Sequence analysis of the canola genome

Jacqueline Batley[1], Michal Lorenc[2], Kaitao Lai[2], Sahana Manoli[2], Jiri Stiller[2], Paul Berkman[2], Adam Skarshewski[2], Lars Smits[2], Megan McKenzie[1], Emma Campbell[1], Michael Imelfort[2], Harsh Raman[3], Bart Lambert[4], Benjamin Laga[4] and David Edwards[2]

[1]School of Land, Crop and Food Sciences and ARC Centre of Excellence for Integrative Legume Research, University of Queensland, Brisbane, 4072, Australia.
[2]School of Land, Crop and Food Sciences and Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, 4072, Australia
[3]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
[4]Bayer Bioscience NV, Technologiepark 38 - 9052 Zwijnaarde, Gent, Belgium.

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, providing an unprecedented insight into genome structure and variation to support Brassica crop improvement.

INTRODUCTION

The application of second generation DNA sequencing technology is rapidly changing Brassica genome research. DNA sequencing technology has changed dramatically in recent years, revolutionising both human and plant genomics. Second generation DNA sequencing technologies can produce more than 200 billion nucleotides of sequence data in a single run (Imelfort and Edwards, 2009) and data production continues to increase rapidly. The completed genome sequences of Brassicas and Arabidopsis provides the opportunity to conduct re-sequencing and comparative genomic analysis of individuals and assist in the identification and characterisation of sequence variants. This crop genome sequencing data can be applied for genome analysis leading to crop improvement (Edwards and Batley, 2009).

The use of modern molecular genetics tools has increased the speed of breeding new varieties and permits the application of newly available genome sequence information for crop improvement. In crop species, genetic variation analyses predominantly focus on single nucleotide polymorphisms (SNPs) for marker-trait association. An appreciation of how this variation affects phenotypic variation in plants is now possible through the improvements in available technologies. Characterisation of SNP density in plants can assist in the understanding of recent selection pressures on plant genomes, the genomic components that contribute to adaptation and the identification of genes that have been the target of selection.

The 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. We aim to analyse the Brassica genomes to identify genes, novel and mapped genetic markers and develop methods for the association of agronomic traits with underlying genomic variation.

MATERIALS AND METHODS

Resequencing Brassica Genomes

Illumina GAIIx and Hi-Seq paired end and mate paired sequence data has been generated for eight B. napus varieties and compared to reference Brassica genomes using custom bioinformatics pipelines. SOAP (Li et al., 2008) is highly efficient for mapping B. napus paired reads against a B. napus reference genome.

SNP Identification

The SNP discovery is performed in a stepwise manner, using the custom developed SGSautoSNP. SNPs are predicted from the aligned data and SNP confidence calculated based on a combination of redundancy, coverage and distribution of base calls between samples. Additional genomic variation such as indels, translocations and inversions can be predicted from the alignment of read pairs to the genome and the identification of paired read mapping variation. The total map of genetic variation is maintained in a custom database and viewed using standard genome feature format (GFF) in compatible genome viewers such as GBrowse (Arnaoudova et al., 2009) or Biomatters Geneious (Drummond et al., 2009).

Assessing SNP Density across the Genomes

The resulting identified genomic variation has been integrated with annotated genome features and previously mapped genetic markers to link agronomic traits, associated markers and candidate gene information on a genome wide scale. A custom pipeline has been developed for the assessment of SNP density across the genome, including specific regions associated with expressed genes.

RESULTS AND DISCUSSION

SNP Identification

More than 100,000 SNPs have been identified across eight varieties of B. napus. The number of predicted SNPs was distributed evenly across the 19 chromosomes, with variation as expected according to chromosome length. The SNP base changes were recorded. As the directionality of the change cannot be inferred from the data, polymorphisms were grouped alphabetically, ie. A>G and G>A are grouped as A>G. A greater number of transitions (A>G or C>T) than transversions (A>C, A>T, C>G or G>T) were identified. This is in accordance with previous computational and laboratory based SNP discovery studies (Deutsch et al., 2001, Duran et al., 2009) and reflects the high frequency of C to T mutation following methylation (Coulondre et al., 1978). SNP density varied across the chromosomes suggesting evidence of selection.

CONCLUSIONS

The sequencing and re-sequencing of canola varieties has identified genome wide variation represented by more than 100,000 high confidence single nucleotide polymorphisms. Bioinformatics tools have been produced and applied to interrogate and annotate this abundant data, and genome wide variation has been integrated with genetic maps and phenotypic information. The Brassica genomes provide an insight into the evolution of these important crop plants and tools to advance breeding of improved varieties with enhanced agronomic traits.

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.

REFERENCES

Arnaoudova E G, Bowens PJ et al. (2009). Visualizing and sharing results in bioinformatics projects: GBrowse and GenBank exports. BMC Bioinformatics **10**.

Coulondre C, Miller JH, Farabaugh PJ and Gilbert, W. (1978) Molecular basis of base substitution hot spots in Escherichia coli. Nature, **274**: 775–780

Deutsch S, Iseli C, Bucher P, Antonarakis SE and Scott HS. (2001) A cSNP map and database for human chromosome 21. Genome Res. **11**: 300–307

Duran C, Appleby N, Vardy M, Imelfort M, Edwards D and **Batley J**. 2009. Single Nucleotide Polymorphism Discovery in Barley using AutoSNPdb. Plant Biotechnology Journal **7**: 326-333

Drummond A J, Ashton B et al. (2009). Geneious v4.6., from http://www.geneious.com/

Edwards D and Batley J. (2009) Plant genome sequencing: applications for crop improvement. Plant Biotech. J. **7**:1–8.

Imelfort M and Edwards D. (2009) Next generation sequencing of plant genomes. Briefings in Bioinformatics **10**: 609–618

Li R, Li Y, et al. (2008) SOAP: short oligonucleotide alignment program. Bioinformatics **24**:713–714