ORIGINAL ARTICLE

Towards the genetic architecture of seed lipid biosynthesis and accumulation in *Arabidopsis thaliana*

CM O'Neill[1], C Morgan[1], C Hattori[1], M Brennan[2], U Rosas[3], H Tschoep[4], PX Deng[1], D Baker[1], R Wells[1] and I Bancroft[1]

[1]Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich, UK; [2]Monsanto Company, St Louis, MO, USA; [3]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY, USA and [4]SESVANDERHAVE NV/SA, Tienen, Belgium

We report the quantitative genetic analysis of seed oil quality and quantity in six *Arabidopsis thaliana* recombinant inbred populations, in which the parent accessions were from diverse geographical origins, and were selected on the basis of variation for seed oil content and lipid composition. Although most of the biochemical steps involved in lipid biosynthesis are known and the key genes have been identified, the regulation of the processes that results in the final oil composition and total amount is not understood. By using physically anchored markers it was possible to compare results across populations. A total of 219 quantitative trait loci (QTLs) were identified, of which 81 were significant at $P<0.001$. Some of these colocalise with QTLs identified previously, but many novel QTLs were also identified. The results highlight the importance of studying traits in multiple populations, which will lead to a better understanding of the contribution that natural variation makes to the genetic architecture of a phenotype.
Heredity (2012) **108,** 115–123; doi:10.1038/hdy.2011.54; published online 6 July 2011

Introduction

Seed oils are both components of the human diet and feedstock for industrial products, where they have the potential to substitute for mineral oil. They are energy-rich compounds and, for many angiosperm species are important for reproductive success as they provide much of the carbon and energy needed during germination and establishment (Rees and Long, 1992). Seed storage oils are composed mainly of triacylglycerols. They are synthesised in two stages, with acyl chain synthesis in the plastid being followed by further modification in the endoplasmic reticulum, resulting in the presence of a range of different fatty acids. Many studies of the regulation of lipid metabolism have been conducted in *Arabidopsis thaliana* (Ohlrogge and Browse, 1995), where the availability of its genome sequence (The Arabidopsis Genome Initiative, 2000) enables the use of genomic approaches (White and Benning, 2001; Ruuska *et al.*, 2002). Early mutagenesis experiments in *A. thaliana* produced a series of mutants in which fatty acid synthesis genes were inactivated, leading to the identification of the major genes involved (Okuley *et al.*, 1994; James *et al.*, 1995; Katavic *et al.*, 1995; Focks and Benning, 1998; Poirier *et al.*, 1999). Although most of the biochemical steps involved in lipid biosynthesis are known and the key genes have been identified (Beisson *et al.*, 2003), the regulation of the processes that results in the final lipid composition is not understood. The observation that there is a coordinated transcription of many genes involved in the pathways is indicative of the existence of regulatory loci in addition to the components of the biochemical pathways already identified (Ruuska *et al.*, 2002).

The genetic networks underlying complex processes are typically under the control of many genes. The impact of allelic variation of individual genes upon the resulting phenotype may be small, making them difficult to identify (Glazier *et al.*, 2002). The method most commonly used for inferring the presence and position of such genes in the genome is based upon analysis as quantitative trait loci (QTLs). Such quantitative genetic approaches provide estimates of several genetic parameters that underlie phenotypic variation, including the number of loci, the type and magnitude of their effects, interactions between genes (epistasis), and gene by environment interactions.

The analysis of natural variation can provide an approach to the identification of genes in *A. thaliana* (Koornneef *et al.*, 2004; Tonsor *et al.*, 2005) and has been successfully applied to a range of traits, such as downy mildew resistance (Nemri *et al.*, 2010), seed dormancy (Bentsink *et al.*, 2010) and flowering time (Shindo *et al.*, 2005). Much of this work will be further aided with the output of the 1001 genomes sequencing project (http://www.1001genomes.org/). As storage lipid accumulation may be regarded as an adaptive trait (Linder, 2000),

116

and even small modifications of seed oil content and/or fatty acid composition potentially confer an advantage in specific environmental conditions, natural variation may be useful for the identification of regulatory mechanisms. A survey of 360 accessions of *A. thaliana* for seed oil content, very long chain fatty acids and polyunsaturated fatty acids revealed extensive natural variation for these traits (O'Neill *et al.*, 2003).

As a largely inbreeding species, recombinant inbred (RI) populations can readily be developed and maintained for *A. thaliana*. These represent 'immortal' populations, as each line is practically homozygous and can be propagated indefinitely by selfing. Currently, 24 RI populations are available from the Arabidopsis Stock Centres, with more under development (http://www.inra.fr/vast/RILs.htm). The genetic control of seed oil content has been studied in an RI population derived from a cross between accessions Cvi-0 and *Landsberg erecta*, with four QTLs being identified, explaining 43% of the genetic variation for that trait (Hobbs *et al.*, 2004). In the same study, QTLs were also identified for the content of the fatty acids palmitic (16:0), stearic (18:0), oleic (18:1), linoleic (18:2), linolenic (18:3) and eicosenoic (20:1) in seed oil. However, studies conducted in a single mapping population, sample only a fraction of the natural allelic variation available within a gene pool and therefore identify only a subset of the loci controlling the traits (Symonds *et al.*, 2005).

Studies using many mapping populations derived from a diverse array of accessions will enable progress towards an understanding of the global genetic architecture of complex traits (Tonsor *et al.*, 2005). We recently developed six new RI populations of *A. thaliana* (O'Neill *et al.*, 2008). The 12 parental accessions used for the development of these populations had been selected so as to represent the range of phenotypic variation identified in the species during our initial survey (O'Neill *et al.*, 2003). We used this set of populations to take a QTL approach to identifying further components of the global genetic architecture of seed lipid biosynthesis and accumulation.

Materials and methods

Plant materials and growth conditions
A total of six RI populations CA (Cvi-0 × Ag-0), KB (Kondara × Br-0), NG (Nok-3 × Ga-0), SG (Sorbo × Gy-0), TJ (Ts-5 × Mz-0) and WC (Wt-5 × Ct-1) were used in this study. These populations were previously described and characterised (O'Neill *et al.*, 2008). Each population consisted of a set of 94 lines. In all 10 seeds from each line of the six populations were sown on Arabidopsis soil mix: Levington's M2 compost with 4 mm grit (8:1) and the insecticide intercept (to protect against aphids) in P40 trays (O'Neill *et al.*, 2003). After sowing, seeds were transferred to a 5 °C chamber, with an 8 h photoperiod for 6 weeks to be vernalised. The plants were then moved to either a glasshouse (GH) or a controlled environment room (CER). Both provided a 22 °C day/18 °C night, 16 h photoperiod with supplementary light provided by sodium lamps. When the first true leaves were fully developed, seedlings were pricked out to Arabidopsis soil mix in 7 cm pots.

All populations, including their parents, were grown using randomised block designs. A minimum of three replicate plants of both the parents were grown in each block, however, because of space limitations, only one plant of each line of the RI populations was grown in each block. Three of the populations (SG, KB and TJ) were grown as single experiments in the GH in three randomised blocks, while two of the populations (CA and WC) were grown in the CER on two separate occasions in two randomised blocks, which were split into two equal sub-blocks consisting of half of the RI population and also grown in the GH in a randomised trial with two blocks. The sixth population, NG, was grown under CER conditions in three randomised blocks. In all experiments the plant trays were rotated 180° each week to reduce positional effects. Watering was stopped when the oldest siliques had turned brown. Seed was collected from the dried mature plants.

Seed lipid analysis
The seed oil content (percent w/w) was determined using NMR spectroscopy. Samples weighing either 50 mg or 200 mg were analysed using a benchtop Oxford MQC NMR machine (Oxford Instruments, Oxford, UK). Analysis was performed in accordance with ISO 10565 guideline (International Standard Organisation, 1993). The oil content of the seeds was expressed as % dry weight, as calculated by the instrument. The machine measurements were calibrated using canola standards supplied by the Canadian Grain Council. Seed fatty-acyl composition was determined according to the method of James and Dooner (1990). Samples were extracted in hexane and analysed using gas chromatography. The conditions allowed the detection of the seed fatty acid methyl esters: 16:0, 18:0, 18:1, 18:2, 18:3, 20:0, 20:1, 20:2 and 22:1. For peak identification, an oil reference standard for low erucic acid rapeseed (Sigma-Aldrich Company Ltd., Gillingham, UK) was also analysed. To quantify the overall levels of fatty acid elongation and desaturation in lines, we calculated ratios for carbon chain length and desaturation. The chain length ratio (CL) is the sum of areas under peaks for all 20 and 22 carbon fatty acids divided by the sum of areas under peaks for all 16 and 18 carbon fatty acids. The desaturation ratio (DS) is the sum of areas under peaks for polyunsaturated 18 carbon fatty acids divided by the sum of areas under peaks for saturated and monounsaturated 18 carbon fatty acids.

The population trait data presented in Figures 1a and b are the mean line values as used for QTL mapping for the KB, NG, SG and TJ populations. For the CA and WC populations, normalised average values were used owing to the significant differences in the mean, minimum and maximum trait values for the three occasions, particularly between the GH and CER experiments. Therefore, to facilitate comparison, all values were normalised to the grand mean calculated as the average of the mean for each of the three occasions. This normalisation was done by multiplying each individual value within an occasion by the ratio of that occasion mean to the grand mean.

Statistical analysis
A minimum of three maternal and paternal plants were grown as control material on each growth occasion.
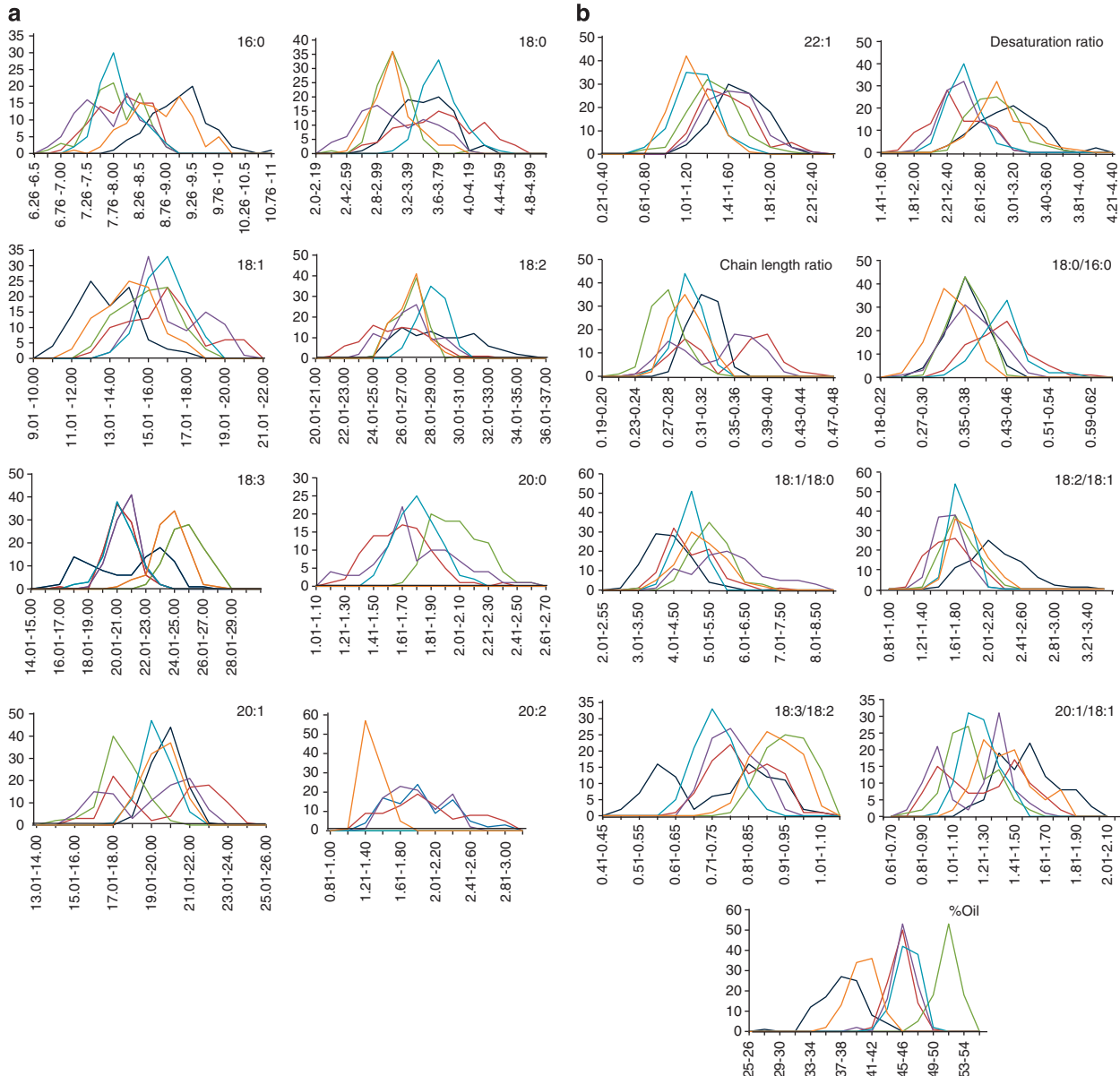
Figure 1 (**a, b**) Line plots of the six RI populations showing the distribution of fatty acids, expressed as the percentage of total seed fatty acid content. DS, CL and fatty acid ratios are also shown. Seed oil content is expressed as percentage of dry weight. The trait values are grouped in bins along the X axis, and the number of plants per bin are given on the Y axis. The line colours are used to denote the populations as follows: CA, Dark blue; KB, Red; NG, Green; SG, Purple; TJ, Light blue; WC, Orange. CA and WC are the normalised means of the CER and the GH experiments.

The KB, NG, SG and TJ parental material was analysed using analysis of variance. To take account of the two different growth environments, the CA and WC parental data were analysed using an unbalanced regression analysis. To determine whether the three data sets for the CA and WC populations could be combined for QTL mapping purposes and trait display, they were analysed using analysis of variance. The results of this analysis showed interactions between the GH- and CER-grown material in both data sets; therefore the GH and the CER data were mapped separately. All statistical analyses including randomisation procedures were carried out using GenStat (VSN International Ltd, Hemel Hempstead, UK) (Payne et al., 2009).

Marker development and assay
To increase the genetic mapping resolution over that described in earlier maps (O'Neill et al., 2008), new Insertion-Deletion markers were identified using the MSQT website (Warthman et al., 2007) and mapped onto the existing maps. The final maps have on average 75 markers. The complete marker set used for the individual populations is available on http://www.jic.bbsrc.ac.uk/staff/ian-bancroft/. Genomic plant DNA was extracted using the high throughput Qiagen Dneasy 96 preparation method by following the manufacturer's instructions (Qiagen GmbH, Hilden, Germany). The markers were amplified using PCR and the products were resolved on an ABI 3730 capillary sequencer to

Figure 2 The physical positions of QTLs identified for fatty acid contents, ratios and oil contents across the six Arabidopsis RI populations. CA and WC are described under both CER and GH environments. The positions were determined using the marker with the most significant LOD value for each QTL. The map was constructed using the physical positions of the markers used to generate the genetic maps. The populations and environments are colour coded. Coincident QTLs share the same line and where there are many a vertical line is used to indicate the grouping with the midpoint of the line at the marker position. QTLs were considered to be coincident if the marker positions were 0.5 Mb or less apart. The chemical symbols used represent the following: 16:0, palmitic acid; 18:0, oleic acid; 18:1, stearic acid; 18:2, linoleic acid; 18:3, linolenic acid; 20:0, arachidic acid; 20:1, eicosenoic acid; 20:2, eicosadienoic acid; 22:1, erucic acid; oil, % of total dry seed weight; 18:0/16:0, the ratio of oleic acid and palmitic acid; 18:1/18:0, the ratio of stearic acid and oleic acid; 18:2/18:1, the ratio of linoleic acid and stearic acid; 18:3/18:2, the ratio of linolenic acid and linoleic acid; CL (calculated as described in Material and methods) and DS (calculated as described in Material and methods).

determine allele sizes (Genome Laboratory Services, John Innes Centre, Norwich, UK).

Genetic mapping and QTL analysis

Genetic linkage maps for each RI population were constructed using the Kosambi mapping function in the JoinMap version 3 (van Ooijen and Voorrips, 2001). To identify the regions of the genome containing QTLs, we used interval mapping as implemented by the software package MapQTL version 5 (van Ooijen et al., 2002). Genome wide significance thresholds of 5% for the Logarithm of Odds ratio (LOD) values were determined using the permutation command within MapQTL (van Ooijen, 2004). The number of permutations was set at 10 000. The maximum LOD values presented were always for the marker with the maximum LOD value. As, all of the markers used were physically anchored within the Arabidopsis genome, it was possible to use

the base coordinates of the marker at a QTL peak and position the associated QTL on a physical Arabidopsis map (Figure 2). Markers that differed by 0.5 Mb or less were treated as coincident.

The mean trait values of the three replicate plants grown, in single environments, for the KB, NG, SG and TJ populations were used for QTL mapping and to estimate broad-sense heritabilities. As the statistical analysis of the CA and WC populations showed $G \times E$ interactions between the GH- and CER-grown material these were therefore mapped separately.

Results

Seed lipid characteristics of the parents of the mapping populations

Plants from the accessions used to generate the six populations were grown alongside the populations and

Table 1 Percentage seed fatty acid and oil contents in the population parental accessions

	Cvi-0	Ag-0	s.e.d.	Kondara	Br-0	s.e.d.	Nok-3	Ga-0	s.e.d.	Sorbo	Gy-0	s.e.d.	Ts-5	Mz-0	s.e.d.	Wt-5	Ct-1	s.e.d.
16:0	8.9	9.3	0.15	7.6	8.5	0.06	8.6	7.4	0.15	7.0	8.7	0.11	10.4	9.1	0.13	9.5	7.9	0.15
18:0	2.8	3.4	0.16	3.0	4.0	0.05	3.2	3.1	0.04[NS]	2.4	3.2	0.05	3.6	4.1	0.09	2.8	3.1	0.07
18:1	12.2	14.1	0.31	16.4	15.9	0.21	16.4	14.7	0.31	19.7	14.0	0.25	14.9	19.5	0.27	14.1	17.4	0.36
18:2	32.7	26.6	0.80	30.8	22.8	0.26	27.3	27.7	0.36[NS]	30.8	24.6	0.47	24.9	27.0	0.35	25.3	29.0	0.44
18:3	21.1	22.9	0.54	20.6	21.2	0.17	23.7	25.1	0.29	20.5	22.0	0.30	20.4	17.1	0.29	25.0	21.2	0.30
20:0	–	–	–	1.5	1.9	0.08	2.1	2.4	0.04	1.2	2.2	0.05	2.0	1.9	0.05[NS]	–	–	–
20:1	18.2	20.8	0.34	16.6	22.4	0.17	18.2	18.2	0.66[NS]	15.6	21.7	0.34	21.1	19.1	0.38	20.7	19.1	0.32
20:2	2.6	1.5	0.09	2.1	1.8	0.03	–	–	–	1.6	1.9	0.03	1.5	1.2	0.03	1.4	1.4	0.06[NS]
22:1	1.5	1.5	0.78[NS]	1.4	1.4	0.04[NS]	1.4	1.7	0.14[NS]	1.2	1.8	0.04	1.4	1.0	0.05	1.3	1.1	0.06
CL	0.3	0.3	0.01	0.3	0.4	0.00	0.3	0.3	0.01[NS]	0.2	0.4	0.01	0.4	0.3	0.01	0.3	0.3	0.01
DS	3.6	2.9	0.11	2.7	2.2	0.05	2.6	3.0	0.04	2.3	2.7	0.04	2.5	1.9	0.04	2.9	2.4	0.08
18:0/16:0	0.3	0.4	0.02	0.4	0.5	0.01	0.4	0.4	0.01	0.3	0.4	0.00	0.3	0.4	0.01	0.3	0.4	0.01
18:1/18:0	4.5	4.4	0.31[NS]	5.4	4.0	0.04	5.2	4.8	0.08	8.4	4.4	0.20	4.2	4.8	0.16	5.1	5.6	0.17
18:2/18:1	2.7	1.9	0.10	1.9	1.4	0.03	1.7	1.9	0.03	1.6	1.8	0.03	1.7	1.4	0.02	1.8	1.7	0.06[NS]
18:3/18:2	0.6	0.9	0.03	0.7	0.9	0.01	0.9	0.9	0.01	0.7	0.9	0.02	0.8	0.6	0.02	1.0	0.7	0.01
20:1/18:1	1.5	1.5	0.05[NS]	1.0	1.4	0.02	1.1	1.3	0.05	0.8	1.6	0.04	1.4	1.0	0.03	1.5	1.1	0.03
Oil	35.4	38.4	1.21	42.7	47.2	0.35	51.3	51.6	0.7[NS]	44.8	45.3	0.27[NS]	38.1	43.6	1.08	39.2	40.8	0.53

Abbreviations: CL, chain length ratio; DS, desaturation ratio; s.e.d., standard error of the difference between two means; NS, not significant.

examined for both their seed fatty acid composition and total oil content. The results are summarised for all parents in Table 1. The parental accessions of the CA population were Cvi-0 and Ag-0. The most striking difference in the fatty acid profiles between these accessions was in the relative amounts of 18:2, Cvi-0 had an average of 32.7% compared with an average of 26.6% in Ag-0. This difference was reflected in the 18:2/18:1 and the desaturation ratios, both of which were higher in Cvi-0 with values of 2.7 and 3.6, respectively as compared with 1.9 and 2.9 found in Ag-0. The KB population was derived from a cross between Kondara and Br-0. The 18:2 content of these accessions was different; Kondara had an average of 30.8% compared with 22.8% in Br-0. The reverse was the case for 20:1 content for which Br-0 had 22.4% and Kondara had just 16.6%. These differences affected both fatty acid desaturation levels as shown by the ratio 18:1/18:0 (5.4 for Kondara; 4.0 for Br-0) and 18:3/18:2 (0.7 for Kondara; 0.9 for Br-0), and chain elongation CL (0.3 for Kondara; 0.4 for Br-0) and 20:1/18:1 (1.0 for Kondara; 1.4 for Br-0). The NG parental accessions, Nok-3 and Ga-0, were broadly similar for the traits studied. The SG population was generated from a cross between Sorbo and Gy-0. These accessions clearly differed for many of the traits studied. The most marked differences were for 18:1/18:0, where Sorbo had a value twice that of Gy-0 at 8.4 and 4.4, respectively. For 20:1/18:1 Sorbo had a ratio of just 0.8, while Gy-0 had a value twice that at 1.6. Gy-0 had greater amounts of the longer carbon chain fatty acids and a resultant higher CL of 0.4, while Sorbo had less giving a lower ratio value of 0.2. The reverse was the case for 18:1 and 18:2 for which Sorbo had the higher levels. For the TJ population the parental accessions were Ts-5 and Mz-0. The key differences between their seed lipid profiles were in the fatty acid chain elongation processes, such as 18:1, 22:1 and 20:1/18:1, where Ts-5 and Mz-0 had values of 14.9 and 19.5%, 1.4 and 1.0%, and 1.4 and 1.0, respectively. Steps in the desaturation process also showed differences such as the DS, 18:3/18:2, where Ts-5 and Mz-0 had values of 2.5 and 1.9, and 0.8 and 0.6, respectively. Wt-5 and Ct-1 are the parental accessions of the WC population. The key differences between the two parents were in their 18:0/16:0 and 20:1/18:1 ratios, which were 0.3 and 0.4, 1.5 and 1.1 for Wt-5 and Ct-1, respectively.

Seed lipid characteristics of the populations

The seed lipid characteristics of the six populations were determined and values compared across the populations (Figures 1a and b, Supplementary Table 1). The CA population frequently showed the most extreme values found for many of the traits studied; thus, 18:1, 18:3, 18:1/18:0, 18:3/18:2 and oil tended to have lower values, while 16:0, 18:2, 18:2/18:1, 20:1/18:1 and DS had higher values. In KB the range of values for a number of traits exceeded those of the other populations; thus, 18:2, 18:2/18:1, 20:1/18:1 and DS had some of the lowest values found, while 18:0, 18:1, 20:1, 20:2, 18:0/16:0 and CL had some of the highest values found across the six populations. NG displayed the lowest 20:1 and CL values, and the highest 18:3, 20:0, 18:3/18:2 and oil values found across the six populations. SG also exhibited traits with the lowest and highest values found among the six populations. Those traits with low values were 16:0, 18:0, 18:2, 20:0, 20:1, DS, 18:0/16:0, 18:2/18:1 and 20:1/18:1, while those displaying some of the highest values found were 18:1, 20:0, 22:1, CL and 18:1/18:0. In TJ, those traits with some of the lowest values were 18:3, 22:1, DS and 18:3/18:2, while those at the higher end of the range were 18:0, 18:2 and 18:0/16:0. WC exhibited some of the lowest values found across the six populations for 18:0, 18:1, 20:2, 22:1, CL, 18:0/16:0 and oil with higher values for 16:0, 18:3, DS, 18:3/18:2 and 20:1/18:1.

Improving the linkage maps of the RI populations

Although shown to be adequate for QTL analysis, the initial linkage maps developed for these six populations (O'Neill et al., 2008) contained a number of relatively large gaps. To improve the coverage and genetic mapping resolution, new Insertion-Deletion markers were identified using the MSQT website (http://www.msqt.weigelworld.org/). The final maps have on average 75 markers from a pool of 185 markers, which

were shared between the six populations. Details of the complete marker set used for the individual populations are available on http://www.jic.bbsrc.ac.uk/staff/ian-bancroft/.

QTL analysis in the CA population

As the CA population was grown in two different environments (CER and GH), QTL analysis was carried out separately for each environment. Details of the QTLs identified in each environment are given in Figure 2, Supplementary Figures 1 and 2 and Supplementary Tables 2 and 3. In all 23 QTLs were detected in the CER and 22 in the GH material, of which 12 are likely to be coincident under both conditions, while the remainder are environment-specific. QTLs significant at $P < 0.001$ were found for 16:0, 18:0, 18:2, 18:3, 20:2, 18:0/16:0, 18:2/18:1 and 18:3/18:2. Both 16:0 and 18:0 had strong, coincident QTLs in the region of markers nga63 and f16j7trb (3.2–3.8 Mb) on linkage group (LG) 1. The QTL with the greatest LOD value (32.7) was identified for 18:3 under GH conditions, and accounted for 80.1% of the variation found for this trait. This QTL was positioned on LG 2 around the marker t32f646516 (13.8 Mb). Coincident QTLs were observed for 18:2, 20:2, and 18:3/18:2. These QTLs explained a large proportion of the phenotypic variance. Other strong QTLs were identified for 20:2, 22:1, 18:2/18:1 and the DS, all at the marker nga162 (4.6 Mb) on LG 3. Additional, less significant, QTLs detected in CA are described in Supplementary Tables 2 and 3.

QTL analysis in population KB

A total of 37 QTLs were detected in the KB population (Figure 2, Supplementary Figure 3 and Supplementary Table 4), of which 19 were significant at $P < 0.001$. Two QTLs for 16:0 content mapped to LG 1, the first at nga63 (3.2 Mb) had a LOD score of 10.0, while the second, at MS_At1_24.7 (24.7 Mb), had a LOD score of 6.8. Neither QTL co-located with any other trait. A large QTL detected for 18:0 on LG 2 at MS_At2_17.5 (17.5 Mb) co-located with QTLs for 20:0, 18:0/16:0 and 18:1/18:0. Strong QTLs for the traits 18:1, 18:2, 20:2, 18:2/18:1, 18:3/18:2 and DS mapped onto LG 3 around nga162 (4.6 Mb). A QTL affecting 18:1/18:0 that co-located with 16:0 and 20:0 was found at MS_At2_14.2 (14.2 Mb) on LG 2. The QTL with the greatest LOD value (23.6) was detected for 20:1 on LG 4 around nga1139 (16.4 Mb) and accounted for 71.1% of the variation. QTLs were also found at this marker for 18:0, 18:1, 20:2, 22:1, 20:1/18:1, CL and DS. A QTL for oil content (LOD 6.5, 28.2% of variation explained) that co-located with 18:2, 18:2/18:1, 18:3/18:2 and DS was mapped towards the end of LG 2 at 90j19t7 (19.6 Mb). Additional, less significant, QTLs detected in KB are described in Supplementary Table 4.

QTL analysis in population NG

In all 20 QTLs were detected in the NG population (Figure 2, Supplementary Figure 4 and Supplementary Table 5), of which 8 were significant at $P < 0.001$. The QTL with the greatest LOD value (12.5) was found for 18:1 on LG 5 at the marker nga151 (4.7 Mb) and explained 45.7% of the variation. QTLs were also identified at this marker for DS, while 20:0, 22:1, 18:1/18:0, 18:2/18:1 and 20:1/18:1 all had QTLs at the adjacent marker f14f1847646

(3.8 Mb). Other QTLs that co-located were detected for 20:0, 18:0/16:0, 18:1/18:0 and 20:1/18:1 at marker t27k12-sp6 (16 Mb) on LG 1, while a QTL for 18:0 was detected at the adjacent marker f140426482 (13.2 Mb). Additional, less significant, QTLs detected in NG are described in Supplementary Table 5.

QTL analysis in population SG

A total of 39 QTLs were detected in the SG population (Figure 2, Supplementary Figure 5 and Supplementary Table 6), of which 24 were significant at $P < 0.001$. Two major QTLs for 16:0 mapped to LG 1, the first at MS_At1_18.2 (18.2 Mb) had a LOD value of 10.2 and co-located with QTLs for 20:0 and 18:1/18:0 and the second, at t12p18ind8-8 (23.5 Mb), had a LOD value of 13.2. A third QTL for 16:0 content (LOD 4.2) mapped to LG 2 at MS_At2_16.9 (16.9 Mb) and co-located with QTLs for 20:0 and 18:1/18:0. A major QTL for 18:0 (LOD 14.9) on LG 2 at MS_At2_18.5 (18.5 Mb) accounted for 52.2% of the variation. A QTL (LOD 11.8) for 18:0/16:0 mapped to the same marker. QTLs for 18:0, 18:0/16:0 and 18:1/18:0 mapped to LG 3 at MS_At3_16.9 (16.9 Mb), while a QTL for 18:3/18:2 (LOD 6.1) mapped to the adjacent marker MS_At3_17.3 (17.3 Mb). QTLs were also found for 18:2, 20:2, DS and 18:2/18:1 on LG 3 at nga162 (4.6 Mb). The QTL with the greatest LOD value (31.8) for a single fatty acid detected in SG was on LG 4 (16.2 Mb). This affected 20:1 content and accounted for 80.7% of the variation. QTLs for other traits co-located at this marker: 18:1, 20:0, 18:3, 20:2, 22:1, CL, DS, 18:2/18:1 and 20:1/18:1, while QTLs were detected at the adjacent marker, nga1139 (16.4 Mb), for 18:2 and 18:1/18:0. Additional, less significant, QTLs detected in SG are described in Supplementary Table 6.

QTL analysis in population TJ

In all 37 QTLs were detected in the TJ population (Figure 2, Supplementary Figure 6 and Supplementary Table 7), of which 11 were significant at $P < 0.001$. The QTL with the greatest LOD value (LOD 10.3) in TJ, affected 16:0 and mapped to LG 1 at MS_At1_22.9 (22.9 Mb). This QTL explained 39.7% of the variation. QTLs for both 18:0 and 18:0/16:0 were detected at the adjacent marker MS_At1_22.1 (22.1 Mb). Another QTL for 16:0 (LOD 5) towards the top of LG 1 at MS_At1_4.3 (4.3 Mb) accounted for 21.8% of the variation and co-located with QTLs for 18:0 and 18:0/16:0. A third QTL (LOD 5.3) that affected 16:0 mapped to LG 5 at mth12 (24.1 Mb). QTLs for 18:1 and DS also mapped to this marker. A QTL (LOD 4.6) affecting 18:1 content was found on LG 1 at aths0392 (10.9 Mb) and accounted for 20.3% of the variation. Co-locating QTLs were found for 18:3 and DS, while QTLs for 18:0, 18:0/16:0, and 18:3/18:2 were detected at the adjacent marker nga392 (9.8 Mb). QTLs for 18:2 (LOD 5.1), and 18:3/18:2 (LOD 3.4), lay on LG 3 at the marker msd2129380 (7.7 Mb). Additional, less significant, QTLs detected in TJ are described in Supplementary Table 7.

QTL analysis in population WC

A total of 28 QTLs were detected in the WC CER-grown material (Figure 2, Supplementary Figure 7 and Supplementary Table 8) and 13 were detected in the GH-grown material (Figure 2, Supplementary Figure 8

and Supplementary Table 9). Of these, six QTLs were significant at $P < 0.001$, four in the CER environment and two in the GH environment. Only one QTL, for 16:0, was common to both the environments and lay on LG 1 at MS_At1_22.9 (22.9 Mb). In the CER it had a LOD value of 4.4 and explained 19.8% of the variation, but in the GH data it had a LOD value of 8.3 and accounted for 33.3% of the variation. A QTL for oil content with a LOD value of 4.7 was found at the same position in the CER grown material. The two remaining strong QTLs detected within the CER material were associated with 20:2 (LOD 4.3) on LG 1 at athsrp54a (5.2 Mb) and 18:2/18:1 (LOD 6.1) on LG 5 at ciw14 (2.1 Mb). In the GH material, the second of the two highly significant QTLs detected in this material was associated with 18:0/16:0 (LOD 6) and mapped to LG 1 at nga111 (27 Mb). Additional, less significant, QTLs detected in WC under the two environments are described in Supplementary Tables 8 and 9.

Discussion

The genetic variation described in this paper for seed lipid traits captured within six phenotypically variable *A. thaliana* RI populations provides an excellent starting point to study allelic variation of the major components of the biosynthetic pathway that lead to phenotypic differences and also to identify new genes important for the regulation of storage lipid biosynthesis. The crosses included two Asian accessions (Kondara and Sorbo), nine European accessions (Ag-0, Br-0, Ct-1, Ga-0, Gy-0, Mz-0, Nok-3, Ts-5 and Wt-5) and one from the Cape Verde Islands (Cvi-0). Differences between the population parents for the traits studied were in most cases reflected within each of the populations (Table 1 and Supplementary Table 1), thus demonstrating the suitability of these populations for identifying seed lipid biosynthesis QTLs and leading to the high number of QTLs detected. In general, it was noted that the best determination of QTLs was observed in those traits with high heritability (Supplementary Table 1). QTLs are detected as a result of allelic differences and also in the way these alleles interact with the environment leading to a Genotype × Environment interaction. We observed this in two of our populations: CA and WC. These populations was grown under both CER and GH conditions, for CA just 12 QTLs were found in common between the two conditions and only one in the case of WC. To enable cross-population comparisons to be made for all of the traits studied, the physical coordinate of the marker with the maximum LOD value for each QTL was used to position all of the QTLs detected onto a single Arabidopsis physical map (Figure 2).

Many of the QTLs detected in this study mapped to regions of the genome containing genes already known to be involved in seed lipid biosynthesis. Although these genes represent obvious candidates for controlling the coincident QTLs, many of the QTLs identified had confidence intervals with high numbers of underlying genes, therefore the involvement of unidentified genes cannot be ruled out. Two loci that affect 16:0 mapped to LG 1 at ca 3.2 Mb and ca 22.9 Mb. The first locus was detected in three populations CA, KB and TJ. This QTL was present in CA grown under both GH and CER conditions. The candidate gene underlying this

QTL is the acyl-acyl carrier protein thioesterase, *FATB* (At1g08510 at 2.69 Mb), which is responsible for the export of 16:0-ACP from the plastid (Bonaventure *et al.*, 2003). The second locus was detected in five populations: CA, KB, SG, TJ and WC. Two of these (CA and WC) were grown under both GH and CER conditions. The 3-ketoacyl carrier protein synthase III (KASIII; At1g62640 at 23.1 Mb) is an underlying candidate gene and is responsible for the initial condensation step in fatty acid biosynthesis. QTLs affecting 18:0 were found on LG 2 at 17.5 Mb in KB and at 18.5 Mb in SG. Previous studies (Lightner *et al.*, 1994) have shown that the stearoyl-ACP-desaturase at the locus *FAB2* (At2g43710 at 18.1 Mb) converts 18:0-ACP to 18:1-ACP, and so represents an underlying candidate gene. A QTL affecting 18:1 on LG 4 at ca 16.4 Mb was detected in the three populations: KB, SG and WC (CER material only). A candidate gene in this region is the *fatty acid elongase 1* gene (*FAE1*; At4g34520 at 16.49 Mb). A QTL associated with 18:2, on LG 3 at ca 4.6 Mb, was detected in both the KB and SG populations. The locus *FAD2* (At3g12120 at 3.86 Mb) coincides with the position of this QTL. This is a candidate for control of the QTL as the oleate desaturase it encodes is responsible for the synthesis of 18:2 fatty acids in the endoplasmic reticulum (Okuley *et al.*, 1994). A QTL for 18:3, detected in both the CA CER and CA GH grown material at 13.8 Mb on LG 2, coincides with the position of the locus *FAD3* (At2g29980 at 12.7 Mb), a candidate gene as it encodes a linoleate desaturase. Although 20:0 accounts for only 1–2% of the total seed fatty acids in Arabidopsis seeds, we detected 11 QTLs for this fatty acid. A QTL was identified on LG 2 in both KB and SG at 17.5 Mb and 16.9 Mb, respectively. The underlying candidate gene is the stearoyl-ACP-desaturase *FAB2* (At2g43710 at 18.1 Mb). At 15–20% of the total seed fatty acids, 20:1 is a major component of seed oil. However, only five QTLs were identified for its content. Two of these were in common between KB and SG. They had high LOD values (23.6 and 31.8, respectively) and accounted for 71.1–80.7% of the variation found. Both mapped to LG 4 at 16.2–16.4 Mb, coinciding with the position of the candidate gene *FAE1* (At4g34520 at 16.4 Mb). Three QTLs were detected for 20:2 content. The first was positioned around 13.8 Mb on LG 2 in CA (GH and CER material) and co-located with a QTL detected for 18:3. A candidate gene for control of this QTL is *FAD3*. The second QTL mapped to LG 3 at 4.6 Mb in CA (again in both environments) and also in KB and SG. This QTL co-located with one detected for 18:2 in both KB and SG. A candidate gene for control of this QTL is *FAD2*. The situation for 22:1 is similar to that of the other 20 carbon fatty acids, with a common QTL identified in KB and SG at 16.2–16.4 Mb on LG 4, with *FAE1* as an underlying candidate gene.

A number of QTLs that we detected cannot be assigned easily to candidate genes, as they mapped to the regions of the genome without identifiable candidate genes as described in the previous paragraph. In addition the QTL span, as determined by the 2 LOD interval, was in most cases large and extended over a high number of genes, which made it difficult to speculate on the identity of the genes responsible in a meaningful manner. A strong QTL for 16:0 was detected in SG on LG 1 at 18.2 Mb, though it was not completely resolved from the adjacent QTL at 23.5 Mb. It had a LOD value of

122

10.2 and accounted for 39.5% of the variation found for this trait. Another strong QTL found to affect 16:0 was identified in TJ on LG 5 at 24.1 Mb, had a LOD value of 5.3, and explained 22.8% of the variation. This QTL was also detected for 18:1 and DS. A QTL (LOD 5.1) found to affect 18:0 content was detected on SG LG 3 at 16.9 Mb and accounted for 22.2% of the variation. A number of QTLs identified for 18:1 cannot be easily assigned to known lipid metabolism genes. The first of these was detected in WC, under both the GH and CER conditions, on LG 1 at ca 5.2–6.4 Mb. A second QTL detected for 18:1 content in both TJ and WC (CER only) was mapped on LG 5 and had a peak maximum at 20.1 Mb. A third QTL was also identified in TJ and WC on LG 5, with peak maxima at 24.1 or 25.2 Mb in the respective populations. Two QTLs were detected for 18:2 that do not coincide with candidate genes. The first was a weak QTL detected in both KB and the WC GH grown material, positioned at the end of LG 2 at 19.6 Mb. This QTL co-located with a QTL for oil content identified in both the CA and KB populations. The second QTL for 18:2 was found on TJ LG 3 at 7.7 Mb. For this QTL, the two LOD confidence interval was 10 to 37 cM and narrowly excluded the position of a candidate gene, FAD2. Two QTLs were detected for 18:3 that do not coincide with candidate genes. The first was detected on LG 5 in both KB and WC (CER only), with maximum LOD at ca 20 Mb. The second QTL for 18:3 was relatively strong (LOD 6.9) and unique to TJ. It explained 28.7% of the variation and was positioned on LG 1 at 10.9 Mb. Of the QTLs detected for 20:0 that do not coincide with candidate genes, those with the greatest LOD values mapped in NG and SG, to LG 1, with peak maxima at 15.5 and 18.2 Mb, respectively. Another QTL for 20:0 was identified in NG and KB on LG 5 at 3.8 or 5.3 Mb. A strong QTL for 20:2 with a LOD value of 4.3 was found on WC (CER only) LG 1 at 5.2 Mb.

Three QTLs that affect seed oil content were identified in at least two of the populations studied. They do not map to genomic regions known to contain genes involved in the control of oil content. The first mapped in CA (CER only) and KB, to LG 2 at ca. 19.5 Mb. QTLs for 18:2, DS, 18:2/18:1 and 18:3/18:2 co-located in KB. The KB oil QTL had a LOD value of 6.5 and accounted for 28.2% of the variation. A QTL for oil content was also detected in a similar position in the Cvi-0 × Ler RI population (Hobbs et al., 2004). A preliminary study of an F$_2$ population derived from a cross between Kondara and Br-0, revealed an association between seed oil content and a marker towards the distal end of LG 2 (O'Neill et al., 2003). A further two QTLs that influence oil content were identified in more than one population, and both mapped to LG 5. The first was detected in KB and WC (GH only) at ca. 22–24 Mb. The second was detected in CA (GH only) and WC (GH only) and was a little further along LG 5 at 26.8 Mb. QTLs for 16:0, 18:0, 18:1 and 18:3 were detected in similar genomic regions in a number of the populations. The major flowering time genes LFY, MAF2, and MAF5 map to this region of the genome. Previous work using these populations mapped flowering-time QTLs (O'Neill et al., 2008) and found that NG, SG, TJ and WC all have QTLs in the region 20.1–26.8 Mb on LG 5. Chen et al. (2010) found that flowering time and oil content are negatively correlated in oilseed rape. A further explanation for the underlying

basis of at least the upper QTL could be the steroleosin gene HSD1 (At5g50600), which is found at 20.6 Mb on LG 5. Two QTLs for oil content that were unique to TJ were detected. The first mapped to LG 1 at 27 Mb, had a LOD value of 3.2 and accounted for 14.7% of the variation. Genes in this region of the genome with products involved in lipid biosynthesis include KASII (At1g74960 at 28.1 Mb) and GLABRA 2 (At1g79840 at 30 Mb). A second QTL, unique to TJ, for oil content mapped to LG 2 at 11.6 Mb, had a LOD value of 3.7 and accounted for 16.7% of the variation. The diacylglycerol acyltransferase gene (At2g19450) that catalyses the final step of the triacylglycerol synthesis pathway is located close to this, at 8.4 Mb (Zou et al., 1999).

In this work we have detected QTLs for seed oil content and composition in six RI populations which were generated from a diverse set of parental accessions. A proportion of the QTLs identified may be controlled by variation at the loci encoding known lipid biosynthesis genes, whereas many are likely to be controlled by genes not previously associated with seed lipid biosynthesis. Future investigation of the novel QTLs identified in this work will help extend our knowledge of the regulatory mechanisms involved in seed lipid biosynthesis.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

References

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

Beisson F, Koo AJK, Ruuska S, Schwender J, Pollard M, Thelen JJ et al. (2003). Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a Web-based database. *Plant Physiol* **132**: 681–697.

Bentsink L, Hanson J, Hanhart CJ, Blankestijn-de Vries H, Coltrane C, Keizer P et al. (2010). Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways. *Proc Natl Acad Sci USA* **107**: 4264–4269.

Bonaventure G, Salas JJ, Pollard MR, Ohlrogge JB (2003). Disruption of the FATB gene in Arabidopsis demonstrates an essential role of saturated fatty acids in plant growth. *Plant Cell* **15**: 1020–1033.

Chen G, Geng J, Rahman M, Liu X, Tu J, Fu T et al. (2010). Identification of QTL for oil content, seed yield and flowering time in oilseed rape (*Brassica napus*). *Euphytica* **175**: 161–174.

Focks N, Benning C (1998). wrinkled1: a novel, low-seed-oil mutant of Arabidopsis with a deficiency in the seed-specific regulation of carbohydrate metabolism. *Plant Physiol* **118**: 91–101.

Glazier AM, Nadeau JH, Aitman TJ (2002). Finding genes that underlie complex traits. *Science* **298**: 2345–2349.

Hobbs DH, Flintham JE, Hills MJ (2004). Genetic control of storage oil synthesis in seeds of Arabidopsis. *Plant Physiol* **136**: 3341–3349.

International Standardization Organization (1993). *Oilseeds. Simultaneous Determination of Oil and Moisture Contents. Method Using Pulsed Nuclear Magnetic Resonance Spectrometry.*

International Standard ISO 10565:1993(E). International Standardization Organization: Geneva, Switzerland.

James DW, Dooner HK (1990). Isolation of EMS-induced mutants in Arabidopsis altered in seed fatty acid composition. *Theor Appl Genet* **80**: 241–245.

James DW, Lim E, Keller J (1995). Directed tagging of the Arabidopsis Fatty-Acid Elongation-1 (FAE1) gene with the maize transposon Activator. *Plant Cell* **7**: 309–319.

Katavic V, Reed DW, Taylor DC, Giblin EM, Barton DL, Zou J et al. (1995). Alteration of seed fatty acid composition by an ethyl methanesulfonate-induced mutation in *Arabidopsis thaliana* affecting diacylglycerol acyltransferase activity. *Plant Physiol* **108**: 399–409.

Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004). Naturally occurring genetic variation in *Arabidopsia thaliana*. *Annu Rev Plant Biol* **55**: 141–172.

Lightner J, Wu J, Browse J (1994). A mutant of Arabidopsis with increased levels of stearic acid. *Plant Physiol* **106**: 1443–1451.

Linder C (2000). Adaptive evolution of seed oils in plants: accounting for the biogeographic distribution of saturated and unsaturated fatty acids in seed oils. *Am Nat* **156**: 442–458.

Nemri A, Atwell S, Tarone AM, Huang YS, Studholme DJ, Nordborg M et al. (2010). Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc Natl Acad Sci USA* **107**: 10302–10307.

Ohlrogge J, Browse J (1995). Lipid biosynthesis. *Plant Cell* **7**: 957–970.

Okuley J, Lightner J, Feldman K, Yadav N, Lark E, Browse J (1994). Arabidopsis FAD2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *Plant Cell* **6**: 147–158.

O'Neill CM, Gill S, Hobbs D, Morgan C, Bancroft I (2003). Natural variation for seed oil composition in *Arabidopsis thaliana*. *Phytochemistry* **64**: 1077–1090.

O'Neill CM, Morgan C, Kirby J, Tschoep H, Deng PX, Brennan M et al. (2008). Six new Recombinant Inbred populations for the study of quantitative traits in *Arabidopsis thaliana*. *Theor Appl Genet* **116**: 623–634.

Payne RW, Murray DA, Harding SA, Baird DB, Soutar DM (2009). *GenStat for Windows (12th Edition) Introduction*. VSN International: Hemel Hempstead, UK.

Poirier Y, Ventre G, Caldelari D (1999). Increased flow of fatty acids toward β-oxidation in developing seeds of Arabidopsis deficient in diacylglycerol acyltransferase activity or synthe-

sizing medium-chain-length fatty acids. *Plant Physiol* **121**: 1359–1366.

Rees M, Long MJ (1992). Germination biology and the ecology of annual plants. *Am Nat* **139**: 484–508.

Ruuska SA, Girke T, Benning C, Ohlrogge JB (2002). Contrapuntal networks of gene expression during Arabidopsis seed filling. *Plant Cell* **14**: 1191–1206.

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M et al. (2005). Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*. *Plant Physiol* **138**: 1163–1173.

Symonds VV, Godoy AV, Alconada T, Botto JF, Juenger TE, Casal JJ et al. (2005). Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic variation for trichome density. *Genetics* **169**: 1649–1658.

Tonsor SJ, Alonso-Blanco C, Koornneef M (2005). Gene function beyond the single trait: natural variation, gene effects and evolutionary ecology in *Arabidopsis thaliana*. *Plant, Cell Environ* **28**: 2–20.

Van Ooijen JW (2004). *MapQTL Version 5, Software for the Mapping of Quantitative Trait Loci in Experimental Populations*. Kyazma B.V.: Wageningen, Netherlands.

van Ooijen JW, Boer MP, Jansen RC, Maliepaard C (2002). *MapQTL 4.0: Software for the Calculation of QTL Positions*. Plant Research International: Wageningen, The Netherlands.

van Ooijen JW, Voorrips RE (2001). *JoinMap Version 3.0: Software for the Calculation of Genetic Linkage Maps*. Plant Research International: Wageningen, The Netherlands.

Warthman N, Fitz J, Weigel D (2007). MSQT for choosing SNP assays from multiple DNA alignments. *Bioinformatics* **23**: 2784–2787.

White JA, Benning C (2001). Genomic approaches towards the engineering of oil seeds. *Plant Physiol Biochem* **39**: 263–270.

Zou JT, Wei YD, Jako C, Kumar A, Selvaraj G, Taylor DC (1999). The *Arabidopsis thaliana* TAG1 mutant has a mutation in a diacylglycerol acyltransferase gene. *Plant J* **19**: 645–653.