BRASSICA SEED IMPROVEMENT USING GENOMICS TOOLS

Faouzi BEKKAOUI[1] and Wilf KELLER[2]

[1]Genome Prairie, NRC-PBI, 110 Gymnasium Place, Saskatoon, SK, S7N 0W9
Email: Faouzi.Bekkaoui@nrc-cnrc.gc.ca

[2]NRC, Plant Biotechnology Institute, 110 Gymnasium Place, Saskatoon, SK, S7N 0W9
Email: Wilf.Keller@nrc-cnrc.gc.ca

Genomics is revolutionising the speed with which biological information can be accumulated thanks to the use of high throughput molecular and bioinformatics technologies. The complete DNA sequencing of the *Arabidopsis* genome is helping plant biologists to unravel the remarkable "machinery" of how plants grow and develop. This article describes a genomic research project focusing on the study of seed development and composition of Brassica species. Because of the close phylogenetic relationship between *Arabidopsis* and Brassica species, scientists are taking advantage of the DNA sequence of *Arabidopsis* to improve our knowledge of Brassica seed development and composition. In addition, others genomics tools described briefly in this article are also employed within the canola genomic project. It is expected that the knowledge acquired in this project will facilitate the development of new canola varieties with improved agronomical and nutritional traits.

Key words: Brassica, canola, genomic, seed, development, metabolism, composition, gene, bioinformatic, ESTs, microarrays

Introduction

Conventional breeding methods have led to significant improvement in Brassica seed composition (Röbbelen, Downey, & Ashri 1989). In particular, the reduction of the glucosinolates and erucic acid levels in seeds has allowed canola to become a major world oilseed crop.Many agronomical and nutritional traits are controlled by more than one gene, most of which are still to be characterized (Borevitz & Nordborg 2003;Chaudhury et al. 2001). It is estimated that the normal seed development in *Arabidopsis* is controlled by approximately 750 genes (McElver et al. 2001). Various genomic approaches are used to assign function of the unknown genes and would ultimately lead to the understanding of the genetic control of complex traits.

With the contribution of Genome Canada funding, the National Research Council of Canada (NRC) and Agriculture and Agri-Food Canada (AAFC) have started a genomic research project to study seed development and composition in Brassica oilseeds crops. This three-year project involves 40 full time scientists in addition to the senior scientists. The goal of the project is to develop and employ genomic tools in the studies of seed development and composition in Brassica oilseeds crops. The long-term aim of this project is to develop crop varieties with improved quality that will provide access to new value-added and niche markets.

Canola seed development

In canola, the seed accounts for about 15 to 35 per cent of the total dry matter produced. Crop value is determined by key characteristics such as the size and number of seeds, the rate of seed growth, the chemical composition of the seeds (the identity and amounts of oils, carbohydrates etc.) and the efficiency and speed of germination of crops after sowing. These characteristics are a consequence of developmental processes. However, there are substantial gaps in our understanding of the many interdependent biochemical and molecular processes that control development in plants, including crop species such as canola.

The seed phase of the plant life cycle commences with fertilization, followed by embryogenesis, maturation, dehydration, imbibition and terminates with germination (Chaudhury et al. 2001;Goldberg, de Paiva, & Yadegari 1994). Despite significant advances in recent years (especially from *Arabidopsis* embryogenesis research (TAIR 2003), there are many knowledge gaps remaining in our understanding of embryogenesis and seed development (Bove, Jullien, & Grappin 2002;Lohe & Chaudhury 2002). For example, little is known of the pattern elaboration processes of the mid to late stages of embryo development that contribute to determining the embryo/seed size.

The detailed study in our project will involve the correlation of transcript, protein and metabolite profiles with an emphasis on understanding the regulation of seed development. Initially a baseline dataset will be collected for seed development in canola. This will be used to make informative comparisons with development in other Brassica species.

Canola seed composition

Brassica seeds synthesize three major storage reserves, proteins, complex carbohydrates and triacylglycerols (oil). The storage reserves are deposited during seed development at the maturation stage of seed development. In oilseed crops, such as canola, the oil is the most valuable component of the harvested seed and a relative increase in oil content is highly desirable. In high yielding oil crops, modifications to lipid biosynthesis that create oil profiles ideal for particular food or industrial uses would add additional value to those crops. Although Canola has very low level of saturated fats compared to other vegetable oils, it is desirable to reduce the saturates even lower from 7% currently, ideally to 3%, and raise mono-unsaturates to levels found in olive oil and slightly reduce the poly-unsaturates, bringing them from 8% to approximately 4%. This change would increase shelf life of the oil. A better understanding of the lipid biosynthetic pathways leading to the fatty-acids would be most useful in further improving the oil composition of canola through either conventional or genetic engineering methods.

Canola meal is not currently as digestible as soybean meal. Canola protein could be improved through the application of genomic technologies and could eventually enter into the human food market. Major crop plants also produce an extensive array of secondary metabolites. Some of these compounds are of potential value but accumulate at sub-optimal levels because of either low metabolic rates for relevant pathways, feedback inhibition, or shunting reactions that divert metabolism to the production of less desirable derivatives. Examples of such compounds include antioxidants and vitamins such as tocopherols (vitamin E), ubiquinone and carotenoids (vitamin A). Other crop secondary metabolites are undesirable anti-feedants. For example, phytate is a seed phosphate storage compound that renders phosphate inaccessible to animal absorption, necessitating phosphate supplements in animal feeds and causing environmental phosphate pollution by animal waste. Another antinutritional substance, sinapine, has a bitter flavour that causes poor palatability by livestock and fish.

Genomics methodologies and objectives of the project

Functional genomics, the process that assigns a biological function to a particular gene sequence, has entered the high-throughput stage with the completion of the DNA sequencing of two model plants *Arabidopsis* (The Arabidopsis Genome Initiative 2000) and rice (Goff et al. 2002). Different methodologies have been developed within the functional genomics technological platform such as transcriptomics (ESTs and DNA microarrays), gene knockout by RNA interference or T-DNA insertion, proteomics and metabolomics. It is a combination of these methodologies that will infer gene function rather than the usage of a single technology (Ge, Walhout, & Vidal 2003;Holtorf, Guitton, & Reski 2002). These methodologies are briefly described here with some of the associated objectives of our project.

Expressed sequence tags (ESTs) are a partial sequence of a complementary DNA (cDNA) clone. Complete cDNA or gene sequence is not necessary to assign the function of a cDNA. ESTs are prepared from specific developmental stages or types of tissues. The ESTs generated from a cDNA library will reflect the overall biochemical and physiological stage of the specific tissue or state. This project is conducting a global survey of gene expression, and its consequences for seed development, from early embryo development to the initiation of germination in canola and related species. At least 25,000 ESTs will be generated within the project. These ESTs are then compared to known sequences of *Arabidopsis* using bioinformatic tools to assign putative function.

DNA Microarray or Gene-Chip technology is a new tool for measuring the levels of activity for thousands of genes (Buckhout & Thimm 2003;Stears, Martinsky, & Schena 2003). It is a technology that allows generation of data with a higher throughput and a greater precision compared to traditional filter blotting techniques. *Arabidopsis* microarrays with the 27,000 DNA sequences representing the whole genome are available commercially. The comprehensive coverage of whole-genome microarrays offers an effective way to assign putative functions to genes on the basis of their expression pattern. *Arabidopsis* DNA microarrays have bee used recently to annotate 30% of all expressed genes (Yamada et al. 2003). *Arabidopsis* microarrays are used in our project to study seed development. Brassica microarrays with a subset of expressed genes will be prepared and also used to identify genes that are involved with specific processes of seed development. We will be also developing a targeted DNA microarray to study genes involved in plant hormone regulation and microspore derived embryos.

Genome sequencing and microarray experiments generate a significant amount of data that is used to determine the structure and function of genes. These data are analyzed using bioinformatics approaches, which include the development and application of computer programs and the development of databases (Baxevanis & Ouellette 2001;Ouzounis & Valencia 2003). With the assistance of bioinformaticians, the project uses the necessary infrastructure including high-performance computers and specialized algorithms to conduct the bioinformatic work. For example sequence alignment of the ESTs generated in this project is conducted using an-in-house computer cluster to assign putative function of a given EST.

Gene knockouts lines are important because they provide a direct route to determining the function of a gene product. Most other approaches to gene function are correlative and do not necessarily prove a causal relationship between gene sequence and function. *Arabidopsis* knockout lines have been generated using a genome-wide insertional mutagenesis method (Alonso et al. 2003). These lines have been transformed with Agrobacterium containing a T-DNA construct that will interrupt randomly a gene. These lines are available publicly to researchers and can be searched on the internet (TAIR 2003). To prove definitively that the insertional mutation causes the phenotype, one must complement the mutation by introducing a wild-type copy of the gene into mutant plants by using transgenic technology. Another new method for silencing genes is RNA-interference (RNAi) (Elbashir et al. 2001;Fire et al. 1998).Blocking the mRNA of a gene would give an indication of what that gene does. This technique is achieved by introducing double stranded RNA (dsRNA) matching the gene sequence into cells. The introduction of dsRNA triggers specific RNA degradation. RNAi can be delivered by transient methods (such as particle bombardment or *Agrobacterium* infiltration) or by stable ones (such as the introduction of amplicon transgenes). This process facilitates targeted post-transcriptional gene silencing and has recently been used to study the function of several genes (Arenz & Schepers 2003). RNAi and knock out mutants are used in our project in particular for assign function of genes involved in lipid biosynthesis and seed storage proteins.

ESTs and DNA microarrays provide information on the gene function however they do not give any information on the protein complement of the genome, because of post-translational modifications. Therefore, the necessity to study the full complement of proteins produced by a particular genome (proteomics). The traditional and most widely used technique for proteomics is two-dimensional electrophoresis (2-DE). 2-DE has the ability to separate simultaneously thousands of proteins to homogeneity, enabling subsequent characterization. Other techniques used in proteomics include tandem mass spectrometry (MS/MS; liquid chromatography (LC)-MS/MS; matrix assisted laser desorption/ionization-time of flight-MS (MALDI-TOF-MS), and the use of protein microarrays. We are using these tools to study proteins involved in embryo development.

Metabolomics or the large-scale phytochemical analysis of plants is another field derived from the functional genomics filed (Sumner, Mendes, & Dixon 2003). Gas and liquid chromatography coupled to mass spectrometry are used to process a high sample numbers for the identification and quantitation of small-molecular-weight metabolites. We are also developing analytical methodology suitable for simultaneous quantitation of all classes of hormones and thecompounds in their biosynthetic pathways using various analytical tools such as HPLC-MS/MS (Chiwocha et al. 2003;Zhou et al. 2003) As with ESTs and microarrays, bioinformatics plays a critical role for the data analysis of proteomics and metabolomics experiments.

Conclusion

We expect that the results of this project will help in the understanding of complex traits, which will ultimately lead to novel varieties of canola with unique traits such as varieties with large seeds or with specific oil composition. Another example would be canola seeds that will mature more rapidly and will likely be more tolerant to frost damage. Seeds with a reduced seed coat thickness that will be more amenable to processing for food or feed would also be desirable.

With respect to seed composition research, new varieties that will accumulate novel or improved proteins or lipids or carbohydrate at high levels may be generated. Genetic markers will be developed to select new varieties of canola with high protein. Varieties of canola with a reduced level of anti-nutritional compounds that can make canola an ideal feed for animals could be developed. Some of the potential products could be ideally suited to high-value niche markets. While transgenic modifications will be essential to the realisation of some of these improvements, other improvements will be achieved through the identification of rare natural variation and the accelerated introduction of this variation into crop varieties using classical breeding techniques.

Acknowledgement.
This project is a joint collaboration of several scientists from AAFC and NRC. The complete list of principal investigators and support scientists can be found at http://www.genomeprairie.ca/canola/. We would like to thank Genome Canada and Genome Prairie for their financial contribution towards the "Enhancing Canola through Genomic" project.

References

Alonso J.M. et al. (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* **301**, 653-657.

Arenz C. & Schepers U. (2003) RNA interference: from an ancient mechanism to a state of the art therapeutic application? *Naturwissenschaften* **90**, 345-359.

Baxevanis,A.D. & Ouellette,B.F. (2001) Bioinformatics: a Practical Guide to the Analysis of Genes and Proteins.

Borevitz J.O. & Nordborg M. (2003) The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol* **132**, 718-725.

Bove J., Jullien M., & Grappin P. (2002) Functional genomics in the study of seed germination. *Genome Biol.* **3**, REVIEWS1002.

Buckhout T.J. & Thimm O. (2003) Insights into metabolism obtained from microarray analysis. *Curr.Opin.Plant Biol.* **6**, 288-296.

Chaudhury A.M. et al. (2001) Control of early seed development. *Annu.Rev.Cell Dev.Biol.* **17**, 677-699.

Chiwocha S.D. et al. (2003) A method for profiling classes of plant hormones and their metabolites using liquid chromatography-electrospray ionization tandem mass spectrometry: an analysis of hormone regulation of thermodormancy of lettuce (Lactuca sativa L.) seeds. *Plant J.* **35**, 405-417.

Elbashir S.M. et al. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494-498.

Fire A. et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811.

Ge H., Walhout A.J., & Vidal M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**, 551-560.

Goff S.A. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**, 92-100.

Goldberg R.B., de Paiva G., & Yadegari R. (1994) Plant Embryogenesis: zygote to seed. *Science* **266**, 605-614.

Holtorf H., Guitton M.C., & Reski R. (2002) Plant functional genomics. *Naturwissenschaften* **89**, 235-249.

Lohe A.R. & Chaudhury A. (2002) Genetic and epigenetic processes in seed development. *Curr.Opin.Plant Biol.* **5**, 19-25.

McElver J. et al. (2001) Insertional mutagenesis of genes required for seed development in Arabidopsis thaliana. *Genetics* **159**, 1751-1763.

Ouzounis C.A. & Valencia A. (2003) Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics.* **19**, 2176-2190.

Röbbelen,R., Downey,K., & Ashri,A. (1989) Oil crops of the world : their breeding and utilization, McGraw-Hill, New Yok.

Stears R.L., Martinsky T., & Schena M. (2003) Trends in microarray analysis. *Nat.Med.* **9**, 140-145.

Sumner L.W., Mendes P., & Dixon R.A. (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817-836.

TAIR. The Arabidopsis Information Resource. http://www.arabidopsis.org/

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the fowering plant Arabidopsis thaliana. *Nature* **408**, 769-815.

Yamada K. et al. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science* **302**, 842-846.

Zhou R. et al. (2003) Rapid extraction of abscisic acid and its metabolites for liquid chromatography-tandem mass spectrometry. *J.Chromatogr.A* **1010**, 75-85.