
Article:
Harper, Andrea Louise orcid.org/0000-0003-3859-1152, He, Zhesi orcid.org/0000-0001-8335-9876, Langer, Swen et al. (6 more authors) (2020) Validation of an Associative Transcriptomics platform in the polyploid crop species Brassica juncea by dissection of the genetic architecture of agronomic and quality traits. The Plant journal. pp. 1-9. ISSN 1365-313X

https://doi.org/10.1111/tpj.14876

---

the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) doi: 10.1111/tpj.14876

Validation of an Associative Transcriptomics platform in the polyploid crop species *Brassica juncea* by dissection of the genetic architecture of agronomic and quality traits

Andrea L. Harper[1,*], Zhesi He[1], Swen Langer[1], Lenka Havlickova[1], Lihong Wang[1], Alison Fellgett[1], Vibha Gupta[2], Akshay Kumar Pradhan[2] and Ian Bancroft[1]

[1]*Department of Biology, University of York, Heslington, York YO10 5DD, UK, and*

[2]*Centre for Genetic Manipulation of Crop Plants, University of Delhi South Campus, New Delhi 110021, India*

Received 12 February 2019; revised 22 May 2020; accepted 1 June 2020.

*For correspondence (e-mail: andrea.harper@york.ac.uk).

ABSTRACT

The development of more productive crops will be key to addressing the challenges that climate change, population growth and diminishing resources pose to global food security. Advanced 'omics techniques can help to accelerate breeding by facilitating the identification of genetic markers for use in marker-assisted selection. Here, we present the validation of a new Associative Transcriptomics platform in the important oilseed crop *Brassica juncea*. To develop this platform, we established a pan-transcriptome reference for *B. juncea*, to which we mapped transcriptome data from a diverse panel of *B. juncea* accessions. From this panel, we identified 355 050 single nucleotide polymorphism variants and quantified the abundance of 93 963 transcripts. Subsequent association analysis of functional genotypes against a number of important agronomic and quality traits revealed a promising candidate gene for seed weight, *BjA.TTL*, as well as additional markers linked to seed colour and vitamin E content. The establishment of the first full-scale Associative Transcriptomics platform for *B. juncea* enables rapid progress to be made towards an understanding of the genetic architecture of trait variation in this important species, and provides an exemplar for other crops.

Keywords: associative transcriptomics, *Brassica juncea*.

INTRODUCTION

Improving crop productivity is becoming increasingly important as the pressures of climate change, population growth and diminishing resources mount, although recent advances in genomic technologies are providing new tools for crop improvement. A key aim in pre-breeding of new crops is the identification of the genetic bases for trait variation to identify beneficial alleles and to develop molecular markers for accelerating their introduction into elite cultivars. Increasingly, genetic diversity panels are being utilised to meet this aim because they represent ideal resources for genome-wide association studies, which exploit historical recombination between molecular markers and trait-controlling loci. The approach of identifying molecular markers in linkage disequilibrium with trait-controlling loci is now an established tool in many species, including plants (Atwell *et al.*, 2010; Cockram *et al.*, 2010; Tian *et al.*, 2011; Zhao *et al.*, 2011), although the more recent development of Associative Transcriptomics (AT) (Harper *et al.*, 2012), often provides a more powerful approach for studying plant species because it utilises transcribed sequences as a means of reducing the sequence complexity that can otherwise confound the analysis of large, repetitive or polyploid species. RNA-sequencing data are used to measure transcript abundance, as well as identify gene sequence variants [including hemi-single nucleotide polymorphisms (SNPs), which are unique to polyploid species] and integrating the results of trait associations with these different types of marker can reveal valuable additional information about the causative variation underlying complex traits (Harper *et al.*, 2012).

The Brassicaceae family includes *Arabidopsis thaliana*, the first plant for which a high-quality genome sequence was available (Arabidopsis Genome Initiative, 2000), and the *Brassica* crops. *Brassica juncea* is an allopolyploid, arising from several hybridization events between closely-related diploid species *Brassica rapa* and *Brassica nigra*, from which the *B. juncea* A and B genomes, respectively, are derived (Chen *et al.*, 2013; Kaur *et al.*, 2014). *Brassica juncea* has also been resynthesised using either the original diploid progenitors (Bansal *et al.*, 2009; Bansal *et al.*,

2012), or by combining A and B genomes from allotetraploid species *Brassica napus* and *Brassica carinata* (Gupta *et al.*, 2015). *Brassica juncea* varieties are grown as vegetable and oilseed mustard crops in China, canola crops in Canada and Australia, and as mustard condiments in Europe, Canada and Australia, and are an important oilseed crop in India (Chen *et al.*, 2013).

As a result of their importance as crops and their utility in studying the evolution of polyploid genomes, *Brassica* species have been used extensively in genomics studies (Song *et al.*, 1995; O'Neill and Bancroft, 2000; Pires *et al.*, 2004; Yang *et al.*, 2006; Town *et al.*, 2006; Cheung *et al.*, 2009). A draft genome sequence has been obtained for *B. juncea*; however, at approximately 922 Mb, the genome of *B. juncea* is relatively large (Yang *et al.*, 2016). To address this problem, rapid and cost-effective transcriptome-based technologies, using RNA-sequencing, have been developed and applied for SNP discovery (Trick *et al.*, 2009), linkage mapping and genome characterization (Bancroft *et al.*, 2011), and transcript quantification (Higgins *et al.*, 2012). Indeed, AT was first developed in *B. napus*, with a very small genetic diversity panel enabling the implication of orthologues of *HAG1* in the control of seed glucosinolate content (Harper *et al.*, 2012).

The breeding objectives for *B. juncea* include the simultaneous improvement of agronomic performance, such as yield, lodging, maturity and seed size traits, and quality traits such as high oil content in oilseed varieties alongside a positive nutritional composition. In the present study, we present an updated pan-transcriptome reference for *B. juncea*, which we utilise in the AT analysis of a conventional *B. juncea* diversity panel, as an exemplar for the identification of markers associated with agronomic and quality traits in this crop.

RESULTS

Establishment of the AB pan-transcriptome reference

To support research in *B. napus*, a pan-transcriptome platform representing the A and C genomes has been established to represent the nascent *B. napus* genome (i.e. before gene fractionation and homoeologous exchanges seen in modern cultivars including synthetics) (He *et al.*, 2015, 2017). To develop an equivalent platform for *B. juncea*, we started with pseudomolecules for the A genome as represented by the *B. rapa* Chiifu v2.0 genome sequence (Cai *et al.,* 2017) and for the B genome as represented by the *B. nigra* YZ12151 genome sequence (Yang *et al.*, 2016), then corrected by high-density transcriptome SNP-based linkage mapping as described previously for *B. napus* (He *et al.*, 2015; Havlickova *et al.*, 2018). The published *B. rapa* Chiifu v2.0 and *B. nigra* YZ12151 CDS gene models were mapped onto the respective genome sequence pseudomolecules using BLAST (https://blast.ncbi.nlm.nih.gov) to identify the highest-scoring significant hit (threshold *e*-value 1×10^{-30}). This resulted in the mapping and ordering of 47 656 *B. rapa* CDS models to the A genome and 41 053 *B. nigra* CDS models to the B genome. In total, 79 644 CDS models were annotated in the *B. juncea* T84-66 genome (Yang *et al.*, 2016). Of these, 3423 CDS models that had been anchored to the 18 *B. juncea* pseudomolecules were mapped onto the respective (*B. rapa* and *B. nigra*-based) genome sequence pseudomolecules by BLAST (threshold *e*-value 1×10^{-30}). *Brassica juncea* CDS models mapping redundantly with CDS models derived from *B. rapa* and *B. nigra* (threshold *e*-value 1×10^{-30}) were excluded, resulting in the addition of 1954 and 1469 CDS models to the A and B genomes respectively. Next, CDS models from the *B. juncea* T84-66 genome sequence that did not have significant (threshold *e*-value 1×10^{-30}) BLAST hits in the (*B. rapa* and *B. nigra*-based) genome sequence pseudomolecules were interpolated based on the positions of flanking gene models that did map to the respective A or B genome. This was carried out by combining the sorted location on the *B. juncea* T84-66 chromosome of the *B. juncea* T84-66 CDS models with the mapped location of flanking genes on the *B. rapa* or *B. nigra*-based pseudomolecules. We recognised, by high-density transcriptome SNP-based linkage mapping, that a major source of error for the interpolation was discernible as a breakdown of collinearity with the positions of orthologues of interpolated and flanking genes in the genomes of other members of the Brassicaceae. We therefore introduced a filter to exclude interpolation to positions that were not collinear in either *Arabidopsis thaliana* or *Thellungiella parvula* (which represents well the genome organisation of the diploid progenitor of the mesohexaploid Brassiceae tribe) (Murat *et al.,* 2015). This resulted in the addition of 817 and 1014 further CDS models to the A and B genomes, respectively. The final AB pan-transcriptome resource therefore comprises 93 963 hypothetically ordered CDS models (Data S1) (50 427 in the *Brassica* A genome and 43 536 in the *Brassica* B genome). This represents 14 319 more than the 79 644 CDS models annotated in the published *B. juncea* T84-66 pseudomolecules (Yang *et al.*, 2016) and 5254 more than had been identified in the *B. rapa* and *B. nigra* pseudomolecules.

To assess the accuracy with which the AB ordered pan-transcriptome platform represents the genome organisation of *B. juncea*, Genome-Ordered Graphical Genotypes (GOGGs) (He and Bancroft, 2018) were produced for 106 lines of the *B. juncea* VHDH mapping population (Paritosh *et al.,* 2014) based on transcriptome re-sequencing using paired-end 100-base read length mRNA sequencing data produced using the HiSeq 4000 platform (Illumina Corp., San Diego, CA, USA). Read mapping statistics are shown in Data S2. After reducing complexity by retention of only one mapped marker per gene model and manual removal of mis-mapped markers, the final linkage map comprises 8265 mapped SNP markers, including 5085 in the A genome and 3180 in the B genome. GOGGs produced for the VHDH population using this core set of markers are illustrated in Figure 1, with map details provided in Data S3.

A detailed analysis of the collinearity of the *Brassica* A and B genomes was conducted, using the hypothetically ordered CDS gene models comprising the AB pan-transcriptome resource. First, we defined 25 579 pairs of homoeologous genes in the A and B genomes by identifying those producing reciprocal top hits in a BLASTN sequence similarity search between gene models in the two genomes, as listed in Data S4. Each member of the pair is associated with a position within its respective genome, and so we could analyse collinearity in detail (Figure S1). The result is consistent with previous analyses (He *et al.*, 2017) confirming that the *Brassica* A and B genomes show extensive segmental collinearity, although this is more disrupted by genome rearrangements than is observed between the A and C genomes.

Functional genotypes for the CGAT genetic diversity panel

A diversity panel of 204 *B. juncea* inbred accessions was used, as described previously (He *et al.*, 2017). The panel is named CGAT after the joint UK Biotechnology and Biological Sciences Research Council (BBSRC) – India Department of Biotechnology (DBT) research initiative 'Crop Genomics and Technologies' that funded its genotyping and analysis.

Functional genotypes were produced for the diversity panel based on leaf RNA, with paired-end 100-base read length mRNA sequencing data produced using the HiSeq 4000 platform. The sequence reads were mapped to the CDS gene model-based *Brassica* AB pan-transcriptome reference, which comprises 93 963 gene models and has an aggregate length of 103 647 589 bases (Data S2). An average of 56 million reads were generated per accession, with 32.5 million being mapped across the reference sequence, representing 57-fold coverage of the 54.8% of the transcriptome to which mRNAseq reads were mapped. SNPs were identified and gene expression quantified (Data S5). Across the panel of 204 lines, 355 050 SNPs were scored, of which the majority (62%) were simple-SNPs, which is a smaller proportion than that found in *B. napus* studies (Trick *et al.*, 2009). In total, 171 196 SNPs were found to be suitable for association mapping after minor allele frequencies < 0.05 were ignored. Significant expression [> 0.4 reads per kilobase of transcript, per million mapped reads (RPKM)] was detected for 48 975 CDS models (52% of all CDS models in the AB pan-transcriptome reference), of which 25 698 belonged to the A genome and 23 277 to the B genome. The functional genotypes are available from the York

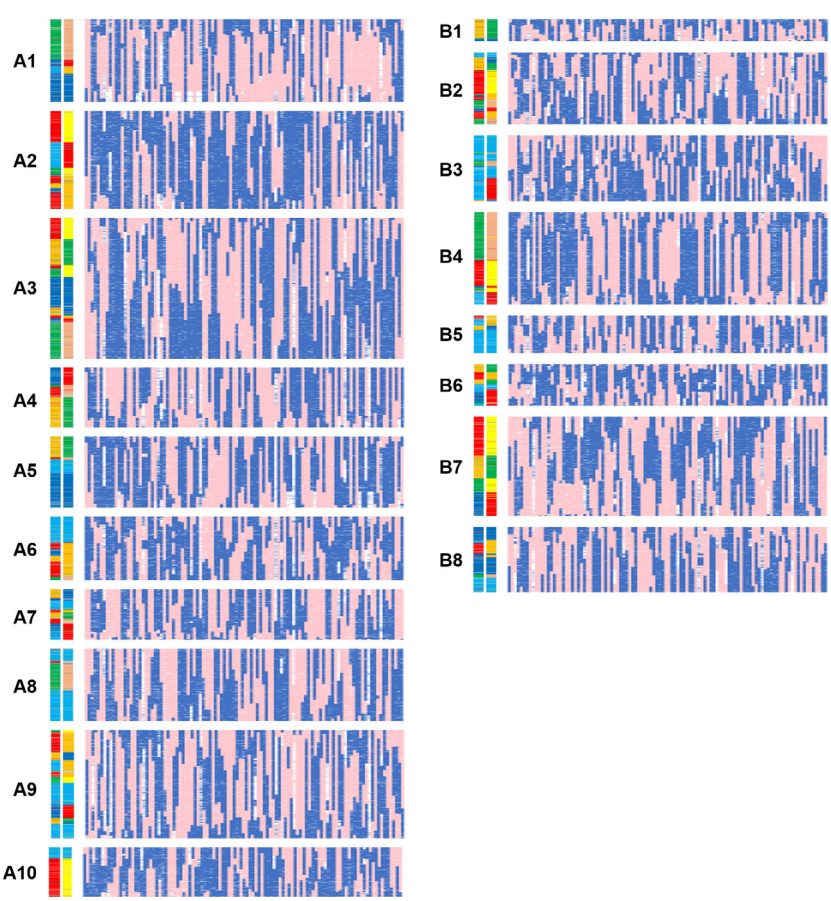

Figure 1. Genome-ordered Graphical Genotypes for the re-ordered A and B genomes. Transcriptome single nucleotide polymorphisms were scored across 106 line of the VHDH linkage mapping population using CDS gene model reference sequences. Alleles were colour-coded pink for the maternal parent (Varuna) and blue for the paternal parent (Heera) and displayed by the order within the chromosome assemblies (A1 to A10; B1 to B8) of the gene models in which the polymorphisms were scored. The collinearity of the *Brassica* A and B genomes relative to the genomes of *Arabidopsis thaliana* and *Thellungiella parvula* is indicated in the two multicoloured columns (*A. thaliana* to the left and *T. parvula* to the right) based on sequence similarity between *Brassica* and *A. thaliana*/ *T. parvula* CDS gene models. The chromosome of the top BLAST hits are indicated by colour coding: light blue for chromosome 1; orange for chromosome 2; dark blue for chromosome 3; green for chromosome 4; red for chromosome 5; and, for *T. parvula* only, yellow for chromosome 6 and pink for chromosome 7.

Oilseed Rape Knowledgebase (http://www.yorknowledgebase.info).

Genetic architecture of the population

The population structure of the diversity panel was inferred using the kernel-pca and optimisation method incorporated in PSIKO (Popescu *et al.*, 2014). This analysis revealed two population clusters, which broadly reflect the origins of the accessions in the panel, with individuals from India having higher membership of cluster 1, and individuals originating from outside India (Canada, Australia, Russia and European samples) having greater membership of cluster 2 (Figure 2). We confirmed these results using STRUCTURE (Pritchard *et al.*, 2000), which also identified $K = 2$ as the most likely number of populations, and assigned individuals to the two clusters in much the same way (Pearson's $r = 0.995$, $P < 0.001$).

Phenotyping

A number of seed and architecture traits were analysed to enable improvement of the nutritional qualities of the seed, and a potential increase in seed yield. All trait data are provided in Data S6, summary statistics are shown in Table 1 and correlations for all traits are shown in Figure S2.

Tocopherols are important nutritional constituents of mustard rape oil as a result of their vitamin E activity. The range of tocopherol isoforms in the *B. juncea* panel (Figure S3), showed a pattern slightly different from that previously observed in a similar *B. napus* diversity panel, where α- and γ-tocopherol were highly significantly correlated (Pearson's $r = -0.49$, $P < 0.001$) and the mean γ- to α-tocopherol ratio was 1.33 (Havlickova *et al.*, 2018). In the case of *B. juncea*, the total amount of tocopherols was slightly lower on average ($t = 1.97$, $P < 0.001$) despite having a slightly wider range of values (171.5–487.1 versus 197.01–445.53 mg kg^{-1} in the *B. napus* study) and γ-tocopherol also tended to more abundant than observed in *B. napus*, which was reflected in a significantly higher γ:α-tocopherol ratio of 4.3 ($t = 1.97$, $P < 0.001$). In addition, α- and γ-tocopherol were only very weakly negatively correlated (Pearson's $r = -0.17$, $P = 0.033$) in our *B. juncea* panel, and only γ-tocopherol showed any correlation with population stratification (Pearson's $r = 0.419$, $P < 0.001$).

Important yield and architecture traits were also measured (Table 1), many of which were significantly correlated (Figure S2), with some relationships suggesting potential yield component trade-offs, presumably as a result of limited resource allocation. In general, plants with a longer vegetative period produced more branches, and

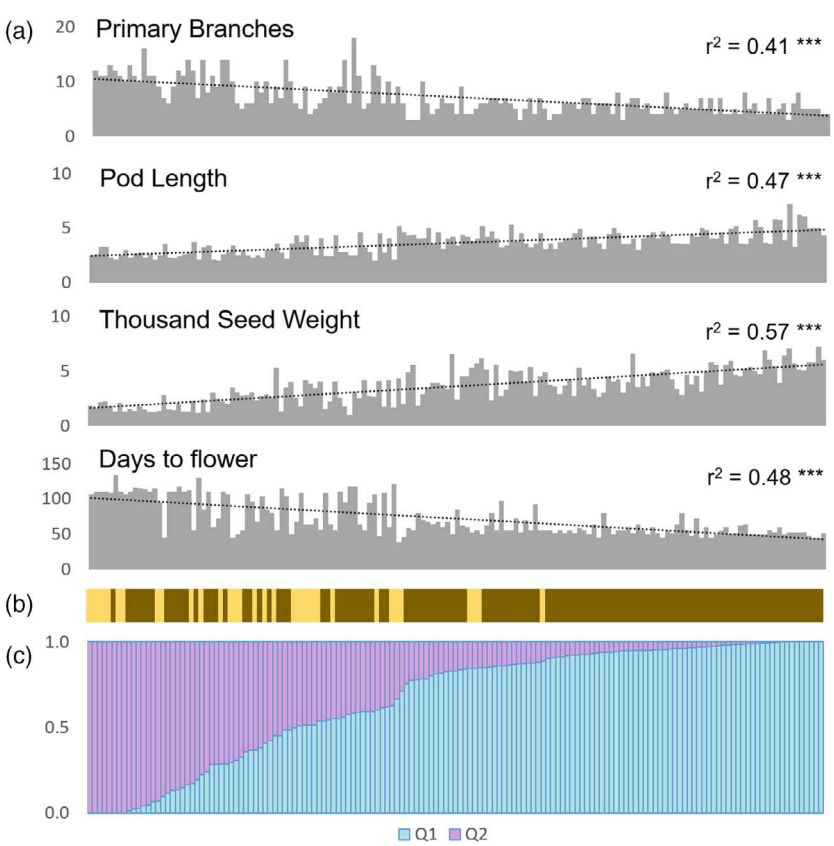

Figure 2. Population structure and morphological traits for 151 *Brassica juncea* accessions
(a) Histograms for thousand seed weight and three of the traits most highly correlated with it: pod length, the number of primary branches and days to flowering. Coefficient of determination r^2 is provided for the correlation between each trait and population cluster Q1
(b) A bar representing seed colour. (c) Histogram showing the two population clusters as determined by PSIKO. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

Table 1 Summary statistics for traits measured for 151 accessions of *Brassica juncea*

	Mean	Median	Min	Max	Range	SD	Narrow-sense heritability (h^2)
Thousand seed weight (TSW; g)	3.62	3.59	0.98	7.2	6.22	1.53	78.4
Plant height (PH; cm)	201.70	197	100	303	203	36.13	66
Primary branches (PB)	7.10	6	3	18	15	3.08	64.8
Secondary branches (SB)	12.28	12	3	23	20	3.93	22.3
Pods/main raceme (PMR)	46.61	49	20	95	75	13.52	49.5
Pod length (PL; cm)	3.67	3.6	2	7.2	5.2	1.00	39.3
Seeds/pod (SP)	12.45	13	7	19	12	2.14	28.7
Oil % (OC)	38.64	39	30	49	19	2.88	73
Days to flower (DF)	71.75	60	38	134	96	24.91	78
Days to maturity (DM)	156.98	154.5	105	195	90	14.08	100
α-tocopherol (AT)	63.68	58.6	14.6	163.2	148.6	23.59	100
γ-tocopherol (GT)	235.92	232.3	101.4	412.6	311.2	57.03	97.7
δ-tocopherol (DT)	5.75	5.3	2.6	12.9	10.3	2.00	100
Total tocopherol (TTC)	305.35	300.2	171.5	487.1	315.6	58.90	100
γ:α-tocopherol (GA)	4.30	3.67	0.88	17.58	16.70	2.22	100

had fewer pods on their branches, as well as fewer, smaller seeds in each pod and a reduced oil content. Most of these traits were also strongly correlated with the population structure of the panel, where accessions with higher proportion membership to cluster 1 (mainly Indian) tended to have shorter growing seasons, higher reproductive to vegetative biomass ratios, and brown over yellow seed coats (Figure 2, Data S6). Narrow-sense heritability (as estimated by GAPIT; Lipka *et al.*, 2012) was generally high for the measured traits (Table 1), indicating that much of the variance observed in these traits can be attributed to additive genetic factors. However, the number of secondary branches and pod characteristics (number of seed/pod, pod length, number of pods on main raceme), showed lower heritability (< 50%), suggesting a considerable environmental influence on these traits.

Principal components analysis

Because of the frequent correlations between many of these traits, principal components analysis (PCA) was used to reduce the complexity of nine highly correlated traits [Thousand seed weight (TSW), Plant height (PH), Primary branches (PB), Secondary branches (SB), Pods/main raceme (PMR), Pod length (PL), Seeds/pod (SP), Days to flower (DF) and Days to maturity (DM)]. Two principal components representing 99% of the variation in these traits were also used for association analysis to reveal master regulators of these large-scale developmental processes (Data S6).

Associative transcriptomics

Significant marker-trait associations were identified for a number of these traits using Associative Transcriptomics (Data S7–10 and Figure S4, 5, 7). We identified a strong candidate, which we have named *BjA.TTL* (after the *Arabidopsis thaliana* gene *TITAN-LIKE)*, underlying a SNP peak on chromosome A1 for TSW (Figure 3) which we decided to use to predict seed weight. Endosperm development involves a number of nuclear divisions that occur in the absence of cytokinesis. Mutations in the orthologue of *TTL* (AT4g24900) in *A. thaliana* lead to greatly enlarged endosperm nuclei, indicating that it is a regulator of endosperm development (Lu *et al.*, 2012). Analysis of other Arabidopsis endosperm mutants such as *HAIKU*, confirmed the key role that endosperm development has in the control of seed size (Garcia *et al.*, 2003), and the contribution of endosperm development to seed size has also been observed in *B. oleracea* (Stoute *et al.*, 2012). Interestingly, an orthologue of Arabidopsis gene *CRWN1* (AT1G67230.1), which also affects nucleus and cell size can be found beneath another peak on chromosome A2 (Figure S4). We therefore suspect that increases in TSW are caused by increases in seed size due to changes in endosperm development and nuclear organisation.

To test the predictive power of this association, the SNP most highly associated with TSW (Cab024377.1:1320:T) was used to successfully predict the seed weight difference for a small number of unrelated *B. juncea* accessions for which we had both SNP data and seed available (Figure S7a), assuming either complete (Pearson's $r = 0.52$, $P = 0.009$) or partial dominance at this locus (Pearson's $r = 0.43$, $P = 0.03$). The seeds from accessions in the test panel which were either homo- or heterozygous for the increasing allele at this SNP position, exhibited a significant 1.8-fold increase in seed weight compared to those with the decreasing allele, which was even greater than the 1.3-fold increase predicted from the association panel. Consistent with the candidate's involvement in endosperm development, we also observed a similar trend for seed size. For example, accessions with alternate alleles,

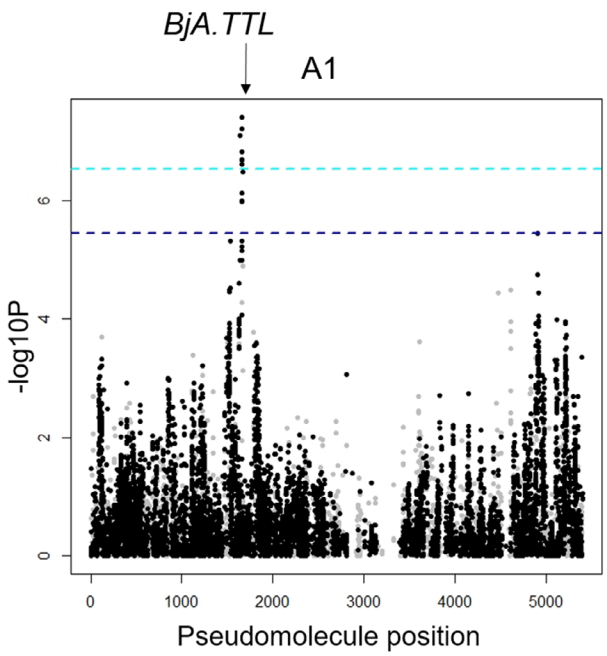

Figure 3. Single nucleotide polymorphism peak identified on chromosome A1, and the position of the candidate gene *BjA.TTL*. Black points have been successfully assigned to the A genome, whereas grey points are unassigned. The dashed lines represents the 0.05 significance thresholds after false discovery rate adjustment (blue) Bonferroni correction (cyan).

Mohan-8 and 27125 which had a 4.6-fold difference in seed weight, also exhibited a 2.2-fold difference in seed area (Figure S7b).

As the prevalent Vitamin E constituent in seed from our *B. juncea* panel, we also identified markers associated with the proportion of γ-tocopherol using Associative Transcriptomics. This analysis revealed strong SNP and GEM peaks (Figure S5), with the top GEM marker (Cab038799.1) found to be a good predictor for the proportion of γ-tocopherol in the test panel accessions (Pearson's r = 0.43, p = 0.002).

A strong association peak was also detected on chromosome A3 for a simple categorical seed colour trait (Figure S6). In this case, using the SNP with the strongest association with this trait, Cab016066.1:252:G, we were able to assign the correct seed colour of individuals in the test panel with 79% accuracy, contributing an additional locus to the seed colour QTL already reported in *Brassica* species (Wang *et al.*, 2017; Zhao *et al.*, 2019; Zhang *et al.*, 2019).

Analysis of the principal components of nine highly correlated traits revealed some interesting associations, although none surpassed the significance threshold after correction for multiple testing. Despite this, further investigation of the single peak associated with PC2 on chromosome B4, revealed a promising candidate. The most highly associated SNP (BniB029491:1532:A), is located extremely close to an orthologue of Arabidopsis *MED8* (AT2G03070.1), a subunit of the Mediator complex which has been implicated in the control of organ size, flowering time and stress responses. Interestingly, expression of this gene in *B. juncea* is highly correlated with expression of *BjA.TTL* (Pearson's $r = 0.53$, $P = 3.024 \times 10^{-12}$), suggesting that the two may be acting in the same pathway.

While we were able to identify significant marker-trait associations for the above traits, several of the architecture traits are likely to be highly polygenic and/or subject to strong genotype by environment interactions. Identification of marker-trait associations in these cases may be optimised by incorporating additional field measurements from multiple seasons/locations and/or through expansion of the diversity panel.

DISCUSSION

Improved agronomy and quality are key objectives for breeding many crops, including *B. juncea*. The analysis of many traits of agronomic importance may be confounded by the tendency for the trade-offs which exist in crops with naturally indeterminate habits. Efficient genetic dissection of such traits may be improved with methods such as AT, as the high precision and resolution of the marker-trait associations enable more subtle dissection of complex phenotypes. In polyploids especially, such methods also rely on high quality references, such as the pan-transcriptome reference presented here, as the basis for SNP discovery and gene expression quantification, as they enable more accurate mapping of homoeologous reads to the appropriate genome and make identification of candidate genes more straightforward. Once assembled, AT platforms may be used to identify markers associated with a wide range of traits, making them valuable resources for the breeding and research communities.

In this study, we introduce the first AT platform for *B. juncea* composed of 204 genetically diverse accessions, complete with genotype information, which may be used for a broad range of association studies, without the need for additional genotyping. This resource offers a large range of potential applications such as the identification of causative genes, uncovering unknown pathways, regulatory genes or transcription factors, screening of available germplasm for allelic variants, and the development of molecular markers for marker-assisted breeding.

Our resource provides 355 050 SNP markers, equivalent to one SNP every 0.3 kb across our *B. juncea* AB pan-transcriptome reference. Although the number of SNPs can be even greater when using whole-genome resequencing, as shown by Yang *et al.* (2016), the advantage of transcriptome resequencing using RNAseq is the availability of transcript abundance data: in our case for 52% of the genes present in the AB pan-transcriptome reference sequence.

As an exemplar, we analysed a range of complex agronomic and quality traits of relevance in *B. juncea*. These analyses revealed promising candidate genes and markers suitable for marker-assisted selection. For example, a single SNP marker showed a 1.8-fold increase in seed weight when used to predict the seed phenotypes of a small test panel, and allowed us to propose a candidate gene for this trait, which we have named *BjA.TTL*. Similarly, a SNP marker was also used to predict seed colour, with 79% efficiency, and a single GEM was successfully used to predict the level of γ-tocopherol in the seed.

By assembling and developing functional genotypes (i.e. comprising both gene sequence variation and gene expression variation) for a diversity panel representing species-wide genetic diversity, we have established a resource for the whole *B. juncea* research community to use. Furthermore, the success of the approach of Associative Transcriptomics for the identification not only of linked markers but of candidates for causative genes serves as an exemplar for plant and crop science more broadly.

EXPERIMENTAL PROCEDURES

Brassica juncea plants used in this study

Linkage mapping to establish the genome order in *B. juncea* was undertaken using the VHDH population. This doubled haploid (DH) mapping population was derived by microspore culture following crossing of parent accessions Varuna (a well-adapted Indian variety) and Heera (a canola quality mustard), as described in (Pradhan *et al.*, 2003). Frozen leaf material from this population was grown in the field (see below for conditions), and leaf material made available for RNA extraction.

A diversity panel comprising 204 *B. juncea* accessions with origins in Asia, North America and Europe was assembled and used for transcriptome sequencing (Data S6). Seeds were sown on Levington professional F2 compost and grown in long day (16/8 h, 20°C/14°C) glasshouse conditions. Second true leaves from each of four plant replicates per accession were harvested when they reached ~3 cm in diameter, as close to the mid point of the light period as possible. Leaves were pooled into a single sample per accession and immediately frozen in liquid nitrogen. Frozen leaf samples were stored at −80°C.

Field trials and trait measurements

The lines were planted in the field in Delhi, India during the mustard growing season (October to March). Each line was planted in five rows with a row length of 3 m. Time to flowering (FT), time to maturation (TM), plant height (PH), the number of primary branches (PB), the number of secondary branches (SB), the number of pods on the main raceme (PMR), the length of those pods (PL; cm), the number of seeds/pod (SP) and hundred seed weight (TSL; g), were measured from 10 competitive plants, with a mean of these 10 observations used as the trait value. In addition, the oil content of the seed (OC; %) was estimated using Near Infrared Spectroscopy (NIRS) (Mika *et al.*, 2003), and the seed colour recorded. Narrow-sense heritability was estimated within GAPIT (Lipka *et al.*, 2012) for all traits, using the SNP markers and genetic relatedness between individual *B. juncea* accessions.

Tocopherols

The α-,γ- and δ-tocopherol (the sum of which formed total tocopherol) were extracted from a homogenous mixture of 80 mg *B. juncea* seeds and analysed by normal-phase HPLC, as described previously (Fritsche *et al.*, 2012). Modified mobile phase A was heptane (Rathburn Chemicals Co., http://rathburn.co.uk), phase B was heptane:dioxane (90:10, v/v; Sigma-Aldrich, https://www.sigmaaldrich.com). The internal standard, α-tocopherol acetate (Sigma-Aldrich), was added to each sample at a concentration of 25.4 μM (12 μg mL^{-1}).

Establishment of the AB pan-transcriptome reference

The transcriptome reference sequence was developed essentially as described previously (He *et al.*, 2017), utilising CDS models from published genome sequences for *B. rapa* Chiifu v2.0 (Cai *et al.*, 2017), *B. nigra* YZ12151 (Yang *et al.*, 2016), and *B. juncea* (Yang *et al.*, 2016). BLASTN was used to identify the highest-scoring significant hit (threshold e-value 1E-30) when mapping CDS models to respective genome sequence pseudomolecules. Methods adapted from He *et al.* (2015), were employed to order and interpolate *B. juncea* specific CDS models. In order to eliminate false interpolation, a filter was introduced to exclude interpolation to positions that were not collinear in either *A. thaliana* or *T. parvula*.

Functional Genotypes

RNA was extracted from each of the pooled leaf samples using the Omega Biotek E.Z.N.A Total RNA kit, and 100-base paired-end transcriptome sequences generated Illumina HiSeq 2500 platform as described previously (He *et al.*, 2017). Maq was used for mapping with default parameters, meaning that reads with no more than two mismatches with summed Q ≥ 70 were mapped. Sequence reads were aligned to the *B. juncea* AB pseudomolecules, transcript abundance scored for each CDS model, and SNPs called by the meta-analysis of alignments as described previously (Harper *et al.*, 2012).

Associative Transcriptomics

The SNP dataset was entered into the program PSIKO (Popescu *et al.*, 2014) to assess the population structure and produce a Q matrix, which was composed of two population clusters. The SNP genotypes, Q matrix and trait scores for 151 accessions were incorporated into a compressed mixed linear model (Zhang *et al.*, 2010) implemented in the GAPIT R package (Lipka *et al.*, 2012), with missing data imputed to the major allele. The kinship matrix used in this analysis was also generated by GAPIT. GEM associations were calculated by a fixed effect linear model in R with RPKM values and the Q matrix inferred by PSIKO as the explanatory variables and damage score the response variable. Coefficients of determination (r2), regression coefficients, constants and significance values were calculated, and a genomic inflation correction factor was applied.

ACKNOWLEDGEMENTS

We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z) for the generation of mRNA sequencing data. This work was supported by UK Biotechnology and Biological Sciences Research Council (BB/L011751/1, BB/L002124/1, BB/R019819/1) and Department of Biotechnology (DBT), India (BT/IN/Indo-UK/CGAT/11/AKP/2014-15).

AUTHOR CONTRIBUTIONS

ALH, IB and APK. conceived and planned the project. ALH, SL, LH, LW, AF and VG performed experiments. ALH and ZH performed data analysis. ALH, IB and AKP wrote the manuscript and all authors reviewed it.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Sequence data from this article can be found in the SRA data library under accession number PRJNA507350. Germplasm are available by request from A. K. Pradhan[2].

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Homoeology relationships between *Brassica* A and B genomes.

Figure S2. Correlation plots for seed quality and morphological characteristics of 151 *B. juncea* accessions.

Figure S3. Tocopherol content in the seed of 151 *B. juncea* accessions. Proportion of the total tocopherol content from α-, γ- and δ-tocopherol is shown.

Figure S4. Manhattan plots displaying the results of Associative Transcriptomics analysis for thousand seed weight in 151 *B. juncea* accessions. (a) SNP analysis. (b) GEM analysis. Dashed lines show 0.05 significance thresholds calculated by false discovery rate (blue) and Bonferroni correction (cyan).

Figure S5. Manhattan plots displaying the results of Associative Transcriptomics analysis for proportion γ-tocopherol in seed for 151 *B. juncea* accessions. (a) SNP analysis. (b) GEM analysis. Dashed lines show 0.05 significance thresholds calculated by false discovery rate (blue) and Bonferroni correction (cyan).

Figure S6. Manhattan plots displaying the results of Associative Transcriptomics analysis for seed colour in 151 *B. juncea* accessions. (a) SNP analysis. (b) GEM analysis. Dashed lines show 0.05 significance thresholds calculated by false discovery rate (blue) and Bonferroni correction (cyan).

Figure S7. Predictions of seed weight based on SNP Cab024377.1:1320:T. (a) Seed weight data for the 151 accession diversity panel, and a test panel of 24 unrelated accessions. (b) Accessions with alternate alleles, Mohan-8 and 27125, which had a 4.6-fold difference in seed weight, also exhibit approximately a 2.2-fold difference in seed area.

Data S1. Ordered list of CDS gene model-based *Brassica* AC pan-transcriptome.

Data S2. Read mapping statistics.

Data S3. SNP marker scores for VHDH population arranged as Genome-Ordered Graphical Genotypes.

Data S4. Homoeologous genes in the A and B genomes.

Data S5. SNP alleles and RPKM data for 151 *B. juncea* accessions.

Data S6. *B. juncea* accessions, population structure and trait data.

Data S7. Associative Transcriptomics results for thousand seed weight.

Data S8. Associative Transcriptomics results for γ-tocopherol.

Data S9. Associative Transcriptomics results for Seed Colour.

Data S10. Associative Transcriptomics results for principal component 2.

REFERENCES

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, **408**, 796–815.

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J. *et al.* (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, **465**, 627–631.

Bancroft, I., Morgan, C., Fraser, F. ***et al.*** (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat. Biotechnol.* **29**, 762–766.

Bansal, P., Kaur, P., Banga, S.K. and Banga, S.S. (2009) Augmenting genetic diversity in Brassica juncea through its resynthesis using purposely selected diploid progenitors. *Int. J. Plant. Breed.* **3**, 41–45.

Bansal, P., Banga, S. and Banga, S.S. (2012) Heterosis as Investigated in Terms of Polyploidy and Genetic Diversity Using Designed Brassica juncea Amphiploid and Its Progenitor Diploid Species. *PLoS One*, **7**, e29607.

Cai, C., Wang, X., Liu, B., Wu, J., Liang, J., Cui, Y., Cheng, F. and Wang, X. (2017) Brassica rapa Genome 2.0: A reference upgrade through sequence re-assembly and gene re-annotation. *Mol. Plant.* **10**, 649–651. https://doi.org/10.1016/j.molp.2016.11.008

Chen, S., Wan, Z., Nelson, M.N. ***et al.*** (2013) Evidence from genome-wide simple sequence repeat markers for a polyphyletic origin and secondary centers of genetic diversity of Brassica juncea in China and India. *J. Hered.* **104**, 416–427.

Cheung, F., Trick, M., Drou, N. ***et al.*** (2009) Comparative analysis between homoeologous genome segments of Brassica napus and its progenitor species reveals extensive sequence-level divergence. *Plant Cell*, **21**, 1912–1928.

Cockram, J., White, J., Zuluaga, D.L. ***et al.*** (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc. Natl. Acad. Sci. USA*, **107**, 21611–21616.

Fritsche, S., Wang, X., Li, J. ***et al.*** (2012) A candidate gene-based association study of tocopherol content and composition in rapeseed (Brassica napus). *Front. Plant Sci.* **3**, 129.

Garcia, D., Saingery, V., Chambrier, P., Mayer, U., Jürgens, G. and Berger, F. (2003) Arabidopsis haiku mutants reveal new controls of seed size by endosperm. *Plant Physiol.* **131**, 1661–1670.

Gupta, M., Gupta, S., Kumar, H., Kumar, N. and Banga, S.S. (2015) Population structure and breeding value of a new type of Brassica juncea created by combining A and B genomes from related allotetraploids. *Theor. Appl. Genet.* **128**, 221–234.

Harper, A.L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P. and Bancroft, I. (2012) Associative transcriptomics of traits in the polyploid crop species Brassica napus. *Nat. Biotechnol.* **30**, 798–802.

Havlickova, L., He, Z., Wang, L., Langer, S., Harper, A.L., Kaur, H., Broadley, M.R., Gegas, V. and Bancroft, I. (2018) Validation of an updated Associative Transcriptomics platform for the polyploid crop species Brassica napus by dissection of the genetic architecture of erucic acid and tocopherol isoform variation in seeds. *Plant J.* **93**, 181–192.

He, Z., Cheng, F., Li, Y., Wang, X., Parkin, I.A.P., Chalhoub, B., Liu, S. and Bancroft, I. (2015) Construction of *Brassica* A and C genome-based ordered pan-transcriptomes for use in rapeseed genomic research. *Data in Brief.* **4**, 357–362.

He, Z., Wang, L., Harper, A.L., Havlickova, L., Pradhan, A.K., Parkin, I.A.P. and Bancroft, I. (2017) Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* **15**, 594–604.

He, Z. and Bancroft, I. (2018) Organisation of the genome sequence of the polyploid crop species *Brassica juncea*. *Nat. Genet.*, **50**, 1496–1497.

Higgins, J., Magusin, A., Trick, M., Fraser, F. and Bancroft, I. (2012) Use of mRNA-seq to discriminate contributions to the transcriptome from the constituent genomes of the polyploid crop species Brassica napus. *BMC Genom.* **13**, 247.

Kaur, P., Banga, S., Kumar, N., Gupta, S., Akhatar, J. and Banga, S.S. (2014) Polyphyletic origin of Brassica juncea with B. rapa and B. nigra (Brassicaceae) participating as cytoplasm donor parents in independent hybridization events. *Am J. Bot.* **101**, 1157–1166.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics*, **28**, 2397–2399.

Lu, X., Li, Y., Su, Y., Liang, Q., Meng, H., Li, S., Shen, S., Fan, Y. and Zhang, C. (2012) An Arabidopsis gene encoding a C_2H_2-domain protein with alternatively spliced transcripts is essential for endosperm development. *J. Exp. Bot.* **63**, 5935–5944.

Mika, V., Tillmann, P., Koprna, R., Nerusil, P. and Kucera, V. (2003) Fast prediction of quality parameters in whole seeds of oilseed rape (Brassica napus L.). *Plant Soil Environ.* **49**, 141–145.

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Crollius, H.R. and Salse, J. (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biol.* **16**, 262. https://doi.org/10.1186/s13059-015-0814-y

O'Neill, C.M. and Bancroft, I. (2000) Comparative physical mapping of segments of the genome of Brassica oleracea var. alboglabra that are homoeologous to sequenced regions of chromosomes 4 and 5 of Arabidopsis thaliana. *Plant J.* **23**, 233–243.

Paritosh, K., Gupta, V., Yadava, S.K., Singh, P., Pradhan, A.K. and Pental, D. (2014) RNA-seq based SNPs for mapping in Brassica juncea (AABB): synteny analysis between the two constituent genomes A (from B. rapa) and B (from B. nigra) shows highly divergent gene block arrangement and unique block fragmentation patterns. *BMC Genom.* **15**, 396. https://doi.org/10.1186/1471-2164-15-396

Pires, J.C., Zhao, J., Schranz, M.E., Leon, E.J., Quijada, P.A., Lukens, L.N. and Osborn, T.C. (2004) Flowering time divergence and genomic rearrangements in resynthesized Brassica polyploids (Brassicaceae). *Biol. J. Linn. Soc. Lond.* **82**, 675–688.

Popescu, A.-A., Harper, A.L., Trick, M., Bancroft, I. and Huber, K.T. (2014) A novel and fast approach for population structure inference using kernel-PCA and optimization. *Genetics*, **198**, 1421–1431.

Pradhan, A.K., Gupta, V., Mukhopadhyay, A., Arumugam, N., Sodhi, Y.S. and Pental, D. (2003) A high-density linkage map in Brassica juncea (Indian mustard) using AFLP and RFLP markers. *Theor. Appl. Genet.* **106**, 607–614.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

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

Stoute, A.I., Varenko, V., King, G.J., Scott, R.J. and Kurup, S. (2012) Parental genome imbalance in Brassica oleracea causes asymmetric triploid block. *Plant J.* **71**, 503–516.

Tian, F., Bradbury, P.J., Brown, P.J. *et al.* (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162.

Town, C.D., Cheung, F., Maiti, R. *et al.* (2006) Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *Plant Cell*, **18**, 1348–1359.

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

Wang, J., Xian, X., Xu, X., Qu, C., Lu, K., Li, J. and Liu, L. (2017) Genome-Wide Association Mapping of Seed Coat Color in Brassica napus. *J. Agric. Food Chem.* **65**, 5229–5237.

Yang, J., Liu, D., Wang, X. *et al.* (2016) The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nat. Genet.* **48**, 1225–1232.

Yang, T.-J., Kim, J.S., Kwon, S.-J. *et al.* (2006) Sequence-level analysis of the diploidization process in the triplicated FLOWERING LOCUS C region of Brassica rapa. *Plant Cell*, **18**, 1339–1347.

Zhang, Y., Sun, Y., Sun, J., Feng, H. and Wang, Y. (2019) Identification and validation of major and minor QTLs controlling seed coat color in Brassica rapa L. *Breed Sci.* **69**, 47–54.

Zhang, Z., Ersoz, E., Lai, C.-Q. *et al.* (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360.

Zhao, H., Basu, U., Kebede, B., Qu, C., Li, J. and Rahman, H. (2019) Fine mapping of the major QTL for seed coat color in Brassica rapa var. Yellow Sarson by use of NIL populations and transcriptome sequencing for identification of the candidate genes. *PLoS One*, **14**, e0209982.

Zhao, K., Tung, C.-W., Eizenga, G.C. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467.