


# Current Biology

Volume 31  
Number 5  
March 8, 2021



 CellPress

# Current Biology

## Deeply Altered Genome Architecture in the Endoparasitic Flowering Plant *Sapria himalayana* Griff. (Rafflesiaceae)

### Highlights

- *Sapria* lost 44% of conserved plant genes
- Patterns of gene loss in *Sapria* are convergent with other parasitic plants
- Most genes are highly streamlined, but others contain exceptionally long introns
- Extinct host associations are revealed from numerous horizontally transferred genes

### Authors

Liming Cai, Brian J. Arnold, Zhenxiang Xi, ..., Sarah Mathews, Timothy B. Sackton, Charles C. Davis

### Correspondence

tsackton@g.harvard.edu (T.B.S.),  
cdavis@oeb.harvard.edu (C.C.D.)

### In Brief

Cai et al. report the first genome of the endoparasitic plant *Sapria*, representing the most extreme form of plant parasitism. Alongside the loss of vegetative features, *Sapria* has lost 44% of conserved plant genes. The genome also demonstrates widespread evidence of horizontal transfer, revealing a dynamic history of former host associations.





## Article

# Deeply Altered Genome Architecture in the Endoparasitic Flowering Plant *Sapria himalayana* Griff. (Rafflesiaceae)

Liming Cai,<sup>1,2</sup> Brian J. Arnold,<sup>3</sup> Zhenxiang Xi,<sup>4</sup> Danielle E. Khost,<sup>3</sup> Niki Patel,<sup>5</sup> Claire B. Hartmann,<sup>6</sup> Sugumaran Manickam,<sup>7</sup> Sawitree Sasirat,<sup>8</sup> Lachezar A. Nikolov,<sup>9</sup> Sarah Mathews,<sup>10</sup> Timothy B. Sackton,<sup>3,\*</sup> and Charles C. Davis<sup>1,2,11,\*</sup>

<sup>1</sup>Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

<sup>2</sup>Harvard University Herbaria, 22 Divinity Avenue, Cambridge, MA 02138, USA

<sup>3</sup>FAS Informatics Group, Harvard University, 38 Oxford Street, Cambridge, MA 02138, USA

<sup>4</sup>Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China

<sup>5</sup>Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, Unit 3043, Storrs, CT 06269, USA

<sup>6</sup>Bauer Core Facilities, Division of Science, Harvard University, 52 Oxford Street, Cambridge, MA 02138, USA

<sup>7</sup>Rimba Ilmu Botanic Garden, Institute of Biological Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia

<sup>8</sup>Queen Sirikit Botanic Garden, PO Box 7 Mae Rim, Chiang Mai 50180, Thailand

<sup>9</sup>Department of Molecular, Cell and Developmental Biology, and Molecular Biology Institute, University of California, Los Angeles, 610 Charles E Young Drive East, Los Angeles, CA 90095, USA

<sup>10</sup>Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70808, USA

<sup>11</sup>Lead Contact

\*Correspondence: [tsackton@g.harvard.edu](mailto:tsackton@g.harvard.edu) (T.B.S.), [cdavis@oeb.harvard.edu](mailto:cdavis@oeb.harvard.edu) (C.C.D.)

<https://doi.org/10.1016/j.cub.2020.12.045>

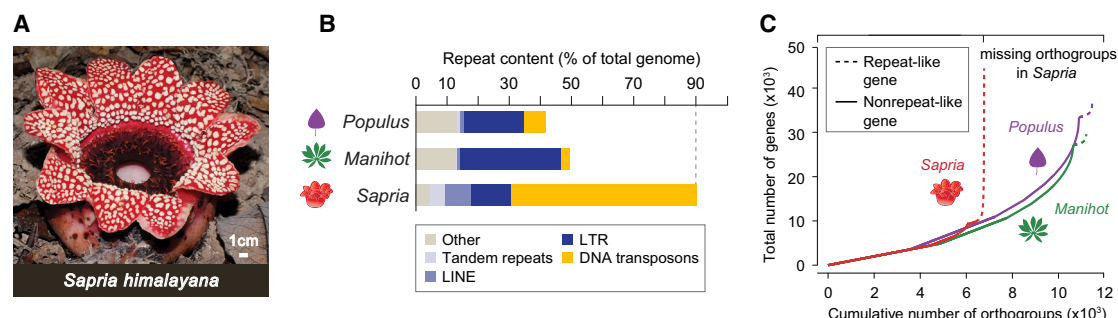
## SUMMARY

Despite more than 2,000-fold variation in genome size, key features of genome architecture are largely conserved across angiosperms. Parasitic plants have elucidated the many ways in which genomes can be modified, yet we still lack comprehensive genome data for species that represent the most extreme form of parasitism. Here, we present the highly modified genome of the iconic endophytic parasite *Sapria himalayana* Griff. (Rafflesiaceae), which lacks a typical plant body. First, 44% of the genes conserved in eurosids are lost in *Sapria*, dwarfing previously reported levels of gene loss in vascular plants. These losses demonstrate remarkable functional convergence with other parasitic plants, suggesting a common genetic road-map underlying the evolution of plant parasitism. Second, we identified extreme disparity in intron size among retained genes. This includes a category of genes with introns longer than any so far observed in angiosperms, nearing 100 kb in some cases, and a second category of genes with exceptionally short or absent introns. Finally, at least 1.2% of the *Sapria* genome, including both genic and intergenic content, is inferred to be derived from host-to-parasite horizontal gene transfers (HGTs) and includes genes potentially adaptive for parasitism. Focused phylogenomic reconstruction of HGTs reveals a hidden history of former host-parasite associations involving close relatives of *Sapria*'s modern hosts in the grapevine family. Our findings offer a unique perspective into how deeply angiosperm genomes can be altered to fit an extreme form of plant parasitism and demonstrate the value of HGTs as DNA fossils to investigate extinct symbioses.

## INTRODUCTION

Species in the flowering plant clade Rafflesiaceae represent the most extreme form of parasitism achieved by plants.<sup>1</sup> Flowers of these species emerge directly from their hosts, subtended by a few rudimentary bracts, but otherwise exhibit no evidence of an obvious plant body. Instead, the vegetative phase of these endophytic holoparasites persists only as a reduced mycelium-like body, with which they obtain nutrients from their obligate hosts in the grapevine family Vitaceae.<sup>1,2</sup> It is thus even more striking that members of Rafflesiaceae produce the largest flowers in the world. Their flowers mimic carrion and deceive the flies that pollinate them.<sup>3,4</sup> Fertilization appears to

be rare, owing to high bud mortality and strongly skewed sex ratios,<sup>5</sup> but when it does occur, successful fertilization may yield more than a quarter million tiny seeds embedded in their immobile, woody fruits.<sup>6</sup> This unusual reproductive mode, combined with their reliance on specific hosts, has likely led to multiple founder events and local extinctions, as reflected by their low allelic diversity.<sup>5</sup> In addition, molecular investigations of Rafflesiaceae have identified intriguing initial findings regarding their genome evolution, including the hypothesized complete loss of their plastid genome<sup>7</sup> and multiple host-to-parasite horizontal gene transfers (HGTs).<sup>8,9</sup> The combination of these factors makes Rafflesiaceae of particular interest for further comparative genomic investigation, especially to address the



**Figure 1. *Sapria himalayana* and Its Highly Repetitive Genome**

(A) *Sapria himalayana* Griff. Image courtesy of L. Worthington (CC BY-SA 2.0).

(B) Repeat content of *Sapria* and close free-living relatives *Manihot esculenta* and *Populus trichocarpa*. Repeat type is color coded according to legend.

(C) Cumulative number of genes within orthogroups in *Sapria*, *Manihot*, and *Populus*. Orthogroups are ordered from small to large for nonrepeat-like (solid line) and repeat-like (dotted line) genes. Very few highly abundant repeat-like genes contribute to the high gene number in *Sapria* despite missing nearly half of the orthogroups that are present in *Manihot* or *Populus*.

See also Figures S1 and S2 and Data S1B and S2A.

lack of a genome assembly for an endophytic plant parasite genome.

## RESULTS AND DISCUSSION

### Genome Assembly of *Sapria himalayana*

Here, we present a 1.28-Gb genome assembly (N50 = 4.3 Mb) of *Sapria himalayana* Griff., a species of Rafflesiaceae that parasitizes three distantly related *Tetrastigma* species (Vitaceae) in Southeast Asia.<sup>10</sup> Estimates of genome size in *Sapria* range from 1.69 to 2.54 Gb based on k-mer distributions (Figure S1) and from 3.2 to 3.5 Gb based on flow cytometry (Figure S1; Data S1A). Our genome was assembled using Chromium technology (10X Genomics, Pleasanton, CA, USA) and nanopore long-read sequence data (Oxford Nanopore Technologies, Oxford, UK).<sup>11</sup> The GC content of the assembly is 24% and exhibits a clear bimodal distribution in scaffolds (Figure S2A). The first peak consists of AT-rich repeat motifs; the second includes mostly gene-rich scaffolds with higher GC content. We estimate that repeat motifs account for 89.6% of the genome (Figure 1). In particular, the two DNA transposon families *CMC–EnSpm* and *hAT–Ac* alone comprise 29.8% and 27.3% of the genome, respectively (Data S1B). The extent of repetitive elements in the *Sapria* genome represents a significant challenge to obtaining a contiguous and complete assembly. Our assembly, although only 40% of the estimated genome size, is a nearly complete representation of the single-copy portion of the genome. More than 99.0% of the Illumina reads map to our assembly, and between 96.2% and 100% of the single-copy regions of the genome are assembled based on kmer analysis and read coverage (see STAR Methods for additional details). We predict a total of 55,179 gene models, of which 42,512 were validated by transcriptome sequences, plant protein databases, or the Pfam database. These protein-coding genes exhibit conserved GC content (mean 41.2%) and exon length (mean 214 bp) compared to other angiosperms, but they encode smaller proteins (mean 265 amino acids) and have fewer introns (mean 3.1; Figure S2). The high number of predicted genes is largely driven by a small number of abundant orthogroups

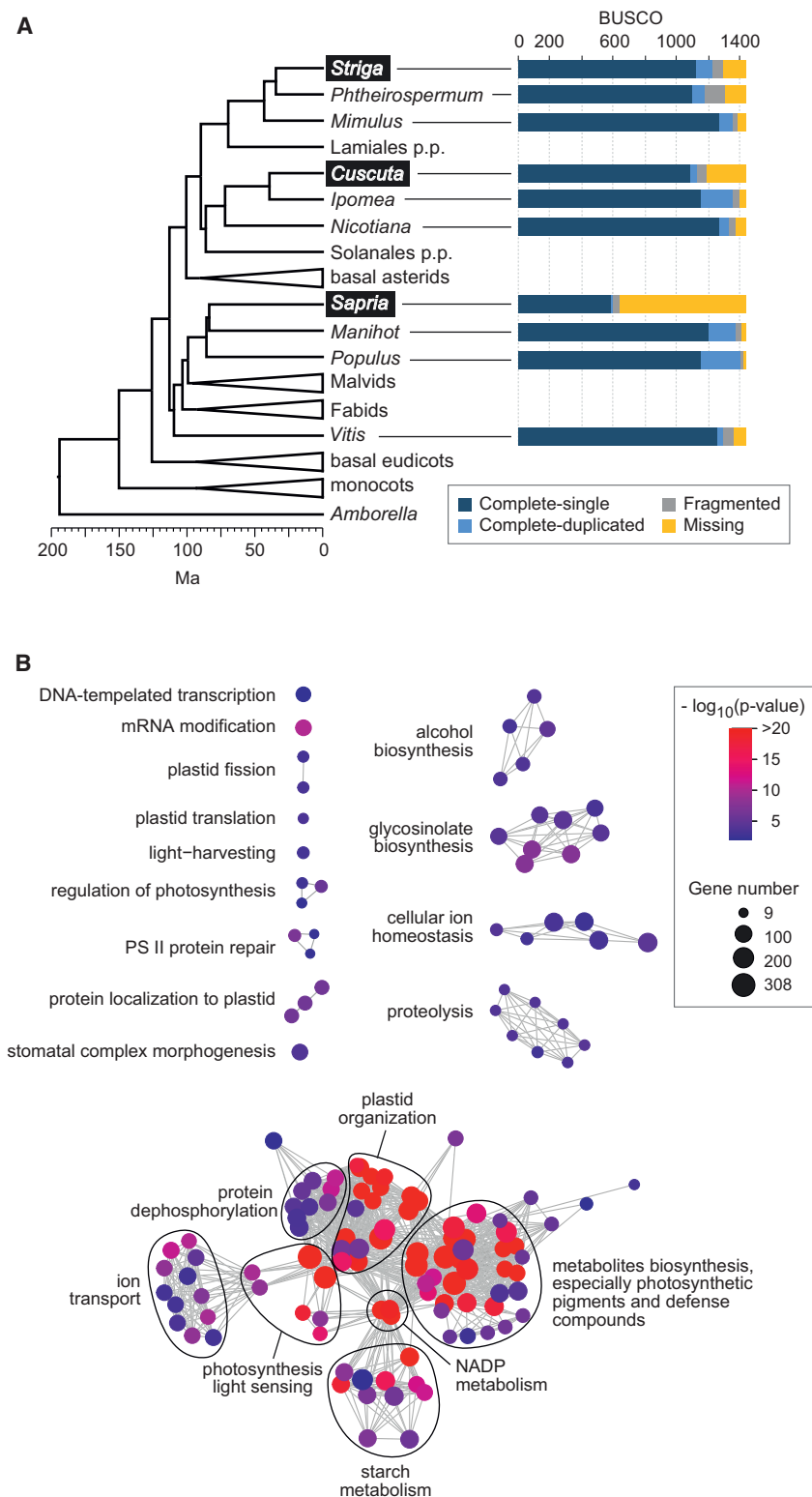
consisting largely of transposable elements (TEs). Specifically, 736 (10.9%) orthogroups containing TE-like domains account for 82.6% ( $n = 35,136$ ) of the validated *Sapria* genes (Figure 1C). The most abundant orthogroup alone contains 4,562 copies of a mitochondrial gag-Pol-related retrotransposon. Using a likelihood-based method,<sup>12</sup> we additionally inferred orthogroup size evolution in *Sapria* under a phylogenetic framework and identified gene expansion in 710 (10.5%) medium-sized orthogroups (<100 gene copies per species). These orthogroups are involved in a broad range of functions, including mitochondria and nuclear chromosome organization, DNA metabolic process, and cell cycle process (Data S1C).

*Sapria* presents a compelling opportunity to understand the connection between its extremely derived phenotype and genome architecture. Across multiple dimensions, the genome reflects its unique biology in gene content, intron size, and presence of HGTs, which also greatly challenges our understanding of plant genome architecture.

### Unprecedented Level of Gene Loss

Consistent with its extraordinary reduction in morphology and physiological modification, we document unprecedented gene loss in *Sapria*, totaling nearly half (44.4%) of the 10,880 orthogroups that are universally conserved across eurosids. This magnitude of gene loss dwarfs any previously reported case in angiosperm parasites.<sup>13–15</sup> Specifically, the extent of gene loss in *Sapria* is nearly four times greater than that of the hemiparasite *Striga* (witchweed, Orobanchaceae; 9.3% gene loss)<sup>15</sup> and two times greater than the hemi-to-holoparasite *Cuscuta* (dodder, Convolvulaceae; 15.7% gene loss; Figure 2).<sup>13,14</sup> Of the conserved genes lost in *Sapria*, 13.2% ( $n = 642$ ) are commonly lost in all three of these independently evolved parasitic clades. These convergently lost genes are enriched in functions involving photosynthesis, defense, and stress response (Figure 3; Data S1C) and likely represent a common genetic response to the shared physiological modifications underlying the transition from an autotrophic to a heterotrophic lifestyle.

Within these convergently reduced functional categories, a far greater number of genes are lost in *Sapria* when compared to



**Figure 2. Extreme Gene Loss in *Sapria himalayana***

(A) A simplified phylogeny of flowering plant genomes sampled in this study. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessments for *Sapria* and two additional independently evolved obligate parasitic plants, *Striga asiatica* and *Cuscuta australis*, are highlighted. These BUSCOs are contrasted with their free-living close relatives, plus *Vitis vinifera*, a close relative of the Rafflesiaceae hosts. *Sapria* has substantially more missing BUSCOs (shown in yellow) compared with all taxa.

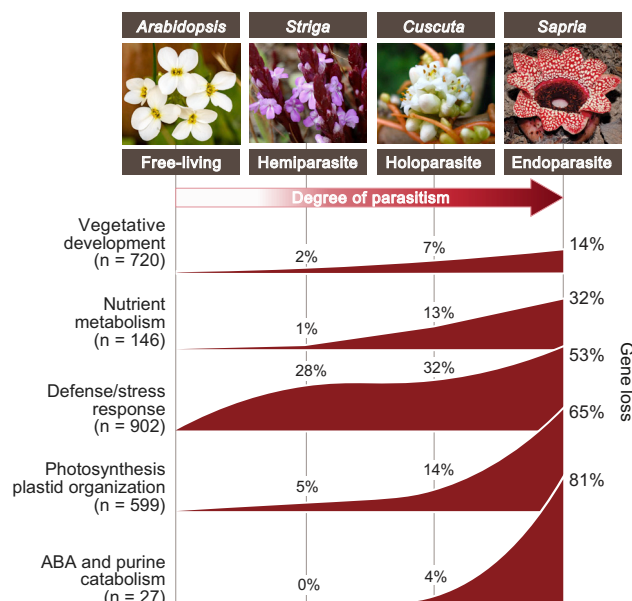
(B) Gene ontology (GO) enrichment analysis of missing conserved orthogroups in *Sapria*. Each circle represents an enriched biological process and is colored by the statistical significance of the enrichment ( $\log_{10}$  p value). The size of the circle indicates the number of genes associated with the term. See also [Data S1E](#) and [S2](#).

and precursor metabolites, such as terpenoid and phyloquinone ([Data S1E](#)). But perhaps the most extreme example involves the complete loss of the plastid genome hypothesized by Molina et al.,<sup>7</sup> which would represent the sole case of such loss across more than a quarter million land plant species. This hypothesis has remained controversial, owing to the presence of nuclear-encoded plastid-targeting products in Rafflesiaceae.<sup>16</sup> In *Sapria*, however, we find convincing evidence for the loss of the plastid genome. Among the 290 scaffolds with sequence similarity to plastid genes, none are of plastid origin, given their GC content, read coverage, and phylogenetic affinity ([Data S1F](#)). Plastid genome loss is further supported by the nearly complete loss of nuclear genes that regulate plastid organization and function, including plastid fission, localization, transcription, membrane biosynthesis, photosynthetic pigment biosynthesis, and plastid-targeted TAT protein transportation ([Figure 2](#); [Data S1E](#)).

We also identify enriched losses in functional categories not previously documented in parasitic plants, including biosynthesis of the plant hormone abscisic acid (ABA), protein degradation, and purine metabolism ([Data S1E](#)). For ABA production, 18 of 27 genes associated with its biosynthesis in *Arabidopsis* are absent in *Sapria* ([Figure 3](#)). Such loss is likely indicative of inability to produce ABA because of (1) nearly complete loss in the biosynthesis

of its precursor carotenoid; (2) loss of the plastid, which is the site of carotenoid biosynthesis; (3) loss of key genes in ABA biosynthesis, including *AAO* and *NCED*;<sup>17</sup> and (4) widespread gene loss in stress-response pathways, which are the major

other parasites, especially for genes related to photosynthesis and plastid organization ([Figure 3](#)). The loss of photosynthesis in *Sapria* has led to significant reduction in genes that regulate the biosynthesis of energy storage molecules, such as starch and fatty acid,



**Figure 3. Convergent Patterns of Gene Loss in Parasitic Plants**

Increased percentage of gene loss across key functional categories in the hemiparasite *Striga asiatica*, holoparasite *Cuscuta australis*, and endoparasite *Sapria himalayana*. Number of genes associated with each functional category in *Arabidopsis thaliana* is listed to the left. Fractional quantification of loss for each species is estimated based on the GO annotation of *Arabidopsis thaliana* for each lost orthogroup. Images courtesy of S. Geyer (CC BY-NC 2.0), J. Pail (CC BY-NC 2.0), P. Leautaud (CC BY-NC 2.0), and L. Worthington (CC BY-SA 2.0). See also Data S1D.

targets of ABA.<sup>18</sup> Within the protein degradation pathway, significant gene reduction is observed in the SCF-ubiquitin-proteasome-mediated protein lysis and the endopeptidase Clp-mediated protein lysis (Data S1E). These losses may be associated with the reduced requirement for nutrient recycling and abiotic stress response, because significant loss is also observed in nitrate assimilation and amino acid transmembrane transport pathways (Data S1E). Finally, within the purine metabolic pathway, only one homolog of nucleoside diphosphate kinase (*NDPK1*) and two homologs of nucleoside-triphosphatase (*AYP1* and *AYP2*) are retained in *Sapria* versus at least four and six homologs encoding these enzymes in *Arabidopsis*, respectively. These reductions may restrict the native production of ribonucleotides and suggest the uptake of these resources from the host.

### Extreme Intron Length Disparity

Genome streamlining, or the tendency toward reduction in non-coding DNA, has been widely reported in obligate intra- and intercellular parasites, including in bacteria,<sup>19</sup> protozoa,<sup>20</sup> and nematodes,<sup>21</sup> potentially reflecting common selective pressures to reduce metabolic cost and cell size faced by all parasites. Likewise, the majority of genes in *Sapria* are highly compact despite its large genome. *Sapria* has even fewer introns on average (3.1 per gene) than *Genlisea aurea* (3.5 per gene), which is a carnivorous plant with the smallest angiosperm genome (63.6 Mb) published to date.<sup>22</sup> In *Sapria*, at least 18.7% of the genes have lost all introns that are otherwise present in both of

its closest free-living relatives, *Manihot* and *Populus*. The abundance of these intron-free genes indicates that retroprocessing-mediated gene conversion is likely an important mechanism for intron deletion in these species.<sup>23</sup>

In *Sapria*, highly compact genes (maximum intron length <150 bp) are significantly enriched for housekeeping functions, such as DNA and RNA metabolism, stem cell maintenance, and reproduction (Data S1G). This contrasts sharply with other free-living plants whose intron size is largely decoupled from gene function.<sup>24</sup> Such gene streamlining may convey a selective advantage via more efficient transcription (the energy cost hypothesis *sensu* Castillo-Davis et al.<sup>25</sup>), which may be especially advantageous for parasites that rely on their host for energy and chemical resources. In particular, losses in purine metabolic pathways and nitrogen uptake in *Sapria* may limit accessibility to the cellular ribonucleotide pool and thus further restrict transcription. In contrast, both *Manihot* and *Populus* exhibit only a small number of gene ontology (GO) terms enriched for highly compact genes (Data S1H and S1I), suggesting a primarily stochastic process of gene streamlining in these free-living relatives. Surprisingly, genes related to the development of stomata, leaves, and roots are also enriched in the highly compact genes in *Sapria*, despite the apparent loss of these readily identifiable vegetative features.<sup>3</sup>

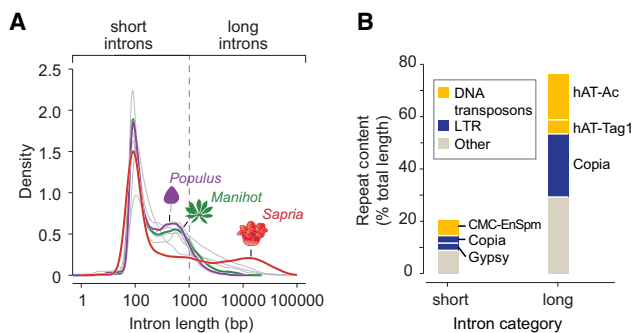
Despite a widespread signal of gene streamlining, a substantial proportion of the genes in *Sapria* contain among the longest introns documented in plants (Figure 4A). Among the eleven eukaryotic plant genomes in our study, no more than 12.3% of introns exceed 1 kb. In contrast, 27.5% of introns in *Sapria* are longer than 1 kb (Figure 4A). The longest intron verified with our transcriptome data is 97.8 kb. Due to these outliers, the average intron length in *Sapria* (1,527 bp) is even longer than that of the Norway spruce (998 bp), which has an enormous 19.6 Gb genome.<sup>26</sup> For introns longer than 1 kb, 74% of the total length consists of TEs (Figure 4B). A similar role of TEs in intron expansion has previously been reported in grapes and gymnosperms.<sup>24,26</sup>

Several hypotheses have been proposed to explain intron length evolution in plants, including the energy cost hypothesis we suggest above<sup>25</sup> to explain gene streamlining in *Sapria*. Here, we observe a significant positive correlation between maximum intron size and  $d_N/d_S$  ratio (Spearman's correlation  $p = 7.2e-9$ ) that is consistent with the mutation burden hypothesis proposed by Lynch.<sup>27</sup> According to Lynch,<sup>27</sup> genes under relaxed selection are more likely to tolerate amino-acid-altering mutations due to intron expansion and imprecise transcript splicing. Hence, a positive correlation between intron length and  $d_N/d_S$  is expected. Finally, we found no correlation between orthologous intron sizes in comparisons between *Sapria-Manihot* or *Sapria-Populus* ( $p > 0.30$  Spearman's rank correlation test), suggesting that putative relaxed selection in genes with long introns is not an ancestral feature of Malpighiales but has instead evolved independently in Rafflesiaceae.

### HGT

Parasitic plants acquire novel genomic components via HGT facilitated by intimate physical associations with their hosts. This phenomenon was initially reported in Rafflesiaceae,<sup>28–30</sup> where numerous host-to-parasite gene transfers were detected





**Figure 4. Intron Length Disparity and Intron Composition in *Sapria himalayana***

(A) Intron length distribution of *Sapria* (red) and eleven eurosid species sampled in this study (gray). The closest free-living relatives of *Sapria*—*Manihot esculenta* and *Populus trichocarpa*—are highlighted in green and purple, respectively.

(B) Repeat content for short (<1 kb) and long (>1 kb) introns. For short introns, only 20.0% of the total size consists of repeats. In contrast, 74.3% of the total size of long introns consists of repeats. The most dominant repeat types (>2% total length) are labeled.

See also Figure S5 and Data S1.

in nuclear transcribed genes and mitochondrial genomes. To further explore HGT in the *Tetrastigma*-Rafflesiaceae host-parasite system, we additionally generated a *de novo* genome assembly of *Tetrastigma voinieranum* Pierre ex Pit., a close relative of the hosts of *Sapria*,<sup>10</sup> using nanopore data. We characterized donor and recipient lineages of HGT using a phylogenomic pipeline and also applied a genome scan approach to illuminate fine-scale insertions of HGT in both genic and intergenic regions (Figure S3). In our phylogenomic analysis, we included a total of 55 species with expanded sampling in the parasite and host lineages (Figures 5 and S4; Data S1J): three transcriptomes from Rafflesiaceae (*Rafflesia cantleyi* Solms-Laubach, *Rafflesia tuan-mudae* Becc., and *Rhizanthus zippelii* (Blume) Spach<sup>8</sup>); eighteen transcriptomes from Vitaceae;<sup>31,32</sup> and 33 published genomes spanning the angiosperm phylogeny. Together, these species represent all three genera in Rafflesiaceae and 12 of the 14 genera in Vitaceae. In our genome scan assessment, we generated an alignment of ten complete plant genomes (Data S1J) to identify HGT regions that are highly similar between the host and parasite but absent in the closest relatives of the parasite (Figure S3). Pairwise divergences were calculated between *Sapria* and its Malpighiales relatives (*Manihot* and *Populus*) or host (*Tetrastigma*) within sliding windows of 100 aligned bases (Figures 5A and 5B). We further used nanopore reads to verify HGT candidates inferred from both assessments and removed candidates that are potentially subject to natural host DNA contamination or chimeric assembly (Figure S3).

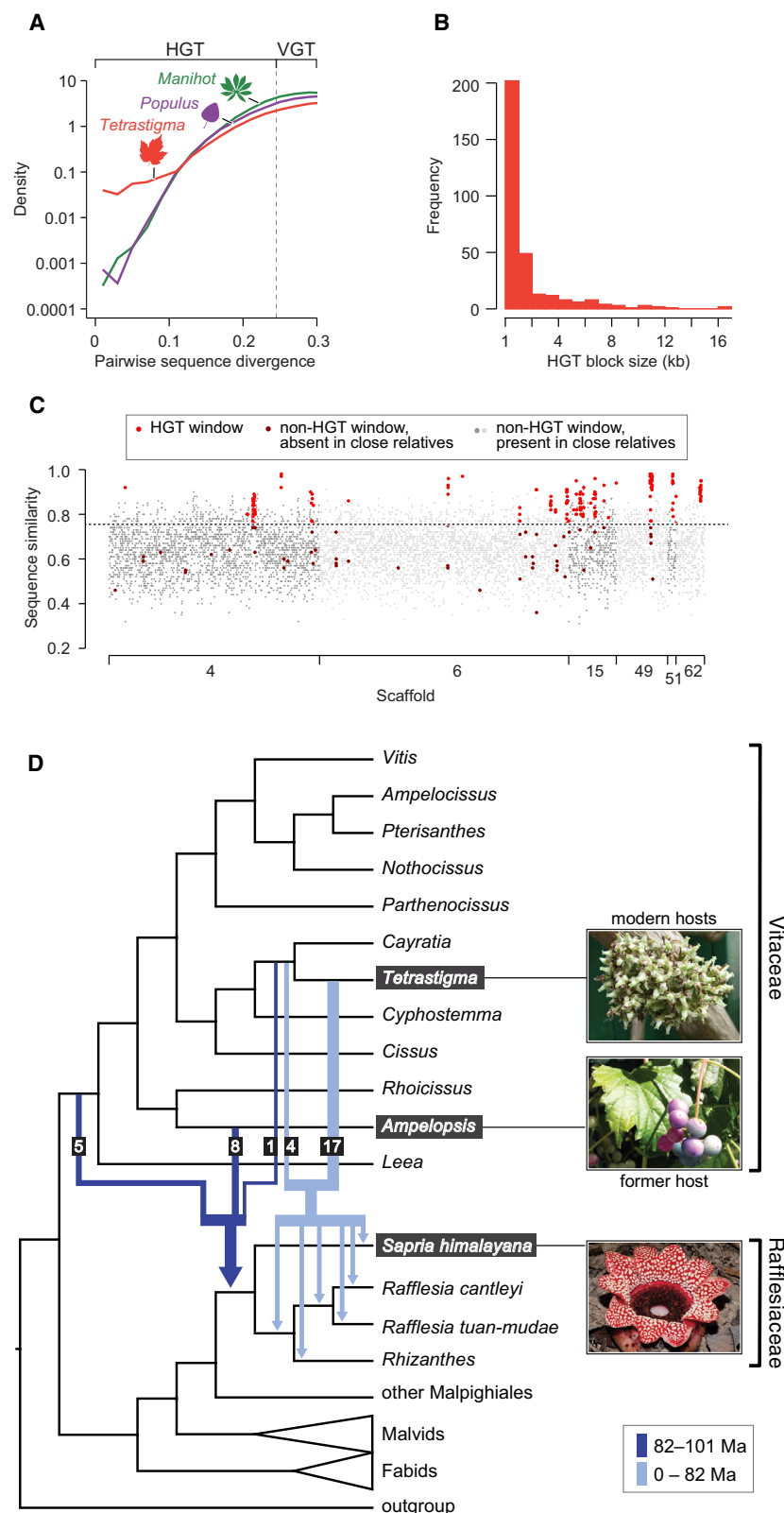
Our results corroborate and greatly expand previous findings that Rafflesiaceae represent active areas of HGT.<sup>34</sup> Our phylogenomic approach identified HGT in 568 genes and pseudogenes from 81 orthogroups, totaling 1.2% of the 6,552 orthogroups examined. Our genome scan approach identified HGTs in 314 genomic blocks that account for 1.2% (100.6 kb) of uniquely aligned *Sapria* sequences. Despite the fact that these two methods were designed to identify non-overlapping types of

HGTs (Figure S3), they independently converge on similar estimated levels of HGT and indicate that at least 1.2% of the low-copy regions of the *Sapria* genome are attributable to HGT. These HGTs range from 100 bp to 16.5 kb in length (median = 559 bp; Figure 5B), and 62% of them are intergenic. Introns were detected in all but two HGT genes where the donor sequences contained introns, supporting previous findings that the uptake of naked foreign DNA is the primary source of HGT.<sup>26,35</sup>

The codon properties of HGTs are highly atypical when compared to *Manihot* and *Tetrastigma*, which is not observed in the vertically transferred genes (VGTs) of *Sapria* (Figure 6A). Surprisingly, HGTs exhibit coding properties more similar to *Sapria*'s closest relative (*Manihot*), despite their closer phylogenetic relationship to *Tetrastigma* (paired t test  $p < 2.2 \times 10^{-16}$ ; Figure 6A). This deviation may be attributed to shifts in mutational and selective biases when HGTs relocate to a novel genomic landscape.<sup>36,37</sup> Such shifts in codon usage have been well characterized in microscopic pathogens<sup>38</sup> and are hypothesized to be driven by the reliance of host tRNA for translation. The adaptive property of codon usage in HGTs is further demonstrated by their prevalent use of optimal codons. On average, 64.8% of codons in HGTs are optimal, although only 48.9% in VGTs are optimal (Figure 6B). We hypothesize that the difference in the use of optimal codons reflects the genealogical history of HGTs and VGTs—native VGTs are more subject to the accumulation of suboptimal codons, due to the small population sizes of *Sapria*,<sup>6</sup> than HGTs. Finally, the frequent use of optimal codons in HGTs is also indicative of high translational efficiency and expression levels.<sup>39</sup> Though the expression profiles of HGTs have yet to be characterized in *Sapria*, it has been reported in the parasitic plant *Cuscuta* that most HGTs are highly expressed in haustoria and play a potentially important role in parasitism.<sup>35</sup> As a result, both genealogical history and translational selection may contribute to divergent codon usage patterns of HGTs and VGTs.

Several HGTs may perform critical functions in their recipient species. We specifically examined the functions of 27 HGTs shared by all Rafflesiaceae. Ten of these encode proteins related to defense or stress response (Data S1J), including defensin and chitinase.<sup>41,42</sup> We also identified independent acquisitions of a phosphomethylpyrimidine synthase gene, *thiC*, in the common ancestor of Rafflesiaceae and acquired secondarily more recently in *Sapria* (Data S1K). Given the essential role of *thiC* in the biosynthesis of pyrimidine<sup>43,44</sup> and the loss of its native copy in *Sapria* (Data S2B), we hypothesize that horizontally transferred genetic materials in Rafflesiaceae may complement genes lost during the evolution of their endoparasitic habit. Among the putatively functional HGTs identified from both the phylogenomic and genome scan assessments, we also find significant enrichment in biological processes that involve beta-glucan biosynthesis, amino acid transport, and methylation (Data S1L). Moreover, twelve of the 81 HGT-derived orthogroups in the phylogenomic analysis encode proteins with retrotransposon- or transposon-related functions (Data S1K).

HGTs also provide a unique opportunity to investigate host-parasite dynamics. Molecular divergence time estimation indicates that crown Rafflesiaceae is older than *Tetrastigma*,<sup>9</sup> suggesting the existence of former host(s) predating their modern



**Figure 5. Horizontal Gene Transfer (HGT) and Host Shift History in Rafflesiaceae**

(A) Distribution of pairwise sequence divergence between *Sapria*–*Tetrastigma*, *Sapria*–*Manihot*, and *Sapria*–*Populus* inferred from the genome scan analysis. Highly similar, low-divergence regions are enriched in the *Sapria*–*Tetrastigma* comparison. Dotted line depicts the sequence divergence cutoff applied to distinguish HGT versus vertical gene transfer (VGT) windows.

(B) Length distribution of HGTs inferred from the genome scan analysis.

(C) Manhattan plot of pairwise sequence similarity between *Sapria* and *Tetrastigma*. Genomic coordinates are displayed along the x axis. Pairwise sequence similarities between *Sapria* and *Tetrastigma* are displayed on the y axis. Bright and dark red dots represent aligned windows that are unique between *Sapria* and *Tetrastigma* and not present in *Manihot* or *Populus*. The bright red windows have sequence similarity higher than the threshold 0.755 and are thus identified as HGT windows. Six scaffolds with the highest proportion of HGT windows are shown.

(D) Phylogenomic assessment of HGT depicting history of host shifts. Phylogeny of Vitaceae follows Wen et al.<sup>31</sup> Dark and light blue arrows indicate direction of ancestral and recent HGTs, respectively. The number of genes supporting each set of transfers is indicated near its respective branch. Relative dating of HGTs follows age estimates of stem and crown group Rafflesiaceae.<sup>33</sup> Images courtesy of Aqiao HQ (CC BY-SA 2.0), Jean (CC BY 2.0), and L. Worthington (CC BY-SA 2.0).

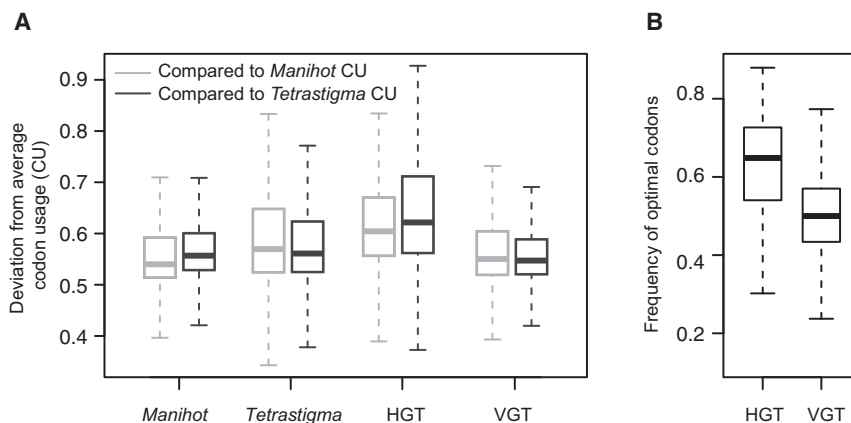
See also [Figures S3, S4, and S6](#) and [Data S1K, S1M, and S2](#).

genus. Among them, the recipients of HGTs in 24 orthogroups are restricted to various crown Rafflesiaceae clades and likely represent younger HGT events. On the other hand, HGTs in 17 orthogroups are ancestral to all Rafflesiaceae genera and thus represent older HGT events. For 21 of the 24 younger HGTs, the donor is either their modern hosts *Tetrastigma* or the most recent common ancestor (MRCA) of *Tetrastigma* and *Cayratia* (Figure 5D; Data S1M). This result strongly supports the long-standing association of the modern Rafflesiaceae–*Tetrastigma* symbiosis. However, none of the 17 ancestral HGTs are associated with *Tetrastigma*. The most common donor among these ancestral HGTs is *Ampelopsis*, which is implicated in eight of the 17 HGTs (Data S1M). Interestingly, this genus has numerous extant species that overlap with the modern distribution of Rafflesiaceae and could plausibly have served as former hosts. Finally, one ancestral HGT was inferred to be associated with the

association with *Tetrastigma*. To test this hypothesis, we further investigated a subset of 41 HGT orthogroups with ample taxon sampling in Vitaceae to identify the putative host lineage to

MRCA of *Tetrastigma* and *Cayratia* (Figure 5D), potentially reflecting the initial formation of the modern symbiotic interaction in stem group Rafflesiaceae. The differential host associations





**Figure 6. Codon Usage Bias in *Sapria himalayana***

(A) Comparison of codon usage (CU) in horizontally transferred genes (HGTs) and VGTs. For each gene, deviation from the average CU of *Manihot* (gray) and *Tetragstigma* (black) is quantified by the Measure Independent of Length and Composition.<sup>40</sup> Pairwise comparisons of *Manihot* and *Tetragstigma* are also included as controls.

(B) Frequency of optimal codons in HGTs and VGTs. The optimal synonymous codon is determined using ribosomal protein coding genes in *Sapria*.

implied by these HGTs demonstrate a dynamic history of host shifting and illustrates the utility of HGTs for assessing symbiotic interactions through time.

## Conclusions

*Sapria* demonstrates levels of genome modification that are unparalleled among plants and even most eukaryotes. Evolution in this lineage was rapid and likely occurred in a context of regressive evolution that tolerates mildly deleterious features, possibly as a result of small population sizes stemming from their extreme parasitic lifestyle and reproductive mode across tens of millions of years. In sharp contrast with other eukaryotic parasites, such as apicomplexans,<sup>45,46</sup> microsporidians,<sup>47</sup> and parasitic nematodes,<sup>48</sup> in which gene loss is coupled with varying degrees of genome size reduction, *Sapria* combines remarkable levels of gene loss with the maintenance of a genome that is substantially larger than their closest free-living relatives *Manihot* and *Populus*. This decoupling of gene content and expansion of genome size requires further exploration and highlights the value of investigating highly atypical species in the Tree of Life, which help to illuminate key insights about how our commonly held assumptions about genome architecture can be deeply modified.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- **METHOD DETAILS**
  - Taxon sampling and DNA extraction
  - Genome size estimation
  - Genome sequencing
  - Genome assembly
  - Genome assembly completeness assessment for *Sapria*
  - RNASeq and transcriptome assembly
  - Repeat masking and gene model prediction

- Orthogroup clustering and gene loss analysis
- Verification of gene loss in intergenic regions and pseudogene identification
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - Functional enrichment of missing orthogroups
  - Plastid genome loss
  - Gene expansion
  - Investigation of intron size
  - Functional enrichment analysis of highly compact genes
  - Intron expansion and selection pressure analysis
  - Genome scan assessment of HGT
  - Phylogenomic assessment of HGT
  - HGT validation with expanded taxon sampling
  - BLAST-based assessment of HGT
  - HGT validation using nanopore reads
  - Functional enrichment analysis of HGTs
  - Codon usage bias of HGTs
  - Modern and former host associations

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.cub.2020.12.045>.

## ACKNOWLEDGMENTS

We thank C. dePamphilis for valuable comments. We thank Y. Satoko for sharing genomic data of *Striga asiatica* and *Phtheirospermum japonicum*. For field support, we thank S.-Y. Wong, K.M. Wong, and the staff at Queen Sirikit Botanic Garden. For living plant material of *Tetragstigma*, we thank the University of Connecticut greenhouse. R. Hellmoss helped with graphic design. This work was largely conducted on the traditional territory of the Wampanoag and Massachusetts peoples; we also acknowledge the Iban, Semai, Jahai, and Hmong people, on whose traditional territory samples were collected for this project and other related studies in our group. This research was supported by National Science Foundation (NSF) ATOL grant (DEB-0622764), NSF grant DEB-1120243, and startup funds from Harvard University to C.C.D. Sawitree Sasirat passed away before the publication of this manuscript, but was instrumental in facilitating fieldwork and securing samples and permitting in Thailand; we dedicate this manuscript to her.

## AUTHOR CONTRIBUTIONS

C.C.D. conceived of the original idea for the project and supervised and worked to fund the research. L.C., B.J.A., T.B.S., and C.C.D. developed and

designed the analyses. C.C.D., S.M., S.S., and L.A.N. collected the plant material. L.C. extracted DNA for genome sequencing. N.P. conducted flow cytometry assessments (unless otherwise noted). C.B.H. sequenced the genomes. L.C., T.B.S., and D.E.K. assembled the genome. L.C. and D.E.K. annotated the repetitive elements. Z.X. and L.A.N. conducted the transcriptome sequencing and assembly. L.C. performed most analyses, including gene predictions, orthogroup clustering, and phylogenomic assessment of HGT. B.J.A. performed the genome scan assessment of HGT. L.C. and C.C.D. wrote the initial version of the manuscript; T.B.S., B.J.A., and S.M. provided substantial effort revising the manuscript; and final polishing was supported by previously listed authors plus L.A.N. All authors approved the final manuscript text.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 21, 2020

Revised: December 11, 2020

Accepted: December 23, 2020

Published: January 22, 2021

## REFERENCES

- Nikolov, L.A., and Davis, C.C. (2017). The big, the bad, and the beautiful: biology of the world's largest flowers. *J. Syst. Evol.* 55, 516–524.
- Nikolov, L.A., Tomlinson, P.B., Manickam, S., Endress, P.K., Kramer, E.M., and Davis, C.C. (2014). Holoparasitic Rafflesiaceae possess the most reduced endophytes and yet give rise to the world's largest flowers. *Ann. Bot.* 114, 233–242.
- Nais, J. (2001). *Rafflesia* of the World (Sabah Parks).
- Davis, C.C., Endress, P.K., and Baum, D.A. (2008). The evolution of floral gigantism. *Curr. Opin. Plant Biol.* 11, 49–57.
- Barkman, T.J., Klooster, M.R., Gaddis, K.D., Franzone, B., Calhoun, S., Manickam, S., Vessabutr, S., Sasirat, S., and Davis, C.C. (2017). Reading between the vines: hosts as islands for extreme holoparasitic plants. *Am. J. Bot.* 104, 1382–1389.
- Nickrent, D.L. (2020). Parasitic angiosperms: how often and how many? *Taxon* 69, 5–27.
- Molina, J., Hazzouri, K.M., Nickrent, D., Geisler, M., Meyer, R.S., Pentony, M.M., Flowers, J.M., Pelsner, P., Barcelona, J., Inovejas, S.A., et al. (2014). Possible loss of the chloroplast genome in the parasitic flowering plant *Rafflesia lagascae* (Rafflesiaceae). *Mol. Biol. Evol.* 31, 793–803.
- Xi, Z., Bradley, R.K., Wurdack, K.J., Wong, K., Sugumaran, M., Bomblies, K., Rest, J.S., and Davis, C.C. (2012). Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics* 13, 227.
- Xi, Z., Wang, Y., Bradley, R.K., Sugumaran, M., Marx, C.J., Rest, J.S., and Davis, C.C. (2013). Massive mitochondrial gene transfer in a parasitic flowering plant clade. *PLoS Genet.* 9, e1003265.
- Habib, S., Dang, V.-C., Ickert-Bond, S.M., Zhang, J.-L., Lu, L.-M., Wen, J., and Chen, Z.-D. (2017). Robust rhylogeny of *Tetrastigma* (Vitaceae) based on ten plastid DNA regions: implications for infragenetic classification and seed character evolution. *Front. Plant Sci.* 8, 590.
- Lu, H., Giordano, F., and Ning, Z. (2016). Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* 14, 265–279.
- Han, M.V., Thomas, G.W., Lugo-Martinez, J., and Hahn, M.W. (2013). Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* 30, 1987–1997.
- Sun, G., Xu, Y., Liu, H., Sun, T., Zhang, J., Hettenhausen, C., Shen, G., Qi, J., Qin, Y., Li, J., et al. (2018). Large-scale gene losses underlie the genome evolution of parasitic plant *Cuscuta australis*. *Nat. Commun.* 9, 2683.
- Vogel, A., Schwacke, R., Denton, A.K., Usadel, B., Hollmann, J., Fischer, K., Bolger, A., Schmidt, M.H.W., Bolger, M.E., Gundlach, H., et al. (2018). Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nat. Commun.* 9, 2515.
- Yoshida, S., Kim, S., Wafula, E.K., Tanskanen, J., Kim, Y.-M., Honaas, L., Yang, Z., Spallek, T., Conn, C.E., Ichihashi, Y., et al. (2019). Genome sequence of *Striga asiatica* provides insight into the evolution of plant parasitism. *Curr. Biol.* 29, 3041–3052.e4.
- Ng, S.-M., Lee, X.-W., Mat-Isa, M.-N., Aizat-Juhari, M.A., Adam, J.H., Mohamed, R., Wan, K.-L., and Firdaus-Raih, M. (2018). Comparative analysis of nucleus-encoded plastid-targeting proteins in *Rafflesia cantleyi* against photosynthetic and non-photosynthetic representatives reveals orthologous systems with potentially divergent functions. *Sci. Rep.* 8, 17258.
- Milborrow, B.V. (2001). The pathway of biosynthesis of abscisic acid in vascular plants: a review of the present state of knowledge of ABA biosynthesis. *J. Exp. Bot.* 52, 1145–1164.
- Li, J., Hettenhausen, C., Sun, G., Zhuang, H., Li, J.-H., and Wu, J. (2015). The parasitic plant *Cuscuta australis* is highly insensitive to abscisic acid-induced suppression of hypocotyl elongation and seed germination. *PLoS ONE* 10, e0135197.
- Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242–1245.
- Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J., et al. (2007). Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* 317, 1921–1926.
- Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., et al. (2008). Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. USA* 105, 14802–14807.
- Leushkin, E.V., Sutormin, R.A., Nabieva, E.R., Penin, A.A., Kondrashov, A.S., and Logacheva, M.D. (2013). The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. *BMC Genomics* 14, 476.
- Mourier, T., and Jeffares, D.C. (2003). Eukaryotic intron loss. *Science* 300, 1393.
- Jiang, K., and Goertzen, L.R. (2011). Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*). *BMC Res. Notes* 4, 52.
- Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002). Selection for short introns in highly expressed genes. *Nat. Genet.* 31, 415–418.
- Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584.
- Lynch, M. (2002). Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* 99, 6118–6123.
- Davis, C.C., and Wurdack, K.J. (2004). Host-to-parasite gene transfer in flowering plants: phylogenetic evidence from Malpighiales. *Science* 305, 676–678.
- Nickrent, D.L., Blarer, A., Qiu, Y.-L., Vidal-Russell, R., and Anderson, F.E. (2004). Phylogenetic inference in Rafflesiales: the influence of rate heterogeneity and horizontal gene transfer. *BMC Evol. Biol.* 4, 40.
- Barkman, T.J., McNeal, J.R., Lim, S.H., Coat, G., Croom, H.B., Young, N.D., and Depamphilis, C.W. (2007). Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evol. Biol.* 7, 248.
- Wen, J., Xiong, Z., Nie, Z.-L., Mao, L., Zhu, Y., Kan, X.-Z., Ickert-Bond, S.M., Gerrath, J., Zimmer, E.A., and Fang, X.-D. (2013). Transcriptome

sequences resolve deep relationships of the grape family. *PLoS ONE* 8, e74394.

32. One Thousand Plant Transcriptomes Initiative (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, 679–685.
33. Pelser, P.B., Nickrent, D.L., van Ee, B.W., and Barcelona, J.F. (2019). A phylogenetic and biogeographic study of *Rafflesia* (Rafflesiaceae) in the Philippines: Limited dispersal and high island endemism. *Mol. Phylogenet. Evol.* 139, 106555.
34. Davis, C.C., and Xi, Z. (2015). Horizontal gene transfer in parasitic plants. *Curr. Opin. Plant Biol.* 26, 14–19.
35. Yang, Z., Wafula, E.K., Kim, G., Shahid, S., McNeal, J.R., Ralph, P.E., Timilsena, P.R., Yu, W.B., Kelly, E.A., Zhang, H., et al. (2019). Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants. *Nat. Plants* 5, 991–1001.
36. Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129, 897–907.
37. Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32–42.
38. Bahir, I., Fromer, M., Prat, Y., and Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* 5, 311.
39. Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 151, 389–409.
40. Supek, F., and Vlahoviček, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics* 6, 182.
41. Stotz, H.U., Thomson, J.G., and Wang, Y. (2009). Plant defensins: defense, development and application. *Plant Signal. Behav.* 4, 1010–1012.
42. Sharma, N., Sharma, K.P., Gaur, R.K., and Gupta, V.K. (2011). Role of Chitinase in plant defense. *Asian J. Biochem.* 6, 29–37.
43. Chatterjee, A., Li, Y., Zhang, Y., Grove, T.L., Lee, M., Krebs, C., Booker, S.J., Begley, T.P., and Ealick, S.E. (2008). Reconstitution of *ThiC* in thiamine pyrimidine biosynthesis expands the radical SAM superfamily. *Nat. Chem. Biol.* 4, 758–765.
44. Vander Horn, P.B., Backstrom, A.D., Stewart, V., and Begley, T.P. (1993). Structural genes for thiamine biosynthetic enzymes (*thiCEFGH*) in *Escherichia coli* K-12. *J. Bacteriol.* 175, 982–992.
45. Katinka, M.D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P., et al. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453.
46. Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., et al. (2015). Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* 4, e06974.
47. Wadi, L., and Reinke, A.W. (2020). Evolution of microsporidia: an extremely successful group of eukaryotic intracellular parasites. *PLoS Pathog.* 16, e1008276.
48. Kikuchi, T., Eves-van den Akker, S., and Jones, J.T. (2017). Genome evolution of plant-parasitic nematodes. *Annu. Rev. Phytopathol.* 55, 333–354.
49. Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. *Genome Res.* 27, 757–767.
50. Coombe, L., Zhang, J., Vandervalk, B.P., Chu, J., Jackman, S.D., Birol, I., and Warren, R.L. (2018). ARKS: chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* 19, 234.
51. Warren, R.L., Yang, C., Vandervalk, B.P., Behsaz, B., Lagman, A., Jones, S.J.M., and Birol, I. (2015). LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *Gigascience* 4, 35.
52. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, arXiv:1303.3997v2. <https://arxiv.org/abs/1303.3997>.
53. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079.
54. Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32, 2103–2110.
55. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100.
56. Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9, e112963.
57. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* 27, 722–736.
58. Chakraborty, M., Baldwin-Brown, J.G., Long, A.D., and Emerson, J.J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* 44, e147.
59. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287.
60. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875.
61. Krueger, F. (2012). Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (reduced representation bisulfite-seq) libraries. [http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
62. Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123.
63. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421.
64. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* 27, 1691–1692.
65. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842.
66. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., and Zeng, Q. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652.
67. Haas, B., and Papanicolaou, A. (2012). Transcoder. <https://github.com/TransDecoder/TransDecoder>.
68. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., and Smit, A.F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* 117, 9451–9457.
69. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics Chapter 4*. Unit 4.10.
70. Campbell, M.S., Holt, C., Moore, B., and Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Curr. Protoc. Bioinformatics* 48, 4.11.1–4.11.39.
71. Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5, 59.
72. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439.



73. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121.
74. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157.
75. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2020). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* **48** (D1), D1104–D1113.
76. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
77. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534.
78. Armstrong, J., Hickey, G., Diekhans, M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., and Johnson, J. (2019). Progressive alignment with Cactus: a multiple-genome aligner for the thousand-genome era. *bioRxiv*, 730531.
79. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780.
80. Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612.
81. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973.
82. Stegemann, S., Keuthe, M., Greiner, S., and Bock, R. (2012). Horizontal transfer of chloroplast genomes between plant species. *Proc. Natl. Acad. Sci. USA* **109**, 2434–2438.
83. Zhang, H.-B., Zhao, X., Ding, X., Paterson, A.H., and Wing, R.A. (1995). Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184.
84. Doyle, J., and Doyle, J. (1990). Isolation of plant DNA from fresh tissue. *Focus* **12**, 39–40.
85. Dolezel, J., Binarova, P., and Lucretti, S. (1989). Analysis of nuclear DNA content in plant cells by flow cytometry. *Biol. Plant.* **31**, 113–120.
86. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770.
87. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.
88. Laine, V.N., Gossmann, T.I., van Oers, K., Visser, M.E., and Groenen, M.A.M. (2019). Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC Genomics* **20**, 19.
89. Cai, L., Xi, Z., Amorim, A.M., Sugumaran, M., Rest, J.S., Liu, L., and Davis, C.C. (2019). Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytol.* **221**, 565–576.
90. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* **47** (D1), D427–D432.
91. Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., et al. (2014). Data access for the 1,000 Plants (1KP) project. *GigaScience* **3**, 17.
92. Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., et al. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524.
93. Mower, J.P., Stefanović, S., Hao, W., Gummow, J.S., Jain, K., Ahmed, D., and Palmer, J.D. (2010). Horizontal acquisition of multiple mitochondrial genes from a parasitic plant followed by gene conversion with host mitochondrial genes. *BMC Biol.* **8**, 150.
94. Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W., Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S., and Stevens, P.F.; The Angiosperm Phylogeny Group (2016). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20.
95. Magallón, S., and Castillo, A. (2009). Angiosperm diversification through time. *Am. J. Bot.* **96**, 349–365.
96. Yang, Y., and Smith, S.A. (2014). Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* **31**, 3081–3092.
97. Hickey, G., Paten, B., Earl, D., Zerbino, D., and Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342.
98. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.
99. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermini, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589.
100. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522.
101. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.
102. Roney, J.K., Khatibi, P.A., and Westwood, J.H. (2007). Cross-species translocation of mRNA from host plants into the parasitic plant dodder. *Plant Physiol.* **143**, 1037–1043.
103. Westwood, J.H., Roney, J.K., Khatibi, P.A., and Stromberg, V.K. (2009). RNA translocation between parasitic plants and their hosts. *Pest Manage. Sci.* **65**, 533–539.
104. Kim, G., LeBlanc, M.L., Wafula, E.K., dePamphilis, C.W., and Westwood, J.H. (2014). Plant science. Genomic-scale exchange of mRNA between a parasitic plant and its hosts. *Science* **345**, 808–811.
105. Elek, A., Kuzman, M., and Vlahoviček, K. (2019). coRdon: codon usage analysis and prediction of gene expressivity. *Bioconductor* 3.8. <https://github.com/BioinfoHR/coRdon>.

## STAR★METHODS

## KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Deposited Data</b>		
Raw sequencing reads and assemblies for <i>Sapria himalayana</i> Griff., <i>Tetrastigma voinierianum</i> Pierre ex Pit., <i>Rafflesia tuan-mudae</i> Becc., and <i>Rhizanthus zippelii</i> (Blume) Spach	This paper	NCBI GenBank BioProject: PRJNA686196
Illumina reads of transcriptomes from <i>Sapria himalayana</i> Griff. and <i>Rafflesia cantleyi</i> Solms-Laubach	Xi et al., 2012 <sup>8</sup>	NCBI GenBank: SRA052224
Genome assemblies from 33 angiosperm species	GenBank, Phytozome v13, etc.	See details in <a href="#">Data S1J</a>
Illumina reads of transcriptomes from 15 Vitaceae species	Wen et al., 2013 <sup>31</sup>	NCBI GenBank: SRA081731
<b>Experimental Models: Organisms/Strains</b>		
<i>Sapria himalayana</i> Griff., <i>Rafflesia tuan-mudae</i> Becc., and <i>Rhizanthus zippelii</i> (Blume) Spach	Wild populations collected from Sarawak, Malaysia, and Thailand	Collection permission from the Controller of National Parks and Nature Reserves, Sarawak Forestry Department (NCCD.907.4.4(Jld.VI)-52 and Park permit No. 25/2011) and the National Research Council of Thailand (NCRT No. 00028498)
<i>Tetrastigma voinierianum</i> Pierre ex Pit.	Living collection from the University of Connecticut greenhouse	Accession# 199200473
<b>Software and Algorithms</b>		
SuperNova v2.1.1	Weisenfeld et al. <sup>49</sup>	<a href="https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome">https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome</a>
ARKS v1.0.3	Coombe et al. <sup>50</sup>	<a href="https://github.com/bcgsc/arks">https://github.com/bcgsc/arks</a>
LINKS v1.8.5	Warren et al. <sup>51</sup>	<a href="https://github.com/bcgsc/LINKS">https://github.com/bcgsc/LINKS</a>
BWA v0.7.17	Li <sup>52</sup>	<a href="http://bio-bwa.sourceforge.net/bwa.shtml">http://bio-bwa.sourceforge.net/bwa.shtml</a>
SAMtools v1.3.1	Li et al. <sup>53</sup>	<a href="http://www.htslib.org/">http://www.htslib.org/</a>
miniasm v0.2	Li et al. <sup>54</sup>	<a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a>
minimap2 v2.9	Li <sup>55</sup>	<a href="https://github.com/lh3/miniasm">https://github.com/lh3/miniasm</a>
Pilon v1.18	Walker et al. <sup>56</sup>	<a href="https://github.com/broadinstitute/pilon">https://github.com/broadinstitute/pilon</a>
CANU v1.3	Koren et al. <sup>57</sup>	<a href="https://github.com/marbl/canu">https://github.com/marbl/canu</a>
Quickmerge	Chakraborty et al. <sup>58</sup>	<a href="https://github.com/mahulchak/quickmerge">https://github.com/mahulchak/quickmerge</a>
clusterProfiler	Yu et al. <sup>59</sup>	<a href="https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html">https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html</a>
GMAP v2019.06.10	Wu et al. <sup>60</sup>	<a href="https://github.com/juliangehring/GMAP-GSNAP">https://github.com/juliangehring/GMAP-GSNAP</a>
TrimGalore v0.5.0	Krueger et al. <sup>61</sup>	<a href="https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/">https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/</a>
Abyss v2.0.2	Simpson et al. <sup>62</sup>	<a href="https://github.com/bcgsc/abyss">https://github.com/bcgsc/abyss</a>
BLAST v2.2.29	Camacho et al. <sup>63</sup>	<a href="https://blast.ncbi.nlm.nih.gov/Blast.cgi">https://blast.ncbi.nlm.nih.gov/Blast.cgi</a>
BamTools v2.3.0	Barnett et al. <sup>64</sup>	<a href="https://github.com/pezmaster31/bamtools">https://github.com/pezmaster31/bamtools</a>
BEDTool2 v2.26.0	Quinlan et al. <sup>65</sup>	<a href="https://github.com/arq5x/bedtools2">https://github.com/arq5x/bedtools2</a>
Trinity v2.6.6	Grabherr et al. <sup>66</sup>	<a href="https://github.com/trinityrnaseq/trinityrnaseq/wiki">https://github.com/trinityrnaseq/trinityrnaseq/wiki</a>
Transdecoder v5.3.0	Haas and Papanicolaou <sup>67</sup>	<a href="https://github.com/TransDecoder/TransDecoder/wiki">https://github.com/TransDecoder/TransDecoder/wiki</a>
RepeatModeler2 v1.0.11	Flynn et al. <sup>68</sup>	<a href="http://www.repeatmasker.org/RepeatModeler/">http://www.repeatmasker.org/RepeatModeler/</a>
RepeatMasker v4.0.8	Tarailo-Graovac et al. <sup>69</sup>	<a href="http://www.repeatmasker.org/">http://www.repeatmasker.org/</a>

(Continued on next page)

**Continued**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MAKER v2.31.10	Campbell et al. <sup>70</sup>	<a href="https://www.yandell-lab.org/software/maker.html">https://www.yandell-lab.org/software/maker.html</a>
SNAP v2006-07-28	Korf <sup>71</sup>	<a href="https://github.com/KorfLab/SNAP">https://github.com/KorfLab/SNAP</a>
AUGUSTUS v3.3	Stanke et al. <sup>72</sup>	<a href="http://bioinf.uni-greifswald.de/augustus/">http://bioinf.uni-greifswald.de/augustus/</a>
HMMER v3.2.1	Mistry et al. <sup>73</sup>	<a href="http://hmmer.org/">http://hmmer.org/</a>
OrthoFinder v2.2.7	Emms et al. <sup>74</sup>	<a href="https://github.com/davidemms/OrthoFinder">https://github.com/davidemms/OrthoFinder</a>
CAFÉ v4	Han et al. <sup>12</sup>	<a href="https://hahnlab.github.io/CAFE/src_docs/html/index.html">https://hahnlab.github.io/CAFE/src_docs/html/index.html</a>
PlantRegMap	Tian et al. <sup>75</sup>	<a href="http://plantregmap.gao-lab.org/">http://plantregmap.gao-lab.org/</a>
KEGG	Kyoto Encyclopedia of Genes and Genomes	<a href="https://www.genome.jp/kegg/">https://www.genome.jp/kegg/</a>
PAML v4.8	Yang <sup>76</sup>	<a href="http://abacus.gene.ucl.ac.uk/software/paml.html">http://abacus.gene.ucl.ac.uk/software/paml.html</a>
IQTREE v2.0.5	Minh et al. <sup>77</sup>	<a href="http://www.iqtree.org/">http://www.iqtree.org/</a>
Cactus v1.1.0	Armstrong et al. <sup>78</sup>	<a href="https://github.com/ComparativeGenomicsToolkit/cactus">https://github.com/ComparativeGenomicsToolkit/cactus</a>
MAFFT v.7.299	Katoh et al. <sup>79</sup>	<a href="https://mafft.cbrc.jp/alignment/software/">https://mafft.cbrc.jp/alignment/software/</a>
pal2nal v.14	Suyama et al. <sup>80</sup>	<a href="http://www.bork.embl.de/pal2nal/">http://www.bork.embl.de/pal2nal/</a>
trimAL	Capella-Gutierrez et al. <sup>81</sup>	<a href="http://trimal.cgenomics.org/">http://trimal.cgenomics.org/</a>
coRdon	Stegemann et al. <sup>82</sup>	<a href="https://github.com/BioinfoHR/coRdon">https://github.com/BioinfoHR/coRdon</a>
Other custom python, R, and bash scripts	This paper	<a href="https://github.com/lmcai/Sapria_genomics">https://github.com/lmcai/Sapria_genomics</a>

## RESOURCE AVAILABILITY

### Lead Contact

Further information and requests for resources should be directed to Charles C. Davis ([cdavis@oeb.harvard.edu](mailto:cdavis@oeb.harvard.edu)).

### Materials Availability

The Illumina and nanopore sequencing data for *Sapria himalayana*, *Tetrastigma voinierianum*, *Rafflesia tuan-mudae*, and *Rhizanthus zippelii* have been deposited to the NCBI Sequence Read Archive (SRA) under GenBank Bioproject PRJNA686196. The genome assemblies and gene annotations of *Sapria* and *Tetrastigma* have been deposited to GenBank Bioproject PRJNA686196. All code, including command lines and custom scripts, used in the current study is deposited in GitHub repository ([https://github.com/lmcai/Sapria\\_genomics](https://github.com/lmcai/Sapria_genomics)).

### Data and Code Availability

The Illumina and nanopore sequencing data for *Sapria himalayana*, *Tetrastigma voinierianum*, *Rafflesia tuan-mudae*, and *Rhizanthus zippelii* have been deposited to the NCBI Sequence Read Archive (SRA) under GenBank Bioproject PRJNA686196. The genome assemblies and gene annotations of *Sapria* and *Tetrastigma* have been deposited to GenBank Bioproject PRJNA686196. All code, including command lines and custom scripts, used in the current study is deposited in GitHub repository ([https://github.com/lmcai/Sapria\\_genomics](https://github.com/lmcai/Sapria_genomics)).

## METHOD DETAILS

### Taxon sampling and DNA extraction

Floral material of *Sapria himalayana* Griff. (referred to as *Sapria* hereafter) was gathered for DNA extraction from wild populations residing within the Queen Sirikit Botanic Garden with permission from the National Research Council of Thailand (NCRT No. 00028498). Voucher specimens of these materials are accessioned at the Harvard University Herbaria. The floral material was flash-frozen in the field and stored at  $-80^{\circ}\text{C}$  prior to extraction. The frozen plant material was carefully dissected before DNA extraction to retain only innermost floral tissue devoid of host material. High molecular weight (HMW) DNA extractions were conducted following a modified nuclei enrichment protocol from Zhang et al.<sup>83</sup> using 2.1 g frozen tissue. This enrichment effectively reduced the abundance of organellar genomes and removes secondary metabolites. DNA was subsequently extracted from enriched nuclei using the CTAB method<sup>84</sup>. The resulting HMW DNA was then cleaned using KAPA pure beads (Roche, Indiana, USA) at 0.5 X volume and treated using the Short Read Eliminator Kit (Circulomics Inc., Maryland, USA) to remove short length fragments. The complete protocol of this HMW DNA extraction is provided in [Data S3](#).



HMW DNA from the host species, *Tetrastigma voinierianum* (Sallier) Pierre ex Gagnep. (referred to as *Tetrastigma* hereafter), was extracted from ca. 1.5 g fresh leaf material collected from the University of Connecticut greenhouse (Accession# 199200473) following the protocol above.

### Genome size estimation

***Sapria* genome size estimation using flow cytometry**—The nuclear genome size of *Sapria* was estimated based on flow cytometry and k-mer assessments. Flow cytometry estimation was conducted using three field frozen samples from a single population of *Sapria*. Nuclei were isolated from one gram of tissue from each sample. Tissues were carefully dissected with a razor blade in 1.2 mL of LB01 buffer<sup>85</sup> and treated with RNase (50 µg/ml) on ice for one minute. The resulting solution was passed through a 30 µm filter to retain only nuclei for characterization. The flowthrough was stained with Sybr Green I (5 µg/ml; Sigma Aldrich, Darmstadt, Germany) and incubated on ice in the dark for 30 minutes. Tissues from *Hordeum vulgare* Morex were similarly dissected and prepared to be used as both external and internal standards. Samples were analyzed using a Coulter Epics XL flow cytometer (Beckman Coulter Genomics), equipped with a UV lamp. Genome size was estimated using the following formula: Sample 2C value (DNA pg) = Reference 2C value x (sample 2C mean peak position/reference 2C mean peak position). The 2C value for the genome was estimated to be  $6.24 \pm 0.189$  pg. Detailed results are reported in Data S1A and Figure S1.

***Sapria* genome size estimation using k-mer distribution**—We used Jellyfish v.2.2.10<sup>86</sup> to count k-mers for the Illumina data using k-mer sizes of 20, 27, and 35. The size of the genome and single-copy regions were subsequently determined following the methods described in the genome size estimation tutorial from the University of Connecticut (<https://bioinformatics.uconn.edu/genome-size-estimation-tutorial/>). The genome size of *Sapria* was estimated to be 1.69–2.54 Gb based on the k-mer distribution, and the size of single-copy region was estimated to be 470–940 Mb (Figure S1).

***Tetrastigma* genome size estimation using flow cytometry**—The total nuclear genome size of *Tetrastigma* was estimated based on flow cytometry at the Flow Cytometry Core Lab at the Benaroya Research Institute (Seattle, WA, USA). The 2C value for the genome was estimated to be  $5.00 \pm 0.128$  pg. Detailed results are reported in Data S1A and Figure S1.

### Genome sequencing

***Sapria* genome sequencing**—We combined linked-read Chromium genome sequencing from 10X genomics and long-read sequencing from Oxford NanoPore Technologies (ONT) to assemble the genome of *Sapria*. HMW DNA from a single *Sapria* individual was used for 10X genomic library construction at the Bauer Core Facility at Harvard University (<https://bauercore.fas.harvard.edu/>). The 10X genomic library was prepared following the manufacturer's protocols and sequenced using the Illumina HiSeq 2500 System on a high output v4 flow cell using paired end 125 bp reads. The Illumina RTA v1.18.54 was used for basecalling and the generation of fastq files with deactivated adaptor trimming. A total of 162.5 Gb from 1.30 billion Illumina reads were obtained for this effort. For nanopore sequencing, DNA libraries from three individuals were prepared using the ligation sequencing kit SQK-LSK109 following the manufacturer's protocols and sequenced using ONT's MinION instrument at the Bauer Core Facility at Harvard University. A total of 36.5 Gb from 1.1 million reads were obtained. The N50 of the raw nanopore reads was 38 kb.

***Tetrastigma* genome sequencing**—HMW DNA from a single individual of the host species was used for nanopore library preparation. The resulting library was sequenced following the nanopore protocols described above. A total of 45.8 Gb from 3.0 million nanopore reads were obtained (N50 = 45 kb). We additionally constructed an Illumina paired-end genome sequencing library using the KAPA HyperPlus Kit (Roche, Indiana, USA) to facilitate assembly polishing. The Illumina library was sequenced on the Illumina NovaSeq Platform using an SP flowcell and paired end 150 bp reads. A total of 152.3 Gb from 1.22 billion Illumina reads were obtained.

### Genome assembly

***Genome assembly of Sapria***—Illumina reads from the 10X library were first barcoded using Long Ranger and assembled using SuperNova v2.1.1<sup>49</sup>. The resulting assembly was further scaffolded with the barcoded Illumina reads using ARKS v1.0.3<sup>50</sup>. The long-read nanopore data were additionally used to build superscaffolds with LINKS v1.8.5<sup>51</sup> by iterative scaffolding. After 18 rounds of scaffolding with increased distance between k-mer pairs from 1 kb to 40 kb (e.g., -d 1000), no further improvements were detected. The scaffold N50 of the final assembly was 4.3 Mb. Illumina reads were mapped to the final assembly using BWA v0.7.17<sup>52</sup> and the resulting bam file was sorted by SAMtools v1.3.1<sup>53</sup>. 99.0% of reads were mapped back to the assembly.

***Genome assembly of Tetrastigma***—We constructed two nanopore *de novo* assemblies of *Tetrastigma* to facilitate the subsequent detection of horizontally transferred genes using our phylogenomic and genome scan analyses, respectively (Figure S3). For our phylogenomic assessments, we aimed for a more complete but potentially fragmented assembly for better gene sampling. For our genome scan assessment, we aimed for a more continuous but less complete assembly. Nanopore reads were filtered prior to assembly to remove reads shorter than 10 kb.

To generate a more complete assembly of *Tetrastigma* for our phylogenomic analysis, we used the minimap-miniasm nanopore *de novo* assembly pipeline<sup>54</sup>. An all-by-all alignment of filtered nanopore reads was generated using minimap2 v2.9<sup>55</sup> under the nanopore reads alignment settings (-ax map-ont). We then used miniasm v0.2<sup>54</sup> to generate the assembly and polished it iteratively three times with Illumina reads using Pilon v1.18<sup>56</sup>. The final assembly was 3.11 Gb in size with an N50 value of 495 kb. We assessed the completeness of this genome using the 1440 plant Benchmarking Universal Single-Copy Orthologs (BUSCOs)<sup>87</sup> and identified 86.1% of these BUSCOs. We used this assembly for subsequent gene annotation and phylogenomic analysis.

To generate a *Tetrastigma* assembly with better contiguity that is best suited for our genome scan analysis, we created a second nanopore *de novo* assembly using CANU v1.3<sup>57</sup> and merged it with the Pilon-polished miniasm assembly described above. To generate the CANU assembly, the nanopore reads were first corrected, trimmed, and assembled using CANU. We set the genome size parameter to ‘2500 m’ based on our flow cytometry results (Data S1A; Figure S1) and set the reads error rate parameter to be 0.025 to accommodate the more error-prone nanopore reads. We then created a merged consensus assembly using the Quickmerge program<sup>58</sup> with two iterative rounds of merging. The first round used the Pilon-polished miniasm assembly as the reference and the CANU assembly as the query, with the following parameters: -hco 5.0 -c 1.5 -l 495000 -ml 20000. The second round used the Pilon-polished assembly as a reference with the merged assembly from round one as a query with the following parameters: -hco 5.0 -c 1.5 -l 613000 -ml 20000. The final merged assembly was 2.80 Gb in size with an N50 value of 736 kb. 79.1% of the plant BUSCOs were identified in this assembly.

### Genome assembly completeness assessment for *Sapria*

**BUSCO assessment**—The completeness of our *Sapria* assembly was assessed by mapping against 303 eukaryotic BUSCOs and 1,440 plant BUSCOs. Our assembly contained 84.5% of conserved eukaryotic BUSCOs but only 44.6% of the plant BUSCOs (Figure 2). To further determine whether the 798 undetectable plant BUSCOs were enriched in certain functional categories, we conducted a Gene Ontology (GO) analysis of these missing plant BUSCOs using the R package clusterProfiler<sup>59</sup>. For each plant BUSCO, we determined the best matching ortholog from *Arabidopsis thaliana* by BLAST. Of the 1,440 total BUSCOs assessed, 1,438 of them had significant hits to *Arabidopsis*. GO analysis of missing plant BUSCOs were subsequently assessed using a custom background of 1,438 *Arabidopsis* orthologs based on Fisher’s Exact Test in combination with the False Discovery Rate correction and a *p*-value threshold of 0.01. A total of 109 GO terms, mostly related to photosynthesis and plastid organization, were significantly enriched (*p*-value < 0.01; Data S1N).

To test whether functional biases in missing plant BUSCOs might result from an incomplete genome, we subsampled the contigs of the chromosomal-level genome assembly of the free-living species *Manihot esculenta* Crantz to simulate 64% missing data. This percentage was derived from the estimated percentage of missing data in our 1.28 Gb *Sapria* assembly, assuming a maximal genome size of 3.5 Gb based on flow cytometry. The subsampled *Manihot* genome had an equivalent amount of missing BUSCOs (56%), however, only two GO terms related to ribosome binding were significantly enriched (*p*-value < 0.01). This level of enrichment in missing BUSCOs was far less than the 109 observed in *Sapria* (see above). To test the statistical significance of this result, we generated 1000 sets of randomly subsampled *Arabidopsis* BUSCOs, each containing 798 (55.4%) genes, and conducted GO enrichment analyses using clusterProfiler. The median number of enriched GO terms from the simulation was two and the maximum number was 69. This result demonstrates that while incomplete assembly can, as expected, lead to a large fraction of missing BUSCOs, the marked functional bias revealed by GO enrichment analysis in *Sapria* is highly unlikely to arise from incomplete assembly.

**Reads mapping assessment to determine genome completeness**—We mapped Illumina reads, nanopore reads, and transcriptomes to our *Sapria* assembly using BWA, minimap2, and GMAP v2019.06.10<sup>60</sup>, respectively. The resulting .bam files were sorted using SAMtools, and 99.0%, 97.0%, and 99.2% reads could be mapped, respectively. In combination with our simulation results above, these very high mapping levels further confirm the completeness of our assembly. To explore the 6.59 million unmapped Illumina reads, we generated *de novo* assemblies for these reads following the methods described by Laine et al.<sup>88</sup>. Briefly, unmapped reads were extracted using SAMtools and filtered using TrimGalore v0.5.0<sup>61</sup> to remove reads shorter than 36 bp or those with quality scores lower than 5 (–q 5–length 36). Filtered reads were assembled with Abyss v2.0.2<sup>62</sup> using a k-mer size of 20. The resulting assembly was 114 Mb and the N50 was 607 bp. A total of 109.3 kb units were longer than 300 bp and were compared to the NCBI non-redundant nucleotide database using BLAST v2.2.29<sup>63</sup> to identify the closest known matching sequence. Eighty contigs were mapped to reference genes in NCBI with E-values less than 1e-5 (Data S1O). These contigs were most frequently mapped to genes from *Vitis vinifera*, a close relative of the host, suggesting possible low levels of host DNA in the parasite (see section HGT validation using nanopore reads to address this concern).

**Estimating the size of single-copy regions**—For *Sapria*, we measured the size of the single-copy regions in the assembly and compared it to the size of the single-copy regions in the genome estimated from the k-mer distribution. Single-copy regions were defined as DNA sequences present in single (or low) copy in the genome, which primarily constitute coding sequences for structural genes. To summarize the single-copy regions in the assembly, read coverage for each site was calculated using BamTools v2.3.0<sup>64</sup>. The median read coverage was 60.2 X (Figure S2B). We then used BEDTool2 v2.26.0<sup>65</sup> to aggregate regions with coverage less than 120 X (twice the average read coverage) to determine the size of single-copy regions. This led to an estimation of 904 Mb of single-copy sequences in our assembly. We next compared this size estimate to our estimate based on the k-mer distribution (see section Genome size estimation above). Using multiple k-mer sizes, we estimated the single-copy region of the genome to be 470–940 Mb (Figure S1). These estimations suggest that between 96.2% (904/940) to 100% of the single-copy regions are assembled.

### RNASeq and transcriptome assembly

To expand our taxon sampling within Rafflesiaceae for gene model prediction and downstream analysis of horizontal gene transfer (HGT), we generated two new transcriptomes from *Rafflesia tuan-mudae* Becc. and *Rhizanthus zippelii* (Blume) Spach and included previously published RNASeq data from *Sapria himalayana* and *Rafflesia cantleyi* Solms-Laubach in Xi et al.<sup>8</sup>. Total RNA extraction and cDNA library preparation for *R. tuan-mudae* and *R. zippelii* followed Xi et al.<sup>8</sup>. These species were collected and field frozen under

Sarawak Forest Department permitting number NCCD.907.4.4(Jld.VI)-52 and Park permitting number 25/2011. Total RNA was extracted and synthesized to cDNA libraries. The resulting cDNA libraries were sequenced using the Genome Analyzer II (Illumina, Inc.) with paired-end 150 bp read lengths at the Bauer Core Facility at Harvard University. Transcriptome assembly and coding sequence prediction followed Cai et al.<sup>69</sup> using Trinity v2.6.6<sup>66</sup> and Transdecoder v5.3.0<sup>67</sup>.

### Repeat masking and gene model prediction

*Annotation of Sapria*—RepeatModeler2 v1.0.11<sup>68</sup> was used to generate a species-specific repeat library for *Sapria*. RepeatMasker v4.0.8<sup>69</sup> was subsequently applied to annotate and mask assemblies based on the species-specific repeat libraries. Following masking, gene models were inferred using MAKER with RNA-seq and protein evidence. For RNA-seq evidence, we used published and newly generated transcripts from the four species of Rafflesiaceae listed above (*Sapria himalayana*, *Rafflesia cantleyi*, *Rafflesia tuan-mudji* and *Rhizanthus zippelii*), complete proteomes from three published eurosid genomes (*Vitis vinifera*, *Manihot esculenta*, and *Populus trichocarpa*), and the UniProtKB/Swiss-Prot database (download date Oct 9, 2019). Protein-coding genes were subsequently predicted using three iterative rounds of MAKER v2.31.10<sup>70</sup> each involving: (i) creation of initial gene models from transcript and homologous proteins using est2genome and protein2genome; (ii) refinement of this initial model with *ab initio* gene predictors SNAP v2006-07-28<sup>71</sup> and AUGUSTUS v3.3<sup>72</sup>; and (iii) final models generated using SNAP trained on gene models output from step (ii). We used the default parameters in MAKER except that we adjusted the 'split\_hit' parameter to 100,000 to accommodate the long introns in *Sapria* (see also main text). A total of 55,179 genes were predicted by MAKER. Putative functions for each gene were inferred by comparing their coding sequences to the Pfam protein domain database v33<sup>90</sup> using HMMER v3.2.1<sup>73</sup>. A total of 12,667 genes do not have significant Pfam hits (E-value < 1e-5) and exhibit significant sequence similarity (BLASTn E-value < 1e-5) to the species-specific repeat library. These gene models are identified as low confidence annotations. The remaining 42,512 gene models were validated by either transcriptome, known plant proteins, or the Pfam database.

*Annotation of Tetrastigma*—Repeat annotation and gene model prediction for *Tetrastigma voinierianum* followed the same pipeline as outlined above for *Sapria*. The Pilon-polished miniasm assembly of *Tetrastigma* was used for *de novo* repeat content identification using RepeatModeler. For MAKER gene prediction, we used the published transcriptomes from *Tetrastigma voinierianum* and *Tetrastigma obtectum*<sup>91</sup>, the proteomes from *Vitis vinifera*, and the UniProtKB/Swiss-Prot database to facilitate the initial round of evidence-based gene prediction. We then applied two rounds of *ab initio* gene prediction as described above to refine gene models. The final annotation contains 49,376 gene models.

### Orthogroup clustering and gene loss analysis

Orthogroups for investigating gene loss were created with OrthoFinder v2.2.7<sup>74</sup> from predicted proteins of *Sapria* and complete proteomes of 34 sequenced flowering plant genomes (Figure S4; Data S1J). We carefully selected taxa representing all major flowering plant clades, including *Amborella trichopoda*, the sister group of all other angiosperms, and *Cinnamomum micranthum* (a magnoliid dicot), two monocots (*Oryza japonica* and *Sorghum bicolor*), two early diverging eudicots representatives (*Aquilegia coerulea* and *Nelumbo nucifera*), sixteen asterids, and eleven eurosids, including three free-living relatives of *Sapria* in Malpighiales (*Jatropha curcas*, *Manihot esculenta* and *Populus trichocarpa*) plus one confamilial relative of the host species (*Vitis vinifera*). We identified 10,880 orthogroups present in at least three of the five eurosid species, *Arabidopsis thaliana*, *Populus trichocarpa*, *Manihot esculenta*, *Gossypium raimondii*, and *Glycine max*. We considered these orthogroups to be conserved across eurosids and subsequently used them for assessing gene loss in *Sapria*. Gene copy number for each species for the 10,880 conserved orthogroup cluster is reported in Data S2A.

As a comparison, we also examined gene loss in two independently evolved parasitic plant species: *Cuscuta australis* and *Striga asiatica*. Because these two species belong to the asterid clade, we first identified conserved orthogroups across asterids by similarly requiring at least three of the five asterid species, *Mimulus guttatus* DC., *Solanum tuberosum* L., *Ipomoea nil* (L.) Roth, *Coffea canephora* Pierre ex A.Froehner, and *Helianthus annuus* L., to be present in these orthogroups. We identified a total of 10,687 conserved asterid orthogroups, within which 999 (9.3%) were missing in *Striga* and 1,580 (15.7%) were missing in *Cuscuta*.

### Verification of gene loss in intergenic regions and pseudogene identification

To address concerns of spurious gene loss due to annotation quality, we searched protein sequences from *Arabidopsis thaliana*, *Manihot esculenta*, *Populus trichocarpa*, and *Vitis vinifera* against the intergenic regions in our *Sapria* genome assembly. Intergenic regions were extracted using BEDTools. Proteomes from the above species were aligned against the intergenic sequences using tblastn with an E-value threshold of 1e-10. Among the 4,828 missing conserved orthogroups, 759 have BLAST hits to the intergenic regions. These BLAST hits contain paralogs, pseudogene fragments, and gene models that were not annotated. This small percentage of additional possible orthogroups (7.0% of the total conserved orthogroups) within intergenic regions does not diminish the scale of reported gene loss in *Sapria*.

We then used these tblastn results to characterize pseudogenes using the MAKER-P protocol<sup>92</sup>. This wrapper script was run with default settings except that the intron length threshold was set to 100 kb to accommodate long introns in *Sapria* (see main text and the section [Intron expansion and selection pressure analysis](#) below). Because these pseudogenes were included in our phylogenomic analysis, we used a custom script to filter them by length (> 150 bp) and gene collinearity. A total of 71,747 pseudogenes met these criteria.



## QUANTIFICATION AND STATISTICAL ANALYSIS

### Functional enrichment of missing orthogroups

We performed a GO analysis using PlantRegMap<sup>75</sup> to determine enriched terms in the missing orthogroups of *Sapria*. For the 4,828 missing orthogroups we analyzed, orthologs from *Arabidopsis* were selected as the foreground gene set. The background gene set contained *Arabidopsis* orthologs from the 10,880 conserved orthogroups. A *p*-value threshold of 0.05 was used to identify significantly enriched terms based on Fisher's Exact Test in combination with the False Discovery Rate correction. A less stringent *p*-value is used here to capture as many statistically significant terms as possible for result discussion. A total of 570 terms in biological processes, cellular components, and molecular functions were significantly enriched (*p*-value < 0.05) and 374 terms have *p*-value < 0.01 (Data S1E). The R package clusterProfiler and enrichplot<sup>59</sup> were used to visualize our results (Figure 2). Our GO enrichment result remained consistent when excluding the 759 orthogroups that had BLAST hits in intergenic regions, although with slightly less enriched terms (*n* = 538, *p* value < 0.05; Data S1P). We also used the orthologs from *Glycine*, *Manihot*, and *Populus* to assess the generality of these findings. The results were also highly consistent among different species (Data S1Q–S1S). To identify metabolic networks that would be impacted by the missing genes, orthologs from *Arabidopsis* representing all missing conserved orthogroups in *Sapria* were searched against the KEGG pathway maps using the KEGG Mapper webserver (<https://www.genome.jp/kegg/mapper.html>).

### Plastid genome loss

To test the hypothesis that the plastid genome was lost in Rafflesiaceae<sup>7</sup>, we compared the reference plastid genome from *Manihot esculenta* (GenBank NC\_010433.1) against the *Sapria* assemblies using BLASTn with an E-value threshold of 1e-5. A total of 290 scaffolds contained plastid-like sequences and were selected for further investigation. Their lengths, GC contents, and average Illumina read coverages are reported in Data S1F. None of the scaffolds were localized to the plastid genome assuming a variety of standards applied to assess the presence of a plastid genome<sup>7,93</sup>: plastid genome size < 200 kb, GC% < 38%, and read coverage > 200 X (nuclear genome median coverage 60.2 X). In addition, all of the putative plastid gene fragments clade with non-Malpighiales species. The nearly complete loss of nuclear genes that regulate plastid functions further support the hypothesis that the plastid genome is indeed lost in Rafflesiaceae.

### Gene expansion

To investigate orthogroup expansion in *Sapria*, we applied a modified birth-death model implemented in CAFE v4, which accommodates errors in orthogroup size estimation<sup>12</sup>. We generated a dated phylogeny for the 29 eudicot genomes based on APG IV<sup>94</sup>, Magallón et al.<sup>95</sup>, and Pelser et al.<sup>33</sup>. Orthogroups were first filtered using the python script 'cafetutorial\_clade\_and\_size\_filter.py' from CAFE to remove orthogroups containing more than 100 copies per species. We then inferred the ancestral orthogroup size for each lineage under the free lambda model. We subsequently performed a GO analysis on the 710 orthogroups that expanded in *Sapria* after its divergence from other Malpighiales species using PlantRegMap<sup>75</sup>. We used orthologs from *Arabidopsis* in these expanded orthogroups as the foreground gene set. A *p*-value threshold of 0.01 was used to identify significantly enriched terms based on Fisher's Exact Test in combination with the False Discovery Rate correction. The GO enrichment result is reported in Data S1C.

### Investigation of intron size

We used the evidence-based annotation from the first iteration of the MAKER genome annotation to characterize intron size in *Sapria*. We chose this annotation because it is based on the alignment of transcripts and proteins to the assembly, and therefore represents the most accurate gene structure prediction. Furthermore, the distribution of intron length from this evidence-based annotation matches very well with that estimated from the uniquely mapped *Sapria* transcripts (Figure S5A), which represents the most unbiased annotation of intron positions. Other versions of the annotation, with *ab initio* evidence from SNAP and AUGUSTUS, are not appropriate for this purpose because both programs cannot annotate long introns above 10 kb (Figure S5A).

We used a custom python script (available on Github, see code availability statement) to extract the number, length, and position of introns from the genome annotation gff file. Both the mean intron number of 3.1 per gene and the median protein length of 265 amino acids (AA) are smaller than those in other plants (Figure S2). To alleviate concerns of biased estimation of intron number due to truncated gene annotation, we compared the protein length for each orthogroup of *Sapria* to the average protein length of the remaining 34 non-Rafflesiaceae species sampled in our orthogroup clustering (Figure S5B; Data S1J). Our results confirmed the completeness of genes in *Sapria* (Figure S5B).

To characterize intron size and presence in *Sapria*, we leveraged the cross-species protein alignments from the evidence-based MAKER annotation and focused especially on protein alignments from *Manihot* and *Populus*, which represent the closest free-living relatives of *Sapria*. First, we used a custom python script to extract the positions and lengths of introns for all proteins in the reference species *Manihot* and *Populus* for comparative purposes. Here, the position of the intron was measured by peptide coordinates (e.g., between the 23th and 24th AA). Second, for each *Manihot* or *Populus* protein that was aligned to the *Sapria* genome, we extracted both the intron position and intron length using the same method. We then compared intron positions in *Sapria* to one reference species at a time (*Populus* or *Manihot*) in a reciprocal manner as described below and recorded the number of introns that were lost and gained in *Sapria*. To conservatively infer intron gain and loss and mitigate false positives due to protein alignment error, we only recorded intron gains when an intron in *Sapria* was absent within  $\pm 5$  AA (upstream and downstream) in the reference species. Similarly,

we only recorded intron loss when an intron in the reference species was absent within  $\pm 5$  AA in the *Sapria* genome. Finally, for each predicted gene model in *Sapria*, when multiple proteins from the same reference species could be mapped, we chose the protein alignment that was most similar to the gene structure in *Sapria*, thus invoking the least number of intron gains and losses. We further removed spurious protein alignments by requiring  $> 80\%$  of the reference protein length to be aligned. As a result, 5,485 gene models from the evidence-based annotation have valid protein alignments from both *Manihot* and *Populus*. Among them, 1,492 (27.2%) genes experienced intron loss when compared to *Manihot* and *Populus* and 1,024 (18.7%) genes became intron-free despite having introns in both *Manihot* and *Populus*.

### Functional enrichment analysis of highly compact genes

In *Sapria*, 63.5% of the genes (14,693 of 23,618) from our evidence-based MAKER annotation have introns shorter than 150 bp. To further characterize the functional significance of these highly compact genes in *Sapria*, we performed GO enrichment analysis using PlantRegMap. For each compact gene in *Sapria*, we identified its one-to-one best matching orthologs in *Populus* using BLASTn searches. When a *Sapria* gene had multiple *Populus* BLAST hits, the gene with the best E-value and most aligned bases was selected. A *p*-value threshold of 0.01 was used to identify significantly enriched GO terms based on the Fisher's Exact Test with a False Discovery Rate correction. A total of 322 GO terms were enriched in three categories when compared against *Populus* orthologs (Data S1G). We repeated the analysis using *Manihot* instead of *Populus* to test for sensitivity of reference species. A total of 199 GO terms were enriched when using *Manihot* as reference (Data S1T).

### Intron expansion and selection pressure analysis

We investigated the relative contribution of TEs to intron expansion in *Sapria* by summarizing the total length of TEs identified by RepeatModeler in both longer ( $> 1$  kb) and shorter introns ( $< 1$  kb; as a control) using BEDTools (Figure 4B).

To further test whether intron expansion was associated with relaxed selection in particular genes, we inferred the  $d_N/d_S$  ratio (non-synonymous substitutions rate/synonymous substitutions rate) for each gene. We sampled six species—*Sapria*, *Manihot*, *Jatropha*, *Populus*, and two outgroup species *Arabidopsis* and *Glycine*—to infer the  $d_N/d_S$  ratio using PAML v4.8<sup>76</sup>. We used a phylogeny-guided method from Yang et al.<sup>96</sup> to reconstruct single-copy orthogroups and additionally removed orthogroups containing horizontally transferred genes in *Sapria* (see below on HGT identification). Protein sequences from each orthogroup were aligned using MAFFT and then converted into the corresponding codon alignments using pal2nal. The maximum likelihood (ML) phylogeny for each orthogroup was inferred using IQ-TREE v2.0.5<sup>77</sup> under the default settings. Detailed description of orthogroup establishment, alignment, and phylogenetic reconstruction is provided in Data S3. CODEML from the PAML package was subsequently applied to infer  $d_N$  and  $d_S$  statistics under the free-ratio model based on the codon alignment and phylogeny of each orthogroup. Finally, we tested the correlation between the maximum intron length and  $d_N/d_S$  ratio using Spearman's rank correlation test in R (function cor.test). The result indicated a significant correlation (*p*-value  $7.2e-9$ , Spearman's rank correlation  $\rho = 0.102$ ).

### Genome scan assessment of HGT

Because HGTs initially create regions of low divergence between host and parasite, we developed a sliding-window genome scan assessment involving pairwise divergence to investigate fine-scale patterns involving more recent HGTs (Figure S3). We aligned the *Sapria* and *Tetrastigma* assembly, along with the complete genomes of two closest relatives of *Sapria*—*Manihot* and *Populus*—using the genome aligner Cactus v1.1.0<sup>78</sup>. Notably, we used the Quickmerge assembly of *Tetrastigma* for this analysis given its better continuity. With this alignment, we used HAL tools<sup>97</sup> to extract pairwise Multiple Alignment Format (MAF) blocks in which no duplicates existed within the host or parasite genomes (e.g., *Tetrastigma* and *Sapria*, respectively) to obtain uniquely mapped regions. These alignment blocks were processed with custom python scripts (available on Github, see code availability statement) to create windows that each contained 100 aligned bases. Only aligned positions were considered when defining these windows, and positions containing gaps or N's were ignored. Thus, although windows contained the same number of aligned bases to calculate divergence, they varied in length in terms of reference genome coordinates if particular regions of the alignment contained more gapped positions (e.g., a window spanning the *Sapria* genome begins at position 1 and ends at position 120, containing 100 aligned positions and 20 gapped positions). Divergence within each window was calculated as the proportion of aligned bases that differed between the *Tetrastigma* and *Sapria* genomes.

To create a data-driven threshold of divergence to classify windows, we calculated the divergences between *Tetrastigma* and *Sapria* for genes classified as HGT or VGT (vertical gene transfer) using our phylogenomic approach described below (Figure S6A). The Receiver Operating Characteristic (ROC) curve analysis (using the R package pROC with Youden's J statistic<sup>98</sup>) suggested a divergence threshold of 0.245 to distinguish these two classes of genetic regions, such that alignment windows with divergences below 0.245 are candidates for HGT.

While low divergence windows are a hallmark of recent HGT, this signature may also arise from strong purifying selection on functionally important loci such as conserved non-coding elements. To rule out this confounding factor, we focused on windows that not only exhibited low divergence between *Sapria* and *Tetrastigma* but also contained no aligned bases in the two close relatives of *Sapria*, *Manihot*, and *Populus*. This makes purifying selection an unlikely explanation for low divergence and also renders a single acquisition of DNA from *Tetrastigma* a more parsimonious explanation than two independent losses.

Our pairwise divergence analysis applied windows of 100 aligned bases, but a single HGT block may involve longer genomic regions that include multiple, nearby low-divergence windows. Indeed, outlier windows of low divergence between *Sapria* and

*Tetrastigma* appeared to be physically clustered (Figure 5B) as would be expected when larger blocks of DNA are transferred. To identify an appropriate threshold distance to group nearby outlier windows, we measured the distance between nearby low-divergence windows. We found that the majority (88%) of low-divergence windows were no more than 500bp away from at least one other outlier window, and distance thresholds longer than 500bp failed to group an appreciable number of additional windows (Figure S6B). After we grouped outliers using this distance threshold, we used the coordinates of the left- and rightmost outlier windows to construct the distribution of HGT lengths.

### Phylogenomic assessment of HGT

Our genome scan approach above was designed to detect recent HGT events and may miss ancient transfers. To detect older events, we performed a large-scale phylogenomic analysis similar to previous efforts to detect HGT in Rafflesiaceae (Figure S3)<sup>8</sup>. We sampled 38 species for phylogenomic investigation, including the addition of transcriptomes from three Rafflesiaceae species (Data S1J). We also added pseudogenes from *Sapria* to characterize HGT genes likely to be nonfunctional. All 71,747 pseudogenes were compared to five randomly chosen sequences from each orthogroup using BLASTn for computational efficiency. We used an E-value threshold of 1e-40 to assign orthologs. We then selected orthogroups containing at least ten species for our phylogenomic assessment of HGT. Specifically, each selected orthogroup should include at least one sequence from Rafflesiaceae (gene or pseudogene) and another from a free-living Malpighiales species (*Manihot*, *Jatropha*, and/or *Populus*). A total of 6,552 orthogroups were selected for further phylogenomic analysis. Protein sequences for each orthogroup were aligned with MAFFT v.7.299<sup>79</sup> using the iterative refinement algorithm E-INS-i. The resulting protein alignments were converted into the corresponding codon alignments using pal2nal v.14<sup>80</sup>. If a pseudogene was present, we then added the pseudogene sequence to the codon alignment using ‘mafft-add-long’ in MAFFT. The alignments were trimmed using trimAL<sup>81</sup> to remove sites with more than 85% gaps (-gt 0.15).

AML phylogeny was inferred for each orthogroup using IQ-TREE. Optimal substitution models for each alignment were determined by ModelFinder<sup>99</sup> using the Bayesian Information Criterion (BIC) within IQ-TREE. Branch support was assessed with ultrafast bootstrap approximation (UF-BP)<sup>100</sup> from 3000 replicates along with the ‘-bnni’ option to reduce the risk of overestimating branch support due to model violations. We additionally assessed branch support with the SH-like approximate likelihood ratio test for each branch (SH-aLRT)<sup>101</sup> using 2000 bootstrap replicates in IQ-TREE. Customized scripts (available on Github, see code availability statement) were subsequently implemented to identify HGT candidates where Rafflesiaceae species were placed outside Malpighiales. One hundred fifty-two orthogroups containing Vitaceae-associated HGT candidates were identified and selected for more in-depth phylogenetic validation with expanded taxon sampling (see next section).

### HGT validation with expanded taxon sampling

We expanded taxon sampling within the host clade, Vitaceae, and applied stringent branch support thresholds to further validate the 152 HGT candidates identified using our phylogenomic method outlined above. Raw sequencing reads from 15 published Vitaceae transcriptomes<sup>31</sup> were downloaded from GenBank: SRA081731. Transcriptome assembly followed the pipeline described above following Cai et al.<sup>89</sup>. We also obtained two transcriptomes of Vitaceae from the oneKP project<sup>91</sup>. To reduce computational burden for orthogroup clustering of the additional species, these 17 transcriptomes were compared to five randomly chosen sequences from each orthogroup using BLASTn. We used an E-value threshold of 1e-40 to assign orthologs. Orthologous sequences from additional Vitaceae species were added to the alignment of each orthogroup by MAFFT using the ‘mafft-add-keeplength’ command. Phylogenetic inference followed the strategy described above using IQ-TREE.

To filter for confidently placed HGTs, we required each HGT candidate to nest within Vitaceae with high branch support (80 SH-aLRT and 80 UF-BP), have a minimum gene length of 150 bp, and a maximum branch length of 0.825 to mitigate long-branch artifacts. This branch length cutoff was determined by the 93% quantile of the branch length distribution of all HGT candidates (Figure S6C), where a sharp decline in frequency was observed. Lower nodal support was observed for ancient HGTs shared by more than two Rafflesiaceae genera. For these candidates, we applied a less stringent branch support threshold of 50 SH-aLRT and 50 UF-BP. A total of 90 orthogroups passed these stringent filtering criteria.

### BLAST-based assessment of HGT

Some HGT candidates exhibited higher sequence divergence between *Sapria* and *Tetrastigma* and did not have orthologs in other Malpighiales species to be assessed using our genome scan or phylogenetic methods. To investigate these genes, we used the BLAST results from OrthoFinder to identify HGT candidates that only had BLAST hits (E-value < 1e-5) in Rafflesiaceae and Vitaceae but not in other plant lineages. This analysis identified six additional HGT orthogroups (Data S1K).

### HGT validation using nanopore reads

It has long been recognized in plants that DNA and RNA molecules are trafficked between host and parasite<sup>82,102–104</sup>, and represent a source of natural contamination. This source of natural contamination may confound HGT detection when host derived genome sequences form chimeric assemblies. To eliminate these concerns, we verified that each candidate HGT was fully integrated into the *Sapria* genome using nanopore reads. Nanopore reads represent continuous, single DNA molecules and can therefore rule out artifactual chimeric assemblies that integrate natural sources of host contamination. Here, we aligned nanopore reads to the *Sapria* assembly using minimap2. We then validated HGT candidates by requiring them to be contained within a continuously aligned nanopore read and reside at least 500 bp away from the end of each read (Figure S3). Second, we also required putative HGTs to be



flanked on either side by 500 bp of non-HGT sequence within a scaffold. For the HGTs identified by our phylogenomic assessment and BLAST-based assessment, 88.2% of the genomic regions (574 of 705) from 87 orthogroups (81 from phylogenomic assessment and 6 from BLAST evidence) were well-nested within long reads. For our genome scan assessment, 94% of HGT windows (793 of 846) met these criteria.

### Functional enrichment analysis of HGTs

For each putative HGT identified using our genome scan, phylogenomic, and BLAST methods, its one-to-one best matching ortholog from *Vitis vinifera* was identified using BLASTn. Putatively enriched functions of these HGTs were then assessed using the background of all *Vitis* genes using PlantRegMap.

### Codon usage bias of HGTs

We compared the codon usage of 568 HGTs and 3007 VGTs to *Manihot* and *Tetrastigma*, respectively. We applied Measure Independent of Length and Composition (MILC) to quantify the distance in codon usage between a *Sapria* gene and the reference gene set<sup>40</sup> using the function 'MILC' in the R package coRdon<sup>105</sup>. Coding sequences from *Manihot* or *Tetrastigma* were used as references, respectively. We also calculated the frequency of optimal codons (Fop,<sup>39</sup>) in HGTs and VGTs to characterize their codon usage adaptiveness. The 303 ribosomal protein coding genes identified by Pfam in *Sapria* were used as the reference set to quantify Fop in HGTs and VGTs using the function Fop in coRdon.

### Modern and former host associations

We filtered HGTs from our phylogenomic assessment based on taxon sampling, gene tree topology, and gene tree support to infer modern and former host associations. First, we selected orthogroups containing at least five Vitaceae genera to place the host lineage. Second, we selected orthogroups whose gene tree topologies largely reflected species tree relationships identified by Wen et al.<sup>31</sup>. Third, we required any Rafflesiaceae transgenes to be placed with its putative donor(s) with at least 50 SH-aLRT and 50 UF-BP. A total of 42 orthogroups were deemed suitable for host association investigation following these filtering criteria (Data S1M; Figure 5). Phylogenies of these 42 orthogroups are provided in Data S2B.