Finishing Strategies, Sequence Validation and Genomic Analyses:
lessons from a GC- and repeat-rich bacterial genome

Warren RL, Morin RD, Yang G, Stott JM, Butterfield YB, Shin H, Smailus D, Schein JE, Siddiqui AS, Holt R, Jones SJM, Marra MA, McLeod MP, Mohn WW, Fukuda M, Davies JE, Eltis LD

Canada's Michael Smith Genome Sciences Centre
www.bcgsc.ca

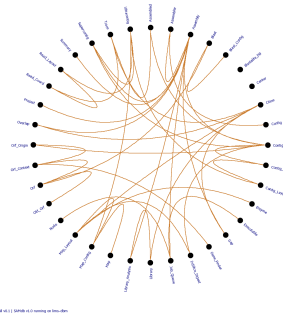
University of British Columbia
Department of Microbiology and Immunology

1. Introduction

Rhodococcus sp. RHA1 is a soil bacterium of environmental and industrial importance due to its ability to metabolize numerous organic substrates and its potential to produce secondary metabolites. Random Whole Genome Sequencing (WGS) of plasmid and fosmid clones yielded 9X coverage of the ~9.7 Mb genome.

In an effort to analyze the RHA1 genome, we have designed and implemented a system to manage whole genome shotgun sequences and whole genome sequence assembly (WGA) data flow. The Sequence Assembly Manager (SAM) consists primarily of a MySQL relational database and PERL applications designed to manipulate and coordinate the analysis of sequence information and to view and report genome assembly progress through its CGI web interface *sam.pl*.

2. SAM MySQL Relational Database

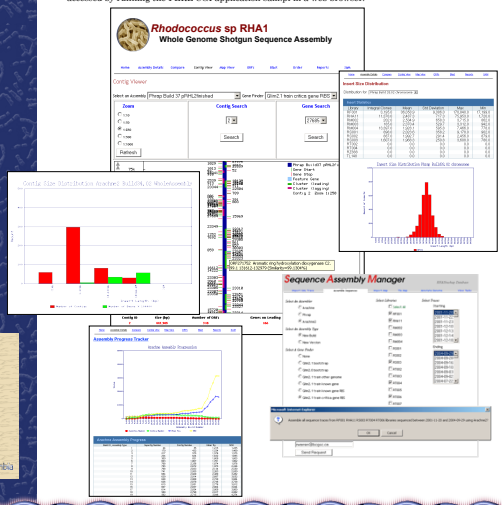


The key component of SAM is a MySQL relational database that stores **sequencing data**, **assembly information**, **fingerprint map data**, **genome analysis data**, **gene predictions** and **annotation information**.

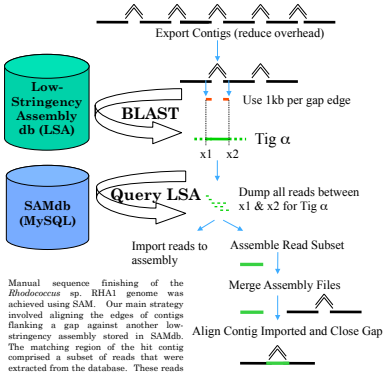
On the left, a MySQL schema of the Sequence Assembly Manager database (SAMdb) generated using schemaball v. 0.1

3. Visualization of WGA using sam.pl

The main purpose of SAM is to organize WGA in a structured way such that users can assess WGA quality rapidly using an intuitive web interface that allows easy evaluation of the assembly quality and its progress. SAM renders WGA data generic, upon which genome analysis tools and viewers for different sequence assemblers were designed. The images shown below are generated dynamically on the web and accessed by running the PERL CGI application sam.pl in a web browser.

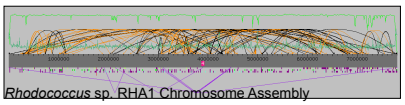


4. Strategy used to Finish the RHA1 Genome



Manual sequence finishing of the *Rhodococcus* sp. RHA1 genome was achieved using SAM. Our main strategy involved aligning the edges of contigs flanking a gap against another low-stringency assembly stored in SAMdb. The matching region of the hit contig comprised a subset of reads that were extracted from the database. These reads were reassembled and used to close the gap. At its peak throughput, this technique allowed the closure of 70 sequence gaps in a two week period.

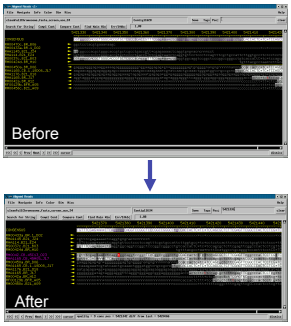
5. Resolving Repeats



The RHA1 genome contains numerous repeats that represented an enormous challenge for finishing. Most repeats were resolved during the final assembly. This was achieved using the Arachne assembler to create supercontigs for ultimate use as mini-assembly bins for Phrap. Manual finishing is labour-intensive, which is why we sought to automate the process.

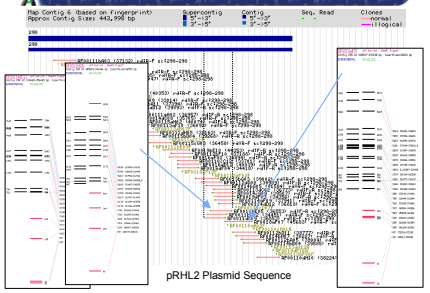
The above figure, taken from Consed's Assembly View, shows the entire 7.8 Mb RHA1 chromosome. Orange and black curves indicate direct and inverted repeats, respectively, of 0.5-20 kb and less than 10% mismatch.

6. Resolving Sequence Hardstops



Due to the GC-rich nature of the RHA1 genome, several clones could not be sequenced with our "standard" reaction mix. The use of dGTP reaction mix in a 5% DMSO solution helped tremendously, but did not yield data for regions with unusually high GC-content (>75%). Ultimately, 72% of these recalcitrant clones were sequenced using the Amersham Sequence Finishing kit, of which 93% were assembled. Analysis of selected traces revealed that this kit yielded longer traces with much less GC compression.
The above screenshots from Consed show a sequence hardstop before and after merging sequences obtained using the Amersham Sequence Finishing Kit*.

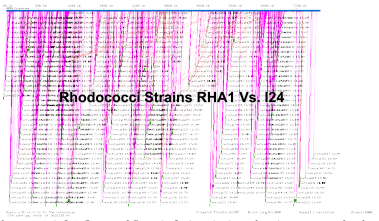
7. Sequence Validation using Fingerprint Maps in SAM



We had previously fingerprinted, end-sequenced and mapped a fosmid clone library. This was used to compute a supercontig layout and orientation for each sequence map contig (Figure above). The Map View feature of sam.pl provides an effective means of high throughput quality control by: (1) visualizing the clone tiling path; and (2) displaying *in silico* digestion band patterns predicted from DNA sequences. The latter may then be compared to the experimental fingerprint for each fosmid clone that projects onto the sequence assembly.

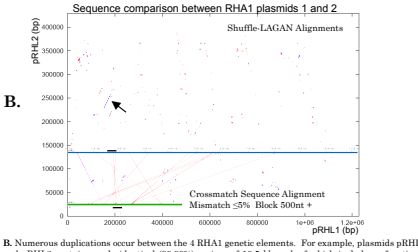
8. Sequence Comparison

A.



A. Rhodococcus strain I24 (Integrated Genomics Inc.) contigs (green boxes) were compared to the RHA1 genome (navy box) using Crossmatch. One kb matches with up to 20% mismatch were plotted using the PERL GD module. Alignments rarely have less than 10% mismatch. Of 6 Mb of I24 sequence, 3.6 Mb did not align to the RHA1 genome using these parameters.

B.



B. Numerous duplications occur between the 4 RHA1 genetic elements. For example, plasmids pRHL1 and pRHL2 contain nearly identical (99.99%) regions of 19.5 kb each of which includes a functional operon for phthalate catabolism (black bars, Crossmatch Alignment). This duplication is identified by shuffle-LAGAN, a glocal alignment algorithm designed to find rearrangements (black arrow, dotplot).

PROJECT FUNDING
GenomeCanada
GenomeBritishColumbia

Acknowledgments

I thank Martin Krzywinski for helpful discussions about PERL programming (& schemaball) and Greg Taylor for his initial work on the large chromosome linear invertrons.

References

Altschul SF et al. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.
*Amersham Sequence Finishing Kit: ©Amersham Biosciences Corp. (www.amershambiosciences.com)
Batzoglou S et al. 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12:177-89
Brudno M et al. 2003 Glocal alignment: finding rearrangements during alignment. Bioinfo. 19:54i-62i
Gordon D et al. 1998. Consed: A Graphical Tool for Sequence Finishing. Genome Res. 8:195-202.
Green P. 1996. http://bozeman.mbt.washington.edu/phrap.docs/phrap.html
Warren R et al. 2004. Functional Characterization of a Catabolic Plasmid from PCB-Degrading *Rhodococcus* sp. RHA1. J. Bacteriology (scheduled to appear in November).

Software

SAM is available for download @ http://www.bcgsc.ca/bioinfo/software/sam.pl
Schemaball is available for download @ http://mkweb.bcgsc.ca/schemaball