Advanced Closure Techniques for Whole Genome Alignment

FF0056

Heather Forberger, Xinyue Liu, Bradley Toms, Luke Tallon, Hoda Khouri, Nadia Fedorova

J. Craig Venter Institute

J. Craig Venter INSTITUTE

Abstract

The emphasis on sequencing multiple strains of the same organism have been on the rise, as learning about conserved regions has proven to be key for many genomic applications such as vaccine development. Following the completion of shotgun sequencing, a genome is assembled via the Celera assembler and scaffolded. The resulting assembly has many physical gaps. By taking advantage of the similarities between strains and using previously finished strains as a reference, we can expedite the finishing of a genome, by applying various new bioinformatics and automated techniques.

Physical/unlinked gaps typically consume a large portion of the closure process. As in theory, each end would have to be compared via PCR from whole genomic DNA to every other physical end within the genome to discover its proper orientation within the complete DNA strand. Although in recent years, the implementation of multiplex PCR has increased the efficiency of pairing ends, it is still an involved procedure. Genomes that have many unclonable regions will result in higher number of physical gaps, increasing the complexity and time spent orienting scaffolds.

We are going to explore several advanced techniques for aligning and closing these physical ends in a more direct manner using a reference genome. We have developed a bioinformatics pipeline using Perl, shell scripts, and Primer3 to achieve high-throughput design of PCR primer pairs at physical gap ends, screening out valid pairs based on their MUMMER alignment with the reference genome. Furthermore, 454 sequencing and subsequent hybrid assembly, as well as optical mapping show to be useful methods in which the strain of interest can be used as its own reference for Sanger sequence only contig alignment.

Although these techniques will prove to significantly reduce the time and labor associated with aligning physical ends, it may never work at 100%, as variations in sequence are typically present between strains. Some of the common differences include, but are not limited to insertions, deletions and recombination of DNA sequences. We will discuss the results of various trials conducted with each technique as well as the effectiveness in saving time and money in genome closure. This poster will reflect the advancements for each of these techniques, as well as their appropriate future applications.

JCVI Finishing Criteria

A genome is considered finished and ready if it satisfies the following criteria:

- A continuous consensus of DNA sequence
- No ambiguous consensus basepairs
- At least 2X sequence coverage over the entire genome.
- Both strands of one clone are sequenced two different clones
- Same clone sequenced with dual chemistry
- At least 2X clone coverage over the entire genome
- Complete confidence in all repetitive areas

What Causes Gaps?

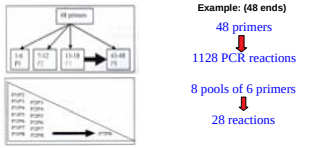
Gaps in a genome's consensus sequence can be caused by various factors, which include, but are not limited to the following:

- Non-clonable regions toxic areas to cloning vector (i.e. *e. coli*)
- Unsequencable regions hair pin loops or homopolymers caused by high GC or AT genome
- Non-random libraries
- Poor genomic DNA quality (Sheared, degraded, or contaminated DNA)
- Assembler issues (Repetitive areas)

Pairing Physical Ends

When resolving physical gaps, in theory one must design a primer from every physical end and react each primer with all other physical ends in the genome, using combinatorial PCR. This can become quite cumbersome in large genomes or difficult to clone genomes that contain upwards of 100 physical ends. In recent years, physical gaps have been closed, using a technique called Pipette Optimal Multiplex PCR (POMP). POMP follows a simple formula in which primers are grouped into pools of 7-16 primers each. These pools are then reacted with one another, significantly reducing the number of PCR reactions needed to mate physical ends. The resulting pool pairs can then be deconvoluted to determine the specific end that created the product.

Figure 1: POMP PCR pooling and reaction statistics



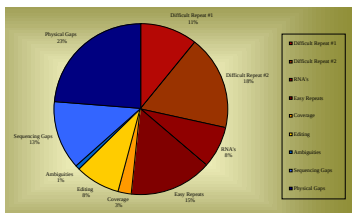


Figure 2: Typical Distribution of Genome Finishing Effort

As you can see, even using POMP, physical gaps monopolize much of the time spent closing a genome. However, with the advancement of technology, we have found several methods by which we can align these physical ends with a more direct approach. The implementation of these new techniques has helped significantly with whole genome alignment, which in turn has made the process of genome closure much more efficient and cost effective.

These concepts are as follows:

➢ **Comparative Analysis to a Reference:** By taking advantage of similarities between related genomes and using previously finished genomes as a reference, we are able to significantly cut the closure time of a project. By aligning the scaffolds to the FastA file of a reference genome, we can use direct PCR of pre-determined pairs to align ends, as well as design walking primers for 2x standard coverage of the gaps.

➢**454 Sequencing:** Because 454 sequencing does not have the cloning bias that plasmid cloning typically has it often is able to get sequencing reads where plasmid cloning cannot. This is advantageous because you can link many ends at once with minimal effort by comparing the 454 assembly and the shotgun assembly.

➢**Optical Mapping:** Optical maps, utilize restriction enzymes on whole genomic DNA that has been immobilized on an Optical Chip. The cut fragments are then stained and imaged to locate the cleavage sites and subsequently measured and mapped according to their orientation within the whole genome.

Comparative Analysis Using a Reference

Recently, we have developed a streamlined pipeline of scripts to automate the process of aligning an unfinished genome to a reference to expedite the closure of physical ends. The process is as follows:

- Align closure genome to reference using NUCmer
- Designed physical end primers using Primer3
- Pairs ends using script to order in 2, 96-well blocks
- NUCmer primers back to reference to design walking primers in both directions every 500 bp
- Set up PCR primers in bulk with a multi-channel pipette to react with Invitrogen High-Fidelity Supermix or Accuzyme
- Clean PCR products for end sequencing, using SAP
- Using the Hamilton STAR robot, cherry pick walking primers to match respective PCR products
- Reassemble with new reads
- Continue walking unclosed PCRs with original products

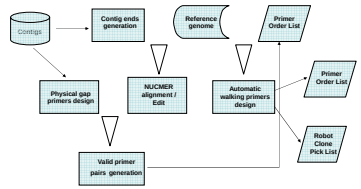


Figure 3: Automated reference genome alignment pipeline

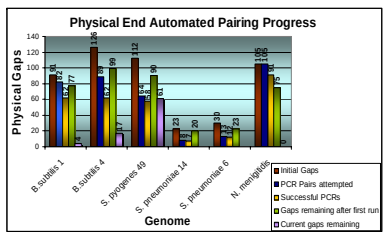


Figure 4: The above chart shows progress made using our automated pipeline to align physical ends and subsequently close physical gaps. Note that some genomes are still in the finishing phase, but are actively using PCR products produced using our physical end alignment scripts.

454 Sequencing

An approximately 2Mb genome was used as our example in our initial study. Both large and small insert libraries were made for this genome during shotgun. However, due to it's commonly toxic nature when cloned into *E. coli*, we began the closure process with 255 physical ends. We begin this process with an experimental automated POMP PCR, however, due to a high error rate, we were unable to get reliable data. We then aligned the remaining gaps using NUCmer, however, we were left with ~40 gaps, at which point we introduced the 454 sequences, in comparison.

454 Data:

➢One full run of regular 454 sequencing & half-plate of paired-end reads
- Received 240 contigs & 34 scaffolds
- Aligned to Sanger contigs with NUCmer

➢Generated a hybrid assembly
- Shredded 454 contigs - Celera Assembler

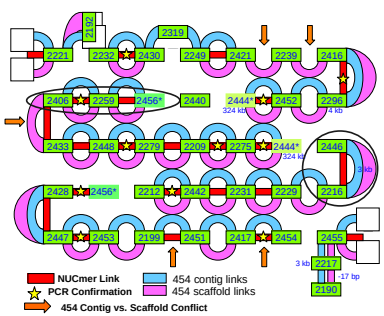


Figure 5: Whole Genome alignment of *S. pneumoniae* The Genome was first aligned to a completed reference genome. Gaps were then confirmed by both PCR and 454 sequencing data.

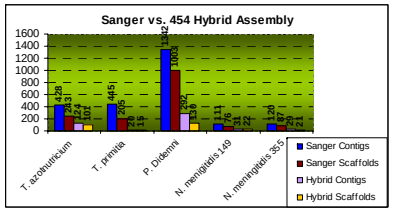


Figure 6: The above chart shows the significant reduction in contigs and scaffolds, when 454 shreds are incorporated with Sanger sequences. However, 454 shreds are still not 100% accepted for "gold standard" finishing.

Optical Mapping

Optical mapping (offered by OpGen) is a process that uses restriction enzymes to cut the DNA and the pieces are then labeled with a fluorescent dye that are read by fluorescence microscopy. This method typically uses whole genomic DNA with multiple chromosomes that are difficult to sequence completely and thus orient properly.

Optical Mapping follows a 6 step process:

- **Capture:** Whole genomic DNA is extracted and placed on a Optical Chip™, resulting in long, single DNA molecules.
- **Immobilize:** Using electrostatic forces, the DNA is immobilized on the chip, in multiple.
- **Interrogate:** Using restriction enzymes, the DNA is digested into fragments, to be characterized
- **Stain:** A stain is applied to the fragments that allows the restriction fragments and cleavage sites to be imaged by robotic fluorescence microscopy.
- **Scan:** The images are then scanned, to reveal gaps, corresponding to restriction sites in the DNA. These pieces are then measured and converted into an Optical Map™
- **Assemble:** A "consensus" map is assembled, typically giving >50x coverage. This map is then compared to the currently sequence fragments at TIGR and the findings returned with the correct orientation of each scaffold.

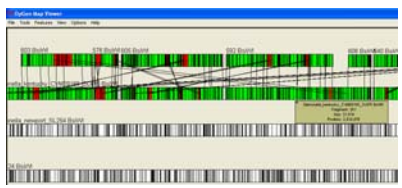


Figure 7.1: OpGen optical map of *S. enterica*, showing alignment of large contigs to completed reference strains, via restriction enzyme loci.

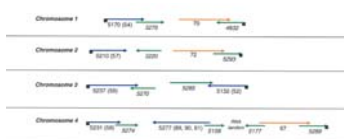


Figure 7.2: From the files received from OpGen, the closure team can then create the above graphical analysis of the scaffold orientation, that can be used to either close the remaining gaps or as a pseudomolecule for annotation.

Conclusions

➢ **Comparative Analysis to a Reference:** Having a previously sequenced genome of a similar nature, greatly reduces the amount of time spent pairing physical ends. A reference should be utilized as much as possible, for not only physical gaps, but forecasting possible repeats, plasmids, phages, walking large gaps all at once, locating rRNAs. However, keep in mind that there can be differences between strains, such as deleteions, insertions and recombinations.

➢ **454 Sequencing:** Because 454 sequencing uses DNA from the exact organism and strain as our Sanger sequences, the sequences provide very accurate linking data, when aligned and or assembles with our CA assemblies. This creates a pseudo-reference genome to align physical ends, as well as confirm areas of low coverage.

➢ **Optical Mapping:** Optical mapping offers a comprehensive overview of the alignment of a large genome that does not contain BACs. With the repetitive nature centromeres in genomes containing multiple chromosomes, in some cases this may be the only way in which to align the respective arms.

Future Recommendations

➢ **Comparative Analysis to a Reference:** We will continue to optimize the pipeline for aligning and sequencing physical ends. We would also like to continue to explore new enzymes that will most suit PCRs of unknown length and content at a cost effective rate. We will continue to investigate other ways to use completed genomes to close related genomes as much possible, keeping in mind that no 2 genomes will ever be 100% alike.

➢ **454 Sequencing:** Although we are now able to assemble and use 454 shreds in our whole genome closure process, there is still much improvement to be made to the length and quality of each read from the 454 before it can be considered "gold standard."

➢**Optical Mapping:** Although optical mapping can be very useful in aligning a genome, it is a very expensive process. It is only recommended in situations that have few other options, in most cases, large, multi-chromosomal genomes. However, it does have the advantage of using the same DNA as a reference and is very accurate in aligning larger genomes with few options for complete orientation otherwise.

References

➢Brown, Terence A. Genomes. 2nd ed. New York: John Wiley and Sons Inc., 2002.

➢ Fleischmann R.D., et. al. Whole -Genome Sequencing of *Haemophilus influenza* Rd. Science (269): 496-512. 1995.

➢Luo MC, Thomas C, You FM, Hsiao J, Ouyang, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82: 378-389

➢ OpGen Optical Map™ www.opgen.com

➢ Sutton G.G., et. al. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Science and Technology. 1(1):9-19. 1995 al

➢ Venter, J.C.; Remington, K.; Heidelberg, J.F.; et. al. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science*, 304, 66-74.

Acknowledgements

- J. Craig Venter Science Foundation
- TIGR Closure Department
- Herve Tettelin, Julie Dunning-Hotopp