The MADS-box transcription factor PHERES1 controls imprinting in the endosperm by binding to domesticated transposons

Rita A. Batista[1], Jordi Moreno-Romero[1,2], Joram van Boven[1], Yichun Qiu[1], Juan Santos-González[1], Duarte D. Figueiredo[1,3], Claudia Köhler[1*]

1. Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences and Linnean Centre for Plant Biology, Uppsala, Sweden
2. Present address: Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus UAB, Bellaterra, Barcelona, Spain
3. Present address: Institute for Biochemistry and Biology, University of Potsdam, Germany

* Corresponding author (claudia.kohler@slu.se)

Abstract

MADS-box transcription factors (TFs) are ubiquitous among eukaryotes, and classified into two groups: type I or SRF (Serum Response Factor)-like, and type II or MEF2 (Myocyte Enhancing Factor2)-like[1]. In flowering plants, type I MADS-box TFs are associated with reproductive development and many are active in the endosperm, a nutritive tissue supporting the embryo[2]. Deregulation of these genes has been frequently linked to failure of endosperm development and seed inviability, both in the Brassicaceae[3–6], and in crop species like tomato and rice[7,8]. Nevertheless, a mechanistic explanation for these observations, clarifying the role of MADS-box TFs in endosperm development, remains to be established. Here we show that the imprinted *Arabidopsis thaliana* MADS-box TF PHERES1 (PHE1)[9] has a central role in endosperm development as a master regulator of imprinted gene expression, especially of paternally expressed genes (PEGs), which have been previously implicated in endosperm development[5,10–12]. Control of imprinted gene expression by PHE1 is mediated by parental asymmetry of epigenetic modifications in PHE1 DNA-binding sites, conferring different accessibilities to maternal and paternal alleles. Importantly, we show that the CArG-box-like DNA-binding motifs used by PHE1 to access gene promoters are carried by RC/Helitron transposable elements (TEs), providing an example of molecular domestication of these elements. Hence, this work shows that TEs are intrinsically linked to imprinting: not only by enforcing specific epigenetic landscapes[13–15], but also by serving as important sources of *cis*-regulatory elements. Moreover, it provides an example of how TEs can widely distribute TF binding sites in a plant genome, allowing to recruit crucial endosperm regulators into a single transcriptional network.

Main

PHE1 was the first described PEG in *Arabidopsis*[9]. To identify the genes regulated by PHE1, we performed a ChIP-seq experiment using a line expressing *PHE1::PHE1-GFP* **(Supplementary Table 1-2)**. The 1942 identified PHE1 target genes were enriched for Gene Ontology (GO)-terms associated with development, metabolic processes, and transcriptional regulation, showing that many PHE1 targets are themselves transcriptional regulators **(Extended Data Fig. 1a)**. Among PHE1 targets are several known regulators of endosperm development like *AGL62*, *YUC10*, *IKU2*, *MINI3*, and *ZHOUPI*, revealing a central role for PHE1 in regulating endosperm development

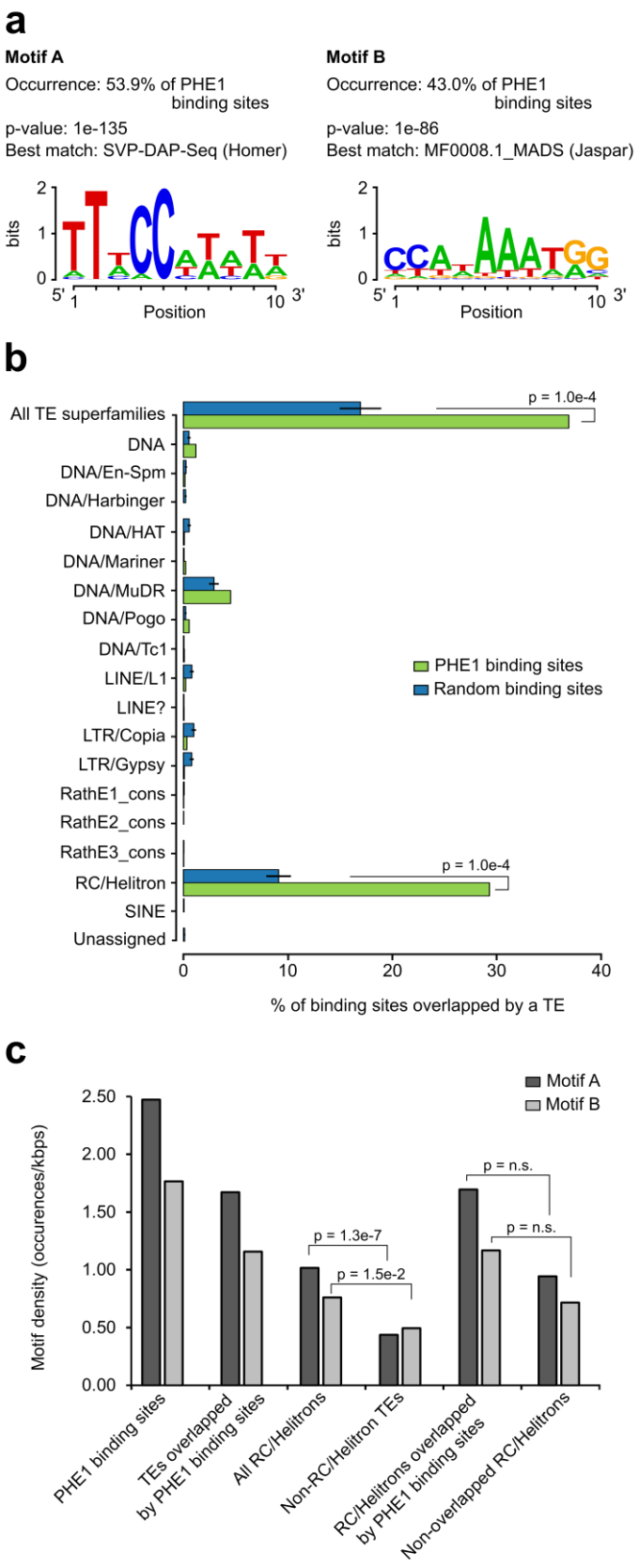


Fig. 1 | RC/Helitrons carry PHE1 DNA-binding motifs. a, CArG-box-like DNA-binding motifs identified from PHE1 ChIP-seq data. **b,** Fraction of PHE1 binding sites (green) overlapping transposable elements (TEs.) Overlap is expressed as the percentage of total binding sites where spatial intersection with features on the y-axis is observed. A set of random binding sites (blue, see Methods) is used as control. P-values were determined using Monte Carlo permutation tests (see Methods). Bars represent ± sd, (n = 2494, PHE1 binding sites, Random binding sites). **c,** Density of PHE1 motifs in different genomic regions of interest. P-values were determined using χ^2 tests.

(Extended Data Table 1). Importantly, type I MADS-box family genes were over-represented among PHE1 targets **(Extended Data Fig. 1b)**, pointing to a high degree of cross-regulation among members of this family.

Our data revealed that PHE1 uses two distinct DNA-binding motifs. Motif A was present in about 53% of PHE1 binding sites, while motif B could be found in 43% of those **(Fig. 1a)**. PHE1 motifs closely resemble type II CArG-boxes, the signature motif of MADS-box TFs[16] **(Fig. 1a, Extended Data Fig. 2)**, suggesting that DNA-binding properties between type I and type II members are conserved.

Strikingly, PHE1 binding sites significantly overlapped with TEs, preferentially with those of the RC/Helitron superfamily (29% in PHE1 binding sites, versus 9% in random binding sites) **(Fig. 1b)**, and in particular with ATREP10D elements **(Extended Data Fig. 3)**. We thus addressed the question whether RC/Helitrons contain sequence properties that promote PHE1 binding. Indeed, a screen of genomic regions for the presence of PHE1 DNA-binding motifs, revealed significantly higher motif densities within all RC/Helitrons, compared to other TE superfamilies **(Fig. 1c)**. Interestingly, even though motif densities were higher in RC/Helitrons overlapped by PHE1 binding sites than in non-overlapped RC/Helitrons, this difference was not significant **(Fig. 1c)**. This reveals that the enrichment of PHE1 DNA-binding motifs is a specific feature of RC/Helitrons, suggesting that the exaptation of these TEs as *cis*-regulatory regions facilitates TF binding and modulation of gene expression. Furthermore, we found that RC/Helitron sequences associated with both PHE1 DNA-binding motifs could be grouped into multiple clusters (127 for motif A and 97 for motif B) based on aligned homologous sequences around the motifs. Thus,

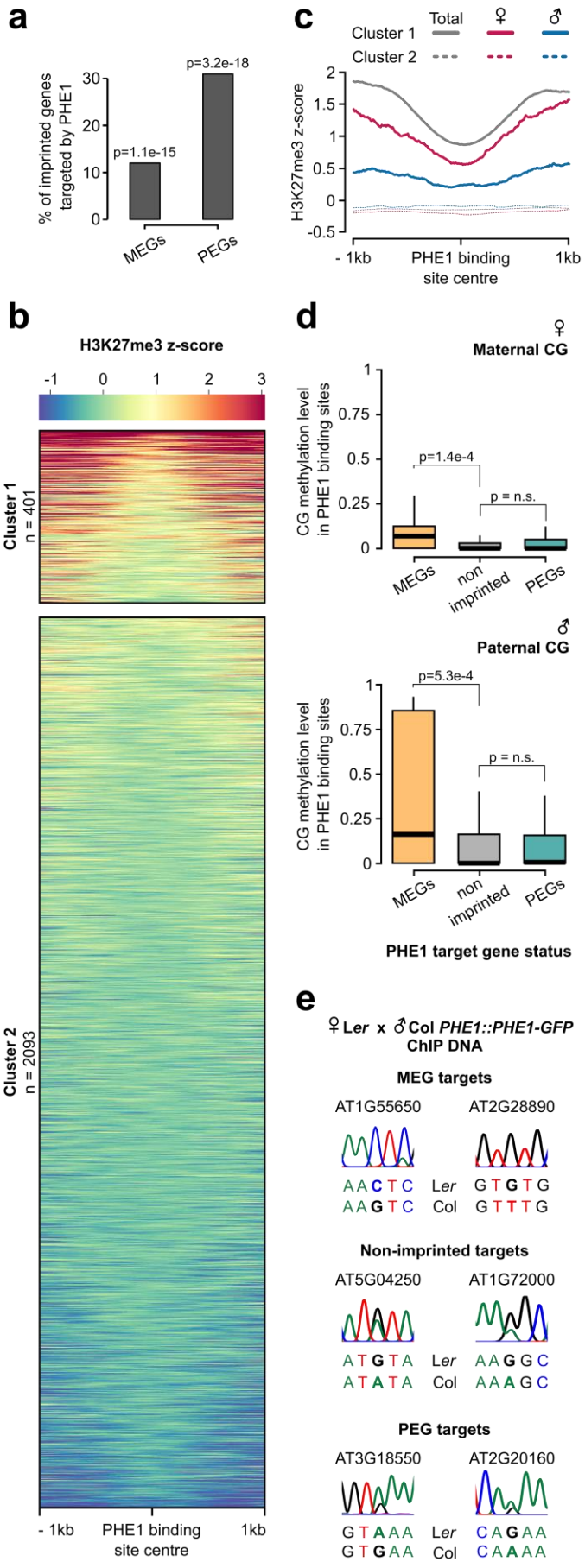


Fig. 2 | Parental asymmetry of epigenetic marks in imprinted gene promoters conditions PHE1 binding. a, Fraction of imprinted genes targeted by PHE1. P-values were determined using the hypergeometric test **b,** Heatmap of endosperm H3K27me3 distribution along PHE1 binding sites. Each horizontal line represents one binding site. Clusters were defined based on the pattern of H3K27me3 distribution (see Methods) **c,** Metagene plot of average maternal (♀, pink), paternal (♂, blue) and total (grey) endosperm H3K27me3 along PHE1 binding sites. **d,** CG methylation levels in maternal (♀, upper panel) and paternal (♂, lower panel) alleles of PHE1 binding sites associated with MEGs (yellow), PEGs (green) and non-imprinted (grey) PHE1 targets. P-values were determined using two-tailed Mann-Whitney tests. **e,** Sanger sequencing of imprinted and non-imprinted gene promoters bound by PHE1. SNPs for maternal (Ler) and paternal (Col) alleles are shown (n = 1 biological replicate).

this suggests that multiple ancestral RC/Helitrons played a role in acquiring these motifs, which were subsequently amplified in the genome through transposition.

Among the PHE1 target genes we detected a significant enrichment of imprinted genes, with 12% of all maternally expressed genes (MEGs) and 31% of all PEGs being targeted **(Fig. 2a)**. Imprinted gene expression relies on parental-specific epigenetic modifications, which are asymmetrically established during male and female gametogenesis, and inherited in the endosperm[17,18]. Demethylation of repeat sequences and TEs – occurring in the central cell, but not in sperm – is a major driver for imprinted gene expression[17,18]. In MEGs, DNA hypomethylation of maternal alleles leads to their expression, while DNA methylation represses the paternal allele[17,18]. In PEGs, the hypomethylated maternal allele undergoes trimethylation of lysine 27 of histone H3 (H3K27me3), a repressive histone modification established by the Fertilization Independent Seed-Polycomb Repressive Complex2 (FIS-PRC2). This renders the maternal alleles inactive, while the paternal allele is expressed[14,17,18].

Given the significant overrepresentation of imprinted genes among PHE1 targets, we assessed how the epigenetic landscape at those loci correlates with

DNA-binding by PHE1. We surveyed levels of endosperm H3K27me3 within PHE1 binding sites and identified two distinct clusters **(Fig. 2b)**. Cluster 1 was characterized by an accumulation of H3K27me3 in regions flanking the centre of the binding site, while the centre itself was devoid of this mark. Cluster 2 contained binding sites largely devoid of H3K27me3. The distribution of H3K27me3 in cluster 1 was mostly attributed to the deposition of H3K27me3 on the maternal alleles, while the paternal alleles were devoid of this mark **(Fig. 2c)** – a pattern usually associated with PEGs[14]. Consistently, genes associated with cluster 1 binding sites had a more paternally-biased expression in the endosperm when compared to genes associated with cluster 2 **(Extended Data Fig. 4a)**. This is reflected by more PEGs and putative PEGs being associated with cluster 1 **(Extended Data Fig. 4b)**. We also identified parental-specific differences in DNA methylation, specifically in the CG context: PHE1 binding sites associated with MEGs had significantly higher methylation levels in paternal alleles than in maternal alleles **(Fig. 2d)**.

We hypothesized that the differential parental deposition of epigenetic marks in PHE1 binding sites can result in differential accessibility of these regions, which might therefore impact transcription in a parent-of-origin-specific manner. To test this, we performed ChIP using a L*er* maternal plant and a Col *PHE1::PHE1-GFP* pollen donor, taking advantage of SNPs between these two accessions to discern parental preferences of PHE1 binding. Using Sanger sequencing, we determined the parental origin of enriched ChIP-DNA in MEG, non-imprinted, and PEG targets **(Fig. 2e, Extended Data Fig. 5)**. While binding of PHE1 was biallelic in non-imprinted targets **(Fig 2e)**, only maternal binding was detected in the tested MEG targets, supporting the idea that CG hypermethylation of paternal alleles prevents their binding by PHE1 **(Fig. 2e, Extended Data Fig. 6a)**. Interestingly, we observed biallelic binding in PEG targets **(Fig. 2e)**. Even though the maternal PHE1 binding sites in PEGs were flanked by H3K27me3 **(Fig. 2b-c)**, correlating with transcriptional repression of maternal alleles, the absence of this mark within the binding site centres seems to be permissive for maternal PHE1 binding. We speculate that the accessibility of this site might be important to mediate recruitment of H3K27me3 in the central cell and/or for maintenance of this mark during endosperm development **(Extended Data Fig. 6a)**.

Previous studies have shown that PEGs are often flanked by RC/Helitrons[19], a phenomenon suggested to lead to the parental asymmetry of epigenetic marks in these genes[13,14]. Consistent with our finding that PHE1 binding sites overlapped with RC/Helitrons **(Fig. 1b)**, we found that PHE1 DNA-binding motifs were contained within these TEs significantly more frequently in PEGs than in non-imprinted genes **(Extended Data Fig. 6b)**. Furthermore, we detected the presence of homologous RC/Helitrons containing PHE1 binding motifs in the promoter regions of several PHE1-targeted PEG orthologs **(Extended Data Fig. 7)**, indicating ancestral insertion events. The presence of these RC/Helitrons correlated with paternally-biased expression of the associated orthologs, providing further support to the hypothesis these TEs contribute to the gain of imprinting, especially of PEGs. Thus, besides facilitating the asymmetry of epigenetic marks, these TEs can contribute to the generation of novel gene promoters that ensure the timely endosperm expression of PEGs, under the control of PHE1, and possibly other type I MADS-box TFs **(Extended Data Fig. 6a)**.

Among the PEGs targeted by PHE1 were *ADM*, *SUVH7*, *PEG2*, and *NRPD1a* **(Extended Data Table 1)**. Mutants in all four PEGs suppress the abortion of triploid (3x) seeds generated by paternal excess interploidy crosses[10,20]. Furthermore, we found that close to 50% of highly upregulated genes in 3x seeds are targeted by PHE1 **(Fig. 3a)**, suggesting this TF might play a central role in mediating the strong gene deregulation observed in these seeds. If this is true, removal of PHE1 is expected to alleviate gene deregulation and prevent 3x seed inviability. To test this hypothesis, we generated a *phe1* CRISPR/Cas9 mutant in the *phe2* background, since both genes are likely

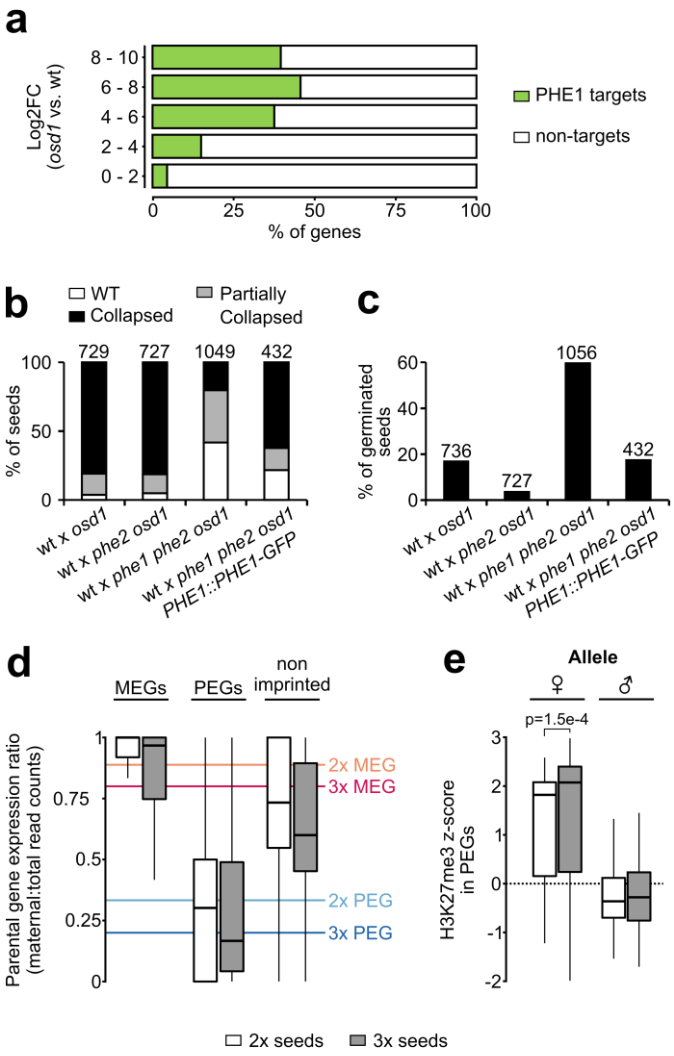


Fig. 3 | PHE1 establishes 3x seed inviability of paternal excess crosses. a, Target status of upregulated genes in paternal excess crosses. Highly upregulated genes in 3x seeds are more often targeted by PHE1 ($p<2.2e-16$, χ^2 test). **b-c,** Seed inviability phenotype (b) of paternal excess crosses in wild-type (wt), *phe1 phe2*, and *phe1* complementation lines, with respective seed germination rates (c). Remaining control crosses are shown in Extended Data Fig. 8b-c. *n* = numbers on top of bars (seeds). **d,** Parental expression ratio of imprinted genes in the endosperm of 2x (white) and 3x seeds (grey). Solid lines indicate the ratio thresholds for definition of MEGs and PEGs, in 2x and 3x seeds. **e,** Accumulation of H3K27me3 across maternal (♀) and paternal (♂) gene bodies of PEGs in the endosperm of 2x and 3x seeds (white and grey, respectively). H3K27me3 accumulation in MEGs and non-imprinted genes is shown in Extended Data Fig. 9. P-value was determined using a two-tailed Mann-Whitney test.

redundant[21] **(Extended Data Fig. 8a)**. We introduced *phe1 phe2* into the *omission of second division 1 (osd1)* mutant background, which produces diploid gametes at high frequency[22]. Wild-type (wt) and *phe2* maternal plants pollinated with *osd1* pollen form 3x seeds that abort at high frequency[23] **(Figure 3b, Extended Data Fig. 8b)**. In contrast, *phe1 phe2 osd1* pollen strongly suppressed 3x seed inviability, reflected by the increased germination of 3x *phe1 phe2* seeds **(Fig. 3c, Extended Data Fig. 8c)**. This phenotype could be reverted by introducing the *PHE1::PHE1-GFP* transgene paternally **(Fig. 3b-c, Extended Data Fig. 8b-c)**. Notably, 3x seed rescue was mostly mediated by *phe1*, as the presence of a wt *PHE2* allele in 3x seeds (wt x *phe1 phe2 osd1*) led to comparable rescue levels than when having no wt *PHE2* allele present (*phe2* x *phe1 phe2 osd1*) **(Fig. 3b-c, Extended Data Fig. 8b-c)**. Importantly, 3x seed rescue was accompanied by reestablishment of endosperm cellularisation **(Extended Data Fig. 8d-e)**, and reduced expression of PHE1 target genes **(Extended Data Fig. 8f)**.

Loss of FIS-PRC2 function causes a similar phenotype to that of paternal excess 3x seeds, correlating with largely overlapping sets of deregulated genes[3,5]. Since FIS-PRC2 is a major regulator of PEGs in *Arabidopsis* endosperm[14], we addressed the question whether imprinting is disrupted in 3x seeds. To assess this, we analysed the parental expression ratio of imprinted genes in the endosperm of 2x and 3x seeds. Surprisingly, imprinting was not disrupted in 3x seeds **(Fig. 3d)**, consistent with similar levels of H3K27me3 on the maternal alleles of PEGs in the endosperm of 2x and 3x seeds **(Fig. 3e, Extended Data Fig. 9)**. Collectively, these data show that the major upregulation of imprinted gene expression in 3x seeds is due to increased transcription of the active allele, likely mediated by PHE1 and other MADS-box TFs, with maintenance of the imprinting status.

In summary, this work reveals that the MADS-box TF PHE1 is a major regulator of imprinted genes in the *Arabidopsis* endosperm, and that this TF establishes a reproductive barrier in response to interploidy hybridizations. We furthermore show that deregulated PEGs in 3x seeds remain imprinted, but that the active allele becomes strongly overexpressed, correlating with

increased PHE1 activity in these seeds[5]. Importantly, we reveal a novel role for RC/Helitrons in the regulation of imprinted genes by showing that they contain PHE1 DNA-binding sites. Our data favour a scenario where these elements have been domesticated to function as providers of *cis*-regulatory sequences that facilitate transcription of imprinted genes. Thus, this study provides an example of TE-mediated distribution of TF binding sites throughout a flowering plant genome, adding support to the long-standing idea that transposition facilitates the formation of *cis*-regulatory architectures required to control complex biological processes[24–26]. We speculate that this process may have contributed to endosperm evolution by allowing the recruitment of crucial developmental genes into a single transcriptional network, regulated by type I MADS-box TFs. The diversification of the mammalian placenta has been connected with the dispersal of hundreds of placenta-specific enhancers by endogenous retroviruses[27], suggesting that the convergent evolution of the endosperm in flowering plants and the mammalian placenta have been promoted by TE transpositions.

Methods

Plant material and growth conditions

Arabidopsis thaliana seeds were sterilized in a closed vessel containing chlorine gas, for 3 hours. Chlorine gas was produced by mixing 3 mL HCl 37% and 100 mL of 100% commercial bleach. Sterile seeds were plated in ½ MS-medium (0.43% MS salts, 0.8% Bacto Agar, 0.19% MES hydrate) supplemented with 1% Sucrose. When required, appropriate antibiotics were supplemented to the medium. Seeds were stratified for 48 h, at 4°C, in darkness. Plates containing stratified seeds were transferred to a long-day growth chamber (16h light / 8h dark; 110 µmol $s^{-1}m^{-2}$; 21ºC; 70% humidity), where seedlings grew for 10 days. After this period, the seedlings were transferred to soil and placed in a long-day growth chamber.

Several mutant lines used in this study have been previously described: *osd1-1*[22], *osd1-3*[28] and *pi-1*[29]. The *phe2* allele corresponds to a T-DNA insertion mutant (SALK_105945). Phenotypical analysis of this mutant revealed no deviant phenotype relative to Col wt plants (data not shown). Genotyping of *phe2* was done using the following primers (PHE2 fw 5'-AAATGTCTGGTTTTATGCCCC-3', PHE2 rv 5'-GTAGCGAGACAATCGATTTCG-3', T-DNA 5'-ATTTTGCCGATTTCGGAAC-3').

Generation of phe1 phe2

The *phe1 phe2* double mutant was generated using the CRISPR/Cas9 technique. A 20-nt sgRNA targeting *PHE1* was designed using the CRISPR Design Tool[30]. A single-stranded DNA oligonucleotide corresponding to the sequence of the sgRNA, as well as its complementary oligonucleotide, were synthesized. BsaI restriction sites were added at the 5' and 3' ends, as represented by the underlined sequences (sgRNA fw 5'-ATTGCTCCTGGATCGAGTTGTAC-3'; sgRNA rv 5'-AAACGTACAACTCGATCCAGGAG-3'). These two oligonucleotides were then annealed to produce a double stranded DNA molecule.

The double-stranded oligonucleotide was ligated into the egg-cell specific pHEE401E CRISPR/Cas9 vector[31] through the BsaI restriction sites. This vector was transformed into the *Agrobacterium tumefaciens* strain GV3101, and *phe2/-* plants were subsequently transformed using the floral-dip method[32].

To screen for T1 mutant plants, we performed Sanger sequencing of *PHE1* amplicons derived from these plants, and obtained with the following primers (fw 5'-AGTGAGGAAAACAACATTCACCA-3'; rv 5'-GCATCCACAACAGTAGGAGC-3'). The selected mutant contained a homozygous two base pair deletion that leads to a premature stop codon, and therefore a truncated $PHE1_{1\text{-}50aa}$ protein. In the T2 generation, the segregation of pHEE401E allowed to select plants that did not contain this vector and that were double homozygous *phe1 phe2* mutants. Genotyping of the *phe1* allele was done using primers fw 5'-AAGGAAGAAAGGGATGCTGA-3' and rv 5'-

TCTGTTTCTTTGGCGATCCT-3', followed by RsaI digestion.

Seed imaging

Analysis of endosperm cellularisation status was done following the Feulgen staining protocol described previously[33]. Imaging of Feulgen-stained seeds was done in a Zeiss LSM780 NLO multiphoton microscope with excitation wavelength of 800 nm, and acquisition between 520 nm – 695 nm.

RT-qPCR

RNA extraction, cDNA synthesis and qPCR analyses for *AGL* and *PEG* expression were performed as described previously[33]. Two biological replicates per cross were used. Primer sequences for *YUC10*, *AGL62*, and *PP2A* were described previously[33]. For the remaining genes, primer sequences were as follows: *AGL48* (fw 5'- TTCGCGATCCACCAGTGTTT-3', rv 5'-GACCGCCTCCTACAAAACCA-3'), *AGL90* (fw 5'-TTGGTGATGAGTCGTTTTCCGA-3', rv 5'-TCATATTCGCATTTGCGTCCG-3').

Chromatin immunoprecipitation (ChIP)

To find targets of PHE1, we performed a ChIP experiment using a reporter line containing the PHE1 protein tagged with GFP. This reporter line, which we denote as *PHE1::PHE1-GFP*, contains the *PHE1* promoter, its coding sequence and its 3' regulatory sequence, and is present in a Col background[34]. Given the presence of the 3' regulatory regions[35], this reporter behaves as a paternally expressed gene, similarly to the endogenous *PHE1* gene. Crosslinking of plant material was done by collecting 600 mg of 2 days after pollination (DAP) *PHE1::PHE1-GFP* siliques, and vacuum infiltrating them with a 1% formaldehyde solution in PBS. The vacuum infiltration was done for two periods of 15 min, with a vacuum release between each period. The crosslinking was then stopped by adding 0.125 mM glycine in PBS and performing a vacuum infiltration for a total of 15 min, with a vacuum release each 5 min. The material was then ground in liquid nitrogen, resuspended in 5mL Honda buffer[36], and incubated for 15 min with gentle rotation. This mixture was filtered twice through Miracloth and one time through a CellTrics filter (30 µm), after which a centrifugation for 5 min, at 4°C and 1,500 g was performed. The nuclei pellet was then resuspended in 100 µL of nuclei lysis buffer[36], and the ChIP protocol was continued as described before[36]. ChIP DNA was isolated using the Pure Kit v2 (Diagenode), following the manufacturer's instructions.

For the parental-specific PHE1 ChIP the starting material consisted of 600 mg of 2DAP siliques from crosses between a *pi-1* mother (L*er* ecotype) and a *PHE1::PHE1-GFP* father (Col ecotype). The male sterile *pi-1* mutant was used to avoid emasculation of maternal plants. Crosslinking of plant material, nuclei isolation, ChIP protocol, and ChIP-DNA purification were the same as described before.

To assess parental-specific H3K27me3 profiles in 3x seeds, the INTACT system was used to isolate 4DAP endosperm of seeds derived from L*er pi-1* x Col *INTACT osd1-1* crosses, as described previously[36,37]. *osd1* mutants were used for their ability to generate unreduced male gametes, leading to the formation of 3x seeds. ChIPs against H3 and H3K27me3 were then performed on the isolated endosperm nuclei, following the previously described protocol[36].

The antibodies used for these ChIP experiments were as follows: GFP Tag Antibody (A-11120, Thermo Fisher Scientific), anti-H3 (Sigma, H9289), and anti-H3K27me3 (Millipore, cat. no. 07-449). All experiments were performed with two biological replicates.

Library preparation and sequencing

PHE1::PHE1-GFP ChIP libraries were prepared using the Ovation Ultralow v2 Library System (NuGEN), with a starting material of 1 ng, following the manufacturer's instructions. These libraries were sequenced at the SciLife Laboratory (Uppsala, Sweden), on an Illumina HiSeq2500 platform, using 50 bp single-end reads.

Library preparation and sequencing of H3K27me3 ChIPs in 3x seeds was done as described previously[37].

Both datasets were deposited at NCBI's Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/), under the accession number GSE129744.

qPCR and Sanger sequencing of parental-specific PHE1 ChIP

Purified ChIP DNA and its respective input DNA obtained from the parental-specific PHE1 ChIP were used to perform qPCR. Positive and negative genomic regions for PHE1 binding were amplified using the following primers: AT1G55650 (fw 5'-CGAAGCGAAAAAGCACTCAC-3'; rv 5'-CCTTTTACATAATCCGCGTTAAA-3'), AT2G28890 (fw 5'-TTTGTGGTTGGAGGTTGTGA-3'; rv 5'-GTTGTTCGTGCCCATTTCTT-3'), AT5G04250 (fw 5'-AATTGACAAATGGTGTAATGGT-3'; rv 5'-CCAAAGAATTTGTTTTTCTATTCC-3'), AT1G72000 (fw 5'-AACAAATATGCACAAGAAGTGC-3'; rv 5'-ACCTAGCAAGCTGGCAAAAC-3'), AT3G18550 (fw 5'-TCCTTTTCCAAATAAAGGCATAA-3'; rv 5'-AAATGAAAGAAATAAAAGGTAATGAGA-3'), AT2G20160 (fw 5'-TCCTAAATAAGGGAAGAGAAAGCA-3'; rv 5'-TGTTAGGTGAAACTGAATCCAA-3'), negative region (fw 5'-TGGTTTTGCTGGTGATGATG-3', rv 5'-CCATGACACCAGTGTGCCTA-3'). HOT FIREPol EvaGreen qPCR Mix Plus (ROX) (Solis Biodyne) was used as a master mix for qPCR amplification in a iQ5 qPCR system (Bio-Rad).

For Sanger sequencing, positive genomic regions for PHE1 binding were amplified by PCR using the Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific), in combination with the primers described above. Amplified DNA was purified using the GeneJET PCR Purification kit (Thermo Fisher Scientific) and used for Sanger sequencing. The chromatograms obtained from Sanger sequencing were then analysed for the presence of SNPs.

Bioinformatic analysis of ChIP-seq data

For the *PHE1::PHE1-GFP* ChIP, reads were aligned to the *Arabidopsis* (TAIR10) genome using Bowtie version 1.2.2[38], allowing 2 mismatches (-v 2). Only uniquely mapped reads were kept. ChIP-seq peaks were called using MACS2 version 2.1.1, with its default settings[39]. Input samples served as control for their corresponding GFP ChIP sample. Each biological replicate was handled individually, and only the overlapping peak regions between the two replicates were considered for further analysis. These regions are referred throughout the text as PHE1 binding sites **(Supplementary Table 1)**. Peak overlap was determined with BEDtools version 2.26.0[40]. Each PHE1 binding site was annotated to a genomic feature and matched with a target gene using the peak annotation feature (annotatePeaks.pl) provided in HOMER version 4.9[41] **(Supplementary Table 2)**. Only binding sites located less than 3 kb away from the nearest transcription start site were considered.

PHE1 DNA-binding motifs were identified from *PHE1::PHE1-GFP* ChIP-seq peak regions with HOMER's findMotifsGenome.pl function, using the default settings. P-values of motif enrichment, as well as alignments between PHE1 motifs and known motifs were generated by HOMER.

Read mapping, coverage analysis, purity calculations, normalisation of data, and determination of parental origin of reads derived from H3K27me3 ChIPs in 3x seeds was done following previously published methods[14] **(Supplementary Table 3)**.

Analysis of PHE1 target genes

Significantly enriched Gene Ontology terms within target genes of PHE1 were identified using AtCOECIS[42], and further summarized using REVIGO[43].

Enrichment of specific transcription factor families within PHE1 targets was calculated by first normalizing the number of PHE1-targeted TFs in each family, to the

total number of TFs targeted by PHE1. As a control, the number of TFs belonging to a certain family was normalised to the total number of TFs in the *Arabidopsis* genome. The Log_2 fold change between these ratios was then calculated for each family. Significance of the enrichment was assessed using the hypergeometric test. Annotation of transcription factor families was done following the Plant Transcription Factor Database version 4.0[44]. Only TF families containing more than 5 members were considered in this analysis.

To determine which imprinted genes are targeted by PHE1, a custom list consisting of the sum of imprinted genes identified in different studies[13,19,45–47] was used.

To determine the proportion of genes overexpressed in paternal excess crosses that are targeted by PHE1, a previously published transcriptome dataset of 3x seeds was used[48].

Spatial overlap of TEs and PHE1 binding sites

Spatial overlap between PHE1 ChIP-seq peak regions (binding sites) and TEs was determined using the regioneR package version 1.8.1[49], implemented in R version 3.4.1[50]. As a control, a mock set of binding sites was created, to which we refer to as random binding sites. This random binding site set had the same total number of binding sites and the same size distribution as the PHE1 binding site set. Using regioneR, a Monte Carlo permutation test with 10000 iterations was performed. In each iteration the random binding sites were arbitrarily shuffled in the 3 kb promoter region of all *Arabidopsis thaliana* genes. From this shuffling, the average overlap and standard deviation of the random binding site set was determined, as well as the statistical significance of the association between PHE1 binding sites and TE superfamilies/families.

BedTools version 2.26.0[40] was used to determine the fraction of PHE1 binding sites targeting MEGs, PEGs, or non-imprinted genes where a spatial overlap between binding sites, RC/Helitrons and PHE1 DNA-binding motifs is simultaneously observed. The hypergeometric test was used to assess the significance of the enrichment of PHE1 binding sites where this overlap is observed, across different target types.

Calculation of PHE1 DNA-binding motif densities

To measure the density of PHE1 DNA-binding motifs within different genomic regions of interest, the fasta sequences of these regions were first obtained using BEDtools. HOMER's scanMotifGenomeWide.pl function was then used to screen these sequences for the presence of PHE1 DNA-binding motifs, and to count the number of occurrences of each motif. Motif density was then calculated as the number of occurrences of each motif, normalized to the size of the genomic region of interest. Chi-square tests of independence were used to test if there were any associations between specific genomic regions and PHE1 DNA-binding motifs. This was done by comparing the proportion of DNA bases corresponding to PHE1 DNA-binding motifs in each genomic region.

Identification of homologous PHE1 DNA-binding motifs carried by RC/Helitrons

To assess the homology of PHE1 DNA-binding motifs and associated RC/Helitron sequences, pairwise comparisons were made among all sequences, using the BLASTN program. The following parameters were followed: word size = 7, match/mismatch scores = 2/-3, gap penalties, existence = 5, extension = 2. The RC/Helitron sequences were considered to be homologous if the alignment covered at least 9 out of the 10 bp PHE1 DNA-binding motif sites, extended longer than 30 bp, and had more than 75% identity. Because the mean length of intragenomic conserved non-coding sequences is around 30 bp in *A. thaliana*[51], we considered this as the minimal length of alignments to define a pair of related motif-carrying TE sequences. We identified 1107 RC/Helitrons sequences out of 1232 that were associated with PHE1 A motifs and shared homology with at least one other sequence carrying a PHE1 A motif. For PHE1 B motifs, 675 out of 849 sequences had shared homology with at least one other sequence. The pairwise homologous sequences were

then merged in higher order clusters, based on shared elements in the homologous pairs.

Phylogenetic analyses of PHE1-targeted PEG orthologs in the Brassicaceae

Amino acid sequences and nucleotide sequences of PHE1-targeted PEGs were obtained from TAIR10. The sequences of homologous genes in the Brassicaceae and several other rosids were obtained in PLAZA 4.0 (https://bioinformatics.psb.ugent.be/plaza/)[52], BRAD database (http://brassicadb.org/brad/)[53], and Phytozome v.12 (https://phytozome.jgi.doe.gov/)[54].

For each PEG of interest, the amino acid sequences of the gene family were used to generate a guided codon alignment by MUSCLE with default settings[55]. A maximum likelihood tree was then generated by IQ-TREE 1.6.7 with codon alignment as the input[56]. The implemented ModelFinder was executed to determine the best substitution model[57], and 1000 replicates of ultrafast bootstrap were applied to evaluate the branch support[58]. The tree topology and branch supports were reciprocally compared with, and supported by another maximum likelihood tree generated using RAxML v. 8.1.2[59].

We selected PEGs that had well supported gene family phylogeny with no lineage-specific duplication in the *Arabidopsis* and *Capsella* clades, and where imprinting data were available for all *Capsella grandiflora* (Cgr), *C. rubella* (Cru), and *Arabidopsis lyrata* (Aly) orthologs of interest[13,60,61]. We then obtained the promoter region, defined as 3 kb upstream of the translation start site, of the orthologs and paralogs in Brassicaceae and rosids species These promoter sequences were searched for the presence of homologous RC/Helitron sequences, as well as for putative PHE1 DNA-binding sites contained in these TEs.

Epigenetic profiling of PHE1 binding sites

Parental-specific H3K27me3 profiles[14] and DNA methylation profiles[48] generated from endosperm of 2x seeds were used for this analysis. Levels of H3K27me3 and CG DNA methylation were quantified in each 50 bp bin across the 2 kb region surrounding PHE1 binding site centres using deepTools version 2.0[62]. These values were then used to generate H3K27me3 heatmaps and metagene plots, as well as boxplots of CG methylation in PHE1 binding sites. Clustering analysis of H3K27me3 distribution in PHE1 binding sites was done following the k-means algorithm as implemented by deepTools. A two-tailed Mann-Whitney test with continuity correction was used to assess statistical significance of differences in CG methylation levels.

Parental gene expression ratios in 2x and 3x seeds

To determine parental gene expression ratios in 2x and 3x seeds, we used previously generated endosperm gene expression data[20]. In this dataset, L*er* plants were used as maternal plants pollinated with wt Col or *osd1* Col plants, allowing to determine the parental origin of sequenced reads following the method described before[63].

Parental gene expression ratios were calculated as the number of maternally-derived reads divided by the sum of maternally- and paternally-derived reads available for any given gene. Ratios were calculated separately for the two biological replicates of each cross (L*er* x wt Col and L*er* x *osd1* Col), and the average of both replicates was considered for further analysis. The MEG and PEG ratio thresholds for 2x and 3x seeds indicated in Figure 3d were defined as a four-fold deviation of the expected read ratios, towards more maternal or paternal read accumulation, respectively. The expected read ratio for a biallelically expressed gene in 2x seeds is 2 maternal reads : 3 total reads, while for 3x paternal excess seeds this ratio is 2 maternal reads : 4 total reads. Deviations from these expected ratios were used to classify the expression of published imprinted genes[13,19,45–47]as maternally or paternally biased in 3x seeds, according to the direction of the deviation. As a control, the parental bias of these imprinted genes was also assessed in 2x seeds.

Parental expression ratios of genes associated with H3K27me3 clusters

Previously published endosperm gene expression data, generated with the INTACT system, was used for this analysis[64]. Parental gene expression ratios were determined as the mean between ratios observed in the L*er* x Col cross and its reciprocal cross. As a reference, the parental gene expression ratio for all endosperm expressed genes was also determined. A two-tailed Mann-Whitney test with continuity correction was used to assess statistical significance of differences between parental gene expression ratios.

H3K27me3 accumulation in imprinted genes

Parental-specific accumulation of H3K27me3 across imprinted gene bodies in the endosperm of 2x[14] and 3x seeds (this study) was estimated by calculating the mean values of the H3K27me3 z-score across the gene length. Imprinted genes were considered as those genes previously identified in different studies[13,19,45–47]. A two-tailed Mann-Whitney test with continuity correction was used to assess statistical significance of differences in H3K27me3 z-score levels.

Statistics

Sample size, statistical tests used, and respective p-values are indicated in each figure or figure legend, and further specified in the corresponding Methods sub-section.

Data availability

ChIP-seq data generated in this study is available at NCBI's Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/), under the accession number GSE129744. Additional data used to support the findings of this study are available at NCBI's Gene Expression Omnibus, under the following accession numbers: H3K27me3 ChIP-seq data from 2x endosperm[14] - GSE66585; Gene expression data in 2x and 3x endosperm[20] - GSE84122; Gene expression data in 2x and 3x seeds and parental-specific DNA methylation from 2x endosperm[48] - GSE53642; Parental-specific gene expression data of 2x INTACT-isolated endosperm nuclei[64] - GSE119915.

Acknowledgments

We thank Qi-Jun Chen for providing the pHEE401E CRISPR/Cas9 vector. We are grateful to Cecilia Wärdig for technical assistance. Sequencing was performed by the SNP&SEQ Technology Platform, Science for Life Laboratory at Uppsala University, a national infrastructure supported by the Swedish Research Council (VRRFI) and the Knut and Alice Wallenberg Foundation. This research was supported by a grant from the Swedish Research Council (to C.K.), a grant from the Knut and Alice Wallenberg Foundation (to C.K.), and support from the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine.

Author contributions

R.A.B., J.M-R., Y.Q, D.D.F and C.K. performed the experimental design. R.A.B., J.M-R., J.V.B. and Y.Q. performed experiments. R.A.B., J.M-R., Y.Q., J.S-G., and C.K. analysed the data. R.A.B and C.K. wrote the manuscript. All authors read and commented on the manuscript.

Competing interests

The authors declare no competing interests

Materials and correspondence

The materials generated in this study are available upon request to C.K. (claudia.kohler@slu.se).

References

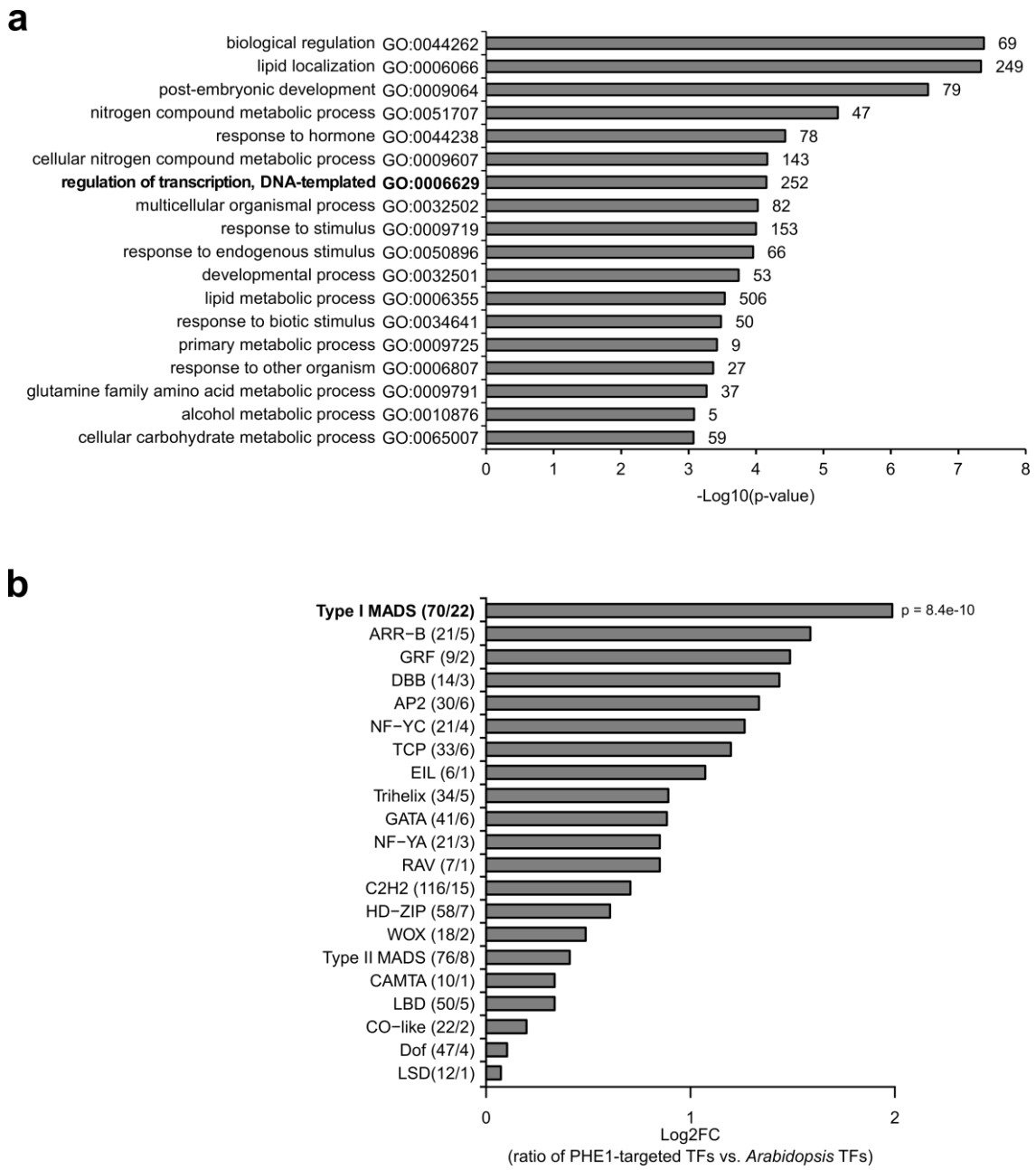
1. Gramzow, L. & Theissen, G. A hitchhiker's guide to the MADS world of plants. *Genome Biology* **11**, 214 (2010).
2. Bemer, M., Heijmans, K., Airoldi, C., Davies, B. & Angenent, G. C. An atlas of type I MADS box gene expression during female gametophyte and seed development in Arabidopsis. *Plant Physiology* **154**, 287–300 (2010).
3. Tiwari, S. *et al.* Transcriptional profiles underlying parent-of-origin effects in seeds of Arabidopsis

thaliana. *BMC Plant Biology* **10**, 72 (2010).

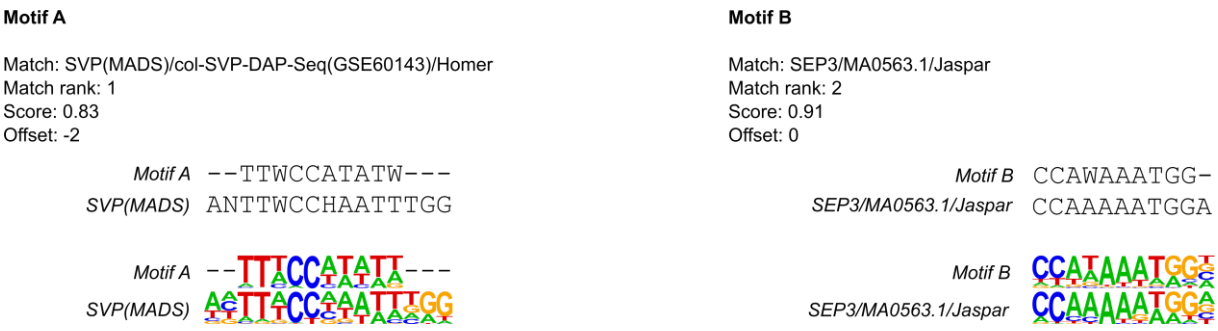
4. Lu, J., Zhang, C., Baulcombe, D. C. & Chen, Z. J. Maternal siRNAs as regulators of parental genome imbalance and gene expression in endosperm of Arabidopsis seeds. *PNAS* **109**, 5529–5534 (2012).
5. Erilova, A. *et al*. Imprinting of the Polycomb group gene MEDEA serves as a ploidy sensor in Arabidopsis. *PLoS Genetics* **5**, e1000663 (2009).
6. Rebernig, C. A., Lafon-Placette, C., Hatorangan, M. R., Slotte, T. & Köhler, C. Non-reciprocal Interspecies Hybridization Barriers in the Capsella Genus Are Established in the Endosperm. *PLoS Genetics* **11**, e1005295 (2015).
7. Roth, M., Florez-Rueda, A. M. & Städler, T. Differences in Effective Ploidy Drive Genome-Wide Endosperm Expression Polarization and Seed Failure in Wild Tomato Hybrids. *Genetics* genetics.302056.2019 (2019). doi:10.1534/genetics.119.302056
8. Ishikawa, R. *et al.* Rice interspecies hybrids show precocious or delayed developmental transitions in the endosperm without change to the rate of syncytial nuclear division. *Plant Journal* **65**, 798–806 (2011).
9. Köhler, C., Page, D. R., Gagliardini, V. & Grossniklaus, U. The Arabidopsis thaliana MEDEA Polycomb group protein controls expression of PHERES1 by parental imprinting. *Nature Genetics* **37**, 28–30 (2005).
10. Wolff, P., Jiang, H., Wang, G., Santos-González, J. & Köhler, C. Paternally expressed imprinted genes establish postzygotic hybridization barriers in Arabidopsis thaliana. *eLife* **4**, (2015).
11. Figueiredo, D. D., Batista, R. A., Roszak, P. J. & Köhler, C. Auxin production couples endosperm development to fertilization. *Nature Plants* **1**, 15184 (2015).
12. Jullien, P. E. & Berger, F. Parental genome dosage imbalance deregulates imprinting in Arabidopsis. *PLoS Genetics* **6**, e1000885 (2010).
13. Pignatta, D. *et al*. Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. *eLife* **3**, (2014).
14. Moreno- Romero, J., Jiang, H., Santos- González, J. & Köhler, C. Parental epigenetic asymmetry of PRC2- mediated histone modifications in the Arabidopsis endosperm. *The EMBO Journal* **35**, 1298–1311 (2016).
15. Gehring, M., Bubb, K. L. & Henikoff, S. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324**, 1447–51 (2009).
16. Folter, S. de & Angenent, G. C. trans meets cis in MADS science. *Trends in Plant Science* **11**, 224–231 (2006).
17. Gehring, M. Genomic Imprinting: Insights From Plants. *Annual Review of Genetics* **47**, 187–208 (2013).
18. Rodrigues, J. A. & Zilberman, D. Evolution and function of genomic imprinting in plants. *Genes and Development* **29**, 2517–2531 (2015).
19. Wolff, P. *et al.* High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Arabidopsis Endosperm. *PLoS Genetics* **7**, e1002126 (2011).
20. Martinez, G. *et al.* Paternal easiRNAs regulate parental genome dosage in Arabidopsis. *Nature Genetics* **50**, 193–198 (2018).
21. Villar, C. B. R., Erilova, A., Makarevich, G., Trösch, R. & Köhler, C. Control of PHERES1 Imprinting in Arabidopsis by Direct Tandem Repeats. *Molecular Plant* **2**, 654–660 (2009).
22. d'Erfurth, I. *et al.* Turning Meiosis into Mitosis. *PLoS Biology* **7**, e1000124 (2009).
23. Kradolfer, D., Wolff, P., Jiang, H., Siretskiy, A. & Köhler, C. An Imprinted Gene Underlies Postzygotic Reproductive Isolation in Arabidopsis thaliana. *Developmental Cell* **26**, 525–535 (2013).
24. Britten, R. J. & Davidson, E. H. Repetitive and Non-Repetitive DNA Sequences and a Speculation on the Origins of Evolutionary Novelty. *The Quarterly Review of Biology* **46**, 111–138 (1971).
25. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**, 397–405 (2008).
26. Hirsch, C. D. & Springer, N. M. Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta* **1860**, 157–165 (2017).
27. Chuong, E. B., Rumi, M. A. K., Soares, M. J. & Baker, J. C. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics* **45**, 325–329 (2013).
28. Heyman, J. *et al.* Arabidopsis ULTRAVIOLET-B-INSENSITIVE4 Maintains Cell Division Activity by Temporal Inhibition of the Anaphase-Promoting Complex/Cyclosome. *The Plant Cell* **23**, 4394–4410 (2011).
29. Goto, K. & Meyerowitz, E. M. Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. *Genes & Development* **8**, 1548–1560 (1994).
30. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
31. Wang, Z.-P. *et al.* Egg cell-specific promoter-controlled CRISPR/Cas9 efficiently generates homozygous mutants for multiple target genes in Arabidopsis in a single generation. *Genome Biology* **16**, 144 (2015).

32. Clough, S. J. & Bent, A. F. Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant Journal* **16**, 735–743 (1998).
33. Batista, R. A., Figueiredo, D. D., Santos-González, J. & Köhler, C. Auxin regulates endosperm cellularization in Arabidopsis. *Genes & Development* (2019). doi:10.1101/gad.316554.118
34. Weinhofer, I., Hehenberger, E., Roszak, P., Hennig, L. & Köhler, C. H3K27me3 profiling of the endosperm implies exclusion of polycomb group protein targeting by DNA methylation. *PLoS Genetics* **6**, 1–14 (2010).
35. Makarevich, G., Villar, C. B. R., Erilova, A. & Köhler, C. Mechanism of PHERES1 imprinting in Arabidopsis. *Journal of Cell Science* **121**, 906–912 (2008).
36. Moreno-Romero, J., Santos-González, J., Hennig, L. & Köhler, C. Applying the INTACT method to purify endosperm nuclei and to generate parental-specific epigenome profiles. *Nature Protocols* **12**, 238–254 (2017).
37. Jiang, H. *et al.* Ectopic application of the repressive histone modification H3K9me2 establishes post-zygotic reproductive isolation in Arabidopsis thaliana. *Genes & Development* **31**, 1272–1287 (2017).
38. Langmead, B. Aligning short sequencing reads with Bowtie. *Current Protocols in Bioinformatics* 11.7.1-11.7.14 (2010). doi:10.1002/0471250953.bi1107s32
39. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–2 (2010).
41. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
42. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. & Van de Peer, Y. Unraveling Transcriptional Control in Arabidopsis Using cis-Regulatory Elements and Coexpression Networks. *Plant Physiology* **150**, 535–546 (2009).
43. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
44. Jin, J., Zhang, H., Kong, L., Gao, G. & Luo, J. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research* **42**, D1182–D1187 (2014).
45. Gehring, M., Missirian, V. & Henikoff, S. Genomic analysis of parent-of-origin allelic expression in Arabidopsis thaliana seeds. *PLoS ONE* **6**, e23687 (2011).
46. Hsieh, T. F. *et al.* Regulation of imprinted gene expression in Arabidopsis endosperm. *PNAS* **108**, 1755–62 (2011).
47. Schon, M. A. & Nodine, M. D. Widespread Contamination of Arabidopsis Embryo and Endosperm Transcriptome Data Sets. *The Plant Cell* **29**, 608–617 (2017).
48. Schatlowski, N. *et al.* Hypomethylated Pollen Bypasses the Interploidy Hybridization Barrier in Arabidopsis. *The Plant Cell* **26**, 3556–3568 (2014).
49. Gel, B. *et al.* RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2015).
50. Core Team R. *R: A language and environment for statistical computing. R Foundation for Statistical Computing* (2017). doi:ISBN 3-900051-07-0
51. Thomas, B. C., Rapaka, L., Lyons, E., Pedersen, B. & Freeling, M. Arabidopsis intragenomic conserved noncoding sequence. *PNAS* **104**, 3348–53 (2007).
52. Van Bel, M. *et al.* PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* **46**, D1190–D1196 (2018).
53. Wang, X., Wu, J., Liang, J., Cheng, F. & Wang, X. Brassica database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. *Database* **2015**, bav093 (2015).
54. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**, D1178–D1186 (2012).
55. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).
56. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).
57. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587–589 (2017).
58. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**, 518–522 (2018).
59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

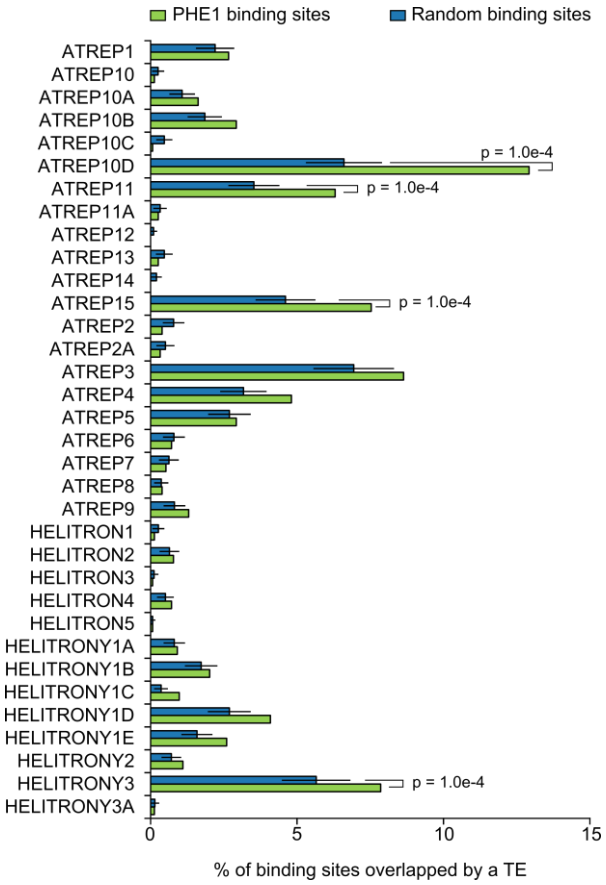
60. Klosinska, M., Picard, C. L. & Gehring, M. Conserved imprinting associated with unique epigenetic signatures in the Arabidopsis genus. *Nature Plants* **2**, (2016).
61. Lafon-Placette, C. *et al.* Paternally expressed imprinted genes associate with hybridization barriers in Capsella. *Nature Plants* **4**, 352–357 (2018).
62. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* **44**, W160–W165 (2016).
63. Moreno-Romero, J., Del Toro-De León, G., Yadav, V. K., Santos-González, J. & Köhler, C. Epigenetic signatures associated with imprinted paternally expressed genes in the Arabidopsis endosperm. *Genome Biology* **20**, 41 (2019).
64. Del Toro-De León, G. & Köhler, C. Endosperm-specific transcriptome analysis by applying the INTACT system. *Plant Reproduction* **32**, 55–61 (2018).
65. Kang, I.-H., Steffen, J. G., Portereiko, M. F., Lloyd, A. & Drews, G. N. The AGL62 MADS domain protein regulates cellularization during endosperm development in Arabidopsis. *The Plant Cell* **20**, 635–47 (2008).
66. Roszak, P. & Köhler, C. Polycomb group proteins are required to couple seed coat initiation to fertilization. *PNAS* **108**, 20826–31 (2011).
67. Figueiredo, D. D., Batista, R. A., Roszak, P. J., Hennig, L. & Köhler, C. Auxin production in the endosperm drives seed coat development in Arabidopsis. *eLife* **5**, 1–23 (2016).
68. Zhao, Y. Auxin Biosynthesis: A Simple Two-Step Pathway Converts Tryptophan to Indole-3-Acetic Acid in Plants. *Molecular Plant* **5**, 334–338 (2012).
69. Garcia, D. *et al.* Arabidopsis haiku mutants reveal new controls of seed size by endosperm. *Plant Physiology* **131**, 1661–70 (2003).
70. Luo, M., Dennis, E. S., Chaudhury, A., Peacock, W. J. & Berger, F. MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in Arabidopsis. *PNAS* **102**, 17531–17536 (2005).
71. Xing, Q. *et al.* ZHOUPI controls embryonic cuticle formation via a signalling pathway involving the subtilisin protease ABNORMAL LEAF-SHAPE1 and the receptor kinases GASSHO1 and GASSHO2. *Development* **140**, 770–9 (2013).
72. Yang, S. *et al.* The endosperm-specific ZHOUPI gene of Arabidopsis thaliana regulates endosperm breakdown and embryonic epidermal development. *Development* **135**, 3501–9 (2008).
73. Grossniklaus, U., Vielle-Calzada, J. P., Hoeppner, M. a & Gagliano, W. B. Maternal control of embryogenesis by MEDEA, a polycomb group gene in Arabidopsis. *Science* **280**, 446–450 (1998).
74. Kiyosue, T. *et al.* Control of fertilization-independent endosperm development by the MEDEA polycomb gene in Arabidopsis. *PNAS* **96**, 4186–4191 (1999).
75. Wang, G., Jiang, H., Del Toro de León, G., Martinez, G. & Köhler, C. Sequestration of a Transposon-Derived siRNA by a Target Mimic Imprinted Gene Induces Postzygotic Reproductive Isolation in Arabidopsis. *Developmental Cell* **46**, 696-705.e4 (2018).



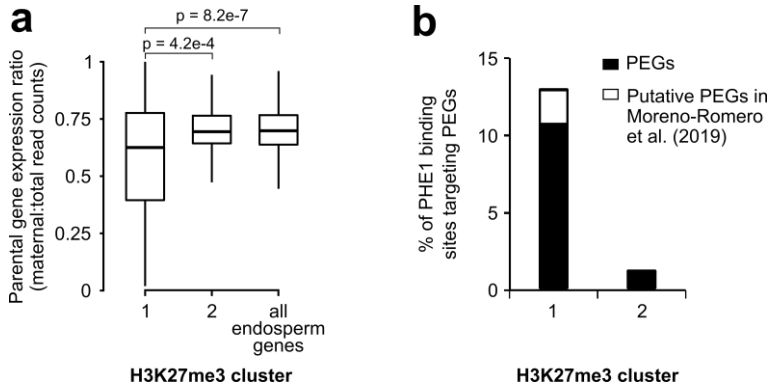
Extended Data Fig. 1 | Transcription factor genes are enriched among PHE1 targets. a, Enriched biological processes within PHE1 target genes. Numbers on bars indicate number of PHE1 target genes within each GO-term. **b,** Enrichment of transcription factor (TF) families among PHE1 target genes (see Methods). Numbers in parenthesis indicate the total number of *Arabidopsis* genes belonging to a certain transcription factor family, and the total number of genes in that family targeted by PHE1, respectively. P-value was determined using the hypergeometric test.



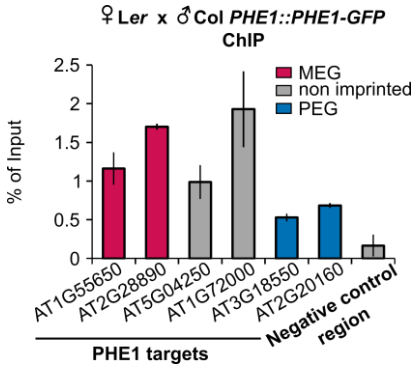
Extended Data Fig. 2 | PHE1 DNA-binding motifs show similarity to type II MADS-box CArG-boxes. Alignment between PHE1 DNA-binding motifs and known motif matches.



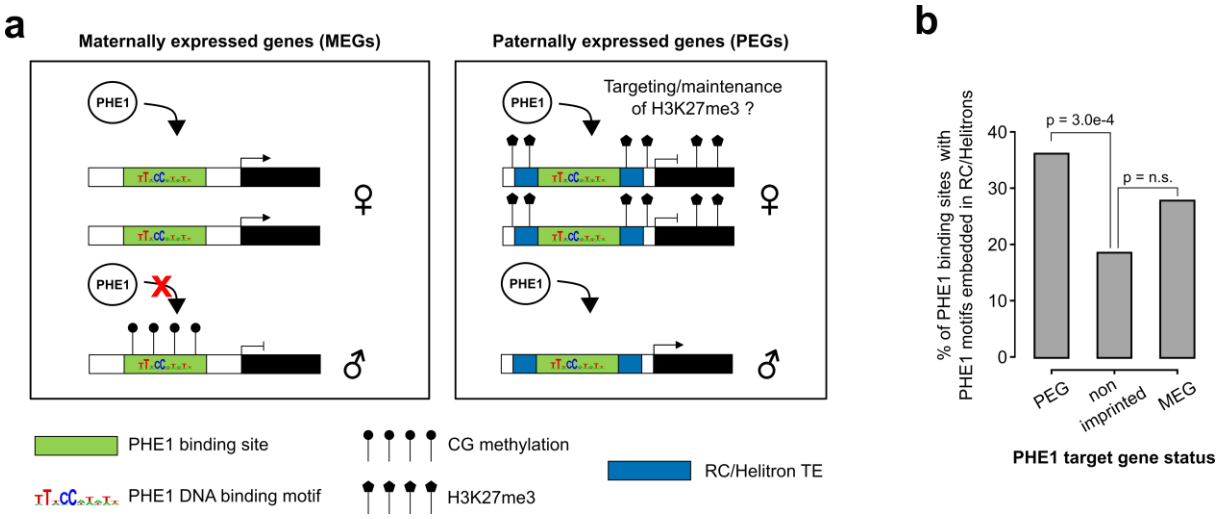
Extended Data Fig. 3 | RC/Helitron family overlap with PHE1 binding sites. Fraction of PHE1 binding sites (green) overlapping with different RC/Helitron TE families. Overlap is expressed as the percentage of binding sites where spatial intersection with the families specified on the y-axis is observed. A set of random binding sites (blue, see Methods) is used as control. P-values were determined using Monte Carlo permutation tests (see Methods). Bars represent ± sd, (n = 2494, PHE1 binding sites, Random binding sites).



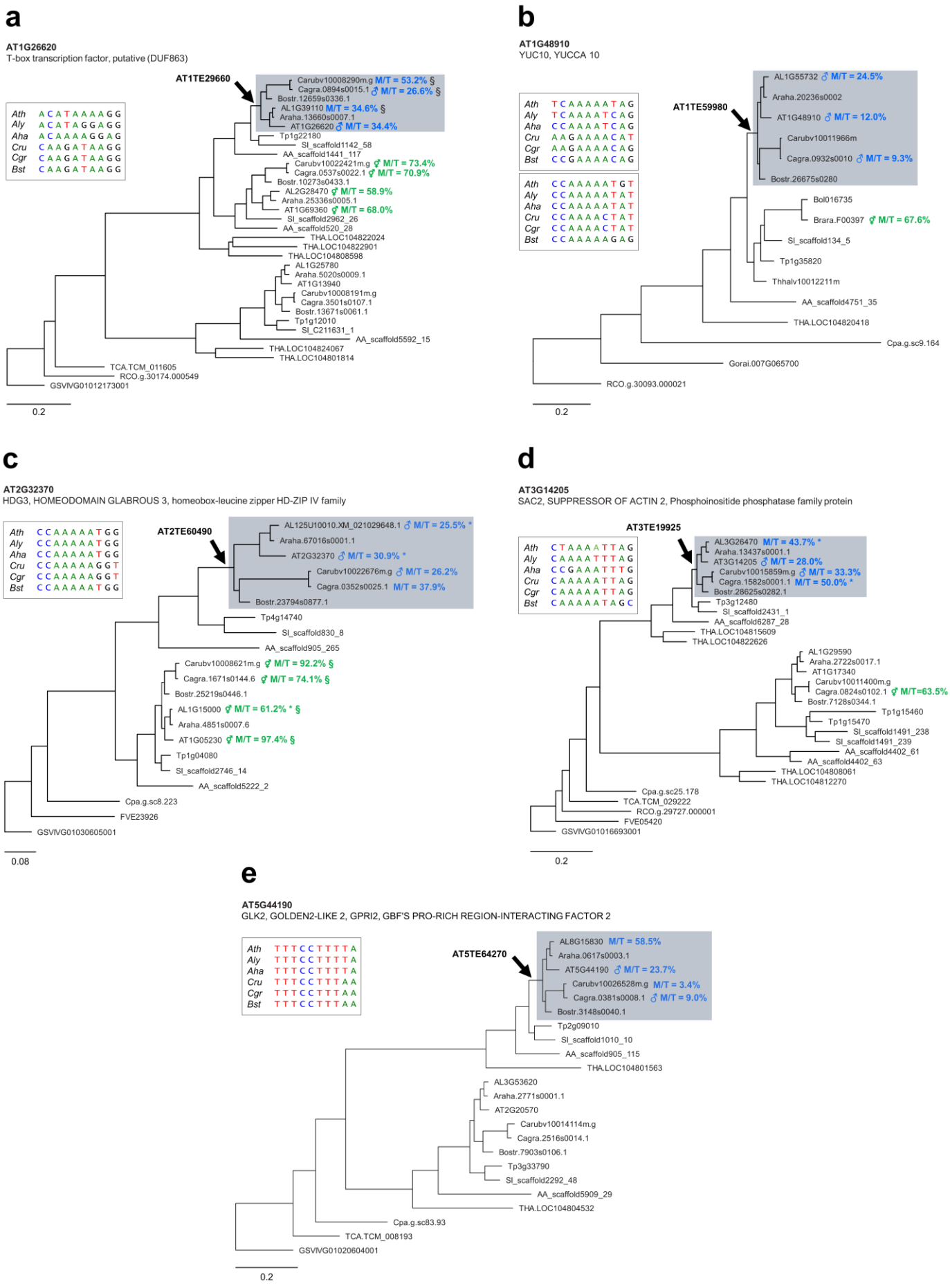
Extended Data Fig. 4 | Characterization of H3K27me3 clusters. a, Parental gene expression ratio of PHE1 targets associated with Cluster 1 and Cluster 2 binding sites, and all endosperm expressed genes. P-values were determined using two-tailed Mann-Whitney tests. **b,** Fraction of Cluster 1 and Cluster 2 binding sites that target PEGs and putative PEGs[63]. (a-b) H3K27me3 clusters are those defined in Fig. 2b.



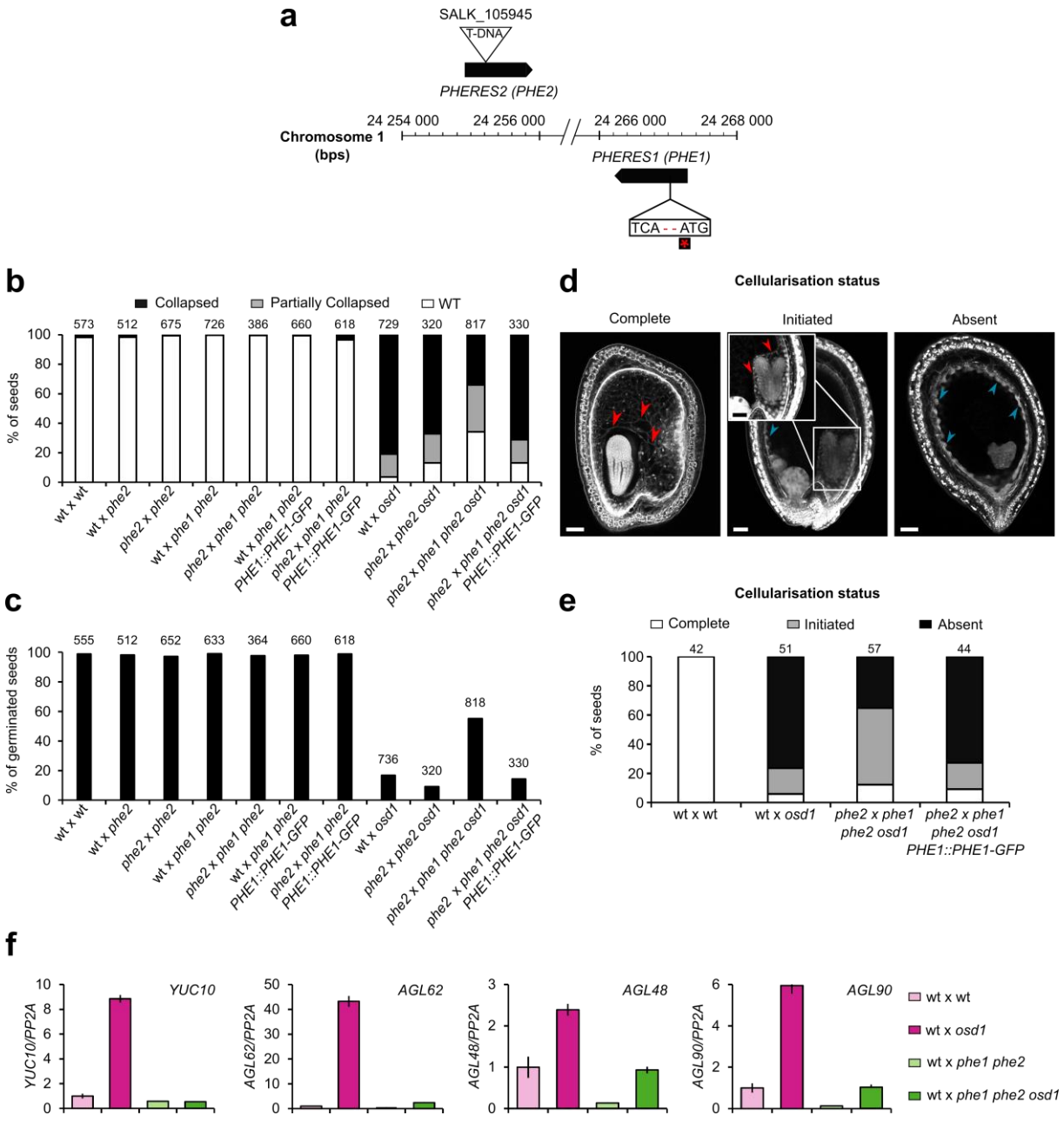
Extended Data Fig. 5 | Parental-specific PHE1 ChIP. qPCR of purified ChIP-DNA. Enrichment is shown as % of input, in regions associated with MEGs (pink), PEGs (blue), and non-imprinted (grey) PHE1 target genes. Bars represent ± sd. Data from one representative biological replicate is shown (n = 2 biological replicates).



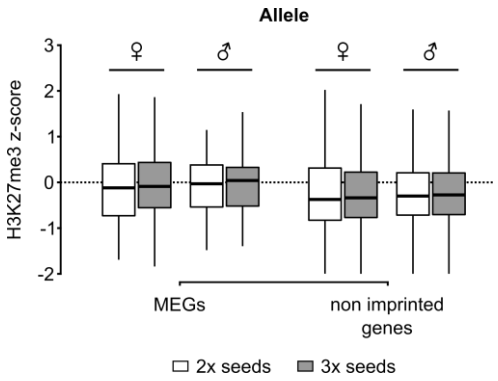
Extended Data Fig. 6 | Control of imprinted gene expression by PHE1. a, Schematic model of imprinted gene control by PHE1. Maternally expressed genes (left panel) show DNA hypermethylation of paternal (♂) PHE1 binding sites. This precludes PHE1 accessibility to paternal alleles, leading to predominant binding and transcription from maternal (♀) alleles. In paternally expressed genes (right panel), RC/Helitrons found in flanking regions carry PHE1 DNA-binding motifs, allowing PHE1 binding. The paternal PHE1 binding site is devoid of repressive H3K27me3, facilitating binding of PHE1 and transcription of this allele. H3K27me3 accumulates at the flanks of maternal PHE1 binding sites, while the binding sites remain devoid of this repressive mark. PHE1 is able to bind maternal alleles, but fails to induce transcription. We hypothesize that the accessibility to maternal PHE1 binding sites might be important for deposition of H3K27me3 during central cell development (possibly by another type I MADS-box transcription factor). It may furthermore be required for maintenance of H3K27me3 during endosperm division. **b,** Fraction of PHE1 binding sites targeting MEGs, PEGs, or non-imprinted genes where a spatial overlap between a RC/Helitron and a PHE1 DNA-binding motif is observed (as illustrated in a, right panel). P-values were determined using the hypergeometric test.



Extended Data Fig. 7 | Ancestral RC/Helitron insertions are associated with gain of imprinting in the Brassicaceae. a-e, Phylogenetic analyses of PHE1-targeted PEGs and their homologs. Each panel represents a distinct target gene, and its corresponding homologs in different species. The genes in grey background have homologous RC/Helitron sequences in their promoter region. The arrow indicates the putative insertion of an ancestral RC/Helitron. The identity of the RC/Helitron identified in *A. thaliana* is indicated. These *A. thaliana* RC/Helitrons contain a PHE1-DNA binding motif and are associated with a PHE1 binding site. The inset boxes represent the alignment between the *A. thaliana* PHE1 DNA-binding motif and similar DNA motifs contained in RC/Helitrons present at the promoter regions of orthologous genes. When available, the imprinting status of a given gene is indicated by the presence of ♂ (PEG), or ♀ (non-imprinted), and reflects the original imprinting analyses done in the source publications (see Methods). The maternal:total read ratio (M/T) for each gene is also indicated. §: potential contamination from maternal tissue. *: accession-biased expression. Scale bar represents substitution per site for the ML tree. The tree is unrooted. Gene identifier nomenclatures: AT, *Arabidopsis thaliana*; AL, *Arabidopsis lyrata*; Araha, *Arabidopsis halleri*; Bostr, *Boechera stricta*; Carubv, *Capsella rubella*; Cagra, *Capsella grandiflora*; Tp, *Schrenkiella parvula*; SI, *Sisymbrium irio*; Bol, *Brassica oleracea*; Brapa, *Brassica rapa*; Thhalv, *Eutrema salsugineum*; AA, *Aethionema arabicum*; THA, *Tarenaya hassleriana*; Cpa, *Carica papaya*; TCA, *Theobroma cacao*; Gorai, *Gossypium raimondii*; RCO, *Ricinus communis*; FVE, *Fragaria vesca*; GSVIVG, *Vitis vinifera*.



Extended Data Fig. 8 | Rescue of 3x seed inviability in *phe1 phe2*. a, Schematic representation of the *phe1 phe2* mutant. CRISPR/Cas9 was used to generate a premature stop codon in *PHE1*, represented by the red asterisk. *PHE1* mutagenesis was done in the *phe2* background, due to the proximity and likely redundancy of both genes (see Methods). **b-c,** Seed inviability phenotype (b) of wild-type (wt) and paternal excess crosses, in wt, *phe1 phe2*, and *phe1* complementation lines, with respective seed germination rates (c). **d-e**, Status of endosperm cellularisation in wt and paternal excess seeds. (b-c,e) *n* = numbers on top of bars (seeds). **f,** Expression of PHE1 target genes in rescued 3x seeds. Bars represent ± sd. Data from one representative biological replicate is shown (*n* = 2 biological replicates).



Extended Data Fig. 9 | Distribution of H3K27me3 in 2x and 3x seeds. Accumulation of H3K27me3 in maternal (♀) and paternal (♂) gene bodies of MEGs and non-imprinted genes in the endosperm of 2x and 3x seeds (white and grey, respectively).

Extended Data Table 1 | PHE1 target genes previously implicated in endosperm development.

Gene ID	Imprinting status	Description of function
AGL62	Non imprinted	Type I MADS-box TF involved in endosperm proliferation and seed coat development[11,65–67]
TAR1	PEG	Tryptophan (Trp) aminotransferase that catalyses the first step of the Trp-dependent auxin biosynthetic pathway[68]. Involved in endosperm proliferation and cellularisation[11,33]
YUC10	PEG	Flavin monooxygenase that catalyses the last step of the Trp-dependent auxin biosynthetic pathway[68]. Involved in endosperm proliferation and cellularisation[11,33]
IKU2	Non imprinted	Encodes a leucine-rich repeat receptor kinase protein that, together with *MINI3*, is part of the IKU pathway controlling seed size. *iku2* mutants show reduced endosperm growth and early endosperm cellularisation[69,70].
MINI3	Non imprinted	WRKY TF that, together with *IKU2*, is part of the IKU pathway controlling seed size. *mini3* mutants show reduced endosperm growth and early endosperm cellularisation[70].
ZHOUPI	Non imprinted	Encodes a bHLH TF expressed in the embryo surrounding region of the endosperm. It is essential for embryo cuticle formation and endosperm breakdown after its cellularisation[71,72].
MEA	MEG	Subunit of the FIS-PRC2 complex, responsible for depositing H3K27me3 at target loci, including PEGs[14,63]. Loss of MEA and, consequently, paternally biased expression of PEGs, leads to a 3x seed like phenotype[73,74].
ADM	PEG	Interacts with SUVH9 and AHL10 to promote H3K9me2 deposition in TEs, influencing neighbouring gene expression. Mutations in *ADM* lead to rescue of the 3x seed abortion phenotype[10,23,37].
SUVH7	PEG	Encodes a putative histone-lysine N-methyltransferase. Mutations in *SUVH7* lead to rescue of the 3x seed abortion phenotype[10].
PEG2	PEG	Encodes a putative SNARE-like superfamily protein, which is not translated in endosperm. *PEG2* transcripts act as a sponge for siRNA854, thus regulating UBP1 abundance[75]. Mutations in *PEG2* lead to rescue of the 3x seed abortion phenotype[10,75].
NRPD1a	PEG	Encodes the largest subunit of RNA POLYMERASE IV, involved in the RNA-directed DNA methylation pathway. Mutations in *NRPD1a* lead to rescue of the 3x seed abortion phenotype[20].