Advancing Science with DNA Sequence

Comparison of *454* Sequencing Platform with Traditional Sanger Sequencing: a Case Study with *de novo* Sequencing of Prochlorococcus marinus NATL2A Genome

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Feng Chen, Joseph Alessi, Edward Kirton, Vasanth Singan, and Paul Richardson

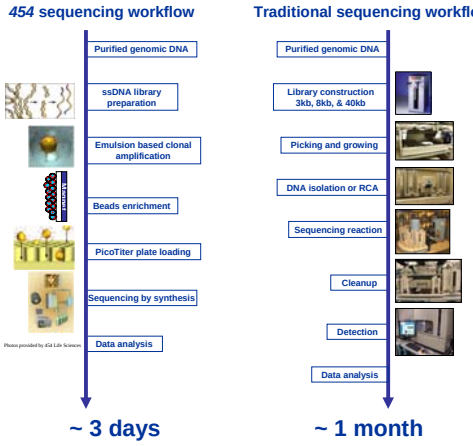
Introduction

The Department of Energy Joint Genome Institute (www.jgi.doe.gov) in Walnut Creek, CA is a high throughput DNA sequencing facility with a current throughput of approximately 3 billion basepairs per month. A major effort at JGI is the sequencing of microbial genomes of relevance to the DOE missions of carbon sequestration, bioremediation and energy production. The JGI Microbial Program is responsible for the generation of over 300 microbial genomes . JGI is running about 70 *ABI* sequencers on a 24/7 schedule and about 40 GE *MegaBACE 4500* sequencers on a 24/5 schedule. Our current whole genome shotgun sequencing strategy is to sequence 3kb and 8kb shotgun libraries to a combined 8x draft coverage and to sequence fosmid ends to 1x sequence coverage, which is equivalent to about 30x clone coverage.

454 Life Sciences recently developed a new scalable, highly parallel sequencing system with significantly greater throughput than our current Sanger sequencing systems. It is an integrated system of emulsion PCR amplification of hundreds of thousands of DNA fragments linked to high throughput parallel pyrosequencing in picoliter-sized wells. It is capable of producing 25 million bases in one 4-hour run. The high throughput nature of *454* sequencing system is attractive especially to microbial whole genome shotgun sequencing. For each organism, three libraries are required for current Sanger sequencing strategy, requiring a significant amount of effort and resources. In contrast, 454 system does not require traditional cloning and when the size of many microbial genome are under 5 million bases, one 454 run will give 5-8x sequencing coverage.

However, the quality of 454 sequencing reads and the resulting assembly from 454's Newbler assembler is not well characterized. The other limitation is the short readlength from *454* system, it is about 100 bp. *Prochlorococcus marinus* NATL2A (genomic DNA was kindly provided by Penny Chisolm and Claudia Steglich, Department of Biology, MIT) is 1.84 million bases in size and has been fully sequenced at JGI with traditional Sanger sequencing technology (Genbank GI number 72381840). We sequenced it with one run on *454* platform, and generated over 36 million bases of data. The 332,387 reads of average length 109 bp were assembled with *454's Newbler* assembly tool which generated contigs of consensus sequence. The assembly results from the *454* system was compared with previously finished Sanger sequencing results. For the purpose of direct comparison, we didn't modify *454's* assembly experimentally or computationally. Informatics tools were developed internally to facilitate the statistic analysis and visualization. We are looking at the error rate at both read and contig levels. We are also looking at the coverage and the depth of the entire genome with *454* sequencing results. *Newbler* assemly tool from *454 Life Sciences* also produced a certain number of misassemblies identified by comparing *Newbler* assembly with the FASTA format sequence downloaded from Genbank.

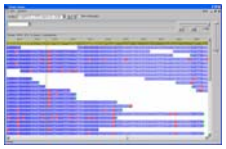
Methods



Method for *454* Data Processing

A sequencing run on the *454* machine produces a set of images corresponding to nucleotide incorporation in each reaction cycle in each well. The subsequent image and signal processing produces arrays of intensity measurements which correspond to the flow order of the bases. The nucleotide sequences of the reads are reported, although they are not used by the assembly software. If the read sequences are of interest, their quality may be improved by using included "training" software which leverages a reference sequence to calibrate intensity thresholds used in base-calling. Both the untrained and trained read sequences were evaluated in this study.

The included *Newbler* software assembles the read data in "flow space" rather than "nucleotide space"; that is, it considers the light intensity measurements, not predicted read sequences. The assembler is designed to correct the systematic errors associated with pyrosequencing, such as homopolymer-related errors (e.g. errors in determining the number of nucleotides incorporated by judging the intensity of the light emitted). The quality of the contigs was evaluated and, as is shown here, the systematic errors were largely eliminated. The use of other contig-assembly software without these special features, such as *Phrap*, would be inappropriate for *454* data (data not shown).



The *Newbler* assembler produces the ubiquitous *fasta*, *qual*, and *ace* files, making the data amenable to manipulation by existing tools. At this time, however, the *ace* file is not parseable by *Consed*, although it can usually be read by TIGR's *CIView*.

Evaluation of the Accuracy of *454* Sequences

A single sequencing run on the *454* instrument produced 332,387 reads, averaging 109bp in length. Total number of bases from this run is 36,230,183. *Newbler* assembled these reads into 76 large contigs totaling 1,814,332 bp, and 4,731 small contigs totaling 438,438 bp. Large contigs are defined by *454* as those contigs consisting of at least 800 NT flows and in this data set the smallest was 697 bp in length. The average depth of coverage for large contigs is about 14x.

To assess the accuracy of *454* sequences, both raw and trained (i.e. by using the known sequence as a reference for calibration) reads, plus large and all assembled contigs were aligned to the previously finished genome using *sim4*. The single best hit for each sequence was selected and differences from the reference sequence were extracted and classified as one of several types of error. For this analysis, a string of two or more occurrences of a NT were considered a homopolymer.

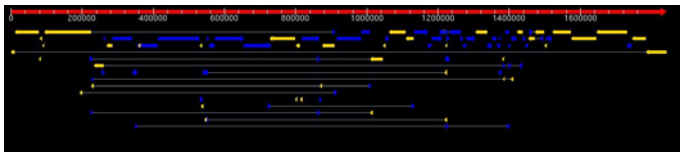
Acknowledgements

- PGF Sequencing Production Department
- JGI Microbe Program
  - Eugene Goltsman and Alla Lapidus
- Department of Biology, MIT
  - Penny Chisolm and Claudia Steglich

TYPE OF ERROR	EXAMPLE	UNTRAIN. READS	TRAINED READS	ALL CONTIGS	LARGE CONTIGS
Miscalled base	aaaCttt c wrong	111,420	8,515	127	77
Insertion	tattcaa extra t	165,096	77,988	359	59
Deletion	tgt-aag missing nt	97,954	30,184	44	16
Short homopolymer	aaa-taa missing a	53,150	73,102	224	174
Long homopolymer	gataaaat extra a	331,797	165,370	325	199
Homopolymer indel	gggaag g should be a	1,361	808	1	1
Homopolymer incomplete extension	aaaaactagg extra a	196,649	160,357	573	310
HMP incompl exten with indel	aactagg a should be g	30,015	5,851	74	48
N's	n	6,639	11,798	0	0
Other – not easily recognized	tct-aaa mixed case	59,734	1,761	13	1
Total number of errors		**1,053,815**	**535,734**	**1,740**	**885**
Total number of bases aligned		**34,299,400**	**31,001,733**	**1,822,055**	**1,802,556**
Overall quality score		**< Q15**	**< Q18**	**Q30**	**Q33**

The average quality score for the reads calculated by *454* software was Q26.4 (where Q=-10 log p), which corresponds to an error rate of 0.23%. The average error rate for reads from our analysis is about 3%, which agrees with what reported in *454's* Nature paper (Margulies, M. etal. *Nature* 437, 376-380(2005)) at about 4%.

The accuracy of base-calls in the contigs is quite high. Most errors are associated with homopolymers, where the length of the tract is misjudged or when they are not fully extended, causing residual signals on the next flow of that nucleotide (incomplete extension). The overall quality of large contigs is Q33, which would be adequate for genomic sequencing projects if the quality at each position was accurately estimated, however, at this time, the *ace* file *Newbler* generates assigns values of Q97 to most bases not in a homopolymeric tract. We hope future version of *Newbler* will provide more useful quality estimates. It is noteworthy that the error rate of consensus from our analysis agrees with *454's* Nature paper (0.05% from our study vs. 0.04% from *454's* report). The difference could be owing to about 3 times more coverage in *454's* dataset. Based on this study, we believe that *454* platform can provide high quality overall assembly results at reasonable depth of coverage and can be used to make *de novo* sequencing of relatively small microbial genomes faster and cheaper.

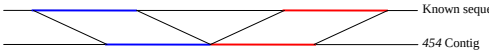


Evaluation of *Newbler* Contig Assemblies

While the sequencing run produced 36 Mbp of data, only 31.6 Mbp could be assembled, those unassembled were partial, singleton, repeat or outlier. All contigs totalled 2,252,770 bp, among these, 1,822,055 were aligned to the reference sequence, giving 98.87% coverage of the whole genome. Large contigs totalled 1,814,332 bp, of which, 1,802,556 were aligned to reference sequence, giving 97.81% coverage. Newbler assembly results in 76 large contigs with N50 contig sized at 56,313 bp.

	Length	G+C%
Not covered by 454	21,882	39.4
Reference Sequence	1,842,899	35.1

Given the short length of the reads (around 109 bp), *454's Newbler* assembler is expected to perform poorly with repeat regions. The current genome was chosen, in part, because it is known to be rather free of repeats. In order to check for contig misassemblies, the *454* contigs were aligned to the finished genome using *BLAST* and the results checked for inconsistencies. There were two major misassemblies, both of the type depicted below in which distal similar regions were incorrectly joined, to the exclusion of the intervening sequence which fell into a separate contig. The *Newbler* assembler is expected to have difficulties assembling genomes with many repeats. However, a strategy which combines scaffold information from traditional clone end-sequencing with *454* data may solve this problem.

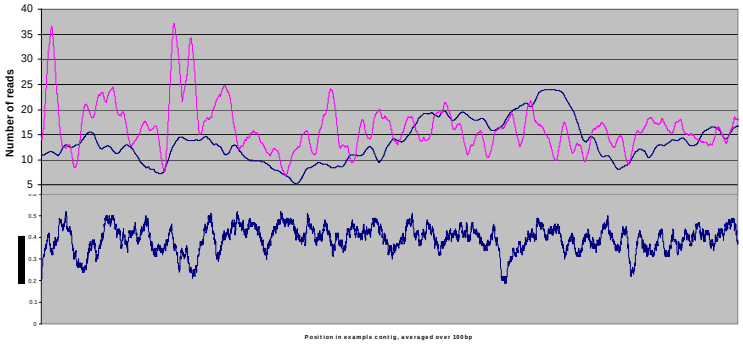


Distribution of *454* Read Coverage

Top: Reads coverage for 9056bp contig, averaged over 100bp sliding window

ABI 454

Bottom: G+C content for the same contig, similarly using 100bp window



By parsing the *ace* file and extracting the coordinates of each read on the *454* large contigs and the *ABI* contigs, we were able to determine the read coverage over each position. We observed greater variability in the distribution of *454* reads over the genome than *ABI* reads. This may partly be a result of the emulsion step following bead preparation, resulting in multiple beads with the same DNA fragment in different wells.

READ COVERAGE	ABI	454
Max. read coverage	33	167
Mean read coverage	14.3	13.7
Std. dev. in coverage	4.1	7.4
# low coverage regions	1,270	4,537
Len. of low cov. regions	316,990	434,542
# high coverage regions	1,957	3,543
Len. high cov. regions	239,133	185,601

The sequences of the regions of high and low *ABI* and *454* coverage were extracted and evaluated. It appears that the *ABI* and *454* methods do not have the same biases in their read coverage. This suggests that a genome sequencing strategy using both technologies would be effective.

Our group is currently carrying out *in silico* experiments to determine the optimal ratio of *454:ABI* sequencing and developing a strategy that would take advantage of *454's* low operating costs and effectively overcome the limitations associated with short read lengths and lack of positional information.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

LBNL-59003 Poster