Establishing Optimal Coverage Levels for Different Sequencing Platforms

Eugene Goltsman, Brian Foster, Alex Copeland, Mingkun Li, Kurt Labutti, Alla Lapidus
DOE-JGI, Walnut Creek

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Sequencing-by-synthesis (SBS) technologies have provided new cost effective ways of obtaining high quality draft assemblies as well as fully finished genomes. The result is also a change in the complex dynamics of shotgun assembly which makes it necessary to re-evaluate the process, especially when complete finishing is the goal. So far, no single SBS technology has proven itself fully capable of providing all the data necessary for efficient and streamlined finishing, therefore, a combination of SBS and Sanger data is still generally preferred. JGI currently utilizes both *454*/Roche and Solexa/Illumina pyrosequencing in addition to the Sanger big-dye shotgun method. In this study we evaluated how various key assembly features are affected by changes in the levels of sequence from different platforms and attempted to arrive to a combination optimal for cost-efficient finishing. We also looked at how stable and predictable this optimum is in genomes with different GC and repeat composition. The criteria we picked are believed to be the best indicators available upon the initial draft assembly of what kind of additional effort, and how much of it, is needed to finish a microbial genome.

METHODS

Genomes:

Natronanaerobium thermophilum JW/NM	- 36% GC, 3.2 mb
Leptothrix cholodnii SP-6	- 67% GC, 4.9 mb
Thermoanaerobacter ethanolicus X514	- 34% GC, 2.5 mb
Brachybacterium faecium DSM 04810	- 72% GC, 3.6 mb
Exiguobacterium sibiricum 255-15	- 49% GC, 3.0 mb
Slackia heliotrinireducens DSM 20476	- 60% GC, 3.2 mb
Cryptobacterium curtum DSM 15641	- 51% GC, 1.6 mb
Sanguibacter keddieii DSM 10542	- 71% GC, 4.2 mb

Sequencing data:

Hundreds of assemblies were generated for each genome by progressively increasing the amount of input sequence data and then re-assembling with PGA. Prior to assembling, all Sanger shotgun reads had vector and low quality regions trimmed using the Lucy software. Reads from one full *454* run (unpaired) were first assembled with Newbler; then, the resulting contigs over 500bp were fragmented and transformed into overlapping "pseudo-reads" with Newbler consensus quality scores preserved as read quality scores. Solexa data will be used in further analysis for "polishing" purposes only. In this study we add the cost of 2 lanes of Solexa as a constant factor, assuming that this will be sufficient to polish an average microbial genome. At this point in time we are not relying on Solexa reads to resolve gaps or layout problems.

Models:

We followed three configuration models:

A. 8kb + *454*: 0x-5x from 8kb plasmid library + 1 run *454*
B. 8kb + 40kb + *454*: 0x-5x from 8kb plasmid library + 1x from 40kb fosmid library + 1 run of *454*
C. Sanger-only: 0-5x of 8kb library + 1x of 40kb Fosmid library

Finisheability criteria:

In the resulting assemblies we focused on the following criteria to measure the amount of effort needed to achieve full closure:

- **Captured gaps** - at least 2 spanning clones exist, providing linkage between contigs and a ready template for primer walking.
- **Uncaptured gaps** - no spanning clones, meaning linkage cannot be easily established, requiring combinatorial PCR followed by sequencing of the products.
- **Uncaptured repeats** - same as above, but sequencing is not needed. These are assumed to cause misassembly and require expensive manual intervention.
- **Total gap size** - measured by aligning the draft assembly to the finished genome. Allows to estimate how much custom sequencing is needed.
- **Largest gap** - are there gaps that are too large for PCR?

Costs:

Costs were estimated based on the information as of January 2008 and are as follows: single run on 454 FLX: $17,000; Solexa data (3 lanes): $3,300; Sanger shotgun read: $1; custom primer walk: $4*; PCR reaction: $6*; Sanger library preparation: $300.

* An additional "redo" factor was applied to the calculations to reflect the failure rate in primer walking and PCR reactions.

RESULTS

Uncaptured Gaps:

Uncloned areas greatly complicate finishing due to the absence of templates that could be used for further sequencing and the lack of linkage between contigs. Therefore, ordering of contigs and gap closure can only be done after successful combinatorial PCR or optical mapping of the chromosome, both of which are time consuming, labor-intensive processes. As apparent in Figure 1, the presence of a 40kb insert fosmid library together with a 454 run (model B) eliminates most uncaptured gaps, and adding coverage from an 8kb library has little effect. When the *454* data is removed and we are left with only Sanger-type sequences (model C), the values increase significantly along with the variance, which reflects the unstable nature of shotgun cloning libraries.

Uncaptured Repeats:

When long, nearly identical repeat copies fail to get spanned by uniquely anchored clone-mates, there is a high potential for misassembly. For the purposes of this analysis, uncaptured repeats are assumed to require combinatorial PCR in order to establish the correct layout, and captured repeats are assumed to assemble correctly right away and not require additional effort. According to our results, a 40kb fosmid library sequenced to 1x is by itself enough to bring the number of uncaptured repeats to a manageable minimum, and the addition of 4x from the 8kb library eliminates them completely (Fig. 1B & 1C). Due to cloning bias, many gaps may remain uncaptured, but, since cloning bias and repeats normally don't coincide, repeat resolution is possible even with biased libraries.

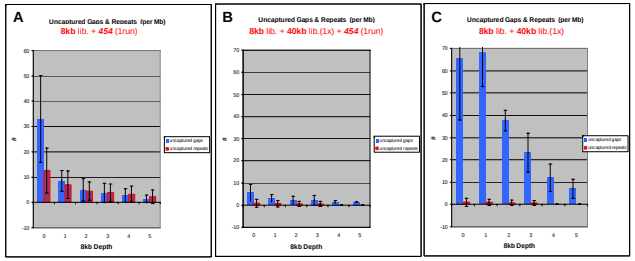

Fig. 1. Uncaptured Gaps and Uncaptured Repeats
All values are for an average genome and are normalized for genome size. Models B and C show very few uncaptured repeats compared to model A thanks to the large insert library. Due to cloning bias, however, many gaps are uncaptured in model C where *454* data has been taken out.

Gap size:

The presence of *454* data is shown to dramatically reduce the size and number of gaps in the assembly. One *454* run assembled alone leaves uncovered roughly as much of the genome as does 5x of the 8kb library combined with 1x of the 40kb (Fig. 2A & 2C). The main factor contributing to this is the lack of cloning bias in *454* datasets, resulting in a more Poisson-like sequencing depth distribution.

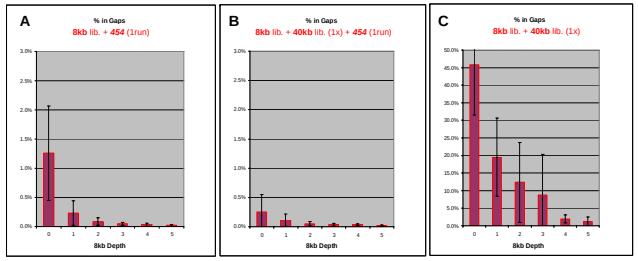

Fig. 2. Percentage of genome not in contigs

Cost Analysis:

The above metrics begin to make full sense only when all the associated costs are taken into account and compared side by side. Since, in theory, finishing costs should decrease with continuous addition of shotgun data, we expect to find an optimal balance between finishing costs and shotgun costs, at which point the total cost of the genome should be at the minimum (Diagram 1). As seen in Figure 3, model B shows the absolute minimum cost of all the assemblies at 2x of 8kb depth. The minimum in model A is in the vicinity of 3x. Since model A is more finishing-intensive due to the absence of the fosmid library, it would likely be preferred by a lab with a fully loaded and optimized finishing pipeline, while model B would fit better where one wants to minimize finishing and rely on the shotgun data instead.

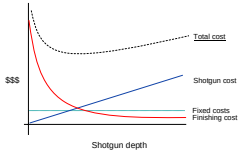

Diagram 1. Relationships between cost factors in a typical microbial assembly.
Finishing costs drop quickly as shotgun data is added, but eventually reach a plateau where additional shotgun sequence doesn't improve the assembly. Shotgun costs increase linearly with depth, and certain costs remain fixed (i.e., library construction).

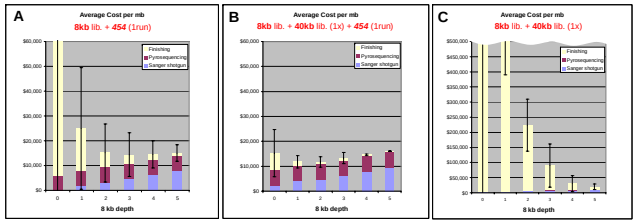

Fig. 3. Total cost breakdown
The absolute minimum cost is in model B, at 2x of 8kb Sanger depth. Variance is also smallest in model B, which can be attributed to the mutually supporting effect of two Sanger libraries. Model A shows higher finishing costs due to the absence of a long-insert library, but is comparable to B at higher 8kb depth.
(different scale is used for model C)

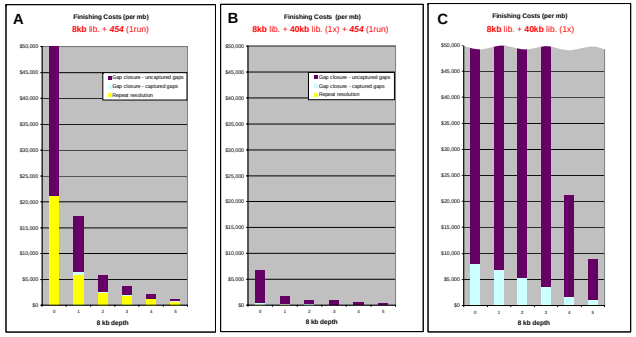

Fig. 4. Finishing cost breakdown
In all three models, the greatest contributors to the cost of finishing are uncaptured gaps which result in an exponential cost increase as Sanger depth goes down. Model A also shows significant costs associated with repeat resolution, which is attributed to the absence of the 40kb paired-end library.

The analysis allows to establish a range of depths to target when sequencing using paired-end Sanger libraries combined with unpaired *454* data. While the former provides mapping information crucial in resolving repeats, the latter allows to cover the genome more uniformly. The most cost-effective configuration overall is model B with around 2x of 8kb library sequence added to 1x of 40 kb data and one run of *454*. This optimum will change as improvements in the paired-end 454 technology will ultimately allow to further reduce or completely eliminate the Sanger technology from the process.

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