Balancing Sanger and 454 Data for Microbial Sequencing

Authors
Goltsman, Eugene
Kirton, Ed
Lapidus, Alla
et al.

Publication Date
2007-03-20

Balancing Sanger and 454 data for Microbial Sequencing

Eugene Goltsman, Ed Kirton, Alla Lapidus, Paul Richardson

DOE-JGI, Walnut Creek

The Joint Genome Institute now uses the 454 sequencing technology in virtually every genome it sequences. Aside from its well known advantages in throughput volume and cost, the other important advantage of 454 and other *sequencing-by-synthesis (SBS)* technologies, is the lack of cloning-related artifacts and pitfalls. One of such pitfalls has been strong cloning bias in plasmid libraries - an extremely problematic issue during genome closure/finishing . Another one is the potential for culture variants in the library, which greatly confounds assembly layout and consensus building process. However, since in other criteria this technology today is still inferior to paired-end Sanger sequencing, it is not feasible at this point to replace the latter entirely by SBS data. This leads to a need to balance these two types of data in our sequencing process.

In this study we arrived at the optimal target coverage level (a.k.a. depth) for Sanger shotgun data from 8kb pMCL200 libraries, given a fixed amount of complementing 454 coverage. We were also able to estimate how stable and predictable this optimum is in different types of genomes and how it respond to the addition of fosmid library data. The result of this and forthcoming analyses will allow JGI to free up a large portion of its sequencing capacity for greater genome throughput.

METHODS

Approach and Criteria:

To establish this optimal coverage, we decided to proceed by way of empirical analysis of a multitude of assembled genomic datasets. In selecting proper criteria for assessing assembly quality , we looked for factors, apparent in the assembly, that could well reflect the amount of further refinement needed. This refinement comes in the form of additional semi-automated custom sequencing of weak areas and gaps plus manual and labor-intensive curation, where all automatic approaches have failed. It is known that at a certain point, a form of saturation is reached where adding more shotgun data doesn't have any significant effect on the amount of further finishing work needed. This point (or range) relates directly to the optimal pMCL200 library coverage we are looking for, and, since addition of 454 data changes where this point is, we need to re-establish it. We picked the following criteria by which to estimate this saturation point and evaluate its stability:

- Uncaptured Gaps in the Assembly
- Consensus Sequence Quality: areas requiring custom finishing
- Data variability: GC composition, cloning bias

Genomes:

- *Burkholderia multivorans* - 7.6 mb, 66%GC
- *Parvibaculum lavamentivorans* – 4.3 mb, 62%GC
- *Pseudomonas putida* - 6.6 mb, 62%GC
- *Halothermotrix orenii* – 2.7 mb, 38%GC

Datasets:

Multiple iterative Phrap assemblies were generated for each genome with varying combinations of data from different sequence sources:

- Pseudo-contigs were generated from 6-lane 454 runs (~11-13X read depth) and added to input datasets one lane at a time (only 6x of 8kb and 2 lanes of 454)
- Each genome's 8 kb dataset was divided into 10 non-overlapping and randomly selected sets of reads, which were then added to the assembly one set at a time.
- Assemblies were done both with and without 40kb fosmid libraries. Fosmids are used here mainly for layout mapping and linkage and can span a lot of gaps which would be uncaptured otherwise. (Not enough fosmid data was present in H.orenii and thus weren't included in the analysis)

From resulting assemblies we recorded the following:

- Average depth of read coverage, based on the Q20 read length in the 8kb library input sequences.
- Number of scaffolds in the assembly - this roughly translates into the number of uncaptured gaps. The Bambus program was used to generate scaffolds.
- Number of projected finishing reactions, as suggested by the customized Consed/Autofinish software. We assumed that any other reactions "unforeseen" by Autofinish will require manual design and constitute a relatively small and constant portion.

The 8kb depth in the assemblies was calculated using only the high quality portion of each read (Q20+). The total number of plates originally requested to be sequenced for each genome was estimated using the following relationship:

$$plates = (target\_coverage * genome\_size) / (2 * 384 * read\_length * efficiency)$$

Since *genome size* and *read_length* at the point of sequencing are merely estimates, the resulting coverage deviated somewhat from the 8x targeted.

RESULTS

Uncaptured Gaps:

To estimate the degree of finishing difficulty we first looked at the number of contig scaffolds which normally reflects the number of uncaptured gaps in the assembly. These uncloned areas greatly complicate the finishing process due to the absence of templates useable for any further sequencing. Since clone-mate information is used to link neighboring contigs, linkage in these cases can only be established via combinatorial PCR or optical mapping of the chromosome, both of which are time consuming, labor-intensive processes.

In the example of *Burcholderia multivorans* (Figure 1a), the number of uncaptured gaps initially decreases sharply as more 8kb data is added, but quickly reaches saturation level at the depth of about 4x. This suggests that certain areas of the genome are likely uncloneable within the current library, and simply adding more coverage does not help. With respect to this criteria alone, we can consider the sequencing depth of 4x to be close to optimal.

Example: Burkholderia multivorans

Fig. 1a. Uncaptured Gaps vs. 8kb depth: *Burkholderia multivorans – 66% GC*

6 datasets with increasing amount of 454 depth, were used. Two additional plots show the result of adding appx.1x of 40kb-insert fosmid reads (dotted line). Number of scaffolds reaches the minimum plateau between 3x and 5x of 8kb depth in fosmid-free datasets. Addition of fosmids results in a capture of virtually all gaps with 6 lanes 454 data. 8kb data seems to have no real effect when fosmids are present

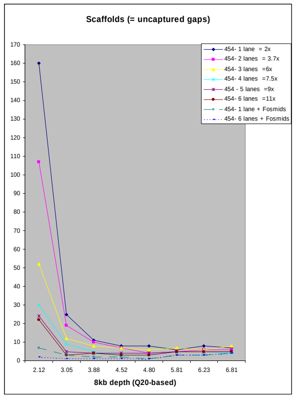
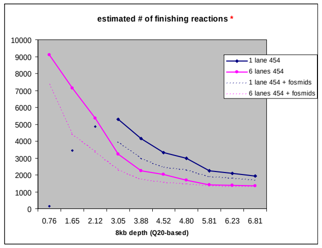


Fig. 1b. Finishing Reactions vs. 8kb depth: *Burkholderia multivorans – 66% GC*

Two 454 datasets (1 and 6 lanes = 2x and 11x of 454 read depth), with and without additional 1x of 40kb fosmid data (dotted lines) were used. Saturation occurs at around 6x of 8kb depth in fosmid-free datasets and at around 5x when fosmids are added.



* This number is based on a single round of Autofinish. Normally, additional rounds are done after adding the results of the first round. These additional rounds can add 25-30% more reactions to the total. What's important here is the shape of the curve, which should be independent of the actual number of reactions that will be done.

Suggested finishing/polishing reactions:

The second criteria we looked at is the projected number of custom sequencing (standard, semi-automated) to be done in the Finishing stage (Figure 1b). Here, we used Consed-Autofinish to suggest the reactions. As the sequence coverage increases, fewer and fewer reactions are necessary, but due to various genome-specific anomalies, some regions with insufficient clone and sequence data persist despite the increase in depth. This causes the observed flattening of the curve, meaning that return from every new shotgun sequence is approaching zero. This, again, allows to approximate the saturation point which is for this genome between 5x and 6x. The increase in 454 depth moves this point closer to the lower margin, and the addition of low-depth fosmid data lowers it even more. Because of the much greater number of areas that need finishing-polishing, compared to the number of uncaptured gaps, this point of relative saturation is more fuzzy and depends on many other, less tangible, variables (see Discussion). However, a conservative estimate can be made that will still result in great reduction of wasteful sequence.

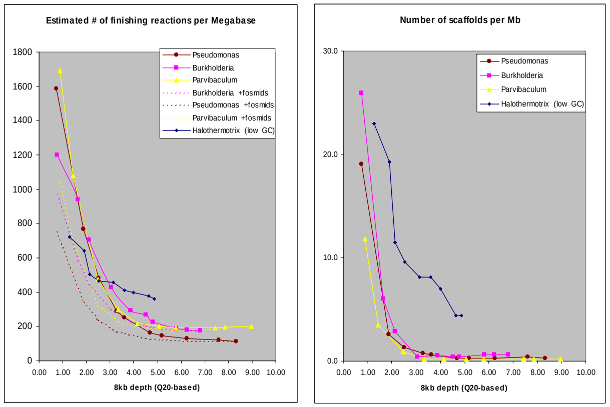
Genomes Compared

We observed strong agreement between the three high-GC genomes in terms of the occurrence of saturation in both plot types (Figure 2), and we expect this correlation to hold true with more genomes of high GC content added to the analysis in the future. The results we observed go in agreement with our earlier observations and assembly analyses.

Genomes with low GC composition are more likely to exhibit strong cloning bias and other cloning and sequencing anomalies. For comparison, we looked at the behavior of the low-GC genome of *Halothermotrix orenii* **.** There is a noticeable increase in values for both criteria, as well as a more rigged curve shape. Most likely, this comes from insufficient or non-normal clone coverage due to cloning bias, and this prevents successful scaffolding, gap closure, and polishing.

Summary of results: All genomes

Fig. 2 Genomes compared. For each criteria, the genomes under study are combined. Datasets with 11x of 454 depth are shown. With both criteria, there is a strong correlation between the high-GC genomes, and the highest saturation point is at 5x. The low-GC Halothermotrix orenii has a significantly larger number of both uncaptured gaps and projected reactions, and the saturation point could be beyond its maximum depth. Fosmid data was excluded in the case of H.orenii.



DISCUSSION

The analysis allowed to establish a preliminary target depth level for 8kb pMCL200 plasmid libraries in the presence of 454 data. Based on the genomes studied, 6x can be considered a safe depth level to reach saturation of clone coverage and of consensus sequence quality in high-GC genomes. This depth can be further reduced with improvements in the 454 technology, such as more predictable individual error probabilities and more consistent overall error rates. This topic is part of an ongoing study at JGI.

We also concluded that significant benefits from fosmid libraries are still present, and that at this point, they should remain in the datasets along with the 8kb plasmids.

Further work aimed at reducing the target depth should include addition of more genomes of low GC composition, time and cost factors of custom polishing, and reliable error probabilities in the 454 data. Because finishing efforts directly relate to the consensus quality, accurate 454 error probabilities can dramatically reduce the need for polishing.

LBNL-62590 Poster

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 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. DE-AC52-06NA25396