Massive Mitochondrial Gene Transfer in a Parasitic Flowering Plant Clade

Zhenxiang Xi1,3, Yuguo Wang2,3,*, Robert K. Bradley4,5, M. Sugumaran6, Christopher J. Marx1,7, Joshua S. Rest8, Charles C. Davis1*

1 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts, United States of America, 2 Ministry of Education Key Laboratory for Biodiversity Science and Ecological Evolution, Institute of Biodiversity Science, Fudan University, Shanghai, China, 3 Department of Ecology and Evolutionary Biology, School of Life Science, Fudan University, Shanghai, China, 4 Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 5 Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 6 Rimba Ilmu Botanic Garden, Institute of Biological Sciences, University of Malaya, Kuala Lumpur, Malaysia, 7 Faculty of Arts and Sciences Center for Systems Biology, Harvard University, Cambridge, Massachusetts, United States of America, 8 Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York, United States of America

Abstract
Recent studies have suggested that plant genomes have undergone potentially rampant horizontal gene transfer (HGT), especially in the mitochondrial genome. Parasitic plants have provided the strongest evidence of HGT, which appears to be facilitated by the intimate physical association between the parasites and their hosts. A recent phylogenomic study demonstrated that in the holoparasite Rafflesia cantleyi (Rafflesiaceae), whose close relatives possess the world’s largest flowers, about 2.1% of nuclear gene transcripts were likely acquired from its obligate host. Here, we used next-generation sequencing to obtain the 38 protein-coding and ribosomal RNA genes common to the mitochondrial genomes of angiosperms from R. cantleyi and five additional species, including two of its closest relatives and two host species. Strikingly, our phylogenetic analyses conservatively indicate that 24%–41% of these gene sequences show evidence of HGT in Rafflesiaceae, depending on the species. Most of these transgenic sequences possess intact reading frames and are actively transcribed, indicating that they are potentially functional. Additionally, some of these transgenes maintain synteny with their donor and recipient lineages, suggesting that native genes have likely been displaced via homologous recombination. Our study is the first to comprehensively assess the magnitude of HGT in plants involving a genome (i.e., mitochondria) and a species interaction (i.e., parasitism) where it has been hypothesized to be potentially rampant. Our results establish for the first time that, although the magnitude of HGT involving nuclear genes is appreciable in these parasitic plants, HGT involving mitochondrial genes is substantially higher. This may represent a more general pattern for other parasitic plant clades and perhaps more broadly for angiosperms.

Introduction
Recent studies have suggested that plant genomes have undergone potentially rampant horizontal gene transfer (HGT) [1,2], especially in the mitochondrial genome [3–7]. Parasitic plants have provided the strongest evidence of HGT [8–12], which appears to be facilitated by the intimate physical association between the parasites and their hosts [8,10,12–15]. One parasitic plant clade that appears to be prone to HGT is Rafflesiaceae sensu stricto, which belong to the order Malpighiales [8,16–18] and whose members possess the largest flowers in the world. Rafflesiaceae are endophytic holoparasites, which lack leaves and stems. This family includes the genera Rafflesia (~28 species), Rhizanthes (four species), and Sapria (three species), and provides one of the best opportunities to investigate HGT in plants because (i) the parasites have a very narrow host specialization range on members of the grapevine family (Tetrastigma spp., Vitaceae), (ii) complete genome sequences, including fully annotated mitochondrial and plastid genomes, are available for close relatives of the parasites (Ricinus communis, Euphorbiaceae) [19,20] and their hosts (Vitis vinifera, Vitaceae) [21–23], and (iii) the hosts and parasites are separated by at least 115 million years of evolution (Figure 1A) [24–27]. These factors make it easier to distinguish transgenes from native genes in Rafflesiaceae using phylogenomic tools.

A recent phylogenomic study demonstrated that in Rafflesia cantleyi, about 2.1% of nuclear gene transcripts were likely acquired from its obligate host [28]. This study, however, did not include a thorough investigation of the mitochondrial genome. Here, we comprehensively sequenced 38 mitochondrial genes from R. cantleyi and five additional species, including two of its closest relatives and two host species. Our results reveal an extraordinarily high degree of HGT in the mitochondrial genome of Rafflesiaceae...
Author Summary

Recent studies have suggested that plant genomes have undergone potentially rampant horizontal gene transfer (HGT), especially in the mitochondrial genome. Here, using phylogenomic approaches, we demonstrate that as much as ~40% of the mitochondrial genes in the parasitic plant species Rafflesiaceae are acquired from their hosts via HGT. These transgenes are likely functional in their recipient species and in some cases appear to have displaced native copies in the same genomic location. These results establish for the first time that, although the magnitude of HGT involving nuclear genes is appreciable in parasitic plants, HGT involving mitochondrial genes is substantially higher.

Results/Discussion

The mitochondrial provenance of our sequenced genes

We used next-generation sequencing to comprehensively sequence the mitochondrial genomes of three species that span the crown node of Rafflesiaceae: Rafflesia cantleyi, Rafflesia tuan-mudae, and Spathia himalayana (Figure 1A, see also Table S1). We then extracted the 38 mitochondrial genes from our de novo assembled contigs that ranged in size from 2 to 54 kilobases (kb). These 38 protein-coding and ribosomal RNA genes are present in the mitochondrial genomes of both Rcinus and Vitis, and are also common to most angiosperms [29]. We included 35, 33, and 59 gene sequences from R. cantleyi, R. tuan-mudae and S. himalayana, respectively, for further analyses. While repetitive sequences made assembly of the entire chromosome impractical, high sequence coverage (Table S1) ensured that we have sequenced all coding regions in these mitochondrial genomes.

Several lines of evidence suggest that all gene sequences we assembled here are localized to the mitochondrial genome of Rafflesiaceae. First, the genome libraries for two of our three Rafflesiaceae species (i.e., R. cantleyi and S. himalayana) were prepared from fresh tissue using sucrose gradient centrifugation, which are enriched for plant organelles [30]. Since the plastid genome has apparently been lost in Rafflesiaceae [31], our libraries are heavily enriched for mitochondria. Second, plastid, mitochondrial and nuclear genes in plant cells differ widely in copy number: plastid genes are generally present in hundreds to thousands of copies per cell, mitochondrial genes in tens to hundreds of copies per cell, while nuclear genes are usually present in only two copies per cell [7,32]. To investigate if gene sequences assembled here have copy numbers that correspond with a mitochondrial localization, we compared gene copy number here to 1,305 genes previously determined to be localized to the nuclear genome of R. cantleyi [28] and R. tuan-mudae [18]. Our results demonstrate that copy numbers for all putative mitochondrial gene sequences in Rafflesiaceae are one to two orders of magnitude greater than for nuclear genes (Table S2), with means of 155-, 68-, and 160-fold greater for R. cantleyi, R. tuan-mudae, and S. himalayana, respectively (p-value<2.2×10^{-16}, Welch’s t test).

These copy numbers are consistent with a mitochondrial localization, but not high enough to suggest localization in, or the existence of, a plastid genome in Rafflesiaceae. Third, when comparing assembled gene sequences from R. cantleyi with our previously published complementary DNA (cDNA) library [28], we identified cytosine-to-uracil (C-to-U) RNA editing in seven genes (i.e., atp1, atp4, atp6, cox2, nad1, nad4, and nad5), which is a common characteristic of mitochondrial genes [33]. These results collectively indicate that these gene sequences are most likely localized to the mitochondrial genome of Rafflesiaceae, although complete assembly of these mitochondrial genomes will be required to definitively confirm our results.

Extraordinarily high, and variable, rates of HGT in the mitochondrial genome of Rafflesiaceae

To estimate the magnitude of HGT in the mitochondrial genome of Rafflesiaceae, we also sequenced the same 38 mitochondrial genes from three species of Vitaceae: Tetrastigma eruciatum, which is the host of S. himalayana [34], Tetrastigma rafigenale, which is the host of R. cantleyi and R. tuan-mudae [34,35], and Lea guineensis (Figure 1A, see also Table S1). The latter represents the earliest diverging lineage of Vitaceae [36,37], which allows us to determine if putative transgenic sequences from Rafflesiaceae are phylogenetically nested within the host clade Vitaceae.

Our newly sequenced mitochondrial gene sequences from the six species were then analyzed using maximum likelihood (ML) with homologous sequences from 27 other seed plants whose mitochondrial genomes have been sequenced and fully annotated (Figure 1A, see also Table S3; Arabidopsis thaliana [Brassicaceae], Beta vulgaris [Amaranthaceae], Bona hygrotestica [Gesneriaceae], Brassica napus [Brassicaceae], Carica papaya [Caricaceae], Citrullus lanatus [Cucurbitaceae], Cucumis sativus [Cucurbitaceae], Cucurbita pepo [Cucurbitaceae], Cycas taitungensis [Cycadaceae], Danucus carota [Apiaceae], Lotus japonicus [Fabaceae], Malus domestica [Rosaceae], Milletilia pinnata [Fabaceae], Minusulus guttatus [Phrymaceae], Nicotiana tabacum [Solanaceae], Orzya sativa [Poaceae], Phoenix dactylifera [Areaceae], Raphanus sativus [Brassicaceae], Ricinus communis, Silene latifolia [Caryophyllaceae], Solanum bicolor [Poaceae], Spindrella polycephala [Araceae], Tripalumum dactyloides [Poaceae], Triticum aestivum [Poaceae], Vigna radiata [Fabaceae], Vitis vinifera, and Zea mays [Poaceae]). These reference species represent a broad sampling of most major flowering plant clades [38]. Each Rafflesiaceae gene sequence was placed into one of three categories – i.e., VGT, HGT, or unassigned – on the basis of its phylogenetic position and ML bootstrap percentage (BP) support following Xi et al. [28]. We applied two BP thresholds to categorize each gene sequence. Our more conservative estimate applied a 70 BP threshold; this BP threshold has been shown to correspond to a very high probability that the clade is real [39]. Here, gene sequences whose placements were consistent with accepted species’ relationships (i.e., Rafflesiaceae gene sequences were sister to their closest relative Ricinus with ≳ 70 BP; [17,18]) were scored as VGT; HGT was inferred when gene sequences were placed elsewhere with ≳ 70 BP; and gene sequences with < 70 BP were left unassigned. To explore if our estimates of HGT were sensitive to our thresholds, we also categorized these gene sequences by applying a less conservative threshold using ≳ 50 BP.

Our phylogenetic analyses of the 38 mitochondrial genes indicated that, for the 30 autotrophic species included here (i.e., 27 reference species, two Tetrastigma species, and Lea; Figure 1A), phylogenetic placements largely agreed with accepted relationships between families [38,40] using both the 70 and 50 BP
Mitochondrial Gene Transfer in Rafflesiaaceae

A

Rafflesia cantleyi
Rafflesia tuan-mudae
Sapria himalayana
Ricinus communis
Brassica napus
Raphanus sativus
Arabidopsis thaliana
Carica papaya
Cucumis sativus
Cucurbita pepo
Citrus lanatus
Malus domestica
Lotus japonicus
Vigna radiata
Milletia pinnata
Tetrastigma cruciatum
Tetrastigma rafflesiae
Vitis vinifera
Leea guineensis
Boea hygrometrica
Mimulus guttatus
Nicotiana tabacum
Daucus carota
Beta vulgaris
Silene latifolia
Tripsacum dactyloides
Zea mays
Sorghum bicolor
Oryza sativa
Triticum aestivum
Phoenix dactylifera
Spirodela polyrhiza
Cycas taitungensis

B

Rafflesia cantleyi
Rafflesia tuan-mudae
Sapria himalayana
Ricinus communis

stem group Tetrastigma
crown group Tetrastigma

100 80 60 40 20 0 Ma

69.5 65.3
Mitochondrial Gene Transfer in Rafflesiaceae

Figure 1. Phylogenetic relationships and divergence times of the three Rafflesiaceae species and two Tetrastigma species included in this study. Phylogenetic relationships (A) and divergence times (B). Holoparasitic Rafflesiaceae (red) is a member of the order Malpighiales, and its obligate host Tetrastigma (blue) is a member of the Vitaceae family. The approximate divergence time between the parasite and host clade is 115 Ma [24–27]. Mitochondrial genome sequences generated in this study are marked with asterisks, and the node age error bars (95% highest posterior density intervals) are shown in green. The accepted phylogenetic relationships are based on APG III [38] and Qiu et al. [40], and the divergence times of Rafflesiaceae and Tetrastigma are based on Bendiksby et al. [48] and Chen et al. [49], respectively.

doi:10.1371/journal.pgen.1003265.g001

thresholds (Figures S1A and S2). The only three exceptions were for atp1 where Brassicaceae (i.e., Arabidopsis+Brassica+Raphanus) was placed sister to the asterids (i.e., Boese+Mimulus+Nicotiana) with 81 BP (a similar topology was also identified by Nickrent et al. [10]); for atp4 where Brassicaceae was placed sister to Fabaceae (i.e., Lotus+Melittia+Vigna) with 93 BP; and for cox1 where Brassicaceae was placed sister to Caryophyllales (i.e., Beta+Silene) with 87 BP (Figure S1A). These results indicate that applying both the 70 and 50 BP thresholds yields very low false positive estimates of HGT in these autotrophic species, for which we expect little or no HGT to occur.

In contrast, in the three holoparasitic Rafflesiaceae species, 11 mitochondrial genes demonstrated evidence for one or more cases of HGT using our more conservative 70 BP threshold: of the 21 gene sequences with ≥70 BP in R. cantleyi, five gene sequences (24%) showed evidence of HGT, 5 of 19 (26%) in R. tuan-mudae, and 11 of 27 (41%) in S. himalayana. Furthermore, vertical placements of these putative transgenic sequences were rejected in 18 of 21 cases using the approximately unbiased (AU) test (Table 1). For the less conservative 50 BP threshold, the number of mitochondrial genes that showed evidence of HGT increased to 16; however, the relative frequencies of HGT are nearly identical with those above: 29% in R. cantleyi (7 of 24), 32% in R. tuan-mudae (7 of 22), and 47% in S. himalayana (16 of 34). This indicates that our less conservative threshold does not increase false positive rates. Thus, given the consistency of our estimates of HGT using both thresholds, we treat these transgenes collectively in the discussion below unless otherwise indicated. Two additional findings support the reliability of our HGT inferences: first, the phylogenetic placements of these transgenic sequences were not obviously biased by C-to-U RNA editing (see Figure S1B for phylogenoms with RNA editing sites excluded from our alignments); second, seven of our large assembled contigs contained both transgenes and native genes (Figure 2), indicating that these transgenes were clearly integrated into the mitochondrial genome of Rafflesiaceae. Therefore, rates of mitochondrial HGT in Rafflesiaceae appear to be extraordinarily high, and well above the false positive rates established from the 30 autotrophic species included here.

Of the 11 mitochondrial genes that showed evidence of HGT using our more conservative threshold, four (i.e., cob, cox5, rpl5, and rps7) maintained both horizontally and vertically transferred homologs, and seven included only transgenic sequences (i.e., atp4, cox1, cox2, rps1, rps7, rps13, and sdh3) (Table 1). An additional five mitochondrial genes showed evidence of HGT using our less conservative threshold, one of them (atp9) maintained both horizontally and vertically transferred homologs, and four included only transgenic sequences (i.e., atp1, rpl2, rps14, and sdh4) (Table 1). Of those genes that included only transgenic copies in Rafflesiaceae all had homologs present in the mitochondrial genome of Ricinus, which suggests that they were likely present as native copies ancestrally in Rafflesiaceae and were subsequently displaced by transgenic homologs. One example is illustrated by our assembled contig containing the genes nad5 exons A and B and rps7 (Figure 2A). In Rafflesiaceae, nad5 exon B was identified as a native sequence (84 BP) in our phylogenetic analyses, while rps7 was identified as a transgene (96 BP). The phylogenetic placement of nad5 exon A within Vitaceae is also consistent with HGT, but support for this placement is <50 BP (Figure 2A, see also Figure S1C). However, the synteny of nad5 exons A and B is conserved among Ricinus, Vitis, and all three Rafflesiaceae species, suggesting that the native copy of nad5 exon A in Rafflesiaceae may have been displaced by a horizontally transferred DNA fragment via homologous recombination [41]. This hypothesis is further supported by the fact that nad5 exon A is immediately adjacent to the well-placed transgene rps7 in Rafflesiaceae, which exactly matches the synteny of Vitis but not Ricinus. To better locate the recombination breakpoint, we analyzed the intron region between nad5 exons A and B. This ~1-kb region is highly conserved across angiosperms and can be easily aligned for phylogenetic analysis. We found that nad5 intron A/B was clearly identified as a native sequence (82 BP; Figure 2A, see also Figure S1C), therefore, the breakpoint is likely very close to the junction of nad5 exon A and intron A/B. Although the integration of foreign DNA via homologous recombination is common in bacteria [41], reports of this phenomenon are rare for plants (e.g., atp1 gene [42] and rps11 gene [4,1]). Such direct homologous recombination, which is likely facilitated by the intimate physical association between Rafflesiaceae and their hosts [14,15] combined with the frequent fusion of plant mitochondria [1,43–45], may obviate the need to invoke a transposable element, bacterium, or virus for catalyzing the insertion of a DNA fragment from donor to recipient in plants.

Additionally, in the mitochondrial genome of S. himalayana, we found evidence of HGT involving 14 genes that were potentially of plastid origin using our more conservative threshold, only one of which was also identified in R. tuan-mudae (Table 1, see also Figure S1D). Thirteen of these genes support the conclusion that they were acquired via host-to-parasite HGT, because in each case S. himalayana is placed sister to, or nested within, Vitaceae with ≥70 BP. Only atp4 from S. himalayana and R. tuan-mudae were placed elsewhere phylogenetically, sister to Daucus with 94 BP. Furthermore, for six of these 14 genes (i.e., atpB, atpD, psaB, psaA, psbC, and psbD), the transgenic sequences from S. himalayana were sister to the mitochondrial and not the plastid homologs from Vitaceae (Figure S1D). These six genes, plus three additional plastid genes (i.e., atpA, ndhB, and rbcL), have been shown to be incorporated into the mitochondrial genome of Vitis [23]. Together, these results suggest that the majority of these plastid genes were likely acquired via HGT from the host mitochondrial genome, instead of from its plastid genome.

Finally, seven of our assembled contigs demonstrated that synteny was maintained between transgenes from Rafflesiaceae and genes from the close relative of their hosts, Vitis, whose mitochondrial and plastid genomes were both fully annotated (i.e., rps7+nad5 exon A, psbC+psbD, sdh3+psaL+psaF+cob+cox1, rpoC1+rcpC2, and sdh4+cox3; Figure 2 and Figure S3). This, combined with our finding that two transgenes bear introns (i.e., cox1 and rpl2), firmly supports our previous suggestion that transgenes in Rafflesiaceae are likely transferred as larger DNA fragments versus shorter mRNAs [28].
Table 1. Horizontally transferred (HGT) sequences identified in the mitochondrial genomes of *Rafflesia cantleyi*, *Rafflesia tuan-mudae*, and *Sapria himalayana* with associated statistics.

<table>
<thead>
<tr>
<th>Gene</th>
<th>species</th>
<th>Gene identity</th>
<th>HGT donor lineage</th>
<th>BP</th>
<th>AU test</th>
<th>Gene length (base pairs)</th>
<th>Reading frame</th>
<th>RPKM</th>
<th>C-to-U RNA editing site</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mitochondrial genes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>atp1</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>54</td>
<td>NS</td>
<td>1520</td>
<td>I</td>
<td>2370</td>
<td>1039</td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>54</td>
<td>NS</td>
<td>1520</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>54</td>
<td>NS</td>
<td>1514</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atp4</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Cucurbitaceae</td>
<td>83</td>
<td>NS</td>
<td>526</td>
<td>I</td>
<td>6560</td>
<td>168, 337</td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Cucurbitaceae</td>
<td>83</td>
<td>NS</td>
<td>526</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Cucurbitaceae</td>
<td>83</td>
<td>NS</td>
<td>520</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atp9</td>
<td><em>R. cantleyi</em></td>
<td>VGT</td>
<td>–</td>
<td>94</td>
<td>–</td>
<td>204</td>
<td>I</td>
<td>1225</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>VGT</td>
<td>–</td>
<td>94</td>
<td>–</td>
<td>204</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>VGT</td>
<td>–</td>
<td>94</td>
<td>–</td>
<td>204</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>65</td>
<td>NS</td>
<td>195</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cob</td>
<td><em>R. cantleyi</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>1166</td>
<td>I</td>
<td>941</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>1166</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>1166</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>99</td>
<td>0.0004</td>
<td>1179</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox1</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Brassicaceae+Caryophyllales</td>
<td>71</td>
<td>0.0360</td>
<td>1542</td>
<td>I</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Brassicaceae+Caryophyllales</td>
<td>71</td>
<td>0.0360</td>
<td>1542</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Brassicaceae+Caryophyllales</td>
<td>71</td>
<td>0.0360</td>
<td>1542</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>100</td>
<td>&lt;0.0001</td>
<td>1461</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox2</td>
<td><em>R. cantleyi</em></td>
<td>–</td>
<td>–</td>
<td>&lt;50</td>
<td>–</td>
<td>770</td>
<td>I</td>
<td>2183</td>
<td>14, 61, 153, 233</td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>–</td>
<td>–</td>
<td>&lt;50</td>
<td>–</td>
<td>770</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>–</td>
<td>–</td>
<td>&lt;50</td>
<td>–</td>
<td>771</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>96</td>
<td>&lt;0.0001</td>
<td>246</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cox3</td>
<td><em>R. cantleyi</em></td>
<td>VGT</td>
<td>–</td>
<td>78</td>
<td>–</td>
<td>798</td>
<td>I</td>
<td>1810</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>VGT</td>
<td>–</td>
<td>78</td>
<td>–</td>
<td>798</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>85</td>
<td>0.0063</td>
<td>337</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rpl2</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>51</td>
<td>NS</td>
<td>408</td>
<td>I</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>51</td>
<td>NS</td>
<td>408</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>69</td>
<td>NS</td>
<td>963</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>85</td>
<td>0.0030</td>
<td>543</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps1</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>100</td>
<td>&lt;0.0001</td>
<td>342</td>
<td>I</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>100</td>
<td>&lt;0.0001</td>
<td>342</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>100</td>
<td>&lt;0.0001</td>
<td>494</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps4</td>
<td><em>R. cantleyi</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>838</td>
<td>I</td>
<td>2358</td>
<td>50, 61</td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>838</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>VGT</td>
<td>–</td>
<td>96</td>
<td>–</td>
<td>1040</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>81</td>
<td>0.0075</td>
<td>143</td>
<td>I</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>81</td>
<td>0.0075</td>
<td>143</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps7</td>
<td><em>R. cantleyi</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>96</td>
<td>0.0060</td>
<td>328</td>
<td>I</td>
<td>850</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>R. tuan-mudae</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>96</td>
<td>0.0060</td>
<td>328</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>96</td>
<td>0.0060</td>
<td>186</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Vitaceae</td>
<td>86</td>
<td>0.0033</td>
<td>264</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rps14</td>
<td><em>S. himalayana</em></td>
<td>HGT</td>
<td>Tetrastigma</td>
<td>68</td>
<td>0.0104</td>
<td>303</td>
<td>I</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Transgenes are expressed in Rafflesiaceae

Most previously reported mitochondrial transgenes in plants appear to be non-functional, i.e., they have been shown to be either introns (e.g., [3,6,8,46,47]) or pseudogenes (e.g., [7,9]). However, among all mitochondrial transgenes identified here, six of seven sequences in *R. cantleyi* and *R. tuan-mudae*, and 13 of 16 sequences in *S. himalayana* maintain their reading frames (Table 1). To further understand if these transgenes are expressed, we re-examined the recently published transcriptome of *R. cantleyi* [28] to quantify gene expression levels of these mitochondrial genes. Our results indicate that all transgenes in *R. cantleyi* show evidence of expression (Table 1). Furthermore, although native genes in *R. cantleyi* show higher overall levels of gene expression than transgenes (Figure S1), this difference is not significant (p-value = 0.19, Welch’s t test). Thus, transgenes are actively transcribed in this species, suggesting that they have functional promoters and likely play a role in cellular function.

Timing of HGT events

Our broad phylogenomic assessment of mitochondrial genome provides a unique opportunity to determine if HGT we identified in Rafflesiaceae is relatively ancient or more recent. For five genes that show evidence of HGT (i.e., *atp1*, *atp4*, *cox1*, *rps7*, and *atp6*; Figure S1A and S1D), it is most parsimonious to infer that they each result from an ancient HGT event. The more ancient origin is supported by the fact that transgenic sequences from *Rafflesia* and *Sapria* form a clade. Furthermore, we found that some of these transgenes maintained synteny between *Rafflesia* and *Sapria* (e.g., *atp1* and *rps7*; Figure 2A and 2B). Therefore, these gene transfers appear to have been relatively ancient and likely occurred after the origin of stem group Rafflesiaceae (95% highest posterior density [HPD] interval of 83.1–109.5 Ma; [17,48]) and before the origin of crown group Rafflesiaceae (69.5–95.9 Ma; [17,48]). Both of these estimated clade ages, accounting for 95% HPD intervals, are outside the age of stem group *Tetrasigma* (36.4–65.3 Ma; [49]), and well outside the age of crown group *Tetrasigma* (25.7–49.3 Ma; [49]) (Figure 1B).

This raises the distinct possibility that Rafflesiaceae has had former host associations with other plant lineages (perhaps within Vitaceae, but also outside of the family), which may have served as past donors of transgenes. We have previously referred to this as the ghost of HGT’s past [50]. In support of this possibility, none of the more ancient transgenic sequences we identified grouped with their current hosts *Tetrasigma*, as would be expected if these species served as hosts. Two genes (i.e., *atp1* and *rps7*) involved in these more ancient HGT events are sister to Vitaceae, suggesting that close relatives of *Tetrasigma* may have served as past transgenic donors. In three other cases (i.e., *atp4*, *cox1*, and *atp6*), however, transgenic sequences do not group closely to Vitaceae (e.g., Cucurbitaceae and *Daucus*; Table 1), indicating different transgenic donors (no evidence of gene conversion, which would confound phylogenetic placements, was detected in these genes using the OrgConv package [51] with p-value < 0.001). To our knowledge, this is the first evidence that Rafflesiaceae may have previously
parasitized different host species, which served as transgenic donors in the past. Further taxon sampling of these genes by co-authors Z.X. and Y.W. is underway and should allow us to determine those previous host donors more precisely.

For the remaining 28 instances of HGT, it is most likely that these were the result of more recent gene transfers. Transgenic sequences in these cases are found exclusively in either *Rafflesia* (i.e., *rps4*) or *Sapria* (e.g., *atp9* and *cob*), or if identified in both

---

**Figure 2.** Gene organization of three assembled contigs for Rafflesiaceae (*Rafflesia cantleyi*, *Rafflesia tuan-mudae*, and *Sapria himalayana*), *Ricinus communis* (Euphorbiaceae), and *Vitis vinifera* (Vitaceae). (A–C) The green and red boxes indicate *Ricinus*-like and *Vitis*-like genes, respectively. Pseudogenes are represented by striped boxes, and the sequence length (in kilobases [kb]) is indicated to the right of each assembled contig. Gene organization of *Ricinus* and *Vitis* mitochondrial genomes (mtDNA) follows Rivarola et al. [20] and Goremykin et al. [23], respectively.

doi:10.1371/journal.pgen.1003265.g002
Rafflesia and Sapria they do not form a clade (i.e., rpl2 and rps1) (Figure S1A and S1D). Evidence of such recent HGT is especially prevalent in S. himalayana: 17 of its transgenic sequences are sister to, or nested within, Tetrastigma (Table 1). Moreover, our phylogenetic analyses indicate that some of these sequences may have resulted from multiple independent gene transfers involving the same gene because transgenic sequences from Rafflesiaceae do not form a clade. In some cases, these gene transfers appear to involve multiple transgenic sequences within a single species for the same gene (i.e., cos1 in Sapria, which possesses two distinct transgenic sequences that appear to have been transferred independently). In other cases, gene transfers involve multiple transgenic sequences in different species for the same gene (i.e., rpl2 and rps1, which show independent transfer events for Rafflesia and Sapria). These more recent HGT events are further supported by syntenic: transgenic sequences involving the same gene from Rafflesia and Sapria are located at different positions in the mitochondrial genome (e.g., rpl2; Figure 2A and 2C). Why some genes exhibit repeated HGT is fertile ground for future investigation.

Conclusion

Our study is the first to comprehensively assess the magnitude of HGT in plants involving a genome (i.e., mitochondria) and a species interaction (i.e., parasitism) where it has been hypothesized to be potentially rampant. These results reveal a high degree of HGT in the mitochondrial genome of Rafflesiaceae involving genes that were likely acquired from its host at various time intervals. We previously established that in R. cantleyi, about 2.1% of nuclear gene transcripts have likely been acquired from its host via HGT [28]. In contrast, our study conservatively indicates that 24–41% of the mitochondrial gene sequences show evidence of HGT in Rafflesiaceae, depending on the species. These results establish for the first time that although the magnitude of HGT involving nuclear genes is appreciable, HGT involving mitochondrial genes in these parasitic plants is an order of magnitude higher. This elevated rate of HGT involving the mitochondrial genome may represent a more general pattern for other parasitic plant clades, and perhaps more broadly for angiosperms.

Materials and Methods

Molecular techniques and next-generation sequencing

For R. cantleyi and S. himalayana, mitochondria were isolated from ~30 grams of fresh material from flower buds using the sucrose gradient centrifugation protocols of Jansen et al. [30]. DNA extracted from purified mitochondria was amplified with the REPLI-g Midi Kit (Qiagen, Inc.). When we were unable to acquire fresh material, total genomic DNA (gDNA) was extracted from silica-dried material using the DNeasy Plant Mini kit (Qiagen, Inc.), and treated with RNase A at 60°C for 1.5 hours to remove any residual RNA contamination. For each species, an Illumina library with the insert size of 350±50 bp was prepared from five micrograms of DNA following the protocols of Bentley et al. [52]. All libraries were sequenced on the Genome Analyzer II Systems Biology at Harvard University (Table S1).

Sequence assembly and alignment

Illumina reads were assembled de novo in ABySS v1.2.1 [53] using default parameters (Table S1). The assembled contigs were annotated against published mitochondrial and plastid genomes from 27 seed plants (Table S3) with BLASTN v2.2.23 [54] using an e-value ≤10^{-5}. Gene sequences from all species were then queried against themselves using BLASTN v2.2.23. BLASTN hits with an e-value ≤10^{-10} were passed to MCL v8.0-312 [55] for Markov clustering. Only those gene clusters that included at least Ccys/Spirodela (for outgroup rooting), Rafflesia/Sapria (sequences under investigation), Ricinus (close relative of Rafflesiaceae), and Vitis (close relative of Tetrastigma) were retained. The nucleotide sequences of each gene were first aligned using MAFFT v6.624 [56], and then manually inspected and realigned if necessary.

Gene copy number estimation

To assess gene copy number and corresponding genomic compartment localization of our assembled gene sequences, we mapped the Illumina gDNA reads from R. cantleyi [28], R. tuan-mudae, and S. himalayana to gene sequences identified here and to the 1305 nuclear genes identified from R. cantleyi [28] and R. tuan-mudae [18] using Bowtie v0.12.7 [57] (Table S2). To avoid complications with intron regions, we first divided each Illumina read into multiple 25 bp fragments following Kim and Salzberg [50], and then mapped each 25-mer with zero mismatches and unique mapping.

Phylogenetic analyses and alternative topology tests

Our ML analyses were conducted for all genes using RAxML v7.2.8 [59] with the GTR+I+T nucleotide substitution model. The best-scoring ML tree and BP for each gene were obtained using the rapid bootstrap algorithm [60] with 500 replicates (Figures S1 and S2). For nad5, we also performed ML analyses on three gene regions separately (i.e., nad5 exon A, intron A/B, and exon B; Figure 2A, see also Figure S1C), which allowed us to determine the location of homologous recombination more accurately (see above).

Alternative topology tests were performed in an ML framework using the approximately unbiased (AU) test [61] as implemented in scaleboot v0.3-3 [62] (Table 1). To generate constrained ML trees for genes that show evidence of HGT, we enforced all transgenic and native (when present) sequences from Rafflesiaceae to be monophyletic with Ricinus, and then conducted ML searches using these constraints.

Gene expression level analyses

To estimate the gene expression level in R. cantleyi, the Illumina cDNA reads from Xi et al. [28] were mapped onto the assembled R. cantleyi mitochondrial gene sequences using Bowtie v0.12.7 [57] as described above. cDNA reads that mapped onto each gene sequence were then summed and further normalized to reads per kilobase per million reads (RPKM [63]; Table 1, see also Figure S4).

Contamination and the determination of HGT in Rafflesiaceae

Tremendous care was taken to avoid and/or detect host or lab contamination during our sample preparation and data analyses. First, our DNA sample preparation and genome library sequencing of Rafflesiaceae were performed separate from any work involving Tetrastigma; thus, laboratory contamination of our Rafflesiaceae DNAs with Tetrastigma is unlikely. Second, the plastid genome has apparently been lost in Rafflesiaceae [31]. If there were any host contamination, the host’s plastid gene sequences should be easily detected in our sequence data. This was not the case. Third, the mitochondrial genome sequences of R. cantleyi and R. tuan-mudae were generated from two different sources, i.e., a fresh flower bud using sucrose gradient centrifugation and silica-dried perigone lobes using total gDNA extraction,
respectively (Table S1). If one of these samples, or genome libraries, were contaminated, we would not expect to have identified the identical set of transgenes from these samples. Similarly, for *S. himalayana*, all transgenes identified from the genome library prepared using sucrose gradient centrifugation were verified in our second library of this species that was prepared from total gDNA (Table S2). Fourth, most transgenic sequences identified here possess some amount of sequence divergence when directly compared with homologs from their current host. For example, all 15 transgenic sequences from *Rafflesia* show some degree of sequence divergence when directly compared with homologs from their host species, *T. rafflesiae* (mean DNA sequence distance = 0.042189). Similarly, 26 of 30 transgenic sequences from *Sapria* show some degree of sequence divergence when directly compared with homologs from their host species, *T. cruciatum* (mean DNA sequence distance = 0.020265) (Figure S1A and S1D). These sequence distances are significantly greater (p-value < 0.01, Welch’s t test) than those between the two included *Tetrastigma* species (mean DNA sequence distance = 0.001984). This is despite the fact that these two *Tetrastigma* species have diverged from each other at least 10 Ma [49]. Furthermore, three transgenic sequences from *Rafflesia* and 13 transgenic sequences from *Sapria* contain nonsense mutations (Table 1). These results strongly indicate that some period of evolution has elapsed since the time of HGT. Fifth, all seven transgenes from *R. cantleyi* show evidence of gene expression based on its transcriptome (Table 1), and levels of expression are not significantly different between transgenes and native genes (p-value = 0.19, Welch’s t test; Figure S4). Lastly, and perhaps most importantly, seven of our assembled contigs contain both transgenes and native genes (Figure 2) indicating that these transgenes are clearly integrated into the mitochondrial genome of Rafflesiaceae.

**Supporting Information**

Figure S1  Phylograms of all horizontally transferred genes in the mitochondrial genomes of *Rafflesia cantleyi*, *Rafflesia tuan-mudae*, and *Sapria himalayana*. Maximum likelihood bootstrap percentages (BP) were summarized from 500 bootstrap replicates, and only BP values greater than 50% are shown. Gene sequences from Rafflesiaceae and the host *Tetrastigma* are highlighted in red and blue, respectively. V indicates sequences of vertical transmission. Number of aligned characters (chars) and scale bar (substitutions per site) are shown for each gene. (PDF)

Figure S2  Phylograms of all vertically transferred and unassigned genes in the mitochondrial genomes of *Rafflesia cantleyi*, *Rafflesia tuan-mudae*, and *Sapria himalayana*. Maximum likelihood bootstrap percentages (BP) were summarized from 500 bootstrap replicates, and only BP values greater than 50% are shown. Gene sequences from Rafflesiaceae and the host *Tetrastigma* are highlighted in red and blue, respectively. V indicates sequences of vertical transmission. Number of aligned characters (chars) and scale bar (substitutions per site) are shown for each gene. (PDF)

Figure S3  Gene organization of three assembled contigs (A–C) for *Sapria himalayana* (Rafflesiaceae) and *Vitis vinifera* (Vitaceae). The red boxes indicate *Vitis*-like genes. Pseudogenes are represented by striped boxes, and the sequence length (in kilobases [kb]) is indicated to the right of each assembled contig. (PDF)

Figure S4  Boxplot of gene expression levels of horizontally (HGT) and vertically (VGT) transferred mitochondrial gene sequences in *Rafflesia cantleyi*. The number of gene sequences for each category is shown in parentheses; RPKM = reads per kilobase per million reads. (PDF)

Table S1  Newly sequenced species in this study with associated assembly statistics. (PDF)

Table S2  Estimated gene copy number for *Rafflesia cantleyi*, *Rafflesia tuan-mudae*, and *Sapria himalayana*. The Illumina genomic DNA reads from *R. cantleyi*, *R. tuan-mudae*, and *S. himalayana* were mapped onto the 38 mitochondrial genes, 14 genes of plastid origin, and 1,305 nuclear genes from *R. cantleyi* and *R. tuan-mudae*. Illumina reads that mapped onto each gene were then summed and further normalized to reads per base pair. (XLS)

Table S3  Data sources of gene sequences included in our phylogenetic analyses. (PDF)

Acknowledgments

We thank Renate Hellmisi, Matthew Klooster, Kenneth Wurdack, and members of the Davis laboratory for advice and discussion. We would also like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

Author Contributions

Conceived and designed the experiments: ZX RKB CJM JSR CCD. Performed the experiments: ZX YW CCD. Analyzed the data: ZX YW RKB CJM JSR CCD. Contributed reagents/materials/analysis tools: MS. Wrote the paper: ZX YW RKB JSR CCD.

References


