Supplementary materials for

Regulatory divergence of flowering time genes in the allopolyploid *Brassica napus*

D. Marc Jones[1,2], Rachel Wells[1], Nick Pullen[1], Martin Trick[2], Judith A. Irwin[1*], Richard J. Morris[1,2†]

[1] *Crop Genetics, John Innes Centre, Norwich Research Park, Colney Lane, Norwich. NR4 7UH. United Kingdom.*

[2] *Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney Lane, Norwich. NR4 7UH. United Kingdom.*

*Corresponding author: judith.irwin@jic.ac.uk

†Corresponding author: richard.morris@jic.ac.uk

Supplementary results

A self-organising map based approach corroborates the finding that *Brassica napus* copies of *Arabidopsis thaliana* flowering time genes have diverged in their regulation.

A self-organising map (SOM) based approach was employed to detect regulatory divergence between *B. napus* copies of flowering time genes. The advantage of this approach, over the

WGCNA approach discussed in the main text, is that the regulatory module assignments are not binary, allowing for more subtle patterns to be detected. A SOM is a construct that groups together expression traces into clusters. The sampling procedure (Supplementary Figure 8a) returns an empirical probability of two expression traces mapping to the same SOM cluster. In addition, clustering probabilities can be calculated for a single gene which represent the uncertainty in the expression measurements quantified for that gene. In this case the clustering probability calculated is referred to as a self-clustering probability. Visualising the clustering probabilities determined by the SOM based method is complicated by the bimodal distribution the probabilities follow. Supplementary Figure 10 reveals a peak in self-clustering probabilities at 0.05 but also at ~1.0. This bimodal structure is a result of some genes only being expressed at a single time point. When these genes are resampled, their normalised expression trace remains the same, leading to a high self-clustering probability. To visualise probabilities from across this distribution, a soft threshold is applied to the probabilities. After the threshold is applied, the higher the clustering coefficient, the more similar two expression traces will tend to be. Genes are assigned to regulatory modules using heatmaps of clustering coefficients. The different patterns of regulatory module assignment are described in the main text.

This method was applied to *B. napus* flowering time genes. The occurrences of the different regulatory module assignment patterns were counted for both apex (Supplementary Figure 8b) and leaf (Supplementary Figure 8c) expression data. The null hypothesis used in the WGCNA analysis was that copies of genes would not show expression divergence (dashed lines in Figure 6, main text). The *redundant* pattern in the SOM analysis is equivalent to this null hypothesis (Figure 7a, main text). Like the results from the WGCNA analysis, this null hypothesis is not true for any flowering time genes with five or more copies in the *B. napus* leaf (Supplementary Figure 8c) or six or more copies in the apex (Supplementary Figure 8b). As with the *redundant* pattern, the *unique* pattern of regulatory module assignment becomes less frequent as the number of *B. napus* copies of a gene increases (Supplementary Figure 8b and 8c). This agrees with the WGNCA analysis, where the number of genes lying on the solid line in Figure 6 in the main text (equivalent to the *unique* pattern in the SOM analysis) decreases at higher numbers of copies.

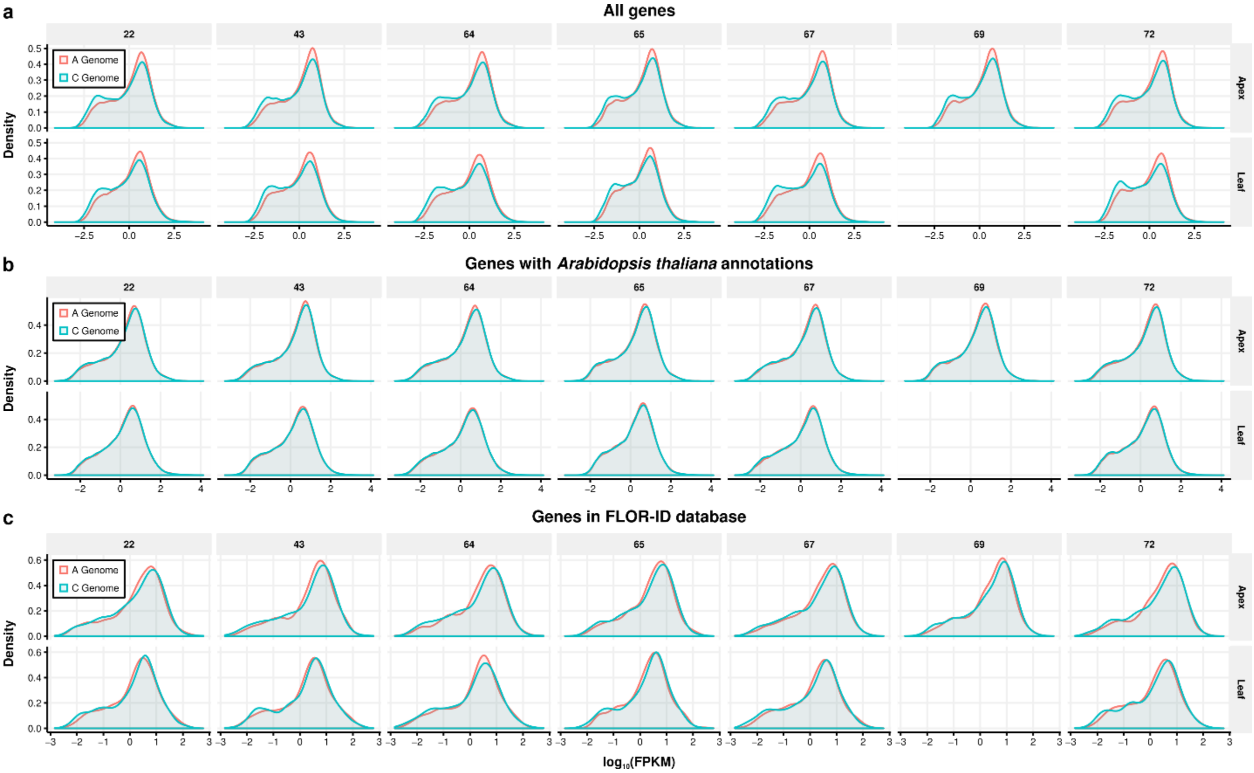
WGCNA cannot detect *gradated* and *mixed* patterns patterns of regulatory module assignment. In the apex and leaf, *mixed* and *gradated* patterns are seen at a lower frequency than *distinct* patterns, revealing that genes exhibiting intermediary regulatory behaviour relative to the other copies of that gene are observed less frequently than genes occupying distinct regulatory modules. Gene copies with intermediate regulatory behaviour may indicate that somecopies are more susceptible to regulatory cross-talk than others. The low number of *gradated* patterns observed when three genes copies are present in both tissues suggests that these genes tend to have expression traces that are detectably different to one another. *Distinct* patterns are more prevalent than *unique* patterns at three gene copies; the majority contain one copy with an expression trace divergent to the expression traces of the other two copies.

We could integrate homoeologue information for the three copy genes exhibiting a *distinct* pattern of regulatory module assignment to ask whether genes tended to be within the same regulatory modules as their homoeologue. In the apex, this is the case, with 59% of genes located in the same module. More generally, we find that of the genes in the apex (leaf) where homoeologue information is available, 69% (64%) of genes are assigned to the same module as the homoeologue, 18% (19%) of genes are assigned to a different module and 12% (16%) of genes have homoeologues which cannot be clustered. Homoeologues that cannot be clustered arise when the clustering coefficient calculated using the self-clustering probability of a gene is below 0.5, or the homoeologue is not expressed in that tissue.

We then asked whether the relatively large number of *distinct* patterns at four gene copies was due to homoeologous copies of genes displaying similar expression traces. For the genes for which homoeologue information was available, we find the majority (76% in apex, 72% in leaf) of genes are in the same regulatory module as their homoeologue.

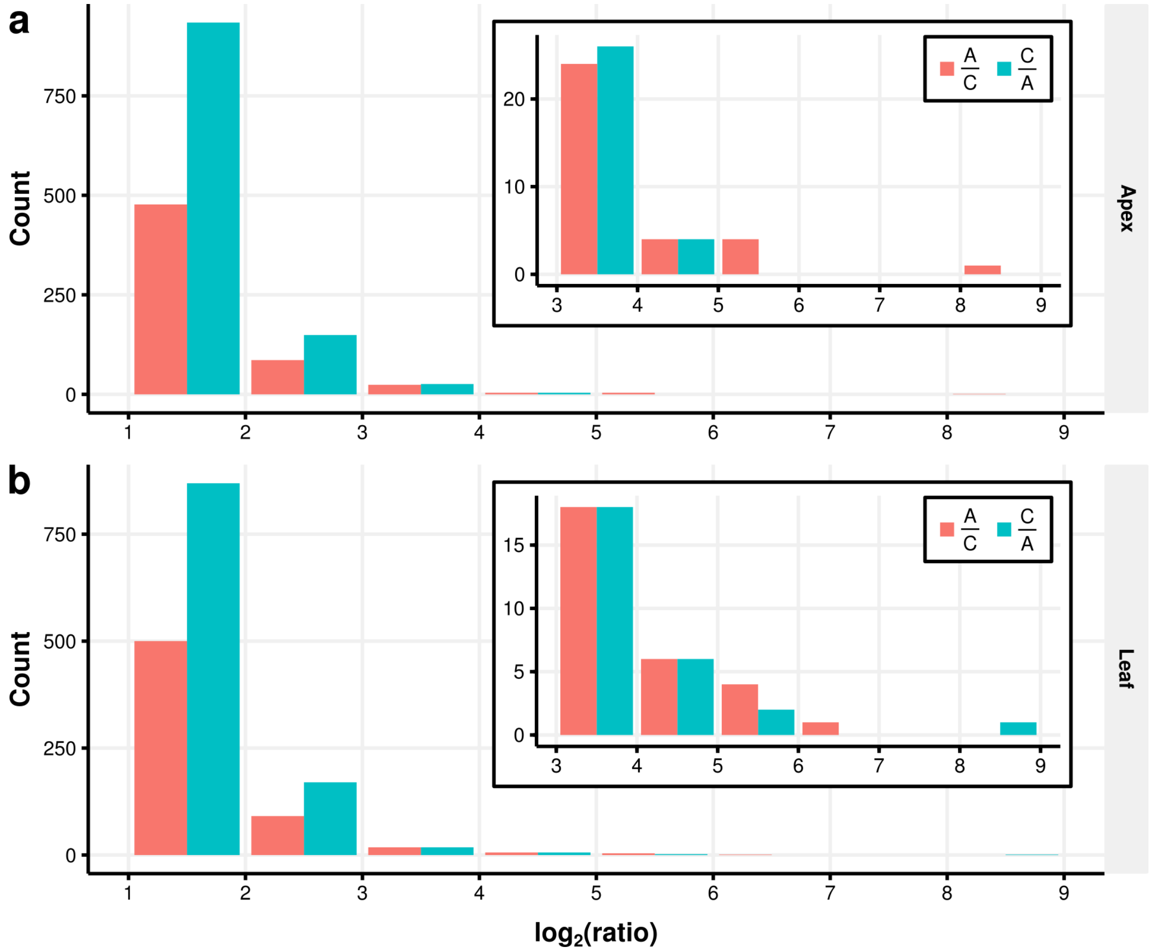
The SOM analysis corroborates many of the key findings of the WGCNA analysis in a manner which takes into account the uncertainty in our data. Namely, that expression divergence between copies is widespread and that as the number of copies of a gene in the genome increases, the likelihood of observing regulatory divergence between those copies increases. Additionally, the SOM analysis reveals that some copies of flowering time genes exhibit a *gradated* pattern of regulatory module assignment, representing subtle differences in regulation. This may be the result of regulatory cross-talk between the copies, or represents subtle functional differences that have consequences for the control of flowering time in *Brassica napus*.

Supplementary figures



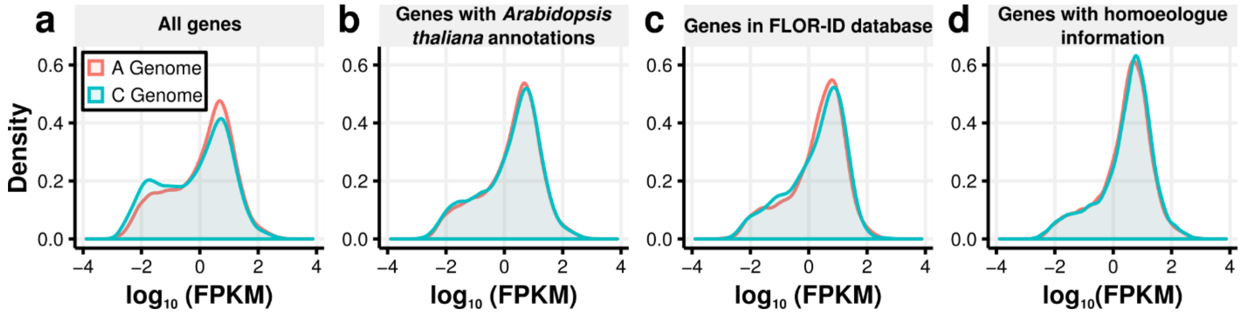
Supplementary figure 1 – Expression differences between A and C genomes are consistent across different tissues and time points.

Density plots of transformed expression levels ($\log_{10}$(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *Brassica napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated *Arabidopsis thaliana* gene, and **c** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene that is present in the FLOR-ID database[1].



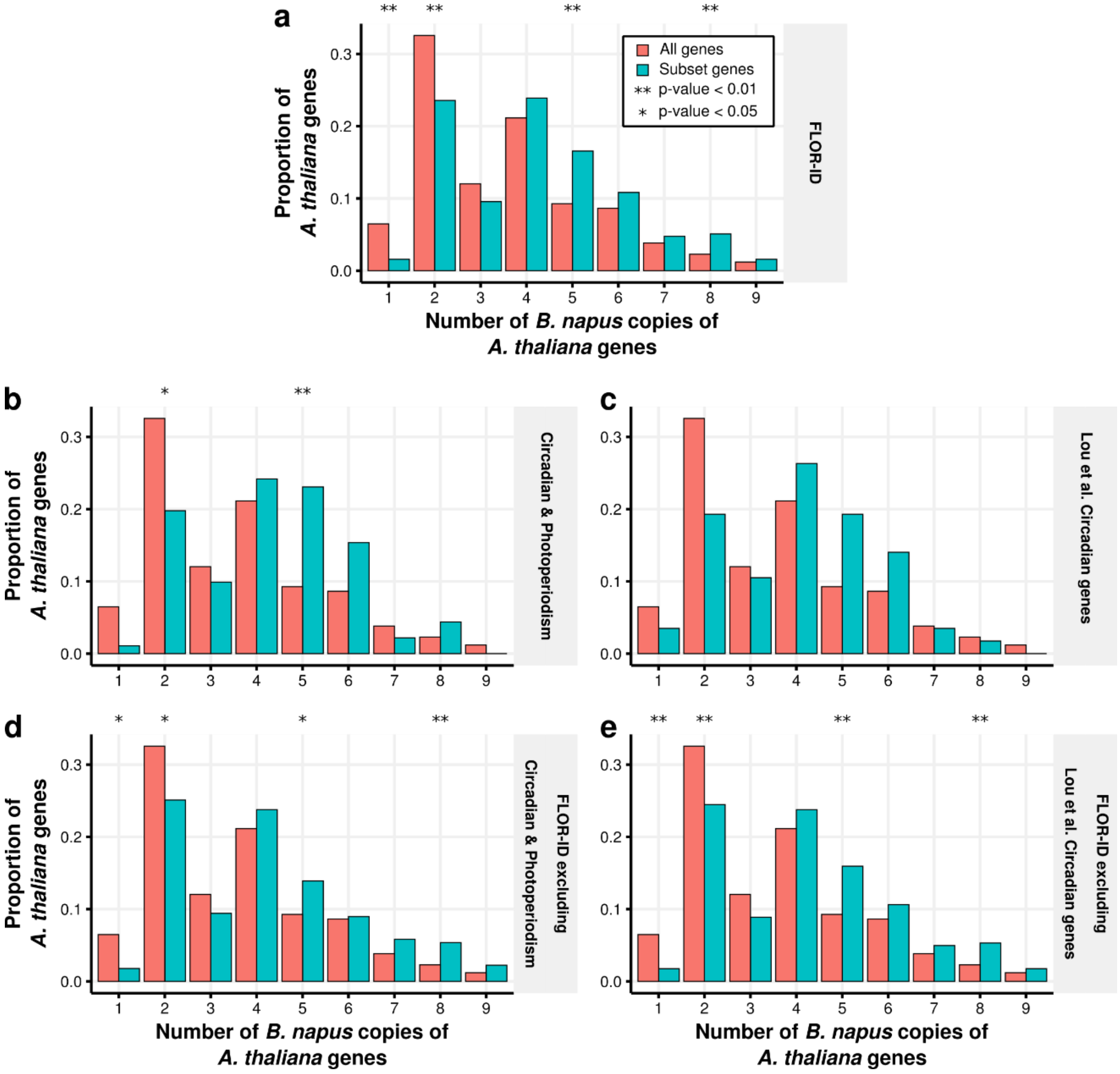
Supplementary figure 2 – Distributions of the fold expression differences between homoeologue pairs exhibiting biased expression

Homoeologue pairs are defined as exhibiting biased expression towards a particular genome if the gene on that genome has an FPKM level at least 2-fold higher than its homoeologue. The fold differences in FPKM level between homoeologues were calculated and $\log_2$ transformed. The values were binned and the number of pairs in each bin are plotted. If the homoeologue pairs exhibit biased expression towards the A genome, then the fold ratio was calculated with the A genome homoeologue FPKM value as the numerator (red bars). Likewise, if the pairs exhibit biased expression towards the C genome then the fold ratio was calculated with the C genome homoeologue FPKM value as the numerator (blue bars). The FPKM values from the day 22 time point were used. The inset of each graph corresponds to the counts above a $\log_2$(ratio) value of 3 plotted on a different count scale.



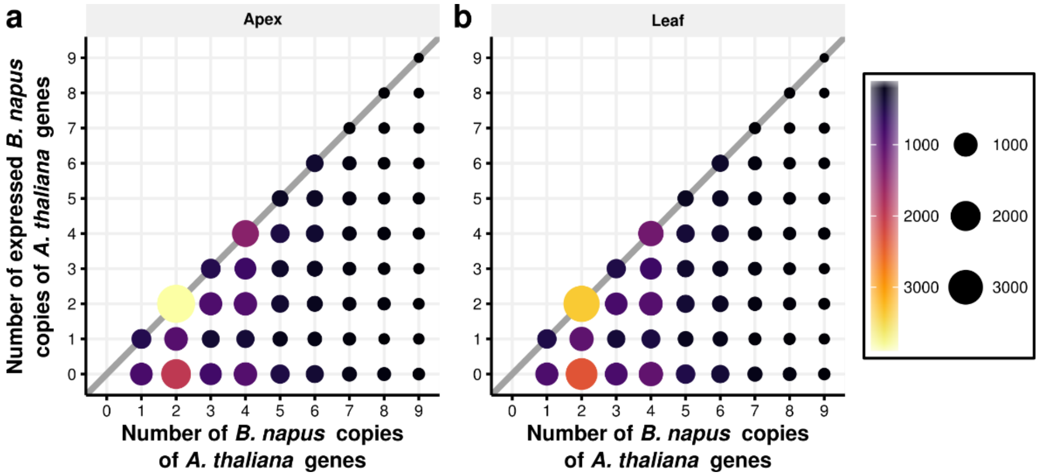
Supplementary figure 3 – Genes for which homoeologue information is available have fewer genes within the very low region of expression

Density plots of transformed expression levels ($\log_{10}$(FPKM)) calculated using different subsets of genes. The data used to generate the density plots consisted of expression data from: **a** all annotated *B. napus* genes, **b** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene, **c** *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene that is present in the FLOR-ID database[1], and **d** *B. napus* genes for which homoeologue information is available. These plots are generated using apex expression data from the time point taken at day 22.



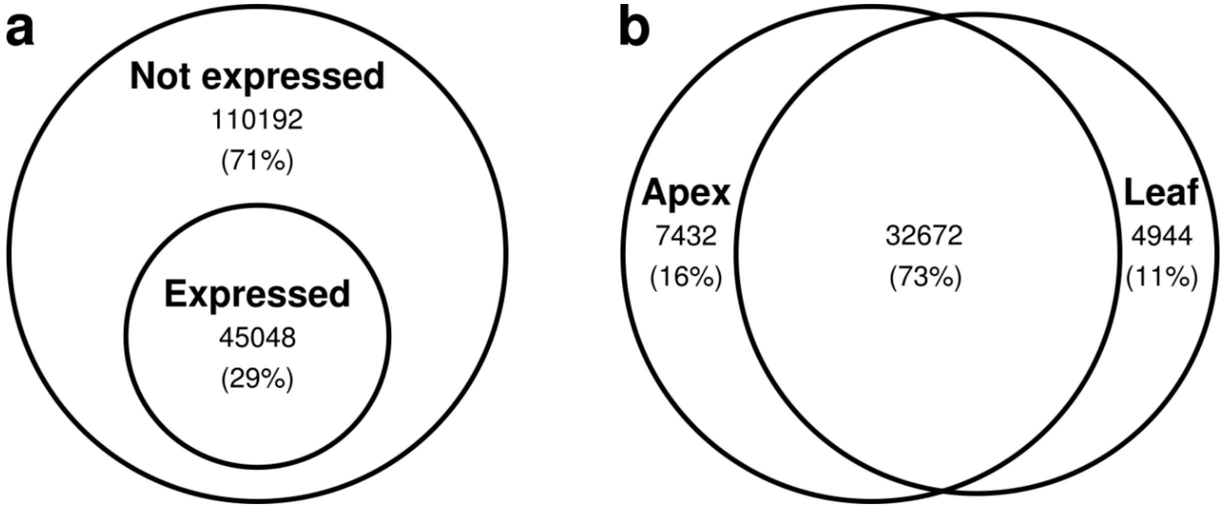
Supplementary figure 4 – The observed retention of flowering time genes is not explained by genes associated with the circadian rhythm alone

The proportions of Arabidopsis genes that have particular numbers of homologues identified in OSR, comparing all genes to a number of different gene subsets. False discovery corrected *p-values* are computed in the same way as Figure 1 in the main text. The gene subsets compared to all genes in each of the plots are as follows: **a** All FLOR-ID genes[1]. **b** FLOR-ID genes annotated as involved with the "Circadian" or "Photoperiodism" pathways. **c** The list of circadian genes used by Lou et al. (2012) to demonstrate gene retention in *B. rapa*[2]. **d** FLOR-ID genes with genes annotated as involved with the "Circadian" or "Photoperiodism" pathways removed. **e** FLOR-ID genes with genes used in the study by Lou et al. (2012) removed[2].



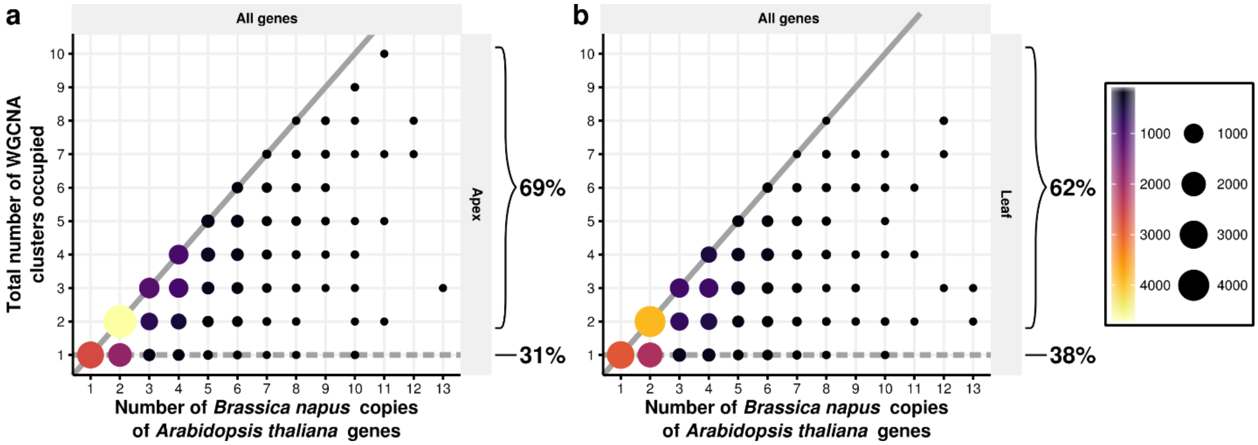
Supplementary figure 5 – Not all annotated *B. napus* copies of *A. thaliana* genes are expressed.

a and **b** depict the relationships when expression data from the apex and leaf are used respectively. The size and colour of the circles indicates the number of data points at that position in the graph. The thick diagonal line indicates *A. thaliana* genes that have *B. napus* orthologues that are all expressed during the developmental transcriptome. All *B. napus* genes that show sequence conservation to an annotated *A. thaliana* gene were used to generate these results.



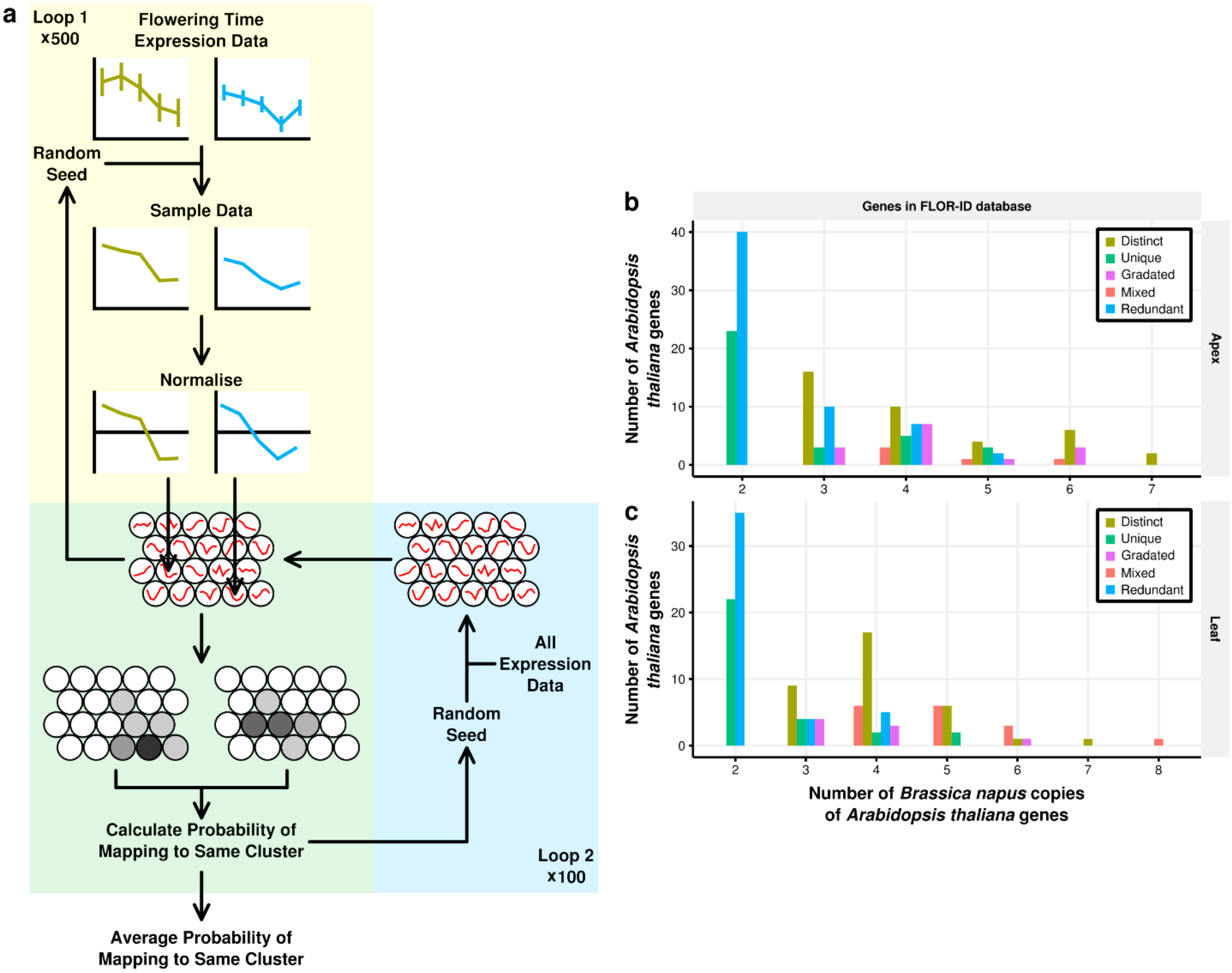
Supplementary figure 6 – Euler and Venn diagrams showing the percentage of expressed genes and the percentage of genes expressed in the apex and leaf samples

Brassica napus genes were classified as expressed if the expression of the genes exceeded 2.0 FPKM at at least one time point during the developmental time series. **a** Genes expressed in at least one tissue of the *Brassica napus* genes compared to the number of annotated genes in the Darmor-*bzh* reference genome. **b** The number of genes expressed specifically in the apex and the leaf and the number of genes that are expressed in both tissues.



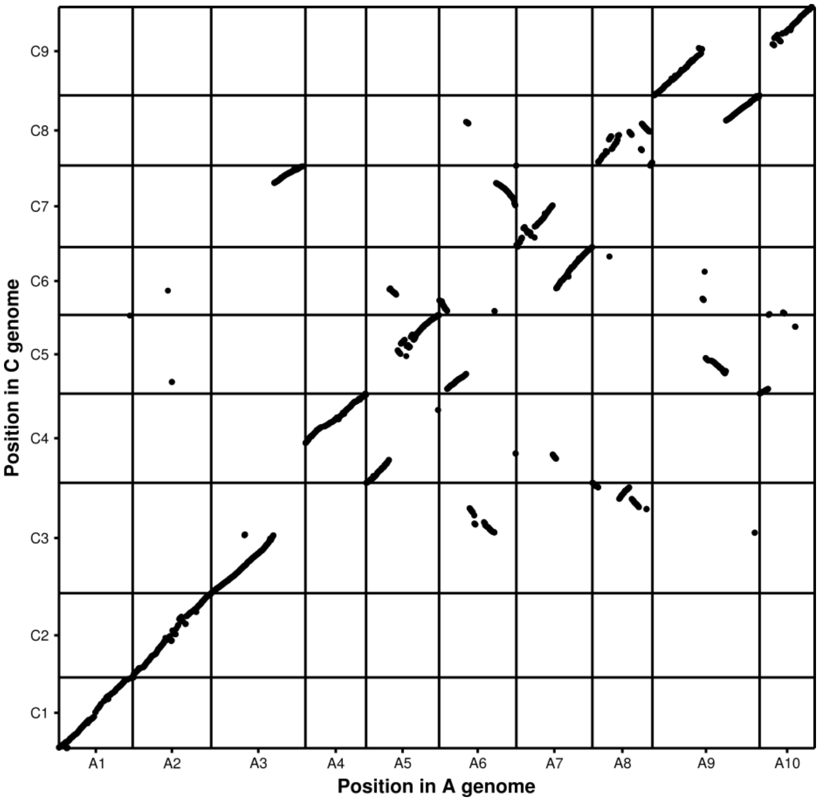
Supplementary figure 7 – Many gene copies are assigned to different regulatory modules in *B. napus.*

B. napus genes were included in this analysis when they i) Have expression above 2.0 FPKM in at least one time point in the developmental time series, and ii) Show sequence conservation to an annotated *A. thaliana* gene. **a** and **b** depict the relationships when expression data from the apex and leaf are used respectively. The size and colour of the circles indicates the number of data points at that position in the graph. The thick lines on each graph represent two potential extremes. The dashed line represents the null hypothesis that all *B. napus* copies of an *A. thaliana* gene are assigned to the same WGCNA cluster. The solid line represents the *A. thaliana* genes that have *B. napus* copies that are each assigned to separate WGCNA clusters. The percentages indicated on the graph indicate the percentage of data points which agree and the percentage which do not agree with the null hypothesis. All *B. napus* genes showing sequence conservation to an annotated *A. thaliana* gene were used to generate these results.



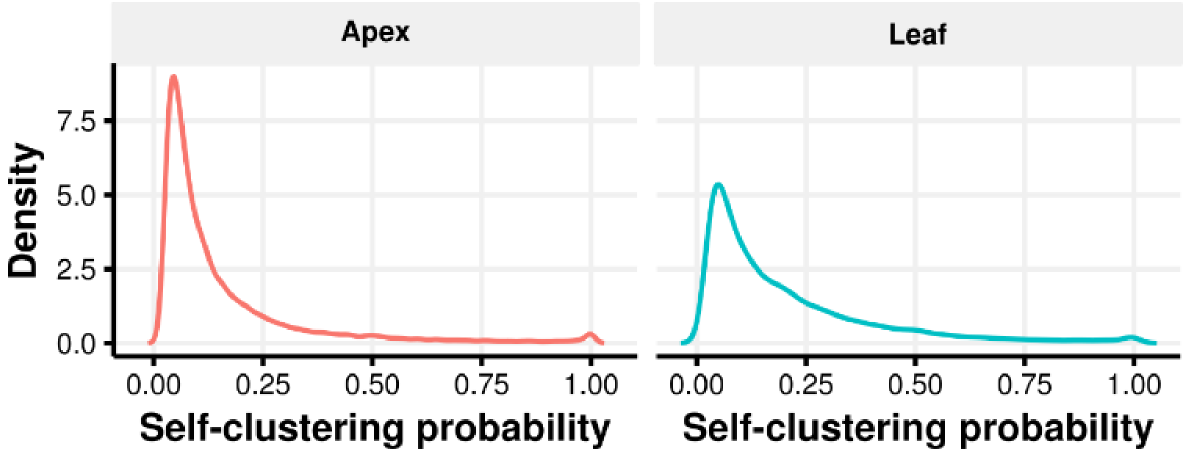
Supplementary figure 8 – Self-organising map (SOM) based assessment of expression trace divergence uncovers widespread regulatory differences and subtle patterns of divergence.

a Schematic of the SOM based clustering approach. The approach consists of two overlapping sampling loops. In loop 1, expression data from flowering time gene copies is sampled assuming a Gaussian error model. Sampled expression traces are zero mean and unit variance normalised and mapped to the SOM. This procedure is repeated 500 times to give two density plots of where in the SOM the copies map. These density plots are used to calculate the probability of the copies mapping to the same SOM cluster. As SOM clustering has a random component, loop 2 consists of regenerating the SOM using all expression data and calculating the probability of copies clustering to the same cluster 100 times. Using this, an average probability of mapping to the same cluster is calculated. **b & c** The relationships between the number of expressed *B. napus* copies of *A. thaliana* genes and the number of different types of regulatory module assignment patterns exhibited by those gene copies. This relationship is calculated using expression data from the apex (**b**) and the leaf (**c**). The different regulatory patterns are illustrated and explained in Figure 7 of the main text.



Supplementary figure 9 – Locations of identified homoeologues pairs in the *B. napus* genome

Homoeologue pairs were identified as detailed in the main text (Methods). The locations of these pairs give a representation of the chromosomal rearrangements that have occurred between the A and C genomes.



Supplementary figure 10 – A bimodal distribution of self-clustering probabilities necessitates the use of a threshold to visualise the probabilities

Self-clustering probabilities are calculated as detailed in the main text (Methods). The density curves presented here represent the self-clustering probabilities calculated from a single SOM. The clustering coefficient threshold was taken by determining the self-clustering probability that corresponded to the peak of the density curve. This threshold was calculated for each SOM and averaged to give the final threshold: apex threshold = 0.053; leaf threshold = 0.056.

Supplementary tables

Date sampled	Days post sowing	Days vernalised	Days post vernalisation	Tissue Type	
				Leaf	Apex
2014-05-29	22	0	-	2	2
2014-06-19	43	21	-	2	2
2014-07-10	64	42	-	2	2
2014-07-11	65	42	1	1	1
2014-07-13	67	42	3	2	2
2014-07-15	69	42	5	0	1
2014-07-18	72	42	8	2	2

Supplementary table 1 – Sampling and sequencing scheme for the developmental time series

The numbers in the rightmost two columns indicate the number of biological pools sampled for that time point within each tissue.

Days post sowing	Apex			Leaf		
	Both expressed	A genome 2-fold higher	C genome 2-fold higher	Both expressed	A genome 2-fold higher	C genome 2-fold higher
22	136	11 (8.1%)	19 (14.0%)	109	8 (7.3%)	14 (12.8%)
43	149	15 (10.1%)	24 (16.1%)	118	12 (10.2%)	16 (13.6%)
64	147	12 (8.2%)	20 (13.6%)	114	11 (9.6%)	13 (11.4%)
65	145	13 (9.0%)	25 (17.2%)	108	10 (9.3%)	16 (14.8%)
67	138	14 (10.1%)	19 (13.8%)	112	7 (6.3%)	12 (10.7%)
69	139	11 (7.9%)	18 (12.9%)	-	-	-
72	142	15 (10.6%)	21 (14.8%)	112	5 (4.5%)	14 (12.5%)

Supplementary table 2 – Number of genes expressed 2-fold higher than their homoeologue for all flowering time gene homoeologue pairs.

As for Table 1 in the main text, calculated using homoeologue pairs which showed sequence similarity to *A. thaliana* flowering time genes from the FLOR-ID database[1]. The geometric mean of the fold difference of the C genome gene relative to the A genome homoeologue for all flowering time homoeologue pairs is 1.10 in the apex and 1.04 the leaf.

Gene	Forward Primer (5' – 3')	Reverse Primer (5' – 3')	Amplicon Length
TFL1 A10	GTCTCCAATGGCCATGAGT	GTGCCGGGGATGTTCATG	179
TFL1 Cnn	GTCATGAACATCCCCGGC	GATCATTCTCGATCGCAAATTCA	196
TFL1 C2	CTGATGTTCCAGGTCCTAGC	TGGGGAGATATCGATAACATGTC	197
TFL1 C3	GAGGTGGTGAGCTATGAGTTG	CTGGGCGTTAAAGAAGACAGCA	189
GAPDH	AGAGCCGCTTCCTTCAACATCATT	TGGGAACACGGAAGGACATTCC	112

Supplementary table 3 – qPCR primer sequences

		Sequencing Run 1				Sequencing Run 2			
Tissue	Days post sowing	Total reads (millions)	Mapped reads (millions / percentage of total)	Multiply mapping reads (millions / percentage of mapped)	Reads mapped to over 20 positions (ten thousand / percentage of mapped)	Total reads (millions)	Mapped reads (millions / percentage of total)	Multiply mapping reads (millions / percentage of mapped)	Reads mapped to over 20 positions (ten thousand / percentage of mapped)
Apex	22	75.6	61.8 (81.8%)	8.3 (13.4%)	20.7 (0.3%)	41.9	34.3 (81.9%)	4.7 (13.8%)	7.8 (0.2%)
Apex	43	71.5	56.8 (79.4%)	7.4 (13.1%)	17.8 (0.3%)	31.7	25.3 (79.8%)	3.4 (13.6%)	5.3 (0.2%)
Apex	64	70.5	57.4 (81.4%)	7.5 (13.0%)	21.6 (0.4%)	28.7	23.3 (81.2%)	3.2 (13.8%)	149.4 (6.4%)
Apex	65	67.6	54.6 (80.7%)	7.2 (13.2%)	26.5 (0.5%)	NA	NA	NA	NA
Apex	67	78.6	63.5 (80.8%)	8.4 (13.2%)	36.3 (0.6%)	30.5	25.1 (82.3%)	3.5 (13.9%)	5.6 (0.2%)
Apex	69	66.2	54.4 (82.2%)	7.3 (13.5%)	30.7 (0.6%)	NA	NA	NA	NA
Apex	72	59.7	48.6 (81.4%)	6.4 (13.2%)	35.2 (0.7%)	31.5	25.8 (81.8%)	3.6 (14.1%)	4.5 (0.2%)
Leaf	22	68.2	54.7 (80.2%)	8.4 (15.4%)	9.5 (0.2%)	33.9	28.0 (82.5%)	4.4 (15.7%)	3.7 (0.1%)
Leaf	43	50.5	41.5 (82.1%)	6.2 (15.0%)	11.1 (0.3%)	33	26.4 (80.1%)	4.0 (15.1%)	4.6 (0.2%)
Leaf	64	73.9	60.7 (82.1%)	8.8 (14.4%)	10.2 (0.2%)	35.5	29.1 (82.1%)	4.3 (14.8%)	3.7 (0.1%)
Leaf	65	45.7	37.6 (82.2%)	5.5 (14.6%)	5.4 (0.1%)	NA	NA	NA	NA
Leaf	67	81.8	67.1 (82.1%)	10.0 (14.9%)	9.4 (0.1%)	35.7	28.8 (80.7%)	4.4 (15.4%)	3.5 (0.1%)
Leaf	72	49	40.3 (82.1%)	5.8 (14.5%)	5.8 (0.1%)	32.2	26.2 (81.2%)	3.9 (15.1%)	3.9 (0.1%)

Supplementary table 4 – Sequencing statistics for the two sequencing runs carried out to generate the developmental transcriptome

Reads were mapped to the Darmor-*bzh* reference genome[3] using TopHat[4] as described in the main text (Methods). The percentage of mapped reads is given as the percentage of the total reads. Multiply mapped reads are defined as reads that mapped to multiple places in the genome with an equal probability. The percentages of multiply mapped reads and the percentage of reads mapping to more than 20 position in the genome are calculated as a total of the reads that were mapped to the genome, and not a percentage of the total reads.

References

1. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in Arabidopsis thaliana. *Nucleic Acids Res.* **44,** D1167–D1171 (2016).
2. Lou, P. *et al.* Preferential Retention of Circadian Clock Genes during Diploidization following Whole Genome Triplication in Brassica rapa. *Plant Cell* **24,** 2415–2426 (2012).
3. Chalhoub, B. *et al.* Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* **345,** 950–953 (2014).
4. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).