High resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in chickpea and canola

Authors

Philipp E. Bayer[1,2], Pradeep Ruperao[1,2,3], Annaliese Mason[1,4], Jiri Stiller, Chon-Kit Kenneth Chan, Satomi Hayashi, Yan Long, Jinling Meng, Tim Sutton, Paul Visendi, Rajeev K. Varshney[3], Jacqueline Batley[1,4], David Edwards[1,2]

1 School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia

2 Australian Centre for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia

3 International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, Andhra Pradesh, India

4 Centre for Integrative Legume Research, University of Queensland, Brisbane, Australia

4 School of Plant Biology, University of Western Australia, Perth, Australia

Corresponding author:

David Edwards: dave.edwards@uq.edu.au

Tel: +61 (0)7 3346 7084 Fax: +61 (0) 7 3365 1176

School of Agriculture and Food Sciences, The University of Queensland, Brisbane 4072, Australia

Running title: Gene conversions in *Brassica* and Chickpea

Keywords: Genotyping by sequencing, Genomics, SNP calling, Genome structure, *Brassica*, Chickpea, *Cicer*

Word count: 6,685

Summary

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. With the reducing cost of sequence data generation and the increasing availability of reference genomes, there are opportunities to develop high density genotyping methods based on whole genome re-sequencing. We have established a skim-based genotyping by sequencing method for crop plants and have applied this approach to genotype segregating populations of *Brassica napus* canola and chickpea. By comparing progeny genotypes with those of the parental individuals, it is possible to identify crossover and gene conversion events at high resolution. Our results identify the positions of recombination events with unprecedented resolution and support recent studies on the frequency of recombination in *Arabidopsis*.

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 from the assembly of draft genome sequences to genome diversity analysis. Genomic information is becoming increasingly available for *Brassica* and chickpea species. Proprietary genome sequences for *B. napus* and its diploid progenitor species were produced in 2009 (http://www.brassicagenome.net/), a public *B. rapa* genome was published in 2011 (Wang et al., 2011), the genome of *B. oleracea* (CC) was published recently (Liu et al., 2014; Parkin et al., 2014), and the genome of *B. napus* (AACC) is expected to become available in the next 12 months. Draft references of both kabuli and desi chickpea genomes were also published in 2013 (Jain et al., 2013; Varshney et al., 2013b). 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).

Single nucleotide polymorphisms (SNPs) now dominate genome analysis. International collaborations have facilitated the discovery and publication of large numbers of canola SNPs, and the recent production of a 60,000 SNP Illumina Infinium genotyping array has improved the ability to genotype this species. SNP discovery in canola has increasingly applied NGS data. Durstewitz et al. identified 604 SNPs using 100 EST-based amplicons in *B. napus* (Durstewitz et al., 2010). Similarly, Hu et. al. identified 655 SNPs from

201,190 Roche 454 ESTs derived from two *B. napus* cultivars ZY036 and 51070 (Hu et al., 2012), while Trick et. al. identified 41,593 SNPs between the *B. napus* cultivars Tapidor and Ningyou 7 based on 20 million Illumina RNA-Seq reads (Trick et al., 2009).

Several studies have identified SNPs in *Cicer arietinum*. Nayak et al. used ESTs to identify 71 SNPs between *C. arietinum* and *Medicago truncatula* and applied these SNPs together with SSRs to generate a genetic map comparing *C. arietinum* to *M. truncatula* (Nayak et al., 2010). Another study generated Illumina NGS data for two chickpea genotypes ICC4958 and ICC1882 and aligned these reads with a transcript assembly in order to identify 4,543 SNPs (Azam et al., 2012). Hiremath et al. used 4 data types (Illumina sequencing, amplicon sequencing, ESTs, candidate gene sequencing) to identify 2,486 chickpea SNPs and applied these to generate a genetic map (Hiremath et al., 2012). A further study developed an Illumina GoldenGate SNP genotyping assay for the cultivars ICC4958 and a wild progenitor *C. reticulatum* PI489777. A total of 768 high-confidence SNPs were identified, of which 697 could be successfully genotyped (Gaur et al., 2012). Lastly, Agarwal et al. sequenced transcripts of kabuli chickpea and aligned Illumina reads from kabuli, desi and wild populations with 454 transcripts (Wijnker et al., 2013). Using this method, they identified 1,986 SNPs between kabuli and desi, and 37,954 SNPs between kabuli and wild germplasm.

Genotyping by sequencing (GBS) is a generic name for a range of methods which use next generation DNA sequencing (NGS) data for the genotyping of SNPs in populations. There are several methods applied in GBS, but most use

some form of reduced representation to decrease the quantity of DNA sequence data required. A common GBS method, restriction site associated DNA (RAD) sequencing, is a reduced complexity approach which samples regions of the genome, identifies SNPs and calls genotypes. RAD-sequencing, and its modified forms, use restriction enzymes to cut the DNA at specific sites in the genome. The ligation of adapter DNA fragments permits the amplification and sequencing of the short regions surrounding the restriction sites and the discovery and genotyping of DNA polymorphisms in these regions.

RAD based GBS has the benefit that it does not require a reference genome; however it has been demonstrated to lead to some bias in the genotyping results. This includes the non-random distribution of restriction sites across genomes, and allele drop-out due to the presence of a polymorphism in the restriction site itself (Gautier et al., 2012). The bias introduced by the chosen restriction enzymes can partially be alleviated by using combinations of restriction enzymes (Hohenlohe et al., 2012), however some bias remains due to variation in restriction site densities across genomes. RAD-sequencing has been applied in *Lolium perenne* to generate a genetic map and identify Quantitative Trait Loci (QTLs) (Hegarty et al., 2013). In eggplant, RAD-sequencing was able to call 10,000 SNPs, 1,600 InDels and 1,800 putative SSRs (Barchi et al., 2011). In *Hordeum vulgare* RAD sequencing was applied to identify 530 SNPs of which 445 were used to generate a genetic map (Chutimanitsakun et al., 2011). Truong et al. used RAD sequencing to identify 1,409 SNPs in *A. thaliana* and 5,583 SNPs in lettuce, while in *B. napus,* (Bus et al., 2012) identified 20,835 SNPs and 125 InDels from 636,179 RAD tags

grouped in 113,221 RAD clusters. This last study has been, to our knowledge, the only RAD GBS study performed in *B. napus*.

Recombination is the major source of genetic variation, through the shuffling of sets of genes to produce novel allelic combinations. Both crossover and non-crossover based recombination is observed in crop genomes and both are initiated by double strand breaks (DSBs) during meiosis. Cells repair DSBs in two ways: either by reciprocal exchange of DNA between homologous chromosome arms (crossovers) or by non-reciprocal exchange of genetic material (non-crossovers). A subset of non-crossovers are gene conversions, where fragments from homologous regions in the partner chromosome are used as a template to connect the two pieces of chromosome, thus fixing the DSB (for a review, see (Mezard et al., 2007)).

The extent and positions of crossovers is well-known in many plant species, but the frequency of gene conversion is mostly unknown in crops (Gaut et al., 2007). Recombination events in plants have been mapped using genetic markers. In maize, recombination has been mapped in regions around four target genes (Yao et al., 2002) and over the entire genome using linkage maps (Farkhari et al., 2011). An older study used restriction fragment length polymorphisms (RFLPs) to map recombination in *B. nigra* (wild Catania x CrGC no. 2, 88 individuals) (Lagercrantz and Lydiate, 1995). There are several studies investigating recombination in *B. napus*. One of the earliest studies used RFLPs to map crossovers and non-crossovers in 4 different double haploid populations (RV289, TO1147, MF216, RV128), identifying 99 gene conversions and 9 crossovers (Udall et al., 2005). Nicolas et al. (Nicolas et al.,

2007) used molecular markers to study three pairs of homeologous regions (N1/N11, N3/N13 and N9/N18) in haploid x euploid *B. napus* cultivars Darmor-bzh and Yudal, and found that homoeologous recombination is greatly enhanced in haploids compared to euploids, suggesting that naturally occurring haploids are a source of chromosomal rearrangements. Another study used 13,551 SSRs to create a recombination map for *B. napus* (Westar x Zhonygou 821). The recombination map identified 1,663 crossovers in 58 double haploid lines (Sun et al., 2007).

Among plants, the distribution of gene conversions has only been mapped in *A. thaliana* (Yang et al., 2012), where they re-sequenced 40 F2 individuals from the *A. thaliana* lines Columbia and Landsberg erecta to compare the distribution of crossovers and gene conversions with the two parental individuals. This study identified more than 3000 gene conversions and 73 crossovers per plant. Interestingly, the majority (72.6%) of smaller crossover blocks (10kb to 500kb) 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 re-arrangements 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. This way, they estimated 1-3 gene conversions and 10 crossovers per meiosis, a much smaller number than presented in (Yang et al., 2012). One of the limitations of most forms of genotyping, including RAD GBS, is that only a restricted set of the total SNPs are assayed across a population. While this is efficient for the identification of major crossover recombinations

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. Towards this, 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 canola population derived from a cross between the cultivars Tapidor and Ningyou 7, as well as a chickpea RIL population. 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 hypothesize that an inflated number of crossovers and gene conversions is caused by errors in the reference assembly, and after stringent filtering, the numbers of gene conversions and crossovers are similar to that reported by (Wijnker et al., 2013).

Results

Brassica napus

A total of 78.8 and 46.0 Gbp of whole genome sequence data, representing 69.6x and 40.6x genome coverage, was generated for the parental cultivars Tapidor and Ningyou 7. After mapping these reads to the *B. rapa* and *B. oleracea* draft genomes assemblies using SOAPaligner (Li et al., 2009), SGSautoSNP (Lorenc et al., 2012) identified a total of 880,809 intervarietal SNPs, 840,264 (95%) of which were distributed across the 19

pseudomolecules, and the remaining 40,545 SNPs located on unplaced contigs (see Table 1). For the relationship between coverage and called alleles, see figure 1.

Illumina genome sequence data was generated for 92 individual progeny of a Tapidor x Ningyou 7 population, with an average coverage of 1.3x per individual and ranging from 0.1x to 7.36x. For details of individual coverage, see table S1. After mapping these reads to the reference, an average of 634,114 alleles could be called per individual, with a minimum of 13,776 and a maximum of 1,210,781 alleles called per individual.

An estimate was made of the frequency of miscalled alleles due to sequence errors. Across the 92 individuals, 48,017(0.082%) of called alleles 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 0.041% (one in 2400).. There was a wide variation in heterozygosity between individuals and 25 individuals demonstrating high heterozygosity were removed from further analysis.

Out of the 880,809 SNPs, 96,268 did not segregate in the population and were removed. After removal of these monomorphic SNPs, a very large number of crossovers and gene conversions could be identified (see Supplementary Table 3)

Before filtering, the A genome exhibited on average 152 crossovers and 655 gene conversions per individual. TN80 exhibited the smallest number of gene

conversions (72), TN21 exhibited the smallest number of crossovers (80), conversely TN98 showed the highest number of gene conversions (1837), and TN82 showed the highest number of crossovers (522).

The C genome exhibited on average 128.64 crossovers and 353.62 gene conversions per individual. TN80 exhibited the smallest number of gene conversions (91), TN21 exhibited the smallest number of crossovers (43), and TN65 showed the highest number of gene conversions (1530), and TN100 showed the highest number of crossovers (344).Close examination of these results suggested that many were due to structural differences in the reference genomes and a further filter was applied to remove all overlapping gene conversions and crossovers (see methods). This removed a total of 17,950 crossovers and 66,6054 gene conversions from the 67 individuals. More gene conversions and crossovers were removed from the A-genome than from the C-genome. From the A-genome, 10,190 crossovers (95.49%) and 43,391 gene conversions (98.76%) were removed, while from the C-genome, 8,619 crossovers (95.35%) and 23,214 gene conversions (97.97%) were removed.

After filtering, we identified 2187 crossovers, 32.6 per individual or 1.7 per chromosome. This ranged from 0 in TN2 to 234 in TN100. In addition, we identified 1021 gene conversions, 15.23 per individual, or 0.80 per chromosome and individual. These ranged from 11 in TN2 to 19 in TN38 (see Supplementary Table 4). For an overview of chromosome A1 before and after filtering, see figures 2 and 3.

Chickpea

A total of 7.2 Gbp and 5.9 Gbp was generated or the two chickpea cultivars PI489777 (wild-type) and ICC4958, which represents an estimated coverage of 9.7x and 7.9x. SGSautoSNP identified 555,346 SNPs, of which 448,619 were distributed over the 8 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 (minimum: 0.13x, maximum: 1.54x, average: 0.58x). Mapping these reads to the reference led to an average 144,836 called alleles per individuals (minimum: 36,648 in RIL12, maximum: 262,308 in RIL43). A subset of SNPs (39,590, 7%) were monomorphic in this population and discarded, leaving 515,756 SNPs (see

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	20,469,451	17,421	0.851	1.305

contigs				

Table **2**). Out of a total of 6,662,458 called alleles, 223,746 (3.3%) exhibited heterozygosity, with 10 individuals exhibiting high heterozygosity. These were removed from subsequent analyses.

Crossovers and gene conversions were predicted following the same approach as for *Brassica*, and the results presented in Table **2**.

Before filtering, crossovers totalled 3737, an average of 103.8 per individual, while gene conversions totalled 4200 or 116.67 per individual. After filtering, the number of gene conversions ranged from 4 in RIL18 to 20 in RIL29, and crossovers ranged from 0 in RIL16 to 54 in RIL29. The number of crossovers totalled 200, and the number of gene conversions totalled 246 (see Supplementary Table 5). For an overview of chromosome 1 before and after filtering, see figures 3 and 4.

Discussion

Here we present the application of a skim-based genotyping by sequencing (SkimGBS) method to *B. napus* and chickpea populations to assess the frequency and scale of recombination. SGSautoSNP has been previously demonstrated to predict SNPs in *B. napus* with an accuracy of >95% (Hayward et al., 2012a). By combining this SNP discovery method with genotyping, we can assess the segregation of SNPs in a population. A total of 7 and 10% of SNPs were monomorphic in *C. arietinum* and *B. napus* respectively and

subsequently removed from the analysis, 2-5% more than expected. There is the possibility that not all of these SNPs have been erroneously predicted. Since we confirmed these SNPs with a low-coverage population it could be that some of the removed SNPs are located in regions which are underrepresented in the sample of population reads. Since we cannot distinguish between false negatives and true negatives in this case, we removed both.

SkimGBS was able to genotype a greater number of SNPs than previous approaches in these species. For example, in *B. napus* we called an average of 328,950 alleles per individual compared to 2,604 genotyped using RAD Seq (Bus et al., 2012). The relatively high rates of sequence error found in next generation DNA sequence data is a potential source of genotype miscalling. We estimate that 0.041% (one in 2400 of genotypes in our analysis are erroneously called due to sequence error. As we require two adjacent SNPs to call a gene conversion and need both SNPs to be at least 20bp 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 number 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 during population development and so these individuals were removed from the analysis. Due to very low coverage in some individuals, it could be that the DH-population contains more heterozygous individuals than observed – in these individuals there may have been not enough reads aligning to call the number of heterozygous alleles required for filtering.

Due to the low coverage of the sequence based genotyping, many alleles were not called and so we used sideways imputation to predict these missing alleles, more than doubling the number of average number alleles from 303,336 to 738,309 per individual in *Brassica*. While imputation allows for improved visualization of haplotype blocks, imputation is not required to determine haplotype blocks and gene conversion events. There was relatively low correlation between the number of aligned reads and number of both crossovers (-0.33) and gene conversions (-0.55) (Supplementary Table 6 and 7) suggesting that we were able to capture the majority of recombination events, and that not all SNPs need to be genotyped. There was a high correlation (0.88) between the number of aligned reads and the number of heterozygous alleles for an individual. For a heterozygous allele to be observed at least two reads have to align to the locus and due to the low coverage of skimGBS, many heterozygous alleles may be missed. However, as heterozygosity is likely to occur in regions it may be possible to collate information from several adjacent SNPs to define a region of heterozygosity.

Following cleaning of monomorphic SNPs and imputation of genotypes, we were able to predict the frequency and positions of gene conversions and crossovers in the population. Initial results suggested that gene conversions outnumbered crossovers in *B. napus*, with a frequency similar to that observed by Yang et al. (2012) in *Arabidopsis*. A subsequent paper by Wijnker et al (2013) suggested that small genomic re-arrangements 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

reduced to levels observed in Arabidopsis by Wijnker et al (2013). After filtering, *B. napus* exhibits an average of 0.93 gene conversions and 1.72 crossovers per individual and chromosome, and *C. arietinum* exhibits 0.85 gene conversions and 1.69 crossovers per individual and chromosome, very similar to the 1-3 gene conversions and 10 crossovers per meiosis (or 0.2-0.6 gene conversions and 2 crossovers per chromosome per individual) observed by Wijnker et al (2013). Interestingly, we observed a difference in erroneous recombination frequency 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 chickpea 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 assembly of the recent Brassica C genome is of greater quality than the A genome assembly which was published three years earlier(Parkin et al., 2014). While the chickpea genome reference is not perfect (Ruperao et al., 2014) this relatively simple genome, produced using the latest sequencing chemistry and assembly methods is likely to have fewer misassembled regions than the Brassica genomes.

There is also the possibility that the method presented here removes too many crossovers. This can only be alleviated by improving the reference assembly.

Previous studies suggest a greater number of crossover events towards telomeres. Areas that have a high frequency of gene conversion events but relatively few crossovers might exhibit a higher tendency to form double strand breaks (DSB). In human genomes, DSBs and recombination hot-spots exhibit

specific sequence motifs or by sequences capable of forming non-B DNA structures (Chen et al., 2007). In *A. thaliana*, recombination hotspots seem to be biased towards a high AT content and away from methylated DNA, and carry at least two distinct sequence motifs (Wijnker et al., 2013).

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.

This study demonstrates for the first time high resolution skim GBS in two important crops and identified 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, the 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 is likely to lead to an increase in the application of GBS for crop improvement.

Materials and Methods

SkimGBS is a two stage method that requires a reference genome sequence and 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.

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 X Ningyou 7 individuals from the TNDH mapping population previously described (Qiu et al., 2006) (see Supplementary Table 1 for a full overview). The chickpea population consisted of 46 double-haploid PI489777 x ICC4958 individuals (see Supplementary Table 2) and reads were aligned to the published kabuli reference genome (Varshney et al., 2013a). Both parental and offspring reads were aligned using SOAPaligner v2.21 (Li et al., 2009), using only reads that map uniquely (setting: '-r 0').

SNPs for the parental genomes were called using SGSAutoSNP (Lorenc et al., 2012). A custom script ('snp_genotyping_all.pl') compared the progeny read alignments with parental genotypes to assign genotypes. SNP positions that exhibited only one parental genotype (monomorphic) were removed using a custom Python script (scriptname.py). Gene conversion events (GCs) have previously been defined as being shorter than 10 kb in length and longer than

20 bp (Yang et al., 2012). Additionally, we define a gene conversion block to have at least 2 alleles. 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 ('fuzzy_recombination_filter.py') For each individual, the total number of gene conversions, crossover events and the number of nucleotides covered by these was counted, as well as the distribution of recombination and gene conversion events. The Shapiro-Wilk test and Spearman's rank correlation coefficient test were performed using R v3.0.1 using the functions shapiro.test() and cor(). The distribution of recombination events was plotted using Python v2.7.

References

Azam, S., Thakur, V., Ruperao, P., Shah, T., Balaji, J., Amindala, B., Farmer, A.D., Studholme, D.J., May, G.D., Edwards, D., Jones, J.D. and Varshney, R.K. (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (Cicer arietinum; Fabaceae), a crop species without a reference genome. *American journal of botany* **99**, 186-192.

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

Bus, A., Hecht, J., Huettel, B., Reinhardt, R. and Stich, B. (2012) High-throughput polymorphism detection and genotyping in Brassica napus using next-generation RAD sequencing. *Bmc Genomics* **13**.

Chen, J.M., Cooper, D.N., Chuzhanova, N., Ferec, C. and Patrinos, G.P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nature reviews. Genetics* **8**, 762-775.

Chutimanitsakun, Y., Nipper, R.W., Cuesta-Marcos, A., Cistue, L., Corey, A., Filichkina, T., Johnson, E.A. and Hayes, P.M. (2011) Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. *Bmc Genomics* **12**, 4.

Durstewitz, G., Polley, A., Plieske, J., Luerssen, H., Graner, E.M., Wieseke, R. and Ganal, M.W. (2010) SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species Brassica napus. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* **53**, 948-956.

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

Edwards, D., Batley, J. and Snowdon, R.J. (2013) Accessing complex crop genomes with next-generation sequencing. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **126**, 1-11.

Farkhari, M., Lu, Y., Shah, T., Zhang, S., Naghavi, M.R., Rong, T. and Xu, Y. (2011) Recombination frequency variation in maize as revealed by genomewide single-nucleotide polymorphisms. *Plant Breeding* **130**, 533-539 %U http://dx.doi.org/510.1111/j.1439-0523.2011.01866.x.

Gaur, R., Azam, S., Jeena, G., Khan, A.W., Choudhary, S., Jain, M., Yadav, G., Tyagi, A.K., Chattopadhyay, D. and Bhatia, S. (2012) High-throughput SNP discovery and genotyping for constructing a saturated linkage map of chickpea (Cicer arietinum L.). *DNA research : an international journal for rapid publication of reports on genes and genomes* **19**, 357-373.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J. and Anderson, L.K. (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nature reviews. Genetics* **8**, 77-84.

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhue, C., Pudlo, P., Cornuet, J.M. and Estoup, A. (2012) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular ecology*.

Hayward, A., Mason, A., Dalton-Morgan, J., Zander, M., Edwards, D. and Batley, J. (2012a) SNP discovery and applications in Brassica napus. *Plant biotechnology journal* **39**, 1-12.

Hayward, A., Vighnesh, G., Delay, C., Samian, M.R., Manoli, S., Stiller, J., McKenzie, M., Edwards, D. and Batley, J. (2012b) Second-generation sequencing for gene discovery in the Brassicaceae. *Plant biotechnology journal* **10**, 750-759.

Hegarty, M., Yadav, R., Lee, M., Armstead, I., Sanderson, R., Scollan, N., Powell, W. and Skot, L. (2013) Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (Lolium perenne (L.)). *Plant biotechnology journal*.

Hiremath, P.J., Kumar, A., Penmetsa, R.V., Farmer, A., Schlueter, J.A., Chamarthi, S.K., Whaley, A.M., Carrasquilla-Garcia, N., Gaur, P.M., Upadhyaya, H.D., Kavi Kishor, P.B., Shah, T.M., Cook, D.R. and Varshney, R.K. (2012) Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. *Plant biotechnology journal* **10**, 716-732.

Hohenlohe, P.A., Bassham, S., Currey, M. and Cresko, W.A. (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philos T R Soc B* **367**, 395-408.

Hu, Z.Y., Huang, S.M., Sun, M.Y., Wang, H.Z. and Hua, W. (2012) Development and application of single nucleotide polymorphism markers in the polyploid Brassica napus by 454 sequencing of expressed sequence tags. *Plant Breeding* **131**, 293-299.

Jain, M., Misra, G., Patel, R.K., Priya, P., Jhanwar, S., Khan, A.W., Shah, N., Singh, V.K., Garg, R., Jeena, G., Yadav, M., Kant, C., Sharma, P., Yadav, G., Bhatia, S., Tyagi, A.K. and Chattopadhyay, D. (2013) A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *The Plant Journal* **74**, 715-729.

Lagercrantz, U. and Lydiate, D.J. (1995) RFLP mapping in Brassica nigra indicates differing recombination rates in male and female meioses. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* **38**, 255-264.

Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967.

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T.J., Hu, Q., Wang, X., Yue, Z., Li, H., Yang, L., Wu, J., Zhou, Q., Wang, W., King, G.J., Pires, J.C., Lu, C., Wu, Z., Sampath, P., Wang, Z., Guo, H., Pan, S., Yang, L., Min, J., Zhang, D., Jin, D., Li, W., Belcram, H., Tu, J., Guan, M., Qi, C., Du, D., Li, J., Jiang, L., Batley, J., Sharpe, A.G., Park, B.S., Ruperao, P., Cheng, F., Waminal, N.E., Huang, Y., Dong, C., Wang, L., Li, J., Hu, Z., Zhuang, M., Huang, Y., Huang, J., Shi, J., Mei, D., Liu, J., Lee, T.H., Wang, J., Jin, H., Li, Z., Li, X., Zhang, J., Xiao, L., Zhou, Y., Liu, Z., Liu, X., Qin, R., Tang, X., Liu, W., Wang, Y., Zhang, Y., Lee, J., Kim, H.H., Denoeud, F., Xu, X., Liang, X., Hua, W., Wang, X., Wang, J., Chalhoub, B. and Paterson, A.H. (2014) The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications* **5**, 3930.

Lorenc, M., Hayashi, S., Stiller, J., Lee, H., Manoli, S., Ruperao, P., Visendi, P., Berkman, P., Lai, K., Batley, J. and Edwards, D. (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* **1**, 370-382.

Mezard, C., Vignard, J., Drouaud, J. and Mercier, R. (2007) The road to crossovers: plants have their say. *Trends in genetics : TIG* **23**, 91-99.

Nayak, S.N., Zhu, H., Varghese, N., Datta, S., Choi, H.K., Horres, R., Jungling, R., Singh, J., Kishor, P.B., Sivaramakrishnan, S., Hoisington, D.A., Kahl, G., Winter, P., Cook, D.R. and Varshney, R.K. (2010) Integration of novel SSR and gene-based SNP marker loci in the chickpea genetic map and establishment of new anchor points with Medicago truncatula genome. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **120**, 1415-1441.

Nicolas, S.D., Le Mignon, G., Eber, F., Coriton, O., Monod, H., Clouet, V., Huteau, V., Lostanlen, A., Delourme, R., Chalhoub, B., Ryder, C.D., Chevre, A.M. and Jenczewski, E. (2007) Homeologous recombination plays a major role in chromosome rearrangements that occur during meiosis of Brassica napus haploids. *Genetics* **175**, 487-503.

Parkin, I.A., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D., Nixon, J., Krishnakumar, V., Bidwell, S.L., Denoeud, F., Belcram, H., Links, M.G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A.S., Pires, C.J., Barker, G., Moore, J., Walley, P.G., Manoli, S., Batley, J., Edwards, D., Nelson, M.N., Wang, X., Paterson, A.H., King, G., Bancroft, I., Chalhoub, B. and Sharpe, A.G. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome biology* **15**, R77.

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

Ruperao, P., Chan, C.K., Azam, S., Karafiatova, M., Hayashi, S., Cizkova, J., Saxena, R.K., Simkova, H., Song, C., Vrana, J., Chitikineni, A., Visendi, P., Gaur, P.M., Millan, T., Singh, K.B., Taran, B., Wang, J., Batley, J., Dolezel, J., Varshney, R.K. and Edwards, D. (2014) A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies. *Plant biotechnology journal*.

Sun, Z., Wang, Z., Tu, J., Zhang, J., Yu, F., McVetty, P.B. and Li, G. (2007) An ultradense genetic recombination map for Brassica napus, consisting of 13551 SRAP markers. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* **114**, 1305-1317.

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

Udall, J.A., Quijada, P.A. and Osborn, T.C. (2005) Detection of chromosomal rearrangements derived from homologous recombination in four mapping populations of Brassica napus L. *Genetics* **169**, 967-979.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S., Baek, J., Rosen, B.D., Tar'an, B., Millan, T., Zhang, X., Ramsay, L.D., Iwata, A., Wang, Y., Nelson, W., Farmer, A.D., Gaur, P.M., Soderlund, C., Penmetsa, R.V., Xu, C., Bharti, A.K., He, W., Winter, P., Zhao, S., Hane, J.K., Carrasquilla-Garcia, N., Condie, J.A., Upadhyaya, H.D., Luo, M.C., Thudi, M., Gowda, C.L., Singh, N.P., Lichtenzveig, J., Gali, K.K., Rubio, J., Nadarajan, N., Dolezel, J., Bansal, K.C., Xu, X., Edwards, D., Zhang, G., Kahl, G., Gil, J., Singh, K.B., Datta, S.K., Jackson, S.A., Wang, J. and Cook, D.R. (2013a) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nature biotechnology* **31**, 240-246.

Varshney, R.K., Song, C., Saxena, R.K., Azam, S., Yu, S., Sharpe, A.G., Cannon, S.B., Baek, J., Tar'an, B., Millan, T., Zhang, X., Rosen, B., Ramsay, L.D., Iwata, A., Wang, Y., Nelson, W., Farmer, A.D., Gaur, P.M., Soderlund, C., Penmetsa, R.V., Xu, C., Bharti, A.K., He, W., Winter, P., Zhao, S., Hane, J.K., Carrasquilla-Garcia, N., Condie, J.A., Upadhyaya, H.D., Luo, M., Singh, N.P., Lichtenzveig, J., Gali, K.K., Rubio, J., Nadarajan, N., Thudi, M., Dolezel, J., Bansal, K.C., Xu, X., Edwards, D., Zhang, G., Kahl, G., Gil, J., Singh, K.B., Datta, S.K., Jackson, S.A., Wang, J. and Cook, D. (2013b) Draft genome sequence of kabuli chickpea (*Cicer arietinum*): genetic structure and breeding constraints for crop improvement. *Nature Biotechnology*.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J.C., Paterson, A.H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A.G., Park, B.S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D.,

Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G.J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I.A., Batley, J., Kim, J.S., Just, J., Li, J., Xu, J., Deng, J., Kim, J.A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Wang, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M.G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P.J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.J., Choi, S.R., Lee, T.H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., Zhang, Z. and Brassica rapa Genome Sequencing Project, C. (2011) The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics* **43**, 1035-1039.

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J.R., Rawat, V., Rowan, B.A., de Jong, D.F., de Snoo, C.B., Zapata, L., Huettel, B., de Jong, H., Ossowski, S., Weigel, D., Koornneef, M., Keurentjes, J.J. and Schneeberger, K. (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *eLife* **2**, e01426.

Yang, S., Yuan, Y., Wang, L., Li, J., Wang, W., Liu, H., Chen, J.Q., Hurst, L.D. and Tian, D. (2012) Great majority of recombination events in Arabidopsis are gene conversion events. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 20992-20997.

Yao, H., Zhou, Q., Li, J., Smith, H., Yandeau, M., Nikolau, B.J. and Schnable, P.S. (2002) Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6157-6162.

Tables

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

Name	Total length	SNP count	SNPs / kbp	Ts / Tv
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

Table 2: Predicted SNPs in chickpea (*C. arietinum*) between PI4958 and ICC489777.

Name	Length	SNP count	SNPs / Kbp	Ts/Tv
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

Figure legends

Figure 1: Relationship between the number of called alleles and number of aligned reads for each of the 92 *Brassica napus* DH individuals.

Figure 2: Crossover map for *Brassica napus* chromosome A1 before filtering of overlapping recombinations. Red: genotype Tapidor, blue: genotype Ningyou, white: missing. Each line is one individual.

Figure 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.

Figure 4: Recombination map for *Cicer arietinum* chromosome A1 before filtering of overlapping recombinations. Red: genotype ICC4958, blue: genotype PI489777 (wild-type, white: missing. Each line is one individual.

Figure 5: Recombination map for *Cicer arietinum* chromosome A1 after filtering of overlapping recombinations. Red: genotype ICC4958, blue: genotype PI489777 (wild-type, white: missing. Each line is one individual.

Figures

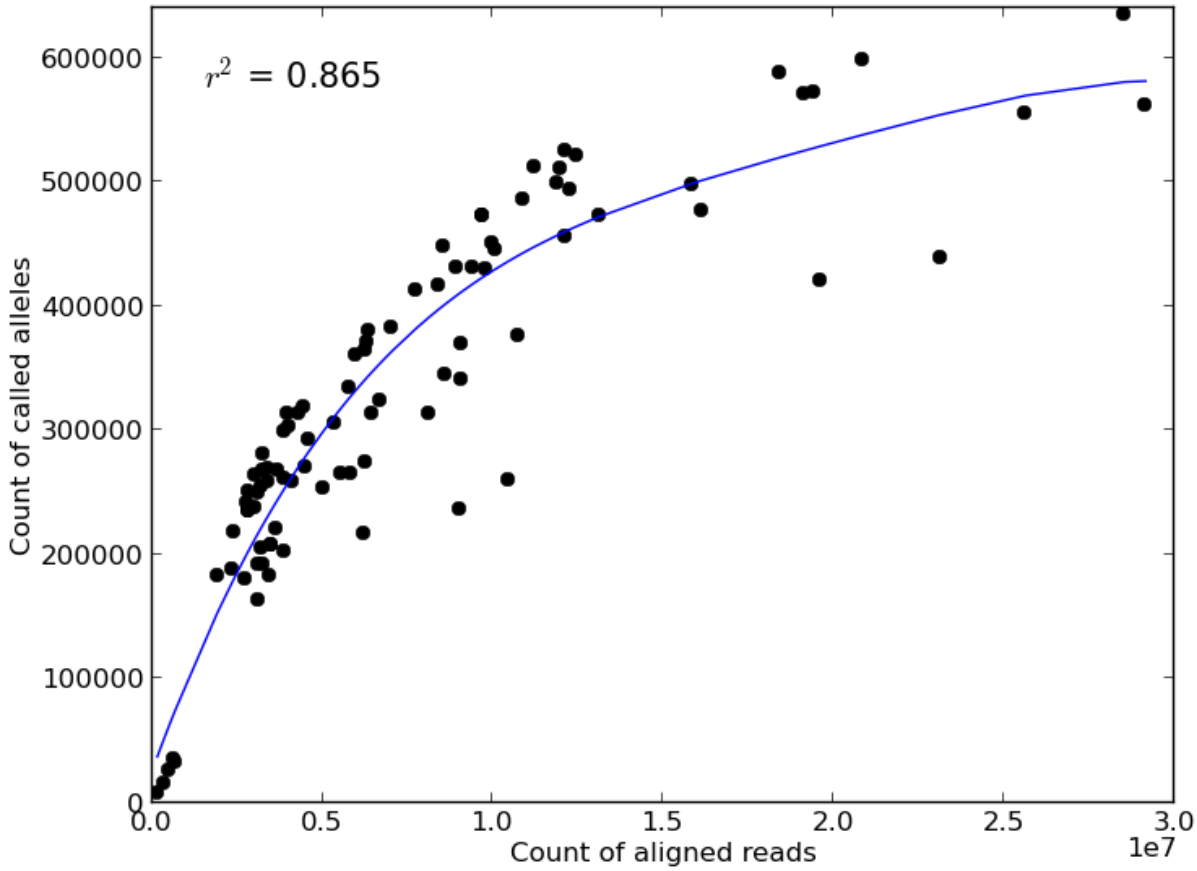

Figure 1: Relationship between the number of called alleles and number of aligned reads for each of the 92 *Brassica napus* DH individuals.

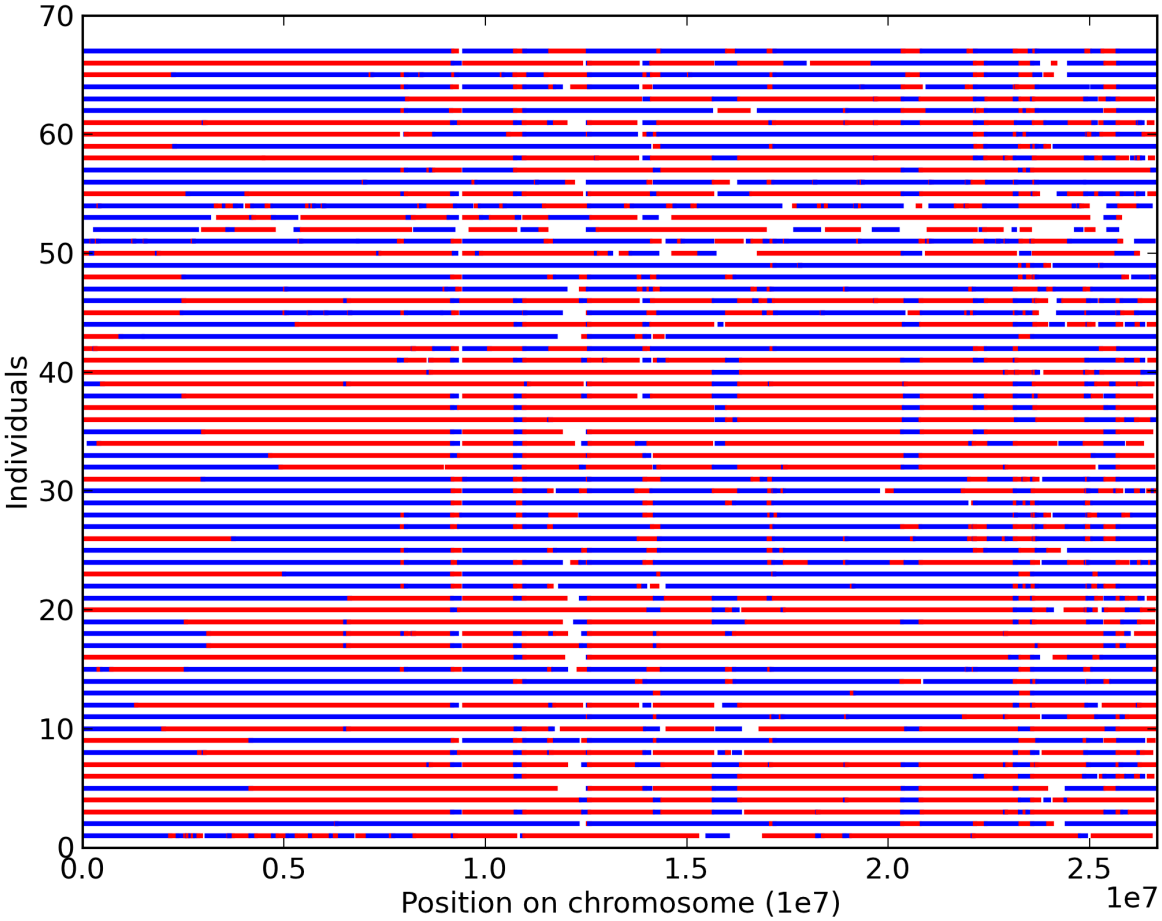

Figure 2: Crossover map for *Brassica napus* chromosome A1 before filtering of overlapping recombinations. Red: genotype Tapidor, blue: genotype Ningyou, white: missing. Each line is one individual.

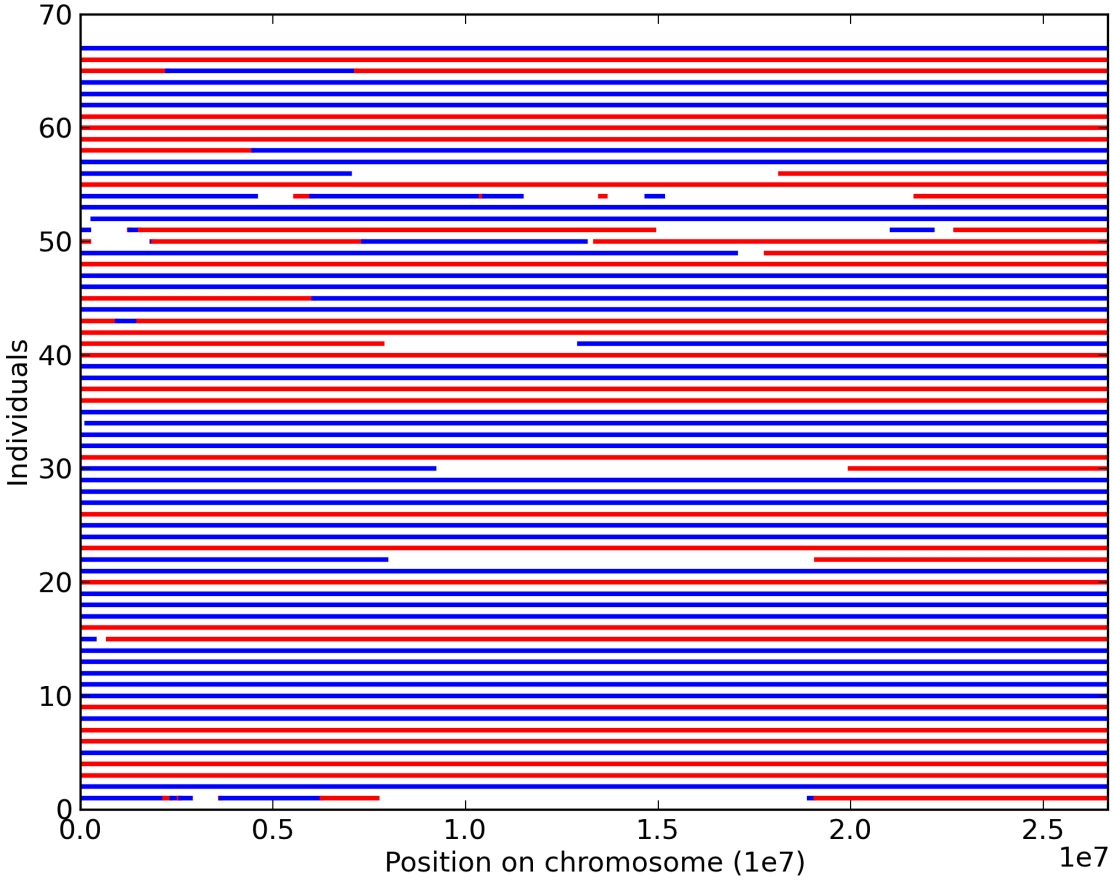

Figure 3: Recombination map for *Brassica napus* chromosome A1 after filtering of overlapping recombinations. Red: genotype Tapidor, blue: genotype Ningyou, white: missing. Each line is one individual.

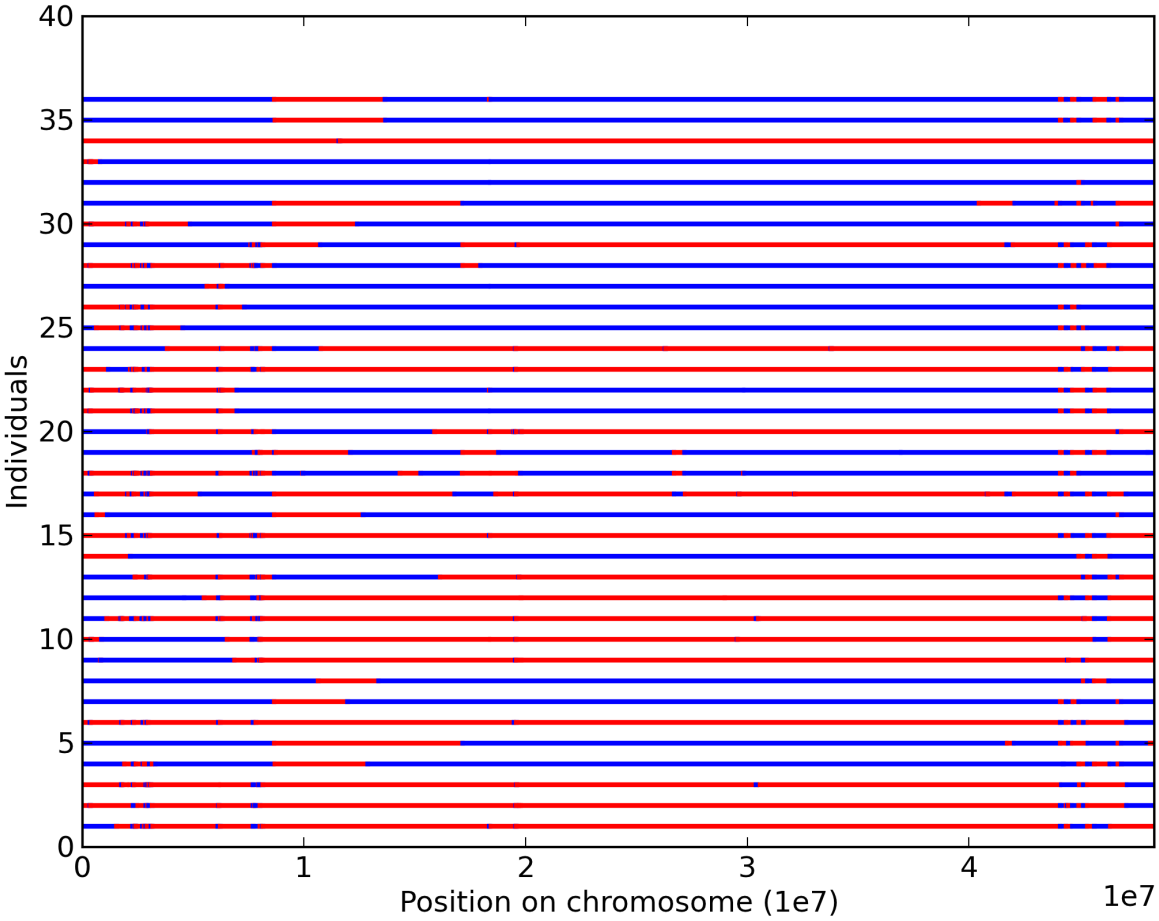

Figure 4: Recombination map for *Cicer arietinum* chromosome A1 before filtering of overlapping recombinations. Red: genotype ICC4958, blue: genotype PI489777 (wild-type, white: missing. Each line is one individual.

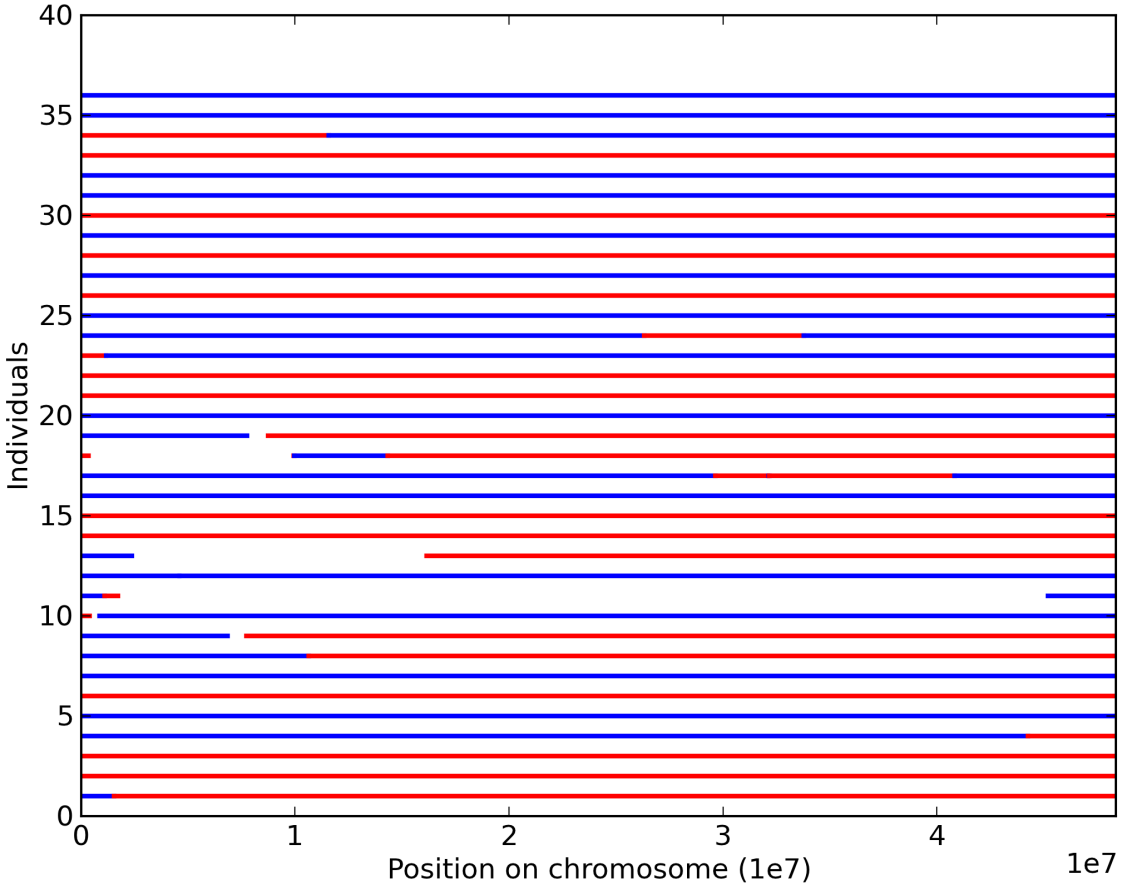

Figure 5: Recombination map for *Cicer arietinum* chromosome A1 after filtering of overlapping recombinations. Red: genotype ICC4958, blue: genotype PI489777 (wild-type, white: missing. Each line is one individual.