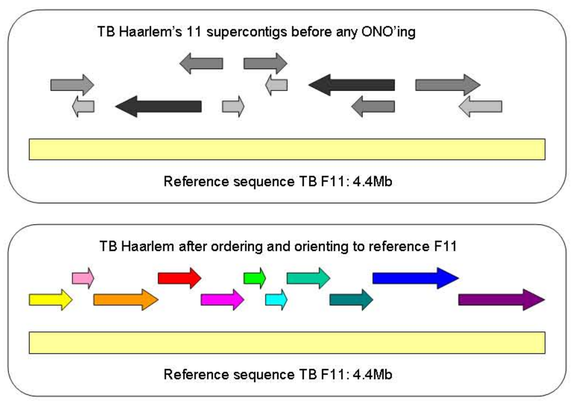
Challenges in Entire Genome Finishing: TB Haarlem

Margaret Priest, Kurt LaButti, Michael G FitzGerald

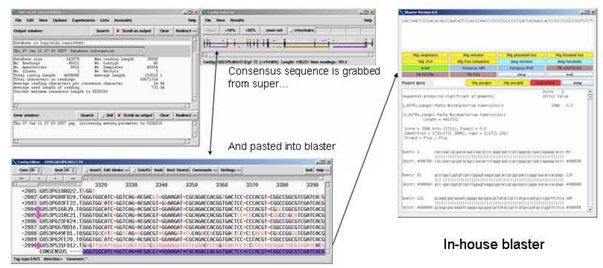
Since the completion of the Human and Mouse genomes our finishing efforts have shifted from BAC tiling to whole genome shotgun assemblies, focusing less on mammals and more on pathogens. We are finishing multiple isolates of pathogenic bacteria as part of an effort to further understand issues like pathogenicity and drug resistance. The resurgent pathogen Mycobacterium tuberculosis is part of this program. There are three currently finished M. tuberculosis genome sequences available, H37Rv (lab strain, Sanger), CDC1551 (TIGR) and F11 (Broad). The latest strain the Broad Institute is finishing is TB Haarlem. Here we will report on efforts to finish Haarlem and compare that to the process and results for F11. Issues including effect of shotgun quality and necessity of new informatics tools will be reviewed.

Whole Genome Finishing Process for TB Haarlem

The advent of whole genome shotgun sequencing has allowed completion of numerous bacterial and other genomes. Continued increases in sanger sequencing efficiency along with the arrival of higher throughput technologies sets the stage for full sequencing of numerous bacterial isolates. We construct shotgun sequencing libraries directly from genomic DNA, dramatically reducing mapping and library efforts. . Our current TB projects are assembled from a mix of pOT, pJAN and fosmid end reads with 4kb, 10kb and 40kb average insert sizes respectively. Sophisticated assembly tools like Arachne allow reconstruction of genomic sequence from paired reads. Several iterations of assemblies with differing coverage can be created with Arachne and then analyzed for gap count, LQRs, etc. Finishing work proceeds directly on the "optimal" genome assembly, utilizing various strategies of walks, transpositions and whole genome pcr.



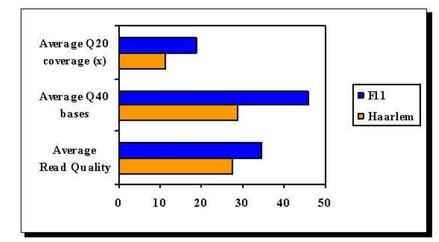
If the genome is not in a single scaffold, each supercontig will have to be ordered and oriented. When a reference sequence is not present, finishers must rely on links between parent reads of the available templates inside each supercontig to orient gaps to each other. Combinatorial PCR could also be used. For TB Haarlem, there were already several reference sequences to use for ordering and orienting supercontigs.



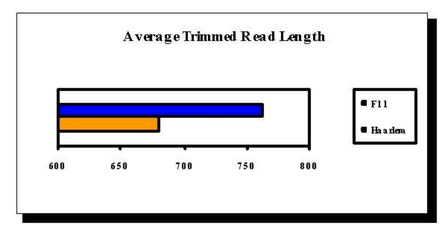
TB Haarlem is still a work in progress, but above is a screenshot from the latest gap assembly of TB Haarlem after the 11 supers were consolidated into one assembly (each super is denoted by a different colored consensus tag for ease of ONO'ing when future finishing work is loaded). The diagram demonstrates how consensus is grabbed and pasted into our in-house Blaster tool, and aligned to one of three reference TB strains available. This was an iterative manual process. The iterative process of blasting can be scripted; however because of the strict nature of blast, the simple fact that there will be LQRs within the contig in addition to possible misassemblies rules out the idea of complete automation of the ONO task. In addition, the process by which assemblies are picked for finishing, based on best statistics, is not always perfect. In some scenarios, assemblies that may have less gaps may have different LQRs than other iterations, leading to unincorporated reads that translates to loss of linking information between supercontigs. This can lead to redundant supers which can be weeded out through blasting.

When conventional templates become exhausted for walks and transpositions, whole genome PCR is saved as a last resort. It has an added complexity in that the searching for unique oligos needs to be unique across an entire genome, vs. BAC finishing where uniqueness is restricted to ~200kb of sequence. The above Blaster tool has become invaluable in this regard. If some gaps become exhausted of templates, the gap edges in TB Haarlem can be blasted against the other strains to screen candidate oligos for uniqueness. It may be challenging to select unique oligos, however, for future genomes where the microbe contains many duplications and there is no reference sequence to blast against.

TB Haarlem Specific Issues



In addition to the overall lower quality scores per read, F11 was linked across 45 gaps, in one super. Haarlem, as a direct result of the lower quality and average read length, was missing linking information and ended up in 11 supers with a total of 132 gaps.



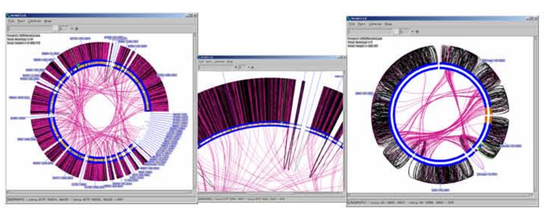
Trimmed reads are clipped at beginning around vector/primer annealing sites and at end when quality wanes

Further, the total GC content for both assemblies was compared before these assemblies entered finishing. At first glance, each strain had roughly the same distribution of bases. However, when genomes are 4.4MB of sequence, small differences in percentages can be significant. A small fraction of GC content was captured in F11 library that was not in Haarlem, 65.58% GC content vs. 65.51% respectively. This represents about 3kb of extra GC content that was not captured in Haarlem assembly.

Other artifacts of lower quality libraries include redundancy. Of the 11 uncaptured gaps that translated into 11 separate supers, there were 2 supers that really belonged inside larger contigs but because of the low quality and GC rich nature of the sequence, were not incorporated. In addition, a third super of 11kb truly belonged inside another supercontig, but due to a small collapsed repeat, was not assembled correctly.

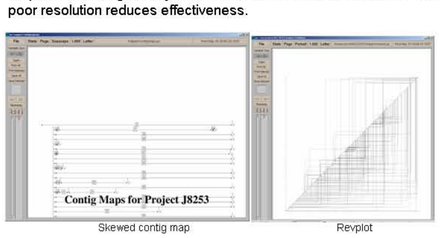
Overall, the numbers point to lower quality sequences across Haarlem's genome which contributed to higher gap counts that tended to have higher GC rich sequence. This has a direct impact on the amount of finishing involved between strains. Most F11 gaps were small enough to close with resequencing, while Haarlem had very few gaps which were resolved through resequencing. Workflows that were better suited towards GC rich sequence were also utilized to a higher extent.

Assembly Analysis Challenges With Large Arachne Assemblies



TB-Haarlem orchid file; ~4.4Mb — Loss of resolution — BAC orchid file; ~200kb

In the case of entire genome finishing such as TB Haarlem, misassembly detection will be more challenging using the conventional tools currently in our arsenal. For instance, the tool Orchid as seen to the left, is optimized for BAC sized projects. Any assembly larger than BAC dimensions provides too little resolution to detect incorrectly mated reads. Separating true assembly problems from noise becomes difficult with increasing genome size and sequence coverage. As you can see from the screenshot of TB-Haarlem's orchid file, this noise combined with poor resolution reduces effectiveness.



Skewed contig map — Revplot

Ordering and orienting contigs, a usually straightforward process, becomes more daunting with entire genome finishing. The increase in numbers of contigs and contiglets can reduce the effectiveness of some visual tools such as the postscript file to the right. There are too many contigs in TB Haarlem for the contigmaps file to display correctly. The screenshot on the right displays the mated pairs in a Revplot, which can be difficult to obtain any useful data from.

Future Directions of Whole Genome Finishing

So far, the benefits of whole genome finishing seem to be efficiency, you can see the big picture, and when reference sequence is available one can have an idea of gap count and nature of sequence at these gaps. A potential future challenge will be managing multiple strains and avoiding contamination in the lab. For instance, if there is a plate swap, there will be no way to know which strain is which since that particular microbe is undergoing the finishing process simultaneously.

In addition, new sequencing technologies including 454 can expedite the process of whole genome finishing. If integrated properly, new technology can take advantage of reducing costs by filling in basic sequence which was not captured through shotgun sequencing, at a much higher-throughput, though some platforms have issues with mononucleotide repeats and extreme GC content repeats are not captured. Error correction is needed in those circumstances.

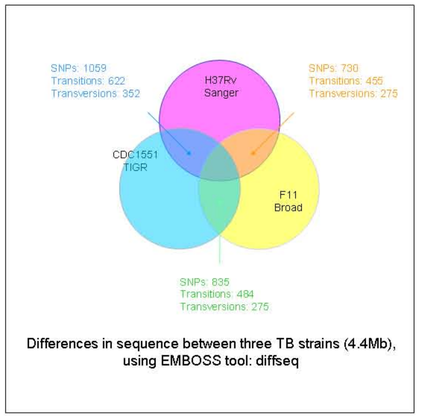
Developing ways to integrate these new sequencing data is currently in the works. See talks by Sean Sykes and Harindra Arachchi for further information on Broad's application of new sequencing technology.

For the computer finishing process, working on an entire genome requires tools that are better suited for more efficient analysis of the sheer increase in volume of information. We hope to have the following tools in development soon:

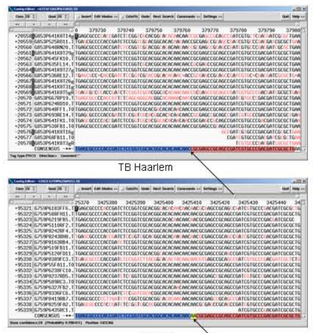
- Auto alignment tool
- A tool similar to Orchid to be able to view bad links for the entire genome, with equal resolution to Orchid
- Script for pulling out unlinked read names and creating mates file in order to pull these reads in via a trace server, to help link supercontigs when there is more than one super
- More automated fishing script, where one blasts new sequence extension (via loading finishing reads) in a trace server to pull in potential unincorporated reads plus their mates
- And finally, new sequencing technology quality scores that can have similar translation to the standard sanger quality scores

A Tidbit of TB comparative analysis

The largest high quality contig currently in TB Haarlem was compared against all three strains in our blaster, using EMBOSS tool diffseq. Already one can see some interesting differences between strains. Transversion is a substitution between a pyrimidine/purine; transition indicates the sequence change from common to variant allele like pyrimidine to pyrimidine or purine to purine. An analysis of these sequence differences between strains will assist in the understanding of the evolution within a genome, and may lead to the design of improved therapeutics.



Differences in sequence between three TB strains (4.4Mb), using EMBOSS tool: diffseq



An example of a 2bp insertion/deletion comparing Haarlem with F11 strain