CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

# The Genome of *Cucurbita argyrosperma* (Silver-Seed Gourd) Reveals Faster Rates of Protein-Coding Gene and Long Noncoding RNA Turnover and Neofunctionalization within *Cucurbita*

Josué Barrera-Redondo[1], Enrique Ibarra-Laclette[2], Alejandra Vázquez-Lobo[3], Yocelyn T. Gutiérrez-Guerrero[1], Guillermo Sánchez de la Vega[1], Daniel Piñero[1], Salvador Montes-Hernández[4], Rafael Lira-Saade[5,*] and Luis E. Eguiarte[1,*]

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Circuito Exterior s/n Anexo al Jardín Botánico, 04510 Ciudad de México, Mexico

[2]Departamento de Estudios Moleculares Avanzados, Instituto de Ecología A.C., Carretera Antigua a Coatepec No. 351, Col. El Haya. C.P., Xalapa, Veracruz 91070, Mexico

[3]Centro de Investigaciones en Biodiversidad y Conservación, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Col. Chamilpa, Cuernavaca, Morelos 62209, Mexico

[4]Campo Experimental Bajío, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias (INIFAP), Km 6.5 Carretera Celaya-San Miguel de Allende, Celaya, Guanajuato 38110, Mexico

[5]UBIPRO, Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México, Av. de los Barrios #1, Col. Los Reyes Iztacala, Tlanepantla, Edo. de Mex 54090, Mexico

***Correspondence: Rafael Lira-Saade (rlira@unam.mx), Luis E. Eguiarte (fruns@unam.mx)**

https://doi.org/10.1016/j.molp.2018.12.023

## ABSTRACT

**Whole-genome duplications are an important source of evolutionary novelties that change the mode and tempo at which genetic elements evolve within a genome. The *Cucurbita* genus experienced a whole-genome duplication around 30 million years ago, although the evolutionary dynamics of the coding and noncoding genes in this genus have not yet been scrutinized. Here, we analyzed the genomes of four *Cucurbita* species, including a newly assembled genome of *Cucurbita argyrosperma*, and compared the gene contents of these species with those of five other members of the Cucurbitaceae family to assess the evolutionary dynamics of protein-coding and long intergenic noncoding RNA (lincRNA) genes after the genome duplication. We report that *Cucurbita* genomes have a higher protein-coding gene birth–death rate compared with the genomes of the other members of the Cucurbitaceae family. *C. argyrosperma* gene families associated with pollination and transmembrane transport had significantly faster evolutionary rates. lincRNA families showed high levels of gene turnover throughout the phylogeny, and 67.7% of the lincRNA families in *Cucurbita* showed evidence of birth from the neofunctionalization of previously existing protein-coding genes. Collectively, our results suggest that the whole-genome duplication in *Cucurbita* resulted in faster rates of gene family evolution through the neofunctionalization of duplicated genes.**

**Key words:** *Cucurbita argyrosperma*, comparative genomics, molecular evolution, neofunctionalization, long noncoding RNA, whole-genome duplication

**Barrera-Redondo J., Ibarra-Laclette E., Vázquez-Lobo A., Gutiérrez-Guerrero Y.T., Sánchez de la Vega G., Piñero D., Montes-Hernández S., Lira-Saade R., and Eguiarte L.E.** (2019). The Genome of *Cucurbita argyrosperma* (Silver-Seed Gourd) Reveals Faster Rates of Protein-Coding Gene and Long Noncoding RNA Turnover and Neofunctionalization within *Cucurbita*. Mol. Plant. **12**, 506–520.

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

## INTRODUCTION

*Cucurbita* is a genus with global agronomic relevance (Lira et al., 2016; Paris, 2016) and is one of the angiosperm genera with the highest number of independent domestication events (Nee, 1990; Zheng et al., 2013; Castellanos-Morales et al., 2018). Recent advances in the study of *Cucurbita* spp. genomes revealed a recent whole-genome duplication around 30 million years ago (Mya) in the common ancestor of the genus (Montero-Pau et al., 2017; Sun et al., 2017).

Genome duplications are important sources of evolutionary novelties in plants, since redundant elements in a genome can develop novel functions in a process called neofunctionalization (Ganfornina and Sánchez, 1999; Magadum et al., 2013). It is expected that a lineage that experienced a recent whole-genome duplication would have different gene evolution dynamics compared with other closely related species with non-duplicated genomes (Ponting et al., 2009; Magadum et al., 2013).

Even though the genomic footprints of a whole-genome duplication are strong in *Cucurbita*, the numbers of predicted protein-coding genes in *Cucurbita* genomes are roughly similar to those in other genomes of the Cucurbitaceae family (Huang et al., 2009; Garcia-Mas et al., 2012; Guo et al., 2012; Montero-Pau et al., 2017; Sun et al., 2017; Urasaki et al., 2017; Wu et al., 2017). This apparent lack of duplicated coding genes could be the result of ''pseudogenization'' processes, implying a loss of redundant genes throughout the evolution of the *Cucurbita* genomes, either by the accumulation of mutations resulting in loss of function or fractionation due to intrachromosomal recombination (Sun et al., 2017). However, duplicated coding genes can also evolve to perform novel functions through positive selection (Wang et al., 2015). These novel functions are not necessarily limited to the emergence of new protein-coding elements, since protein-coding genes can also evolve into regulatory elements as noncoding RNAs (Chen and Rajewsky, 2007). Most of the transcriptional activity in eukaryotes generates noncoding RNAs (Smith and Mattick, 2017), whose abundance correlates positively with organismal complexity, whereas the abundance of protein-coding genes does not scale with complexity (Liu et al., 2013). A particular category of noncoding transcripts called long noncoding RNAs (lncRNAs) seem to play critical roles in eukaryotic development and differentiation (Smith and Mattick, 2017).

lncRNAs are a heterogeneous group of noncoding RNAs larger than 200 nucleotides that lack coding potential (Mercer et al., 2009; Ulitsky, 2016). lncRNAs act as master regulatory genes, mainly through the recruitment of chromatin modifiers in the nucleus, such as DNA methyltransferases and histone posttranslational modifiers (Fatica and Bozzoni, 2014; Smith and Mattick, 2017), although lncRNAs can also act in the cytoplasm through sequence complementarity to other RNA molecules and the modulation of mRNA stability (Fatica and Bozzoni, 2014). In plants, lncRNAs are involved in several biological functions, including organ development, flowering and vernalization, phosphate homeostasis, photomorphogenesis, response to biotic and abiotic stress conditions such as heat stress and response to phytopathogens, alternative splicing of protein-coding genes, nodule formation, and cell-wall synthesis (Chekanova, 2015; Liu et al., 2015).

Studies regarding the evolutionary dynamics of lncRNAs are limited despite their evident importance in plant biology, due to a traditional focus on protein-coding genes in genome-wide studies (Ulitsky, 2016; Nelson et al., 2017). Furthermore, evolutionary analyses of these genes have been limited due to a lack of conservation at both the sequence level and the secondary structure level (Ulitsky, 2016), and research on their origin and evolution is still scarce (see Necsulea et al., 2014; Nelson et al., 2016; Zhao et al., 2018). Several hypotheses have been proposed to explain the emergence of new lncRNAs, such as the neofunctionalization of duplicated protein-coding genes, co-option of transposable elements in the genome, duplication followed by neofunctionalization from other lncRNAs, and *de novo* emergence (Kapusta et al., 2013). Previous studies have suggested that whole-genome duplications can lead to faster rates of evolution in lncRNA families (Ponting et al., 2009; Nelson and Shippen, 2015).

This study focuses on the effects of the *Cucurbita*-wide genome duplication on the evolutionary dynamics of both protein-coding and long intergenic noncoding RNA (lincRNA) genes in the *Cucurbita* genus. We propose that the *Cucurbita* genomes have faster gene evolutionary dynamics, including higher rates of gene birth and death, than the genomes of other members of the Cucurbitaceae family due to this whole-genome duplication. We expected that species belonging to the *Cucurbita* genus might have experienced several recent lincRNA birth events due to the whole-genome duplication (Ponting et al., 2009; Nelson and Shippen, 2015). We also analyzed the possibility that the duplication of protein-coding genes and posterior neofunctionalization may be a source of new lincRNA genes in *Cucurbita* (Kapusta et al., 2013). We explored these hypotheses by analyzing four *Cucurbita* genomes, including our novel genome assembly of *Cucurbita argyrosperma* ssp. *argyrosperma*, commonly known as cushaw or silver-seed gourd in English and ''calabaza pipiana'' in Spanish (Lira et al., 2016), by comparing the coding and noncoding genes in these genomes with those in other genome assemblies reported for the Cucurbitaceae family.

## RESULTS

### *Cucurbita argyrosperma* Genome and Transcriptome

We sequenced the genome of *C. argyrosperma* ssp. *argyrosperma* using three sequencing platforms: Illumina HiSeq2000, Illumina MiSeq, and PacBio RS II (see Supplemental Methods for detailed information on the genome and transcriptome sequencing). We obtained 38.4 Gb of data from HiSeq2000, 13.1 Gb from MiSeq, and 11.4 Gb from PacBio RS II. After applying quality filters to the data (see Supplemental Methods for detailed parameters) and filtering organelle reads, we obtained ~120× high-quality sequence coverage with the Illumina reads and ~31× coverage with the PacBio reads. We estimated the genome size of *C. argyrosperma* to be ca. 238 Mb using KmerGenie (Chikhi and Medvedev, 2014).

The chloroplast genome was assembled into a single circular contig of 157 623 bp and had the typical structures of a

| | |
|---|---|
| Assembly size | 228 814 150 bp |
| No. of scaffolds | 920 |
| Longest scaffold | 2 746 581 bp |
| $N_{50}$ of scaffolds | 620 880 bp |
| $L_{50}$ of scaffolds | 103 |
| No. of scaffolds >1 kbp | 920 (100.0%) |
| No. of scaffolds >10 kbp | 903 (98.2%) |
| No. of scaffolds >100 kbp | 455 (49.5%) |
| No. of contigs | 1481 |
| Longest contig | 2 172 140 bp |
| $N_{50}$ of contigs | 463 388 bp |
| $L_{50}$ of contigs | 132 |
| No. of contigs >1 kbp | 1481 (100.0%) |
| No. of contigs >10 kbp | 1417 (95.7%) |
| No. of contigs >100 kbp | 493 (33.3%) |
| CG content | 36.22% |
| Illumina read coverage | 120× |
| PacBio read coverage | 31× |
| No. of protein-coding genes | 28 298 |
| Protein-coding gene average size | 3457 bp |
| Protein-coding gene median size | 2627 bp |
| No. of tRNAs | 4387 |
| No. of long noncoding intergenic RNAs | 6124 |

**Table 1. *Cucurbita argyrosperma* ssp. *argyrosperma* Genome Assembly Statistics.**

chloroplast genome: a large single-copy region, a small single-copy region, and two inverted repeats (Daniell et al., 2016). The mitochondrial genome was assembled into 17 scaffolds composed of 1 062 053 bp and showed several instances of chloroplast sequence insertions, as previously described for the mitochondrial genome of *Cucurbita pepo* (Alverson et al., 2010).

We assembled the *C. argyrosperma* nuclear genome into 920 scaffolds (1481 contigs), with an $N_{50}$ of 620 880 bp (Table 1). The total length of the assembled scaffolds was ~229 Mbp, around 96% of the estimated size of the genome and similar to the assembly size of previously reported *Cucurbita* genomes (Montero-Pau et al., 2017; Sun et al., 2017). Genome completeness was assessed by finding single-copy orthologous genes conserved in embryophytes (1440) using BUSCO (Simão et al., 2015). We found complete sequences for 93.2% (1342) of the BUSCO genes and fragmented sequences for 0.9% (13) of the BUSCO genes within the *C. argyrosperma* genome assembly, suggesting a high level of assembly completeness. We also found that 80.5% of the Illumina reads and 100% of the PacBio reads used for genome assembly mapped against the assembled scaffolds, indicating that most of the sequenced genome is present in the nuclear assembly.

We sequenced the *C. argyrosperma* transcriptome using Illumina HiSeq2000, obtaining 51 Gb of RNA sequencing (RNA-seq) data. We mapped 90.69% of the transcriptome reads back to either the nuclear or the organelle assemblies, indicating a high level of genome completeness. The transcriptome was assembled both *de novo* (Grabherr et al., 2011) and using a genome-guided assembly (Pertea et al., 2015) to aid in the prediction of gene models. We predicted 4387 tRNAs and 28 298 protein-coding genes within the genome assembly, numbers similar to those reported in *C. pepo*, *Cucumis sativus*, and *Cucumis melo* (Huang et al., 2009; Garcia-Mas et al., 2012; Montero-Pau et al., 2017) (Table 1). 78.9% of the protein-coding genes were functionally annotated using InterProScan (Jones et al., 2014; Supplemental Data 4), where 51.7% of the genes could be assigned with at least one gene ontology (GO) term (Ashburner et al., 2000; The Gene Ontology Consortium, 2017).

We predicted ~78 Mbp of transposable elements (TEs) within the *C. argyrosperma* genome, corresponding to 34.1% of the genome assembly. This proportion of TEs is similar to those found within the genomes of *C. pepo* (93 Mbp, 37.8% of the genome assembly), *Cucurbita maxima* (107 Mbp, 40.3%), and *Cucurbita moschata* (106 Mbp, 40.6%) (Montero-Pau et al., 2017; Sun et al., 2017). Of the 78 Mbp of TEs, 93.6% correspond to RNA transposons, with most being LTR retrotransposons (49.09%) and LARD retrotransposons (29.36%). Just 1.95% of the observed TEs correspond to DNA transposons, and 4.44% correspond to unidentifiable TEs (Supplemental Table 1). The dominance of LTR retrotransposons within the genome of *C. argyrosperma* is similar to that in *C. pepo* (50.7%) (Montero-Pau et al., 2017), *C. moschata* (62.9%), and *C. maxima* (69.9%) (Sun et al., 2017), revealing that TE families remained relatively stable within the *Cucurbita* genus.

## Phylogeny and Evolution of Protein-Coding Gene Families

We compared the protein-coding genes of *C. argyrosperma* with those of *C. pepo* (Montero-Pau et al., 2017), *C. moschata*, and *C. maxima* (Sun et al., 2017); as well as other genera in the Cucurbitaceae family, *C. sativus* (Huang et al., 2009), *C. melo* (Garcia-Mas et al., 2012), *Citrullus lanatus* (Levi et al., 2011), *Lagenaria siceraria* (Wu et al., 2017), and *Momordica charantia* (Urasaki et al., 2017), to assess protein-coding gene family expansions and contractions within the Cucurbitaceae family, as well as within the *Cucurbita* genus. We used *Fragaria vesca* (Edger et al., 2018) and *Juglans regia* (Martínez-García et al., 2016) as outgroups.

We retrieved 23 247 protein-coding gene families, of which only 11 961 families included at least one homolog conserved in two or more different species; the remaining families were exclusive to a single species (Supplemental Figure 1). We found 698 gene families that remained multicopy within *Cucurbita* after the whole-genome duplication, and these families were functionally enriched ($p < 0.01$) in intracellular protein transport. We also found 858 gene families that remained a constant size throughout *Cucurbita* and the other Cucurbitaceae species, although we found no functional enrichment within these families. We identified 369 gene families as single-copy orthologs conserved in all Cucurbitaceae and outgroup species, which were used to obtain a time-calibrated phylogeny needed for gene family evolution analyses. The resulting species tree had approximate likelihood-ratio test (Ansimova and Gascuel, 2006) support

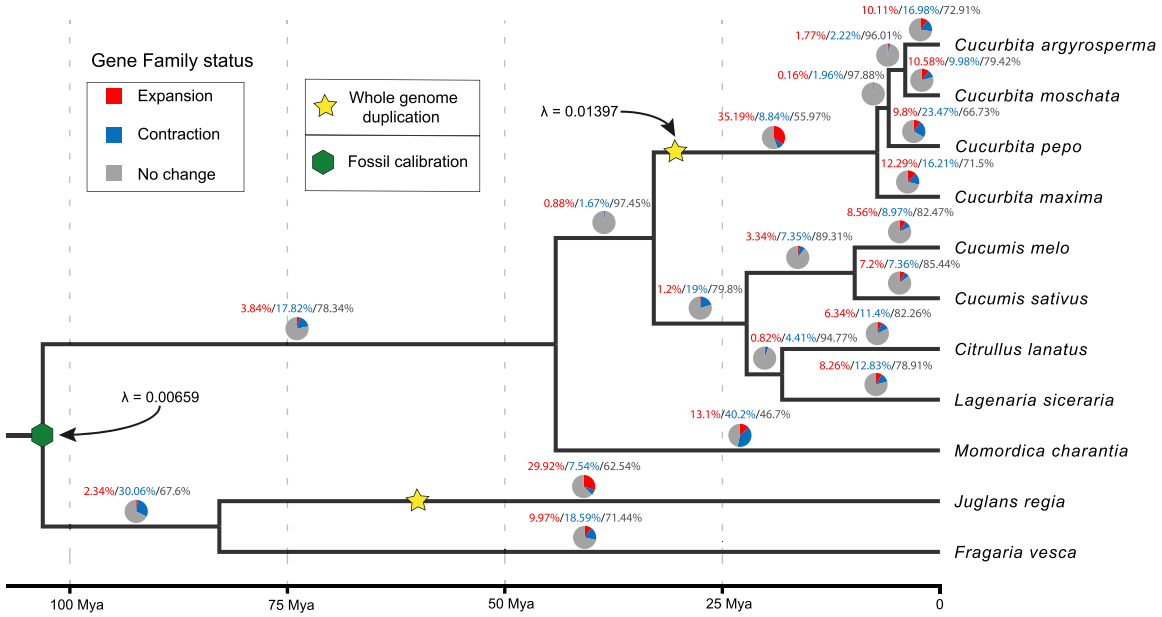


**Figure 1. Dated Phylogeny of the Cucurbitaceae Family with Protein-Coding Gene Family Expansions and Contractions per Branch.**
The phylogeny was generated with 369 single-copy orthologous genes. Fossil evidence was used to calibrate the basal node of the tree (green hexagon). The pie charts and the percentages at every branch of the tree indicate whether a gene family expanded (red), contracted (blue), or remained the same size (gray). The yellow stars indicate the estimated ages of the whole-genome duplication events in the *Cucurbita* genus (Montero-Pau et al., 2017) and in *Juglans regia* (Luo et al., 2015). The black arrows indicate the change from a basal gene birth/death rate (λ) in the most recent common ancestor of the phylogeny to a faster gene birth/death rate after the whole-genome duplication in *Cucurbita*. Every node in the phylogeny has an approximate likelihood-ratio test (aLRT) support value of 100%.

values of 100% at every node. The dated phylogeny with hishest posteior densities (HPS ± 95% confidence interval) obtained using mcmctree (Yang, 2007) supports a divergence time between *C. argyrosperma* and its sister species *C. moschata* of around 3.98 ± 1.7 Mya, while the divergence between *Cucurbita* and Benincaseae (*C. sativus* + *C. melo* + *C. lanatus* + *L. siceraria*; Schaefer et al., 2009) happened around 32.9 ± 11 Mya (Figure 1), concordant with the expected age of the whole-genome duplication event in *Cucurbita*, approximately 30 ± 4 Mya (Montero-Pau et al., 2017). The crown node of the included Cucurbitaceae species was dated at 44.1 ± 14 Mya.

We performed likelihood-ratio tests to compare the likelihood score of a global gene birth–death rate parameter (λ) across the tree against multiple λ values throughout the phylogeny. A model with a change in λ within *Cucurbita* had a significantly higher log-likelihood (−193 746.504) than a single λ (−198 328.178) throughout the tree (Figure 1 and Supplemental Figure 2). After accounting for genome assembly and annotation error rates, we found that the gene birth–death rate was twice as high in *Cucurbita* (λ = 0.01397) than in the rest of the phylogeny (λ = 0.00659).

We detected phylogenetic inconsistencies in gene content within *Cucurbita*, with some genomes containing ~32 000 protein-coding genes and others containing ~28 000 genes (Supplemental Table 2). To account for possible errors in gene prediction, we repeated the gene family analysis using only high-quality protein-coding gene predictions with annotation edit distances (eAED) lower than 0.5 (Yandell and Ence, 2012; Campbell et al., 2014) to eliminate low-quality gene models. We found a similar number of high-quality gene models in all *Cucurbita* genomes, which is much closer to the total number of predicted genes in *C. pepo* and *C. argyrosperma* (Supplemental Table 2). Even after discarding low-quality gene models, λ was still twice as high in *Cucurbita* (0.01188) than in the rest of the phylogeny (0.00566) (Supplemental Figures 3 and 4).

We found significantly rapid changes in gene family sizes ($p < 0.01$) throughout most of the branches within the phylogeny (Figure 1). We found that the branch leading to the crown node of the *Cucurbita* genus and the terminal branch of *C. argyrosperma* had unusually high rates of gene family evolution (Figure 1). Even though just a small number of gene families showed significantly rapid ($p < 0.01$) levels of change in the branch leading to the crown node of *Cucurbita* (six gene families), this branch had the second highest number of gene family changes within the entire phylogeny (Figure 1). Surprisingly, the terminal branch of *M. charantia* showed the highest number of gene family changes in the whole phylogeny (Figure 1), although there were only a few gene families with significantly rapid changes (27 gene families). Furthermore, the proportion of gene families that either expanded or contracted in the terminal branches of *Cucurbita* was higher compared with the proportion of gene families that either expanded or contracted in the terminal branches of Benincaseae (Figure 1).

The terminal branch of *C. argyrosperma* had an unusually high number of gene families with significantly rapid expansions/contractions (327 families). However, most of the rapidly evolving gene families within *C. argyrosperma* underwent contractions (78.3%), rather than expansions (21.7%). After performing

| GO ID | GO term | FDR *p* value |
|---|---|---|
| **Significantly expanded protein-coding gene families** | | |
| GO:0007018 | Microtubule-based movement | <1.729E−27 |
| GO:0006270 | DNA replication initiation | 7.8E−16 |
| GO:0006855 | Drug transmembrane transport | 4.6E−05 |
| GO:0007010 | Cytoskeleton organization | 0.00346 |
| **Significantly contracted protein-coding gene families** | | |
| GO:0042545 | Cell-wall modification | <1.729E−27 |
| GO:0006979 | Response to oxidative stress | <1.729E−27 |
| GO:0055114 | Oxidation–reduction process | <1.729E−27 |
| GO:0009733 | Response to auxin | <1.729E−27 |
| GO:0030244 | Cellulose biosynthetic process | 2.2E−22 |
| GO:0006508 | Proteolysis | 1.2E−17 |
| GO:0006887 | Exocytosis | 3.9E−06 |
| GO:0006855 | Drug transmembrane transport | 4.7E−05 |
| GO:0003333 | Amino acid transmembrane transport | 0.00048 |
| GO:0005992 | Trehalose biosynthetic process | 0.00064 |
| GO:0048544 | Recognition of pollen | 0.00064 |
| GO:0005975 | Carbohydrate metabolic process | 0.00453 |
| GO:0071577 | Zinc II ion transmembrane transport | 0.00939 |

**Table 2. Enriched Biological Functions of Rapidly Evolving Protein-Coding Gene Families in *C. argyrosperma*.**

a GO enrichment analysis, we found four overrepresented biological functions associated with the significantly expanded families in *C. argyrosperma* (Table 2), including microtubule-based movement in families mainly composed of proteins with kinesin motor domains, and drug transmembrane transport in families mainly composed of villin/gelsolin proteins. We also found 13 overrepresented biological functions associated with the significantly contracted families in *C. argyrosperma* (Table 2), including cell-wall modification in families mainly composed of pectinesterases, response to oxidative stress, oxidation–reduction processes, recognition of pollen, exocytosis, and several processes associated with transmembrane transport. Curiously, drug transmembrane transport was enriched in both significantly expanded families and significantly contracted families, which were composed mainly of multi-antimicrobial extrusion proteins.

## lincRNA Prediction and Analysis

We used Evolinc-I (Nelson et al., 2017) to predict lincRNAs within the genome assembly of *C. argyrosperma*, as well as the genomes of *C. maxima*, *C. moschata*, *C. pepo*, *C. melo*, *C. sativus*, *C. lanatus*, and *L. siceraria*. The predicted lincRNAs were compared against the protein-coding gene transcripts of each genome to determine the percentage of lincRNAs produced from the neofunctionalization of duplicated protein-coding sequences. We also compared the predicted lincRNAs against the RepBase (Bao et al., 2015) sequences from eudicots to determine the percentage of lincRNAs produced from the neofunctionalization of TEs within each genome.

Since most of the species transcriptomes used to predict lincRNAs had differences in the organs sequenced, as well as differences in sequencing depth and library construction (see Supplemental Table 3), each species had a different number of predicted lincRNAs (Supplemental Table 2), and these numbers could not be directly compared. However, we expect that the proportion of lincRNAs derived from protein-coding genes and TEs relative to the total number of predicted lincRNAs in a genome remains relatively constant, despite differences in the RNA-seq strategy. Hence, we compared this proportion between *Cucurbita* species and the other species within Cucurbitaceae. Despite the differences in the number of predicted lincRNAs, the percentage of protein-coding-derived lincRNAs was roughly similar between *Cucurbita* species, while there was more variance in this proportion between the other cucurbits (Figure 2). We found a higher percentage of protein-coding-derived lincRNAs in *Cucurbita* species than in other cucurbits (Figure 2; $p = 0.041$), which fits the coding-to-noncoding neofunctionalization hypothesis (Kapusta et al., 2013). However, the proportion of TE-derived lincRNAs was lower in the *Cucurbita* genus compared with the other taxa of the Cucurbitaceae family (Figure 2; $p = 0.013$).

We analyzed the evolution of lincRNA families across the Cucurbitaceae family by using the *C. argyrosperma* predicted lincRNAs as queries to search for homologs within all the analyzed genomes. We compared the lincRNA homologs of each lincRNA family against the protein-coding genes and the Evolinc-I predicted lincRNAs for each species to assess the relationship between lincRNAs and protein-coding genes, as well as to assess the transcriptional potential of the lincRNA homologs. Since the evolutionary proximity of *C. argyrosperma* to the other *Cucurbita* species can lead to erroneous inferences about lincRNA family expansion in this genus, we also used the predicted lincRNAs of *C. lanatus* as sequence queries in the

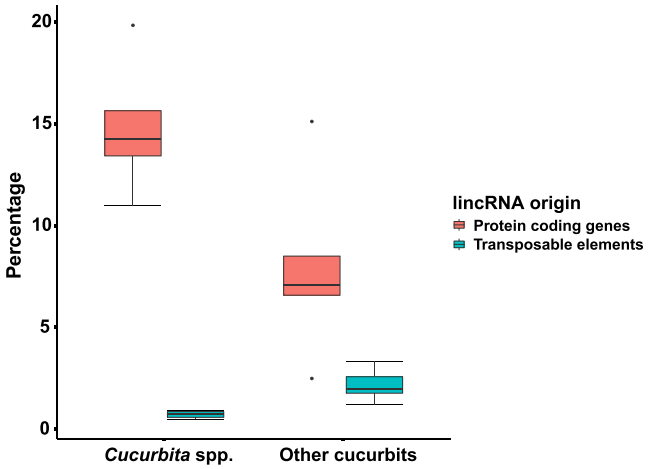


**Figure 2. Differences in the Proportion of lincRNAs Associated with Protein-Coding Transcripts (Red) and Transposable Elements (Blue) between Four *Cucurbita* Genomes (*C. argyrosperma*, *C. moschata*, *C. maxima*, and *C. pepo*) and Five Genomes of Other Cucurbit Species (*C. sativus*, *C. melo*, *C. lanatus*, *L. siceraria*, and *M. charantia*).**

*Cucurbita* spp. show a higher proportion of protein-coding gene-derived lincRNAs compared with other members of the same family ($p = 0.041$), whereas the proportion of transposable element (TE)-derived lincRNAs is lower in *Cucurbita* compared with the other cucurbit species ($p = 0.013$).

lincRNA family analysis to assess whether the patterns observed within *Cucurbita* were determined by the effect of lincRNA turnover throughout the phylogeny (Ulitsky, 2016).

We retrieved 5466 *C. argyrosperma* lincRNA families, of which 67.7% showed evidence of protein-coding gene neofunctionalization throughout their phylogenies, while only 32.3% were exclusively composed of noncoding elements. In contrast to the *C. argyrosperma* lincRNA gene families, we found that 34.3% of the 5231 *C. lanatus* lincRNA families had evidence of protein-coding neofunctionalization, while 65.7% were exclusively composed of noncoding elements. To assess the level of lincRNA conservation across the Cucurbitaceae family, we only used the subset of lincRNAs whose families were solely composed of noncoding elements, since protein-coding-derived lincRNAs can be traced to homologous protein-coding genes in distantly related taxa, therefore leading to overestimation of the conservation of lincRNAs throughout the phylogeny. This analysis showed that the conservation of lincRNAs between *C. argyrosperma* and the rest of the analyzed species steadily declines with phylogenetic distance, with an average of 2.27% of lincRNA homologs lost per million years (Figure 3A).

75.9%–92.4% of the lincRNAs found in *C. argyrosperma* had at least one homolog within the genomes of the other *Cucurbita* species, with an average lincRNA loss rate of about 2.4% per million years. Just 4.3%–8.8% of the lincRNA families had homologs within the genomes of species belonging to Benincaseae, whereas 6.5% of the lincRNA families had homologs within the genome of *M. charantia*, which is a higher percentage than that observed in some species belonging to Benincaseae. Just 1.2% of the lincRNA families in *J. regia* and 0.3% of those in in *F. vesca* were retained, and only five lincRNA homologs were conserved in both species.

Despite the general trend between phylogenetic distance and lincRNA loss, the average rate of lincRNA loss between *C. argyrosperma* and Benincaseae increased to 2.8% per million years and then declined to around 2.1% per million years in *M. charantia* and to 0.96% per million years in the outgroup species. The decline in lincRNA conservation with respect to phylogenetic distance could also be observed for *C. lanatus* lincRNAs, with an average loss rate of 2.54% per million years in the whole phylogeny (Figure 3B), although the percentage of retained lincRNAs in *J. regia* and *F. vesca* was higher (2.3% and 0.7% respectively). The average rate of lincRNA loss between *C. lanatus* and *Cucurbita* was also around 2.8% per million years. However, the lincRNA loss rate within Benincaseae was approximately 3.4% per million years, the highest rate observed in the study. The loss rate also declined in *M. Charantia* to around 2% per million years and to 0.95% per million years in the outgroup species.

Some *C. argyrosperma* lincRNA families showed a high degree of conservation within Cucurbitaceae, that is, every species had at least one representative gene within the family (1016 families). While most of these conserved lincRNA families had at least one protein-coding gene within its phylogeny (95.17%), a small percentage of the families showing a high degree of conservation were composed of putatively noncoding elements (4.82%). We found that seven of these putatively noncoding families were present as single-copy orthologs within Cucurbitaceae, and the members of these families were mostly predicted independently as lincRNAs with Evolinc-I, that is, using transcriptional evidence (Figure 4A). The five lincRNAs that were shared between *C. argyrosperma* and both outgroup species showed complex evolutionary histories, with several instances of duplications and losses, and none of them were conserved in all the analyzed species (e.g., Supplemental Figure 5).

Our analyses show a high rate of lincRNA family birth within the *Cucurbita* genus (Figure 4B). 55.94% of the lincRNA families were exclusively found in the *Cucurbita* genus. Of these gene families, 78.87% were present in all four *Cucurbita* species and just 1.7% were exclusive to *C. argyrosperma*. We found a similar pattern for the *C. lanatus* lincRNA families, where 64.61% were exclusively found in Benincaseae. However, 47.24% of these lincRNA families were exclusive to *C. lanatus*, and only 10.47% were present in all four species.

Many of the *C. argyrosperma* lincRNA families (61.2%) showed signals of gene duplication within the *Cucurbita* genus. Many of these families showed symmetric expansion within *Cucurbita* (1948 families), that is, expansion where the number of lincRNA genes remained constant within *Cucurbita* but at least two-fold higher with respect to the rest of the Cucurbitaceae species (Figure 4C). This pattern is not a product of the phylogenetic distance between *Cucurbita* species, as it could also be seen within the *Cucurbita* clade when analyzing the *C. lanatus* lincRNA families (53 families, Supplemental Figure 6). Most of the lincRNA families with symmetric expansion within *Cucurbita* also showed evidence of protein-coding neofunctionalization (62.26%). The phylogenies of some of these families showed a duplication event corresponding to the whole-genome duplication in *Cucurbita*, where a protein-coding family gave rise to a lincRNA clade

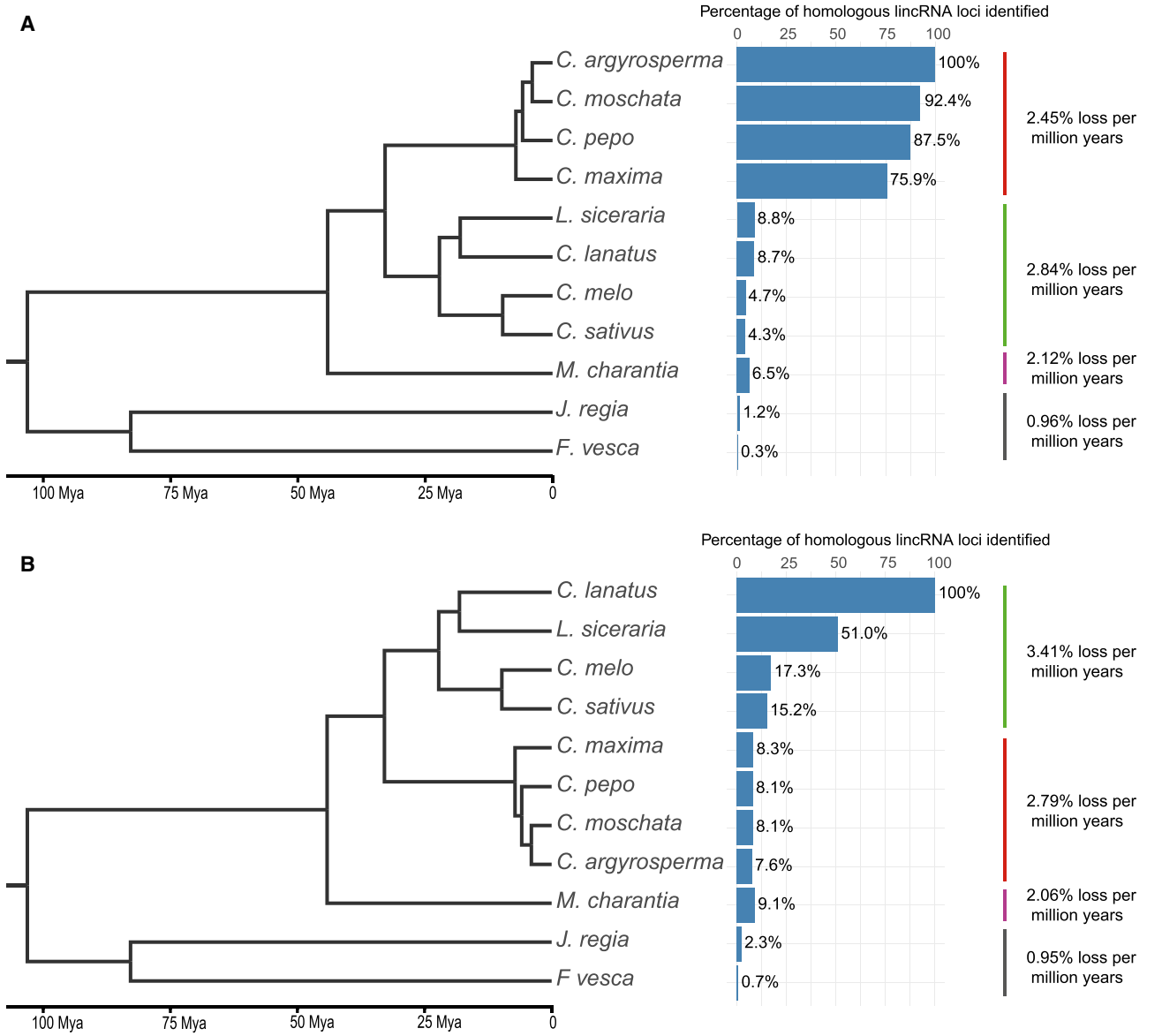


**Figure 3. lincRNA Conservation across the Cucurbitaceae Family.**
Each column alongside the phylogeny represents the percentage of homologs found within each cucurbit genome for the subset of lincRNAs without homology to the predicted protein-coding genes in the genomes of **(A)** *Cucurbita argyrosperma* and **(B)** *Citrullus lanatus*. The percentage of conserved lincRNAs between species declines drastically as the phylogenetic distance becomes larger, due to a high rate of lincRNA turnover (gene birth/death rate). Only a small fraction of lincRNAs are conserved between Cucurbitaceae and the outgroup species. The average rate of lincRNA loss per million years (right) is shown for the following clades: *Cucurbita* (red), Benincaseae (green), *Momordica charantia* (purple), and the outgroup species (gray).

(Figure 4D). After carefully inspecting one of these lincRNA families (Carg_TCONS _00015392), we found high levels of sequence conservation between the *Cucurbita* lincRNAs, although no conservation of the length, the ORF, or the codon structure of its protein-coding homolog was observed (Figure 5A). We also found a thermodynamically stable secondary structure in the lincRNA (Figure 5B and 5D), whereas the homologous protein-coding transcript showed lower structure stability, as well as signs of many equally stable structures throughout the transcript (Supplemental Figure 7), despite both the lincRNA and protein-coding transcript having similar dinucleotide frequencies (Supplemental Table 4). The structural stability of this lincRNA is even higher than that of one of the highly conserved lincRNAs found in single copy throughout the Cucurbitaceae (Carg_TCONS_00063022; Figure 5C and 5E).

## DISCUSSION

The protein-coding gene content within the *Cucurbita* species has remained relatively constant, despite the whole-genome duplication that happened around 30 ± 4 Mya (Montero-Pau et al., 2017). However, our results indicate that this genome duplication event had a profound effect on the gene evolutionary dynamics within *Cucurbita*, namely, higher rates of protein-coding gene family evolution and higher rates of

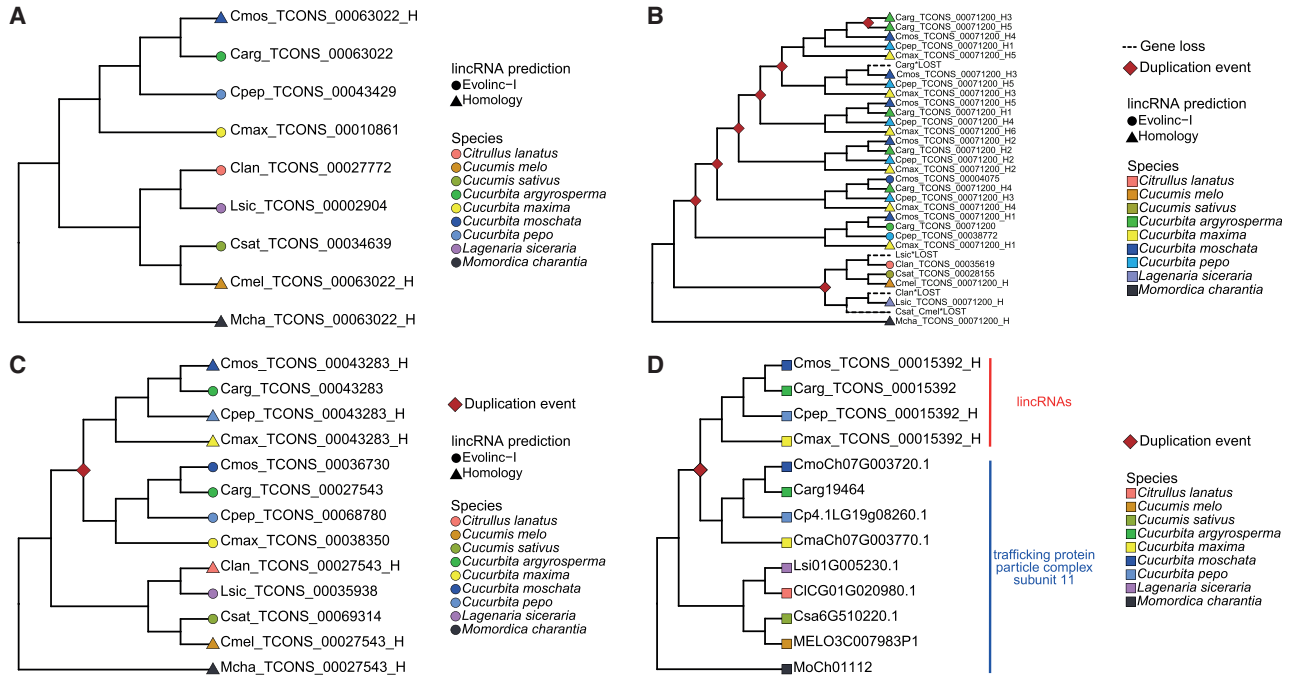


**Figure 4. Patterns of lincRNA Family Evolution throughout the *Cucurbita* Genus.**

**(A)** Presence of single-copy orthologous lincRNAs with a high degree of conservation throughout the Cucurbitaceae family suggests these lincRNAs have an essential biological function.

**(B)** Sudden duplication bursts (red diamonds) within the *Cucurbita* genus, as well as multiple gene losses (dotted branches), reveal a high rate of lincRNA turnover.

**(C)** Duplication of lincRNA families associated with the whole-genome duplication in *Cucurbita* (red diamond).

**(D)** Neofunctionalization of protein-coding genes (blue bar) into the novel Carg_TCONS_00015392 lincRNAs (red bar) after the whole-genome duplication in the *Cucurbita* genus (red diamond). Some lincRNAs were independently predicted based on homology using Evolinc-II (triangles in terminal nodes) and based on transcriptomic evidence using Evolinc-I (circles in terminal nodes), further supporting the transcription of these genes. The colors at the terminal nodes indicate the species to which each gene belongs.

coding-to-noncoding gene neofunctionalization. This idea is further supported by the concordance between the estimated age of the whole-genome duplication and the divergence time between *Cucurbita* and Benincaseae, as observed in our dated phylogeny. The divergence times estimated in this study mostly fall within the 95highest posterior density of previous phylogenetic studies, although our estimated ages within the *Cucurbita* genus are slightly older (Schaefer et al., 2009; Castellanos-Morales et al., 2018).

The whole-genome duplication within *Cucurbita* seems to be responsible for the observed acceleration in the rate of protein-coding gene family evolution, as almost half of the gene families experienced either expansions or contractions in the branch leading to the crown node of the *Cucurbita* genus, with a higher proportion of expansions than contractions. Furthermore, the terminal branches in the *Cucurbita* genus also showed larger proportions of gene family expansions and contractions compared with most of the other cucurbit taxa. Finally, the rate of gene family birth/death was two times higher in the *Cucurbita* clade compared with the rest of the phylogeny. All this evidence points toward a higher rate of gene family evolution in *Cucurbita* after the whole-genome duplication event that happened around 30 Mya (Montero-Pau et al., 2017). These patterns were observed despite performing statistical corrections for genome assembly and annotation errors during our gene family analyses, and were also observed after filtering low-quality gene models, suggesting that our results are robust to possible errors in gene model predictions. Furthermore, the number of gene models obtained after filtering low-quality gene models suggests that the real number of protein-coding genes in *Cucurbita* may be closer to 28 000 (Montero-Pau et al., 2017) than 32 000 genes (Sun et al., 2017).

The terminal branch of *J. regia* showed a high proportion of gene family expansion. These gene families have been previously shown to be involved in the biosynthesis of nonstructural polyphenols (Martínez-García et al., 2016). The genome of *J. regia* also shows evidence of a whole-genome duplication that happened around 60 Mya (Luo et al., 2015; Martínez-García et al., 2016), further suggesting that whole-genome duplications are correlated with higher rates of gene family evolution.

The results of GO enrichment analysis of the rapidly evolving families suggest that several biological functions changed in *C. argyrosperma* in relation to other *Cucurbita* species. For instance, we found a significant contraction in gene families associated with the recognition of pollen as well as pectinesterases, which are involved in pollen tube growth during pollination (Tian et al., 2006). Both kinesin motor protein and villin/gelsolin protein families, which expanded in *C. argyrosperma*, are also involved in pollen tube growth during pollination (Su et al., 2007; Li et al., 2012). Contraction and expansion of these families could be

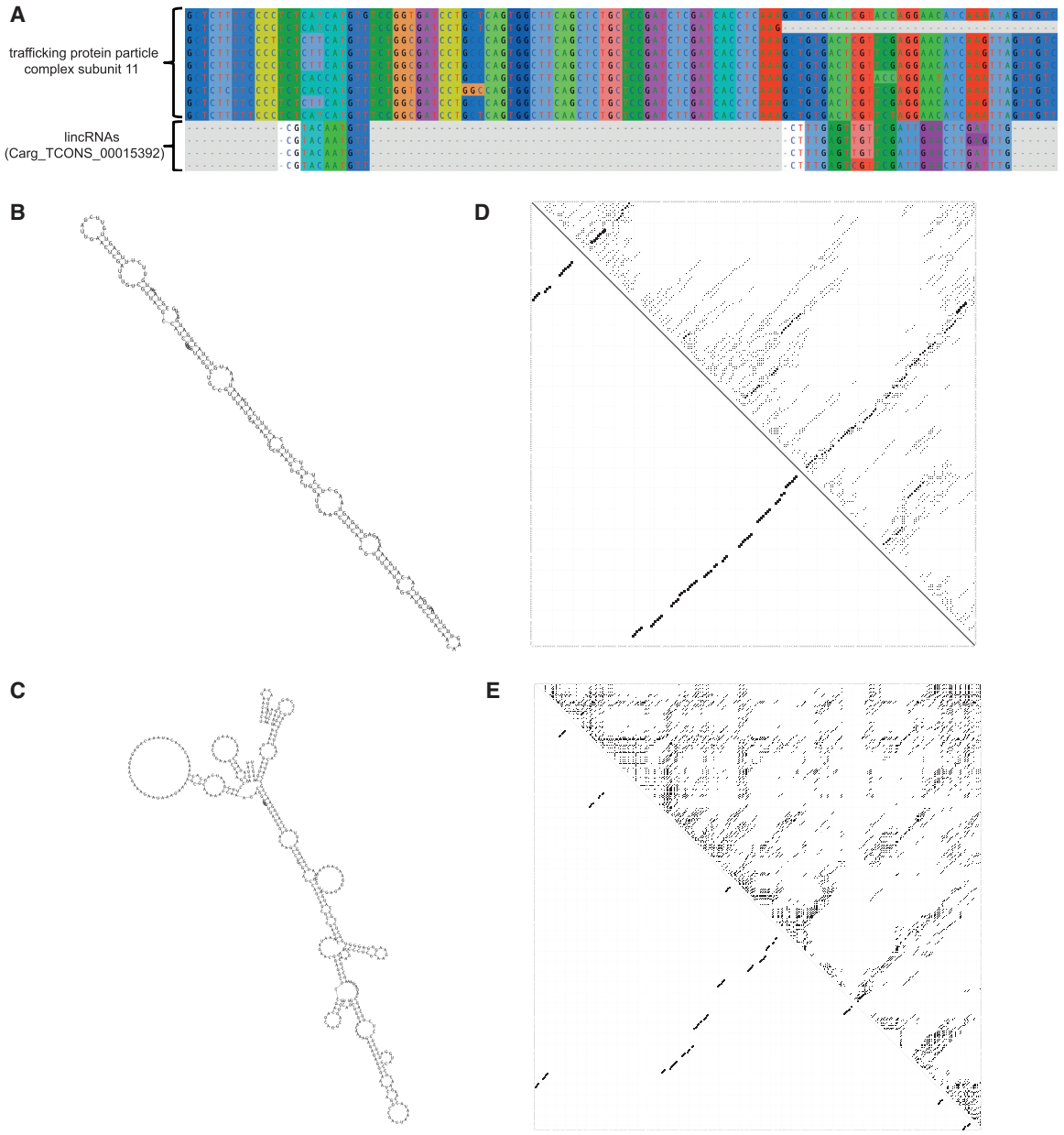


**Figure 5. Manual Assessment of Primary and Secondary Structure in Some lincRNAs Found within the Genome of *Cucurbita argyrosperma*.**

**(A)** Part of a multiple sequence alignment between a protein-coding gene-derived lincRNA (Carg_TCONS_00015392; down) and the homologous protein-coding gene transcripts (trafficking protein particle complex subunit 11; up). Each codon in the alignment is shown in a different color. Carg_TCONS_00015392 lincRNAs do not contain an open reading frame or show codon conservation, but orthologous lincRNA sequences are conserved between species.

**(B–C)** Minimum free energy (MFE) structural predictions of **(B)** the protein-coding-derived lincRNA Carg_TCONS_00015392 and **(C)** the highly conserved lincRNA Carg_TCONS_00063022.

**(D–E)** RNA base-pairing probability matrices showing the MFE structural prediction (below the diagonal) and all possible suboptimal pairings (above the diagonal) for **(D)** Carg_TCONS_00015392 and **(E)** Carg_TCONS_00063022. Higher probabilities are represented as larger dots within the matrices.

associated with changes in reproductive isolation, since reproductive barriers that prevent hybridization are more stringent in *C. argyrosperma* than in its sister species, *C. moschata* (Hurd et al., 1971). Pectinesterases are also involved in cell-wall modification during fruit ripening, and changes in the pectinesterase family could explain the differences in the smoothness of the fruit flesh between *C. argyrosperma* and *C. moschata*. The reduction in the number of pectinesterases within *C. argyrosperma* could have had an impact in its domestication, since fewer genes were available for artificial selection to act upon, possibly restricting its use to seed consumption, unlike the rest of the domesticated *Cucurbita* species whose ripened fruit flesh is commonly consumed (Lira et al., 2016). A future comparison between the genomes of *C.*

*argyrosperma* ssp. *argyrosperma* and *C*. *argyrosperma* ssp. *sororia* will reveal whether this reduction in gene family size happened before or after the domestication of this species.

Several contracted families were functionally enriched in exocytosis and transmembrane transport functions, which are usually involved in the release of secondary metabolites, hormones, and numerous other compounds (Hedrich and Marten, 2006). Interestingly, different multi-antimicrobial extrusion protein families either expanded or contracted within the genome of *C*. *argyrosperma*. This suggests an adaptive transition between different families of multi-antimicrobial extrusion proteins, perhaps related to changes in the geographic distribution of *C*. *argyrosperma* from its ancestors (Castellanos-Morales et al., 2018). Since multi-antimicrobial proteins are usually involved in the removal of cytotoxic compounds (Eckardt, 2001), the levels of these compounds could change alongside the geographic distribution of the species, acting as selective pressures in these gene families.

Some of the observed evolutionary dynamics within the lincRNA families in Cucurbitaceae can be explained under a high-turnover model of lincRNA evolution, such as the decline in lincRNA conservation as a function of phylogenetic distance and sudden duplication bursts. These dynamics were observed in both the *C*. *argyrosperma* and *C*. *lanatus* lincRNA families, suggesting that gene duplication is a common mechanism of lincRNA birth within the Cucurbitaceae family.

The average rate of lincRNA loss observed in the Cucurbitaceae family is similar to that observed in the Brassicaceae family, around 2.47% per million years (Nelson et al., 2016). However, considerable variation can be observed within each clade in both phylogenies. In the case of Cucurbitaceae, the loss rate ranged from 2.8% per million years between *Cucurbita* and Benincaseae to 2.1% per million years between *Cucurbita* and *M. charantia* and 3.4% per million years within Benincaseae. The differences in loss rate between Brassicaceae species are even more drastic, ranging from 4.3% per million years between *Arabidopsis* and *Capsella* to 2% per million years between *Arabidopsis* and *Brassica* (Nelson et al., 2016). Even though the total number of shared lincRNAs decreases with phylogenetic distance, the loss rate seems to decrease between distantly related taxa. In the case of Brassicaceae, the loss rate declined to 1.5% per million years between *Arabidopsis* and Cleomaceae, which diverged 64 Mya (Nelson et al., 2016). In the case of Cucurbitaceae, we found that the loss rate declined to 0.96% per million years between Cucurbitaceae and the outgroup species. This decline could be explained by a survivor bias, whereby the most biologically important genes are conserved throughout distantly related taxa, thus slowing the rate of lincRNA loss per million years. The rate of lincRNA loss within Tetrapoda seems to be slower than in plants, as 3% of the lincRNAs in humans are also present in chickens, which diverged 300 Mya (Necsulea et al., 2014). The loss rate between *Cucurbita* and Benincaseae was higher than the rate within *Cucurbita*, as expected by the effect of the whole-genome duplication (Nelson and Shippen, 2015). However, the high loss rate within Benincaseae was unexpected, since it was higher than that observed between *Cucurbita* and Benincaseae. We propose that the acceleration in the rate of lincRNA turnover within Benincaseae was caused by the multiple changes in karyotype number throughout Benincaseae (Huang et al., 2009; Levi et al., 2011; Garcia-Mas et al., 2012; Wu et al., 2017), which are considered genomic disturbances that can accelerate this rate (Nelson and Shippen, 2015). This hypothesis is supported by the larger proportion of conserved lincRNAs between *C*. *argyrosperma* and *M*. *charantia* than between *C*. *argyrosperma* and *Cucumis*, which can be explained by additional genomic disturbances within the Benincaseae family.

The decline in lincRNA conservation throughout the Cucurbitaceae phylogeny could be explained by high levels of gene birth and death, making the search for homology between distantly related species futile, as the vast majority of these genes either arose before the divergence of both taxa or became extinct in one of the lineages (Ulitsky, 2016). It is also possible that lincRNAs have high rates of nucleotide substitution due to positive selection (Smith and Mattick, 2017), which hinders the search for homologous sequences as species become more distantly related. Given that lncRNAs are involved in epigenetic regulation through several mechanisms (Mercer et al., 2009), the possibility of positive selection acting on such dynamic genes may be an important factor in adaptive radiation (Smith and Mattick, 2017).

The observation of highly conserved lincRNAs, as well as the evidence of their transcription suggests they have an important biological function within the Cucurbitaceae family. However, pinpointing the specific biological functions of lincRNAs is still difficult without experimental data. The high rate of turnover in lincRNA families limits the inference of biological functions from distantly related plants such as *Arabidopsis thaliana*, in which experimental validation of gene functions is more common (Ulitsky, 2016). Future experimental studies should focus on the functional characterization of these lincRNAs.

The proportion of Evolinc-II lincRNA families with evidence of protein-coding neofunctionalization was higher than initially expected based on our direct comparison between lincRNAs and protein-coding genes. Such events are not exclusive of the *Cucurbita* crown node, as they are rather common throughout the Cucurbitaceae family, although they are particularly frequent within the *Cucurbita* clade. This suggests that the neofunctionalization of protein-coding genes into novel lincRNAs is more common than initially suspected (Kapusta et al., 2013), and may be a recurrent source of noncoding genes in the Cucurbitaceae family (Chen and Rajewsky, 2007) alongside other sources of lincRNA genes, such as neofunctionalization from TEs (Kapusta et al., 2013).

The proportion of lincRNA families with evidence of protein-coding neofunctionalization predicted from comparisons with the *C*. *argyrosperma* lincRNAs was higher compared with those predicted from comparisons with the *C*. *lanatus* lincRNAs. Furthermore, the proportion of protein-coding gene-derived lincRNAs calculated from the direct comparison between coding transcripts and lincRNAs was significantly higher in *Cucurbita* compared with the proportion in other cucurbit species, whereas the proportion of TE-derived lincRNAs was lower in *Cucurbita* with respect to the rest of the Cucurbitaceae family. These results suggest that the whole-genome duplication in *Cucurbita* acted as a genomic disturbance that altered lincRNA family birth dynamics,

with more lincRNAs being derived from protein-coding genes than from TEs (Kapusta et al., 2013; Nelson and Shippen, 2015). This is consistent with the apparent conservation of TE proportions between *Cucurbita* species, suggesting that TEs did not play an important role in lincRNA family evolution within *Cucurbita* after the whole-genome duplication, unlike in other taxa such as vertebrate species, where TEs play an important role in lincRNA birth (Kapusta et al., 2013). Our results also differ from those obtained in cotton, where a larger proportion of lincRNAs were homologous to TEs than to protein-coding loci (Zhao et al., 2018). This suggests that the evolutionary dynamics of lincRNAs can change drastically between different plant taxa. After the whole-genome duplication within *Cucurbita*, many duplicated protein-coding genes that were functionally redundant were co-opted into novel lincRNAs (Ponting et al., 2009; Kapusta et al., 2013).

The observed differences in length between Carg_TCONS_00015392 and its protein-coding homologs, as well as the disruption of the ORF and the lack of codon structure, suggest this lincRNA is not a cryptic protein-coding transcript but a true noncoding element in *Cucurbita*. Furthermore, the level of sequence conservation between species, as well as the thermodynamic stability of its predicted secondary structure, suggests that this lincRNA may be functional (Smith and Mattick, 2017). Even though secondary structure alone is insufficient evidence to support the hypothesis that a noncoding RNA is functional, especially considering that some lincRNAs have more than one functional structure (Smith and Mattick, 2017), secondary structures in functional noncoding RNAs are expected to be more stable than those in other sequences with similar compositions (Clote et al., 2005). This can be observed when comparing the stability between Carg_TCONS_00015392 and its protein-coding homolog, both of which have similar dinucleotide frequencies but different structural stabilities. This structural stability is also present in the highly conserved lincRNA Carg_TCONS_00063022. The stability of the predicted secondary structures in both lincRNAs suggests that they are functional (Clote et al., 2005), unlike the secondary structure in the transcript of the protein-coding gene homologous to Carg_TCONS_00015392, which appears to be random. Whether all predicted lincRNAs behave similarly or whether they are indeed functional remains to be experimentally validated.

We propose that the whole-genome duplication within the *Cucurbita* genus allowed for faster rates of gene family evolution, since functional redundancy within the genome facilitated the co-option of complex genetic elements, such as previously existing genes, into new functions. During the fractionation process after the whole-genome duplication (Sun et al., 2017), a substantial fraction of the duplicated protein-coding genes with redundant functions either diverged (acquiring novel functions as protein-coding genes, thereby increasing the rate of gene family birth/death), or neofunctionalized into noncoding elements, such as lincRNAs.

## METHODS

### Biological Samples and DNA/RNA Extraction

We obtained seeds from a cultivated individual of *C. argyrosperma* ssp. *argyrosperma* collected in the region of Tepec, Jalisco (see Supplemental Data 1 for detailed methods and data on fruit selection). One of the seeds was germinated in a greenhouse, and plants were grown to maturity, when flower buds started to develop. We selected one of the germinated plants and extracted total DNA from fresh leaves (Doyle and Doyle, 1987) for genome sequencing.

For transcriptome sequencing, we extracted total RNA from leaves, stems, roots, male flower buds, and tendrils using the RNeasy Plant Mini Kit (Qiagen) according to the manufacturer's protocol. Each RNA sample was precipitated in salty ethanol (260 mM lithium chloride and 66% EtOH).

The plant used for whole-genome sequencing and transcriptome sequencing was deposited in the National Herbarium of Mexico (MEXU) under accession number SMH-JMG-627. The details of DNA and RNA sequencing are available in Supplemental Methods.

### Genome and Transcriptome Assembly

The chloroplast genome of *C. argyrosperma* was assembled with NOVOPlasty (Dierckxsens et al., 2016), and the mitochondrion genome was assembled using the Organelle-PBA pipeline (Soorni et al., 2017). Both organelles were scaffolded using SSPACE-longread (Boetzer and Pirovano, 2014). Gap-filling and base corrections were performed with Pilon (Walker et al., 2014). The nuclear genome was assembled with a hybrid approach, using Platanus (Kajitani et al., 2014) and DBG2OLC (Ye et al., 2016). We used Minimap and Racon (Vaser et al., 2017) to obtain a consensus sequence assembly, then base corrections were made using Pilon. Scaffolding was done using BESST (Sahlin et al., 2014) and SSPACE-longread. Gap closing was performed with GapFiller (Boetzer and Pirovano, 2012) and a final base correction was done with Pilon. See Supplemental Methods for a detailed description of the organelle and nuclear genome assemblies and a description of the transcriptome assembly.

The Illumina and PacBio sequence reads were mapped against the genome using BWA *mem* (Li, 2013) and BlasR (Chaisson and Tesler, 2012), respectively, to assess the completeness of the genome assembly. We mapped the transcriptome reads against the nuclear and organelle genomes using Hisat2 (Kim et al., 2015) to assess assembly completeness. The percentage of reads that mapped to the assembly was calculated using *flagstat* within SAMtools (Li et al., 2009).

### Prediction of Transposable Elements and Protein-Coding Gene Models

We used the REPET package (Flutre et al., 2011) to predict *de novo* the TEs within the *C. argyrosperma* genome assembly, generating a library of non-redundant consensus sequences. These consensus sequences were classified according to Wicker's classification system (Wicker et al., 2007) using PASTEC (Hoede et al., 2014) within the REPET pipeline (repeat library available in Supplemental Data 2). The repeat library was used to annotate and mask the TEs within the genome assembly with RepeatMasker (Smit et al., 2013).

MAKER3 (Cantarel et al., 2008) was used to predict protein-coding gene models in the *C. argyrosperma* genome assembly. We incorporated AUGUSTUS (Stanke et al., 2006) GeneMark-ES (Lomsadze et al., 2005) and SNAP (Korf, 2004) as *ab initio* gene predictors within MAKER3. We also used EvidenceModeler (Haas et al., 2008) to obtain additional gene models within MAKER3. We incorporated tRNAscan-SE (Lowe and Eddy, 1997) within the MAKER3 pipeline to predict tRNA genes. See Supplemental Methods for a detailed description of the prediction of protein-coding gene models. The protein-coding genes predicted within *C. argyrosperma* were functionally annotated with InterProScan (Jones et al., 2014). The annotation table is available in Supplemental Data 4.

### Phylogenetic and Protein-Coding Gene Family Analyses

The details of the phylogenetic analyses can be found within Supplemental Methods. We performed an all-VS-all BLASTp (Camacho et al., 2009) analysis to identify protein-coding gene families with MCL (Enright et al., 2002), using an inflation parameter of 3. All gene families were aligned with MAFFT (Katoh et al., 2002).

We generated the species phylogeny with PhyML (Guindon et al., 2010), using SMS (Lefort et al., 2017) to determine the best amino acid substitution model for our sequence alignment. To obtain a dated phylogeny, we used a Bayesian Markov chain Monte Carlo approach with approximate likelihood calculation, as implemented in mcmctree (Yang, 2007). The mcmctree trace files are available in Supplemental Data 3.

We assessed changes in protein-coding gene family sizes across the dated phylogeny using CAFE v4.0.2 (De Bie et al., 2006). We initially estimated the gene birth–death parameter λ to assess gene family evolution using a subset of gene clusters that had <100 differences in gene content between any pair of species and used it to calculate significant changes ($p$ value <0.01) in gene family size at every branch in the dated phylogeny for every gene family.

We compared three different λ schemes: (a) a change in λ within Cucurbitaceae, (b) a change in λ within *Cucurbita*, and (c) two changes in λ, one within Cucurbitaceae and another within *Cucurbita*. After finding the best scheme of λ parameters within the phylogeny, we estimated error models for genome assembly and annotation errors (Han et al., 2013), and used those models to analyze significant gene family expansions and contractions throughout the tree. We performed the clustering, molecular clock, and gene family analyses using the same methodology as described above using a subset of high-quality protein-coding gene models obtained after filtering MAKER predictions with eAED values lower than 0.5 (input files, CAFE scripts, and final outputs are available in Supplemental Data 5). We performed GO enrichment analyses with topGO using the *weight01* method (Alexa et al., 2006). Significantly enriched terms were assessed after performing Fisher's exact tests and performing false discovery rate (FDR) adjustments of the $p$ values (Benjamini and Hochberg, 1995).

### Prediction and Analysis of Long Noncoding RNAs

We used Evolinc-I (Nelson et al., 2017) to predict lincRNAs within the genome-guided transcriptome assembly of *C. argyrosperma*. In brief, Evolinc-I predicted as lincRNAs all transcripts longer than 200 bp that did not overlap with any of the predicted protein-coding genes within the genome and did not contain an ORF longer than 100 amino acids (Nelson et al., 2017). lincRNAs were also predicted for the genomes of *C. maxima*, *C. moschata*, *C. pepo*, *C. melo*, *C. sativus*, *C. lanatus*, and *L. siceraria* using the same methodology as described above, using RNA-seq data available in the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra). SRA accessions used for each species are available in Supplemental Table 3. The lincRNAs of *M. charantia* were extracted from the gff3 file available under the NCBI RefSeq genome accession PRJNA433137 (Urasaki et al., 2017). We used BLASTn (Camacho et al., 2009) with a cutoff of 50% coverage and 30% identity to define similarity due to sequence homology between lincRNAs, protein-coding transcripts, and TEs. We performed Student's $t$-tests with the *stats* package in R (R Core Team, 2016) to assess statistical differences in the proportion of protein-coding gene-derived lincRNAs and TE-derived lincRNAs between *Cucurbita* and the other cucurbit species.

We used Evolinc-II (Nelson et al., 2017) to assess the evolution of lincRNA families across the Cucurbitaceae family. Homologous sequences for each query lincRNA were filtered using an e-value of $<1e^{-20}$ after a reiterative reciprocal BLAST search against each genome. We used MAFFT (Katoh et al., 2002) to align each lincRNA family and RAxML (Stamatakis, 2014) to generate gene family trees. Each gene family tree was compared against the species tree to detect duplication or loss of lincRNAs across the phylogeny using Notung (Chen et al., 2000). Since there is a possibility that the predicted lincRNAs are protein-coding genes that could not be predicted with MAKER3, we wanted to exclude as many false positives as possible to obtain a conservative estimation of the number of lincRNA families that evolved from protein-coding genes. Thus, we eliminated 648 possible spurious lincRNAs (that is, protein-coding genes that were mistakenly predicted as lincRNAs) where the only noncoding elements within the gene family belonged to the query species, and the other elements were protein-coding genes predicted from the other species. Likewise, we defined lincRNA families with evidence of protein-coding gene neofunctionalization as those with noncoding elements from two or more different species within the phylogeny, as the rate of parallel misidentification of lincRNA genes in several species should be low.

Finally, we defined putatively noncoding families as those that lacked any protein-coding gene within the phylogeny, that is, they were composed exclusively of lincRNA genes. All lincRNA family alignments and phylogenies are available in Supplemental Data 6 and 7. Rates of lincRNA loss were calculated as the percentage of lincRNAs in the query species without a homolog in another species divided by the mean divergence time between the two taxa. The secondary structure predictions and the base-pairing probability matrix of the lincRNAs were calculated with RNAfold from the ViennaRNA package (Hofacker et al., 1994; Lorenz et al., 2011).

## ACCESSION NUMBERS

The raw sequence reads of all the genomic and transcriptomic data are available in the NCBI SRA database (accession SRP157098). The nuclear and organelle genome assemblies; as well as the protein-coding gene, tRNA, and lincRNA predictions of *C. argyrosperma* are available in the CoGe database (IDs 53608, 52005, 52006) and in the Figshare database (https://doi.org/10.6084/m9.figshare.7728608.v1). The predicted lincRNAs of the other species are available in the CoGe database (IDs 52078–52084). The dated species tree and the single-copy ortholog alignment used for the phylogenetic analyses are available in TreeBase (submission ID S23151).

## SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

## FUNDING

This study was funded by Comisión Natural para el Conocimiento y Uso de la Biodiversidad (CONABIO) KE004 "Diversidad genética de las especies de *Cucurbita* en México e hibridación entre plantas genéticamente modificadas y especies silvestres de *Cucurbita*," CONABIO PE001 "Diversidad genética de las especies de *Cucurbita* en México. Fase II. Genómica evolutiva y de poblaciones, recursos genéticos y domesticación" (both awarded to R.L.-S. and L.E.E.), Consejo Nacional de Ciencia y Tecnología (CONACyT) Investigación Científica Básica 2011.167826 "Genómica de poblaciones: estudios en el maíz silvestre, el teosinte (*Zea mays* ssp. *parviglumis* y *Zea mays* ssp. *mexicana*)" (awarded to L.E.E.), and CONACyT Problemas Nacionales through grant number 247730 (awarded to D.P.). J.B.-R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 583146 from CONACyT. The sabbatical leave of L.E.E. at Dr. Peter Tiffin's laboratory, Department of Plant and Microbial Biology, University of Minnesota, was supported by the program PASPA-DGAPA, UNAM. The Evolinc-II analyses were carried out using CONABIO's computing cluster, which was partially funded by Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) through the grant "Contribución de la Biodiversidad para el Cambio Climático" to CONABIO.

## AUTHOR CONTRIBUTIONS

J.B.-R., E.I.-L., A.V.-L., S.M.-H., R.L.-S., and L.E.E. jointly conceived and designed the study. G.S.d.l.V. and S.M.-H. collected the biological samples and carried out morphological analyses to select the parental individual of *C. argyrosperma* used for whole-genome sequencing and assembly. J.B.-R., G.S.d.l.V., and A.V.-L. carried out the molecular laboratory work. J.B.-R. performed the genome assembly and gene predictions, and carried out most of the analyses. E.I.-L. trained the AUGUSTUS model for gene prediction and ran REPET for TE prediction. Y.T.G.-G. did the functional annotation of the protein-coding gene families. E.I.-L. and Y.T.G.-G. helped in several bioinformatics analyses. J.B.-R. drafted the manuscript. D.P., R.L.-S., and L.E.E. obtained the funding for the study. All authors revised the final version of the manuscript.

## ACKNOWLEDGMENTS

This manuscript is presented in partial fulfillment of the requirements to obtain a PhD degree by J.B.-R. in the Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). We acknowledge the Doctorado en Ciencias Biomédicas for the support provided during the development of this project. Special thanks to Valeria Souza for supporting this research, and the technical support of Laura Espinosa-Asuar and Erika Aguirre-Planter. We thank Xitlali Aguirre-Dugua for depositing the plant accession in the herbarium, and all the Laboratorio de Evolución Molecular y Experimental at Instituto de Ecología, UNAM. We also acknowledge the support of the Facultad de Estudios Superiores Iztacala, UNAM, and of the Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Campo Experimental Bajío. The authors thank Rodrigo García Herrera, head of the Scientific Computing Department at LANCIS, Instituto de Ecología, UNAM, for running the HTC infrastructure we used for part of the analyses. The authors acknowledge the technical support of MSc Emanuel Villafán and the resources for high-performance computing that the Institute of Ecology (INECOL) made available for conducting the genome assembly and annotation reported in this paper. L.E.E. acknowledges the support of the program PASPA-DGAPA, UNAM to conduct his sabbatical year. No conflict of interest declared.

Received: August 27, 2018
Revised: December 12, 2018
Accepted: December 28, 2018
Published: January 7, 2019

## REFERENCES

**Alexa, A., Rahnenführer, J., and Lengauer, T.** (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics **22**:1600–1607.

**Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K., and Palmer, J.D.** (2010). Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol. Biol. Evol. **27**:1436–1448.

**Ansimova, M., and Gascuel, O.** (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. **55**:539–552.

**Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.** (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**:25–29.

**Bao, W., Kojima, K.K., and Kohany, O.** (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA **6**:11.

**Benjamini, Y., and Hochberg, Y.** (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**:289–300.

**Boetzer, M., and Pirovano, W.** (2012). Toward almost closed genomes with GapFiller. Genome Biol. **13**:R56.

**Boetzer, M., and Pirovano, W.** (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics **15**:211.

**Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L.** (2009). BLAST plus: architecture and applications. BMC Bioinformatics **10**:421.

**Campbell, M.S., Holt, C., Moore, B., and Yandell, M.** (2014). Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinform. **48**:4.11.1–4.11.39.

**Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., and Yandell, M.** (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. **18**:188–196.

**Castellanos-Morales, G., Paredes-Torres, L.M., Gámez, N., Hernández-Rosales, H.S., Sánchez-de la Vega, G., Barrera-Redondo, J., Aguirre-Planter, E., Vázquez-Lobo, A., Montes-Hernández, S., Lira-Saade, R., et al.** (2018). Historical biogeography and phylogeny of *Cucurbita*: insights from ancestral area reconstruction and niche evolution. Mol. Phylogenet. Evol. **128**:38–54.

**Chaisson, M.J., and Tesler, G.** (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics **13**:238.

**Chekanova, J.A.** (2015). Long non-coding RNAs and their functions in plants. Curr. Opin. Plant Biol. **27**:207–216.

**Chen, K., and Rajewsky, N.** (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. **8**:93–103.

**Chen, K., Durand, D., and Farach-Colton, M.** (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. J. Comput. Biol. **7**:429–447.

**Chikhi, R., and Medvedev, P.** (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics **30**:31–37.

**Clote, P., Ferré, F., Kranakis, E., and Krizanc, D.** (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. RNA **11**:578–591.

**Daniell, H., Lin, C.-S., Yu, M., and Chang, W.-J.** (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. **17**:134.

**De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W.** (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics **22**:1269–1271.

**Dierckxsens, N., Mardulyn, P., and Smits, G.** (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. **45**:gkw955.

**Doyle, J.J., and Doyle, J.L.** (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.

**Edger, P.P., VanBuren, R., Colle, M., Poorten, T.J., Wai, C.M., Niederhuth, C.E., Alger, E.I., Ou, S., Acharya, C.B., Wang, J., et al.** (2018). Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. Gigascience **7**:1–7.

**Eckardt, N.A.** (2001). Move it on out with MATEs. Plant Cell **13**:1477–1480.

**Enright, A.J., Van Dongen, S., and Ouzounis, C.A.** (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30**:1575–1584.

**Fatica, A., and Bozzoni, I.** (2014). Long non-coding RNAs: new players in cell differentiation and development. Nat. Rev. Genet. **15**:7–21.

**Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H.** (2011). Considering transposable element diversification in *de novo* annotation approaches. PLoS One **6**:e16526.

**Ganfornina, M.D., and Sánchez, D.** (1999). Generation of evolutionary novelty by functional shift. BioEssays **21**:432–439.

**Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., Gonzalez, V.M., Henaff, E., Camara, F., Cozzuto, L., Lowy, E., et al.** (2012). The genome of melon (*Cucumis melo* L.). Proc. Natl. Acad. Sci. U S A **109**:11872–11877.

**Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.** (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. **29**:644–652.

**Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O.** (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. **59**:307–321.

**Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W.J., Zhang, H., Zheng, Y., Mao, L., Ren, Y., Wang, Z., et al.** (2012). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat. Genet. **45**:51–58.

**Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R.** (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. **9**:R7.

**Han, M.V., Thomas, G.W.C., Lugo-Martinez, J., and Hahn, M.W.** (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol. Biol. Evol. **30**:1987–1997.

**Hedrich, R., and Marten, I.** (2006). 30-year progress of membrane transport in plants. Planta **224**:725–739.

**Hoede, C., Arnoux, S., Moisset, M., Chaumier, T., Inizan, O., Jamilloux, V., and Quesneville, H.** (2014). PASTEC: an automatic transposable element classification tool. PLoS One **9**:1–6.

**Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P.** (1994). Fast folding and comparison of RNA secondary structures. Monatsh. F. Chem. **125**:167–188.

**Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., Lucas, W.J., Wang, X., Xie, B., Ni, P., et al.** (2009). The genome of the cucumber, *Cucumis sativus* L. Nat. Genet. **41**:1275–1281.

**Hurd, P.D., Jr., Linsley, E.G., and Whitaker, T.W.** (1971). Squash and gourd bees (*Peponapis*, *Xenoglossa*) and the origin of the cultivated *Cucurbita*. Evolution **25**:218–234.

**Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al.** (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

**Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al.** (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. **24**:1384–1395.

**Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C.** (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. **9**:e1003470.

**Katoh, K., Misawa, K., Kuma, K., and Miyata, T.** (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**:3059–3066.

**Kim, D., Langmead, B., and Salzberg, S.L.** (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

**Korf, I.** (2004). Gene finding in novel genomes. BMC Bioinformatics **5**:59.

**Lefort, V., Longueville, J.-E., and Gascuel, O.** (2017). SMS: smart model selection in PhyML. Mol. Biol. Evol. **34**:2422–2424.

Levi, A., Hernandez, A., Thimmapuram, J., Donthu, R., Wright, C., Ali, C., Wechter, W.P., Reddy, U., and Mikel, M. (2011). Sequencing the genome of the heirloom watermelon cultivar charleston gray. Plant and Animal Genome Conference. P047.

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup** (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics **25**:2078–2079.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv 1303.3997.

**Li, J., Xu, Y., and Chong, K.** (2012). The novel functions of kinesin motor proteins in plants. Protoplasma **249**:S95–S100.

**Lira, R., Eguiarte, L., Montes, S., Zizumbo-Villarreal, D., Colunga-GarcíaMarín, P., and Quesada, M.** (2016). *Homo sapiens-Cucurbita* interaction in Mesoamerica: domestication, dissemination and diversification. In Ethnobotany of Mexico, R. Lira, A. Casas, and J. Blancas, eds. (New York: Springer-Verlag), pp. 389–402.

**Liu, G., Mattick, J.S., and Taft, R.J.** (2013). A meta-analysis of the genomic and transcriptomic composition of complex life. Cell Cycle **12**:2061–2072.

**Liu, X., Hao, L., Li, D., Zhu, L., and Hu, S.** (2015). Long non-coding RNAs and their biological roles in plants. Genomics Proteomics Bioinformatics **13**:137–147.

**Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M.** (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. **33**:6494–6506.

**Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L.** (2011). ViennaRNA package 2.0. Algorithms Mol. Biol. **6**:26.

**Lowe, T.M., and Eddy, S.R.** (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

**Luo, M.-C., You, F.M., Li, P., Wang, J.-R., Zhu, T., Dandekar, A.M., Leslie, C.A., Aradhya, M., McGuire, P.E., and Dvorak, J.** (2015). Synteny analysis in Rosids with a walnut physical map reveals slow genome evolution in long-lived woody perennials. BMC Genomics **16**:707.

**Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R.** (2013). Gene duplication as a major force in evolution. J. Genet. **92**:155–161.

**Martínez-García, P.J., Crepeau, M.W., Puiu, D., Gonzalez-Ibeas, D., Whalen, J., Stevens, K.A., Paul, R., Butterfield, T.S., Britton, M.T., Reagan, R.L., et al.** (2016). The walnut (*Juglans regia*) genome sequence reveals diversity in genes coding for the biosynthesis of non-structural polyphenols. Plant J. **87**:507–532.

**Mercer, T.R., Dinger, M.E., and Mattick, J.S.** (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. **10**:155–159.

**Montero-Pau, J., Blanca, J., Bombarely, A., Ziarsolo, P., Esteras, C., Martí-Gómez, C., Ferriol, M., Gómez, P., Jamilena, M., Mueller, L., et al.** (2017). *De novo* assembly of the zucchini genome reveals a whole genome duplication associated with the origin of the *Cucurbita* genus. Plant Biotechnol. J. **12**:3218–3221.

**Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H.** (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature **505**:635–640.

**Nee, M.** (1990). The domestication of *Cucurbita* (Cucurbitaceae). Econ. Bot. **44**:56–68.

**Nelson, A.D.L., and Shippen, D.E.** (2015). Evolution of TERT-interacting lncRNAs: expanding the regulatory landscape of telomerase. Front. Genet. **6**:1–6.

**Nelson, A.D.L., Devisetty, U.K., Palos, K., Haug-Baltzell, A.K., Lyons, E., and Beilstein, M.A.** (2017). Evolinc: a tool for the identification and evolutionary comparison of long intergenic non-coding RNAs. Front. Genet. **8**:1–12.

**Nelson, A.D.L., Forsythe, E.S., Devisetty, U.K., Clausen, D.S., Haug-Batzell, A.K., Meldrum, A.M., Frank, M.R., Lyons, E., and Beilstein, M.A.** (2016). A genomic analysis of factors driving lincRNA diversification: lessons from plants. G3 (Bethesda) **6**:2881–2891.

**Paris, H.S.** (2016). Genetic resources of pumpkins and squash, *Cucurbita* spp. In Genetics and Genomics of Cucurbitaceae, R. Grumet, N. Katzir, and J. Garcia-Mas, eds. (Cham, Switzerland: Springer), pp. 111–154.

**Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L.** (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. **33**:290–295.

**Ponting, C.P., Oliver, P.L., and Reik, W.** (2009). Evolution and functions of long noncoding RNAs. Cell **136**:629–641.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.

**Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., and Arvestad, L.** (2014). BESST—efficient scaffolding of large fragmented assemblies. BMC Bioinformatics **15**:281.

**Schaefer, H., Heibl, C., and Renner, S.S.** (2009). Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. Proc. R. Soc. B Biol. Sci. **276**:843–851.

**Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M.** (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker. *Open4.0*. http://www.repeatmasker.org.

**Smith, M.A., and Mattick, J.S.** (2017). Structural and functional annotation of long noncoding RNAs. In Bioinformatics. Methods in Molecular Biology, J.M. Keith, ed. (New York: Humana Press), pp. 65–85.

**Soorni, A., Haak, D., Zaitlin, D., and Bombarely, A.** (2017). Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics **18**:49.

**Stamatakis, A.** (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313.

**Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S.** (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics **7**:62.

**Su, H., Wang, T., Dong, H., and Ren, H.** (2007). The villin/gelsolin/fragmin superfamily proteins in plants. J. Integr. Plant Biol. **49**:1183–1191.

**Sun, H., Wu, S., Zhang, G., Jiao, C., Guo, S., Ren, Y., Zhang, J., Zhang, H., Gong, G., Jia, Z., et al.** (2017). Karyotype stability and unbiased fractionation in the Paleo-Allotetraploid *Cucurbita* genomes. Mol. Plant **10**:1293–1306.

**The Gene Ontology Consorsium.** (2017). Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. **45**:D331–D338.

**Tian, G.W., Chen, M.H., Zaltsman, A., and Citovsky, V.** (2006). Pollen-specific pectin methylesterase involved in pollen tube growth. Dev. Biol. **294**:83–91.

**Ulitsky, I.** (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nat. Rev. Genet. **17**:601–614.

**Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., Fukushima, M., Suzuki, S., Tarora, K., Tamaki, M., et al.** (2017). Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. DNA Res. **24**:51–58.

**Vaser, R., Sovic, I., Nagarajan, N., and Sikic, M.** (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. **27**:737–746.

**Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al.** (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

**Wang, J., Chu, S., Zhu, Y., Cheng, H., and Yu, D.** (2015). Positive selection drives neofunctionalization of the UbiA prenyltransferase gene family. Plant Mol. Biol. **87**:383–394.

**Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al.** (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**:973–982.

**Wu, S., Shamimuzzaman, M., Sun, H., Salse, J., Sui, X., Wilder, A., Wu, Z., Levi, A., Xu, Y., Ling, K.-S., et al.** (2017). The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. Plant J. **92**:963–975.

**Yandell, M., and Ence, D.** (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. **13**:329–342.

**Yang, Z.** (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**:1586–1591.

**Ye, C., Hill, C.M., Wu, S., Ruan, J., and Ma, Z.S.** (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. Sci. Rep. **6**:31900.

**Zhao, T., Tao, X., Feng, S., Wang, L., Hong, H., Ma, W., Shang, G., Guo, S., He, Y., Zhou, B., et al.** (2018). LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. Genome Biol. **19**:195.

**Zheng, Y.H., Alverson, A.J., Wang, Q.F., and Palmer, J.D.** (2013). Chloroplast phylogeny of *Cucurbita*: evolution of the domesticated and wild species. J. Syst. Evol. **51**:326–334.