High-throughput genotyping for species identification and diversity assessment in germplasm collections

Annaliese S. Mason[a,b], Jing Zhang[c,d], Reece Tollenaere[a,b], Paula Vasquez Teuber[a,b], Jessica Dalton-Morgan[a,b], Liyong Hu[c], Guijun Yan[d,e], David Edwards[a,d], Robert Redden[f], Jacqueline Batley[a,b,d]*

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

[c] Ministry of Agriculture (MOA) key laboratory of Huazhong Crop Physiology, Ecology and Production, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan, 430070, China.

[d] School of Plant Biology, Faculty of Science and [e] The UWA Institute of Agriculture, The University of Western Australia, Perth, 6009, WA, Australia

[f] Australian Grains Genebank, Department of Environment and Primary Industries, Horsham, 3401, VIC , Australia

* corresponding author; jacqueline.batley@uwa.edu.au; Tel: +61 (0) 7 334 69534; Fax:+ 61 (0)7 336 59556

Keywords: molecular genotyping, Brassicaceae, germplasm collections, genebanks, genetic resources, Illumina Infinium SNP array

Running title: High throughput germplasm genotyping

Abstract

Germplasm collections provide an extremely valuable resource for breeders and researchers. However, misclassification of accessions by species often hinders the effective use of these collections. We propose that use of high-throughput genotyping tools can provide a fast, efficient and cost-effective way of confirming species in germplasm collections, as well as providing valuable genetic diversity data. We genotyped 180 Brassicaceae samples sourced from the Australian Grains Genebank across the recently released Illumina Infinium Brassica 60K SNP array. Of these, 76 were provided on the basis of suspected misclassification and another 104 were sourced independently from the germplasm collection. Presence of the A and C genomes combined with principle components analysis clearly separated *B. rapa*, *B. oleracea*, *B. napus*, *B. carinata* and *B. juncea* samples into distinct species groups. Several lines were further validated using chromosome counts. Overall, 18% of samples (32/180) were misclassified on the basis of species. Within these 180 samples, 23/76 (30%) supplied on the basis of suspected misclassification were misclassified, and 9/105 (9%) of the samples randomly sourced from the Genebank were misclassified. Surprisingly, several individuals were also found to be the product of interspecific hybridisation events. The SNP (Single Nucleotide Polymorphism) array proved effective at confirming species, and provided useful information related to genetic diversity. As similar genomic resources become available for different crops, high-throughput molecular genotyping will offer an efficient and cost-effective method to screen germplasm collections worldwide, facilitating more effective use of these valuable resources by breeders and researchers.

Introduction

Natural genetic diversity in crop species is a key resource for agricultural improvement. Genetic variation for cold and heat tolerance, drought and disease resistance as well as other environmental stresses exists in most natural species, but is often lost through domestication and selection for yield and yield-related traits in crops (Day 1973; Hyten *et al.* 2006; Simmonds 1962; Zamir 2001). In order to preserve this useful genetic diversity for later introgression back into crop cultivars and for targeted breeding attempts in crop improvement, "genebanks" and diversity collections exist around the world (Tanksley& McCouch 1997). These collections preserve wild accessions, landraces and cultivars collated from local and international sources, often comprising tens to thousands of lines. Seeds are donated by breeders, collectors and research institutions, and lines are maintained as a resource for future generations.

Brassica comprises the largest number of domesticated crop species of any genus, and includes leaf vegetables, oilseeds, condiments and root vegetable crops; such as rapeseed, mustards, cabbage, turnips, broccoli and cauliflower. Numerous species in the wider Brassicaceae can also be hybridised with key crop species within *Brassica*, including the wild radishes (*Raphanus*), woad (*Isatis*) and white mustard (*Sinapis*), as well as the *Brassica* C genome clade of *B. cretica, B. hilarionis*, *B. incana* and *B. macrocarpa*, among others (FitzJohn *et al.* 2007; Harberd& McArthur 1980; Prakash *et al.* 1999; Warwick *et al.* 2003). This potential for hybrid introgression from wild relatives coupled with the extant genetic diversity in the non-cultivated forms of key crop species makes *Brassica* a major feature of genebank collections worldwide. The six cultivated *Brassica* species share an interesting genomic relationship, with three diploids (*B. rapa*, 2n = AA = 20; *B. nigra*, 2n = BB = 16 and *B. oleracea*, 2n = CC = 18) and a set of three allotetraploids each containing two of the three diploid genomes (*B. juncea*, 2n = AABB = 36; *B. napus*, 2n = AACC = 38 and *B. carinata*, 2n = BBCC = 34) (Morinaga 1934; U 1935). Allotetraploid *B. napus* is one of the most agriculturally significant crop species within this genus, with rapeseed and canola contributing to oil production for food and biofuel. However, canola is also the least diverse, with major genetic bottlenecks as a result of only a limited number of hybridisation events between diploid progenitors to form the allotetraploid (Palmer *et al.* 1983), coupled with rigorous selective pressure to achieve "canola-quality" oil for human consumption and enhance yield with the recent emphasis on breeding of oilseeds in this domesticated crop (Cowling 2007). No known wild forms of this species exist (Dixon 2007). Hence, *B. napus* in particular is a critical crop species for genetic improvement via introgression of diversity from both wild and domestic diploid relatives, particularly those with which it shares the A and C genomes (*B. rapa*, *B. oleracea*, *B. juncea* and *B. carinata*). Several past breeding attempts have demonstrated the efficacy of this approach in introgressing disease resistance from related species (Navabi *et al.* 2010b; Rygulla *et al.* 2007; Saal *et al.* 2004).

A major problem with genebank collections is ensuring the accurate identification of species. Many genebanks do not have the resources to assess every line gifted to them for genetic diversity, correct origin and correct species identification. To date, attempts to identify species in germplasm collections have all relied on low-throughput molecular marker genotyping approaches (Dangl *et al.* 2001; Ferriol *et al.* 2003; Lee *et al.* 2014; Martin *et al.* 1997; Pradhan *et al.* 2011). However, generation of inexpensive high-throughput molecular marker data is now becoming routine for many genera. We show how the recently released Illumina Infinium Brassica 60K SNP array can be used for rapid species identification in the *Brassica* genus, revealing cases of species misclassification, providing useful genetic diversity information and confirming genome composition in this major agricultural genus.

Materials and Methods

Germplasm

A total of 188 experimental samples (176 lines) were genotyped for this experiment (Supplementary Table 1). A set of 77 samples with suspected species attribution errors and another set of 111 independently-obtained samples were sourced from the Australian Grains Genebank (Supplementary Table 1). Forty two additional samples (37 lines) of confirmed species origin were also included in the analysis as controls (Supplementary Table 1). These comprised 22 *B. napus* lines (commercially available canola cultivars from Australia and China), four *B. juncea* lines (“JN9-04”, “Purple Leaf Mustard”, “Domo” and “Lethbridge”), two *B. carinata* lines (“195923” and “94024”, breeding lines in Australia of Ethiopian origin), two *B. oleracea* lines (sequenced accession “TO1000” and commercially available cauliflower “Snowball”), two *B. rapa* lines (sequenced South Korean cultivar “Chiifu” and a commercial “Pak Choy” variety) and five *Raphanus sativus* lines (commercial radish varieties “Cherry Belle”, “Long Scarlet”, “Mila”, “Saxa” and “Scarlet Globe”).

Genotyping and statistical analyses

DNA was extracted according to methodology detailed in Fulton et al. (1995). All DNA samples were hybridized to an Illumina Infinium Brassica 60K array SNP array released for the *Brassica napus* genome (http://illumina.com; 52157 SNPs) according to manufacturer’s instructions. SNP (Single Nucleotide Polymorphism) chips were scanned using an Illumina HiScanSQ and data visualised using Genome Studio V2011.1 (Illumina, Inc., San Diego, CA, USA). A cluster file provided by Agriculture and Agri-Food Canada, Saskatoon, Canada was used to cluster SNPs into genotype groupings (e.g. GG, GT and TT allele calls, which were converted into 0, 1 and 2 scores for subsequent analysis). SNP locations were determined through BLAST comparison with the public *B. rapa* and *B. oleracea* reference genome sequences (Parkin *et al.* 2014; Wang *et al.* 2011); Supplementary Table 2. Percentage SNP calls for each genome were calculated for each sample and this information used to determine the presence or absence of the A and C genomes in the sample.

Hierarchical clustering and principle components analysis (PCA) were carried out using R version 3.0 (The R Project for Statistical Computing). Dendrograms were generated using n = 1000 bootstrap iterations to validate branches, using the "pvclust" function in R package "pvclust". Dendrogram "Height" represents squared Euclidean distance between samples. Missing values were replaced with means for each SNP across the population using R package "gam", function "na.gam.replace". PCA was carried out and output graphs generated using the "dudi.pca" function in R package "ade4".

Chromosome counting

Seeds from five experimental lines were germinated on petri dishes under laboratory conditions before harvesting root tip meristems. Root tips were collected and chromosome spreads prepared according to protocols detailed in Mason et al. (2014), using DAPI (4,6-diamidino-2-phenylindole) as a fluorescent stain. Pictures were taken on a Nikon Eclipse E600 microscope with digital camera.

Results

Presence and absence of the A and C genomes

The Illumina Infinium Brassica 60K array comprises 52 157 SNPs. Of these, 10 634 (20.4%) were removed as unreliable or non-specific (consistently amplifying alleles at more than one locus) on the basis of information provided by the Illumina Infinium Brassica 60K cluster file. Of the remaining 41 523 SNPs, 44.5% (18 471) were physically located on the *B. rapa* genome (Wang *et al.* 2011) and 53.4% (22 155) on the *B. oleracea* genome (Parkin *et al.* 2014). Approximately 12% of these A-genome SNPs also amplified C-genome alleles (in *B. carinata* and *B. oleracea* controls with no A genome), and approximately 23% of these C-genome SNPs also amplified A-genome alleles (in *B. rapa* and *B. juncea* controls with no C genome). *Raphanus sativus* samples amplified 13% of alleles on average, with no difference in amplification between the A and C genome SNPs (p = 0.2, Student's t-test).

A set of 43 control samples (3 *B. rapa*; 6 *B. juncea*, 23 *B. napus,* 2 *B. oleracea*, 4 *B. carinata* and 5 *Raphanus sativus*) were run on the Illumina Infinium 60K SNP array. Amplification of A and C genome alleles was assessed in these samples. Clear groups could be distinguished on the basis of A and C genome presence or absence in the controls (Supplementary Figure 1); these groups corresponded to the expected genome presence/absence for each species sample. Of the 188 samples in the experimental population, 59 samples could be classed as “A genome only”, 16 samples could be classed as “C genome only”, 101 samples could be classed as “A + C genomes” and two samples could be classed as “neither A or C genome present” (Figure 1). An additional seven samples were considered to have failed due to poor quality amplification (removed from further analysis and not included in Figure 1), and another three samples were considered anomalous (included in Figure 1). Two of these samples (R14 and J16) were included in subsequent “A genome only” analyses, and one sample (I2) was discarded from further analysis, leaving 180 samples. On this basis alone, 29/180 of the samples (16%) could be identified to belong to a different species than the one in the genebank records (Supplementary Table 1, Figure 1). Presence of both the A and C genomes also provided a unique identifier for *Brassica napus* samples: 83% of samples (95/115) thought to be *B. napus* were actually *B. napus* (Supplementary Table 1, Figure 1).

A robust cut-off for sample quality was >75% amplification (an allele call for >75% of SNPs in the A and/or C genome rather than no call reliably indicated genome presence) or <35% amplification in each genome (an allele call for <35% of SNPs in the A and/or C genome reliably indicated genome absence). Samples with 32-57% A and C genome amplification (Supplementary Table 1) also showed random patterns of allele calls and missing data across chromosomes, indicative of unreliable and poor quality SNP data. One of the three samples considered to be anomalous was a putative *B. nigra* sample (I2) that showed 36% A genome and 41% C genome amplification (Figure 1); this may be due to misclassification of this sample coupled with poor quality amplification. The second sample (J16) considered to be anomalous showed 70% A genome amplification and 39% C genome amplification (Figure 1). The third sample considered to be anomalous (putative *B. rapa* sample R14) had 89% A genome presence and 49% C genome presence: on closer inspection of the SNP data, this individual showed presence of some C genome chromosome segments (27 Mbp of C1, all of C2, 7 Mbp of C5, 24 Mbp of C6, 30 Mbp of C7 and 39 Mbp of C8). Although material was not available from the individual genotyped, the presence of only 20 chromosomes was confirmed in other individuals from this same line by chromosome counting. Anomalous samples J16 and R14 were retained in our analysis, and sample I2 was discarded.

Phylogenetic groupings for species identification

Hierarchical clustering and principle components analysis were performed to separate *B. juncea* and *B. rapa* individuals and *B. carinata* and *B. oleracea* individuals. The *B. juncea* and *B. rapa* group (as deduced from genome presence/absence to have only the A genome) comprised 9 controls and 61 experimental individuals. Of the 18 471 SNPs physically mapping to the A genome, 11 983 were polymorphic and amplified in ≥ 90% of the individuals in the population, and were hence used for subsequent analysis. Hierarchical clustering allowed separation of *B. rapa* and *B. juncea* lines, but although species-specific clades were apparent, 100% confidence was not achieved for clade separation (Figure 2; numbers in green and red represent the number of times each branch was in the same position over the 1000 iterations, hence P<0.05 = 95 or greater). PCA provided clear separation between *B. rapa* and *B. juncea*, with the first two axes separating two *B. rapa* clades and separating these two groups from *B. juncea* clade, contributing to 18.4% and 13.9% of the variance respectively (Figure 3). Sixty-eight axes were generated, with 48.7% of the variance explained by the first five axes of the PCA.

The *B. carinata* and *B. oleracea* group as identified by presence of only the C genome consisted of 6 control samples and 16 experimental samples. Of the SNP markers mapped to the C genome, 12 794 were polymorphic and amplified in ≥ 90% of the individuals in the population, and were hence used for subsequent analysis. Although the *B. carinata* clade fell within the wider *B. oleracea* group, these individuals formed a smaller subgroup with 100% confidence for clade identity using hierarchical clustering (Figure 4). Principle components analysis also showed very clear separation of *B. oleracea* and *B. carinata* samples (first and second axes 41.3% and 13.0% of the variance respectively) and extremely tight grouping of *B. carinata* samples relative to the *B. oleracea* types (Figure 5).

Overall, 18% of samples (32/180) were misclassified on the basis of species (Table 1). Of the samples suspected to be misclassified, 23/76 (30%) were indeed a species different to the one listed by the Australian Grains Genebank. Of the samples otherwise sourced from the Australian Grains Genebank, 9/104 (9%) were misclassified on the basis of species. *B. napus* was observed to be mistaken for each of *B. rapa*, *B. juncea* and *B. carinata*; *B. juncea* was mistaken for *B. rapa* and *B. napus* and *B. rapa* was mistaken for *B. juncea* (Table 1). A complete set of source, species and cultivar/landrace/wild type classifications from the Australian Grains Genebank with confirmed species identifications and SNP genome amplification and heterozygosity results is provided in Supplementary Table 1. Lines were supplied by the Australian Grains Genebank with the label "Advanced cultivar", "Breeder's Line", "Traditional Cultivar/Landrace", "Wild" or "Unknown". Of the 75 samples in the "Advanced cultivar" category, 9 were misclassified (12%). "Traditional cultivar/landraces" had 2/22 samples misclassified (9%) and "Breeder's Line" samples had 2/21 samples misclassified (10%). The single "Wild" sample was also misclassified. "Unknown" samples were misclassified 21 % of the time (11/61).

Genetic diversity

Genome diversity within the A genome was assessed in *B. napus*, *B. juncea* and *B. rapa* lines using 13 292 polymorphic SNPs amplifying in ≥ 90% of the individuals. Percentage heterozygosity for each individual within the A genome was also calculated using the entire set of A-genome specific SNPs (Supplementary Table 1). C genome diversity was assessed in *B. napus*, *B. oleracea* and *B. carinata* lines using 18 076 SNPs amplifying in ≥ 90% of the individuals and not monomorphic in the population. Percentage heterozygosity for each individual within the C genome was also calculated using the whole set of C-genome specific SNPs (Supplementary Table 1).

Brassica rapa samples putatively from India and Bangladesh based on provenance of samples R05 and R21 (Supplementary Table 1, leftmost clade in Figure 3) formed a clearly distinct subgroup when compared to other samples originating from Europe and the rest of Asia. This grouping was not apparent in the first two axes of the PCA of A-genome diversity including the *B. napus* samples (Figure 6). Two outliers were observed on the basis of A-genome diversity using PCA: J06 and J08 (Figure 6), which were both reported to be *B. juncea* from China but showed presence of both the A and C genomes; however, using hierarchical clustering analysis these individuals fell within the *B. juncea* clade (Supplementary Figure 2). Both individuals had very high A genome heterozygosity (40 and 49%) but lower C genome heterozygosity (7 and 21%; Supplementary Table 1).

As previously observed (Figure 5), the *B. carinata* clade formed a group of tightly-related lines nestled within the *B. oleracea* samples using hierarchical cluster analysis (Supplementary Figure 3). All *B. napus* lines fell outside the *B. oleracea*/*carinata* clade except for three: N019a, N038 and N074 (Supplementary Figure 3). Principle components analysis placed N019a within the *B. oleracea* samples, with N038 and N074 in the *B. napus* group but close to *B. oleracea* (Figure 7). N019b, a separately sourced individual of the same accession as N019a, was confirmed to be *B. carinata* due to lack of A genome alleles. N019a had a complete A and C genome, but showed 8.5% heterozygosity in the A genome and 43% C genome heterozygosity, the highest C genome heterozygosity of any experimental *B. napus* sample (Supplementary Table 1). N038 and N074 both had high A- and C-genome heterozygosity (25 – 36% per genome, Supplementary Table 1).

Chromosome counting

Chromosome counts were performed for five experimental lines: N067, N089, R05, R14 and R21 (Figure 8). Putative *B. napus* sample N067 was confirmed to be *B. juncea* (2n = 36 chromosomes) rather than *B. napus* or *B. rapa*, and putative *B. napus* sample N089 was confirmed to be *B. carinata* (2n = 34 chromosomes) rather than *B. napus* or *B. oleracea*. Each of putative *B. rapa* samples R05, R14 and R21 had 2n = 20 chromosomes, confirming that these plants were *B. rapa*.

Discussion

Germplasm collections and genebanks provide an excellent resource for breeders and researchers. However, misclassification of sample genotype and even species is common. Here, we evaluate the use of a high-throughput genotyping technology for the assessment of germplasm collections: the Illumina SNP array, which is increasingly becoming available and cost-effective for many species of interest. We used the Illumina Brassica 60K SNP array for species identification in 180 Brassicaceae samples from the Australian Grains Genebank, a widely used germplasm collection housed in Horsham, Victoria, Australia. The Illumina SNP array provided a quick and effective means to classify species and assess genetic diversity in these samples. A total of 18% of samples were found to be misclassified on the basis of species, and several subpopulations were identified within the various *Brassica* species. A few individuals were also unexpectedly found to result from interspecific hybridisation. This information will prove valuable to future users of this germplasm resource, and validates the use of the Illumina SNP array system for high–throughput genotyping of germplasm collections, particularly in *Brassica*.

Molecular markers have been used to genotype germplasm collections in the past: SRAP and AFLP markers have been used in cucumber (Ferriol *et al.* 2003), RAPD markers have been used in rice (Martin *et al.* 1997) and SSR markers have been used in grape (Dangl *et al.* 2001) and safflower (Lee *et al.* 2014). High-throughput molecular genotyping is now also starting to be used in major crops: a recent study used genotyping-by-sequencing to characterise lines in the USA national maize inbred seed bank (Romay *et al.* 2013). Problems of species identity within germplasm collections are widespread: in rice, 9/62 (15%) of wild *Oryza* accessions were found to be misclassified; 2/41 grape lines were misclassified, and in another *B. nigra* study using SSR markers, 16/60 (27%) accessions were found to not be *B. nigra* (Pradhan *et al.* 2011). However, older marker technologies are generally not high-throughput, and species identification in germplasm collections using molecular markers has remained out of reach in terms of time and cost until now. In *Brassica* in particular, the high level of homoeology between the A and C genomes, and the presence of multiple species sharing these genomes, can make identification of species-specific alleles difficult (Li *et al.* 2013). In our study, the provision of SNP markers already mapped to the reference genome sequences, a resource which is increasingly available for species of interest, allowed much greater resolution and effectively separated the closely related *Brassica* species.

We used both Principal Components Analysis and hierarchical clustering to group individuals based on the SNP data results. Importantly, presence of the A genome only, C genome only or both A and C genomes was first used to discriminate *B. napus* samples from *B. juncea*/*B. rapa* and *B. carinata*/*B. oleracea*, as *B. napus* samples were not always otherwise 100% distinguishable from *B. juncea* or *B. carinata*. Principal Components Analysis proved more effective at separating species with shared genomes than hierarchical clustering in our analysis. As allopolyploid species *B. carinata*, *B. juncea* and *B. napus* result from a few hybridisation events between diploid progenitor species *B. rapa*, *B.*

nigra and *B. oleracea* (Arias *et al.* 2014; Kaur *et al.* 2014), the allopolyploid species form less diverse clades nested within the diversity represented by the diploids. To distinguish between *B. juncea* and *B. rapa* and between *B. carinata* and *B. oleracea*, only shared genome information (A or C genome) was available. Hence, hierarchical clustering, which performs pairwise calculations of similarity between samples, may have been less effective at separating species than Principal Components Analysis, which looks at broader correlations and similarities across the data set. Although hierarchical clustering still showed some utility in discriminating between species (Fig. 2, Fig. 4) single-genome Principal Components Analysis is therefore recommended for this purpose in future studies.

Interestingly, *B. napus* lines in our study were observed to be mistaken for each of *B. carinata*, *B. juncea* and *B. rapa*, but only *B. juncea* was commonly mistaken as *B. napus*. However, more *B. napus* lines were used in this experiment than any other species, hence increasing the chance that misclassification errors would be picked up in *B. napus* relative to the other species. Lines sourced as "Traditional cultivar/landraces" or "Unknown" samples may have been expected to be more commonly misclassified than "Advanced Cultivar" or "Breeding Line" samples. However, although "Unknown" samples comprised by far the largest percentage of misclassified samples (11/25), lines sourced as "Advanced Cultivars" were also likely to be misclassified, with a further 9 samples falling into this category. Some of these may have resulted from mislabelling or contamination during seed collection or during seed regeneration of accessions, particularly in the case of commercially available open pollinated (OP) canola cultivars or lines that have passed through many hands before being donated to the Australian Grains Genebank. However, in many cases accurate phenotypic identification of species misclassification was made by the germplasm curators. Samples suspected to be misclassified by the germplasm bank were three times more likely to actually be misclassified on the basis of species (30% as opposed to 9%). In addition, specific recorded notes or remarks

(Supplementary Table 1) identified the actual species of the sample in a number of instances. For example, N057 was correctly identified as *B. juncea* based on 2010 phenotype data, and likewise N045, N046, N047 and N048 were suspected to be *B. juncea* or *B. rapa* rather than *B. napus* on the basis of phenotype and were confirmed as *B. rapa* by the SNP molecular data. These findings highlight the significance of obtaining phenotypic data wherever possible as a complement to molecular marker results, and support the important role of expert curators in managing germplasm material.

One of the most surprising and interesting results was the presence in the germplasm collection of several individuals clearly originating from interspecific hybridisation events. Although this is a common method for crop improvement in the *Brassica* genus (Chen *et al.* 2011; Navabi *et al.* 2010b; Rygulla *et al.* 2007; Seyis *et al.* 2003; Zou *et al.* 2011), and all species assessed in this experiment are known to be able to hybridise (FitzJohn *et al.* 2007), lines resulting from interspecific hybridisation events seem unlikely candidates for donation to a germplasm collection, at least without explicit labelling. Hence, it seems likely that these events were spontaneous and originated as a result of cross-contamination during seed bulking processes. We observed one very clear case of interspecific hybridisation in putative *B. rapa* individual R14, which contained a partial C genome in addition to a complete A genome. Confusingly, chromosome counting of another individual resulting from the same seed packet revealed only 20 chromosomes, suggesting either that C genome fragments were still present in a heterozygous state or that only some individuals from this line were carrying these introgressions. Indirect but compelling evidence for hybridisation between *B. juncea* and *B. napus* was obtained for individuals J06 and J08: both were classified as *B. juncea* but also showed presence of a complete C genome; both had much higher A genome heterozygosity than average (45 and 49%) but normal C genome heterozygosity (Supplementary Table 1), and both fell outside the *B. juncea* – *B. napus* groups in the PCA. Individual N019a was also a strong candidate for an interspecific hybridisation event between *B. napus* and *B. carinata*: individual N019b from the same Australian Grains Genebank line but sourced separately was conclusively *B. carinata*, N019a clustered within the *B. oleracea/B. carinata* clade in both the PCA and hierarchical clustering analysis and N019a also had disproportionately high C genome heterozygosity (43%) but normal A genome heterozygosity (9%). Individuals J22 and J05 both also contained an A and a C genome, but grouped strongly with *B. carinata* samples using both PCA and hierarchical clustering. These putative interspecific hybridisation events are plausible: accessions in genebanks are often sown in close proximity, and accidental cross-pollination could occur. Hybridisation between the allotetraploid species is relatively easy when carried out by hand pollination (Mason *et al.* 2011) and interspecific hybrids between the allotetraploids are capable of producing seed when self-pollinated (Mason *et al.* 2011) and when back-crossed to the parent species (Chèvre *et al.* 1997; Navabi *et al.* 2010a). Accessions of different *Brassica* species are often grown adjacently during seed regeneration, allowing opportunity for natural cross-pollination to occur.

High-throughput genotyping using molecular resources such as SNP chip arrays and genotyping-by-sequencing is becoming both readily accessible and cost-effective for large sample sizes and complex crop genomes (Edwards& Batley 2010; Edwards *et al.* 2013). As demonstrated in our study by identification of A- and C-genome-specific SNPs, the availability of reference genome sequences can also dramatically increase the effectiveness of standard molecular marker approaches. We provide validation of the Illumina Infinium Brassica 60K SNP array for species classification in germplasm collections, and suggest that similar high-throughput SNP genotyping approaches should be carried out in future in germplasm collections to support these valuable resources for research and breeding.

Acknowledgements

We thank Rowan Bunch for Illumina HiScan SNP chip scanning. The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462 and LP0989200). ASM was supported by an Australian Research Council Discovery Early Career Researcher Award (DE120100668). JB acknowledges funding support from the Australian Research Council (Projects LP110100200, LP130100061, LP130100925, FT130100604 and DP0985953).

374 **Table 1: Species identity as confirmed by SNP molecular genotyping in a set of *Brassica* samples and related species sourced from the Australian Grains Genebank.**

Germplasm collection species	Confirmed species	No. samples	% accuracy overall
B. napus	*B. napus*	95	
B. napus	*B. rapa*	8	
B. napus	*B. juncea*	3	
B. napus	*B. carinata*	9	
	Subtotal	**115**	**83%**
B. rapa	*B. rapa*	20	
B. rapa	*B. juncea*	1	
	Subtotal	**21**	**95%**
B. oleracea	*B. oleracea*	3	
	Subtotal	**3**	**100%**
B. carinata	*B. carinata*	3	
	Subtotal	**3**	**100%**
B. juncea	*B. juncea*	25	
B. juncea	*B. rapa*	2	
B. juncea	*B. napus*	5	
	Subtotal	**32**	**77%**
B. nigra	*B. nigra*	1	
B. nigra	*B. juncea*	1	
Sinapis alba	*Sinapis alba*	1	
Sinapis alba	*B. nigra*	1	
Sinapis alba	*B. carinata*	1	
Raphanus sativus	*B. napus*	1	
	Subtotal	**6**	**33%**
	TOTAL	**180**	**82%**

Figure Legends

Figure 1: Presence of the *Brassica* A and C genomes using SNP markers in a set of Brassicaceae samples sourced from a germplasm collection: 32 putative *B. juncea* samples, 21 putative *B. rapa* samples, 115 putative *B. napus* samples, 3 putative *B. oleracea* samples, 3 putative *B. carinata* samples, 3 putative *B. nigra* samples, 3 putative *Sinapis alba* samples and 1 putative *Raphanus sativus* sample. Three anomalous samples are observed outside the tight genome clusters.

Figure 2: Separation of *Brassica rapa* and *B. juncea* samples using A genome SNP data from the Illumina Infinium Brassica 60K array. Dendrogram generated using default hierarchical clustering in package and function "pvclust" in R v 3.0 using n = 1000 iterations; "au" and "bp" refer to the "approximately unbiased" and "bootstrap probability" p-values for each branch. Control samples from confirmed species genotypes are labelled with "Control_" followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species ("J" for *B. juncea*, "R" for *B. rapa*, "I" for *B. nigra*, "N" for *B. napus* (supplied as *B. napus* but containing only an A genome), and "XS" for non-*Brassica, Sinapis alba* (also containing an A genome)). Individual plants from the same genotype are labelled with the same number but different lowercase letters. Chromosome-counted samples are indicated by red stars.

Figure 3: Separation of *B. rapa* and *B. juncea* samples using Principle Components Analysis (first two axes plotted, explaining 18.2% and 13.7% of the variance respectively). Control samples from confirmed species genotypes are labelled with "Control" followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species ("J" for *B. juncea*, "R" for *B. rapa*, "I" for *B. nigra*, "N" for *B. napus* (supplied as *B. napus* but containing only

an A genome), and “XS” for non-*Brassica, Sinapis alba* (also containing an A genome)). Individual plants from the same genotype are labelled with the same number but different lowercase letters. Red stars indicate chromosome-counted samples. Individual R014 was anomalous (putatively *B. rapa*) with C-genome introgressions in an A-genome background.

Figure 4: Separation of *Brassica oleracea* and *B. carinata* samples using C genome SNP data from the Illumina Infinium Brassica 60K array. Dendrogram generated using default hierarchical clustering in package and function “pvclust” in R v 3.0 using n = 1000 iterations; “au” and “bp” refer to the “approximately unbiased” and “bootstrap probability” p-values for each branch. Control samples from confirmed species genotypes are labelled with “Control_” followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species (“N” for *B. napus* (supplied as *B. napus* but with no A genome), “O” for *B. oleracea*, “C” for *B. carinata* and “XS” for non-*Brassica, Sinapis alba*). Individual plants from the same genotype are labelled with the same number but different lowercase letters. A chromosome-counted sample is indicated with a red star.

Figure 5: Separation of *B. oleracea* and *B. carinata* samples using Principle Components Analysis (first two axes plotted, explaining 41.3% and 13.0% of the variance respectively). Control samples from confirmed species genotypes are labelled with “Control” followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species (“N” for *B. napus* (supplied as *B. napus* but with no A genome), “O” for *B. oleracea*, “C” for *B. carinata* and “XS” for non-*Brassica, Sinapis alba*). Individual plants from the same genotype are labelled with the same number but different lowercase letters. The red star indicates a chromosome-counted sample.

Figure 6: A genome diversity as assessed by Principle Components Analysis of Illumina Infinium 60k Brassica array data in a set of 31 A-genome controls of known species origin and 162 *B. rapa*, *B. juncea* and *B. napus* samples found to contain an A genome and originating from the Australian Grains Genebank. Experimental samples are labelled by a letter representing the supplied species ("J" for *B. juncea* and "R" for *B. rapa*).

Figure 7: C genome diversity as assessed by principle components analysis of Illumina Infinium 60k Brassica array data from a set of 29 C-genome controls of known species origin (2 *B. oleracea*, 4 *B. carinata* and 23 *B. napus*) and 117 *B. carinata*, *B. oleracea* and *B. napus* samples all containing a C genome and originating from the Australian Grains Genebank. Control samples from confirmed species genotypes are labelled with "Control" followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species: "N" for *B. napus*, "O" for *B. oleracea* and "J" for *B. juncea* (supplied as *B. juncea* but containing an A and a C genome and hence actually *B. napus*). Individual plants from the same genotype are labelled with the same number but different lowercase letters.

Figure 8: Chromosome counts for two putative *Brassica napus* plants (N089 and N067) showing 2n = 34 (*B. carinata*) and 2n = 36 (*B. juncea*) respectively; and three *B. rapa* individuals (R05, R14 and R21) showing 2n = 20. Bar = 10 µm

References

Arias T, Beilstein MA, Tang M, McKain MR, Pires JC (2014) Diversification times among *Brassica* (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. *American Journal of Botany* **101**, 86-91.

Chen S, Nelson MN, Chèvre A-M*, et al.* (2011) Trigenomic bridges for *Brassica* improvement. *Critical Reviews in Plant Sciences* **30**, 524-547.

Chèvre AM, Barret P, Eber F*, et al.* (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. *Theoretical and Applied Genetics* **95**, 1104-1111.

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

Dangl GS, Mendum ML, Prins BH*, et al.* (2001) Simple sequence repeat analysis of a clonally propagated species: A tool for managing a grape germplasm collection. *Genome* **44**, 432-438.

Day PR (1973) Genetic variability of crops. *Annual Review of Phytopathology* **11**, 293-312.

Dixon GR (2007) Vegetable Brassicas and related crucifers. In: *Crop production science in horticulture series* (eds. Atherton J, Rees H). CAB International, Oxfordshire, UK.

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

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. *Theoretical and Applied Genetics* **126**, 1-11.

Ferriol M, Pico B, Nuez F (2003) Genetic diversity of a germplasm collection of *Cucurbita pepo* using SRAP and AFLP markers. *Theoretical and Applied Genetics* **107**, 271-282.

FitzJohn RG, Armstrong TT, Newstrom-Lloyd LE, Wilton AD, Cochrane M (2007) Hybridisation within *Brassica* and allied genera: evaluation of potential for transgene escape. *Euphytica* **158**, 209-230.

Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Molecular Biology Reporter* **13**, 207-209.

Harberd DJ, McArthur ED (1980) Meiotic analysis of some species and genus hybrids in the Brassiceae. In: *Brassica Crops and Wild Allies: Biology and Breeding* (ed. Tsunoda S, Hinata, K., Gomez-Campo, C.), pp. 65-87. Japan Scientific Societies Press, Tokyo.

Hyten DL, Song QJ, Zhu YL*, et al.* (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 16666-16671.

Kaur P, Banga S, Kumar N*, et al.* (2014) Polyphyletic origin of *Brassica juncea* with *B. rapa* and *B. nigra* (Brassicaceae) participating as cytoplasm donor parents in independent hybridization events. *American Journal of Botany* **101**, 1157-1166.

Lee GA, Sung JS, Lee SY*, et al.* (2014) Genetic assessment of safflower (*Carthamus tinctorius* L.) collection with microsatellite markers acquired via pyrosequencing method. *Molecular Ecology Resources* **14**, 69-78.

Li HT, Younas M, Wang XF*, et al.* (2013) Development of a core set of single-locus SSR markers for allotetraploid rapeseed (*Brassica napus* L.). *Theoretical and Applied Genetics* **126**, 937-947.

Martin C, Juliano A, Newbury HJ*, et al.* (1997) The use of RAPD markers to facilitate the identification of *Oryza* species within a germplasm collection. *Genetic Resources and Crop Evolution* **44**, 175-183.

Mason AS, Nelson MN, Takahira J*, et al.* (2014) The fate of chromosomes and alleles in an allohexaploid *Brassica* population. *Genetics* **197**, 273-283.

Mason AS, Nelson MN, Yan GJ, Cowling WA (2011) Production of viable male unreduced gametes in *Brassica* interspecific hybrids is genotype specific and stimulated by cold temperatures. *BMC Plant Biology* **11**, 103.

Morinaga T (1934) Interspecific hybridisation in *Brassica* VI. The cytology of F_1 hybrids of *B. juncea* and *B. nigra*. *Cytologia* **6**, 62-67.

Navabi ZK, Parkin IA, Pires JC*, et al.* (2010a) Introgression of B-genome chromosomes in a doubled haploid population of *Brassica napus* x *B. carinata*. *Genome* **53**, 619-629.

Navabi ZK, Strelkov SE, Good AG, Thiagarajah MR, Rahman MH (2010b) *Brassica* B-genome resistance to stem rot (*Sclerotinia sclerotiorum*) in a doubled haploid population of *Brassica napus* x *Brassica carinata*. *Canadian Journal of Plant Pathology-Revue Canadienne De Phytopathologie* **32**, 237-246.

Palmer JD, Shields CR, Cohen DB, Orten TJ (1983) Chloroplast DNA evolution and the origin of amphidiploid *Brassica* species. *Theoretical and Applied Genetics* **65**, 181-189.

Parkin IA, Koh C, Tang H*, et al.* (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea Genome Biology* **15**, R77.

Pradhan A, Nelson MN, Plummer JA, Cowling WA, Yan GJ (2011) Characterization of *Brassica nigra* collections using simple sequence repeat markers reveals distinct groups associated with geographical location, and frequent mislabelling of species identity. *Genome* **54**, 50-63.

Prakash S, Takahata Y, Kirti PB, Chopra VL (1999) Cytogenetics. In: *Biology of Brassica coenospecies* (ed. Gómez-Campo C), pp. 59-106. Elsevier Science B.V., Amsterdam.

Romay MC, Millard MJ, Glaubitz JC*, et al.* (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology* **14**.

Rygulla W, Friedt W, Seyis F*, et al.* (2007) Combination of resistance to *Verticillium longisporum* from zero erucic acid *Brassica oleracea* and oilseed *Brassica rapa* genotypes in resynthesized rapeseed (*Brassica napus*) lines. *Plant Breeding* **126**, 596-602.

Saal B, Brun H, Glais I, Struss D (2004) Identification of a *Brassica juncea*-derived recessive gene conferring resistance to *Leptosphaeria maculans* in oilseed rape. *Plant Breeding* **123**, 505-511.

Seyis F, Snowdon RJ, Luhs W, Friedt W (2003) Molecular characterization of novel resynthesized rapeseed (*Brassica napus*) lines and analysis of their genetic diversity in comparison with spring rapeseed cultivars. *Plant Breeding* **122**, 473-478.

Simmonds NW (1962) Variability in crop plants, its use and conservation. *Biological Reviews of the Cambridge Philosophical Society* **37**, 422-&.

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* **277**, 1063-1066.

U N (1935) Genome-analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Japanese Journal of Botany* **7**, 389-452.

Wang XW, Wang HZ, Wang J*, et al.* (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics* **43**, 1035-1039.

Warwick SI, Simard M-J, Légère A*, et al.* (2003) Hybridization between transgenic *Brassica napus* L. and its wild relatives: *Brassica rapa* L., *Raphanus raphanistrum* L., *Sinapis arvensis* L., and *Erucastrum gallicum* (Willd.) O.E. Schulz. *Theoretical and Applied Genetics* **107**, 528-539.

Zamir D (2001) Improving plant breeding with exotic genetic libraries. *Nature Reviews Genetics* **2**, 983-989.

Zou J, Fu DH, Gong HH*, et al.* (2011) *De novo* genetic variation associated with retrotransposon activation, genomic rearrangements and trait variation in a recombinant inbred line population of *Brassica napus* derived from interspecific hybridization with *Brassica rapa*. *Plant Journal* **68**, 212-224.

Data Accessibility

The Illumina Infinium Brassica 60K SNP array used in this analysis can be obtained from Illumina Inc. (http://www.illumina.com/). Summary information for each Australian Germplasm Genebank accession used in this analysis is provided in Supplementary Table 1. Genotype data and SNP information is provided in Supplementary Table 2 and this data is also available *via* the Dryad data repository (doi:10.5061/dryad.c3g5r). Seeds for each of the lines used can be obtained from the Australian Germplasm Genebank. PCA and hierarchical clustering analyses were performed using the R base software and packages "pvclust", "ade4" and "gam" freely available from the R Project for Statistical Computing (http://www.r-project.org/).

Author Contributions

JB, DE and BR conceptualised the study. JB managed the project. BR, GY, JZ and LH contributed material. RT and JZ grew up seeds and extracted DNA. PVT carried out chromosome counting. JDM ran the SNP chip. ASM analysed the SNP chip data, generated the figures and tables and wrote the paper. JB, DE, BR and GY critically revised the manuscript. All authors have read and approved the final version of the manuscript.

Supporting Information

Supplementary Figure 1: Presence of the *Brassica* A and C genomes in a set of known control samples using SNP markers: 3 *B. rapa* (A genome only); 6 *B. juncea* (A genome only), 23 *B. napus* (A+C genomes)*,* 2 *B. oleracea* (C genome only), 4 *B. carinata* (C genome only) and 5 *Raphanus sativus* (neither genome)*.*

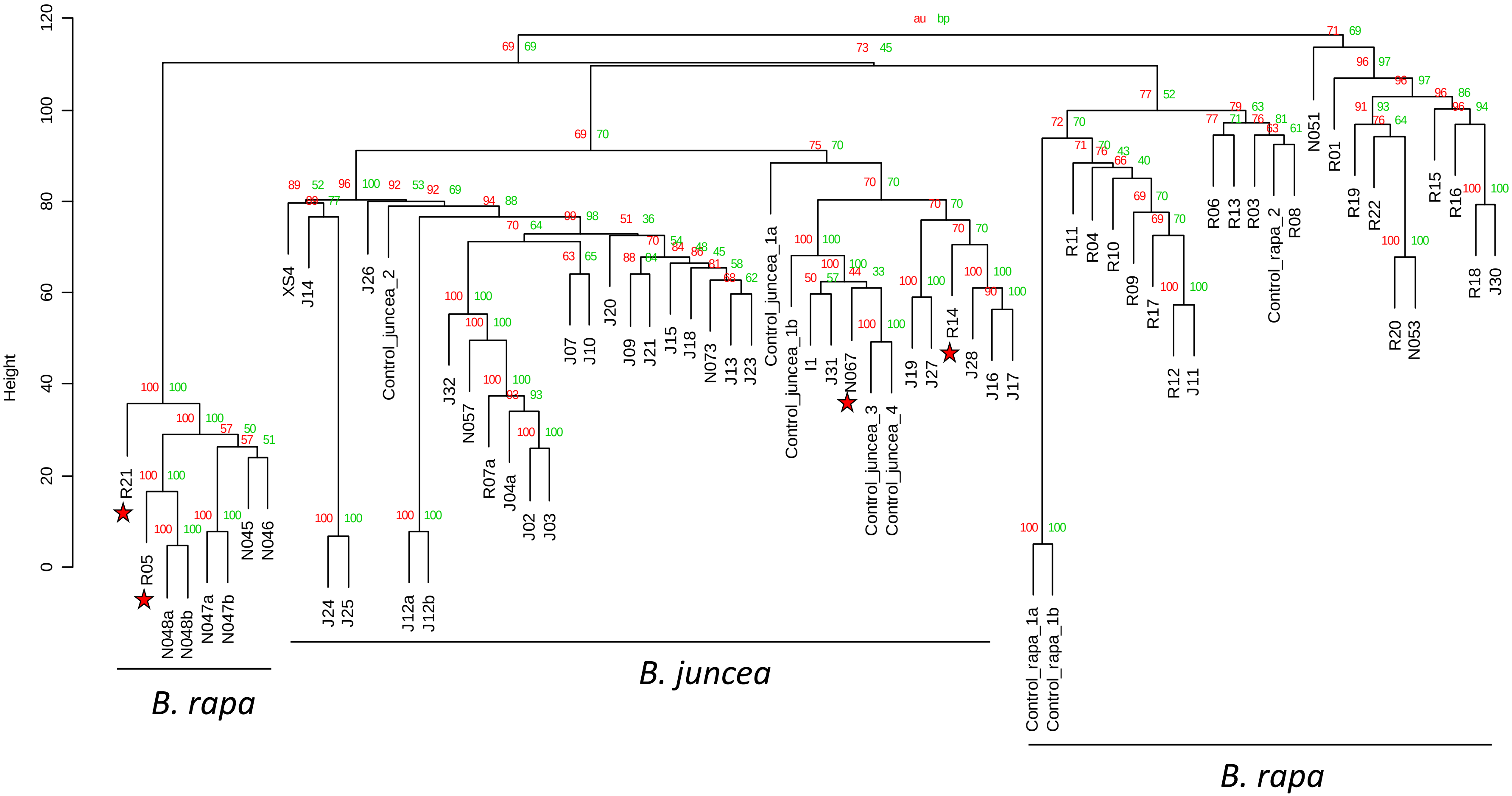
Supplementary Figure 2: A genome diversity as assessed by hierarchical clustering of Illumina Infinium 60k Brassica array data in a set of 31 controls of known species origin and 162 *B. rapa*, *B. juncea* and *B. napus* lines originating from the Australian Grains Genebank. Control samples from confirmed species genotypes are labelled with "Control" followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species ("J" for *B. juncea*, "R" for *B. rapa*, "I" for *B. nigra*, "N" for *B. napus* and "XS" for non-*Brassica, Sinapis alba* (also containing an A genome)). Individual plants from the same genotype are labelled with the same number but different lowercase letters. Red stars indicate chromosome-counted samples. Chromosome-counted lines are indicated by red stars, and samples of interest are indicated using blue four-pointed stars.

Supplementary Figure 3: C genome diversity as assessed by hierarchical clustering of Illumina Infinium 60k Brassica array data from a set of 29 controls of known species origin and 117 *B. carinata*, *B. oleracea* and *B. napus* lines originating from the Australian Grains Genebank. Control samples from confirmed species genotypes are labelled with "Control" followed by the species and a genotype designation; experimental samples are labelled by a letter representing the supplied species ("N" for *B. napus*, "O" for *B. oleracea*, "C" for *B. carinata*, "J" for *B. juncea* (but containing both an A and C genome and hence actually *B. napus*) and "XS" for non-*Brassica, Sinapis alba*).

Individual plants from the same genotype are labelled with the same number but different lowercase letters. Chromosome-counted lines are indicated by red stars, and samples of interest are indicated using blue four-pointed stars.

Supplementary Table 1: Information for the set of 188 experimental samples sourced from the Australian Grains Genebank: sample identification numbers, provided information, genome amplification results and species re-classifications based on SNP analyses.

Supplementary Table 2: SNP molecular genotyping data and information.


Height
0
20
40
60
80
100
120
au
bp
R21
R05
N048a
N048b
N047a
N047b
N045
N046
XS4
J14
J24
J25
J26
Control_juncea_2
J12a
J12b
J32
N057
R07a
J04a
J02
J03
J07
J10
J20
J09
J21
J15
J18
N073
J13
J23
Control_juncea_1a
Control_juncea_1b
I1
J31
N067
Control_juncea_3
Control_juncea_4
J19
J27
R14
J28
J16
J17
Control_rapa_1a
Control_rapa_1b
R11
R04
R10
R09
R17
R12
J11
R06
R13
R03
Control_rapa_2
R08
N051
R01
R19
R22
R20
N053
R15
R16
R18
J30
B. rapa
B. juncea
B. rapa