REVIEW

A user guide to the *Brassica* 60K Illumina Infinium™ SNP genotyping array

Annaliese S. Mason[1] · Erin E. Higgins[3] · Rod J. Snowdon[1] · Jacqueline Batley[2,4] · Anna Stein[1] · Christian Werner[1] · Isobel A. P. Parkin[3]

Received: 11 May 2016 / Accepted: 14 September 2016 / Published online: 20 February 2017

Abstract The *Brassica napus* 60K Illumina Infinium™ SNP array has had huge international uptake in the rapeseed community due to the revolutionary speed of acquisition and ease of analysis of this high-throughput genotyping data, particularly when coupled with the newly available reference genome sequence. However, further utilization of this valuable resource can be optimized by better understanding the promises and pitfalls of SNP arrays. We outline how best to analyze *Brassica* SNP marker array data for diverse applications, including linkage and association mapping, genetic diversity and genomic introgression studies. We present data on which SNPs are locus-specific in winter, semi-winter and spring *B. napus* germplasm pools, rather than amplifying both an A-genome and a C-genome locus or multiple loci. Common issues that arise when analyzing array data will be discussed, particularly those unique to SNP markers and how to deal with these for practical applications in *Brassica* breeding applications.

Introduction

Brassica napus L. (AC genome) is a recent allopolyploid species, less than 7500 years old (Chalhoub et al. 2014), formed through interspecific hybridization between *B. rapa* (A genome) and *B. oleracea* (C genome). These diploid genomes are themselves ancestral polyploids (Lysak et al. 2005) as *B. rapa* and *B. oleracea* have undergone multiple rounds of genome duplication (Cheng et al. 2013; Parkin et al. 2005; Schranz et al. 2006). As a result of this, most genes are present in multiple copies in the *Brassica* genomes. Furthermore, characterization of the *B. napus* genome has demonstrated that there are numerous homoeologous (non-homologous, ancestrally-related region) exchanges between the A and C genomes of *B. napus* (Chalhoub et al. 2014). Crossovers between A- and C-genome chromatids during meiosis I can result in production of gametes with balanced (reciprocal) and unbalanced (non-reciprocal, duplication or deletion) translocation events in the A and C chromosomes. Duplications and deletions may then be fixed in individuals or populations through self-pollination, putatively resulting in both the copy number variation (CNV) and presence absence variation (PAV) previously observed in the *B. napus* genome (Schiessl et al. 2014). This history of recurrent polyploidy and subsequent chromosome rearrangements, along with a high genomic fraction of repetitive elements, contributes to genome complexity and can cause challenges in genotyping *B. napus* cultivars (Fu et al. 2015).

A variety of DNA marker systems have been established in *B. napus* over the last 35 years, ranging from

Communicated by R. K. Varshney.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-016-2849-1) contains supplementary material, which is available to authorized users.

✉ Annaliese S. Mason
annaliese.mason@agrar.uni.giessen.de

[1] Department of Plant Breeding, IFZ for Biosystems, Land Use and Nutrition, Justus Liebig University Giessen, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany

[2] School of Agriculture and Food Sciences and Centre for Integrative Legume Research, The University of Queensland, Brisbane 4072, Australia

[3] Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon S7N0X2, Canada

[4] School of Plant Biology and The UWA Institute of Agriculture, The University of Western Australia, 35 Stirling Highway, Crawley, 6009 Perth, Australia

hybridization-based markers, to PCR-based markers, to sequence-based multi-parallel marker systems (see Obermeier and Friedt (2015) for review). SNPs are single nucleotide differences between the DNA sequences of individuals in a population, and are categorised as transversions (C/G, A/T, C/A and T/G), transitions (C/T or G/A) and insertions/deletions (indels). SNPs represent the most frequent type of genetic polymorphism, and may therefore provide a high density of markers near a locus of interest (Picoult-Newberg et al. 1999). The vast majority of SNPs are bi-allelic, although tri-allelic and tetra-allelic SNPs also exist (Brookes 1999). Multiallelic markers such as simple sequence repeats (SSRs) can be preferable for genetic diversity studies, particularly within species, because of their higher polymorphism and faster mutation rate (Hodel et al. 2016); however, SSRs are often poorly linked to genes (Hong et al. 2007; Mohan et al. 1997). SNPs are currently the preferred markers for many applications due to their high prevalence in the genome and their potential for strong, or even perfect, linkage to traits of interest (Hayward et al. 2012). They have a fine resolution, are highly stable and reproducible (Syvänen 2001) and are amenable to ultra-high-throughput discovery and detection (Anithakumari et al. 2010; Barchi et al. 2011; Batley et al. 2007; Duran et al. 2009a, b). With the development of next generation sequencing, it has become possible to automate the discovery of millions of SNPs (Imelfort et al. 2009), which has the potential to drive genomics-assisted improvement of canola (Hayward et al. 2012).

The Illumina Infinium™ assay is an array technology capable of genotyping between 3000 and 1 million polymorphic sites per sample, with multiple samples assayed simultaneously in a single experiment (Chagné et al. 2015). Following the selection of SNPs to be assayed, probes are designed to the sequence immediately adjacent to the target SNPs. During the assay a whole-genome amplification step, rather than PCR, is used to increase the amount of DNA up to 1000-fold. The amplified DNA is fragmented and captured on a bead array by hybridization to the immobilised SNP-specific primers, followed by single-base extension with hapten-labelled nucleotides representing the SNP allele. The signal for the incorporated hapten-modified nucleotides is amplified by adding fluorescently labelled antibodies in several steps and subsequently detected using a high resolution confocal scanner.

SNP arrays can be disadvantageous for some applications relative to other SNP genotyping approaches: assessment of genetic diversity will necessarily be limited to the variants initially used for the production of the array, which may introduce bias and exclude rare alleles (Ganal et al. 2012). As well, for applications involving large populations but only requiring a few markers, such as marker-assisted selection, other platforms (e.g. Taqman, KASP) may be more suitable. However, whilst genotyping by sequencing (GBS) and whole genome re-sequencing are being utilised in some instances for genotypic analysis, array platforms have several important advantages over GBS approaches. These include rapid data generation (a 2–3-day turnaround) and a relatively easy sample preparation protocol, as well as simple data analysis that does not require significant bioinformatic support. Data reproducibility is also high: the same SNPs are genotyped across all individuals, allowing for straightforward comparison between samples. Table 1 summarises some advantages and disadvantages of SNP array marker systems.

A number of high throughput SNP arrays in *B. napus* have been developed, initially with 6000 to 9000 SNPs. The rapid advances using these arrays (Dalton-Morgan et al. 2014; Delourme et al. 2013; Edwards et al. 2013) led to the development of a community-led *Brassica* 60K Illumina Infinium™ array, containing a total of 52 157 SNPs (Clarke et al. 2016). Since the commercial release of the *Brassica* 60K Illumina Infinium array in 2013 there have been many diverse applications of the array, including molecular karyotyping (Mason et al. 2014), germplasm collection characterization (Mason et al. 2015) and genome-wide association mapping for traits such as seed weight and quality (Li et al. 2014), seed glucosinolate and leaf chlorophyll content (Qian et al. 2016), seed germination and vigor (Hatzig et al. 2015), branch angle (Liu et al. 2016), flowering time (Xu et al. 2016), *Sclerotinia* disease resistance (Wei et al. 2016) and plant height and branch number (Li et al. 2016). The

Table 1 Advantages and disadvantages of the SNP array marker systems

Advantages	Disadvantages
Large amount of marker data generated in one experiment with reasonably good distribution across the genome	High cost and effort of array establishment (identification, validation and implementtation of SNPs)
High information content, including physical position, allelic and insertion/deletion (indel) variation and presence-absence variation	Analyses are only cost-effective in larger sets of individuals and for the complete genome
No need for DNA fragment separation	Screening of specific genomic regions can be difficult due to biased marker distribution
Fast, high-throughput automation for large sample sizes is possible at a moderate cost	Physical marker position can be difficult to determine due to the complexities of a reticulate genome and imperfection of the available genome reference sequence

array has also been applied for high-resolution, biparental QTL mapping, for example to map QTL and candidate genes for fatty acid content (Liu and Li 2014), root morphology (Fletcher et al. 2015), seed glucosinolate content (Qu et al. 2015) and water stress tolerance (Zhang et al. 2015). In this review we discuss some of these applications in the context of the complex *B. napus* genome, and suggest best practices for analyzing and interpreting SNP data from the *Brassica* 60K Illumina Infinium™ array for a range of practical breeding applications.

Visualization of SNP array marker data: recognizing the different types of SNPs and their uses for different applications

Visualizing fluorescence intensities for individual SNP markers and individuals on the Illumina array is helpful in identifying which SNP markers are reliable. This can be readily undertaken using the graphical user interface Illumina GenomeStudio software package (Figure S1), now freely available from Illumina (http://support.illumina.com/array/array_software/genomestudio/downloads.html). Open source alternatives for reading, filtering, normalizing, visualizing marker patterns and calling genotypes from Illumina SNP array data also exist (Morgan 2016; Ritchie et al. 2009). Different algorithms can be used to process raw image data into genotype calls: Illumina uses the "Gencall" algorithm, which shows similar accuracy to "GenoSNP" and "CRLMM" (Ritchie et al. 2011); and new calling algorithms are continually being produced (Li 2016). Regardless of the method used to visualize the data, observation of SNP marker patterns provides valuable information about the relative trustworthiness of individual SNP markers in different populations.

Several common marker patterns can be observed in *Brassica* and particularly *B. napus* for array data, and recognizing each of these patterns and the underlying mechanism involved is essential for selecting appropriate SNP markers for particular analyses. The most straightforward pattern will show three distinct clusters, one for each expected genotype: AA, AB, and BB (e.g. A/A, A/G and G/G for an A/G SNP). The clusters should be well separated in terms of allelic intensity ratios; i.e. the normalized theta values are close to 0, 0.5 and 1 for the three genotypes, respectively (Fig. 1a). This is the classic diploid cluster pattern for segregation of a biallelic marker at a single locus within the experimental population, and this marker type is generally applicable for all analyses.

The high degree of homoeology between the A and C genomes of *B. napus* allows some SNP probes to hybridize to both of the primary homoeologues, i.e. to two different genome locations. When these loci are both polymorphic the resulting pattern in a diversity panel will have five distinct clusters: AAAA, AAAB, AABB, ABBB, and BBBB (Fig. 1b), whereas in an inbred mapping population three clusters would be observed (AAAA, AABB and BBBB) instead of two (Fig. 1a). Observations of heterozygous (AB) SNP calls in homozygous mapping populations (doubled haploids or inbred lines) are often due to this kind of SNP, where "heterozygosity" is actually amplification of one allele from the A genome and one allele from the C genome (Fig. 2). These markers should be eliminated for applications such as linkage mapping analysis, as unlike other marker types, where the detected loci/alleles are separated by size, e.g. simple sequence repeats (SSRs) or restriction fragment length polymorphisms (RFLPs), SNP markers cannot be used to score two loci simultaneously. The A- and C-genome alleles co-hybridise to the probe sequence and are superimposed on the SNP image, rather than separated by locus, and the use of two fluorescent dyes results in only two possible allele types, making the homoeologues impossible to disassociate. For example, if an individual is heterozygous at both loci and has the genotype ABAB, and a second individual is homozygous at both loci with the genotype AABB, each of these individuals has two A alleles and two B alleles. Thus, both individuals will fall into the same central cluster, making the scoring ambiguous. Only individuals homozygous for the same allele at both the A and C genome loci can be scored correctly, i.e. as AAAA or BBBB. Hence, this marker type is of limited use: it is possible to incorporate this information, for example in diversity assessment studies where genome location information may not be critical; however, the inaccuracy in genotype calling for non-homozygotes can introduce additional bias into the data.

A third common marker pattern that can be difficult to detect is due to hybridization of the SNP marker to two different genomic locations, one of which is fixed for a particular allele, and one of which is segregating. These markers are termed "hemi-SNPs" (Trick et al. 2009), and will show three genotype clusters but with shifts in relative fluorescence intensity, usually resulting in only AA and AB or AB and BB calls in the population using default clustering algorithms. However, these clusters can be manually repositioned to call the three genotypes correctly for the segregating locus (Fig. 1c). In the *Brassica* 60K cluster file (Clarke et al. 2016), the clusters were repositioned for the majority of such SNPs based on a diverse set of winter and spring type *B. napus* germplasm, resulting in correct genotype calls for most *B. napus* lines. However, since hemi-SNPs represent two independent loci, a hemi-SNP in one particular population may be a five-cluster SNP in a diversity set (or other mapping population) or vice versa, depending on polymorphism at the two loci being

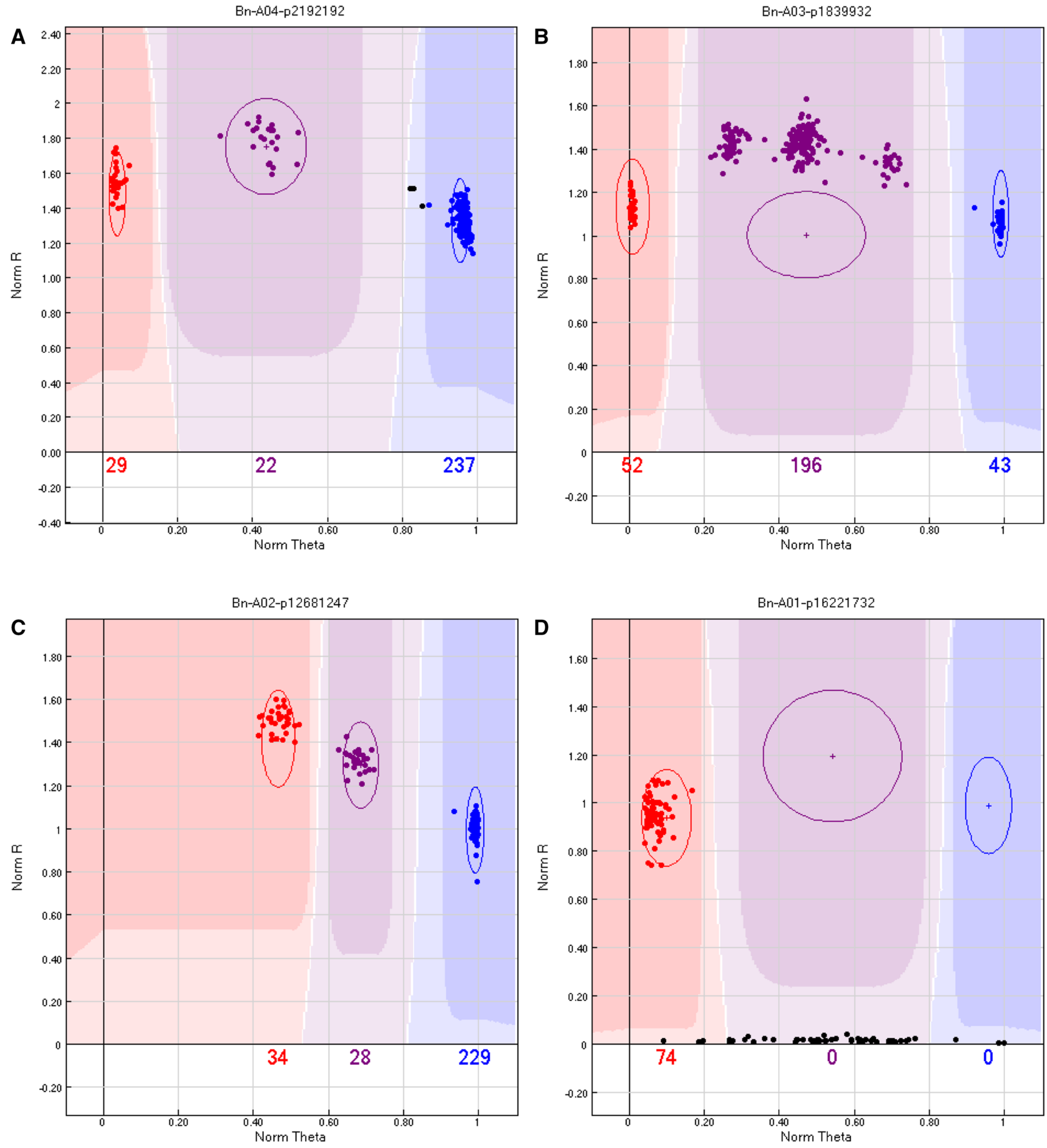

Fig. 1 GenomeStudio images of typical SNP marker patterns. **a** Classic diploid three-cluster SNP assay; **b** SNP marker which detects two homoeologous loci resulting in five distinct clusters; **c** hemi-SNP from hybridization with two loci, only one of which is segregating; **d** dominant SNP indicating the SNP probe only binds to one allele (presence-absence SNP)

amplified, so any particular cluster file may not reflect all germplasm sets accurately.

A number of less frequent marker patterns may also be observed. Some SNPs only detect one allele, resulting in A/- or B/- marker patterns (Figs. 1d, 2). These indel SNPs can be treated as dominant markers, since heterozygotes cluster together with individuals that are homozygous for the amplified allele. Some SNPs also hybridize to multiple genomic locations (three or more; i.e. to paralogous genomic sequences as well as homoeologues);

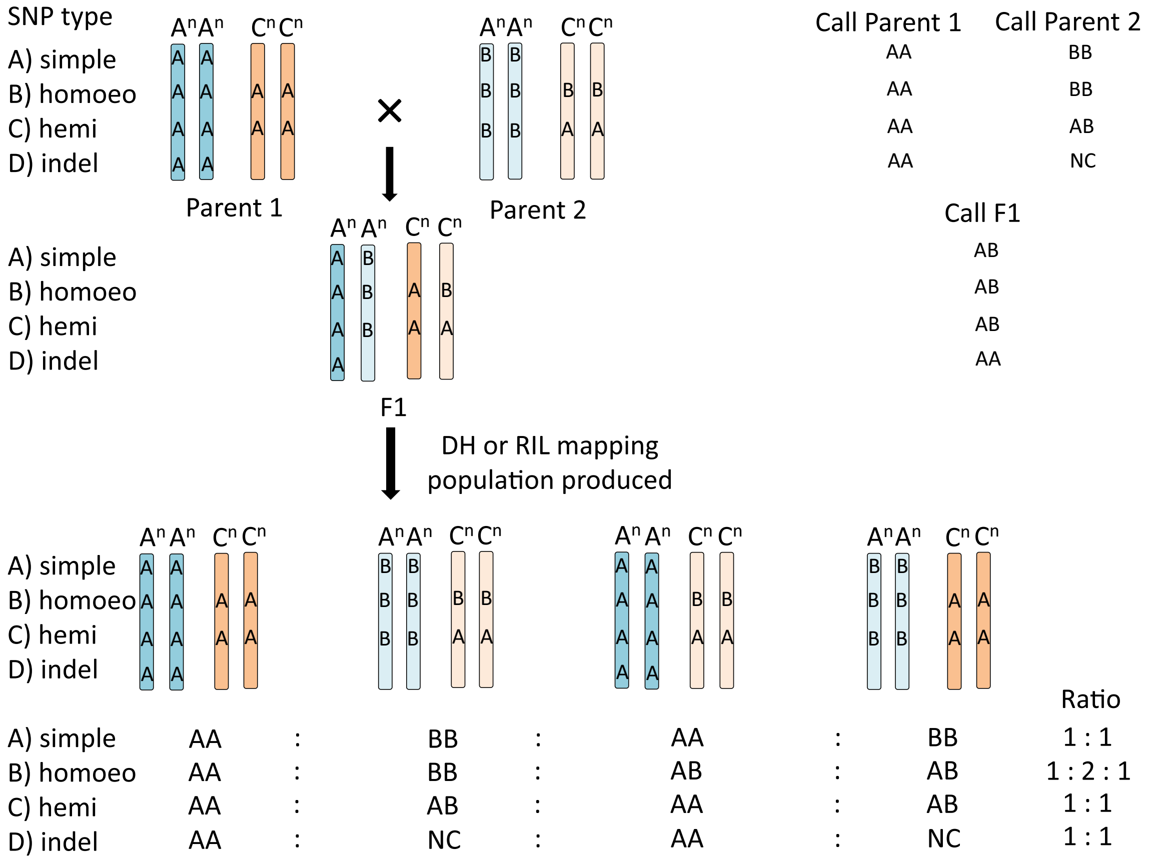

Fig. 2 SNP allele segregation in an inbred or homozygous mapping population of *Brassica napus*, showing chromatid and allele inheritance and population-based ratios for different SNP types. **a** Simple SNPs: normal locus-specific SNPs, polymorphic between parents. Expected segregation ratio AA:BB = 1:1 in a DH mapping population. **b** Homoeo SNPs: homoeologous SNPs, or SNPs amplifying polymorphic loci in both the A and C genomes. Only one type is shown, where each parent has the same homozygous call for both subgenome alleles. Homoeo SNPs are also called "5 cluster" SNPs (AAAA, AAAB, AABB, BBBA, BBBB genotype clusters), but only make 3 clusters in DH populations (AAAA, AABB, BBBB). Expected segregation ratio AA:AB:BB = 1:2:1 in a DH mapping population. **c** Hemi SNPs: a special class of homoeo SNPs (two loci amplified) where one locus is homozygous and the other is polymorphic (genotype clusters AAAA, AAAB, AABB, or just AAAA and AABB in a DH population). Can be manually reclustered to call "AB" genotype group as "BB" in Genome Studio. Expected segregation ratio AA:AB or BB:AB = 1:1 in a DH mapping population. **d** Indel SNPs (insertion/deletion variants): SNPs where an allele is amplified in one parent, but the other parent has no amplification of any allele. Expected segregation ratio AA:NC or BB:NC = 1:1 in a DH mapping population. *DH* doubled-haploid, *RIL* recombinant inbred lines

this can result in either more than five clusters or a spread of marker calls across the entire fluorescence range, with a generally higher frequency of heterozygote and no-calls than other marker types. Occasionally a SNP image will show two distinct groups in the same genotype cluster, separated vertically (i.e. by normalized R values) within the call region. One explanation for this is impaired hybridization of the probe based on a mismatch in the SNP flanking sequence, similar to that seen in microarrays where SNPs in the probe sequence have been shown to cause variation in fluorescence (Alberts et al. 2007). Such samples would have lower fluorescence due to poor hybridization, effectively blocking a portion of the binding sites on the SNP bead.

Parameters for clustering and filtering sample and SNP data in different population types

Several steps are suggested to produce a better data set for further analysis (Figure S2). As a useful shortcut, applying the cluster file developed for the *Brassica* 60K array (Clarke et al. 2016) in GenomeStudio is an excellent first step for analysis of genotype data. The cluster file eliminates SNPs with low or zero amplification and those which had irresolvable cluster patterns in a diverse set of *B. napus* lines, usually due to cohybridization of the A and C genomes. Application of the cluster file to a germplasm set will zero the statistics for those SNPs so they can be easily filtered out using the GenTrain score.

Although it is possible to simply export the genotype calls from GenomeStudio after eliminating poor markers by applying the cluster file, downstream analysis can greatly benefit from additional filtering within GenomeStudio prior to exporting the data. Apart from initial removal of zeroed SNPs, the GenomeStudio GenTrain score is of limited use for selecting high quality markers on the *Brassica* 60K array. However, other metrics can be employed for further analysis and filtering to generate a high quality data set, and these metrics are equally applicable to different software programs used for generating genotype calls from SNP array image files.

Poor samples can bias the SNP statistics, particularly when working with a small number of samples, and should be removed. Poor samples will often have low call rates and higher percentages of heterozygous calls compared to other samples, and will not cluster consistently or may have extreme normalized R values. Looking at sample call rates and raw fluorescence values rather than normalized data will identify samples with poor amplification, usually due to poor quality DNA: these samples should be removed. In addition, when working with doubled haploid or inbred mapping populations, lines with unexpectedly high heterozygote call rates should also be removed. The SNP statistics must then be recalculated based on the remaining samples before proceeding with additional filtering.

The SNP call rate can be used to identify markers where the positions for the AA, AB and BB genotypes defined by a cluster file do not match well with the samples being analyzed: these SNP markers will have low call rates and should be manually re-clustered or removed. Allele ratios can also be applied as a filtering metric prior to exporting the SNP data as genotype calls. For homozygous mapping populations, calculating allele frequencies can differentiate SNP markers as: (1) simple locus-specific SNPs with AA:BB (1:1) ratios; (2) potentially co-hybridising polymorphic SNPs with ratio of 1:2:1 for AA:AB:BB, which should be removed from further analyses; and (3) hemi-SNPs that can be scored as simple SNPs, providing they show a 1:1 ratio of AB:BB or AB:AA (or AA:BB, when repositioned by the cluster file into simple SNPs). Monomorphic or highly biased SNPs have a disproportionately high AA or BB frequency. Setting a minimum call rate and allele frequency is the easiest way to remove these SNP loci but it will also eliminate indel SNPs, which have a disproportionate level of no-calls. Dominant markers can be useful for genetic mapping, so in some cases it may be desirable to identify these markers and keep them as part of the data set (Fig. 2). The presence of hemi-SNPs and indel SNPs should be treated with care. If the frequency of their occurrence fits into expected ratios they can be used as markers for genomic rearrangements: if contiguous blocks of these marker types are present this could indicate a homoeologous exchange or a deletion/duplication event.

The mean theta values for the homozygous AA and BB clusters are useful in screening for single locus SNPs by selecting only those markers where the homozygous clusters have values close to 0 and 1, respectively. Combined with removal of SNPs with an unusually high number of heterozygotes (where appropriate) this is an easy and accurate way to eliminate SNPs that amplify more than one locus, usually the A and C genome homoeologues. The degree to which deviations from these values are tolerated can be adjusted to include more or fewer markers as needed.

The type of analyses and population being studied should always being considered when applying filters: one caveat of filtering by mean theta values is that extreme values will eliminate hemi-SNPs (regardless of repositioning by the cluster file), which may have value in homozygous mapping populations. Hemi-SNPs will have two genotype clusters (technically AABB and AAAA or AABB and BBBB) in a DH mapping population, but these may be output as either AA and AB, BB and AB or AA and BB depending on the clustering algorithm. Although these can be challenging to score since the occurrence of heterozygosity is a common criterion for eliminating SNP markers in such populations, as mentioned the cluster file for the *Brassica* 60K array was manually adjusted to allow a number of such loci to be accurately scored. Removal of multi-locus SNPs for F_2 populations or hybrid lines will be more complicated and probably depend on allele ratios and a very strict no-call threshold to eliminate markers with individuals between the AA, AB and BB clusters, i.e. the AAAB and ABBB groups.

Applying some or all of these filtering steps should produce a smaller, cleaner data set for initial genetic analysis. Following this, manual adjustment of the clusters for individual SNPs in regions of interest or areas with low marker density can be used to incorporate additional targeted loci into the data set.

Physical mapping of SNPs to the reference genome: identifying single-locus SNPs in polyploid *B. napus*

The availability of reference genome sequences for *B. napus* and its two diploid progenitors (Chalhoub et al. 2014; Liu et al. 2014; Parkin et al. 2014; Wang et al. 2011) allows for the possibility of not only positioning the SNP loci based on their genetic location, but also on their exact physical locale down to the nucleotide position of the variant. The ability to define a genomic context for a molecular marker can be invaluable when attempting to develop either a predictive tag for breeding or attempting to determine the causative gene for an agronomic trait. However, the duplicated nature of the *B. napus* genome, and in particular the strong sequence conservation between the constituent A and C diploid genomes, can easily confuse the physical placement of loci, which are defined by only a short probe sequence (50 bp). The SNP loci in the first instance were determined from short read sequence data aligned to the progenitor genomes. Hence, in theory this genomic point of origin should be considered the SNP physical location. The problem in defining these positions mirrors that which arises during the SNP assay. The assay is based on hybridization kinetics between the short probe sequence and the genome, whereas the SNP loci were derived from alignment of short read sequences to the genome. In some instances, multiple equally probable points of contact will occur between the probe and the genome, such that a single true genomic origin for the SNP cannot be faithfully distinguished. Thus, to provide a realistic physical position for the SNP locus, it is necessary to align the SNP probe sequence to the genome and find all possible regions of contact. This analysis identifies those loci that are more likely to provide genome specificity and which hence can be mapped as a single locus in genetic mapping projects.

The probes from the *Brassica* 60K Infinium™ array have been aligned to the *B. napus* genome using both BLASTN (Qian et al. 2014) and BLAT (Clarke et al. 2016) with somewhat similar results. The basic local alignment search tool BLAST (Altschul et al. 1990) has more flexibility when attempting to match sequences, whereas BLAT (Kent 2002), which is more computationally efficient, requires the resultant matches to be effectively identical or near-identical (Kent 2002; Morgulis et al. 2008). Qian et al. (2014) utilized stringent BLAST matching (zero mismatches) to align sequences of all the SNP assays (52,157), identifying 35,162 (67%) single-locus SNP loci. However, it should be noted that they removed 6930 SNPs from the analyses that had no identical hit in the reference *B. napus* genome. These latter loci could represent regions that were not assembled for *B. napus* but had been assembled for the diploid genomes. Alternatively, divergence between *B. napus* genotypes at these loci may have been sufficient to prevent stringent sequence alignment. BLAT alignment was used to determine the physical position of the probe sequences in two *B. napus* genomes: the winter type reference 'Darmor-*bzh*' (Chalhoub et al. 2014) and a spring type 'DH12075' (Parkin et al., unpublished)(Clarke et al. 2016). Based on a percentage identity of at least 85%, 50,255 SNPs were positioned in the spring type and 49,794 were positioned in the winter type genome sequence, equating to 51,172 SNPs that could be matched to one or both *B. napus* genomes. Of these 51,172 SNPs, based on the BLAT scores 22 258 and 23,191 SNPs could be unambiguously positioned in the A and C genomes respectively, while 2138 were placed on either the A or C genomes with equal probability. Additionally, 4570 SNPs could not be positioned on the pseudochromosomes as a result of either missing data in the *B. napus* reference genomes, or the alignment of the SNP sequence to an unanchored scaffold in one or both *B. napus* genomes.

Regardless of the method used to physically position the SNP loci in the *B. napus* genome, users of such information should always be aware that the definitive position of any locus can only be determined through genetic mapping. In some instances SNP loci mapped in one population can have an alternate position in a second population. This effect largely results from homoeologous co-hybridising SNP loci, where the two progenitor genomes are independently polymorphic in the two populations. A highly polymorphic population was used to genetically position 21,766 of the SNP loci from the array, and of these 7% did not map to the region expected based on sequence alignment of the probe sequence (Clarke et al. 2016). The conflict between physical and genetic position could be explained in 3% of the loci by mapping to the homoeologous position in the genome, while the remaining 4% may have either originated from unanchored *B. napus* genome scaffolds or be due to other problems with the reference genome assemblies. The current genome reference assemblies contain misassembled scaffolds and scaffolds incorrectly anchored to the genetic map: such errors are generally due to limitations of genetic anchoring, but are also somewhat inevitable in a genome prone to recombination events between homoeologous regions (Chalhoub et al. 2014; Osborn et al. 2003).

Selecting SNPs for different applications: the relevance of population structure and polymorphism in *B. napus* germplasm

Different criteria need to be applied to select SNPs for different applications. The strongly duplicated nature of the *B. napus* genome means that a large number of SNPs on the *Brassica* 60K SNP array do not behave in a truly locus-specific manner. Hence, for applications requiring a high degree of robustness we recommend the use of marker subsets which are filtered for single-copy BLAST hits in the *B. napus* reference genome. Depending on the population, a stringent BLAST search with zero mismatches and exactly one hit in *B. napus* Darmor-*bzh* v4.2 can be expected to recover approximately 20,000–30,000 polymorphic SNPs with a minor allele frequency (MAF) and call rate suitable for linkage disequilibrium (LD) or genome-wide association studies, genetic diversity analyses or other population genetics investigations (Hatzig et al. 2015; Qian et al. 2016, 2014). The advantage of this approach is that SNPs with a unique position can be ordered according to their chromosome positions in the reference genome, simplifying downstream analyses and providing more positional context to data analyses where linkage or LD relationships are of importance. For marker-based foreground and background selection during breeding the Brassica 60K SNP array provides a useful tool for pre-screening parental lines of breeding populations to identify single-locus SNPs that are (a) diagnostic for specific monogenic traits or major QTL, or (b) evenly distributed across chromosomes at pre-determined intervals. Based on the SNP flanking sequences, the former are highly amenable to conversion into single-marker assays (e.g. Taqman® or KASP®) for cost-efficient marker-assisted selection, whereas the latter can be used to assemble panels containing multiple, evenly-spaced SNPs per chromosome for recognition of background recombinants in a marker-assisted background selection program. Since the large numbers of SNPs provided by a high-density SNP array are generally not needed for marker-assisted selection, it is uncommon for breeders to implement the Brassica 60K SNP array directly for this purpose, but data from pre-screening with the array provides essential information for optimizing fast and cost-effective downstream SNP selection platforms in commercial breeding activities.

For genome-wide association analysis (GWAS) we recommend the use of single-copy SNPs with a minor allele frequency ≥0.05: detecting SNPs with small effect sizes (odds ratio <2) theoretically requires population sizes of >580 individuals when using a minor allele frequency of 0.05 (Gorlov et al. 2008), and false positives at this minor allele frequency may still be relatively common (Tabangin et al. 2009). As mentioned above, one downside of this increased stringency approach (Qian et al. 2014) is the loss of potentially useful SNPs from the analysis (perhaps up to 7000), and the reliance of this method on the accuracy of the Darmor-*bzh* v4.2 reference genome.

Due to the original genotypes from which the SNPs on the *Brassica* 60K SNP array were identified, genotypes from different *B. napus* gene pools can behave somewhat differently in terms of their polymorphism rates or SNP presence-absence. Hence, some SNP panels may be more useful than others for any given study. As an example, the *Brassica* 60K SNP array was used to assay and explore genome-wide diversity and LD patterns in large, similarly-sized panels of winter-type ($n=181$), spring-type ($n=186$) and semi-winter type oilseed rape (n = 186) (Hatzig et al. 2015; Jan et al. 2016; Qian et al. 2014). In these eco-geographically divergent gene pools we identified quite different groups of single-copy, polymorphic, non-heterozygous SNPs that show MAF ≥0.05 and ≤0.03 missing marker calls in each gene pool. Figure 3 and Supplementary Table 1 provide numbers and lists of polymorphic SNPs, respectively, which meet these criteria within and between these three main ecogeographical gene pools of oilseed rape.

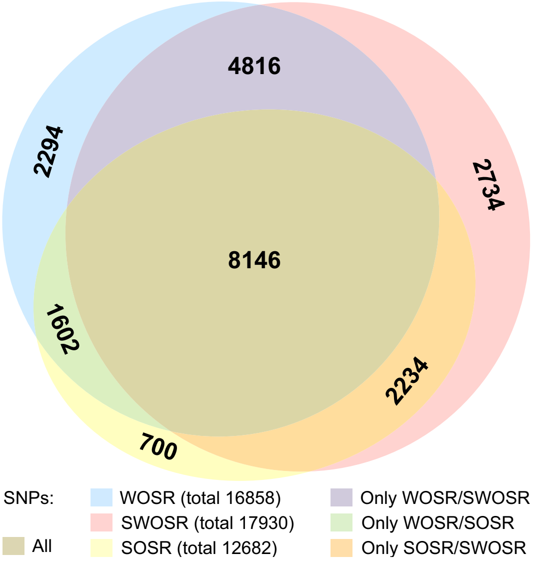

Fig. 3 Proportional Venn diagram describing numbers of single-copy SNPs from the *Brassica* 60K SNP genotyping array showing polymorphism (MAF ≥0.05) in diversity collections representing the three main gene pools of oilseed *Brassica napus*. Only 22,526 SNPs showing a unique BLAST hit to the Darmor-*bzh* reference *B. napus* genome, with zero mismatches of 50 bp flanking oligonucleotide sequence, were considered as single-copy loci for this analysis. The three *colored circles* represent homozygous accessions from genotyped panels of European winter oilseed rape (WOSR; $n=181$, data from Schiessl et al. 2015), Chinese semi-winter oilseed rape (SWOSR; $n=186$, data from Qian et al. 2014) and spring oilseed rape (SOSR; $n=186$; data from Jan et al. 2016). Numbers show the quantity of SNPs showing polymorphism levels above the 0.05 threshold within or between each of the individual pools

Use of these ordered SNP panels, selected for a gene pool of immediate interest, is a useful way to pre-screen for informative SNPs to be used in mapping or GWAS analyses within a specific germplasm collection. Dedicated marker panels can be more appropriate for population genetic analysis or other applications influenced by missing data, while the use of robust, preselected, single-copy SNPs for genetic mapping facilitates confident navigation from QTL confidence intervals to candidate sequences in the *B. napus* genome, for example. On the other hand, clusters of SNPs which show no polymorphism in a specific gene pool may be associated with selection signatures for ecogeographical adaptation (Voss-Fels and Snowdon 2015).

Identification of genomic rearrangements

SNP markers are codominant and hence useful for scoring alleles in F_2 mapping populations. However, doubled-haploid or recombinant inbred line mapping populations, where homozygous individuals are derived from an F_1 hybrid between two divergent homozygous parents, are more common in *B. napus* linkage mapping. Particularly valuable for genetic linkage mapping of *B. napus* populations is the fact that the *Brassica* 60K array provides a large panel of locus specific markers, which helps to interpret mapping results that otherwise would have been overlooked, such as deletions or homoeologous recombination events. A block of markers representing a deletion is more trustworthy than a single marker in a low-density map. In traditional marker systems such as RFLP, AFLP and SSR markers, scoring of indels and heterozygotes has been avoided to a large extent in order to prevent mapping errors. The *Brassica* 60K array, however, provides enough marker data to calculate dense genetic maps or to allow powerful association studies, and inclusion of indel SNPs and hemi-SNPs can add valuable information about specific genomic loci.

When calculating a genetic linkage map using a large marker set and including indel and hemi-SNPs, it is likely that blocks of markers sharing the same mode of variation or markers co-segregating with those from different genomic regions will be observed. These blocks preferentially occur between homoeologous chromosomes (Chalhoub et al. 2014) and could represent deleted or translocated regions in one of the mapping parents (Fig. 4). It has to be kept in mind that co-segregation of a small number of homoeologous markers may also be the result of non-locus-specific SNP marker amplification, or arise due to inappropriate alignment parameters, errors in the reference genome sequence or genotype-specific genomic variation. Hence, a minimal cut-off for the number of contiguous SNPs (at least 3–5) showing the same mode of inheritance is suggested to more accurately identify these regions. Furthermore, residual heterozygosity may potentially occur in one or both of the mapping parents and make precise SNP calling in the segregating population difficult. Here, manual screening for plausibility of individual SNP calls would be required. Unexpected linkage map orders should therefore be examined with high interest and the possibility of a causative genomic rearrangement should be tested, for example with resequencing or FISH technologies. Theoretically such patterns can also be called in genetically diverse genotype panels; however, care must be taken to distinguish between failed calls due to technical causes and genuine deletions. This can be achieved by independent validation, e.g. repeated genotyping or implementation of alternative datasets like genome sequence data (Schmutzer et al. 2015) or transcriptome sequences (Bancroft et al. 2011).

The inclusion of indel- and hemi-SNPs may result in over-estimation of genetic distances. Nonetheless, these high density genetic maps are still advantageous over those established with other marker systems, because physical positions of markers, rearrangements, QTLs and candidate genes are deducible with relative ease, and this can significantly speed up the process of map-based cloning.

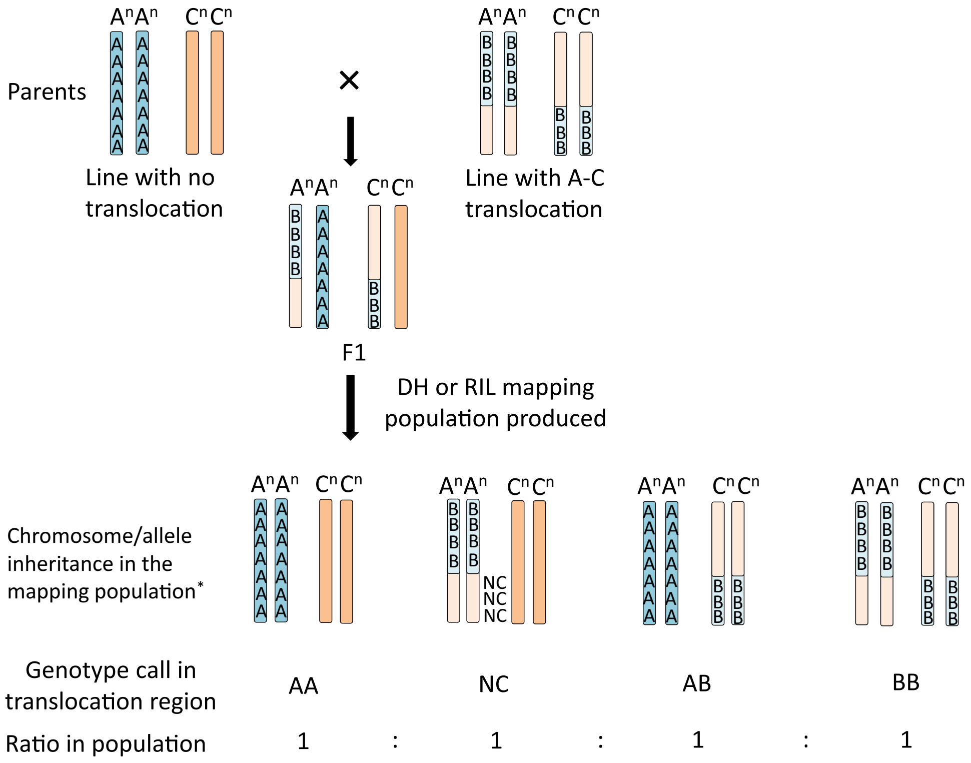
Parents
A^nA^n
C^nC^n
×
Line with no translocation
Line with A-C translocation
F1
DH or RIL mapping population produced
Chromosome/allele inheritance in the mapping population*
NC
NC
NC
Genotype call in translocation region
AA
NC
AB
BB
Ratio in population
1 : 1 : 1 : 1

Fig. 4 Cartoon showing A-genome SNP allele segregation for a single homoeologous A and C genome chromosome pair in a mapping population for which one parent has an A–C reciprocal translocation, showing origin of heterozygous and null allele calls for chromosome segments in progeny individuals. C genome chromosomes are in *orange* (different colors for alleles from different parents) and A genome chromosomes are in *blue* (*different colors* for alleles from different parents). *A* A allele call from parent with no translocation; *B* A allele call from parent with translocation. *NC* no call, failure to amplify any allele at that locus. *AB* amplification of both parental alleles at that locus. *Assuming no crossing over within the translocation region

Tracking genomic introgressions from related species

In the *Brassica* genus, production of interspecific or even intergeneric hybrids has proven to be a particularly useful method to introduce novel genetic diversity and traits into crop species. Previous examples of this method include introgression of resistance to *Sclerotinia* disease from *Erucastrum* and *Diplotaxis* species (Garg et al. 2010), resistance to blackleg (*Leptosphaeria maculans*) from *Brassica nigra* (Chevre et al. 1996), and introgression of resistance to the herbicide triazine from *B. rapa* (Beversdorf et al. 1980). Producing and selecting for these introgressions can be difficult. Producing lines which have a small introgression containing the target genomic region from the donor species, but which maintain an elite germplasm background over the majority of the genome, is desirable to maintain yield and other adapted traits. Hence, after the initial interspecific hybridization event to produce an F_1 hybrid, many rounds of backcrossing to the crop parent are generally undertaken, coupled with selection for the desired introgression trait. Molecular marker-assisted selection can and has been used to facilitate this process and to characterize subsequent introgressions (Barret et al. 1998; Chevre et al. 1997; Delourme et al. 1994; Saal and Struss 2005). Markers linked to the trait of interest in the wild parent reduce the need for phenotyping in every generation, while markers covering the rest of the crop genome show which lines have more or less introgressed segments in non-desirable parts of the genome.

Use of the *Brassica* 60K Illumina Infinium™ array to genotype lines going through this process has both advantages and disadvantages. Because the SNPs on the array are designed to specifically amplify the A and C genomes of *B. napus* there are unlikely to be markers that amplify only the introgressed genome fragments from other species (Mason et al. 2015). For example, although cross-amplification to the *Brassica* B genome does occur in a small subset of markers, this tends to result only in multi-locus SNPs that also amplify one or more A- or C-genome loci in addition to the B genome locus. The second problem is that the SNP array is not currently optimized for detection of dosage or copy number, so deletion of one copy of an A or C genome locus is not detectable. However, if the locus deleted (or replaced by a foreign chromosome segment) is homozygous (both copies are lost), it will present convincingly as a row of "no call" or failed amplification SNPs. The dense, genome-wide coverage of the array allows these regions to be very accurately delineated, a major advantage over other marker systems for tracking introgression events. Again to avoid incorrect identification of failed SNP amplification as introgression regions, at least three, preferably five consecutive SNPs should show the same pattern of missing values. Due to meiotic constraints, it is also unlikely that non-homologous recombination between the crop and donor species chromosomes will occur in an interspecific hybrid to produce genomic introgression regions smaller than a few Mbp, or that interstitial introgressions (within chromosomes, requiring two crossovers) rather than terminal introgressions (at the end of the chromosome) will occur (Mason et al. 2010). Deletions or introgressions present on only one of a pair of homologous chromosomes (i.e. heterozygous events) will not be detectable by the array.

As described previously, chromosome rearrangements between the A and C genomes can be observed using the SNP array. Such events are also expected to occur in greater numbers as a result of interspecific hybridization, as the A and C genomes are usually present as unpaired haploid genomes in the F_1 interspecific hybrid between *B. napus* and the target wild relative, which increases the probability of A-C exchanges (Nicolas et al. 2008). Synthetic *B. napus* is also meiotically unstable, and will present with many A–C exchanges in subsequent generations after formation (Song et al. 1995; Szadkowski et al. 2010). Distinguishing these types of A–C non-homologous recombination events from those due to introgression from the wild species is difficult using the SNP array: both look like strings of "no calls" or failed SNP amplification in one genome. Heterozygous calls may also occur as a result of a homoeologous translocation (more obviously in doubled-haploid or otherwise putatively homozygous lines) as two chromatids are now present that contain this genomic region (one on the A and one on the C genome) (Fig. 4). Full elucidation of genomic introgressions and A–C or other non-homologous recombination events is pending a reliable copy number pipeline that can definitively assess how many copies of each allele are present for each genomic region.

Identification of candidate genes underlying SNP loci

As described above, the *Brassica* 60K Illumina Infinium™ SNP array has become a popular tool for high-resolution genome-wide association studies in genetic diversity panels [e.g. (Hatzig et al. 2015; Li et al. 2016; Liu et al. 2016; Qian et al. 2014, 2016; Schiessl et al. 2015; Wei et al. 2016; Xu et al. 2016)]. With robust marker platforms like a SNP array such analyses provide a powerful tool for comparative QTL mapping and, ultimately, for discovery of potential candidate genes associated to traits of interest.

On the other hand, chromosomes of *B. napus* vary considerably in their patterns of sequence diversity and linkage disequilibrium (LD). This has a number of consequences for interpretation of genetic diversity data and significant SNP-trait associations. In all major *B. napus* gene pools, C-genome chromosomes exhibit considerably higher mean LD conservation than A-genome chromosomes, whereby different chromosomes in different gene pools are particularly affected by a lack of LD decay (Jan et al. 2016; Qian et al. 2014; Schiessl et al. 2015). It can therefore be extremely important to first determine the level of local LD surrounding the trait-associated SNPs (Qian et al. 2016; Voss-Fels and Snowdon 2015) when interpreting GWAS hits in *B. napus*. Chromosome regions with very rapid LD decay confidence intervals often represent recombination hotspots in which small intervals, with possibly only a few potential candidate genes, may show strong associations to a trait of interest. In contrast, if a GWAS hit occurs in a *B. napus* chromosome region with extremely conserved LD, there may be many hundreds or even thousands of genes with strong LD to the trait-associated SNPs. This can make it very difficult to determine potential candidate genes. In other species it has become common to set arbitrary distance thresholds for candidate gene searches, based on an average genome-wide LD decay rate across the whole genome. Adoption of this approach in *B. napus* [e.g. Körber et al. (2015); Bus et al. (2014)] carries the danger of potentially overestimating the true confidence interval in recombination-rich chromosome regions, or potentially strong underestimation in regions with little recombination.

Conclusions

The *Brassica* 60K Illumina Infinium™ array has revolutionized high-throughput genotyping in *Brassica napus, B. rapa* and *B. oleracea*. The availability of this resource has already been of inestimable use for diverse breeding and genetics applications. The SNP array provides a dense set of high-quality markers that can be physically mapped to the reference genomes; despite the high homoeology between the A and C genomes we have identified large subsets of SNPs that show distinct locus specificity within *B. napus* germplasm gene pools. In this review we have also discussed which SNPs should be excluded from analyses and how to clean the SNP data, different SNP types as well as how best to use the array for association and linkage mapping approaches and for tracking genomic introgressions. Possible pitfalls of the array include the high frequency of homoeologous, multi-locus SNPs, and a high false positive error rate in poor quality samples. However, these minor negatives are far outweighed by the high data quality, quantity, high-throughput nature and ease of analysis of this genotyping array.

Author contribution statement Conception and structure: ASM, EH and IAPP. Section writing: JB, EH, IAPP, AS, RJS, ASM. Figure S1, Figure 1—EH, Figure 2 and Figure 4—ASM, Table S1 and Figure 3—CW, Table 1—AS. All authors read and approved the final manuscript.

Acknowledgements ASM is supported by DFG Emmy Noether Grant MA6473/1–1.

Compliance with ethical standards

Conflict of interest The authors declare no conflicts of interest.

References

Alberts R, Terpstra P, Li Y, Breitling R, Nap JP, Jansen RC (2007) Sequence polymorphisms cause many false cis eQTLs. Plos One 2:e622

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Anithakumari AM, Tang JF, van Eck HJ, Visser RGF, Leunissen JAM, Vosman B, van der Linden CG (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Mol Breed 26:65–75

Bancroft I, Morgan C, Fraser F, Higgins J, Wells R, Clissold L, Baker D, Long Y, Meng JL, Wang XW, Liu SY, Trick M (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat Biotechnol 29:762–766

Barchi L, Lanteri S, Portis E, Acquadro A, Vale G, Toppino L, Rotino GL (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC Genomics 12:304

Barret P, Guerif J, Reynoird JP, Delourme R, Eber F, Renard M, Chevre AM (1998) Selection of stable *Brassica napus—Brassica juncea* recombinant lines resistant to blackleg (*Leptosphaeria maculans*). 2. A 'to and fro' strategy to localise and characterise interspecific introgressions on the *B. napus* genome. Theor Appl Genet 96:1097–1103

Batley J, Jewell E, Edwards D (2007) Automated discovery of single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular genetic markers. In: Edwards D (ed) Plant bioinformatics. Methods in molecular biology. Humana Press, USA, pp 473–494

Beversdorf WD, Weiss-Lerman J, Erickson LR, Souza Machado V (1980) Transfer of cytoplasmically-inherited triazine resistance from bird's rape to cultivated oilseed rape (*Brassica campestris* and *B. napus*). Can J Genet Cytol 22:167–172

Brookes AJ (1999) The essence of SNPs. Gene 234:177–186

Bus A, Korber N, Parkin IAP, Samans B, Snowdon RJ, Li JQ, Stich B (2014) Species- and genome-wide dissection of the shoot ionome in *Brassica napus* and its relationship to seedling development. Front Plant Sci 5:485

Chagné D, Bianco L, Lawley C, Micheletti D, Jacobs JME (2015) Methods for the design, implementation, and analysis of Illumina Infinium™ SNP assays in plants. In: Batley J (ed) Plant genotyping: methods and protocols. Springer, New York, pp 281–298

Chalhoub B, Denoeud F, Liu SY, Parkin IAP, Tang HB, Wang XY, Chiquet J, Belcram H, Tong CB, Samans B, Correa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao MX, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan GY, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VHD, Chalabi S, Hu Q, Fan CC, Tollenaere R, Lu YH, Battail C, Shen JX, Sidebottom CHD, Wang XF, Canaguier A, Chauveau A, Berard A, Deniot G, Guan M, Liu ZS, Sun FM, Lim YP, Lyons E, Town CD, Bancroft I, Wang XW, Meng JL, Ma JX, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou YM, Hua W, Sharpe AG, Paterson AH, Guan CY, Wincker P (2014) Early allopolyploid evolution in the post-neolithic *Brassica napus* oilseed genome. Science 345:950–953

Cheng F, Mandakova T, Wu J, Xie Q, Lysak MA, Wang XW (2013) Deciphering the diploid ancestral genome of the mesohexaploid *Brassica rapa*. Plant Cell 25:1541–1554

Chevre AM, Eber F, This P, Barret P, Tanguy X, Brun H, Delseny M, Renard M (1996) Characterization of *Brassica nigra* chromosomes and of blackleg resistance in *B. napus* - *B. nigra* addition lines. Plant Breed 115:113–118

Chevre 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

Clarke WE, Higgins EE, Plieske J, Wieseke R, Sidebottom C, Khedikar Y, Batley J, Edwards D, Meng J, Li R, Lawley CT, Pauquet J, Laga B, Cheung W, Iniguez-Luy F, Dyrszka E, Rae S, Stich B, Snowdon RJ, Sharpe AG, Ganal MW, Parkin IAP (2016) A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. Theor Appl Genet. doi:10.1007/s00122-016-2746-7 **(in press)**

Dalton-Morgan J, Hayward A, Alamery S, Tollenaere R, Mason A, Campbell E, Patel D, Lorenc M, Yi B, Long Y, Meng J, Raman R, Raman H, Lawley C, Edwards D, Batley J (2014) A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes. Funct Integr Genom 14:643–655

Delourme R, Bouchereau A, Hubert N, Renard M, Landry BS (1994) Identification of RAPD markers linked to a fertility restorer gene for the Ogura radish cytoplasmic male sterility of rapeseed (*Brassica napus* L). Theor Appl Genet 88:741–748

Delourme R, Falentin C, Fomeju BF, Boillot M, Lassalle G, Andre I, Duarte J, Gauthier V, Lucante N, Marty A, Pauchon M, Pichon JP, Ribiere N, Trotoux G, Blanchard P, Riviere N, Martinant JP, Pauquet J (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. BMC Genom 14:120

Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D (2009a) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res 37:951–953

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

Edwards D, Batley J, Snowdon R (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126:1–11

Fletcher RS, Mullen JL, Heiliger A, McKay JK (2015) QTL analysis of root morphology, flowering time, and yield reveals trade-offs in response to drought in *Brassica napus*. J Exp Bot 66:245–256

Fu D, Mason AS, Xiao M, Yan H (2015) Effects of genome structure variation, homeologous genes and repetitive DNA on polyploid crop research in the age of genomics. Plant Sci 242:37–46

Ganal MW, Polley A, Graner EM, Plieske J, Wieseke R, Luerssen H, Durstewitz G (2012) Large SNP arrays for genotyping in crop plants. J Biosci 37:821–828

Garg H, Atri C, Sandhu PS, Kaur B, Renton M, Banga SK, Singh H, Singh C, Barbetti MJ, Banga SS (2010) High level of resistance to *Sclerotinia sclerotiorum* in introgression lines derived from hybridization between wild crucifers and the crop *Brassica* species *B. napus* and *B. juncea*. Field Crops Research 117:51–58

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Hatzig SV, Frisch M, Breuer F, Nesi N, Ducoumau S, Wagner MH, Leckband G, Abbadi A, Snowdon RJ (2015) Genome-wide association mapping unravels the genetic control of seed germination and vigor in *Brassica napus*. Front Plant Sci 6. doi:10.3389/fpls.2015.00221

Hayward A, Mason AS, Morgan JD, Zander M, Edwards D, Batley J (2012) Special Issue: Reviews; SNP discovery and applications in Brassica napus. J Plant Biotechnol 39:49–61(식물생명공학회지)

Hodel RDGJ, Segovia-Salcedo MC, Landis JB, Crowl AA, Sun M, Liu XX, Gitzendanner MA, Douglas NNA, Germain-Aubrey CC, Chen SC, Soltis DE, Soltis PS (2016) The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. Appl Plant Sci 4:apps.1600025

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

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

Jan HU, Abbadi A, Lücke S, Nichols RA, Snowdon RJ (2016) Genomic prediction of testcross performance in canola (*Brassica napus*). PLoS One. doi:10.1371/journal.pone.0147769

Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12:656–664

Körber N, Bus A, Li J, Higgins J, Bancroft I, Higgins EE, Parkin IAP, Salazar-Colqui B, Snowdon RJ, Stich B (2015) Seedling development traits in *Brassica napus* examined by gene expression analysis and association mapping. BMC Plant Biol 15:136

Li G (2016) A new model calling procedure for Illumina Bead Array data. BMC Genet 17:90

Li F, Chen BY, Xu K, Wu JF, Song WL, Bancroft I, Harper AL, Trick M, Liu SY, Gao GZ, Wang N, Yan GX, Qiao JW, Li J, Li H, Xiao X, Zhang TY, Wu XM (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Res 21:355–367

Li F, Chen BY, Xu K, Gao GZ, Yan GX, Qiao JW, Li J, Li H, Li LX, Xiao X, Zhang TY, Nishio T, Wu XM (2016) A genome-wide association study of plant height and primary branch number in rapeseed (*Brassica napus*). Plant Sci 242:169–177

Liu LZ, Li JN (2014) QTL Mapping of oleic acid, linolenic acid and erucic acid content in *Brassica napus* by using the high density SNP genetic map. Sci Agric Sin 2014-01

Liu SY, Liu YM, Yang XH, Tong CB, Edwards D, Parkin IAP, Zhao MX, Ma JX, Yu JY, Huang SM, Wang XY, Wang JY, Lu K, Fang ZY, Bancroft I, Yang TJ, Hu Q, Wang XF, Yue Z, Li HJ, Yang LF, Wu J, Zhou Q, Wang WX, King GJ, Pires JC, Lu CX, Wu ZY, Sampath P, Wang Z, Guo H, Pan SK, Yang LM, Min JM, Zhang D, Jin DC, Li WS, Belcram H, Tu JX, Guan M, Qi CK, Du DZ, Li JN, Jiang LC, Batley J, Sharpe AG, Park BS, Ruperao P, Cheng F, Waminal NE, Huang Y, Dong CH, Wang L, Li JP, Hu ZY, Zhuang M, Huang Y, Huang JY, Shi JQ, Mei DS, Liu J, Lee TH, Wang JP, Jin HZ, Li ZY, Li X, Zhang JF, Xiao L, Zhou YM, Liu ZS, Liu XQ, Qin R, Tang X, Liu WB, Wang YP, Zhang YY, Lee J, Kim HH, Denoeud F, Xu X, Liang XM, Hua W, Wang XW, Wang J, Chalhoub B, Paterson AH (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun 5:3930

Liu J, Wang WX, Mei DS, Wang H, Fu L, Liu DM, Li YC, Hui Q (2016) Characterizing variation of branch angle and genome-wide association mapping in rapeseed (*Brassica napus* L.). Front Plant Sci 7:21

Lysak MA, Koch MA, Pecinka A, Schubert I (2005) Chromosome triplication found across the tribe *Brassiceae*. Genome Res 15:516–525

Mason AS, Huteau V, Eber F, Coriton O, Yan G, Nelson MN, Cowling WA, Chèvre A-M (2010) Genome structure affects the rate of autosyndesis and allosyndesis in AABC, BBAC and CCAB *Brassica* interspecific hybrids. Chromosome Res 18:655–666

Mason AS, Batley J, Bayer PE, Hayward A, Cowling WA, Nelson MN (2014) High-resolution molecular karyotyping uncovers pairing between ancestrally related *Brassica* chromosomes. New Phytol 202:964–974

Mason AS, Zhang J, Tollenaere R, Teuber PV, Dalton-Morgan J, Hu LY, Yan GJ, Edwards D, Redden R, Batley J (2015) High-throughput genotyping for species identification and diversity assessment in germplasm collections. Mol Ecol Resour 15:1091–1101

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

Morgan AP (2016) argyle: an R package for analysis of Illumina genotyping arrays. G3-Genes genomes. Genetics 6:281–286

Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA (2008) Database indexing for production Mega BLAST searches. Bioinformatics 24:1757–1764

Nicolas SD, Leflon M, Liu Z, Eber F, Chelysheva L, Coriton O, Chèvre AM, Jenczewski E (2008) Chromosome 'speed dating' during meiosis of polyploid *Brassica* hybrids and haploids. Cytogenet Genome Res 120:331–338

Obermeier C, Friedt W (2015) Applied oilseed rape marker technology and genomics. In: Poltronieri P, Hong Y (eds) Applied Plant Genomics and Biotechnology. Elsevier, Heidelberg, pp 253–295

Osborn TC, Butrulle DV, Sharpe AG, Pickering KJ, Parkin IA, Parker JS, Lydiate DJ (2003) Detection and effects of a homeologous reciprocal transposition in *Brassica napus*. Genetics 165:1569–1577

Parkin IAP, Gulden SM, Sharpe A, Lukens L, Trick M, Osborn TC, Lydiate DJ (2005) Segmental structure of the *Brassica napus* genome based on comparative analysis with *Arabidopsis thaliana*. Genetics 171:765–781

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL, Denoeud F, Belcram H, Links MG, Just J, Clarke C, Bender T, Huebert T, Mason AS, Pires JC, Barker G, Moore J, Walley PG, Manoli S, Batley J, Edwards D, Nelson MN, Wang X, Paterson AH, King G, Bancroft I, Chalhoub B, Sharpe AG (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15:R77

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

Qian LW, Qian W, Snowdon RJ (2014) Sub-genomic selection patterns as a signature of breeding in the allopolyploid *Brassica napus* genome. BMC Genomics 15

Qian L, Qian W, Snowdon RJ (2016) Haplotype hitchhiking promotes trait coselection in *Brassica napus*. Plant Biotechnol J 14:1578–1588

Qu CM, Li SM, Duan XJ, Fan JH, Jia LD, Zhao HY, Lu K, Li JN, Xu XF, Wang R (2015) Identification of candidate genes for seed glucosinolate content using association mapping in *Brassica napus* L. Genes 6:1215–1229

Ritchie ME, Carvalho BS, Hetrick KN, Tavare S, Irizarry RA (2009) R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. Bioinformatics 25:2621–2623

Ritchie ME, Liu R, Carvalho BS, ANZgene, Irizarry RA (2011) Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics 12:68

Saal B, Struss D (2005) RGA- and RAPD-derived SCAR markers for a *Brassica* B-genome introgression conferring resistance to blackleg in oilseed rape. Theor Appl Genet 111:281–290

Schiessl S, Samans B, Huttel B, Reinhard R, Snowdon RJ (2014) Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species *Brassica napus*. Front Plant Sci 5:404

Schiessl S, Iniguez-Luy F, Qian W, Snowdon RJ (2015) Diverse regulatory factors associate with flowering time and yield responses in winter-type *Brassica napus*. BMC Genomics 16:737

Schmutzer T, Samans B, Dyrska E, Lespinasse D, Micic Z, Abel S, Duchscherer P, Breuer F, Abbadi A, Leckband G, Snowdon RJ, Scholz U (2015) Species-wide genome sequence and nucleotide polymorphism datasets from the model allopolyploid plant *Brassica napus*. Sci Data 2:150072

Schranz ME, Lysak MA, Mitchell-Olds T (2006) The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci 11:535–542

Song KM, Lu P, Tang KL, Osborn TC (1995) Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc Natl Acad Sci USA 92:7719–7723

Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942

Szadkowski E, Eber F, Huteau V, Lodé M, Huneau C, Belcram H, Coriton O, Manzanares-Dauleux MJ, Delourme R, King GJ, Chalhoub B, Jenczewski E, Chèvre AM (2010) The first meiosis of resynthesized *Brassica napus*, a genome blender. New Phytol 186:102–112

Tabangin ME, Woo JG, Martin LJ (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc 3(Suppl 7):S41

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

Voss-Fels K, Snowdon RJ (2015) Understanding and utilizing crop genome diversity via high-resolution genotyping. Plant Biotechnol J 14:1086–1094

Wang XW, Wang HZ, Wang J, Sun RF, Wu J, Liu SY, Bai YQ, Mun JH, Bancroft I, Cheng F, Huang SW, Li XX, Hua W, Wang JY, Wang XY, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu BH, Li B, Liu B, Tong CB, Song C, Duran C, Peng CF, Geng CY, Koh CS, Lin CY, Edwards D, Mu DS, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang HB, Wang HP, Belcram H, Zhou HL, Hirakawa H, Abe H, Guo H, Wang H, Jin HZ, Parkin IAP, Batley J, Kim JS, Just J, Li JW, Xu JH, Deng J, Kim JA, Li JP, Yu JY, Meng JL, Wang JP, Min JM, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao MX, Jin MN, Ramchiary N, Drou N, Berkman PJ, Cai QL, Huang QF, Li RQ, Tabata S, Cheng SF, Zhang S, Zhang SJ, Huang SM, Sato S, Sun SL, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li YR, Du YC, Liao YC, Lim Y, Narusaka Y, Wang YP, Wang ZY, Li ZY, Wang ZW, Xiong ZY, Zhang ZH (2011) The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 43:1035–1039

Wei L, Jian H, Lu K, Filardo F, Yin N, Liu L, Qu C, Li W, Du H, Li J (2016) Genome-wide association analysis and differential expression analysis of resistance to *Sclerotinia* stem rot in *Brassica napus*. Plant Biotechnol J 14:1368–1380

Xu L, Hu K, Zhang Z, Guan C, Chen S, Hua W, Li J, Wen J, Yi B, Shen J, Ma C, Tu J, Fu T (2016) Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (*Brassica napus* L.). DNA Res 23:43–52

Zhang J, Mason AS, Wu J, Liu S, Zhang XC, Luo T, Redden R, Batley J, Hu LY, Yan GJ (2015) Identification of putative candidate genes for water stress tolerance in canola (*Brassica napus*). Front Plant Sci 6:1058