J Plant Biotechnol (2012) 39:49-61
DOI:http://dx.doi.org/10.5010/JPB.2012.39.1.049

Review

# SNP discovery and applications in *Brassica napus*

**Alice Hayward · Annaliese S. Mason · Jessica Dalton-Morgan · Manuel Zander · David Edwards · Jacqueline Batley**

Received: 5 March 2012 / Accepted: 15 March 2012


**Abstract** This review summarises the biology, discovery and applications of single nucleotide polymorphisms in complex polyploid crop genomes, with a focus on the important oilseed crop *Brassica napus*. *Brassica napus* is an allotetraploid species, and along with soybean and oil palm is one of the top three most important oilseed crops globally. Current efforts are well underway to *de novo* assemble the *B. napus* genome, following the release of the related *B. rapa* 'A' genome last year. The next generation of genome sequencing, SNP discovery and analysis pipelines, and the associated challenges for this work in *B. napus*, will be addressed. The biological applications of SNP technology for both evolutionary and molecular geneticists as well as plant breeders and industry are far-reaching, and will be invaluable to our understanding and advancement of the *Brassica* crop species.

**Keywords** *Brassica napus,* canola, Single Nucleotide Polymorphism, genetic marker, QTL mapping, genotyping, Illumina®, next generation sequencing, Infinium, Goldengate

## Introduction to Brassicas

The Brassicaceae family comprises some of the most economically important food crops in the world. Of particular significance are the *Brassica* species, which include more agricultural and horticultural crops than any other plant genus (Taiyan et al. 2001). The six major cultivated *Brassica* species exist as either diploids (*B. rapa* (AA genome), *B. nigra* (BB genome) and *B. oleracea* (CC genome)) or allotetraploids, which contain two diploid sets of chromosomes (*B. napus* (AACC), *B. juncea* (AABB) and *B. carinata* (BBCC)). The genetic relationship between these species was first discovered by Morinaga (1934) and visualised by Dr Woo Jang-choon in 1935 (U 1935). These researchers concluded that the three cultivated allotetraploid species were derived from spontaneous interspecific hybridization between the three diploid progenitors.

Four of the 'U' triangle species are cultivated as oilseed crops. The oil extracted from the seed is widely used for both human consumption and industrial purposes. *Brassica napus* (canola/oilseed rape) is the most economically important of these *Brassica* oilseeds, and is responsible for the majority of the world's *Brassica*-derived vegetable oil supply, comprising 14% of total oilseed production globally (CCC 2010; Raymer 2002). Canola oil is used in the production of margarine and cooking oil due to its superior nutritional qualities, which include low saturated fat content and high levels of monounsaturated fat and omega-3 fatty acids. Rapeseed is also a source of oil extraction meal (Friedt and Snowdon 2009) which is used as animal feed due to its high level of protein. World production of rapeseed in 2009 was estimated at 61.6 million metric tons (UNFAO). By far the largest producers of rapeseed oil are China and Canada, with almost half of the world's rapeseed oil originating from these countries (UNFAO). *Brassica napus* varieties are also grown for the stem and leaves, which are commonly used in Eastern cuisine.

*Brassica juncea* has also been bred as an oilseed crop due to its ability to grow in semi-arid territories (Edwards et al. 2007a). The leaves and stem of *B. juncea* are commonly used in African and Asian cooking and the seeds are used to produce brown mustard. *Brassica rapa* is also grown as an oilseed, but is particularly popular in Asia as a cruciferous vegetable (e.g. chinese cabbage, bok choi, pak choi, turnips). *Brassica carinata* (Ethiopian mustard)

A. Hayward · A. S. · J. Dalton-Morgan · M. Zander · J. Batley (✉)
Centre for Integrative Legume Research and School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia
e-mail: J.Batley@uq.edu.au

D. Edwards
Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia

is grown as an oilseed crop and leafy vegetable in Ethiopia, while *B. nigra* (black mustard) is grown as a condiment mustard seed. The most morphologically diverse cruciferous *Brassica* species is *B. oleracea*, which encompasses cabbages, broccoli, cauliflower and brussels sprouts.

*Brassica napus* genome sequencing

The decision of the Multinational *Brassica* Genome Project (MBGP) to sequence *B. rapa* has led to the accumulation of extensive production of *B. rapa* sequence data. The MBGP was initiated in 2002 as a collaboration between *Brassica* researchers around the world (Edwards and Wang, 2012). The Steering Committee selected *B. rapa* as the first *Brassica* species to be fully sequenced due to its relatively small *Brassica* diploid genome (550 Mbp) and the availability of communal BAC libraries and mapping populations. The international project initially allocated chromosomes to groups in Australia, Canada, China, France, Germany, Korea, UK and USA. With the rapid development of second generation sequencing technology and the associated bioinformatics necessary for the assembly of short paired sequence reads (Imelfort et al. 2009a), a strategy based on Illumina® whole genome sequencing was adopted. The first *Brassica* A, C and AC genomes to be sequenced were proprietary and produced by Bayer CropScience, BGI Shenzhen, Keygene N.V and the University of Queensland, Australia (http://tinyurl.com/*Brassica*genome). A subsequent public A genome sequencing project was completed in 2011 (The *Brassica rapa* Genome Sequencing Project Consortium et al. 2011). Additional international projects are ongoing to produce reference genomes for *B. oleracea, B. napus, B. nigra* and *B. juncea*, and it is expected that these genomes will be complete in the next few years. Along with existing genomic data, the *B. rapa* reference sequence now provides a platform upon which diversity and genetic marker analysis can begin to unravel *Brassica* genome evolution and diversity with single-base resolution at the whole-genome scale (Edwards and Batley 2010).

Genetic markers

Over the past two decades, the development and implementation of molecular marker technology has given rise to a greater understanding of genetic diversity in *Brassica* crops, particularly *B. napus*. These markers have been a vital tool for mapping genes of agronomic importance with the goal of implementing marker-assisted breeding of elite crop cultivars (Duran et al. 2009a; Edwards and Batley 2008). The plummeting cost of genome sequencing combined with significant advances in molecular biology and the rapid rise of novel bioinformatics approaches have made marker technology an invaluable and easily accessible tool in genome analyses (Duran et al. 2009c).

The first marker technologies to play a role in *Brassica* mapping studies were protein and isozyme markers. Technological advances led to the development of restriction fragment length polymorphisms (RFLPs), followed by PCR-based markers such as randomly amplified polymorphic DNA (RAPD) and amplified fragment length polymorphisms (AFLP) markers, which contributed greatly to the first linkage maps in *Brassica* crop species (Snowdon and Friedt 2004). The next successful system was simple sequence repeat (SSR) marker technology, also known as microsatellite markers. SSR markers allowed simple, reliable and relatively inexpensive amplification of highly polymorphic microsatellite DNA repeat sequences (Batley et al. 2007b; Hopkins et al. 2007).

Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are currently one of the most popular markers for the fine mapping of heritable traits (Chagné et al., 2007). SNPs are single nucleotide differences between the DNA sequences of individuals in a population. There are three categories of SNPs: transversions (C/G, A/T, C/A and T/G), transitions (C/T or G/A) and insertions/deletions, also known as indels. Most SNPs at any given site are bi-allelic, although tri-allelic and tetra-allelic SNPs also exist.

The high heritability of SNPs makes them an excellent indicator of genetic diversity and phylogeny in crop species with ancient genome duplications, such as *B. napus*. Studies have shown that *B. napus* has a SNP every 600 base pairs of the genome (Edwards et al. 2007c; Fourmann et al. 2002). Given the ~ 1 Gb size of the *B. napus* genome, this would equate to ~ 1.7 million SNPs. This allows for the construction of high-density genetic maps which can provide a scaffold to map undesirable, as well as agronomically important, genetic traits (Duran et al., 2010a; Edwards and Batley 2004; Edwards et al. 2007d).

Genic vs genomic SNPs

The usefulness of SNPs for various applications depends on their genomic location and environment. Genic SNPs are those that have been identified within expressed sequences, most commonly from available EST databases

(Duran et al. 2009d; Erwin et al. 2007; Love et al. 2005; Love and Edwards 2007; Love et al. 2004) and, more recently, next generation transcriptome sequencing data (e.g. mRNAseq) (Batley et al. 2007a; Edwards 2007). These SNPs can be synonymous, resulting in the same amino-acid being translated, or non-synonymous, where a different amino-acid is incorporated into the resulting gene product. Non-synonymous SNPs within transcribed genes can be responsible for phenotypic change, whereby they alter protein structure or function to affect an organism's development or response to environment. Such SNPs that are linked directly to gene function and quantifiable phenotypic change are known as 'perfect' markers.

Due to their location, genic SNPs have an increased likelihood of being deleterious, or associated with a fitness cost for an organism. Genic SNPs are often selected against, which can be observed by the lower frequency of non-synonymous to synonymous base changes in gene regions. SNPs also tend to be more frequent outside of transcribed regions due to the increased 5-methylcytosine (5meC) abundance, resulting in a greater likelihood of C to T mutations over evolutionary time due to amplified cytosine deamination (reviewed in Edwards et al. (2007b)).

The accessibility and redundancy of transcriptome information relative to whole-genome NGS data means that genic SNPs have contributed the majority of genotyping work in *B. napus* to date (Durstewitz et al. 2010; Trick et al. 2009; Westermeier et al. 2009). The exclusive use of genic SNPs does, however, pose some problems. The evolutionary constraints on genic sequences relative to non-coding regions means that analysis of genic SNPs from EST and transcriptome data can lead to an underestimation of true SNP number, providing reduced resolution for genetic diversity studies (reviewed in Edwards et al (2007b)). Furthermore, the use of genic SNPs increases the number of false SNP predictions, especially in polyploid species, as ESTs or NGS transcriptome reads that originate from different homoeologous (between genome) and paralogous (within genome) loci are difficult to differentiate from intervarietal sequences. When using EST data this effect can be minimised by predominately using SNPs found within the 3'-end of ESTs, corresponding to more variable 3'UTR sequences (Bhattramakki and Rafalski 2001; Bhattramakki et al. 2002; Ching et al. 2002).

In *B. napus*, EST data derived from the phenotypically and genetically divergent cultivars Tapidor and Ningyou7 (Qiu et al. 2006) was used to generate 41 593 putative allelic polymorphisms over 15 626 unigenes (Trick et al. 2009). This study found that homoeologous transcripts from the A and C genomes may both be expressed, and differ by an average of only 3.5% (Trick et al. 2009). Moreover, the majority of SNPs (87.5–91.2%) detected in the NGS transcriptome data were 'hemi-SNPs' arising from transcription of homoeoloci. The existence of these duplicated loci, and even of highly conserved gene family members, can greatly compromise the applicability of genic SNPs to downstream applications such as association mapping and LD studies (Batley and Edwards 2007).

An additional consideration in the use of genic SNPs is the bias resulting from the choice of tissue(s) used to generate transcriptome libraries, as not all genes are uniformally spatiotemporally expressed. Moreover, by their nature genic SNPs are limited to actively transcribed or gene-rich regions of the genome. Although a number of data analysis algorithms attempt to normalise for transcript level differences (e.g. Trick et al. (2009)), such differences may prevent identification of SNPs associated with lowly expressed agronomically important genes.

With the recent advances in whole genome sequencing (WGS) technologies, the identification and use of genomic SNPs in *B. napus* is becoming more and more popular and accessible (Batley and Edwards 2009). Genomic SNPs, by nature, can be identified from any sequenced region in the genome, and, given that non-coding regions are less likely to be conserved within and between genomes, minimise problems of gene or genome duplication. Furthermore, the majority of genomic SNPs are evolutionarily neutral, whereby they are free of selective pressures such that their abundance in a population depends on random genetic drift, allowing a more complete estimate of diversity levels (Edwards et al. 2007b).

Nonetheless, genic SNPs have their advantages. Validated SNPs specific for different paralogues enable allocation and comparison of the specific genomic locations of multi-gene family members (Edwards et al. 2007d; Love et al. 2006b). Moreover, the relatively high sequence redundancy and genotype diversity represented within EST databases, combined with the ability to associate identified SNPs to an expressed gene sequence, means these databases remain a rich resource for identifying biologically useful SNPs (Duran et al. 2011; Edwards et al. 2009; Edwards et al. 2007d; Lorenc et al. 2012; Picoult-Newberg et al. 1999). SNPs that are linked to traits under selection are highly valuable for identifying genetic loci that contribute to phenotypic variation based on linkage disequilibrium (LD). In all, the ability to identify a large number of unique, genome-specific SNPs, including both randomly distributed and trait-linked SNPs under selective pressure, greatly

enhances the applicability of SNPs to biological outcomes (Duran et al. 2009b). With the availability of the *Brassica* A genome sequence (The *Brassica rapa* Genome Sequencing Project Consortium et al. 2011) and the *Brassica* B and C genomes currently being assembled, as well as the *B. napus* reference genome, it will soon be possible to rapidly validate unique SNPs within *B. napus*.

## SNP discovery pipelines

### Sequence-independent SNP discovery

Prior to the availability of large sequence databases, methods for SNP discovery were relatively low-throughput and did not require prior sequence knowledge. SNPs were detected using either 1) recognition of restriction enzyme site differences, 2) DNA conformational changes or 3) TILLING (Targeting Induced Local Lesions in Genomes). In 1), SNPs are identified as a single base change in a restriction enzyme recognition site, detected by analysis of fragment size and number after restriction enzyme treatment. A number of molecular markers utilise this method of SNP discovery to produce polymorphic alleles, including restriction fragment length polymorphisms (RFLPs), cleaved amplified polymorphic sequences (CAPS) and derived cleaved amplified polymorphic sequences (dCAPS).

In method 2), properties conferred by DNA sequence differences (e.g. SNPs) allow for discrimination by electrophoresis under certain experimental conditions. Methods using DNA conformational differences for SNP discovery include temperature gradient gel electrophoresis (TGGE) and denaturing gradient gel electrophoresis (DGGE), which rely on temperature and chemical gradients to denature DNA molecules. Single strand conformation polymorphisms (SSCPs) and heteroduplex analysis, which rely on the conformational properties of single stranded DNA and non-identical double-stranded DNA respectively, may also be used to identify SNPs using conformational differences.

The third means of SNP detection in early studies was TILLING, a reverse genetics approach allowing identification of mutations within a specific gene in either natural or mutagenised populations. In TILLING, sequence deviation from a reference sample is detected using a mismatch-specific enzyme. In mutagenised populations, the reference sample is usually the wild type, and in natural populations (EcoTILLING) the reference sample is chosen by the researcher. Base changes (SNPs) may be reliably detected using either method.

### The role of sequencing in SNP discovery

While SNP discovery methods that are not reliant on knowledge of the sequence can give an indication of SNP presence and genetic diversity, separate sequencing steps are required to determine the actual bases present. Early endeavours in genome sequencing often used the BAC by BAC approach, utilising bacterial artificial chromosomes (Mun et al. 2010). In this approach, a library of large genomic fragments (around 120 000 bases) is maintained within bacterial artificial chromosomes. Portions of these large genomic fragments where sequences overlap were used to construct chromosomal sequences. Individual SNPs in a particular region may also be detected using the sequencing of PCR amplicons, either of random amplicons or of amplicons where SNP variation has already been assayed by one of the aforementioned non-sequence-based methods. SNPs can be identified in SSR flanking regions (Mogg et al. 2002), allowing conversion of SSR to SNP markers and providing a valuable source of novel SNPs. Computational methods have been applied to discover large numbers of SNPs from expressed sequence databases (Barker et al. 2003; Batley et al. 2003a).

### Next-generation sequencing and SNP discovery

The advent of next generation sequencing (NGS) technologies, first commercialised in 2005, significantly decreased the cost of whole genome sequencing (Imelfort and Edwards 2009; Imelfort et al. 2009b). This technological advance also enabled the detection of SNPs in a high throughput manner. The initial Roche NGS system (GS20) was capable of sequencing over 20 million base pairs (as 100 bp reads) in just over four hours. Since then sequencing capacity has increased, resulting in both higher read lengths and greater overall sequence output. Once a genome sequence has been assembled, it is easier to perform re-sequencing of other individuals of the same species, as sequence reads can easily be mapped to the assembled reference genome. Another NGS platform, the Illumina® GAIIx, was used to generate more than 50× coverage of the *B. rapa* genome (The *Brassica rapa* Genome Sequencing Project Consortium et al. 2011). Re-sequencing can be used for the discovery of sequence variation, especially SNPs, on a genome-wide scale, and more than 1.5 million *B. napus* SNPs have so far been identified using this approach (D Edwards pers. Comm.).

Sequence variation discovered using next-generation sequencing efforts requires validation to ensure that the

variation was not caused by an error during the sequencing or assembly stages, or by the mis-mapping of NGS reads associated with increased ploidy or multigene families. While higher read depths can provide some level of confidence, it may not always be economical to sequence enough individuals to supply the required depth. Polymorphisms present in only one read are usually disregarded, with a redundancy of two or more reads favoured to diminish the impact of sequencing errors.

Problems posed by crop genomes

Most endeavours in genome sequencing and detection of genetic variation are directed towards crop improvement. Unfortunately, crop species frequently have very large, complicated and often polyploid genomes, such as allohexaploid wheat (*Triticum aestivum*), allotetraploid canola (*B. napus*) and sugarcane (*Saccharum sp.*) with 2n = 100 to 130 chromosomes. While genome size of itself is a hindrance, the complexity originating from the amplification of repetitive elements, as well as whole or partial genome duplication, impedes sequencing efforts. SNP discovery subsequently also relies on higher coverage to ensure confidence in the SNP detection. One way to combat these issues is to apply reduced complexity sequencing approaches that focus on a select portion of the genome rather than the complete genome.

One approach to simplify the process of SNP calling is to attempt sequencing of only a portion of the genome rather than the whole. Expressed sequence tag (EST) sequencing can be a useful tool in simplifying sequencing and gene discovery, narrowing the analysis to genic SNPs. However, SNPs derived from EST data may still be problematic due to gene and genome duplication. An alternative methodology for selecting a portion of the genome to be sequenced is sequence capture. In this method, probes are designed based on a reference genome to "capture" large contiguous or separate genomic regions of interest, enriching a DNA sample for sequencing. Another strategy that is suitable for the large, allohexaploid bread wheat genome has been to isolate individual chromosome arms and sequence these separately. This approach eliminates the majority of the homoeology resulting from multiple genomes as well as decreasing the effective genome size and complexity, and has been used successfully to sequence several chromosome arms of bread wheat (Berkman et al. 2012a; Berkman et al. 2012b; Berkman et al. 2011; Hernandez et al. 2012; Lai et al. 2012). Genomic complexity in polyploid crops such as *B. napus* can also be reduced for SNP discovery using the CRoPS system (complexity reduction of polymorphic sequences). This approach combines AFLP fingerprinting with next generation sequencing, and has been utilised for SNP discovery in maize (Orsouw et al. 2007).

Discovery tools – *in silico* SNP prediction

Bioinformatics applications that make viewing and interpreting the large volume of data created by next generation sequencing easier have become necessary (Lee et al. 2012; Marshall et al. 2010). For SNP discovery in particular, the level of genome coverage required for confident SNP calling creates a large amount of sequence data that must be handled and interpreted. Bioinformatics applications available for viewing this data include Tablet (Milne et al. 2010), MagicViewer (Hou et al. 2010) and Geneious (Drummond et al. 2011). SGSautoSNP (Second Generation Sequencing autoSNP) is an application designed specifically to predict SNPs from whole genome Illumina® shotgun sequence data. This predictive software has been successfully applied to identify more than 1.5 million SNPs in canola, with accuracy greater than 95% (D. Edwards, pers. comm.).

New sequencing and genotyping technologies

The technologies described above are collectively called Next-Generation Sequencing or Second-Generation Sequencing. The advent of further advances has necessitated the need for the terms Next-Next-Generation or Third-Generation Sequencing methods. These terms refer to technologies that utilise 'single-molecule' sequencing (SMS).

Developments that allow for higher indexing levels of sequencing libraries have further decreased the cost of sequencing, as more samples are able to be pooled in the same lane (currently up to 96-fold pooling is available using the NextFlex and Nextera kits). In addition to cost saving, this has allowed for the development of the technique known as Genotyping-By-Sequencing. This technique involves low coverage sequencing of larger groups of samples such as mapping populations, producing an efficient means of large-scale genotyping (including SNP discovery).

The Illumina® GoldenGate and Infinium assays provide excellent utility in SNP screening and detection of genetic variation for both research and breeding purposes (described in more detail below). Last year, an international consortium was established in *B. napus* to design and fund

the development of a public high-density (50K) Illumina® Infinium SNP array. SNP sequence data contributions from international research partners in Australia, Europe, China, North America and South America are expected to lead to the production of this chip mid-2012, with a per-sample expected cost of less than US$100. This development is expected to facilitate association mapping studies, breeding efforts and studies of global gene expression networks in canola, paving the way for predictive breeding (Stokes et al., 2010) and genomic selection (Cowling and Balázs 2010) in this agriculturally important crop species.

## High throughput and large scale genotyping technologies

The methods used to genotype SNPs depend largely on the technology and sequence information available. The end-use is determined by the density and location of the SNPs in the genome, their level of polymorphism in the populations of interest, and their ability to be genotyped across multiple individuals rapidly and cost-effectively.

As discussed above, the most important consideration when identifying and genotyping SNPs in *B. napus* is being able to distinguish genome-specific SNPs within the complex allotetraploid genome. Most current and high-throughput methods of SNP genotyping require some knowledge of the SNP and surrounding sequence, although traditional SNP detection techniques were independent of sequence information. These techniques were valuable for genetic mapping and assessing genetic diversity in species where limited sequence information was available. Where sequence data is available, PCR-based methods such as SNuPE (Single Nucleotide Primer Extension), involving amplification and allele-specific extension of SNP target regions with fluorescently labelled primers followed by screening on polyacrylamide gels, can be used for phosphorimage visualization and quantitation of SNP allele across a large sample set (Batley et al. 2003b). With the declining costs and increasing accessibility of next generation sequencing and genotyping assays, these technologies have been largely superseded by high-throughput genotyping assays such as Amplifluor (Serological Corp.), the Affymetrix Genechips, and the SNPlex, TaqMan and SnaPshot assays from Applied Biosystems (Appleby et al. 2009). Most recently, Illumina® has introduced the Goldengate and Infinium high-throughput genotyping assays (Fan et al. 2006).

### *Illumina® Goldengate Assay*

The Illumina® Goldengate assays can be designed for any species with user-defined sequence information. It involves multiplex, loci-specific PCR amplification followed by an allele-specific primer-extension and ligation reaction using red or green fluorescently labelled primers to distinguish between biallelic SNPs. Reactions are carried out on BeadChip multi-sample arrays and allele-specific fluorescence is detected using an iScan system or a BeadArray Reader produced by Illumina®. These Beadchips can genotype up to 32 samples simultaneously for a maximum of 3072 SNPs. Any number of samples can then be genotyped in multiples of 16 depending on the number of chips purchased, with the potential to generate over 300 000 genotypes in six hours of labour.

In *Brassica* the Illumina® Goldengate assay has been used to successfully screen 384 SNPs identified between *B. rapa* cultivars "chiifu" and "kenshin" across their doubled haploid (DH) mapping progeny, as well as a diverse set of wild and cultivated *Brassica* species (J. Batley pers. com.). Of the 384 SNPs screened, 338 (88%) of the alleles could be reliably discriminated between members of the mapping population and diverse *Brassica* species, with species clearly segregating into expected genome groupings following phylogenetic analyses (J. Batley pers. com.). These SNPs were also anchored to the recently released *Brassica rapa* genome sequence (The *Brassica rapa* Genome Sequencing Project Consortium et al. 2011) to assist in genome contig orientation and assembly. Additional work in *B. napus* used the Goldengate assay to screen 360 accessions with SNPs identified from EST data (Durstewitz et al. 2010). Approximately 84% of these SNPs produced clear genotypes with good cluster separation and peak intensities, with 46% demonstrated to be polymorphic in a panel of 90 winter oilseed rape lines. The majority of the identified SNP markers are expected to provide a useful resource for genetic mapping and genetic analyses (Durstewitz et al. 2010).

### *Illumina® Infinium HD Assay*

The Infinium assay is similar in principle to the Goldengate assay, but uses improved chemistry to cater for a vastly increased SNP number. In this assay, the PCR and extension-ligation steps are replaced with whole genome amplification and an allele-specific hybridisation and extension step. Custom Infinium chips can be purchased for any species of interest to genotype from 3000 to one million SNPs simultaneously. The number of samples accommodated per Infinium BeadChip is determined by the size of the assay, with 4, 12 and 24 samples able to be assayed simultaneously with up to one million, 250 000

and 90 000 SNPs respectively. The Infinium® II assay was recently released, and permits genotyping of an unlimited number of SNP loci simultaneously. As introduced above, a custom chip for *B. napus* has been designed to screen 50 000 genome-specific SNPs simultaneously. This chip will be available to the general research community to support a multitude of research applications, including physical SNP mapping for validation of current *B. napus de novo* genome sequencing and assembly efforts, *Brassica* evolutionary and diversity analyses and trait association/ mapping assays.

## Biological applications of SNPs

In *B. napus*, as for many plant species, SNPs are becoming increasingly popular as genetic markers, traversing the fields of pure genetics research to applied agriculture and plant molecular breeding. The low mutation rate of SNPs makes them valuable for understanding complex genetic traits and genome evolution (Syvänen, 2001). Specifically, SNPs are excellent genetic markers for high-density genetic map construction, physical ordering of chromosome contigs, QTL mapping, association and linkage disequilibrium (LD) studies, and comparative and evolutionary genomics analyses. Moreover, SNPs can be applied to genetic diagnostics, germplasm identification and marker-assisted selection for breeding programs in agriculture. These applications are explained in more detail below.

## Mapping, fine mapping and QTL

The abundance of SNPs and their ability to be discovered and genotyped rapidly and with high throughput makes them valuable markers for genetic mapping. SNPs identified within ESTs or transcriptome NGS data also permit estimates of allele frequencies and allelic association with phenotypes of interest. Quantitative trait loci (QTLs) in crops are regions of the genome that have been associated with traits of agronomic importance through traditional genetic mapping. Genetic mapping involves screening genetic markers in populations derived from two genetically diverse parents in order to link these markers to quantified segregating phenotypic characteristics. QTLs of importance for the canola industry include those for oil yield, oil quality, disease resistance and pod shatter tolerance amongst many others (Kaur et al. 2009; Pilet et al. 1998; Qiu et al. 2006; Smooker et al. 2011).

Once QTLs have been physically assigned to a region on the genome sequence, identifying and genotyping SNPs in these regions enables fine-mapping, or extremely high density mapping, of the QTLs (Choi et al. 2007). SNPs found to be associated with genes are then candidates for qualitative or quantitative trait nucleotides (QTNs), also called perfect markers, whereby different alleles are causally related to phenotypes. This enables the identification of candidate causal loci for further *in vitro* validation. This strategy is currently being applied to the discovery of genes required for resistance to the major fungal pathogen of canola, *Leptosphaeria maculans*, causal agent of blackleg disease (Hayward et al. 2012). Perfect markers also facilitate the development of assays for rapid screening of germplasm for genetic selection, or marker-assisted breeding.

## Association genetics and linkage disequilibrium studies

Linkage disequilibrium (LD) occurs when regions of the genome are inherited together at a frequency that is higher than that expected based on recombination. LD implies a relationship between these 'linked' regions. SNPs in species with moderate or high LD can be found as SNP haplotypes, which are comprised of a number of 'linked' SNP alleles that are always found in particular allelic combinations. These may extend over the length of genes or gene clusters (Edwards et al. 2007d). In this way just a small subset of SNPs may be required to define haplotypes, and these can be screened across breeding germplasm to provide a potentially powerful method for fast-tracking genomic selection, or for genetic improvement in breeding programs (Cowling and Balázs 2010). In *Arabidopsis*, linkage disequilibrium based on high density SNP maps is providing a powerful data set for evolutionary and association genetics studies (Atwell et al. 2010).

Association genetics has become a favoured approach to identify trait–marker relationships within many species. In this method SNPs, or any applicable genetic markers, are screened across populations or a diverse collection of individuals in order to associate alleles with phenotypic traits of interest on the basis of historical recombination and linkage disequilibrium (Cardon and Bell 2001; Flint-Garcia et al. 2003; Oraguzie 2007). Therefore, for species with low levels of LD, association studies have the potential to produce very high map resolution (Neale and Savolainen 2004). In *B. napus* there are collections of a small number of homozygous lines designated 'diversity fixed foundation sets', which aim to capture a large proportion of the genetic diversity available for the species (http://www.oregin.info/). With the availability of the 50K Infinium SNP chip for *B. napus* there will finally be a

large enough collection of polymorphic markers available to undertake a genome-wide association approach for this important crop species, previously hampered by the polyploid nature of the genome (Duran et al. 2010b).

## Marker assisted selection

Marker assisted selection involves the use of molecular markers tightly linked to an allele for a trait of interest. These molecular markers may then be used as a proxy for this desirable phenotype, allowing selection of plants from a segregating population at an early plant growth stage. However, the main issue with this approach hinges on the suitability of the molecular marker in question. To be effective, a molecular marker must be tightly linked to the trait of interest (generally within 1 cM), and high throughput, highly reproducible screening methods must be available for that marker (Mohan et al. 1997). This is often problematic, with many marker types showing either low polymorphism and/or poor genomic distribution (e.g. RFLPs) or poor reproducibility (e.g. RAPDs). Microsatellite (SSR) markers were the most polymorphic, reproducible markers available before the advent of SNPs, but are rarely found in genes (Hong et al. 2007; Mohan et al. 1997). Hence, the linkage between the marker and trait may not be very tight, and recombination could still act to cleave this linkage in some individuals in the population. SNPs however are by far the most prevalent form of genetic variation in the genome (Ching et al. 2002), and are quite often directly responsible for the allelic variation in the phenotypic trait of interest. Hence, the use of SNPs for marker-assisted selection offers two major advantages: firstly, obtaining a tightly linked SNP to a trait of interest is easier than obtaining any other form of molecular marker; and secondly, some of these SNPs may be causative agents for the trait, or "perfect" genetic markers for the phenotype.

In *B. napus*, marker assisted selection has been used for several different purposes, including selection of intervarietal substitution lines (Howell et al. 1996) and enrichment of genomic introgression lines (Zou et al. 2010), selection for major gene disease resistance (Chèvre et al. 1997), selection of yellow seed coat colour (Somers et al. 2001), selection for restored male fertility (Hansen et al. 1997) and improvement of oil quality (Tanhuanpää et al. 1995). These studies were predominantly carried out using SSRs and RAPDs, although more recently SNPs have also been utilised in marker-assisted selection for oil quality (Rahman et al. 2008). However, SNP markers offer greater versatility and utility than other marker types (Snowdon and Friedt 2004), and with the advent of whole genome sequence data and public SNP availability in *B. napus*, SNPs provide the future of marker assisted canola breeding.

## Genetic diversity

Within *Brassica*, and within the six agronomically important U's Triangle species in particular (U 1935), there is great scope for crop improvement via wide hybridisation and introgression of genetic diversity from wild relatives (Chen et al. 2011). *Brassica napus* in particular is a recent allopolyploid resulting from only a few hybridisation events between progenitor species *B. rapa* and *B. oleracea*, and hence contains only a fraction of the genetic diversity present in these diploid relatives (Song and Osborn 1992). Furthermore, rapeseed as a crop has undergone significant reductions in genetic diversity as a result of breeding canola quality oil for human consumption (Cowling 2007).

This lack of genetic diversity creates a major problem for breeding and selection for crop improvement of canola. As well, the large blocks of linkage disequilibrium in canola due to genetic bottlenecks and homozygosity in breeding populations create further problems for breeders. Regions of high LD create linkage drag whereby desirable alleles are inextricably linked to undesirable alleles, and have a strong negative influence on the utility of genetic markers for gene identification and marker assisted selection (Flint-Garcia et al. 2003). SNP markers, as the most common form of genetic variation across the genome, are highly useful in identifying regions of LD as well as regions of low genetic diversity. In the case of genetic introgressions from wild relatives or the diploid progenitor genomes into canola, SNP markers may be used to track chromosome segments and to select for recombination events that break up regions of high allelic LD, on a much finer scale than possible using other forms of molecular marker variation (Fourmann et al. 2002).

## Comparative genomics

The Brassicaceae family is in a uniquely favourable position for studies of comparative genomics (Hopkins et al. 2006), containing a large range of species distributed globally as well as model plant *Arabidopsis thaliana* and the agriculturally significant *Brassica* genus (Hong et al. 2006; Lim et al. 2007; Love et al. 2006a; Lysak and Koch 2011). Whole genome sequencing projects planned for the Brassicaceae in the immediate to near future will enable

SNP discovery on an unprecedented scale, allowing fine mapping and tracing of evolutionary events on a scale ranging from whole genome to chromosome to sequence evolution over time. Phylogenies in the Brassicaceae are still a matter of some dispute, partly due to the recurrent polyploid and hybridisation events which have characterised this family (Lysak and Koch 2011). However, use of SNPs for high-throughput evolutionary analysis is expected to allow the origin and timing of whole genome duplication and hybridisation events and chromosome rearrangements to be determined, along with elucidation of ancestral karyotypes in the Brassicaceae (Schranz et al. 2007). As mentioned above, early high-throughput SNP genotyping efforts across the Brassicaceae suggest that this family is amenable to large-scale cross-species SNP genotyping assays, even using SNPs derived from only a few sequenced species or cultivars (J. Batley pers. com.).

## Conclusion

The applications of SNPs to the study of complex crop genomes, including *B. napus*, are manifold. With rapid advancements in next generation sequencing and genotyping technologies, there are likely to be a number of landmark discoveries in the near future, with potential to greatly broaden our understanding of crop genome evolution and capacity for agricultural improvement. With the development of massively multiplexed sequencing assays, such as 'genotyping by sequencing', SNPs will continue to provide an invaluable resource for *B. napus* biologists and breeders alike due to their abundance, amenability to high-throughput screening and relative ease of discovery and downstream analyses.

## Acknowledgements

The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462, DP0985953 and DE120100668). Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) is gratefully acknowledged.

## References

Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants, in: D Somers, et al. (Eds.), Plant genomics, Methods in Molecular Biology: Humana Press, USA. pp 19-40

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, Meaux Jd, Ecker JR, Faure N, Kniskern JM, Jones JDG, Michael T, Nemri A, Roux F, David E. Salt, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627-631

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using AutoSNP. Bioinformatics 19:421-422

Batley J, Edwards D (2007) SNP applications in plants, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer, New York. pp 95-102

Batley J, Edwards D (2009) Genome sequence data: management, storage and visualization. Biotechniques 46:333-336

Batley J, Jewell E, Edwards D (2007a) Automated discovery of Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeat (SSR) molecular genetic markers, in: D Edwards (Ed.), Plant bioinformatics, Methods in Molecular Biology, Humana Press, USA. pp 473-494

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003a) Mining for Single Nucleotide Polymorphisms and insertions/deletions in maize Expressed Sequence Tag data. Plant Physiol 132:84-91.

Batley J, Mogg R, Edwards D, O'Sullivan H, Edwards KJ (2003b) A high-throughput SNuPE assay for genotyping SNPs in the flanking regions of *Zea mays* sequence tagged simple sequence repeats. Mol Breeding 11:111-120

Batley J, Hopkins CJ, Cogan NOI, Hand M, Jewell E, Kaur J, Kaur S, Li X, Ling AE, Love C, Mountford H, Todorovic M, Vardy M, Walkiewicz M, Spangenberg GC, Edwards D (2007b) Identification and characterisation of Simple Sequence Repeat (SSR) markers from *Brassica napus* expressed sequences. Mol Ecol Notes 7:886-889

Berkman PJ, Lai K, Lorenc MT, Edwards D (2012a) Next generation sequencing applications for wheat crop improvement. Am J Bot 99:365-371

Berkman PJ, Skarshewski A, Manoli S, Lorenc MT, Stiller J, Smits L, Lai K, Campbell E, Kubaláková M, Šimková H, Batley J, Dolezel J, Hernandez P, Edwards D (2012b) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. Theor Appl Genet 124:423-432.

Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, Ling EYS, Stiller J, Smits L, Imelfort M, Manoli S, McKenzie M, Kubaláková M, Šimková H, Batley J, Fleury D, Doležel J, Edwards D (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. Plant Biotech J 9:768-775

Bhattramakki D, Rafalski A (2001) Discovery and application of single nucleotide polymorphism markers in plants, in: RJ Henry (Ed.), Plant genotyping: the DNA fingerprinting of plants, CABI Publishing, Wallingford, Oxon. pp 179-193

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (2002) Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol 48:539-547

Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91-99

CCC (2010) Canola Council of Canada - Official definition of canola, http://www.canolacouncil.org/ind_definition.aspx.

Chagné D, Batley J, Edwards D, Forster JW (2007) Single Nucleotide Polymorphisms genotyping in plants, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer, New York. pp 77-94

Chen S, Nelson MN, Chèvre AM, Jenczewski E, Li Z, Mason AS, Meng J, Plummer JA, Pradhan A, Siddique KHM, Snowdon RJ, Yan G, Zhou W, Cowling WA (2011) Trigenomic bridges for *Brassica* improvement. Crit Rev Plant Sci 30:524-547

Chèvre AM, Barret P, Eber F, Dupuy P, Brun H, Tanguy X, Renard M (1997) Selection of stable *Brassica napus-B. juncea* recombinant lines resistant to blackleg (*Leptosphaeria maculans*). 1. Identification of molecular markers, chromosomal and genomic origin of the introgression. Theor Appl Genet 95:1104-1111

Ching A, Caldwell KS, Jung M, Dolan M, Smith OSH, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Choi SR, Teakle GR, Plaha P, Kim JH, Allender CJ, Beynon E, Piao ZY, Soengas P, Han TH, King GJ, Barker GC, Hand P, Lydiate DJ, Batley J, Edwards D, Koo DH, Bang JW, Park BS, Lim YP (2007) The reference genetic linkage map for the multinational *Brassica rapa* genome sequencing project. Theor Appl Genet 115:777-792

Cowling WA (2007) Genetic diversity in Australian canola and implications for crop breeding for changing future environments. Field Crops Res 104:103-111

Cowling WA, Balázs E (2010) Prospects and challenges for genome-wide association and genomic selection in oilseed *Brassica* species. Genome 53:1024-1028

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2011) Geneious v5.5, Available from http://www.geneious.com

Duran C, Edwards D, Batley J (2009a) Genetic maps and the use of synteny, in: DJ Somers, et al. (Eds.), Plant genomics, Methods in Molecular Biology, Humana Press, USA. pp 41-56

Duran C, Edwards D, Batley J (2009b) Molecular marker discovery and genetic map visualisation, in: D Edwards, et al. (Eds.), Applied Bioinformatics, Springer, USA. pp 165-189

Duran C, Appleby N, Edwards D, Batley J (2009c) Molecular genetic markers: discovery, applications, data storage and visualisation. Curr Bioinform 4:16-27.

Duran C, Boskovic Z, Batley J, Edwards D (2011) Role of bioinformatics as a tool for vegetable *Brassica* species, in: J Sadowski (Ed.), Vegetable *Brassica*s, Science Publishers Inc., New Hampshire. pp 406-418

Duran C, Boskovic Z, Imelfort M, Batley J, Hamilton NA, Edwards D (2010a) CMap3D: A 3D visualisation tool for comparative genetic maps. Bioinformatics 26:273-274

Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D (2009d) AutoSNPdb: an annotated Single Nucleotide Polymorphism database for crop plants. Nucleic Acids Res 37:951-953

Duran C, Eales D, Marshall D, Imelfort M, Stiller J, Berkman P, Clark T, McKenzie M, Appleby N, Batley J, Basford K, Edwards D (2010b) Future tools for association mapping in crop plants. Genome 53:1017-1023

Durstewitz G, Polley A, Plieske J, Luerssen H, Graner EM, Wieseke R, Ganal MW (2010) SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. Genome 53:948-956

Edwards D (2007) Bioinformatics and plant genomics for staple crops improvement, in: MS Kang and PM Priyadarshan (Eds.), Breeding major food staples, Blackwell. pp 93-106

Edwards D, Batley J (2004) Plant bioinformatics: from genome to phenome. Trends Biotechnol 22:232-237

Edwards D, Batley J (2008) Bioinformatics: fundamentals and applications in plant genetics, mapping and breeding, in: C Kole and AG Abbott (Eds.), Principles and practices of plant genomics, Science Publishers Inc., USA. pp 269-302

Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. Plant Biotech J 8:2-9

Edwards D, Wang X (2012) Genome sequencing initiatives, in: D Edwards, et al. (Eds.), Genetics, genomics and breeding of oilseed *Brassica*s, Science Publishers Inc., New Hampshire, USA. pp 152-157

Edwards D, Hansen D, Stajich J (2009) DNA sequence databases, in: D Edwards, et al. (Eds.), Applied bioinformatics, Springer, USA. pp 1-11

Edwards D, Burton WA, Hopkins CJ, Batley J (2007a) Indian mustard, in: C Kole (Ed.), Genome mapping and molecular breeding in plants, Springer, New York. pp 179-210

Edwards D, Forster JW, Chagné D, Batley J (2007b) What are SNPs?, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer, New York. pp 41-52

Edwards D, Batley J, Cogan NOI, Forster JW, Chagné D (2007c) Single Nucleotide Polymorphism discovery, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer, New York. pp 53-76

Edwards D, Batley J, Cogan NOI, Forster JW, Chagné D (2007d) Single Nucleotide Polymorphism discovery, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer, New York

Erwin T, Jewell E, Love C, Lim G, Li X, Chapman R, Batley J, Stajich J, Mongin E, Stupka E, Ross B, Spangenberg GC, Edwards D (2007) BASC: an integrated bioinformatics system

for *Brassica* research. Nucleic Acids Res 35:D870-D873

Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nat Rev Genet 7:632-644

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54: 357-374

Fourmann M, Barret P, Froger N, Baron C, Charlot F, Delourme R, Brunel D (2002) From *Arabidopsis thaliana* to *Brassica napus*: development of amplified consensus genetic markers (ACGM) for construction of a gene map. Theor Appl Genet 105:1196-1206

Friedt W, Snowdon R (2009) Oilseed rape, in: J Vollmann and I Rajcan (Eds.), Oil crops, Springer Science & Business Media.

Hansen M, Halldén C, Nilsson N-O, Säll T (1997) Marker-assisted selection of restored male-fertile *Brassica napus* plants using a set of dominant RAPD markers. Mol Breeding 3:449-456

Hayward A, McLanders J, Campbell E, Edwards D, Batley J (2012) Genomic advances will herald new insights into the *Brassica:Leptosphaeria maculans* pathosystem. Plant Biol 14:1-10

Hernandez P, Martis M, Dorado G, Pfeifer M, Gálvez S, Schaaf S, Jouve N, Šimková H, Valárik M, Doležel J, Mayer KFX (2012) Next generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69:377-386

Hong CP, Piao ZY, Kang TW, Batley J, Yang TJ, Hur YK, Bhak J, Edwards D, Lim YP (2007) Genomic distribution of Simple Sequence Repeats in *Brassica rapa*. Mol Cells 23:349-356

Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A survey of the *Brassica rapa* genome through BAC-end sequence analysis, and comparative analysis with *Arabidopsis thaliana*. Mol Cells 22:300-307

Hopkins C, Mogg R, Gororo N, Salisbury PA, Burton WA, Love C, Spangenberg GC, Edwards D, Batley J (2006) An assessment of genetic diversity within and between *Brassica napus* and *Brassica juncea* lines from international germplasm collections. Acta Hortic 706:115-119

Hopkins CJ, Cogan NOI, Hand M, Jewell E, Kaur J, Li X, Lim GAC, Ling AE, Love C, Mountford H, Todorovic M, Vardy M, Spangenberg GC, Edwards D, Batley J (2007) Sixteen new simple sequence repeat markers from *Brassica juncea* expressed sequences and their cross-species amplification. Mol Ecol Notes 7:697-700

Hou H, Zhao F, Zhou L, Zhu E, Teng H, Li X, Bao Q, Wu J, Sun Z (2010) MagicViewer: integrated solution for next-generation sequencing data visualization and genetic variation detection and annotation. Nucleic Acids Res 38:732-736

Howell PM, Marshall DF, Lydiate DJ (1996) Towards developing intervarietal substitution lines in *Brassica napus* using marker-assisted selection. Genome 39:348-358

Imelfort M, Edwards D (2009) *De novo* sequencing of plant genomes using second-generation technologies. Brief Bioinform 10:609-618

Imelfort M, Duran C, Batley J, Edwards D (2009a) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotech J 7:312-317

Imelfort M, Batley J, Grimmond S, Edwards D (2009b) Genome sequencing approaches and successes, in: D Somers, et al. (Eds.), Plant Genomics, Methods in Molecular Biology, Humana Press, USA

Kaur S, Cogan NOI, Ye G, Baillie RC, Hand ML, Ling AE, Mcgearey AK, Kaur J, Hopkins CJ, Todorovic M, Mountford H, Edwards D, Batley J, Burton W, Salisbury P, Gororo N, Marcroft S, Kearney G, Smith KF, Forster JW, Spangenberg GC (2009) Genetic map construction and QTL mapping of resistance to blackleg (*Leptosphaeria maculans*) disease in Australian canola (*Brassica napus* L.) cultivars. Theor Appl Genet120:71-83

Lai K, Berkman PJ, Lorenc MT, Duran C, Smits L, Manoli S, Stiller J, Edwards D (2012) WheatGenome.info: An integrated database and portal for wheat genome information. Plant Cell Physiol (In Press, accepted October 2011)

Lee H, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D (2012) Bioinformatics tools and databases for analysis of next generation sequence data. Briefings in Functional Genomics 11:12-24

Lim GAC, Jewell EG, Li X, Erwin TA, Love C, Batley J, Spangenberg G, Edwards D (2007) A comparative map viewer integrating genetic maps for *Brassica* and *Arabidopsis*. BMC Plant Biol 7:40

Lorenc M, Boskovic Z, Stiller J, Duran C, Edwards D (2012) Role of bioinformatics as a tool for oilseed *Brassica* species, in: D Edwards, et al. (Eds.), Genetics, genomics and breeding of oilseed *Brassica*s, Science Publishers Inc., New Hampshire. pp. 194-205

Love C, Logan E, Erwin T, Spangenberg G, Edwards D (2006a) Analysis of the *Brassica* A and C genomes and comparison with the genome of *Arabidopsis thaliana*. Acta Hort 706: 99-104

Love C, Robinson A, Lim G, Hopkins C, Batley J, Barker G, Spangenberg GC, Edwards D (2005) *Brassica* ASTRA: an integrated database for *Brassica* genomic research. Nucleic Acids Res 33:W493-W495

Love C, Logan E, Erwin T, Kaur J, Lim GAC, Hopkins C, Batley J, James N, May S, Spangenberg G, Edwards D (2006b) Integrating and interrogating diverse *Brassica* data within an EnsEMBL structured database. Acta Hort 706:77-82

Love CG, Edwards D (2007) Accessing integrated *Brassica* genetic and genomic data using the BASC server, in: D Edwards (Ed.), Plant Bioinformatics, Methods in Molecular Biology, Humana Press, USA. pp 229-244

Love CG, Batley J, Lim G, Robinson AJ, Savage D, Singh D, Spangenberg GC, Edwards D (2004) New computational tools for *Brassica* genome research. Comp Funct Genom 5:276-280

Lysak MA, Koch MA (2011) Phylogeny, genome, and karyotype evolution of crucifers (*Brassica*ceae), in: R Schmidt and I Bancroft (Eds.), Genetics and Genomics of the *Brassica*ceae, Plant Genetics and Genomics: Crops and Models 9, Springer

Science + Business Media, LLC

Marshall DJ, Hayward A, Eales D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Lai K, Duran C, Batley J, Edwards D (2010) Targeted identification of genomic regions using TAGdb. Plant Methods 6:19

MBGP Multinational *Brassica* Genome Project - www.*Brassica*.info

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet - next generation sequence assembly visualization. Bioinformatics 26:401-402

Mogg R, Batley J, Hanley S, Edwards D, O'Sullivan H, Edwards KJ (2002) Characterisation of the flanking regions of *Zea Mays* microsatellites reveals a large number of useful sequence polymorphisms. Theor Appl Genet 105:532-543

Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, Bhatia CR, Sasaki T (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breeding 3:87-103

Morinaga T (1934) Interspecific hybridisation in *Brassica* VI. The cytology of F1 hybrids of *B. juncea* and *B. nigra*. Cytologia 6:62-67

Mun JH, Kwon SJ, Seol YJ, Kim JA, Jin M, Kim JS, Lim MH, Lee SI, Hong JK, Park TH, Lee SC, Kim BJ, Seo MS, Baek S, Lee MJ, Shin JY, Hahn JH, Hwang YJ, Lim KB, Park JY, Lee J, Yang TJ, Yu HJ, Choi SR, Ramchiary N, Lim YP, Fraser F, Drou N, Soumpourou E, Trick M, Bancroft I, Parkin IAP, Batley J, Edwards D, Park BS (2010) Sequence and structure of *Brassica rapa* chromosome A3. Genome Biol 11:R94

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325-330

Oraguzie NC (2007) An overview of association mapping, in: NC Oraguzie, et al. (Eds.), Association mapping in plants, Springer Science + Business Media, New York

Orsouw NJv, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, Schneiders H, Poel Hvd, Oeveren Jv, Verstegen H, Eijk MJTv (2007) Complexity reduction of polymorphic sequences (CRoPS$^{TM}$): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS ONE 2:e1172

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167-174

Pilet ML, Delourme R, Foisset N, Renard M (1998) Identification of loci contributing to quantitative field resistance to blackleg disease, causal agent Leptosphaeria maculans (Desm.) Ces. et de Not., in Winter rapeseed (*Brassica napus* L.). Theor Appl Genet 96:23-30

Qiu D, Morgan C, Shi J, Long Y, Liu J, Li R, Zhuang X, Wang Y, Tan X, Dietrich E, Weihmann T, Everett C, Vanstraelen S, Beckett P, Fraser F, Trick M, Barnes S, Wilmer J, Schmidt R, Li J, Li D, Meng J, Bancroft I (2006) A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. Theor Appl Genet 114:67-80

Rahman M, Sun Z, McVetty PBE, Li G (2008) High throughput genome-specific and gene-specific molecular markers for erucic acid genes in *Brassica napus* (L.) for marker-assisted selection in plant breeding. Theor Appl Genet117:895-904

Raymer PL (2002) Canola: an emerging oilseed crop, in: J Janick and A Whipkey (Eds.), Trends in new crops and new uses, ASHS Press, Alexandria

Schranz ME, Song B-H, Windsor AJ, Mitchell-Olds T (2007) Comparative genomics in the *Brassica*ceae: a family-wide perspective. Curr Opin Plant Biol10:168-175

Smooker AM, Wells R, Morgan C, Beaudoin F, Cho K, Fraser F, Bancroft I (2011) The identification and mapping of candidate genes and QTL involved in the fatty acid desaturation pathway in *Brassica napus*. Theor Appl Genet 122:1075-1090

Snowdon RJ, Friedt W (2004) Molecular markers in *Brassica* oilseed breeding: current status and future possibilities. Plant Breeding 123:1-8

Somers DJ, Rakow G, Prabhu VK, Friesen KRD (2001) Identification of a major gene and RAPD markers for yellow seed coat colour in *Brassica napus*. Genome 44:1077-1082

Song K, Osborn TC (1992) Polyphyletic origins of *Brassica napus*: new evidence based on organelle and nuclear RFLP analyses. Genome 35:992-1001

Stokes D, Fraser F, Morgan C, O'Neill CM, Dreos R, Magusin A, Szalma S, Bancroft I (2010) An association transcriptomics approach to the prediction of hybrid performance. Mol Breeding 26:91-106

Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930-942

Taiyan Z, Lianli L, Guang Y, Al-Shehbaz IA (2001) *Brassica*ceae, in: Z Wu and PH Raven (Eds.), Flora of China, Science Press, Beijing

Tanhuanpää PK, Vilkki JP, Vilkki HJ (1995) Association of a RAPD marker with linolenic acid concentration in the seed oil of rapeseed (*Brassica napus* L.). Genome 38:414-416

The *Brassica rapa* Genome Sequencing Project Consortium, Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun J-H, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park B-S, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IAP, Batley J, Kim J-S, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon S-J, Choi S-R, Lee T-H, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, et al. (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035-1039

Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. Plant Biotech J 7:334-346

U N (1935) Genome-analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. Jap J Bot 7:389-452

UNFAO, http://www.fao.org/index_en.htm, Food and Agriculture Organization of the United Nations

Westermeier P, Wenzel G, Mohler V (2009) Development and evaluation of single-nucleotide polymorphism markers in allotetraploid rapeseed (*Brassica napus* L.). Theor Appl Genet 119:1301-1311

Zou J, Zhu J, Huang S, Tian E, Xiao Y, Fu D, Tu J-x, Fu T-d, Meng J (2010) Broadening the avenue of intersubgenomic heterosis in oilseed *Brassica*. Theor Appl Genet 120:283-290