the plant journal

SEB Society for Experimental Biology

The Plant Journal (2017) doi: 10.1111/tpj.13767

RESOURCE

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

Lenka Havlickova[1], Zhesi He[1], Lihong Wang[1], Swen Langer[1], Andrea L. Harper[1], Harjeevan Kaur[1], Martin R. Broadley[2], Vasilis Gegas[3] and Ian Bancroft[1,*]

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

[2]*Plant and Crop Sciences Division, School of Biosciences, University of Nottingham, Sutton Bonington Campus, Loughborough LE12 5RD, UK, and*

[3]*Limagrain UK Ltd., Joseph Nickerson Research Centre, Rothwell, LN7 6DT, UK*

Received 11 May 2017; revised 6 October 2017; accepted 30 October 2017.
*For correspondence (e-mail ian.bancroft@york.ac.uk).

SUMMARY

An updated platform was developed to underpin association genetics studies in the polyploid crop species *Brassica napus* (oilseed rape). Based on 1.92 × 10^{12} bases of leaf mRNAseq data, functional genotypes, comprising 355 536 single-nucleotide polymorphism markers and transcript abundance were scored across a genetic diversity panel of 383 accessions using a transcriptome reference comprising 116 098 ordered coding DNA sequence (CDS) gene models. The use of the platform for Associative Transcriptomics was first tested by analysing the genetic architecture of variation in seed erucic acid content, as high-erucic rapeseed oil is highly valued for a variety of applications in industry. Known loci were identified, along with a previously undetected minor-effect locus. The platform was then used to analyse variation for the relative proportions of tocopherol (vitamin E) forms in seeds, and the validity of the most significant markers was assessed using a take-one-out approach. Furthermore, the analysis implicated expression variation of the gene *Bo2g050970.1*, an orthologue of *VTE4* (which encodes a γ-tocopherol methyl transferase converting γ-tocopherol into α-tocopherol) associated with the observed trait variation. The establishment of the first full-scale Associative Transcriptomics platform for *B. napus* 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: association genetics, transcriptomics, *Brassica napus*, tocopherol, erucic acid.

INTRODUCTION

As the demand for ever-increasing crop productivity continues against the backdrop of climate change and diminishing resources, crop improvement has become an important driver for advances in genomic technologies in plants. A broad aim of crop science is the identification of the genetic bases for trait variation, including both the identification of beneficial alleles and the development of molecular markers to accelerate their introduction into elite germplasm. Genetic diversity panels, typically comprising past and current cultivars along with wild relatives, are usually available for crop species. Such panels represent ideal resources for genome-wide association studies (GWAS), which exploit historical recombination between molecular markers and loci associated with trait variation. Where recombination between loci is observed proportionately less frequently than expected for unlinked loci (i.e. < 0.5), those loci are said to be in linkage disequilibrium (LD). The approach of identifying molecular markers in LD with loci associated with trait variation is an important tool used in human genetics studies, and has been applied

© 2017 The Authors.
The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

successfully in several plant species (Garrigan and Hammer, 2006; Li *et al.*, 2008; Atwell *et al.*, 2010; Cockram *et al.*, 2010; Tian *et al.*, 2011; Zhao *et al.*, 2011). The recent development of transcriptome-based GWAS, including the technology termed Associative Transcriptomics (AT), in which both gene sequence variation and transcript abundance variation are used to identify associations with trait variation (Harper *et al.*, 2012), greatly increases the range of crops to which GWAS approaches can be applied.

The Brassicaceae family includes *Arabidopsis thaliana*, the first plant for which a high-quality genome sequence was available (AGI, 2000), and the Brassica crops. The diploid species *Brassica rapa* and *Brassica oleracea*, which contain the Brassica A and C genomes, respectively, are closely related, having shared a common ancestor only *c.* 3.7 Mya (Inaba and Nishio, 2002). *Brassica napus* is an allopolyploid arising from the hybridization of these species (U. N. 1935), and the related (homoeologous) regions of the genomes are clearly discernible (Bancroft *et al.*, 2015). A diverse range of *B. napus* crop types have been developed, including oilseed rape, fodders, leafy vegetables and root vegetables. Brassica species have been used extensively in genomics studies, because of their utility in studying the evolution of polyploid genomes (Song *et al.*, 1995; O'Neill and Bancroft, 2000; Pires *et al.*, 2004; Town *et al.*, 2006; Yang *et al.*, 2006; Cheung *et al.*, 2009). A draft genome sequence has been obtained for *B. napus* (Chalhoub *et al.*, 2014); however, at approximately 1.2 Gb, the genome of *B. napus* is relatively large. To address this problem, rapid and cost-effective transcriptome-based technologies, using mRNAseq, 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).

Vegetable oils are a major source of dietary vitamin E (Goffman and Becker, 2002). Vitamin E occurs in the form of tocopherols, which are lipid-soluble antioxidants that accumulate in the chloroplast. Their function is to protect photosystem II from oxidative damage under the influence of free/released lipid peroxyl radicals and singlet oxygen (Quadrana *et al.*, 2013), and in seeds they play a role in preventing the oxidation of polyunsaturated fatty acids (PUFAs). The four forms of tocopherol (α, β, γ and δ) vary in the number and position of methyl substituents attached to the chromanol ring (Munné-Bosch and Alegre, 2002). The most abundant forms of vitamin E in rapeseed oil are γ- and α-tocopherol, with a small proportion of δ-tocopherol (Fritsche *et al.*, 2012; Wang *et al.*, 2012). Besides its nutritional value, α-tocopherol is the most potent vitamin E, whereas the γ- and δ-tocopherol forms are valued for their oil-stabilizing properties (Munné-Bosch and Alegre, 2002), which is particularly relevant for PUFA-rich oils such as rapeseed. Tocopherol content and composition in rapeseed varies widely: values for total tocopherol content (TTC) have been reported to range between 166 and 687 mg kg^{-1}, α-tocopherol content ranges between 59 and 286 mg kg^{-1}, and γ-tocopherol content ranges between 107 and 280 mg kg^{-1}. The ratio between α- and γ-tocopherol has also been reported to range between 0.33 and 2.14 (Dolde *et al.*, 1999; Goffman and Becker, 2002; Fritsche *et al.*, 2012; Wang *et al.*, 2012). Genes involved in the tocopherol biosynthetic pathway have been identified in *A. thaliana* and other model plants (Valentin *et al.*, 2006; Endrigkeit *et al.*, 2009; Li *et al.*, 2012; Figure 1). Quantitative trait loci (QTL) affecting seed tocopherol content and composition have also been reported (Gilliland *et al.*, 2006), but the control of biosynthesis is poorly understood.

The first AT panel reported for *B. napus* (Harper *et al.*, 2012) comprised only 84 accessions and was smaller than is usually required for association studies (Spencer *et al.*, 2009), meaning that it could be used successfully only for traits with a simple genetic basis. In this study, we report the establishment of a full AT platform for the crop species

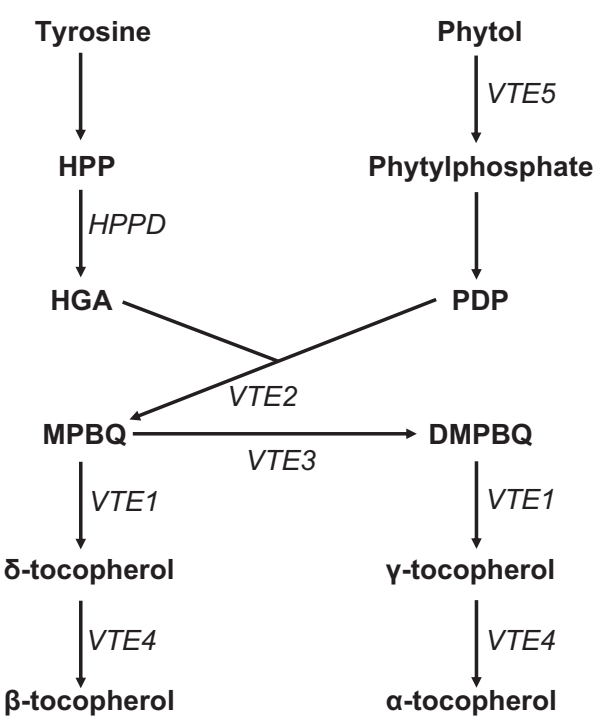

Figure 1. Simplified tocopherol biosynthesis pathway in plants. Abbreviations: DMPBQ, 2,3-dimethyl-5-phytyl-1,4-benzoquinone; HGA, homogentisic acid; HPP, *p*-hydroxyphenylpyruvate; *HPPD*, HPP dioxygenase; MPBQ, 2-methyl-6-phytyl-1,4-benzoquinone; PDP, phytyl-diphosphate; *VTE1*, tocopherol cyclase; *VTE2*, homogentisate phytyltransferase; *VTE3*, MPBQ methyltransferase; *VTE4*, γ-tocopherol methyltransferase; *VTE5*, phytol kinase.

© 2017 The Authors.

B. napus, based on a widely shared genetic diversity panel of 383 accessions, which can be used to address the genetic architecture of a broad range of traits. We validated the resource by using the new platform to analyse a trait that had been analysed previously using the original panel (erucic acid content of seed oil) and a new trait (the relative content of γ and α forms of tocopherol in seeds).

RESULTS

The Renewable Industrial Products from Rapeseed (RIPR) genetic diversity panel

A diversity panel of 383 *B. napus* doubled haploid (DH) or inbred accessions was assembled, with the aim of covering the breadth of genetic variation available in the species. This panel included the breadth of crop types of *B. napus*, and comprised 362 inbred lines previously used by Bus *et al.* (2011) and Harper *et al.* (2012) plus 21 further accessions used by Thomas *et al.* (2016). The list of accessions is shown in Appendix S1. The panel is named RIPR after the research project 'BBSRC Renewable Industrial Products from Rapeseed (RIPR) Programme' that funded its development and genotyping.

Functional genotypes

Functional genotypes were produced for the panel based on leaf RNA, with 100-base read length mRNAseq data produced using the Illumina HiSeq 2000 platform. A total of 1.92×10^{12} bases of sequence data were produced. The sequence reads were mapped to the CDS gene model-based Brassica AC pan-transcriptome reference (He *et al.*, 2015), which comprised 116 098 gene models, has an aggregate length of 118 657 829 bases and for which we provide an updated gene order based on a high-density single-nucleotide polymorphism (SNP) linkage map, as shown in Appendix S2. Sequence read mapping statistics are summarised in Appendix S1. Mean values of 50 165 125 reads were generated per accession, with 32 275 718 being mapped across 61 620 266 bases of the reference sequence, representing 52.1-fold coverage of the 51.9% of the predicted transcriptome to which mRNAseq reads were mapped. SNPs were identified and gene expression quantified. Across the panel of 383 lines, 355 536 SNPs were scored, of which the majority (87.0%) were hemi-SNPs, as found in previous *B. napus* studies (Trick *et al.*, 2009). A total of 127 153 561 allele calls were made, with 9 017 727 (6.6%) missing values. Significant expression (>0.4 reads per kilobase per million mapped reads, RPKM) was detected for 53 889 CDS models (46.4% of all CDS models in the AC pan-transcriptome reference), of which 25 834 belong to the A genome and 28 055 belong to the C genome. The functional genotypes are available from the York Oilseed Rape Knowledgebase (http://www.yorknowledgebase.info/).

Genetic architecture of the population

The 355 536 SNP markers scored across the RIPR panel were used to analyse the relatedness of members of the panel. First, a distance matrix was generated and visualized by the dendrogram shown in Figure 2a. The assigned crop types (Appendix S1) show the expected clustering, as shown in Figure 2b. Next, the population structure of the panel was analysed using PSIKO (Popescu *et al.*, 2014). The highest likelihood is a subpopulation $k = 2$, with mixture across the panel as illustrated in Figure 2c. Finally, LD was calculated across the genome, as summarised in Figure S1, producing a mean value of 0.031 for the population.

Seed erucic acid analysis

Erucic acid is a 22-carbon monounsaturated fatty acid. Its content in rapeseed oil is one of the key determinants of suitability for use as an edible or industrial oil. Detection of the known loci controlling the biosynthesis of erucic acid in seeds was used as a validation study for the first report of AT (Harper *et al.*, 2012). We re-analysed this trait to compare the performance of the original panel with the new RIPR panel. The fatty-acid composition of seeds was determined for 376 lines of the RIPR diversity panel (summarised in Appendix S3). The erucic acid content of seeds varied between 0 and 51%, reflecting the range of crop types represented in the panel, which included modern Canola quality rapeseed varieties as well as crop types for which seed composition was not the subject of an active domesticated selection process (hence representative of 'unimproved' seed composition).

Associative Transcriptomics of erucic acid content

The first stage of validation of the new AT platform for *B. napus* involved the analysis of seed erucic acid content, a trait for which the two main control loci are known and were confirmed previously by AT (Harper *et al.*, 2012). The estimated narrow-sense heritability (h^2) for the erucic acid trait was estimated from the SNP analysis as 0.794. A total of 318 genome-assigned SNP markers above the Bonferroni-corrected significance threshold of $P = 0.05$ (i.e. $-\log_{10} P$ value of 6.7) were detected across association signals on chromosomes A5, A8, A9, A10 and C3 (Appendix S4; Figure S3), as illustrated in Figure 3a. The main loci controlling erucic content (on chromosomes A8 and C3) provide association signals with a significance eight orders of magnitude greater: $-\log_{10} P > 16$, compared with <8 in the previous study. The known control genes, orthologues of *FAE1* (AT4G34520), represented by gene models Cab035983.1 and Bo3g168810.1, are near the centres of these SNP association peaks, within six genes (approximately 42 kb) and nine genes (approximately 56 kb) from the closest significantly associated gene,

© 2017 The Authors.

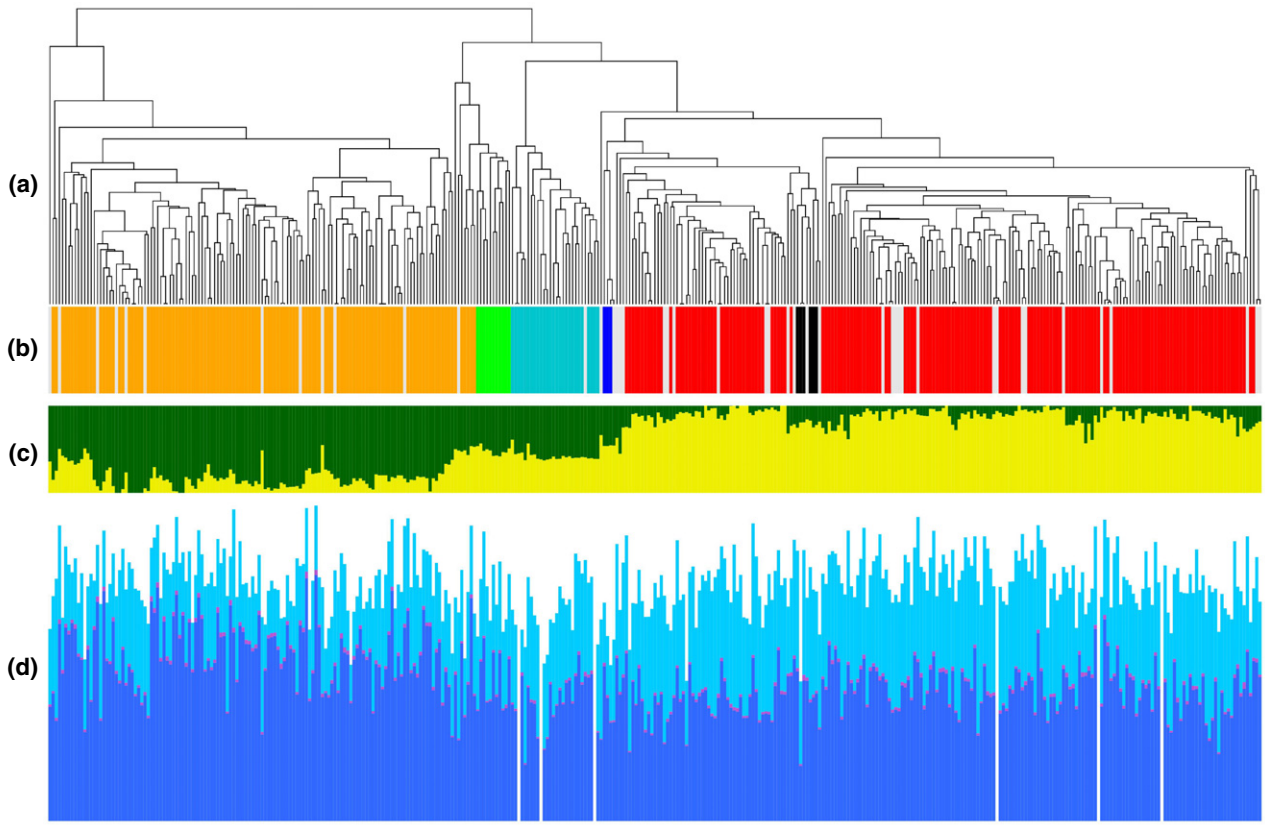

Figure 2. Population structure and trait variation across the Renewable Industrial Products from Rapeseed (RIPR) panel.
(a) Relatedness of accessions in the panel based on 355 536 scored single-nucleotide polymorphisms (SNPs).
(b) Main crop types in the panel, colour-coded: orange for spring oilseed rape; green for semi-winter oilseed rape; light blue for swede; dark blue for kale; black for fodder; red for winter oilseed rape; and grey for crop type not assigned.
(c) Population structure for highest likelihood $k = 2$.
(d) Variation for seed content of α-tocopherol (light blue), γ-tocopherol (dark blue) and δ-tocopherol (magenta).

respectively, according to the reference sequence (Appendix S4). In addition, SNP associations were found for a region of the genome, on chromosome A5, which were not previously detected. This indicates the position of a novel locus with minor effect on the trait. A candidate for the trait control gene in this region is Cab033920.1. This gene is an orthologue of AT2G34770.1, annotated as fatty acid hydroxylase 1, which has a potential role in very long chain fatty-acid biosynthesis. An association signal was also detected for a relatively large region of chromosome A9, which we interpret as corresponding to a seed glucosinolate-controlling locus, which was co-selected in modern low erucic rapeseed cultivars to produce Canola quality seed.

In addition to association analysis using SNP markers, AT also reveals associations between gene expression markers (in the tissue of second true leaves used for the development of functional genotypes) and trait variation. In the case of seed erucic acid content, the main control genes (orthologues of *FAE1*) are transcriptionally inactive in the tissue (leaves) sampled for the production of the functional genotypes. We are still able to detect both SNP and gene expression marker (GEM) association peaks through markers in LD with *FAE1* on A8 and C3, however, as illustrated in Figure 3b. The lower resolution observed for the A8 peaks may reflect the influence of two strong bottlenecks during breeding selection (Hasan *et al.*, 2008) for low glucosinolate content (controlling loci on chromosomes A2, A9, C2 and C9) and zero seed erucic acid content (controlling loci on chromosomes A8 and C3), or perhaps the presence of additional minor effect genes located on A8 that also contribute to the erucic trait. Indeed there are many potential candidate genes in the region that could have an effect, including an orthologue of FAD6 (AT4G30950), which could act to reduce the pool of oleic acid available for elongation to erucic acid. In addition, there is a signature of slightly inflated LD on the first half of A8, which may further contribute to reducing the resolution of association peaks in this region (Figure S1).

The clear signals in the transcript abundance-based association analysis confirms the stability of differential gene expression across the panel, and its utility for the

© 2017 The Authors.

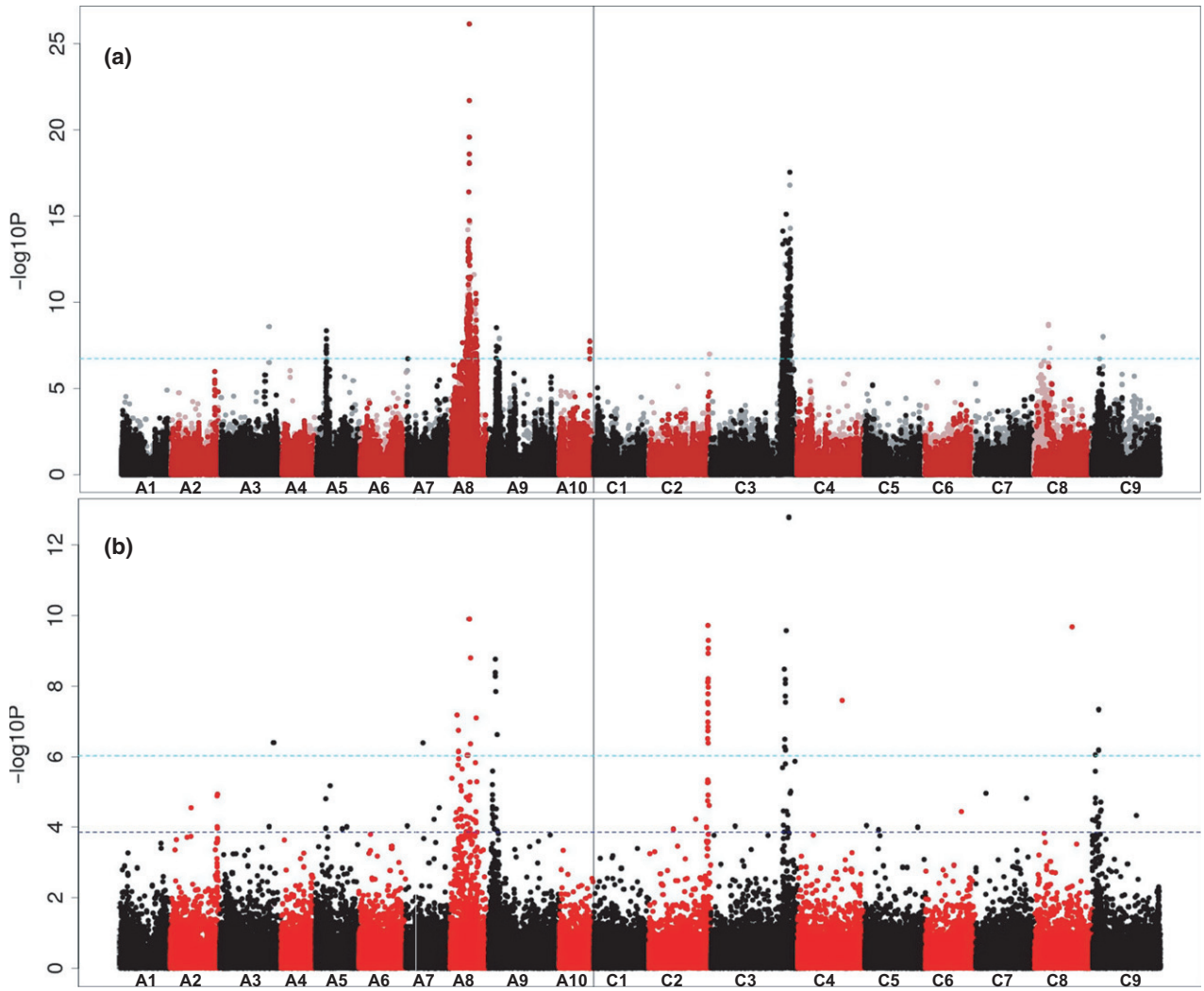

Figure 3. Association analysis.
(a) Transcriptome single-nucleotide polymorphism (SNP) markers with seed erucic acid content. The SNP markers are positioned on the *x*-axis based on the genomic order of the gene models in which the polymorphism was scored, with the significance of the trait association, as –log10*P*, plotted on the *y*-axis. A1–A10 and C1–C9 are the chromosomes of *Brassica napus*, shown in alternating black and red colours to permit boundaries to be distinguished. Hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned are shown as light points, whereas simple SNP markers (i.e. polymorphisms between resolved bases) and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to a genome, are shown as dark points. The broken light-blue horizontal line marks the Bonferroni-corrected significance threshold of 0.05.
(b) Transcript abundance with seed erucic acid content. The gene models are positioned on the *x*-axis based on their genomic order, with the significance of the trait association, as –log10*P*, plotted on the *y*-axis. The broken dark-blue horizontal line marks the 5% false discovery rate.

identification of association signals. Regions of the genome previously associated with seed glucosinolate content (selected alongside erucic content in Canola quality rapeseed) show particularly strong transcript abundance associations, which we interpret as consequences of the extensive structural variation in these regions of the genome (He *et al.*, 2016). The new AT platform generates strong signals because of the large, diverse panel and superior number of markers assigned to homoeologues, properties lacking in the platform reported previously (Harper *et al.*, 2012).

Tocopherol phenotype analysis

We selected tocopherols in seeds as test traits of unknown genetic basis, quantifying α, γ and δ forms. Tocopherols were purified from seeds and quantified for 377 accessions of the RIPR panel. The results are summarised in Appendix S5 and Figure S2. Total tocopherol in seeds varied from 197 to 445 mg kg^{-1}, with the main types being γ-tocopherol (78–347 mg kg^{-1}) and α-tocopherol (51–229 mg kg^{-1}), the relative proportions of which (measured as the γ/α-tocopherol ratio) varied greatly, ranging

© 2017 The Authors.

from 0.485 to 5.00, with δ-tocopherol representing a minor component (1.8–9.9 mg kg^{-1}). Analysis of tocopherol characteristics by crop type showed that γ-tocopherol content tended to be higher in spring crop types and α-tocopherol content tended to be higher in winter crop types, as illustrated in Figure 2d.

Given that the purpose of tocopherols in seed oil is to protect against oxidation, we assessed the diversity panel for correlations of tocopherol traits with the proportions of the fatty acids found in seed oil that are most susceptible to oxidation, the PUFAs linoleic and linolenic. The content of these fatty acids had been determined alongside that of erucic acid (Appendix S3). A weak positive correlation between total tocopherol and PUFA content was, indeed, identified (R^2 = 0.13; $P < 0.001$).

Associative Transcriptomics of tocopherol composition

To undertake AT for tocopherol traits, we analysed the population for loci controlling the proportion of tocopherol occurring in the γ form rather than the α form by using the γ/α ratio as the trait. The SNP-based association analysis, as illustrated in Figure 4a, revealed exceptionally strong associations with markers in a very small region of chromosome C2, along with weaker associations with a few markers in regions of chromosomes A2 and A10. Unlike seed erucic acid, tocopherol composition has not been selected for by *B. napus* breeders. We interpret the very sharp association signal as indicative of this lack of selection, and consider this to be consistent with LD across most of the genome. The association peak on chromosome C2 includes 33 genome-assigned markers above the Bonferroni-corrected significance threshold (alpha = 0.05; $-\log_{10} P$ value of 6.7; Appendix S6; Figure S3). These delineated a genomic region containing 39 genes, including an orthologue of *VTE4*, which encodes γ-tocopherol methyl transferase (γ-TMT), an enzyme that converts γ-tocopherol into α-tocopherol (Figure 1). A homoeologous region including a duplicate copy of the *VTE4* gene within the association peak on chromosome A2 was observed, whereas there was no obvious candidate gene in the region of chromosome A10 showing associations. Four transcript abundance-based markers above the Bonferroni-corrected significance threshold ($-\log_{10}P$ value of 6.03 for GEMs) were identified on chromosomes C2, C5 and C7 (Figure 4b). The identification of the gene *VTE4* as the most highly associated GEM on chromosome C2 demonstrated the ability for AT to efficiently provide candidate genes associated with traits of interest.

To investigate whether the top selected markers are predictive for the γ/α ratio, we performed a set of 'take-one-out' permutations for the SNP and GEM markers identified from association analysis of 377 accessions adapted from Harper *et al.* (2016). Markers above the Bonferroni line (Appendixes S6 and S7) were selected for each round of permutations. For SNP data, the allelic effects of each of these markers was used to predict trait values for the missing accessions based on their scored genotypes. For GEM data, RPKM values were fitted to the regression line to predict trait values. The predicted trait values against the observed traits are illustrated as scatter plots in Figure 5, and confirmed their excellent predictive ability (R^2 = 0.59 for SNPs and R^2 = 0.47 for GEMs between predicted and observed values; $P < 0.001$), which reflect the estimated narrow-sense heritability (h^2) of 0.452 for the γ/α ratio. These SNPs and GEMs can therefore be used as promising markers in marker-assisted breeding.

In order to confirm the role of the *VTE4* orthologue in the associated region of C2 (Bo2g050970.1), we used the transcript quantification data that were obtained alongside the transcriptome SNP data as part of the functional genotypes. As illustrated in Figure 6, these show that the expression level of Bo2g050970.1 in the tissue sampled to produce the functional genotypes (leaves) is negatively correlated with the γ/α ratio (R^2 = 0.41, $P < 0.001$). This is consistent with the predicted γ-TMT activity of the gene encoded by Bo2g050970.1 (i.e. lower expression leading to less conversion of γ-tocopherol to α-tocopherol). There had been no significant associations between SNPs within Bo2g050970.1 and the γ/α ratio, consistent with the basis of the allelic variation being variation in gene expression rather than variation in gene sequence.

DISCUSSION

Association studies are becoming increasingly widely used in crops for identifying molecular markers linked to trait-controlling loci (Rafalski, 2010); however, polyploid crops present additional difficulties that must be overcome, including the intrinsic genome complexity and increased genome structural instability, such as the copy-number variations (CNVs) that affect gene families (Zhang *et al.*, 2013; Renny-Byfield and Wendel, 2014). Such difficulties occur in *B. napus*, as was recently shown by Chalhoub *et al.* (2014) and He *et al.* (2016). Association studies have to meet many demands to maximize the probability of identifying marker–trait associations. In addition to good experimental design, along with access to all the necessary equipment and available funds, there is also the need to choose a permanent and sufficiently large set of diverse and preferably homozygous individuals, the larger size and higher genetic diversity of which providing sufficient power for association analysis (Spencer *et al.*, 2009; Huang and Han, 2014). Once assembled, association panels need to be genotyped with molecular markers to a sufficiently high density to identify polymorphisms in linkage disequilibrium with trait-controlling loci. The development of suitable association panels is challenging for individual research groups, providing a driver for the development of community resources.

© 2017 The Authors.

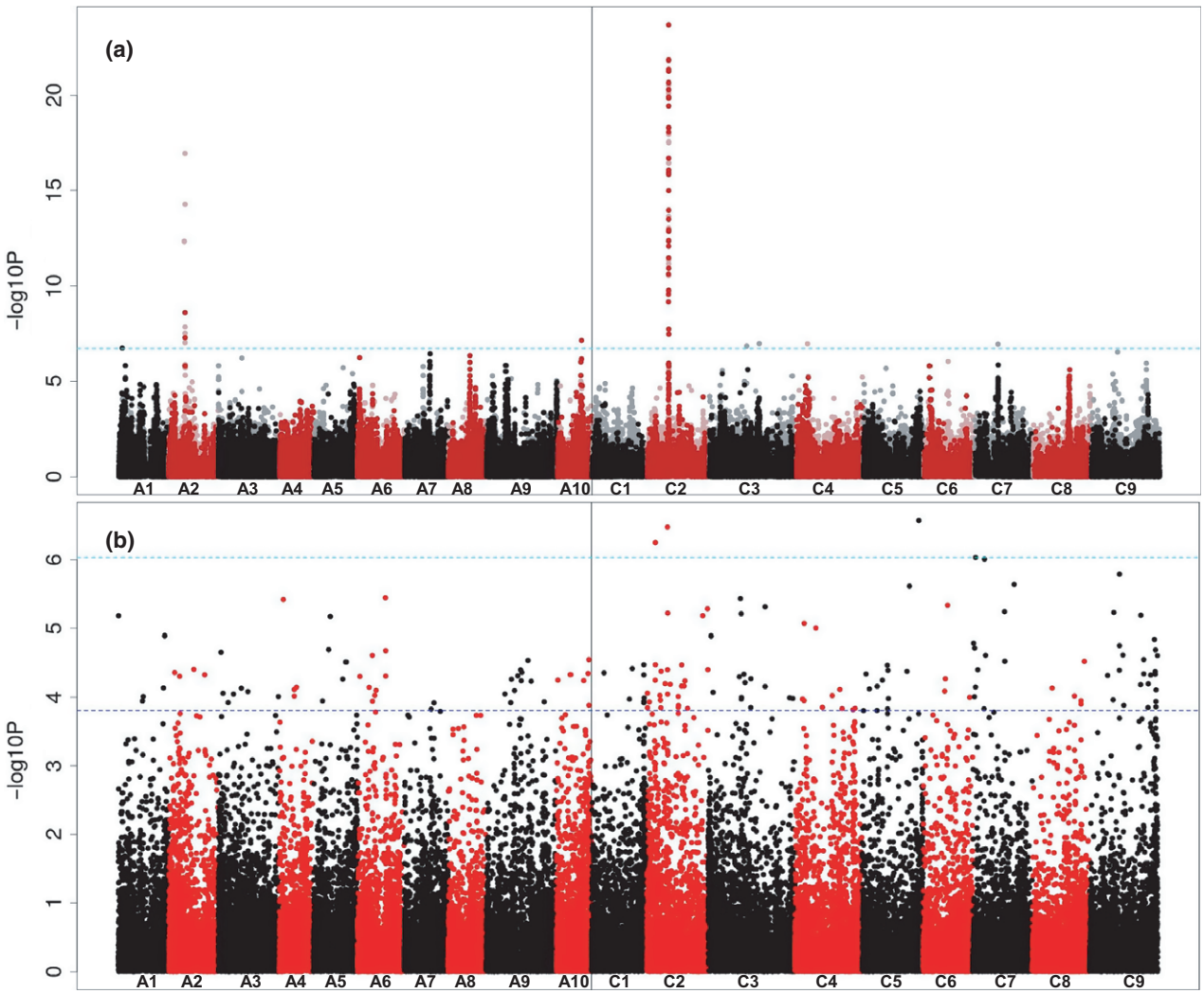

Figure 4. Association analysis.
(a) Transcriptome single-nucleotide polymorphism (SNP) association analysis for seed γ/α-tocopherol ratio. The SNP markers are positioned on the *x*-axis based on the genomic order of the gene models in which the polymorphism was scored, with the significance of the trait association, as −log10*P*, plotted on the *y*-axis. A1–A10 and C1–C9 are the chromosomes of *Brassica napus*, shown in alternating black and red colours to permit boundaries to be distinguished. Hemi-SNP markers (i.e. polymorphisms involving multiple bases called at the SNP position in one allele of the polymorphism) for which the genome of the polymorphism cannot be assigned are shown as light points, whereas simple SNP markers (i.e. polymorphisms between resolved bases) and hemi-SNPs that have been directly linkage-mapped, both of which can be assigned to a genome, are shown as dark points. The broken light-blue horizontal line marks the Bonferroni-corrected significance threshold of 0.05.
(b) Association analysis of transcript abundance with seed γ/α-tocopherol ratio. The gene models are positioned on the *x*-axis based on their genomic order, with the significance of the trait association, as −log10*P*, plotted on the *y*-axis. The broken dark-blue horizontal line marks the 5% false discovery rate.

In this study, we introduce a new genetically diverse AT panel of 383 rapeseed accessions, together with a mapping platform that comprises complete genotype information for this panel, which may be used for a broad range of association studies suitable for re-phenotyping any trait, without the need of additional genotyping. This panel, being made available with all transcriptomic data, offers a large range of potential applications: identifying causative genes, uncovering unknown pathways, identifying regulatory genes or transcription factors, and screening of available germplasm for allelic variants and to support the development of molecular markers for marker-assisted breeding. Our resource provides 355 536 SNP markers, equivalent to one SNP every 0.33 kb across our *B. napus* AC pan-transcriptome reference. The SNP density is much higher than the density of the commercially available 60K Brassica Infinium® SNP array, which only provided 26 841 or 21 117 SNPs for recent *B. napus* GWAS studies (Li *et al.*, 2014; Xu *et al.*, 2016). Although the number of SNPs can even be greater when using whole-genome resequencing, as shown by Huang *et al.* (2013), the advantage of transcriptome resequencing using mRNAseq is the availability

© 2017 The Authors.

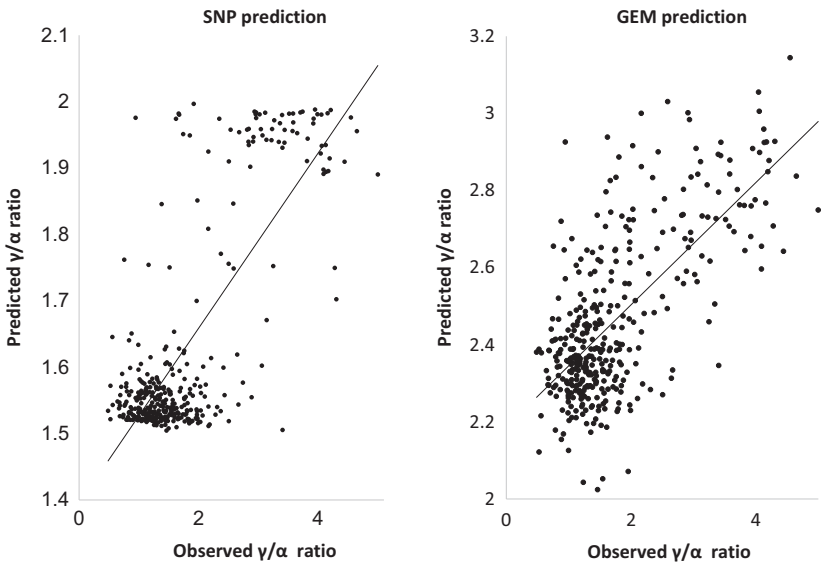

Figure 5. Test of the predictive ability of single-nucleotide polymorphisms (SNPs) and gene expression markers (GEMs) associated with γ/α-tocopherol ratio by 'take-one-out' permutation. The allelic effects of each of 36 SNP markers associated with the γ/α-tocopherol ratio was used to predict the γ/α-tocopherol ratio for the missing accessions. For GEM data, reads per kilobase per million (RPKM) values for each of four GEMs were fitted to the regression line to predict the γ/α-tocopherol ratio. The strong correlation between predicted and observed γ/α-tocopherol ratio values ($R^2 = 0.59$, $P < 0.001$ for SNPs; $R^2 = 0.47$, $P < 0.001$ for GEMs) demonstrates excellent predictive ability.

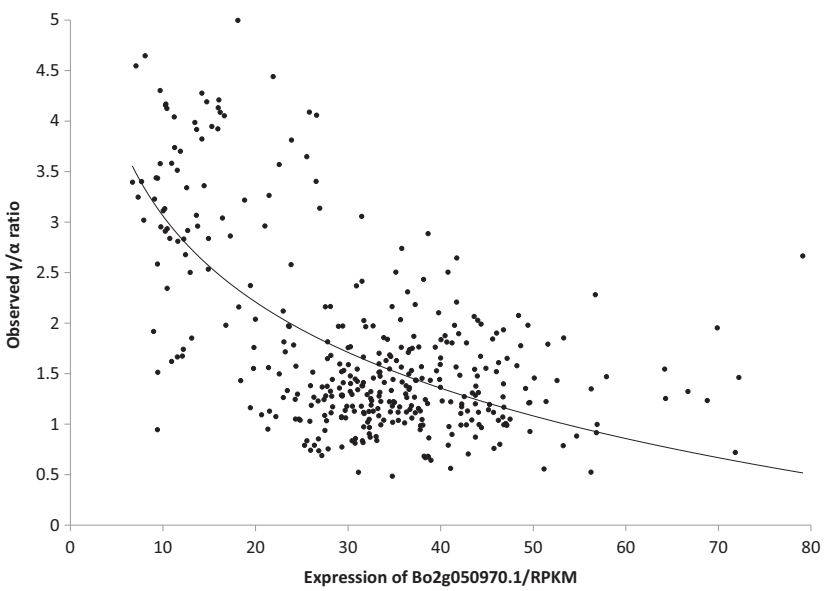

Figure 6. Relationship between the expression of *Bo2g050970.1* in leaves and the γ/α-tocopherol ratio in seeds. The ratio of γ/α-tocopherol measured in seeds was regressed against the transcript abundance in leaves of the VTE4 orthologue *Bo2g050970.1* ($R^2 = 0.26$; $P < 0.001$), measured as reads per kilobase per million aligned reads (RPKM).

of transcript abundance data: in our case for 46% of the genes present in the AC pan-transcriptome reference sequence. In this study, we demonstrate a significant step-change in resolution from our original AT platform based on a panel of 84 accessions, as reported in Harper *et al.* (2012). The unigene-based transcriptome reference sequence used by that platform had relatively poor capability to resolve homoeologous loci, because of its construction based on a Brassica-wide transcriptome assembly and subsequent 'curing' to more closely match the progenitor genomes. In the absence of the ability to map sequence reads unambiguously to the correct homoeologue, most SNPs appear, because of cross-mapping, as 'hemi-SNPs', i.e. where one allele comprises a mixture of two bases (Trick *et al.*, 2009). In the original platform only a small proportion of markers could be assigned with high confidence to a genome, the majority being assigned to both homoeologous positions. The new platform is based mainly on gene models originating from the genome sequences of the progenitor species, and permits more discriminating read-mapping, resulting in a greater proportion of 'simple SNPs' (i.e. where the polymorphism is between resolved single bases only) that can be assigned with confidence to a genome. Where there are association peaks comprising pale points in homoeologous positions to the associations identified, such as those observed in the regions of A2 depicted in Figure 4a, these can be disregarded as homoeologous 'shadows' of the regions genuinely containing causative variation. The discovery of SNPs for particular genes from juvenile leaves can be

© 2017 The Authors.

limited by their transcription in different phenological stages or tissues, but candidate loci/genes associated with traits manifesting in different times or places can be identified, as demonstrated here in the case of *FAE1* and in previous AT studies (Lu *et al.*, 2014; Wood *et al.*, 2017). This is possible because of the presence of variation in genes in LD with the causative gene, resulting in an associated region including the control gene. In addition, the new platform provides much greater resolution of the contributions to the transcriptome of pairs of homoeologous genes. This permitted the efficient detection of association peaks based solely on transcript abundance variation, as illustrated in Figure 3. Moreover, the current platform also allows a deeper insight into the structural changes and functional interactions between *B. napus* AC genomes. Information about respective homologous genes, including their copy number, sequence variation and transcript prevalence provides important information in polyploid research.

In addition to extending previous association studies of the control of seed erucic acid content, a trait selected recently by rapeseed breeders, we applied the platform to a trait not previously selected by breeders or studied extensively: the control of tocopherol (vitamin E) forms accumulated in seeds. We analysed seed tocopherols in 377 rapeseed accessions for their type and content. The profiles presented here showed a high degree of variability for the γ/α-tocopherol ratio (Coefficient of Variance = 53%), displaying distinct patterns for different crop types, that allowed us to identify gene *Bo2g050970.1* (an orthologue of the Arabidopsis gene *VTE4*) on chromosome C2 as a candidate gene, based on inference of gene function from studies of its orthologue in *A. thaliana*. Although there was no evidence of the presence of any specific allelic form of the *VTE4* orthologue associated with γ/α-tocopherol ratio, this gene has been easily identifiable by the presence of SNPs in surrounding genes. This set of tightly linked markers exhibited excellent predictive ability (Figure 5), which we attribute to the broad (species-wide) range of genetic variation represented by the RIPR diversity panel, overcoming the lack of predictive capability that can be encountered when applying markers to test material (Bush and Moore, 2012). The association that we observed between transcript abundance of *Bo2g050970.1* in leaves and the γ/α-tocopherol ratio in seeds is consistent with our understanding that tocopherols are synthesized and localized in plastids and accumulate in all tissues, with generally the highest content in seeds (Sattler *et al.*, 2004). In Arabidopsis, γ-TMT (*VTE4,* AT1G64970) is known to use δ- and γ-tocopherols as substrates to produce β- and α-tocopherols, respectively (Shintani and DellaPenna, 1998), and the effect of the *VTE4* gene from *B. napus* on α-tocopherol content has also been proven by overexpression in *Glycine max* (soya bean) and Arabidopsis (Endrigkeit *et al.*, 2009; Chen *et al.*, 2012).

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 rapeseed 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

Growth of the genetic diversity panel

The panel of 383 *B. napus* accessions is available from the John Innes Centre (https://www.jic.ac.uk). It was planted in a randomized block design of five biological replicates under controlled conditions of two polytunnels at the University of Nottingham, as described by Thomas *et al.* (2016). The accessions comprise inbred derivatives of both recent and historic varieties and some research lines. Plants were bagged before flowering to prevent cross-pollination. Seeds were collected from individual plants at maturity. Seeds from 377 and 376 accessions were used for the tocopherol and erucic acid measurement, respectively. Based on descriptors originally received with the material and analysis of relatedness, they were attributed to one of seven different groups, namely spring oilseed rape (123), semi-winter oilseed rape (11), swede (27), kale (3), fodder (6), winter oilseed rape (169) or crop type not assigned (44), as listed in Appendix S1.

Measurement of fatty-acid content and composition

For the analysis of fatty acid methyl esters (FAMEs), 30 mg of seeds were homogenized in a glass vial with 5 mL of heptane. To the homogenate, 500 μL of 2 M potassium hydroxide was added, left for 1 h and then neutralized with sodium hydrogen sulphate monohydrate. The upper phase was transferred into crimp-cap Chromacol 0.8-ml vials (https://www.thermofisher.com) for analysis using a DANI Master GC fitted with an SGE-BPX70 double column (https://dani-instruments.com).

Measurement of tocopherol content and composition

The α-, γ- and δ-tocopherol (the sum of which formed total tocopherol, TTC) were extracted from a homogenous mixture of 80 mg rapeseed 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}).

SNP identification and transcript quantification for RNA-seq data

The growth conditions, sampling of plant material, RNA extraction and transcriptome sequencing was carried out as described by He *et al.* (2016). The RNA-seq data from each accession line were mapped onto recently developed ordered Brassica A and

© 2017 The Authors.

C pan-transcriptomes (He *et al.*, 2015) as reference sequences (MAQ 0.7.1; Li *et al.*, 2008). SNPs were called by the meta-analysis of alignments as described in Bancroft *et al.* (2011) of mRNAseq reads obtained from each of the *B. napus* accessions. SNP positions were excluded if they did not have a read depth in excess of 10, a base call quality above Q20, missing data below 0.25, and three alleles or fewer. An additional noise threshold was employed to reduce the effect of sequencing errors, whereby ambiguous bases were only allowed to be called if both bases were present at a frequency of 0.2 or above. This resulted in a set of 355 536 SNPs, of which 256 397 had the second most frequent allele in the population, so called here as a minor allele frequency (MAF) > 0.01. The markers were also classified as those that can be assigned with confidence to the genomic position of the CDS model in which they are scored (simple SNPs and hemi-SNPs genetically mapped into the appropriate genome using the Tapidor Ningyou 7 Doubled Haploid (TNDH) mapping population), and those that cannot, as the polymorphism may be in either homoeologue of the CDS model in which they are scored (hemi-SNPs not genetically mapped into the appropriate genome using the TNDH mapping population). Transcript abundance was quantified and normalized as reads per kb per million aligned reads (RPKM) for each sample for 116 098 CDS models of the pan-transcriptome reference. Significant expression (> 0.4 RPKM) was detected for 53 889 CDS models.

Clustering based on SNP genotypes

Clustering and dendrogram visualization on SNP data was performed by an R script developed in-house. R package 'PHANGORN' was used for generating a distance matrix with the JC69 model (Schliep, 2011).

Assessment of linkage disequilibrium

Pairwise LD was calculated and heat maps were produced for each individual chromosome, and these values were then used to calculate the mean LD across the genome. SNPs were removed from the analysis if they were not confirmed by TNDH population (Qiu *et al.*, 2006) that assigned to the A or C genome, and if their minor allele frequency was below 0.01. A single SNP was selected at random from each CDS model to reduce the effect of many linked SNPs in the same gene. Pairwise R^2 LD matrices and heat maps were calculated for each chromosome using the R package LDHEATMAP 0.99-2 (Shin *et al.*, 2006).

Associative Transcriptomic analysis

Association analysis for SNPs and GEMs was performed using R, as previously described (Harper *et al.*, 2012; Sollars *et al.*, 2017), with modifications. In order to deal with the greatly increased sizes of the data sets, PSIKO (Popescu *et al.*, 2014) was used for Q-matrix generation and the GAPIT R package was used with a mixed linear model (Lipka *et al.*, 2012) for GWAS analysis. For Manhattan plots of SNP associations, SNP markers were filtered to include only those with minor allele frequencies of > 0.01: markers that could be assigned with confidence to the genomic position of the CDS model are rendered as dark points and markers that could not be assigned with confidence were rendered as pale points. For GEM association, CDS models were filtered prior to regression to include only those with mean expression across the panel of >0.4 RPKM. The association between gene expression and traits was calculated by fixed-effect linear model in R, with RPKM values and the Q matrix inferred by PSIKO as the explanatory variables, and with trait score as the response variable. R^2 regression coefficients, constants and significance values were outputted for each regression. Genomic control (Devlin and Roeder, 1999) was applied to the GEM analysis to correct for spurious associations, with *P*-value adjustment applied when the genomic inflation factor (λ) was observed to be greater than 1.

Validation of marker association by trait prediction

The predictive power of the best GEMs and SNPs were assessed using a 'take-one-out' approach (Harper *et al.*, 2016) whereby each accession is removed from the SNP or GEM analysis in turn. An in-house R script was performed with adaptation from Harper *et al.* (2016), with a modification of incorporating all SNPs and GEMs above Bonferroni lines. When permutations finish, an R^2 value is calculated from predicted trait values regressed against the observed trait values, which indicates the predictive power of the top selected GEMs and SNPs.

ACCESSION NUMBERS

Sequence data from this article can be found in the SRA data library under accession number PRJNA309367.

ACKNOWLEDGEMENTS

We thank Neil Graham and Rory Hayden at the University of Nottingham for growing plants and seed collection. Next-generation sequencing and library construction was delivered via the BBSRC National Capability in Genomics (BB/J010375/1) programme at The Genome Analysis Centre by members of the Platforms and Pipelines Group. This work was supported by UK Biotechnology and Biological Sciences Research Council (BB/L002124/1), including work carried out within the ERA-CAPS Research Programme (BB/L027844/1).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

SUPPORTING INFORMATION

Supporting data are available. The largest data sets, representing the functional genotypes of the RIPR panel, are accessible via a data distribution website: http://www.yorknowledgebase.info/.The smaller data sets are hosted as supporting information online.

SUPPORTING INFORMATION

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

Figure S1. Genome-wide linkage disequilibrium analysis for the RIPR diversity panel.

Figure S2. Histograms of seed tocopherol composition of the RIPR diversity panel in different crop types.

Figure S3. Quantile–quantile plots from GEM and SNP association analysis for erucic acid and γ/α-tocopherol ratio.

Appendix S1. List of cultivars, crop type classifications and Illumina read mapping statistics.

Appendix S2. Ordered list of CDS gene model-based Brassica AC pan-transcriptome.

Appendix S3. Seed fatty-acid composition of the RIPR diversity panel.

Appendix S4. Markers and genomic regions showing association with variation for erucic acid content.

© 2017 The Authors.

Appendix S5. Seed tocopherol composition of the RIPR diversity panel.

Appendix S6. Markers and genomic regions showing association with variation for γ/α-tocopherol ratio.

Appendix S7. Gene expression markers showing association with variation for γ/α-tocopherol ratio.

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 a common set of *Arabidopsis thaliana* inbred lines. *Nature*, **465**(7298), 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.

Bancroft, I., Fraser, F., Morgan, C. and Trick, M. (2015) Collinearity analysis of Brassica A and C genomes based on an updated inferred unigene order. *Data Brief*, **3**, 51–55.

Bus, A., Körber, N., Snowdon, R.J. and Stich, B. (2011) Patterns of molecular variation in a species-wide germplasm set of Brassica napus. *Theor. Appl. Genet.* **123**(8), 1413–1423.

Bush, W.S. and Moore, J.H. (2012) Chapter 11: genome-Wide Association studies. *PLoS Comput. Biol.* **8**(12):e1002822

Chalhoub, B., Denoeud, F., Liu, S. *et al.* (2014) Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science*, **345**(6199), 950–953.

Chen, D.F., Zhang, M., Wang, Y.O. and Chen, X.W. (2012) Expression of γ-tocopherol methyltransferase gene from *Brassica napus* increased α-tocopherol content in soybean seed. *Biol. Plant.* **56**(1), 131–134.

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**(7), 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**(50), 21611–21616.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**(4), 997–1004.

Dolde, D., Vlahakis, C. and Hazebroek, J. (1999) Tocopherols in breeding lines and effects of planting location, fatty acid composition, and temperature during development. *J. Am. Oil Chem. Soc.* **76**(3), 349–355.

Endrigkeit, J., Wang, X., Cai, D., Zhang, C., Long, Y., Meng, J. and Jung, C. (2009) Genetic mapping, cloning, and functional characterization of the *BnaX.VTE4* gene encoding α-tocopherol methyltransferase from oilseed rape. *Theor. Appl. Genet.* **119**(3), 567–575.

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), 1–24.

Garrigan, D. and Hammer, M.F. (2006) Reconstructing human origins in the genomic era. *Nat. Rev. Genet.* **7**, 669–680.

Gilliland, L.U., Magallanes-Lundback, M., Hemming, C., Supplee, A., Koornneef, M., Bentsink, L. and DellaPenna, D. (2006) Genetic basis for natural variation in seed vitamin E levels in Arabidopsis thaliana. *Proc. Natl Acad. Sci. USA*, **103**(49), 18834–18841.

Goffman, F.D. and Becker, H.C. (2002) Genetic variation of tocopherol content in a germplasm collection of *Brasscia napus* L. *Euphytica*, **125**(2), 189–196.

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.

Harper, A.L., McKinney, L.V., Nielsen, L.R. *et al.* (2016) Molecular markers for tolerance of European ash (*Fraxinus excelsior*) to dieback disease identified using Associative Transcriptomics. *Sci. Rep-UK*, **6**, 19335.

Hasan, M., Friedt, W., Pons-Kühnemann, J., Freitag, N.M., Link, K. and Snowdon, R.J. (2008) Association of gene-linked SSR markers to seed glucosynolate content in oilseed rape (*Brassica napus* ssp. n*apus*). *Theor. Appl. Genet.* **116**, 1035–1049.

He, Z., Cheng, F., Li, Y., Wang, X., Parkin, I.A., 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 Brief*, **4**, 357–362.

He, Z., Wang, L., Harper, A.L., Havlickova, L., Pradhan, A.K., Parkin, I.A.P. and Bancroft, I. (2016) Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* **15**, 594–604.

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.

Huang, X. and Han, B. (2014) Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* **65**, 531–551.

Huang, S., Deng, L., Guan, M., Li, J., Lu, K., Wang, H., Fu, D., Mason, A.S., Liu, S. and Hua, W. (2013) Identification of genome-wide single nucleotide polymorphisms in allopolyploid crop Brassica napus. *BMC Genom.* **14**, 717.

Inaba, R. and Nishio, T. (2002) Phylogenetic analysis of Brassiceae based on the nucleotide sequences of the S-locus related gene, SLR1. *Theor. Appl. Genet.* **105**, 1159–1165.

Li, J.Z., Absher, D.M., Tang, H. *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

Li, Q., Yang, X., Xu, S. *et al.* (2012) Genome-wide association studies identified three independent polymorphisms associated with α-tocopherol content in maize kernels. *PLoS ONE*, **7**(5), e36807.

Li, F., Chen, B., Xu, K. *et al.* (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). *DNA Res.* **21**, 355–367.

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**(18), 2397–2399.

Lu, G., Harper, A.L., Trick, M., Morgan, C., Fraser, F., O'Neill, C. and Bancroft, I. (2014) Associative transcriptomics study dissects the genetic architecture of seed glucosinolate content in *Brassica napus*. *DNA Res.* **21**(6), 613–625.

Munné-Bosch, S. and Alegre, L. (2002) The function of tocopherols in plants. *Crit. Rev. Plant Sci.* **21**(1), 31–57.

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**(2), 233–243.

Pires, C.J., 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.* **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**(4), 1421–1431.

Qiu, D., Morgan, C., Shi, J. *et al.* (2006) A comparative linkage map of oilseed rape and its use for QTL analysis of seed oil and erucic acid content. *Theor. Appl. Genet.* **114**(1), 67–80.

Quadrana, L., Almeida, J., Otaiza, S.N. *et al.* (2013) Transcriptional regulation of tocopherol biosynthesis in tomato. *Plant Mol. Biol.* **81**(3), 309–325.

Rafalski, J.A. (2010) Association genetics in crop improvement. *Curr. Opin. Plant Biol.* **13**(2), 174–180.

Renny-Byfield, S. and Wendel, J.F. (2014) Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* **101**(10), 1711–1725.

Sattler, S.E., Gilliland, L.U., Magallanes-Lundback, M., Pollard, M. and DellaPenna, D. (2004) Vitamin E is essential for seed longevity and for preventing lipid peroxidation during germination. *Plant Cell*, **16**(6), 1419–1432.

Schliep, K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**(4), 592–593.

Shin, J.-H., Blay, S., McNeney, B. and Graham, J. (2006) LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* **16**, Code Snippet 3, 1–9.

Shintani, D. and DellaPenna, D. (1998) Elevating the vitamin E content of plants through metabolic engineering. *Science*, **282**, 2098–2100.

Sollars, E.S.S., Harper, A.L., Kelly, L.J. *et al.* (2017) Genome sequence and genetic diversity of European ash trees. *Nature* **541**, 212–216

© 2017 The Authors.

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**(17), 7719–7723.

Spencer, C.C.A., Su, Z., Donnelly, P. and Marchini, J. (2009) Designing Genome-Wide Association Studies: sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genet.* **5**(5), e1000477.

Thomas, C.L., Alcock, T.D., Graham, N.S. *et al.* (2016) Root morphology and seed and leaf ionomic traits in a *Brassica napus* L. diversity panel show wide phenotypic variation and are characteristic of crop habit. *BMC Plant Biol.* **16**, 214.

Tian, F., Bradbury, P.J., Brown, P.J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T.R., McMullen, M.D., Holland, J.B. and Buckler, E.S. (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**(6), 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**(4), 334–346.

U, N. (1935) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *Jap. J. Bot.* **7**, 389–452.

Valentin, H.E., Lincoln, K., Moshiri, F. *et al.* (2006) The Arabidopsis vitamin E pathway gene5-1 Mutant Reveals a Critical Role for Phytol Kinase in Seed Tocopherol Biosynthesis. *Plant Cell*, **18**(1), 212–224.

Wang, X., Zhang, C., Li, L., Fritsche, S., Enderigkeit, J., Zhang, W., Long, Y., Jung, C. and Meng, J. (2012) Unraveling the genetic basis of seed tocopherol content and composition in rapeseed (*Brassica napus* L.). *PLoS ONE*, **7**(11), e50038.

Wood, I.P., Pearson, B.M., Garcia-Gutierrez, E.G., Havlickova, L., He, Z., Harper, A.L., Bancroft, I. and Waldron, K.W. (2017) Carbohydrate microarrays and their use for the identification of molecular markers dor plant cell wall composition. *Proc. Natl Acad. Sci. USA*, **114**(26), 6860–6865.

Xu, L., Hu, K., Zhang, Z. *et al.* (2016) Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (Brassica napus L.). *DNA Res.* **23**(1), 43–52.

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**(6), 1339–1347.

Zhang, H., Bian, Y., Gou, X. *et al.* (2013) Intrinsic karyotype stability and gene copy number variations may have laid the foundation for tetraploid wheat formation. *Proc. Natl Acad. Sci. USA*, **110**(48), 19466–19471.

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.

© 2017 The Authors.
The Plant Journal published by John Wiley & Sons Ltd and Society for Experimental Biology.,