The Multinational *Brassica* Genome Project.

David Edwards, Jacqueline Batley, German C Spangenberg, Benjamin R Burgess, Tim Erwin, Melanie Hand, Clare Hopkins, Andrew Keniry, Xi Li, Erica Logan, Christopher G Love, Hayley Mountford and Marzena Walkiewicz

Plant Biotechnology Centre, Primary Research Industries Research Victoria, Department of Primary Industries, Bundoora, 3084, Australia

Abstract

The Multinational *Brassica* Genome Project was initiated in January 2003 with the aim of sharing and coordinating *Brassica* genome resources for the benefit of the international *Brassica* research community. One of the primary goals identified, by the project steering committee, was the sequencing of the complete *Brassica* A Genome. The first milestone of this project is now complete, with over 100 000 cloned genomic fragments characterised, enabling the rapid identification of any gene on the A-genome. Phase 2 of the project, to obtain the complete genome sequence, is now under way. Complementary work to shotgun sequence the *Brassica* C genome is also progressing, with sequencing to 0.5 fold coverage completed and a further 4 fold coverage planned. The availability of this sequence information enables the development of gene expression resources that are being applied to understand the genes underlying important agronomic traits. The *Brassica* sequence information has also enabled the identification and characterisation of large numbers of molecular genetic markers and the construction of high-resolution genetic maps. Trait associated genetic markers are being identified for applications in *Brassica* molecular breeding.

Why sequence the *Brassica* genome?

Just as the human genome sequence is aiding our understanding of human development and the diseases which plague mankind, the genome sequence of crops provides an insight into complex heritable traits such as yield, quality and resistance to biotic and abiotic stress. Several plant genome-sequencing efforts are under way around the world and the genomes of rice and the model plant *Arabidopsis thaliana* have already been completed (International Rice Genome Sequencing Project 2005, The *Arabidopsis* Genome Initiative 2000). Through analysis of these genomes, we now have a greater understanding of the complexities of plant evolution that have shaped the crops that we grow today. By sequencing further genomes, we may more readily translate our understanding of model plants for the improvement of crop germplasm.

Brassica genome sequencing update

There are several methods than can be applied for the sequencing of a genome. These can be broken down into two main approaches; The first is called whole genome shotgun sequencing. In this approach, the complete genome, which for each of the *Brassica* genomes amounts to 500-800 million base pairs of DNA, is fragmented into relatively short pieces, each between 5 and 20 thousand base pairs in length. Approximately 1000 base pairs of sequence information is obtained from the end of each fragment by Sanger sequencing, and computational tools are applied to assemble these fragments of sequence into the whole genome sequence. The alternative approach involves fragmenting the genome into much larger pieces, usually around 120 thousand base pairs in length. These fragments are maintained in bacterial hosts as artificial chromosomes (BACs). Analysis of these BACs through genetic mapping, end sequencing and DNA fingerprinting, enables the assembly of a minimal tiling path of BACs representing each of the chromosomes. The complete sequence of each of these BACs can then be obtained using a shotgun approach by

fragmenting them into 2000-5000 base pair pieces and reading approximately 1000 base pairs of sequence from each end of these fragments.

The BAC by BAC approach has several advantages over the whole genome shotgun method. Working with much larger fragments facilitates genome assembly. This is of particular value when the genome contains a large quantity of repetitive DNA. The finished genome sequence is generally of greater quality with fewer gaps or misassembled sections than genomes sequenced by whole genome shotgun sequencing, and the sequence produced can be used as a reference sequence for comparison with other genome sequences. These advantages however, come at a price and there is significant cost in terms of labour, time and consumables, so where a closely related reference genome sequence is available, a whole genome shotgun approach is usually preferred. The whole genome shotgun approach is sometimes considered to be a 'cheap and dirty' method for obtaining genome sequence. It has the advantage that the cost of production is a fraction of the BAC by BAC approach. However, finished assembly of the whole genome is rarely completed without comparison with a closely related reference genome sequence.

The initial *Brassica* genome sequencing project founded in 2003 aims to produce, from BAC clones, "Phase 2" sequence (i.e. fully oriented and ordered sequence but some small sequence gaps and low quality sequences) for the ca. 500 Mb genome of *Brassica rapa* subspecies pekinensis (A-genome). The genome sequence is to be anchored to a reference genetic map by ca. 1000 molecular markers. The first stage of this project is now completed and over 110 000 BACs have been characterised by end sequencing. Seed BACs distributed across the genome are now being sequences with priority given to BACs that are predicted to contain genes for valuable traits. The completion of this project will provide a reference sequence for *Brassica* genomics for comparison with other related genomes and the study of allele diversity in association with heritable traits. A complementary project is also underway to obtain the genome sequence of *Brassica oleracea* (C-genome). To date, 0.5 fold coverage has been produced in a collaboration between The Institute for Genomic Research (TIGR) and the Cold Spring Harbour Laboratory, using a whole genome shotgun approach (Ayele *et al*. 2005). Completion of at least of 4-fold coverage of this genome by an international consortium is expected to coincide with the completion of the A-genome sequencing project. A whole genome shotgun approach was chosen for the C-genome project as this genome could be reliably assembled by comparison with the A-genome reference sequence. Together, these projects should reveal the complete genome for the economically important B. napus (AC-genome) amphidiploid species.

Identification of genes underlying heritable traits

Genome sequencing identifies the DNA sequence for every gene in an organism. However, only a small number of these genes are likely to be responsible for agronomic traits and the function of a large proportion of the identified genes will not be determined through analysis of the DNA sequence alone. Furthermore, allelic variation of genes is responsible for the variation in the heritable phenotype, and genome sequencing will generally only provide one allelic form of each gene. Conversion of the information present in the genome sequences to an understanding of the biology of the organism is a major undertaking in science today. There are essentially three main routes to associate genes with traits.

1, Gene expression information

The expression of genes is carefully regulated at many levels in the cell, from the transcription of the message from the DNA template, through to the activity of the resulting translated protein. The principle form of gene regulation is the transcription and degradation of messenger RNA. Messenger RNA is the template for protein synthesis and is thus required for the expression of a gene. The quantitative abundance of mRNA is usually reflected in the observed phenotype. For

example, genes responsible for the production of the photosynthetic apparatus are highly expressed in leaves, and specific disease resistance genes are only expressed in resistant plant varieties. Thus, measuring the presence and quantity of mRNAs for each gene provides an insight into the potential role of the gene. Microarrays are the most common tool applied to measure the expression of large numbers of genes in parallel. Recent experiments at the Plant Biotechnology Centre have addressed the question of which genes are differentially expressed in partially resistant or susceptible canola on infection with *Leptospheria maculans*. This has identified many *Brassica* genes demonstrating an expression pattern that is associated with defence response to blackleg.

2, Comparison with known gene sequences

Many genes are highly conserved in both sequence and function between organisms. The genes that regulate the core functions of the cell are conserved between plants, fungi and humans. Genes that play more specialised roles eg. certain biochemical pathways, may only be present in a small number of related organisms. The conservation of gene sequence thus enables the prediction of gene function through comparison with related genes. For closely related plants, there is not only conservation in gene sequence, but also gene order along the chromosomes. The genomes of *Brassica* species share significant synteny with the completely sequenced model plant *Arabidopsis* (Parkin *et al.* 2005). Thus, the location of *Brassica* genes may be predicted by comparison with the *Arabidopsis* genome (Love *et al.* 2005). While *Arabidopsis* itself possesses few traits of value for *Brassica* breeders, it may still be used as a valuable model for the identification of *Brassica* genes and therefore alleles that have agronomic value.

3, Genetic association studies

Recent years have seen the expansion of the application of molecular genetic markers for the mapping of agronomic traits. Genetic markers that associate with beneficial alleles may be used to select for these alleles during breeding, without the requirement for extensive field selections. This method is of particular value where the reproducibility of the trait is not reliable each year. A comparison of traits mapped on genetic maps with physical genomic sequences enables the identification of the candidate genes underlying the trait. While detailed fine mapping using large populations is required to focus the genomic region of interest to a few tens of genes, standard methods of trait mapping still permit the localisation of traits to genomic regions representing a few hundred genes.

The three methods described above may each be applied individually. However, the resolution of gene identification is greatest when the methods are applied in unison. For example, genetic mapping of a trait may identify genomic regions represented by 500 genes. Through screening of the annotation of these genes, derived from comparison with previously characterised genes, the number of potential candidate genes may reduce to 100, with perhaps 5-10 strong candidates. Gene expression information may further refine the query. Observation of continued gene segregation with the trait or genetic transformation would then validate candidate genes. Within the *Brassica* Molecular Marker Program, we have applied all three methods for candidate gene identification and are currently in the process of sequencing seed BACs containing genes of potential agronomic importance.

The application of the *Brassica* genome sequence for crop improvement

Genes underlying agronomic traits may be applied to crop improvement by converting them to perfect markers for marker assisted selection. However, greater benefits may be gained through understanding the sequence and potential function of the gene. The sequencing of alternative alleles of the gene across diverse germplasm is likely to yield new allele sequences which could be assessed for trait association, with the most beneficial allele bred into the crop using marker assisted selection. The identification of the gene sequences underlying traits, permits the identification of

closely related genes that may offer improved agronomic characteristics. Furthermore, if the gene is known to be part of a characterised biological pathway, then genes in other parts of the pathways may become targets for germplasm enhancement. The regulation of oil quality is one area for potential improvement using this approach. Many of the genes regulating oil quality have been characterised in other plants. The sequencing of the *Brassica* genome will identify orthologous genes in *Brassica* that could be developed as targets for germplasm enhancement.

References

Ayele M, Haas BJ, Kumar N, Wu H, Xiao Y, Van Aken S, Utterback TR, Wortman JR, White JR and Town CD (2005) Whole genome shotgun sequencing of *Brassica oleracea* and its application to gene discovery and annotation in *Arabidopsis*. Genome Research 15:487-495

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436: 793-800

Love C, Logan E, Erwin T, Spangenberg G and Edwards D (2005) Analysis of the *Brassica* A and C Genomes and Comparison with the Genome of *Arabidopsis thaliana.* Acta Horticulturae (in press)

Parkin IAP, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC and Lydiate DJ (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics (in press)

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408 (6814): 796-815