nature plants

ARTICLES

https://doi.org/10.1038/s41477-019-0458-0

Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants

Zhenzhen Yang[1,2,7], Eric K. Wafula[2], Gunjune Kim[3,8], Saima Shahid[1,2,9], Joel R. McNeal[4], Paula E. Ralph[2], Prakash R. Timilsena[2], Wen-bin Yu[2,10], Elizabeth A. Kelly[1,2], Huiting Zhang[1,2], Thomas Nate Person[5], Naomi S. Altman[6], Michael J. Axtell[1,2], James H. Westwood[3,11]* and Claude W. dePamphilis[1,2,5]*

Horizontal gene transfer (HGT), the movement and genomic integration of DNA across species boundaries, is commonly associated with bacteria and other microorganisms, but functional HGT (fHGT) is increasingly being recognized in heterotrophic parasitic plants that obtain their nutrients and water from their host plants through direct haustorial feeding. Here, in the holoparasitic stem parasite *Cuscuta*, we identify 108 transcribed and probably functional HGT events in *Cuscuta campestris* and related species, plus 42 additional regions with host-derived transposon, pseudogene and non-coding sequences. Surprisingly, 18 *Cuscuta* fHGTs were acquired from the same gene families by independent HGT events in Orobanchaceae parasites, and the majority are highly expressed in the haustorial feeding structures in both lineages. Convergent retention and expression of HGT sequences suggests an adaptive role for specific additional genes in parasite biology. Between 16 and 20 of the transcribed HGT events are inferred as ancestral in *Cuscuta* based on transcriptome sequences from species across the phylogenetic range of the genus, implicating fHGT in the successful radiation of *Cuscuta* parasites. Genome sequencing of *C. campestris* supports transfer of genomic DNA—rather than retroprocessed RNA—as the mechanism of fHGT. Many of the *C. campestris* genes horizontally acquired are also frequent sources of 24-nucleotide small RNAs that are typically associated with RNA-directed DNA methylation. One HGT encoding a leucine-rich repeat protein kinase overlaps with a microRNA that has been shown to regulate host gene expression, suggesting that HGT-derived parasite small RNAs may function in the parasite–host interaction. This study enriches our understanding of HGT by describing a parasite–host system with unprecedented gene exchange that points to convergent evolution of HGT events and the functional importance of horizontally transferred coding and non-coding sequences.

Horizontal gene transfer (HGT), defined as the movement and genomic integration of DNA across species boundaries, is a common mechanism used by bacteria to acquire novel traits such as antibiotic resistance[1], but is far less common in plants and other eukaryotes. Although instances of HGT have been documented among autotrophic plants[2,3], the process most frequently involves parasitic plants[2,4–10] probably due to their intimate feeding connections with their host plants. However, the extent of HGT, the mechanisms of transfer and functional implications of the acquired sequences remain unclear. HGT events in parasitic plants[2,4–10] have been linked to increasing heterotrophy[2,4,7], and these horizontally acquired genes may function in parasitic processes[4,7]. Dodders (genus *Cuscuta*, Convolvulaceae) are obligate, fully heterotrophic parasites that obtain all carbon and other nutrients from their host plants. In addition, they are twining parasites that attach to host stems and are recognized for having open phloem connections[11] with their hosts via the feeding structures known as haustoria, making them ideal systems in which to look for high levels of HGT events[12,13]. Studies of *Cuscuta*–host interactions indicate reciprocal and massive messenger RNA flow (up to 1% of total tissue transcripts) between *Cuscuta* and its host[14]. Furthermore, microRNAs are also transferred from *Cuscuta campestris* to its host and downregulate host mRNAs involved in defence and phloem functions[15]. We hypothesized that such broad RNA exchange could lead to HGT after reverse transcription of exchanged mRNA, followed by genomic integration of the retroprocessed sequences. One signature of a reverse transcription mechanism of HGT would be the loss of introns in HGT genes compared to donor sequences, and so to infer the presence of introns in HGT sequences we generated a nuclear genome and assembly for *C. campestris*[15]. This approach (Supplementary Fig. 1) has the additional advantage of enabling the identification (Supplementary Fig. 1b) of non-functional pseudogenes[16], which could result from reverse transcription-mediated HGT and may not be captured by transcriptome data only. A recent initial survey of HGT in *C. campestris*[12] reported putative HGT events involving 72 genic regions. We note important methodological differences between the approach used therein and our methods, which we believe yield more true HGT events (for example,

[1]Intercollege Graduate Program in Plant Biology, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA. [2]Department of Biology, The Pennsylvania State University, University Park, PA, USA. [3]Department of Plant Pathology, Physiology and Weed Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [4]Department of Ecology, Evolution, and Organismal Biology, Kennesaw State University, Kennesaw, GA, USA. [5]Intercollege Graduate Program in Bioinformatics and Genomics, Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA. [6]Department of Statistics and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA. [7]Present address: Shanghai Institute for Advanced Immunochemical Studies, ShanghaiTech University, Shanghai, China. [8]Present address: Future Technology Corporate R&D, Seoul, Republic of Korea. [9]Present address: Donald Danforth Plant Science Center, St. Louis, MO, USA. [10]Present address: Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, China. [11]Present address: School of Plant and Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.
*e-mail: westwood@vt.edu; cwd3@psu.edu

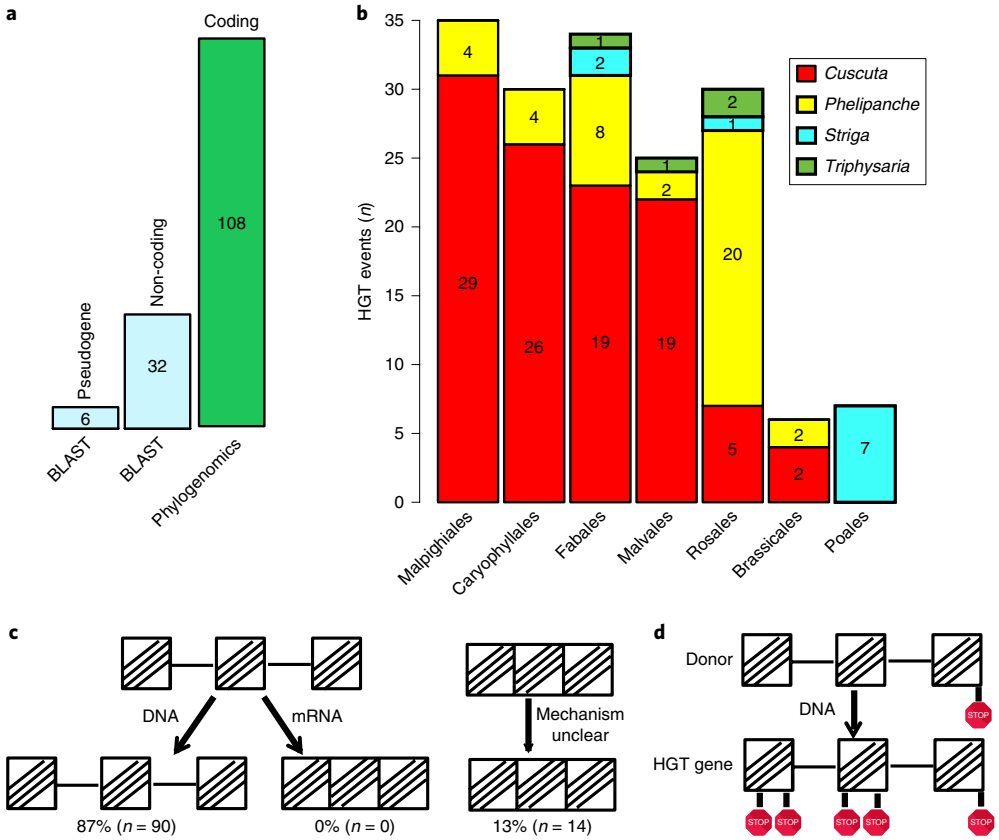


Fig. 1 | Identification and characterization of HGT genes and donors. a, Barplot showing 108 functional HGT (fHGT) events of coding genes by phylogenomics (green), and six non-functional HGT-derived pseudogenes and 32 non-coding elements by genome-based BLAST (light blue) (Supplementary Table 4). **b**, Comparison of HGT results for *Cuscuta* versus Orobanchaceae. Numbers of HGT events are grouped by donor, to reflect donor preference of each parasitic plant. HGTs from Myrtales (three events), Apiales (three events) and Sapindales (one event) are not shown on this barplot. **c**, Percentage breakdown of the number of mRNA- and DNA-mediated HGT events (zero RNA-mediated HGTs). Squares with three diagonal marks are exons, horizontal lines between the squares are introns. **d**, Example of HGT leading to pseudogenes: the presence of introns suggests genomic transfer and accumulation of premature stop codons, indicating that some HGT sequences are subject to pseudogenization.

case 3 in (Supplementary Fig. 2). In this study, we provide systematic approaches for HGT identification and a rigorous pipeline for further validation and analysis of each HGT event. This led to evidence of new phenomena in parasite–host interactions, including: convergence of adaptive HGT events in independent parasite lineages, ancestral HGT before species radiation in a large parasitic lineage and evidence that horizontally acquired sequences are sources of mobile small RNAs that may play a role in parasite–host interactions.

Results

A joint BLAST and phylogenomic approach for transcriptome- and genome-based HGT discovery. To identify coding HGT events in *C. campestris*, we used the phylogenomic approach established in ref. [7], with slightly improved detection strategies (Supplementary Fig. 3), applied to a comprehensive transcriptome from an earlier study of *Cuscuta*[17] (now identified as *C. campestris*[15]) (Supplementary Fig. 1a). After rigorous validation, including confirmatory analyses with transcriptomes from eight additional *Cuscuta* species (Supplementary Fig. 1a), phylogenomic analyses identified 108 high-confidence HGT events (Fig. 1a), twice the total number of events detected in three parasitic plant genera in Orobanchaceae[7] (Fig. 1a, Supplementary Fig. 4 and Supplementary Tables 1 and 2). A stoichiometry plot where HGT genomic contigs have levels of read coverage and genomic copy content similar to vertically transmitted genomic contigs (Supplementary Fig. 5), plus evidence from quantitative PCR with reverse transcription (RT–PCR) of representative HGT sequences (Supplementary Fig. 6), provide additional support that these HGT sequences are not due to contamination. Unlike *Phelipanche*, which shows strong bias toward sequences from host lineages in the Rosales (20/40 are from Rosales), the generalist feeder *Cuscuta* has acquired sequences from a wider range of angiosperm hosts, including Malpighiales, Caryophyllales, Fabales and Malvales (Fig. 1b and Supplementary Tables 2 and 3). Consistent with its current feeding patterns, no HGTs in *Cuscuta* were derived from grasses (Poales) (Fig. 1b).

To complement phylogenomically based HGT identification of protein-coding sequences but, in particular, to explore the possibility of reverse transcription-mediated HGTs, we applied a BLAST-based approach (HGTpropor) to the sequenced nuclear genome to identify pseudogenized or non-coding HGTs (Supplementary Fig. 1b). This method identifies HGTs based on the proportion of strong BLAST alignments (hits) from distantly related organisms compared to closely related taxonomic lineages. This method identified a minimum of 41 host-derived non-coding HGTs (Supplementary Table 2), including (1) HGT-derived pseudogenes from six gene families (OrthoFinder groups; five of six OrthoFinder groups encode retrotransposon-related proteins in the donor genome) and (2) 32 plant-derived non-coding HGT scaffolds, the majority of which (24 out of 32) are Fabales/Brassicales-derived long terminal repeat (LTR) elements including 24 Ty1/Copia and one Gypsy/DIRS1 (Fig. 1a and Supplementary Table 2).

While the method has been tuned to be conservative, it also identifies, when applied to coding regions, a large subset of the HGTs (59 OrthoFinder groups, 74 events) predicted by phylogenomics (Supplementary Figs. 1, 4 and 7a and Supplementary Tables 4 and 5). Moreover, HGTpropor identified two published HGTs (Supplementary Table 6)—albumin 1 (ref. [10] (Supplementary Fig. 8) and SSL[9] (Supplementary Fig. 9).

Because we have tried to establish a high bar for HGT acceptance through stringent criteria and multiple cross-checks, it is likely that we have rejected some true HGTs that lacked sufficient evidence. Moreover, additional HGTs may still be discovered with additional transcriptome and/or genome sequencing of the different *Cuscuta* species, and with continued expansion of the sequenced plant database.

Protein-coding HGTs are under evolutionary constraint. To test whether HGT sequences are evolving under constraint and likely to encode functional proteins, we used the branch test in PAML[18] for evolutionary analyses of the 108 HGT events, in which HGT sequences from *C. campestris* and related *Cuscuta* species on each phylogenetic tree were used as the foreground. The overwhelming majority of protein-coding HGT sequences in *Cuscuta* are evolving either under selective constraint similar to the background levels (69 genes) or have experienced even stronger levels of purifying selection (34 genes) (Supplementary Table 7). Three *Cuscuta* HGT lineages display relaxed purifying selection compared to the background (Supplementary Table 7), and two suggest positive selection (Supplementary Table 7). This analysis provides another line of evidence supporting functionality of the transcribed HGT sequences in *Cuscuta*.

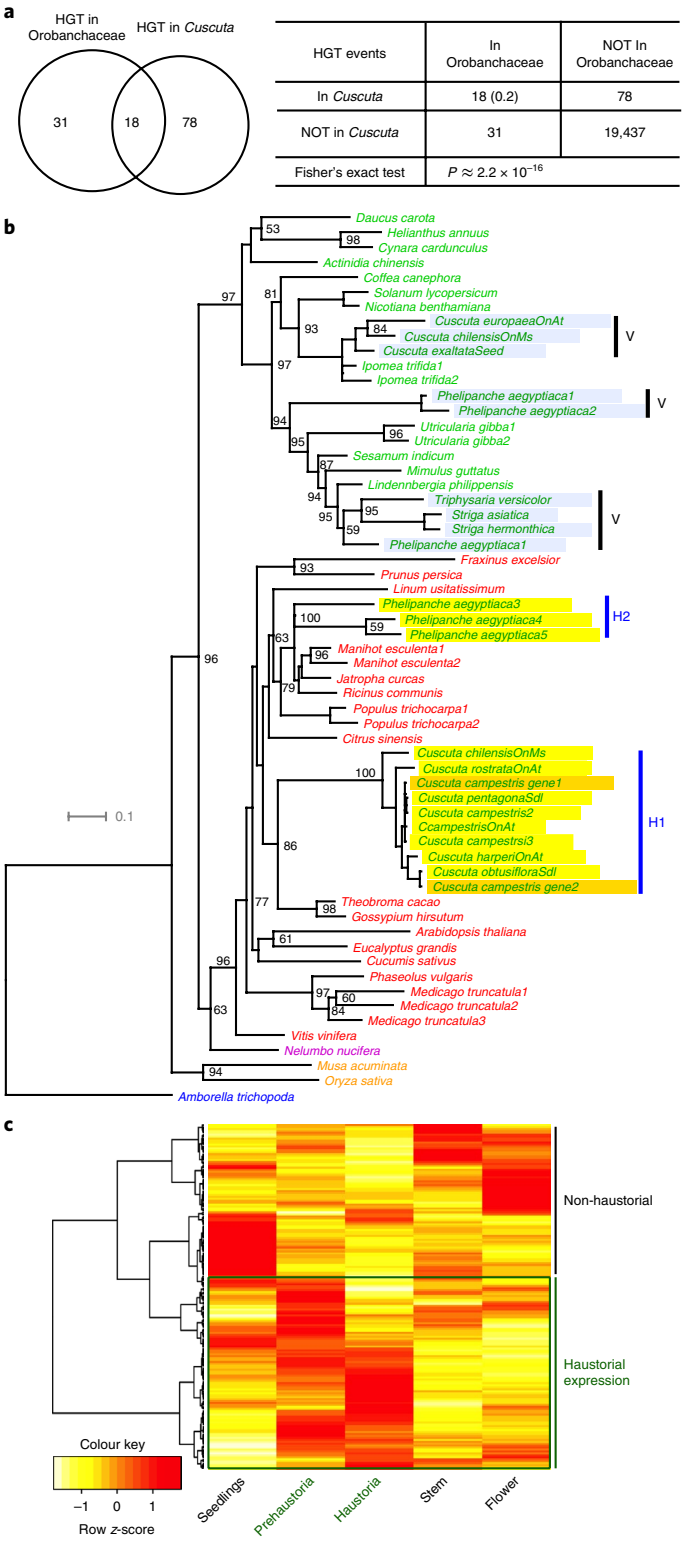
Defence response and mRNA processing are enriched among the HGT genes in *Cuscuta*. To infer the potential roles of the horizontally acquired genes in *Cuscuta*, we performed Gene Ontology[19] enrichment analyses of the 108 transcribed and functional HGTs. We found that defence response genes (Gene Ontology biological process (BP)) are significantly enriched (Supplementary Table 8) (Fisher's exact test, $P = 1.69 \times 10^{-4}$). Other enriched terms included protein phosphorylation ($P = 6.27 \times 10^{-8}$) and leucine-rich repeat domains ($P = 3.36 \times 10^{-8}$), which are also commonly associated with defence responses[20] (Supplementary Table 8). Another enriched GOSlim BP category is amino acid metabolic processes ($P = 0.047$), which may be linked to enriched Interproscan terms such as aminoacyl transfer RNA synthetase ($P = 0.02$) (Supplementary Table 8). Among the non-coding HGT set, retrotransposase and LTR retrotransposon sequences are mostly represented among the HGT pseudogenes and non-coding repetitive elements, respectively (Supplementary Table 2).

Functional and non-functional HGT—DNA not mRNA. Abundant mRNA transfer between *Cuscuta* and its host[14] led us to hypothesize that mobile RNA could be an important source of HGT in this parasite. However, introns were detected in all of the transcribed and evolutionarily constrained HGT genes where the donor sequences contained introns (90 fHGT events) (Fig. 1c). The absence of any detected intron losses argues against a mRNA-mediated mechanism for fHGTs in *Cuscuta*. To investigate this further, we randomly selected ten gene families to compare the intron positions and sequence from the HGT gene, the inferred donor and the vertical relative. Due to their usually short length, and difficulty of homology assessment in gappy alignments, intron phylogenies are typically poorly resolved. Similar to HGT in Orobanchaceae[7], the intron positions are largely highly conserved across the sampled angiosperms, including donor, recipient and related lineages. In one case where the donor and acceptor sites GT and AG could be identified (Supplementary Fig. 10), comparison of the sequences confirmed a match between the donor and parasite, and not the vertical relative, providing structural confirmation of a genomic transfer (Supplementary Fig. 10). We then considered whether site-specific homing introns may have invaded intron-less HGT sequences after insertion. An example of such an invasive sequence that is known in parasitic plants is the Type I homing intron of probable fungal origin[21] that has repeatedly invaded the mitochondrial *coxI* gene of parasitic plants[22,23]. However, a search for functional annotations associated with each of the 729 introns contained in the HGT genes in this study found only a single intron sequence with the annotation term Endonuclease-reverse transcriptase (intron from Seq716_Cp_v0.1_Contig94547_16048). There were no occurrences of other annotation terms that would also have suggested a potential homing intron (see Methods). These lines of evidence all point to *Cuscuta* fHGTs resulting overwhelmingly from transfer of intact genomic DNA, including introns. Such sequences are much more likely than randomly inserted retroprocessed sequences to have recognizable promoter elements in the new genomic environment, greatly increasing the likelihood that a foreign sequence would be transcribed and find function in the recipient parasite[7].

Although genomic integration is the apparent mechanism for fHGT events in *Cuscuta*, we hypothesized that reverse transcription of abundantly transferred mRNA in this species might lead to an accumulation of unexpressed pseudogenes, because reverse-transcribed mRNAs would usually be incorporated into genomes without functional regulatory modules. However, in the majority of HGT-derived pseudogenes we examined (Supplementary Table 2) that mostly encode retrotransposons or retrotransposon-domains, an absence of introns in the donor sequence made it impossible either to support or refute a reverse transcription-mediated hypothesis. In just one instance, we observed conserved introns in an HGT sequence that were similar to those in the donor gene, indicating a genomic transfer (Fig. 1d and Supplementary Fig. 11). The presence of multiple premature stop codons suggests that this sequence is a pseudogene (Fig. 1d and Supplementary Fig. 11). In another case of retrotransposon HGT, we observed both non-functional and potentially fHGT copies coexisting in the *Cuscuta* genome (Supplementary Fig. 12).

Further refuting the possibility of reverse transcription-mediated transfer, we compared our list of HGT genes to a dataset of host-to-*Cuscuta* mobile mRNA and found no enrichment in HGT genes among these mRNAs[14] (Supplementary Table 9). In contrast, we found that HGT genes from 23 orthogroups generated *Cuscuta*-to-host mobile mRNAs (Supplementary Table 10). These include mobile HGT genes that encode disease resistance proteins such as NB-ARC domain-containing disease resistance protein, cell wall-modifying enzymes including beta-glucosidase 24, and leucine-transfer RNA ligase. Although we find no evidence for a role of mobile mRNA as a source of HGT, transferred genes in the parasite express mRNAs that move back into the host and could affect parasite–host interactions (Supplementary Table 8). We conclude that mobile DNA, rather than mobile mRNA, has been the primary contributor to HGT in *Cuscuta*. Questions regarding the fate and function of trans-species mobile mRNAs will need to be addressed in future research.

HGT displays functional and transcriptional convergence in two parasitic lineages. To examine whether certain types of foreign genes may be favoured for retention in parasitic species, we compared the list of *Cuscuta*-coding HGTs to those from Orobanchaceae parasites[7]. Among 96 HGT-derived OrthoFinder groups in *Cuscuta*, phylogenetic trees of 18 OrthoFinder groups support independent HGT events in Orobanchaceae parasites (Fig. 2a). For instance, independent horizontal acquisitions (Shimodaira and Hasegawa (SH) test, $P < 0.01$) of a leucine-tRNA ligase occurred in *Cuscuta* and *Phelipanche* from Malvales and Malpighiales, respectively



HGT events	In Orobanchaceae	NOT In Orobanchaceae
In *Cuscuta*	18 (0.2)	78
NOT in *Cuscuta*	31	19,437
Fisher's exact test	$P \approx 2.2 \times 10^{-16}$	

d

Total	Haustorial expression	Non-haustorial
Unique	42	30
Shared	12 (8.8)	2
Fisher's exact test	One-sided, $P = 0.046$	

Shared genes are enriched for haustorial expression

Fig. 2 | Convergent evolution of expressed HGTs in two independent parasitic lineages. a, Fisher's exact test (total number of genes, $n = 19{,}564$) to assess the significance of co-occurrence of orthogroups containing HGT events in both *Cuscuta* and Orobanchaceae lineages. The expected value for one HGT occurring in both lineages is 0.2. **b**, HGT tree (60 sequences) showing two separate events in *Cuscuta* and *Phelipanche* (SH test, $P < 0.01$). HGT sequences in *Cuscuta* and *Phelipanche* are labelled with yellow highlighting and as H1 and H2, respectively. Gene sequences annotated in the genome are highlighted in orange. Vertically transmitted sequences are labelled with green highlighting and V. Clades H and V contain parasitic species in the genus *Cuscuta* and the Orobanchaceae parasites *P. aegyptiaca*, *S. hermonthica* and *T. versicolor*. **c**, Heatmap showing the transcriptional profile (the expression of each gene from each tissue involves $n = 2$ replicates) of all *Cuscuta* HGTs (178 sequences). Genes with high expression in haustorial tissues are denoted as haustorial expression, and the remainder as non-haustorial. **d**, One-sided Fisher's exact test (total number of genes, $n = 86$) to test for over-representation of HGT genes with haustorial expression occurring in two lineages. **e**, Hypothetical tree illustrating the colour-coding system of the major angiosperm lineages in Figs. 2b and 3g.

(Fig. 2b). The overall probability of an HGT event in any given gene family in *Cuscuta* is approximately 0.49% (96/19,470 orthogroups), and in Orobanchaceae 0.25% (49/19,470). The probability of independent HGT events occurring in both of these lineages is very low ($0.0012\% = 0.49 \times 0.25\%$). Thus, the retention of 18 shared HGT OrthoFinder groups is extremely unlikely (Fisher's exact test, $P \approx 2.2 \times 10^{-16}$) (Fig. 2a). Moreover, at least half of the HGT genes in *Cuscuta* were highly expressed in haustorial tissues (Fig. 2c and Supplementary Table 11). Of the expressed HGT genes shared by both *Cuscuta* and Orobanchaceae, 14 were expressed in haustoria and two were not (Fig. 2d and Supplementary Table 11) (one-sided Fisher's test, $P = 0.046$). A previous study of Orobanchaceae parasites revealed HGT of four haustorial-upregulated tRNA synthetases in *Phelipanche*[7]. We found an enrichment of aminoacyl tRNA synthetases in *Cuscuta* HGTs (Supplementary Table 8). We examined the HGT gene topology and transcriptional profile in *Cuscuta* and Orobanchaceae parasites for the 18 shared HGT trees (Fig. 2a). Six OrthoFinder trees support strong convergent evolution where the HGT genes in both lineages show similar expression and independent transfer events (Supplementary Table 12). Together with their upregulation in haustorial tissues[7] (Supplementary Table 11), this suggests that these HGT gene products could contribute to metabolic capacity and/or parasitic ability of the haustoria in the two parasitic lineages. Thus, we see evidence for widespread convergent acquisition of HGT genes, linking HGT gene function and expression through differential HGT survival and acquisition of haustorial expression.

Despite the striking similarity in events between *Cuscuta* and Orobanchaceae, we see evidence of HGT genes encoding cell wall-modifying enzymes that were captured by an ancestor of all sampled *Cuscuta* species (Supplementary Figs. 13–15), but not by Orobanchaceae parasites[4,7]. Two of these HGT-derived cell wall-modifying enzymes (glycosyl hydrolase and pectin acetylesterase) are expressed primarily in prehaustorial structures (Supplementary Fig. 13), implicating a role in host invasion.

HGTs as sources of mobile small RNAs. Given recent evidence that *Cuscuta*-derived miRNAs target host gene expression[15], we hypothesized that HGT could be a source of host-specific small RNAs. This was supported by the discovery of many HGT genes in *Cuscuta* encoding leucine-rich repeat (LRR) domain-containing proteins (Supplementary Tables 2 and 8), which are known to generate miRNA-dependent short interfering RNAs that may act in silencing cascades[24–26]. We also found several transposable element-related HGTs and repeat-derived HGT fragments (Supplementary Table 2). Repetitive DNA and transposable elements are known to be rich sources of 24-nucleotide (nt) siRNAs that target their silencing via the RNA-dependent DNA methylation pathway[27]. To evaluate this hypothesis, we compared the HGT set to previously annotated small RNA loci in *C. campestris*[15]. We found that 75 out of 200 *C. campestris* protein-coding HGT genes (the combined set of HGT sequences identified from phylogenomic and BLAST approaches, shown in Supplementary Table 2) overlap with 147 small RNA loci (Fig. 3a and Supplementary Table 13), significantly higher than the non-HGT background (right-tailed Chi-square test, $P = 1.01 \times 10^{-5}$) (Supplementary Table 13). This suggests that HGT sequences are more likely to be small RNA sources compared to canonical genes. The majority of the overlapping small RNA loci (119 out of 147) produce small RNAs from both strands (Fig. 3a), indicating these are siRNA loci. In the case of non-coding HGT fragments and pseudogenes, only five out of 43 overlap with small RNA-producing loci (Supplementary Fig. 16b and Supplementary Table 14). Most of the protein-coding, HGT-overlapping small RNA loci (138/147) predominantly generate 24-nt siRNAs (Fig. 3b, Supplementary Fig. 16a and Supplementary Table 13). Besides LRR genes and transposable elements, many other protein-coding HGTs are also associated with 24-nt siRNA-producing regions (Supplementary Fig. 17). Most intriguingly, one of the HGT genes in *Cuscuta*, which encodes a LRR protein kinase (*Cc_v0.1_Contig437972_11875*) highly expressed in haustoria (Fig. 3g), overlaps with a known *MIRNA*, *ccm-MIR12494a*[15] (Fig. 3c). Specifically, the 3' UTR of the HGT gene overlaps with the 3' end of this *MIRNA* locus by 24 base pairs (Fig. 3c,d). *ccm-MIR12494a* has previously been shown to be induced in the haustoria (Fig. 3e), where it triggers cleavage of a host mRNA—*Arabidopsis Heat Shock Factor Binding 4/Schizoriza* (*HSFB4/SCZ*) (Fig. 3f)—followed by secondary siRNA production[15]. In *Arabidopsis*, *SCZ* regulates the determination of stem cell fate in ground tissues[28,29] and de novo formation of roots[30]. The phylogenetic tree supports an HGT in the common ancestor of all *Cuscuta* spp. from a Rosales lineage closely related to *Ziziphus jujuba*[31] (Fig. 3g). Although we cannot be certain of the exact boundaries of this HGT sequence given the fact that this miRNA family is not known outside of *Cuscuta*, we favour the hypothesis that it arose after the HGT event, perhaps evolving from the HGT-inserted sequence itself, given their close proximity. This suggests that a trans-species active miRNA is very closely linked to, if not part of, an ancient HGT event.

Given that HGT genes were foreign when they first entered the *Cuscuta* genome, we speculate that some HGT-associated siRNAs may have been generated as part of a genomic silencing response against the incoming DNA segments, which can often cause local repeat structures that trigger siRNA production. For instance, repetitive pararetroviral sequences integrated into the genome of the flowering plant *Fritillaria imperialis* have been shown to generate 24-nt siRNAs that target these for silencing[32]. It is also possible that some HGT-associated *Cuscuta* small RNAs, such as *ccm-MIR12494a*, may have been acquired as part of adaptation to parasitism.

HGT in *Cuscuta* is ancestral and an ongoing process. Nine horizontally acquired genes were identified in the transcriptomes of a genus-wide set of eight *Cuscuta* taxa sampled in this study, but absent from non-parasitic Convolvulaceae and related families, and are thus inferred to have been acquired in an ancestral *Cuscuta* parasite (Fig. 4 and Supplementary Table 15). By mapping the presence of the detected HGT sequences onto the reconstructed species tree, the number and timing of HGT events occurring at each node of the tree were reconstructed with two methods that accommodate intron loss or missing data[33,34]. In this way, between seven and 11 additional HGT events were inferred as ancestral in our genus-wide sampling (Fig. 4) (Supplementary Table 15). This is very different from previous observations of HGT in Orobanchaceae, where no HGTs were found to be ancestral[7]. The 16–20 ancestral *Cuscuta* HGT genes include the previously discovered albumin1 (refs. [10,12]), LRR protein kinase, cell wall-modifying enzymes such as the glycoside hydrolase family, and transposons (Supplementary Table 16). The largest number of surviving HGT events occurred in an ancestor of *Cuscuta chilensis* and the five other members of subgenus *Grammica* (Fig. 4). Three of the ancestral HGTs encode transposable element gene families (hAT transposons), with the level of sequence identity among *Cuscuta* spp. falling within the range of identities seen in the ancestral protein-coding genes (Supplementary Table 16). In addition, the ratios of synonymous and non-synonymous substitutions per site for the two transposable element-related orthogroups show the proteins evolving under strong and weak purifying selection (Supplementary Table 7). Furthermore, phylogenies of the two transposable element genes (OrthoFinder group 29 in Supplementary Fig. 4.11 and OrthoFinder group 31 in Supplementary Fig. 4.13) are very close to the *Cuscuta* species phylogeny. These lines of evidence together support the inference that these were indeed ancestral insertions as implied by the formal character reconstructions. The distribution of HGT events in

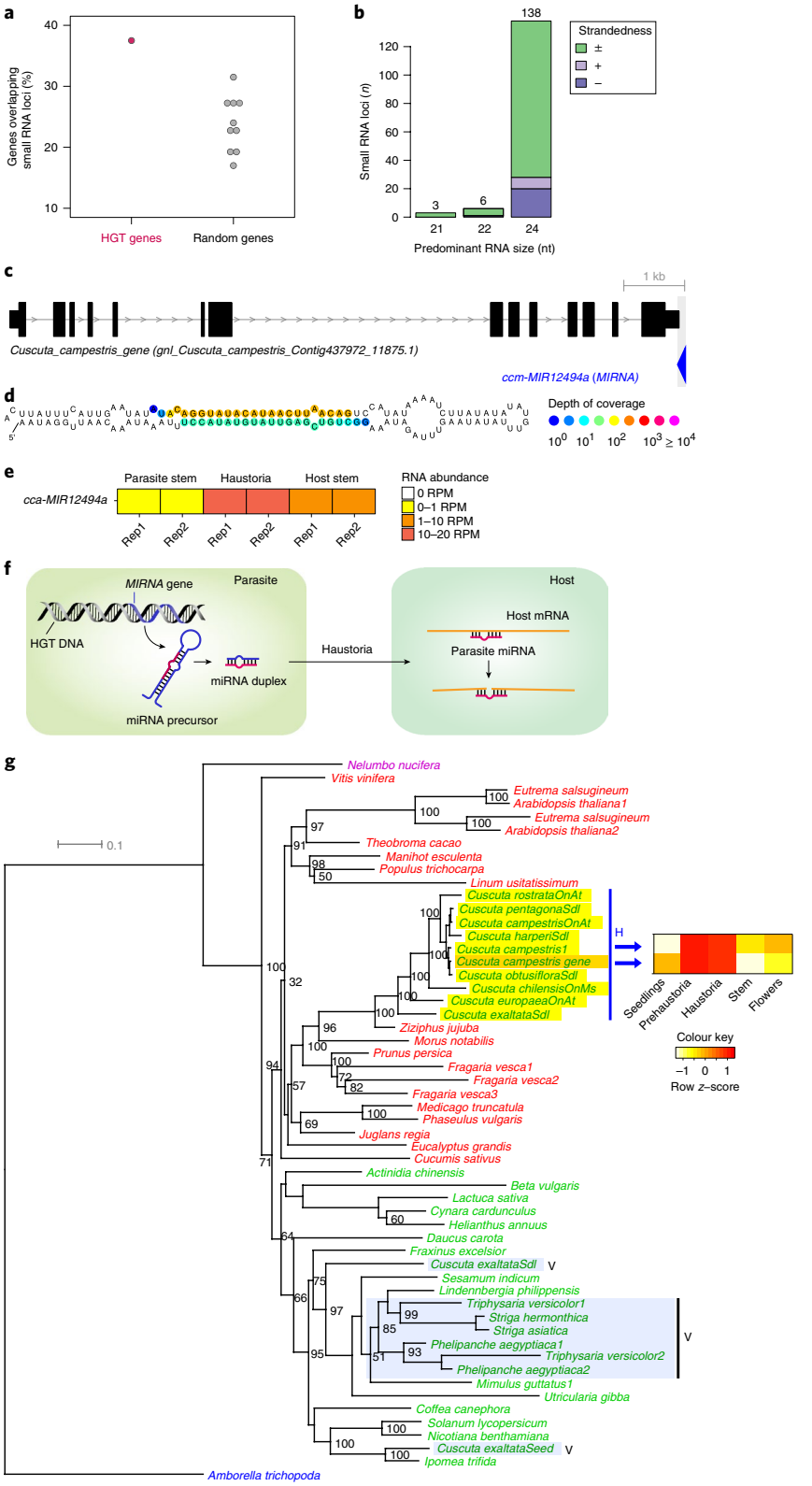


Fig. 3 | *C. campestris* HGT events are a frequent source of small RNAs. a, Scatter plot showing percentage of *C. campestris* HGT genes or random gene sets (ten random samplings of 200 genes per sampling) that enriched for overlap with small RNA-producing loci. **b**, Predominant RNA size and strandedness of HGT-overlapping small RNA loci. Here, the symbols +, – and ± indicate plus, minus and both strands of the genome, respectively. **c**, Structure of a HGT gene which overlaps with a known *MIRNA* locus (*ccm-MIR12494a*). **d**, Small RNA coverage along the hairpin structure of *ccm-MIR12494a*. **e**, Heatmap showing the expression of *ccm-MIR12494a* induced in haustoria. RPM, reads per million. **f**, Diagram illustrating how HGT-overlapping *ccm-MIR12494a* functions in parasite–host interactions. **g**, Phylogenetic tree (55 sequences) supporting the occurrence of HGT in a common ancestor of eight *Cuscuta* species from a Rosales ancestor. HGT sequences are highlighted in yellow and by H; vertically transmitted sequences are highlighted in light blue and by V. Heatmap ($n = 2$ replicates from each tissue) showing the primary haustorial expression of two HGT sequences.

various ancestral and terminal lineages indicates that HGT has been an ongoing process in *Cuscuta*.

Discussion

Our analyses of HGT in *Cuscuta* led to the following conclusions.

(1) As predicted by its heterotrophic habit and open vascular connections, *Cuscuta* has experienced high levels of host-to-parasite horizontal gene transfer. The markedly higher incidence of HGT in *Cuscuta* compared to holoparasitic Orobanchaceae[4,7] could have occurred because *Cuscuta's* numerous stem haustorial connections, distributed throughout the vine, are physically much closer to the floral meristematic tissues than in the root-parasitic Orobanchaceae, making the path shorter for integration of host DNA transported through the haustorium into cells that ultimately enter the parasite germline.

(2) Although large numbers of fHGTs have been discovered through analyses of transcribed gene sets in *Cuscuta*, we also identified numerous HGT-derived pseudogenes, transposon-derived sequences and non-coding elements in the *Cuscuta* genome. Despite the large number of documented HGT events, the genome of *C. campestris* is a fairly modest 581 mega base pairs (bp) per amount of DNA contained within an unreplicated haploid chromosome set (1C)[12], and the amount of xenologous sequence detected to date in this species is small relative to the total genome size. However, other parasitic plants have greatly expanded genomes that are far larger than those of closely related parasitic or non-parasitic relatives[35,36]. Parasites with giant genomes include certain Orobanchaceae (up to 19 Gbp per 1C)[37], *Cuscuta indecora* (32 Gbp per 1C)[35] and certain mistletoes (up to 80 Gbp per 1C, or 27 times larger than a human genome!)[38] (http://data.kew.org/cvalues/). We hypothesize that horizontal transfers of foreign sequences into parasitic plants expose them to host-derived transposable elements that can sometimes expand wildly in their new genomic environment. Most transposable elements, of course, are quiescent due to epigenetic silencing[39] or mutation. If either active or silenced transposable elements in the host genome are acquired via horizontal transfer into the parasite[40], they may in some cases be able to replicate rapidly outside of epigenetic control in the 'naive' parasitic plant genome, allowing them to expand greatly before silencing is achieved. Such an event of HGT-mediated 'escape' into the parasitic plant genome would be highly adaptive for transposons that are silenced but potentially functional. If true, fHGT could be just the 'tip of the iceberg' of the total amount of foreign DNA in the vastly larger genomes of some parasitic plants. Future sequencing of these giant genomes will be needed to test this prediction.

(3) Genes that have been acquired by *Cuscuta* through HGT are enriched for haustorial expression, defence response and amino acid metabolism, suggesting that fHGT contributes to the mechanism of parasite–host interactions. This process of acquiring functional genes that could contribute to parasitic ability began early in the history of *Cuscuta*, with 16–20 horizontally acquired genes having been retained from a common ancestor of extant *Cuscuta* species. Two independent parasitic lineages underwent convergent evolution after acquiring genes of similar functional categories through independent HGT events, with many genes evolving predominantly haustorial expression, supporting the notion that HGT has indeed been used as a mechanism for enhancing parasite–host interactions.

(4) *Cuscuta* HGTs are enriched as sources for endogenous siRNAs associated with silencing. This suggests that incoming HGT sequences might be recognized as foreign, and targeted for silencing. However, despite this targeting, the HGT genes we discovered are clearly expressed and are generally evolving under constraint as functional proteins. The overlap of an HGT event with a miRNA that targets host mRNAs suggests that HGT sequences could also spawn novel regulatory functions to promote parasite success. Collectively, these lines of evidence suggest that, in *Cuscuta*, HGT not only contributes to a parasitic lifestyle but modulates parasite–host interaction by interacting with mobile mRNA/sRNAs.

(5) Our study identifies evidence of cross-talk between mobile DNA (HGT), mobile mRNAs and mobile small RNAs (Fig. 5). We demonstrated that genomic DNA from the host, rather than mRNA, acts as the source of parasite HGTs. In addition, nuclear HGTs in parasitic *Cuscuta* have previously been detected only in the direction from host to parasite, and not from parasite to host as in our analyses (Fig. 5). DNA fragments moving from host to parasite can exist in the form of genes or non-coding elements such as transposable elements and repeats. Following horizontal transfer, these HGTs are subject to both selective retention and evolution by mutation, which can give rise to pseudogenes (Fig. 5).

(6) This study demonstrates that *Cuscuta* is unusually proficient at exchanging nucleic acids with its host. Taken together with mRNA movement between host and *C. campestris*, and the movement of miRNAs from *C. campestris* to host, this unprecedented number of HGT events in *Cuscuta* suggests there are few barriers to the exchange of material that was once thought to reside strictly in a single cell (note that this also includes exchange of proteins). We are increasingly aware of the roles of small RNAs and mRNA in the systemic transmission of signals within a single plant, so the prospect of such exchange occurring between plant species raises many questions. Our study argues strongly for a DNA mechanism of HGT, but offers little insight into the pathway taken by a genomic integrant. Is DNA itself mobile within a typical plant? Is a viral vector involved, or does transfer occur directly from plant to plant only in unusual circumstances, such as grafting or parasitism? In addition, the transfer of DNA between plants is only the first step in a potentially ongoing interaction. The retained DNA sequence may produce an mRNA that directly codes for a protein useful in that cell, or the mRNA may move systemically around the parasite or even back into the host (Fig. 5). Additionally, the HGT may evolve into a miRNA-encoding sequence that can regulate host gene expression and thus facilitate parasitism (Figs. 3 and 5). At each step the exchange of nucleic acids must confer an advantage to the parasite for the genes to be retained, expressed and evolving under constraint, long after the HGT event, so it would appear that *Cuscuta* and other parasites gain substantial benefit from acquiring host genes.

Methods

Phylogenomic reconstruction of *Cuscuta* gene trees. A comprehensive de novo transcriptome dataset from *C. campestris*[15,17] was used for protein-coding gene predictions. Two versions of de novo transcriptome assembly (one from ref. [17] with the authors' names in the format (Cuscuta_pentagona_Ranjan_117086), the other from our own transcriptome assembly (with authors' names in the format, Cpent200559_cp73505_c0_seq1_1911x_f1)) were used to capture all the transcripts from their data. The first assembly will henceforth be denoted as 'Ranjan et al. assembly', whereas our own assembly will be denoted as 'custom assembly'. The *AssemblyPostProcessing* pipeline in PlantTribes (v.1.0.2) (https://github.com/dePamphilis/PlantTribes) was used to predict non-redundant sets of coding regions (>200 nt) and their corresponding amino acid translations for both de novo transcriptome assemblies in the study using ESTScan (v.3.0.3)[41]. In total, 59,949 sequences from the custom assembly were then classified into 9,852 orthogroups with the 26 genome orthogroup classifications. To increase the signal-to-noise ratio for HGT detection, we added an expanded set of potential donor genomes and additional Asteridae genomes that are closely related to *Cuscuta*. These 13 additional genomes include *Ipomea trifida* (sweet potato line Mx23Hm)[42], *Coffea canephora* (coffee)[43], *Fraxinus excelsior* (ash)[44], *Sesamum indicum*[45] (sesame), *Actinidia chinensis* (kiwifruit)[46], *Cynara cardunculus* (artichoke)[47], *Nicotiana benthamiana* (tobacco)[48], *Daucus carota*[49]

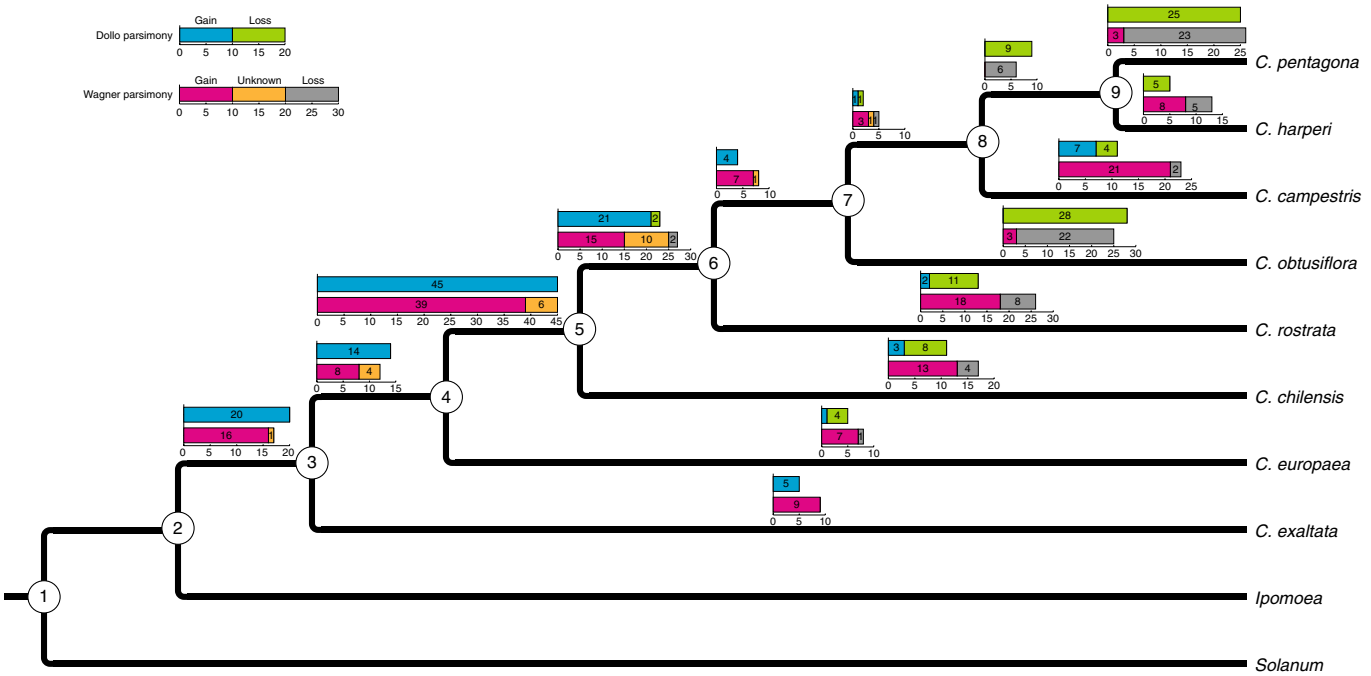


Fig. 4 | HGT is both ancestral and an ongoing process in *Cuscuta*. The HGT states of each ancestral node of the *Cuscuta* species tree[33] were inferred with Dollo and Wagner parsimony (Supplementary Table 15). The inferred numbers of HGT events for each ancestral and terminal node are labelled with two sets of bars, representing each reconstruction method. The inferred numbers of HGT gain-and-loss events (which could represent unexpressed genes in the tissues sampled) are labelled within each bar.

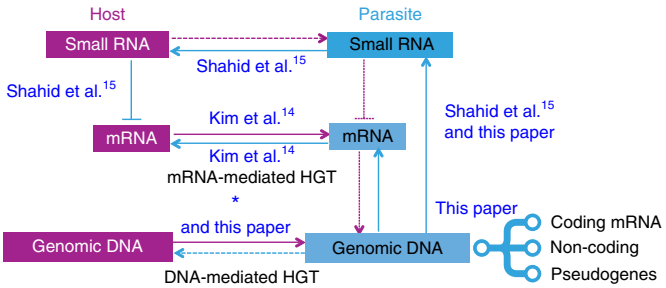


Fig. 5 | HGT pathways in light of interaction with mobile mRNAs and mobile small RNAs. Blue represents parasite, purple represents host. Solid lines represent direct evidence, dashed lines represent potential pathways that currently lack evidence in this system (but see ref. [13] for a probable case of parasite-to-host HGT in plant mitochondrial DNA). Numbers on branches represent supporting reference numbers, and the blue asterisk represents a group of five references (refs. [2,7–10]). DNA-mediated HGT from host to parasite is the predominant form; HGT from parasite to host is rare and seldom results in fHGT. mRNA-mediated HGT has not been observed in *Cuscuta*. HGT sequences exist predominantly in functional protein-coding genes, sometimes as non-coding elements (such as repeats) or evolve gradually into pseudogenes following transfer.

(carrot, Phytozome), *Amaranthus hypochondriacus* (prince's feather, Phytozome), *Kalanchoe marnieriana* (Marnier's kalanchoe, Phytozome), *Linum usitatissimum* (flax, Phytozome), *Capsella grandiflora* (grand shepherd's purse, Phytozome) and *Eutrema salsugineum* (saltwater cress, Phytozome). Three Orobanchaceae parasitic plants were also included—*Triphysaria versicolor* (yellowbeak owl's clover), *Striga hermonthica* (giant witchweed) and *Phelipanche aegyptiaca* (Egyptian broomrape)—with transcriptome assemblies from our previous study[50] but with post-processing improvements as above. Genes from these 16 species were assigned to the 26-genome orthogroup classifications and used to build trees together with *C. campestris* sequences. Two versions of trees were generated, one by filtering out short sequences resulting in gaps in multiple sequence alignments, the other by retaining all sequences in each orthogroup. For the custom assembly, 9,782 trees were generated for sequences from 9,852 orthogroups that contained at least four sequences. An additional 902 trees were built using genes from the Ranjan et al. genome assembly[17] that were classified into a unique set of orthogroups compared to those from the custom assembly.

HGT screening on phylogenetic trees. Screening for HGT was performed using both large-scale phylogenetic trees and BLAST analyses. Phylogenomic screening of HGT followed the same approach employed in ref. [7], but with improvements to enhance handling cases of HGT events involving more than one *Cuscuta* sequence. An improved schema (Supplementary Fig. 3) of all possible HGT scenarios was designed to identify HGT from putative donors from distantly related rosids and distant asterids. Customized scripts were developed to identify nodes with HGT signal and placed at GitHub (https://github.com/dePamphilis/Cuscuta_HGT_ms_code). Further manual curation was performed for each orthogroup to improve the tree by increased taxon sampling to fix long branch artefacts or increase the taxon density when the HGT signal was weak.

BLAST-based HGT screening. *HGTpropor*. We developed a new approach that identifies HGT-derived sequences based on BLAST analyses. This approach is reliant on the assumption that a vertically transmitted sequence will have a high proportion of hits from closely related taxa (close group), whereas a HGT-derived sequence will have a high proportion of hits from distantly related taxa (distant, or distal, group), and that often the hits are from one homogenous distal group. BLAST results were parsed by following several criteria: (1) genes with best non-self hit from non-plant sources were removed; (2) genes with five or fewer *Cuscuta*-only hits were removed; and (3) the remaining genes were examined by their Viridiplantae hits. The detailed procedure (involving manual curation, because some hits to the NCBI non-redundant protein collection (NR) lack taxonomic information such as order or phylum) on parsing of BLAST results from both NR and customized databases is available on GitHub (https://github.com/dePamphilis/Cuscuta_HGT_ms_code). Because the signal is concentrated mainly on the top hits, only the top 50 BLAST hits were examined. We termed this approach HGTpropor, due to its focus on a high proportion of top hits from distantly related taxa. A fairly complete HGT database was built that includes sequences from the NR and a manual selection of taxa (137 species) including 48 genomes and 89 transcriptomes (our own collection or taxa from the 1000 Plant Transcriptome Project[51] and PlantGDB[52]) (Supplementary Table 17). Customized scripts were developed to obtain taxonomic classification (species, genus, family, order, manually curated group) by parsing BLAST hits from both NR and our manually collected taxa. Each taxonomic order was classified into a group based on their relationship with the focal taxa (*C. campestris*)—self, close or distal group (Supplementary Table 18). Classification distinguishes between a genus from self or

close taxa and one from parasitic or non-parasitic taxa. A distal group was split into several smaller evolutionary lineages to decrease the noise. In this study, *Cuscuta* belongs to Convolvulaceae in the order Solanales, which can be either a self (if the hit is to *Cuscuta*) or close (if to another genus, such as *Convolvulus*, *Ipomoea* or *Solanum*). Sequences from any parasitic taxa in the top 50 hits were considered as potential HGT across parasitic plants (Supplementary Table 18). A distal group was further classified into either distal rosid, distal asterid (Asterales, Apiales, Ericales, Berberidopsidales, Santalales and Caryophyllales), distal basal (basal eudicots and basal angiosperms), distal non-seed plant or distal gymnosperm (Supplementary Table 18). Based on this classification, scripts were developed to produce a proportion for each gene from self, close and various distal groups (https://github.com/dePamphilis/Cuscuta_HGT_ms_code). An HGT candidate was characterized by the signature of a high proportion of hits from a distal group but low proportion from a close group. In our analyses, we applied a distal cut-off ratio of 0.7 and close cut-off of 0.2.

HGTector. HGTector, a previously published approach[53] for identification of HGT from BLAST results, was also employed in our study. HGTector uses a normalized score by dividing each hit's bitscore by the self bitscore. Then for each gene, the normalized bitscores from self, close and distal groups are summed to generate three measures for each gene. To decrease the noise from inaccurate open reading frame prediction, we kept BLAST results only for genes with greater than 20 hits. A density distribution was then plotted for each self, close and distal group using the measures for all genes. Cut-offs were determined by taking the midpoint between the first peak and valley from the distribution of self, close and distal groups. HGT genes were identified with a close weight smaller than the close cut-off but a distal weight greater than the distal cut-off. Inspired by our HGTpropor approach, we adapted it by splitting a distal group into several distal groups, requiring each taxon from a distal group to be closely related to the others.

Using a combined BLAST-based identification from HGTector and HGTpropor, we gathered a union of all HGT candidates from each approach (Supplementary Table 19). Candidate HGT genes in *Cuscuta* were classified into 26-genome orthogroup classifications (Supplementary Table 20), and gene trees were built for validation using the approach previously described (see Phylogenomic reconstruction of *Cuscuta* gene trees, above).

Validation of HGT candidates (protein-coding genes). Horizontal gene transfer candidates were validated with an automated tree reconstruction pipeline, plus a scoring system. In this round of validation, all 26 genomes comprising the backbone of the orthogroup classification were included, along with additional genomes and transcriptomes. The additional genomes included 13 sequenced plant genomes (see Phylogenomic reconstruction of *Cuscuta* gene trees, above). The transcriptomes included three Orobanchaceae parasitic taxa described above (*T. versicolor*, *S. hermonthica* and *P. aegyptiaca*), with the aim of identifying HGT events that had occurred in more than one parasitic lineage; eight *Cuscuta* species with de novo transcriptome assemblies of one or two developmental stages from each taxon (*C. campestris* (haustoria on *Arabidopsis thaliana*), *C. chilensis* (haustoria on *Medicago sativa*), *Cuscuta europaea* (haustoria on *A. thaliana* and seeds), *Cuscuta exaltata* (seeds and seedlings), *Cuscuta harperi* (haustoria on *A. thaliana* and seedlings), *Cuscuta obtusiflora* (seeds and seedlings), *Cuscuta pentagona* (seedlings) and *Cuscuta rostrata* (haustoria on *A. thaliana*)); and 11 closely related taxa, to increase the signal/noise ratio (*Atropa belladonna* (1,000 Plants database (1KP) project[51]), *Convolvulus arvensis* (1KP), *Ipomoea indica* (1KP), *Ipomoea quamoclit* (1KP), *Olea europaea* (Trinity assembly of reads from 1KP), *Paulownia fargesil* (Trinity assembly of reads from 1KP), *Ipomoea nil* (plantGDB), *Petunia integrifolia* (plantGDB), *Physalis peruviana* (plantGDB), *Solanum melongena* (plantGDB) and *Wrightia tinctoria* (plantGDB)). The scoring criteria described in ref. [7] were employed to validate the confidence of HGTs in the final versions of the phylogenetic tree.

All medium- and high-confidence HGT orthogroup trees were retained and then subjected to a rigorous validation process. HGT sequences in *C. campestris* were first searched against the v.0.1 *C. campestris* assembly[15] to confirm their presence in the genome. Sequences without BLASTn hits in the genome assembly were eliminated from further consideration. Second, sequences were searched with BLASTn against the NCBI non-redundant nucleotide database (NT) to remove possible contamination from the experimental host. Third, candidate sequences, especially from phylogenomic screening, were searched with BLASTp against NR (downloaded from ftp://ftp.ncbi.nih.gov/blast/db on 7 October, 2015) to ensure that top BLASTp hits were from a distantly related donor (HGTpropor criteria: BLAST hits from closely related taxa were weaker than hits to distal taxa) and that the distal blast bitscore was significantly greater than the self bitscore (HGTector criteria[53]). This is necessary because BLAST searches are exhaustive, whereas the donor inferred from the tree may be limited to species with sequenced genomes or good transcriptomes. Thus if top BLASTp hits of the HGT candidate match the tree-inferred donor or it is a close relative of the tree-inferred donor, it increases the confidence that the HGT candidate represents a true HGT. Finally, an expanded phylogenetic tree was built with sequences from the orthogroup containing the HGT sequences. At this stage, a final list of 105 true HGT OrthoMCL (v.2.0.9) groups was validated from the candidates predicted by a combination of BLAST and trees (Supplementary Table 21). However, this set of 105 true HGT trees contains a largely redundant set of *Cuscuta* sequences because sequences were obtained from two independent *C. campestris* transcriptome assemblies, high-confidence Maker-P gene models from the 0.1 genome assembly (see section Genome assembly and annotation, below) and closely related copies detected from transcriptome assemblies in eight additional *Cuscuta* species.

Therefore, to further improve the trees for the final set of 105 HGT OrthoMCL groups, an OrthoFinder (v.1.1.2) classification was built by selecting 34 representative genomes used for *Cuscuta* HGT identification (Supplementary Table 22). Sequences from the 105 HGT OrthoMCL groups were cleaned and subjected to phylogenetic reconstruction with OrthoFinder groups. These OrthoFinder trees were used to validate the HGT events from the 105 HGT orthogroups.

The final tree was accepted as a high-confidence HGT only if it matched either scheme 1 (HGT sequences nested within the donor clade with two strong supporting nodes) or scheme 2 (HGT sequences sister to the donor clade) (Supplementary Fig. 3), with two strong supporting nodes. When the HGT sequence grouped with a list of donor sequences but without high bootstrap support values, manual addition of more donor hits from a NR BLAST was used in an attempt to improve the tree.

Genome assembly and annotation. The *C. campestris* genome was assembled with SOAPdenovo using 100-bp PE Illumina reads with insert sizes of 200 bp, 340 bp, 480 bp, 3 kb and 5 kb (after read cleaning and assembly, approximately 42× depth of nuclear genome). The de novo assembly was scaffolded and gap-filled using PBJelly (PBSuite v.15.2.20) with approximately 5× coverage of long Pacific Biosciences RSI-filtered sub-reads (8 SMRT cells), resulting in 56,350 scaffolds (≥1 kb) with an N50 of 16.18 kb. We followed the protocol described in ref. [54] to create a *C. campestris*-specific repeat library suitable for repeat masking before protein-coding gene annotation. Genes were predicted using the MAKER-P pipeline (v.2.31.8) with training sets incorporating curated plant proteins from SwissProt and *C. campestris* mRNA sequencing data from multiple tissue types including seed, seedling, stem, haustoria, flower and leaf. A total of 42,494 protein-coding genes were predicted, consisting of all gene models supported by annotation evidence and gene models not supported by annotation evidence but encoding Pfam domains. The 0.1 genome assembly and annotation are available at the Parasitic Plant Genome Project database (http://ppgp.huck.psu.edu/cuscuta.html). A detailed description of the *C. campestris* genome assembly, species-specific repeat database construction and gene annotation will be reported elsewhere.

Selective constraint analyses. Branch tests were performed following the procedure described in ref. [7]. For each of the 108 HGT events, HGT sequences from *C. campestris* and related *Cuscuta* species were used as the foreground, with the remaining sequences used as the background. The foreground and background lineages for each gene tree are clearly indicated on individual supplementary phylogenetic trees, with the foreground sequences indicated by '#1' (a link to pdf trees indicating the foreground and background is available at http://ppgp.huck.psu.edu/data/Cuscuta_HGT_Manuscript_Data/pdf_trees_with_foreground_labeled_for_constraint_analysis.zip).

Ancestral reconstruction of HGT events. Each of the HGT events was coded as one of two states: 0 (without or not detected HGT) or 1 (with HGT) (Supplementary Table 15). Both Dollo and Wagner parsimony methods in the PHYLIP package (v.3.695)[55] with default options (search for best tree-yes; Randomize input order of species-no; Use Threshold parsimony-no; Use ancestral states in input file-no; Sites weighted-no; Terminal type-ANSI) were used to infer the ancestral states of HGTs with an input tree for the eight sampled *Cuscuta* species, which was simplified from published phylogenies[33].

Bioinformatic search for homing introns. Each of the 729 introns contained in the HGT genes in this study was extracted from the *C. campestris* 0.1 genome and further examined with BLASTp searches against databases NCBI/t, NCBI/nr, UniProt/SwissProt, UniProt/TrEMBL and TAIR10. Introns were also searched against a collection of protein family domain databases included in InterProScan (v.5.25.64.0) software and assigned with identified domains, which were also translated into gene ontology terms. The Interproscan command was: 'interproscan.sh -i <INPUT FILE> -f TSV -goterms -iprlookup –o <OUTPUT FILE> t -T temp -t p'. We then queried the text for terms associated with sequences (homing, endonuclease, LAGLIDADG, I-PpoI and I-CreI. GIY-YIG, PD(D/E)xK, His-Cys Box, HNH, nicking enzyme, intein) that could be associated with homing capability[56,57].

Genomic identification of HGT candidates using BLASTN. The genome scaffolds of *C. campestris*[15] were interrogated to search for potential HGT-derived sequences. A BLASTn search was used because certain features, such as pseudogenes and non-coding elements, do not encode proteins and would be missed by a BLASTp-based approach and phylogenomics of predicted peptide sequences of genes. The BLASTn search started with 91,940 genomic scaffolds resulting from the genome assembly against the latest NT database. The BLASTn

commands used were: 'blastn -task megablast -outfmt 6 -query input.fa -db nt_database -out input.blastn.nt -evalue 1e-5'. Next, BLAST results were processed by extraction of the GenInfo Identifier, which was converted to a taxonomic identification from each BLASTn hit. The command to convert GenInfo Identifier to taxonomy ID is: 'efetch -id gi -db nucleotide -format docsum | xtract –pattern DocumentSummary -element TaxId > output'[58]. Custom scripts using the 'Entrez. efetch' command were developed to export taxonomic information from each taxonomy ID. Taxonomic information includes scientific name, kingdom, subclass, order, family and genus. The tabular blast output was then concatenated with each hit's taxonomic information. BLAST results were parsed to retain one significant HSP for each hit, and only Viridiplantae hits were retained. The order information of each hit was used to classify hits into self, close and distal groups. Subsequently, HGTpropor was applied to identify HGT candidates with a high proportion of hits from distal groups and a low proportion of hits from closely related groups. A distal cut-off of 0.9 was used to identify candidates, because our database contained only NT without the addition of sequences from closely related genomes of *Cuscuta*. Using this cut-off, 387, 27 and two HGT candidates greater than 300 bp from rosids, monocots and distal asterids were identified, respectively.

Furthermore, these candidates were evaluated in great detail for HGT validation. A large number of these candidates were identified as artefacts of vertically transmitted sequences, which was revealed by the addition of sequences from closely related genomes, in particular *Nicotiana*, *Solanum* and *Mimulus*. Remaining candidates were analysed with BLASTN and BLAST searches against NR and Phytozome[59]. Online BLAST searches against the NR database used a word size of 16 rather than the default value of 28. Otherwise, the BLAST hits can occasionally be misleading (BLASTp searches of many vertically transmitted *Cuscuta* sequences show top hits from closely related asterid species, yet BLASTn default searches show that top hits are from monocots; word size adjustment in BLASTn allowed the discovery of additional close sequences. This represents how many of the monocot-derived HGT candidates were found to be artefacts, one of which finally proved to be a HGT from a legume donor rather than from a monocot (onion). We found no HGTs from grasses in *Cuscuta*, a result corroborated by the protein-coding sequences.

In terms of rosid-derived HGT candidates, when BLAST hit distributions against NR are similar to those from Phytozome (v.10.1)[59] and when separate BLAST against the Sol Genomics Network[60] yields far lower bitscores than from distal rosids hits from NR (the empirical cut-off for close hit bitscore is 80% lower than the distal bitscore), the HGT candidate is likely to be validated as it also has support from HGTpropor. Next, these candidates were validated with BLASTx to ensure a BLAST output similar to BLASTn. The use of BLASTn and BLASTx can further classify the candidates into protein-coding or non-coding candidates. If the top BLASTn hits of a HGT candidate encode proteins, and when further BLASTx or BLASTp yields hits of similar sequences, these HGT candidates were classified as 'protein-coding genes'. The HGT sequences were then aligned against donor sequences to examine whether they contained premature stop codons. HGT sequences that contained premature stop codons while their donor encoded a functional protein were classified as 'HGT-derived pseudogenes'. On the other hand, if BLASTn hits of HGT sequences were repeats or non-protein-coding features and BLASTx did not yield significant hits, these candidates were likely to be considered as repeats. The non-coding nature of the HGT sequences was further validated when a Phytozome BLASTn against the genome of the top blast hit species yielded the best BLAST hit annotated as 'repeats' or non-genes. As an additional check, all 203 high-confidence HGT regions were aligned to an independently generated *C. campestris* genome[12]; 193 or 203 (95%) of the sequences were also found in that genome with alignments of at least 250 bp and 95% or greater sequence identity (Supplementary Fig. 1). The results of the final list of HGT sequences, and whether they encode proteins or repeats, are shown in Supplementary Table 23.

Identifying small RNA enrichment of *C. campestris* HGT genes. To find HGT genes enriched with small RNAs, we utilized de novo annotated *C. campestris* small RNA loci from a previously published study[15]. We also obtained the expression and alignment data of these loci for parasite stem, haustoria and host *A. thaliana* stem libraries (two biological replicates per sample) from the same study. Only *C. campestris* small RNA loci predominantly expressing 20–24-nt RNAs (13,809 loci) were utilized for all subsequent analyses. strucVis (v.0.3; https://github.com/MikeAxtell/strucVis) was used to visualize small RNA sequencing read coverage along the predicted secondary structures of annotated *C. campestris MIRNAs*. For generation of expression heat maps, raw read count for each of the small RNA loci was normalized based on the total number of processed reads for each of the libraries.

Overlap of *C. campestris* small RNA loci against 200 protein-coding HGT genes annotated in the *C. campestris* v.0.1 genome (Supplementary Table 13) was determined using bedtools v.2.22.0 (ref. [61]) with default settings. Ten independent cohorts of random *C. campestris v0.1* genes equal to the number of HGT genes (that is, 200 genes) were also generated for testing small RNA enrichment of random genes. For each of these cohorts, *C. campestris v0.1* genes overlapping annotated small RNA loci were identified using bedtools as described above.

C. campestris v0.1 parent locus coordinates of the HGT genes was used as reference loci for quantifying the total number and predominant size of small RNAs per HGT gene, using ShortStack v.3.8.4 (ref. [62]) with previously reported alignment data[15]. The ShortStack commands were: 'ShortStack –locifile <HGT_loci.txt>–bamfile <alignments.bam>–genomefile <merged_ath167_cpv0.1_genome.fasta>'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Publicly available data sources are as given in the Methods section of the manuscript. Web links for publicly available datasets are indicated in Supplementary Table 22. *C. campestris* genome assembly and annotations are available from http://ppgp.huck.psu.edu/cuscuta.html. The raw sequence reads for the eight *Cuscuta* taxa sampled in this study (*Cuscuta* species RNA sequencing datasets), *C. campestris* HGT sequences, all multiple sequence alignments and HGT tree files, as well as the supporting trees and alignments for selective constraint analyses (*C. campestris* HGT gene sequences, alignments and phylogenies), are given as supporting data at http://ppgp.huck.psu.edu/data/Cuscuta_HGT_Manuscript_Data/. All HGT sequences extracted from these assemblies are included as supporting data in the posted multiple sequence alignments and as described below. The raw data for Fig. 1a are in Supplementary Table 2 (column C); Fig. 1b in Supplementary Table 3; Fig. 1d in Supplementary Figs. 11 and 12; Fig. 2a,c,d in Supplementary Table 11; Fig. 2b on http://ppgp.huck.psu.edu/data/Cuscuta_HGT_Manuscript_Data/; Fig. 3a,b in Supplementary Tables 13 and 14; Fig. 3c in Supplementary Table 2; Fig. 3g on http://ppgp.huck.psu.edu/data/Cuscuta_HGT_Manuscript_Data/; Fig. 4 in Supplementary Table 15; and Fig. 5 in Supplementary Tables 2, 13 and 14.

Code availability

The customized code and pipeline associated with data analysis are available from https://github.com/dePamphilis/Cuscuta_HGT_ms_code.

Received: 7 September 2018; Accepted: 23 May 2019;
Published online: 22 July 2019

References

1. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
2. Davis, C. C. & Xi, Z. Horizontal gene transfer in parasitic plants. *Curr. Opin. Plant Biol.* **26**, 14–19 (2015).
3. Gao, C. et al. Horizontal gene transfer in plants. *Funct. Integr. Genom.* **14**, 23–29 (2014).
4. Kado, T. & Innan, H. Horizontal gene transfer in five parasite plant species in Orobanchaceae. *Genome Biol. Evol.* **10**, 3196–3210 (2018).
5. Sun, T. et al. Two hAT transposon genes were transferred from Brassicaceae to broomrapes and are actively expressed in some recipients. *Sci. Rep.* **6**, 30192 (2016).
6. Xi, Z. et al. Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genom.* **13**, 227 (2012).
7. Yang, Z. et al. Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. *Proc. Natl Acad. Sci. USA* **113**, E7010–E7019 (2016).
8. Yoshida, S., Maruyama, S., Nozaki, H. & Shirasu, K. Horizontal gene transfer by the parasitic plant *Striga hermonthica*. *Science* **328**, 1128 (2010).
9. Zhang, D. et al. Root parasitic plant *Orobanche aegyptiaca* and shoot parasitic plant *Cuscuta australis* obtained Brassicaceae-specific strictosidine synthase-like genes by horizontal gene transfer. *BMC Plant Biol.* **14**, 19 (2014).
10. Zhang, Y. et al. Evolution of a horizontally acquired legume gene, albumin 1, in the parasitic plant *Phelipanche aegyptiaca* and related species. *BMC Evol. Biol.* **13**, 48 (2013).
11. Kim, G. & Westwood, J. H. Macromolecule exchange in *Cuscuta*-host plant interactions. *Curr. Opin. Plant Biol.* **26**, 20–25 (2015).
12. Vogel, A. et al. Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nat. Commun.* **9**, 2515 (2018).
13. Mower, J. P. et al. Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. *BMC Biol.* **8**, 150 (2010).
14. Kim, G., LeBlanc, M. L., Wafula, E., dePamphilis, C. W. & Westwood, J. H. Genomic-scale exchange of mRNA between a parasitic plant and its hosts. *Science* **345**, 808–811 (2014).
15. Shahid, S. et al. MicroRNAs from the parasitic plant *Cuscuta campestris* target host messenger RNAs. *Nature* **553**, 82–85 (2018).
16. Hepburn, N. J., Schmidt, D. W. & Mower, J. P. Loss of two introns from the *Magnolia tripetala* mitochondrial cox2 gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. *Mol. Biol. Evol.* **29**, 3111–3120 (2012).

17. Ranjan, A. et al. De novo assembly and characterization of the transcriptome of the parasitic weed dodder identifies genes associated with plant parasitism. *Plant Physiol.* **166**, 1186–1199 (2014).
18. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
19. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
20. Park, C. J., Caddell, D. F. & Ronald, P. C. Protein phosphorylation in plant immunity: insights into the regulation of pattern recognition receptor-mediated signaling. *Front. Plant Sci.* **3**, 177 (2012).
21. Cho, Y., Qiu, Y. L., Kuhlman, P. & Palmer, J. D. Explosive invasion of plant mitochondria by a group I intron. *Proc. Natl Acad. Sci. USA* **95**, 14244–14249 (1998).
22. Barkman, T. J. et al. Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evol. Biol.* **7**, 248 (2007).
23. Cho, Y., Mower, J. P., Qiu, Y. L. & Palmer, J. D. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *Proc. Natl Acad. Sci. USA* **101**, 17741–17746 (2004).
24. Li, F. et al. MicroRNA regulation of plant innate immune receptors. *Proc. Natl Acad. Sci. USA* **109**, 1790–1795 (2012).
25. Fei, Q., Xia, R. & Meyers, B. C. Phased, secondary, small interfering RNAs in posttranscriptional regulatory networks. *Plant Cell* **25**, 2400–2415 (2013).
26. Arikit, S. et al. An atlas of soybean small RNAs identifies phased siRNAs from hundreds of coding genes. *Plant Cell* **26**, 4584–4601 (2014).
27. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010).
28. ten Hove, C. A. et al. SCHIZORIZA encodes a nuclear factor regulating asymmetry of stem cell divisions in the *Arabidopsis* root. *Curr. Biol.* **20**, 452–457 (2010).
29. Mylona, P., Linstead, P., Martienssen, R. & Dolan, L. SCHIZORIZA controls an asymmetric cell division and restricts epidermal identity in the *Arabidopsis* root. *Development* **129**, 4327–4334 (2002).
30. Bustillo-Avendano, E. et al. Regulation of hormonal control, cell reprogramming, and patterning during *de novo* root organogenesis. *Plant Physiol.* **176**, 1709–1727 (2018).
31. Liu, M. J. et al. The complex jujube genome provides insights into fruit tree biology. *Nat. Commun.* **5**, 5315 (2014).
32. Becher, H. et al. Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome. *Plant J.* **80**, 823–833 (2014).
33. Costea, M., García, M. A., Baute, K. & Stefanović, S. Entangled evolutionary history of *Cuscuta pentagona* clade: a story involving hybridization and Darwin in the Galapagos. *Taxon* **64**, 1225–1242 (2015).
34. Costea, M., García, M. A. & Stefanović, S. A phylogenetically based infrageneric classification of the parasitic plant genus *Cuscuta* (Dodders, Convolvulaceae). *Syst. Bot.* **40**, 269–285 (2015).
35. McNeal, J. R., Kuehl, J. V., Boore, J. L. & dePamphilis, C. W. Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus Cuscuta. *BMC Plant Biol.* **7**, 57 (2007).
36. Westwood, J. H., Yoder, J. I., Timko, M. P. & dePamphilis, C. W. The evolution of parasitism in plants. *Trends Plant Sci.* **15**, 227–235 (2010).
37. Weiss-Schneeweiss, H., Greilhuber, J. & Schneeweiss, G. M. Genome size evolution in holoparasitic *Orobanche* (Orobanchaceae) and related genera. *Am. J. Bot.* **93**, 148–156 (2006).
38. Zonneveld, B. J. M. New record holders for maximum genome size in eudicots and monocots. *J. Bot.* **2010**, 527357 (2010).
39. Fultz, D., Choudury, S. G. & Slotkin, R. K. Silencing of active transposable elements in plants. *Curr. Opin. Plant Biol.* **27**, 67–76 (2015).
40. Sun, T. et al. Two hAT transposon genes were transferred from Brassicaceae to broomrapes and are actively expressed in some recipients. *Sci. Rep.* **6**, 30192 (2016).
41. Iseli, C., Jongeneel, C. V. & Bucher, P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 138–148 (1999).
42. Hirakawa, H. et al. Survey of genome sequences in a wild sweet potato, *Ipomoea trifida* (H. B. K.) G. Don. *DNA Res.* **22**, 171–179 (2015).
43. Denoeud, F. et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184 (2014).
44. Sollars, E. S. et al. Genome sequence and genetic diversity of European ash trees. *Nature* **541**, 212–216 (2017).
45. Wang, L. et al. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. *Genome Biol.* **15**, R39 (2014).
46. Huang, S. et al. Draft genome of the kiwifruit *Actinidia chinensis*. *Nat. Commun.* **4**, 2640 (2013).
47. Scaglione, D. et al. The genome sequence of the outbreeding globe artichoke constructed *de novo* incorporating a phase-aware low-pass sequencing strategy of F1 progeny. *Sci. Rep.* **6**, 19427 (2016).
48. Sierro, N. et al. The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, 3833 (2014).
49. Iorizzo, M. et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
50. Yang, Z. et al. Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Mol. Biol. Evol.* **32**, 767–790 (2015).
51. Matasci, N. et al. Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
52. Duvick, J. et al. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.* **36**, D959–D965 (2008).
53. Zhu, Q., Kosoy, M. & Dittmar, K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genom.* **15**, 717 (2014).
54. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* **48**, 4.11.1–4.11.39 (2014).
55. Felsenstein, J. *PHYLIP (Phylogenetic Inference Package), Version 3.6* Vol. 5 (Department of Genome Sciences, Univ. of Washington, 2005).
56. Belfort, M. & Bonocora, R. P. Homing endonucleases: from genetic anomalies to programmable genomic clippers. *Methods Mol. Biol.* **1123**, 1–26 (2014).
57. Edgell, D. R. Selfish DNA: homing endonucleases find a home. *Curr. Biol.* **19**, R115–R117 (2009).
58. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52–D57 (2011).
59. Goodstein, D. M. et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186 (2012).
60. Fernandez-Pozo, N. et al. The Sol Genomics Network (SGN)–from genotype to phenotype to breeding. *Nucleic Acids Res.* **43**, D1036–D1041 (2015).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Johnson, N. R., Yeoh, J. M., Coruh, C. & Axtell, M. J. Improved placement of multi-mapping small RNAs. *G3 (Bethesda)* **6**, 2103–2111 (2016).

Acknowledgements

Sequence data are archived at National Center for Biotechnology Information BioProject ID SRP001053, and at http://ppgp.huck.psu.edu. This research was supported by award No. IOS-1238057 to J.H.W. and C.W.deP. from the NSF Plant Genome Research Program; No. 2018-05102 to M.J.A., J.H.W. and C.W.deP. from the United States Department of Agriculture, with additional support to Z.Y. from the Plant Biology and Biology Department graduate programmes at Penn State; and by the National Institute of Food and Agriculture Project (No. 131997) to J.H.W. The authors thank I. Ko for help with PCR experiments and E. Bellis, Y. Zheng and three anonymous reviewers for helpful comments and suggestions.

Author contributions

C.W.deP., J.H.W. and Z.Y. conceived this project. Z.Y. performed major analyses, with additional analyses by E.K.W., S.S., G.K., J.R.M., P.R.T, W.-b.Y. and T.N.P. G.K. and P.E.R. performed experiments. E.A.K. and H.Z. performed RT–PCR. J.R.M. generated transcriptome samples for ancestral inference. M.J.A. and S.S. contributed small RNA analyses. N.S.A. supervised the statistical analyses and conception of HGTpropor. Z.Y. and C.W.deP. wrote the manuscript with contributions from E.K.W., J.H.W., P.E.R., S.S. and M.J.A. All authors read and approved the final manuscript.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41477-019-0458-0.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.H.W. or C.W.d.

Peer review information: *Nature Plants* thanks David Hannapel, Fay-Wei Li, Jianqiang Wu and the other anonymous reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): NPLANTS-18085406B

Last updated by author(s): May 15, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Datasets were obtained from NCBI GeneBank, Phytozome Plant Genomic Resource, and The 1000 Plants (oneKP or 1KP) Initiative. A detailed description of the datasets and their sources is provided in the Materials and Methods section the manuscript.
Data analysis	The customized analysis pipeline code is committed on the GitHub repository, https://github.com/dePamphilis/Cuscuta_HGT_ms_code including a step by step description of how to the data were analyzed. Data processing for phylogenetic trees followed Yang et al. (2016) as available in PlantTribes (version 1.0.2) (https://github.com/dePamphilis/PlantTribes/). Here, we briefly describe the processing flow. Coding sequences (CDS) for each orthogroup were translated and the inferred protein sequences were aligned using MAFFT (version 7.407). Additional coding sequences (CDS) obtained from transcriptomes or added genomes were classified into the corresponding orthogroup and forced onto the translated alignments to obtain protein-based DNA sequence alignments. Resulting DNA alignments were trimmed using trimAl (version 1.4) to remove sites with gaps in more than 90% of the sequences. Sequences in the trimmed alignments that had less than 50% base alignment coverage were removed and the alignment process repeated. Finally, trimmed and filtered FASTA DNA alignments were converted to PHYLIP format and phylogenetic trees were inferred using RAxML (version 7.2.7). InterProscan (v5.25.64.0) was used to annotate Cuscuta gene annotations, orthogroups scaffolds, and HGT orthogroups. PBJelly was used to fill the gap of the C. campestris genome scaffolds. MAKER-P pipeline (version 2.31.8) was used to predict the gene models with training sets including curated plant proteins from SwissProt and C. campestris mRNA-seq data from multiple tissue types. Two sets of orthogroups were created using OrthoMCL (version 2.0.9) and OrthoFinder (version 1.1.2) software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Publicly available data sources are as given in the Materials and Methods section of the manuscript. Web links for publicly available datasets are indicated in Table S22. C. campestris genome assembly and annotations are available from http://ppgp.huck.psu.edu/cuscuta.html. The raw sequence reads for the eight Cuscuta taxa sampled in this study (Cuscuta species RNA-Seq datasets), C. campestris HGT sequences, all multiple sequence alignments and HGT tree files, as well as the supporting trees and alignments for selective constraint analyses (Cuscuta campestris HGT gene sequences, alignments, and phylogenies) are given as supporting data at http://ppgp.huck.psu.edu/cuscuta.html.The customized code and pipeline associated with data analysis are available from http://ppgp.huck.psu.edu/data/Cuscuta_HGT_Manuscript_Data/. Unpublished transcriptome assemblies of seven additional Cuscuta species (C. pentagona, C. harperi, C. obtusiflora, C. rostrata, C. chilensis, C. europaea, C. exaltata) were provided by coauthor Dr. Joel R. McNeal of Kennesaw State University. All HGT sequences extracted from these assemblies are included as supporting data in the posted multiple sequence alignments and as described below. The following figures have associated raw data described in the Materials and Methods section of the manuscript and in the supplementary tables and figures: The raw data for Fig. 1a is in Table S2 (column C), Fig. 1b in Table S3, Fig. 1d in Fig. S11 and Fig. S12; Fig. 2a, 2c, 2d in Table S11, Fig. 2b on github; Fig. 3a, 3b in Table S13 and Table S14; Fig. S3c in table S2, Fig. S3g on github; Fig. 4 in Table S15; Fig. 5 in Table S2, Table S13, and S14.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for each test and design has been explicitly indicated in the respective table or figure legend. Gene expressions data and statistical tests of functional enrichment for HGT categories use sample numbers that have been expressly stated with the respective analyses.
Data exclusions	No data were excluded from the analysis.
Replication	The expression of genes used in Fig. 2b, Fig. 3g, and Fig. S13-S15 was calculated from a published transcriptome study of at least two replicates taken from a published study as indicated in the Materials and Methods. The expression of small RNA loci in Fig. S16 and Fig. S17 also involves small RNA sequencing of two replicates.
Randomization	Randomization is not applicable in our study. Because no new data were generated. We used the expression data directly from a previously published paper by Ranjan et al. 2014 and the small RNA sequencing data from Shahid et al. 2018.
Blinding	Blinding is not applicable in our study because it does not involve subjects which receive different treatments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging