Genome-wide screen of genomic imprinting in endosperm and population-level analysis reveal allelic variation for imprinting in flax

Haixia Jiang^{1, #}, Dongliang Guo^{1, #}, Jiali Ye², Yanfang Gao¹, Huiqing Liu¹, Yue Wang¹, Min Xue¹, Qingcheng Yan¹, Jiaxun Chen¹, Lepeng Duan¹, Gongzhe Li¹, Xiao Li¹ and Liqiong Xie^{1, *}

¹National Center of Melon Engineering and Technology, Molecular Breeding Laboratory, College of Life Science and Technology, Xinjiang University, Urumqi, Xinjiang, China.

²State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Soil and Water Conservation, College of Life Sciences, Northwest A&F University, Yangling, Shaanxi, China.

[#]These authors have contributed equally to this work.

^{*}**Correspondence:** Liqiong Xie, picea@sina.com.

Abstract

Genomic imprinting is an epigenetic phenomenon caused by the biased expression of maternally and paternally inherited alleles. In flowering plants, genomic imprinting predominantly occurs in triploid endosperm and plays a vital role in seed development. In this study, we identified 241 candidate imprinted genes including 143 maternally expressed imprinted genes (MEGs) and 98 paternally expressed imprinted genes (PEGs) in flax (*Linum usitatissimum* L.) endosperm using deep RNA sequencing. The conservation of imprinting in plants is very limited and imprinting clustering is not a general feature. MEGs tends to be endosperm expression specific, while PEGs are non-tissue specific. Imprinted SNPs differentiated 200 flax cultivars into oil flax, oil-fiber dual purpose flax (OF) and fiber flax subgroups, suggesting that genomic imprinting contributes to intraspecific variation in flax. The nucleotide diversity (π) of imprinted genes in oil flax subgroup is significantly higher than that in fiber flax subgroup, indicating that some imprinted genes undergo positive selection during flax domestication from oil flax to fiber flax. Imprinted genes undergo positive selection is related to the functions. Eleven imprinted genes related to seed size and weight were identified using the candidate gene-based association study. Our study provides information for further exploring the function and genomic variation of imprinted genes in flax population.

Keywords: genomic imprinting, flax endosperm, intraspecific variation, seed size, positive selection

Introduction

Genomic imprinting is an epigenetic phenomenon occurring in both mammals and flowering plants (Hutter *et al.*, 2010; Waters *et al.*, 2013). Imprinting occurs in the

placenta and embryo as well as in adult tissues in mammals (Renfree *et al.*, 2012), while in flowering plants, imprinting predominantly occurs in endosperm, and rarely in embryo and seed coat (Yuan *et al.*, 2017; Meng *et al.*, 2018). The diploid embryo transmits genetic information to the next generation, while the triploid endosperm, with 2:1 ratio of maternal to paternal genomes (2m:1p), provides nutrition and signals for embryo development (Sabelli and Larkins, 2009; Mei *et al.*, 2015). The deviation of the 2m:1p ratio in endosperm has adverse effect on seed development (Wang *et al.*, 2018; Scott *et al.*, 1998; Lu *et al.*, 2012; Sekine *et al.*, 2013), implying the important role of imprinting in endosperm development. Imprinted genes were studied only in a limited plant species, including monocots of rice (*Oryza sativa*) (Yuan *et al.*, 2017; Luo *et al.*, 2011), maize (*Zea mays*) (Waters *et al.*, 2013; Meng *et al.*, 2018; Zhang *et al.*, 2011; Waters *et al.*, 2011; Dong *et al.*, 2017), sorghum (*Sorghum bicolor* L. Moench) (Zhang *et al.*, 2016a) and dicots of *Arabidopsis* (Gehring *et al.*, 2011; Wolff *et al.*, 2011; Hsieh *et al.*, 2011), castor bean (*Ricinus communis*) (Xu *et al.*, 2014). In most dicots, the endosperm is transient and consumed by the embryo at the later stage of seed development (Sreenivasulu and Wobus, 2013), while the endosperm of most monocots remains persistently and serves as a source of nutrition for seed germination (Luo *et al.*, 2011). There is no extensive conservation of imprinted genes across species (Waters *et al.*, 2011; Zhang *et al.*, 2016a; Dong, 2017), suggesting different species may require a unique set of imprinted genes for seed development.

In plants, imprinted genes were detected to be involved in the regulation of seed development, seed dormancy and postzygotic reproductive isolation (Sun *et al.*, 2017; Piskurewicz *et al.*, 2016; Kradolfer *et al.*, 2013; Wolff *et al.*, 2015). The loss of function of some imprinted genes leads to seed abortion (Joanis and Lloyd, 2002; Chaudhury *et al.*, 1997; Berger *et al.*, 2006). In *Arabidopsis*, the maternally expressed imprinted gene *MEA* encodes a Polycomb Repressive Complex 2 (PRC2) subunit in Polycomb group (PcG) complex. The seeds with maternal *MEA* allele developed normally, while those carrying maternal *mea* allele were aborted, regardless of the genotypes of paternal allele (Joanis and Lloyd, 2002). Imprinted genes influence seed size by regulating the development of embryo and endosperm (Yuan *et al.*, 2017; Scott *et al.*, 1998; Köhler *et al.*, 2005; Chen *et al.*, 2016). In rice, the loss-of-function mutants of *MEG2* and *MEG3* (MEGs) showed a significant reduction in seed size and weight, and the loss-of-function of *PEG1*, *PEG2*, *PEG3* (PEG) decreased starch content, seed size and yield (Yuan *et al.*, 2017). Regarding to another imprinted gene *OsFIE1* in rice, RNAi lines and homozygous T-DNA insertion mutant *osfie1* lines all showed delayed embryo development and reduction of seeds fertility, grain size, grain weight and aleurone layer cells (Huang *et al.*, 2016). Even a few imprinted genes were shown to be important in seed development, there are many more, of which the function is yet to be determined.

Some imprinted genes were shown to be under positive selection and intraspecific variation features (Hutter *et al.*, 2010; Berger *et al.*, 2012; Pignatta *et al.*, 2014). *MEA*

originated a block duplication 35 to 85 million years ago owing to a whole-genome duplication within the Brassicaceae lineage (Spillane *et al.*, 2007). After duplication, *MEA* underwent positive selection consistent with neo-functionalization and the parental conflict theory (Spillane *et al.*, 2007; Miyake *et al.*, 2009). In maize, conservative imprinting genes increased the substitution rate of nonsynonymous to synonymous (dN/dS) compared with non-conservative imprinting genes and more likely to undergo positive selection (Waters *et al.*, 2013). PEGs were more likely to be under positive selection and rapidly evolve than MEGs in *Arabidopsis thaliana* (Tuteja *et al.*, 2019). Imprinting showed evidence of intraspecific variation in *Arabidopsis* and maize (Waters *et al.*, 2013; Pignatta *et al.*, 2014). The existence of intraspecific variation of imprinting was associated with epigenetic variation (Pignatta *et al.*, 2014).

Flax (*Linum usitatissimum* L.) is an important economic crop due to its stem fiber and seed oil (Cloutier *et al.*, 2012; Guo *et al.*, 2020). Flax is a strict annual self pollination crop with a smaller genome size (~373 Mb) (Wang *et al.*, 2012), which are good for biological research. Cultivated flax were domesticated from a pale flax (*Linum bienne*) for oil usage at 10,000 years ago in the Near East and differentiated into fiber, OF and oil subgroups during the domestication of flax (Guo *et al.*, 2020; Allaby *et al.*, 2005; Fu and Allaby, 2010). Some genes underlying important traits including plant architecture, flowering, dehiscence, oil production and yield underwent strong artificial selection in the domestication process of flax (Guo *et al.*, 2020; Zhang *et al.*, 2020). Whether imprinted genes are also undergone artificial selection during flax domestication has not been reported.

In this study, we have performed RNA-seq analysis of flax endosperm isolated from the reciprocal crosses between Cili2719 and Z11637. Based on the parent-of-origin biased expression of the parental alleles in the endosperm from both crosses, we identified 241 moderately imprinted genes including 143 MEGs and 98 PEGs, 67 strongly imprinted genes including 63 MEGs and 4 PEGs and 19 completely MEGs. The analysis of imprinted genes at population level demonstrated that imprinted genes divided the 200 flax germplasms into oil flax, OF and fiber flax subgroups, which suggested that imprinting promoted intraspecific variation of flax. The nucleotide diversity analysis showed that some imprinted genes were shown to be under positive selection consistent with function. Furthermore, we identified 11 imprinted genes associated to seed size and weight. Our results will provide a theoretical basis for further study of gene imprinting and provide some insights for understanding the diversity of imprinted genes.

Results

Identification of imprinted genes in flax endosperm

To understand the parental origin of gene expression in flax, we performed RNA-seq analysis in endosperm isolated from the F₁ generation of Cili2719×Z11637 (CZ) and

Z11637×Cili2719 (ZC) at 7 days after pollination (DAP). Average 7.21 Gb of 125bp paired-end clean reads were obtained from each biological replicate with the Illumina novaseq6000 platform. The clean reads from CZ and ZC were aligned to parental genomes Cili2719 and Z11637 (Guo *et al.*, 2020) to identify the reads specifically originated from one of parents at each SNP site for allelic expression (Figure S1). In total, 52,794 SNPs in both reciprocal crosses with at least ten reads could be assigned to a parental allele and were used for allele-specific expression analysis in hybrid endosperm. About 37,000 SNP loci showed statistically significant deviation ($p < 0.05$, χ^2 test) from the expected 2m:1p ratio in both CZ and ZC endosperm.

Three different thresholds were used to identify genes showing parent-of-origin biased expression at three different levels (see “Materials and methods”). Among these SNP loci, 498 loci were considered as moderately imprinted loci, including 319 maternally expressed imprinted SNP loci (ME-SNPs) corresponding to 143 MEGs and 179 paternally expressed imprinted SNP loci (PE-SNPs) corresponding to 98 PEGs. And 141 loci were identified as strongly imprinted loci, including 135 ME-SNPs and 6 PE-SNPs which correspond to 63 MEGs and 4 PEGs, respectively. In addition, 36 loci were identified as completely imprinted loci and all of them were ME-SNPs corresponding to 19 MEGs (Figure 1, Table S1). Among 241 imprinted genes, 229 genes were protein-coding genes (Table S2).

Validation of imprinted genes

For experimental confirmation, thirteen genes including five MEGs, five PEGs, and three non-imprinted genes, which represented the whole transcripts by RNA sequencing in this study, were randomly selected to validate the gene expression level of the high throughput sequencing with qRT-PCR analysis. The analysis showed that the gene expression level of the selected genes by qRT-PCR analysis was consistent with the RNA-seq data (Figure S2).

For further verification of the imprinting status, nine MEGs and three PEGs were used to perform RT-PCR on the hybrid endosperm and parents endosperm followed by Sanger sequencing (Figure S3). A 400-800bp fragment of each gene with at least one imprinted SNP site was selected for PCR amplification. The results showed that nine MEGs were predominantly expressed from the maternal alleles and three PEGs were preferentially expressed from the paternal alleles in reciprocal crosses, which were consistent with the RNA-seq data.

Characterization analysis of imprinted genes identified in flax

We carried out gene ontology (GO) analysis for the 229 imprinted protein-coding genes of flax, including 135 MEGs and 94 PEGs (Table S2). The background genes for GO analysis were 12,395 endosperm-expressed genes with at least ten reads could be assigned to a specific allele in both CZ and ZC. Categories with a significant level ($P < 0.05$) were defined as enriched. Compared with the whole transcripts in

endosperm, imprinted genes were significantly enriched in catalytic activity or transferase activity according to their molecular function and metabolic process according to their biological processes (Figure S4A, Table S3).

To evaluate the interspecific conservation of imprinted genes, we compared imprinted genes identified in this study with the imprinted genes reported in *Arabidopsis* (Gehring *et al.*, 2011), rice (Yuan *et al.*, 2017; Luo *et al.*, 2011), maize (Waters *et al.*, 2013; Dong, 2017), sorghum (Zhang *et al.*, 2016a), and castor bean (Xu *et al.*, 2014). The analysis showed that there were 115, 89 and 59 genes found to be homologous in at least one of the five species at different confidence levels ($<1E-10$, $<1E-20$ and $<1E-50$, respectively) (Figure 2, Table S4-S5). At $E\text{-value}<1E-10$, there were 32, 26, 43, 52, 39 flax imprinted genes conserved in *Arabidopsis*, castor bean, rice, sorghum and maize, respectively (Figure 2A, Table S4-S5). And 19, 21, 35, 46, 27 imprinted genes conserved with these plants at $E\text{-value}<1E-20$, respectively (Figure 2B, Table S4-S5). In addition, less imprinted genes 13, 11, 21, 22 and 15 conserved with the five species at $E\text{-value}<1E-50$, respectively (Figure 2C, Table S4-S5). Some imprinted genes had imprinted homologs in up to four species, while no imprinted gene in flax was conserved in all species (Figure 2, Table S4-S5). These results suggested that the conservation of imprinting in plants was quite limited.

Intriguingly, the expression of some conserved flax imprinted genes showed different parental origin in other species. For example, the flax gene *Lus10022747* encoding a serine/threonine-protein kinase WNK5 and its homologues in *Arabidopsis*, castor bean and rice showed maternally preferential expression, while its maize homolog displayed preferentially paternal expression (Table S4). A PAS domain tyrosine kinase family protein-coding gene *Lus10040540* and its homolog of maize were PEGs, but its homologues in castor bean, rice and sorghum were MEGs (Table S4). Among 115 conserved flax imprinted genes, only 59% (58 MEGs and 10 PEGs) remain the same preference of parental expression with other species (Table S4).

GO enrichment analysis of 115 conserved imprinted genes in flax displayed that MEGs significantly enriched in catalytic activity and metabolic process according to their molecular function and biological processes, respectively. PEGs enriched in compound binding according to their molecular function (Figure S4B, Table S6).

Clustering of the flax imprinted genes

To study the genomic distribution of flax imprinted genes, 229 imprinted genes were mapped to fifteen chromosomes for cluster analysis. The 229 imprinted genes were scattered distribution across fifteen chromosomes. By analyzing the genomic distance between the imprinted genes, we found that most of them were not co-localized in a cluster, and only 24 were fall into 12 clusters, where two imprinted genes of each cluster were within 10 kb (Figure 3, Table S7). The finding was similar to the results of *Arabidopsis* (Gehring *et al.*, 2011; Wolff *et al.*, 2011), maize (Waters *et al.*, 2011),

rice (Luo *et al.*, 2011), castor bean (Xu *et al.*, 2014) and sorghum (Zhang *et al.*, 2016a), showing that clustering of imprinted gene is not a common phenomenon in plants.

Endosperm-specific expression of the flax imprinted genes

We analyzed the expression specificity of the imprinted genes in various tissues. The majority of MEGs (60%) preferentially expressed in endosperm, and only a minority (25.5%) of PEGs were preferentially expressing in endosperm (Figure 4A-B). The expression level of endosperm-preferred MEGs (endo-MEGs) and PEGs (endo-PEGs) were significantly ($P < 0.05$) higher than that of all genes, whereas there was no evidence that endo-MEGs or endo-PEGs exhibited unusually high or low expression levels than other MEGs or PEGs which also expressed in other tissues (Figure 4C-D). We also analyzed the tissue specificity of 115 (75 MEGs and 40 PEGs) conserved imprinted genes and 38 (36 MEGs and 2 PEGs) conserved strong imprinted genes. Among 115 conserved imprinted genes, 49 (65.3%) MEGs and 12 (30%) PEGs showed endosperm-preferred expression (Figure S5A-B), and among the 38 conserved strong imprinted genes, 29 MEGs (80.6%) and all PEGs were preferentially expressed in endosperm (Figure S5C), respectively. These results suggested that MEGs and the conserved imprinted genes are more likely to be preferentially expressed in endosperm.

Flax imprinted genes can differentiate flax subgroups

To investigate whether the variation in imprinted genes reflects genetic diversity among 200 natural flax varieties, we detected individual kinship of imprinted SNPs (498) or genome-wide SNPs (674,074) (Guo *et al.*, 2020). The kinships between imprinted SNPs or genome-wide SNPs were significantly correlated ($R^2 = 0.8457$) (Figure 5A). Phylogenetic tree constructed based on all imprinted SNPs, ME-SNPs or PE-SNPs separated 200 accessions into three different subgroups which correspond to oil flax, OF and fiber flax subgroups (Figure 5B, Figure S6). The flax population could also be separated into three subgroups by principle component analysis (PCA) (Figure 5C). These results indicated that the allele frequency of imprinted SNPs was significantly different among subgroups, suggesting that imprinted genes may be selected differently in subgroups and contribute to domestication.

Selective sweep signals in imprinted genes

To test the hypothesis that the diversity of imprinted genes are different in flax population, we collated all SNPs (3,191 SNPs) of 241 imprinted genes and compared the nucleotide diversity between oil flax (78 germplasms) and fiber flax subgroups (51 germplasms) which represented two primary morphotypes of cultivated flax (Zhang *et al.*, 2020). The π values of all imprinted genes, MEGs or PEGs decreased significantly in fiber flax compared with oil flax ($P < 0.0001$, t test) (Figure 6A-C). We focused on two imprinted genes, *Lus10010350* (PEG) and *Lus10024230* (MEG),

which contained more SNPs (31 SNPs and 43 SNPs, respectively) in genomic sequences. The allele frequency distribution of imprinted SNPs in *Lus10010350* was significantly different in the two subgroups, and the alternate allele ‘G’ (position 333406) was primarily found in the oil subgroup and rarely in fiber subgroup (Figure 6D). Significant reduction of π was also observed at the *Lus10010350* locus in fiber subgroup compared with that of oil subgroup and 90.32% of the SNPs were identified with a signature of purifying selection (Figure 6E-F). Similarly, the allele frequency distribution of *Lus10024230* was obvious different between the two subgroups and the alternate allele ‘A’ (position 492363) accounted for 61.67% in oil subgroup but only 6.98% in fiber subgroup (Figure 6G). Compared with oil subgroup, the π value of the SNPs in *Lus10024230* was dramatically decreased in fiber subgroup, and 95.35% of the SNPs were identified with a signature of purifying selection (Figure 6H-I). Taken together, these findings suggested that some imprinted genes may have been subjected to artificial selection during flax domestication.

Candidate gene-based association study for seed size using flax imprinted genes

Imprinted genes played an important role in the regulation of endosperm development and seed size (Yuan *et al.*, 2017; Luo *et al.*, 2000; Guitton *et al.*, 2004; Huang *et al.*, 2017; Kinoshita *et al.*, 1999; Kiyosue *et al.*, 1999). To investigate whether imprinting is associated with seed size in flax, imprinted genes were used to perform candidate gene-based association study of seed size-related traits including seed length (SL), seed width (SW) and 1,000-seed weight (1000-SW). Using the general linear model (GLM) and mixed linear model (MLM) in TASSEL 5.0 (Bradbury *et al.*, 2007), 33 imprinted genes containing 63 associated loci (SNPs) were detected to be associated with seed size and weight. Among them, 11 imprinted genes were repeatedly detected at least two environments or traits (Figure 7A-B, Table S8).

One of the significant signal peaks on chromosome 15 contained 9 repetitive SNPs (Figure 7C, Table S8), which located in the *Lus10010350* (PEG), encoding a bifunctional arginine demethylase and lysine hydroxylase jmjd6 protein. This gene contained 31 SNPs, of which 12 induced nonsynonymous mutations and formed 11 haplotypes (Figure 7D-E). We classified 200 accessions into two groups including haplotype A (reference alleles) and haplotypes B–K (alternate alleles) based on gene structural variation (Figure 7F). We found that flax accessions in haplotypes B–K had significantly longer seed length and width, and larger 1,000-seed weight than those in haplotype A (Figure 7G-J, Figure S7). These results suggested that the PEG *Lus10010350* may be involved in the seed size regulation in flax.

Discussion

Characterization of imprinted genes in flax endosperm

GO analysis for the 229 imprinted protein-coding genes of flax revealed that a majority of imprinted genes were significantly enriched in catalytic activity and

metabolic process (Figure S4A, Table S3), similar to the results in castor bean (Xu *et al.*, 2014) and sorghum (Zhang *et al.*, 2016a), suggesting imprinted genes affected various aspects of endosperm development. However, imprinted genes have limited conservation across plant species, in contrast to those in mammals (Waters *et al.*, 2011; Zhang *et al.*, 2016a; Xu *et al.*, 2014; Dong, 2017; Zhang *et al.*, 2003). In maize and sorghum endosperm, only about 10% and 33% imprinted genes were conserved with other species, respectively (Waters *et al.*, 2011; Zhang *et al.*, 2016a). Among 165 imprinted genes in rice, only 33% of them were conserved in maize (Dong, 2017). Twenty-five (12%) imprinted genes identified in castor bean were conserved in *Arabidopsis*, rice or maize (Xu *et al.*, 2014). In this study, the conservation of flax imprinted genes was evaluated with those in other species (*Arabidopsis*, castor bean, rice, sorghum and maize) at three levels of stringency. There were 115 (50.2%), 89 (38.9%), 59 (25.8%) imprinted genes in flax which were homologous to those in at least one of the five species at E-value<1E-10, 1E-20 and 1E-50, respectively (Figure 2, Table S4-S5) while none imprinted genes of flax were found having imprinted homologs with all five species (Figure 2, Table S4-S5). Those results suggested that some common pathways in different flowering plants may need to be regulated by imprinting to modulate endosperm development but different genes in the pathway are selected to be imprinted in different species. This explains why the conservation of imprinting in plants is quite limited.

An interesting observation was that among 115 conserved imprinted genes, only 68 genes (58 MEGs and 10 PEGs) showed the same origin of parental expression as in other species (Table S4). The remaining 47 genes (17 MEGs and 30 PEGs) had opposite origin of parental expression. For instance, the gene *Lus10041031* is a complete MEG identified in our study, but its maize homolog is a PEG (Waters *et al.*, 2013). *Lus10021926* is a strong PEG in flax, while in *Arabidopsis* and castor bean its homologs are MEGs (Gehring *et al.*, 2011; Xu *et al.*, 2014). The homologous genes of the strong PEG *Lus10016563* are PEG in maize and MEGs in rice and sorghum (Waters *et al.*, 2013; Yuan *et al.*, 2017; Zhang *et al.*, 2016a). This suggested that those genes have opposite mode of parental expression in different species may be subject to gene dosage, a mechanism thought to be important for endosperm development (Wang *et al.*, 2018; Scott *et al.*, 1998; Lu *et al.*, 2012; Sekine *et al.*, 2013). It was noteworthy to note that the gene *Lus10041386* encoding a histone-lysine N-methyltransferase Enhancer of Zeste homolog 2 (EZH2) is homologous to the FERTILIZATION-INDEPENDENT SEED Polycomb Repressive Complex 2 (FIS-PRC2) class gene *MEA* in *Arabidopsis* and the imprinted gene *GRMZM2G157820* (*EZH2*) in maize (Table S4). In rice and other species, other members of PRC2 genes are also imprinted, suggesting that imprinting of PRC2 genes is a conserved mechanism in flowering plants. Several lines of evidence suggested that PRC2 repressed the replication of central-cell nuclear before fertilization likely by the maternally expressed alleles and regulated endosperm proliferation, suggesting a vital role in seed development (Chaudhury *et al.*, 1997;

Kiyosue *et al.*, 1999; Zhang *et al.*, 2005; Ohad *et al.*, 1996; Ohad *et al.*, 1999; Luo *et al.*, 1999; Moreno-Romero *et al.*, 2019).

Imprinted genes are not extensively clustered

Physical clustering of imprinted genes is a conserved feature in mammals (Gehring *et al.*, 2011; Gregg *et al.*, 2010), while there is little evidence of clustering in plant species. Imprinted genes identified in maize (Waters *et al.*, 2011), *Arabidopsis* (Gehring *et al.*, 2011; Wolff *et al.*, 2011), rice (Luo *et al.*, 2011), castor bean (Xu *et al.*, 2014) and sorghum (Zhang *et al.*, 2016a) were not shown to be extensive clustered. Using a clustering criterion consistent with that in *Arabidopsis* (~125 Mb) and castor bean (~350 Mb) which have comparable genome size to flax (~373 Mb) (Gehring *et al.*, 2011; Xu *et al.*, 2014), we found that 24 of 229 flax imprinted genes were fall into 12 clusters (Figure 3, Table S7), similar to the proportion in *Arabidopsis* (Gehring *et al.*, 2011; Wolff *et al.*, 2011), rice (Luo *et al.*, 2011), maize (Waters *et al.*, 2011), castor bean (Xu *et al.*, 2014) and sorghum (Zhang *et al.*, 2016a), suggesting that imprinting clustering may be not a general feature in plants. Whether the clustered imprinted genes are coordinately regulated as those genes in animal clusters remains to be investigated.

Endosperm-specific expression of flax imprinted genes

According to previous reports, imprinted genes in plants were mainly restricted to express in endosperm (Berger *et al.*, 2012). But more and more studies had shown that only some imprinted genes are preferentially expressed in endosperm, while others are also expressed in other tissues (Waters *et al.*, 2013; Waters *et al.*, 2011; Dong, 2017). The proportions of endo-MEGs and endo-PEGs were dramatically different (68% MEGs versus 26% PEGs, 51% MEGs versus 24% PEGs in maize; 50% MEGs versus 16% PEGs in rice; 50% MEGs versus 20% PEGs in sorghum) (Waters *et al.*, 2013; Dong, 2017). In our study, we found 81 endo-MEGs (60%) and only 24 endo-PEGs (25.5%) in flax (Figure 4A-B), similar to the proportion in rice, sorghum and maize (Waters *et al.*, 2013; Zhang *et al.*, 2016a; Dong, 2017). The expression level of endo-MEGs and endo-PEGs was significantly higher than that of all genes (Figure 4C-D). Compared with all imprinted genes (60% endo-MEGs versus 25.5% endo-PEGs), the proportion of endo-MEGs (65%) and endo-PEGs (30%) of conserved imprinted genes increased (Figure S5A-B). In the 38 (36 MEGs and 2 PEGs) conserved strong imprinted genes, 29 MEGs (80.6%) and all PEGs were endosperm-preferred expression (Figure S5C). These findings suggested that MEGs tends to be endosperm preferentially expressed, while PEGs are inclined to non-tissue specific expression. It also implied that the conserved imprinted genes are more likely to be preferentially expressed in endosperm and play an important role in seed development.

Candidate gene-based association study reveals that some imprinted genes are involved in flax seed size regulation

Previous studies have shown that imprinted genes play an important role in seed development by regulating the development of endosperm (Yuan *et al.*, 2017; Köhler *et al.*, 2005; Chen *et al.*, 2016; Luo *et al.*, 2000; Guitton *et al.*, 2004; Huang *et al.*, 2017; Zhang *et al.*, 2016b). Eleven imprinted genes related to seed size and 1,000-seed weight were obtained based on candidate gene-based association study. Among 11 imprinted genes, the gene *Lus10036044* (MEG) encoding a plant AT-rich sequence- and zinc-binding (PLATZ) transcription factor had significantly associated signal peaks on chromosome 1 (Table S8). PLATZ transcription factor is a novel class of plant-based zinc ion and DNA binding proteins, reported to regulate the seed size and weight (Azim *et al.*, 2020). *ZmPLATZ12* (*Fl3*) is a maternally expressed imprinted gene specifically expressing in the starchy cells of endosperm in maize. The semi dominant negative *fl3* mutant resulted in severe defects of endosperm and dramatically reduced the weight of seeds (Li *et al.*, 2017). In rice, the PLATZ transcription factor *GL6* positively controlled grain length through promoting cell proliferation in grains. The null *gl6* mutant led to short grains, whereas overexpression the *GL6* produced large grains (Wang *et al.*, 2019). Another PLATZ gene *SHORT GRAIN6* (*SG6*) determined grain size by regulating the cell division of spikelet hull. The grain size and weight was significantly enlarged in the *SG6* overexpression lines and reduced in *sg6* mutant lines in rice (Zhou and Xue, 2020).

Another candidate gene *Lus10037040* (MEG) located on chromosome 1, which belongs to the MADS-box genes (Table S8). MADS-box genes had important functions in the development of seed by epigenetic mechanism including DNA methylation and histone modifications (Zhang *et al.*, 2016b). In rice, *OsMADS87* (MEG) affected seed size by regulating endosperm cellularization during syncytial stage. Over expression the *OsMADS87* led to larger seeds, and *OsMADS87*-RNAi resulted in smaller seeds (Chen *et al.*, 2016). The MADS-box gene *PHE1* (PEG) regulated seed size in *Arabidopsis thaliana* via influencing the expression of *AGL62* which might affect the endosperm cellularization (Sun *et al.*, 2017). *OsMADS29* regulated seed development though regulating cell degeneration of maternal tissues. *OsMADS29*-RNAi resulted in aborted and/or shriveled seeds with deficient starch accumulation in endosperm (Yang *et al.*, 2012). Heterologous expression the *CnMADS* gene significantly increased the seed size of *Arabidopsis* (Sun, 2018).

Lus10024230 annotated as flavonol synthase (FLS) is also potentially involved in seed size control (Table S8). In the lines of *FLS*-RNAi of tobacco, the pods and seed development was arrested and the height, pods size, pods weight, seeds number were significantly reduced (Mahajan *et al.*, 2011). Furthermore, the alternative alleles at *Lus10010350* (haplotypes B–K) had significantly longer seed length and width, and larger 1,000-seed weight than those in reference allele (haplotype A) in 200 flax accessions (Figure 7G–J, Figure S7). Together, our study identified a few candidate imprinted gene which are potentially involved in seed development and modulate the

seed size. The genetic variation of these genes between flax lines may be harnessed as breeding tool for enhance seed yield.

Intraspecific variation of flax imprinted genes

DNA methylation, histone modification and non-coding small RNAs caused genomic imprinting (Zhang *et al.*, 2016a; Sha, 2008; Hanna and Kelsey, 2017). Epigenetic modification often varied across different individuals of the same species (Pignatta *et al.*, 2014; Xu *et al.*, 2019). In maize, differentially methylated regions (DMRs) were changed in different subgroups and genotypes (Xu *et al.*, 2019; Li *et al.*, 2015). In *Arabidopsis*, DNA methylation and small RNAs differentiated in natural populations and contributed to phenotypic diversity (Pignatta *et al.*, 2014; Schmitz *et al.*, 2011; Becker *et al.*, 2011; Graaf *et al.*, 2015; Schmitz *et al.*, 2013). Genomic imprinting, as the functional product of epigenetic modification, varied within a same species and the intraspecific variation of imprinted genes was associated with epigenetic variation (Waters *et al.*, 2013; Pignatta *et al.*, 2014). In this research, the analysis of phylogenetic tree and PCA for imprinted SNPs showed that imprinted SNPs effectively divided the 200 flax germplasms into oil, OF and fiber flax subgroups (Figure 5, Figure S6), suggesting that genomic imprinting changed in different subgroups and contributed to phenotypic diversity in flax.

Some imprinted genes show evidence of positive selection

According to previous reports, some imprinted genes showed positive selection features (Hutter *et al.*, 2010; Berger *et al.*, 2012). *MEA* as a component of FIS-PRC2 was a very important conserved imprinted gene in seed development underwent positive selection in the out-crossing lineages but not in the self-fertilizing species of *Arabidopsis* (Spillane *et al.*, 2007; Miyake *et al.*, 2009). Conserved imprinted genes displayed higher dN/dS rates than non-conservative imprinted genes between maize, rice and sorghum, suggesting conserved imprinted genes showing greater evidence of positive selection (Waters *et al.*, 2013). Compared with MEGs, PEGs exhibited elevated dN/dS values and more likely to under positive darwinian selection in *Arabidopsis thaliana* (Tuteja *et al.*, 2019). Our data showed that the nucleotide diversity of imprinted genes in oil flax subgroup was significantly higher than that in fiber flax subgroup (Figure 6A-C). The π values of some imprinted genes, such as *Lus10010350* (PEG, Figure 6D-F), *Lus10024230* (MEG, Figure 6G-I) and *Lus10041386* (MEG, Figure S8) were also significant difference between oil and fiber flax subgroup. Our results revealed that imprinted genes have been undergone artificial selection in the process of flax domestication from oil flax to fiber flax (Guo *et al.*, 2020).

By analyzing the nucleotide diversity of imprinted genes in different flax subgroups, we found that the π values of imprinted genes in oil flax subgroup were significantly higher than those in fiber flax subgroup no matter what parental origin they were (Figure 6A-C). Meanwhile, we also discovered that the imprinted genes related to

seed size and weight contained MEGs and PEGs (Table S8). It seemed that MEGs and PEGs were same shaped by selective force in flax population differentiation although the number of MEGs was larger than that of PEGs. So, we expected that imprinted genes undergo positive selection is related to the functions, but not to the parental origin which was different from the previous report (Tuteja *et al.*, 2019). Compared with parental conflict theory, the imprinting under relaxed selection theory that genomic imprinting evolves consistent with neo-functionalization (Rodrigues and Zilberman, 2015) can better explain the intraspecific imprinting variation in flax subgroups.

Materials and methods

Plant Material and Tissue collection

The two parental lines of flax (*Linum usitatissimum* L.) for reciprocal crosses, Cili2719 (C) and Z11637 (Z), were grown at the Miquan Experiment filed in Urumqi, Xinjiang. The large seed line Cili2719 which 1000-seed weight was about 10.5g originated from France and the small seed line Z11637 that 1000-seed weight was about 3.7g originated from the United States. The seeds of Cili2719×Z11637 (CZ) and Z11637×Cili2719 (ZC) were collected at 7 DAP (day after pollination). Endosperm tissues were collected from at least 50 seeds by manual dissection in each replicate and were immediately frozen in liquid nitrogen. Three biological repeats were set up for each line. For phenotyping, the 200 accessions were planted in four environment comprising Dali in Yunnan Province in 2016 (2016DL), Urumqi in Xinjiang autonomous region in 2017 and 2019 (2017UR, 2019UR), and YiLi in Xinjiang autonomous region in 2019 (2019YL). Planting and phenotyping of the 200 accessions were performed using a same strategy as described in our previous study (Guo *et al.*, 2020).

Library construction for RNA-Seq

Total RNA was extracted using a RNAprep Pure Plant Kit (Tiangen Biotechnology of Beijing, <http://www.tiangen.com/>). The quantification and qualification of RNA was checked by 1% agarose gels, NanoPhotometer[®] spectrophotometer (IMPLEN, CA, USA), Qubit[®] RNA Assay Kit in Qubit[®] 2.0 Fluorometer (Life Technologies, CA, USA) and the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The RNA-seq libraries were generated using NEBNext[®] Ultra[™] RNA Library Prep Kit for Illumina[®] (NEB, USA) according to the manufacturer's instructions and the high-throughput sequencing was performed with the Illumina NovaSeq6000 platform. Then, the quality and quantity of these libraries were assessed by using the Agilent Bioanalyzer 2100 system and Q-PCR. A data size of 301.98 million 125bp paired-end raw reads was obtained from CZ and ZC.

Read mapping and gene expression analysis

After removing the reads containing adapter, reads containing poly-N (> 10%) and low quality reads ($Q_{\text{phred}} \leq 20$) from raw data, a total of 288.37 million clean reads (43.26 Gb) were obtained for the following analysis. The clean reads were aligned to flax reference genome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Lusitatissimum) (Wang *et al.*, 2012) using Hisat2 v2.0.4. HTSeq v0.9.1 to count the reads numbers mapped to each gene. And then the expected number of Fragments Per Kilo base of transcript sequence per Millions base pairs sequenced (FPKM) of each gene was calculated based on the length of the gene and reads count mapped to this gene (Trapnell *et al.*, 2010). In the three biological replicates, the gene with an average expression level of FPKM > 1 was identified as "expressed" (Meng *et al.*, 2018).

Identification of imprinted genes

The clean reads from CZ and ZC were aligned to parental genomes Clli2719 and Z11637 from our previous research (Guo *et al.*, 2020) to obtain the reads of C and Z alleles at each SNP site for parental allelic expression analysis. Theoretically, the allelic ratio of the maternal to paternal is 2 to 1 in hybrid endosperm. Based on the 2m:1p ratio, SNP loci with more than 10 alleles reads in reciprocal crosses were used to perform a two-tailed chi square (χ^2) test. Moderately imprinted SNP loci had significant allelic bias ($\chi^2 < 0.05$) and >80% of the transcripts from the maternal allele for maternally expressed imprinted SNP loci or >60% of the transcripts coming from the paternal allele for paternally expressed imprinted SNP loci in both reciprocal hybrids. Strong maternally or paternally expressed imprinted SNP loci were defined as having significant allelic bias ($\chi^2 < 0.01$) and >90% of transcripts derived from the maternal allele or paternal allele, respectively. Complete maternally/paternally expressed imprinted SNP loci had >99% of the transcripts from the maternal/paternal allele (Waters *et al.*, 2013; Meng *et al.*, 2018). And the genes containing at least one imprinted SNP loci were identified as imprinted genes.

Validation of imprinted gene and expression analysis

Thirteen genes were used to perform Quantitative RT-PCR (qRT-PCR) analysis (Table S9) and twelve imprinted genes were detected using a PCR-sequencing method (Table S10) (Meng *et al.*, 2018; Xu *et al.*, 2014). The endosperm cDNA samples at 7-DAP were collected for RNA isolation with three biological repeats for each sample.

The extraction, quantification and identification of total RNA were the same as that of library construction for RNA-Seq. First-strand cDNA synthesis was performed using 5×All-In-One RT MasterMix (with AccuRT Genomic DNA Removal Kit) according to the manufacturer recommended protocol for qRT-PCR and RT-PCR (abm, Cat. No.G492, <http://www.abmGood.com/>). Each qRT-PCR reaction of CZ and ZC was performed by the manufacturer's instructions of EvaGreen Express 2×qPCR MasterMix (abm, Cat. No.MasterMix-ES, <http://www.abmGood.com/>) and

BioRad®CFX96 Real-Time PCR system (Bio-Rad). Relative expression was quantified with the geometric mean of internal reference genes *ETIF1* (eukaryotic translation initiationfactor 1), *GAPDH* (glyceraldehyde 3-phosphate dehydrogenase), and *ETIF5A* (eukaryotic translation initiationfactor 5A) (Hobson and Deyholos, 2013; Huis *et al.*, 2010). For RT-PCR, a 400-800bp amplification fragment of each gene was amplified by different primers with four endosperm cDNA samples: CC (inbred lines of Clli2719), ZZ (inbred lines of Z11637), CZ and ZC of 7-DAP endosperm. The RT-PCR amplified products contained at least one imprinted SNP sites were analyzed on agarose gels and then sequenced.

Functional characterization of imprinted features

Gene annotation of imprinted genes in flax endosperm was downloaded from the reference genome (<https://phytozome.jgi.doe.gov/pz/portal.html>), and GO enrichment analysis was carried out with WEGO (<http://wego.genomics.org.cn/>) (Xu *et al.*, 2014; Ye *et al.*, 2006).

The imprinted genes of flax were investigated for sequence homology in *Arabidopsis* (Gehring *et al.*, 2011), rice (Yuan *et al.*, 2017; Luo *et al.*, 2011), maize (Waters *et al.*, 2013; Dong, 2017), sorghum (Zhang *et al.*, 2016a), and castor bean (Xu *et al.*, 2014) using blast. The peptide sequences of flax imprinted genes were obtained from the flax database in Phytozome v12.1 (<https://phytozome.jgi.doe.gov/pz/portal.html>). Then, the peptide sequences were aligned to the *Arabidopsis* genome (*Arabidopsis thaliana* TAIR10) and high-scoring (E-value<1E-10, 1E-20 and 1E-50) blasts hits were ordered by increasing E-value. If the *Arabidopsis* imprinted genes were identified amongst the blast hits, and the gene with the smallest E-value was recorded. Similarly, candidate genes from this study in flax were aligned to the rice (*Oryza sativa* v7_JGI), maize (*Zea mays* Ensembl-18), sorghum (*Sorghum bicolor* v3.1.1) and castor bean genome (*Ricinus communis* v0.1). The Venn diagrams were drawn by the draw venn diagram online software (<http://bioinformatics.psb.ugent.be/webtools/Venn/>).

For clustering analysis of imprinted genes, a standard was applied that imprinted genes within 10 kb of one another in the flax genome was a candidate cluster (Gehring *et al.*, 2011; Xu *et al.*, 2014). Positions of imprinted genes on chromosome were mapped using the MapChart software (Voorrips, 2002).

The tissue-specific expression analysis of imprinted genes in endosperm

The gene-expression patterns for MEGs and PEGs in various flax tissues in reciprocal hybrids were identified based on RNA-seq analysis. The endosperm and embryo tissues were harvested at 7 DAP and the leaf tissues were collected at 2 weeks after planting. For each sample, three biological replicates were used. The FPKM expression values of all genes and imprinted genes in CZ and ZC were log-transformed. All genes with FPKM>1 in endosperm were used in this study. The heat map and hierarchical clustering of normalized expression levels (FPKM) were

performed with the MeV4.9.0 software (Multi Experiment Viewer, <https://sourceforge.net/projects/mev4/files/mev-tm4/>) (Guo *et al.*, 2020).

Phylogenetic tree and population structure analysis using imprinted SNPs

To test the relationship between SNP variation at population level and population structure, an individual-based neighbor-joining tree was generated based on the all imprinted SNPs (498 SNPs), ME-SNPs (319 ME-SNPs) or PE-SNPs (179 PE-SNPs) by TASSEL's Cladogram function (Bradbury *et al.*, 2007). We compared the kinship of 200 accessions calculated by imprinted SNPs (498) and genome-wide SNPs (674,074) obtained from our previous research (Guo *et al.*, 2020). Principal component analysis (PCA) was conducted based on the imprinted SNPs by using the software TASSEL 5.0 (Bradbury *et al.*, 2007).

Identifying selection signatures of imprinted genes

A set of 78 oil flax and 51 fiber flax accessions which represent two primary morphotypes of cultivated flax were used for selective sweeps analysis. To test the genetic diversity of imprinted genes in different subgroups, 3,191 SNPs were obtained by mapping the imprinted genes to our previously constructed variation map (Guo *et al.*, 2020). The π values of all imprinted genes (241), MEGs (143) and PEGs (98) between oil and fiber subgroups were calculated at the gene level using all SNPs within each imprinted gene by the software DnaSP 5.1 (Librado and Rozas, 2009). Furthermore, the π values were also calculated at the SNP level in oil subgroup and fiber subgroup for detecting the selection signatures in a single imprinted gene using DnaSP 5.1 (Librado and Rozas, 2009).

Candidate gene-based association study for seed size-related traits using flax imprinted genes

To analyze the association between imprinted genes and seed size-related traits, imprinted genes were used to perform candidate gene-based association study by the general linear model (GLM) and mixed linear model (MLM) in TASSEL 5.0 (Bradbury *et al.*, 2007). The 3,191 SNPs in the imprinted genes and seed size-related traits including seed length (SL), seed width (SW) and 1,000-seed weight (1,000-SW) of 200 flax accessions were obtained from our previous research (Guo *et al.*, 2020). For GLM analysis, the top two principal components (PC) were used to generate the population structure matrix and the threshold was set as $0.1/\text{total SNPs}$ ($\log_{10}(P) = -4.50$). For MLM analysis, P matrix and Kinship (K) matrix need to be considered, and the suggestive threshold was set as $1/\text{total SNPs}$ ($\log_{10}(P) = -3.50$). The imprinted genes repeatedly detected for at least two environments or traits were considered to be associated with seed size and weight.

Acknowledgements

We thank Chinese Crop Germplasm Resources Information System (CGRIS), Plant Gene Resource Centre in Canada (PGRC) and United States National Plant Germplasm System (U.S.NPGS) for providing us with germplasm resources. This study was supported by the National Natural Science Foundation of China (32060426), Resource Platform Project of Xinjiang Uygur Autonomous Region of China (PT1808) and Science and Technology Innovation Project for Doctoral Students of Xinjiang University of China (XJUBSCX-2017017). We are grateful to professors Ming Luo and Donghe Xu for constructive comments to improve the manuscript. The authors declare no competing interests of this work.

Conflicts of interest statement

The authors declare no competing financial interests.

Availability of data and materials

Raw resequencing sequence data of the two parental lines Cili2719 and Z11637 of flax for reciprocal crosses are available at NCBI under accession PRJNA590636 (ncbi.nlm.nih.gov/bioproject/PRJNA590636) (Guo *et al.*, 2020).

References

- Allaby RG, Peterson GW, Merriwether DA, Fu Y-B (2005). Evidence of the domestication history of flax (*Linum usitatissimum* L.) from genetic diversity of the *sad2* locus. *Theoretical and Applied Genetics* **112**, 58-65. <https://doi.org/10.1007/s00122-005-0103-3>
- Azim JB, Khan MFH, Hassan L, Robin AHK (2020). Genome-wide characterization and expression profiling of plant-specific PLATZ transcription factor family genes in *Brassica rapa* L. *Plant Breeding and Biotechnology* **8**, 28-45.
- Becker C, Hagmann J, Müller J, Koenig D, Stegle O, Borgwardt K, Weigel D (2011). Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* **480**, 245–249. <https://doi.org/10.1038/nature10555>
- Berger F, Grini PE, Schnittger A (2006). Endosperm: an integrator of seed growth and development. *Current Opinion in Plant Biology* **9**, 664–670. <https://doi.org/10.1016/j.pbi.2006.09.015>
- Berger F, Vu TM, Li J, Chen B (2012). Hypothesis: selection of imprinted genes is driven by silencing deleterious gene activity in somatic tissues. *Cold Spring Harbor Symposia on Quantitative Biology* **77**, 23-29. <https://doi.org/10.1101/sqb.2012.77.014514>
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Chaudhury AM, Ming L, Miller C, Craig S, Dennis ES, Peacock WJ (1997). Fertilization-independent seed development in *Arabidopsis thaliana*. *PNAS* **94**, 4223–4228. <https://doi.org/10.1073/pnas.94.8.4223>
- Chen C, Begcy K, Liu K, Folsom JJ, Wang Z, Zhang C, Walia H (2016). Heat stress

- 633 yields a unique MADS box transcription factor in determining seed size and
- 634 thermal sensitivity. *Plant Physiology* **171**, 606-622.
- 635 <https://doi.org/10.1104/pp.15.01992>
- 636 Cloutier S, Ragupathy R, Miranda E, Radovanovic N, Reimer E, Walichnowski A,
- 637 Ward K, Rowland G, Duguid S, Banik M (2012). Integrated consensus genetic and
- 638 physical maps of flax (*Linum usitatissimum* L.). *Theoretical and Applied Genetics*
- 639 **125**, 1783-1795. <https://doi.org/10.1007/s00122-012-1953-0>
- 640 Dong X 2017. The regulation of genetic imprinting in maize endosperm and its
- 641 stability and conservation. China Agricultural University.
- 642 Dong X, Zhang M, Chen J, Peng L, Zhang N, Wang X, Lai J (2017). Dynamic and
- 643 antagonistic allele-specific epigenetic modifications controlling the expression of
- 644 imprinted genes in maize endosperm. *Molecular Plant* **10**, 442-455.
- 645 <https://doi.org/10.1016/j.molp.2016.10.007>
- 646 Fu Y-B, Allaby RG (2010). Phylogenetic network of *Linum* species as revealed by
- 647 non-coding chloroplast DNA sequences. *Genetic Resources and Crop Evolution* **57**,
- 648 667-677. <https://doi.org/10.1007/s10722-009-9502-7>
- 649 Gehring M, Missirian V, Henikoff S (2011). Genomic analysis of parent-of-origin
- 650 allelic expression in *Arabidopsis thaliana* seeds. *Plos One* **6**, e23687.
- 651 <https://doi.org/10.1371/journal.pone.0023687>
- 652 Graaf Avd, Wardenaar R, Neumann DA, Taudt A, Shaw RG, Jansen RC, Schmitz RJ,
- 653 Colomé-Tatché M, Johannes F (2015). Rate, spectrum, and evolutionary dynamics
- 654 of spontaneous epimutations. *PNAS* **112**, 6676-6681.
- 655 <https://doi.org/10.1073/pnas.1424254112>
- 656 Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C (2010).
- 657 High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse
- 658 Brain. *Science* **329**, 643-648. <https://doi.org/10.1126/science.1190830>
- 659 Guitton A-E, Page DR, Chambrier P, Lionnet C, Faure J-E, Grossniklaus U, Berger F
- 660 (2004). Identification of new members of Fertilisation Independent Seed Polycomb
- 661 Group pathway involved in the control of seed development in *Arabidopsis*
- 662 *thaliana*. *Development* **131**, 2971-2981. <https://doi.org/10.1242/dev.01168>
- 663 Guo D, Jiang H, Yan W, Yang L, Ye J, Wang Y, Yan Q, Chen J, Gao Y, Duan L, Liu H,
- 664 Xie L (2020). Resequencing 200 Flax Cultivated Accessions Identifies Candidate
- 665 Genes Related to Seed Size and Weight and Reveals Signatures of Artificial
- 666 Selection. *Frontiers in Plant Science* **10**, 1682.
- 667 <https://doi.org/10.3389/fpls.2019.01682>
- 668 Hanna CW, Kelsey G (2017). Genomic imprinting beyond DNA methylation: a role
- 669 for maternal histones. *Genome Biology* **18**, 177.
- 670 <https://doi.org/10.1186/s13059-017-1317-9>
- 671 Hobson N, Deyholos MK (2013). Genomic and expression analysis of the flax (*Linum*
- 672 *usitatissimum*) family of glycosyl hydrolase 35 genes. *BMC Genomics* **14**, 344.
- 673 <https://doi.org/10.1186/1471-2164-14-344>
- 674 Hsieh T-F, Shin J, Uzawa R, Silva P, Cohen S, Bauer MJ, Hashimoto M, Kirkbride

- RC, Harada JJ, Zilberman D, Fischer RL (2011). Regulation of imprinted gene expression in *Arabidopsis* endosperm. *PNAS* **108**, 1755-1762. <https://doi.org/10.1073/pnas.1019273108>
- Huang F, Zhu Qh, Zhu A, Wu X, Xie L, Wu X, Helliwell C, Chaudhury A, Finnegan EJ, Luo M (2017). Mutants in the imprinted *PICKLE RELATED 2* gene suppress seed abortion of fertilization independent seed class mutants and paternal excess interploidy crosses in *Arabidopsis*. *The Plant Journal* **90**, 383–395
- Huang X, Lu Z, Wang X, Ouyang Y, Chen W, Xie K, Wang D, Luo M, Luo J, Yao J (2016). Imprinted gene *OsFIE1* modulates rice seed development by influencing nutrient metabolism and modifying genome H3K27me3. *The Plant Journal* **87**, 305–317.
- Huis R, Hawkins S, Neutelings G (2010). Selection of reference genes for quantitative gene expression normalization in flax (*Linum usitatissimum* L.). *BMC Plant Biology* **10**, 71. <https://doi.org/10.1186/1471-2229-10-71>
- Hutter B, Bieg M, Helms V, Paulsen M (2010). Divergence of imprinted genes during mammalian evolution. *BMC Evolutionary Biology* **10**, 116. <https://doi.org/10.1186/1471-2148-10-116>
- Joanis V, Lloyd VK (2002). Genomic imprinting in *Drosophila* is maintained by the products of *Suppressor of variegation* and *trithorax* group, but not *Polycomb* group, genes. *Molecular Genetics and Genomics* **268**, 103-112. <https://doi.org/10.1007/s00438-002-0731-0>
- Kinoshita T, Yadegari R, Harada JJ, Goldberg RB, Fischer RL (1999). Imprinting of the *MEDEA* polycomb gene in the *Arabidopsis* endosperm. *The Plant Cell* **11**, 1945-1952. <https://doi.org/10.2307/3871089>
- Kiyosue T, Ohad N, Yadegari R, Hannon M, Dinneny JR, Wells D, Katz A, Margossian L, Harada JJ, Goldberg RB, Fischer RL (1999). Control of fertilization-independent endosperm development by the *MEDEA* polycomb gene in *Arabidopsis*. *PNAS* **96**, 4186-4191. <https://doi.org/10.1073/pnas.96.7.4186>
- Köhler C, Page DR, Gagliardini V, Grossniklaus U (2005). The *Arabidopsis thaliana* *MEDEA* Polycomb group protein controls expression of *PHERES1* by parental imprinting. *Nature Genetics* **37**, 28-30. <https://doi.org/10.1038/ng1495>
- Kradolfer D, Wolff P, Jiang H, Siretskiy A, Köhler C (2013). An imprinted gene underlies postzygotic reproductive isolation in *Arabidopsis thaliana*. *Developmental Cell* **26**, 525–535. <https://doi.org/10.1016/j.devcel.2013.08.006>
- Li Q, Song J, West PT, Zynda G, Eichten SR, Vaughn MW, Springer NM (2015). Examining the causes and consequences of context-specific differential DNA methylation in maize. *Plant Physiology* **168**, 1262-1274. <https://doi.org/10.1104/pp.15.00052>
- Li Q, Wang J, Ye J, Zheng X, Xiang X, Li C, Fu M, Wang Q, Zhang Z, Wu Y (2017). The maize imprinted gene *Floury3* encodes a PLATZ protein required for tRNA and 5S rRNA transcription through interaction with RNA polymerase III. *The Plant Cell* **29**, 2661-2675. <https://doi.org/10.1105/tpc.17.00576>

- Librado P, Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452. <https://doi.org/10.1093/bioinformatics/btp187>
- Lu J, Zhang C, Baulcombe DC, Chen ZJ (2012). Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of *Arabidopsis* seeds. *PNAS* **109**, 5529-5534. <https://doi.org/10.1073/pnas.1203094109>
- Luo M, Bilodeau P, Dennis ES, Peacock WJ, Chaudhury A (2000). Expression and parent-of-origin effects for *FIS2*, *MEA*, and *FIE* in the endosperm and embryo of developing *Arabidopsis* seeds. *PNAS* **97**, 10637-10642. <https://doi.org/10.1073/pnas.170292997>
- Luo M, Bilodeau P, Koltunow A, Dennis ES, Peacock WJ, Chaudhury AM (1999). Genes controlling fertilization-independent seed development in *Arabidopsis thaliana*. *PNAS* **96**, 296-301. <https://doi.org/10.1073/pnas.96.1.296>
- Luo M, Taylor JM, Spriggs A, Zhang H, Wu X, Russell S, Singh M, Koltunow A (2011). A genome-wide survey of imprinted genes in rice seeds reveals imprinting primarily occurs in the endosperm. *PLoS Genetics* **7**, e1002125. <https://doi.org/10.1371/journal.pgen.1002125>
- Mahajan M, Ahuja PS, Yadav SK (2011). Post-transcriptional silencing of flavonol synthase mRNA in tobacco leads to fruits with arrested seed set. *Plos One* **6**, e28315. <https://doi.org/10.1371/journal.pone.0028315>
- Mei X, Liu C, Yu T, Liu X, Xu D, Wang J, Wang G, Cai Y (2015). Identification and characterization of paternal-preferentially expressed gene *NF-YC8* in maize endosperm. *Molecular Genetics and Genomics* **290**, 1819-1831. <https://doi.org/10.1007/s00438-015-1043-5>
- Meng D, Zhao J, Zhao C, Luo H, Jin W (2018). Sequential gene activation and gene imprinting during early embryo development in maize. *The Plant Journal* **93**, 445-459.
- Miyake T, Takebayashi N, Wolf DE (2009). Possible diversifying selection in the imprinted gene, *MEDEA*, in *Arabidopsis*. *Molecular Biology and Evolution* **26**, 843-857. <https://doi.org/10.1093/molbev/msp001>
- Moreno-Romero J, León GDT-D, Yadav VK, Santos-González J, Köhler C (2019). Epigenetic signatures associated with imprinted paternally expressed genes in the *Arabidopsis* endosperm. *Genome Biology* **20**, 41. <https://doi.org/10.1186/s13059-019-1652-0>
- Ohad N, Margossian L, Hsu Y-C, Williams C, Repetti P, Fischer RL (1996). A mutation that allows endosperm development without fertilization. *PNAS* **93**, 5319-5324. <https://doi.org/10.1073/pnas.93.11.5319>
- Ohad N, Yadegari R, Margossian L, Hannon M, Michaeli D, Harada JJ, Goldberg RB, Fischer RL (1999). Mutations in *FIE*, a WD polycomb group gene, allow endosperm development without fertilization. *The Plant Cell* **11**, 407-415. <https://doi.org/10.1105/tpc.11.3.407>
- Pignatta D, Erdmann RM, Scheer E, Picard CL, Bell GW, Gehring M (2014). Natural

- epigenetic polymorphisms lead to intraspecific variation in *Arabidopsis* gene imprinting. *eLife* **3**, e03198. <https://doi.org/10.7554/eLife.03198>
- Piskurewicz U, Iwasaki M, Susaki D, Megies C, Kinoshita T, Lopez-Molina L (2016). Dormancy-specific imprinting underlies maternal inheritance of seed dormancy in *Arabidopsis thaliana*. *eLife* **5**, e19573. <https://doi.org/10.7554/elife.19573>
- Renfree MB, Suzuki S, Kaneko-Ishino T (2012). The origin and evolution of genomic imprinting and viviparity in mammals. *Philosophical Transactions of the Royal Society B Biological Sciences* **368**, 20120151. <https://doi.org/10.1098/rstb.2012.0151>
- Rodrigues JA, Zilberman D (2015). Evolution and function of genomic imprinting in plants. *Genes and Development* **29**, 2517–2531. <https://doi.org/10.1101/gad.269902.115>
- Sabelli PA, Larkins BA (2009). The development of endosperm in grasses. *Plant Physiology* **149**, 14–26. <https://doi.org/10.1104/pp.108.129437>
- Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011). Transgenerational epigenetic instability is a source of novel methylation variants. *Science* **334**, 369–373. <https://doi.org/10.1126/science.1212959>
- Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, Ecker JR (2013). Patterns of population epigenomic diversity. *Nature* **495**, 193–198. <https://doi.org/10.1038/nature11968>
- Scott RJ, Spielman M, Bailey J, Dickinson HG (1998). Parent-of-origin effects on seed development in *Arabidopsis thaliana*. *Development* **125**, 3329–3341.
- Sekine D, Ohnishi T, Furuumi H, Ono A, Yamada T, Kurata N, Kinoshita T (2013). Dissection of two major components of the post-zygotic hybridization barrier in rice endosperm. *The Plant Journal* **76**, 792–799. <https://doi.org/10.1111/tpj.12333>
- Sha K (2008). A mechanistic view of genomic imprinting. *Annu Rev Genomics Hum Genet* **9**, 197–216. <https://doi.org/10.1146/annurev.genom.122007.110031>
- Spillane C, Schmid KJ, Laoueillé-Duprat S, Pien S, Escobar-Restrepo J-M, Baroux C, Gagliardini V, Page DR, Wolfe KH, Grossniklaus U (2007). Positive darwinian selection at the imprinted *MEDEA* locus in plants. *Nature* **448**, 349–352. <https://doi.org/10.1038/nature05984>
- Sreenivasulu N, Wobus U (2013). Seed-development programs: A systems biology-based comparison between dicots and monocots. *Annual Review of Plant Biology* **64**, 189–217. <https://doi.org/10.1146/annurev-arplant-050312-120215>
- Sun C, Tang T, Tang W, Sui L, Zhang H, Xia Z, Zhao H, Han Y, Lin L (2017). Plant genomic imprinting and its effect on seed development. *Agricultural Science and Technology* **18**, 984–987. <https://doi.org/10.16175/j.cnki.1009-4229.2017.06.006>
- Sun R-h 2018. Characterization and Functional Analysis of MADS-box from coconut (*Cocos nucifera* L.) endosperm. [Hainan University.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Baren MJv, Salzberg SL, Wold BJ, Pachter L (2010). Transcript assembly and quantification by RNA-Seq

reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28**, 511-515. <https://doi.org/10.1038/nbt.1621>

Tuteja R, McKeown PC, Ryan P, Morgan CC, Donoghue MTA, Downing T, O'Connell MJ, Spillane C (2019). Paternally expressed imprinted genes under positive Darwinian selection in *Arabidopsis thaliana*. *Molecular Biology and Evolution* **36**, 1239–1253. <https://doi.org/10.1093/molbev/msz063>

Voorrips RE (2002). MapChart: software for the Graphical Presentation of Linkage Maps and QTLs. *Journal of Heredity* **93**, 77-78. <https://doi.org/10.1093/jhered/93.1.77>

Wang A, Hou Q, Si L, Huang X, Luo J, Lu D, Zhu J, Shangguan Y, Miao J, Xie Y, Wang Y, Zhao Q, Feng Q, Zhou C, Li Y, Fan D, Lu Y, Tian Q, Wang Z, Han B (2019). The PLATZ transcription factor GL6 affects grain length and number in rice. *Plant Physiology* **180**, 2077-2090. <https://doi.org/10.1104/pp.18.01574>

Wang L, Yuan J, Ma Y, Jiao W, Ye W, Yang D, Yi C, Chen ZJ (2018). Rice interploidy crosses disrupt epigenetic regulation, gene expression, and seed development. *Molecular Plant* **11**, 300-314. <https://doi.org/10.1016/j.molp.2017.12.006>

Wang Z, Hobson N, Galindo L, Zhu S, Shi D, McDill J, Yang L, Hawkins S, Neutelings G, Datla R, Lambert G, Galbraith DW, Grassa CJ, Gerald A, Cronk QC, Cullis C, Dash PK, Kumar PA, Cloutier S, Sharpe AG, Wong GK-S, Wang J, Deyholos MK (2012). The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *The Plant Journal* **72**, 461–473. <https://doi.org/10.1111/j.1365-313X.2012.05093.x>

Waters AJ, Bilinski P, Eichten SR, Vaughn MW, Ross-Ibarra J, Gehring M, Springer NM (2013). Comprehensive analysis of imprinted genes in maize reveals allelic variation for imprinting and limited conservation with other species. *PNAS* **110**, 19639–19644. <https://doi.org/10.1073/pnas.1309182110>

Waters AJ, Makarevitch I, Eichten SR, Swanson-Wagner RA, Yeh C-T, Xu W, Schnable PS, Vaughn MW, Gehring M, Springer NM (2011). Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. *The Plant Cell* **23**, 4221-4233. <https://doi.org/10.1105/tpc.111.092668>

Wolff P, Jiang H, Wang G, Santos-González J, Köhler C (2015). Paternally expressed imprinted genes establish postzygotic hybridization barriers in *Arabidopsis thaliana*. *eLife* **4**, e10074. <https://doi.org/10.7554/eLife.10074>

Wolff P, Weinhofer I, Seguin J, Roszak P, Beisel C, Donoghue MTA, Spillane C, Nordborg M, Rehmsmeier M, Köhler C (2011). High-resolution analysis of parent-of-origin allelic expression in the *Arabidopsis* endosperm. *PLoS Genetics* **7**, e1002126. <https://doi.org/10.1371/journal.pgen.1002126>

Xu J, Chen G, Hermanson PJ, Xu Q, Sun C, Chen W, Kan Q, Li M, Crisp PA, Yan J, Li L, Springer NM, Li Q (2019). Population-level analysis reveals the widespread occurrence and phenotypic consequence of DNA methylation variation not tagged by genetic variation in maize. *Genome Biology* **20**, 243. <https://doi.org/10.1186/s13059-019-1859-0>

- Xu W, Dai M, Li F, Liu A (2014). Genomic imprinting, methylation and parent-of-origin effects in reciprocal hybrid endosperm of castor bean. *Nucleic Acids Research* **42**, 6987-6998. <https://doi.org/10.1093/nar/gku375>
- Yang X, Wu F, Lin X, Du X, Chong K, Gramzow L, Schilling S, Becker A, Theißen G, Meng Z (2012). Live and let die - the B_{sister} MADS-box gene *OsMADS29* controls the degeneration of cells in maternal tissues during seed development of rice (*Oryza sativa*). *Plos One* **7**, e51435. <https://doi.org/10.1371/journal.pone.0051435>
- Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, Wang J (2006). WEGO : a web tool for plotting GO annotations. *Nucleic Acids Research* **34**, W293-W297. <https://doi.org/10.1093/nar/gkl031>
- Yuan J, Chen S, Jiao W, Wang L, Wang L, Ye W, Lu J, Hong D, You S, Cheng Z, Yang D, Chen ZJ (2017). Both maternally and paternally imprinted genes regulate seed development in rice. *New Phytologist* **216**, 373–387.
- Zhang J, Qi Y, Wang L, Wang L, Yan X, Dang Z, Li W, Zhao W, Pei X, Li X, Liu M, Tan M, Wang L, Long Y, Wang J, Zhang X, Dang Z, Zheng H, Liu T (2020). Genomic comparison and population diversity analysis provide insights into the domestication and improvement of flax. *iScience* **23**, 100967. <https://doi.org/10.1016/j.isci.2020.100967>
- Zhang M, Li N, He W, Zhang H, Yang W, Liu B (2016a). Genome-wide screen of genes imprinted in sorghum endosperm, and the roles of allelic differential cytosine methylation. *The Plant Journal* **85**, 424-436.
- Zhang M, Zhao H, Xie S, Chen J, Xu Y, Wang K, Zhao H, Guan H, Hu X, Jiao Y, Song W, Lai J (2011). Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. *PNAS* **108**, 20042–20047. <https://doi.org/10.1073/pnas.1112186108>
- Zhang S, Feng D, Tian X, Yang X (2003). Progress on imprinted genes in mammals. *Progress in Biotechnology* **23**, 48-54. <https://doi.org/10.3969/j.issn.1671-8135.2003.12.011>
- Zhang W, Cao S, Jiang L, Zhu S, Wan J (2005). Genomic imprinting and seed development. *Hereditas* **27**, 665-670. <https://doi.org/10.3321/j.issn:0253-9772.2005.04.029>
- Zhang X-J, Jiang W-B, Pang Y-Z (2016b). Advances in the regulation mechanism of plant seed size. *Plant Physiology Journal* **52**, 998-1010. <https://doi.org/10.13592/j.cnki.ppj.2016.0159>
- Zhou S-R, Xue H-W (2020). The rice PLATZ protein SHORT GRAIN6 determines grain size by regulating spikelet hull cell division. *Journal of Integrative Plant Biology* **62**, 847-864. <https://doi.org/10.1111/jipb.12851>

Figure legends

Figure 1. Identification of imprinted genes in flax endosperm. (A) The proportion of parental transcripts in both CZ and ZC was plotted for 52,794 SNPs with at least ten reads could be assigned to a specific allele. The shaded areas indicated moderate

(pink), strong (blue), or complete (arrows) maternally expressed imprinted SNP loci (upper right) or paternally expressed imprinted SNP loci (lower left). (B) The number of moderate (pink), strong (blue), and complete (gray) imprinted SNP loci and imprinted genes in endosperm. ME-SNPs, maternally expressed imprinted SNP loci; PE-SNPs, paternally expressed imprinted SNP loci; MEG, maternally expressed imprinted genes; PEG, paternally expressed imprinted gene.

Figure 2. The conservation of flax imprinted genes between other species. (A) Venn diagram showing overlaps of imprinted genes at E-value<1E-10 between flax and maize, rice, *Arabidopsis*, castor bean, sorghum, respectively. (B) Venn diagram showing overlaps of imprinted genes at E-value<1E-20. (C) Venn diagram showing overlaps of imprinted genes at E-value<1E-50.

Figure 3. Some flax imprinted genes are located in mini-clusters. The MEGs (red) and PEGs (blue) were mapped to the 15 flax chromosomes. Genes clustered within 10 kb are boxed in green.

Figure 4. Expression of flax imprinted genes over various flax tissues in both reciprocal hybrids based on RNA-seq analysis. (A, B) The gene-expression patterns for MEGs (A) and PEGs (B). endo-MEGs, MEGs that expressed preferentially in endosperm; con-MEGs, MEGs that also expressed in other tissues; endo-PEGs, PEGs that expressed preferentially in endosperm; con-PEGs, PEGs that also expressed in other tissues. The normalized values were used for hierarchical clustering and the heat map indicated relative levels of expression. The endosperm and embryo tissues were harvested at 7 DAP and the leaf tissues were collected at 2 weeks after planting. For each sample, three biological replicates were used. (C, D) The Log₁₀(FPKM) values of CZ (C) and ZC (D). The Log₁₀(FPKM) values of CZ (C) and ZC (D). All genes with FPKM>1 in endosperm were used in this study. The values listed above box plots were the number of genes in each group. Asterisks indicate the significance level (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

Figure 5. Imprinted genes differentiate flax subgroups. (A) The correlation of kinships between imprinted SNPs or all SNPs. (B) Phylogenetic tree of 200 flax accessions inferred from imprinted SNPs. (C) PCA plots of all imprinted SNPs, ME-SNPs and PE-SNPs. Fiber flax, oil-fiber dual purpose flax (OF), and Oil flax were represented in red, blue and green colors, respectively.

Figure 6. Distribution of nucleotide diversity (π) within imprinted genes and allele frequency differences of two genes across oil and fiber subgroups. (A-C) Boxplots for nucleotide diversity of all imprinted genes (A), MEGs (B) and PEGs (C) across Oil (green) and Fiber (red) groups. (D, G) The distribution of allele frequency of SNPs located in *Lus10010350* (D, PEG) and *Lus10024230* (G, MEG) in Oil and Fiber subgroups. The alternate alleles and reference alleles were shown in purple and orange, respectively. (E, H) The nucleotide diversity distribution of *Lus10010350* on chromosome 15 (E) and *Lus10024230* on chromosome 12 (H) among Oil and Fiber

subgroups. (F, I) Boxplots for nucleotide diversity of *Lus10010350* (F) and *Lus10024230* (I) among Oil and Fiber subgroups. The difference was analyzed by two-tailed *t* tests.

Figure 7. Imprinted gene-based association study for seed size and weight, and identification of a causal gene for the peak on chromosome 15. (A-B) The overlapping Manhattan plots for seed length, seed width and 1,000-seed weight in four environments (including 2016DL, 2017UR, 2019UR, 2019YL) using GLM (A) and MLM (B) models. Imprinted genes repeatedly detected at least two environments or traits related to seed size and weight were marked by red arrows. (C) Local Manhattan plots for seed width in 2016DL of imprinted gene-based association analysis surrounding the peak on chromosome 15. The position of *Lus10010350* was highlighted by a shaded pink column. (D) Gene structure of *Lus10010350*. (E) DNA polymorphism in *Lus10010350*. (F) Schematic representation of the structural variation in *Lus10010350*. (G-J) Boxplots for 1,000-seed weight based on the haplotypes (Hap.) for *Lus10010350* in 2016DL (G), 2017UR (H), 2019UR (I) and 2019YL (J). In the box plots, the center line represented the median, box limits indicated the upper and lower quartiles, whiskers marked the range of the data and points showed outliers. *n* indicated the number of accessions with the same genotype. The difference between haplotypes was analyzed by two-tailed *t* tests.

Supplementary information

Figure S1. Flow chart for identification of imprinted genes in flax endosperm.

Figure S2. Verification of thirteen genes in flax endosperm based on qRT-PCR analysis. Thirteen genes were chosen for the qRT-PCR analyses. Among these genes, five were MEGs, five were PEGs, and others were not imprinted genes. So, the gene expression level between qRT-PCR and RNA sequencing of these thirteen genes represented the whole types of genes in this study.

Figure S3. Validation of the imprinted genes in flax endosperm by PCR sequencing. Twelve imprinted genes including nine MEGs and three PEGs were selected for validation. Each gene was designed by a pair of primers with a 400-800bp amplification fragment which was a part of the corresponding CDS sequence of CC (endosperm of Cili2719 self-cross), ZZ (endosperm of Z11637 self-cross), CZ (endosperm of Cili2719×Z11637), and ZC (endosperm of Z11637×Cili2719) and the amplification fragment contained at least one imprinted SNP site.

Figure S4. Gene ontology analysis of 229 identified imprinted genes and 115 conserved imprinted genes. (A) Gene ontology analysis of identified imprinted genes. MEGs represented 135 maternally expressed genes (red), PEGs represented 94 paternally expressed genes (purple), All imprinted genes represented 229 moderate imprinted genes (blue), All genes represented all endosperm-expressed genes with at least ten reads could be assigned to a specific allele in both CZ and ZC (yellow). (B)

Gene ontology analysis of conserved imprinted genes. Conserved MEGs represented 75 conserved maternally expressed genes (red), Conserved PEGs represented 40 conserved paternally expressed genes (purple), Conserved imprinted genes represented 115 conserved imprinted genes (blue), All genes were presented as A.

Figure S5. Expression of conserved imprinted genes in different tissues of flax in both reciprocal hybrids CZ and ZC based on RNA-seq analysis. (A, B) The gene-expression patterns for MEGs (A) and PEGs (B) of 115 conserved imprinted genes. (C) The gene-expression patterns for MEGs and PEGs of 38 conserved strong imprinted genes. endo-MEGs, MEGs that expressed preferentially in endosperm; con-MEGs, MEGs that expressed in many tissues; endo-PEGs, PEGs that expressed preferentially in endosperm; con-PEGs, PEGs that expressed in many tissues. The normalized values were used for hierarchical clustering and the heat map indicates relative levels of expression. The endosperm and embryo tissues were harvested at 7 DAP and the leaf tissues were collected at 2 weeks after planting. For each sample, three biological replicates were used.

Figure S6. MEGs and PEGs can differentiate flax subgroups. (A) Phylogenetic tree of 200 flax accessions inferred from ME-SNPs. (B) Phylogenetic tree of 200 flax accessions inferred from PE-SNPs. ME-SNPs, maternally expressed imprinted SNP loci; PE-SNPs, paternally expressed imprinted SNP loci. Fiber flax, oil-fiber dual purpose flax (OF), and Oil flax were represented in red, blue and green colors, respectively.

Figure S7. Boxplots for seed length and seed width based on the haplotypes (Hap.) for *Lus10010350*. (A-B) The seed length (A) and seed width (B) in 2016DL. (C-D) The seed length (C) and seed width (D) in 2017UR. (E-F) The seed length (E) and seed width (F) in 2019UR. (G-H) The seed length (G) and seed width (H) in 2019YL. In the box plots, the center line represented the median, box limits indicated the upper and lower quartiles, whiskers marked the range of the data and points showed outliers. *n* indicates the number of accessions with the same genotype. The difference between haplotypes was analyzed by two-tailed *t* tests.

Figure S8. The nucleotide diversity distribution of *Lus10041386*. (A) The nucleotide diversity distribution of *Lus10041386* on chromosome 15 among Oil and Fiber subgroups. (B) Boxplots for nucleotide diversity of *Lus10041386* among Oil and Fiber subgroups. The difference was analyzed by two-tailed *t* tests.

Table S1. Imprinted genes in both hybrid endosperms and associated SNPs.

Table S2. Functional annotations of 229 imprinted protein-coding genes.

Table S3. Gene ontology enrichment analysis of 229 flax imprinted genes.

Table S4. The conservation of imprinted genes detected in our endosperm samples.

Table S5. The number of conserved flax imprinted genes with other species.

1001 **Table S6.** Gene ontology enrichment analysis of 115 conserved imprinted genes in
1002 flax.

1003 **Table S7.** Clusters of imprinted genes.

1004 **Table S8.** Imprinted genes related to seed size and 1000-seed weight based on
1005 candidate gene-based association study.

1006 **Table S9.** Primers for qRT-PCR.

1007 **Table S10.** Primers for imprinting validation by PCR-sequencing.

1008

Figure 1. Identification of imprinted genes in flax endosperm. (A) The proportion of parental transcripts in both CZ and ZC was plotted for 52,794 SNPs with at least ten reads could be assigned to a specific allele. The shaded areas indicated moderate (pink), strong (blue), or complete (arrows) maternally expressed imprinted SNP loci (upper right) or paternally expressed imprinted SNP loci (lower left). (B) The number of moderate (pink), strong (blue), and complete (gray) imprinted SNP loci and imprinted genes in endosperm. ME-SNPs, maternally expressed imprinted SNP loci; PE-SNPs, paternally expressed imprinted SNP loci; MEG, maternally expressed imprinted genes; PEG, paternally expressed imprinted gene.

Figure 2. The conservation of flax imprinted genes between other species. (A) Venn diagram showing overlaps of imprinted genes at E-value<1E-10 between flax and maize, rice, *Arabidopsis*, castor bean, sorghum, respectively. (B) Venn diagram showing overlaps of imprinted genes at E-value<1E-20. (C) Venn diagram showing overlaps of imprinted genes at E-value<1E-50.

Figure 3. Some flax imprinted genes are located in mini-clusters. The MEGs (red) and PEGs (blue) were mapped to the 15 flax chromosomes. Genes clustered within 10 kb are boxed in green.

Figure 4. Expression of flax imprinted genes over various flax tissues in both reciprocal hybrids based on RNA-seq analysis. (A, B) The gene-expression patterns for MEGs (A) and PEGs (B). endo-MEGs, MEGs that expressed preferentially in endosperm; con-MEGs, MEGs that also expressed in other tissues; endo-PEGs, PEGs that expressed preferentially in endosperm; con-PEGs, PEGs that also expressed in other tissues. The normalized values were used for hierarchical clustering and the heat map indicated relative levels of expression. The endosperm and embryo tissues were harvested at 7 DAP and the leaf tissues were collected at 2 weeks after planting. For each sample, three biological replicates were used. (C, D) The Log₁₀ (FPKM) values of CZ (C) and ZC (D). The Log₁₀ (FPKM) values of CZ (C) and ZC (D). All genes with FPKM>1 in endosperm were used in this study. The values listed above box plots were the number of genes in each group. Asterisks indicate the significance level (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

Figure 5. Imprinted genes differentiate flax subgroups. (A) The correlation of kinships between imprinted SNPs or all SNPs. (B) Phylogenetic tree of 200 flax accessions inferred from imprinted SNPs. (C) PCA plots of all imprinted SNPs, ME-SNPs and PE-SNPs. Fiber flax, oil-fiber dual purpose flax (OF), and Oil flax were represented in red, blue and green colors, respectively.

Figure 6. Distribution of nucleotide diversity (π) within imprinted genes and allele frequency differences of two genes across oil and fiber subgroups. (A-C) Boxplots for nucleotide diversity of all imprinted genes (A), MEGs (B) and PEGs (C) across Oil (green) and Fiber (red) groups. (D, G) The distribution of allele frequency of SNPs located in *Lus10010350* (D, PEG) and *Lus10024230* (G, MEG) in Oil and Fiber subgroups. The alternate alleles and reference alleles were shown in purple and orange, respectively. (E, H) The nucleotide diversity distribution of *Lus10010350* on chromosome 15 (E) and *Lus10024230* on chromosome 12 (H) among Oil and Fiber subgroups. (F, I) Boxplots for nucleotide diversity of *Lus10010350* (F) and *Lus10024230* (I) among Oil and Fiber subgroups. The difference was analyzed by two-tailed *t* tests.

Figure 7. Imprinted gene-based association study for seed size and weight, and identification of a causal gene for the peak on chromosome 15. (A-B) The overlapping Manhattan plots for seed length, seed width and 1,000-seed weight in four environments (including 2016DL, 2017UR, 2019UR, 2019YL) using GLM (A) and MLM (B) models. Imprinted genes repeatedly detected at least two environments or traits related to seed size and weight were marked by red arrows. (C) Local Manhattan plots for seed width in 2016DL of imprinted gene-based association analysis surrounding the peak on chromosome 15. The position of *Lus10010350* was highlighted by a shaded pink column. (D) Gene structure of *Lus10010350*. (E) DNA polymorphism in *Lus10010350*. (F) Schematic representation of the structural variation in *Lus10010350*. (G-J) Boxplots for 1,000-seed weight based on the haplotypes (Hap.) for *Lus10010350* in 2016DL (G), 2017UR (H), 2019UR (I) and 2019YL (J). DL, Dali; UR, Urumqi; YL, YiLi. In the box plots, the center line represented the median, box limits indicated the upper and lower quartiles, whiskers marked the range of the data and points showed outliers. *n* indicated the number of accessions with the same genotype. The difference between haplotypes was analyzed by two-tailed *t* tests.