We have been constructing Titanium Std libraries on a 96-well format using a modified version of the 454 protocol. The sequence quality generated from these libraries are comparable to those constructed with individual tubes.**

2. **We have successfully programmed the BioMek robot to execute the dry test method and perform a wet test run. Once we run the program with samples, we can automate the construction of 96 libraries simultaneously with minimal ergonomic risk.**

3. **We have demonstrated that using a set of molecular barcodes to create and pool libraries for 454 sequencing is a valid strategy to sequence many target DNA and analyze separately. The data presented here shows that the assembly of individual *Glomus intraradices* fosmids do not seem to be affected by the repetitive sequences in the genome. These repeats have been shown to hinder the whole genome shotgun assembly of this genome.**

LLNL-POST-412357

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.

## DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.