Simple Sequence Repeat (SSR) and Single Nucleotide Polymorphism (SNP) Diversity in the Brassicaceae

CILR ARC Centre of Excellence

The University of Queensland Australia

Emma Campbell[1], Megan McKenzie[1], Alice Hayward[1], Reece Tollenaere[1], Paul Berkman[2], Chris Duran[2], Wu Jian[3], Liu Bo[3], Xiaowu Wang[3], Robert Redden[4], David Edwards[2] and Jacqueline Batley[1]*

[1]ARC Centre of Excellence for Integrative Legume Research and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia
[2]Australian Centre for Plant Functional Genomics, School of Land, Crop and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia
[3]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[4]Department of Primary Industries, Private Bag 260, Horsham Vic 3401, Australia
* j.batley@uq.edu.au

Introduction

Single Nucleotide Polymorphisms (SNPs) are the most abundant genetic marker, being both highly stable and reliable for high-throughput discovery and detection. They are predominantly applied, along with Simple Sequence Repeats (SSRs), in modern plant genetic analysis, genetic trait mapping, diversity and association studies, and marker assisted selection. Here we present the application of these markers for genetic diversity analysis in *Brassica* species and wild relatives from an international germplasm collection. An assessment of genetic and genomic diversity within these collections can be used to study the evolutionary history of wild relatives, assign lines to heterotic groups and help narrow the search for new alleles at loci of interest.

SNP Discovery, Validation and Screening

The SNPs in this study were discovered from *Brassica rapa* second generation re-sequencing data. The SNPs were predicted to be real using alignment of short read tags to a reference sequence in TAGdb (http://flora.acpfg.com.au/tagdb). The SNP prediction was validated using direct sequencing. Primers for five validated SNP regions were used for PCR amplification in wild and cultivated Brassicaceae accessions, representing 28 species, obtained from an international germplasm collection. Sequence data was analysed to assess the transferability across the species, identify genetic diversity and estimate phylogeny within the Brassicaceae. The five validated SNPs were also aligned to the *B. oleracea* and *B. nigra* tags in TAGdb to assess genome conservation.

SNP Results

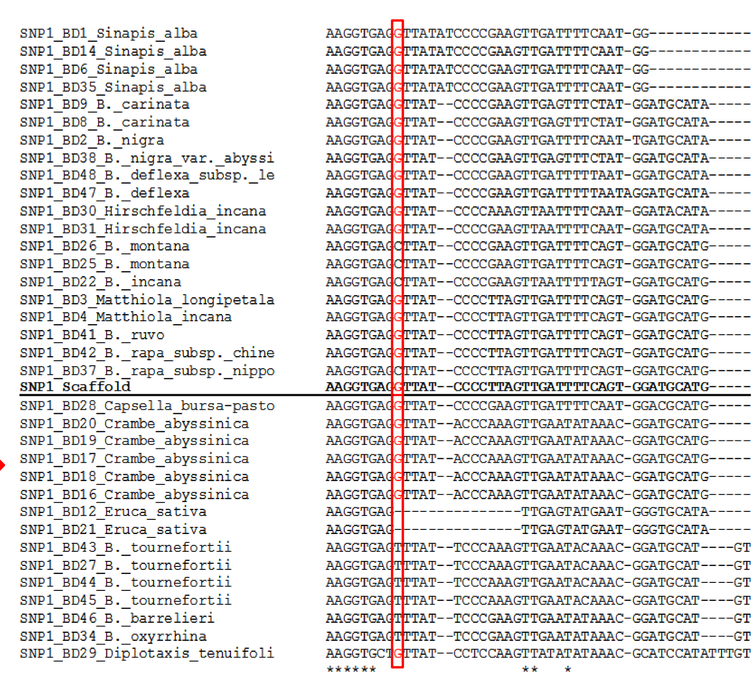
When screened across the Brassicaceae, the SNP loci amplified best in the primary and secondary progenitors of the cultivated *Brassicas*, while only a selection of the tertiary progenitors amplified (data not shown). All loci demonstrated conservation across the cultivated species, through PCR amplification and the identification of tags from the B and C genomes in TAGdb. Between 19 and 43 samples amplified with the 5 primer sets (mean = 70.8%). In general, PCR amplification in the secondary progenitors was weaker than the primary progenitors and cultivated *Brassicas.* Direct sequencing of these products further assessed the level of genome conservation and the suitability of these SNP markers. The sequence conservation varied within the SNP flanking sequence, identifying numerous indels and additional SNPs (Figure 1C).

A. SNP *in silico* validation

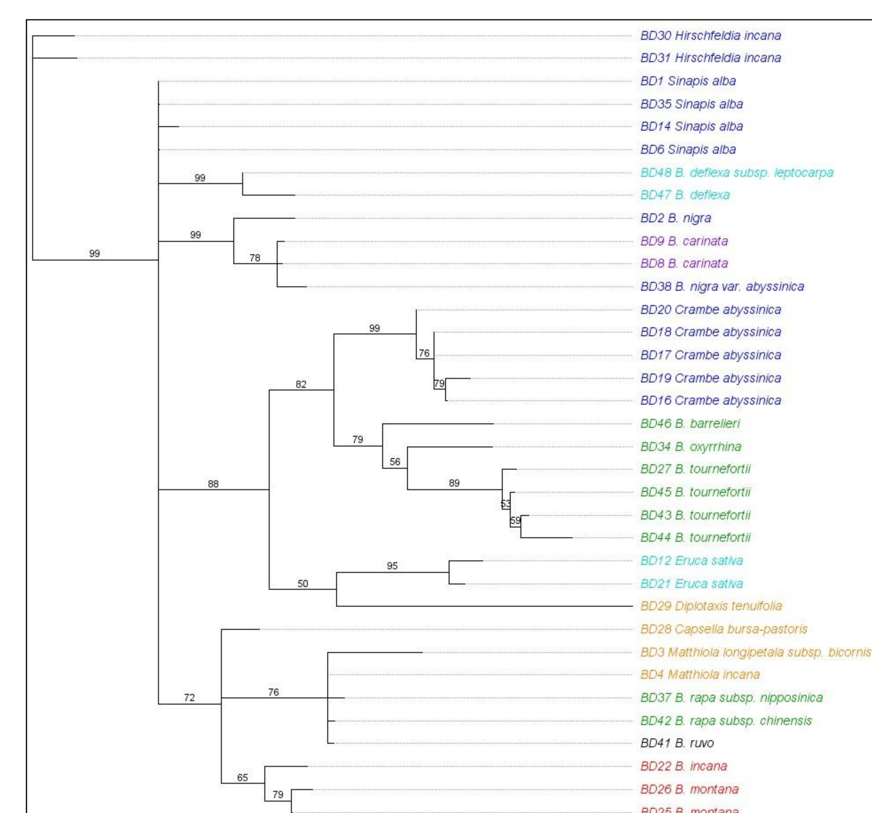
SNP	Primer name	Sequence	Predicted polymorphism	No. *Arabidopsis* hits	No. *B. rapa* tags	SNP Prediction	SNP validation
SNP1	SNP1F	AAAGGTTGCTATTCATGAAGCC		0	108	Real	Real
	SNP1R	AACCTCGAGAGAATGGCAGGT	G/C				
SNP2	SNP2F	CCCTGAGTAGATTATCCCTG		0	117	Real	Real
	SNP2R	TCTCAAGATCTTTACCAAAT	C/G				
SNP3	SNP3F	TCCGGTTACTACCAGCCAAC			92	Real	Real
	SNP3R	TGATGTAGCAATCAATACCTT	T/C	0			
SNP4	SNP4F	GTCGGTGTCTTTCATCTTCA		0	100	Real	Real
	SNP4R	GTTGAACTCAGCAAACTATAC	T/G				
SNP5	SNP5F	ACCATAAACTTGCTGAACGTAT		1	86	Real	Real
	SNP5R	ACTCTCTGTATGTTTGTATAA	G/C				

B. SNP amplification

Taxon	Genebank	Sample type	Gene Pool	SNP1	SNP2	SNP3	SNP4	SNP5
B. barrelieri	Spain	wild	secondary[2]	Y	N	Y	untested	Y
B. carinata	Ethiopia	landrace	allotetraploid	Y	Y	Y	Y	Y
B. carinata	Pakistan	landrace	allotetraploid	Y	Y	Y	Y	Y
B. deflexa	Turkey	wild	secondary[2]	Y	N	Y	untested	Y
B. deflexa subsp. leptocarpa	Iran	wild	secondary[2]	Y	N	Y	untested	Y
B. incana (94671)	Italy	wild	primary[2]	Y	Y	Y	Y	Y
B. incana (94696)	Italy	wild	primary[2]	Y	N	Y	Y	Y
B. juncea	China	advanced cultivar	allotetraploid	Y	Y	Y	Y	Y
B. juncea	Pakistan	breeder's line	allotetraploid	Y	Y	Y	Y	Y
B. maurorum	Algeria	weed	primary[1]	Y	N	N	untested	N
B. montana	Italy	wild	primary[2]	Y	Y	Y	Y	Y
B. montana	France	wild	primary[2]	Y	Y	Y	Y	N
B. nigra	India	landrace	diploid progenitor	Y	N	Y	Y	Y
B. nigra	Ethiopia	landrace	diploid progenitor	N	N	Y	untested	Y
B. nigra var. abyssinica	Ethiopia	traditional cultivar/landrace	diploid progenitor	Y	Y	Y	untested	Y
B. oleracea var. gongylodes	Austria	unknown	diploid progenitor	Y	Y	Y	Y	Y
B. oxyrrhina	Australia	wild	secondary[1]	Y	N	N	untested	N
B. rapa subsp. campestris	Algeria	landrace	diploid progenitor	Y	Y	Y	untested	Y
B. rapa subsp. chinensis	China	traditional cultivar/landrace	diploid progenitor	Y	Y	Y	untested	Y
B. rapa subsp. nipposinica	Japan	traditional cultivar/landrace	diploid progenitor	Y	Y	Y	untested	Y
B. ruvo	Italy	advanced cultivar oleraceae	primary [2]	Y	Y	Y	untested	Y
B. souliei subsp. amplexicaulis	Australia	unknown	secondary[1]	Y	Y	Y	Y	Y
B. spinescens	Algeria	wild	primary [1]	Y	N	Y	untested	Y
B. tournefortii	Australia	wild	secondary[1/2]	Y	Y	Y	untested	Y
B. tournefortii	France	wild	secondary[1/2]	Y	N	Y	untested	Y
B. tournefortii	Spain	wild	secondary[1/2]	Y	N	Y	untested	Y
B. tournefortii	India	wild	secondary[1/2]	Y	N	Y	untested	Y
Capsella bursa-pastoris	Australia	wild	tertiary	Y	Y	Y	untested	Y



C. SNP Flanking sequence analysis



D. Flanking sequence and SNP genotyping dendrogram

Figure 1. The screening process, from SNP validation (A) and testing, (B) to amplification and sequencing of the Brassicaceae (C) and phylogenetic analysis (D). The SNP flanking sequence variation shows samples grouping into their proposed A, B and C genome lineages.

SNP Genotyping

Phylogenetic analysis of the SNP genotyping and SNP flanking sequence data indicate that the Brassicaceae grouped primarily within the proposed A, B and C lineages (Figures 1D and 2). In the case of the SNP genotyping dendrogram (Figure 2) the *Eruca sativa* samples form a small outgroup due to an indel in one of the loci (as seen in Figure 1C).

There is a major cluster of the A genome samples, and the three samples from the C genome also group. The largest cluster seen is primarily composed of B genome samples and some tertiary progenitors. The *B. deflexa* consistently groups with the B genome (data not shown), indicating that this species may have been assigned to the incorrect lineage, which will require further investigation.

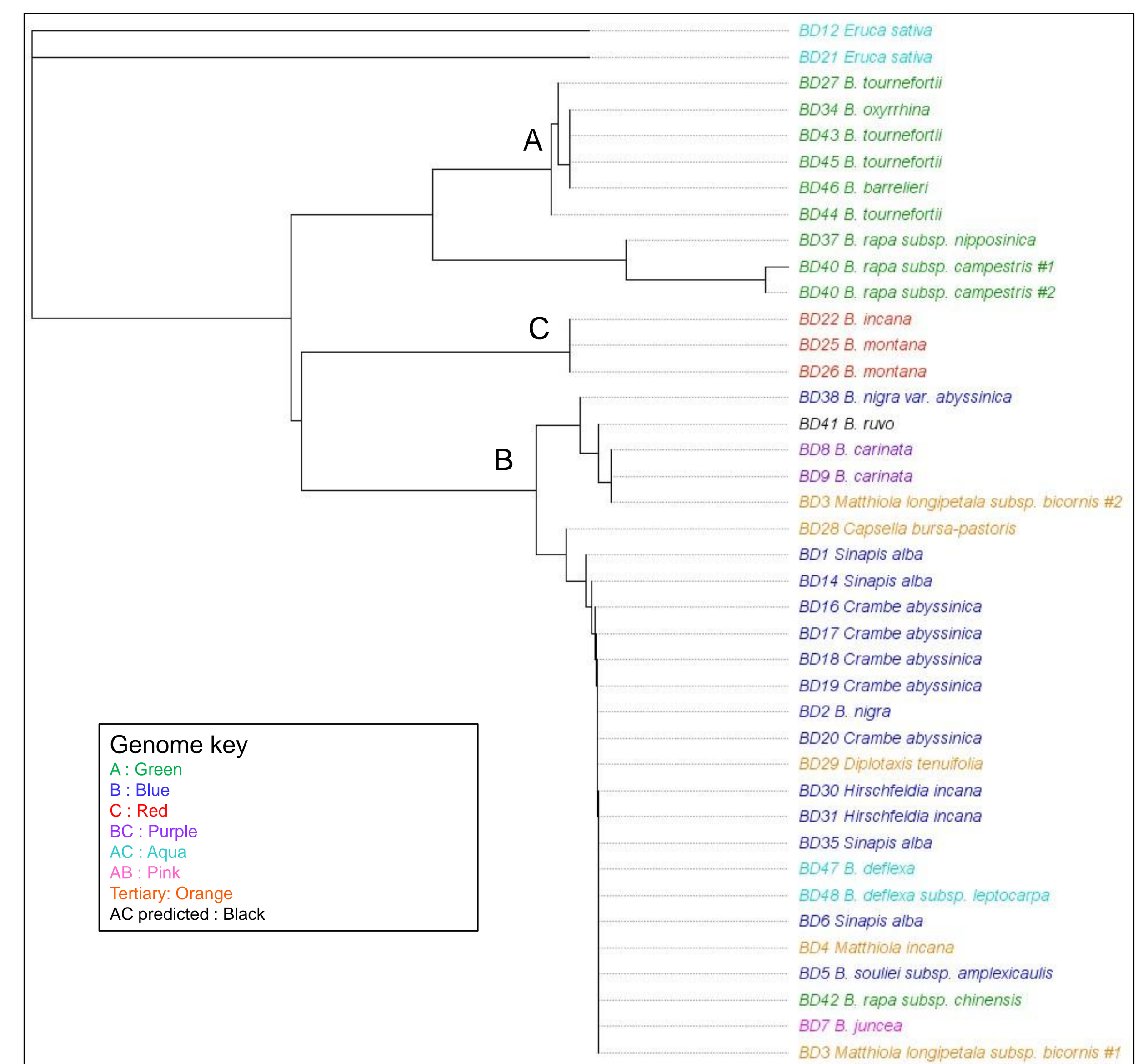


Figure 2. Scored SNP genotypes for the five primers were compared in a phylogenetic tree, the Brassicaceae samples are coloured according to the proposed A, B, C or amphidiploid lineage.

SSRs

Selected SSR markers, representing linkage groups across the A, B and C genomes, were screened against 48 Brassicaceae lineages generating 562 alleles. This data is currently being analysed along with the SNP data to clarify the genetic relationship between the lineages.

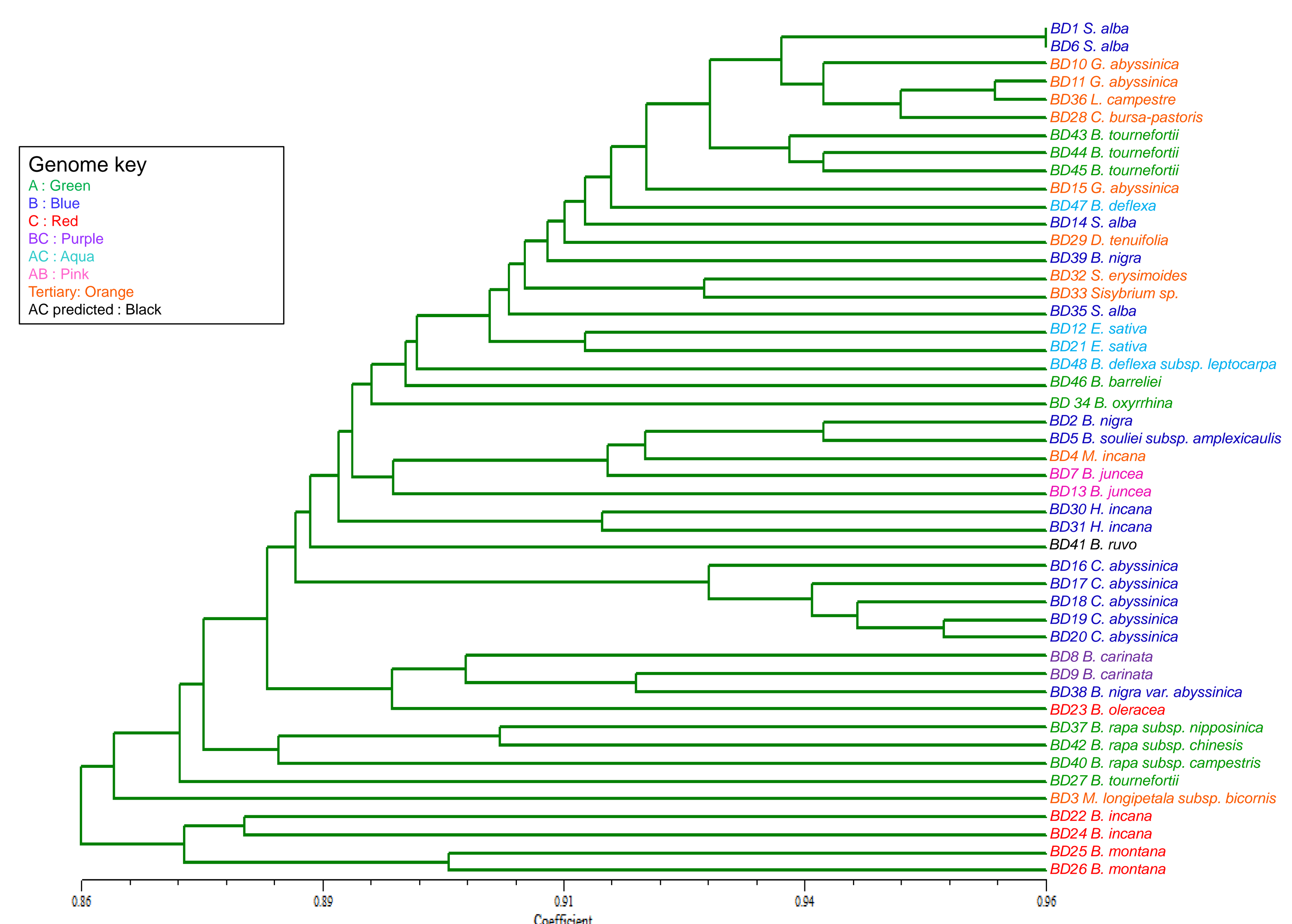


Figure 3. Phylogenetic tree using 562 SSR alleles scored against 48 Brassicaceae samples.

Phylogenetic analysis of the SSR alleles across the Brassicaceae showed the species grouping together in the dendrogram (Figure 3) with some clustering according to genome origin. The samples which do not group as expected will be investigated further to confirm the origin of the samples.

Conclusion

We have demonstrated that SNPs and SSRs can be utilised for genetic and genomic analysis of the Brassicaceae. With a small number of markers relationships between the genomes can begin to be resolved. SNPs from *B. rapa* can be used to successfully genotype the Brassicaceae and a larger panel of markers is currently being genotyped.