WHOLE GENOME SHOTGUN ASSEMBLY AND CHARACTERIZATION OF *RHODOCOCCUS* SP. RHA1, A PCB-DEGRADING ACTINOMYCETE

Warren R, Butterfield, Y, Dosanjh M, Petrescu A, Myhre M, Yang G, Scott JM, Schein JB, Shin H, Latreille P, Khattra J, Smailus D, Siddiqui A, Holt R, Jones S, Marra M, Mohn WW, Fukuda M, Davies J and Eltis LD

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

University of British Columbia
Department of Microbiology and Immunology

DEFINITIONS

N50: Contig length such that 50% of all bases are contained in contigs of length N50 or larger.

Q40 (or Phred40): Quality score for each DNA base. Q40 represents a probability of $10^{-40/10}$ (1:10,000) that the base was inaccurately called.

Supercontig (scaffold): An arrangement of ordered and oriented contigs facilitated by read pairing (clone) information.

Ultracontig: An arrangement of ordered and oriented supercontigs facilitated by a fingerprint map (Fosmid/BAC clone overlap).

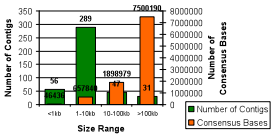
1. Abstract

Rhodococci are non-motile soil actinomycetes of biotechnological interest due to their ability to assimilate a wide-range of organic substrates. *Rhodococcus* sp. RHA1 is of particular interest for its ability to transform PCBs. The GC-rich (67%) genome of *Rhodococcus* sp. RHA1 comprises a chromosome of unknown topology (~8 Mb) and three linear catabolic plasmids: pRHL1 (1.1 Mb), pRHL2 (450 kb) and pRHL3 (332 kb).

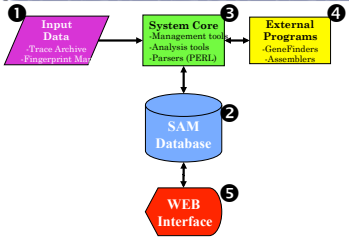
Shotgun sequencing of 76,416 plasmid clones and 9,984 fosmid clones yielded a genome coverage of 8X. We have used a combination of approaches to assemble and finish the sequence. One approach involved using the Arachne assembler to create supercontigs for ultimate use as mini-assembly bins for Phrap. This approach closed a considerable number of sequence gaps in a high throughput fashion using the Autofinish tool from Consed. Assembly analyses were facilitated using our Sequence Assembly Manager (SAM), a tool to manage and visualize whole genome sequence assembly data. Concurrently, we created a 25-contig fingerprint map from 4,973 fosmid clones. The map was used (1) to sort the mini-assemblies into the four RHA1 genetic elements, (2) to resolve a number of misassemblies through transposon sequencing of selected fosmid clones, and (3) to validate our sequence assembly using SAM. Using this approach, we have finished pRHL3 and almost finished the two larger plasmids. In our current assembly, 78 large contigs (>10 kb) comprise 98.9% of the genome (based on a 9.5 Mb genome), and 50% of all bases are contained in 16 contigs of length 210 kb and higher.

Automated annotation has revealed almost 10,000 putative open reading frames (ORFs). Analysis of these ORFs indicate that they include a large number of genes encoding oxygenases (133) as well as genes involved in the metabolism of secondary metabolites. The sequence of RHA1 will provide a valuable resource to understand the metabolism and physiology of rhodococci and actinomycetes.

Phrap Assembly of Arachne 2.0.1 Supercontigs Reads
Contig Size Distribution



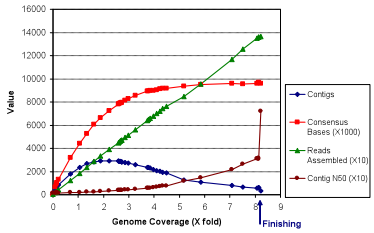
2. SAM: Sequence Assembly Manager



The Sequence Assembly Manager (SAM) consists primarily of a perl CGI web application (**5**) designed to easily manipulate and coordinate the analysis of genomic information and to view and report genome assembly progress. The user interface sits on top of a MYSQL relational database (**2**), created for the storage of trace archive, assembly information, gene predictions and annotations. SAM manages the execution of sub applications (**3**) required for the control of data storage, sequence assembly & gene prediction (**4**), XML file parsing (**1**), custom analysis and visualization of whole genome shotgun assembly data.

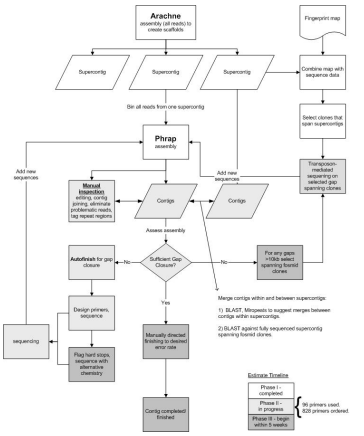
3. WGS Assembly Summary

RHA1 WGS Sequence Assembly Progression



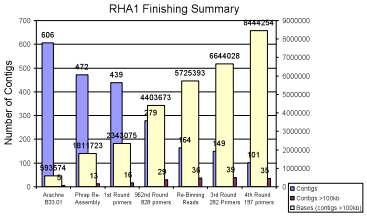
The graph above shows the progress of the whole genome shotgun assembly for *Rhodococcus* sp RHA1. It is interesting to observe that beyond 5X coverage, the addition of nearly 50,000 shotgun reads did not contribute significantly to the total number of good quality (>Q40) consensus bases but resolved a considerable number of gaps (decreased the number of contigs), generated much larger contigs and contributed to the read coverage. During the initial phase of sequence finishing (blue arrow), oligonucleotides were designed from gap-spanning clones to walk off contig ends (blue arrow). This approach has been extremely effective: it reduced the number of contigs by ~80% and more than doubles the contig N50 length.

4. Finishing RHA1



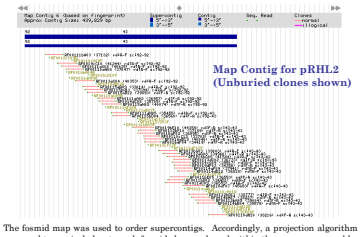
The finishing strategy uses a combination of the Arachne and Phrap assembly tools and the fingerprint map (see flow chart). Gaps have been closed in a high throughput fashion using the Autofinish tool from Consed.

RHA1 Finishing Summary



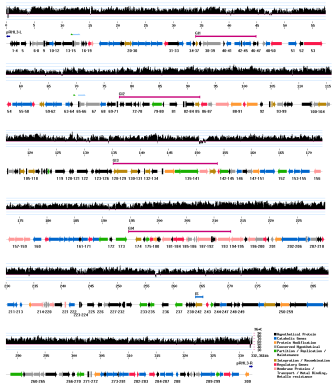
1,206 Primers were designed from selected clones to walk off the end of contigs. This reduced the number of gaps by 75% and increased the number of large contigs (>100 kb) by a factor of 20X (1st, 2nd and 3rd round oligo walks). Manual inspection of the contigs revealed a number of sequence hardstops causing low quality consensus and short clone gaps. Many of these hardstops were resolved in the 4th round of primer walk using an alternate sequencing chemistry (dGTP/DMSO).

5. Sequence Assembly Validation



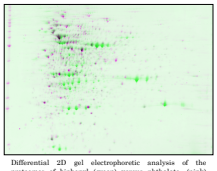
The fosmid map was used to order supercontigs. Accordingly, a projection algorithm was used to precisely locate each fosmid clone end-read within the sequence assembly. An arrangement of sequence ultracontigs is thus deduced from the fingerprint map and represented visually using SAM. This visual tool has been used to: 1) identify fosmid misassemblies and misplaced contigs; 2) merge contigs and supercontigs; 3) orient supercontigs in a given ultracontig; 4) confirm potential sequence breaks based on map contig boundaries; 5) segregate sequence reads from different supercontig bins by genetic element to help finish RHA1 linear plasmids (see flowchart above); and 6) select clones for transposon sequencing in order to resolve clone gaps.

6. Genome Annotation



The figure above depicts pRHL3 ORFs and their functional classification. Putative genes were identified and annotated using integrated automated and manual approaches. Open reading frames (ORFs) were independently predicted by Glimmer2.10 trained on a set of 500 known rhodococcal genes and by GeneMark-prokaryote. Hand curation was facilitated using Acedb and an in-house interface to our pRHL3 annotation MySQL database (mysql.com). ORF positions and annotations were further validated with BLASTP alignments and BLASTX alignments of pRHL3 sequences to nr-SPtrEMBL, NCBI-nr and *Rhodococcus* strain I24 (integratedgenomics.com) protein databases. Automated gene prediction and annotation for the whole genome is being conducted in collaboration with the Oak Ridge National Laboratories (ornl.gov).

7. From Genomics to Proteomics & Beyond



Differential 2D gel electrophoretic analysis of the proteomes of biphenyl (green) versus phthalate- (pink) cells.

Functional genomic investigations of RHA1 are aimed at better understanding the organism's catabolism and response to environmental stresses. A prototype microarray with probes for 2484 ORFs is being used for transcriptional analyses. Parallel proteomic analyses use high-resolution 2D electrophoresis and mass spectrometry (see figure). These functional genomic studies have elucidated several catabolic pathways for aromatic compounds and the roles of particular oxygenases.

The analyses indicate that a similar suite of enzymes is employed for biodegradation of both biphenyl and ethylbenzene. This suite includes at least 3 ring-hydroxylating dioxygenases, 2 dihydrodiol dehydrogenases and 3 ring-cleavage dioxygenases. The use of multiple isozymes may contribute to the superior PCB-degrading capabilities of RHA1. Biphenyl degradation also involves benzoate degradation genes (lower pathway) not involved in ethylbenzene degradation. Interestingly, some of the enzymes involved in benzoate catabolism are also responsible for phthalate catabolism. To further these studies, we are developing a system for targeted gene disruption in RHA1. In particular, putative regulatory genes will be tested by knockout analysis. This investigation has substantially furthered our understanding of an important group of soil bacteria and will facilitate the commercial exploitation of rhodococci and related organisms.

8. Acknowledgements

We thank Bob Fulton at Washington University in St-Louis (GSC) for insightful suggestions and discussions on sequence finishing. Julian Parkhill at the Sanger Institute provided expert advice and comments on our finishing strategy.

References

Altschul, SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES. (2002) ARACHNE: a whole-genome shotgun assembler. Genome Research. 12(1):177-89

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. (1999) Improved microbial gene identification with Glimmer. Nucleic Acids Research 27(23):4636-41

Durbin R., and J. T. Mieg. 1991. A *C. elegans* Database.

Gordon D, Abajian C, and Green P. 1998. Consed: A Graphical Tool for Sequence Finishing. Genome Research. 8:195-202.

Green, P. 1996. http://bozeman.mbt.washington.edu/phrap.docs/phrap.html

Warren R, Hsiao WWL, Kudo H, Myhre M, Dosanjh M, Petrescu A, Kobayashi H, Shimizu S, Miyauchi K, Masai E, Yang G, Stott JM, Schein JE, Shin H, Khattra J, Butterfield Y, Smailus D, Siddiqui A, Holt R, Jones S, Marra M, Mohn WW, Brinkman FSL, Fukuda M, Davies J, and Eltis LD 2004. Functional Characterization of a Catabolic Plasmid from PCB-Degrading *Rhodococcus* sp. RHA1. Journal of Bacteriology

PROJECT FUNDING

GenomeCanada

GenomeBritishColumbia