A novel approach to high throughput microbial genome finishing: Incorporation of 454 sequence data for gap closure in low quality Sanger data

Olga Chertkov, Avinash Kewalramani, Riley Arnaudville, Cliff Han and Thomas Brettin
Los Alamos National Laboratory, Los Alamos NM; Joint Genome Institute, Walnut Creek CA

ABSTRACT

Our recommendations for ideal 454 assembly: The number of contigs should be less than 500 for microbial genomes less than 8 Mb. It may require more than one 454 run per 2 Mb of sequence (especially if genome is more than 4 Mb). It will reduce the number of chimeric reads, reads with high-quality discrepancies due to short insertion/deletions and mutations. These wrong 454 reads appear to be the result of low coverage in 454 assembly as well as the lack data. These wrong 454 reads tend to mess up assembly even with very good (10x) coverage of Sanger data.

Conclusion: Our primary objective of incorporating 454 sequencing data to close low quality gaps in Sanger data has been successful as presented in the results. We were able to pick comparable number of primers, if not less, in an automated fashion during multiple cycles of AutoFinish. The strategy to assign low quality scores in only homopolymeric regions of 454 data is much more efficient than assigning low quality scores to all 454 bases. Also this homopolymer. We have also shown that the 454 sequencing data is fairly reliable and consistent in terms of errors for a variety of genomes, since errors are not dependent on the GC content of the genomes.

Table 5 Finishing progress with Sanger and 454 data

| Project | Sanger Coverage | 454 Coverage | Sanger only | Sanger plus 5X 454
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>10x</td>
<td>5X</td>
<td>44</td>
<td>70</td>
</tr>
<tr>
<td>Test 2</td>
<td>10x</td>
<td>5X</td>
<td>45</td>
<td>75</td>
</tr>
<tr>
<td>Test 3</td>
<td>10x</td>
<td>5X</td>
<td>46</td>
<td>80</td>
</tr>
</tbody>
</table>

Fig. 1. Parity of Pyrimidine homopolymer error rate dependency on homopolymer length.

Fig. 2. Error rate across all motifs in a particular project against the GC% (total GC content of the genome which is varying in the range of 27 to 68%). We do not see any correlation across the whole genome between the GC% and the number of errors associated with the 454 data. Hence we do not feel that the 454 sequencing methodology is biased in the prediction of bases in any way associated with the GC content of the genome within this range.

IGI-LANL has finished approximately 105 microbial genomes, some of them done with 454 and Sanger reads; all finished sequences have been uploaded into public databases. We have seen dramatic improvement in finishing efficiency when we incorporated 454 data in our assemblies. We also tried to identify other types of problematic regions in 454 data, but their number was very low. Only with the growing number of projects we found that the number of chimeric reads and reads with high quality discrepancies due to short insertion/deletions and mutations depends on 454 coverage and number of 454 contigs in assembly.