

# Comparative analysis of *Brassica* genomes using second generation sequencing

Chris Duran<sup>1,†</sup>, Jiri Stiller<sup>1</sup>, Paul Berkman<sup>1</sup>, Megan McKenzie<sup>2</sup>,  
Jacqueline Batley<sup>2</sup> and David Edwards<sup>1,‡</sup>

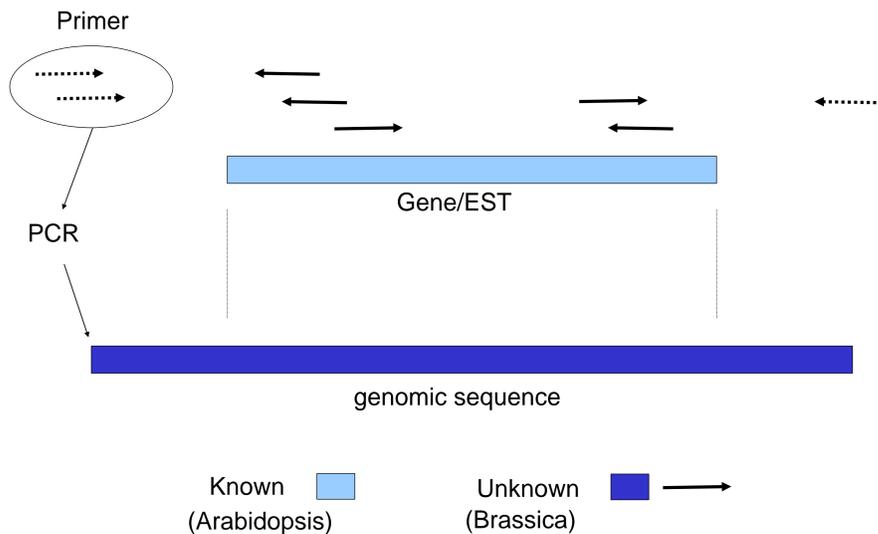
<sup>1</sup>Australian Centre for Plant Functional Genomics, School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia.  
<sup>2</sup>ARC Centre of Excellence for Integrative Legume Research, School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia.  
†c.duran@uq.edu.au ‡Dave.Edwards@uq.edu.au

## Abstract

The ability to produce vast quantities of short paired read DNA sequence data is creating opportunities to characterise complex plant genomes. As well as genome assembly, this data can be used for gene and promoter discovery, genome sequence validation, genome annotation, repetitive element characterisation, comparative genomics and molecular marker discovery. We have generated Illumina GAll paired read sequence data for divergent *Brassica* genomes. Comparative analysis of these genomes identifies regions of sequence conservation and divergence at the whole genome level through to single nucleotide and short indel variation. We present examples of second generation *Brassica* comparative genomics and demonstrate custom analysis and visualisation tools for second generation comparative genomics.

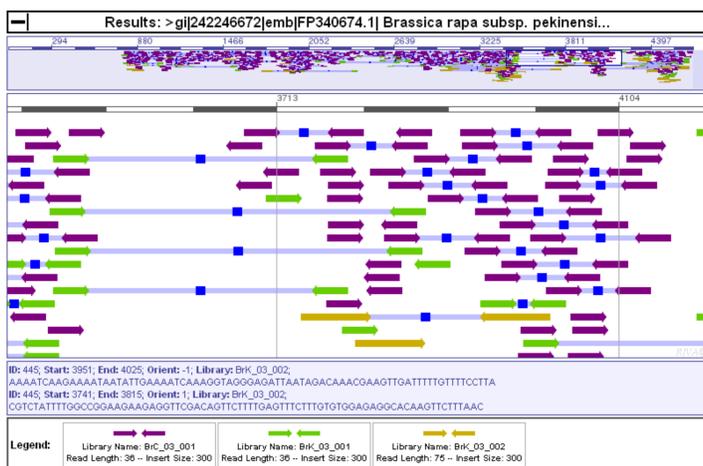
## Gene and promoter discovery

Illumina paired read sequence data allows for the extension of gene sequence, as well as the discovery of gene promoters. By aligning paired reads to a known sequence, using TAGdb (<http://flora.acpfg.com.au/tagdb/>), researchers can identify short sequences that lie upstream and downstream of the query sequence (Figure 1). These sequence tags can be used to design PCR primers to amplify and sequence the intervening region.



**Figure 1.** The application of paired-end read sequences for gene extension and discovery.

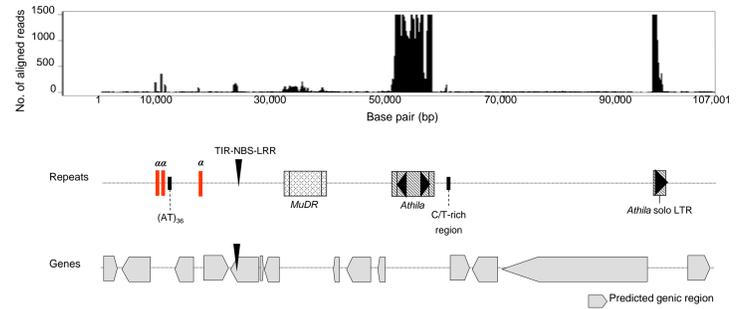
TAGdb presents reads matching the query sequence in two frames (Figure 2) and lists the sequence tags as a table. Directional reads are colour coded by library. Where both ends of a pair match the query, this is joined by a coloured bar.



**Figure 2.** A screenshot of the TAGdb web application, showing the mapping of paired-read data to a *Brassica rapa* BAC sequence.

## Repetitive element characterisation

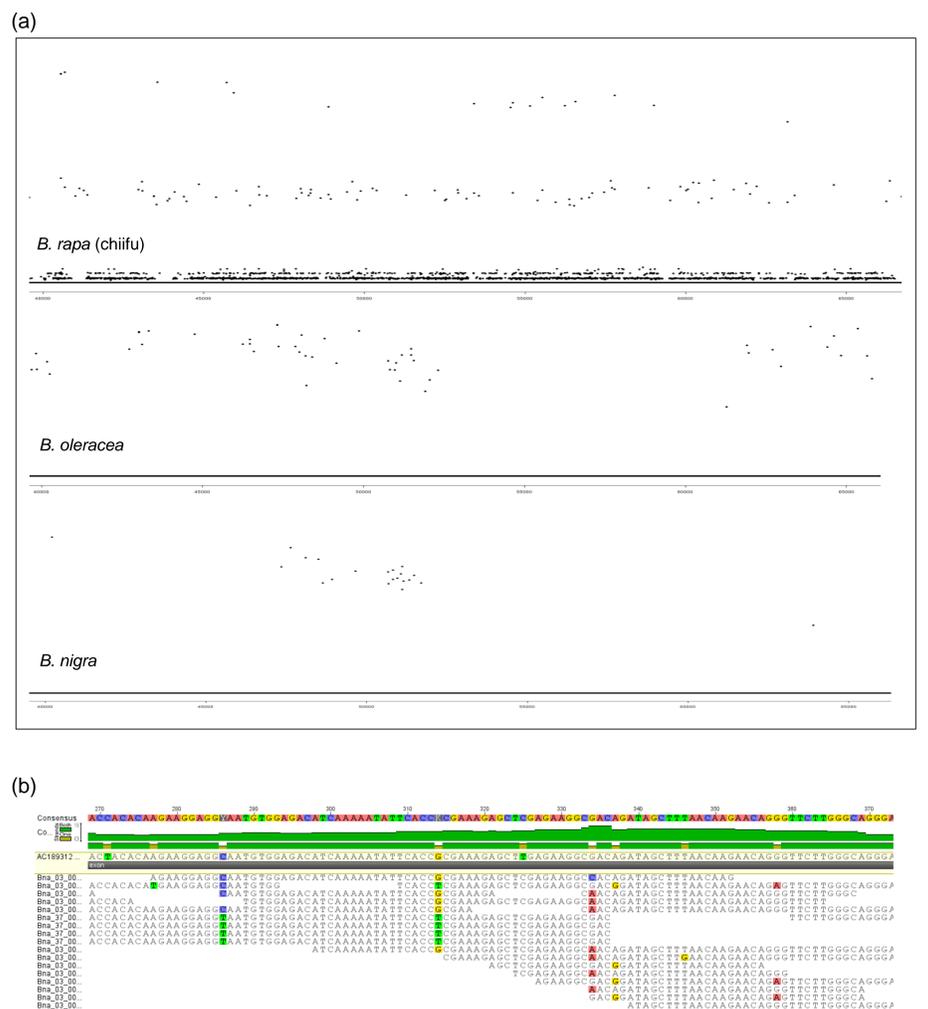
Short read data can be used to identify and characterise repetitive elements in the genome. Sequences that are over-represented in the genome also have higher representation in the sequence data (Figure 3). By comparing the sequence data abundance with annotated repetitive elements, it is possible to identify novel repetitive elements in the genome



**Figure 3.** Regions of high coverage of short reads and their corresponding annotation in a *B. rapa* genomic sequence.

## Comparative analysis of divergent *Brassica* genomes

Short read data can be used to identify regions of conservation and divergence across closely related genomes. Comparing a reference sequence with Illumina sequence reads from a related species demonstrates regions of conservation. Figure 4a shows the mapping of short read sequence data from *B. rapa*, *B. oleracea* and *B. nigra* to a *B. rapa* genomic sequence. There is a high level of coverage with the *B. rapa* data, and increasingly sparse but clustered coverage with data from the more distantly related *B. oleracea* and *B. nigra* genomes. The identification of single nucleotide polymorphisms (SNPs) is also possible. The assembly of redundant overlapping reads enables the *in silico* detection of SNPs (Figure 4b).



**Figure 4.** (a) The horizontal axis represents a *B. rapa* genomic sequence, with short read data from *B. rapa*, *B. oleracea* and *B. nigra* mapped as points central to the mapped read pair. The vertical axis represents the distance between the read pair. (b) A screenshot from the Geneious program, showing polymorphisms between sequences aligned to a reference.

## Conclusion

High throughput, Illumina genome sequence data can be used for a myriad of applications in addition to genome assembly. We have demonstrated application of short read data in the fields of gene discovery, repeat characterisation, genome annotation, marker discovery and comparative genomics. TAGdb, is an online short-read mapping application, available at <http://flora.acpfg.com.au/tagdb/>