Finishing Sanger/454 Hybrid Sequenced Genomes at JCVI

Mary Kim, Nadia Fedorova, Yasmin Mohamoud, Daniela Puiu, Luke J. Tallon

J. Craig Venter INSTITUTE

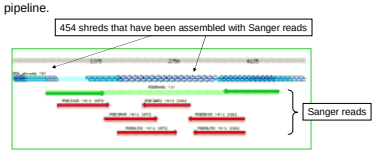
9704 Medical Center Drive, Rockville, MD 20850

Abstract

As new and promising genome-finishing techniques emerge on the scene, they must undergo thorough analysis to determine their full efficacy. At JCVI we strive to test and incorporate new methods with our own to achieve optimal results in finishing genomes. Currently, we generate genome assembly sets from a combination of 454 sequence data and Sanger sequence data (i.e. hybrid sequenced genomes). These genomes proceed through the finishing process to close remaining gaps and resolve misassemblies, hard stops and low quality regions. Both generating a viable assembly set and manipulating the hybrid genome through our database and software tools present an array of challenges and difficulties. Furthermore, modification of our quality standards for finished hybrid sequenced genomes is a need that must be addressed. Finishing such genomes will give us a better understanding of how to work in tandem with new approaches such as 454 technology, and thus help to advance genome-finishing techniques.

Generating the Sanger/454 Hybrid Assembly Set

In order to generate a hybrid genome, JCVI uses data from DNA fragments synthesized by 454 sequencing technology - data that is packaged into Standard Flowgram Format (SFF) files. These SFF files consist of basecalls, flowgrams, and quality scores for each nucleotide in each sequenced read. They are then assembled by 454 Life Sciences's assembler tool, Newbler, in order to generate contigs based on pairwise overlaps and multiple read alignments. Consensus basecalls are produced by averaging the flow signals for each base in the alignment. Because 454 reads cannot be processed by Celera assembler, the 454 Newbler contigs must first be "shredded" into overlapping segments of 600 bp. The shreds are then incorporated with Sanger reads to generate an assembly set suitable for JCVI's genome finishing pipeline.



Unclonable Regions

Hybrid vs. Sanger-only Assembly Sets

As depicted in the table below, the 454 data that was incorporated into four genomes – which are currently being finished at JCVI - significantly reduced the number of physical ends by eliminating cloning bias.

Genome	*Treponema azotonutricium*	*Treponema primitia*	*Streptococcus pyogenes* Alabama49 (Sanger-only)*	*Streptococcus pneumoniae* Taiwan19F-14 (hybrid)*	*Streptococcus pneumoniae* Hungary19A-6 (hybrid)*
# of scaffolds generated only from Sanger reads	243	205	112	n/a	n/a
# of scaffolds generated with both 454 and Sanger data	101	15	n/a	23	30
coverage depth only from Sanger reads	9.53	11.08	9.41	6.11†	5.23†
coverage depth with both 454 and Sanger data	28.37	31.13	n/a	20.94	18.39

* Streptococcus pyogenes is listed as a reference point for the other two Strep strains shown because of similarity in sequence. Although a hybrid version of the pyogenes strain does not exist, the scaffold information created only from Sanger reads gives a rough idea of what the Sanger-only scaffolds would have looked like for the other two strains, and how that compares with the hybrid scaffold information.

† The targeted coverage depth of Sanger reads in the hybrid Strep strains were intentionally set lower than our standard 8x coverage. Previous observations from shotgun sequencing for Strep indicate that obtaining a coverage depth of more than 6x does not significantly aid in closing additional physical gaps (i.e. cloning bias still persists).

Basecall discrepancies

Each consensus base of a contig is determined by calculations using the cumulative quality values of all the underlying reads. Before 454 shreds are incorporated with Sanger reads into assemblies, they are assigned low quality values, giving bias towards Sanger reads when calling the consensus.

This approach is particularly important for areas in which the 454 shreds collectively appear to call a different base or a different number of bases compared to the Sanger sequences.

We are currently observing a high frequency of discrepancies between 454 and Sanger data in homopolymer regions across genomes, where the shreds tend to call one less base. These areas include stretches as short as two bases.

This issue is described in 454's documentation of their sequencing technology, which is based on light emissions for each synthesized base of a DNA fragment[1]. In a homopolymer region, the light emitted from the last base of the stretch may be weaker than the emissions from the preceding bases, and thus would not be accounted for due to a weak signal. We have also observed the opposite case (where Sanger reads call one less base) at a lesser frequency, as well as discrepancies in non-homopolymer regions. All instances are represented on this poster.

454 calls one less base
454 calls a different base
454 calls one more base

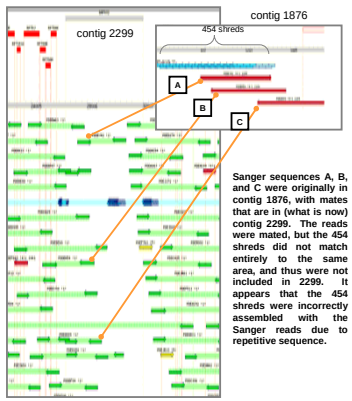
[1] Roche Applied Science, *GS FLX Data Processing Software Manual* (December 2006), 15, 45-46.

Hard Stop Regions

In addition to reducing scaffold number, 454 data has also helped sequence through difficult areas such as hard stops (i.e. 2° structures and homopoly-mer stretches). The data helps to reduce the time and effort devoted to resolving these difficult regions (however, these areas will still be confirmed with Sanger sequence). Shown here are some examples of hard stops where 454 data has provided sequence where the initial Sanger shotgun reads did not:

Misassemblies

As observed in many genomes, repetitive regions in the sequence can potentially cause misassemblies. In resolving these, we primarily rely upon useful information provided by Sanger reads – clone size and mate-pairing. We do not have similar information for the 454 data at this time, and therefore cannot relocate the shreds within a genome with confidence if there seems to be a misassembly. If the shreds disagree with Sanger reads that are correctly sized and oriented with their mates, we then place a bias towards the Sanger reads and disregard the 454 shreds. The following is an example of an assembly with both 454 shreds and Sanger sequences. We were able to move the Sanger reads to their correct location, but the shreds did not match correctly and had to be excluded.



Sanger sequences A, B, and C were originally in contig 1876, with mates that are in (what is now) contig 2299. The reads were mated, but the 454 shreds did not match entirely to the same area, and thus were not included in 2299. It appears that the 454 shreds were incorrectly assembled with the Sanger reads due to repetitive sequence.

Conclusion

As we continue to incorporate new technology and methods into our genome finishing process, we see both progress and areas that need modification and improvement. Based on our work with hybrid sequenced genomes, we observe that 454 data in combination with Sanger reads significantly reduces the number of physical gaps and helps to resolve difficult regions such as hard stops. We also observe, however, a high frequency of discrepancies in homopolymer stretches (and other basecall disagreements to a lesser degree), as well as the potential misassembly of 454 shreds with no clone or mate-pairing information readily available to resolve this occurrence. Due to issues such as these and because of the long-established validity of Sanger data, we currently place a bias towards the Sanger reads in calling the consensus sequence of our genomes and verify areas supported solely by 454 data. However, as 454 technology and other novel sequencing methods continue to develop, we continually strive to transform our finishing pipeline at JCVI to reflect and incorporate these advancements.

Acknowledgements

The authors would like to thank Daniela Puiu for providing a significant portion of the project assembly data and pipeline documentation necessary for this poster. Special thanks also to Luke Tallon for his aid in developing this poster and contributing critical information on specific projects and the finishing pipeline.