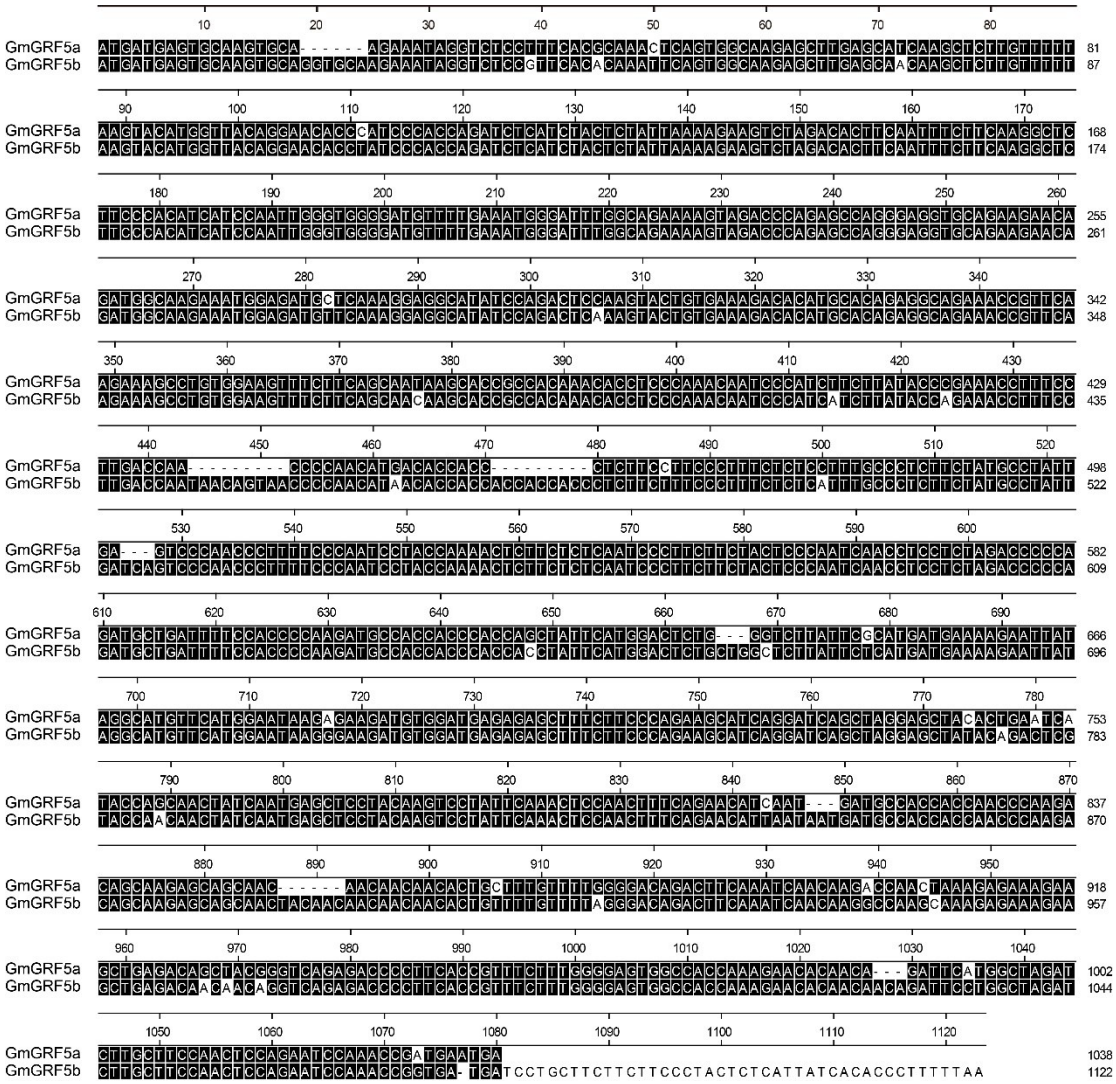
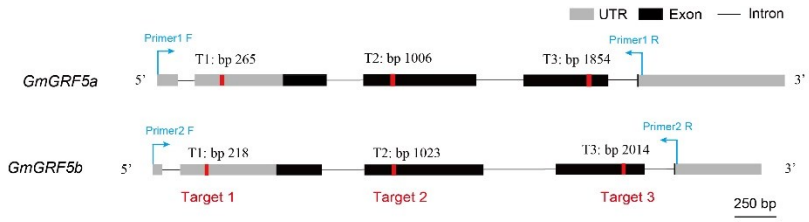


Fig. S1 Alignment of *GmGRF5a* and *GmGRF5b* cDNA sequence. Two GRF5 homologous *GmGRF5a* and *GmGRF5b* share highly similar cDNA sequences.

A



B

Name		Target 1	Target 2	Target 3	The bp number of deletion (-) or insertion (+)
WS82	GmGRF5a	AGCACATTAGAGAAATAGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	
	GmGRF5b	AGCACATTAGAGAAATAGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	
Gmgrf5a^{D}	GmGRF5a	AGCACATTAGAGA---AGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGT------GGGAGTGGCCAC	- 3 bp, -7 bp
	GmGRF5b	AGCACATTAGAGA---AGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	- 3 bp
Gmgrf5a-1	GmGRF5a	AGCACATTAGAGAAATGACA---	----------	----------GGGGAGTGGCCAC	+4 bp, -1581 bp
	GmGRF5b	AGCACATTAGAGAAATAGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	
Gmgrf5b-1	GmGRF5a	AGCACATTAGAGA---AGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	- 3 bp
	GmGRF5b	AGCACATTAGAGAAATAGAGAGG	TGTGAAAGACACATGCAACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	+ 1 bp
Gmgrf5b-2	GmGRF5a	AGCACATTAGAGAAATAGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	
	GmGRF5b	AGCACATTAGAGAA----------	----------	CCGTTTCTTTGGGGAGTGGCCAC	-1625 bp
Gmgrf5a-2 Gmgrf5b-3					
	GmGRF5a	AGCACATTAGAGAAA---------	----------	----------	- 2414 bp
	GmGRF5b	AGCACATTAGAGAA--AGAGAGG	TGTGAAAGACACATG--------	------CTTTGGGGAGTGGCCAC	- 2 bp, -979 bp
Gmgrf5a-3 Gmgrf5b-4					
	GmGRF5a	AGCACATTAGAGAAATAAGAGAGG	TGTGAAAGACACATGCACAGAGG	CCGTTTCTTTGGGGAGTGGCCAC	+1 bp, +1 bp
	GmGRF5b	AGCACATTAGAGAA--AGAGAGG	TGTGAAAGACACATGCA------	------CTTTGGGGAGTGGCCAC	- 2 bp, -977 bp
Gmgrf5a-5 Gmgrf5b-5					
	GmGRF5a	AGCACATTAGAGAAA----------	----------	------CTTTGGGGAGTGGCCAC	- 1577 bp
	GmGRF5b	AGCACATTAGAGAAATA-------	----------	------CTTTGGGGAGTGGCCAC	-1782 bp

Fig. S2 The genotypes of different *GmGRF5* Cas9-edited mutants.

(A) Location of the targets and the primers in *GmGRF5a* and *GmGRF5b*. *GmGRF5a* and *GmGRF5b* share the same three targets (T1, T2, and T3. Their positions in the gene were indicated as the digitals next to them). Two pairs of primers were used to identify genotypes of *GmGRF5a* and *GmGRF5b*, respectively. Bar = 250 bp.

(B) Target sequences and details of gene editing of *GmGRF5a* and *GmGRF5b* in mutants. Red letters refer to the PAM sites; green letters refer to inserted bases; dash lines refer to deleted bases.

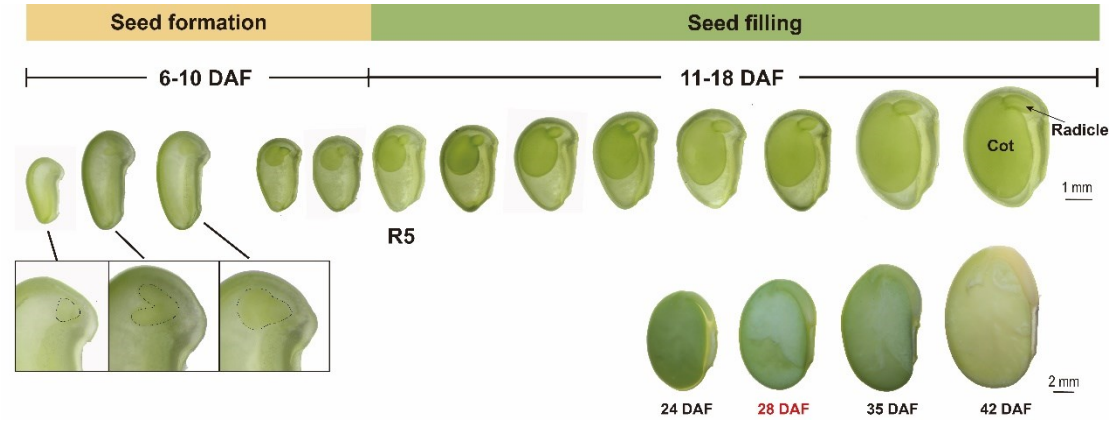


Fig. S3 Soybean seed developmental stages in growth chamber. Soybean plants were grown in soil (PINDSTRUP SPHAGNUM (pH5.5, <10 mm, Latvia): vermiculite = 3:1) in the growth chamber in short day conditions (8 h light/16 h dark), 300 ~ 500 µmol·m^{-2}·s^{-1} light, 26℃. The sampling day for transcriptome, metabolome, and RT-qPCR was 28 DAF (days after flowering).

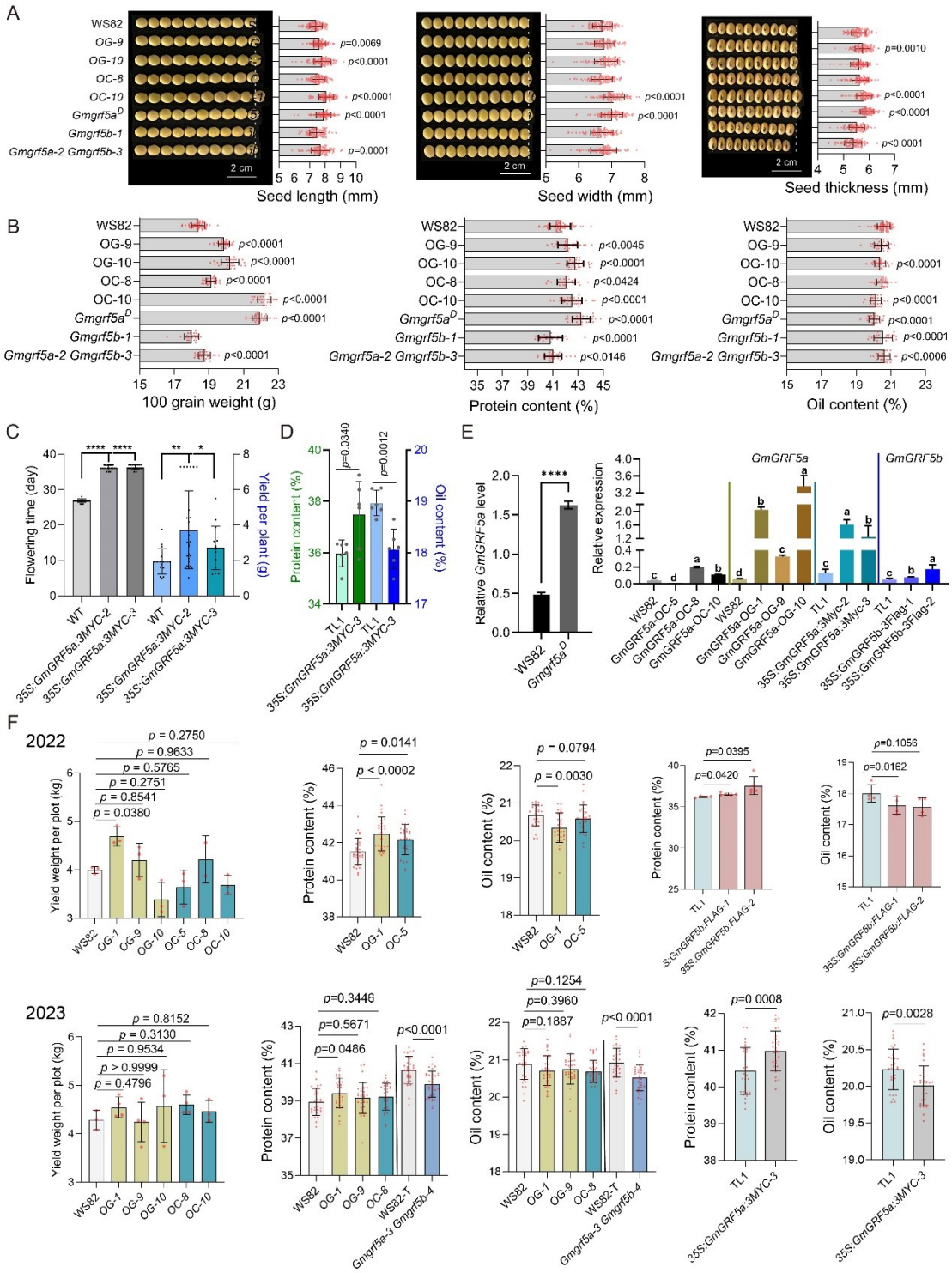


Fig. S4 Phenotypes of *GmGRF5*-overexpressing lines and mutants in chamber and in field. **A**, Seed phenotypes in growth chamber; Values are means ± SD (n = 100). Statistical analysis was performed using One-way ANOVA. **B**, 100-seed weight, yield, and the content of proteins and oils in seeds in growth chamber. **C**, Flowering time (Left Y-axis, black) and yield per plant (right axis, blue) of *35S:GmGRF5a:3MYC* in growth chamber. **D**. The content of proteins (Left Y-axis, green) and oils (right Y-axis, blue) of *35S:GmGRF5a:3MYC* seeds in growth chamber.

E, Exogenous gene expression of the *Gmgrf5a*^{D} mutant and *35S:GmGRF5a:3MYC* and *35S:GmGRF5b:FLAG* plants in growth chamber. **F**, Seed phenotypes of different overexpressing lines and mutants in the field at Shunyi Beijing in 2022 and 2023, respectively. WS82-T was used as a control for *Gmgrf5a-3 Gmgrf5b-4*. Values are means ± SD (n = 3 for Yield weight per plot, n = 30 for the content of proteins and oils). Statistical analysis was performed using Student's *t* test.

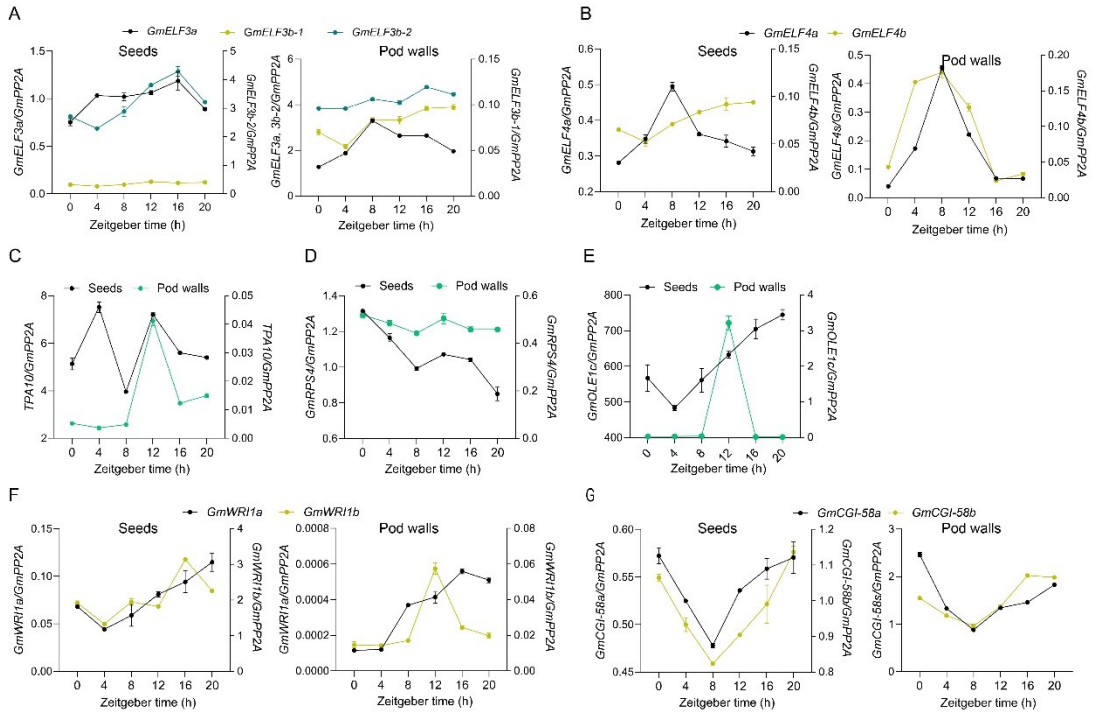


Fig. S5 The circadian expression of clock, protein synthesis, and oil synthesis genes in seeds and pod walls. Samples were harvested at DAF 28 in free running conditions after being entrained in short day conditions in growth chamber. *GmPP2A* is used as a reference gene. Due to significant difference, different genes were plotted on left or right Y-axes. Significant variations in the expression pattern were found between both different homologous genes and different tissues, and most of genes showed circadian expression patterns. Most of protein synthesis and oil synthesis genes, such as *TPA10*, *GmRPS4*, *GmOLE1c*, *GmWRI1a*, and *GmWRI1b*, showed higher expression level in seeds than in pod walls, but *GmCGI-58a* and *GmCGI-58b* was higher in pod walls than in seeds.

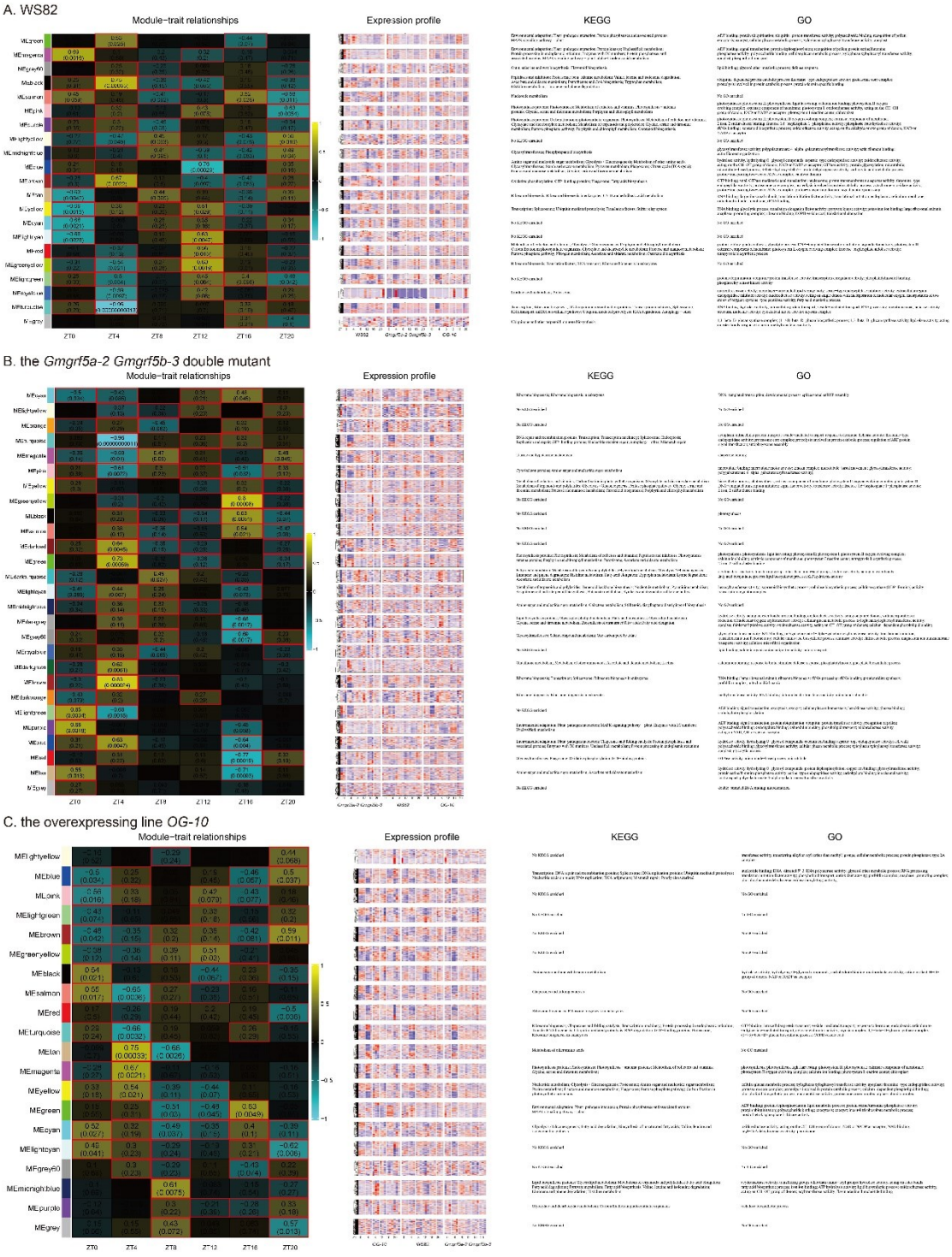


Fig. S6 Module-trait relationship in pod walls of WS82, the *Gmgrf5a-2 Gmgrf5b-3* double mutant, and the overexpressing line *OG-10*. The digital in each block was Pearson's correlation and the digital in bracket was *p* Value. Gene expression profile, KEGG enrichment and GO enrichment were analyzed for the genes in each color module.

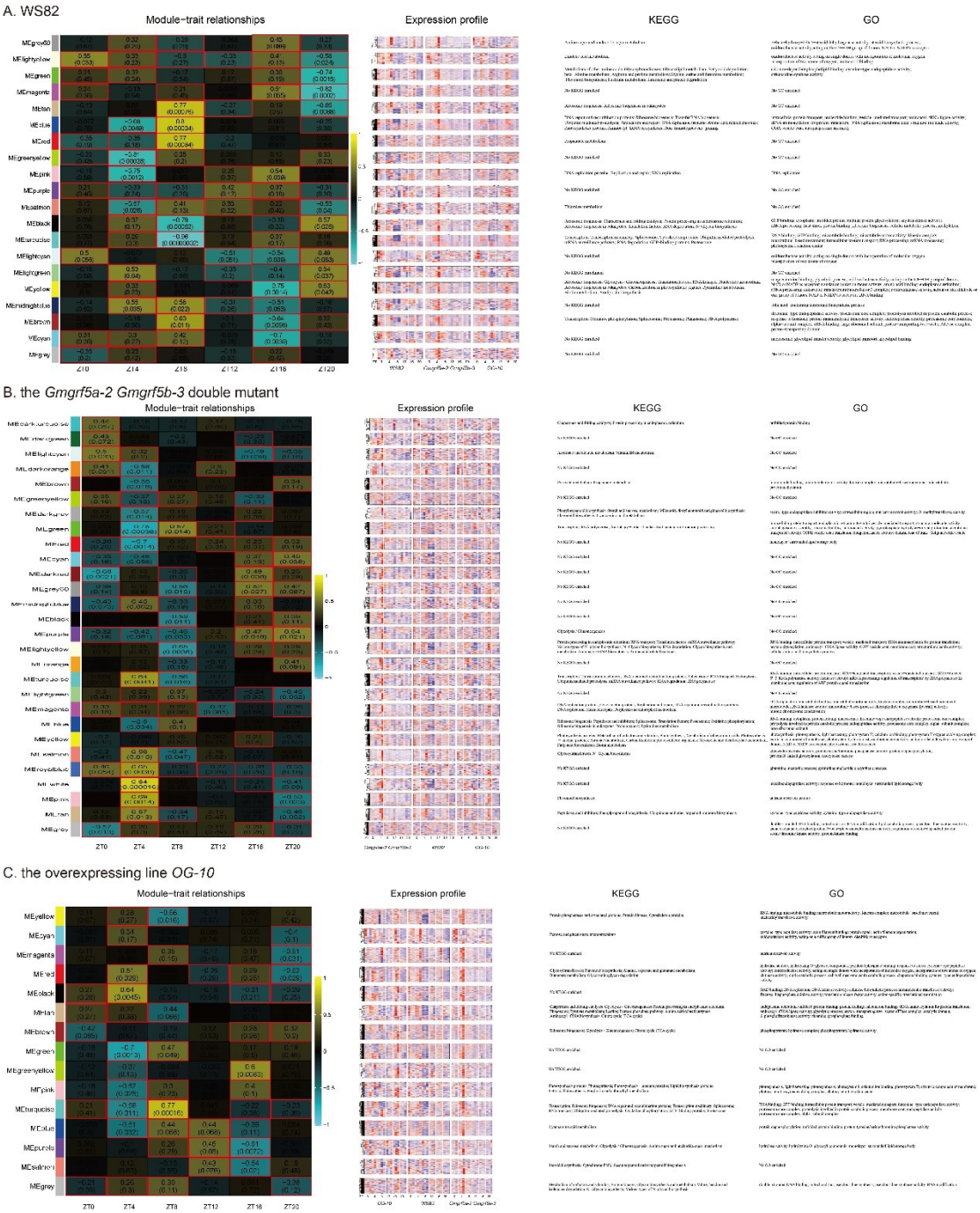


Fig. S7 Module-trait relationship in seeds of WS82, the *Gmgrf5a-2 Gmgrf5b-3* double mutant, and the overexpressing line *OG-10*. The digital in each block was Pearson's correlation and the digital in bracket was *p* Value. Gene expression profile, KEGG enrichment and GO enrichment were analyzed for the genes in each color module.

CIRCADIAN RHYTHM - PLANT

Pod
Seed

FPKM
low
high

PrB PHYB
Far red light
Red light
PfrB PHYB
PrA PHYA
Far red light
Red light
PfrA PHYA

Blue light
CRY

ELF3
PIF3
PIF3

PRR9
PRR7
PRR5

DNA
LHY
CCA1
CHE
TOC1
ZTL
GI
FKF1
GI
COP1
ELF3
+u
GI
COP1
SPA
+u
HY5

+p
CK2α
CK2β

CO
CDF1
+u

DNA
CHS
UV-B protection?

DNA
FT
Flowering

Antenna proteins
Cell elongation

Photomorphogenesis

04712 6/9/23
(c) Kanehisa Laboratories

Fig. S8 The operation of circadian clocks in soybean pod walls and seeds. This figure was constructed based on the KEGG pathway gmx04712 (Circadian rhythm - plant - Glycine max; https://www.kegg.jp/pathway/gmx04712). Green boxes indicate the presence of these genes in soybean. Arrows between upstream and downstream genes represent activation relationships, while T-shaped arrows indicate inhibition. Solid lines denote direct effects, and dashed lines signify indirect effects. The symbols "+u" and "+p" refer to ubiquitination and phosphorylation, respectively. Heatmaps adjacent to each gene box show expression profiles in seeds and pods for the WS82, *grf5*, and *OG-10*, with colors reflecting relative expression differences between seeds and pods. The line plot above displays the time-series expression patterns of these genes, with blue representing pods and red representing seeds.

Fig. S9 The circadian rhythms of putative clock genes in soybean seeds and pod walls based on transcriptome data. Refer to Table S2 for gene ID. General expression characters of these genes were showed as following: Similar pattens to their homologs in *Arabidopsis*, indicating functional conservation of clock genes in soybean; In most cases, different homologs of clock central genes showed much similar trends or patterns, such as *CCA1*/*LHY*, *TOC1*, and evening complex (*ELF3*, *LUX*, and *ELF4*), but with different levels in the same tissues (seeds or pod walls). Other genes, including *CKB4*, *REV*, *LINK* family genes displayed much different expression patterns: phases in pod walls advanced that in seeds, indicating that pod wall clock may control seed clock; compared to that in seeds, genes in the same family in pod walls show closer expression patterns, including the peak phase and general level; the difference in magnitude of variability of gene expression levels in the same tissue is not significant among the homologs.

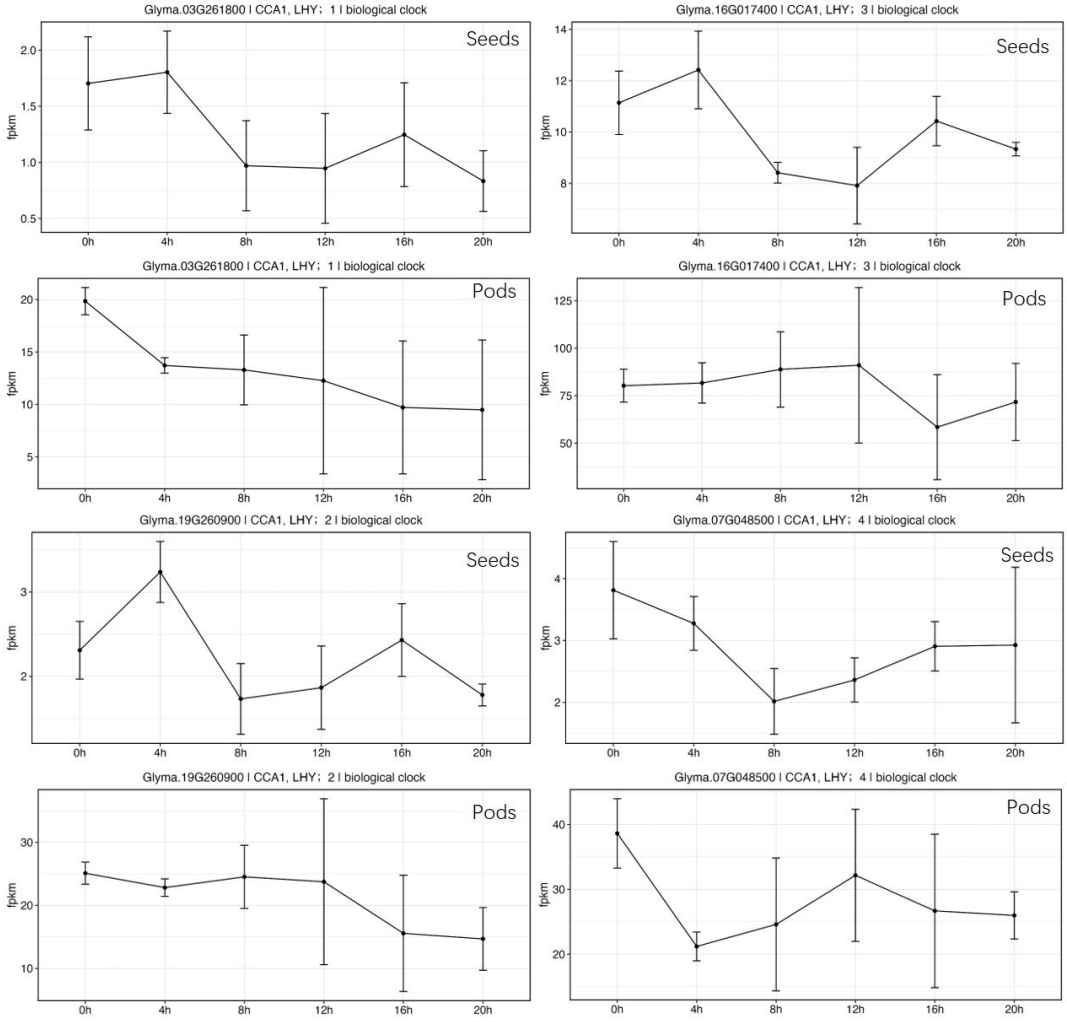


Fig. S9-1 The circadian expression of *CCA/LHY* homologs in seeds and pod walls. The general level of transcripts is higher in light than that in dark and higher in pod walls than that in seeds (nearly tenfold difference between pod walls and seeds). Different homologs display similar expression extents in the same organs, except *Glyma.16g017400*, whose expression level is higher than that of its homologs in seeds. Two peaks in seeds (ZT4 and ZT16), one peak in pod walls (ZT0); but opposite trend

for *Glyma.07G048500*, one peak in light (ZT0), the other in dark (ZT16). In term of the same organ, there is no much difference found among different homologs.

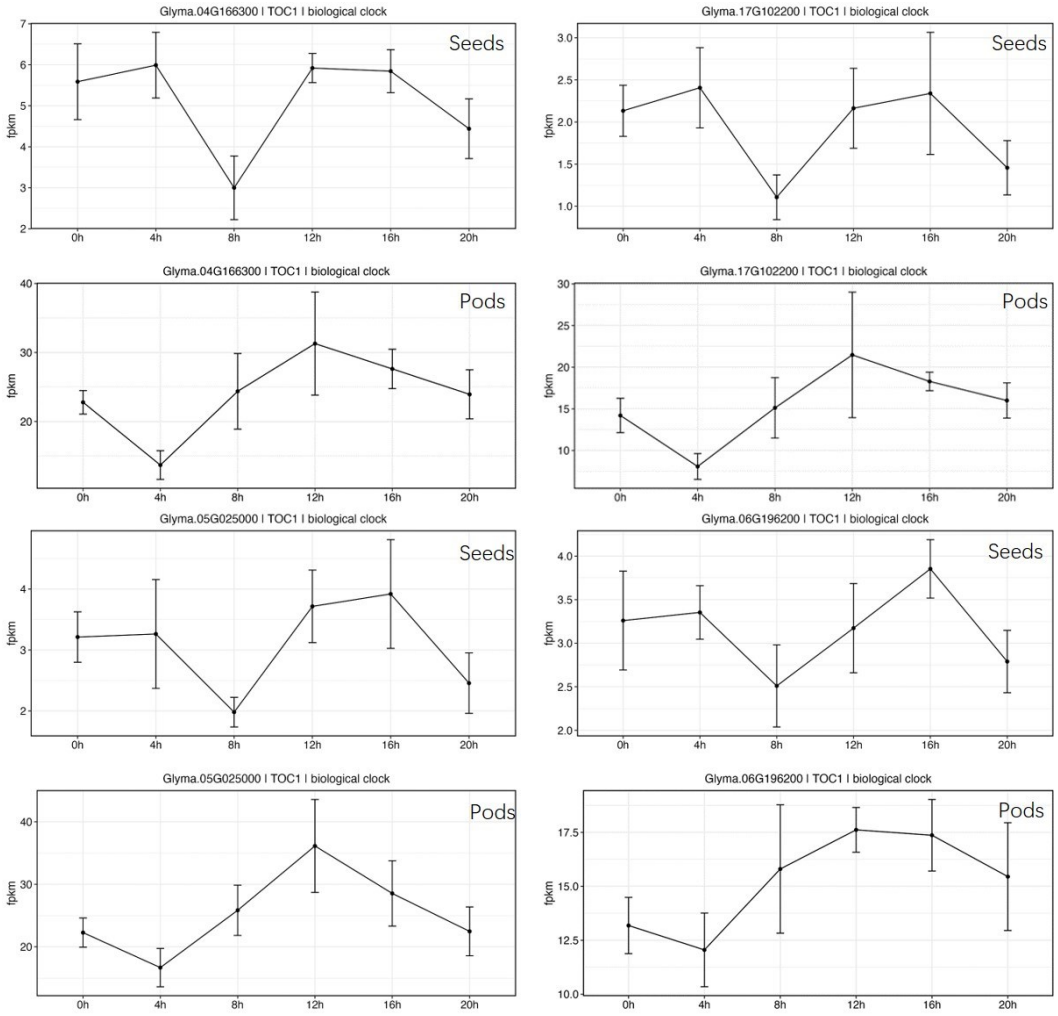


Fig. S9-2 The circadian expression of *GmTOC1* homologs in seeds and pod walls. The general level of transcripts in pod walls is higher in dark than that in light (nearly tenfold difference between pod walls and seeds), however, transcript abundancy in seeds remains at a relative high level, except at ZT8, when is the time point of the valley. Different homologs display similar expression extents in the same organs. Two peaks in seeds (ZT0/ZT4 and ZT12/16), one peak in pod walls (ZT12), indicating *GmTOC1* may have functions in both dark and light. In term of the same organ, there is no much difference found among different homologs.

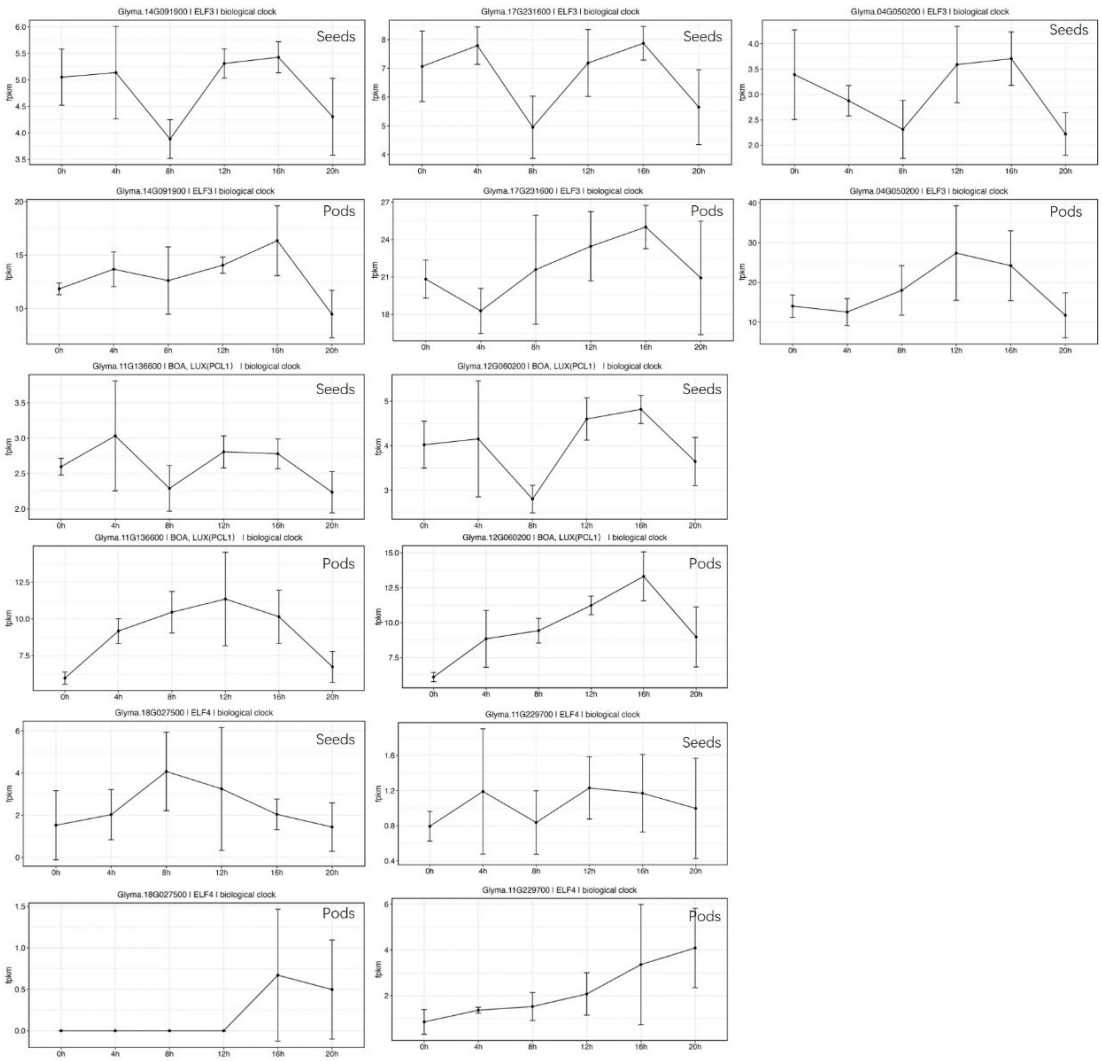


Fig. S9-3 The circadian expression of homologs of evening complex components in seeds and pod walls. For *ELF3* and *LUX* homologs, the general level of transcripts in seeds is quite similar in light and dark conditions (two peaks at ZT4 or ZT0 and ZT16, respectively), but with a valley (the lowest value) at ZT8; In pod walls, the expression peaks happen at ZT12 or ZT16. In term of the same organ, there is a difference found among different homologs. For ELF4, there is one (ZT8, *Glyma.18g027500*) or two (ZT4 and ZT12, *Glyma.11g229700*) peaks in seeds; in pod walls, the peaks appear in the late phase of dark.

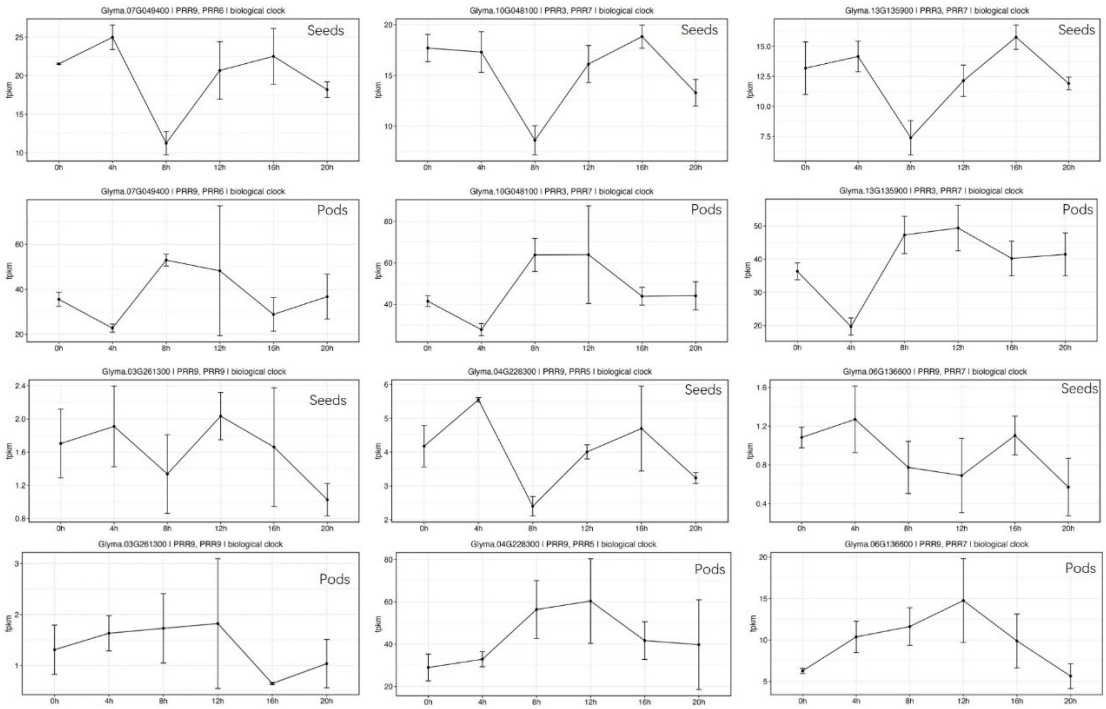


Fig. S9-4 The circadian expression of *GmPRR* homologs in seeds and pod walls. The general level of transcripts in seeds is quite similar in light and dark conditions (two peaks at ZT4 and ZT16, respectively), but with a valley (the lowest value) at ZT8 except of *Glyma.06g136600* with a valley at ZT12. In pod walls, the expression peaks happen at ZT8 or ZT12. In term of the same organ, there is no much difference found among different homologs.

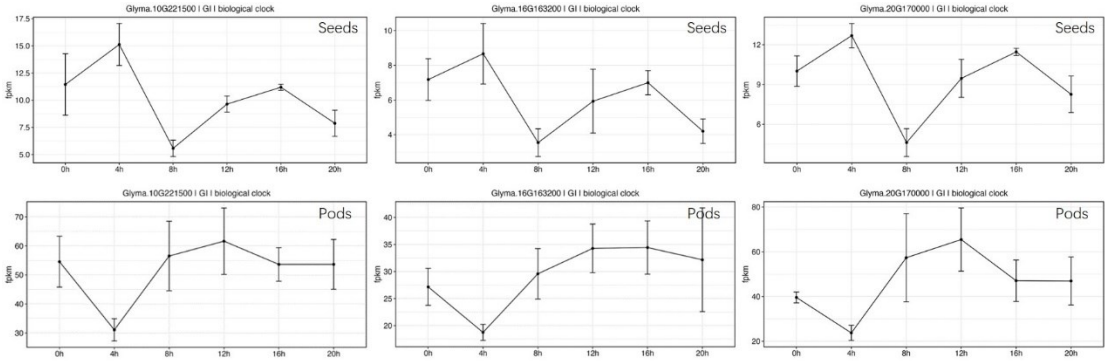


Fig. S9-5 The circadian expression of *GI* homologs in seeds and pod walls. The general level of transcripts in seeds is quite similar in light and dark conditions (two peaks at ZT4 and ZT16, respectively), but with a valley (the lowest value) at ZT8; In pod walls, the expression peaks appear at ZT12.

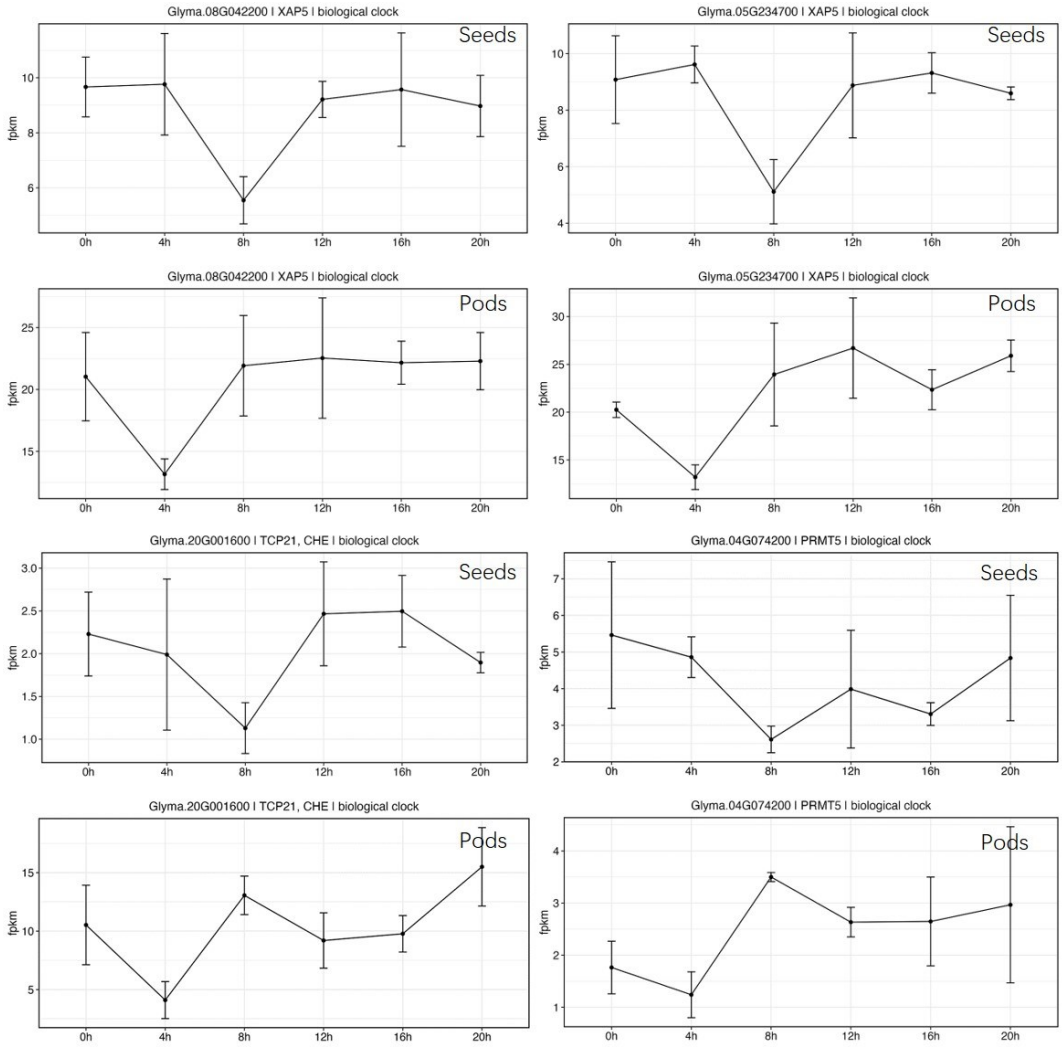


Fig. S9-6 The circadian expression of *XAP5*, *CHE*, and *PRMT5* homologs in seeds and pod walls. *XAP5*, *CHE*, and *PRMT5* homologs share a common expression pattern in seeds, that is, the general level of transcripts in seeds is quite similar in light and dark conditions (two peaks at ZT4 and ZT16 for *XAP5* and *CHE*, or ZT0 for *PRMT5*, respectively), but with a valley (the lowest value) at ZT8; In pod walls, the expression

curve is quite similar to that in seeds, but phases in pod walls advance obviously than that in seeds.

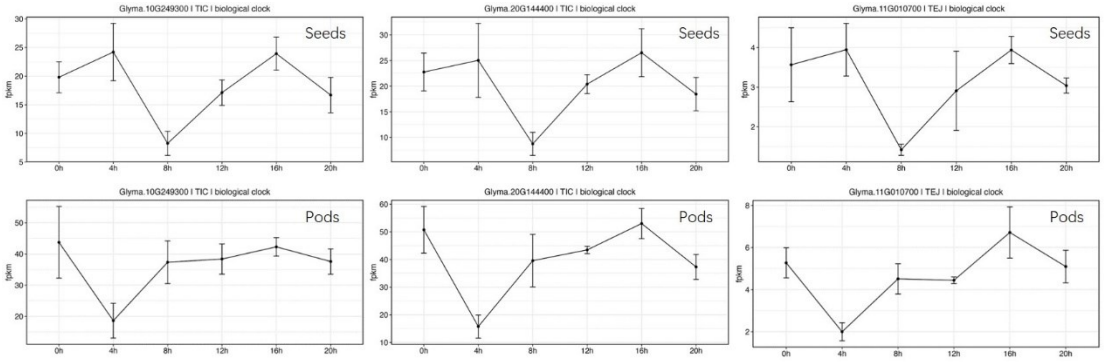


Fig. S9-7 The circadian expression of *TIC* homologs in seeds and pod walls. Three *TIC* homologs share a common expression pattern in both seeds and pod walls, but phases in pod walls advance obviously than that in seeds. The peaks in seeds appear at ZT4 and ZT16, but at ZT0 and ZT16 in pod walls.

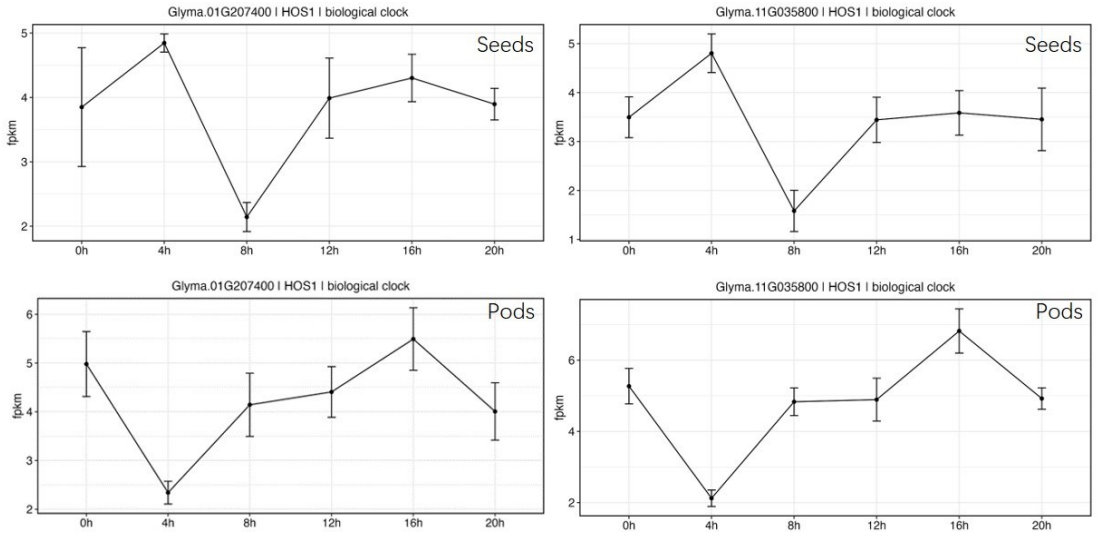


Fig. S9-8 The circadian expression of *HOS1* homologs in seeds and pod walls. Two *HOS1* homologs share a common expression pattern in both seeds and pod walls, but phases in pod walls advance obviously than that in seeds. The peaks in seeds appear at ZT4, but at ZT0 and ZT16 in pod walls.

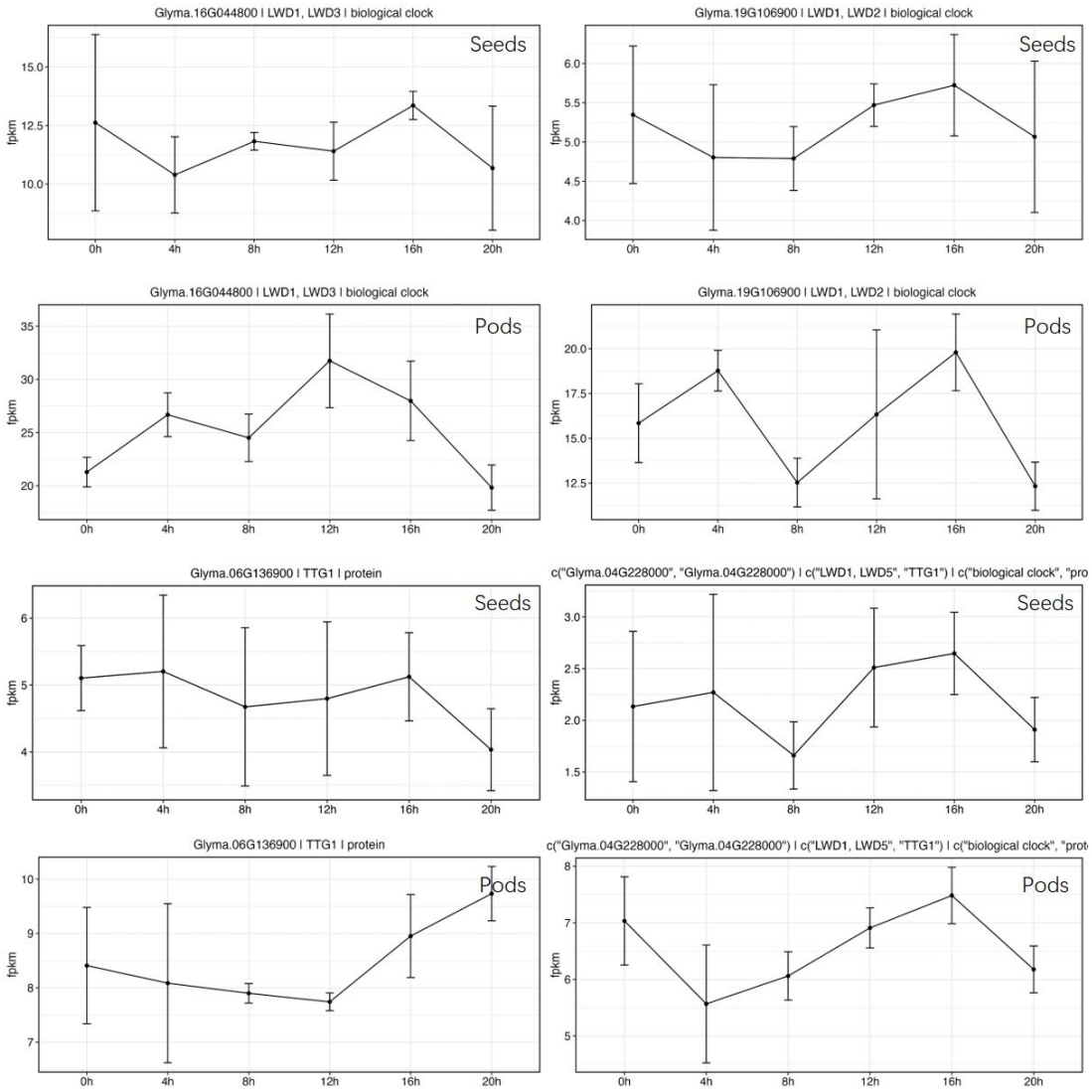


Fig. S9-9 The circadian expression of *LWD* and *TTG1* homologs in seeds and pod walls. Two *LWD* homologs (*Glyma.16g044800* and *Glyma.19g106900*) share a common expression pattern in both seeds and pod walls, but phases advance in pod walls obviously than that in seeds. The peaks in seeds appear at ZT16, but ZT12 in pod walls. For *Glyma.06g136900* and *Glyma.04g228000*, there are no striking peak in seeds, but with peaks at ZT20 and ZT 16, respectively.

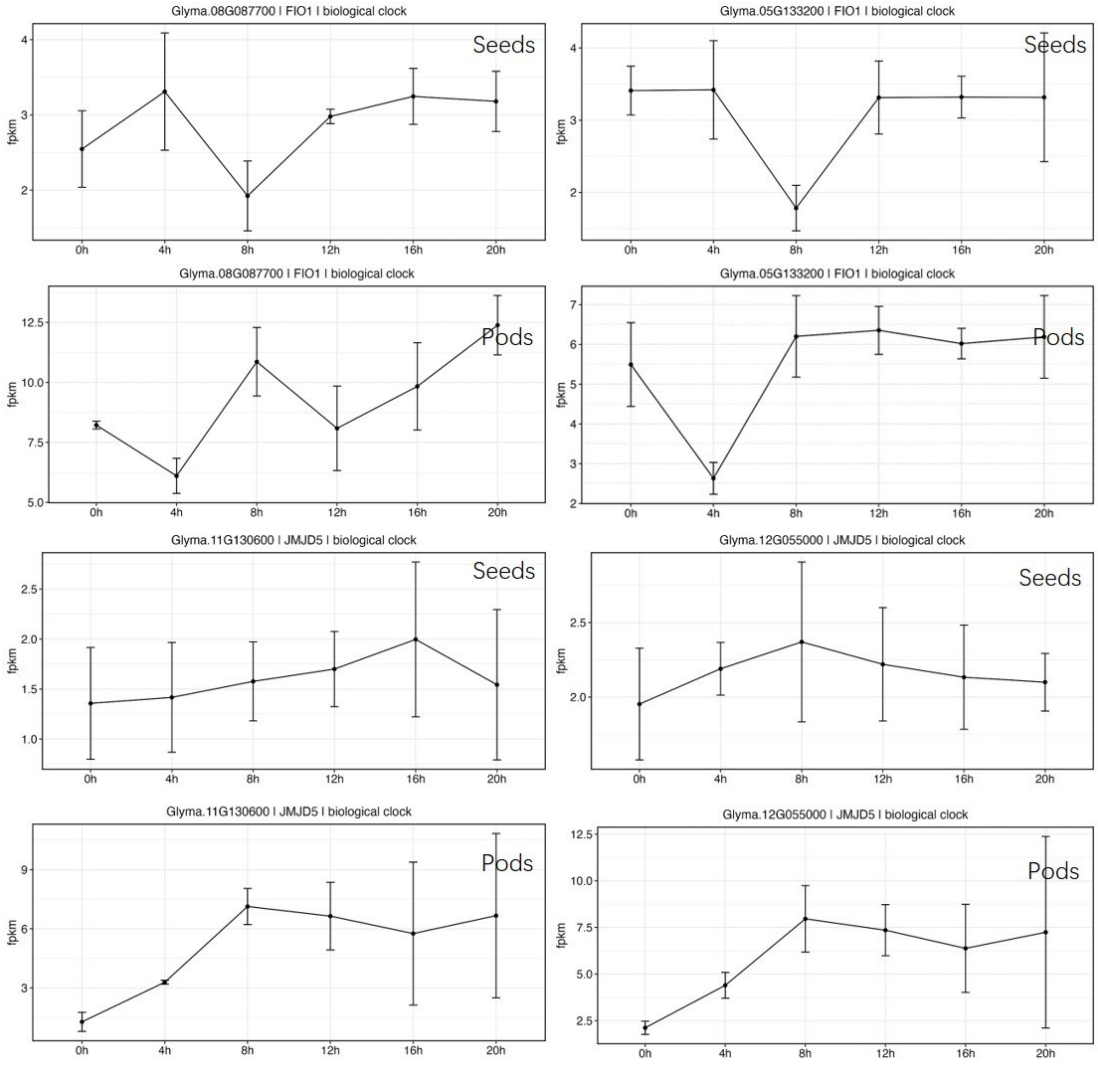


Fig. S9-10 The circadian expression of *FIO1* and *JMJD5* homologs in seeds and pod walls. Two *FIO* homologs share a common expression pattern in both seeds and pod walls, but phases in pod walls advance obviously than that in seeds. The peaks in seeds appear at ZT4 and ZT16, but ZT8 and ZT20 in pod walls. Two *JMJD5* display different patterns in seeds with peaks at ZT16 and ZT8, respectively, but a similar pattern in pod walls with a peak at ZT8.

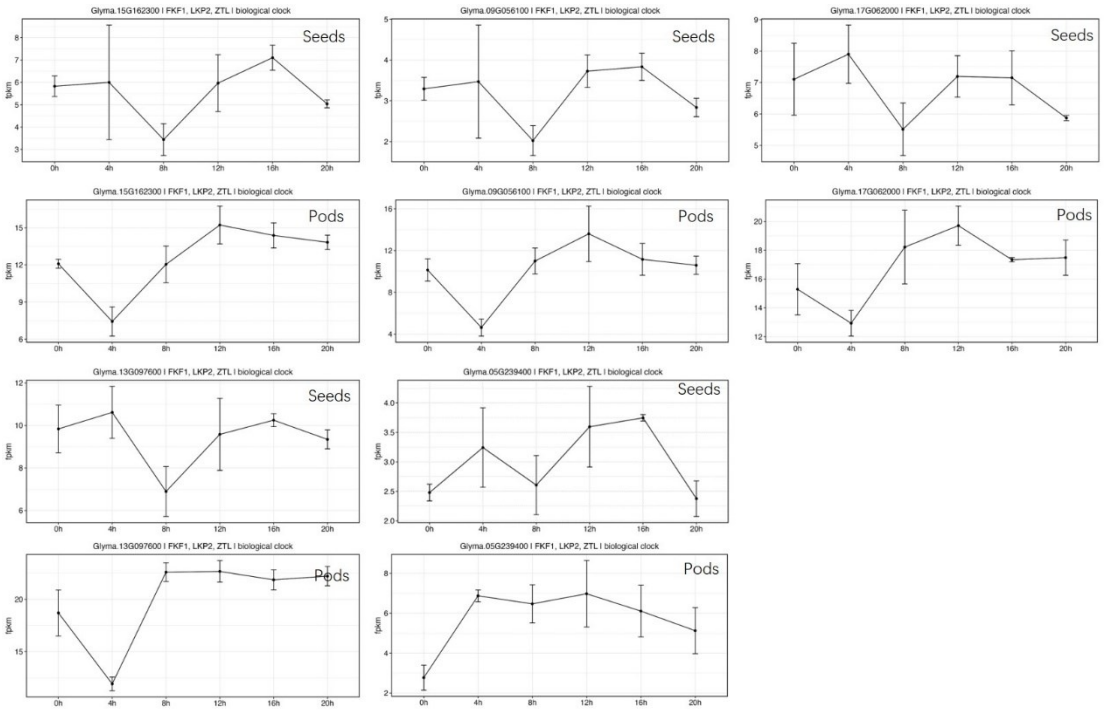


Fig. S9-11 The circadian expression of *FKF*/*LKP2*/*ZTL* homologs in seeds and pod walls. These five *FKF*/*LKP2*/*ZTL* homologs share a common expression pattern in both seeds and pod walls, but phases in pod walls advance obviously than that in seeds. The peaks in seeds appear at ZT4 and ZT16, but ZT12 in pod walls.

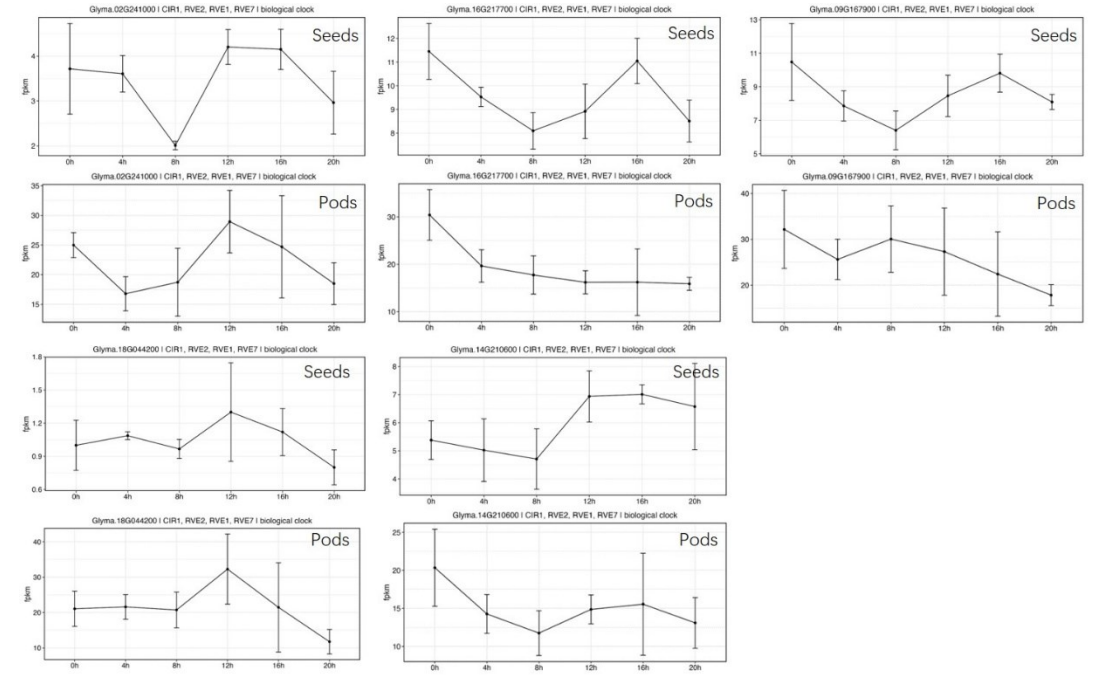


Fig. S9-12 The circadian expression of *REV2* homologs in seeds and pod walls. These genes show more or less similar expression patterns in seeds (peak at ZT16 or ZT12), but in pod walls there is a variation found and can be grouped into two classes: one has *Glyma.16G217700*, *Glyma.09G167900*, and *Glyma.14G210600* (peak at ZT0), the other has *Glyma.02G241000* and *Glyma.18G044200* (peak at ZT12).

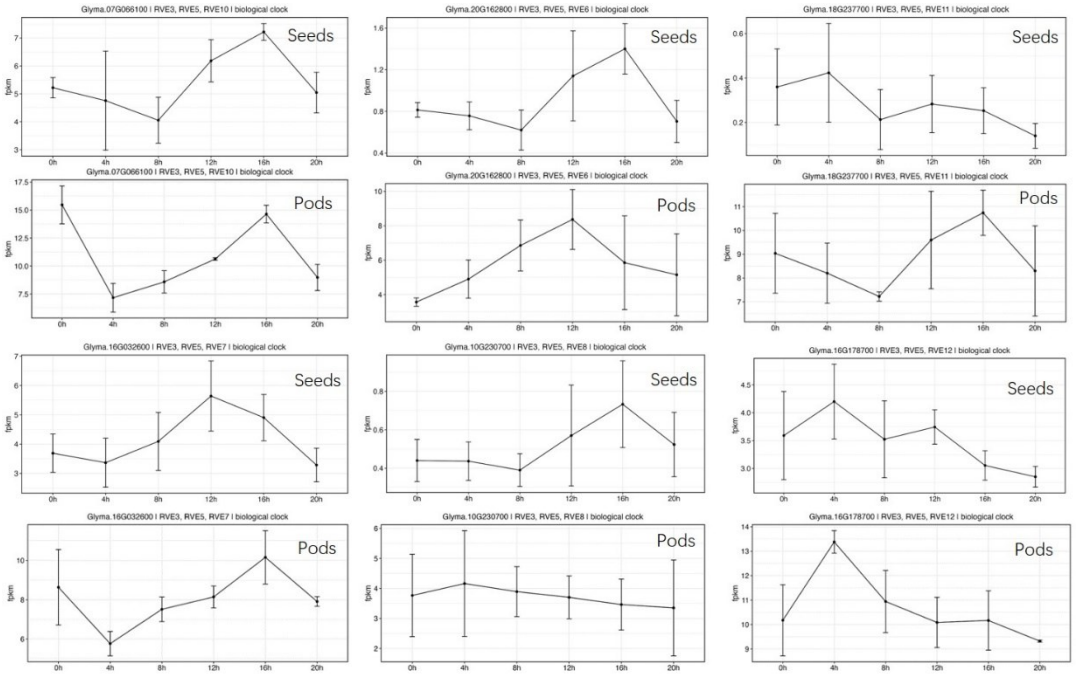


Fig. S9-13 The circadian expression of *REV3* homologs in seeds and pod walls. *Glyma.07G066100*, *Glyma.20G162800*, *Glyma.16G032600*, and *Glyma.10G230700* show similar expression patterns (peak at ZT12 or ZT16) in seeds, but not in pod walls (peak at ZT0, ZT12 or ZT16). *Glyma.18G237700* and *Glyma.16G178700* show similar expression patterns (peak at ZT4) in seeds, but not in pod walls (peak at ZT4 or ZT16).

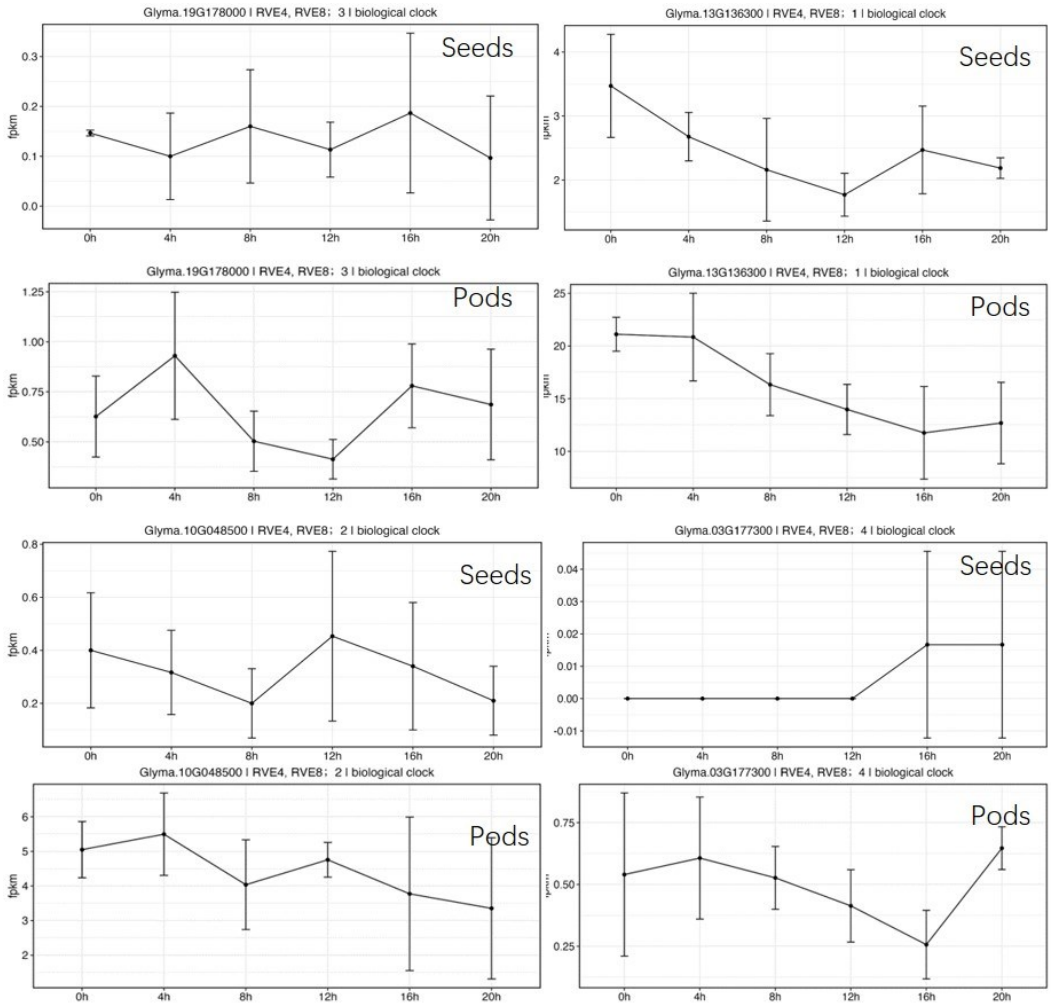


Fig. S9-14 The circadian expression of *REV4/8* homologs in seeds and pod walls. These *REV* homologs do not share common expression patterns in both seeds and pod walls and different homologs exhibit various expression patterns in both seeds and pod walls.

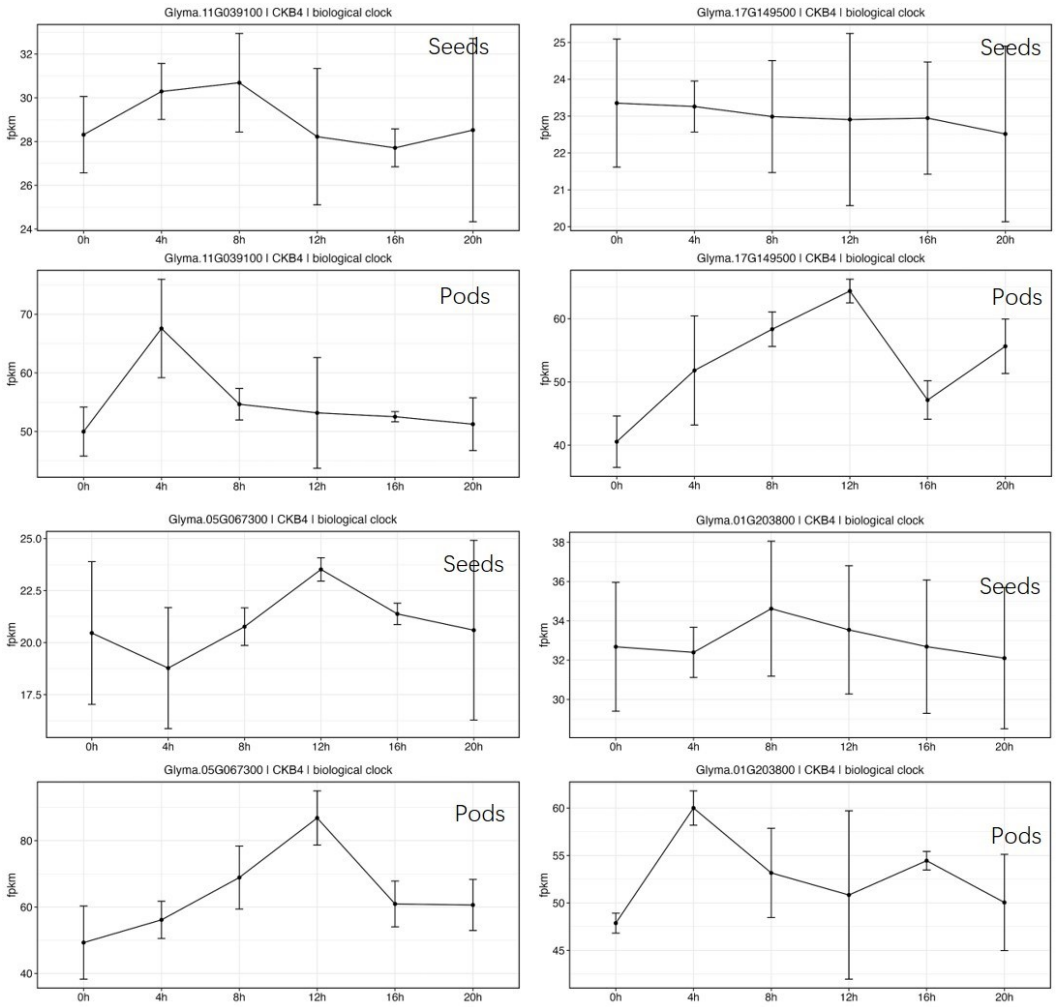


Fig. S9-15 The circadian expression of *CKB4* homologs in seeds and pod walls. These *CKB4* homologs do not share common expression patterns in both seeds and pod walls, and different homologs exhibit various expression patterns in both seeds and pod walls.

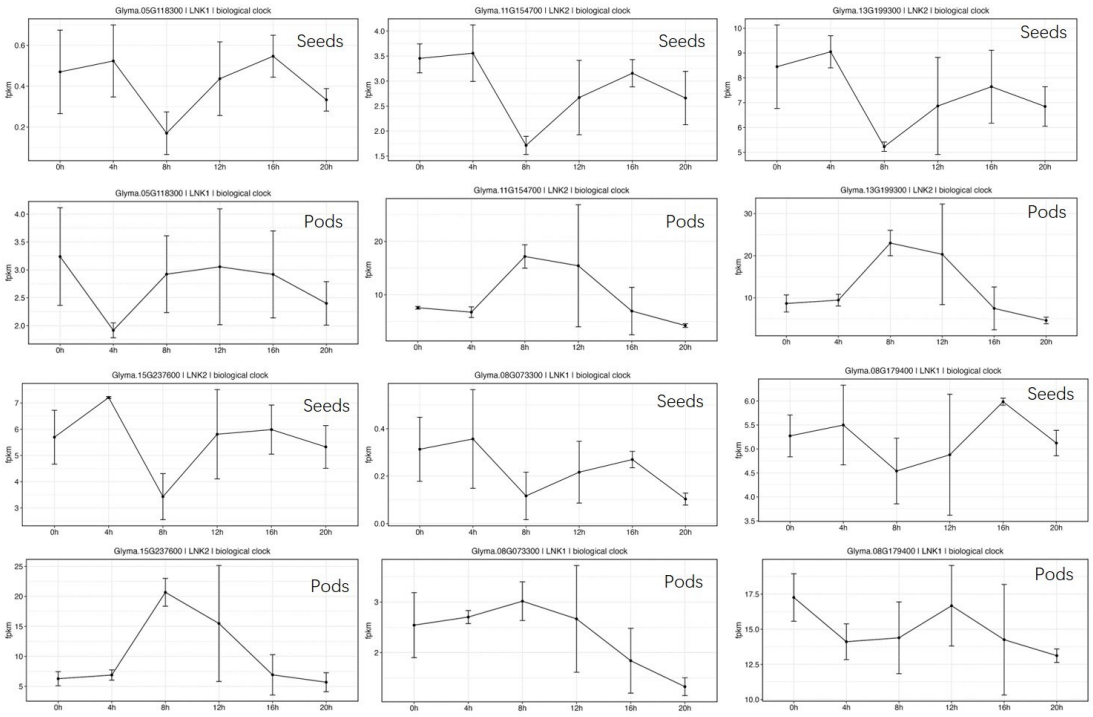


Fig. S9-16 The circadian expression of *LINK* homologs in seeds and pod walls-1. These *LINK* homologs share a common expression pattern in seeds (peaks at ZT4 and ZT16), but not in pod walls (peaks at ZT8 or ZT12, respectively).

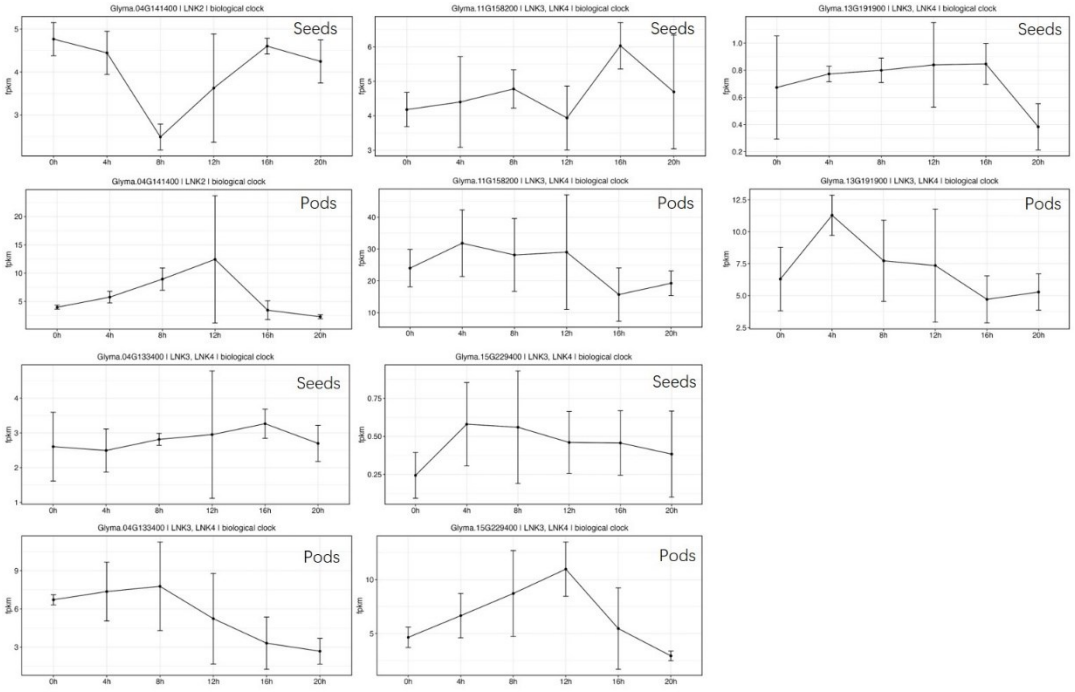


Fig. S9-17 The circadian expression of *LINK* homologs in seeds and pod walls-2. These *LINK* do not homologs share a common expression pattern in seeds (peaks at ZT0, ZT4, or ZT16), but not in pod walls (peaks at ZT4, ZT8, or ZT12, respectively).

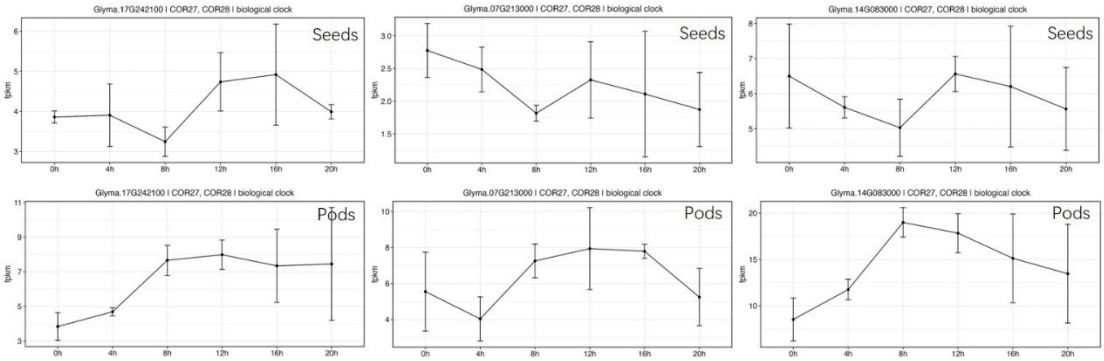


Fig. S9-18 The circadian expression of *COR27/28* homologs in seeds and pod walls-2. These *COR27/28* do not homologs share a common expression pattern in both seeds (peaks at ZT12 or ZT16) and pod walls (peaks at ZT8 or ZT12, respectively).

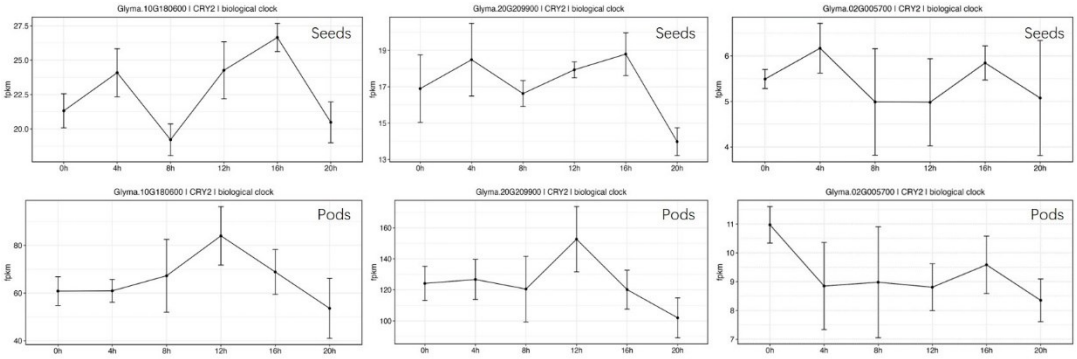


Fig. S9-19 The circadian expression of *CRY2* homologs in seeds and pod walls. *CRY2* homologs share common expression patterns in both seeds (peaks at ZT4 or ZT16) and pod walls (peaks at ZT12).

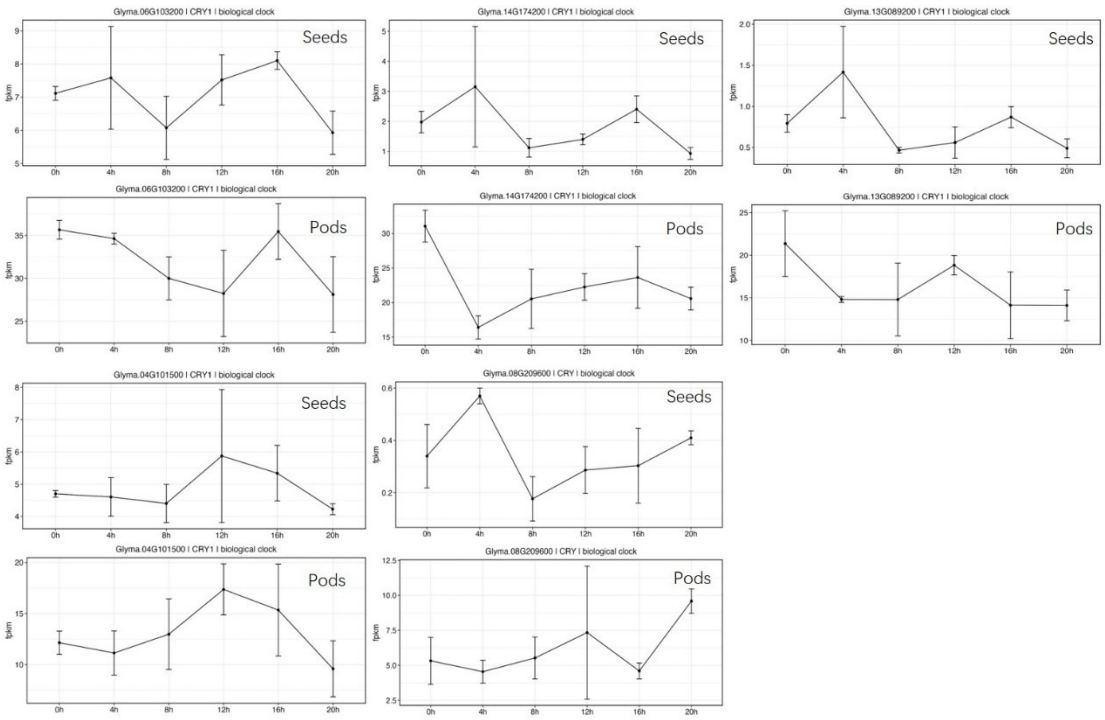


Fig. S9-20 The circadian expression of *CRY1* homologs in seeds and pod walls. Three *CRY1* homologs (*Glyma.06G103200*, *Glyma.13G089200*, *Glyma.14G174200*) share common expression patterns in both seeds (peaks at ZT4 or ZT16) and pod walls (peaks at ZT0, ZT12, or ZT16, respectively). Other two *CRY1* homologs (*Glyma.04G101500* and *Glyma.08G209600*) display different expression in both seeds (peaks at ZT2 or ZT12) and pod walls (peaks at ZT12 or ZT20).

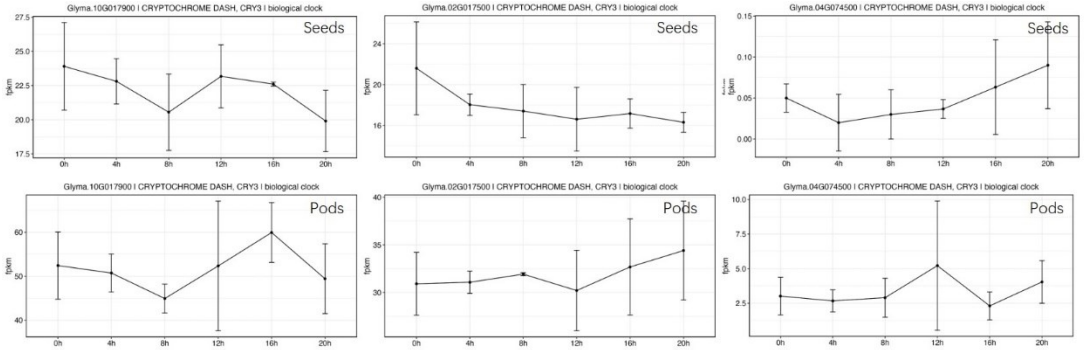


Fig. S9-21 The circadian expression of *CRY3* homologs in seeds and pod walls. *CRY3* homologs do not a share common expression pattern in seeds (peaks at ZT0, ZT12, or ZT20, respectively) and pod walls (peaks at ZT12, ZT16, or ZT20, respectively).

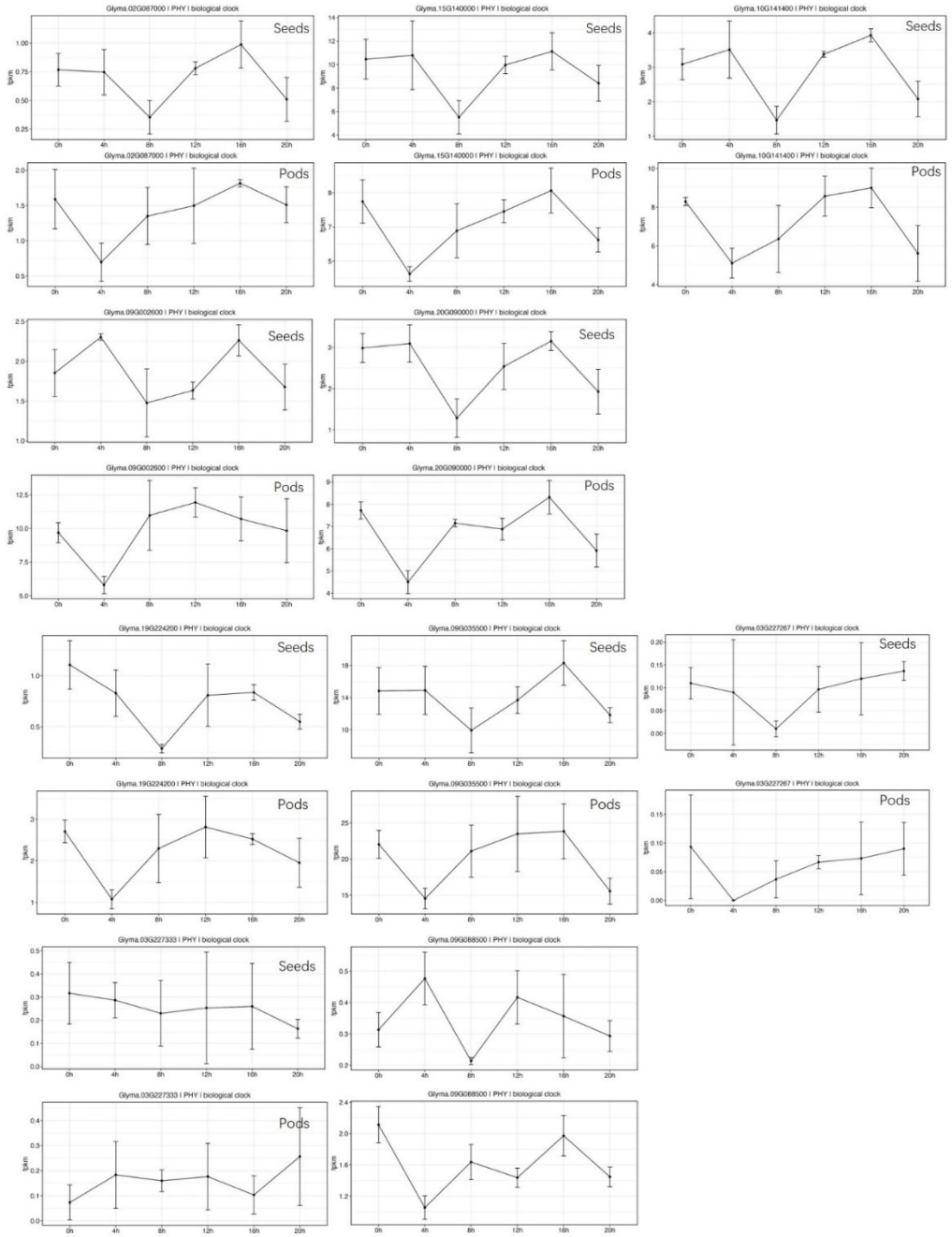


Fig. S9-22 The circadian expression of *PHY* homologs in seeds and pod walls-1. These *PHY* homologs share common expression patterns in seeds (peaks at ZT16) and pod walls (peaks at ZT12 or ZT16).

Fig. S9-23 The circadian expression of *PHY* homologs in seeds and pod walls-2. These *PHY* do not homologs share common expression patterns in seeds and pod walls.

Fig. S10 The circadian expression pattern of soybean genes related to protein and oil synthesis. We extracted expression patterns of following genes from transcriptome data in this study, most of them showed circadian rhythm. However, their circadian rhythms were not robust than that of clock genes (Fig. S8). Refer to Table S2 for gene IDs. Their homologs in *Arabidopsis* are following: *LEC1* (*Leafy Cotyledon1*) (Feeney et al., 2013), *ABI3* (*Abscisic Acid Insensitive3*) and *ABI5* (Kagaya et al., 2005; Verdier and Thompson, 2008; Yeap et al., 2017; Vanhercke et al., 2019), *TuODORANT1* (Luo et al., 2021), *PV72* (a vacuolar sorting receptor) (Shimada et al., 2002), *GmSWEET39* (*Sugars Will Eventually be Exported Transporters 39*) (Zhang et al., 2020), *GOGAT* (*glutamate synthase*) (Yang et al., 2018), *TTG1* (Chen et al., 2015), *PPA1* (*pyrophosphatase*) (Meyer et al., 2012), *GmPPC2* (*PEPC*, *phosphoenolpyruvate* carboxylase, *isoform*) (Yamamoto et al., 2020), *PEI* (Verdier and Thompson, 2008), *PsOMT* (*2-oxoglutarate*/*malate translocator*) (Riebeseel et al., 2010), *FUS3* (*FUSCA3*) (Kagaya et al., 2005), *MYB*, *MYC*, and *MAX* (Gao et al., 2016; Luo et al., 2021), *WRI* (Kanai et al., 2016; Vanhercke et al., 2019), *CGI-58* (James et al., 2010; Vanhercke et al., 2019), *SEIPIN* (Wood et al., 2018; Kong et al., 2019; Vanhercke et al., 2019), *OLE1* (Kong et al., 2019; Vanhercke et al., 2019; Ortiz et al., 2020), *FAD2* (Wood et al., 2018; Vanhercke et al., 2019), *MYB89* (Vanhercke et al., 2019), *GPDH* (Vanhercke et al., 2019), *DGAT* (Bouvier-Nave et al., 2000; Vanhercke et al., 2019; Ortiz et al., 2020).

Following genes displayed strong circadian expression patterns in seeds or pod walls:

- **In seeds:** *Glyma.07G268100* (*LEC1*), *Glyma.10G071700* (*ABI5*), *Glyma.13G153200* (*ABI5*), *Glyma.19G194500* (*ABI5*), *Glyma.18G176100* (*ABI5*), *Glyma.08G357600* (*ABI5*), *Glyma.11G215800* (*TuODORANT1*), *Glyma.01G242800* (*PV72b*), *Glyma.11G001500* (*PV72b*), *Glyma.15G049200* (*SWEET39*), *Glyma.04G236900* (*GOGAT2*), *Glyma.06G127400* (*GOGAT2*), *Glyma.06G266800* (*ILR3*), *Glyma.16G017200* (*PPA1*), *Glyma.12G229400* (*PPC2*), *Glyma.13G270400* (*PPC2*), *Glyma.16G050300* (*FUS3*), *Glyma.01G096600* (*MYC3*), *Glyma.08G227700* (*WRI1*), *Glyma.15G221600* (*WRI1*), *Glyma.04G026500* (*CGI-58*), *Glyma.14G216800* (*CGI-58*), *Glyma.18G242100* (*SEIPIN*), *Glyma.09G250400* (*SEIPIN*), *Glyma.*07G228700 (*MYB89*), *Glyma.02G218700* (*GPDH*), *Glyma.09G065300* (*DGAT*), *Glyma.17G053300* (*DGAT*), and *Glyma.13G106100* (*DGAT*);
- **In pod walls:** *Glyma.11G215800* (*TuODORANT1*. Opposite to that in seeds), *Glyma.04G236900* (*GOGAT2*), *Glyma.06G127400* (*GOGAT2*), *Glyma.16G017200* (*PPA1*), *Glyma.11G070800* (*PPA1*), *Glyma.19G261100* (*PPA1*), *Glyma.04G026500* (*CGI-58*), *Glyma.14G216800* (*CGI-58*), *Glyma.*07G228700 (*MYB89*) and *Glyma.09G065300* (*DGAT*).

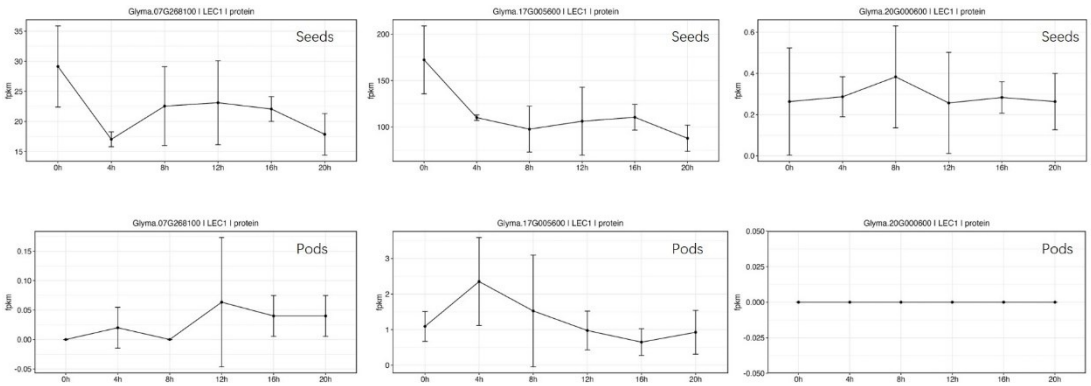


Fig. S10-1 The circadian expression of *LEC* homologs in seeds and pod walls. The *LEC* homologs *Glyma.07G268100* and *Glyma.17G005600* showed relative higher expression in seeds at dawn and in dark than that in light, while *Glyma.20G000600* did not show circadian rhythm. These three genes did not display obvious circadian expression in pod walls, maybe due to their functions in seeds.

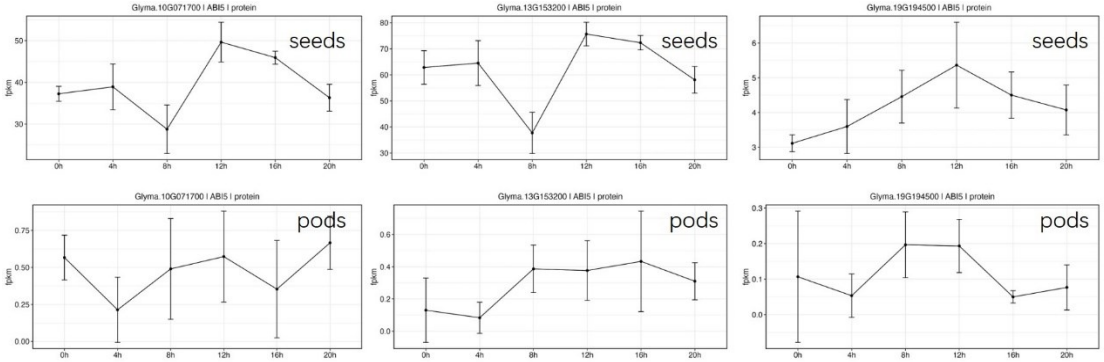


Fig. S10-2 The circadian expression of *ABI5* homologs in seeds and pod walls. *Glyma.10G071700* and *Glyma.13G153200* had a valley at ZT8 in seeds or ZT4 in pod walls, while *Glyma.19G194500* had a peak at ZT8 to ZT12 in seeds or pod walls, indicating they had the characteristics of circadian rhythm. Additionally, the valleys or peaks advanced in pod walls than that in seeds.

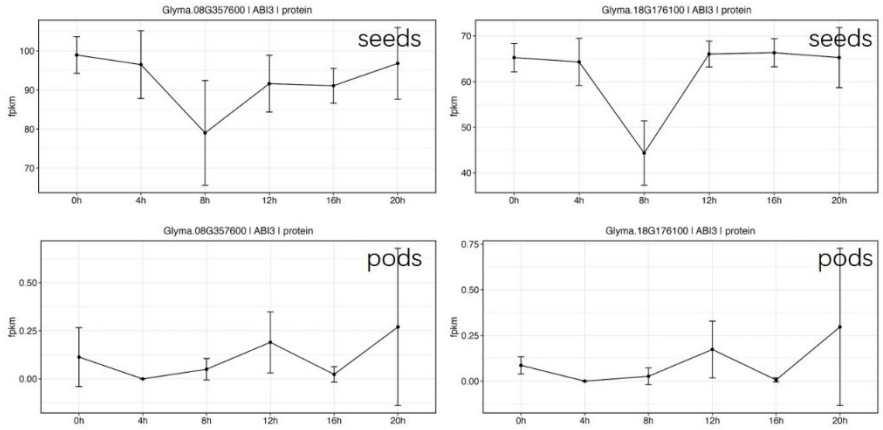


Fig. S10-3 The circadian expression of *ABI3* homologs in seeds and pod walls. These two genes (*Glyma.18G176100* and *Glyma.08G357600*) had obvious valleys at ZT8 (dusk) in seeds, but not in pod walls.

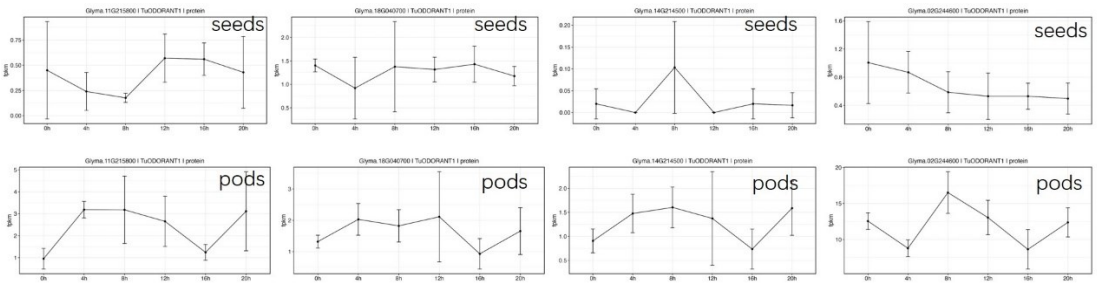


Fig. S10-4 The circadian expression of *TuODORANT1* homologs in seeds and pod walls. *Glyma.11G215800* showed different patterns in seeds and pod walls: low level in the second half of light-phase in seeds while in the first half in pod walls. The peaks of *Glyma.14G214500* were at ZT8 in both seeds and pod walls. *Glyma.02G244600* had a peak at dawn in seeds or at dusk in pod walls. *Glyma.18G040700* did not show significant circadian rhythms in both seeds and pod walls.

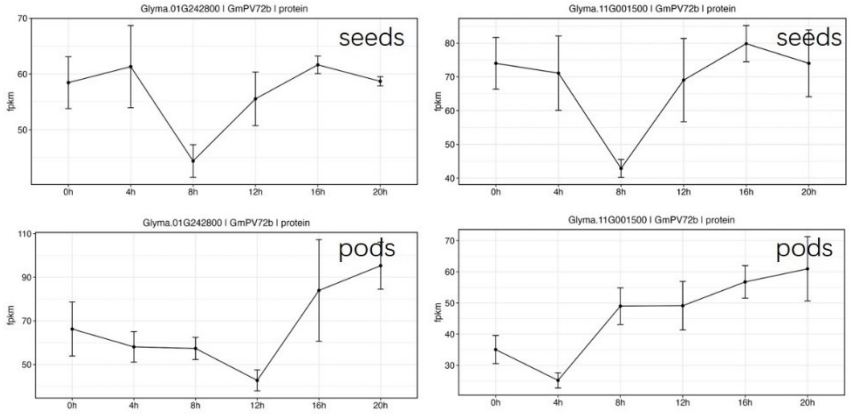


Fig. S10-5 The circadian expression of *PV72b* homologs in seeds and pod walls. *Glyma.01G242800* and *Glyma.11G001500* shared a common pattern in seeds with valleys at ZT8. In pod walls, *Glyma.01G242800* had a valley at ZTT12, while Glyma.11G001500had a valley at ZT4, but their peaks appeared at the end of dark.

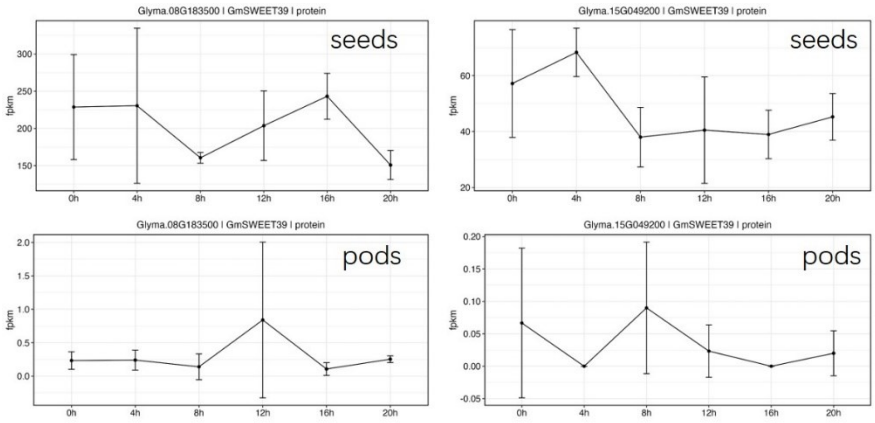


Fig. S10-6 The circadian expression of *SWEET39* homologs in seeds and pod walls. *Glyma.15G049200* and *Glyma.08G183500* may not show circadian rhythm expression in both seeds and pod walls.

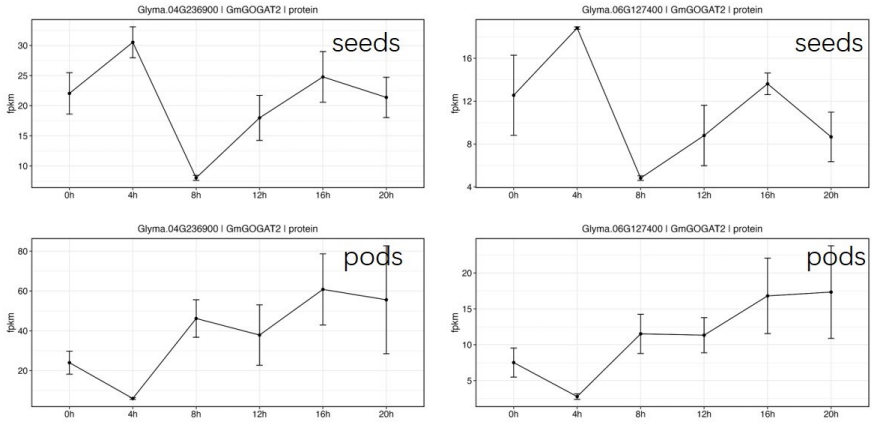


Fig. S10-7 The circadian expression of *GOGAT2* homologs in seeds and pod walls. The expression of *Glyma.04G236900* and *Glyma.06G127400* had similar pattern in seeds and pod walls. In seeds they peaked at ZT4, but in pod walls they peaked at the end of dark.

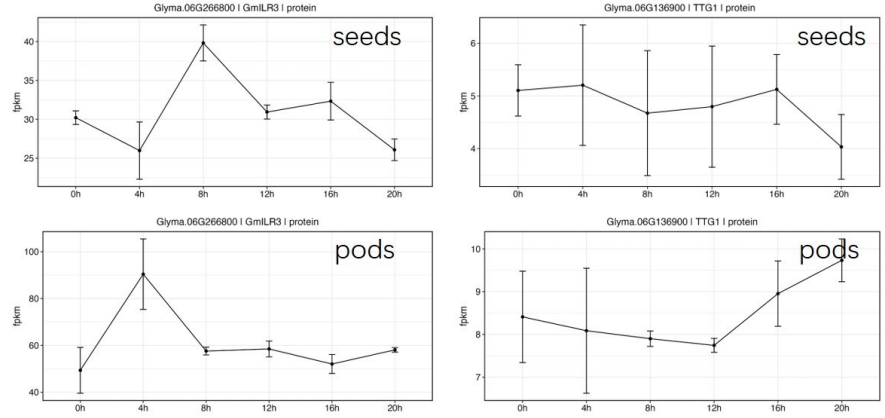


Fig. S10-8 The circadian expression of *ILR3* and *TTG1* homologs in seeds and pod walls. *Glyma.06G266800* peak at ZT8 in seeds, but at ZT4 in pod walls. *Glyma.06G136900* had no circadian rhythm in seeds, but had relative high level in dark in pod walls.

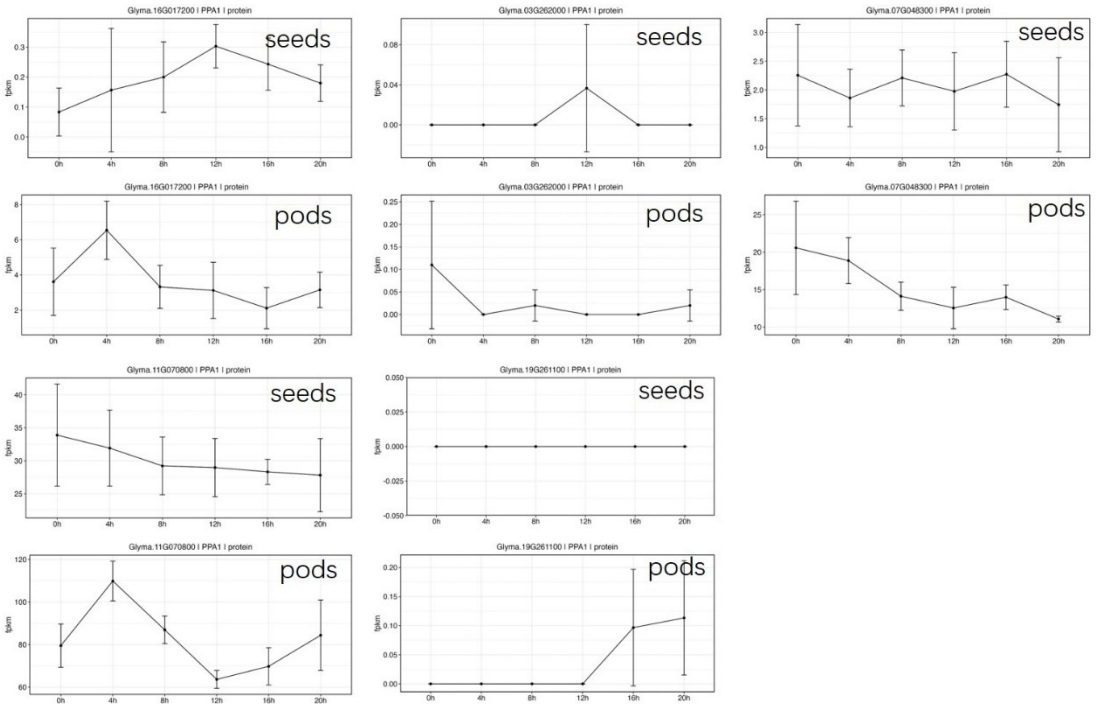


Fig. S10-9 The circadian expression of *PPA1* homologs in seeds and pod walls.
These genes peaked: *Glyma.16G017200* at ZT12 in seeds, but ZT4 in pod walls; *Glyma.03G262000* at ZT12 in seeds, but at ZT0 in pod walls; *Glyma.07G048300* at ZT0 in both seeds and pod walls; *Glyma.11G070800* at ZT0 in seeds, but at ZT4 in pod walls; *Glyma.19G261100* at ZT20 in pod walls, but no peak in seeds.

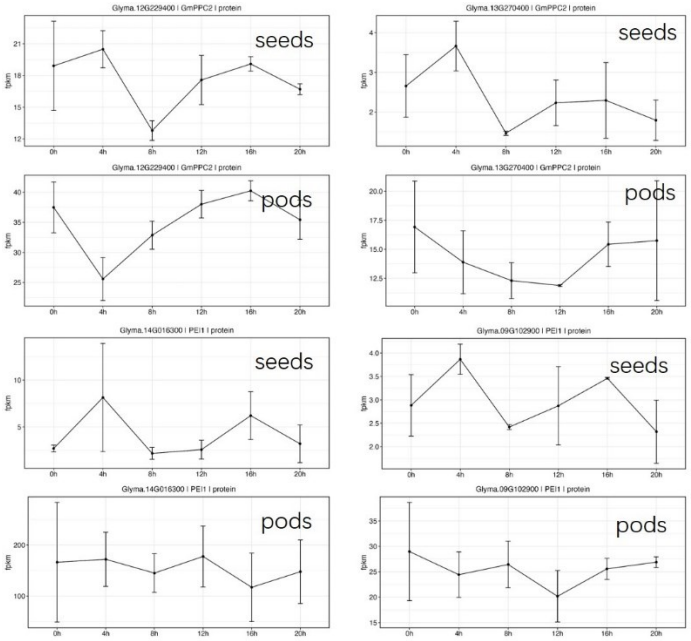


Fig. S10-10 The circadian expression of *PPC2 and PEI1* homologs in seeds and pod walls.
Peaks: *Glyma.12G229400* at ZT4 and ZT16 in seeds, but at ZT0 and ZT16; *Glyma.13G270400* at ZT in seeds, but ZT0 in pod walls; *Glyma.14G016300* at ZT4 in seeds, but no peak in pod walls; *Glyma.09G102900* at 4 and ZT16 in seeds, but ZT0

and ZT20 in pod walls.

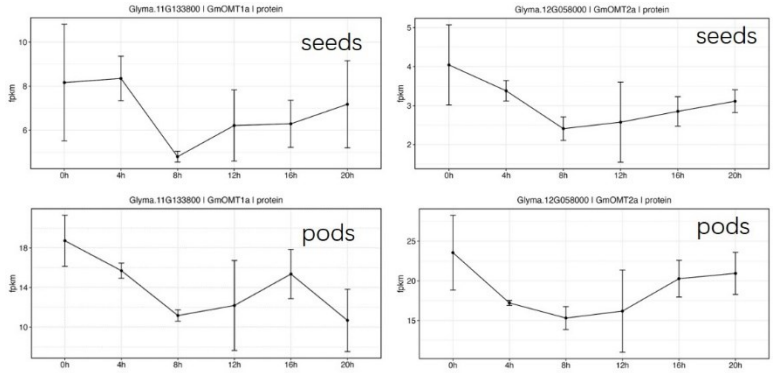


Fig. S10-11 The circadian expression of *OMT1* homologs in seeds and pod walls. *Glyma.11G133800* and *Glyma.12G058000* had peaks at ZT0 in seeds and pod walls.

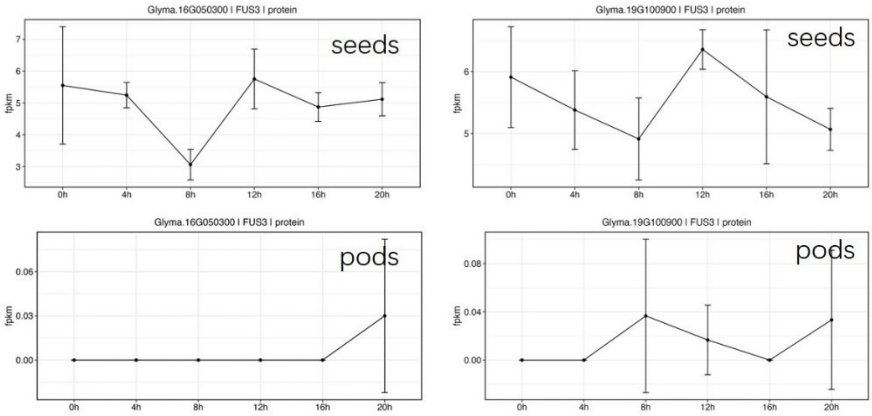


Fig. S10-12 The circadian expression of *FUS3* homologs in seeds and pod walls. *Glyma.19G100900* and *Glyma.16G050300* had peaks at ZT12 in seeds, but no obvious circadian rhythm in pod walls.

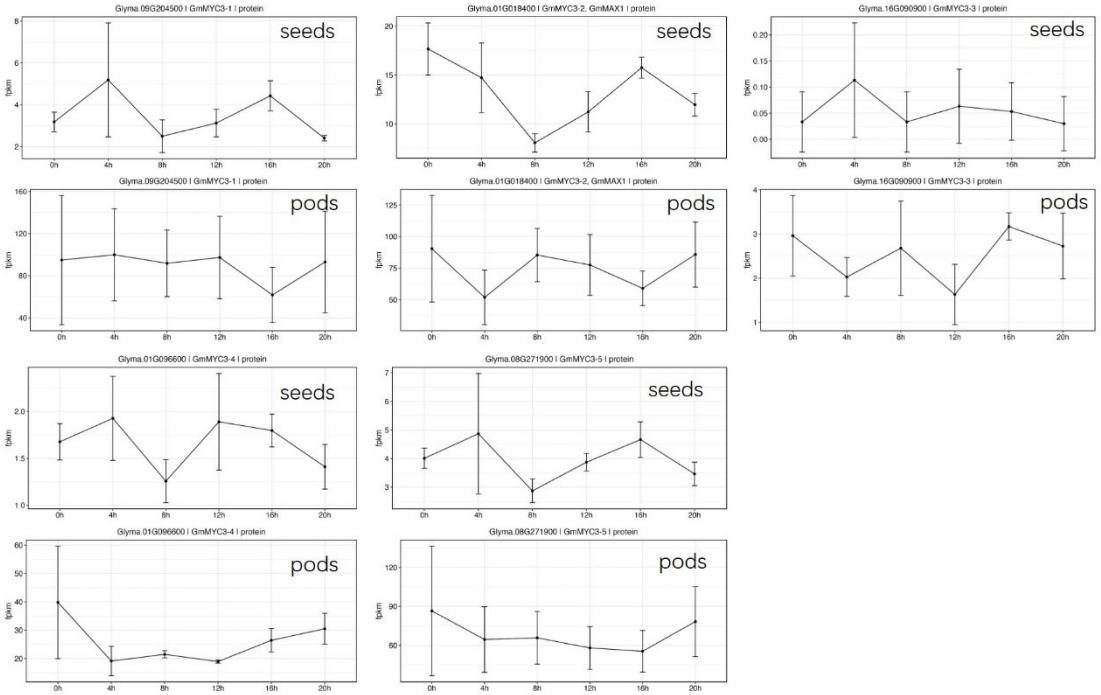


Fig. S10-13 The circadian expression of *MYC3* homologs in seeds and pod walls. *Glyma.09G204500*, *Glyma.16G090900*, and *Glyma.08G271900* had no obvious circadian rhythm in seeds and pod walls. *Glyma.01G018400* peaked at ZT0 and ZT16 in seeds, but may at ZT0 in pod walls. *Glyma.01G096600* peaked at ZT4 and ZT12 in seeds, but at ZT0 in pod walls.

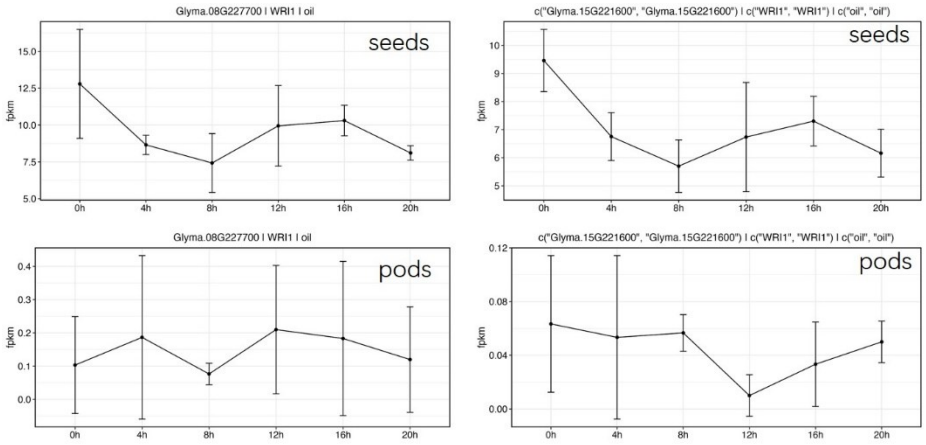


Fig. S10-14 The circadian expression of *WRI1* homologs in seeds and pod walls. *Glyma.08G227700* peaked at ZT0 in seeds, but no circadian rhythm in pod walls. *Glyma.15G221600* peaked at ZT0 in both seeds and pod walls.

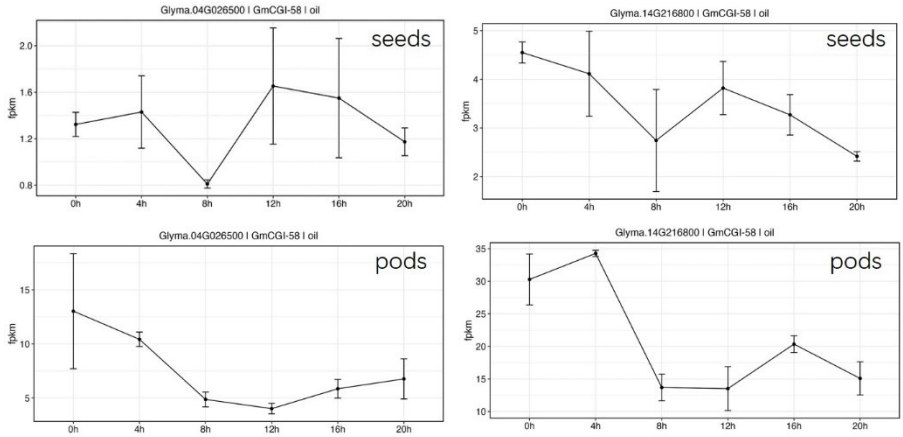


Fig. S10-15 The circadian expression of *CGI-58* homologs in seeds and pod walls. *Glyma.04G026500* had a valley at ZT8 in seeds, but in pod walls it peaked at ZT0 and had relative low level at other time points. *Glyma.14G216800* peaked at ZT0 and ZT12 in seeds, but at ZT4 in pod walls.

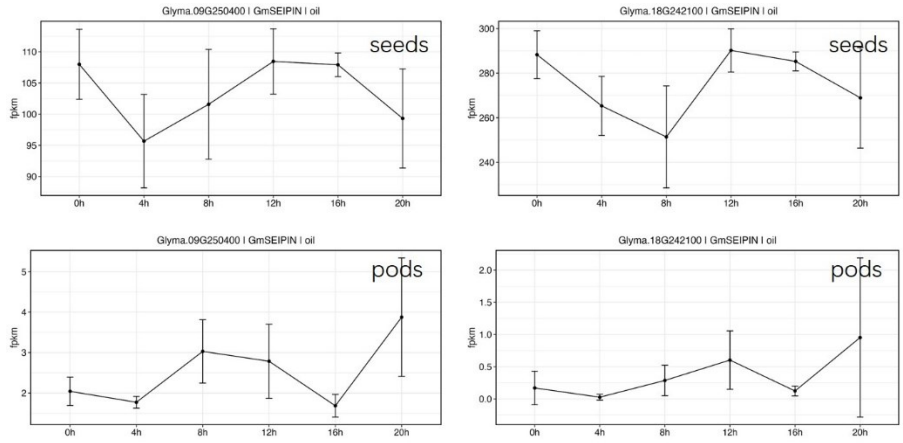


Fig. S10-16 The circadian expression of *SEIPIN* homologs in seeds and pod walls. Peaks: *Glyma.18G242100* at ZT0 and ZT12 in seeds, but at ZT20 in pod walls. *Glyma.09G250400* at ZT0 and ZT12 in seeds, but at ZT20 in pod walls.

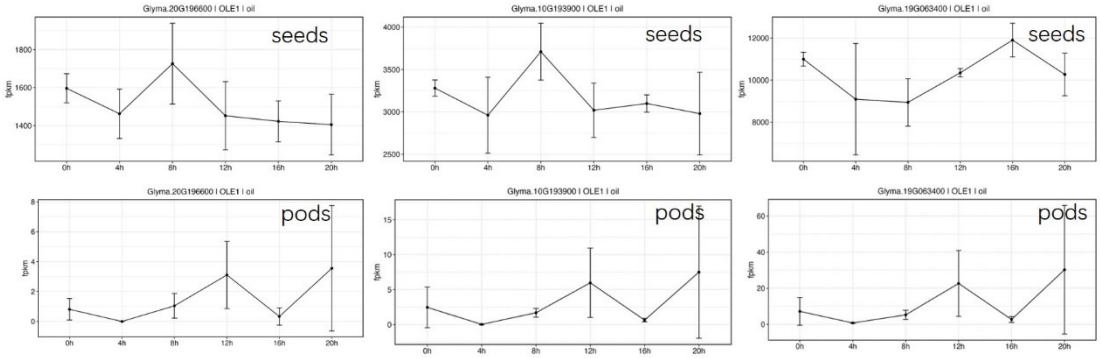


Fig. S10-17 The circadian expression of *OLE1* homologs in seeds and pod walls. Peaks: *Glyma.20G196600* and *Glyma.10G193900* at ZT8 in seeds, but at ZT12 in pods. *Glyma.19G063400* at ZT16 in seeds, but at ZT20 in pod walls.

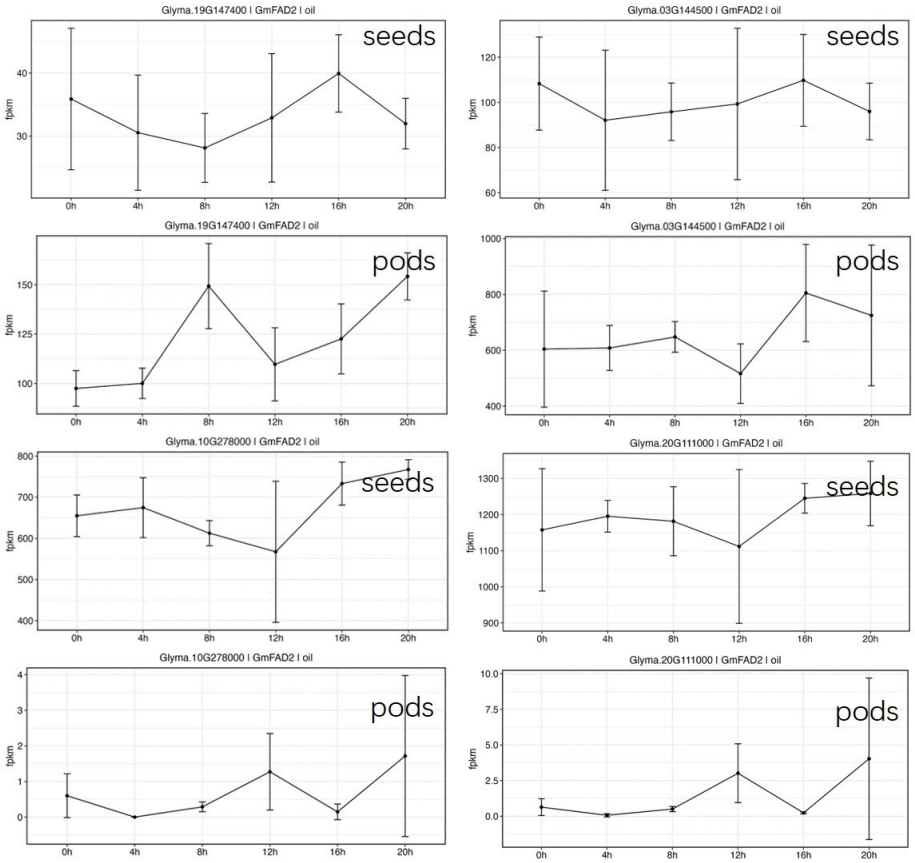


Fig. S10-18 The circadian expression of *FAD2* homologs in seeds and pod walls. *Glyma.19G147400*, *Glyma.03G144500*, *Glyma.10G278000*, and *Glyma.20G111000* peaked at ZT16 in seeds, but ZT8 and ZT16.

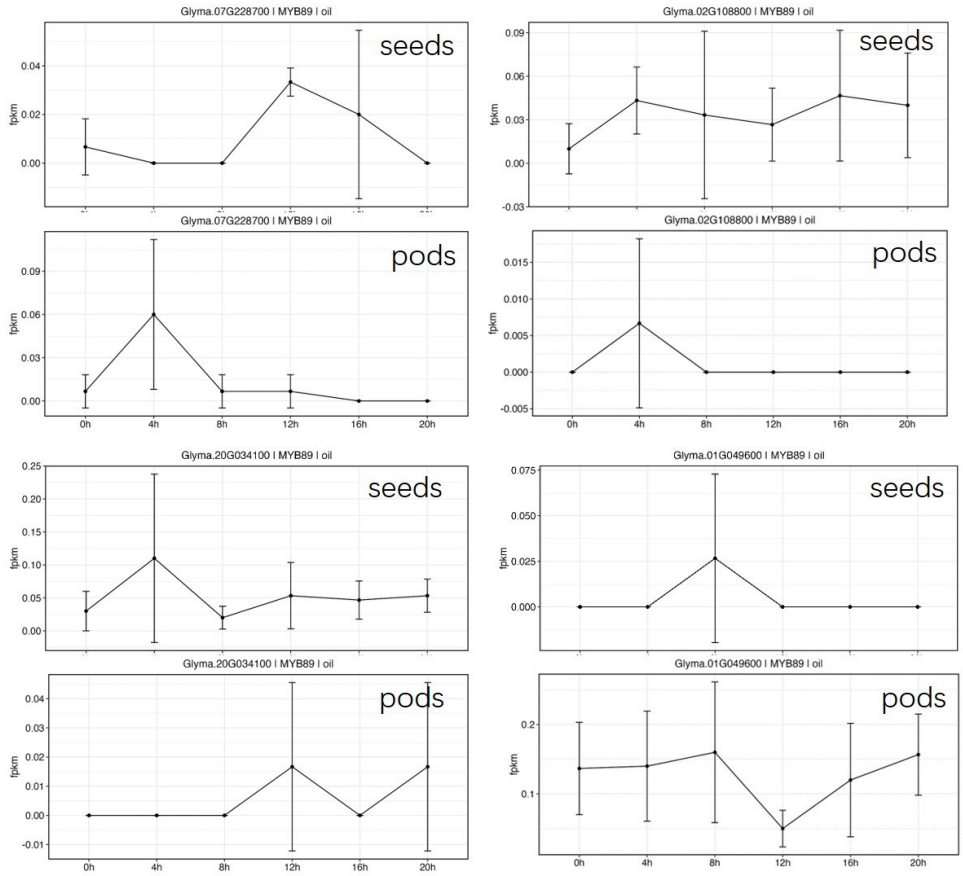


Fig. S10-19 The circadian expression of *MYB89* homologs in seeds and pod walls. Peaks: *Glyma.07G228700*, at ZT12 in seeds but at ZT4 in pod walls; Glyma.02G108800, at ZT4 in both seeds and pod walls; Glyma.20G034100 at ZT4 in seeds, but at ZT12 and ZT20 in pod walls; *Glyma.01G049600* at ZT8 in seeds, but in pod walls had a valley at ZT12.

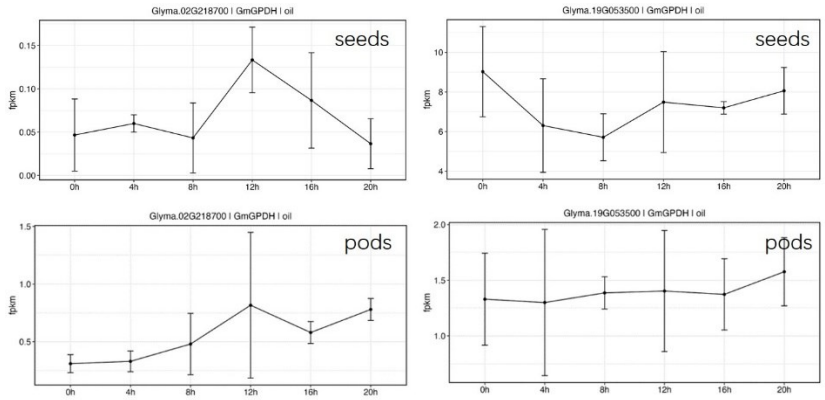


Fig. S10-20 The circadian expression of *GPDH* homologs in seeds and pod walls. *Glyma.02G218700* peaked at ZT 12 in both seeds and pod walls. *Glyma.19G053500* peaked at ZT0 in seeds, but there was no obvious circadian rhythm in pod walls.

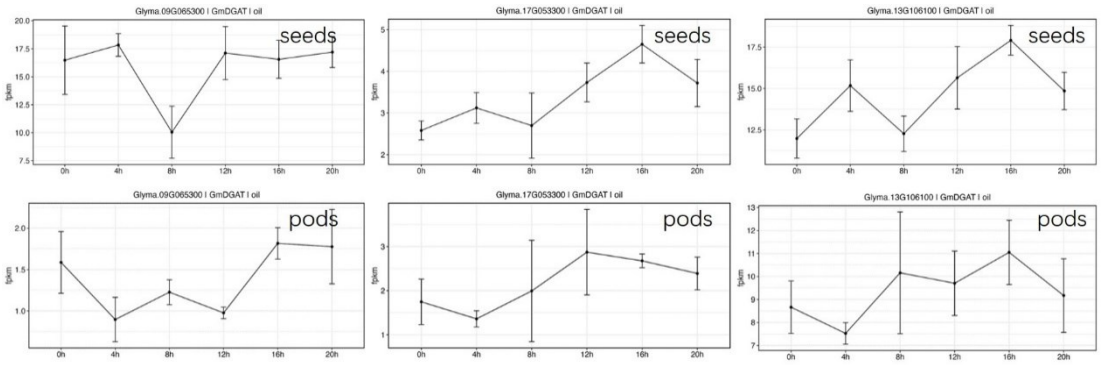


Fig. S10-21 The circadian expression of *DGAT* homologs in seeds and pod walls. *Glyma.09G065300* had a valley in seeds, but kept at low level between ZT4-ZT12 and maintained at high level at other time. *Glyma.17G053300* peak at ZT16 in seeds, but at ZT16 in pod walls. *Glyma.13G106100* peaked at ZT16 in both seeds and pod walls.

Reference：

Bouvier-Nave, P., Benveniste, P., Oelkers, P., Sturley, S.L., and Schaller, H. (2000). Expression in yeast and tobacco of plant cDNAs encoding acyl CoA:diacylglycerol acyltransferase. Eur J Biochem **267,** 85-96.

Chen, M., Zhang, B., Li, C., Kulaveerasingam, H., Chew, F.T., and Yu, H. (2015). TRANSPARENT TESTA GLABRA1 Regulates the Accumulation of Seed Storage Reserves in Arabidopsis. Plant Physiol **169,** 391-402.

Feeney, M., Frigerio, L., Cui, Y., and Menassa, R. (2013). Following vegetative to embryonic cellular changes in leaves of Arabidopsis overexpressing LEAFY COTYLEDON2. Plant Physiol **162,** 1881-1896.

Bouvier-Nave, P., Benveniste, P., Oelkers, P., Sturley, S.L., and Schaller, H. (2000). Expression in yeast and tobacco of plant cDNAs encoding acyl CoA:diacylglycerol acyltransferase. Eur J Biochem **267,** 85-96.

Chen, M., Zhang, B., Li, C., Kulaveerasingam, H., Chew, F.T., and Yu, H. (2015). TRANSPARENT TESTA GLABRA1 Regulates the Accumulation of Seed Storage Reserves in Arabidopsis. Plant Physiol **169,** 391-402.

Feeney, M., Frigerio, L., Cui, Y., and Menassa, R. (2013). Following vegetative to embryonic cellular changes in leaves of Arabidopsis overexpressing LEAFY COTYLEDON2. Plant Physiol **162,** 1881-1896.

Gao, C., Qi, S., Liu, K., Li, D., Jin, C., Li, Z., Huang, G., Hai, J., Zhang, M., and Chen, M. (2016). MYC2, MYC3, and MYC4 function redundantly in seed storage protein accumulation in Arabidopsis. Plant physiology and biochemistry : PPB **108,** 63-70.

James, C.N., Horn, P.J., Case, C.R., Gidda, S.K., Zhang, D., Mullen, R.T., Dyer, J.M., Anderson, R.G., and Chapman, K.D. (2010). Disruption of the Arabidopsis CGI-58 homologue produces Chanarin-Dorfman-like lipid droplet accumulation in plants. Proc Natl Acad Sci U S A **107,** 17833-17838.

Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., Yamamoto, A., and Hattori, T. (2005). Indirect ABA-dependent regulation of

seed storage protein genes by FUSCA3 transcription factor in Arabidopsis. Plant Cell Physiol **46,** 300-311.

Kanai, M., Mano, S., Kondo, M., Hayashi, M., and Nishimura, M. (2016). Extension of oil biosynthesis during the mid-phase of seed development enhances oil content in Arabidopsis seeds. Plant biotechnology journal **14,** 1241-1250.

Kong, Q., Yuan, L., and Ma, W. (2019). WRINKLED1, a "Master Regulator" in Transcriptional Control of Plant Oil Biosynthesis. Plants-Basel **8**.

Luo, G., Shen, L., Song, Y., Yu, K., Ji, J., Zhang, C., Yang, W., Li, X., Sun, J., Zhan, K., Cui, D., Wang, Y., Gao, C., Liu, D., and Zhang, A. (2021). The MYB family transcription factor TuODORANT1 from Triticum urartu and the homolog TaODORANT1 from Triticum aestivum inhibit seed storage protein synthesis in wheat. Plant biotechnology journal **19,** 1863-1877.

Meyer, K., Stecca, K.L., Ewell-Hicks, K., Allen, S.M., and Everard, J.D. (2012). Oil and Protein Accumulation in Developing Seeds Is Influenced by the Expression of a Cytosolic Pyrophosphatase in Arabidopsis. Plant Physiology **159,** 1221-1234.

Ortiz, R., Geleta, M., Gustafsson, C., Lager, I., Hofvander, P., Lofstedt, C., Cahoon, E.B., Minina, E., Bozhkov, P., and Stymne, S. (2020). Oil crops for the future. Curr Opin Plant Biol **56,** 181-189.

Riebeseel, E., Hausler, R.E., Radchuk, R., Meitzel, T., Hajirezaei, M.R., Emery, R.J., Kuster, H., Nunes-Nesi, A., Fernie, A.R., Weschke, W., and Weber, H. (2010). The 2-oxoglutarate/malate translocator mediates amino acid and storage protein biosynthesis in pea embryos. Plant J **61,** 350-363.

Shimada, T., Watanabe, E., Tamura, K., Hayashi, Y., Nishimura, M., and Hara-Nishimura, I. (2002). A vacuolar sorting receptor PV72 on the membrane of vesicles that accumulate precursors of seed storage proteins (PAC vesicles). Plant Cell Physiol **43,** 1086-1095.

Vanhercke, T., Dyer, J.M., Mullen, R.T., Kilaru, A., Rahman, M.M., Petrie, J.R., Green, A.G., Yurchenko, O., and Singh, S.P. (2019). Metabolic engineering for enhanced oil in biomass. Prog Lipid Res **74,** 103-129.

Verdier, J., and Thompson, R.D. (2008). Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. Plant Cell Physiol **49,** 1263-1271.

Wood, C.C., Okada, S., Taylor, M.C., Menon, A., Mathew, A., Cullerne, D., Stephen, S.J., Allen, R.S., Zhou, X.R., Liu, Q., Oakeshott, J.G., Singh, S.P., and Green, A.G. (2018). Seed-specific RNAi in safflower generates a superhigh oleic oil with extended oxidative stability. Plant biotechnology journal **16,** 1788-1796.

Yamamoto, N., Sugimoto, T., Takano, T., Sasou, A., Morita, S., Yano, K., and Masumura, T. (2020). The plant-type phosphoenolpyruvate carboxylase Gmppc2 is developmentally induced in immature soy seeds at the late maturation stage: a potential protein biomarker for seed chemical composition. Biosci Biotech Bioch **84,** 552-562.

Yang, H., Gu, X., Ding, M., Lu, W., and Lu, D. (2018). Heat stress during grain filling

affects activities of enzymes involved in grain protein and starch synthesis in waxy maize. Scientific reports **8,** 15665.

Yeap, W.C., Lee, F.C., Shabari Shan, D.K., Musa, H., Appleton, D.R., and Kulaveerasingam, H. (2017). WRI1-1, ABI5, NF-YA3 and NF-YC2 increase oil biosynthesis in coordination with hormonal signaling during fruit development in oil palm. Plant J **91,** 97-113.

Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M.L., and An, Y.C. (2020). Selection of GmSWEET39 for oil and protein improvement in soybean. PLoS genetics **16,** e1009114.

James, C.N., Horn, P.J., Case, C.R., Gidda, S.K., Zhang, D., Mullen, R.T., Dyer, J.M., Anderson, R.G., and Chapman, K.D. (2010). Disruption of the Arabidopsis CGI-58 homologue produces Chanarin-Dorfman-like lipid droplet accumulation in plants. Proc Natl Acad Sci U S A **107,** 17833-17838.

Kagaya, Y., Okuda, R., Ban, A., Toyoshima, R., Tsutsumida, K., Usui, H., Yamamoto, A., and Hattori, T. (2005). Indirect ABA-dependent regulation of seed storage protein genes by FUSCA3 transcription factor in Arabidopsis. Plant Cell Physiol **46,** 300-311.

Kanai, M., Mano, S., Kondo, M., Hayashi, M., and Nishimura, M. (2016). Extension of oil biosynthesis during the mid-phase of seed development enhances oil content in Arabidopsis seeds. Plant biotechnology journal **14,** 1241-1250.

Kong, Q., Yuan, L., and Ma, W. (2019). WRINKLED1, a "Master Regulator" in Transcriptional Control of Plant Oil Biosynthesis. Plants-Basel **8**.

Luo, G., Shen, L., Song, Y., Yu, K., Ji, J., Zhang, C., Yang, W., Li, X., Sun, J., Zhan, K., Cui, D., Wang, Y., Gao, C., Liu, D., and Zhang, A. (2021). The MYB family transcription factor TuODORANT1 from Triticum urartu and the homolog TaODORANT1 from Triticum aestivum inhibit seed storage protein synthesis in wheat. Plant biotechnology journal **19,** 1863-1877.

Meyer, K., Stecca, K.L., Ewell-Hicks, K., Allen, S.M., and Everard, J.D. (2012). Oil and Protein Accumulation in Developing Seeds Is Influenced by the Expression of a Cytosolic Pyrophosphatase in Arabidopsis. Plant Physiology **159,** 1221-1234.

Ortiz, R., Geleta, M., Gustafsson, C., Lager, I., Hofvander, P., Lofstedt, C., Cahoon, E.B., Minina, E., Bozhkov, P., and Stymne, S. (2020). Oil crops for the future. Curr Opin Plant Biol **56,** 181-189.

Riebeseel, E., Hausler, R.E., Radchuk, R., Meitzel, T., Hajirezaei, M.R., Emery, R.J., Kuster, H., Nunes-Nesi, A., Fernie, A.R., Weschke, W., and Weber, H. (2010). The 2-oxoglutarate/malate translocator mediates amino acid and storage protein biosynthesis in pea embryos. Plant J **61,** 350-363.

Shimada, T., Watanabe, E., Tamura, K., Hayashi, Y., Nishimura, M., and Hara-Nishimura, I. (2002). A vacuolar sorting receptor PV72 on the membrane of vesicles that accumulate precursors of seed storage proteins (PAC vesicles). Plant Cell Physiol **43,** 1086-1095.

Vanhercke, T., Dyer, J.M., Mullen, R.T., Kilaru, A., Rahman, M.M., Petrie, J.R.,

Green, A.G., Yurchenko, O., and Singh, S.P. (2019). Metabolic engineering for enhanced oil in biomass. Prog Lipid Res **74,** 103-129.

Verdier, J., and Thompson, R.D. (2008). Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. Plant Cell Physiol **49,** 1263-1271.

Wood, C.C., Okada, S., Taylor, M.C., Menon, A., Mathew, A., Cullerne, D., Stephen, S.J., Allen, R.S., Zhou, X.R., Liu, Q., Oakeshott, J.G., Singh, S.P., and Green, A.G. (2018). Seed-specific RNAi in safflower generates a superhigh oleic oil with extended oxidative stability. Plant biotechnology journal **16,** 1788-1796.

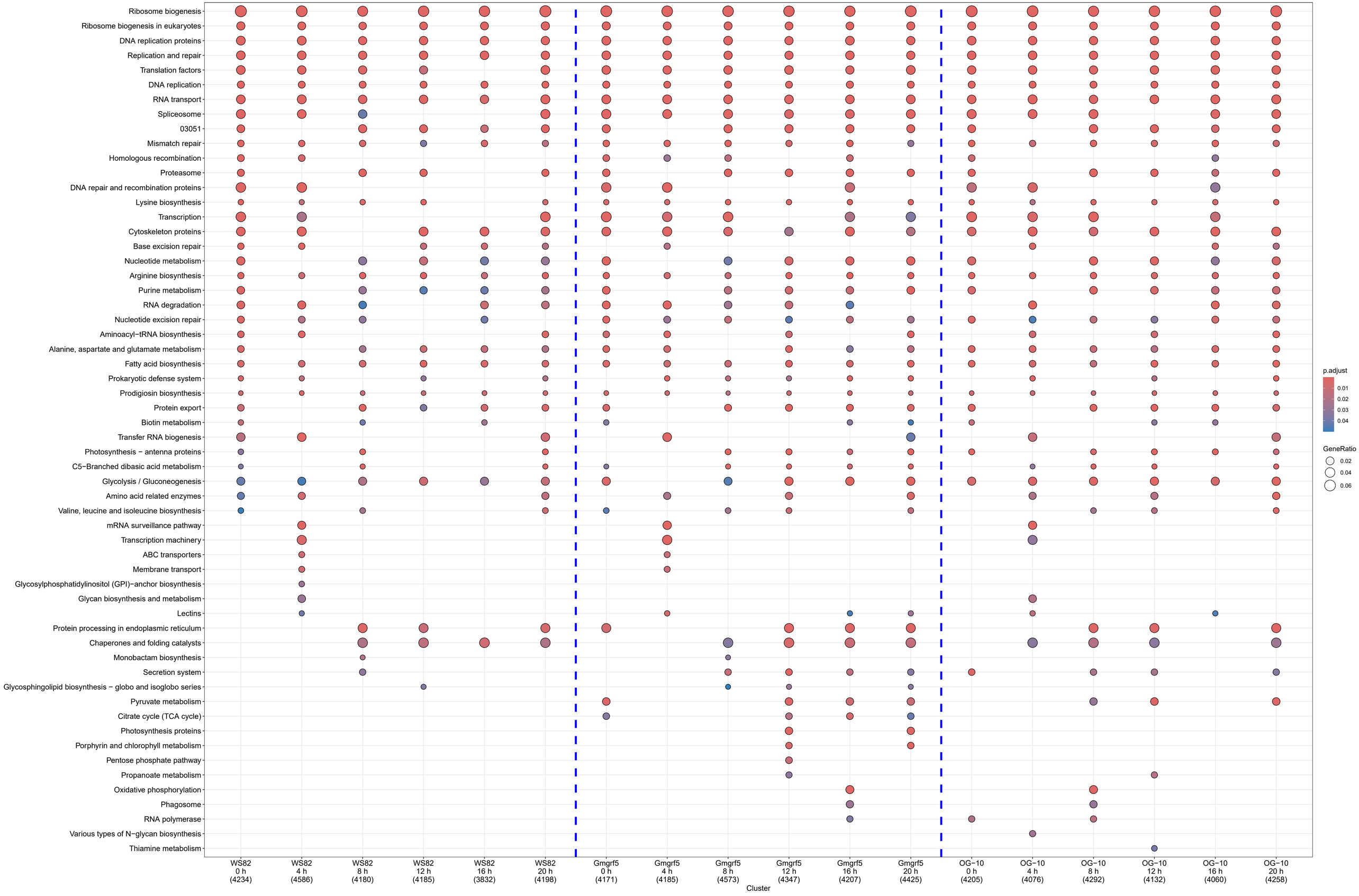
Yamamoto, N., Sugimoto, T., Takano, T., Sasou, A., Morita, S., Yano, K., and Masumura, T. (2020). The plant-type phosphoenolpyruvate carboxylase Gmppc2 is developmentally induced in immature soy seeds at the late maturation stage: a potential protein biomarker for seed chemical composition. Biosci Biotech Bioch **84,** 552-562.

Yang, H., Gu, X., Ding, M., Lu, W., and Lu, D. (2018). Heat stress during grain filling affects activities of enzymes involved in grain protein and starch synthesis in waxy maize. Scientific reports **8,** 15665.

Yeap, W.C., Lee, F.C., Shabari Shan, D.K., Musa, H., Appleton, D.R., and Kulaveerasingam, H. (2017). WRI1-1, ABI5, NF-YA3 and NF-YC2 increase oil biosynthesis in coordination with hormonal signaling during fruit development in oil palm. Plant J **91,** 97-113.

Zhang, H., Goettel, W., Song, Q., Jiang, H., Hu, Z., Wang, M.L., and An, Y.C. (2020). Selection of GmSWEET39 for oil and protein improvement in soybean. PLoS genetics **16,** e1009114.

Fig. S11 A merged figure of the top 50 enriched KEGG pathway in transcriptome of WS82, Gmgrf5 mutant, and OG-10 (up-regulated in seeds compared to pod walls)



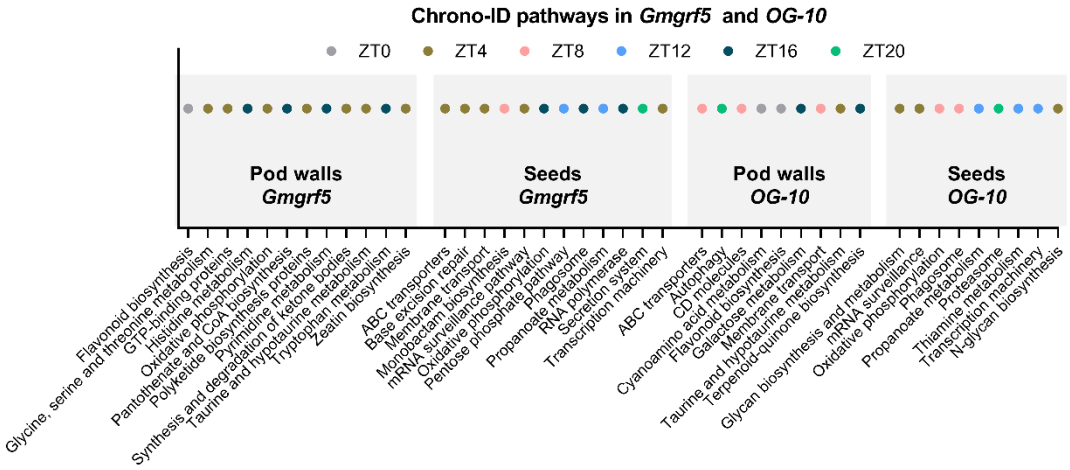
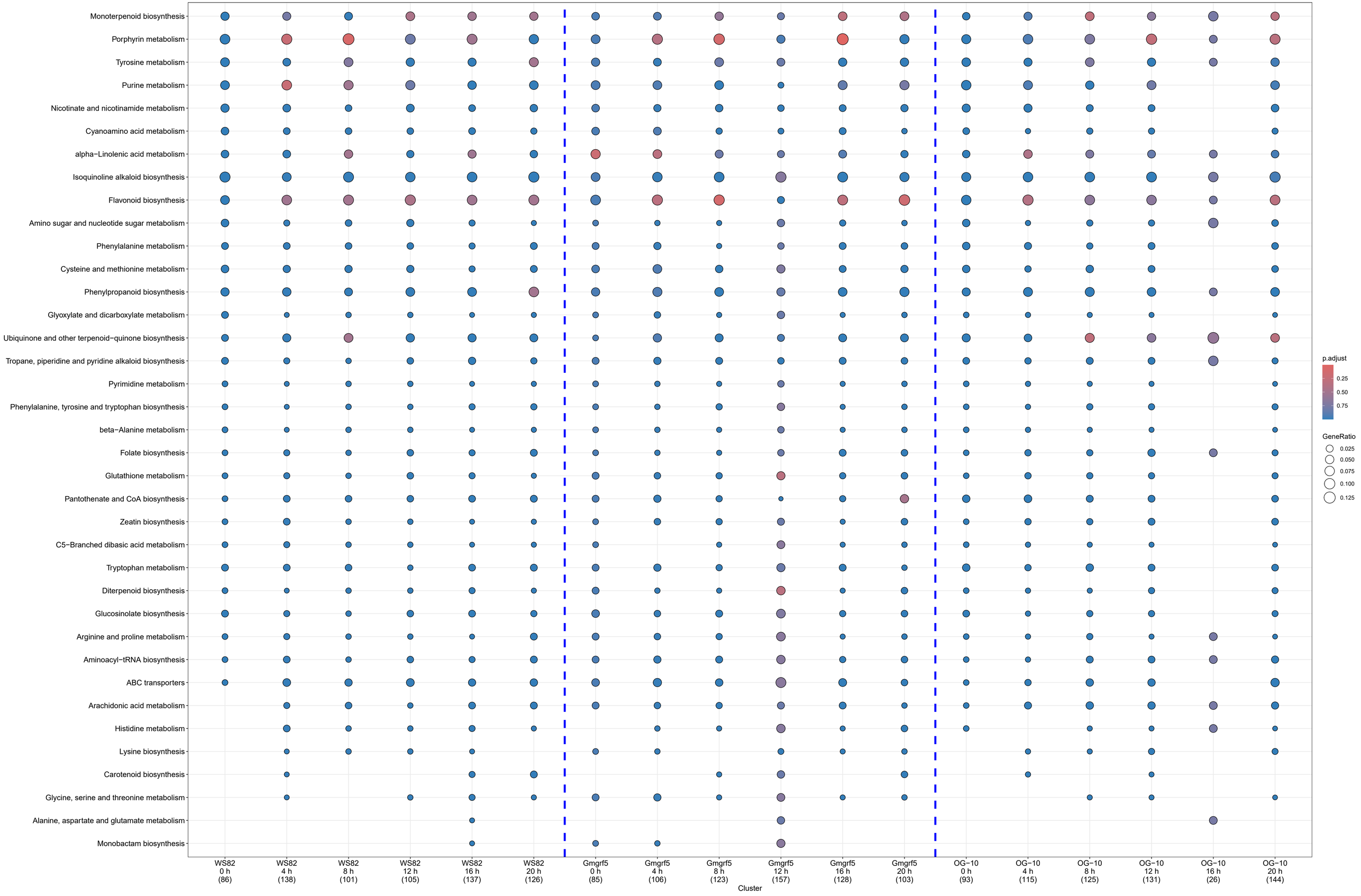


Fig. S12 Chrono-ID pathways in the *Gmgrf5* mutant and a *GmGRF5a*-overexpressing line *OG-10*.

Fig. S13 A merged figure of the top 50 enriched KEGG pathways in metabolome of WS82, Gmgrf5 mutant and OG-10 (up-regulated in seeds compared to pod walls)



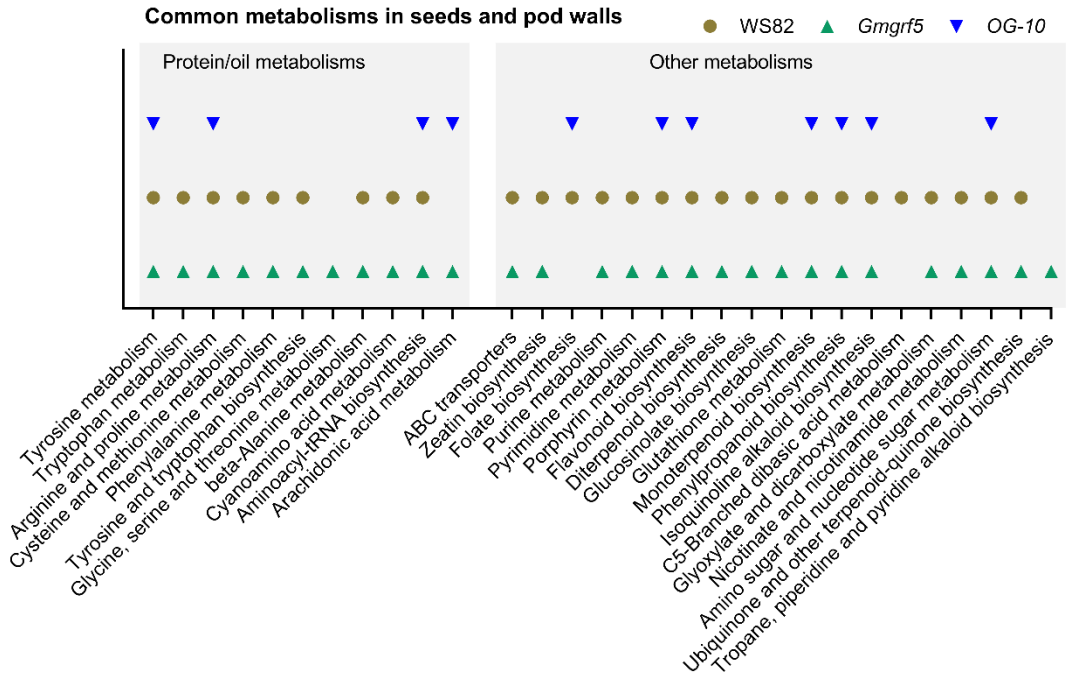


Fig. S14 Common metabolism pathways found in seeds and pod walls. Protein and oil relative metabolisms, including tyrosine, tryptophan, arginine, proline, cysteine, methionine, phenylamine, beta-alanine, and cyanoamino acid metabolism, and aminonacyl-tRNA biosynthesis were found in both seeds and pod walls of WS82 plants, but *GmGRF5* mutation or overexpression led to different changes. Seeds and pod walls also shared some other metabolism pathways including metabolisms of secondary metabolites, transporters, hormones, nucleotides, but *GmGRF5* also disturbed them.

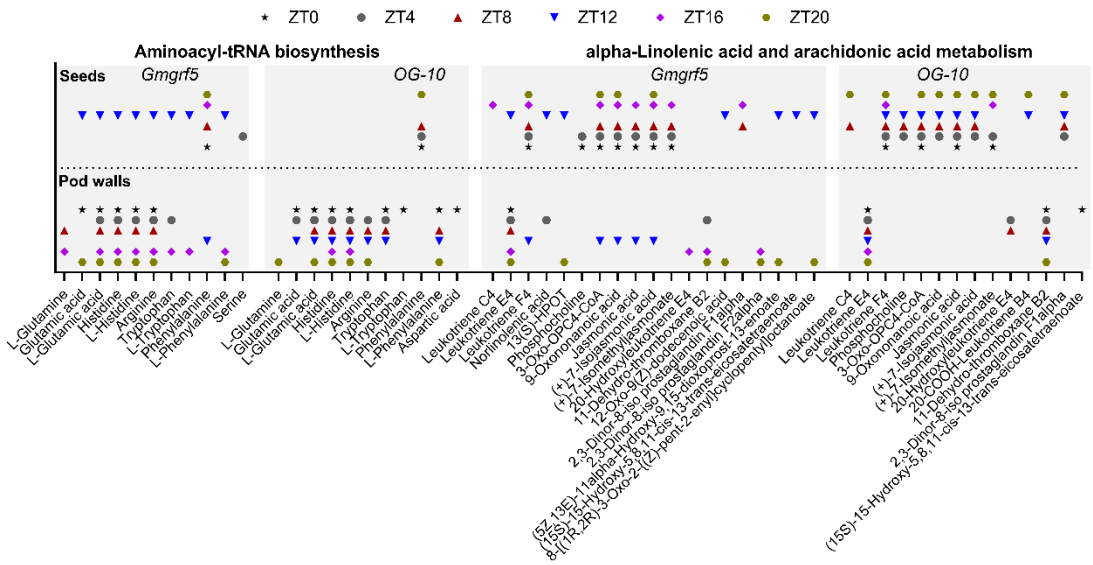


Fig. S15 Chrono-enrichment of amino acids and lipid acid related compounds in seeds and pod walls in *Gmgrf5* and *OG-10*. L-Amino acids accumulated in pod walls in different ZTs except of ZT12, while lipid acid related compounds accumulated in seeds at different ZTs of *Gmgrf5* mutant and *OG-10* overexpressing line as in WS82 (Fig. 3E). Distinguished from WS82, *Gmgrf5* mutant seeds also accumulated L-amino acids at ZT12, but no L-amino acids found in *OG-10* seeds.

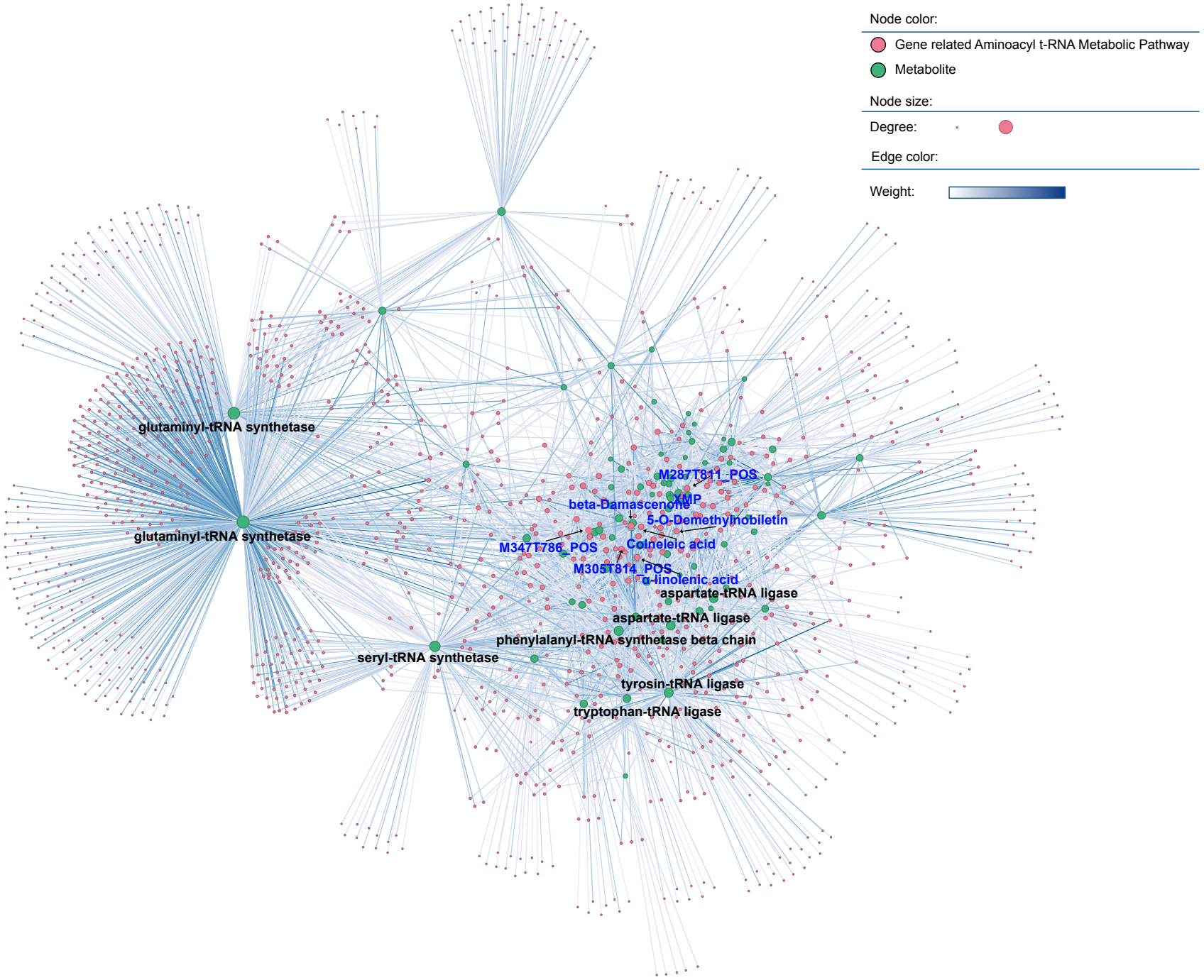

Node color:
Gene related Aminoacyl t-RNA Metabolic Pathway
Metabolite
Node size:
Degree:
Edge color:
Weight:
glutaminyl-tRNA synthetase
glutaminyl-tRNA synthetase
seryl-tRNA synthetase
M287T811_POS
XMP
beta-Damascenone
5-O-Demethylnobiletin
M347T786_POS
Colneleic acid
M305T814_POS
α-linolenic acid
aspartate-tRNA ligase
aspartate-tRNA ligase
phenylalanyl-tRNA synthetase beta chain
tyrosin-tRNA ligase
tryptophan-tRNA ligase

Figure S16 Correlation network of aminoacyl-tRNA synthetase pathway genes and metabolites. In the network, salmon-colored nodes represent genes, while green nodes represent metabolites. Node size reflects the number of edges (degree) connected to corresponding node, meaning larger nodes indicate a greater number of significant correlations with either metabolites or genes. The color of the edges represents the weight, indicating the strength of the correlation between each metabolite and gene, with darker colors denoting stronger correlations. The top eight genes and top eight metabolites with the most connections are labeled; genes are marked with black labels, while metabolites are labeled in blue.

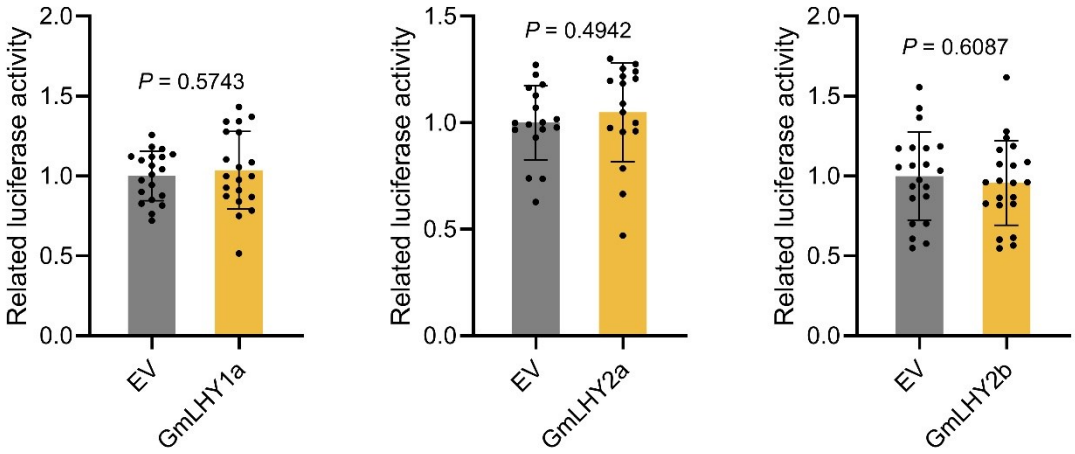


Fig. S17 *GmLHY1a*, *GmLHY2a*, and *GmLHY2b* did not activate *GmGRF5a* promoter. Single luciferase reporter assay showing that *GmLHY1a*, *GmLHY2a*, and *GmLHY2b* did not induce the transcriptions driven by *GmGRF5a* promoter in *Nicotiana benthamiana* leaves. The empty vector pSoy2-GFP was used as a control. Values are means ± SD (n = 20). Statistical analysis was performed using the two-sided Student's *t*-test.

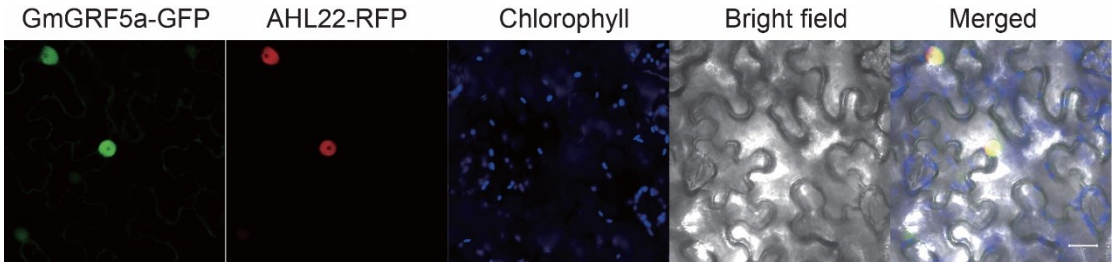


Fig. S18 GmGRF5a proteins localized in nuclei. Subcellular localization of GmGRF5a:GFP fusion proteins in *Nicotiana benthamiana* leaves. AHL22-RFP was used as a marker for nuclear localization. Scale bars, 20 μm.

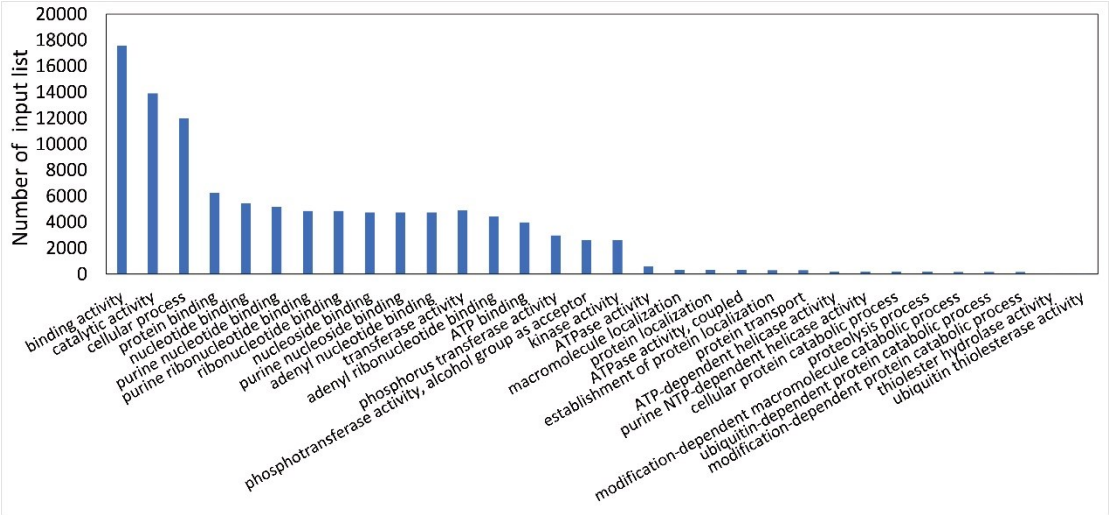


Fig. S19 KEGG analysis of DAP-seq target genes. Results showed that target genes of GmGRF5a include wide functional genes.

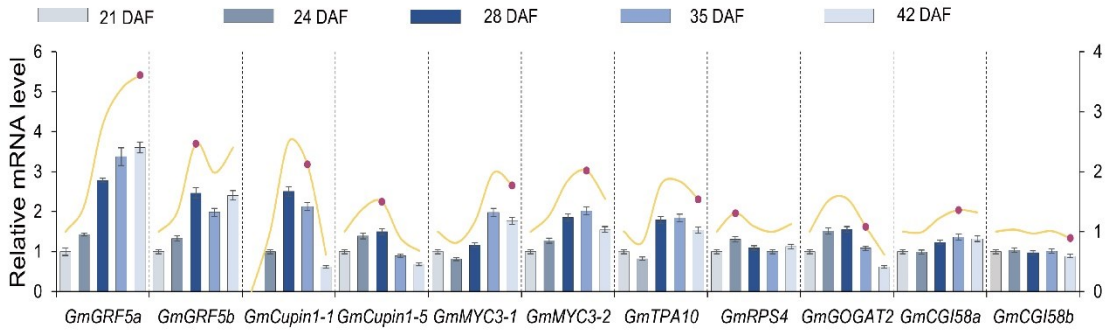


Fig. S20 The seed developmental expression patterns of GmGRF5a, GmGRF5b, and genes related to protein and oil synthesis. DAF, days after flowering. Red dots on trend lines indicate the checking stages for different genes in Fig. 5F. Left Y-axis for columns, and right Y-axis for curve lines. The curve line figure shares the same set of data with the column figure and shows the trend in gene expression.

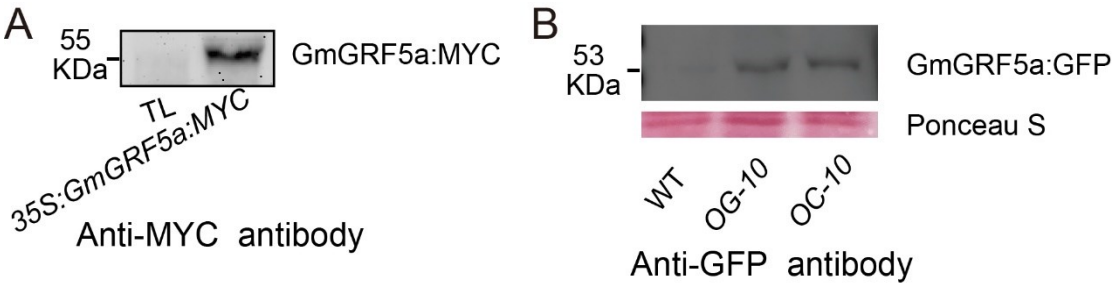


Fig. S21 Identification of transgenic lines by Western blots. A. Immunoblots showing the level of GmGRF5a:6×MYC fused proteins in wild type plants (TL) and *35S:GmGRF5a:MYC* overexpressing line. Nuclear proteins were extracted from 21 DAE leaves (the first fully-opened trifoliate leaves) of soybean plants grown in short days (8 h light/16 h dark) and immunoprecipitated with and probed by an anti-MYC antibody. **B.** Immunoblots showing the level of GmGRF5a:GFP fused proteins in WS82, *OG-10*, and *OC-10* plants. Nuclear proteins were extracted from 21 DAE leaves (the first fully-opened trifoliate leaves) of soybean plants grown in short days (8 h light/16 h dark) and immunoprecipitated with and probed by an anti-GFP antibody. Ponceau S was used as the loading control.

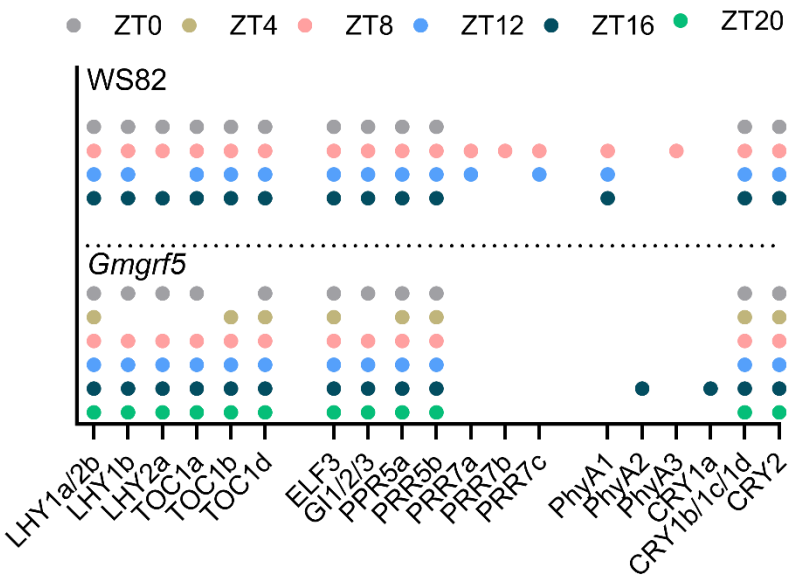


Fig. S22 Circadian expression of clock genes in pod walls of WS82 and *Gmgrf5* mutant. *GmGRF5* mutation results in expression of clock genes at ZT4 and ZT20. The expression of red light and blue light receptors (*PhyA*s and *CRY*s) is also changed by *GmGRF5* mutation.

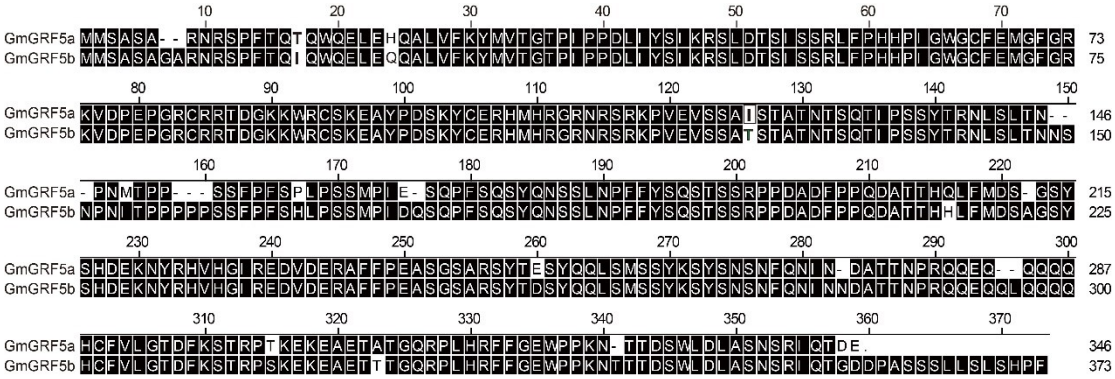


Fig. S23 Alignment of GmGRF5a and GmGRF5b protein sequence.

Fig. S24 The haplotypes of GmGRF5a and GmGRF5b. Five sections (Fig. S24A to S24E), which are haplotype data from five soybean collections and summarized in Fig. 6B.

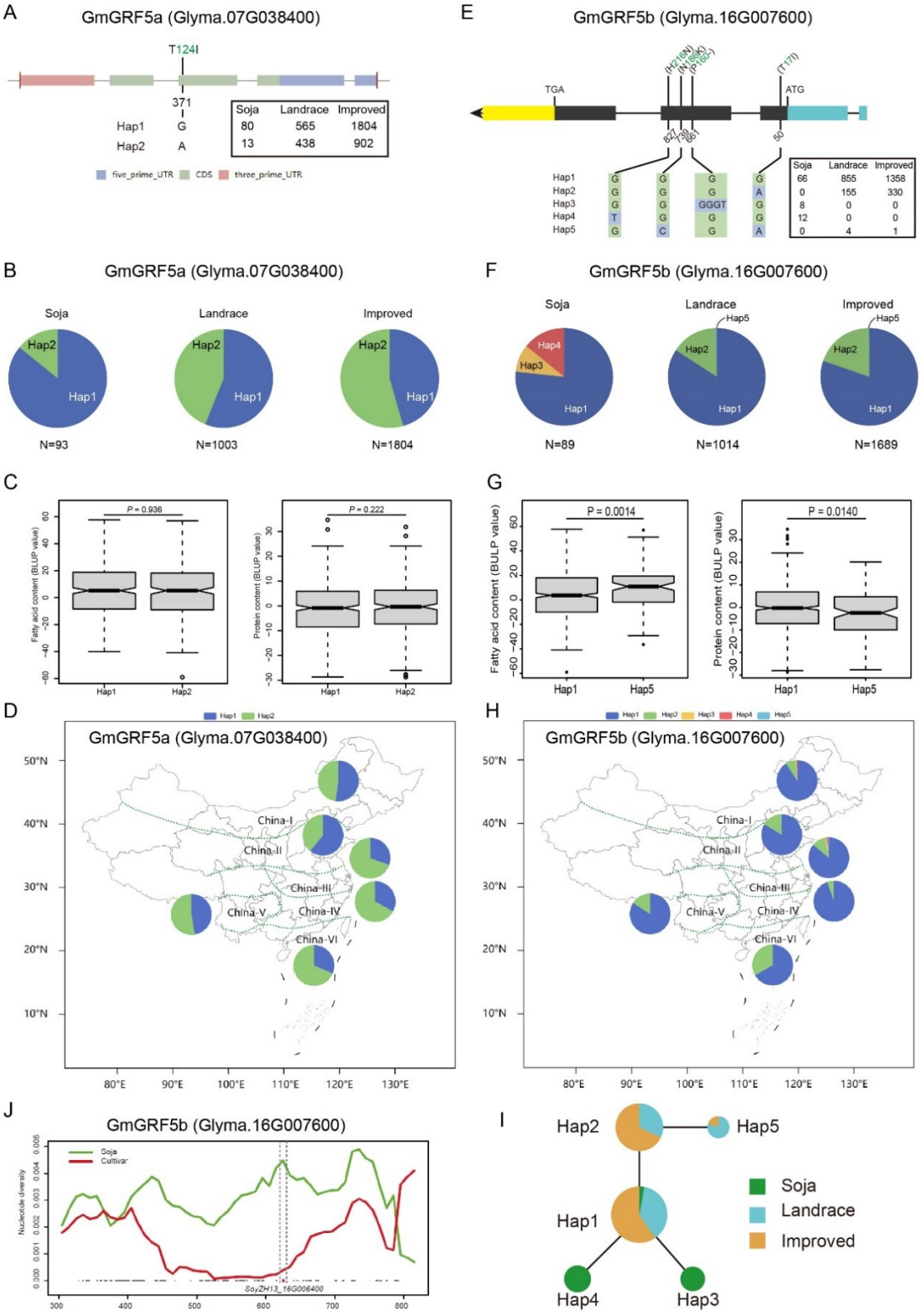


Fig. S24A Domestication analysis of *GmGRF5a* and *GmGRF5b* with Collection I (Fig. 6B). *GmGRF5a* and *GmGRF5b* are orthologous genes in soybean with 93.0%

CDS identity. However, haplotype analysis by missense substitutions of the two genes showed quite different population history. *GmGRF5a* had two haplotypes (**A**) which were shared in wild, landrace and improved soybean, with frequency change of Hap1 from 13.1% to 54.3%, and Hap2 from 86.9% to 45.7% (**B**). Fatty acid and protein contents showed no significant difference between Hap1 and Hap2 (**C**). Although selective signature was not obvious for *GmGRF5a*, it reflected geographic distribution divergence. Soybean from north China (China-I/II) contained higher frequency of Hap2 than south China (China-III/IV/VI) (**D**). The pattern of *GmGRF5b* haplotype was different from that of *GmGRFa*. *GmGRF5b* in wild soybean contained 4 haplotypes, but 3 of which were disappeared during domestication (**E**). Meanwhile, the Hap2 didn't exist in wild but came to a relatively high frequency in landrace (15.3%) and cultivar (19.5%) (**F**). Haplotypes of *GmGRFb* didn't reflect geographic divergence, but had significant difference of fatty acid and protein content (**G** and **H**). Hap2 of *GmGRF5b* was derived from Hap1 and had higher fatty acid and lower protein content, which met the tendency of soybean domestication (**G** and **I**). Selective test by nucleotide diversity (π) showed *GmGRF5b* was embedded in a sharply diversity reduced region between wild and cultivar soybean (**J**). Take together, *GmGRF5b* was putative to be fiercely affected by artificial selection for fatty acid and protein content.

Methods:

Population genetics

We used the variation data from former report of ~3,000 soybean accessions (Liu et al., 2020) to calculate the population genetics parameter F_{ST} and nucleotide diversity (π). For analysis of *GmGRF5a* (*Glyma.07G038400*, *SoyZH13_07G037100*), we picked up samples of landrace and cultivar soybeans and divided them into groups by Eco-regions (I~VI) of soybean of China. We identified I/II as North China and III/IV/VI as South China. Then F_{ST} was calculated by South against North, and π was calculated by each of the region. For analysis of *GmGRF5b* (*Glyma.16G007600*, *SoyZH13_16G006400*), we divided samples into *G. soja* and cultivar soybeans. π was calculated for each of the group. The F_{ST} and π was calculated by *VCFTOOLS* (0.1.16) with parameters: "--fst-window-size 50000 --fst-window-step 10000" and "--window-pi 50000 --window-pi-step 10000".

Haplotype analysis

Haplotype analysis of *GmGRF5a* (*Glyma.07G038400* and *SoyZH13_07G037100*) and *GmGRF5b* (*Glyma.16G007600* and *SoyZH13_16G006400*) was done by *HapSnap* toolkit of SoyOmics (https://ngdc.cncb.ac.cn/soyomics). Variations with missense substitution were kept. Variation quality control was set with "MAF ≥ 0.0001". Haplotype network was created by *NETWORK* (10.2.0.0) with median joining model. Fatty acid content and protein content were the BLUP values from SoyOmics. Significant test for haplotypes of single gene was done by Student's *t*-test. Multiple comparison for combined haplotypes from the two genes was done by LSD test.

Reference:

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., Huang, X., Li, Y., Zhang, M., Wang, Z., Zhu, B., Han, B., Liang, C., and Tian, Z. (2020). Pan-Genome of Wild and Cultivated Soybeans. Cell **182,** 162-176 e113.

A1

Glyma.07G038400										Number
Hap1	C	G	G	G	T	C	C	T	G	258
Hap2	C	G	G	G	A	T	A	C	A	68
Hap3	A	A	T	A	A	T	A	C	A	34
CDS Position in Wm82	-970	-733	-717	-223	-164	-137	-54	371	759	
Amino Acid Position in Wm82	Promoter							124	253	
Amino Acid in Hap1								1	Q	
Amino Acid in Hap2 and Hap3								T	Q	

A2

Glyma.16G007600										Number
Hap1	G	T	A	T	C	C	T	A	T	58
Hap2	A	A	G	C	T	G	G	C	C	46
Hap3	A	A	G	C	T	G	T	C	C	29
CDS Position in Wm82	-2453	-2074	-1893	-1768	-1685	-1629	-1317	-551	50	
Amino Acid Position in Wm82	Promoter								17	
Amino Acid in Hap1									I	
Amino Acid in Hap2 and Hap3									I	

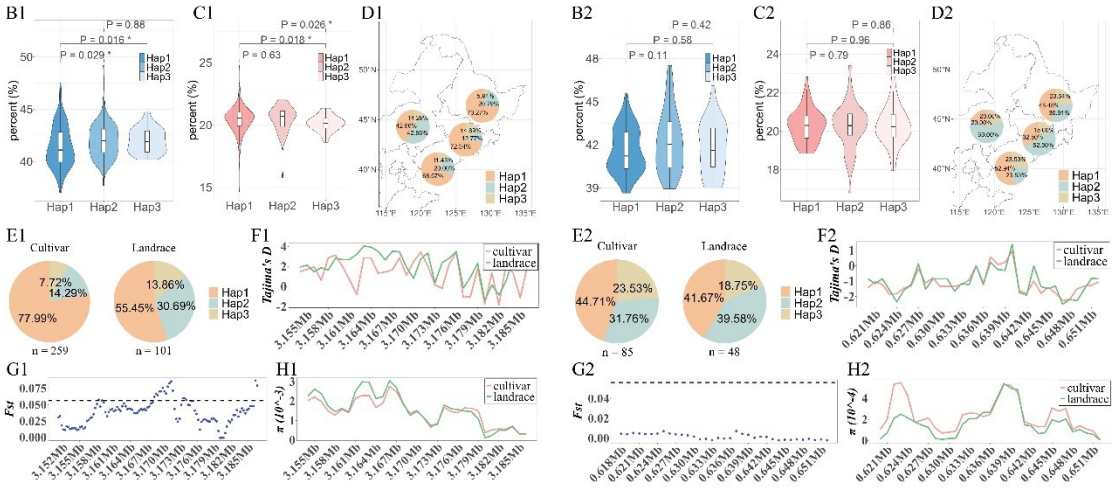


Fig. S24B Dominant haplotype distributions of *GmGRF5a* (Glyma.07g038400) and *GmGRF5b* (Glyma.16G007600) and their correlation with crude oil and crude protein in Collection II (Fig. 6B). **A1** and **B1**, Haplotypes of *Glyma.07g038400* and *Glyma.16G007600*. **B1** and **B2**, Crude protein content. **C1** and **C2**, Crude oil content. **D1** and **D2**, Geographical distribution of dominant haplotypes. **E1** and **E2**, Distribution of dominant haplotypes in cultivars and landraces. **F1** and **F2**, *Tajima's D*, *pi*, and *Fst*. *Tajima's D* (**F1** and **F2**), *Fst* (**G1** and **G2**) and *pi* (**H1** and **H2**) analysis for the SNP diversity of *GmGRF5a* and *GmGRF5b*. Subsequently, leveraging a dataset comprising 547 previously resequenced soybean accessions, including 365 cultivars and 182 landraces, haplotype analyses were conducted for *GmGRF5a* and *GmGRF5b*. A total of three dominant haplotypes from *GmGRF5a* and three from *GmGRF5b* were identified (**A1** and **A2**). To ascertain the effect of these dominant haplotypes, association analysis was performed between the dominant haplotypes and the Best Linear Unbiased Estimates (BLUE) of seed protein and oil content of different accessions, grown in Harbin in 2018, 2019, 2021, and 2022. The results revealed significant differences in seed protein and oil contents among *Glyma.07G038400* dominant haplotypes (**B1** and **C1**), whereas no differences were observed for *Glyma.16G007600* dominant haplotypes (B2 and C2). Soil environment and climatic conditions are crucial factors influencing soybean yield and quality. In the northeastern region of China, there is a spatial distribution pattern with the northern areas having higher soil depth, soil organic matter content, and soil pH values compared to the southern areas [1]. Interestingly, *Glyma.07G038400* dominant haplotypes exhibited systematic variations across germplasm types and regions. For example, $Glyma.07G038400^{Hap1}$ showed a decreased proportion among Heilongjiang-Jilin-Liaoning-Inner Mongolia (**D1**) and a reduced proportion between cultivars and landraces (**E1**). Conversely, $Glyma.07G038400^{Hap2}$ and $Glyma.07G038400^{Hap3}$ displayed an increased proportion between cultivars and landraces (**E1**). In contrast, *Glyma.16G007600* dominant haplotypes did not exhibit consistent patterns in

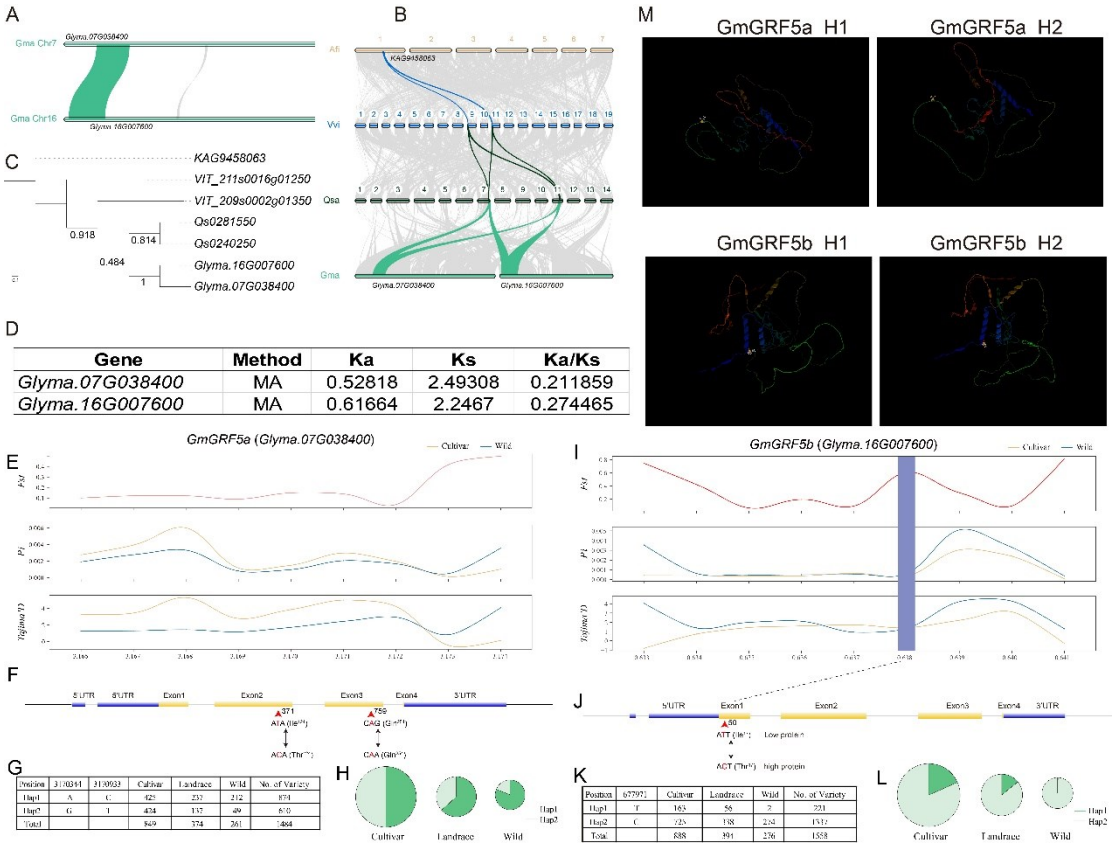
germplasm types or regions (**D2** and **E2**). Furthermore, Select Sweep analysis was conducted for the regions of chromosome 7 (3.155-3.185 Mb) and chromosome 16 (0.618-0.651 Mb). In the region of *Glyma.07G038400*, nucleotide diversity slightly decreased from cultivars to landraces. *Tajima's D* values were positive for both germplasm types, suggesting possible balancing selection. *FST* values significantly exceeded the threshold (*FST* = 0.0561), indicating clear differentiation between cultivars and landraces (**F1**). For Glyma.16G007600's region, nucleotide diversity slightly increased from cultivars to landraces; average *Tajima's D* values approached 0, and *FST* values were close to 0 (**F2**). These findings suggest that *Glyma.07G038400* underwent selection during domestication, while *Glyma.16G007600* did not exhibit discernible signs of selection.

Materials and Methods for Haplotype Analysis and Select Sweep Analysis

Haplotype analysis was conducted on *Glyma.07G038400* and *Glyma.16G007600* in a set of 547 soybean re-sequenced accessions, which following our previously procedure [2]. Haplotypes with a frequency exceeding 5% in the population (Number > 27.35) were identified as dominant haplotypes. The Best Linear Unbiased Estimates (BLUE) for seed oil and protein contents of these soybean accessions, planted in Harbin in 2018, 2019, 2021, and 2022, were determined. Association analysis between the dominant haplotypes and BLUE values was performed using student *t*-tests. Vcftools v0.1.17 was employed to calculate *pi* (window-*pi* 3000, window-*pi*-step 1000), *Tajima's D* (TajimaD 1000), and *FST* (*Fst*-window-size 3000, *Fst*-window-step 1000), with the top 5% of genome-wide *FST* values used as the threshold.

Reference:

1. Li, C., Zhao, Y., Yang, C., Zhang, C., Yun, W., Zhang, F., and Cheng, F. (2023). Characteristics of cultivated land soil conditions in northeast china based on cultivated land resource quality. Chinese Journal of Soil Science *55*, 10.19336/j.cnki.trtb.2023072801. In Chinese. 10.19336/j.cnki.trtb.2023072801.
2. Qi, Z.M., Guo, C.C., Li, H.Y., Qiu, H.M., Li, H., Jong, C., Yu, G.L., Zhang, Y., Hu, L.M., Wu, X.X., et al. (2023). Natural variation in Fatty Acid 9 is a determinant of fatty acid and protein content. Plant biotechnology journal. 10.1111/pbi.14222.



Gene	Method	Ka	Ks	Ka/Ks
Glyma.07G038400	MA	0.52818	2.49308	0.211859
Glyma.16G007600	MA	0.61664	2.2467	0.274465

Fig. S24C Domestication analysis of *GmGRF5a* and *GmGRF5b* with Collection III (Fig 6B). Tracing the evolutionary history of *Glyma.07G038400* (*GmGRF5a*) and *Glyma.16G007600* (*GmGRF5b*) revealed synteny blocks between Chromosome 7 and 16 of *G. max* (**A**). Collinearity analysis was conducted on *G. max*, *Q. sappnaria*, *V. vinifera* and *A. fimbriata* identifying orthologous genes (*Qs0240250*, *Qs0281550*, *VIT_209s002g01350*, *VIT_211s0016g01250* and *KAG9458063*) corresponding to *Glyma.07G038400* and *Glyma.16G007600* (**B**). Phylogenetic analysis of these genes suggested that *Glyma.07G038400* and *Glyma.16G007600* might be duplicated from *α*-WGD event in *G. max* (**C**). To determine the primacy of *Glyma.07G038400* or *Glyma.16G007600*, *Ks* value of these genes and *KAG9458063* were calculated to be 2.49 and 2.24, respectively (**D**). Furthermore, the *Ks* value between these two genes was found to be 0.10. Using the formula (T = Ks/2r), it is suggested that *Glyma.16G007600* emerged from *Glyma.07G038400* via the α event at ~7.84 Mya, based on the *Ks*. The results suggested that *Glyma.07G038400* was the most recent ancestor of *Glyma.16G007600*. To elucidate the evolution of *Glyma.07G038400* and *Glyma.16G007600* during domestication, 36 and 27 SNPs in the upstream and downstream 3-kb regions surrounding *Glyma.07G038400* and *Glyma.16G007600*, respectively, were analyzed to detect the selection signals based on F_{ST}, π and Tajima's D (**E *and* I**). Analysis revealed an absence of selection signals within the regions surrounding *Glyma.07G038400* (**A**). Similarly, no distinct differentiation was discernible among haplotypes across improved cultivars, landrace, and wild soybean (**G** and **H**). However, a distinct selective sweep was evident in the genomic regions of *Glyma.16G007600*, as indicated by all three values (**I**). Subsequent analysis of the

haplotypes of the selected locus in both cultivated and wild soybean revealed an association with protein content. A SNP (T to C) at position 50 in the coding region led to a codon change from ATT (Ile) to ACT (Tre) (**F** and **G**), affecting the QLQ domain crucial for protein-protein interactions (**M** and **Fig. 6**). Predominantly, wild soybean exhibited the H2 haplotype, and it has been documented that wild soybean possess higher protein content than their cultivated soybean [1], suggesting a correlation between H2 and elevated protein content (**G**, **H**, **K**, and **L**) and *Glyma.16G007600* were selected during domestication. **M**, Predicted structures of various haplotypes within the *GmGRF5a* and *GmGRF5b* proteins by AlphaFold2. The resulting monomeric structures were annotated with a color-coded confidence scale, where blue indicates regions of high confidence and red signifies areas of low confidence.

Methods

Collinearity block analyses

Glycine max (soybean) experienced three rounds of whole-genome duplication (WGD): the first (~120 million years (Mya) ago) shared with eudicot (γ event), the second (~59 mya) shared with Fabaceae (β event), and the third (~13 mya) following its divergence from *Phaseolus vulgris* (α event) [2-4]. Consequently, orthologous gene pairs between soybean genome and those of other angiosperms genome were identified through collinearity analysis using the JCVI software package with default settings (https://github.com/tanghaibao/jcvi). For this study, the genome of *Quilaja sappnaria* (the closest phylogeny of Fabaceae) [5], *Vitis vinifera* (undergoing only γ event) [6] and *Aristolochia fimbriata* (lacking WGD event) [7] were selected to analyze the divergence of gene pair *Glyma.07G038400* (*GmGRF5a*) and *Glyma.16G007600* (*GmGRF5b*). Homologous coding sequences were aligned using a codon-based model in MUSCLE, implemented in MEGA11 [8]. Phylogenetic analysis was constructed using the neighbor-joining model method in MEGA11, with 1000 bootstrap replicates for robustness [8]. The *Ks* (synonymous substitution) values of the homologous gene pairs were computed using KaKs_Calculator 3.0 [9]. The divergence time between duplication genes was estimated using the formula T = Ks/2r, where r is the neutral substitution rate ($r = 6.5*10^{-9}$ mutations per site year) [10].

Detection of nucleotide polymorphism and F_{ST}

Estimating the natural variation and F_{ST} of *Glyma.07G038400* (*GmGRF5a*) and *Glyma.16G007600* (*GmGRF5b*) in both cultivated and wild soybeans, we collected 1608 soybean genomes, comprising 121 cultivars, 407 landraces and 280 wild genomes from NCBI SRA database (https://www.ncbi.nlm.nih.gov). Read mapping and variant calling were performed using the standard pipeline methods, as detailed in Zhou et al. [11]. Subsequently, the 3-kb regions upstream and downstream of *Glyma.07G038400* (*GmGRF5a*) and *Glyma.16G007600* (*GmGRF5b*) were analyzed to calculate the nucleotide polymorphism (π), Tajima's D, and F_{ST} using a 1 kb sliding window approach in VCFtools (v0.1.14) [12]. SNPs in CDS regions were subjected to haplotype analysis. Then, AlphaFold2 was utilized to predict the protein structures at sites exhibiting selection signals within exons [13]. Structural model figures were generated

with CLC Sequence Viewer 8.0 software (CLC Bio, Aarhus, Denmark). Additionally, the telomere-to-telomere gap-free assembly of soybean genome [14,15] and the assembly by Yucheng Liu *et al.* [16] were used to extract the sequences of *Glyma.07G038400* (*GmGRF5a*) and *Glyma.16G007600* (*GmGRF5b*), assessing mutation accuracy through the re-sequence data.

Reference:

1. Dong, Y.S., Zhuang, B.C., Zhao, L.M., Sun, H., and He, M.Y. (2001). The genetic diversity of annual wild soybeans grown in China. Theoretical and Applied Genetics *103*, 98-103. DOI 10.1007/s001220000522.
2. Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature *463*, 178-183. 10.1038/nature08670.
3. Lavin, M., Herendeen, P.S., and Wojciechowski, M.F. (2005). Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Systematic Biology *54*, 575-594. 10.1080/10635150590947131.
4. Wu, S., Han, B., and Jiao, Y. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. Molecular Plant *13*, 59-71.
5. Group, T.A.P., Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S., and Stevens, P.F. (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society *181*, 1-20. 10.1111/boj.12385.
6. Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature *449*, 463-467. 10.1038/nature06148.
7. Qin, L., Hu, Y., Wang, J., Wang, X., Zhao, R., Shan, H., Li, K., Xu, P., Wu, H., Yan, X., et al. (2021). Insights into angiosperm evolution, floral development and chemical biosynthesis from the Aristolochia fimbriata genome. Nature Plants *7*, 1239-1253. 10.1038/s41477-021-00990-2.
8. Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution *38*, 3022-3027. 10.1093/molbev/msab120.
9. Zhang, Z. (2022). KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-Coding Sequences. Genomics, Proteomics & Bioinformatics *20*, 536-540. 10.1016/j.gpb.2021.12.002.
10. Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proceedings of the National Academy of Sciences *93*, 10274-10279. doi:10.1073/pnas.93.19.10274.
11. Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y.,

Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology *33*, 408-414. 10.1038/nbt.3096.

12. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158. 10.1093/bioinformatics/btr330.
13. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583-589. 10.1038/s41586-021-03819-2.
14. Wang, L., Zhang, M., Li, M., Jiang, X., Jiao, W., and Song, Q. (2023). A telomere-to-telomere gap-free assembly of soybean genome. Molecular Plant *16*, 1711-1714.
15. Zhang, C., Xie, L., Yu, H., Wang, J., Chen, Q., and Wang, H. (2023). The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. Molecular Plant *16*, 1715-1718.
16. Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., and Shi, M. (2020). Pan-genome of wild and cultivated soybeans. Cell *182*, 162-176. e113.

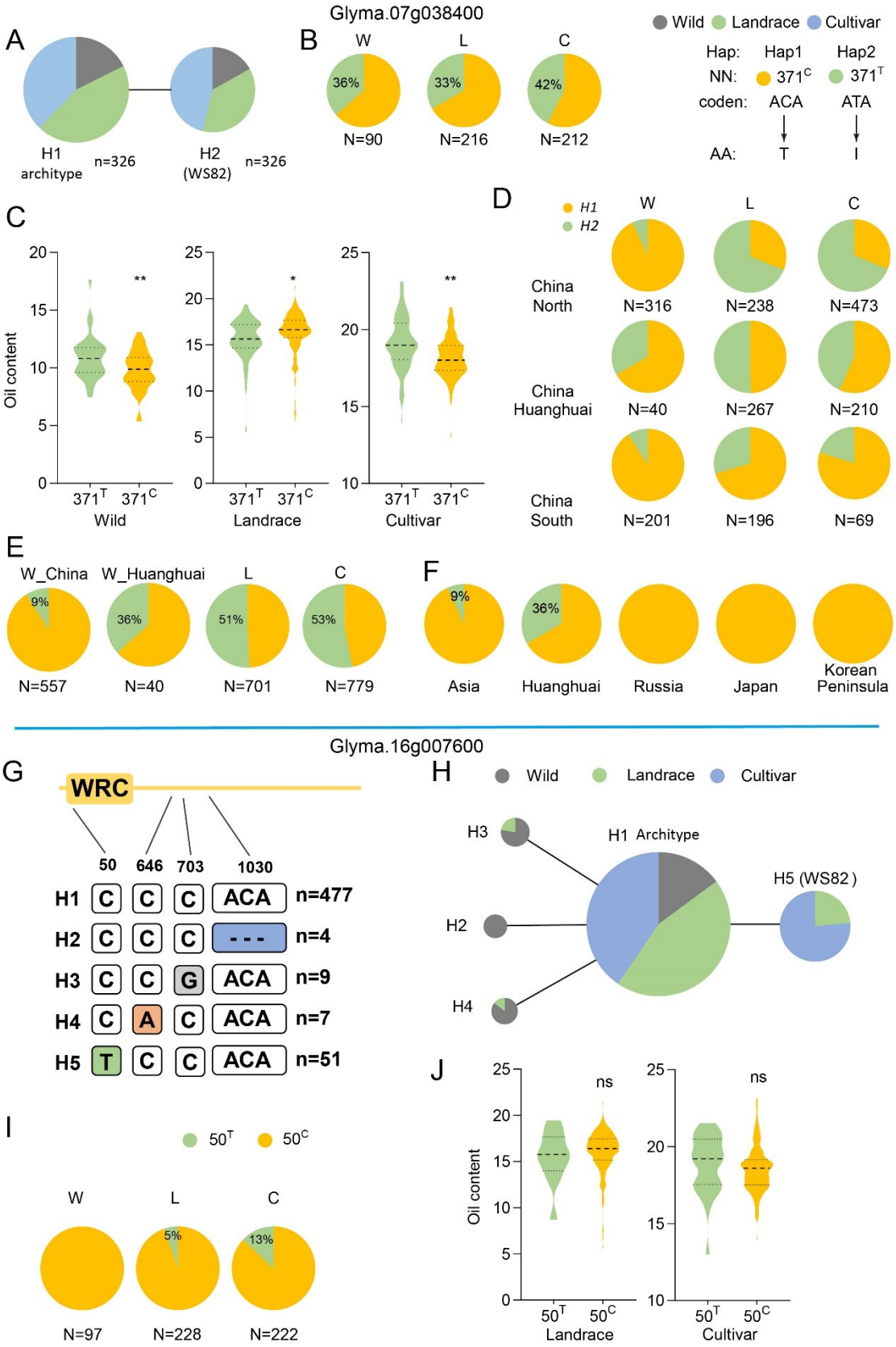


Fig. S24D Domestication analysis of *GmGRF5a* and *GmGRF5b* with Collection IV (Fig 6B). **A**, Haplotype origins of *Glyma.07G038400.* Grey for wild soybean accessions, green for landraces, blue for improved cultivars. **B**, Proportions of

Glyma.07G038400^{H1} and *Glyma.07G038400*^{H2} in the three germplasm groups in China: wild soybeans (W), landraces (L), cultivars (C). **C**, Oil content of different haplotypes of *Glyma.07G038400* in the three germplasm groups in China. * indicates statistically significant determined by Student's *t*-test ($P < 0.05$). **D**, Proportions of *Glyma.07G038400*^{H1} and *Glyma.07G038400*^{H2} in the three germplasm groups in China. W_China for wild soybean in the whole China, W_Huanghuai for wild soybean in Huanghuai area. **E**, Geographical proportions of *Glyma.07G038400*^{H1} and *Glyma.07G038400*^{H2} within each of the three germplasm groups in China. **F**, Proportions of *Glyma.07G038400*^{H1} and *Glyma.07G038400*^{H2} in wild soybeans across the world. **G**, Summary of the main haplotypes of the *Glyma.16G007600*. **H**, Haplotype origins of *Glyma.16G007600.* Grey for wild soybean accessions, green for landraces, blue for improved cultivars. **I**, Proportions of *Glyma.16G007600*^{H1} and *Glyma.16G007600*^{H2} in the three germplasm groups in China. **J**. Oil content of different haplotypes of *Glyma.16G007600* in the three germplasm groups in China: wild soybeans (W), landraces (L), cultivars (C). * indicates statistically significant determined by Student's *t*-test ($P < 0.05$).

Results

GmGRF5a polymorphisms in the CDS formed two high-confidence haplotypes, H1 and H2 (Fig. A). H1 is the archetype, H2 is formed by the 539th base substitution (T to C), and H3 is formed by the 371st base substitution (C to T). The 371st base substitution makes differences in oil content in all three subgroups (Fig. B). However, in subgroups of wild soybeans and cultivars, 371^{T} carriers have a higher oil content, while in landraces, 371^{C} carriers have a higher oil content, suggesting that the natural mutation of *GmGRF5a* has an effect on oil content (**Fig. C**). The proportions of 371^{T} carriers in landraces and cultivars both have a significant rising trend from south to north, implying that *GmGRF5a* has a potential role in regulating flowering time or latitude adaptation (**Fig. D**). We also found a rising trend in the proportions of 371^{T} carriers from wild soybean in the whole of China (9 %) or Huanghuai region alone (36 %) to landrace (51 %) and cultivar (53 %) in China. In addition, we found that 371^{T} carriers mainly exist in the Huanghuai region of China, with a small proportion in the north and south, but do not exist in other areas where wild soybeans exist, including Russia, Japan and the Korean Peninsula (**Fig. E and F**). These results suggested that *GmGRF5b* may undergo a selection during not only domestication but also the diversification of soybean adapting to high latitudes and low latitudes.

Analysis of polymorphisms in the coding sequence of *GmGRF5b* defined a primary archetype H1, and four minor but high confidence (n>3) haplotypes, H2-H5 (Fig. **G** and **H**). Among the four minor haplotypes, H5 has relatively more carriers and is formed by the 50th base substitution (C to T). This mutation is present in landraces (5%) and cultivars (13%) but not in wild soybeans (**Fig. I**). Although the mutation occurs in the coding region of the WRC domain, the oil content of H5 (50^{c}) has no significant difference with H1-H4 (50^{T}) (**Fig. J**), implying that this mutation did not cause functional differences in landraces and cultivars. These results showed that *GmGRF5b*

is a conserved gene during the evolutionary history.

Methods

In this study, we utilized a variety of published re-sequencing data and VCF files (PRJNA394629, PRJNA608146, PRJNA859249, PRJNA743225, PRJNA776405, PRJNA859249, PRJNA1033042) [1, 2, 3, 4, 5] available in the NCBI database. The paired-end resequencing reads from the accessions in this study were mapped to the reference genome (Gmax_Wm82_a2_v1) using the BWA software with default parameters[5]. The filtration of duplicate sequencing reads, along with SNP and InDel calling, was carried out as described in previous studies[1]. High confidence mutations (with MAF > 0.05 and max missing < 0.01) were identified using VCFtools (version 0.1.16)[6]. The annotation was accomplished using ANNOVAR (version -0400)[7].

Reference

1. Lu, S. *et al.* Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. *Nat Genet* **52**, 428-436 (2020).
2. Dong, L. *et al.* The genetic basis of high-latitude adaptation in wild soybean. *Curr Biol* **33**, 252-262 e4 (2023).
3. Dong, L. *et al.* Genetic basis and adaptation trajectory of soybean from its temperate origin to tropics. *Nat Commun* **12**, 5445 (2021).
4. Dong, L. *et al.* Parallel selection of distinct Tof5 alleles drove the adaptation of cultivated and wild soybean to high latitudes. *Mol Plant* **15**, 308-321 (2022).
5. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
6. Petr Danecek, et al. The Variant Call Format and VCFtools. Bioinformatics. 27, 2156–2158, (2011)
7. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research*, 38:e164, (2010)

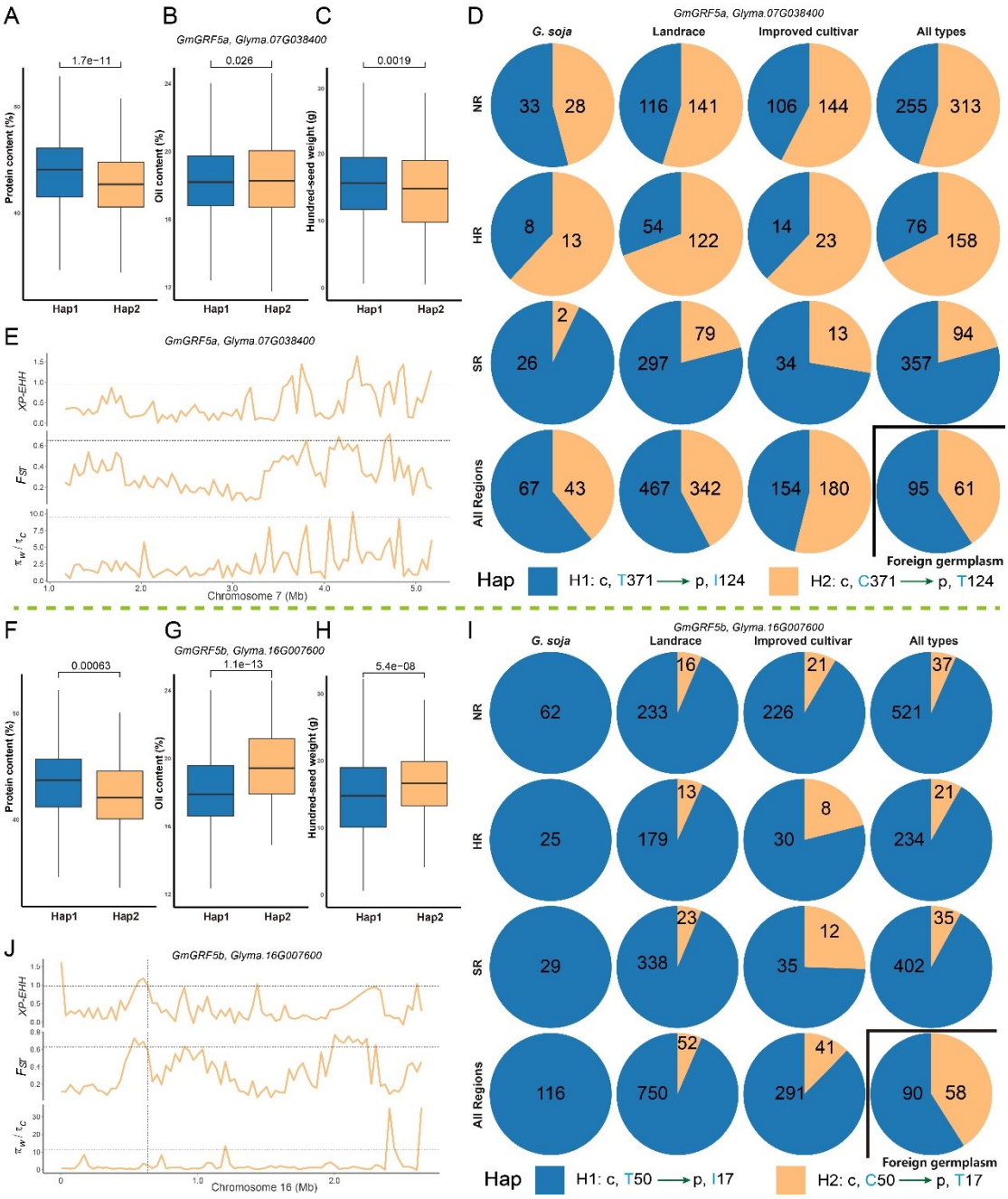


Fig. S24E *GmGRF5* haplotypes in geographical distribution and evolution and its relationship with yield and quality in Collection V (Fig 6B). (**A** to **E**) for *GmGRF5a*, (**F** to **J**) for *GmGRF5b*. Variations of protein content (**A**, **F**), oil content (**B**, **G**) and hundred-seed weight (**C**, **H**) between two haplotypes of *GmGRF5a* (**A to C**) and (**F** to **H**). The distribution of two haplotypes among *G. soja*, landrace, improved cultivar, and All types from NR, HR, SR, All Regions, and foreign germplasm (**D**, **I**). NR, HR and SR indicated northern region, Huanghuai region and southern region of China, respectively. π_w/π_c, F_{ST}, and XP-EHH values were calculated between *G. soja* and improved cultivar (**E**, **J**).

Result

In the soybean varieties of *GmGRF5a* and *GmGRF5b* genes, one nonsynonymous

variation suitable for haplotype analysis was identified in each gene. We further investigated the differences in protein content, fat content, and hundred-seed weight among different haplotypes, as well as the distribution of these haplotypes across various ecological zones and among different soybean varieties.

The two haplotypes of *GmGRF5a* exhibited significant differences in protein content, fat content, and hundred-seed weight (**A** to **C**). Specifically, $GmGRF5a^{H2}$ was characterized as a higher fat content, but lower protein content and hundred-seed weight compared to $GmGRF5a^{H1}$. The haplotype frequency of foreign germplasm is very close to that of landrace (**D**). The region harboring *GmGRF5a* showed a pronounced selection signal as indicated by its XP-EHH values (**E**). However, this area did not show significant genetic differentiation. The similar nucleotide diversity (π) values between wild and cultivated varieties suggested comparable diversity in both groups. These findings implied that the *GmGRF5a* gene may not have undergone substantial selection during domestication.

Compared to $GmGRF5b^{H1}$, $GmGRF5b^{H2}$ displayed higher fat content and hundred-seed weight, but lower protein content (**F** to **H**). Notably, $GmGRF5b^{H2}$ was absent in *G. soja* across all ecological regions. Furthermore, there was a significant increase in the frequency of $GmGRF5b^{H2}$ from landrace to improved cultivar, particularly in the Huanghuai region (by 23.90%) and in the Southern region (by 19.39%) (**I**). The proportion of $GmGRF5b^{H2}$ in foreign germplasm was as high as 39.19%, indicating that it may be more susceptible to selection outside of China. Analysis using the Cross Population Extended Haplotype Homozygosity (XP-EHH) revealed a significant selection signal in the region containing *GmGRF5b* (**J**). Additionally, elevated F_{ST} values suggested considerable genetic differentiation in this region. The lower nucleotide diversity (π) observed in cultivated varieties also implied potential selective pressure on the genomic region encompassing *GmGRF5b*.

These results suggested that different haplotypes of *GmGRF5a* and *GmGRF5b* may be associated with variations in protein content, fat content, and hundred-seed weight. $GmGRF5b^{H2}$ underwent significant selection in breeding programs, while $GmGRF5a^{H2}$ showed less evident selection but still presented a higher frequency in cultivated varieties and foreign germplasm.

Discussion

Targeted breeding selection for different haplotypes is beneficial for improving soybean quality and yield

Soybean varieties demonstrate a trend of decreasing fat content and increasing protein content from higher to lower latitudes, resulting in Northern and Southern varieties being more predisposed to high oil and high protein traits, respectively [1,2]. Multiple genes related to soybean protein or fat content identified by previous researchers are simultaneously associated with 100 seed weight, which is an important yield trait in soybeans [3-6]. These results suggest that soybean oil content may be selected along

with yield.

The two haplotypes of *GmGRF5a* experienced limited selection during the breeding process. Additionally, we observed distinct distributions of *GmGRF5a* haplotypes across different ecological regions. Targeted breeding selection for *GmGRF5a* could help improve the protein or fat content of soybeans, especially for the development of high protein varieties in northern China and high oil varieties in southern China. $GmGRF5b^{H2}$, which was associated with higher fat content and hundred-seed weight, was only present in cultivated soybeans. Thus, *GmGRF5b* may be an essential gene related to domestication. The substantial increase in the proportion of $GmGRF5b^{H2}$ from landrace to improved cultivar, especially in the Huanghuai and Southern regions, indicated significant selection of this gene in breeding programs. A higher proportion of $GmGRF5b^{H2}$ in foreign germplasm suggested that this haplotype may be more susceptible to stronger selection outside of China to improve yield and oil content. Enhancing the proportion of $GmGRF5b^{H2}$ could be a strategy for breeding high-oil, high-yield soybeans, especially in northern China.

Methods

The genotypic data for the 1690 accessions used in this study has been previously reported in earlier research [7]. Information on population grouping, along with data on protein content, fat content, and hundred-seed weight, was obtained from the soybean resource catalog (www.cgris.net).

We filtered SNPs with missing data > 20% or MAF < 5%. For calculations of nucleotide diversity (π) and fixation index (F_{ST}), vcftools [8] was utilized, employing a window size of 20k and a step size of 2k. We further calculated the ratio of π between *G. soja* and improved cultivars (π_w/π_c). The R package REHH 2.0 [9] was used to compute the Cross Population Extended Haplotype Homozygosity (XP-EHH). Sequences within the top 10% of π_w/π_c, F_{ST} and XP-EHH values were identified as regions under selection. Student's *t*-tests were implemented to ascertain *P* values.

Reference:

1. Guo, B., Sun, L., Jiang, S., Ren, H., Sun, R., Wei, Z., Hong, H., Luan, X., Wang, J., Wang, X., et al. (2022). Soybean genetic resources contributing to sustainable protein production. Theor Appl Genet *135*, 4095-4121. 10.1007/s00122-022-04222-9.
2. Abdelghany, A.M., Zhang, S., Azam, M., Shaibu, A.S., Feng, Y., Li, Y., Tian, Y., Hong, H., Li, B., and Sun, J. (2020). Profiling of seed fatty acid composition in 1025 Chinese soybean accessions from diverse ecoregions. The Crop Journal *8*, 635-644. https://doi.org/10.1016/j.cj.2019.11.002.
3. Cai, Z., Xian, P., Cheng, Y., Yang, Y., Zhang, Y., He, Z., Xiong, C., Guo, Z., Chen, Z., Jiang, H., et al. (2023). Natural variation of GmFATA1B regulates seed oil content and composition in soybean. Journal of Integrative Plant Biology *65*, 2368-2379. https://doi.org/10.1111/jipb.13561.

4. Duan, Z., Zhang, M., Zhang, Z., Liang, S., Fan, L., Yang, X., Yuan, Y., Pan, Y., Zhou, G., Liu, S., and Tian, Z. (2022). Natural allelic variation of GmST05 controlling seed size and quality in soybean. Plant Biotechnology Journal *20*, 1807-1818. https://doi.org/10.1111/pbi.13865.
5. Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y.-C., Liu, Z., Frommer, W.B., Ma, J.F., Chen, L.-Q., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. National Science Review *7*, 1776-1786. 10.1093/nsr/nwaa110.
6. Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., Song, Q., Pantalone, V.R., Song, B.-H., Yu, D., and An, Y.-q.C. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. Nature Communications *13*, 3051. 10.1038/s41467-022-30314-7.
7. Li, Y.-H., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., Li, Y., Xing, G., Wang, J., Gu, Y., et al. (2023). Genome-wide signatures of the geographic expansion and breeding of soybean. Science China Life Sciences *66*, 350-365. 10.1007/s11427-022-2158-7.
8. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158. 10.1093/bioinformatics/btr330.
9. Gautier, M., Klassmann, A., and Vitalis, R. (2017). rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Molecular Ecology Resources *17*, 78-90. https://doi.org/10.1111/1755-0998.12634.

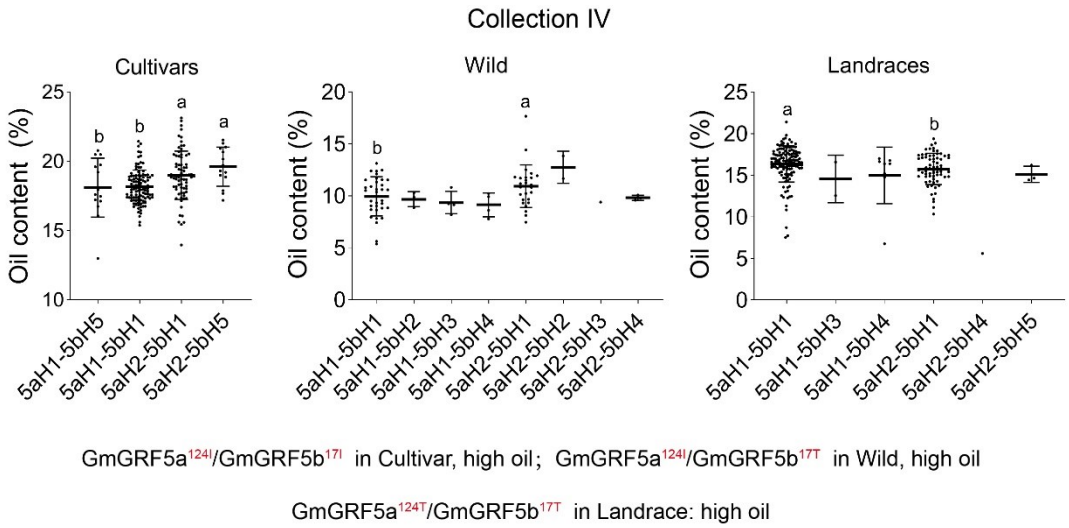


Fig. S25 The effect of combination of different haplotypes of *GmGRF5a* and *GmGRF5b* on protein and oil contents in soybean seeds. Haplotypes are referred to Fig. S24. Oil contents of each allelic combinations of *GmGRF5a* and *GmGRF5b* haplotypes in wild soybeans, landraces, and cultivars of Collection IV (Fig. 6B). The lower and upper box edges corresponded to the first and third quartiles (the twenty-fifth and seventy-fifth percentiles); the horizontal line indicated the median value. One-way analysis of variance (ANOVA) followed by Tukey's post hoc test ($p < 0.05$). Data sets with a number less than 10 were not included in the statistics.

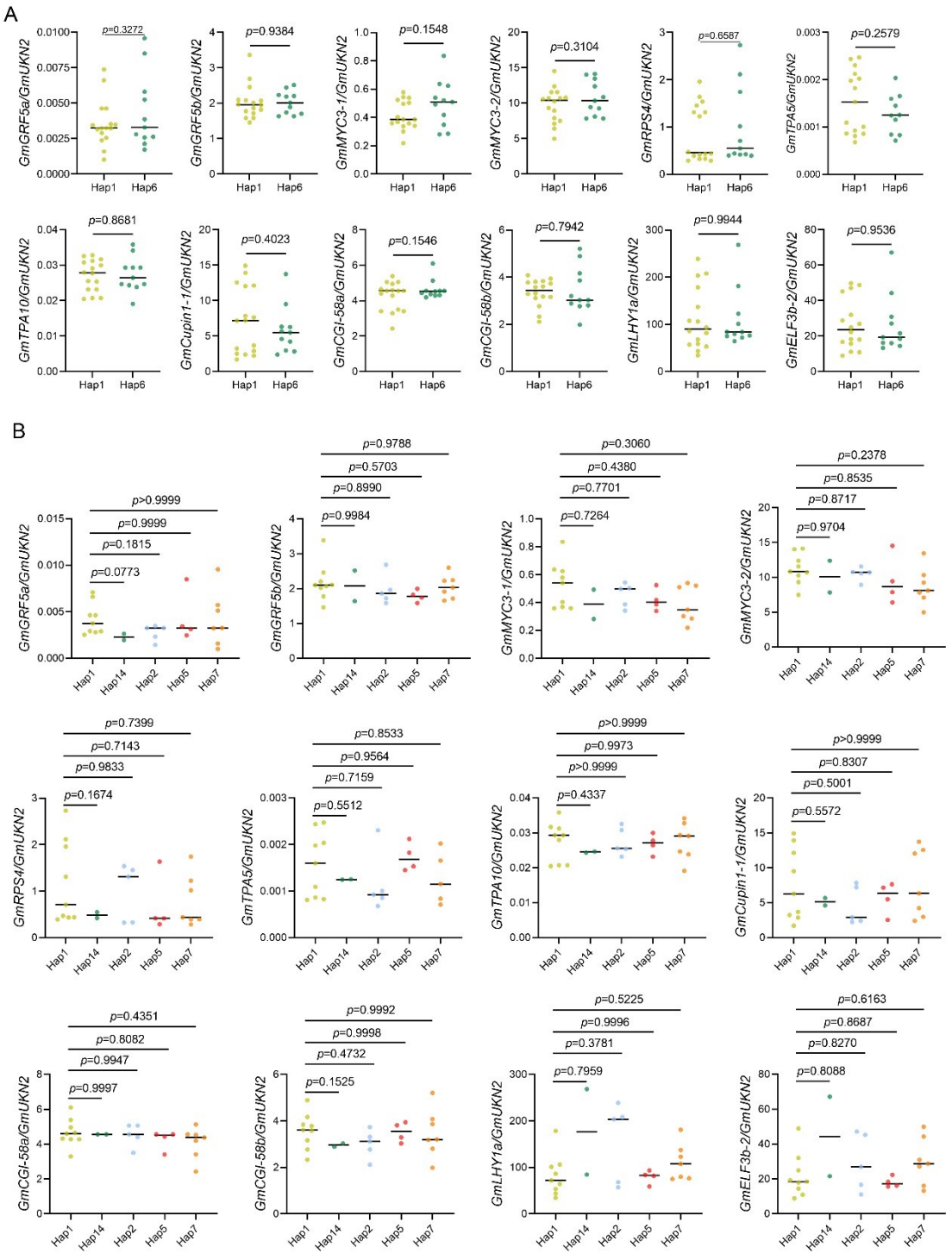


Fig. S26 Gene expression in germplasms with promoter Haplotypes min Collection II. A, mRNA level of *GmGRF5a*, *GmGRF5b*, and its related gene in two Haps (Hap1 and Hap6) of *GmGRF5a*. **B**, mRNA level of *GmGRF5*, *GmGRF5b*, and its related gene in five Haps (Hap1, Hap14, Hap2, Hap5 and Hap7) of *GmGRF5b*. All Haps (Fig. S24B) were in promoters. The samples of the first unrolled trifoliate leaves were harvested at 21 DAE (day after emergence) in growth chamber. *GmUKN2* is employed as a reference gene. Values are means ± SD (n = 3 biological repeats). Statistical analysis was performed using Student's *t* test (**A**) and One-way ANOVA (**B**).