

High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*

Philipp E. Bayer · Pradeep Ruperao · Annaliese S. Mason · Jiri Stiller ·
Chon-Kit Kenneth Chan · Satomi Hayashi · Yan Long · Jinling Meng · Tim Sutton ·
Paul Visendi · Rajeev K. Varshney · Jacqueline Batley · David Edwards

Received: 18 December 2014 / Accepted: 24 February 2015 / Published online: 10 March 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract

Key message We characterise the distribution of crossover and non-crossover recombination in *Brassica napus* and *Cicer arietinum* using a low-coverage genotyping by sequencing pipeline SkimGBS.

Abstract The growth of next-generation DNA sequencing technologies has led to a rapid increase in sequence-based genotyping for applications including diversity assessment, genome structure validation and gene–trait association.

Communicated by L. Jiang.

All scripts used in this study are available at <http://www.appliedbioinformatics.com.au/index.php/SkimGBS>.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-015-2488-y) contains supplementary material, which is available to authorized users.

P. E. Bayer · P. Ruperao · A. S. Mason · C.-K. K. Chan ·
S. Hayashi · P. Visendi · J. Batley · D. Edwards (✉)
School of Agriculture and Food Sciences, University
of Queensland, Brisbane 4072, Australia
e-mail: dave.edwards@uwa.edu.au

P. E. Bayer · P. Ruperao · C.-K. K. Chan · P. Visendi · D. Edwards
Australian Centre for Plant Functional Genomics, School
of Agriculture and Food Sciences, University of Queensland,
Brisbane, Australia

P. E. Bayer · R. K. Varshney · J. Batley · D. Edwards
School of Plant Biology, University of Western Australia,
Perth, Australia

P. Ruperao · R. K. Varshney
International Crops Research Institute for the Semi-Arid Tropics
(ICRISAT), Hyderabad, Telangana, India

A. S. Mason · S. Hayashi · J. Batley
Centre for Integrative Legume Research,
University of Queensland, Brisbane, Australia

We have established a skim-based genotyping by sequencing method for crop plants and applied this approach to genotype-segregating populations of *Brassica napus* and *Cicer arietinum*. Comparison of progeny genotypes with those of the parental individuals allowed the identification of crossover and non-crossover (gene conversion) events. Our results identify the positions of recombination events with high resolution, permitting the mapping and frequency assessment of recombination in segregating populations.

Introduction

Recent years have seen a massive rise in the volumes of data generated by next-generation sequencing technologies (NGS), which now form the basis of most genome studies, ranging

J. Stiller
Commonwealth Scientific and Industrial Research Organisation
(CSIRO) Plant Industry, Queensland Bioscience Precinct,
Brisbane, QLD, Australia

Y. Long · J. Meng
National Key Laboratory of Crop Genetic Improvement,
Huazhong Agricultural University, Wuhan, China

T. Sutton
South Australian Research and Development Institute,
GPO Box 397, Adelaide, SA 5001, Australia

T. Sutton
Australian Centre for Plant Functional Genomics, School
of Agriculture, Food and Wine, University of Adelaide, PMB1,
Glen Osmond, SA 5064, Australia

from the assembly of draft genome sequences to genome diversity analysis. Genomic information is becoming increasingly available for *Brassica* and *Cicer* species. Information on a public *B. rapa* genome was published in 2011 (Wang et al. 2011) and of two genomes of *B. oleracea* was published recently (Liu et al. 2014; Parkin et al. 2014) together with that of *B. napus* (Chalhoub et al. 2014). Draft references of both kabuli and desi *Cicer arietinum* genomes were also published in 2013 (Jain et al. 2013; Varshney et al. 2013). The availability of these reference genomes enables the discovery of sequence-based molecular markers and their association with agronomic traits for applied crop improvement (Edwards and Batley 2010; Edwards et al. 2013; Hayward et al. 2012b).

Recombination is one of the major sources of genetic variation, shuffling sets of genes to produce novel allelic combinations. Both reciprocal exchange between homologous chromosomes (crossover) and non-reciprocal exchange (non-crossover) can occur, and both are initiated by the repair of double-strand breaks (DSBs) in DNA during meiosis, reviewed by Chen et al. (2007). A subset of non-crossover events referred to as gene conversions result in fragments from homologous regions in the partner chromosome being used as a template for DSB repair (Mezard et al. 2007). The positioning of crossovers is well known in many plant species through cytological observation and through recombination mapping, but the frequency of gene conversion is mostly unknown in crops (Gaut et al. 2007). Recombination events in plants have frequently been genetically mapped using molecular markers (Farkhari et al. 2011; Yao et al. 2002). These studies showed variation in recombination frequencies for the same recombination bins across different populations. In general, pericentromeric regions showed the lowest frequency of recombination, and telomeric regions showed the highest frequency of recombination.

One study used 13,551 SRAP markers to produce a recombination map for *B. napus* (Westar × Zhonyou 821). The study identified 1663 crossovers in 58 double-haploid lines, which corresponds to 1.51 crossovers per individual per chromosome (Sun et al. 2007). Other studies used genotyping by sequencing (GBS) using restriction site-associated genomic DNA (RAD tags) to produce genetic maps. In maize and barley, an early GBS approach was able to map 200,000 and 25,000 sequence tags (Elshire et al. 2011). Another approach used two different restriction enzymes to reduce genomic complexity and was able to map 34,000 SNPs in barley and 20,000 SNPs in wheat (Poland et al. 2012).

The distribution of recombination has been mapped in *A. thaliana*, where 40 F2 individuals from lines Columbia and Landsberg erecta were resequenced to assess the distribution of crossovers and gene conversions (Yang et al. 2012). This study identified more than 3000 gene conversions and 73 crossovers per plant. Interestingly, the majority (72.6 %) of smaller crossover blocks (10–500 kb) were found in pericentromeric regions,

while larger crossover blocks were distributed evenly among all chromosomes. A follow-up study repeated the analysis, but under the assumption that genomic rearrangements confound the mapping of short reads (Wijnker et al. 2013). The researchers removed all markers near putative duplicated regions to counter the errors introduced by mismapped reads, leading to an estimated one to three gene conversions and ten crossovers per meiosis, a much smaller number than presented in Yang et al. (2012). A recent study hypothesised that the large number of gene conversions found in *Arabidopsis* in (Yang et al. 2012) are due to false-positive SNPs caused by copy-number variation, mismappings due to duplicated regions caused by transposable elements and tandem repeats (Qi et al. 2014). After removal of these erroneous SNPs, 11 crossovers could be identified in two plants, which is equal to an average of 1.1 crossovers per chromosome. A smaller number of non-crossovers could be identified with five non-crossovers in one plant and one non-crossover in the other.

One of the limitations of most forms of genotyping is that only a restricted set of the total SNPs are assayed across a population. While this is efficient for the identification of major recombination events used for genetic mapping, the resolution is fixed by the restriction site density. With the decreasing cost of NGS data generation and the increasing availability of reference genomes, it is becoming cost-effective to generate whole genome sequence data for GBS applications. To this end, we have developed a novel GBS approach called skim-based genotyping by sequencing (skimGBS) which uses low-coverage whole genome sequencing for high-resolution genotyping. We demonstrate the application of this approach to genotype a double-haploid (DH) canola population derived from a cross between the cultivars Tapidor and Ningyou 7, as well as a population of *C. arietinum* recombinant inbred lines (RILs). Using this approach, it was possible to generate genome-wide recombination maps and to assess and compare the frequency of crossover and gene conversion events. We hypothesise that inflated numbers of crossovers and gene conversions are caused by errors in the reference assembly, and that the correct number of gene conversions and crossovers should be similar to those reported by Wijnker et al. (2013) and Qi et al. (2014).

Materials and methods

SkimGBS is a two-stage method that requires a reference genome sequence, genomic reads from parental individuals and individuals of the population. Firstly, the parental reads are mapped to the reference genome and SNPs are called using SGSautoSNP (Lorenc et al. 2012). Subsequent mapping of the progeny reads to the same reference and comparison with the parental SNP file enables the calling of the parental genotype. According to the SGSautoSNP protocol, read data were not trimmed or filtered.

For *B. napus*, two reference sequences relating to the *B. napus* diploid progenitors were used for mapping reads: the A-genome (Wang et al. 2011) and the C-genome (Parkin et al. 2014). The *Brassica* population consisted of 92 double-haploid Tapidor × Ningyou 7 individuals from the TNDH mapping population previously described (Qiu et al. 2006) (Supplementary Table 1). The *C. arietinum* population consisted of 46 PI489777 × ICC4958 F9-F10 RILs (Supplementary Table 2) (Gaur et al. 2012) and reads were aligned to the published kabuli reference genome (Varshney et al. 2013). Both parental and offspring reads were aligned using SOAPaligner v2.21 (Li et al. 2009), using only reads that map uniquely (setting: ‘-r 0’, maximum mismatch: standard of 2) with a generous insert size (0–1000). Only reads aligning in pairs were used in subsequent analyses. The genomic reads of both populations have been deposited in the Short Read Archives and are collected in two BioProjects at <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA274890> and <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA274892>.

SNPs for the parental genomes were called using SGSautoSNP (Lorenec et al. 2012). A custom script (‘snp_genotyping_all.pl’) compared the progeny read alignments with parental genotypes to assign genotypes. SNPs that had more than 85 % missing alleles were removed with a simple Python script (‘RemoveEmptySNPs.py’). We implemented a simple method of sideways imputation (‘impute-FlapjackAlleles.py’), which assumes that recombinations do not occur. For example, if there are two missing alleles in between two Ningyou genotypes in an individual, then the two missing alleles are imputed with the Ningyou genotype.

In both populations, some individuals exhibited a much higher heterozygosity than the others in the population: 10 individuals in the *C. arietinum* population and 25 in the *B. napus* population. These were removed from all subsequent steps of analysis.

Gene conversion events have previously been defined as being shorter than 10 kb in length and longer than 20 bp (Yang et al. 2012). Additionally, we defined a gene conversion block as containing at least two SNPs. It follows from this definition that crossover events are longer than 10 kb. Crossovers and gene conversions that shared their start or endpoints within the resolution offered by the skimGBS data were removed using a custom script (‘fuzzyRecombinationFilter.py’). For each individual, the total number of gene conversions, crossover events and the number of nucleotides covered by these were counted, as well as the distribution of recombination and gene conversion events. The Shapiro–Wilk test, *t* test and Spearman’s rank correlation coefficient test were performed using R v3.0.1 using the functions `shapiro.test()`, `t.test()` and `cor()`. The distribution of recombination events was plotted using Python v2.7. The centromere positions for *Brassica* were derived (Cheng et al. 2013; Parkin et al. 2014).

Results

Brassica napus

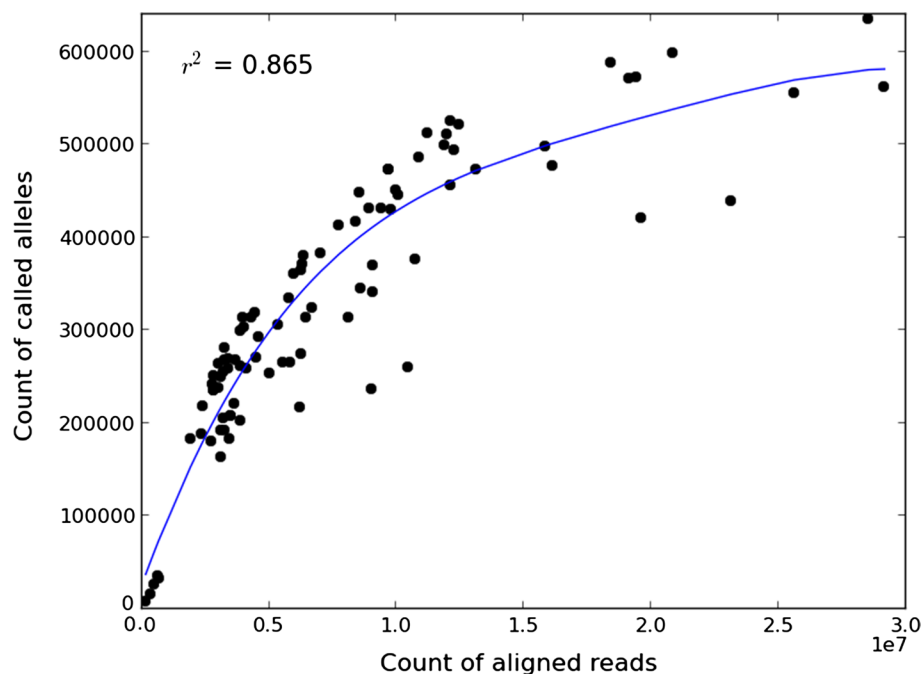
A total of 78.8 and 46.0 Gbp of whole genome sequence data, representing 69.6× and 40.6× genome coverage, were generated for the parental cultivars Tapidor and Ningyou 7. After mapping these reads to the *B. rapa* cv. Chiifu and *B. oleracea* cv. TO1000 draft genome assemblies using SOAPaligner (Li et al. 2009), SGSautoSNP (Lorenec et al. 2012) identified a total of 880,809 intervarietal SNPs. Of these, 840,264 (95 %) were distributed across the 19 pseudomolecules and the remaining 40,545 SNPs were located on unplaced contigs (see Table 1).

Illumina genome sequence data were generated for 92 individual progeny of a Tapidor × Ningyou 7 population, with an average coverage of 1.3× per individual and ranging from 0.1× to 7.36× (Supplementary Table 1). After read mapping to the reference, an average of 313,590.25 alleles could be called per individual, with a minimum of 7000 and a maximum of 602,133 alleles called per individual. The relationship between coverage and called SNPs in the population is shown in Fig. 1. This shows that with about 1.5 Gbp of coverage, the majority of SNPs are called.

Table 1 Predicted SNPs in *B. napus* between the cultivars Tapidor and Ningyou

Name	Total length	SNP count	SNPs/kbp	Transitions/transversions
C01	43,764,888	77,050	1.761	1.368
C02	52,886,895	37,911	0.717	1.336
C03	64,984,695	55,852	0.859	1.3
C04	53,719,093	79,794	1.485	1.346
C05	46,902,585	21,285	0.454	1.307
C06	39,822,476	33,729	0.847	1.323
C07	48,366,697	30,097	0.622	1.318
C08	41,758,685	36,288	0.869	1.329
C09	54,679,868	37,333	0.683	1.33
Unplaced C contigs	45,028,525	23,124	0.514	1.339
A01	26,743,757	38,432	1.437	1.261
A02	27,848,229	46,564	1.672	1.247
A03	32,230,199	54,180	1.681	1.213
A04	20,227,473	47,345	2.341	1.259
A05	23,942,034	43,674	1.824	1.269
A06	26,273,342	59,662	2.271	1.272
A07	22,305,923	32,768	1.469	1.25
A08	21,233,127	24,178	1.139	1.242
A09	37,197,712	56,757	1.526	1.28
A10	17,624,801	27,365	1.553	1.257
Unplaced A contigs	20,469,451	17,421	0.851	1.305

Fig. 1 Relationship between the number of called SNPs and number of aligned reads for each of the 92 *Brassica napus* DH individuals



An estimate was made of the frequency of miscalled alleles due to sequence errors. Across the 92 individuals, 0.065 % of called alleles (19,219) were different from either of the parental alleles and presumed to be due to sequence error. As these errors represent two possible non-parental alleles, we estimate the frequency that a sequence error calls the incorrect parental allele to be 0.032 % (1 in 3000).

Due to the low coverage of the population, for many SNPs only few alleles could be called in the population. After removal of SNPs with more than 85 % missing alleles, 794,837 SNPs (90.2 %) with an average number of alleles of 306,982 remained. Sideway imputation raised the total number of alleles in the population from 28,242,426 to 62,903,177, with an average per individual of 683,730.

After removal of the SNPs with high, too many missing alleles in the population, a very large number of crossovers and gene conversions could be identified (see Supplementary Table 3). There was a wide variation in heterozygosity between individuals and 25 individuals with high heterozygosity were removed from further analysis.

After the initial analysis, the A-genome exhibited on average 151.18 crossovers and 697.85 gene conversions per individual. In the A-genome, TN80 exhibited the smallest number of gene conversions (76), TN21 exhibited the smallest number of crossovers (85), TN98 showed the highest number of gene conversions (19,047), and TN100 showed the highest number of crossovers (536).

Similarly, the C-genome exhibited on average 115.53 crossovers and 374.85 gene conversions per individual. TN80 exhibited the smallest number of gene conversions

(93), TN21 the smallest number of crossovers (37), TN65 the highest number of gene conversions (1628), and TN100 the highest number of crossovers (337). Close examination of these results suggested that many were due to structural differences, either due to differences in the reference cultivar to Tapidor and Ningyou, or due to misassemblies in the reference genomes, and so a filter was applied to remove all overlapping gene conversions and crossovers. A total of 16,943 crossovers and 70,984 gene conversions from the 67 individuals were removed, with more gene conversions and crossovers removed from the A-genome than from the C-genome.

From the A-genome, 9655 crossovers (95.32 %) and 46,245 gene conversions (98.91 %) were removed compared to removal of 7288 crossovers (94.15 %) and 24,649 gene conversions (98.14 %) from the C-genome. The difference in crossovers and gene conversions between the A-genome to the C-genome was statistically significant (two-tailed Student's *t* test, crossovers $p < 0.001$, gene conversions $p < 0.00001$).

After filtering, we identified 927 crossovers, 13.84 per individual or 0.7 per chromosome. This ranged from 0 in TN2 to 249 in TN100. In addition, we identified 977 gene conversions, 14.58 per individual, or 0.76 per chromosome and individual. These ranged from 10 in TN18 to 20 in TN7 (see Supplementary Table 4). An overview of chromosome A1 before and after filtering is presented in Figs. 2 and 3.

After filtering, chromosome C4 had the highest number of crossovers per individual, ranging from 0 to 40, with an average of 1.42, whilst chromosome C5 had the lowest average number of crossovers at 0.28. The number of

Fig. 2 Recombination map for *Brassica napus* chromosome A1 before filtering of overlapping recombinations. *Red* genotype Tapidor, *blue* genotype Ningyou, *white* missing. Each *line* is one individual; from *top* to *bottom*: TN9, TN99, TN98, TN97, TN94, TN93, TN90, TN8, TN89, TN88, TN87, TN86, TN85, TN83, TN82, TN80, TN7, TN78, TN76, TN75, TN74, TN73, TN65, TN5, TN57, TN54, TN4, TN48, TN47, TN46, TN45, TN44, TN43, TN42, TN41, TN40, TN3, TN39, TN38, TN37, TN36, TN35, TN34, TN32, TN31, TN30, TN2, TN29, TN28, TN27, TN26, TN25, TN24, TN22, TN21, TN20, TN1, TN19, TN18, TN17, TN16, TN15, TN14, TN12, TN11, TN10 and TN100 (colour figure online)

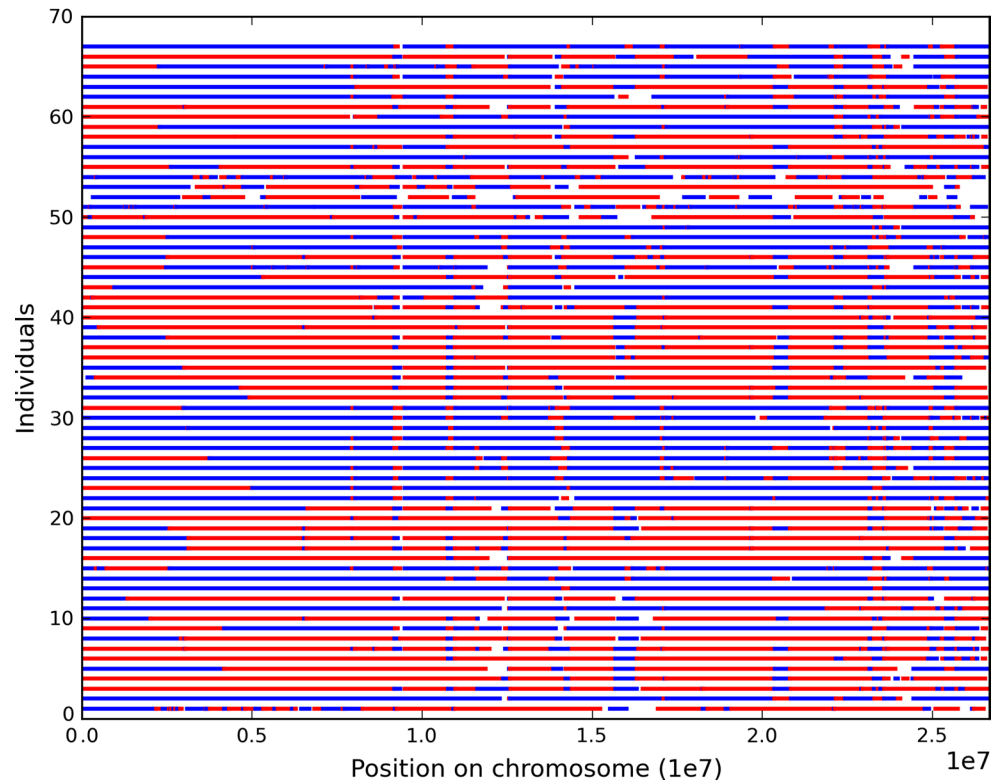
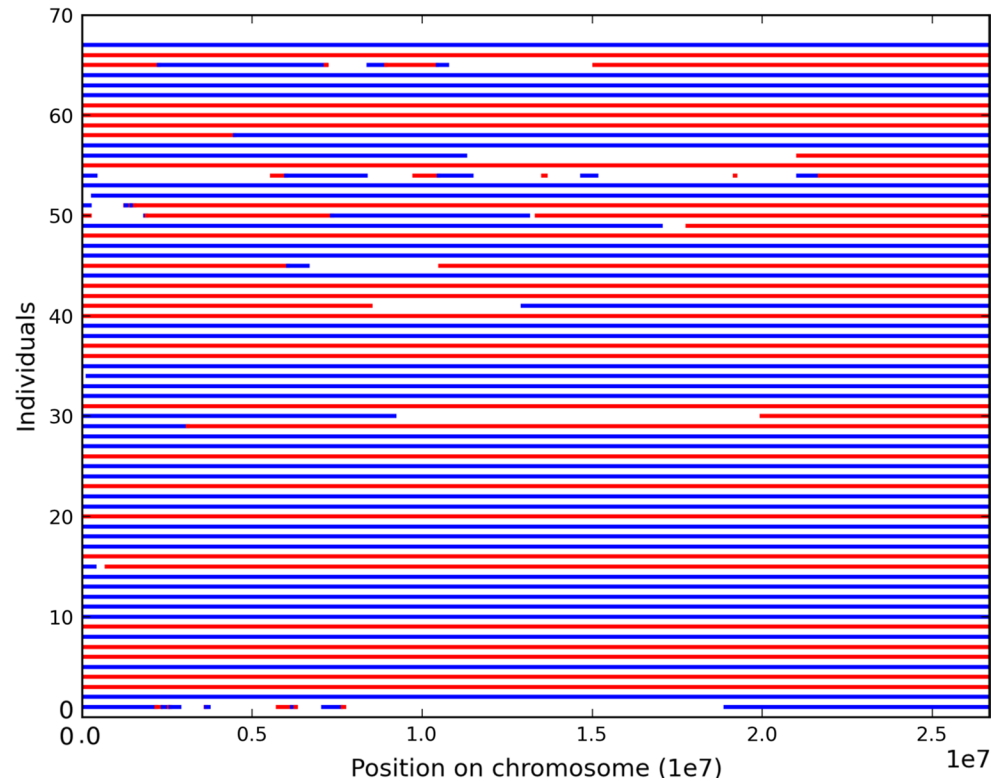


Fig. 3 Recombination map for *Brassica napus* chromosome A1 after filtering of overlapping recombinations. *Red* genotype Tapidor, *blue* genotype Ningyou, *white* missing. Each *line* is one individual; from *top* to *bottom*: TN9, TN99, TN98, TN97, TN94, TN93, TN90, TN8, TN89, TN88, TN87, TN86, TN85, TN83, TN82, TN80, TN7, TN78, TN76, TN75, TN74, TN73, TN65, TN5, TN57, TN54, TN4, TN48, TN47, TN46, TN45, TN44, TN43, TN42, TN41, TN40, TN3, TN39, TN38, TN37, TN36, TN35, TN34, TN32, TN31, TN30, TN2, TN29, TN28, TN27, TN26, TN25, TN24, TN22, TN21, TN20, TN1, TN19, TN18, TN17, TN16, TN15, TN14, TN12, TN11, TN10 and TN100 (colour figure online)



gene conversions per chromosome was very similar with all chromosomes carrying zero to two gene conversions (see Supplementary Table 5). After filtering, there was no

difference in crossovers or gene conversions when comparing the A-genome with the C-genome (two-tailed Student's *t* test, crossovers $p > 0.7$, gene conversions $p > 0.6$). The

distribution of crossovers on chromosomes was plotted (Supplementary Figs. 1–19).

Cicer arietinum

A total of 7.2 and 5.9 Gbp of sequence data were generated from the two *C. arietinum* cultivars PI489777 (wild-type) and ICC4958, which represent an estimated coverage of 9.7× and 7.9×, respectively (Supplementary Table 2). A total of 555,346 SNPs were identified using SGSautoSNP (Lorenc et al. 2012), of which 448,619 (80.7 %) were distributed over the eight chromosomes and 106,727 were located on unplaced contigs. A total of 20.9 Gbp of Illumina paired read sequence data were generated for 46 progeny individuals (between 0.13× and 1.54×, with an average of 0.58×). Mapping these reads to the reference led to between 37,444 (RIL12) and 268,431 (RIL43) called SNPs, with an average of 147,363 per individual. 43,722 SNPs had too many missing alleles (>85 %) and were discarded (Table 2). Sideway imputation added 15,221,101 alleles, leading to a total of 21,999,801 called alleles, an average of 478,256 per individual.

Out of a total of 6,778,700 called alleles, 6440 (0.09 %) exhibited heterozygosity, with 10 individuals exhibiting high heterozygosity. These individuals were removed from subsequent analyses.

Crossovers and gene conversions were predicted following the same approach as for *Brassica*. Prior to filtering, crossovers totalled 3960, an average of 110 per individual, while gene conversions totalled 4675, or 129.86 per individual. After filtering, the number of gene conversions ranged from 5 (RIL4) to 22 (RIL29), and crossovers ranged from 0 (RIL7) to 60 (RIL29). There were 219 crossovers and 256 gene conversions (see Supplementary Table 6). An overview of chromosome 1 before and after filtering is presented in Supplementary Figs. 20 and 21.

Table 2 Predicted SNPs in *C. arietinum* between accessions PI4958 and ICC489777

Name	Length	SNP count	SNPs/Kbp	Transitions/transversions
Ca1	48,359,943	62,154	1.28	1.681
Ca2	36,634,854	45,836	1.25	1.732
Ca3	39,989,001	59,818	1.49	1.687
Ca4	49,191,682	69,229	1.40	1.645
Ca5	48,169,137	63,406	1.31	1.7
Ca6	59,463,898	75,323	1.26	1.698
Ca7	48,961,560	52,550	1.07	1.664
Ca8	16,477,302	20,303	1.23	1.561
Unplaced contigs	186,473,055	106,727	0.57	1.726

After filtering, out of the eight *C. arietinum* chromosomes, chromosome 6 had the highest average number of crossovers with 1.44 and chromosome 3 had the lowest average number of crossovers with 0.19 (ranging from 0 to 3 in individuals). The number of gene conversions varied over the chromosomes, from an average of 0.69 in chromosome 7 to 1.0 in chromosome 5 (see Supplementary Table 7). The distribution of all crossovers was plotted (Supplementary Figs. 22–29).

Discussion

Here, we present the application of a skim-based genotyping by sequencing (skimGBS) method in *B. napus* and *C. arietinum* to assess the frequency and distribution of recombination. SGSautoSNP has been previously used to successfully predict SNPs in *B. napus* with an accuracy of >95 % (Hayward et al. 2012a) and in wheat with an accuracy of 93 % (Lai et al. 2014). By combining this SNP discovery method with skimGBS, we can assess the segregation of SNPs in a population. All scripts for the skimGBS pipeline are available at: <http://www.appliedbioinformatics.com.au/index.php/SkimGBS>.

We demonstrate that SkimGBS can be used to genotype a greater number of SNPs than previous approaches in these species. For example, in *B. napus* without imputation we genotyped an average of 147,363 SNPs per individual compared to 2604 using RAD-based GBS (Bus et al. 2012). SkimGBS was able to call more SNPs than earlier GBS approaches (Elshire et al. 2011; Poland et al. 2012), since there is no genomic complexity reduction steps in the SkimGBS pipeline.

The relatively high rates of sequence error found in next-generation DNA sequence data are a potential source of genotype miscalling. We estimate that 0.032 % of SNPs (one in 3000 of genotypes) are erroneously called in our analysis due to sequence error. Erroneous SNPs may also be predicted due to mismapping of reads to the reference genome. Due to the low coverage of the population, for about 10 % of the SNPs more than 85 % of the alleles in the population were missing. We removed these SNPs from further steps as these may have a negative impact on the sideway imputation.

We used only non-repetitively aligning reads to minimise the number of SNPs from homeologous regions (Lai et al. 2014; Lorenc et al. 2012). As we require two adjacent SNPs to call a gene conversion and need both SNPs to be at least 20 bp apart, we estimate the frequency of miscalled gene conversions due to sequence error to be negligible. We observed that some individuals in the *B. napus* population had a relatively high frequency of heterozygous alleles. This was unexpected as the population was produced as

double haploids and so should be homozygous. We expect that the heterozygous individuals were due to pollen flow between lines during population development, and so these individuals were removed from the analysis.

Due to the low coverage of the sequence-based genotyping, some alleles were not called and so we used side-way imputation to predict these missing alleles, increasing the average number of alleles from 306,982 to 683,730 per individual in *Brassica*. While imputation allows for improved visualisation of haplotype blocks, it is not required to determine haplotype blocks or recombination events. There was weak to moderate correlation between the number of aligned reads and number of both crossovers (-0.22) and gene conversions (-0.54) (Supplementary Tables 8 and 9), suggesting that the majority of recombination events were captured. There was a high correlation (0.81) between the number of aligned reads and the number of heterozygous SNPs for an individual. This is due to the fact that higher coverage is required to observe a heterozygous SNP. For a heterozygous SNP to be observed, at least two reads have to align to the locus, and due to the low coverage of skimGBS many heterozygous SNPs may be missed.

Initial results suggested that gene conversions outnumbered crossovers in *B. napus* and *C. arietinum*, with the ratio of gene conversions to crossovers similar to that observed in *Arabidopsis* by Yang et al. (2012). A subsequent paper by Wijnker et al. (2013) suggested that small genomic rearrangements may lead to false high counts of gene conversion events. After filtering to remove genotypes around potentially rearranged regions, the number of gene conversions and crossovers in our study reduced to levels observed in *Arabidopsis* by Wijnker et al. (2013). After filtering, an average of 0.76 gene conversions and 0.73 crossovers per individual and per chromosome were detected in *B. napus*. The number of crossover and gene conversion events per meiosis differs between species due to various factors, and is dependent on the number and size of chromosomes present. Also, only 50 % of the total recombination events occurring in the F1 meiosis can be detected by progeny testing, as recombination events occur between only two of the four chromatids comprising a homologous chromosome pair, and only one chromatid is subsequently retained in gamete production. However, these results are in the same range as the 1–3 gene conversions and two crossovers per meiosis (or 0.2–0.6 gene conversions and 2 crossovers per chromosome per individual) detected by Wijnker et al. (2013) and to the average of one crossover per chromosome and 0.2–1 gene conversion per chromosome reported by Qi et al. (2014) using similar methods.

The RIL population of *C. arietinum* exhibited a similar number of crossovers and gene conversions to *B. napus*. One individual in the population had a much higher number

of gene conversions than the rest of the population, leading to the possibility that this individual skewed the average. Some chromosomes show a greater abundance towards the telomeres, but others exhibit a more even distribution (Supplementary Figs. 22–29).

In both populations there are individuals that after filtering exhibit either a much higher non-crossover rate (RIL29 in the *Cicer* population) or crossover rate (TN100 in the *Brassica* population) than the rest of the population. It could be that these were missed in the filtering step of non-homozygous individuals, or that the reads for these individuals were actually from several different individuals, leading to a larger number of recombinations.

Interestingly, we observed a difference in erroneously called recombination events between the three genomes used as references in this study, with more errors in the *Brassica* A-genome than the C-genome, and fewer again in the *C. arietinum* genome. This corresponds with genome assembly quality and likelihood of misassembled regions. The *Brassica* diploid genomes are highly complex, sharing a whole genome triplication (Liu et al. 2014; Parkin et al. 2014; Wang et al. 2011), and the recent assembly of the *Brassica* C-genome (Parkin et al. 2014) is of greater quality than the A-genome assembly which was published 3 years earlier (Wang et al. 2011). While the *C. arietinum* genome reference carries some misassembled regions (Ruperao et al. 2014), this relatively simple genome, with no recent genome duplications, and produced using the latest sequencing chemistry and assembly methods, is likely to have fewer misassembled regions than the *Brassica* genomes.

Previous studies suggest that lower numbers of recombination events occur around centromeres and a greater number of crossover events occur towards telomeres (Farkhari et al. 2011; Helms et al. 1992; Roberts 1965). In human genomes, DSBs and recombination hotspots exhibit specific sequence motifs: for example, polypurine and polypyrimidine tracts are overrepresented in regions of gene conversions (Chen et al. 2007). In *A. thaliana*, recombination hotspots seem to be biased towards regions with a high AT content, located away from methylated DNA and carrying at least two distinct sequence motifs (Wijnker et al. 2013). Other studies such as that by Drouaud et al. (2013) showed distinct recombination hotspots and related the results to proteins such as MSH4.

In addition to predicted recombination, we observed regions of the genome which demonstrated an alternative haplotype structure compared to the surrounding regions across all individuals. These regions reflect major differences in structure between the reference genomes used for read mapping and the genomes of the sequenced population. While these positions were removed from the analysis of recombination in this study, they offer the potential

to validate genome structural assemblies and characterise differences in genome structure at a high resolution. Due to the early draft status of both genome annotations (compared to the highly validated annotations for *A. thaliana*), we did not compare the distribution of recombination to genetic content.

This study demonstrates high-resolution skimGBS in two important crops and identifies gene conversion and crossover recombination with high precision. The skim GBS approach is flexible, with relatively little data required for trait association, while increasing the volume of sequence data enables fine mapping of recombination events, detailed characterisation of gene conversions as well as the potential to validate genome assemblies and identify structural variations. The continued decline in the cost of generating genome sequence data should lead to an increase in the application of skimGBS for crop improvement.

Author contribution statement PEB and PR ran the pipeline and analysed the results. PEB wrote parts of the SkimGBS pipeline and co-wrote the manuscript. ASM provided valuable critique and discussion of recombinations and their positions. JS and CKC wrote parts of the SkimGBS pipeline. SH, YL, JM, TS, PV and RV generated the DH and RIL populations, provided genetic material and genomic data and provided genome references. JB and DE conceived the study and co-wrote the manuscript.

Acknowledgments The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462, LP110100200 and DP0985953), the Australian India Strategic Research Fund (AISRF) Grand Challenge fund (GCF010013), CGIAR Generation Challenge Programme (Theme Leader Discretionary grant), the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) and the Center of Excellence in Genomics (CEG) of ICRISAT. The part of this work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of the CGIAR Consortium.

Conflict of interest The authors declare no conflict of interest.

References

- Bus A, Hecht J, Huettel B, Reinhardt R, Stich B (2012) High-throughput polymorphism detection and genotyping in *Brassica napus* using next-generation RAD sequencing. *BMC Genom* 13:281
- Chalhoub B et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345:950–953. doi:10.1126/science.1253435
- Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8:762–775. doi:10.1038/nrg2193

- Cheng F, Mandakova T, Wu J, Xie Q, Lysak MA, Wang X (2013) Deciphering the diploid ancestral genome of the Mesohexaploid *Brassica rapa*. *Plant Cell* 25:1541–1554. doi:10.1105/tpc.113.110486
- Drouaud J et al (2013) Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. *PLoS Genet* 9:e1003922. doi:10.1371/journal.pgen.1003922
- Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J* 8:2–9. doi:10.1111/j.1467-7652.2009.00459.x
- Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126:1–11. doi:10.1007/s00122-012-1964-x
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi:10.1371/journal.pone.0019379
- Farkhari M, Lu Y, Shah T, Zhang S, Naghavi MR, Rong T, Xu Y (2011) Recombination frequency variation in maize as revealed by genomewide single-nucleotide polymorphisms. *Plant Breed* 130:533–539
- Gaur R et al (2012) High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (*Cicer arietinum* L.). *DNA Res* 19:357–373. doi:10.1093/dnares/dss018
- Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat Rev Genet* 8:77–84. doi:10.1038/nrg1970
- Hayward A, Mason A, Dalton-Morgan J, Zander M, Edwards D, Batley J (2012a) SNP discovery and applications in *Brassica napus*. *J Plant Biotechnol* 39:1–12
- Hayward A et al (2012b) Second-generation sequencing for gene discovery in the Brassicaceae. *Plant Biotechnol J* 10:750–759. doi:10.1111/j.1467-7652.2012.00719.x
- Helms C et al (1992) Closure of a genetic linkage map of human chromosome 7q with centromere and telomere polymorphisms. *Genomics* 14:1041–1054
- Jain M et al (2013) A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *Plant J* 74:715–729. doi:10.1111/tbj.12173
- Lai K et al (2014) Identification and characterization of more than 4 million intervarietal SNPs across the group 7 chromosomes of bread wheat. *Plant Biotechnol J* 13:97–104. doi:10.1111/pbi.12240
- Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967. doi:10.1093/bioinformatics/btp336
- Liu S et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat Commun* 5:3930. doi:10.1038/ncomms4930
- Lorenc M et al (2012) Discovery of single nucleotide polymorphisms in complex genomes using SGSautoSNP. *Biology* 1:370–382. doi:10.3390/biology1020370
- Mezard C, Vignard J, Drouaud J, Mercier R (2007) The road to crossovers: plants have their say. *Trends Genet* 23:91–99. doi:10.1016/j.tig.2006.12.007
- Parkin IA et al (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biol* 15:R77. doi:10.1186/gb-2014-15-6-r77
- Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:e32253. doi:10.1371/journal.pone.0032253
- Qi J, Chen Y, Copenhaver GP, Ma H (2014) Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc Natl Acad Sci USA* 111:10007–10012. doi:10.1073/pnas.1321897111

- Qiu D et al (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. doi:[10.1007/s00122-006-0411-2](https://doi.org/10.1007/s00122-006-0411-2)
- Roberts PA (1965) Difference in behaviour of Eu- and Hetero-chromatin—crossing-over. *Nature* 205:725–726. doi:[10.1038/205725b0](https://doi.org/10.1038/205725b0)
- Ruperao P et al (2014) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant Biotechnol J* 12:778–786. doi:[10.1111/pbi.12182](https://doi.org/10.1111/pbi.12182)
- Sun Z, Wang Z, Tu J, Zhang J, Yu F, McVetty PB, Li G (2007) An ultradense genetic recombination map for *Brassica napus*, consisting of 13551 SRAP markers. *Theor Appl Genet* 114:1305–1317. doi:[10.1007/s00122-006-0483-z](https://doi.org/10.1007/s00122-006-0483-z)
- Varshney RK et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat Biotechnol* 31:240–246. doi:[10.1038/nbt.2491](https://doi.org/10.1038/nbt.2491)
- Wang X et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* 43:1035–1039. doi:[10.1038/ng.919](https://doi.org/10.1038/ng.919)
- Wijnker E et al. (2013) The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* 2: e01426 doi:[10.7554/eLife.01426](https://doi.org/10.7554/eLife.01426)
- Yang S et al (2012) Great majority of recombination events in *Arabidopsis* are gene conversion events. *Proc Natl Acad Sci USA* 109:20992–20997. doi:[10.1073/pnas.1211827110](https://doi.org/10.1073/pnas.1211827110)
- Yao H, Zhou Q, Li J, Smith H, Yandea M, Nikolau BJ, Schnable PS (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proc Natl Acad Sci USA* 99:6157–6162. doi:[10.1073/pnas.082562199](https://doi.org/10.1073/pnas.082562199)