Producing high-quality statistical software

I. Milne, F. Wright, C. Hackett & J. McNicol

BioSS works at the interface between the development and application of quantitative methodologies. We are strongly committed to the dissemination of state of the art quantitative methods to a wide range of recipients, including the scientific community, research students, government and the bioindustries. Here we describe two user-friendly software applications produced for this purpose.

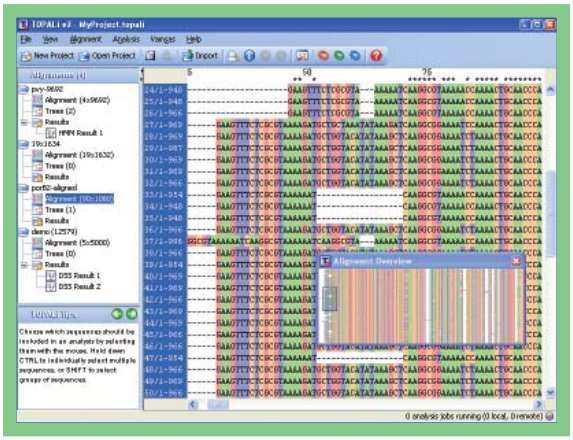


Figure 1 A 90seq by 1080bp DNA multiple alignment viewed in zoom and overview modes.

TOPALi (Fig. 1 and 2) Conventional phylogenetic tree estimation methods, applied to DNA multiple alignments, assume that all sites have the same evolutionary history. This assumption is violated if recombination has occurred among any sequences. Recombination produces mosaic sequences, which may cause errors in phylogenetic tree estimation. If recombination is possible, a check for mosaic sequences is essential prior to phylogenetic analysis.

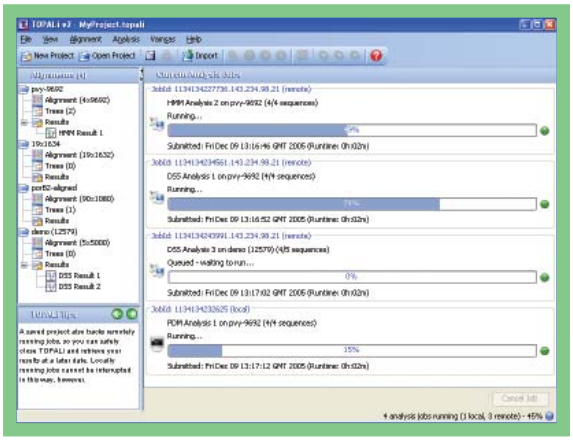


Figure 2 TOPALi allows for multiple analysis jobs to be run simultaneously on remote clusters.

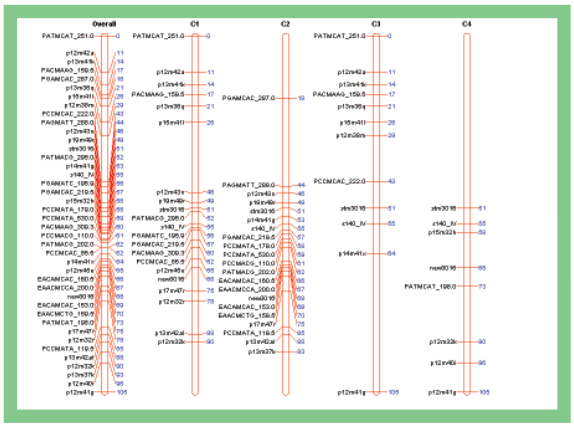


Figure 3 Linkage map of potato linkage group IV for the SCRI breeding clone 12601ab1.

Our software, TOPALi, provides an interface to three methods of recombination detection (Difference of Sums of Squares, Probabilistic Divergence Measures, and a Hidden Markov Model) that look for changes in phylogenetic tree topologies as we move a window along an alignment. These methods differ in the number of sequences that can be analysed and in their computational speed. TOPALi provides a complete graphical analysis tool for detecting recombinants in DNA multiple alignments. All tasks can be automated, requiring minimal user-intervention.

Recent development of TOPALi (in collaboration with the University of Dundee and the European Bio-

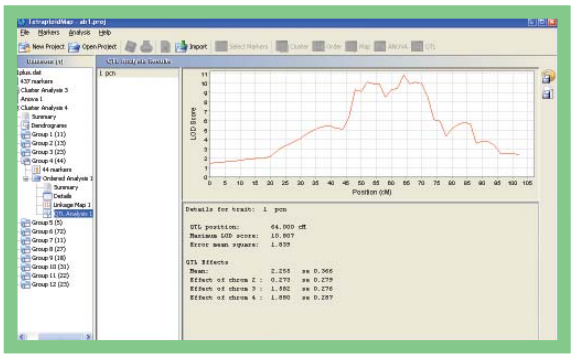
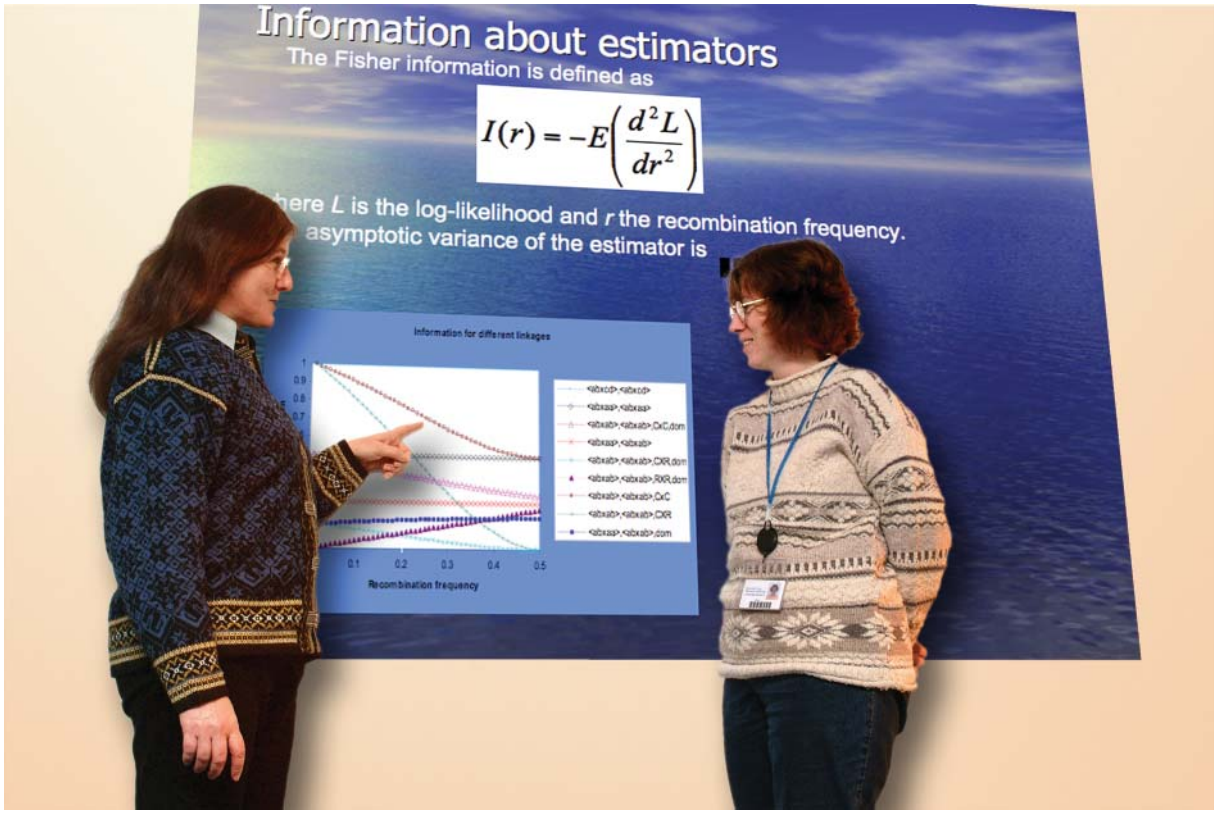


Figure 4 Lod profile for a QTL affecting PCN resistance on linkage group IV of 12601ab1.



informatics Institute in Cambridge) has concentrated on increasing its efficiency by programming many of its analyses for parallel computation, and accessing and running them via one or more high-performance computing clusters. As part of this work, new ways of interacting with the statistical programs running remotely have been designed, primarily through the use of web services. The ultimate goal is for the user to benefit from accessing distributed, high-performance facilities (such as the 28 CPU cluster at SCRI) from their normal desktop environment via a seamless high-quality graphical interface.

TetraploidMap (Fig. 3 and 4) The TetraploidMap program has been developed for calculating linkage maps and QTL mapping for autotetraploid species, such as potato. The program can be used to analyse both dominant and codominant markers scored for two parents and their offspring. Cluster analysis is used to separate the molecular markers into linkage groups. Recombination frequencies are calculated between all pairs of markers within a linkage group using the EM algorithm, and simulated annealing is used to order the markers.

Phenotypic traits, such as yield or disease levels, can be analysed by two methods. A simple marker regression method analyses each marker in turn to detect the markers that are most closely associated with the trait. Interval mapping can then be carried out for the linkage groups containing these markers to find the most likely QTL location and its mode of action (simplex, duplex, dominant etc.).

TetraploidMap is based on methodology developed by BioSS in collaboration with SCRI scientists since 1996, partly funded by a grant from the BBSRC GAIT initiative. It is also being used by groups in Europe and North and South America working on several autotetraploid species, including potato, leek and alfalfa.