Sequencing and re-sequencing Brassica genomes

Chris Duran[1], Jiri Stiller[1], Mike Imelfort[1], Dominic Eales[1], Chang Pyo Hong[1], Paul Berkman[1], Daniel Marshall[1], Megan Vardy[1], Harsh Raman[2], Jacqueline Batley[3], David Edwards[1]

[1]School of Land, Crop and Food Sciences and Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, 4072, Australia, [2]EH Graham Centre for Agricultural Innovation (an alliance between NSW Department of Primary Industries and Charles Sturt University), Wagga Wagga Agricultural Institute, Wagga Wagga, 2650, Australia and [3]School of Land, Crop and Food Sciences and ARC Centre of Excellence for Integrative Legume Research, University of Queensland, Brisbane, 4072, Australia.

ABSTRACT

Brassica genomes are relatively large and complex due to historic duplication events, the amplification of families of transposable elements, and polyploidisation. The development of second generation DNA sequencing methods is rapidly changing plant genome research and we are applying this technology for the analysis of the Brassica genomes. We have generated genome sequence data for several Brassica species and developed tools for the analysis of this data. These tools can be applied for gene and molecular marker discovery to support Brassica crop improvement.

Key words: Genome sequencing - genetic map visualisation - marker database - gene discovery - gene annotation.

INTRODUCTION

The application of second generation DNA sequencing technology is rapidly changing Brassica genome research. The sequence of the A genome is almost complete, produced by whole genome assembly of Illumina GAII short reads. The C genome is also rapidly being sequenced with expected completion during 2010. These reference diploid genomes will provide the basis for sequencing the amphidiploid canola genome and for genetic diversity analysis across Brassica species. The cost of producing genome sequence data continues to fall and we have produced a large quantity of reference genomic data for each of the diploid Brassica genomes.

These Brassica sequencing projects are generating volumes of data that cannot be easily analysed using traditional bioinformatics methods and this creates a set of unique challenges that do not exist with traditional long-read sequencing. We have been developing a number of tools to interrogate and analyse this sequence information to accelerate research in Brassica crop improvement. These tools have applications in the areas of integrative genomics, gene discovery and gene annotation. This paper demonstrates three of the tools we have developed for these applications in genomic research: CMap3D, TAGdb and BAC and gene annotator (BGA). The tools are accessible to all researchers via the web (http://acpfg.imb.uq.edu.au/). CMap3D is a stand-alone application which can be downloaded from this site.

MATERIALS AND METHODS

CMap3D

CMap3D has been developed using a client/server approach to ensure compatibility with current CMap databases. The CMap3D viewer accesses comparative map data and displays it as maps in 3D space. The CMap3D Viewer is a stand-alone client available for Windows, Linux and OSX. The CMap3D client first connects to a centralised repository listing server which provides the client with a list of available and compatible CMap repositories and their details. The client then communicates directly with the repository server to request and retrieve the required data. The client software uses the HTTP protocol for data transfer, to minimise institution network security conflicts.

The Brassica CMap repository currently hosts 23 map sets, representing 318 linkage groups. There are a total of 6902 marker annotations for these linkage groups, representing

approximately 4899 unique markers, making this resource a comprehensive database for Brassica comparative genetics and genomics.

TAGdb

TAGdb stores paired-end sequencing data produced from Illumina Solexa sequencers in binary form as a BLAST-enabled database (Altschul et al., 1990). A web-based interface allows researchers to upload or input a FASTA formatted nucleotide sequence up to 5000 base pairs long. The user may select one or more short-paired read libraries for comparison with their input sequence. The input sequence is aligned with short-reads of significant identity using MEGABLAST (Zhang et al., 2000). Each submitted job has a unique identifier and an email is sent to the user once the job has completed.

Visiting the job page shows the short-read alignment. TAGdb presents data using IGLOO, a map tiling engine developed in-house using OpenLayers (http://www.openlayers.org). The input sequence is represented horizontally, with reads represented by arrows, the colour of the arrow reflecting the sequence library. Arrows are connected with a line if they represent paired reads, and these lines have a heat-mapped confidence marker in the centre with blue indicating a good representation of the expected insert length and read orientation. A red marker indicates poor confidence for the given representation.

The TAGdb currently hosts short-read sequence data for all three Brassica genomes: *Brassica rapa*, (with multiple libraries representing chiifu and kenshin cultivars), *Brassica nigra* and *Brassica oleracea*.

BGA

The web-based interface allows researchers to upload a FASTA formatted nucleotide sequence file. This file may contain a single entry representing a Bacterial Artificial Chromosome (BAC) size sequence or multiple entries represented a set of gene sequences such as expressed sequence tags.

The genomic sequence data is processed using RepeatMasker (http://www.repeatmasker.org) to identify and mask repeat sequences. Several databases of repeats are available for selection by the user and additional repeat databases may be hosted on request. Genscan (Burge and Karlin, 1997) performs gene structure prediction, and these genes are annotated for potential function using BLAST (Altschul et al., 1990) comparison with the UniProt (Bairoch et al., 2009) and Genbank (Benson et al., 2009) databases, as well as reference genomes such as Arabidopsis.

RESULTS

Integrating and visualising the genetic and physical maps using CMap3D

One of the greatest challenges in biotechnology is the linking of observed heritable phenotyopes with the underlying causal genetic variation (Edwards and Batley, 2004). Many traits of agronomic importance have been mapped to regions of genetic maps as quantitative trait loci (QTL). However linking regions on genetic maps to physical sequenced genomes requires corresponding features between datasets and a means of visualisation/interpretation. Being able to compare genetic maps provides valuable information on the number of genetic loci corresponding to complex traits across populations which may aid the selection process. The comparison of genetic maps with sequenced genomes allows the identification of the genes underlying traits, providing insights into the biological mechanisms responsible for the trait. CMap (http://www.gmod.org/cmap) is a popular tool for the comparison and visualisation of genetic maps and sequenced genomes. It has been successfully applied for intra- and inter-species comparison within and between species, including Brassica (Lim et al., 2007). We have extended this CMap database to include physical map data for Brassica and Arabidopsis and have developed a tool to visualise this data in three-dimensional space, overcoming some of the limitations of the original CMap viewer. The implementation of CMap3D allows users to directly compare multiple genetic maps simultaneously and adapt their visualisation of maps and features on-the-fly (see Fig. 1). The ability to directly compare and align traits located on genetic maps with sequenced genomes will permit the identification and characterisation of Brassica traits at the genome level when the A genome sequence becomes available. This tool is being used to assist in the assembly of the Brassica A genome sequencing project and

permits the validation of the Brassica genome assembly through comparison with the reference *B. rapa* genetic map (Choi et al., 2007). A reference genome sequence for Brassica provides the foundation for detailed molecular variation analysis and the association of genes with agronomic traits.

Gene and promoter discovery using TAGdb

The latest whole genome sequencing technology can be applied for gene, promoter and marker discovery without the requirement to assemble the complete genome. We have generated short paired read data for several Brassica genomes and developed TAGdb to interrogate this data. TAGdb is a database of paired-end short read data that provides alignment with uploaded query sequences, with graphical and tabulated results (see Fig. 2). Where abundant reads are available, these can be used for local assembly and SNP discovery. Where there are fewer sequence reads, the sequence tags can be used to amplify and sequence the genomic region using the polymerase chain reaction (PCR) technique. TAGdb permits the identification of gene orthologs and homologs from each of the Brassica genomes and can be used to identify allelic variation between cultivars which may be responsible for phenotypic variation.

Gene and genome annotation using the BAC and Gene Annotator

The BAC and Gene Annotator (BGA) has been developed to allow users to annotate all the As, Gs Cs and Ts, produced by gene and genome sequencing projects, in real time over the web. The user submits the sequence for processing and sets the annotation parameters; the sequence is then automatically annotated using the gene discovery and annotation pipeline; finally, the results are displayed through an interactive web browser (see Fig. 3). BGA can also perform comparative analysis to identify orthologous regions of fully sequenced genomes

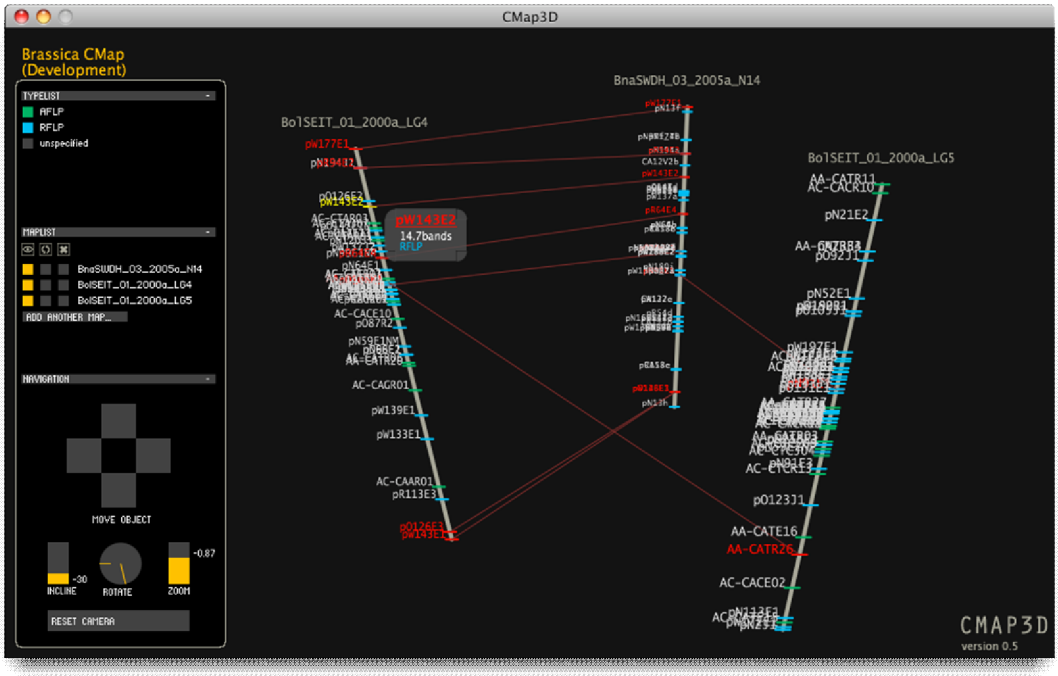

Fig.1. CMap3D interface showing correspondence between two Brassica genetic map linkage groups and an Arabidopsis physical map, representing chromosome 4

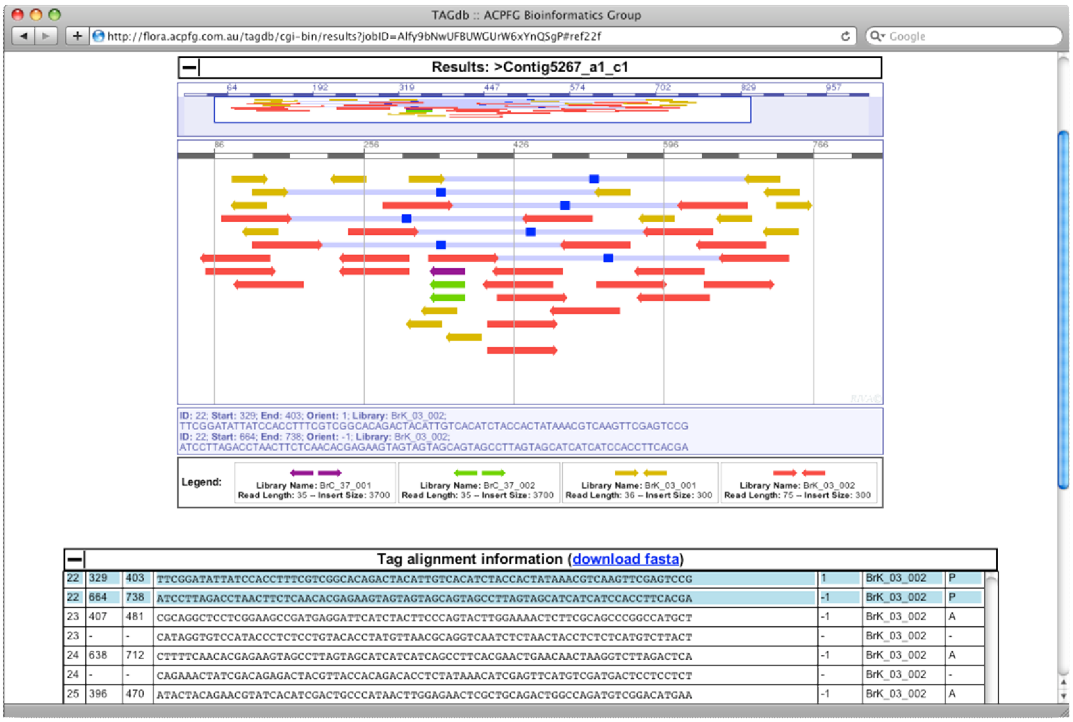

Fig.2. Screenshot of TAGdb showing the positions of tags mapping to a Brassica UniGene

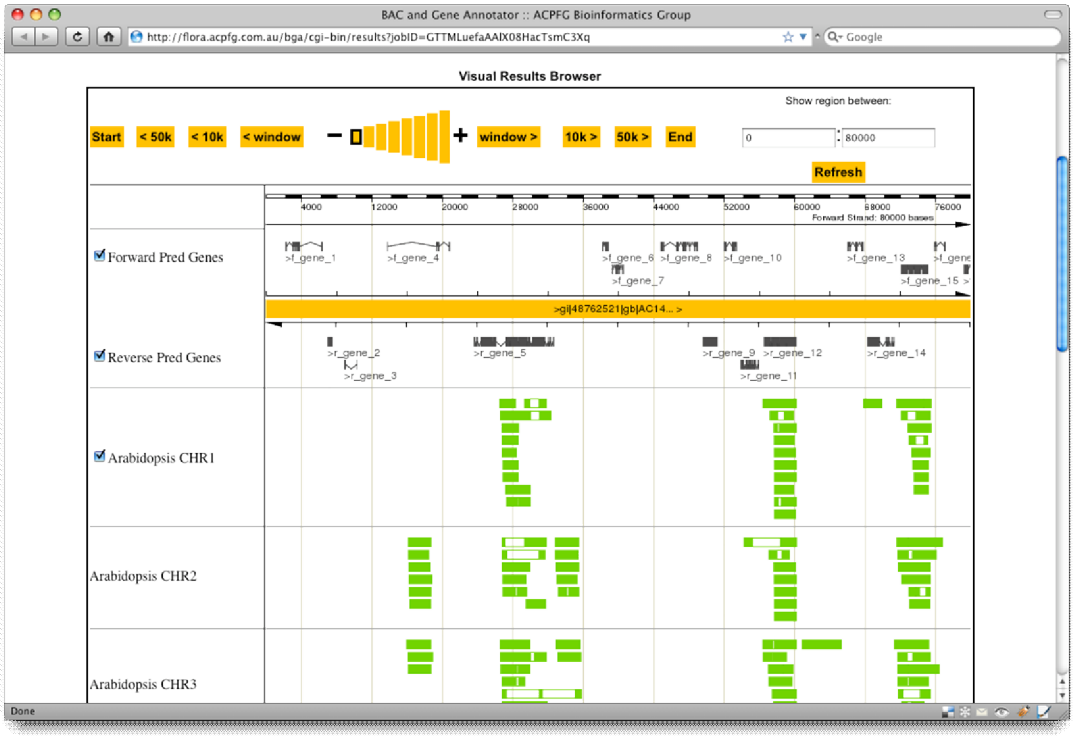

Fig.3. Screenshot of BAC and Gene Annotator showing positional information on a Brassica BAC of predicted genes and regions of homology to Arabidopsis chromosomes

DISCUSSION

DNA sequencing technology is undergoing a revolution with the introduction of second generation DNA sequencing such as the Illumina genome analyser. This flood of data provides opportunities to gain a greater understanding of crop genetics and genomics while providing challenges for bioinformatics researchers who need to manage and interpret the data. We have developed tools for the management and interpretation of genome sequence data with the aim of making this available for crop improvement programs. We have generated second generation sequence data for each of the three Brassica diploid genomes and established a public database for the mining of this vast dataset. To complement this, a web based DNA sequence annotation pipeline has been developed with Brassica specific parameters for gene discovery in Brassica crops. Finally, we have developed a Brassica genetic marker database and comparative map viewer to provide a link between the genetically mapped traits and sequenced genome information. These tools and data are publicly available without restriction through the web at: http://acpfg.imb.uq.edu.au/

ACKNOWLEDGEMENTS

The authors would like to acknowledge funding support from the Grains Research and Development Corporation (Project DAN00117) and the Australian Research Council (Projects LP0882095, LP0883462 and DP0985953). Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF), the Australian Partnership for Advanced Computing (APAC) and Queensland Facility for Advanced Bioinformatics (QFAB) is gratefully acknowledged.