Management and Visualization of Whole Genome Shotgun Assemblies

René Warren, Yaron Butterfield, Steven Jones, Marco Marra

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

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

Abstract

We have designed and implemented a sequence data management system integrated with our genome annotation pipeline to deal with genome sequence assembly data flow. The Sequence Assembly Manager (SAM) consists primarily of a perl CGI web application 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 relational database, created for the storage of trace archive and assembly information. SAM manages the execution of sub applications required for the control of data storage, sequence assembly, file parsing, custom analysis and visualization of whole genome shotgun assembly data. The software includes a tool to compare sequence assemblies to fingerprint maps and this has already been useful in the identification of sequence misassemblies.

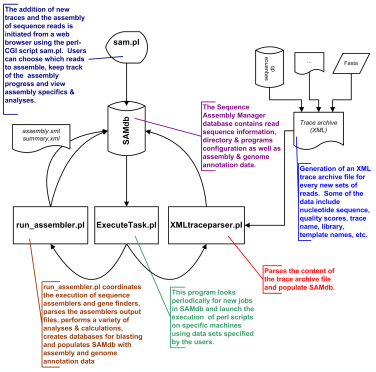
A main advantage of this system is the ease and flexibility at which genome assemblies could be performed using all the sequence data available, or a subset of it. We concurrently use Phrap and Arachne for all our genome assemblies and the system has been designed to incorporate easily any new assemblers as they become available. Modular programs have been written to parse the output files that each assembler generates, analyze their content and populate the database accordingly, with assembly -specific information. This includes general and specific information about contigs, supercontigs, sequence composition, coverage and build statistics along with custom calculations such as the identification of gap-spanning and overlapping clones, clone insert size distribution and library-specific read distribution within the assembly. The assembly data can be viewed on the web as a whole, or by comparing data obtained from different assemblers and assemblies. At any point, the progress of the assembly can be easily followed visually, through graphical representation of the data.

In order to facilitate the annotation of our assembly data, we plan to integrate existing components of our gene annotation pipeline to the assembly database. We have already developed an extensive set of independent tools to ease gene discovery and annotation, assign gene ontology and visually locate annotated genes in the context of their sequence assembly.

This tool will prove useful to other groups as well, especially when genome information from multiple projects must be administered simultaneously. It allows for more flexibility, control and ease of interpretation of sequence assemblies. When used in conjunction with sequence visualization software such as Consed, SAM has been an asset to help redirect further sequencing efforts.

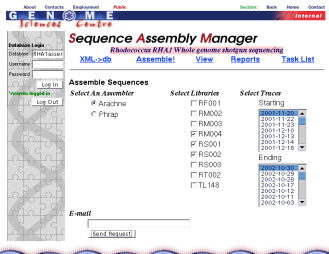
The data shown here pertains to the whole genome shotgun sequence of Rhodococcus sp RHA1, one of the best characterized PCB-degraders.

SAM Data Flow

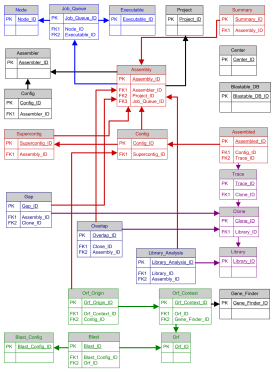


Custom Sequence Assemblies

Whole Genome Shotgun (WGS) sequence assemblies of specific sequences are facilitated by sam.pl, a web application (figure below). Users can custom-assemble sequence reads by choosing the assembler, libraries and the dates when the traces were obtained. Once selected, the task is added to a job queue where it waits to be executed. The assembly will be performed once ExecuteTask.pl runs.



SAM Database



A SAM database is created for every new WGS sequence assembly project undertaken at the GSC. For any new project, the content of a few configuration tables must be set to appropriate values in order for the system to work properly (depicted in black in the above figure). Trace archive information is then added to the database as the initial step of data acquisition (purple tables). Once a sequence assembly is complete, general build information and calculations based on current assembly statistics (e.g. Poisson Distribution, N50) are included. The results of assembly analyses, gene finding and gene annotation are then added to the database as they become available.

Assembly Analysis and Visualization

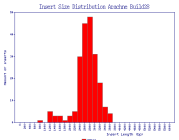
A variety of sequence assemblies analyses are automatically performed to help us assess the quality of the assembly, plan further sequencing efforts and evaluate the quality of genomic libraries. In order to help close sequence gaps, we generate a list of gap -spanning clones after every WGS assembly (table below).

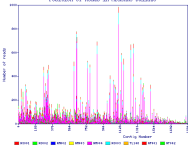
RATIONALE: For every paired-end reads originating from the same clone, we assess whether the reads are located on different contigs. If they do, they likely span a gap between these contigs. Gap length and contig ordering is estimated by looking at the read position and orientation in their respective contigs. Any confidence that a sequence gap truly exists is reinforced by the increasing number of gap-spanning clones identified for any two contig pairs. Clones from these gaps that belong to genomic library of smaller insert size could be chosen for transposon-mediated sequencing.

Gap Contig 1105-1106 (4 members)

Clone	Gap Size (bp)	Comment	Likely Tig Orientation
RM004498C07	2,108	-	
RM00460aA11	1,910	-	
RM00460aC04	1,845	-	
fl50013bH08	-11,421	misassembly?	

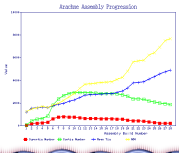
The position of paired-end reads within the assembly provide valuable **information on library quality and/or potential sequence misassemblies**. The figure on the right shows the insert size distribution for all the clones found on contiguous sequences for the genomic library RM002. In this example, we expect most clones to have insert lengths ~2.5kb. A shift in the distribution would indicate that the overall insert size for that library is different than what was originally intended. The presence of marginal clones positioned outside the distribution are usually due to misassembled reads and are easily identified.





The position of every single sequence read in the assembly allows the **identification of regions under-represented by certain genomic libraries** and assess, at a glance, the contribution of every read to the sequence assembly. Shown on the left is the the read distribution trend for Arachne Build 28 of RHA1 WGS assembly.

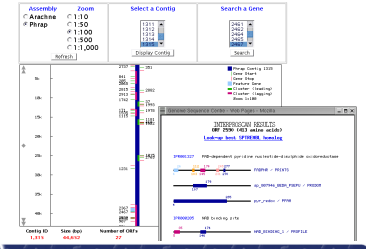
The perl-CGI sam.pl comprises a set of tools to graphically follow the assembly progress (figure on the right) and compare different assemblies (not shown). The assembly progression pages produce dynamic graphs based on user's selections. Every build completed by a specific assembler can be compared to the next in a single eye's view, thus permitting to evaluate more easily the **global trend of the assembly**. The assembly comparison tool is useful to monitor more closely the progress made since previous builds or investigate how two assemblers compare against one another on any given set of data.



Automated Genome Annotation

Putative genes are currently identified from the sequence assembly using Glimmer, a public microbial gene identification system. ORF coding sequences are extracted from their parental contigs and translated using the universal amino acid code. After screening for genes with good sequence quality (rationale below), genes are annotated using the high-throughput alignment program blastp from NCBI and the protein -domain finder InterProScan (EBI). The position of every ORF identified is represented visually in the context of its sequence assembly (figure below), facilitating the identification of gene clusters and operons.

RATIONALE: In order to confidently identify open reading frames (ORFs), gene coordinates are used to extract contig quality scores for every base encompassed by an ORF on a given contig. Only ORFs with overall base quality greater than phred 30 and with less than 0.25% phred 20 bases are kept.

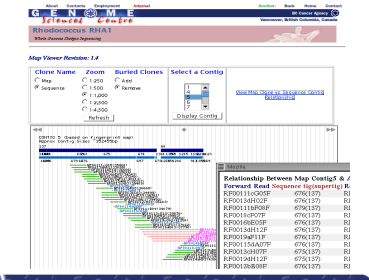


Sequence Assembly Vs. Fingerprint Map

Provided that a fingerprint map is available, a list of overlapping clones for every map contig can be used for ordering sequence supercontigs by finding the precise location of every clone end reads within sequence assemblies. An arrangement of sequence ultracontigs is thus deduced from the fingerprint map and represented visually using a perl-CGI application (below). A similar analysis, based on the sequence assembly is also performed. In this latter experiment, we localize fosmid clones within the fingerprint map for every fosmid end reads present in any given sequence supercontigs (not shown).

Together, these analyses have allowed us to perform 32 map contig merges for the current WGS sequence assembly of RHA1, decreasing the total number of fingerprint contigs from 60 to 28. These tools were also used to choose a set of 96 fosmid clones for further sequencing (transposon-mediated sequencing). Comparing the map and the sequence assembly will become increasingly important as we approach the finishing stage for this project. It will allow us to identify & resolve sequence misassemblies quickly, with more confidence, identify misplaced contigs, merge contigs, merge supercontigs, orient supercontigs in a given ultracontig and confirm potential sequence breaks based on map contig boundaries.

RATIONALE: From a list of known overlapping clones comprised in given map contigs (deduced by fingerprinting), we have localized their corresponding paired-end reads within our arachne sequence assembly. Merging of supercontigs into ultracontigs was permitted only if two or more reverse-forward links from overlapping clones spanned the gaps between two given supercontigs.



Acknowledgments

Heesun Shin, Chris Fjell (RHA1 fingerprint map)

Robert Kirkpatrick, Jerry Liu, Ran Guin & Andy Chan (Helpful Discussions)

References

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

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using PHRED.I.Accuracy assessment. Genome Research. 8:175-85

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. Nature Genetics. 3(3):266-72

Green P. 1994. PHRAP documentation. http://www.phrap.org/

Zdobnov EM, Apweiler R. 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17(9):847-8