bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 **Widespread imprinting of transposable elements and young genes in the maize**
2 **endosperm**
3
4 Sarah N Anderson[1]*, Peng Zhou[2], Kaitlin Higgins[1], Yaniv Brandvain[2], Nathan M Springer[2]
5
6 [1]Department of Genetics, Development, and Cell Biology; Iowa State University; Ames, IA
7 [2]Department of Plant and Microbial Biology; University of Minnesota; St. Paul, MN
8
9 * Corresponding author: Sarah N Anderson (sna@iastate.edu)
10
11 ORCIDs: 0000-0002-1671-2286 (SNA); 0000-0001-5684-2256 (PZ); 0000-0001-7759-4386
12 (KH); 0000-0002-3392-0220 (YB); 0000-0002-7301-4759 (NMS)
13
14
15 **Abstract**
16 Fertilization and seed development is a critical time in the plant life cycle, and coordinated
17 development of the embryo and endosperm are required to produce a viable seed. In the
18 endosperm, some genes show imprinted expression where transcripts are derived primarily
19 from one parental genome. Imprinted gene expression has been observed across many
20 flowering plant species, though only a small proportion of genes are imprinted. Understanding
21 the rate of turnover for gain or loss of imprinted expression has been complicated by the
22 reliance on single nucleotide polymorphisms between alleles to enable testing for imprinting.
23 Here, we develop a method to use whole genome assemblies of multiple genotypes to assess
24 for imprinting of both shared and variable portions of the genome using data from reciprocal
25 crosses. This reveals widespread maternal expression of genes and transposable elements with
26 presence-absence variation within maize and across species. Most maternally expressed
27 features are expressed primarily in the endosperm, suggesting that maternal de-repression in
28 the central cell facilitates expression. Furthermore, maternally expressed TEs are enriched for
29 maternal expression of the nearest gene. Read alignments over maternal TE-gene pairs
30 indicate fused transcripts, suggesting that variable TEs contribute imprinted expression of
31 nearby genes.
32

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

33 **Main Text**

34 Imprinted genes showing parent-of-origin based patterns of expression were first identified in
35 maize[1] and have since been identified in a variety of flowering plants. In plants, imprinted
36 expression is primarily observed in the endosperm, which is a nutritive tissue of the seed that is
37 formed when the diploid central cell is fertilized by one of the two sperm cells delivered by the
38 pollen tube. The central cell is epigenetically distinct from most vegetative cells in the plant due
39 to DNA demethylation targeted primarily to Transposable Elements (TEs) [7–9]. This
40 demethylation acts as a primary imprint that distinguishes the female and the male alleles in the
41 endosperm. Maternal and paternal alleles are further distinguished through differential
42 accumulation of histone modifications such as H3K27me3 [10,11] which often marks the maternal
43 allele of paternally expressed genes (PEGs) while maternally expressed genes (MEGs) often
44 show differences in DNA methylation alone [12].
45

46 Imprinting has been studied at the genomic level in many plant species [2–6]. While some genes
47 with conserved imprinting across species contribute to the establishment of imprinting [13],
48 several studies have observed substantial turnover of imprinting for many genes, either within a
49 single species or across species [14,15]. However, understanding the rate of turnover and the
50 source of the imprinted expression pattern has been challenging due in part to methodological
51 inconsistencies across studies and the limitations of available SNPs for allele calls. In
52 Arabidopsis, applying consistent methods and cutoffs across studies reduces apparent
53 variability in imprinting calls [16,17], however many genes cannot be assessed due to a lack of
54 informative SNPs. A lack of SNPs can be due to identical sequence or unalignable regions
55 resulting from large structural changes or presence-absence variation (PAV) of whole genes or
56 features. In maize, many genes and TEs exhibit PAV among genotypes [18–20]. This limits the
57 ability to use SNP-based allele-specific expression analyses to study imprinting, especially for
58 transposons and variable genes. In this study, we develop an alternative approach that relies

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

59 upon comparisons of expression in reciprocal crosses to assess the imprinting of both
60 conserved and variable genes and TEs across maize genotypes with whole genome
61 assemblies, revealing imprinting for many transposable elements and variable genic sequences.
62
63 Reciprocal crosses for every pairwise contrast between three maize genotypes with whole
64 genome assemblies (B73 [21], W22 [22], and PH207 [23]) were performed, and 14 days after
65 pollination, endosperm was isolated in triplicate for RNA-sequencing (Table S1). Two
66 approaches were applied to identify imprinted expression (Figure 1A). The traditional approach
67 for calling imprinting uses Single Nucleotide Polymorphisms to call Allele Specific Expression
68 (SNP-ASE) followed by comparison of biases across reciprocal crosses (methods). The SNP-
69 ASE ratio is calculated by assigning SNP-containing reads to one allele and determining the
70 proportion of informative reads from each allele, providing an estimate of the expression of two
71 alleles within a single sample. We developed and implemented an alternative approach where
72 reads are aligned to concatenated genome files and the Reciprocal Expression Ratio (RER)
73 was calculated to describe the ratio of expression for features in each genome when inherited
74 maternally versus paternally. Unlike SNP-ASE, the RER is a comparison of expression of a
75 feature in reciprocal crosses and cannot be calculated for a single sample. Calculations of RER
76 rely on the ~15% of reads that map uniquely to a single location in the concatenated genomes
77 (Table S1). While many reads map equally well to both genomes and are therefore discarded,
78 unique mapping reads are only found in places of the genome with variants distinguishing the
79 alleles (SNPs or indels) or in regions unique to one genome. After assigning unique reads to
80 features including genes and TEs using HTseq, RER was calculated by dividing the expression
81 level (RPM) when inherited maternally by the sum of expression when maternally or paternally
82 inherited. Given that endosperm is composed of two copies of the maternal genome and one
83 copy of the paternal genome, the null expectation for a transcript's expression is that it will be
84 twice as highly expressed when inherited from the maternal parent compared to the paternal

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

85 parent. For both SNP-ASE and RER, the average value representing a biparentally expressed
86 gene is 0.67, allowing direct comparison of the methods. A comparison of SNP-ASE and RER
87 reveals general agreement between these two approaches for genes that could be analyzed
88 with SNPs, with the majority of genes expressed at the ratio expected by dosage (Figure 1B).
89 Many of the genes showing disagreement between methods in Figure 1B result from genotype-
90 biased expression which exhibits a strong bias in SNP-ASE for a single sample but doesn't
91 result in bias for RER (Figure S1). To further assess accuracy of RER, expression patterns for
92 three MEGs and three PEGs with conserved imprinting status in maize, rice, and Arabidopsis [4]
93 were assessed (Figure 1C, Table S2). In most cases with informative reads, clear parental bias
94 in the expected direction was observed for all genes (Figure 1C).
95

96 While both methods can be used to define imprinting for shared genes distinguishable by SNPs,
97 only the RER method can capture imprinting for portions of the genome that exhibit PAV. This
98 provides new opportunities to study parent-of-origin biased gene expression for TEs and
99 variable genes. The distribution of RER values was assessed across contrasts for different
100 feature types (Figure S2), and the proportion of each set that showed parentally-biased
101 expression was summarized based on RER (Figure 1D). This revealed that across all contrasts,
102 genes conserved within maize rarely exhibit parent-of-origin biased expression (Figure 1D,
103 Figure S2). On average, < 3% of expressed genes that are present in all three maize genotypes
104 in this study show a strong parental bias (Figure 1D). For genes that are variable among maize
105 lines, a higher proportion (> 6%) of expressed genes show high parental bias, with this set
106 representing genes that are accessible using RER but not SNP-ASE. Strikingly, > 11% of
107 expressed TEs show a strong parental bias, with the majority of strongly biased TEs expressed
108 maternally (Figure 1D).
109

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

110 In order to identify imprinted transcripts, we applied the lfcThreshold option within DESeq2 to
111 test for significance (adjusted p-value < 0.05) over the expected 2:1 gene dosage across
112 reciprocals using three biological replicates. To increase the stringency of imprinting calls,
113 significant hits were further filtered by RER values. Maternally Expressed Genes (MEGs) and
114 Maternally Expressed TEs (matTEs) were filtered for RER > 0.9, while Paternally Expressed
115 Genes (PEGs) were filtered for RER < 0.1. It can be difficult to remove all maternal tissues
116 when isolating endosperm tissue and therefore it is important to limit potential false-positive
117 calls of maternal expression that may result from genes expressed in the maternal seed coat [24].
118 Previously published RNA-seq data [25] was used to filter out genes whose maternal expression
119 could result from seed coat contamination rather than maternal expression in the endosperm.
120 Pericarp-preferred genes were defined where the mean expression in pericarp was >2-fold
121 higher than the expression in endosperm (Figure S3). After implementing these criteria and
122 filters, we identified an average of 182 total imprinted genes across all hybrid combinations, with
123 an average of 112 MEGs and 70 PEGs in each (Figure 2A).
124
125 The imprinted genes discovered in each genome were compared to assess the consistency of
126 imprinting. A comparison of imprinted features in the B73 x W22 reciprocal hybrid endosperm
127 tissue identifies 17 MEGs, 39 PEGs, and 4 matTEs that were consistently imprinted in both
128 genomes (Figure 2B, 3B). A subset of the genes that do not exhibit consistent imprinting are
129 shared between the two genomes. For example, there are 26 MEGs observed only in B73 and
130 11 only observed in W22 despite the fact that both genomes retain a syntenic ortholog for these
131 genes. For the majority of these shared genes with variable imprinting, the lack of overlap is due
132 to cutoff stringency or lack of coverage rather than true turnover of imprinting (Figure S4). There
133 are many additional cases where imprinted genes are only present in one genome. For PEGs,
134 variable genes represent the minority of non-conserved imprinted genes, with only 13 of 34 B73
135 PEGs that are not imprinted in W22 variable across genomes. In contrast, for the majority of

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

136 MEGs with inconsistent imprinting (i.e. 107 of 133 B73 genes in the B73 by W22 contrast), the
137 genes themselves are absent from the other genome. Similar patterns are observed for the B73
138 by PH207 contrast, though a higher proportion of genes are shared in this contrast. The large
139 number of maternally expressed transcripts with variability in maize suggests that imprinting of
140 non-conserved elements may be far more prevalent than previously detected due to the
141 limitations of SNP-based allele calls.

142

143 To understand additional features of imprinted genes, we focused on the B73 genes that were
144 called imprinted in at least one contrast, which included 202 MEGs and 111 PEGs. B73 was
145 selected as the central genotype because it has substantially more expression datasets,
146 syntenic gene information, and functional gene annotations than other genomes. For the genes
147 identified as imprinted, we compared several characteristics relative to genes that were
148 expressed but were not classified as imprinted. First, genes were assessed for variability across
149 maize inbred lines by defining conserved genes as those with syntenic orthologs in B73, W22,
150 and PH207 and variable genes as those without a corresponding gene in at least one genome
151 [26]. This revealed a clear enrichment for variable genes among MEGs (p-value < 0.001, chisq
152 test), but not PEGs, compared to genes that are not imprinted but have enough unique reads to
153 be assessed for imprinting (Figure 2C). We then expanded our evolutionary distance and
154 assessed how many genes in each set are syntenic with other grasses as defined by having a
155 syntenic ortholog in sorghum, rice, foxtail millet, and brachypodium. For genes without
156 imprinting, the majority (62%) are syntenic with other grasses. However, MEGs are highly
157 depleted for syntenic genes (19%) and PEGs show a minor depletion (50%, p-value < 0.05,
158 chisq test). Next, the expression pattern across B73 development was assessed using
159 published RNA-seq data [25]. Since imprinting can arise from either silencing of one parental
160 allele specifically in the endosperm or de-repression of one parental allele in the endosperm, the
161 pattern of expression across tissues was defined as either constitutive or endosperm-preferred

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

162 (see methods, figure S4). While only 3% of non-imprinted genes are expressed preferentially in
163 the endosperm, 77% of MEGs and 32% of PEGs show this expression pattern (Figure 2C, p-
164 value < 0.001, chisq test). Many of the MEGs (38%) have no assigned GO term, a 2.8-fold
165 enrichment compared to genes that are not imprinted (p-value < 0.001, chisq test). Since TEs
166 are a common source of new genes and a driver of gene content variation among maize lines,
167 we intersected our imprinted genes with annotated TEs, identifying 26 MEGs and 1 PEG
168 completely within an annotated transposable element. While MEGs and PEGs are annotated as
169 genes in the B73v4 annotation, transcription of a locus does not imply the creation of a
170 functional gene product. While evolutionarily conserved genes with synteny to other grasses
171 may be the best candidates for real genes capable of conferring phenotypes [27], variable genes
172 can be important for functions such as disease resistance [28].
173
174 To further investigate the imprinting of TEs themselves, the RER method was used to define
175 imprinted TEs, with an average of 95 matTEs identified across contrasts (Figure 3A). There are
176 a small number of paternally expressed TEs, however these were excluded from further
177 analyses due to the low number detected and potential technical complications (Figure 3A, S2).
178 Consistent with the large amount of TE variability among genotypes, the majority of imprinted
179 TEs were unique to one genome (Figure 3B). There are 145 maternally expressed TEs in B73
180 relative to at least one other genotype, including 72 LTR retrotransposons, 52 Helitrons, 9 TIR
181 transposons, and 2 LINEs (Figure 3C). The vast majority of these TEs (93%) represent specific
182 TE insertions that are polymorphic among the three maize genotypes [20]. Given the high tissue-
183 specificity of TE expression observed previously [29], the tissue-specific expression patterns for
184 matTEs were also assessed. We found that 92% of matTEs are expressed preferentially in the
185 endosperm, suggesting that imprinting is established through de-repression of the maternal
186 allele preferentially in the endosperm and that this is the only stage of development for
187 expression of these elements (Figure 3C, S5). Since TE families have the potential for

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

188 coordinated expression responses among members, the families for matTEs were assessed.
189 matTEs are in 84 families, with only one Helitron family containing more than 5 imprinted
190 elements. This family, DHH00002 (DHH2), contains 44 maternally expressed members and is
191 the only Helitron family in B73 that is predicted to have autonomous members. Since prior work
192 has suggested that Helitrons are responsible for creating imprinting by moving PHE1 binding
193 sites around the genome [30], the proportion of DHH2 Helitrons with predicted motifs was
194 assessed (Figure S6). We found that matTEs of this family are more likely to have a binding site
195 than elements that are not detected in our analysis, though the distribution is similar to family
196 members that are not imprinted so it is unlikely that PHE1 sites alone are sufficient to confer
197 imprinting of DHH2 Helitrons.
198
199 TEs have been proposed as a source of variation in imprinted gene expression, and this dataset
200 allows for investigation into the relationship between imprinted genes and TEs. For every matTE
201 in B73, the closest gene was identified and assessed for imprinting. For 13% of matTEs, the
202 nearest gene is a MEG, which is a significant enrichment (p-value < 0.001, binomial test) and
203 11.6 times more common than expected based on the proportion of expressed genes that are
204 called MEGs (Figure 3D). In contrast, there were no identified examples of matTEs where the
205 closest gene is a PEG. There were 19 matTEs where the closest gene is one of 15 MEGs
206 (Table S3). In the majority of cases, the TE overlapped (N = 7) or was upstream of the gene (N
207 = 10). We identified only two cases of the TE located downstream of the gene, and one of these
208 genes also overlapped a matTE. The asymmetry between upstream and downstream TE
209 relationships suggests that the orientation likely matters for determining which TEs are able to
210 influence gene expression patterning. In all cases, the developmental expression patterns of the
211 genes and the nearby TEs match. To understand the nature of transcripts, read alignments for
212 matTE-MEG pairs were visualized with IGV. In all cases, reads aligning to both the matTE and
213 corresponding MEG mapped to the same strand without clear separation in read alignments,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

214 suggesting that many of these clusters may actually represent single transcripts overlapping
215 multiple features (Figure 3E).
216
217 In summary, we developed the RER method to use information from shared and variable
218 portions of maize whole genome assemblies to identify imprinted expression of genes and TEs
219 in maize. This revealed imprinting of many genes that were undetectable by traditional methods
220 that rely on diagnostic SNPs between parental alleles. The majority of maternally expressed
221 features (genes and TEs) represent young portions of the genome that are variable within maize
222 and non-syntenic with other grasses. We also observe strong enrichment for MEGs near
223 maternally expressed TEs, further supporting the connection between turnover of imprinting and
224 presence-absence variation of TEs. In mammals, imprinting in the placenta has been proposed
225 to result from different defense mechanisms used by male and female germlines to reduce
226 retrovirus proliferation in the germ line [31], and turnover of imprinting could have a similar host
227 defense explanation in plant endosperm. In plants, there are genes with conserved imprinting
228 across plant species that support theories of parental conflict [32] or dosage [33], however the
229 majority of imprinted loci are variable within and across species. By studying imprinting using
230 whole genome assemblies, we are able to better understand the turnover of imprinted
231 expression of both shared and variable portions of plant genomes.
232

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

233 **Materials and Methods**

234 **Materials**

235 Three maize inbred lines, B73, W22, and PH207, were grown in the field in Saint Paul, MN in
236 the summer of 2018. Reciprocal crosses between each pair of genotypes were performed. Ears
237 were collected 14 days after pollination and endosperm was isolated using manual dissection,
238 with approximately 10 kernels per ear pooled for each biological replicate. Paired-end, stranded
239 RNA-seq libraries were created using the Illumina TruSeq Stranded mRNA kit and sequencing
240 was performed with the Illumina HiSeq 2500 at the University of Minnesota Genomics Center.
241 On average, > 45 million reads were generated per library (Table S1).

242

243 **Sequence alignments for RER**

244 Concatenated genome files were created for each pairwise contrast of parental genomes and
245 assemblies used included B73v4 [21], W22 [22], and PH207 [23]. When necessary, chromosome
246 designations were altered to ensure non-redundant sequence names across parents. Hisat2
247 index files were created using genome sequences only for each contrast. Gene annotations and
248 disjoined filteredTE annotations available at https://github.com/SNAnderson/maizeTE_variation
249 were combined by first subtracting exon regions from the TE annotations and then combining
250 full gene and TE annotations for each genome. Concatenated annotation files were then
251 created for each pairwise contrast using the same chromosomal designation as for the genome
252 files. RNA-seq reads were trimmed using cutadapt [34] and aligned to the concatenated genomes
253 corresponding to the parents using hisat2 [35]. Unique-mapping reads to the concatenated
254 genome files were then assigned to features (genes and TEs) using HTseq [36]. Counts to each
255 feature were normalized as reads per million using library size estimates derived from the SNP-
256 ASE method (described below). RER for each annotation (gene and TE) was calculated by
257 dividing the mean expression when inherited maternally by the sum of the expression when
258 inherited maternally and paternally.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

259

260 **Sequence alignments for SNP-ASE**

261 In parallel to the above method of mapping reads, we also ran the standard, SNP-based allele
262 specific expression pipeline by mapping reads to the B73 AGPv4 reference assembly using a
263 variant-aware aligner HiSat2 trained with a set of known SNPs as described in [37]. The number
264 of reads supporting each parental genotype were used to calculate the proportion of maternal
265 reads for each gene. For comparison across mapping methods, genes were filtered for only
266 those with at least 10 informative reads in both methods. SNP-ASE ratios were calculated for
267 each gene in each direction of the reciprocal cross separately by dividing the number of reads
268 matching the maternal allele by the total number of informative reads. Genes with parent-
269 specific expression were defined as those with a SNP maternal ratio > 0.85 in one direction and
270 < 0.15 in the reciprocal direction.

271

272 **Defining imprinting**

273 To define imprinted features using RER, count tables for genes and TEs in each library were
274 loaded into R. For each of the three reciprocal crosses performed in triplicate, DESeq2 [38] was
275 applied using the lfcThreshold=1 and altHypothesis="greaterAbs" options to identify features
276 with significant deviations from the 2:1 expected expression difference based on dosage. Each
277 contrast includes features from both parental genomes, so maternal and paternal expression
278 was determined by the direction of the differential expression plus the genome where the
279 feature was annotated. Significant features were further filtered to only strong cases of
280 imprinting where RER was > 0.9 for MEGs and matTEs and < 0.1 for PEGs. To create the final
281 list of imprinted features, maternal features with pericarp-preferred expression were filtered out
282 (see Tissue Dynamics).

283

284 **Tissue Dynamics**

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

285 The expression profile of genes and TEs was analyzed for B73 features using previously
286 published analysis [29] using data from [25]. To filter out genes where expression is higher in the
287 pericarp than the endosperm and could thus result in inaccurate imprinting calls [24], expression
288 was compared for 14 dap seeds (the time point used in this study) and 18 dap pericarp. Genes
289 with expression over twice as high in the pericarp over the endosperm were excluded from MEG
290 calls. W22 and PH207 genes corresponding to genes expressed higher in the pericarp were
291 also excluded from MEG calls. No matTEs were identified as potential contaminants using this
292 method. Expression data across all tissues was also used to identify endosperm-preferred
293 expression. Endosperm-preferred expression was defined as genes and TEs where the sum of
294 expression in endosperm and wole seed libraries (26% of libraries) was more than 60% of the
295 sum of expression across all libraries.
296
297 **Descriptors**
298 To identify genes that are shared between genome assemblies and annotations, the file
299 gene_model_xref_v4.txt was downloaded from MaizeGDB [26] on 2020/01/22. This file is B73-
300 based and genes with a single corresponding gene in either the pairwise contrast (for venn
301 diagrams) or in both W22 and PH207 (all other analyses) were defined as conserved in maize
302 while remaining genes were defined as variable. This file was also used to define genes that are
303 syntenic with other grasses, with syntenic genes being defined as any gene with a syntenic
304 ortholog in foxtail millet, rice, brachypodium, and sorghum. To identify the nearest gene to each
305 matTE, bedtools closest was used and distances between TE and gene were reported relative
306 to the orientation of the gene.
307
308 **Data Availability**
309 RNA-seq data files have been uploaded to NCBI SRA under BioProject ID PRJNA623806.
310 Scripts and data files used to process results are available at

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

311 https://github.com/SNAnderson/Imprinting2020 and
312 https://github.com/kmhiggins/Imprinting_2020.
313
314 **Acknowledgements**
315 We thank Peter Hermanson for technical assistance. This work was funded by grants from
316 USDA-NIFA 2016-67013-24747 (SNA, and NMS), NSF IOS-1546899 (PZ and NMS), NSF
317 DGE-1545453 (KH), and from the Minnesota Agricultural Experiment Station (MIN 71-068). The
318 Minnesota Supercomputing Institute (MSI) at the University of Minnesota provided
319 computational resources that contributed to this research.
320

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

321 **References**

322 1. Kermicle, J. L. Dependence of the R-mottled aleurone phenotype in maize on mode of
323 sexual transmission. *Genetics* **66**, 69–85 (1970).
324 2. Hsieh, T.-F. *et al.* Regulation of imprinted gene expression in Arabidopsis endosperm.
325 *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1755–1762 (2011).
326 3. Luo, M. *et al.* A genome-wide survey of imprinted genes in rice seeds reveals imprinting
327 primarily occurs in the endosperm. *PLoS Genet.* **7**, e1002125 (2011).
328 4. Waters, A. J. *et al.* Parent-of-origin effects on gene expression and DNA methylation in the
329 maize endosperm. *Plant Cell* **23**, 4221–4233 (2011).
330 5. Zhang, M. *et al.* Genome-wide screen of genes imprinted in sorghum endosperm, and the
331 roles of allelic differential cytosine methylation. *Plant J.* **85**, 424–436 (2016).
332 6. Hatorangan, M. R., Laenen, B., Steige, K. A., Slotte, T. & Köhler, C. Rapid Evolution of
333 Genomic Imprinting in Two Species of the Brassicaceae. *Plant Cell* **28**, 1815–1827 (2016).
334 7. Gehring, M., Bubb, K. L. & Henikoff, S. Extensive demethylation of repetitive elements
335 during seed development underlies gene imprinting. *Science* **324**, 1447–1451 (2009).
336 8. Ibarra, C. A. *et al.* Active DNA demethylation in plant companion cells reinforces
337 transposon methylation in gametes. *Science* **337**, 1360–1364 (2012).
338 9. Park, K. *et al.* DNA demethylation is initiated in the central cells of Arabidopsis and rice.
339 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 15138–15143 (2016).
340 10. Weinhofer, I., Hehenberger, E., Roszak, P., Hennig, L. & Köhler, C. H3K27me3 profiling of
341 the endosperm implies exclusion of polycomb group protein targeting by DNA methylation.
342 *PLoS Genet.* **6**, (2010).
343 11. Moreno-Romero, J., Jiang, H., Santos-González, J. & Köhler, C. Parental epigenetic
344 asymmetry of PRC2-mediated histone modifications in the Arabidopsis endosperm. *EMBO*
345 *J.* **35**, 1298–1311 (2016).
346 12. Zhang, M. *et al.* Genome-wide high resolution parental-specific DNA and histone

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

347 methylation maps uncover patterns of imprinting regulation in maize. *Genome Res.* **24**,
348 167–176 (2014).
349 13. Luo, M., Bilodeau, P., Dennis, E. S., Peacock, W. J. & Chaudhury, A. Expression and
350 parent-of-origin effects for FIS2, MEA, and FIE in the endosperm and embryo of developing
351 Arabidopsis seeds. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10637–10642 (2000).
352 14. Waters, A. J. *et al.* Comprehensive analysis of imprinted genes in maize reveals allelic
353 variation for imprinting and limited conservation with other species. *Proc. Natl. Acad. Sci. U.*
354 *S. A.* **110**, 19639–19644 (2013).
355 15. Pignatta, D. *et al.* Natural epigenetic polymorphisms lead to intraspecific variation in
356 Arabidopsis gene imprinting. *Elife* **3**, e03198 (2014).
357 16. Wyder, S., Raissig, M. T. & Grossniklaus, U. Consistent Reanalysis of Genome-wide
358 Imprinting Studies in Plants Using Generalized Linear Models Increases Concordance
359 across Datasets. *Sci. Rep.* **9**, 1320 (2019).
360 17. Picard, C. L. & Gehring, M. Identification and Comparison of Imprinted Genes Across Plant
361 Species. *Methods Mol. Biol.* **2093**, 173–201 (2020).
362 18. Springer, N. M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and
363 presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734 (2009).
364 19. Hirsch, C. N. *et al.* Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*
365 **26**, 121–135 (2014).
366 20. Anderson, S. N. *et al.* Transposable elements contribute to dynamic genome content in
367 maize. (2019) doi:10.1101/547398.
368 21. Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies. *Nature*
369 **546**, 524–527 (2017).
370 22. Springer, N. M. *et al.* The maize W22 genome provides a foundation for functional
371 genomics and transposon biology. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0158-0.
372 23. Hirsch, C. N. *et al.* Draft Assembly of Elite Inbred Line PH207 Provides Insights into

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

373 Genomic and Transcriptome Diversity in Maize. *Plant Cell* **28**, 2700–2714 (2016).
374 24. Schon, M. A. & Nodine, M. D. Widespread Contamination of Arabidopsis Embryo and
375 Endosperm Transcriptome Data Sets. *Plant Cell* **29**, 608–617 (2017).
376 25. Stelpflug, S. C. *et al.* An Expanded Maize Gene Expression Atlas based on RNA
377 Sequencing and its Use to Explore Root Development. *Plant Genome* **9**, (2016).
378 26. Portwood, J. L., 2nd *et al.* MaizeGDB 2018: the maize multi-genome genetics and
379 genomics database. *Nucleic Acids Res.* **47**, D1146–D1154 (2019).
380 27. Schnable, J. C. Genome evolution in maize: from genomes back to genes. *Annu. Rev.*
381 *Plant Biol.* **66**, 329–343 (2015).
382 28. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for
383 identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2011).
384 29. Anderson, S. N. *et al.* Dynamic Patterns of Transcript Abundance of Transposable Element
385 Families in Maize. *G3* **9**, 3673–3682 (2019).
386 30. Batista, R. A. *et al.* The MADS-box transcription factor PHERES1 controls imprinting in the
387 endosperm by binding to domesticated transposons. *Elife* **8**, (2019).
388 31. Haig, D. Retroviruses and the placenta. *Curr. Biol.* **22**, R609–13 (2012).
389 32. Haig, D. Coadaptation and conflict, misconception and muddle, in the evolution of genomic
390 imprinting. *Heredity* **113**, 96–103 (2014).
391 33. Dilkes, B. P. & Comai, L. A differential dosage hypothesis for parental effects in seed
392 development. *Plant Cell* **16**, 3174–3180 (2004).
393 34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
394 *EMBnet.journal* **17**, 10–12 (2011).
395 35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
396 requirements. *Nat. Methods* **12**, 357–360 (2015).
397 36. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-
398 throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

399 37. Zhou, P., Hirsch, C. N., Briggs, S. P. & Springer, N. M. Dynamic Patterns of Gene
400 Expression Additivity and Regulatory Variation throughout Maize Development. *Mol. Plant*
401 **12**, 410–425 (2019).
402 38. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
403 RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

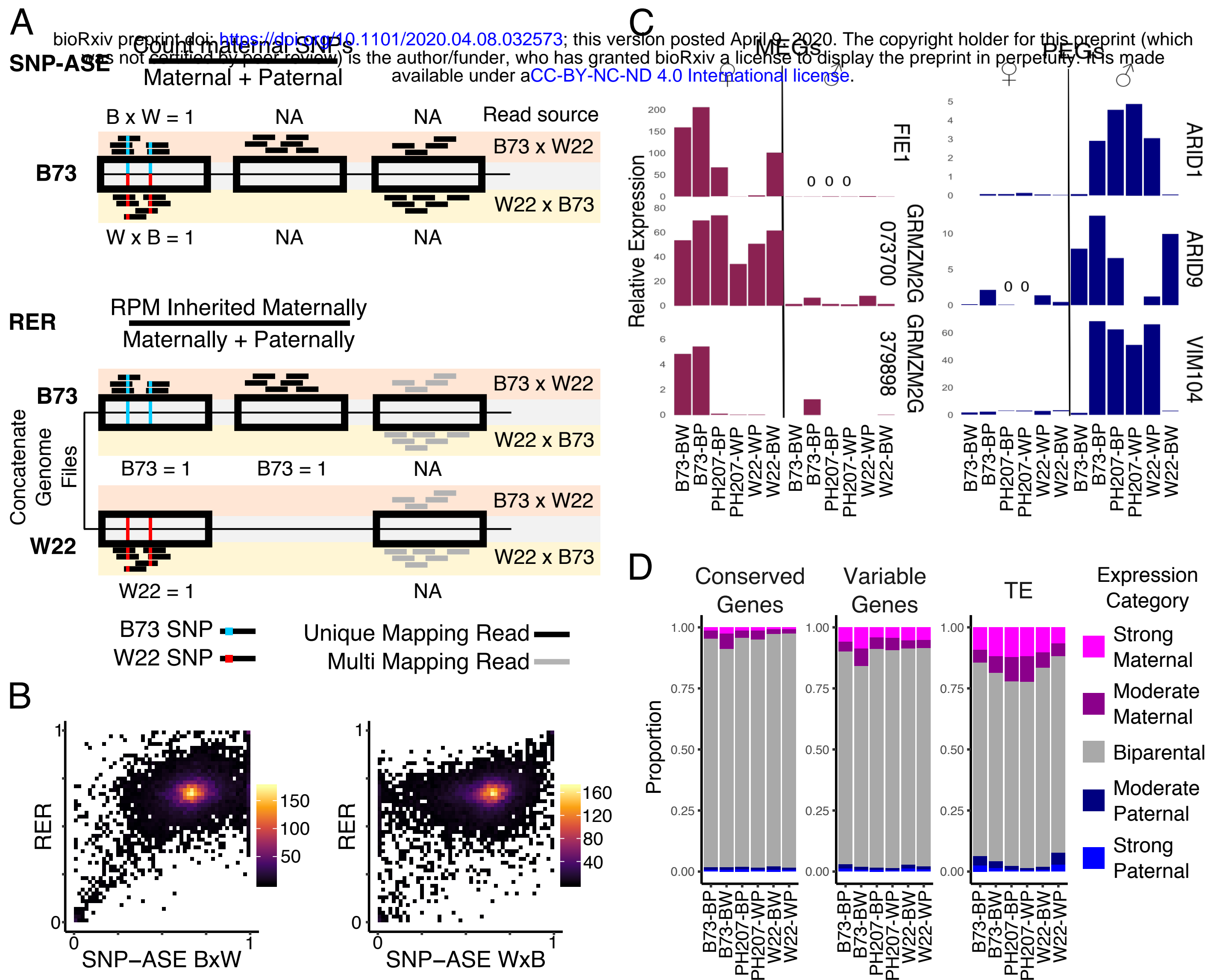

Figure 1: Assessing imprinted expression pattern in maize. A) The method for defining imprinting using SNP-ASE versus RER. in SNP-ASE, reads mapping to a SNP-corrected reference genome are assigned to alleles based on the SNP supported. In RER, reads are assigned to a concatenated reference genome and retained at unique positions. Both methods can be used to assess imprinting for shared genes distinguished by SNPs, but only RER can assess imprinting for PAV features. B) Comparison of SNP-ASE and RER for B73 genes accessible using both methods in the B73 x W22 cross, with values plotted showing the average across three biological replicates. SNP-ASE is assessed for each direction of reciprocal crosses separately while RER is calculated with reciprocals. The heat represents the number of genes in each pixel of the plot. C) Expression across all contrasts for genes with conserved imprinting in maize, rice, and Arabidopsis (Waters et al 2011) using the RER method. Bar height represents the mean expression across replicates. Symbols above the plot show whether the gene was inherited maternally or paternally. Gene IDs for these genes are listed in Table S2. D) The distribution of RER values for different features across contrasts. RER cutoffs for strong maternal and strong paternal are > 0.9 and < 0.1, respectively, and cutoffs for moderate maternal and paternal are > 0.8 or < 0.2, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2

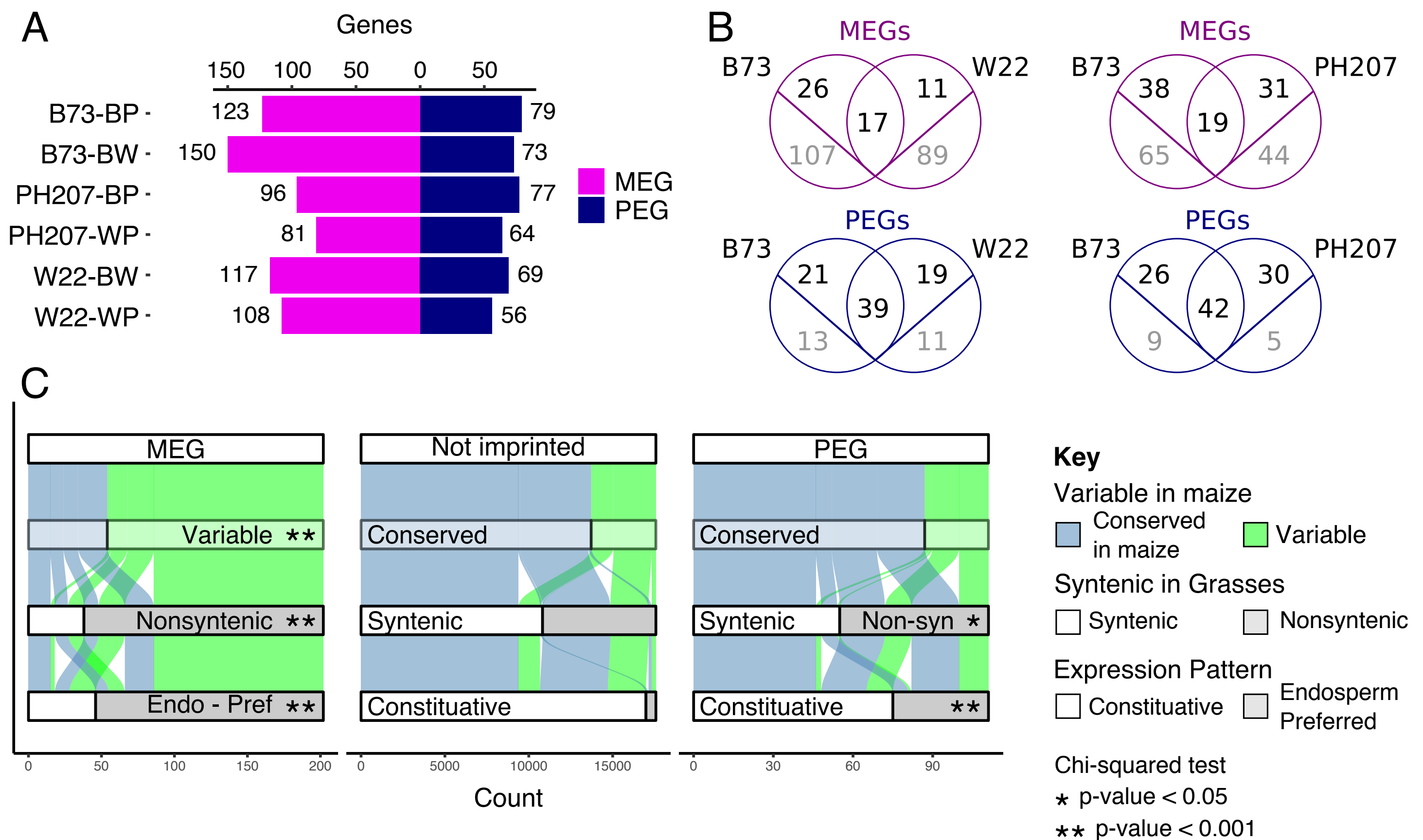

Figure 2: Imprinting of genes defined by RER. A) The number of imprinted genes identified across contrasts using the RER method (see methods). MEGs are shown in magenta and PEGs are shown in blue. B) The overlap between imprinted genes across pairwise contrasts. Genes that are shared between genotypes that could be assessed for imprinting are shown in black above the line while imprinted genes unique to one genome are shown in gray below the line. C) Comparison of features for MEGs, PEGs, and non-imprinted B73 genes. Genes are defined as conserved when they are shared with all genotypes present in this study, syntenic when a syntenic ortholog exists in sorghum, rice, foxtail millet, and brachypodium, and endosperm-preferred if expression is primarily restricted to the endosperm (Figure S4). Asterisks denote significance relative to the Not Imprinted set (chi-squared test).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.08.032573; this version posted April 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3

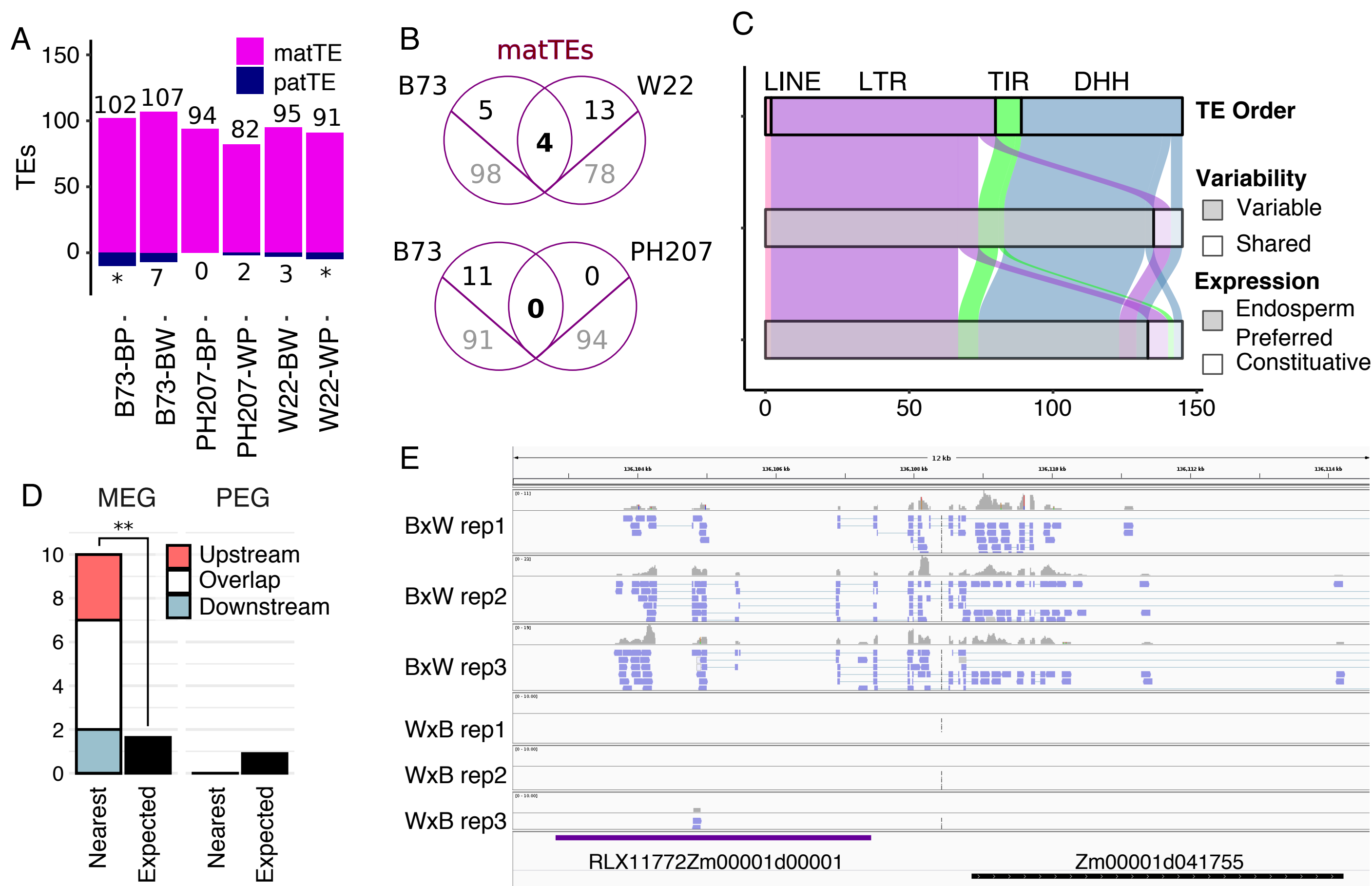

Figure 3: Imprinted TEs defined by RER. A) The number of imprinted TEs across contrasts. matTEs are marked in magenta and paternally expressed TEs are marked in navy. Asterisks denote contrasts where paternally expressed TEs could not be defined (Figure S2). B) The overlap between matTEs across pairwise contrasts. TEs that are shared between genotypes that could be assessed for imprinting are shown in black above the line while imprinted TEs unique to one genome are shown in gray below the line. C. Features of matTEs in B73. TE orders are abbreviated: DHH = Helitron, TIR = terminal inverted repeat transposon, LTR = LTR retrotransposon, and LINE = long interspersed nuclear element. TE variability is defined by prior work (Anderson et al. 2019). Endosperm-preferred expression is described by patterns across development (Figure S4). D) Imprinting status of closest gene to matTEs. Expected number is based on the number of MEGs and PEGs that were assessed for imprinting. ** p-value < 0.001 (binomial test) E) IGV view showing a representative example of reads aligning to a matTE near a MEG. Reads are colored by the strand of alignments, where blue = forward strand.