Genomics of homoploid hybrid speciation: diversity and transcriptional activity of long terminal repeat retrotransposons in hybrid sunflowers

Sebastien Renaut¹, Heather C. Rowe¹, Mark C. Ungerer³ and Loren H. Rieseberg¹,²

¹Biodiversity Research Centre and Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4
²Department of Biology, Indiana University, 1001 East Third St., Bloomington, IN 47405, USA
³Division of Biology, Kansas State University, 426 Ackert Hall, Manhattan, KS 66506, USA

Hybridization is thought to play an important role in plant evolution by introducing novel genetic combinations and promoting genome restructuring. However, surprisingly little is known about the impact of hybridization on transposable element (TE) proliferation and the genomic response to TE activity. In this paper, we first review the mechanisms by which homoploid hybrid species may arise in nature. We then present hybrid sunflowers as a case study to examine transcriptional activity of long terminal repeat retrotransposons in the annual sunflowers Helianthus annuus, Helianthus petiolaris and their homoploid hybrid derivatives (H. paradoxus, H. anomalus and H. deserticola) using high-throughput transcriptome sequencing technologies (RNAseq). Sampling homoploid hybrid sunflower taxa revealed abundant variation in TE transcript accumulation. In addition, genetic diversity for several candidate genes hypothesized to regulate TE activity was characterized. Specifically, we highlight one candidate chromatin remodelling factor gene with a direct role in repressing TE activity in a hybrid species. This paper shows that TE amplification in hybrid lineages is more idiosyncratic than previously believed and provides a first step towards identifying the mechanisms responsible for regulating and repressing TE expansions.

1. Introduction

Botanists have long recognized that new species may arise as a consequence of hybridization between genetically differentiated lineages [1–5]. Hybrid speciation occurs most commonly via duplication of a hybrid genome or allopolyploidy [3,6,7]. Genome duplication solves the two main challenges associated with hybrid speciation: hybrid sterility and the preservation of fit gene combinations [2,8,9]. Early-generation plant hybrids often exhibit not only reduced fertility due to abnormal meiotic pairing [10–12], but also increased vigour or heterosis [13–15]. The doubling of a hybrid’s chromosomal complement restores normal pairing and fertility [13,16] and reduces recombination between homeologous chromosomes, thereby fixing heterotic gene combinations [17]. In addition, changes in ploidy confer partial reproductive isolation between the new hybrid and parental populations [3,18], contributing to their divergent evolutionary trajectories.

Hybrid speciation can occur without a change in ploidy (homoploid hybrid speciation), but the conditions are much more restrictive than for allopolyploidy [19–21]. In homoploid hybrid speciation, ecological and fertility selection are expected to lead to the establishment of fit hybrid segregants [3,22]. However, long-term stabilization of these hybrid segregants requires reproductive isolation, as fit gene combinations will be disrupted by gene flow with parental genotypes or with other hybrids [23]. Reproductive isolation may occur through behavioural [24], habitat [25–28], pollinator [29], karyotypic [30,31], geographical [32] and/or...
known to generate novel phenotypic variation that can be hybrid speciation are less clear, although both processes are expression alteration and TE proliferation in homoploid tive isolation of the new hybrid lineage [31]. The roles of gene hybrid speciation, as they can contribute directly to reproduc-sion and evolution of homoploid hybrid species. Karyotypic changes have the most obvious role in changes on the origin and evolution of homoploid hybrid species often differ from their parental species in karyotype [47,48], gene mode [9,30,31,40–47]. Homoploid hybrid species often differ (i) providing the first rigorously documented example of homoploid hybrid speciation in plants, and (ii) showing how molecular markers (in this case, allozymes) allow critical evaluation of the evolutionary outcomes of hybridization. Building on the Stephanomeria example (recently confirmed by Sherman & Burke [35]), numerous examples of plant homoploid hybrid species were convincingly demonstrated in the 1990s (reviewed in [36]), followed by a deluge of animal examples in the 2000s (reviewed in [8,37–39]).

The publication of unambiguous examples of homoploid hybrid speciation has stimulated interest in the genomic feature large genome sizes relative to parental species. While the hybrid species of Helianthus annuus and Helianthus petiolaris. These hybrid species have different geographical and temporal origins, and two species appear to have multiple origins [32–34]. All five species are diploid (2n = 34), self-incompatible and native to central and western North America (figure 1). The parental species have widespread overlapping distributions across the central and western USA [55,56]. Helianthus annuus is found in mesic, clay-based soils, whereas H. petiolaris occurs in drier, sandier soils [10]. By contrast, the three hybrid species are extremophiles, occurring in sand dune (H. anomalus), desert floor (H. deserticola) and salt marsh (H. paradoxus) habitats [57]. Genetic and ecological studies indicate that transgressive phenotypic variation (i.e. variation outside the range of their parental species) and the hybrid gene combinations underlying this variation allowed the hybrid species to colonize these extreme habitats [25,26,58,59]. In addition to habitat isolation, the hybrid lineages are reproductively isolated by large-scale karyotypic changes [31,48,60]. Analyses of the sizes and distribution of parental chromosomal segments in the three hybrid species further indicate that reproductive isolation probably arose quickly during speciation [61,62] and that the process is surprisingly repeatable [40,58].

Intriguingly, the sunflower homoploid hybrid species feature large genome sizes relative to parental species. While the genomes of H. petiolaris and H. annuus are approximately 3.3

Figure 1. Phylogenetic network based on a random subset of 11,522 high-quality SNPs genotyped for all individuals.
and 3.5 Gb, respectively, the genomes of the hybrid species range from approximately 5.3 Gb in *H. deserticola* and *H. paradoxus* to 5.6 Gb in *H. anomalus* [63], with the difference in size largely accounted for by proliferation of TEs in each of the separate hybrid lineages [49,50]. Numerous classes of TEs exist in plant genomes, and long terminal repeat (LTR) retrotransposons are the most abundant and variable. Related to infectious retroviruses, these elements transpose through an RNA intermediate and thus individual elements can give rise to numerous daughter copies capable of inserting elsewhere in the genome [64]. Regulation of TEs in host genomes is mainly controlled by epigenetic mechanisms [65], and under most circumstances the vast majority of them are suppressed and rendered inactive. Only under specific conditions, such as during hybridization or stress, can a breakdown in gene silencing mechanisms reactivate these elements [66,67]. In many plant species, including the three sunflower homoploid hybrid species, the high replicative capacity of these elements has been associated with genome expansion [49,68].

The origins of *H. deserticola*, *H. paradoxus* and *H. anomalus* via hybridization between the same two parental species allow for unique comparative analysis of LTR retrotransposon activity and proliferation, because elements in the hybrid species are necessarily derived from the parental species genomes. While both major classes of LTR retrotransposons (i.e. *Gypsy* and *Copia*) have undergone proliferation events, the dynamics differ among species. For example, *Gypsy* sequences exhibit clear patterns of large-scale proliferation in all three sunflower hybrid species [49], whereas *Copia* sequences show differential patterns of proliferation, with *H. paradoxus* having experienced larger-scale proliferation of these sequences compared with *H. deserticola* and *H. anomalus* [68]. Interestingly, proliferation of LTR retrotransposons, while of massive scale in the sunflower hybrid species, does not appear to be a common feature of contemporary *H. annuus* × *H. petiolaris* natural hybrid populations [69]. Both *Gypsy* and *Copia* sequences remain transcriptionally active, however, in the parental species (*H. annuus* and *H. petiolaris*), in early-generation *H. annuus* × *H. petiolaris* hybrid genotypes generated through controlled crosses and found naturally, and in the homoploid hybrid species themselves [69,70].

The objectives of this study are to examine transcriptional activity of LTR retrotransposons (hereafter referred as TEs) within and between the annual sunflowers *H. annuus*, *H. petiolaris* and their homoploid hybrid derivatives using high-throughput transcriptome sequencing technologies (RNASeq). In addition, we characterized genetic diversity in these species for candidate genes hypothesized to regulate TE activity. RNASeq advances previous studies of TE activity in hybrid sunflowers by providing precise identification of TE variants and their estimated transcriptional activity. Expanded sampling of homoploid hybrid taxa reveals that abundant variation in TE transcript accumulation occurs within species. Comparison of hybrid species’ TE transcript levels to those of parental species indicates that relatively few of the TEs examined (less than 5%) differ in expression between hybrids and parents, and that only a few TEs are overexpressed in multiple hybrid species relative to parent species. Analyses of sequence and expression diversity of candidate TE regulatory loci suggest the influence of divergent selection on these loci in hybrid species lineages, possibly contributing to differences in observed patterns of TE activity. Finally, we highlight one candidate gene displaying multiple hallmarks consistent with a direct role in repressing TE activity in one of the hybrid species.

### Table 1. Summary of the TEs identified by Gill et al. [74], 914 contigs. The group ‘Other elements’ is further broken into subgroups based on BLAST results.

<table>
<thead>
<tr>
<th>Transposable element family</th>
<th>no. elements</th>
<th>mean length of elements (base pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gypsy</strong></td>
<td>100</td>
<td>6603</td>
</tr>
<tr>
<td><strong>Copia</strong></td>
<td>37</td>
<td>6018</td>
</tr>
<tr>
<td><strong>Other elements</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-LTR retro</td>
<td>2</td>
<td>2961</td>
</tr>
<tr>
<td>unclass retro</td>
<td>35</td>
<td>3971</td>
</tr>
<tr>
<td>uncharacterized</td>
<td>17</td>
<td>9121</td>
</tr>
<tr>
<td>DNA (pingpong)</td>
<td>1</td>
<td>1341</td>
</tr>
<tr>
<td>top hit non-TE</td>
<td>44</td>
<td>6047</td>
</tr>
<tr>
<td>no hits</td>
<td>678</td>
<td>3280</td>
</tr>
</tbody>
</table>

2. Material and methods

(a) Plant collection and transcriptome sequencing

Achenes (single seeded fruits) representing 14 Helianthus annuus, 14 H. petiolaris, eight H. petiolaris × H. annuus F1 hybrids, three H. anomalous, seven H. deserticola and eight H. paradoxus spanning the range of each species were acquired either from USDA collections, previous sampling efforts or laboratory crosses for F1 hybrids (figure 1 and the electronic supplementary material, table S1). For each individual, we extracted RNA from young leaves and stems using a modified TRIzol reagent protocol (Invitrogen, Carlsbad, CA). All reads were sequenced on an Illumina (San Diego, CA) GAII or HiSeq next-generation sequencing platform (paired end reads, 2 × 100 bp, non-normalized libraries). Note that H. petiolaris and H. annuus individuals were sequenced as part of a larger study on genomic islands of divergence in wild sunflowers and are reported in detail in Renaut et al. [71]. Raw sequences from the three hybrid species (H. anomalus, H. deserticola and H. paradoxus) have been described and made publicly available previously [72]. Finally, the F1 hybrid sequences were previously described in Rowe & Rieseberg [73].

(b) Reference datasets and alignments

We used a reference dataset of 914 TEs (table 1) derived from the sequences of 96 randomly chosen BACs and described in details in Gill et al. [74]. This set of candidate TE nucleotide sequences was annotated (tBLASTx) using Uniprot Protein NR database (release-2013_01), and queried against GO databases (blast2GO, [75], electronic supplementary material, table S3). Additionally, candidate TEs were aligned to a published [76] Helianthus TE reference set (BLAST, e-value < 10−10, only best hit retained). The combined results were used to classify sequences as LTR retrotransposon *Copia*, LTR retrotransposon *Gypsy* or Other element (table 1).

Reads were aligned against a reference *H. annuus* transcriptome using the Burrows–Wheeler aligner (bwa, aln and sampe commands [77]). The transcriptome reference consisting of 51 468 contigs (51.3 M base pairs) is available on dryad (www.datadryad.org) and described in Renaut et al. [71]. Reads were aligned to the reference set of 914 TEs using the same approach. Aligned files (bam format) were sorted using samtools sort utility.
Table 2. Summary statistics of sequence alignments for the six species analysed here.

<table>
<thead>
<tr>
<th>species</th>
<th>no. reads aligned in millions (% total reads)</th>
<th>mean (95% CI) no. reads aligned per reference contig</th>
<th>no. contigs with &gt; 2 reads aligned</th>
<th>no. reads aligned in thousands (% total reads)</th>
<th>mean (95% CI) no. reads aligned per TE</th>
<th>no. contigs with &gt; 2 reads aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. annuus</td>
<td>17.0 (62%)</td>
<td>331 (284 – 378)</td>
<td>26 139</td>
<td>34 (0.11%)</td>
<td>37.6 (18 – 57)</td>
<td>460</td>
</tr>
<tr>
<td>H. petiolaris</td>
<td>24.8 (66%)</td>
<td>464 (376 – 552)</td>
<td>24 782</td>
<td>38 (0.08%)</td>
<td>34 (7 – 62)</td>
<td>421</td>
</tr>
<tr>
<td>F₁ hybrids</td>
<td>15.8 (55%)</td>
<td>306 (297 – 317)</td>
<td>32 850</td>
<td>34 (0.12%)</td>
<td>42 (13 – 71)</td>
<td>445</td>
</tr>
<tr>
<td>H. anomalus</td>
<td>28.8 (66%)</td>
<td>598 (553 – 643)</td>
<td>24 345</td>
<td>71 (0.15%)</td>
<td>77 (34 – 121)</td>
<td>497</td>
</tr>
<tr>
<td>H. deserticola</td>
<td>23.5 (64%)</td>
<td>442 (393 – 490)</td>
<td>24 560</td>
<td>78 (0.19%)</td>
<td>85 (52 – 118)</td>
<td>517</td>
</tr>
<tr>
<td>H. paradoxus</td>
<td>21.0 (64%)</td>
<td>428 (364 – 492)</td>
<td>24 351</td>
<td>73 (0.19%)</td>
<td>77 (50 – 104)</td>
<td>471</td>
</tr>
</tbody>
</table>

(c) Gene and transposable element expression

Raw estimates of transcript accumulation were obtained using BEDTOOLS (coverageBed) to determine the number of sequence reads mapping to each contig [78]. Comparisons of the accumulation of individual transcripts between sample groups per species were conducted within the R [79] package DESeq [80]. This program normalizes raw read counts based on aligned library size but not sequence length, as all comparisons are performed within a given transcript using a modified Fisher’s exact test of data fit to a binomial distribution. Adjusted p-values (q-values < 0.05 [81]) were used to determine statistical significance of comparisons. We then normalized TE transcript estimates first by the total number of reads aligned for each sample and then by the length of the reference gene or TE (fragments per kilobase per million fragments mapped (FPKM) value [82]). TE expression estimates (normalized read counts per sample) were summed across element classes Copia, Gypsy and Other elements to create TE expression phenotypes. Species differences in these aggregate TE expression phenotypes were assessed via linear modelling in R.

A previously identified set of 107 LTR retrotransposons with insertion age estimates for the H. annuus genome [76] were matched to our own reference set of TEs by BLAST search. We then calculated whether element age was correlated with levels of transcript accumulation in any of the five sunflower species.

(d) Evaluation of candidate transposable element regulatory genes

We constructed a reference dataset of candidate genes presumed to be involved in repressing TEs in *Arabidopsis thaliana*. First, DNA sequences of *A. thaliana* genes assigned to the Gene Ontology term GO:0016441 (post-transcriptional gene silencing, 131 genes) were obtained from The Arabidopsis Information Resource (www.arabidopsis.org). Literature searches identified 14 additional genes with empirical evidence of involvement in TE repression [65,83–85]. These genes were then compared (BLASTX, e-value < 1 x 10^-10) with all 51 468 genes in the *H. annuus* reference transcriptome. Two hundred and forty-five genes in the reference transcriptome matched these criteria and were subjected to analysis as candidate TE regulatory genes.

We used the weighted gene co-expression network analysis (WGCNA) package in R to cluster these candidate regulatory genes by principal component analysis [86]. Genes showing no transcript variance were excluded from the analysis. First, we evaluated the strength of module membership for individual genes (correlation of individual gene transcript accumulation estimate with module eigenvector). We then evaluated the correlation of individual candidate regulatory genes and co-expressed gene modules with TE expression phenotypes (Pearson correlation coefficient with Bonferroni-corrected p-values). Genes significantly correlated with TE expression phenotypes and with Pearson correlation coefficients greater than 0.4 were retained for sequence diversity analyses and hereafter referred to as TE regulator genes.

(e) Variant calling, population genetics and selection

Because relationships among populations may not conform to a tree-like bifurcating pattern owing to introgression and shared ancestral polymorphisms, we performed a phylogenetic network analysis using the neighbour-net method implemented in SPLITSTREE4 [87]. We used SAMTOOLS (mpileup and bcftools [88]) to call single nucleotide polymorphisms (SNPs) using information from all samples for a random set of 1000 genes. From this, we compiled an artificial nucleotide sequence comprising 11 522 high-quality (overall missing data less than 10%) SNPs. We then used these markers to generate a phylogenetic network in SPLITSTREE4 using default parameters (figure 1).

We used SAMTOOLS (mpileup and bcftools [88]) to identify SNPs between each pair of samples (table 2). SNPs with more than 20% missing data were removed. We also filtered out SNPs as described previously in Renaut et al. [71]. Briefly, SNPs with low expected heterozygosity (He < 0.2) were removed given that they probably represent either sequencing errors or rare alleles with little information content for interspecific comparisons. We also filtered out SNPs with very high observed heterozygosity (Ho > 0.6) because they probably represent paralogous sequence variants. From this curated dataset, FST values [89] were calculated for each marker and each species pair, using the R package HIERFSTAT [90]. We also calculated genetic diversity (θ) using sites [91].

We predicted open reading frames in our reference transcriptome and tested whether the ratio of non-synonymous to synonymous fixed differences was greater than the ratio of non-synonymous to synonymous polymorphisms using a g-test [92]. As an extension of this approach, we estimated the average proportion of amino acid substitutions driven by positive selection (alpha [93]). Based on empirical distributions of FST values that showed a bimodal distribution (see electronic supplementary material, figure S1), SNPs with an FST value greater or equal to 0.8 were considered as substitutions (D). We calculated alpha per species pair, first for all polymorphic genes, then for the subset of TE regulator genes. Significance values were calculated by resampling (with replacement, number of resamples equal to number of TE regulator genes) for each species pair.
Figure 2. Abundance of transcript reads from individual samples aligned to putative transposable element sequences classified as (a) Gypsy-like, (b) Copia-like or (c) Other elements. Sample groups (‘species’) are arrayed along the x-axes, with each point representing an individual sample. y-Axes indicate transcript estimates normalized by library size. ANOVA revealed significant differences among species per sample groups (Gypsy: $F = 6.2$, p-value = $1.7 \times 10^{-4}$; Copia: $F = 7.8$, p-value = $2.0 \times 10^{-3}$; Other elements: $F = 7.7$, p-value = $2.2 \times 10^{-3}$). Within a given element class, species per sample groups labelled with the same letter (A, B, C) do not significantly differ (pairwise t-test, p-value > 0.05).

Table 3. Interspecific comparisons of transcriptional activity for 914 putative TEs identified in H. annuus genomic DNA sequence. Comparisons were performed between parental and hybrid species only. Diagonal (grey box): number of putative TEs showing evidence of transcription (>=2 read pairs aligned); above the diagonal: number of putative TEs showing significant (q-value < 0.05) differences in pairwise comparison of transcriptional activity; below the diagonal: proportion of significant comparisons where the hybrid shows higher transcription levels. Note that because only significant comparisons are shown here, the proportions reported for H. anomalus in the lower diagonal are based on very few comparisons (1 and 3), which accounts for apparent differences relative to figure 2.

<table>
<thead>
<tr>
<th>N</th>
<th>H. annuus</th>
<th>H. petiolaris</th>
<th>H. anomalus</th>
<th>H. deserticola</th>
<th>H. paradoxus</th>
<th>H. annuus × petiolaris F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. annuus</td>
<td>860</td>
<td>20</td>
<td>1</td>
<td>19</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>H. petiolaris</td>
<td>n.a.</td>
<td>849</td>
<td>3</td>
<td>23</td>
<td>30</td>
<td>69</td>
</tr>
<tr>
<td>H. anomalus</td>
<td>1.00</td>
<td>1.00</td>
<td>815</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>H. deserticola</td>
<td>0.58</td>
<td>0.83</td>
<td>n.a.</td>
<td>866</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>H. paradoxus</td>
<td>0.58</td>
<td>0.87</td>
<td>n.a.</td>
<td>n.a.</td>
<td>851</td>
<td>n.a.</td>
</tr>
<tr>
<td>H. annuus × petiolaris F1</td>
<td>0.08</td>
<td>0.49</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>764</td>
</tr>
</tbody>
</table>

3. Results

(a) Alignments and summary statistics

Between 17 and 29 million reads (approx. 65% of all reads) aligned to the reference transcriptome of 51,468 contigs (table 2). A much smaller proportion of reads aligned to the TE reference of 914 contigs (0.1–0.2%). Nevertheless, this represents approximately 55,000 reads aligned per individual to the TE reference (table 3), which is sufficient to produce quantitative estimates of TE expression variation.

Based on a concatenated sequence of 11,522 SNPs, we plotted a phylogenetic network of all individuals (figure 1). While most individuals cluster as expected, H. anomalus did not form a distinct cluster. Limited sampling of this species or multiple hybrid origins [52] may contribute to the poor clustering (figure 1).

(b) Gene and transposable element expression

Of 914 candidate TE sequences extracted from H. annuus genomic sequence, only 14 showed no evidence of transcription across all 54 samples (electronic supplementary material, table S3). However, for another 104 putative TEs, no sample contained more than four reads aligning to the reference sequence. Fewer than 10% of putative TEs showed mean expression across all samples greater than one FPKM, with sample groups ranging from 7.1% of putative TEs transcribed at greater than one FPKM in H. petiolaris to 13.6% in H. deserticola.

TE expression phenotypes (aggregate transcript levels for elements assigned to Gypsy, Copia, or Other elements) show positive transgression in hybrid species, with both H. paradoxus and H. deserticola displaying transcript levels significantly higher than either parental species for all three TE categories (figure 2). The third hybrid species, H. anomalus, exhibited positive transgression for Copia, but not for Gypsy or Other elements. By contrast, F1 hybrids express intermediate (or additive) transcript levels with respect to the parental species (figure 2). Intraspecific variation in TE transcript accumulation varied among species (Levene’s test of equality of variances, $F_{3,48} = 3.0$, 4.0 and 2.8, p-value = 0.02, 0.004 and 0.03 for the Gypsy, Copia and Other elements categories, respectively). In addition, variance appeared higher in H. paradoxus and H. deserticola (but not H. anomalus) than in the parental species or F1 hybrids (figure 2).

Pairwise comparisons of transcript accumulation per individual TEs among all groups revealed that 782 (85%) showed no significant difference in inferred transcript accumulation (number of aligned reads) for any comparison (table 3). The
number of TEs showing significant differences between the parental species and their hybrid species derivatives ranged from as low as one between *H. annuus* and *H. anomalus* (presumably owing to low sample size for *H. anomalus*) to as high as 30 between *H. petiolaris* and *H. paradoxus* (table 3). As expected, the majority of significant expression changes were due to increased expression in the hybrid species. While comparisons between F1s and parental species revealed a greater number of significant differences in TE expression, less than half of these TEs showed increased transcript in the F1 hybrids relative to the parental species.

In addition, comparisons were performed within *H. annuus* (between *H. annuus* and subspecies *H. annuus texanus*), and within *H. petiolaris* (between subspecies *H. petiolaris petiolaris* and *H. petiolaris fallax*), with no significant subspecies differences in transcription of putative TEs identified (Fisher’s exact test, *q*-value > 0.05 for all comparisons, data not shown).

Element age from a previously identified set of 107 LTR retrotransposons [76] does not appear to correlate with levels of transcript accumulation in any of the five species examined here (data for *H. annuus* in the electronic supplementary material, figure S2, all other comparisons similar and n.s.).

(c) Evaluation of candidate transposable element regulatory genes

Analysis of transcript patterns for 236 candidate genes within WGCNA indicated that these formed two clusters of 167 and 28 genes, with 41 genes remaining unclustered (electronic supplementary material, table S2). The larger cluster (labelled turquoise; figure 3) showed strong correlation with TE expression phenotypes (*Pearson’s* *r* (*p*-value): *Gypsy* 0.84 (2 × 10^−15), *Copia* 0.81 (8 × 10^−14), Other elements 0.74 (2 × 10^−15)). Cluster membership (correlation of individual locus transcript level with the cluster eigenvalue) was strongly and positively correlated with TE expression phenotypes (*Pearson’s* *r* (*p*-value): *Gypsy* 0.91 (5.5 × 10^-85), *Copia* 0.71 (6.4 × 10^-72), Other elements 0.83 (1.1 × 10^-45)).

A set of 170 genes significantly correlated with TE expression phenotypes and with Pearson correlation coefficients greater than 0.4 were retained for sequence diversity analyses (grey area in figure 3). These are hereafter referred to as *TE regulatory genes*.

(d) Variant calling, population genetics and selection

We compared population genetic aspects of these *TE regulatory genes* identified through WGCNA with a larger set of expressed genes. The total number of SNPs identified for each of the seven species pairs was between 162 and 224 thousand (mean = 191 thousand per species pair, approx. three SNPs per kb of reference sequence). By contrast, few SNPs were identified in the putative TE sequences (mean = 63 per species pair, approximately 0.02 SNPs per kb of reference sequence).

Next, we calculated global *F*_{ST} for each of seven species pairs (figure 4 and the electronic supplementary material, figure S1). While *F*_{ST} varied among species pairs (Kruskal–Wallis rank-sum test, *χ²* (6, *N* = 3 139 052) = 252 035, *p*-value < 2 × 10^−16 for the species pair effect), estimates remained similar whether based on all genes or the subset *TE regulatory genes* (Kruskal–Wallis rank-sum test, *χ²* (1, *N* = 3 139 052) = 0.05, *p*-value = 0.83 for the gene category effect).

(e) Identification of a candidate transposable element repressor gene

Here, we present one promising gene as an example of how combining different lines of evidence can suggest candidates for TE regulation. First, a BLAST search revealed that this candidate gene is similar to a *chromatin remodelling factor of the CHD3 group* in Ambidopsis thaliana, a member of a conserved group of negative transcriptional regulators (figure 7a) [94,95]. Additionally, this transcribed sequence is overexpressed in hybrid species compared with parental species (figure 7b), shows strong correlation with aggregate *Gypsy* transcript levels (figure 7b), and possesses four non-synonymous fixed mutations that differentiate *H. paradoxus* from both parental species, in addition to one synonymous mutation that differentiates *H. petiolaris* from *H. petiolaris* (figure 7c).
4. Discussion

Hybridization in plants can reactivate dormant TEs, contributing to genome expansion and restructuring [96]. Yet, for viable hybrid populations to persist, mosaic genomes exposed to novel TEs must prevent selfish elements from proliferating uncontrollably [97]. The fact that the hybrid species show unique cases of transgressive TE expression suggests that...
these lineages have potential redundancies for preventing the expansion of some elements, while perhaps having lost the capacity to suppress others. Fundamentally, it implies an evolutionary arms race between the host and its genetic para-sites where both players must adapt to one another in order to reproduce and thrive.

Here, sampling homoploid hybrid sunflower taxa revealed abundant variation in TE transcript accumulation. High-throughput transcriptome sequencing (RNAseq) allowed simultaneous quantification of gene expression for a large number of genes, identification of polymorphic sites and measurement of genetic divergence in loci potentially involved in TE repression. Patterns of transgressive TE expression in these hybrids suggest that relatively few elements are highly transcribed, and that few elements show consistent differences in expression across species. Combining TE transcript patterns with model species annotation identified TE regulatory candidate genes that show patterns of sequence diversity consistent with evolution under divergent natural selection.

(a) Expression divergence

The majority of sequences in our reference set of 914 putative TEs, extracted from H. annuus genomic sequence, showed some level of transcriptional activity within our set of 54 transcriptome samples from five sunflower species and one set of first generation interspecific hybrids. As observed for systems such as Arabidopsis spp. (thaliana and lyrata), maize and rice, a relatively small proportion of elements appear to account for the bulk of transcription [98,99]. The hybrid species showed an overall trend towards higher accumulation of putative TE transcripts, both in summed levels of TE transcript observed and the number of individual elements contributing transcripts. F1 hybrids showed aggregate TE transcript levels similar to those of parental species, and for individual TEs showing significant differences in transcript levels between F1 and parental species, F1 transcript estimates were often lower (figure 2 and table 3). Once possible explanation for this pattern is the concerted action of dominant alleles from both parental species to reduce TE transcription in F1 hybrids, and if this were the case, homologous recombination should break up co-adapted allele combinations in the next generation. Additionally, both parental species contain substantial levels of genetic variation that may combine to produce effects on TE expression that were not observed in this or prior studies. While it will be extremely insightful to expand our studies to include later-generation hybrids and a broader sampling of existing variation, at this point, and in combination with prior observations of early-generation H. annuus−H. petiolaris hybrids [70], our results argue that increased TE activity in hybrid species is not an immediate or necessary consequence of hybridization [69,70].

These results are in general accord with Ungerer & Kawakami [70] who found non-additive expression of Gypsy elements for H. deserticola and H. paradoxus, although here this pattern extends to additional classes of elements, probably owing to the increased sensitivity of RNAseq. For the group of putative elements not specified as Gypsy or Copia (i.e. Other elements), H. petiolaris actually showed significantly lower TE-associated transcript levels than any other sample group (figure 2). Building a genomic reference specific for H. petiolaris may help to clarify whether this difference reflects more strict control of TE transcription in H. petiolaris or the presence of divergent TEs not present in the H. annuus genome.

While mean values for aggregate TE expression phenotypes demonstrate that the hybrid species accumulate higher
levels of TE-associated transcripts than parental species or F1 hybrid genotypes, the sunflower hybrid species were quite variable both in overall TE transcript levels and the identity of elements contributing to these aggregate TE expression phenotypes. This suggests that the mechanism(s) regulating transcription of TEs in these hybrid species are not highly specific to particular elements and that intraspecific variation in TE regulation is present. It is currently unclear whether elevated transcriptional activity of TEs in the hybrid species underlies differences in genomic copy number of these sequences in the hybrid versus parental sunflower taxa and whether amplification of these elements and genome expansion in the hybrid species is an ongoing phenomenon. Additional experiments in the five sunflower species designed to assay insertional activity of these sequences will be required to address this issue. For example, copy number variation could be assessed through genome resequencing efforts [100,101], which are currently underway for several species of sunflowers.

Moreover, while the initial set of candidate genes gathered from the literature represents a good starting point for further study, the specific genetic mechanisms that lead to TE expression differences remain nebulous. Do elevated expression levels for particular variants in the hybrid species simply reflect higher copy numbers for these sequences, or might expression differences result from epigenetic TE silencing mechanisms [65,102] that are differentially effective among these species? It is also unknown what triggered the amplification of TEs and what maintains elevated transcriptional activity of these sequences in contemporary populations of these hybrid species. Detailed analyses of genomic sequence from these hybrid species may provide insights into how observed variation in transcription contributes to variation in genome size and structure.

(b) Genetic divergence
Following quantification of expression divergence, we calculated several genetic parameters to compare the complete
of genetic and phenotypic novelty. This paper shows that TE amplification in hybrid lineages is more idiosyncratic than previously documented and provides a first step towards identifying the gene(s) and evolutionary mechanisms responsible for regulating and repressing TE expansions.

In the future, it would be useful to expand the study of TE evolution both horizontally—to assess whether our findings in Helianthus hybrid lineages can be extended to other organismal groups such as the Stephanoheria system studied by Gottlieb—and vertically to functionally validate the candidate transcription regulator genes found in this study. Both kinds of studies offer the opportunity to assess the repeatability of genomic changes in hybrid evolution and may provide clues regarding potential abiotic or epigenetic factors that trigger TE amplification in the first place.

Several other puzzles about TE evolution in hybrid lineages remain to be solved. Most importantly, we need to clarify the importance of hybridization (or the interaction of divergent parental genomes) versus other evolutionary processes in the TE expansions. Why, for example, do we see little evidence of TE amplification in contemporary F1 sunflower hybrids? A similar result from studies of interspecific Arabidopsis F1 crosses [104] suggests that upregulation of TEs is not a general phenomenon. Empirical evidence is also required to assess causality between the TE expansion reported here and the large-scale karyotypic and phenotypic changes observed in the homoploid hybrid species.

Thirty years ago, Gallez & Gottlieb [34] argued that the availability of electrophoretic techniques to identify large numbers of genetic loci would allow evolutionary biologists to estimate the likelihood of homoploid hybrid speciation. They were right, although it took hundreds (or thousands) of markers made available by DNA-sequence-based marker technologies to finally fulfill their prophecy. They also predicted that the most interesting examples of homoploid hybrid speciation would ‘be those in which the diploid parents appear to be strongly distinct in a genetic sense’. The many fascinating genomic changes that have accompanied homoploid hybrid speciation in sunflowers, which involve highly divergent parental species, supports this claim, although further studies of the genomic consequences of homoploid hybrid speciation across a continuum of parental divergence will be required to fully validate this second prediction.

Acknowledgements. We thank Navdeep Gill for providing us an advance access to the H. annuus TE reference.

Funding statement. This work was supported by an NIH Postdoctoral Fellowship to H.C.R., an NSERC Postdoctoral Fellowship to S.R., an NSERC grant no. (327475) to L.H.R., and NSF grant no. DEB-0742993 to M.C.U.

References

10. Heiser Jr CB. 1947 Hybridization between the sunflower species Helianthus annuus and