

MINERVA, ARCTURUS, ARTEMIS: The Development of new finishing tools at the Wellcome Trust Sanger Institute Pathogen Sequencing Unit

Danielle Walker, Karen L. Mungall, Mandy Sanders, Carol Churcher, Andrew Barron, Mark Simmonds, Pathogen Sequencing Unit
The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA. U.K.

Abstract

The techniques and software used in Finishing have made considerable changes and progression over the last few years. The Sanger Institute's Pathogen Sequencing Unit has committed to sequencing more comparative projects and significantly larger genomes. The changes associated with these, for a finisher, are the ability to use comparative genomes when finishing project and the introduction of the Minerva/Arcturus interface and assembly data management system.

Completed comparative genome sequences have been of particular use in reducing the amount of time it takes to finish a project and in highlighting differences between related organisms. Artemis: a DNA sequence viewer and annotation tool, and ACT (Artemis Comparison Tool): a DNA sequence comparison viewer, have been used in annotation for some time and are now being used with great success as finishing tools. An example of this has been the continuing work at the Sanger Institute finishing several different *Streptococcus suis* SC84 using Sanger capillary shotgun data and 454 Shotgun data is also utilising these comparative tools.

There has been good progress made over the last year on the Minerva interface, which has been used successfully on several large projects using the Arcturus assembly data management system which was developed in house. Minerva is a Java application that will enable visualisation of a current genome assembly and scaffold. Of note recently, work on *Plasmodium knowlesi*, *Candida parapsilosis* and various *Haemonchus contortus* clones has successfully utilised this system. We envisage that it will be useful for prospective Helminth projects which are significantly larger than the genomes currently in finishing.

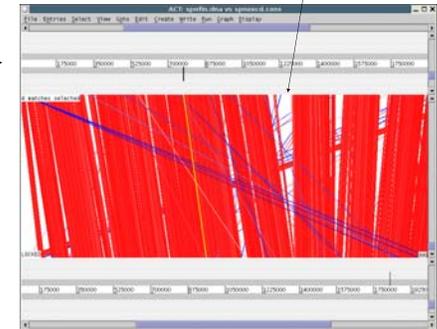


Artemis

Completed comparative genome sequences have been of paramount importance in reducing the amount of time it takes to finish a project, and in highlighting differences between related organisms. Below are two ACT comparisons showing closely related organisms and more distant relations.

Sequence insertion in *S.pneumoniae* type 23F

Finished sequence of *Streptococcus pneumoniae* type 23F

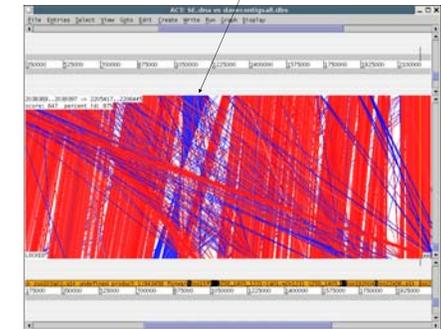


Unfinished sequence of *Streptococcus pneumoniae* OXC141

ACT comparison showing *Streptococcus pneumoniae* type 23F compared to *Streptococcus pneumoniae* OXC141. Both of these organisms are causative agents of many human diseases including pneumonia and meningitis. The coloured regions show synteny between the two genomes and the clear regions significant differences in the genetic sequence.

This area represents an inverted area of sequence

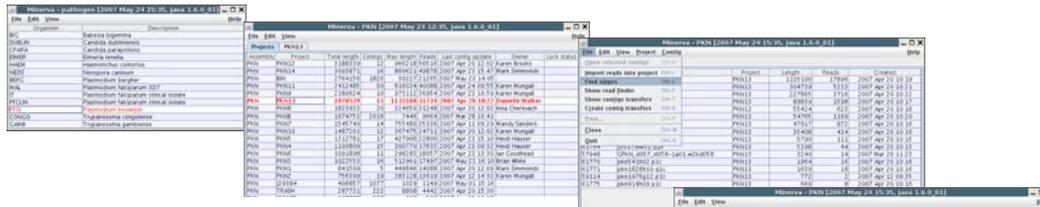
Finished sequence of *Streptococcus equi*



Unfinished sequence of *Streptococcus zooepidemicus*

ACT comparison showing *Streptococcus equi* compared to *Streptococcus zooepidemicus*. These two organisms are less closely related, however both affect the same host organism, horses. *S.equi* causes strangles, and *S.zooepidemicus* causes Inflammatory Airway Disease. The coloured regions show synteny between the two genomes and the clear regions significant differences in the genetic sequence.

Minerva & Arcturus



Arcturus, an assembly data management system, is based on the open-source MySQL database system. The core software is written in Perl and consists of a set of scripts which allow Arcturus to be integrated into the existing assembly pipeline, and to interact with existing finishing tools. Arcturus gives finishers the flexibility to work on large genomes, biologists the ability to track and update gene annotations on unfinished genomes, and project managers the ability to track the cost and progress of sequencing projects.

A graphical user interface known as Minerva is currently under development. Minerva is a Java application that will enable visualisation of a current genome assembly and scaffold. There has been good progress made over the last year on the Minerva interface, which has been used successfully on several large projects.

Minerva's current features allow you to:

- View the current state of one or more projects in any assembly.
- Examine the contigs in any project.
- Search for oligo sequences in the current contigs and/or the unassembled reads.
- Query the status of any read by name.
- Request the transfer of one or more unassembled reads into a particular project.
- Request the transfer of contigs between projects.

The Pathogen Sequencing Unit finishes a wide range of projects to varying finishing criteria. Minerva will be able to facilitate these changing needs. It can be used to draft projects at a scaffolding level, and to complete projects to a high finishing standard.

The diagrams shown give a representation of the tools available to use in Minerva, using a *Plasmodium knowlesi* chromosome (PKN13) as the example.