Current Biology

Genome sequence of Striga asiatica provides insight into the evolution of plant parasitism

--Manuscript Draft--

Manuscript Number:	CURRENT-BIOLOGY-D-19-01058R1
Full Title:	Genome sequence of Striga asiatica provides insight into the evolution of plant parasitism
Article Type:	Research Article
Corresponding Author:	Ken Shirasu, Ph.D. RIKEN Yokohama, JAPAN
First Author:	Satoko Yoshida
Order of Authors:	Satoko Yoshida
	Seungill Kim
	Eric K. Wafula
	Jaakko Tanskanen
	Yong-Min Kim
	Loren Honaas
	Zhenzhen Yang
	Thomas Spallek
	Caitlin E. Conn
	Yasunori Ichihashi
	Kyeongchae Cheong
	Songkui Cui
	Joshua P. Der
	Heidrun Gundlach
	Yuannian Jiao
	Chiaki Hori
	Juliane K. K. Ishida
	Hiroyuki Kasahara
	Takatoshi Kiba
	Myung-Shin Kim
	Namjin Koo
	Anuphon Laohavisit
	Yong-Hwan Lee
	Shelley Lumba
	Peter McCourt
	Jenny C. Mortimer
	J. Musembi Mutuku
	Takahito Nomura
	Yuko Sasaki-Sekimoto

	Yoshiya Seto
	Yu Wang
	Takanori Wakatake
	Hitoshi Sakakibara
	Taku Demura
	Shinjiro Yamaguchi
	Koichi Yoneyama
	Ri-ichiroh Manabe
	David C. Nelson
	Alan H. Schulman
	Michael P. Timko
	Claude W. dePamphilis
	Doil Choi
	Ken Shirasu, Ph.D.
Abstract:	Parasitic plants in the genus Striga, commonly known as witchweeds, cause major crop losses in sub-Saharan Africa and pose a threat to agriculture worldwide. An understanding of Striga parasite biology, which could lead to agricultural solutions, has been hampered by the lack of genome information. Here we report the draft genome sequence of Striga asiatica with 34,577 predicted protein-coding genes, which reflects gene family contractions and expansions that are consistent with a three-phase model of parasitic plant genome evolution. Striga seeds germinate in response to host-derived strigolactones (SLs) and then develop a specialised penetration structure, the haustorium, to invade the host root. A family of SL receptors has undergone a striking expansion, suggesting a molecular basis for the evolution of broad host range among Striga spp. We found that genes involved in lateral root development in non-parasitic model species are coordinately induced during haustorium development in Striga, suggesting a pathway that was partly co-opted during the evolution of the haustorium. In addition, we found evidence for horizontal transfer of host genes as well as retrotransposons, indicating gene flow to S. asiatica from hosts. Our results provide valuable insights into the evolution of parasitism and a key resource for the future development of Striga control strategies.

Cover Letter

RIKEN Center for Sustainable Resource Science

1-7-22, Suehiro-cho, Tsurumi-ku
Yokohama-shi, Kanagawa, 230-0045 Japan

Phone : +81-45-503-9574 Fax : +81-45-503-9573
E-MAIL : ken.shirasu@riken.jp

2019 July 22

Dear Anne,

We would like to resubmit our manuscript entitled “Genome sequence of Striga asiatica provides insight into the evolution of plant parasitism” by Yoshida et al.

We have revised our manuscript accordingly to your email. Mainly, we moved 3 supplemental figures into main ones. We also created STAR methods section to include necessary information. DataS2 and DataS3 were swapped as appeared in the next. In addition we added Highlights and Blurb as well as Key Resource Table as separate word files.

I hope that the revised version is an acceptable form but please let me know if required more changes.

Yours faithfully,

Ken Shirasu

Group Director

Response to Reviewers

N/A

Genome sequence of *Striga asiatica* provides insight into the evolution of plant parasitism

Satoko Yoshida[1, 2,3], Seungill Kim[4,5], Eric K. Wafula[6], Jaakko Tanskanen[7,8], Yong-Min Kim[9], Loren Honaas[6], Zhenzhen Yang[6], Thomas Spallek[1,10], Caitlin E. Conn[11], Yasunori Ichihashi[1,12], Kyeongchae Cheong[4,5], Songkui Cui[1,3], Joshua P. Der[13], Heidrun Gundlach[14], Yuannian Jiao[6], Chiaki Hori[1,15], Juliane K. Ishida[1], Hiroyuki Kasahara[1,16], Takatoshi Kiba[1,17], Myung-Shin Kim[4,5], Namjin Koo[9], Anuphon Laohavisit[1], Yong-Hwan Lee[4,18], Shelley Lumba[19], Peter McCourt[19], Jenny C. Mortimer[1,20,21], J. Musembi Mutuku[1,22], Takahito Nomura[23], Yuko Sasaki-Sekimoto[24], Yoshiya Seto[25,26], Yu Wang[27], Takanori Wakatake[1,28], Hitoshi Sakakibara[1,17], Taku Demura[1,2], Shinjiro Yamaguchi[25,29], Koichi Yoneyama[23], Ri-ichiroh Manabe[30], David C. Nelson[11,31], Alan H. Schulman[7,8], Michael P. Timko[27], Claude W. dePamphilis[6], Doil Choi[4,5], Ken Shirasu[1, 28]

[1]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, 230-0045 Japan

[2] Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan

[3] Institute for Research Initiatives, Division for Research Strategy, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan

[4]Interdisciplinary Program in Agricultural Genomics, Seoul National University, Seoul 151-742, Korea.

[5] Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-742, Korea.

[6] Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

[7]Production Systems, Luke Natural Resources Institute Finland, FI-00790 Helsinki, Finland

[8]Luke/BI Plant Genomics Laboratory, Institute of Biotechnology and Viikki Plant Science Centre, University of Helsinki, FI-00014 Helsinki, Finland

[9]Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejon 305-806, Korea

[10] Institute of Plant Physiology and Biochemistry, University of Hohenheim, 70599

Stuttgart, Germany
[11] Department of Genetics, University of Georgia, Athens, GA 30602, USA
[12] RIKEN BioResource Research Center, Tsukuba, Ibaraki 305-0074, Japan.
[13] Department of Biological Science, California State University, Fullerton, CA 92831, USA
[14] Plant Genome and Systems Biology (PGSB), Helmholtz Center Munich, Neuherberg, D-85764, Germany
[15] Research Faculty of Engineering, Hokkaido University, Sapporo, 060-8628, Japan
[16] Institute of Global Innovation Research, Tokyo University of Agriculture and Technology, Fuchu, 183-8509 Japan
[17] Department of Applied Biosciences, Graduate School of Bioagricultural Sciences, Nagoya University, Chikusa, Nagoya, 464-8601, Japan
[18] Department of Agricultural Biotechnology, Seoul National University, Seoul, Korea.
[19] Department of Cell & Systems Biology, University of Toronto, Toronto, ON M5S-3B2, Canada
[20] Joint BioEnergy Institute, Emeryville, CA, 94608 USA
[21] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[22] International Livestock Research Institute (BecA-ILRI) Hub, 00100 Nairobi, Kenya
[23] Center for Bioscience Research and Education, Utsunomiya University, Utsunomiya 321-8505, Japan
[24] School of Life Science and Technology, Tokyo Institute of Technology, 226-8501, Yokohama, Kanagawa, Japan
[25] Graduate School of Life Sciences, Tohoku University, Sendai, Miyagi, 980-8577, Japan
[26] Department of Agricultural Chemistry, School of Agriculture, Meiji University, Kawasaki, Kanagawa, 214-8571, Japan
[27] Department of Biology, University of Virginia, Charlottesville, VA 22903, USA
[28] Graduate School of Science, The University of Tokyo, Bunkyo, Tokyo 113-0033, Japan
[29] Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[30] Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Kanagawa 230-0045, Japan
[31] Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

Correspondence Author and Lead Contact:
Ken Shirasu (ken.shirasu@riken.jp)

Summary

Parasitic plants in the genus *Striga*, commonly known as witchweeds, cause major crop losses in sub-Saharan Africa and pose a threat to agriculture worldwide. An understanding of *Striga* parasite biology, which could lead to agricultural solutions, has been hampered by the lack of genome information. Here we report the draft genome sequence of *Striga asiatica* with 34,577 predicted protein-coding genes, which reflects gene family contractions and expansions that are consistent with a three-phase model of parasitic plant genome evolution. *Striga* seeds germinate in response to host-derived strigolactones (SLs) and then develop a specialised penetration structure, the haustorium, to invade the host root. A family of SL receptors has undergone a striking expansion, suggesting a molecular basis for the evolution of broad host range among *Striga* spp. We found that genes involved in lateral root development in non-parasitic model species are coordinately induced during haustorium development in *Striga*, suggesting a pathway that was partly co-opted during the evolution of the haustorium. In addition, we found evidence for horizontal transfer of host genes as well as retrotransposons, indicating gene flow to *S. asiatica* from hosts. Our results provide valuable insights into the evolution of parasitism and a key resource for the future development of *Striga* control strategies.

Introduction

Striga is a genus of parasitic plants in the Orobanchaceae family that includes major agricultural weeds. *S. asiatica* and *S. hermonthica* infect grain crops such as sorghum, millet, maize, upland rice, and sugarcane, causing $US billions of annual yield losses[1–3]. *Striga* has evolved unique parasitic adaptations that make infestations extremely difficult to eradicate[3]. A single *Striga* plant produces more than 100,000 small (~200 µm) seeds, which can be wind-dispersed for a long distance. The seeds can lie dormant for decades, surviving extreme conditions, until they perceive host-derived germination stimulants, such as strigolactones (SLs)[4,5]. Once germinated, *Striga* roots grow towards the host and detect compounds derived from the host cell wall, which induce the development of a specialised organ called the haustorium at the tip of the radicle[6,7]. The haustorium invades the host root, and connects its xylem with that of the host to

assimilate water and nutrients. In addition, genetic materials from the hosts are also transferred into *Striga*, but the extent and the precise mechanism of horizontal gene transfer remain elusive[8–10].

Results and Discussion

The structure and evolution of the *Striga* genome

The genome of the *S. asiatica* strain that invaded the United States in 1950s[2] was sequenced and assembled using a combination of Illumina-based whole-genome shotgun technology and Sanger-based BAC library end sequencing. The Kmer-based estimation of the *S. asiatica* genome size is approximately 600 megabase pairs (Mb), and 472 Mb of the genome was assembled with an N50 scaffold size >1.3 Mbp (contig N50 > 16.2 kbp and 393 x read coverage, Data S1A), in which a total of 34,577 genes was predicted (For detail, see Data S2A,B).

Global gene family phylogenetic analysis and genome structure/synteny analysis with the closely related nonparasitic plant *Mimulus (Erythranthe) guttatus* (Figure 1) both indicate that the *S. asiatica* genome retains evidence of at least two whole genome duplication events (WGD, Figures 2A-D, Data S2C). We examined the divergence patterns of synonymous substitution rates (K_s) for Lamiales-wide duplicate genes identified by an integrated syntenic and phylogenomic analysis. Comparison of gene trees for 1,440 orthologous single-copy genes showed that the length for the branch leading to *S. asiatica* was longer than that leading to *Mimulus* suggesting that *S. asiatica* has experienced a more rapid molecular evolution than *Mimulus* (Figure 1). We identified two significant duplication components in *S. asiatica* at mean $K_s \approx 0.47$ (younger) and 1.22 (older) as well as one significant component for *Mimulus* at rescaled mean $K_s \approx 0.94$ (Figure 2B). The older *Striga* K_s peak and the single peak of the *Mimulus* K_s distribution represent a shared ancestral WGD event for Lamiales (Figure 2C). As expected the *S. asiatica* peak is shifted to the right (a higher K_s value) because of the accelerated rate of evolution for *S. asiatica*. The prominent younger peak in the *Striga* K_s distributions represents a duplication event that occurred after the divergence of lineages leading to *S. asiatica* and *Mimulus.*

Parasitic plant evolution is thought to progress through three phases: *Phase I*, evolutionary gain of a haustorium; *Phase II*, loss of functions that are supplemented by a

host resource; and *Phase III*, specialisation of the parasitic relationship[11,12] (Figure 2D). Shifts of gene expression (in scope and/or specificity) and changes in the global functional gene profile presumably accompany innovation during parasite evolution. Thus, we examined shifts of parasite gene expression and function by genome-scale comparative analyses to identify the signatures of each phase. Using the list of *S. hermonthica* "haustorium" orthogroups defined in Yang et al. 2015[13], with a parallel analysis that identifies genes with tissue-specific expression in *Arabidopsis*, we found that haustorial genes are significantly enriched for tissue-specific orthogroups in *S. asiatica* (Data S1B). Concordant with Yang et al. 2015[13], this pattern was strongest for pollen orthogroups. This suggests that haustorium innovation during *Phase I* may have involved co-option of genes with tissue-specific gene expression.

Next, we identified functions associated with shifts in gene content by reconstructing each orthogroup (approximate gene family) in a common ancestor of *Striga* and *Mimulus*, as well as successively earlier common ancestors (Data S1C, Data S2C). Among the 10,248 orthogroups, approximately ~23% showed changes in gene numbers inferred for the *Striga* lineage (647 contractions, 1,742 expansions, 456 losses, and, 152 gains; Data S1D and E, Data S3). The relative age of genes in contracted orthogroups was significantly older (two-tailed Mann-Whitney U test, p-values < 2.2e-16) than genes in expanded families (Figure 2D and Data S1E). In addition, the expanded gene families show higher non-synonymous/synonymous substation (Kn/Ks) ratios compared to the contracted gene families (Student's t-test, p-value<4.7e-10, Figure S1), suggesting that the expanded gene families are under more relaxed selection pressure. The relatively younger expanded gene families, apparently gained largely as a result of the *Striga* WGD (Figures 3B, E), potentially provided a source of genes to encode specialised traits in the parasite.

Significant (Benjamini corrected P<0.05) signatures of gene family contractions were detected in two photosynthesis-related KEGG pathways (Data S1F and G). Additionally, an analysis of Gene Ontology (GO) terms among contracted lineages showed several photosynthesis-related cellular compartment (CC) terms and biological process (BP) terms were significantly over-represented (Data S1H, Data S2C, Figure 2D). These contractions are consistent with *Striga's* high reliance on host carbon[14,15]. Furthermore, significantly enriched GO BP terms associated with leaf anatomy and function were detected among contracted lineages, consistent with the anatomical and

functional reductions in *Striga* leaves. In addition to the well-documented gene losses in parasitic plant plastomes[12,16], these changes indicate a complementary reduction in reliance on photosynthesis-related gene function[17] representing *Phase II*.

Perhaps the clearest support for Searcy's *Phase II* are substantial contractions in gene families annotated with GO BP terms that relate to abiotic and biotic stimulus response including virtually all plant hormones (Data S2C, Data S1H and I, Figure 2D). This includes one in four significant GO BP terms that are seven times more numerous in contracted lineages than expanded ones. This pattern of loss points to an increasingly insensitive parasite sensing apparatus that is likely supplemented by the host. Concordant with this evolutionary signature, empirical evidence suggests that *Striga* lost abscisic acid sensitivity to regulate water loss machinery and maintains constitutively open stomata even under drought conditions[18,19] contributing to a net carbon loss in the host leaves [20].

The transition from *Phase II* to *Phase III* may in some cases be blurred from a functional standpoint because, for instance, the host plant could complement water stress response pathways, while decreasing water potential in the parasite could be adaptive[9]. Indeed, significantly enriched water relations terms can be found among both expanded and contracted lineages, yet orthogroup contractions dominate water relation signatures indicating that altered water relations may largely, but not exclusively, represent older *Phase II* losses. In GO CC profiles, contractions are biased towards structural and photosynthesis related genes families – consistent with *Phase II* complementation. However, the newer and expanded gene families are significantly biased towards endocytosis and intracellular transport, suggesting that *Phase III* innovations contribute to host resource acquisition processes. The expansions in cellular transport machinery may help explain how *Striga* obtains photosynthates derived host resources even though direct phloem connections are lacking[15,20].

Host recognition – evolution of SL receptors

As an obligate pathogen, *Striga* requires nutrients from a host within a few days after germination. One unique aspect of the specialised relationship with the host (*Phase III*) in the *Striga* parasitic lifestyle is the ability to germinate after sensing SLs, which indicate presence of a host[5]. In *Arabidopsis*, *D14* and *KAI2/D14L* are ancient paralogues that encode receptors for SLs and the karrikins (smoke-derived compounds that stimulate

germination of many nonparasitic plants), respectively[21,22]. *KAI2*, which controls seed germination in *Arabidopsis*, has undergone higher than normal gene duplication in several parasite genomes in the Orobanchaceae[23–25]. A divergent subclade of *KAI2* paralogues (*KAI2d*) has evolved SL perception, which facilitates host-detection in seeds. The super-orthogroup that contains the *KAI2* genes was expanded strikingly in *S. asiatica* (Data S1J, Data S2D). We found that the *S. asiatica* genome encodes 21 *KAI2* paralogues, and that 17 of these are in the *KAI2d* class (Figure 3A). Most of the *KAI2d* genes in *S. asiatica* are highly expressed in the seed as well as in seedling stages (Figure 3B). Two other paralogues, *KAI2c1* and *KAI2c2*, cluster with highly conserved *Arabidopsis* (*AtKAI2*) and *Mimulus* proteins (*MgKAI2c*). The intermediate group contain two *KAI2i* paralogues, which are sister to the expanded *KAI2d* clade. *Mimulus KAI2i* (*MgKAI2i*) is branched from the ancestral node of the *Striga KAI2d* and *KAI2i*, suggesting that *Striga KAI2d* genes evolved out of the intermediate group. In addition, seven *KAI2* pseudogenes are also found in the genome, providing further evidence for highly dynamic evolution of the *KAI2* gene family (Figure 3C). *KAI2* paralogues and pseudogenes are often found on the same scaffold (Figure 3C, D). All *KAI2* genes retain a single intron at a conserved position. Tandem *KAI2* paralogues typically share the same orientation, consistent with localised *KAI2* duplication by unequal recombination. Interestingly, *KAI2i*, which is ancestral to *KAI2d* genes, is located next to *Striga*-specific *KAI2d7* and *KAI2d8* (Scaffold 62, Figure 3C, D), suggesting that the *Striga*- specific *KAI2d* clade originally may have been derived by the tandem duplication of *KAI2i*. If different *KAI2d* paralogues have specificity for distinct types of SLs, then the rapid evolution of the *KAI2d* clade likely enabled *Striga* seeds to recognise a wide range of hosts[23–25]. We noted that the high level of expression of many *KAI2d* homologues have a high level of expression at the seedling stage, suggesting that the host-derived SL may influence other functions beyond germination.

Development of the invading organ, the haustorium

Immediately after germination, *Striga* grows towards the host and detects cell wall-derived compounds[6]. This initiates a drastic developmental reprogramming, resulting in the formation of a haustorium that invades the host root (Figure 4A). To investigate gene expression dynamics during haustorium development, RNA-seq analysis was performed with the most devastating *Striga* species, *S. hermonthica* (Data S1K-M, Data

S2E). Principal component analysis (PCA) and self-organising map (SOM) clustering were used to classify the transcripts into twelve clusters, each with a distinct expression pattern specific to one or more developmental stages (Figures 4A, B). The GO enrichment analysis of these clusters (Benjamini and Hochberg corrected $P < 0.05$; Figure 3c, Data S1N) projected a similar sequence of molecular events during *Striga* parasitism. Clusters 2, 3, and 6 showed expression patterns specific to the seed; transcripts in these clusters are enriched for GO terms related to post-embryonic development and to embryonic development towards the end of seed dormancy (Benjamini and Hochberg corrected $P < 0.05$; Figure 4C). The seedling-specific cluster 12 showed enrichment in defence responses as well as in transcriptional regulatory activity (Benjamini and Hochberg corrected $P < 0.05$; Figure 4C). This suggests that the seedling has already started to change its transcriptional profile to enable parasitisation of host plants, i.e., the primary haustorium formation may be coupled with seed germination in *S. hermonthica*. Our SOM analysis allowed us to capture a subsequent peak of gene expression from seedling to 7 d, represented by clusters 9, 1, 5, 4, 8, 7, 11, 10, in that order (Figure 4C). The temporal expression patterns of several selected genes were confirmed by RT-qPCR upon host and nonhost interactions (Figure S2, Data S2E). While the early gene expression was induced by DMBQ treatments as well as host and nonhost interactions, the expression of middle- and late-stage genes was not seen in the interaction with nonhost *Lotus japonicus* (Figure 4D, Data S2E). Because *S. hermonthica* is able to penetrate tissues of nonhost *Arabidopsis* and *L. japonicus*, but not establish xylem connections with *L. japonicus*[26], the early genes are likely to be important for haustorium formation and host penetration, while the genes involved in the middle to late stages of haustorial development may associate with xylem connection formation and/or host materials acquisition. *In situ* hybridisation analysis highlights the tissue specific expression of such genes. An early-stage gene, encoding the peroxidase, exclusively expressed at the intrusive cells that are aligned at host-parasite interface (Figures 4E, F), whereas various 7-d-specific genes are highly expressed in the hyaline body (Figures 4G-J), a specific parenchymatic tissue whose characteristics include dense cytoplasm, organelle-rich structure, and high metabolic activity[27]. The hyaline body is proposed to function as a sink for host materials, and the high expression of catabolic enzymes such as proteases within this tissue may contribute to such a function. The middle and late genes include the recruitment of catalytic activity-related genes (especially hydrolases) during host

penetration, transport-related genes during host nutrient acquisition, and signal transduction-related genes during resource allocation. In fact, among the identified 1,292 CAZyme (Carbohydrate-active enzyme)-categorised genes[28], 252 are differentially expressed during invasion stages (Figure 5, Data S1O and P, Data S2E). Specifically, enzymes targeting primary cell wall components, such as those degrading pectin, are highly upregulated (Figures 5C, D). In addition, many proteases are upregulated at late stages of infection.

Comparative studies of development in an evolutionary context have been routinely employed to understand developmental mechanisms and to deduce how the regulatory changes in gene expression contribute to morphological diversity[29]. Since our genome analysis indicated potential sub-functionalisation and/or co-option of existing genes from tissue-specific gene families (*Phase I*), we hypothesised that parasitic plants may have employed a pre-existing developmental program to produce the haustorium. One such program is lateral root formation, as this also creates new xylem connections in roots. Out of the known 18 lateral root development (LRD) genes in *Arabidopsis*[30], we identified respectively 18 and 17 LRD orthologues in the *S. asiatica* genome and the *S. hermonthica* transcriptome (Data S1Q, Data S2E). Among these genes, *SLR/IAA14*, *ARF19*, and *LAX3* orthologues are specifically expressed during the early stage of haustorium development (Figure 6A, Figure S3). *SLR/IAA14* and *ARF19* function as a module to regulate the expression of the auxin influx carrier *LAX3*, which localises auxin accumulation during LRD[31] (Figure 6B). Thus, the *SLR/IAA14-ARF19-LAX3* component might be utilised to initiate auxin accumulation during *Striga* haustoria formation. We also detected another putative target of the *SLR/IAA14-ARF19* module, the *LBD18* orthologue, which is highly expressed in the early stage (Figure 6A). *Arabidopsis LBD18* activates cell proliferation in the lateral root primordia[32]. Correspondingly, cell proliferation is highly active in haustoria[33], suggesting that the *LBD18* orthologue might have a conserved function to coordinate the spatial pattern of cell proliferation during haustorium formation. In the later stages of haustoria formation, such as 3 d and 7 d, we observed the upregulation of *ARF5* and of *ARF8* homologues (Figure 6A). *ARF5* follows *SLR/IAA14-ARF19* expression to control lateral root organogenesis[34], whereas *ARF8* activates lateral root meristem in response to nitrogen availability[35]. Therefore, these genes might be involved in the later stages of haustorium formation when host penetration occurs and vasculature connections are formed. Note that no up-regulation of

two other LRD related genes, *ABERRANT LATERAL ROOT FORMATION 4* (*ALF4*) and *ARABIDOPSIS CRINKLY 4* (*ACR4*), were detected in *S. hermonthica* haustoria, but, surprisingly, their orthologues (*ALF4*: LOC_Os08g19320; *ACR4*: LOC_Os03g43670) were upregulated in host plants 1 day after infection (Figure 6A). As *ACR4* expression is dependent on *SLR/IAA14-ARF19* to specify LRD cell identity in *Arabidopsis*[36] and *ALF4* functions in maintaining the mitotically competent state of the pericycle cells in LRD[37], *ACR4* and *ALF4* might link the interaction between *S. hermonthica* and its host. Taken together, certain LRD genes in *S. asiatica* and *S. hermonthica* are activated during haustorium formation and, interestingly, the expression orders follow developmental timeframes similar to those during LRD in *Arabidopsis* (Figure 6B), suggesting that haustorium formation, which confers parasite function in parasitic plants, might be evolved partly through the recruitment of parasitic plant and host LRD programs.

Horizontal Gene Transfer

Genetic materials such as mRNAs are transferred from hosts to parasitic plants[38]. The transferred material may also be integrated into the germ line of the parasites[8,39]. To understand the extent of such horizontal gene transfer (HGT) events, the *S. asiatica* genome was compared with other dicot and monocot genomes to find *Striga* genes that clustered with monocot orthologues. We identified 34 potential HGT candidates in the *S. asiatica* genome (Figure 7, Data S1R, Data S2F). Two of the HGT candidate genes are aligned in tandem in an approximately 30 kbp region in the genome of *S. asiatica.* The orthologues of the two genes, including introns and untranslated regions, are also located in tandem in the genomes of two Poaceae, *Panicum hallii*, and *Setaria italica*, suggesting transfer of a large (~100 kb in *P. hallii*) genomic segment from host to parasite (Figures 7A, B). Phylogenetic analyses showed that the two *S. asiatica* genes clustered only with Poaceae genes, supporting HGT from host to parasite (Figures 7C, D). Interestingly, a few other genomic regions contain multiple HGT genes in close proximity (Data S2F), although the syntenic regions are not found in the Poaceae genomes, possibly due to rearrangement of the host genome after the gene transfer. These data suggest that the inter-species transfer of large genomic fragments may have occurred multiple times.

Because transposable elements were previously reported as HGT targets[10], we conducted phylogenetic analyses for all the reverse transcriptase (*rt*) domains in *S. asiatica*, and for representative *rt* sequences from both eudicots and monocots (Data S2F).

Our analyses included 35,690 from *Copia* and 54,973 from *Gypsy* elements in the publicly available plant genome sequences. Clusters containing both *S. asiatica rt* sequences and monocot sequences were analysed further. Three putative HGT events were identified. One of these, comprising ~80 total *rt* sequences, includes 29 *S. asiatica rt* in a cluster with 48 diverse *Sorghum bicolor rt*, suggesting a direct horizontal transfer from *S. bicolor,* a natural host of *Striga* (Figure 7E), and subsequent amplification of *rt* sequences in the *Striga* genome. Two other trees, in which *S. asiatica rt* sequences are found nested within an exclusively Poaceae clade, having their closest orthologues respectively in *Oryza* and *Z. mays* or in *Oryza* and *S. bicolor,* suggest additional transfers from Poaceae hosts to *Striga* (Figure S4). These results indicate that *Striga* acquired genetic materials from its hosts with higher frequency compared to the autotrophic angiosperms, which may have influenced the parasite's evolution and adaption.

Outlook

Striga remains the greatest biological constraint to food production in its endemic areas in Africa, and thus its genomic and transcriptomic sequences are important tools for understanding its parasitic strategies and for developing efficient, knowledge-based management programs. In addition, the genome information provides a basis for understanding the origin of parasitism during the course of evolution. Similar to recently published stem parasites dodder (*Cuscuta* spp) genomes[39,40], *Striga* evolved rapidly compared to autotrophic species, acquired genes from their hosts *via* horizontal gene transfer, and recruited root developmental programs for haustorial formation. Both parasites have lost genes related to environmental sensing, leaf developmental processes, and photosynthesis, as predicted for the degratory phase of parasite evolution, but *Striga* frequently retains portions of reduced gene families, reflecting its status as a leafy hemiparasite that is photosynthetically competent while being highly dependent on host-derived carbon. Detailed comparisons of nuclear genomes from fully heterotrophic Orobanchaceae, and other parasitic plants with different levels of host dependency will deliver further insights into the evolution of parasitism.

Acknowledgements

This work is partially supported by MEXT KAKENHI (No, 24228008, 15H05959, 17H06172 to

K.S., No. 18H02464 and 18H04838 to Sa.Y., No. 15K18589 to Y.I., 17H06474 to Sh.Y., 17H06473 to H.S., and 25891029 and 17K15142 to S.C.), JSPS Postdoctoral Fellowship (to J.M.M. and C.H.), JSPS Research Fellowship for Young Scientist (to T.W.), the RIKEN Special Postdoctoral Researchers Program (to Y.I., J.C.M. and T.S.), NSF Postdoctoral Research Fellowship (no. 1711545 to C.C.), Academy of Finland Project 266430 (to A.H.S.), NSF IOS-1737153 and IOS-1740560 (to D.C.N.), NSF IOS-1238057 (to C.W.D. and M.P.T.), MEXT Cell Innovation program (to R.M. and Y.H. in CLST RIKEN) and NSERC (to S.L.). This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A6A3A04004014) to S.K., by a grant from the Agricultural Genome Center of the Next Generation Biogreen 21 Program of RDA (Project No. PJ013153) to D.C. and RIKEN president fund to K.S. HiSeq2000 sequencing was performed by Genome Network Analysis Support facility in RIKEN CLST. Computational analysis was partially performed using the super computer system of National Institute of Genetics, Research Organization of Information and Systems. Strigol and *S. hermonthica* seeds were kindly provided by Drs. K. Mori and A. G. E. Babikar, respectively. M.P.T. and Y.W. were supported in part by grants from the NSF (IOS-1238057 and IOS-1213059) and the Kirkhouse Trust. J.C.M.'s work was funded as part of the DOE Joint BioEnergy Institute supported by the U. S. Department of Energy (DE-AC02-05CH11231).

Author Contributions

K.S. conceived the project, designed the content, and organised the manuscript. M.T. provided plant materials. Sa.Y., T.S. and R.M. performed data generation and sequencing analysis. S.K., Y.-M.K., K.C., M.-S.K., Y.-H.L., and D.C. performed *de novo* genome assembly, *E.W. and C.W.D* performed genome scale annotation and duplication analysis. Sa.Y., T.S., Y.I., J.M.M., A.L., J.I., T.W., H.K., T.K., H.S., T.N., Y.S., Sh.Y., K.Y., Y.S.-S., C.E.C., D.N. S.L., P.M., C.H., J.C.M., and T.D. performed gene annotation and individual gene family analysis. Y.W. and M.P.T. analysed transcriptional factors. E.W., L.H., Z.Y., J.D. and C.W.D. performed comparative genome analysis, phases of parasite evolution, and whole genome duplication analysis. J.T., H.G., and A.H.S. performed TE annotation and searched for horizontally transferred TEs. S.C. performed *in situ* hybridisation. Y.I. and Sa.Y. analysed the transcriptome data. Sa.Y. Y.I., T.S., S.C., S.K., Y.-M.K., D.C., S.L., P.M.. C.W.D., M.P.T. A.H.S., and K.S. wrote the manuscript.

Declaration of interests

The authors declare no competing financial interests.

Figure legends

Figure 1. The maximum likelihood species tree

Phylogenetic tree of 26 representative plant species (Data S1C) was estimated from the concatenated data matrix for 1,440 single copy orthogroup genes obtained from BUSCO classification. Bootstrap values were 100% for each node.

Figure 2. The *Striga asiatica* genome

A. Syntenic scaffolds of *Striga* (blue), *Mimulus* (orange) and *Vitis* (grey). **B.** Ks plots of *Striga* and *Mimulus* genes. Orange and blue colours represent an older and a recent ploidy event, respectively. **C.** Schematic phylogenetic tree presenting whole-genome duplication events that occurred during the lineage evolution of *Striga*. Gamma is the genome triplication shared by core eudicots, *Striga* and *Mimulus* share a WGD (M+S), and *Striga* has experienced an independent WGD. **D.** Three-phase model of parasite evolution, showing gene categories with expression shifts, expanded and contracted in the *Striga* genome relative to a reconstructed ancestor of *Striga* and *Mimulus*. See Data S2 for details. **E .** Ks plots of expanded and contracted *Striga* genes. Age of contracted genes categorises significantly older than expanded genes categories. See also Figure S1.

Figure 3. The evolution of strigolactone (SL) receptor genes in *S. asiatica*

A. Maximum-likelihood phylogeny of predicted amino acid sequences of *KAI2/D14-LIKE* homologues in *S. asiatica* and *S. hermonthica* together with other non-parasitic species. The tree was generated based on the JTT-matrix based model. Bootstrap values above 50% are shown at the bases of branches. The scale shows inferred number of evolutionary changes per amino acid. Conserved, intermediate, and divergent clades are shown in blue, green, and red, respectively. **B.** Scaled expression levels of *S. asiatica KAI2* genes at indicated stages. **C.** Local similarities detected between the genomic regions containing *KAI2/D14-LIKE* (blue for *KAI2c*, purple for *KAI2i* and orange for *KAI2d*), *D14* (green), *DLK2* (yellow) homologues and/or their pseudogenes (grey). Locally aligned genomic regions among scaffolds (blastZ score>15000) are connected with solid lines. Orange and yellow lines represent regions containing *KAI2* or psuedo-*KAI2*, and *DLK2* homologues, respectively. Grey lines connect locally similar regions

outside *KAI2/D14/DLK2* genes. Nucleotide numbers in the scaffold are written beside the scaffold. **D.** Schematic representation of tandemly duplicated *KAI2* homologues in the genome. See Data S2 for details.

Figure 4. Transcriptional reprogramming in haustorium development

A. Developmental stages used for the transcriptome analysis of *S. hermonthica*. Seeds, preconditioned seeds; seedlings, 48 h after 10 nM strigol [41] treatment; 1 d, whole *S. hermonthica* seedlings 1 day after rice infection; 3 d and 7 d, *S. hermonthica* haustoria attached to rice tissues at 3 and 7 days after rice infection. Scale bar, 100 μm. **B.** The expression profile of each transcript is represented in PCA space with SOM node memberships indicated by different colours. A total of twelve clusters showing expression patterns specific to one or more stages were defined. The percentage shown along the x- or y- axis represents the percentage of variance explained by each component. **C.** Heat map of normalised gene expression of each transcript separated by SOM clustering with selected enriched GO terms ($P < 0.05$). **D.** Expression heatmap of stage-specific *S. hermonthica* genes in interaction with host (*O. sativa*) and nonhosts (*Arabidopsis*, *Lotus japonicus*, and *Phtheirospermum*) interactions. **E - J** *In situ* hybridisation on haustorial sections of *S. hermonthica* at 1 day (**E-F**) and 7 days (**G-J**) after rice infection. The hybridised signal (blue) represents the localisation of the transcript of an early-expressing gene encoding peroxidase (**E**) and late-expressing genes encoding subtilase 1 (**G**), LRR kinase (**H**) or cytokinin oxidase/dehydrogenase (**I**). The sense probe of peroxidase (**F**) and *subtilase1* (**J**) was used as a negative control. H: host plant, P: parasite. Scale bar, 200 μm. See also Figure S2.

Figure 5. CAZyme classification of the *S. hermonthica* transcriptome.

A. Clustering and heatmap of the differentially expressed genes containing CAZyme motifs. **B**. Number of significantly upregulated contigs containing each class of CAZyme motifs. Contigs carrying AA and GH motifs are highly upregulated at 3 d and 7 d after host interaction. **C, D**. Expression patterns of CE8 family containing pectin methyl esterases (**C**) and GH28 family containing polygalacturonases (**D**).

Figure 6. Expression patterns of genes involved in lateral root development

A. Heat map of scaled gene expression of each transcript of the *LRD*-related genes in *S.*

hermonthica. **B.** Schematic models comparing the haustorium development in *Striga* and the lateral root developmental (LRD) program in *Arabidopsis*. Expressed genes/orthologues are represented at their expressional time points. Arrows are assumed by the identified interactions in the *Arabidopsis* LRD pathway. During the haustorium formation, the corresponding *Striga* LRD orthologues showed a similar sequential expression pattern as those found in the LRD development in *Arabidopsis*. See also Figure S3.

Figure 7. Horizontal gene transfers between host and *Striga*

A. Comparison of genomic regions between *P. hallii*, *S. asiatica* and *S. italica*. The regions that show high similarity (LastZ score >5000) are connected with sky-blue lines. Coding sequences are shown as dark-blue boxes and untranslated regions are shown as pink boxes. **B.** A dot plot comparing an approx. 60 kb region in *S. asiatica* scaffold555 and either 100 kb region of *P. hallii* chromosome 3 (left) or 60 kb region of *S. italica* scaffold 3 (right) visualized by nucmer program in nummer[42] (default option). Similarity percentages are shown as rainbow colour scale. **C.** Phylogenetic tree of a hypothetical protein (555T52903) that previously was found as horizontally transferred gene in *Striga hermonthica* ESTs[8]. **D.** Phylogenetic tree of an Arginin-tRNA synthetase-like protein (555T52910). **E.** Phylogenetic trees of nucleotide sequences for reverse transcriptase in horizontally transferred retrotransposons from a host (*Sorghum*) to *S. asiatica*. The trees were unrooted and based on the maximum-likelihood method. Local support values are shown for branches. *Striga* genes are shown in red, and genes from grass species are shown in blue. HGT events are highlighted with yellow. See also Figure S4.

STAR METHODS

Lead Contact and Materials Availability

Further information and requests for materials, resources and reagents, including mosquito lines, should be directed to and will be fulfilled by the Lead Contact, Ken Shirasu (ken.shirasu@riken.jp)

Experimental Model and Subject Details

Seeds of the *S. asiatica* US strain were originally obtained from the USDA Methods Development Center (Whiteville, N.C.) and the seeds from a single plant after six rounds of self-fertilization were used as starting materials. The seeds were surface sterilised with 5% commercial bleach solution (containing final sodium hypochlorite concentration at approx. 0.3%) for 5 min and washed with excess amount of sterile water at least 5 times. The sterile seeds were preconditioned on GM media (full strength of MS salts, 0.01% Myo-inositol, 1% Sucrose, 0.5% Phytagel (Sigma)) for 10 days and the germination was induced by adding 10 nM strigol [41]. The germinated *S. asiatica* seedlings were transferred to new GM media and grown *in vitro* in a 26°C chamber at a long-day (16-h light/8-h dark) condition. For *S. asiatica* shoot propagation, the shoots were cut and transferred to new GM media every month. When *S. asiatica* shoots were transferred into the new GM media, multiple shoots were induced.

S. asiatica and *S. hermonthica* infection to rice (*Oryza sativa*, c.v. Koshihikari) was performed in the rhizotron system as previously published[43]. *S. hermonthica* seed and seedling samples were collected after preconditioning on glass-fibre filter paper (Watman GF/A) for 10 days, and before and after 10 nM strigol treatment for 2 days, respectively. *S. hermonthica* samples for 1-day post infection were carefully removed from rice roots using forceps. For the 3- and 7-day post infection samples, haustorial parts (include host tissues) were carefully excised using razor blades. For the control, rice roots without *S. hermonthica* infection were also harvested at the same day as 7-d samples. All samples were collected in triplicates of independent experiments. *S. asiatica* haustorium samples were harvested by excising the infected parts with a razor blade together with rice roots. For shoot and root samples, the sterile *S. asiatica* seeds were germinated on MS media containing sucrose and grown *in vitro* for one month.

Method Details

Whole genome shotgun sequencing, assembly and annotation of *S. asiatica*. The genomic DNA for Illumina library preparation was obtained from *S. asiatica* shoots derived from a single plant. The genomic DNA for BAC library was prepared from the siblings of the plant. The genomic DNA was extracted by using Phytopure DNA extraction kit (GE healthcare) according to manufacturer's instructions. The Illumina paired-end (PE) and mate-pair (MP) libraries were prepared using the TruSeq DNA Sample Prep Kit (Illumina, San Diego, CA) and Mate-Pair Library Prep Kit (Illumina, San Diego, CA) from according to the manufacturer's instructions. A bacterial artificial chromosome (BAC) library with an average length of 120 kbp was prepared with CopyControl pCC1BAC^{TM} vector by Amplicon Express Ltd (Washington, USA) and the BAC-end sequencing was performed in the Kazusa DNA Research Institute (Kisarazu, Japan). Whole genome shotgun (WGS) sequencing and BAC-end sequencing were done through Illumina HiSeq 2000 and Sanger ABI3730x1 platforms. Raw sequence data were filtered for bacterial genome contamination, PCR-duplicated reads and low quality reads were error-corrected. Paired-end Illumina reads were merged by FLASH to make longer single reads and the genome assembly and scaffolding were performed by Platanus[44] and by SSPACE[45]. The gene model predictions were performed using MAKER pipeline[46] using *S. asiatica* RNA sequencing described below. Details of read processing, assembly and annotation are described in Data S2A and B.

RNA sequencing. Total RNA was extracted from shoots and roots using the RNAeasy Plant Kit (Qiagen). Illumina PE libraries were constructed using the TruSeq RNA Sample Prep Kit (Illumina) and sequenced by an Illumina HiSeq2000 for 101 cycles per run. The obtained *S. asiatica* RNA sequences were quality-filtered and then used for the gene annotation pipeline and validation of the assembly. *S. hermonthica* sequences were quality trimmed with the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) using the fastq_quality_trimmer with option –l 60 and –t 30 and assembled by CLC genomics workbench (ver. 5) after removing host gene contamination (for details, see Data S2E). The sequence reads were mapped on *S. hermonthica de novo* assembled contigs concatenated with rice cDNAs by bowtie2. The contigs that are mapped with rice control reads were excluded from the subsequent analysis to avoid contamination of rice sequences. The normalised FPKM values were calculated by RSEM program (for details,

see Data S2E). After selecting genes in the upper 75% and 50% quartile of coefficient of variation for the expression across samples, scaled expression values within tissues were used to cluster these genes for a multilevel 3 x 4 hexagonal self-organising map (SOM). The outcome of SOM clustering was visualised in PCA space where PC values were calculated based on gene expression across samples (R stats package, prcomp function). GO enrichment analysis of contigs detected in SOM was performed using the GOSeq Bioconductor package[47] with Benhamini and Hochberg multiple hypothesis testing correction.

Genome comparative analysis. Maximum likelihood species tree for the 26 representative plant genomes were estimated using a concatenated matrix of trimmed codon alignments for genes from 1,440 BUSCO single copy orthogroups with RAxML[48] (Figure 1). Protein coding genes from 26 plant genomes (Data S1C) including *S. asiatica* were classified into orthogroups using the Orthofinder version 1.1.8 algorithm[49]. We further performed a second iteration of MCL[50] to connect distantly related orthogroups into superorthogroups as described in Wall et. al., 2009[51]. Amino acid sequence alignments for each orthogroup were generated with PASTA [52] using a maximum of five iterative refinements. Corresponding DNA codon alignments were trimmed using the heuristic automated method implemented in trimAl version 1.4.rev8[53]. Approximately-maximum likelihood (ML) analyses were conducted using FastTree version 2.1.10 [54], searching for the best ML tree with the GTR and GAMMA models. The unrooted FastTree phylogenies were traversed and rooted with the most distant taxa the orthogroup using rooting functions implemented in ETE Toolkit, a python phylogenetic framework [55]. The trees were examined for gene duplications in *Striga* and *Mimulus* and the detected duplications were scored using a scoring strategy similar to that described by Jiao et al., 2011[56]. A synonymous mutation (K_s) value for each duplicated sequence pair was calculated using the ML method implemented in CODEML[57] with a minimum alignment length of 300 bp. Structural syntenic analyses were performed with the SynMap tool[58] of the CoGe comparative genomics platform[59]. The genomes of *Mimulus* and *Vitis* were compared to the genome of *Striga* with the chaining algorithm DAGChainer[60]with a maximum distance of 20 genes between gene matches, and a minimum of 5 genes to seed a syntenic region. Scaffolds and contigs of *Striga* were ordered and oriented based on

their syntenic path to both *Mimulus* and *Vitis*. Parsimony method in DupliPHY[61] was used for reconstruction of the presence and size of each gene family in the common ancestor of *S. asiatica* and of the closely related non-parasite *Mimulus guttatus* as well as of other successively earlier ancestors. The numbers of evolutionary events were estimated using gene counts in each orthogroup or superorthogroup at each node of the 26-genome species tree. The tissue-specific orthogroups were defined using *Arabidopsis* microarray expression data[62]. These data are a curated summary of more than 5,000 microarray experiments conducted using the Agilent ATH1 GeneChip®. Further details are described in Supplementary Information Section 3. Comparison of the genomic regions containing *KAI2* paralogues was performed by GEvo tool in CoGe. The 60 kb regions containing each *KAI2*, *D14* or *DLK2* paralogue were submitted to GEvo with blastZ threshold score 15000. The data is visualised with Circos plot (http://circos.ca). Duplication origins of these loci were predicted as described in Supplementary Information Section 3.3.1.

RT-qPCR. Total RNAs were extracted as described above. cDNAs were synthesised using ReverTra Ace qPCR RT Kit (Toyobo, Japan) and quantitative PCRs were conducted using THUNDERBIRD SYBR qPCR kit (Toyobo, Japan) in Mx3000P qPCR system (Agilent Technologies). RT-qPCR was performed in three segments. Segment 1 consisted of 1 min at 95°C for one cycle, segment 2 consisted either of 15 s at 95°C and 30 s at 60°C for 40 cycles, or 15 s at 95°C, 30 s at 55°C and 30 s at 72°C for 40 cycles and segment 3 consisted of 1 min at 95°C, 30 s at 55°C, and 30 s at 95°C for one cycle. The primer sequences used are listed in Data S1S.

***In situ* hybridisation**. Preparation of DIG labelled RNA probe was performed as described previously[63]. The probe fragments were amplified by PCR from the cDNA library of rice infected with *S. hermonthica* using the primers listed in Data S1S. Sense or antisense probes with the length of 600-900 bp were generated using the T7 or SP6 polymerase (Roche) and DIG-UTP mix (Roche). The haustorial tissues attached with host rice were fixed in the freshly prepared PFA fixation buffer composed of 4% (w/v) paraformaldehyde in 1×PBS buffer (130 mM NaCl, 7 mM Na_2HPO_4, 3 mM NaH_2PO_4, pH 7.4 adjusted by NaOH). The samples were dehydrated by incubation in 1xPBS for 2.5 h and the concentration of ethanol was gradually increased at 4°C (30% for 1 h, 50% for 1h, 70% for overnight, 85% for 1h, 95% for overnight and 100% for 3 h). Samples were then permeabilised by incubation in gradually increasing concentrations of Histo-Clear

in ethanol at room temperature (Histo-Clear and ethanol mixture of 1:3 for 1 h, 1:1 for 1h, 3:1 for 1 h and 100% of Histo Clear for 2 h) and in a 1:1 mixture of Histo-Clear and paraffin for 1 h at 60°C. Paraffin was changed 6 times before being embedded on wooden blocks. We followed the steps of *in situ* hybridisation as described previously[63] with minor modifications; a concentration of 10 µg/ml^{-1} of the probes was used and the use of levamisole in the detection solution was omitted. The images of *in situ* hybridised samples were taken using the light microscopy BX-51 (Olympus).

Identification of horizontally transferred genes and retrotransposons. To analyse the *S. asiatica* genome for genes horizontally transferred from grass host species, the *S. asiatica* annotation was subjected to a BLASTp search with the threshold e-value 1e-10 against a database of combined predicted proteins from the genome of 28 different plant species, including *Striga* host plants, rice, sorghum, foxtail millet, and maize. *S. asiatica* proteins having at least one hit to grass species in their top 20 hits were selected, and modified Alien Index (AI) values[64] were calculated with the following formula: Modified AI = log((Best E-value for dicots) + 1e-200) - log((Best E-value for grasses) + 1e-200). Genes having modified AI >30 and genes that did not have a dicot hit were selected for further analysis. Using the RAxML program, maximum-likelihood phylogenetic trees were drawn of BLASTp-hit homolog genes from the 28-species database as well as from the non-redundant (nr) database. Manual investigation of the phylogenetic trees found 34 positive HGT candidate genes, which were assigned into 20 orthogroups by orthoMCL analysis. A few of HGT candidates are near each other in the genome, and therefore the genomic regions were compared using CoGE with the GEvo function.

For identification of horizontally transferred retrotransposons, superfamily *Copia* and *Gypsy* elements were retrieved, using LtrHarvest[65] and LtrDigest[66], from the genome sequences of *S. asiatica* and those of the monocots *Sorghum bicolor, Zea mays, Oryza sativa* ssp. *japonica* and ssp. *indica, O. rufipogon,* and *O. glaberrima* and the eudicots *Glycine max, Solanum tuberosum,* and *Vitis vinifera*. The *rt* sequences were clustered and the *S. asiatica rt* sequences that were found in clusters mixed with those of other genomes were treated further. These were characterised by exonerate-search[67] using known *rt* sequences from GypsyDB[68] and clustered by homology search against each other (BLASTn -evalue 1e-20) and subsequently by silix-software[69] (silix -i 0.60 -r 0.70). The resulting clusters were aligned with the clustal-omega[70] and prank-ms[67]

multiple aligners and phylogenetic trees were constructed by FastTree (fasttree –nt –gtr –gamma)[54]. The details of HGT analysis are described in Data S2F.

Quantification and Statistical Analysis

Statistical analyses for GO enrichment was performed either chi square test or fisher's exact test with Benjamini and Hochberg correction for multiple samples. Other statistical analyses were performed with two-tailed Mann-Whitney U test, Student's t-test, or one-way ANOVA combined with the post hoc Tukey-Kramer test as indicated in the text or figure legends. Error bars represent SEM.

Data and Code Availability

S. asiatica genome and transcriptome sequence data are deposited in DDBJ as accession number DRA007962 and DRA008308. The *S. hermonthica* RNA-seq data are available as accession numbers DRA008615 and DRA003608 in DDBJ. *S. hermonthica* and *S. gesnerioides* genome sequence raw reads are deposited in Genbank as accession number PRJNA551337 and PRJNA551339, respectively. *S. asiatica* genome assembly and annotation, *S. hermonthica* transcriptome assembly and annotation, and horizontally transferred retrotransposon sequences are available at Dryad data repository (http://datadryad.org/reource/doi:10.5061/dryad.53t3574).

All bioinformatic analyses were performed with open-source or commercially available software. Perl, Python or R scripts were used for run each software according to software manuals.

Supplemental Figures and data

Figure S1. Kn/Ks ratios between *Striga* and *Mimulus* orthologues in expanded and contracted gene families. Related to Figure 2

Figure S2. Stage-specific gene expression in *S. hermonthica*. Related to Figure 4

Figure S3. Expression patterns of lateral root development gene orthologues in *S. asiatica* during host infection. Related to Figure 6

Figure S4. Phylogenetic tree of RT domains of HGT candidate retrotransposons. Related

to Figure 7.

Table S1. Primers used in this study. Related to STAR method.

Data S1. Data summary tables. Related to Figure 1-7 and STAR method.

Data S2. Supplemental information. Related to Figure 1-7 and STAR method.

Data S3. Gene lists in each orthogroup in 26 plant species. Related to Figure 2 and STAR method.

References

1. Ejeta, G. (2007). The Striga scourge in Africa: A growing pandemic. In Integrating new technologies for Striga control, towards ending the witch-hunt., G. Ejeta and J. Gressel, eds. (London, UK.: World Scientific Publishing Co.), pp. 3–16.

2. Iverson, R.D., Westbrooks, R.G., Eplee, R.E., and Tasker, A. V. (2011). Overview and status of the witchweed (Striga asiatica) eradication program in the Carolinas. In Invasive plant management issues and challenges in the United States : 2011 overview, A. R. L. and R. G.Westbrooks, ed. (Washington, DC), pp. 51–68.

3. Spallek, T., Mutuku, M., and Shirasu, K. (2013). The genus *Striga*: a witch profile. Mol Plant Pathol *14*, 861–869.

4. Cook, C.E., Whichard, L.P., Turner, B., and Wall, M.E. (1966). Germination of witchweed (*Striga Lutea* Lour) - isolation and properties of a potent stimulant. Science *154*, 1189–1190.

5. Al-babili, S., and Bouwmeester, H.J. (2015). Strigolactones, a novel carotenoid-derived plant hormone. Ann Rev Plant Biol *66*, 161–186.

6. Joel, D.M. (2013). The haustorium and the life cycles of parasitic Orobanchaceae. In Parasitic Orobanchaceae, L. J. Joel, Daniel M., Gressel, Jonathan, Musselman, ed. (Springer), pp. 21–24.

7. Chang, M., and Lynn, D.G. (1986). The haustorium and the chemistry of host

recognition in parasitic angiosperms. J. Chem. Ecol. *12*, 561–579.

8. Yoshida, S., Maruyama, S., Nozaki, H., and Shirasu, K. (2010). Horizontal gene transfer by the parasitic plant *Striga hermonthica*. Science *328*, 1128.

9. Westwood, J.H., Yoder, J.I., Timko, M.P., and dePamphilis, C.W. (2010). The evolution of parasitism in plants. Trends Plant Sci *15*, 227–235.

10. Yang, Z., Zhang, Y., Wafula, E.K., Honaas, L.A., Ralph, P.E., and Jones, S. (2016). Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. Proc Natl Acad Sci U S A *113*, E7010-7019.

11. Searcy, D., and MacInnis, A.J. (1970). Measurements by DNA renaturation of the genetic basis of parasitic reduction. Evolution (N. Y). *24*, 796–806.

12. dePamphilis, C.W. (1995). Genes and genomes. In Parasitic Plants, M. Press and J. Graves, eds. (Chapman and Hall).

13. Yang, Z., Wafula, E.K., Honaas, L.A., Zhang, H., Fernandez-aparicio, M., Huang, K., and Bandaranayake, P.C.G. (2015). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. Mol. Biol. Evol. *32*, 767–790.

14. Westwood, J. (2013). The physiology of the established parasite-host association. In Parasitic Orobanchaceae, D. M. Joel, J. Gressel, and L. J. Musselman, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg).

15. Rogers, W.E., and Nelson, R.R. (1962). Penetration and nutrition of *Striga asiatica*. Phytopathology *52*, 1064–1070.

16. Wicke, S., Muller, K.F., de Pamphilis, C.W., Quandt, D., Wickett, N.J., Zhang, Y., Renner, S.S., and Schneeweiss, G.M. (2013). Mechanisms of functional and physical genome reduction in photosynthetic and nonphotosynthetic parasitic plants of the broomrape family. Plant Cell *25*, 3711–3725.

17. Wickett, N.J., Honaas, L.A., Wafula, E.K., Das, M., Huang, K., Wu, B., Landherr, L., Timko, M.P., Yoder, J., Westwood, J.H., *et al.* (2011). Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising

254 conservation of chlorophyll synthesis. Curr. Biol. *21*, 2098–104.

255 18. Press, M.C., Tuohy, J.M., and Stewart, G.R. (1987). Gas exchange characteristics
256 of the sorghum-striga host-parasite association. Plant Physiol. *84*, 814–819.

257 19. Fujioka, H., Samejima, H., Suzuki, H., Mizutani, M., Okamoto, M., and
258 Sugimoto, Y. (2019). Aberrant protein phosphatase 2C leads to abscisic acid
259 insensitivity and high transpiration in parasitic Striga. Nat. Plants *5*, 258–262.

260 20. Press, M.C., Smith, S., and Stewart, G.R. (1991). Carbon Acquisition and
261 Assimilation in Parasitic Plants. Funct. Ecol. *5*, 278–283.

262 21. Waters, M.T., Nelson, D.C., Scaffidi, A., Flematti, G.R., Sun, Y.K., Dixon,
263 K.W., and Smith, S.M. (2012). Specialisation within the DWARF14 protein
264 family confers distinct responses to karrikins and strigolactones in Arabidopsis.
265 Development *139*, 1285–1295.

266 22. Scaffidi, A., Waters, M.T., Sun, Y.K., Skelton, B.W., Dixon, K.W., Ghisalberti,
267 E.L., Flematti, G.R., and Smith, S.M. (2014). Strigolactone hormones and their
268 stereoisomers signal through two related receptor proteins to induce different
269 physiological responses in Arabidopsis. Plant Physiol. *165*, 1221–1232.

270 23. Conn, C.E., Bythell-douglas, R., Neumann, D., Yoshida, S., Whittington, B.,
271 Westwood, J.H., Shirasu, K., Bond, C.S., Dyer, K.A., and Nelson, D.C. (2015).
272 Convergent evolution of strigolactone perception enabled host detection in
273 parasitic plants. Science *349*, 540–543.

274 24. Tsuchiya, Y., Yoshimura, M., Sato, Y., and Kuwata, K. (2015). Probing
275 strigolactone receptors in Striga hermonthica with fluorescence. Science (80-.).
276 *349*, 864–868.

277 25. Toh, S., Holbrook-smith, D., Stogios, P.J., Onopriyenko, O., Lumba, S.,
278 Tsuchiya, Y., Savchenko, A., and Mccourt, P. (2015). Structure-function analysis
279 identifies highly sensitive strigolactone receptors in Striga. Science *350*, 203–
280 208.

281 26. Yoshida, S., and Shirasu, K. (2009). Multiple layers of incompatibility to the
282 parasitic witchweed, *Striga hermonthica*. New Phytol *183*, 180–189.

27. Visser, J.H., Inge, D., and Kollmann, R. (1984). The “hyaline body” of the root parasite *Alectra orobanchoides* benth. (Scrophulariaceae)—Its anatomy, ultrastructure and histochemistry. Protoplasma *121*, 146–156.

28. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, 490–495.

29. Peter, I., and Davidson, E. (2011). Evolution of gene regulatory networks that control embryonic development of the body plan. Cell *144*, 970–985.

30. Lavenus, J., Goh, T., Roberts, I., Guyomarc’h, S., Lucas, M., De Smet, I., Fukaki, H., Beeckman, T., Bennett, M., and Laplaze, L. (2013). Lateral root development in *Arabidopsis*: Fifty shades of auxin. Trends Plant Sci. *18*, 1360–1385.

31. Swarup, K., Benková, E., Swarup, R., Casimiro, I., Péret, B., Yang, Y., Parry, G., Nielsen, E., De Smet, I., Vanneste, S., *et al.* (2008). The auxin influx carrier LAX3 promotes lateral root emergence. Nat. Cell Biol. *10*, 946–54.

32. Berckmans, B., Vassileva, V., Schmid, S.P.C., Maes, S., Parizot, B., Naramoto, S., Magyar, Z., Kamei, C.L. a., Koncz, C., Bogre, L., *et al.* (2011). Auxin-dependent cell cycle reactivation through transcriptional regulation of *Arabidopsis* E2Fa by lateral organ boundary proteins. Plant Cell *23*, 3671–3683.

33. Ishida, J.K., Yoshida, S., Ito, M., Namba, S., and Shirasu, K. (2011). *Agrobacterium rhizogenes*-mediated transformation of the parasitic plant Phtheirospermum japonicum. PLoS One *6*, e25802.

34. De Smet, I., White, P.J., Bengough, a G., Dupuy, L., Parizot, B., Casimiro, I., Heidstra, R., Laskowski, M., Lepetit, M., Hochholdinger, F., *et al.* (2012). Analyzing lateral root development: how to move forward. Plant Cell *24*, 15–20.

35. Gifford, M.L., Dean, A., Gutierrez, R. a, Coruzzi, G.M., and Birnbaum, K.D. (2008). Cell-specific nitrogen responses mediate developmental plasticity. Proc. Natl. Acad. Sci. U. S. A. *105*, 803–808.

36. Smet, I. De, Vassileva, V., Rybel, B. De, Levesque, M.P., Grunewald, W.,

Damme, D. Van, Noorden, G. Van, Naudts, M., Isterdael, G. Van, Clercq, R. De, *et al.* (2008). iReceptor-Like Kinase ACR4 Restricts Formative Cell Divisions in the Arabidopsis Root Ive. Science (80-.). *322*, 594–597.

37. DiDonato, R.J., Arbuckle, E., Buker, S., Sheets, J., Tobar, J., Totong, R., Grisafi, P., Fink, G.R., and Celenza, J.L. (2004). Arabidopsis ALF4 encodes a nuclear-localized protein required for lateral root formation. Plant J. *37*, 340–353.

38. Kim, G., LeBlanc, M.L., Wafula, E.K., DePamphilis, C.W., and Westwood1, J.H. (2014). Genomic-scale exchange of mRNA between a parasitic plant and its hosts. Science *345*, 808–811.

39. 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*.

40. 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*, 4–11.

41. Hirayama, K., and Mori, K. (1999). Synthesis of (+)-Strigol and (+)-Orobanchol, the germination stimulants, and their stereoisomers by employing lipase-catalyzed asymmetric acetylationas the key step. European J. Org. Chem. *1999*, 2211–2217.

42. Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, C.A. and S.L.S., Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*.

43. Yoshida, S., and Shirasu, K. (2009). Multiple layers of incompatibility to the parasitic witchweed, Striga hermonthica. New Phytol *183*, 180–189.

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

45. Boetzer, M., Henkel, C. V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

46. 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.* (2013). MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. Plant Physiol. *164*, 513–524.

47. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

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

49. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 1–14.

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

51. Wall, P.K., Leebens-Mack, J., Chanderbali, A.S., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L.P., Hu, Y., Carlson, J.E., *et al.* (2009). Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics *10*, 347.

52. Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J. Comput. Biol. *22*, 377–386.

53. 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.

54. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One *5*.

55. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. *33*, 1635–1638.

56. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.

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

58. Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics *12*, 102.

59. Tang, H., and Lyons, E. (2012). Unleashing the Genome of Brassica Rapa. Front. Plant Sci. *3*, 1–12.

60. Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: A tool for mining segmental genome duplications and synteny. Bioinformatics *20*, 3643–3646.

61. Ames, R.M., Money, D., Ghatge, V.P., Whelan, S., and Lovell, S.C. (2012). Determining the evolutionary history of gene families. Bioinformatics *28*, 48–55.

62. Zimmermann, P., Bleuler, S., Laule, O., Martin, F., Ivanov, N. V, Campanoni, P., Oishi, K., Lugon-moulin, N., Wyss, M., Hruz, T., *et al.* (2014). ExpressionData - A public resource of high quality curated datasets representing gene expression across anatomy , development and experimental conditions. BioData Min. *7*, 18.

63. Hejátko, J., Blilou, I., Brewer, P.B., Friml, J., Scheres, B., and Benková, E. (2006). In situ hybridization technique for mRNA detection in whole mount Arabidopsis samples. Nat. Protoc. *1*, 1939–1946.

64. Gladyshev, E.A., Meselson, M., and Arkhipova, I.R. (2008). Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science (80-.). *320*, 1210–1214.

65. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC

Bioinformatics *9*, 18.

66. Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. *37*, 7002–7013.

67. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

68. Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., *et al.* (2011). The Gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. *39*, 70–74.

69. Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics *12*, 116.

70. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., *et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
(+)-Strigol	Prof. Kenji Mori (Tokyo Univ.)	
MS salts	Wako Chemical, Kyoto, Japan	# 392-00591
Phytagel	Sigma & Aldrich	# P8169
Critical Commercial Assays		
Nucleon Phytopure DNA extraction kit	GE healthcare	#RPN8510
TruSeq DNA Sample Prep kit	Illumina	#15026486
Mate-Pair library Prep kit	Illumina	# PE-930-1003
RNAeasy Plant Kit	Qiagen	# 74904
TruSeq RNA Sample Prep kit	Illumina	# RS-122-2001
ReverTra Ace qPCR RT Kit	Toyobo	# FSQ-201
THUNDERBIRD SYBR qPCR kit	Toyobo	# QPS-201
Deposited Data		
S. asiatica genome sequence raw data	DRA007962	
S. asiatica transcriptome sequence raw data	DRA008308	
S. hermonthica transcriptome sequence raw data	DRA008615, DRA003608	
S. hermonthica genome sequence raw data	PRJNA551337	
S. gesnerioides genome sequence raw data	PRJNA551339	
S. asiatica genome assembly and annotation	xxxx	
S. hermonthica transcriptome assembly	xxxx	
S. asiatica BAC-end sequences	xxxx	
Experimental Models: Organisms/Strains		
Striga asiatica	Provided from Prof. Mike Timko (U. Virginia, VA, USA)	UVA1
Striga hermonthica	Provided from Prof. Abdel G. E. Babiker (Environment and Natural Resources and Desertification Research Institute, Sudan)	
Oryza sativa (japonica, c.v. Koshihikari)	Rice Genome Resource Center (RGRC), Tsukuba, Japan	
Arabidopsis thaliana (ecotype: Col-0)	Arabidopsis biological resource center (ABRC)	Col-0
Lotus japonicus (ecotype: MG-20)	Legume base (https://www.legumebase.brc.miyazaki-u.ac.jp/lotus/)	
Oligonucleotides		

See Data S1S		
Software and Algorithms		
FLASH	https://ccb.jhu.edu/software/FLASH/	
Platanus	[44]	
SSPACE	[45]	
MAKER-P	[46]	
CLC assembly cell	https://filgen.jp/Product/BioScience21-software/CLC/index11-g.htm	
CLC genomic workbench	https://filgen.jp/Product/BioScience21-software/CLC/index11-g.htm	
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	
RSEM	http://deweylab.github.io/RSEM/	
Orthofinder	[49]	
RaxML	[48]	
FastTree	[54]	
trimAl	[53]	
PASTA	[52]	
CODEML	[57]	
CoGE	https://genomevolution.org	
DAGChainer	[60]	
DupliPHY	[61]	
RaxML	[52]	
LtrHarvest	[65]	
LtrDigest	[66]	
Other		
Striga asiatica genome and predicted genes and protein in multifasta format, annotation in gff3 file format.	https://datadryad.org/	doi:10.5061/dryad.53t3574
Striga hermonthica transcriptome assembly and predicted protein sequences in multifasta format, and functional annotation and GO information	https://datadryad.org/	doi:10.5061/dryad.53t3574
Retrotransposon sequences and phylogenetic trees appeared in Figure 7E and Figure S4	https://datadryad.org/	doi:10.5061/dryad.53t3574

Figure 1

Click here to access/download;Figure;figures-01.tif

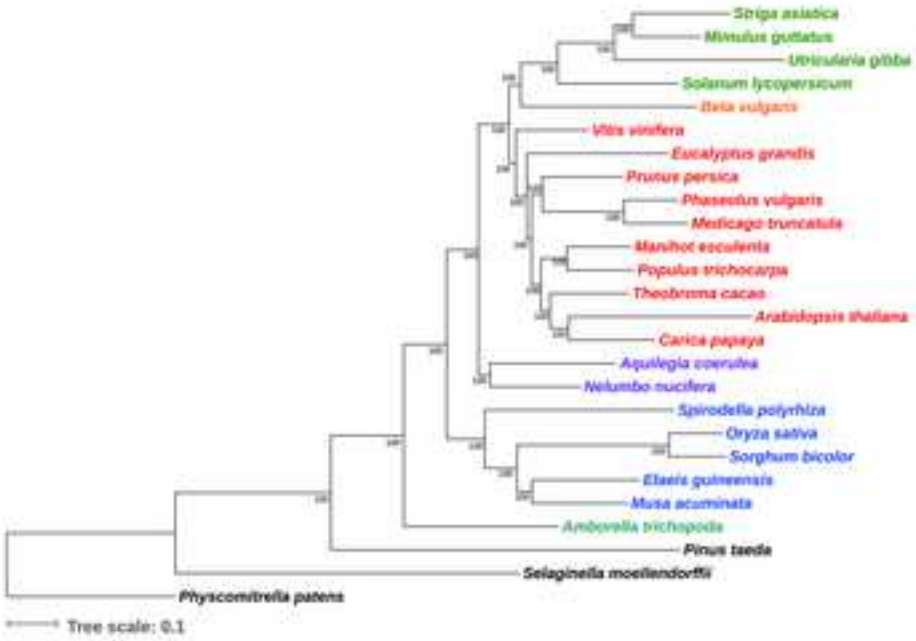
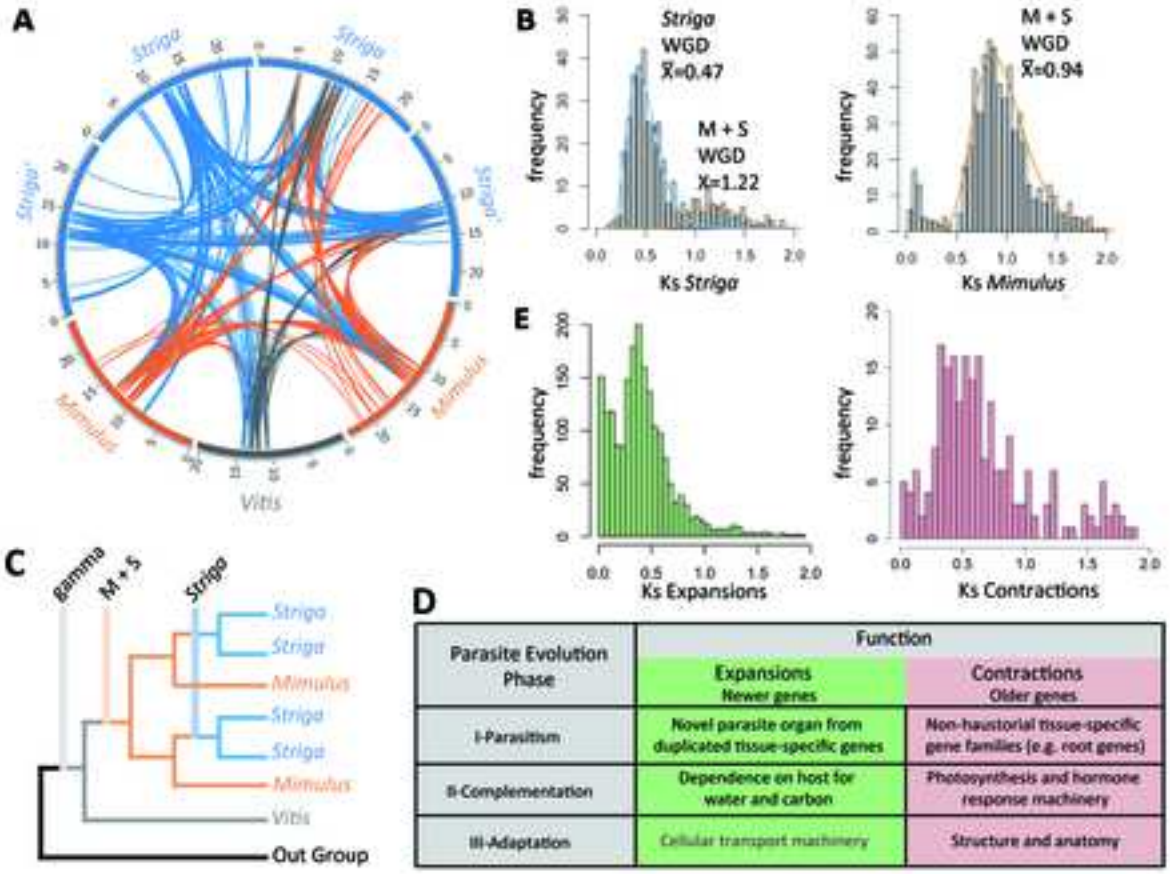


Figure 2



Parasite Evolution Phase	Function	
	Expansions Newer genes	Contractions Older genes
I-Parasitism	Novel parasite organ from duplicated tissue-specific genes	Non-haustorial tissue-specific gene families (e.g. root genes)
II-Complementation	Dependence on host for water and carbon	Photosynthesis and hormone response machinery
III-Adaptation	Cellular transport machinery	Structure and anatomy

Figure 3

Click here to access/download;Figure;figures-03.tif

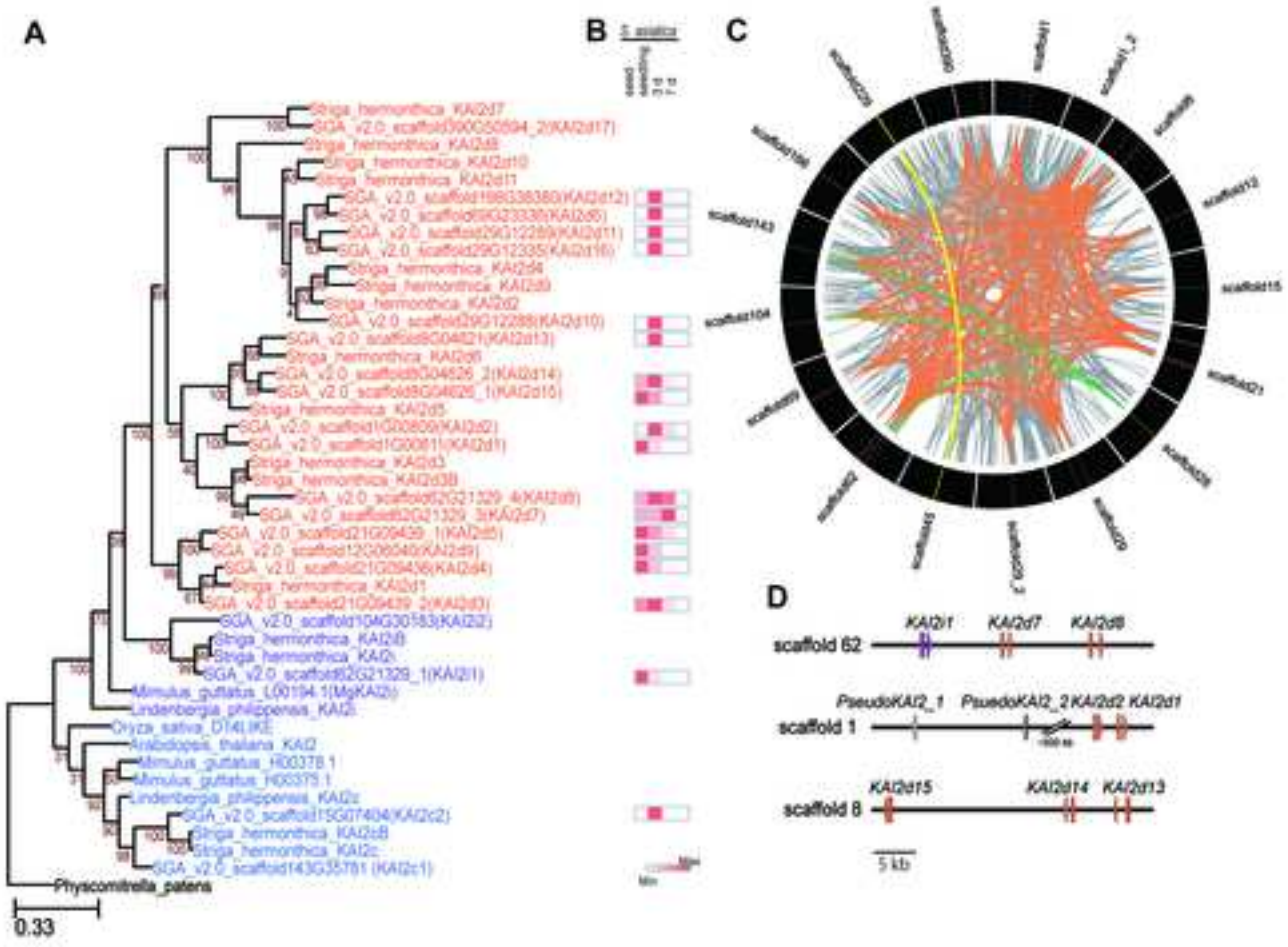


Figure 4

Click here to access/download;Figure;figures-04.tif

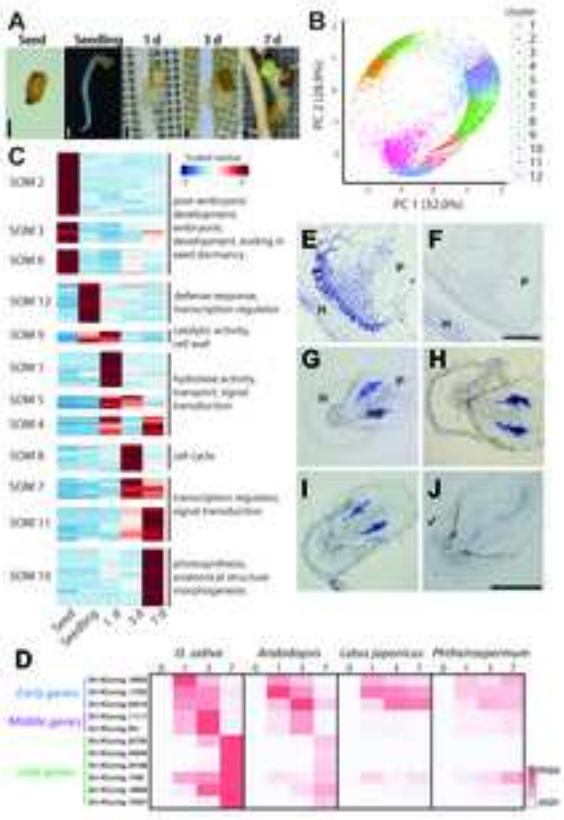


Figure 5

Click here to access/download;Figure;figures-05.tif

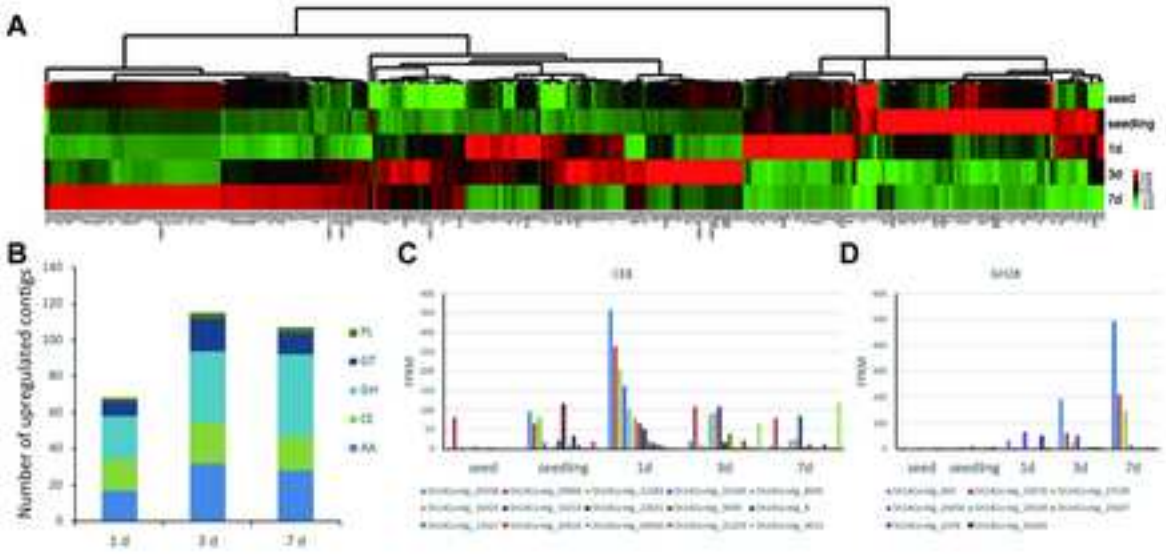


Figure 6

Click here to access/download;Figure;figures-06.tif

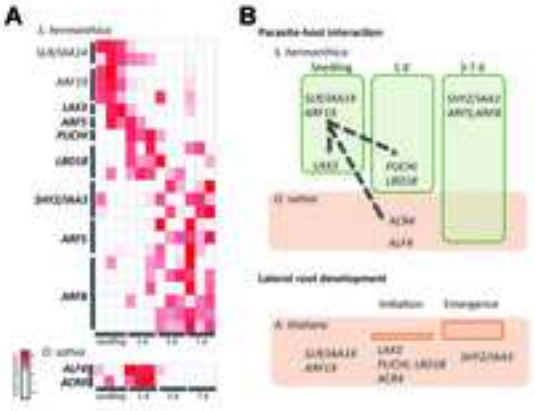
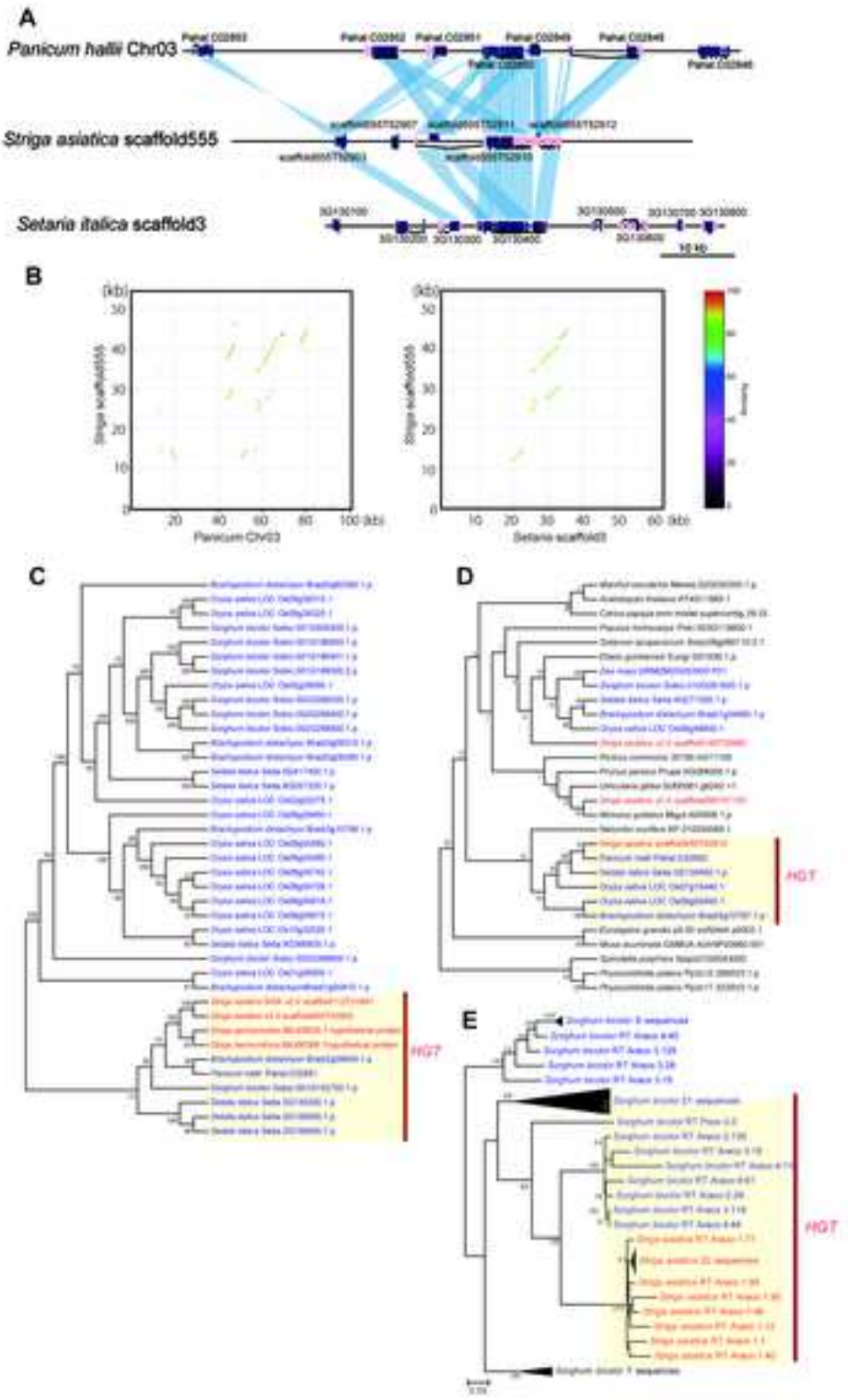


Figure 7

Click here to access/download;Figure;figures-07.tif



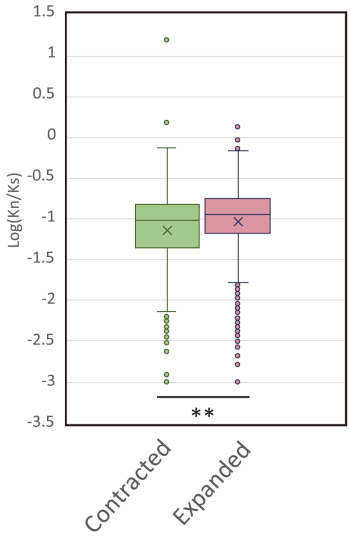


Figure S1. Kn/Ks ratios between *Striga* and *Mimulus* orthologues in expanded and contracted gene families. Related to Figure 2.

Ratios of non-synonimous and synonymous substitutions between *S. asiatica* and *M. guttatus* orthologous genes present in syntenic regions were calculated and plotted depending on their evolutional categories in *S. asiatica* genome. Expanded gene families show significantly higher Kn/Ks ratio compared to contracted gene families (Student's t-test, $p<0.00001$).

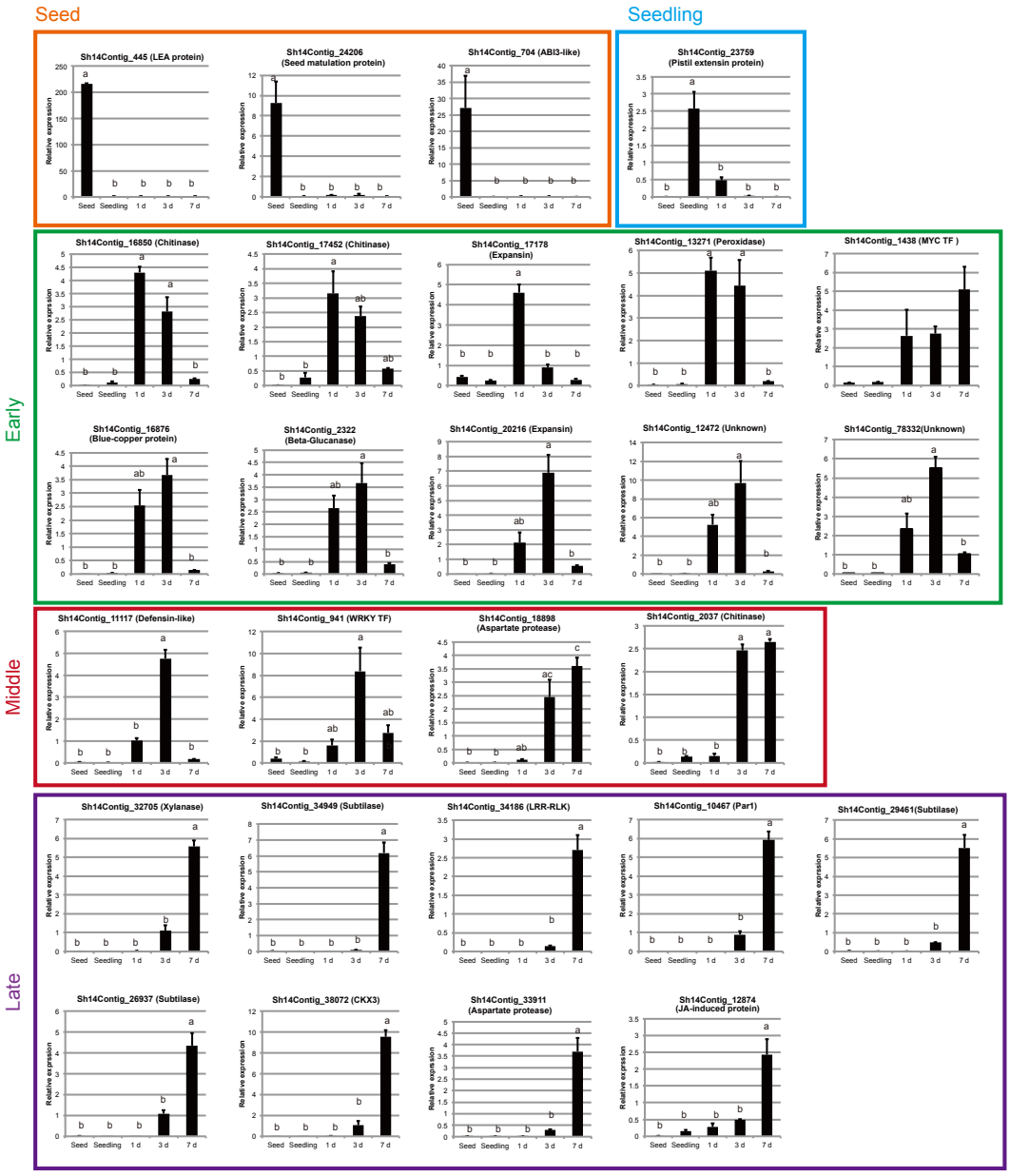


Figure S2. Stage-specific gene expression in *S. hermonthica.* Related to Figure 4.

RT-qPCR confirmation of stage-specific expression of selected genes. Relative expression values were normalised by expression of an internal control gene (CHYLOPHILIN). Each value was obtained as the mean of three biological replicates with SE. Statistically significant differences were tested by Tukey's test and shown in different alphabetic characters ($P<0.05$).

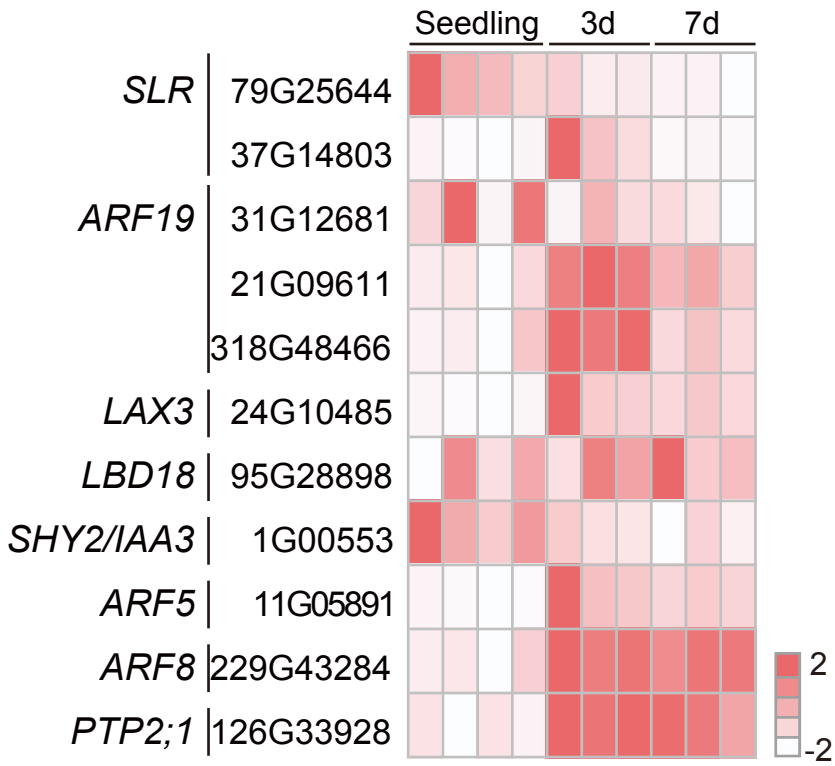


Figure S3. Expression patterns of lateral root development gene orthologues in *S. asiatica* during host infection. Related to Figure 6.

The relative expression levels of the LRD genes in *S. asiatica* were measured by RT-qPCR. Seedlings were sampled at 2 d after strigol treatment, and for 3 d and 7 d samples, the *S. asiatica* plants were harvested at 3 d and 7 d post infection of rice roots (cv. Koshihikari).

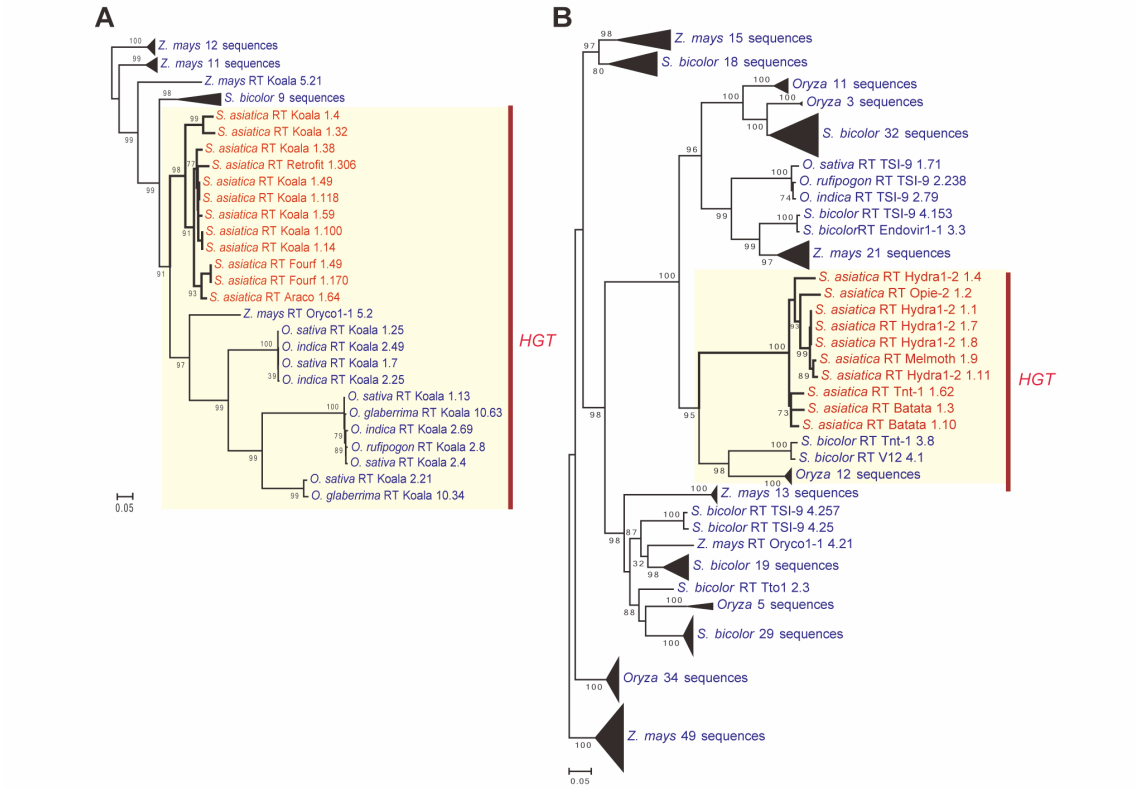


Figure S4. Phylogenetic tree of RT domains of HGT candidate retrotransposons. Related to Figure 7.

A, B. Unrooted phylogenetic trees for two RT sequences from S. asiatica *Copia* nested to Poaceae sequences drawn by FastTree (v.2.1.10). Local support values are shown at each node, and were calculated by the Shimodaira-Hasegawa test on the three alternate topologies (NNIs). Clades involving the horizontal transfer events are highlighted with pink background.

Table S1. Primers used in this paper. Related to STAR method.

Annotation/Contig ID	Forward primer sequence	Reverse primer sequence	Gene annotation
Primers used for RT-qPCR in S. asiatica			
SGA2.0.scaffold15G07404	TCAATTTGGCCGTGCAATG	CATGGAAACTGAGGGTCATCT	SaKAI2c1
SGA2.0.scaffold62G21329_1	GAGGTCATCAACACCGAAGG	CCGGCCACCCAAAGAAT	SaKAI2i1
SGA2.0.scaffold1G00812	CCTCACTCACTTGTCTGCAAT	AATACCTCCGTCGGAACCT	SaKAI2d1
SGA2.0.scaffold1G00810	AGTCACTCGTCTGCAATGTC	GTGGGAAGTCCGTCGTG	SaKAI2d2
SGA2.0.scaffold21G09436	GGTCACGATTCCTGTGATTCT	AAAGGAGAATGGGCACCTAAA	SaKAI2d3
SGA2.0.scaffold21G09439_1	GAGGTCACGGTTCCAATCAT	CGAAGACTTGTCAAATCCTAATGG	SaKAI2d4
SGA2.0.scaffold21G09439_2	CCCGTGATTCTCCGTCATATAAA	AAGGAGAATGCGCACCTAAA	SaKAI2d5
SGA2.0.scaffold69G23336	ATACATATCGGACCGACACCGGA	TAGACGGTGCTAATTACTTTAGC	SaKAI2d6
SGA2.0.scaffold62G21329_3	GTCGTATGATATCGGGCCTTGAC	CACCTCCACCACAGACTTAC	SaKAI2d7
SGA2.0.scaffold62G21329_4	CACTCCTCCGCCACATAAAT	AGCGTTACCAAACAAGCTCTA	SaKAI2d8
SGA2.0.scaffold12G06040	GTCAAGTCCTAATGGTGGGT	TCCGATCATTCTTCGCCATATAA	SaKAI2d9
SGA2.0.scaffold29G12288	ACGTGACAAGTTATGCTTTAGGA	GTTATTGGCCGGTGCTAGTTA	SaKAI2d10
SGA2.0.scaffold29G12289	AGCACCTCTTACTGTTACTCTTG	GGGCTTGGTTTGATGTCATTAG	SaKAI2d11
SGA2.0.scaffold166G38380	CTGCTTCCACACCGACTG	GTTGGATCGGTTCATCGTCATA	SaKAI2d12
SGA2.0.scaffold8G04626_1	CATCGCGGATCAGTGAAGAT	TAACTACCACACACACACACTC	SaKAI2d13
SGA2.0.scaffold8G04626_2	AAGACAGGACATCGAGGTTTAG	CACACACATACACTCACACTTTC	SaKAI2d14
SGA2.0.scaffold8G04621	GCCACATCAGACAAGACATCA	CACACACACACACACTCTCTC	SaKAI2d15
SGA2.0.scaffold29G12335	TCATAAACCCGGTGTTGCTC	CTACAAGATGTCCTGGCGTATAG	SaKAI2d16
SGA.2.0.scaffold11G05891	TTGGAGGCCTTGTGTACTATTT	GGGAGAGATTCGGATAGTTTGG	ARF5
SGA.2.0.scaffold229G43284	GGCTATCAGAACCCTCTGTATG	CCAATGTCCAATGACCTACCA	ARF8
SGA.2.0.scaffold79G25644	CTGCTGATTCCGACCCAAA	TTGGCAGGTGGTTTCATAACC	SLR/IAA14
SGA.2.0.scaffold37G14803	AGGCGATGTAATGAACGAGAA	TCTCCAACTAACATCCAATCCC	SLR/IAA14
SGA.2.0.scaffold1G00553	CGGAACGCGAAGGCTATAAA	CGAGCCTCTCATGATCCTTAATC	SHY2/IAA3
SGA.2.0.scaffold24G10485	AGGTTGCCAGTGGTTATTCC	GTGATCCAACGGTCGAGTTTAT	LAX3
SGA.2.0.scaffold126G33928	CATTGGGTTTGCCGTGTTC	GCCTTGTCCTTCCCGTAAAT	PIP2;1
SGA.2.0.scaffold95G28898	ATGCGGTGGTCACGATATG	TGCTGGAGGGCAAAGATG	LBD18
SGA.2.0.scaffold92G28259	CCATCGGAAGTTCAGCAGAT	ACTTCCGAGATTAAGCCGTTATTA	ARF5
SGA.2.0.scaffold162G37941	GGAAGAGGGAATGCAGCTT	AGTACATAGGTTAAGACCCATCTTT	ARF5
SGA.2.0.scaffold31G12681	AGTGAAGGCACCTGCATAAA	ATGCTTGGAAAGTCCACTATGA	ARF19
SGA.2.0.scaffold21G09611	ACTCCGCTCGTTAATATTCATG	GGTTTGGGTAGTTCGGGATTT	ARF19
SGA.2.0.scaffold318G48466	TGAGCTTGGATGGCGATTT	GGAAGAAGAATAAGTTGGCATTGT	ARF19
SGA.1.0.scaffold382G00010	GTAATGGGACTGGTGGAGAATC	CCCTGCATTTGCCATTGATAATA	SaCyclophilin (for *S. asiatica* internal control)
SGA.2.0.scaffold119G32689	GTGGGAAGACTAAACCGCCT	GATACACTCTCGCAGAGCCG	SaRPS2(for *S. asiatica* internal control)
Primers used for in situ hybridisation in S. hermonthica			
Sh14Contig_26937	TACAGGGACCTCCTCCTCCT	TTTTAGGAGGGCAACAATGC	Subtilase1
Sh14Contig_34949	AAGCACGATCGACAGGAGGTT	ACCAGTCGGGATGTGCACTT	Subtilase2
Sh14Contig_33911	AATCCG GCTGTACCTTTCCT	CTGGTCCGTGGAAGTCTGAT	Aspartate protease
Sh14Contig_34186	TGTGCATACTGCCATGTCTG	TGGTGTGGCTTATGTCCAGA	LRR kinase
Sh14Contig_38072	ATTCCACGTGGGGACAATCC	TTGACGGTGTGGACAGTCTG	Cytokinin dehydrogenase
Sh14Contig_13271	AGACGGCTATCCCAAACCAA	GCCGAAGAATTTCAACGCGA	Peroxidase
Primers used for qRT-PCR in S. hermonthica			
Sh14Contig_11117	CCCATCACCAAATCATTACTGC	CGTATGCATGGCTTCTCAAAAT	defensin-like protein
Sh14Contig_20216	TCCAGAGCTTGAATCTGGTGAA	TCGGCAAACTGAAGAATTTACG	LRR kinase
Sh14Contig_38072	ATGGCGAAGGTCTTGTTTGTTT	AATTCCGTTTTTGGCCCTAAGT	cytokinin dehydrogenase
Sh14Contig_10467	TTGAGATGGCTAGGGAAAGGAC	TCCCCTAATAGCAAAGCAAAGC	photoassimilate responsive protein Par1
Sh14Contig_18898	CAGTACGGAGCCTCCAAGTTCT	CACCCCACATCATGACATCTTT	aspartate protease
Sh14Contig_2037	ACTGGATTGGATCGGGTATGAC	CATTGACAGCCCAGAAGAAGTG	mammalian chitinase
Sh14Contig_12874	CCCCTTACCCTCATGTTATCCA	TGTAGACGATTGCCTCCTTTGA	Jasmonate-induced protein
Sh14Contig_32705	ACGGCCCAGCTATATTTTGAGA	CTTGGTGGGATTTCCACTCTTC	endo-beta xylanase c
Sh14Contig_13817	CATTGTCGTCCTCGTCATTGAT	AGGTGGACAAGACGAAGAAAGG	polyphenol oxidase
Sh14Contig_445	GGAAACTAGATCCGACCCGTTA	CATAAACCCCACAACAGAACGA	LEA protein
Sh14Contig_24206	GGATTCAGATCGACAAGATCCA	GCCTAGATCGTCCTTGTTCTCG	seed maturation protein
Sh14Contig_704	TGCACCTCTCAAGCTAGCCATA	GAAAAACGAGCAAAAGCCACTT	abi3-like
Sh14Contig_23759	GCTGGAGAGGAAAACCAAGAAA	ATCAAGAACACCCGGCAATATC	120 kda pistil extensin-like protein
Sh14Contig_941	ATTTCTGGCTCGTGCATCTGTA	TCGACAATCTTGAGGACGGATA	WRKY transcription factor
Sh14Contig_1438	GAAATTTCGCACGAATTCCCTA	GTTTTTCATGCTGCTACGGTTG	MYC transcription factor
Sh14Contig_16850	CCTGCCCTCGATTTACTCACTG	GCCACAGTAGTCATCGGTTGTG	class iv chitinase
Sh14Contig_2322	GAAGTGGCCTCGTACATCAACC	GTGAAGAGCGCGTAGTCCAAGT	beta-glucanase
Sh14Contig_33911	ATTATTGTTGTGGCTGGCTGCT	ATTCCACTCTCGGCAATTTTCA	aspartate protease
Sh14Contig_13271	ATGGGTGCCGGGATTGTCTC	CCGTGGGCAGTTGAAGGTCG	proxidase precursor
Sh14Contig_17452	ACCGCGCGGACATTATCGTA	GACGTACGGCCAGATCGTGA	chitinase
Sh14Contig_78332	GGCCCTCTCGGCTTCATAGC	CGAGAATAACGTTGGGGTGCCT	no hit
Sh14Contig_17178	TGCGGTGGCCATAGAGTACG	CCACTTTCCAAACGGAACCCC	beta-expansin
Sh14Contig_16876	GCGCGACACAATTTGGTACCTGTT	ATGTCCCGGCCTTATTTAGCGTCA	blue copper protein
Sh14Contig_12472	TTACCATAACCGTCAAGCGCAAGC	ACTCCGTCAGCTCCATACAAACCA	unknown protein
Sh14Contig_20216	GAAACGATGTTAACGCGTGCGGAA	TGGCCCGAGCATATATCCAACGAA	expansin b1
Sh14Contig_26937	GTGTCGATAAGCCCAACGAT	CACCACAAGAAGCTGGGATT	Subtilase1
Sh14Contig_29461	ACCAGGTTCCCTTTCTCCTG	CATGCCTTTGCGGATTCTAT	Subtilase3
Sh14Contig_34949	AGGAGACCAGCTGGCTATGA	GCCGTCTCGTATTTCTCGTC	Subtilase2
Sh_Cyclophylin_1	TCGCCGACGAGAACTTTGTGAAGA	TCTTCGCGGTGCAGATGAAGAACT	cyclophylin (for rice interaction internal control)
Sh_Cyclophylin_2	GTCGTGATGGAGCTTTTCGC	CCTTGTAGTGGAGGGGCTTG	cyclophylin (for nonhost interaction internal control)

Data S2

Table of contents:

A. *S. asiatica* genome sequencing, assembly, and repeat masking

A.1 Plant materials and sequencing

In the 1950s, *S. asiatica* was accidentally introduced into the US and its eradication program cost about $US250 million. We used the seeds of the *S. asiatica* US strain originally obtained from the field collections made in 1992 at the USDA Methods Development Center (Whiteville, N.C.). The Illumina pair-end (PE) libraries and the mate-pair (MP) libraries (3 kb and 10 kb) were prepared and sequenced. A bacterial artificial chromosome (BAC) library with an average length of 120 kbp was prepared by Amplicon Express ltd (Washington, USA). Both ends of total 27,648 BAC clones corresponding 4.6x physical coverage were sequenced by a Sanger sequencer (ABI 3730xl; in the Kazusa DNA Research Institute, Kisarazu, Japan) and 50,513 clean (QV 20<) sequence reads were obtained with average length 549 bp. A total of 216.4 Gb (366.8 X) of *Striga* genome sequences was generated using whole genome shotgun (WGS) sequencing by Illumina HiSeq2000 and BAC-end sequencing by a Sanger platform (Table A.1).

Table A.1. Generated genome sequences of *S. asiatica*.

Sequencing data	Insert size	Total length (Gb)	Sequencing depth (X)	Physical coverage (X)	Average read length (bp)
Illumina reads	400 bp	126.7	214.7	171.1	251
	3 kbp	41.5	70.3	1,044.1	101
	9-10 kbp	64.0	108.5	5,371.3	101
Sanger BAC-end	120 kbp	0.03	0.1	4.6	549
Total		232.2	392.9		

A.2. Raw data processing

Assembly of a large genome is highly complicated and sophisticated due to extensive error correction and filtering of contaminated sequences demanding enormous computational resources[S1]. To remove nonessential sequences while retaining a proper amount of data for genome assembly, we performed data pre-processing analyses before assembly. Firstly, identical prokaryotic reads in the raw data (98% identity and 50% coverage) were detected and eliminated using the CLC NGS assembly cell (CLCBio,

Denmark) with publicly available bacterial genomes as reference. Secondly, duplicated reads by PCR amplification during data generation were removed by using the CLC NGS assembly cell. Low-quality regions were also removed using strict parameters (cut-off quality value as 25 and 70% coverage). Lastly, an error correction process was performed using Jellyfish[S2] and Quake[S3]. The K-mer distribution analysis indicates various information such as low frequencies, sequencing depth, degree of heterozygosity, and genome size[S4]. To examine low frequencies as error candidates, Illumina PE reads were used for 17-mer K-mer analysis using Jellyfish (Figure A.1). Compared to the K-mer charts from *S. hermonthica* and *S. gesnerioides*, the *S. asiatica* genome showed less heterozygosity and smaller estimated genome size (Figures A.2B, C, Table A.2). The low frequency reads were trimmed in PE, MP and BAC data using Quake. After filtering, a total of 84.6 Gb (143 X) *Striga* genome sequences were used for *de novo* assembly (Table A.3).

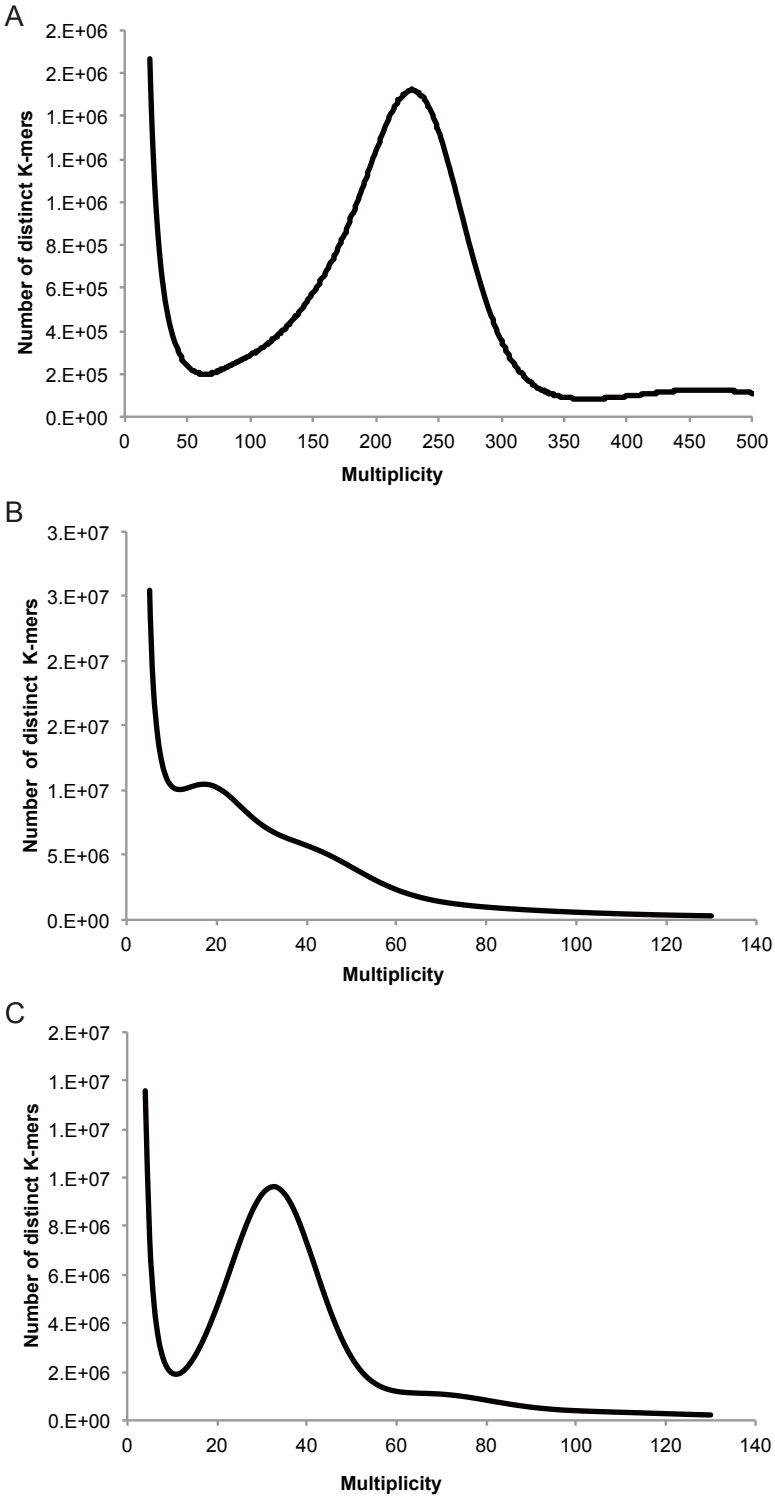


Figure A.1. The 17-mer distribution of *S. asiatica*, *S. hermonthica* and *S. gesnerioides* genomes. The frequencies of unique 17-mers were counted by the Jellyfish program. a. *S. asiatica*, b. *S. hermonthica*, c. *S. gesnerioides*. The 17 mers with low frequencies (less than 20 in *S. asiatica* (A) and less than 4 in *S. hermonthica* and *S. gesnerioides* (B, C)) were removed as they were considered as error sequences.

Table A.2. Generated genome sequence of *S. hermonthica* and *S. gesnerioides*.

Species	Sequencing data	Insert size	Read number (M reads)	Total read length (Gbp)	Average read length (bp)
S. hermonthica	Illumina HiSeq2000 reads	150 bp	183	18.4	101
		150 bp	180	18.2	101
S. gesnerioides		180 bp	321.6	32.5	101
		500 bp	191.3	19.3	101

Table A.3. Statistics of pre-processed *S. asiatica* genome sequences.

Insert Size	#Library	RawData[a]	Step1[b]	Step2[c]	Step3[d]	Step4[e]	Filtered Data[f]
400 bp	2	126.7 Gb	126.2 Gb	124.2 Gb	72.7 Gb	71.2 Gb	71.2 Gb
3 kbp	1	41.5 Gb	41.2 Gb	17.9 Gb	15.4 Gb	8.56 Gb	8.56 Gb
9-10 kbp	2	64.0 Gb	63.4 Gb	8.3 Gb	6.4 Gb	4.8 Gb	4.8 Gb
BAC-end	1	0.03 Gb	0.03 Gb	0.03 Gb	0.03 Gb	0.02 Gb	0.02 Gb
Total	7	232.2 Gb	230.1 Gb	150.4 Gb	94.5 Gb	84.6 Gb	84.6Gb (143 X)

[a] Original raw data.
[b] Raw data, which removed bacterial genome.
[c] For each generation, amount of data after removing duplicated reads.
[d] For step2, amount of data after trimming low quality.
[e] For step3, remained data after error correction using quake.
[f] Final raw data that used for genome assembly.

A.3 Genome assembly, scaffolding and gap-closing

Genome assembly is one of the major challenges in the plant community. Especially, the construction of a high quality genome is very difficult on account of repeat sequences, heterozygosity and ploidy in plant genomes[S5]. To overcome those problems and to de novo assemble a solid genome, we developed our in-house pipeline (Figure A.2). First, initial contigs were meticulously constructed to ensure a high-quality genome. To generate longer initial contigs, overlapped forward and backward reads of 400 bp PE library were merged to single reads by FLASH[S6]. These longer single reads and the remaining paired reads of 400 bp library contributed to an assembly of high quality initial contigs.

Owing to the optimisation of parameters such as a K-mer, various versions of initial contigs were generated by using different K-mer values and the best version was selected for scaffolding. Estimation of the actual insert length is another critical process because the insert distance of both the sides is an important factor for accurate scaffolding. Insert length calculations of PE, MP, and BAC-end libraries were fulfilled through reference-guided assembly for initial contigs and scaffolds (Table A.4). The insert length of the PE library was 459 bp and the insert distance of MP and BAC-end libraries were reasonably decided accordingly. Scaffolding processing was performed by Platanus[S7] and SSPACE[S8]. We first determined the K-mer value for scaffolding of the PE to BAC-end library (Table A.5), and found that the serial scaffolding processes generated longer scaffolds using optimised K-mer value. To extend the length of scaffolds, we used SSPACE, which fulfilled serial scaffolding with stringent parameters using MP and BAC sequences for the scaffolds generated by Platanus. Lastly, the remaining gaps were filled by Gapcloser (http://soap.genomics.org.cn/down/GapCloser_release_2011.tar.gz) and Platanus[S9] using reads of PE and MP libraries. As a consequence, a total of 471.6 Mb (80 % of 590 Mb) including 24.7 Mbp of gap sequences was assembled and the N50 values of the scaffolds and contigs were found to be 1.3 Mb and 16.2 kbp, respectively (Table A.6). In particular, 90% of the assembled genome was covered by 406 scaffolds.

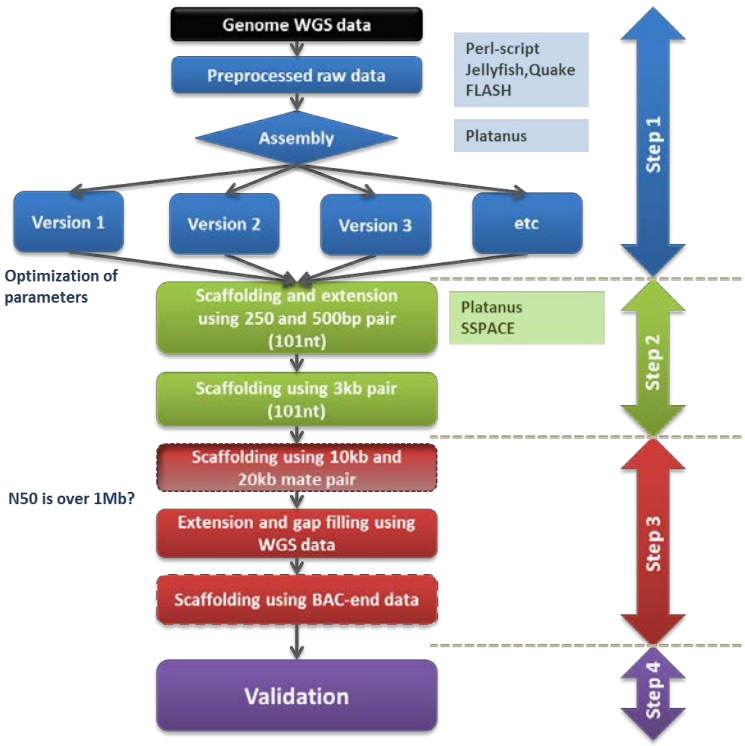


Figure A.2. Flow chart of *S. asiatica* genome assembly pipeline.

Table A.4. Estimation of insert length of PE, MP, and BAC-end libraries.

Data	EstimatedSize (bp)	Mapped as Paired	Range (99.9 %)	Range (99.0 %)	Range (95 %)
400 bp	459.3	88.1%	126-611	224-599	368-570
3 kbp	3,125	34.7%	637-4200	2436-4021	2598-3767
9 kbp	9,031	6.7%	26-13501	379-12730	5253-12606
10 kbp	10,030	15.34%	546-14473	1005-12916	7731-12715
BAC-end	100,778	23.35%	292-149592	3124-147416	16180-140318

Table A.5. Statistics of *S. asiatica* genome assembly.

Step	Software	N50 (bp)	Total Number	Total Length (Mb)
Initial contig	Platanus	2,281	692,284	557.6
Scaffold	Platanus	1,183,906	20,051	468.1
Final scaffold	SSPACE /Gapcloser /platanus	1,308,318	13,847	471.6
Final contig		16,191	65,272	446.9

Table A.6. Detailed statistics of *S. asiatica* genome assembly.

	Scaffold	Contig
N10	3,881,260 bp (11th)	52,892 bp (613th)
N20	2,669,643 bp (27th)	36,107 bp (1,650th)
N30	2,266,381 bp (46th)	26,813 bp (3,100th)
N40	1,838,224 bp (69th)	20,786 bp (4,998th)
N50	1,308,318 bp (99th)	16,191 bp (7,436th)
N60	1,014,510 bp (141th)	12,417 bp (10,595th)
N70	741,222 bp (196th)	9,260 bp (14,766th)
N80	498,068 bp (272th)	6,345 bp (20,574th)
N90	222,000 bp (406th)	3,373 bp (30,028th)
Max / Min	5,868,886 bp / 500 bp	196,100 bp / 201 bp
Total length / number	471.6 Mb / 13,847 ea	446.8 Mb / 65,237 ea

A.4 Assessment of *S. asiatica* genome assembly

Genome assembly validation is an essential process to assess genome assembly quality. To compare the BAC clone sequences with the *de novo* assembly, we sequenced paired-end libraries constructed from 9 BAC clones with Illumina HiSeq2000 sequencer at approximately 2,000 coverage (2.19 Gbp). The obtained short reads were assembled with Edena assembler[S10] and the gap regions were filled by Sanger sequencing. To confirm the sequence alignment between BAC contigs and scaffolds, we performed BLAST analysis for BAC contigs and scaffolds and BAC contigs were matched to the scaffolds based on a 98% identity (Table A.7). Although the *S. asiatica* genome assembly was identified by most of BAC contigs, some unclear or unconfirmed regions for BAC contigs were also present. To analyse the BLAST result in detail, we visualised each sequence alignment between the scaffolds and BAC contigs (Figure A.3). The results showed that the detected unmatched regions were caused by gap regions, resulting in exaggerated and ambiguous scaffolding. Consequently, despite several unclear results, our assembled *S. asiatica* genome was evaluated as a high quality genome by BLAST and visualisation using BAC contigs.

Table A.7. Summary for assessment of *S. asiatica* genome assembly using BAC contigs, assembled transcripts and filtered raw sequences.

Data set	Number (Length) of data	Average length of data	Analysis method	Identity (%)	Coverage			
					Matched	>70 %	>80 %	>90 %
BAC contigs	209 (0.87 Mb)	4,168.5 bp	Calculating query (data) coverage using BLASTN	95	202 (97%)	201 (96%)	198 (95%)	192 (92%)
				98	201 (96%)	198 (95%)	194 (93%)	187 (89%)
				99	194 (93%)	192 (92%)	187 (89%)	180 (86%)
Assembled transcripts	43,709 (40.13Mb)	918.09 bp		95	43,056 (99%)	42,308 (97%)	41,864 (96%)	40,576 (93%)
				98	42,736 (98%)	41,793 (96%)	41,251 (94%)	39,722 (91%)
				99	42,122(96%)	40,599(93%)	39,761(91%)	37,595(86%)
Extended single reads	84 M (31 Gb)	374.3 bp	Calculating mapped reads as paired	98	80 M (96.1%)	-	-	-
Filtered PE reads	159 M (84.6 Gb)	204.3 bp		98	139 M (87.9%)	-	-	-

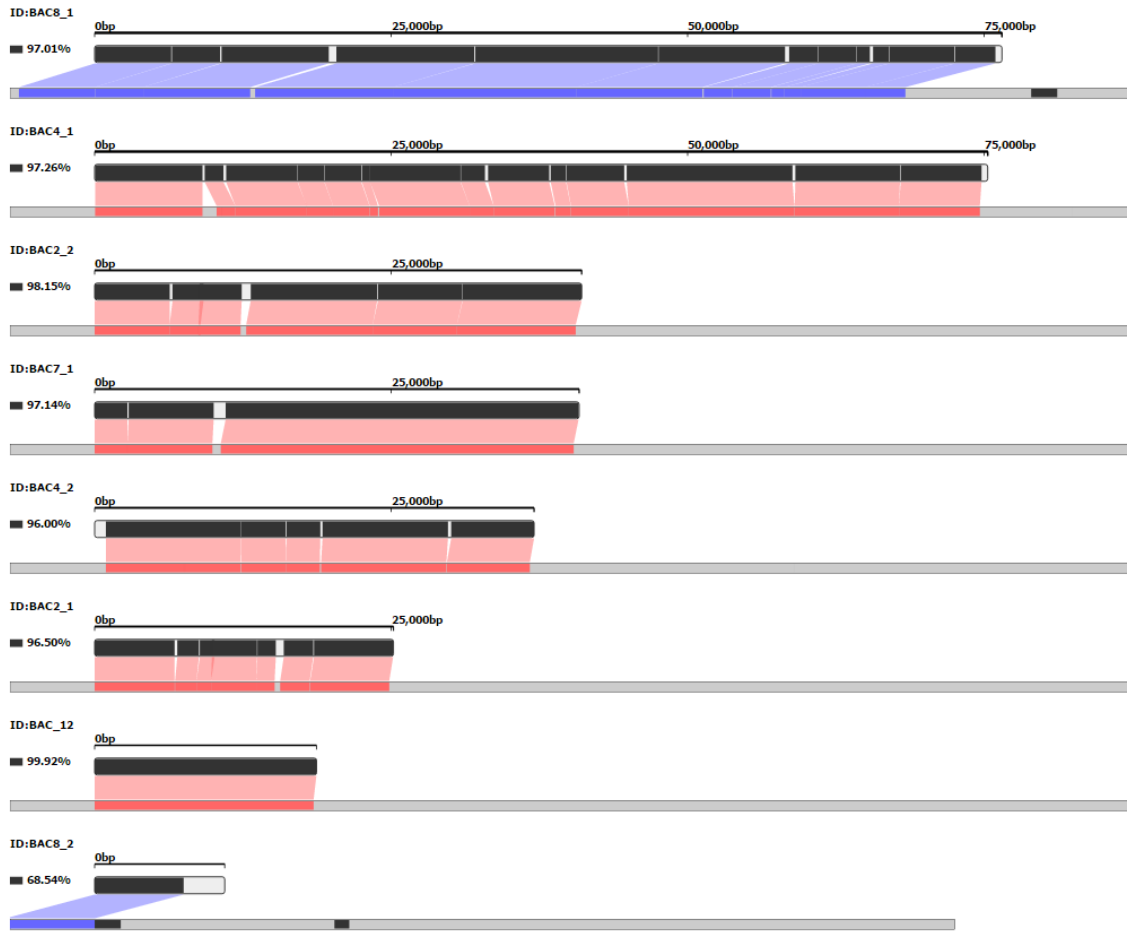


Figure A.3. Representative validation result of *S. asiatica* genome assembly against 8 longest BAC contigs.
The upper bars indicate BAC contigs and the lower bars mean scaffolds. In the upper bars, black and white represent matched and unmatched regions to the scaffold, respectively. In lower bars, red and blue indicate matched regions to forward and backward strand. Black represents gap sequences and grey represents unmatched regions of scaffold.

Table A.8. *S. asiatica* RNA sequencing reads

Sample	Insert size	Number of library	Total sequence read number	Total length	Read length
Leaf	180 bp	1	135 M	13.6 Gbp	101 bp
Root	180 bp	1	135 M	13.6 Gbp	101 bp
Shoot	180 bp	1	99 M	5.0 Gbp	101 bp
7 d Haustoria	180 bp	4	168 M	16.8 Gbp	101 bp

We performed additional validation of the assembled genome using *de novo* assembled transcriptome and the filtered raw sequences shown in Table A.8. The RNAs were extracted from *S. asiatica* shoots and roots that were axenically grown on MS media for 1 month and Illumina PE libraries were constructed using TruSeq RNA sample prep kit (Illumina) for an insert size of 180 bp. Total two libraries were sequenced by Illumina HiSeq2000 sequencer for 101 cycles per run (Table A.8). The RNA sequences were *de novo* assembled using CLC Assembly Cell (CLC bio, Aarhus, Denmark). This resulted in 43,709 contigs with average length of 918 bp. Through BLASTN analyses, 38,557 (88.2%) contigs were found in the assembled genome with cut-off values over 98% identity and 80% coverage (Table A.7). Furthermore, we confirmed that the 91.9% and 85.4% of filtered PE and total reads were mapped as paired in single scaffold(s) by reference-guided alignment using CLC Assembly Cell (CLC bio, Aarhus, Denmark).

A.5 Annotation of transposable elements (TEs)

Most of the DNA of large eukaryotic genomes is composed of repetitious sequences, primarily transposable elements (TEs). In large plant genomes, TEs can comprise 80% or more of the total genomic DNA, most of that derived from retrotransposons (Class I TEs). Repeat analysis was performed by RepeatModeler and RepeatMasker (http://www.repeatmasker.org) in the assembled *S. asiatica* genome. First, a repetitive element library was constructed by combining the results from Repet-pipeline (https://urgi.versailles.inra.fr/Tools/REPET), LTRharvest/LTRdigest from genometools (http://genometools.org/), our own pipeline, and a library of LTR retrotransposons. Then, RepeatMasker was used to mask TEs in the *S. asiatica* genome through classified repeat libraries.

The total repetitive fraction comprises 48.8% of the genome assembly, with all TEs forming 44.1% of the assembly and 90.3% of the repeats (Table A.9). Together, the retrotransposon sequences (83.7% of the repetitive DNA) constitute 40.9% of the genome assembly very similar to the 45.2% TEs and 39% retrotransposons of the *Phaseolus vulgaris* 473 Mb assembly[S11,S12], almost identical in size to the *S. asiatica* assembly here. By comparison, the retrotransposons form 21.4% in *B. distachyon*, 26% in rice, and over 82% in barley. The DNA (Class II) transposons together form only 3.2% of the *S. asiatica* genome. Hence, *Striga* fits well into the overall picture for vascular plants, in which retrotransposons abundance explains much of genome size variation[S13].

The DNA transposons (DXX, codes according to Wicker et al.[S14], form only 3.2% of the genome assembly, with the MULE-MuDR (DTM) and hAT (DTA) families of terminal-inverted-repeat (DTX) elements being the mar ones identified. The *Helitron* (DHH) elements, which replicate by rolling-circle amplification[S15], are the second most abundant group of Class II retrotransposons behind the MULE-MuDRs, forming 3 Mbp from 3,330 copies.

The LTR retrotransposons[S14] form the overwhelming majority (85.5% by coverage, 84.6% by number) of all TEs, with LINE and SINE retrotransposons only as minor players (respectively 6.1% and 0.1% by coverage). Of the LINEs that can be further characterized, L1 comprises 28% of the LINEs and is the dominant superfamily of this order in *S. asiatica*, as in the case in many plants[16], although 71% of the LINEs cannot be identified to the superfamily level. Among the LTR retrotransposons, superfamilies *Gypsy* and *Copia* respectively comprise 8.4% and 5.2% of the genome assembly, but the non-autonomous LARD[S17] and TRIM{Formatting Citation} retrotransposons appear relatively abundant in *Striga*, occupying respectively 6.3% and 1.3% of the genome space.

Although members of both the *Copia* and *Gypsy* superfamilies display an average age of 1.1 million years (MY) and few elements are older than 3 MY, their age profiles are very different (Figure A.4). *Gypsy* elements of 0.5 to 1.0 MY are relatively more common, with only eight elements (1.8% of all) aged 0.025 MY or younger present. By contrast, *S. asiatica* displays an abundance (30, 6.9% of all) of *Copia* elements younger than 0.025 MY and a broad but fairly even distribution of older elements. The data thus suggest a very recent burst of amplification among *Copia* elements and one at least 0.5 MY ago in the *Gypsy* superfamily. A very recent *Copia* burst and an older (~2 MY) *Gypsy* one were likewise seen in the model monocot *B. distachyon*[S19], although in that species a broad decline in abundance over time was seen for *Copia*. As a result of the insertion of the 49 retrotransposons younger than 0.025 MY, 342 Kbp has been added to the genome (0.06% of its total size).

Retrotransposons replicate by a life cycle in which a reverse-transcribed RNA integrates into the chromosome, thereby increasing the genome size[S20]. Two mechanisms counter growth in the genome size through retrotransposon integration. One is the homologous intra-strand LTR:LTR recombination, which removes the DNA intervening between the LTRs and leaves behind a solo LTR, and the other is a piecemeal loss through recurrent small deletions[S13,S21]. The 2180 full-length *Gypsy* and *Copia* of the *S. asiatica* genome comprise only 11% of the total LTR retrotransposon coverage, mirroring the extent of the element loss. *Gypsy* elements comprise 1.6-fold more of the genome than do *Copia* ones, but the ratio drops to 1.07 for full-length elements. This indicates that *Gypsy* elements have been differentially lost, consistent with their higher overall age and the more recent amplification of *Copia* elements. Therefore, LTR retrotransposons removal by recombination has played a major role in maintaining the compactness of the *Striga* genome. For the following gene prediction, the genome sequences that were masked by using only classified repeat sequences (except unknown TEs) were used to avoid the unexpected masking of some essential gene families.

Table A.9. Annotated repeat abundances in *S. asiatica*. The major represented classes, superfamilies, and subgroups of transposable elements as determined by automated annotation and classification, as well as other major repeat types, are presented.

	% of genome assembly	Sum (Mbp)	% all TEs (bp)	Number	% all TEs (number)	Number full-length	% Full-length	Average length (bp)
All repeats	**48.83**	**230.101**						
Mobile Elements	**44.10**	**207.809**	**100.00**	**250 653**	**100.00**	**14 206**	**5.67**	n/a
Class I: Retroelement (RXX)	**40.87**	**192.598**	**92.68**	**229 146**	**91.42**	**10 869**	**4.74**	n/a
LTR Retrotransposon (RLX)	37.70	177.656	85.49	212 009	84.58	6 773	3.19	n/a
Gypsy (RLG)	8.41	39.621	19.07	31 075	12.40	1144	3.68	1350.7
Copia (RLC)	5.20	24.523	11.80	24 998	9.97	1 036	4.14	1018.8
LARDs (RLX)	6.31	29.730	14.31	45 608	18.20	1659	3.64	717.2
TRIMs (RLX)	1.28	6.041	2.91	8 512	3.40	704	8.27	723.9
unclassified LTR (RLX)	16.48	77.645	37.36	101 540	40.51	2 196	2.16	811.9
non-LTR Retrotransposon (RXX)	2.76	12.992	6.25	10 874	4.34	1304	11.99	n/a
LINE (RIX)	2.70	12.705	6.11	9 978	3.98	852	8.54	1347.9
L1 (RIL)	0.75	3.551	1.71	2 255	0.90	215	9.53	1623.5
RTE (RIT)	0.03	0.128	0.06	265	0.11	53	20.00	467.2
Unknown (RIX)	1.91	9.024	4.34	7 449	2.97	584	7.84	1297.1
SINE (RSX)	0.06	0.286	0.14	896	0.36	452	50.45	324.0
Class II: DNA Transposon (DXX)	**3.23**	**15.211**	**7.32**	**21 507**	**8.58**	**3 337**	**15.52**	**737.4**
DNA Transposon Superfamily (DTX)	2.48	11.678	5.62	17 609	7.03	3 124	17.74	688.4
MULE-MuDR (DTM)	0.82	3.862	1.86	3 840	1.53	661	17.21	1005.7
hAT (DTA)	0.41	1.920	0.92	5 303	2.12	1321	24.91	362.1
PIF-Harbinger (DTH)	0.19	0.873	0.42	1 075	0.43	202	18.79	792.1
CACTA (DTC)	0.13	0.589	0.28	1 327	0.53	288	21.70	464.5
Unclassified (DTX)	0.89	4.203	2.02	5 301	2.11	369	6.96	848.2
Maverick (DMM)	0.01	0.040	0.02	74	0.03	16	21.62	623.26
MITE (DXX)	0.41	1.950	0.94	6 263	2.50	2 792	44.58	325.9
Helitron (DHH)	0.64	3.012	1.45	3 330	1.33	141	4.23	954.77
unclassified DNA transposon (DXX)	0.10	0.480	0.23	494	0.20	56	11.34	1034.9
Class I/Class II ratio		**12.66**		**10.65**			**0.31**	
***Gypsy/Copia* ratio**		**1.62**		**1.24**		**1.10**	**0.89**	
Other	**4.73**	**22.292**	n/a	**188 850**	n/a			

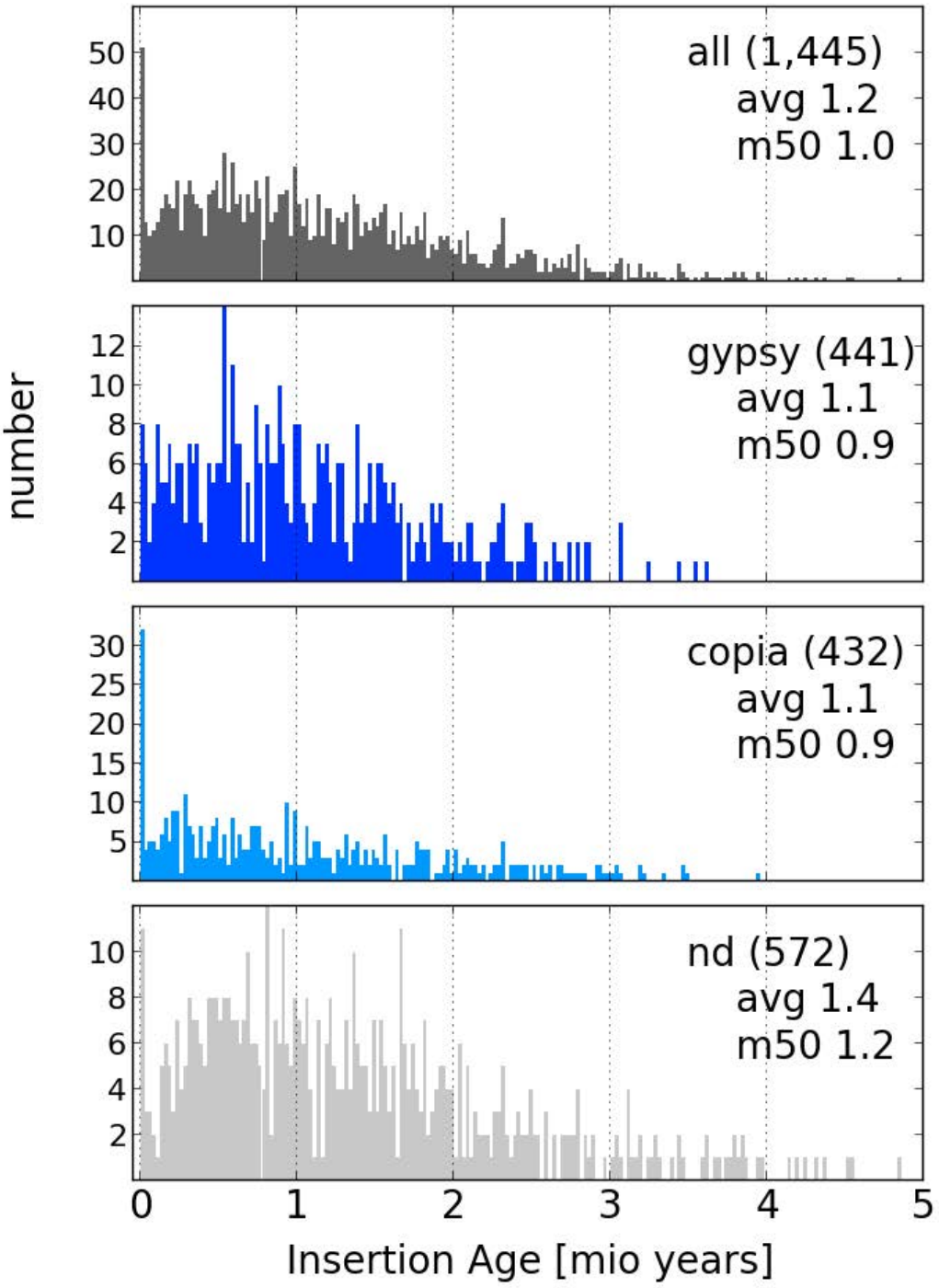


Figure A.4. Retroelement ages in the *S. asiatica* genome.
The age distribution and abundance of intact superfamily *Gypsy* (blue bars) and *Copia* (turquoise), all (grey), and unclassified (nd) LTR retrotransposons grouped in age classes of 0.025 MY.

B. Gene annotation

B.1 Assembly cleaning

To exclude any extraneous DNA sequences in the *S. asiatica* nuclear genome assembly, we mapped reads from Illumina PE libraries back onto the assembly and computed the read depth of all scaffolds and contigs using CLC Assembly Cell. Additionally, the taxonomic and source attribution of 100 best-matching sequences in the NCBI nt database to the *S. asiatica* scaffolds and contigs were determined using Megablast (e-value < 1e-10). In a plant genome assembly, high read depth contigs mainly belong to chloroplast genome (cpDNA), mitochondrial genome (mtDNA), and nuclear repeat sequences, and lower read depth contigs belong to the nuclear genome. We removed from the assembly scaffolds and contigs that had all their best-matching sequences in the nt database attributed to plant organelles and were also of high read depth (> 100x). Other likely plant cpDNA and mtDNA sequence in the assembly that did not meet these criteria were not removed from the assembly because it has been shown that chloroplast and mitochondrial DNA can be transferred into nuclear chromosomes of diverse eukaryotes including plants[22]. The remaining scaffolds and contigs that had their best-matching sequences in the nt database attributed to non-embryophytes were set aside as likely contaminants. In total, 200 out 13,847 assembled sequences were determined to be contaminants and excluded from the genome assembly.

B.2 Annotation-specific repeat masking library

A custom annotation-specific repeat library (database) was created for masking the genome assembly to enable high-quality gene prediction and genome structural analysis. Novel genomes often have new classes of repeats that are not present in Repbase. Therefore, generic genome masking using Repbase[S23,S24] in conjunction with RepeatMasker (http://www.repeatmasker.org) prior to gene prediction and whole genome comparative alignment is not sufficient. It is essential to identify, annotate, and mask repeats including interspersed repeats, low-complexity regions, and transposable elements to avoid prediction of spurious gene models and confounding alignments by repeat-mediated artifacts[S25–S27]. We followed the protocol described by Campbell et al., 2013[S26] to create a *S. asiatica*-specific repeat library suitable for repeat masking prior to protein-coding gene annotation. Briefly, the genome assembly was first searched with structural approaches to collect consensus miniature inverted-repeat transposable elements (MITEs) and long terminal repeat retrotransposons (LTRs) using MITE-Hunter[S28] and LTRharvest/ LTRdigest[S29,S30] respectively. LTRs were filtered to remove false positives and elements with nested insertions. The genome was then masked using collected LTRs and MITEs and additional *de novo* repetitive sequences predicted by RepeatModeler (http://www.repeatmasker.org/RepeatModeler) from the unmasked regions of the

genome. All collected repeat sequences were searched against plant proteins from UniRef[S31] where elements with significant hits to genes were excluded from the repeat masking library.

B.3 RNA sequencing and assembly

Total RNAs were extracted from tissue samples (leaf, shoot, root, and haustoria) of *S. asiatica* according to the protocol described by Yoshida et al., 2010[S32]. RNA-Seq libraries were prepared using TruSeq RNA Sample Prep Kit (Illumina) for an insert size of 180 bp and sequenced using 101-bp paired-end sequencing on the Illumina HiSeq 2000 platform (Table A.8). Raw reads were trimmed to remove low-quality bases as well as embedded adaptor sequences and filtered to discard short read fragments using Trimmomatic v0.33[S33]. FastQC v0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to assess the overall sequence quality before and after trimming. Cleaned reads from each tissue sample were *de novo* assembled using Trinity[S34] with the default parameters. The resulting transcriptome assemblies were post-processed with PlantTribes AssemblyPostProcessor (https://github.com/dePamphilis/PlantTribes) to select contigs with potential coding regions to use as evidence for gene annotation.

B.4 Gene prediction, quality assessment, and functional assignment

Gene models were predicted with the MAKER pipeline (release 2.31.8)[S35] using tissue-specific RNA-Seq assemblies of *S. asiatica* described above and RNA-Seq assemblies of plant parasite developmental stages described in Westwood et al. (2012)[S36] for related species of *Orobanchaceae* obtained from the Parasitic Plant Genome Project[S37] as transcript evidence. Further cross-species protein homology evidence was supplied by proteomes derived from the annotations for *M. guttatus* v2.0 as represented in Phytozome v11[38] and a set of canonical plant (embryophytes) proteins from UniProt/SwissProt release 2017_04[39]. Repetitive and low complexity regions of the genome assembly were masked with RepeatMasker in MAKER using the custom annotation repeat library developed for *S. asiatica*. Genes were predicted using SNAP[S40] and Augustus[S41] which were trained for *S. asiatica* using MAKER with bootstrap training to iteratively improve the performance of *ab initio* gene predictors[S26,42]. Gene models from each round of MAKER run were used to seed the next round of SNAP and Augustus training. Selected gene models for Augustus training were required to meet the following criteria: (1) must have greater than 75% evidence support, (2) the length of both 5' and 3' UTRs must be at least 200 bp, (3) at least 80% of the splice sites must be confirmed with RNA-Seq alignment evidence, (4) at least 80% of the exons must match both RNA-Seq and protein alignment evidence, (5) the length of the protein sequence produced by the predicted mRNAs must be approximately 450 amino acids, the average plant protein size[S43], and (6) the training set genes must be divergent enough (< 50% identity) and not overlap each other.

Out of the 5,666 scaffolds and contigs (>= 1kb) used in the MAKER annotation, 1,553 were annotated with genes. The final *S. asiatica* post-processed gene annotation set consisted of the all gene models supported by annotation evidence, and gene models not supported by annotation evidence but encode Pfam domains. A total of 34, 575 coding protein were predicted, 91% of which have an annotation edit distance (AED) <0.5. AED is a quantitative measure of gene annotation quality based on annotation evidence with values ranging from 0 (perfect agreement) to 1 (no support)[44]. To evaluate the completeness of the *S. asiatica* genome, we examined the presence and completeness of 1,440 land plants (embryophytes) benchmarking universal single-copy orthologues (BUSCO)[S12] in *S. asiatica* compared to 25 other sequenced plant genomes in the orthogroup classification described below. Evaluation of *S. asiatica* BUSCO genes suggests 87.1% are complete genes, 4.0% are fragmented, and 8.9% are missing; these presence and completion rates are comparable to other taxa in the classification (Table B.1). Provisional functional descriptions for the gene models were assigned using the AHRD (https://github.com/groupschoof/AHRD), a pipeline for lexical analysis and selection of the best functional descriptor for gene products following BLASTP searches against UniProt/SwissProt, UniProt/TrEMBL, and TAIR10[S45] databases. Additionally, gene models were also annotated with protein family domains as detected by InterProScan[S46], and identified domains were directly translated into gene ontology terms.

Table B.1. The presence and completeness of universally conserved single copy land plants genes in *Striga* (BUSCO) genome compared to 25 other annotated representative plant genomes.

Species	Complete	Fragmented	Missing
Arabidopsis thaliana	99.3	0.3	0.4
Carica papaya	71.9	13.8	14.3
Theobroma cacao	97.6	1.1	1.3
Eucalyptus grandis	92.3	2.8	4.9
Phaseolus vulgaris	96.4	1.0	2.6
Medicago truncatula	93.7	1.9	4.4
Prunus persica	98.9	0.8	0.3
Manihot esculenta	95.3	2.6	2.1
Populus trichocarpa	97.6	1.3	1.1
Vitis vinifera	90.0	4.1	5.9
Solanum lycopersicum	95.5	2.8	1.7
Utricularia gibba	79.8	5.1	15.1
Mimulus guttatus	94.4	1.7	3.9
Striga asiatica	87.1	4.0	8.9
Beta vulgaris	93.2	2.4	4.4
Nelumbo nucifera	75.2	10.3	14.5
Aquilegia coerulea	95.7	2.0	2.3
Oryza sativa	95.6	2.5	1.9
Sorghum bicolor	98.3	1.0	0.7
Musa acuminata	86.8	4.7	8.5
Elaeis guineensis	42.4	18.8	38.8
Spirodella polyrhiza	79.6	10.7	9.7
Amborella trichopoda	84.4	6.0	9.6
Pinus taeda	19.8	6.8	73.4
Selaginella moellendorffii	61.7	5.5	32.8
Physcomitrella patens	67.9	3.3	28.8

We obtained 34,577 protein coding gene predictions with similar intron-exon structures with o ther plant species (Table B.2 and Figure B.1).

Table B.2 Metrics of the *S. asiatica* gene models.

	Protein coding	Total CDS length (bp)	Avg CDS length (bp)	Avg exon length (bp)	Avg intron length (bp)
Striga asiatica	34,577	38,151,497	1,103	206	632
Mimulus gutattus[a]	28,140	33,563,049	1,193	240	390
Capsicum annumm[b]	34,914	35,254,530	1,009	286	541
Solanum lycopersicum[c]	34,771	35,972,459	1,057	179	533
Arabidopsis thaliana[d]	27,206	24,861,465	1,212	265	164
Oryza sativa[e]	28,236	78,281,992	1,081	312	414

[a]Representative CDS of *Mimulus guttatus* v2.0 (phytozome 10.0) were used.
[b]PAG (Pepper Genome Annotation) 1.5 were used.
[c]The ITAG pre-2.3 pre-release data were used.
[d]All protein-coding transcripts were included, with the exception of TEs and pseudogenes.
[e]All protein-coding transcripts (MSU Release 6.3) were included, with the exception of TEs, pseudogenes, organellar insertions, and small genes.

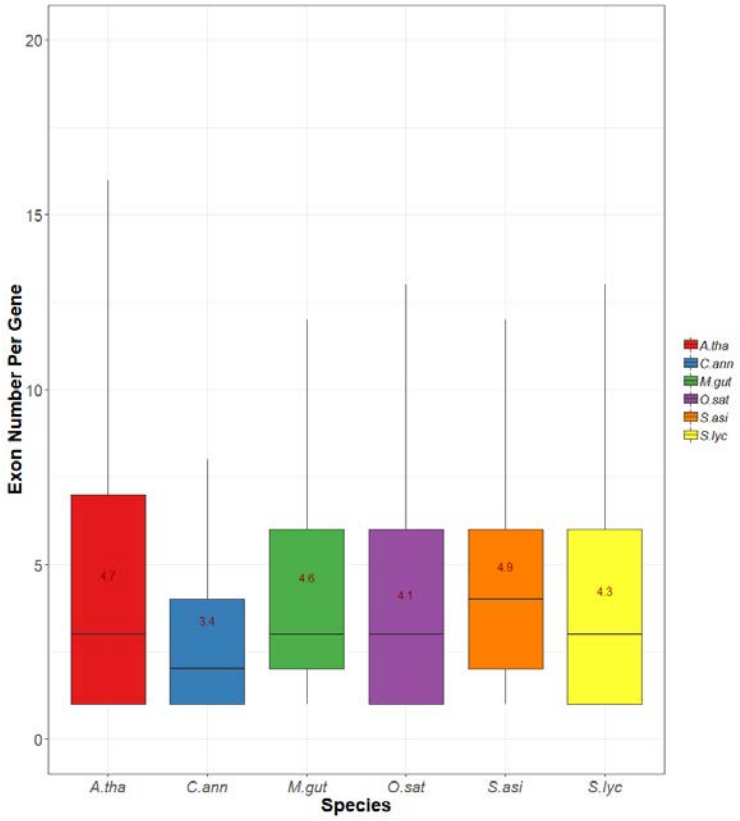


Figure B.1. Average exon numbers per gene. Average exon numbers of gene were calculated with representative CDS and were shown as box plot. (A. tha, *Arabidopsis thaliana*; C ann, *Capsicum annuum*; M. gut, *Mimulus guttatus*; O. sat, *Oryza sativa*; S. asi; *Striga asiatica*; S. lyc, *Solanum lycopersicum*)

C. Genome comparative analysis

C.1 Global gene family classification

Complete sets of protein-coding genes from 26 plant genomes (Data S1C) were classified into gene lineages (*i.e.*, orthogroups) using OrthoFinder version 1.1.8 algorithm[S47]. We selected taxa that represent all of the major land plant lineages for which genome sequence data were available, including ten *rosids* genomes (*Arabidopsis thaliana, Carica papaya, Theobroma cacao, Eucalyptus grandis, Manihot esculenta, Populus trichocarpa, Prunus persica, Phaseolus vulgaris, Medicago truncatula, Vitis vinifera*), one basal core-eudicot (*Beta vulgaris), four asterids (Striga asiatica, Mimulus guttatus, Utricularia gibba, Solanum lycopersicum*), two basal *eudicots* (*Aquilegia coerulea, Nelumbo nucifera*), five *monocots (Oryza sativa, Sorghum bicolor, Elaeis guineensis, Musa acuminata, Spirodella polyrhiza), one basal angiosperm (Amborella trichopoda*), one *gymnosperm* (*Pinus taeda*), one *lycophyte* (*Selaginella moellendorffii*), and one *moss* (*Physcomitrella patens*). A total of 18,110 orthogroups containing at least two genes were identified, 9,936 of which contain at least one gene from *Striga* (Data S3). Of the 34,575 annotated genes in *Striga*, 25,126 (72.7%) were classified in an orthogroup, and the remaining 9,449 (27.3%) genes are considered singletons, a clustering rate that is comparable to other taxa in the classification (Table C.1). Complete details for each orthogroup, including gene counts and functional annotations, are reported in Data S1S. We further performed a second iteration of MCL[S48] to connect distant, but potentially related orthogroups into larger hierarchical gene families (i.e., super-orthogroups) as described in Wall et. al., 2009[S49]. We used 10 MCL stringencies with inflation values 1.2 to 5.0 to cluster gene families into super-orthogroups to broadly represent all traditionally defined gene families characterized by functional domains. An average of 3,491 super-orthogroups were circumscribed for 10 MCL stringencies, of which at least 65% contain *Striga* genes (Table C.2).

Table C.1. Orthogroup classification summary for 663,272 validated annotated protein-coding genes in the 26 representative sequenced plant genomes.

Species	Number of Validated and Cleaned Annotated Genes	Number of orthogroups	Number of genes in orthogroups	Percentage of genes in orthogroups	Number of singleton genes	Percentage of singleton genes
Manihot esculenta	32,966	10,368	29,259	88.8	3,707	11.2
Populus trichocarpa	41,207	10,633	36,029	87.4	5,178	12.6
Phaseolus vulgaris	27,388	10,305	26,135	95.4	1,253	4.6
Medicago truncatula	50,869	10,922	39,619	77.9	11,250	22.1
Prunus persica	26,772	10,289	23,852	89.1	2,920	10.9
Arabidopsis thaliana	27,369	9,782	24,523	89.6	2,846	10.4
Carica papaya	27,528	10,221	21,978	79.8	5,550	20.2
Theobroma cacao	29,171	10,387	24,802	85.0	4,369	15.0
Eucalyptus grandis	36,288	9,958	31,195	86.0	5,093	14.0
Vitis vinifera	26,315	9,827	21,791	82.8	4,524	17.2
Striga asiatica	34,575	9,936	25,126	72.7	9,449	27.3
Mimulus guttatus	28,079	10,173	26,131	93.1	1,948	6.9
Utricularia gibba	27,206	9,102	21,220	78.0	5,986	22.0
Solanum lycopersicum	34,476	10,422	28,586	82.9	5,890	17.1
Beta vulgaris	27,911	10,026	21,794	78.1	6,117	21.9
Aquilegia coerulea	29,869	10,310	25,255	84.6	4,614	15.4
Nelumbo nucifera	26,643	9,795	23,775	89.2	2,868	10.8
Sorghum bicolor	34,118	11,115	27,239	79.8	6,879	20.2
Oryza sativa	41,411	11,216	29,734	71.8	11,677	28.2
Musa acuminata	36,514	9,707	29,770	81.5	6,744	18.5
Elaeis guineensis	29,667	10,054	26,638	89.8	3,029	10.2
Spirodella polyrhiza	19,572	9,371	17,372	88.8	2,200	11.2
Amborella trichopoda	26,802	10,003	19,588	73.1	7,214	26.9
Pinus taeda	27,596	5,768	23,770	86.1	3,826	13.9
Selaginella moellendorffii	22,251	7,907	17,057	76.7	5,194	23.3
Physcomitrella patens	32,853	8,348	21,037	64.0	11,816	36.0

Table C.2. Summary table of MCL Super-Orthogroup classification using minimum BLASTP E-value between all pairs of orthogroups.

MCL Stringency (inflation values)	Number of Super-Orthogroups	Number of Super-Orthogroups with *Striga* Genes	Percentage of Super-Orthogroups with *Striga* Genes
1.2	1,610	535	33.23
1.5	2,561	1,486	58.02
1.8	3,006	1,931	64.24
2.0	3,204	2,127	66.39
2.5	3,547	2,457	69.27
3.0	3,833	2,710	70.70
3.5	4,044	2,885	71.34
4.0	4,229	3,030	71.65
4.5	4,367	3,121	71.47
5.0	4,511	3,219	71.36
AVERAGE	**3,491**	**2,350**	**64.77**

C.2 Whole genome duplication history

We integrated the results of three complementary approaches to diagnose the history of genome duplication in *Striga* and the closely related nonparasitic plant *Mimulus*. Sequence alignments and phylogenetic analyses were described in STAR Methods.

C.2.1 Identification of *Striga* and *Mimulus* gene duplication events

Trees of each orthogroup were examined for gene duplications (terminal or shared with other taxa) and the detected duplications were scored using a scoring strategy [S50]. We scored orthogroups that showed at least one shared *Lamiales* (*Striga, Mimulus and Utricularia*) gene duplication with support values of at least 0.500 (50%) for the *Lamiales* duplication node and for one of the two internal *Lamiales* branches (arbitrarily defined as the “right” or “left” branch).

Striga and *Mimulus* genes were classified with respect to their likely duplication origins (Table C.3) with MCScanX[S51], an algorithm for detection of gene synteny and collinearity. Using default parameters, we classified genes within a single genome as singletons, dispersed duplicates, proximal duplicates, tandem duplicates, and WGD/segmental duplicates. WGD/segmental duplicates were inferred by the anchor genes in collinear blocks, with blocks defined by a minimum of five anchor genes. A total of 889 and 1521 orthogroups preserved duplicate copies of *Striga* (supported by 1,605 *Striga* anchor genes) and *Mimulus* (3,493 *Mimulus* anchor genes), respectively (Data S1T and S1U). We further identified 323 orthogroups (supported by 475 *Striga* and 608 *Mimulus* anchor genes) with *Lamiales* gene duplications that were supported with both *Striga* and *Mimulus* syntenic anchor

genes (Data S1V).

Table C.3. A summary of *Striga* and *Mimulus* genes classified into their likely duplication origins.

Species	Singleton	Dispersed	Proximal	Tandem	WGD/ Segmental
Striga asiatica	7,997	17,121	1,467	1,181	6,809
Mimulus guttatus	4,248	11,295	1,730	3,366	7,440

C.2.2 Duplicated gene divergence

We sought evidence for genome duplications in *Striga* by examining the divergence patterns of synonymous substitution rates (Ks) for *Lamiales* duplicate genes identified by the integrated syntenic and phylogenomic analysis. The best reciprocal paralogous matches for both *Striga* and *Mimulus* were identified using all-against-all BLASTP searches of their respective *Lamiales* duplicate genes. To determine the variation in synonymous substitution rates between the *Striga* and *Mimulus* lineages, we estimated a RAxML[S52] maximum likelihood species tree for the 26 representative plant genomes using a concatenated matrix of trimmed codon alignments for genes from 1,440 BUSCO single copy orthogroups (Figure 1). We determined that the length for the branch leading to *Striga* was longer than that leading to *Mimulus*, indicating that the lineage including the parasite *Striga* had experienced more rapid molecular evolution than its non-parasitic sister taxon *Mimulus*. A follow-up inspection of conserved single copy gene trees and spot inspection of phylograms from larger gene families including those with WGD synteny orthologs showed that *Striga* genes were in fact consistently on branches somewhat longer than their *Mimulus* orthologs. These results suggest that this was a *bona fide* description of a tendency for *Striga* branches to have evolved faster than those of *Mimulus*. Therefore, we expect this accelerated rate of evolution for *Striga* to be reflected in the estimated significant duplication components in which the shared event(s) with *Mimulus* would the shifted to higher K_s values. The EMMIX software[S53] was used to fit a mixture model of multivariate normal components to K_s data sets following the procedure described in Jiao et al., 2011[50]. The frequency of gene pairs with K_s divergences in each interval size of 0.05 within the range of 0 to 2.0 was plotted for *Striga* and *Mimulus* paralogs (Figure C.1). The K_s distributions identify two significant duplication components in *Striga* at mean $K_s \approx 0.47$ and mean $K_s \approx 1.22$, and one significant component for *Mimulus* at mean $K_s \approx 0.94$. Inspection of representative gene trees indicated that the peak of the older component in *Striga* K_s distribution corresponds to the peak of the single component in the *Mimulus* K_s distribution. The larger K_s value for *Striga* compared to *Mimulus* suggests a higher rate of synonymous substitutions in *Striga* as previously described. Taken together, these analyses suggest that the prominent younger peak in the *Striga* K_s distributions represents a duplication event

in the *Striga* lineage that occurred after the divergence of lineages leading to *Striga* and *Mimulus*, and the older peak represents a duplication event in the common ancestral genome of the three *Lamiales* taxa (*Striga, Mimulus,* and *Utricularia*).

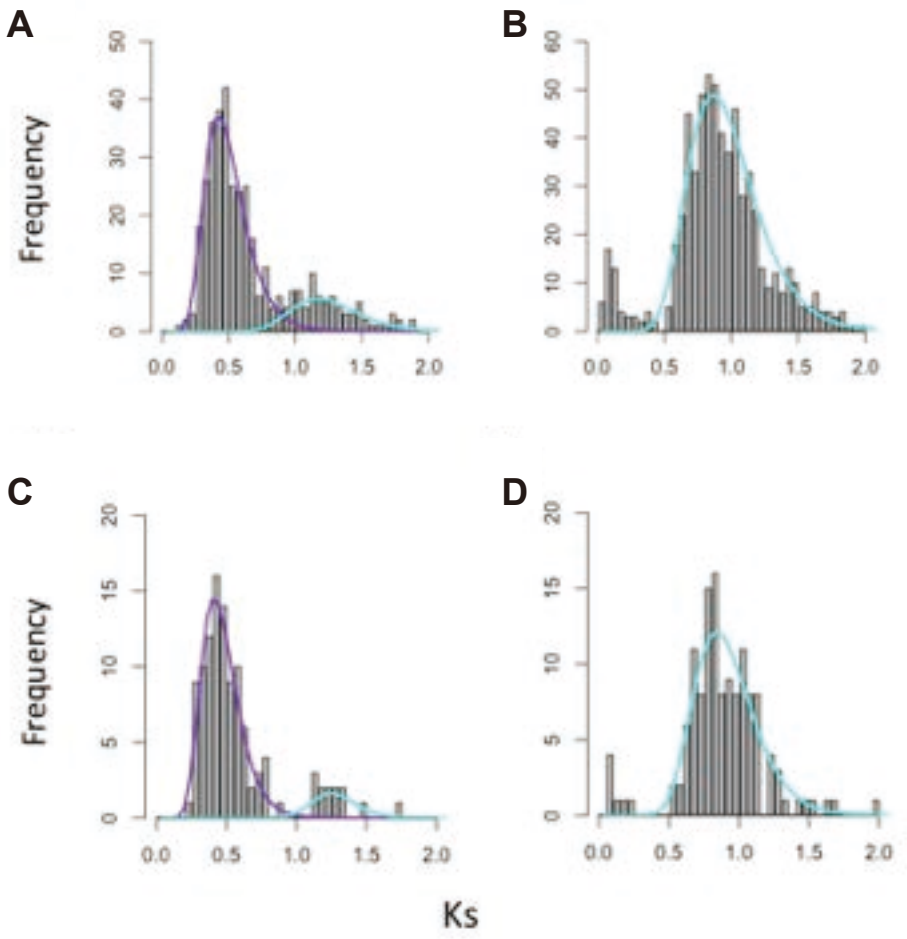


Figure C.1. K_s distributions of *Lamiales*-wide duplicate gene pairs in *Striga* and *Mimulus* identified by the integrated syntenic and phylogenomic analysis (Data S1T, S1U and S1V). Coloured lines superimposed on K_s distributions represent significant duplication components identified by likelihood mixture model. Plots show "colour/mean/proportion" where colour is the component (curve) colour, mean is the mean divergence of gene pairs assigned to the identified component, and proportion is fraction of duplicate pairs assigned to the identified component. **A.** Pairwise K_s distribution for 1,605 *Striga* genes from duplications within orthogroups, and on syntenic blocks anchored by *Striga* genes. Two statistically significant components: purple/0.47/0.80 and cyan/1.22/0.20. **B.** Pairwise K_s distributions for 3,493 *Mimulus* genes from duplications within orthogroups, and on syntenic blocks anchored by *Mimulus* genes. One statistically significant component: cyan/0.94/0.92. **C.** Pairwise K_s distribution for 475 *Striga* genes from duplications within orthogroups, and on syntenic blocks anchored by both *Striga* and *Mimulus* genes. Two statistically significant components: purple/0.45/0.88 and cyan/1.27/0.12. **D.** Pairwise K_s distributions for 608 *Mimulus* genes from duplications within

orthogroups, and on syntenic blocks anchored by both *Striga* and *Mimulus* genes. One statistically significant component: cyan/0.89/0.95. Negative exponential curves identified by maximum likelihood mixture model the in the *Mimulus* plots that represent the background distribution of paralogs due to normal gene births and deaths in a genome are not shown.

C.2.4 Genome Structure and Synteny

Structural syntenic analyses were performed with the SynMap tool of the CoGe comparative genomics platform[S54]. The genomes of *Mimulus* and *Vitis* were compared to the genome of *Striga* with the chaining algorithm DAGChainer[S55]. We specified a maximum distance of 20 genes between gene matches and required a minimum of five genes to seed a syntenic region. Scaffolds and contigs of *Striga* were ordered and oriented based on their syntenic path to both *Mimulus* and *Vitis*.

The self-self dot plot of *Striga* syntenic blocks (FigureC.2A) shows evidence (on the diagonal axis) of extensive collinear blocks, distributed throughout the genome, indicating at least one round of ancient polyploidy. However, there are numerous syntenic signals off the diagonal, which suggest a second, older polyploidy event. The overlaid color scheme that corresponds to the synonymous mutation (K_s) age distribution histogram (Figure C.2B) as calculated by CODEML identifies that the majority of genes comprising syntenic regions are from one age distribution (purple) and numerous others (off-diagonal) are from an older age distribution (cyan). This pattern is also evident in the cross-species dot plots of *Striga-Mimulus* (Figure C.3) and *Striga-Vitis* (Figure C.4) that show a relatively recent WGD (purple) superimposed on an older polyploidy event (cyan). Taken together, the structure and synteny results suggest that the *Striga* genome reflects two rounds of ancient polyploidy. The histogram of *Striga* K_s values derived from syntenic blocks shows a bimodal makeup in its K_s distribution with peaks around $\log_{10}$ transformed values of -0.3 ($K_s \approx 0.5$, younger peak) and 0.09 ($K_s \approx 1.2$, older peak) indicated in purple and cyan respectively (Figure C.2B). The purple peak that represents the larger population of duplicate pairs is evidence that they are derived from a younger evolutionary event than the smaller population represented by the cyan peak.

Previous studies have shown that the *Mimulus* lineage reflects only one WGD (that is most probably shared with *Utricularia gibba*) following their divergence from the *Vitis* lineage, which has not had any polyploidy event since the eudicot-wide paleohexaploidy event (also known as *gamma*)[S56,S57]. Therefore, there is a 1:2 mapping of orthologous syntenic regions between *Mimulus* and *Vitis*, as was reported by Ibarra-Laclette et al., 2013[S56]. The *Striga-Mimulus* and *Striga-Vitis* ortholog plots show many large purple syntenic regions superimposed on many smaller and older cyan syntenic regions highlighting two different age classes of syntenic blocks (Figures C.3 and C.4). The younger syntenic blocks are orthologous blocks, while older paralogous blocks were

detected as well. The duplication peaks of *Striga-Mimulus* and *Striga-Vitus* orthologs are around $\log_{10}$ transformed values of 0.04 ($K_s \approx 1.0$) and 0.3 ($K_s \approx 2.0$) respectively.

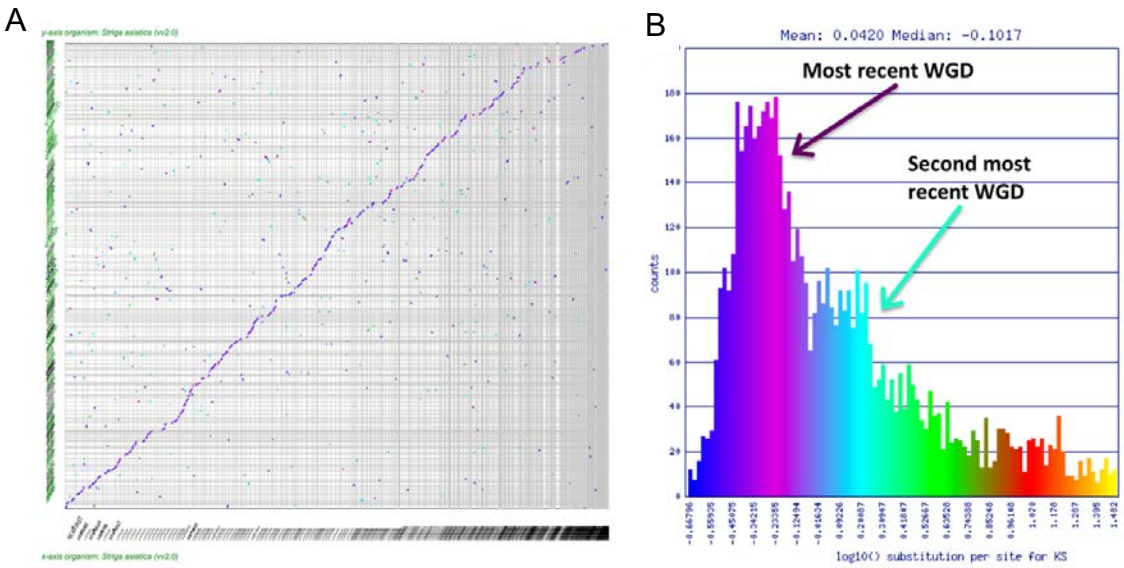


Figure C.2. Syntenic analysis of *Striga* against itself showing evidence of at least two WGD events. **A.** Self-self syntenic dot plot where contigs are ordered and oriented by syntenic path assembly. Syntenic gene pairs colored by their K_s values show two age distributions. Purple syntenic paralogs are younger than cyan. **B.** Histogram of $\log_{10}$ transformed K_s values of syntenic gene pairs identified in (A) shows a bimodal distribution with the younger syntenic gene pairs in purple and older ones in cyan. Results can be regenerated: https://genomevolution.org/r/11ncl

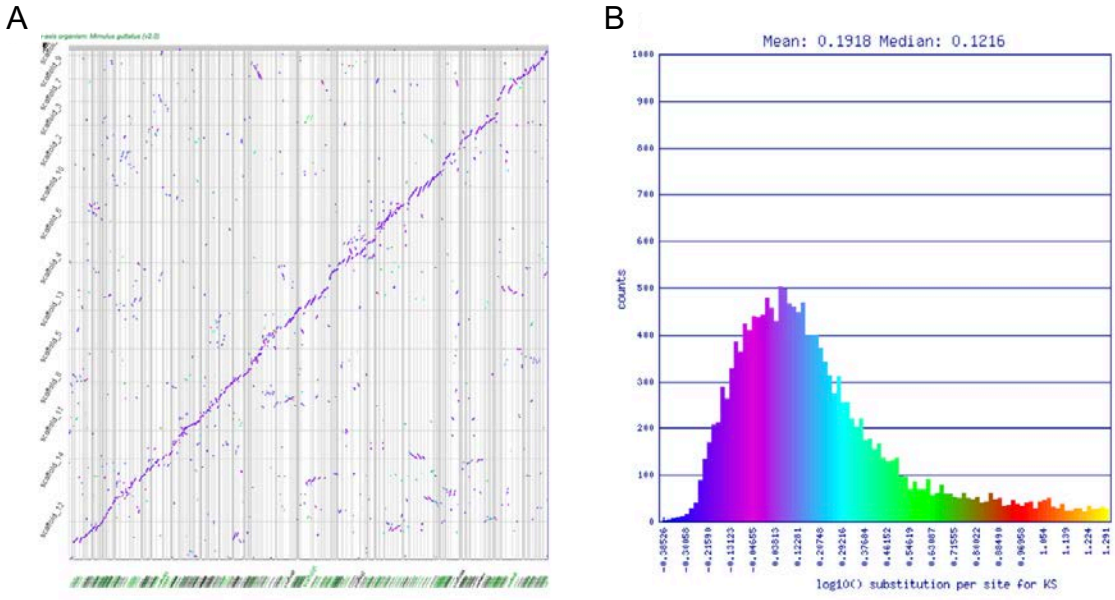


Figure C.3. Syntenic analysis of *Striga* and *Mimulus*. **A.** Syntenic dot plot of orthologous *Striga* (y-axis) versus *Mimulus* (x-axis) with *Striga* contigs ordered and oriented based on their syntenic path to

Mimulus. Syntenic gene pairs colored by their K_s values could reflect a mixture of two age distributions. Purple syntenic orthologs are younger than cyan. **B.** Histogram of $\log_{10}$ transformed K_s values of syntenic gene pairs identified in (A). Results can be regenerated: https://genomevolution.org/r/11nki

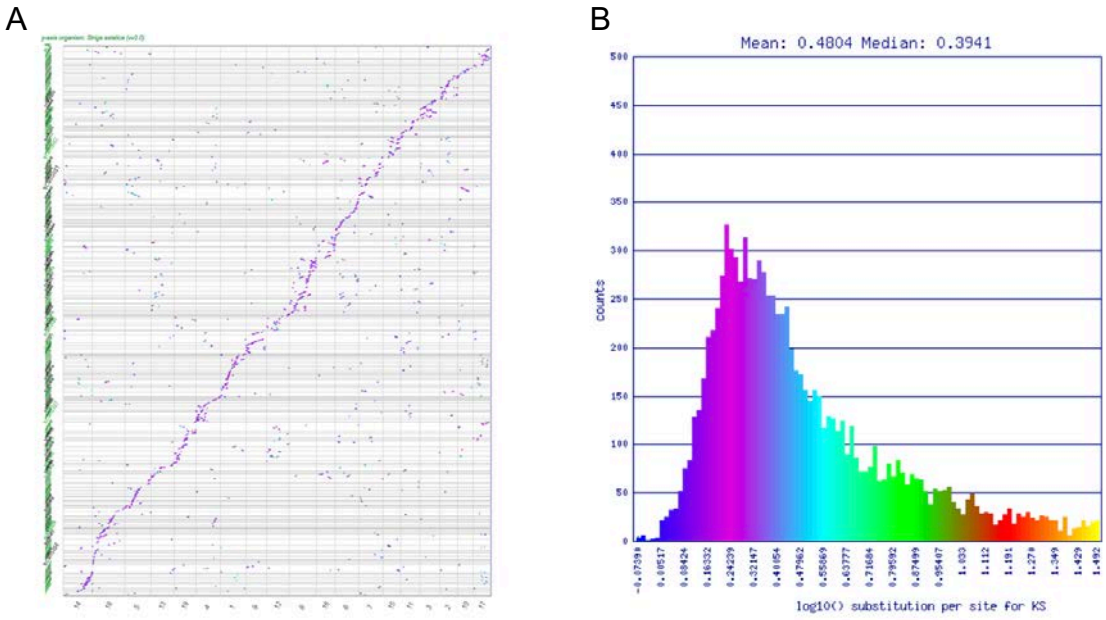


Figure C.4. Syntenic analysis *Striga* versus *Vitis*. **A.** Syntenic dot plot *Striga* (y-axis) versus *Vitis* (x-axis) with *Striga* contigs ordered and oriented based on their syntenic path to *Vitis*. Syntenic gene pairs colored by their K_s values could reflect a mixture of two age distributions. Purple syntenic orthologs are younger than cyan. **B.** Histogram of $\log_{10}$ transformed K_s values of syntenic gene pairs identified in (A). Results can be regenerated: https://genomevolution.org/r/11nl5

C.2.5 Microsynteny analysis

High-resolution analysis of microsyntenic regions was performed using CoGe's GEvo tool[S58], which permits comparison of multiple genomic regions. The whole genome syntenic ortholog dot plot (Figure C.3A) shows that most of the *Striga* genome is syntenic with at least one region of *Mimulus*. An example of one of several regions identified that showed 1x *Mimulus* to 2x *Striga* shows fractionated gene content, as expected following a polyploidy event (Figure C.5A)[S59]. An earlier WGD in the common ancestor of *Mimulus* and *Striga* would, therefore, create syntenic blocks comprised of 2x *Mimulus* regions and 4x *Striga* regions (Figure C.5B). A close-up view of these regions (Figure C.5C) shows evidence of 4 *Striga* and 2 *Mimulus* collinear anchor genes that are present on the duplication node of the gene family tree in Figure 1. We further identified a *Vitis* region from the ortholog collinear block that is syntenic to the shared *Striga* and *Mimulus* regions shown in

Figure C.5. The regenerated microsynteny plot (Figures C.6 and C.7) shows this *Vitis* region syntenic to the two *Mimulus* and four *Striga* regions as is expected following their divergence after the core eudicot-wide paleohexaploidy event. Taken as a whole, all three sets of analyses indicate *Striga*-specific WGD event and an earlier WGD event in the common ancestor of *Striga* and *Mimulus*.

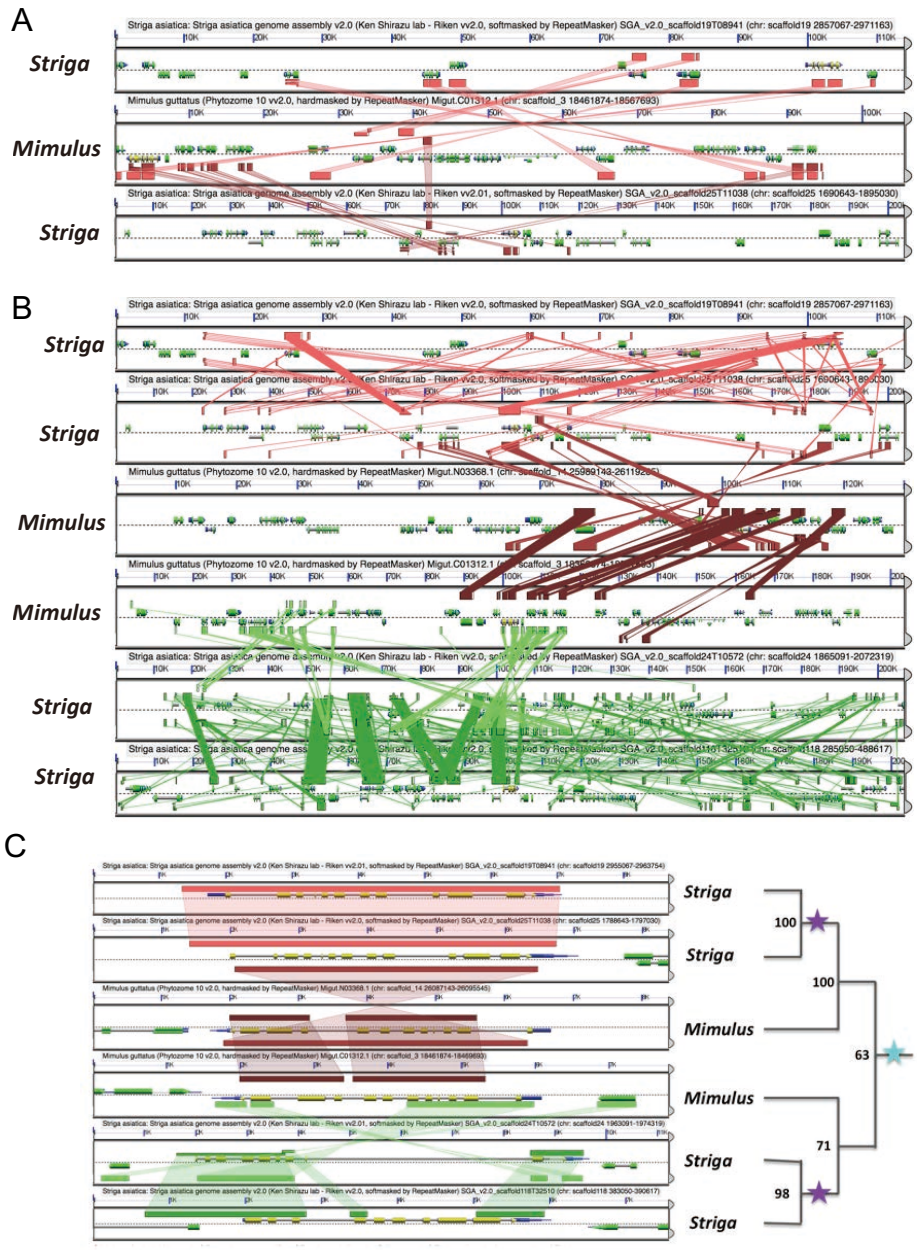


Figure C.5. Microsynteny analysis of two syntenic *Striga* regions and one *Mimulus* region **A.** Example microsynteny analysis of two syntenic *Striga* regions and one *Mimulus* region showing evidence of fractionated gene content. **B.** Syntenic regions in (A*)* with one additional region of *Mimulus* and two additional regions of *Striga*. **C.** Evidence of 4x *Striga* to 2x *Mimulus* collinear anchor genes present on the duplication node of a gene family tree (Figure 1). Cyan star represents duplication in a common ancestor of *Mimulus* and *Striga*, and purple star represent duplication in the *Striga* lineage. Results can

be regenerated following the links below: A. https://genomevolution.org/r/11obn, B. https://genomevolution.org/r/11obq, *c*. https://genomevolution.org/r/11q3g

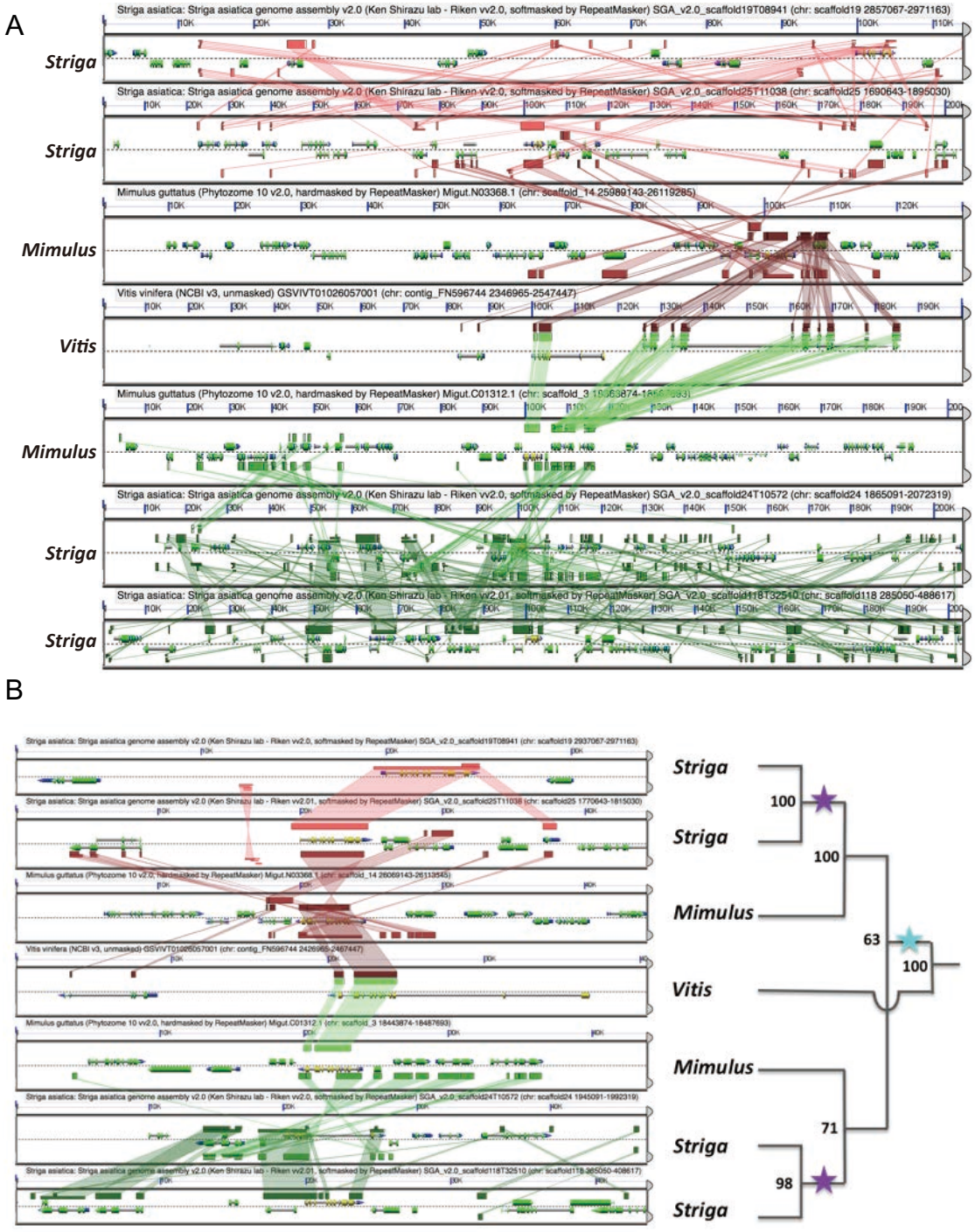


Figure C.6. Microsynteny among four *Striga* and two *Mimulus* syntenic regions **A.** Example of microsynteny among four *Striga* and two *Mimulus* syntenic regions shown in Figure C.5, and one *Vitis* region. **B.** Evidence of 4x *Striga* to 2x *Mimulus* to 1x *Vitis* collinear anchor genes present on the duplication node of a gene family tree (Figure 1). Cyan star represents duplication in common ancestor of *Striga* and *Mimulus*, and purple star represents duplication in the *Striga* lineage. Results can be regenerated following the links below: A. https://genomevolution.org/r/11ufe, B.

https://genomevolution.org/r/11ue2

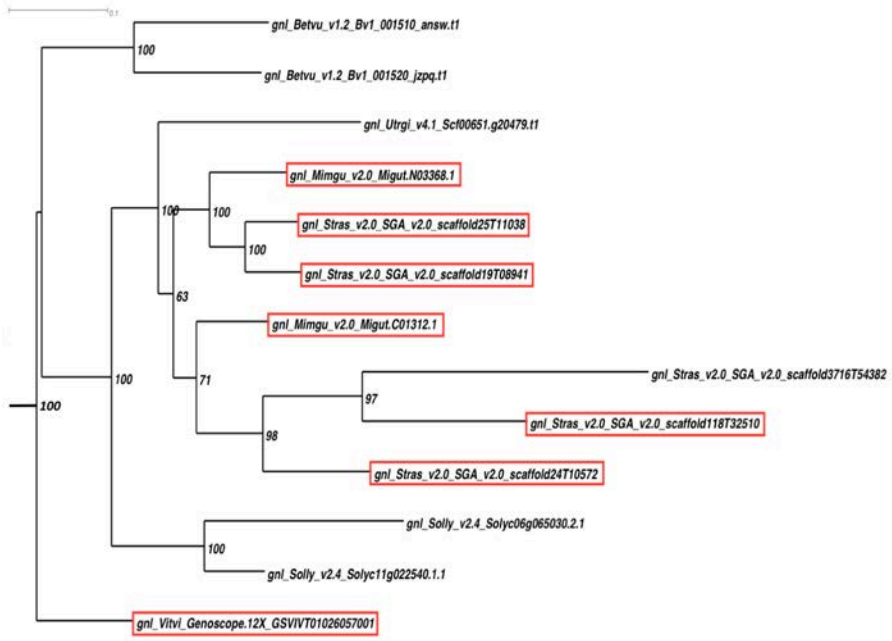


Figure C.7. Example a subtree of RAxML ML gene family tree (orthogroup 460) shows the duplication of anchor genes located on homologous *Striga*, *Mimulus*, and *Vitis* syntenic blocks. Anchor genes present on the syntenic blocks are surrounded in red boxes.

C.3 Ancestral gene family reconstruction

Searcy[S60,S61] proposed that gains of parasitic ability, then losses of functions supplemented by the host, and finally gains of highly specialized traits would characterize the evolutionary transition to heterotrophy in parasitic angiosperms. Therefore, the relative timing of evolutionary events, and thus the age of affected gene families, should follow a predictable pattern. The supplementary functions should be more broadly shared with the parasite and host and therefore older, while newer, lineage-specific functions should provide specialized adaptations to the parasite.

We used the parsimony method in DupliPHY[S62] to reconstruct the presence and size of each gene family in the common ancestor of *Striga asiatica* and the closely related non-parasite *Mimulus guttatus* as well as other successively earlier common ancestors. We used a table with the number of genes observed in each orthogroup (approximate gene family) from the 26-genomes orthogroup circumscription (Data S1S), and the corresponding species tree inferred from hundreds of single copy genes (Figure 1) as input for DupliPhy. The gene family evolutionary dynamics estimated at each node of the 26-genome species tree is shown in Data S1W. Among 10,248 orthogroups with representative

asterid taxa, ~23% showed a significant change in gene numbers between *Striga* and its ancestral node shared with *Mimulus*. We estimated 647 contractions, 1,742 expansions, 456 losses, and, 153 gains (Data S1D).

The relative age of genes in contracted orthogroups was significantly older (two-tailed Mann-Whitney U test, p-values < 2.2e-16) than genes in expanded families (Figure 2D and Data S1E). In support of Searcy's hypothesis, the older, contracted gene families include plant genes whose functions are more likely to align with vestigial parasite functions. The relatively younger expanded gene families, apparently gained largely as a result of the younger *Striga* WGD (Figure 2D and Data S1E), also support this hypothesis by providing a more recent source of genes to encode specialized traits in the parasite.

C.4. Selective pressure on protein-coding genes in the *Striga* genome

Selection pressure on each *Striga* protein sequence was estimated by calculating the ratios the rate of non-synonymous substitutions (*Kn*) to the rate of synonymous substitutions (*Ks*) between *Striga* and *Mimulus* orthologous genes present in syntenic genomic regions using CoGE Synmap function. Among 10,055 orthologous pairs, 40 were detected as under positive selection (*Kn/Ks* > 1.0, Data S1X). These genes include transcriptional factors, hormone response genes, genes involved in ubiquitin-proteasome pathway and histone deacetylase, indicating positive selection in on genes encoding components of signal transduction pathways. These results are consistent to the findings in the genome analysis of *Cuscuta australis*, a stem parasite in the Convolvulaceae family; GO terms of "response to hormones", "DNA methylation" and "regulation of transcription" are enriched in positively selected genes [S63], implying commonality between independently evolved parasitic species. Moreover, the average *Kn/Ks* ratio in evolutionary expanded gene families in the *Striga* genome is significantly higher than that in the contracted gene families (Student's t-test, p< 5e-10, Figure S1), suggesting that expanded gene families are under more relaxed purifying selection pressure than contracted gene families. Such relaxed selection pressure together with gene duplication may lead to neofunctionalization of the duplicated genes and contribute acquisition of new phenotypes, such as parasitism.

C.5. Evolutionary events related to parasitism

C.5.1 Evaluation of Searcy hypothesis

Important facets of the Searcy hypothesis are the function and also the source of genes leveraged by the parasite during the three phases of parasitic evolution[S64]. During ***Phase I***, genetic innovation is required for the evolution of the haustorium either by the acquisition of new genetic material or by

modification of existing genetic material. ***Phase II*** is characterized by loss of genes whose encoded functions were made redundant by resources acquired from the host (e.g., the carbon and water– see below). ***Phase III*** predicts that obligate parasites would add genetic material associated with further adaptations to the parasitic lifestyle. *Striga*, an obligate parasite, should show evidence of all three phases of parasite evolution.

To test these predictions, we created an annotation platform for estimating the function of *Striga* genes by comparison with established functions of orthogroup members. We leveraged these functional annotations as input for analysis of functional biases in specified genes sets relative to the remainder of the genome[S65]. This allowed us to estimate the function of each gene family and functional group that underwent significant changes during *Striga* evolution (Figure 2E, Data S1S).

Presumably, specificity of gene expression is correlated with tissue- or organ-specific function; therefore, changes in gene number for tissue-specific orthogroups can be used as a proxy for changes in tissue function. Thus, we defined a set of tissue-specific orthogroups using microarray expression data[66]. These data are a curated summary of more than 5,000 microarray experiments conducted using the Agilent ATH1 GeneChip®. Updated ATH1 annotations from Gene Networks in Seed Development website (http://seedgenenetwork.net) were used to update the gene expression matrix with current probe annotations. We further screened for Arabidopsis genes with orthogroup assignments. Z-scores were calculated for each gene[S67], and a z-score cut-off of 2 was determined empirically to select gene sets for which >95% of the genes had a Z-score >2 in only one tissue category. Of the <5% that were not stage-specific, roughly half were represented in sub-stages, e.g., stamen and pollen. This score cutoff was also generally sufficient to generate lists with values amenable to the Chi-Square test (i.e., expected values >5 per cell). The *Arabidopsis thaliana* gene identifiers and orthogroups were extracted for tissue-specific genes and were appended to the 26 plant genomes orthogroup classification to identify orthogroups with genes that have tissue-specific expression (Data S1D). The orthogroup lists were tested for proportionality against the background pattern of orthogroup evolution in *Striga asiatica* (Data S1Y).

C.5.2. Haustorium innovation- Phase I

Recently it was reported by Yang et al. 2015[S37] that gene families with preferential haustorium expression were derived from duplicated genes whose orthologs have preferential root or pollen gene expression in non-parasitic angiosperms. We classified Orobanchaceae (*Striga hermonthica, Phelipanche aegyptiaca,* and *Tryphisaria versicolor*) "haustorial" genes identified in that study into the

26-genome orthogroups and examined assignments to orthogroups with tissue-specific genes. Concordant with the results in Yang et al. (2015) we observed the recruitment of tissue-specific genes for haustorial development in parasitic plants. Haustorial genes were enriched for orthogroups with a tissue-specific expression pattern. Testa, hypocotyl, and root were identified as likely sources for haustorial genes, but most predominantly pollen (Data S1B). These results suggest that during ***Phase I***, haustorium innovation is underpinned with neo-sub-functionalization of existing (and duplicated) genes from tissue-specific gene families. Curiously, most of the tissue-specific gene families (except seedling, leaf and embryo) were also enriched for contracted orthogroups (Data S1Y); this may represent ***Phase II*** – loss of parasite functions via host complementation.

C.5.3 Functional complementation – Phase II

Gene family contractions characterize patterns of gene family evolution in *Striga asiatica*, and conspicuously orthogroups with highly tissue-specific expression are enriched for contracted gene families (Data S1Y). We expected to see contractions in "root" specific Orthogroups since *Striga* completely lacks a proper root system[S68], yet these data suggest that the pattern of functional complementation by the host extends to other parasite functions beyond the more obvious changes like loss of a functional root system. Consistent with the relatively normal outward appearance of *Striga* leaves, leaf-specific orthogroups lacked strong evidence of evolutionary shifts. Evolutionary losses of leaf and root genes in the leafless and rootless holoparasites *Monotropa* (a mycoheterotroph) and *Cuscuta* have been reported[S63,S69,S70]. However, even the leafy green hemiparasite *Striga* is heavily dependent upon the host for carbon, and entirely heterotrophic as a seedling and during its extensive subterranean growth phase[S68,S71,S72]. Therefore, we should see evidence for losses of *photosynthesis-related* genes.

It has been shown that the plastid genomes of parasitic plants undergo wholesale gene loss, accelerated sequence evolution, and genome reduction, including the loss of photosynthesis genes in holoparasites[S64,S73]. These observations support ***Phase II*** of the Searcy hypothesis that vestigial parasite functions, like carbon assimilation, are supplemented by host photosynthesis, and through time are lost by parasitic plants due to the relaxed constraint of genes involved in the pathway. A recent study[74] defined a list of photosynthesis genes used to survey changes in the photosynthetic apparatus in three species of parasitic *Orobanchaceae*, including *Striga hermonthica*. Concordant with the findings in Wickett et al.[S74], we found that most gene families representing chlorophyll synthesis and photosynthesis pathways are present. However, some of these gene families encoding proteins involved in heme and protoporphyrin IX (in the chlorophyll biosynthesis pathway), as well as light harvesting, showed signatures of contraction (Data S1G). By contrast, the nuclear-encoded photosystem gene

families were intact compared to the ancestral state (shared with *Mimulus*, Data S1G).

Additional ***Phase II*** signatures of gene loss in the genome of *Striga asiatica* include overrepresentation among contracted orthogroups of the KEGG pathways "photosynthesis-antenna proteins" (Benjamini P=0.0021) and "carbon fixation in photosynthetic organisms" (Benjamini P=0.0419) (Data S1F). Among contracted orthogroups, the GO Biological Process (BP) terms "protein-chromophore linkage" (Benjamini P=6.6e-5), "carbon fixation" (Benjamini P=0.0015), and "photosynthesis, light harvesting in photosystem I" (Benjamini P=0.0023) were significantly enriched (Data S1H). A similar theme of photosynthesis-related losses is also observed in GO Cellular Compartment (CC) terms "plastoglobule" (Benjamini P=2.5e-5), "light-harvesting complex" (Benjamini P=0.0021), "photosystem 1" (Benjamini P=0.0256) and "thylakoid" (Benjamini P=0.0471) that were enriched among contracted orthogroups (Data S1H). These losses may explain the reduced photosynthetic efficiency of *Striga*[S68, S71], even though *Striga* still maintains low levels of photosynthetic flux that result in carbon fixation[S68].

Leaves of *Striga* have undifferentiated mesophyll[S75], a low number of plastids per cell[S76], low chlorophyll concentration[S77], an insensitive apparatus for regulating water loss[S78], and likely a negative net carbon gain in leaves[S75]. Consistent with these reductions in anatomy and function of *Striga* leaves GO BP terms "leaf development" (Benjamini P=7.5e-4), "regulation of stomatal movement" (Benjamini P=0.0298), "transpiration" (Benjamini P=0.0346), and "vasculature development" (Benjamini P=0.0339) are overrepresented among contracted orthogroups (Data S1H). This indicates that genes encoding elements of the transpirational apparatus of *Striga asiatica* are also under relaxed constraint. Indeed, the insensitive water loss apparatus[S78] and abnormally high nighttime foliar carbon emission due to constitutively open stomata[S75,S79] show that *Striga* has limited capability to regulate water loss. It has been shown that the closely related holoparasite *Phelipanche* expresses a full complement of chlorophyll synthesis genes, but not photosystem genes[S74]. Additional roles for chlorophyll (and other tetrapyrroles), like retrograde plastid-nuclear signaling[S80] may explain conservation of these pathways in obligate parasites that have diminished photosynthetic capability. Together with our results, this suggests that the primary function of the *Striga* leaf is not carbon assimilation.

A clear and dominant signal in the ancestral gene family reconstruction is the contraction of cellular response machinery. ~28% of all overrepresented GO BP terms in contracted orthogroups, compared to ~4% in the expanded orthogroups, were "response" to abiotic or biotic stimuli including virtually all major plant hormones (Data S1H). Also included were numerous "signaling" terms that also implicate hormone response/action (Data S1H). Furthermore, the KEGG pathways "plant hormone

signal transduction" (Benjamini P=1.2e-10) and "plant-pathogen interaction" (Benjamini P=0.0169) were also enriched among contracted orthogroups. Consistent with Searcy's prediction of complementation by the host plant of vestigial parasite functions, these data along with the reported insensitivity to water stress (thus implicating ABA [S78]) show that the parasite may have increased its reliance on the host to sense and respond to its environment. This shift would reduce the energetic burden to perceive and integrate environmental cues while at the same time promoting parasite wellness over a stressed host plant. The same applies to biotic stresses – parasites could leverage host responses and defense strategies to biotic stress without expending its own resources. This might even expand the parasite niche by leveraging locally adapted defense responses. These data reveal a wide pattern of loss of sensing and response systems that provides strong support to the Searcy hypothesis.

Functions that are lost and complemented by the host during ***Phase II*** may also be targets for ***Phase III*** specialization of the parasite-host relationship. For instance, alteration in water movement functions may span evolutionary events in ***Phases II*** and ***III*** because the host plant could complement water stress response pathways while decreased water potential[S68], constitutive transpiration[S73,S81] and other alterations to the water relations apparatus such as host vessel element invasion by parasitic oscula[S82] could be adaptive. We can parse evolutionary shifts within a common process into the respective phases based on the timing of these events. For instance, the GO BP term "response to water" (Benjamini P=9.29e-4) is enriched in expanded orthogroups that have been shown to be significantly younger than contracted ones. This would suggest these expanded orthogroups represent ***Phase III*** signatures, even though orthogroup contractions dominate water relation signatures.

C.5.4 Parasite adaptation – Phase III

During the transition to obligate parasitism, it was suggested by Searcy[S60,S61] that parasitic plants would adapt to the parasitic lifestyle by accruing new genetic information. We have shown that the WGD in *Striga asiatica* is a source for gene family expansion. It is, therefore, possible that new and highly derived genes sourced from the *Striga* lineage-specific WGD encode genes that underpin highly adapted parasite traits, especially in the novel haustorium. The primary function of the haustorium is to connect the parasite to its host, and implicit in this function is the acquisition of host resources and regulation of host defenses. Heide-Jørgensen and Kuijt[S83,S84] observed that the haustorium of the closely related *Triphysaria versicolor* contained transfer-like cells. Because evidence of phloem continuity in *Striga* is lacking, we hypothesized that ***Phase III*** innovation may include cellular machinery such as endocytosis and vesicle mediated transport that would facilitate acquisition of host resources, perhaps in haustorial-interface transfer cells. It is clear that the high proportion of heterotrophic carbon, especially in unemerged *Striga* seedlings at virtually 100%, would require a

highly efficient means of obtaining host carbon[S72]. Our survey of functions in expanded orthogroups revealed that GO BP terms “vesicle-mediated transport” (Benjamini P=1.04e-6) and “Golgi vesicle budding” (Benjamini P=1.29e-4) were enriched. Furthermore, the GO CC terms “Golgi membrane” (Benjamini P=1.05e-15), “trans-Golgi network” (Benjamini P=4.99e-8), “endosome” (Benjamini P=4.90e-8), “cis-Golgi network” (Benjamini P=2.22e-4), “early endosome membrane” (Benjamini P=0.0054), “clathrin-coated vesicle membrane” (Benjamini P=0.0273), “trans-Golgi network membrane” (Benjamini P=0.0321), and “Golgi cisterna membrane” (Benjamini P=0.0437) and KEGG pathway “endocytosis” (Benjamini P=5.39e-4) were enriched among expanded orthogroups（Data S 1F and H）. This suggests that relatively young and significantly expanded orthogroups that encode inter- and intra-cellular transport genes may represent ***Phase III*** innovations related to host resource acquisition.

Host-induced gene silencing from host plants to *Orobanchaceae* parasites[S85,S86] provides a potential mechanism for parasite resistance involving RNA movement from host to parasite. Previous work has revealed massive mRNA transfer between parasite plant *Cuscuta* and host[S87]. However, the mechanism(s) of RNA transport in these systems remain unknown. Clues that RNA transfer may occur in *Striga* as well are found in enriched GO Molecular Function terms that are unique in expanded orthogroups that included “mRNA binding” (Bonferroni P=7.1e-16), “RNA binding” (Bonferroni P=2.3e-11), “nucleic acid binding” (Bonferroni P=4.4e-7), “poly(A) binding” (Bonferroni P=6.9e-4), and “single stranded RNA binding” (Bonferroni P=0.0015) (Data S1H). These orthogroups encode nucleic acid binding proteins that could be part of a mechanism for RNA transfer between parasitic plants and host plants, perhaps similar to phloem localized RNA binding proteins that likely facilitate mRNA translocation via phloem in plants[S88].

D. Analyses of selected gene families

D.1 Plant hormone related genes

D.1.1 Auxin

Genes related to auxin biosynthesis, transport, receptor and signalling were manually assessed for their presence in the *S. asiatica* genome using BLAST programs from the annotated CDS sequences and the genome sequence. All known auxin-related genes are conserved in the *S. asiatica* genome (Data S1I). However, several gene families including major auxin responsible genes[S89], such as the small auxin up RNA (SAUR), GH3, and IAA, are assigned to contracted orthogroups (Data S1I), suggesting the auxin responses may have been simplified during parasitism evolution. *Striga* as an obligate parasite

has lost their root systems, although adventitious root-like structures emerge to form secondary haustoria. Contraction of auxin responsive genes may reflect loss of structures and physiologies that support an autotrophic plant life style.

D.1.2 Cytokinin

Genes involved in cytokinin biosynthesis, perception and signalling were manually assessed for their presence in the *S. asiatica* genome. We found that all tested genes are conserved (Data S1I). A number of cytokinin metabolism genes, which encode cytokinin oxidase/dehydrogenase (CKX), were highly expressed during infection. The expression of a CKX-encoding gene in the haustorium at 7-d after host interaction was confirmed by RT-qPCR and *in situ* hybridisation (Figure 4H) The hyaline body-specific expression of CKX suggests that cytokinin is degraded in this tissue. In *Arabidopsis*, expression of CKX gene is induced by cytokinin accumulation to remove the excess amount of cytokinin[S90]. Thus it is possible that the coordinated expression of IPT and CKX functions to control the cytokinin content in the haustorium.

D.1.3 Abscisic acid (ABA)

In contrast to non-parasitic plants, *S. hermonthica* stomata remain open in drought-stressed leaves and display reduced sensitivity to applied ABA[S91,S92]. This evolved response is most likely to maximize transfer of water and/or nutrients from the host even under dry conditions. Previous studies showed *S. hermonthica* synthesizes ABA, and consistent with this, all the genes involved in ABA synthesis and catabolism were identified in the *S. asiatica* genome[S93,S94](Data S1I). ABA transporters such as ABCGs and AITs were highly conserved in *S. asiatica*[S95], suggesting that ABA can be transported from vascular tissues into stomata in *S. asiatica*.

All core ABA signaling components (PYR/PYL receptors, PP2Cs, SnRK2s) were also present. Although all three ABA receptor subfamilies (I, II and III) were represented in *S. asiatica*, there appeared to be a preponderance of subfamily I receptors, which are the most sensitive receptors to ABA[S96,S97]. The *S. asiatica* genome contains 9 class A PP2C-encoding genes. One of the PP2C genes contains mutations near a conserved tryptophan residue, as reported in PP2C1 gene in *S. hermonthica*, is likely acting as a dominant negative regulator for ABA signaling to keep high transpiration in *Striga*[S92](Figure D.1). In addition, although SnRK2-targeted ABF transcription factor sequences exist in the *Striga* genome, the alignment for ABI5 is very poor. *ABI5* plays a key role in late seed maturation and germination and a potentially non-functional *ABI5* in *S. asiatica* could lead to ABA insensitivity[S98].

With respect to guard cell function, core ABA signaling outputs to a collection of ion channels[S99]. Sequences for KAT1 and KUP6 potassium channels were identified in *S. asiatica,* but only two SLAC family anion channels possessed complete domains. By contrast, *Arabidopsis* has five SLAC-like genes. A loss-of-function mutation in SLAC1 resulted in reduced stomatal closure in response to ABA[S100]. It is possible that the absence of three SLAC-like genes could contribute to insensitivity of *Striga* stomata to close in the presence of ABA.

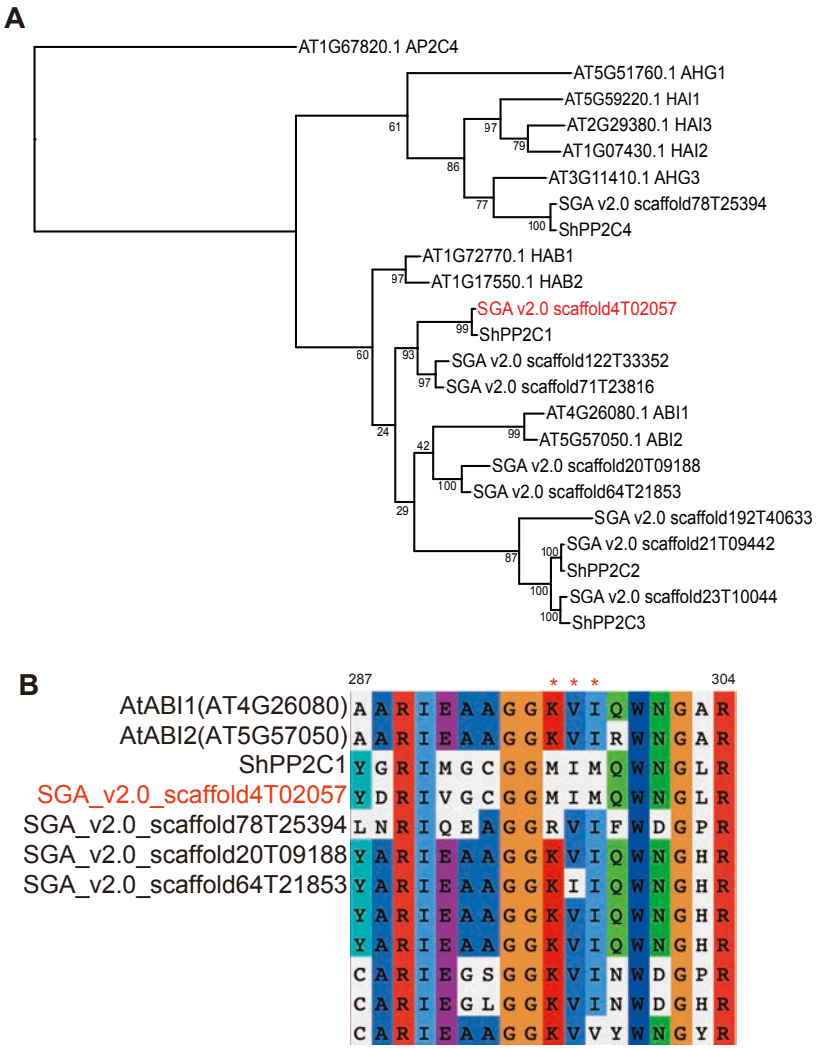


Figure D.1. Class A PP2C gene family in the *S. asiatica* genome.
A. Maximum likelihood tree of amino acid sequences of ClassA PP2C genes from *Arabidopsis*, *S. hermonthica* and *S. asiatica*. AT1G67820 sequence was used as a root. Bootstrap values of 100 replicates are indicated at each node. **B.** Amino acid alignment of *Arabidopsis* ABI genes, *S. hemonthica* PP2C1, and *S. asiatica* class A PP2C genes. Mutations that confer interruption of ABA signalling are shown by asterisks. Dominant negative PP2C protein in *S. asiatica* is highlighted by red letters.

D.1.4 Ethylene

Besides SLs, ethylene is also able to induce *Striga* seed germination[S101]. In fact, ethylene gas was used for suicidal germination strategy in order to eradicate *S. asiatica* from North and South Carolina in USA[S102]. To understand ethylene responses in *Striga* spp., the number of genes involved in ethylene signaling and biosynthesis were investigated using reciprocal blast searches. The *Arabidopsis* genome has 5 ethylene receptor encoding genes, *ETR1*, *ETR2*, *ERS1*, *ERS2* and *EIN4* and the receptor-mediated signal is transduced via *CTR1* and *EIN2* to the nuclear-localised EIN3/EILs transcriptional regulators. The EIN2 C-terminal end leads to the stabilisation of EIN3/EILs by degradation of F-BOX proteins, EBF1 and EBF2, that negatively regulate ethylene responses[S103]. The *S. asiatica* genome contains all ethylene signaling and biosynthesis genes, except *ETP1* and *ETP2* (Data S1I). The F-box proteins ETP1 and ETP2 negatively regulate EIN2 via the 26S proteasome-mediated degradation in *Arabidopsis*[S104]. However, the amino acid sequences of ETP homologues are not well conserved among species[S105]. Thus, it is less likely that the loss of ETP genes reflect the unique ethylene response in *Striga* spp. The key transcription factor *EIN3/EIN3-like (EIL)* family was in contracted orthogroups. On the other hand, the *S. asiatica* genome contains 5 orthologues of *CTR* gene, a key negative regulator of ethylene signaling, showing expansion by orthogroup analysis. This may suggest that some of physiological responses against ethylene were modified during *Striga* evolution.

D.1.5 Jasmonic acid (JA) and salicylic acid (SA)

JA and SA are two major defence-related plant hormones. We have examined the presence of JA and SA-related genes in the *S. asiatica* genome (Data S1I). Genes related to JA and SA biosynthesis as well as signalling genes are all conserved in the *S. asiatica* genome.

D.2 Strigolactone (SL)-related genes

D.2.1 SL biosynthesis genes

SLs are well known as germination stimulants for *Striga*. It has been questioned whether *Striga* can produce active SLs by themselves. Mutants and enzyme analyses of various plant species identified key genes encoding SL biosynthesis pathway. SLs are derived from carotenoids. DWARF27 (D27)[S106], catalyses the isomerization of all-*trans*-ß-carotene to 9-cis-ß-carotene, which is sequentially cleaved by carotenoid cleavage dioxygenase7 (CCD7/MAX3) and carotenoid cleavage dioxygenase8 (CCD8/MAX4)[S107,S108] to yield carlactone (CL), a common precursor of SLs[S109]. Carlactone is

further oxidized by cytochrome P450 enzyme (CYP711A1/MAX1) to produce bioactive SLs. The biosynthesis pathway from carotenoid to carlactone is supposed to be widely conserved among plants, while the later steps can be more diversified. Rice genome encodes five MAX1-homogolue genes and two of these proteins sequentially catalyse carlactone to 4-deoxyorobanchol and 4-deoxyorobanchol to orobanchol[S110], which has canonical SL structure with four rings. *Arabidopsis* genome encodes only one MAX1 protein that catalyses CL to calactonoic acid (CLA)[S111]. CLA is further methylated by unknown methyltransferase to produce methyl carlactonoate (MeCLA), and an oxidoreductase-like protein LATERAL BRANCHING OXIDOREDUCTASE (LBO) converts MeCLA into bioactive non-canonical SLs in *Arabidopsis*[S112]. *S. asiatica* genome encode one each of SL-biosynthesis gene orthologues (Figure D.2). Highly conserved amino acids among angiosperms, suggesting the ability of *Striga* to synthesise SLs, consistent with a previously published report[S113].

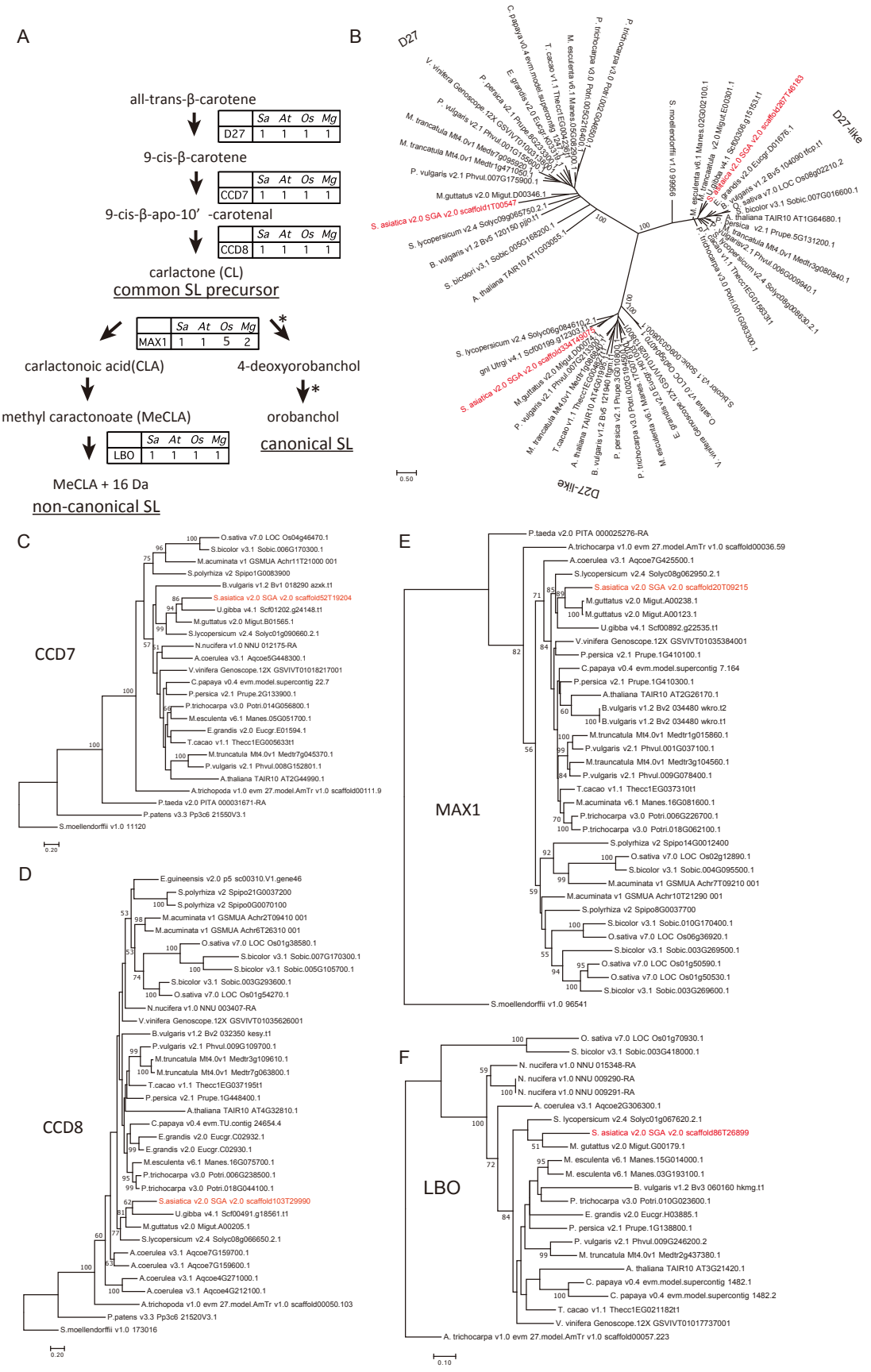

A
all-trans-β-carotene
D27
9-cis-β-carotene
CCD7
9-cis-β-apo-10’ -carotenal
CCD8
carlactone (CL)
common SL precursor
MAX1
carlactonoic acid(CLA)
4-deoxyorobanchol
methyl caractonoate (MeCLA)
orobanchol
LBO
canonical SL
MeCLA + 16 Da
non-canonical SL
B
D27
D27-like
C
CCD7
D
CCD8
E
MAX1
F
LBO

Figure D.2. SL biosynthesis genes in *S. asiatica* genome
A. Canonical and noncanonical SL biosynthesis pathway and corresponding enzymes in each steps modified from Brewer et al. 2016[S112]. Tables show the number of genes encoding each enzyme in indicate species (Sa: *Striga asiatica*, At: *Arabidopsis thaliana*, Os: *Oryza sativa*, Mg: *Mimulus guttatus*). Asterisks indicate MAX1 homologues regulating steps found in rice [S110]. Non-canonical SL biosynthesis pathway mediated by LBO was found in Arabidopsis. **B-F.** Maximum likelihood tree of amino acid sequences of SL-biosynthesis genes from various plant species. B, D27 homologues, C, CCD7 (D17/MAX3) homologues, D, CCD8 (D10/MAX4) homologues, E, MAX1 homologues, F, LBO homologues.

D.2.2 SL signalling genes

Perception and signalling of SLs and karrikins are known to be regulated by an F-box protein (D3 in rice and MAX2 in *Arabidopsis*), α/β hydrolase (D14 and D14-LIKE in rice, AtD14 and KAI2 in *Arabidopsis*) and D14/KAI2 interacting repressor proteins known as D53 in rice[S114–S118]. Genes encoding homologues of these proteins were identified in *S. asiatica* genome. One copy of *D3/MAX2* homologue is found in *S. asiatica* genome and *S. hermonthica* transcriptome. Eleven genes and five contigs are assigned as *D53* homologues in *S. asiatica* genome and *S. hermonthica* transcriptome, respectively (Figure D.3A). D53 is an SL signalling component that forms complexes with MAX2 and D14. SL induces degradation of D53 and promotes the SL signalling pathway resulting in the suppression of bud outgrowth. The *Arabidopsis* homologues of *D53* belong to a family containing 8 genes including *SMAX1*, the suppressor of *MAX2*[S119]. Mutation in *SMAX1* restores the seed germination and the seedling morphogenesis phenotypes of *max2*, but it does not affect lateral root formation or axillary bud outgrowth[S119]. Recent analysis reported that SMAX1-LIKE genes SMXL6, SMXL7 and SMXL8 regulate SL-dependent axillary bud outgrowth in *Arabidopsis*, indicating that the SMAX1 and SMXL6,7,8 regulate karrikin and SL dependent phenotype respectively[120]. Phylogenetic analysis indicates that all the 4 genes in *S. asiatica* genome are clustered with and *SMXL6,7,8,* and 7 genes with *SMAX1*. The transcriptome assembly of *S. hermonthica* contains at least 2 genes that cluster with *SMAX1* and one gene in the *D53* clade. Expression patterns of *SMAX1* homologues in *S. hermonthica* suggest that the *MAX2* homologue and two *SMAX1* homologues are expressed in seeds and seedling stages (Figure D.3A and B). The proteins encoded by these genes possibly interact with highly duplicated *KAI2* homologues to ensure proper SL signalling, leading to *Striga* germination.

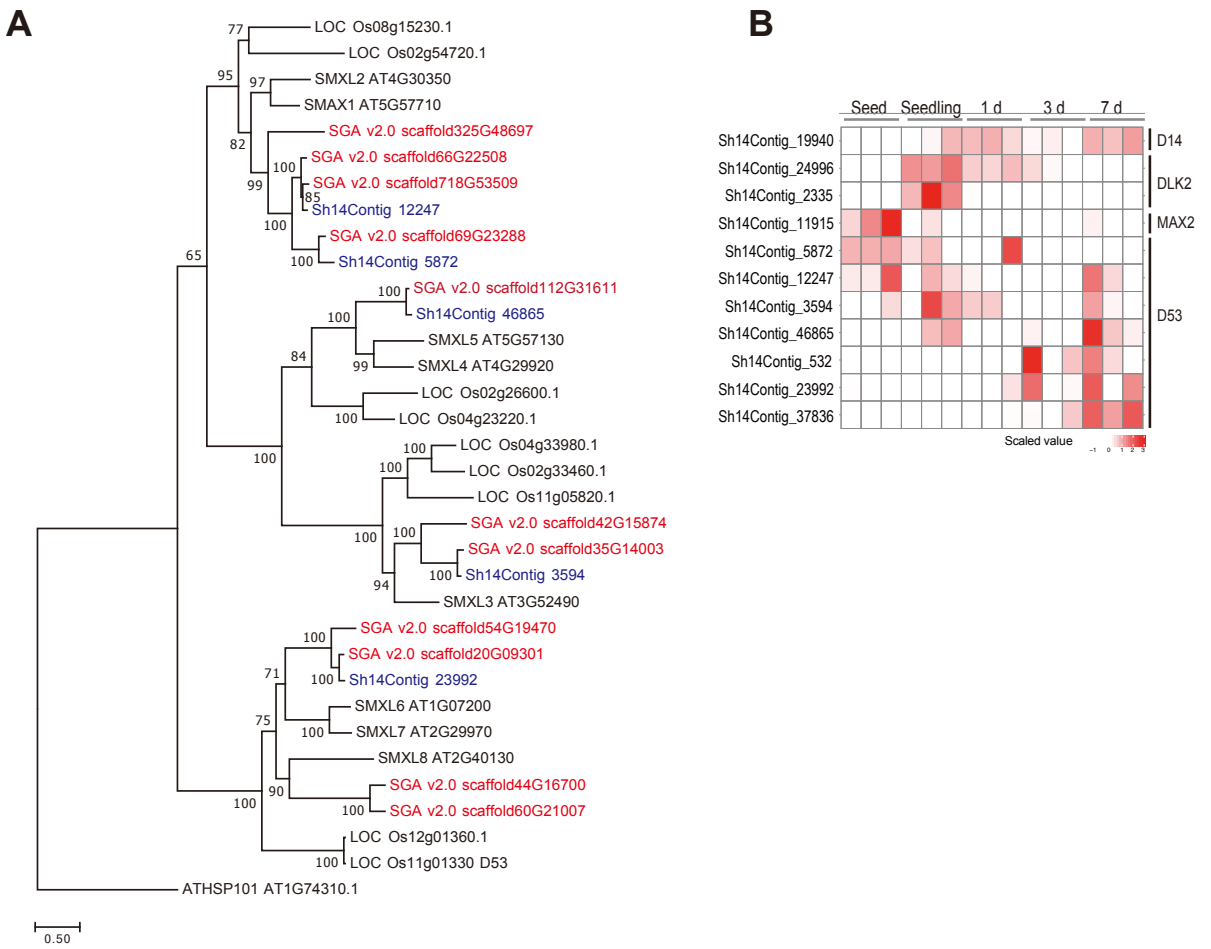


Figure D.3. SL signaling genes in *Striga* spp.
A. Maximum-likelihood phylogenetic tree of D53 homologues in *S. asiatica* (red), *S. hermonthica* (blue), *A. thaliana* and *O. sativa* was drawn. Numbers indicate bootstrap values at each node. **B.** Expression patterns of MAX2, D14, DLK2 and D53 homologues in *S. hermonthica*. The relative expression levels calculated by RNAseq analysis are shown as scaled heatmap.

D.2.3 Genomic distribution of *KAI2* homologues in *S. asiatica*.

KAI2 homologues were searched by BLAST analysis using *Arabidopsis* KAI2 (At4g37470) protein sequence as a query against the annotated data and the assembled genome. The incomplete or chimeric annotations were manually corrected. In total, 21 *KAI2* homologues were found in the *S. asiatica* genome. In addition, we identified 7 *KAI2* sequences that do not encode a full-length protein due to frameshifts, large insertion, or premature termination codons (Data S1J) and defined those as pseudogenes. We also found one *D14* and two *DLK2* homologues in the *S. asiatica* genome. In total, 31 loci on 16 scaffolds contain *D14/KAI2*-related sequences. These 16 scaffolds were compared with each other and with the *M. guttatus* genome using the DAGChainer[S55] function in SynMap of

CoGE[54] (http://www.genomevolution.org). With the default setting (-D 20, -A 5) of DAGChainer, a strong syntenic relationship was detected between the *M. guttatus* genomic region containing *MgKAI2c* and the *S. asiatica* regions containing *KAI2c1* (Figure D.4). The regions containing the intermediate type *KAI2i* do not show syntenic relations between *S. asiatica* and *M. guttatus*. The *S. asiatica* genome contains two *KAI2i* genes, and the *KAI2i_2* containing region (scaffold104) showed strong syntenic relationships with *M. guttatus* scaffold1. However, the *KAI2i* gene is missing in *M. guttatus* scaffold1, suggesting loss of *KAI2i* gene in *M. guttatus* or local acquirement of *KAI2i* in the *S. asiatica* genome (Figure D.5). The *S. asiatica* regions containing the *KAI2d* genes do not show syntenic relationship between each other, suggesting that the *KAI2d* genes are locally duplicated. Similarly, similarities among the *KAI2d* loci are not restricted only to protein-coding sequences but are extended to 5' and 3' regions and introns (Figure D.6). For example, *KAI2d6* and *KAI2d12* are aligned with 97.78% identity in 2,947 bp, which includes 431 bp upstream of the start codon, an 88 bp intron and 763 bp downstream of the stop codon, in addition to the open reading frame. Such high similarity may indicate that 5', 3' and intron sequences harbour conserved regulatory functions, or alternatively, that the gene duplications occurred relatively recently.

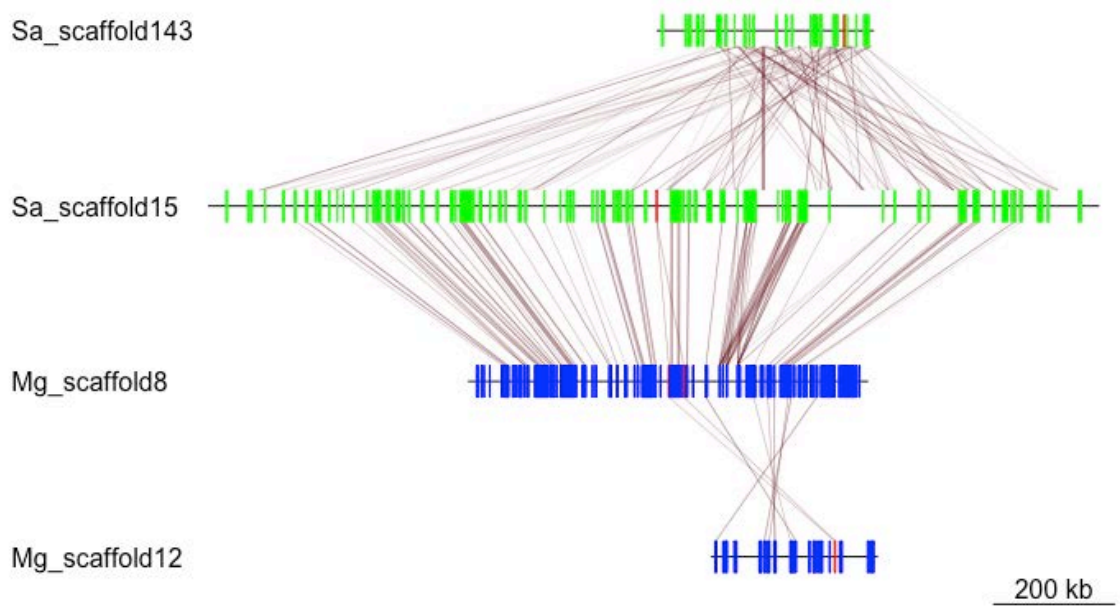


Figure D.4. Syntenic relationships among four genomic regions containing *M. guttatus* KAI2c, *S. asiatica KAI2c1* and *KAI2c2*, respectively.
Genomic fragments of *S. asiatica* scaffold143 (20019-969482, containing *KAI2c1*) and *S. asiatica* scaffold15 (1596215-3032540, containing *KAI2c2*), *M. guttatus* scaffold8 (1786420-2787771, containing *MgKAI2c1* and *MgKAI2c2*), *M. guttatus* scaffold12 (1214175-1498127, containing *MgKAI2i*) are compared with blastZ program in GEvo website. The regions showing similarities (score>3000) are connected with solid lines. Green and grey bars represent protein coding and intron sequences, respectively. Highly syntenic relationships are confirmed between *M. guttatus* scaffold8 and *S. asiatica* scaffold15.

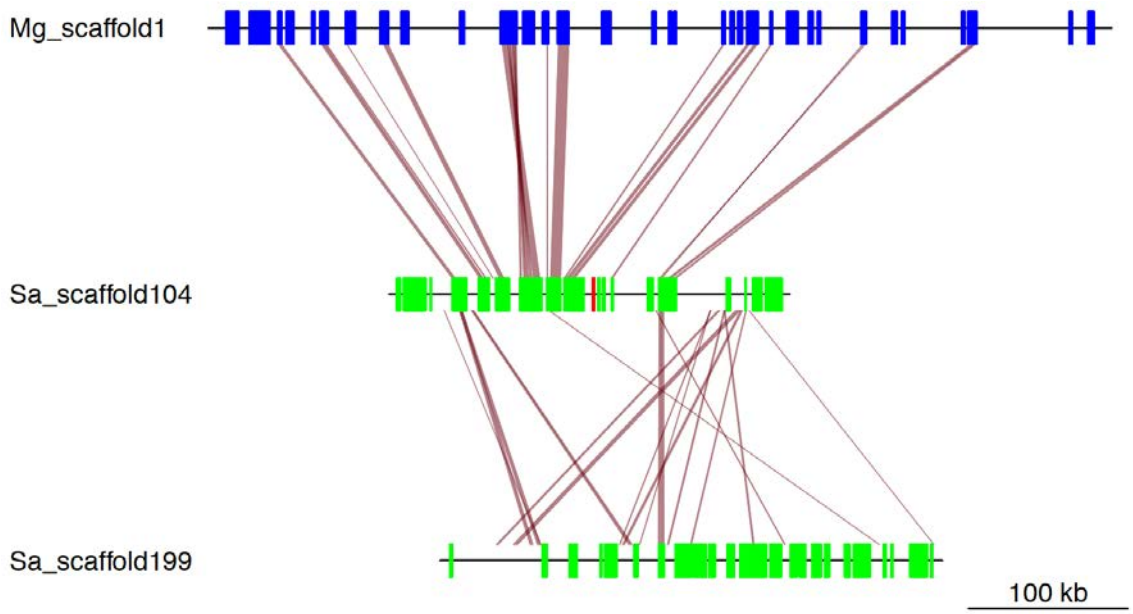


Figure D.5. Syntenic relationship among genomic regions containing *KAI2i* in *S, asiatica* and scaffold 1 in *M. guttatus*.
Genomic fragments of *S. asiatica* scaffold104 (287875-640765, containing *KAI2i_2*) and scaffold199 (1354464-1606599, syntenic but not containing KAI2 related sequences) and syntenic *M. guttatus* region (Scffold1) are compared with blastZ program in GEvo website. The regions showing similarities (score>3000) are connected with solid lines. Green and red boxes represent protein-coding and *KAI2* encoding sequences, respectively.

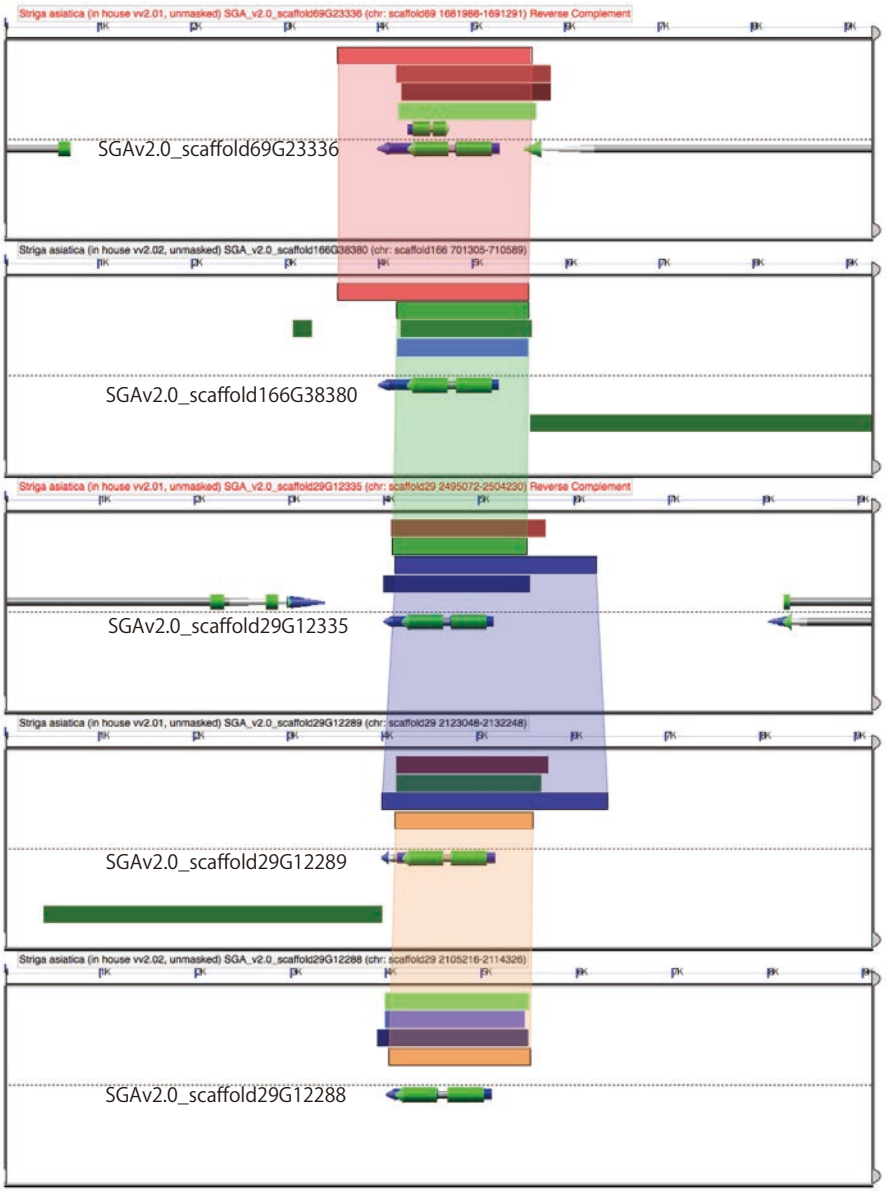


Figure D.6. Comparison of genomic regions of divergent *KAI2d* genes in the *S. asiatica* genome. Genomic regions containing 4 kbp up and downstream of 5 *S. asiatica KAI2d* genes are compared with each other with blastZ program in GEvo (score threshold 10000). High-scoring segment pairs (HSPs) are highlighted with various colors and connected with wedges.

E. *S. hermonthica* transcriptome

E.1 RNA sequencing

S. hermonthica seeds, seedlings and 1, 3 and 7 days after rice infection samples were harvested and subjected to RNA sequencing analysis. Illumina PE libraries were constructed for an insert size of 180 bp and sequenced by Illumina HiSeq2000 sequencer. The read numbers are shown in Table E.1. The PE reads were quality trimmed and filtered, and the reads that have a PE structure after filtering were used for the assembly and mapping.

Table E.1. Total sequence read number from each library of *S. hermonthica* RNAseq. Sequences mapped on rice sequences were shown in brackets.

	Sequence read number (rice sequence number)		
Stages	**Library #1**	**Library #2**	**Library #3**
Seed	57,723,072	55,923,544	44,677,364
Seedling	63,209,994	49,996,818	50,356,660
1 d	12,053,490 (2,054,386)	71,482,012 (29,775,460)	64,136,650 (15,680,526)
3 d	22,427,084 (3,806,778)	58,984,488 (20,418,255)	61,172,636 (18,056,724)
7 d	58,477,954 (8,866,956)	66,170,692 (14,062,472)	62,925,914 (8,022,996)
Rice root	(73,603,526)	(50,425,126)	(45,299,274)

E.2 *de novo* assembly and annotation

The filtered reads were mapped against rice (c.v. Nipponbare) cDNAs and genome (MSU Rice Genome Annotation Project ver. 7) using CLC Genomics Workbench (ver 5) with options of length 0.7 and similarity 0.98, and the sequences unmapped to both rice cDNA and genome were considered as *S. hermonthica* sequences (Table E.1). These unmapped sequences were *de novo* assembled in two rounds using CLC Genomics Workbench (ver. 5) (Figure E.1). In the first round, each library was assembled with the word size 24, and for the second round the resultant was assembled with the word size 64. The assembly was further assembled by CAP3[S123] program with option –o50 –p95 followed by clustering with CD-Hit-EST[S124] ver.4.5.4 (threshold 0.95). This procedure yielded 81,560 contigs. The assessment of the assembly quality is shown in Data S1K. The median contig length (N50) values is 1.3 kb and is similar to the average insert length in the *S. hermonthica* full-length-enriched cDNA library[S32], suggesting a high quality of this cDNA assembly. Homologues for 81% of the *Arabidopsis* proteins were also covered in the assembly (tBLASTn threshold e value 1e-10), which is similar to the

S. hermonthica Sanger EST sequences[S32]. The assembly was annotated for gene ontology (GO) terms using Blast2GO[S125] software and the slim GO terms were assigned using map2slim script (http://search.cpan.org/~cmungall/go-perl/scripts/map2slim).

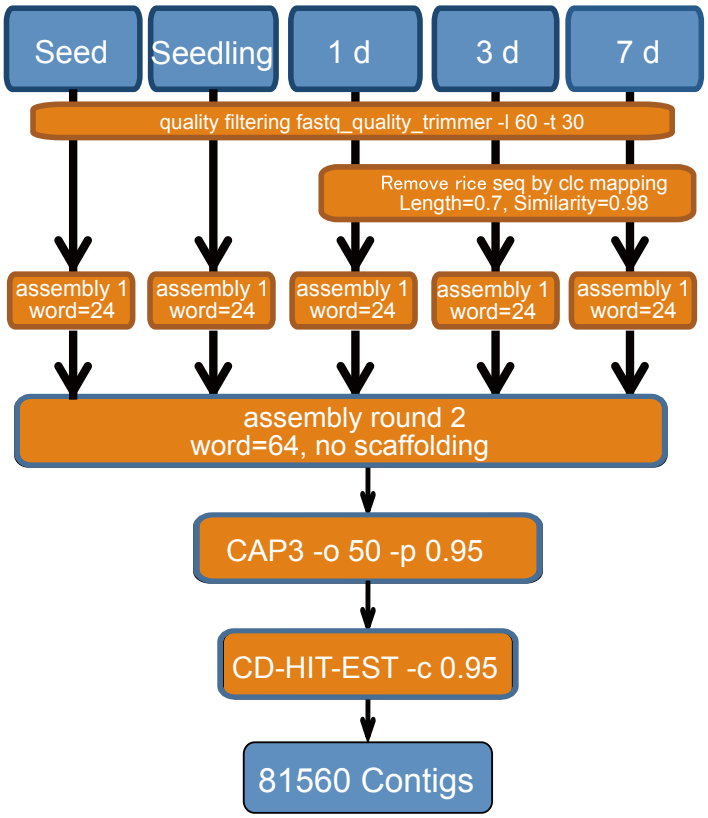


Figure E.1. *S. hermonthica* RNA-seq *de novo* assembly procedures.
Each library was quality filtered by fastx toolkit and mapped on rice cDNA and genome sequences subsequently using CLC genomics workbench. Rice-removed sequences were assembled using CLC genomics workbench *de novo* assembly function and the resultant sequences were assembled again with word size 64. After cap3 and CD-HIT-EST clustering, the sequences shorter than 300 bp were eliminated, resulting 81,560 contigs.

E.4 Read mapping and calculation of expression values

For sequence mapping, the *S. hermonthica* transcriptome assembly and rice cDNAs (MSU Rice Genome Annotation Project ver. 7) were concatenated and used for a reference sequence to be able to detect expression of both organisms. The filtered Illumina sequence reads were mapped on the concatenated sequences using bowtie2[S126] with the default setting. After this mapping step, read counts of *S. hermonthica* and rice were analysed separately. The *S. hermonthica* contigs having more than 10 counts of total mapped reads from sequences obtained from rice control samples were eliminated from the subsequent analysis to avoid the cross-mapping problem. The cDNAs with total mapped reads less than 40 were also removed as lowly expressed genes. After these filtering, 52,669 contigs remained for calculation of expression values. The reads mapped to the *S. hermonthica* reference sequence were normalised with trimmed mean of M-value (TMM) method[S127] and normalised-FPKM (fragments per kilobase of exon per million fragments mapped) values were calculated using the RSEM program[S128] (Data S1L).

E.5 Gene clustering and detection of differentially expressed genes

In order to investigate gene expression dynamics during parasite development of *S. hermonthica*, a principal component analysis (PCA) was performed using normalised FPKM values (Data S1L). The multiple dimensional scaling (MDS) plot shows that the three biological replicates of each stage samples do not a have big variation. In addition, the "seeds", "seedling" and "1 d" samples, and "3 d" and "7 d" samples make distinct clusters, respectively, suggesting that the transcriptomic transitions occurs from seeds to seedlings, and 1 d to 3 d after infection (Figure E.2A). Principal component 1 (PC1), representing a sequential gene expression pattern along parasite development, explained 32.0% of variation in our dataset (Figure E.2B). This suggests that a large part of expressed genes are regulated in a manner consistent with the development of the plant. PC2 explained 28.9% of variation in our dataset and represented the specific gene expression of "seedling" and "1 d" (Figure E.2B), which included shoot tissues whereas the other samples did not, reflecting the methodological effects of our sampling. Normalised FPKM values of *S. hermonthica* were used for a gene expression clustering method[S129]. After selecting genes in the upper 75% and 50% quartile of coefficient of variation for the expression across samples, scaled expression values within tissues were used to cluster these genes for a multilevel 3 x 4 hexagonal self-organising map (SOM)[S130]. One hundred training interactions were used during clustering, over which the alpha learning rate decreased from 0.0035 to 0.002. The final assignment of genes to winning units forms the basis of the gene clusters. The outcome of SOM clustering was visualised in PCA space where PC values were calculated based on gene expression across samples (R stats package, prcomp function). GO enrichment analysis of contigs detected in SOM was performed using the GOSeq Bioconductor package[S131](Data S1N).

Differentially expressed genes were detected by the DESeq package[S132] based on mapped read count data using scripts available in the Trinity software (r2012_10_05) with threshold fold change 4 times and p-value less than 0.001 (Data S1M). All vs all comparison resulted in 10,768 contigs were differentially regulated during *S. hermonthica* seed development and parasitism. MA plots visualise differentially expressed genes (Figure E.3). The most dynamic expression change occurs during germination, because comparison between pre-conditioned seeds and germinated seedlings show many significantly up and down-regulated genes, compared to those among other infection stages. During the infection processes hundreds of genes were differentially regulated. Compared to seedlings (which are before haustorium formation) the 1-d, 3-d and 7-d infection samples contain 375, 727 and 843 upregulated genes and 91, 330 and 695 downregulated genes respectively. There were 111 common upregulated genes and 56 down-regulated genes across all infectious stages (Figure E.4). These numbers are lower than stage-specific genes; *i.e.* 7-d specifically up- and down-regulated genes are 459 and 405, respectively (Figure E.4). These results together with SOM mapping analysis (Figure 4), suggest the occurrence of dynamic changes of expressed gene sets occur during the stages of parasitism, which presumably reflects the developmental shift of the parasite from autotrophic to heterotrophic life cycles.

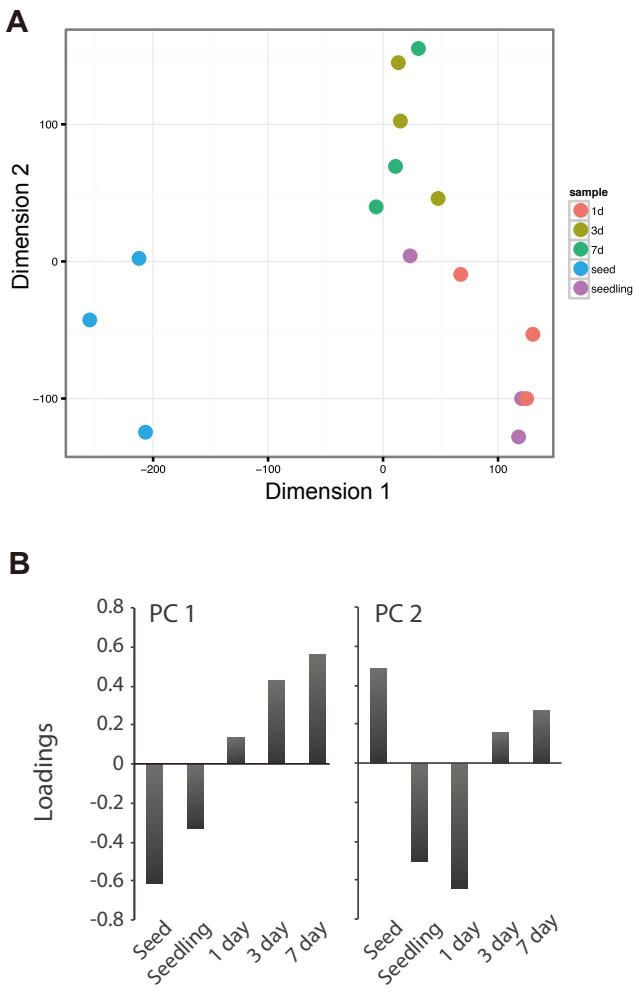


Figure E.2. Assessment of sample variations and principal component analysis. **A.** A multidimensional scaling (MDS) plot assessing the variations among samples shows that biological replicates cluster together. **B.** Loadings of PC1 and PC2 with variance explained. PC1 represented developmental expression pattern, whereas PC2 represented "seedling" and "1 day" specific gene expression.

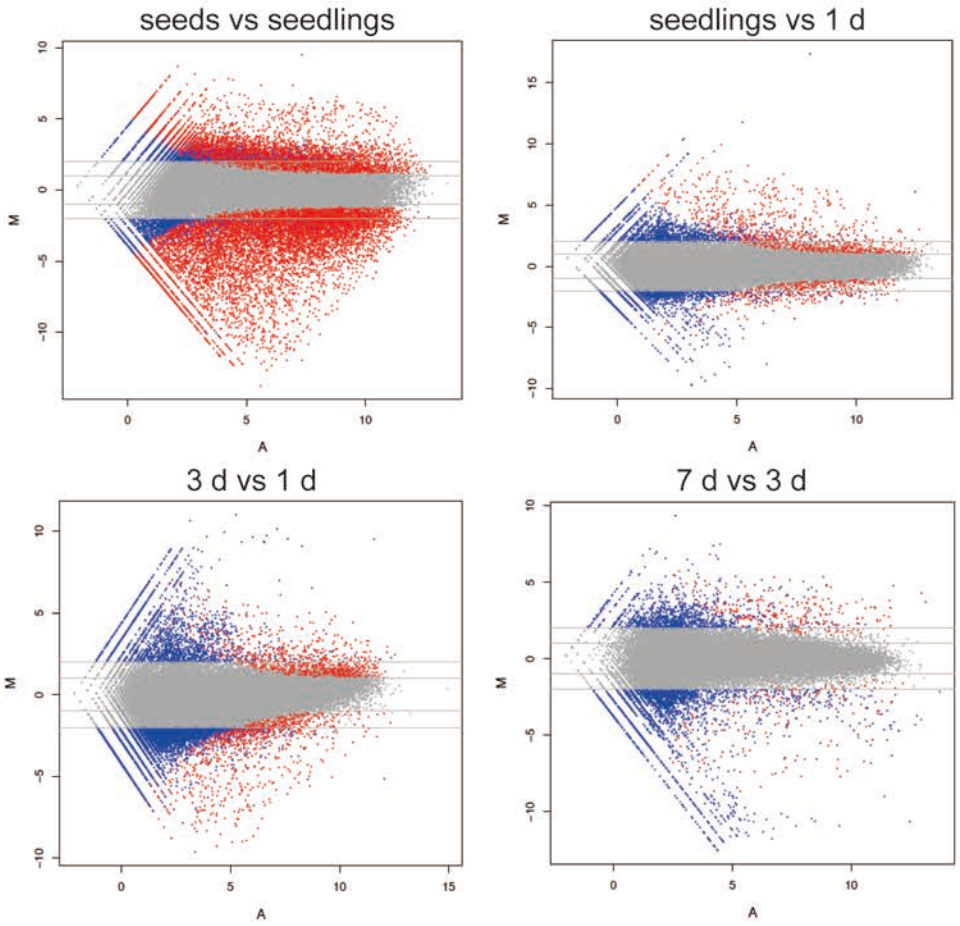


Figure E.3. Differential expression analysis of *S. hermonthica* RNA-seq.

MAplot for comparison of the two indicted stages. M means log ratios and A indicates mean average scale. Red indicates contigs with more than 2 log2 fold changes with adjusted p value=<0.001 and blue dots indicate contigs with more than 2 log2 fold changes but p value>0.001. Grey dots indicate contigs with no expression changes.

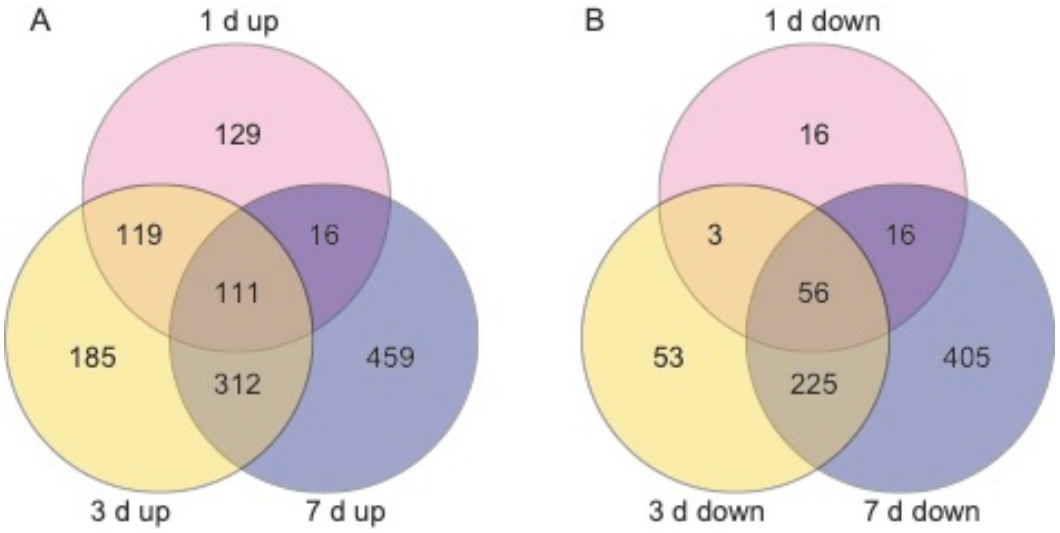


Figure E.4. Differentially expressed genes during infection stages.
Number of differentially expressed contigs compared to before infection (seedling) are shown as Venn diagrams. A, Upregulated genes. B, Downregulated genes.

E.6 Stage-specific gene expression

The RNA-seq analysis during *S. hermonthica* infection suggests that the status of the parasite dramatically changes during infection. Therefore, we aimed to identify marker genes expressed at the particular stages of *Striga* parasitism. To determine the stages of parasitism stages, the rates of xylem bridge formation between *S. hermonthica* and rice roots were analysed. Rice-parasitising *S. hermonthica* samples at 1, 3 and 7 days after host interaction were stained with Safranin-O following protocols previously published[S121] (Figure E.5). In addition, *S. hermonthica* samples were embedded in Technovit 7100 and were observed after cross sectioning and double staining with Safranin-O and Fastgreen[S121] (Figure E.5C-E). At 1 day after host interaction, *S. hermonthica* forms haustorium and invades rice roots, but no xylem differentiation between the host and the parasite was observed. This stage was defined as “early” stage. *S. hermonthica* starts forming a xylem bridge at 3 days after interaction. We often observed a construction of xylem bridge from both the host interaction site and from the parasite stele. However, only 5% of parasites were able to connect vasculatures at this time point (“middle” stage). At 7 days after infection, approx. 60% of parasites succeed to connect vasculatures and the development of hyaline body[S133] is evident in cross section. Therefore, we designated this stage as the “late” stage. To identify stage-specific gene markers, we have selected 30 genes specifically expressed at infection stages and seed or seedlings, and performed RT-qPCR (Figure S2). To avoid cross amplification of host cDNAs, all primers were tested for “rice only” samples and no amplification was observed. We used constitutively expressing Cyclophilin encoding genes as an internal control. Each gene is expressed specifically at one or two stages, experimentally confirming

the RNA-seq data. These genes will subsequently be used as expression markers for assessing *Striga* parasitism.

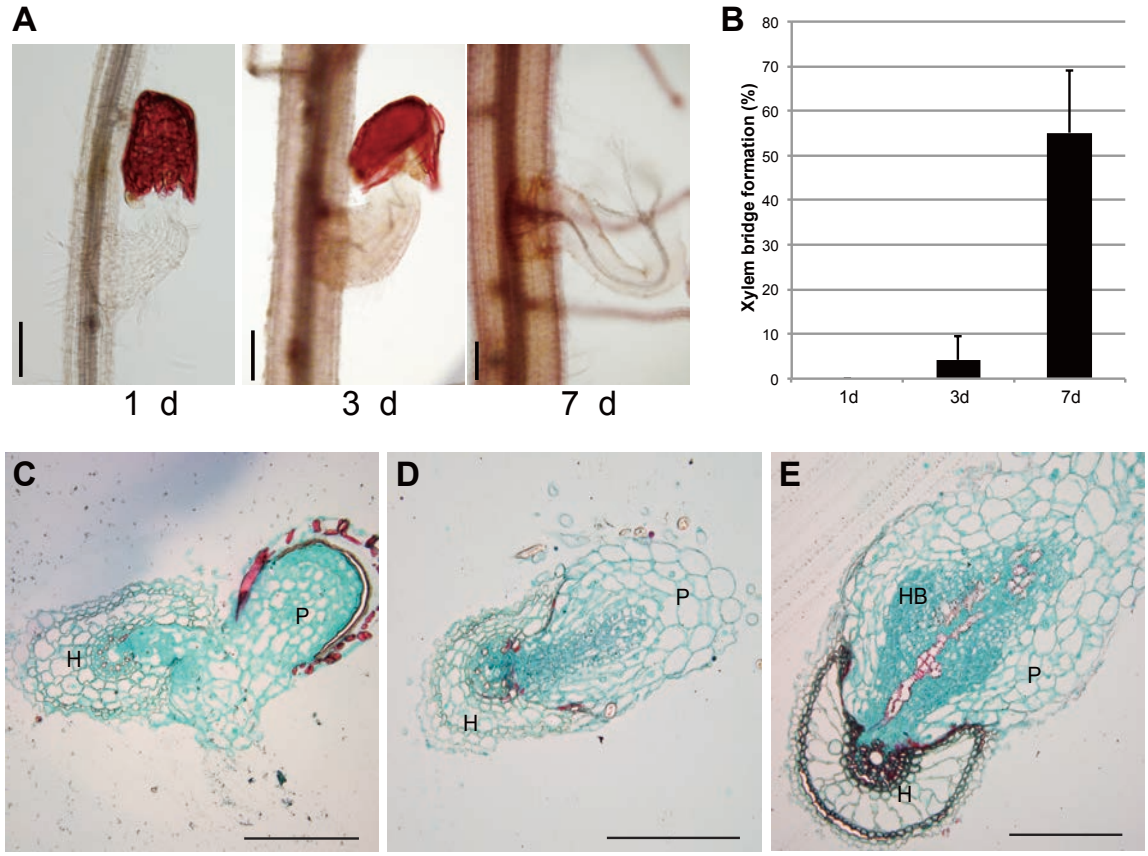


Figure E.5. Vascular connection and infection stages of *S. hermonthica* at 1, 3 and 7 days after host interaction.
A. Xylem connection between *S. hermonthica* and host rice roots. Rice parasitising *S. hermonthica* were stained with Safranin-O at 1, 3 and 7 days after host interaction. At 3 d, xylems are elongated from the host interacting region and from parasite stele bidirectionally, but connection was not established. **B.** Rates of xylem bridge formation at 1, 3 and 7 days after rice interaction. The number of *Striga* seedling with complete xylem connection are counted after Safranin-O staining. n > 200 with more than 15 rice plants. **C-E.** The cross sections of *S. hermonthica* infected rice root stained by fast green and Safranin-O at 1- (C), 3-(D) and 7-(E) day post infection. H, rice (cv. koshihikari); P, *S. hermonthica*; HB, hyaline body. Bar scale, 200 μm.

E.7 Gene expression in nonhost interactions

To further confirm that the expression of genes reflects the parasitism processes, we tested expression patterns of the above genes in nonhost interactions. We previously reported that *L. japonicus* is a nonhost for *S. hermonthica* and the infection stops at the cortical cell layers, and thus no vascular connection was observed in this interaction. On the other hand, *Arabidopsis* is also a nonhost but the vascular connection can be established in this interaction[S121] *S. hermonthica* was infected to *L. japonicus* and *A. thaliana* in a rhizotron chamber and 1-, 3- and 7-day samples were harvested. The *S.*

hermonthica seedlings treated with the haustorium-inducing chemical 2,6-dimethoxy-*p*-benzoquinone (DMBQ) (10 μM) for 2 days were also analysed. All the primers were tested to ensure that there was no amplification of *L. japonicum* or *Arabidopsis* root cDNAs. The early responsive genes were upregulated with DMBQ and with *L. japonicum* but genes that were induced after 3 days in rice interaction did not express in the interaction with *L. japonicum* (Figure 4D). In *Arabidopsis* interaction, the late marker genes were induced during similar times as in rice interactions, indicating that the expression of middle and late-stage genes are associated with stele penetration and haustorium development after vascular connection.

E.8 Analyses of Carbohydrate-Active enzymes (CAZyme)

Upon invasion into the host roots, the *Striga* haustorium must make its way through the root tissue until it can locate and join with the host xylems[S82]. Thus, it is likely that cell wall degrading/modifying enzymes are active in *Striga* invasion. To identify the cell wall-modifying enzymes from *Striga* spp., annotated proteins from the *S. asiatica* genome and the *S. hermonthica* transcriptome assembly were classified with carbohydrate-active enzyme (CAZy) database[S134] using dbCAN, a web-based annotation tool[S135]. In *S. asiatica*, 1223 predicted genes were assigned to at least one CAZyme classification with 1407 motifs including 350 glycoside hydrolases (GHs), 34 polysaccharide lyases (PLs), 486 glycosyltransferases (GTs), 222 carbohydrate esterases (CEs), 147 auxiliary activities (AAs), and 151 carbohydrate binding modules (CBMs) (Data S1Z). Using the same method, 1,533 and 1,609 genes were assigned for at least one CAZyme classification in *Arabidopsis* and *M. guttatus*, respectively. Comparing each CAZyme class, none was found to be particularly over-represented in the *S. asiatica* genome (Data S1AA). Therefore, the acquisition of host invading function is probably not due to the duplication or acquisition of particular CAZymes. Next, proteins from the *S. hermonthica* transcriptome were classified with CAZyme motifs. In total 1,212 contigs were assigned at least one CAZyme motif and a total 1,292 of CAZyme motifs were found (Data S1O). Among them, 252 contigs were differentially regulated during the infection stages compared to the seedling stage. Clustering analysis showed that various CAZyme-encoding genes were expressed throughout the infection stages (Figure 5A). CAZyme classification revealed that motifs assigned for AA and GH were upregulated at 3 and 7 days after host interaction (Figure 5B, Figure E.6). The detailed numbers of significantly upregulated genes in each stage are listed in Data S1P. Plant cells form two types of cell walls, the primary and the secondary cell walls. In general, the primary cell walls are synthesised in growing cells and are composed dominantly of cellulose (15-40% dry weight), pectic polysaccharides (30-50%), and xyloglucans (20-30%)[S136]. In grass species, however, primary cell walls contain arabinoxylans and

mixed-linkage glucans[S136]. In contrast, the secondary cell walls are formed in growth-ceased mature cells and are laid down on the inside of the primary wall. The secondary wall is typically composed of cellulose (35%-45% dry weight in grasses, 45%-50% in dicots), xylans (35%-45% in grasses, 45%-50% in dicots) and lignin (35%-45% in grasses, 45%-50% in dicots), providing rigidity and strength to the plant cells[S137]. A third pectin-rich layer, called the middle lamella, is formed at cytokinesis, and it makes up the outer layer of the wall, cementing cells together[S138]. Among the GH families in *S. hermonthica*, 12 families (GH1, GH3, GH5, GH9, GH10, GH16, GH17, GH18, GH19, GH28, GH35 and GH79) have at least 2 upregulated contigs during infection. The family containing the highest number of contigs is GH28 (Figure 5 and Figure E.6), a family encoding polygalacturonases that degrade pectin-derived polygalacturonan. Consistently, the carbohydrate esterase (CE) 8, which de-methyl esterifies the pectin resulting in a polygalacturonase susceptible form, increases its expression preceding GH28 (Figure 5 and Figure E.6B).

The top10 highly expressed contigs in each stage are from the auxiliary activities (AA) family. For example, AA2 and AA7 classes were highly expressed at 3 d and 7 d stages, respectively. The AA2 class contains peroxidases, some of which function in lignin degradation[S139], but our phylogenetic analysis indicates that the highly expressed AA2 proteins are class III peroxidases (Figure E.7) that are involved in various biotic or abiotic stress responses and in developmental processes including lignification[S140].

A

Top 10 highly expressed CAZyme genes in each stage

Seedling			1 d			3 d			7 d		
CAZy class	contig id	FPKM	CAZy class	contig id	FPKM	CAZy class	contig id	FPKM	CAZy class	contig id	FPKM
AA2	Sh14Contig_17836	592.45	GH19	Sh14Contig_18268	1089.73	CBM43	Sh14Contig_14886	2316.59	AA7	Sh14Contig_49223	2665.69
AA2	Sh14Contig_12100	585.52	AA2	Sh14Contig_12100	1029.86	AA2	Sh14Contig_22273	902.98	AA7	Sh14Contig_38072	2258.40
GH16	Sh14Contig_3854	444.04	GH16	Sh14Contig_9437	772.60	AA2	Sh14Contig_20298	603.37	AA2	Sh14Contig_32201	1526.44
AA1	Sh14Contig_16671	438.37	AA2	Sh14Contig_7471	769.68	GH19	Sh14Contig_18268	425.56	AA2	Sh14Contig_10092	1447.08
AA6	Sh14Contig_25657	435.55	AA2	Sh14Contig_20298	734.31	AA2	Sh14Contig_6979	404.80	AA7	Sh14Contig_4974	1443.76
GH16	Sh14Contig_9437	427.39	CBM43	Sh14Contig_14886	732.33	AA6	Sh14Contig_16015	353.96	CE16	Sh14Contig_38107	1262.94
AA1	Sh14Contig_19739	418.32	GH19, CBM18	Sh14Contig_20722	729.63	AA2	Sh14Contig_26853	340.39	CE1	Sh14Contig_25970	1254.78
GT75	Sh14Contig_1878	409.91	GH19, CBM18	Sh14Contig_16850	720.55	AA2	Sh14Contig_11737	317.84	CE16	Sh14Contig_1153	755.60
CE16	Sh14Contig_2808	397.77	GT75	Sh14Contig_1878	687.81	GH19, CBM18	Sh14Contig_20722	315.88	GH35	Sh14Contig_32472	584.55
AA2	Sh14Contig_13316	395.35	CE16	Sh14Contig_10983	672.97	AA2	Sh14Contig_10092	286.94	GH28	Sh14Contig_834	494.62

B

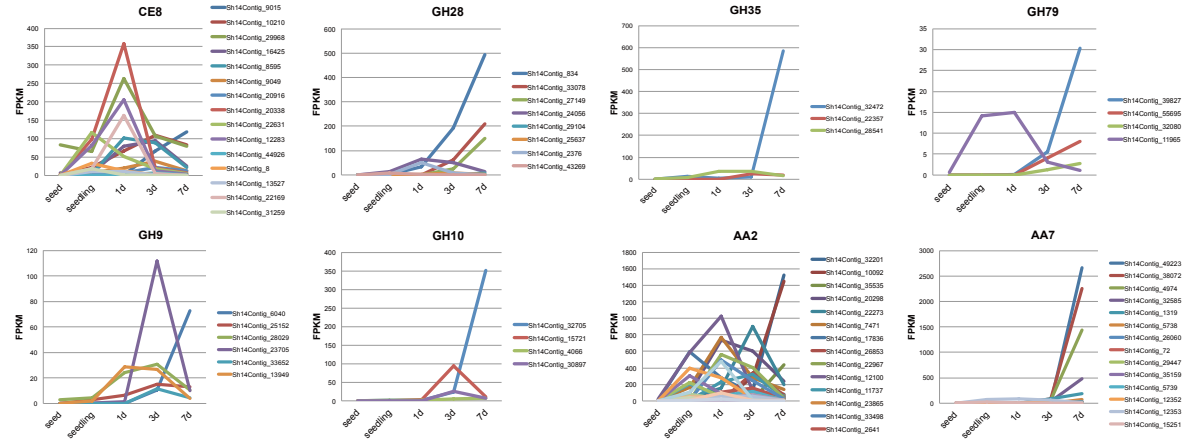


Figure E.6. Expression patterns of upregulated CAZyme-encoding contigs.
A. Top10 Highly expressed contigs classified into CAZyme class. The normalized FPKM values are shown. **B.** The charts showing expression patterns calculated from RNA-seq data. CE8, pectin methylesterase, GH28, polygalacturonase, GH35, β-galactosidase and GH79, β-glucuronidase, GH9, endo-1,3-β-glucanase, GH10, xylanase, AA2, peroxidase, and AA7, oligosaccharide oxidase family.

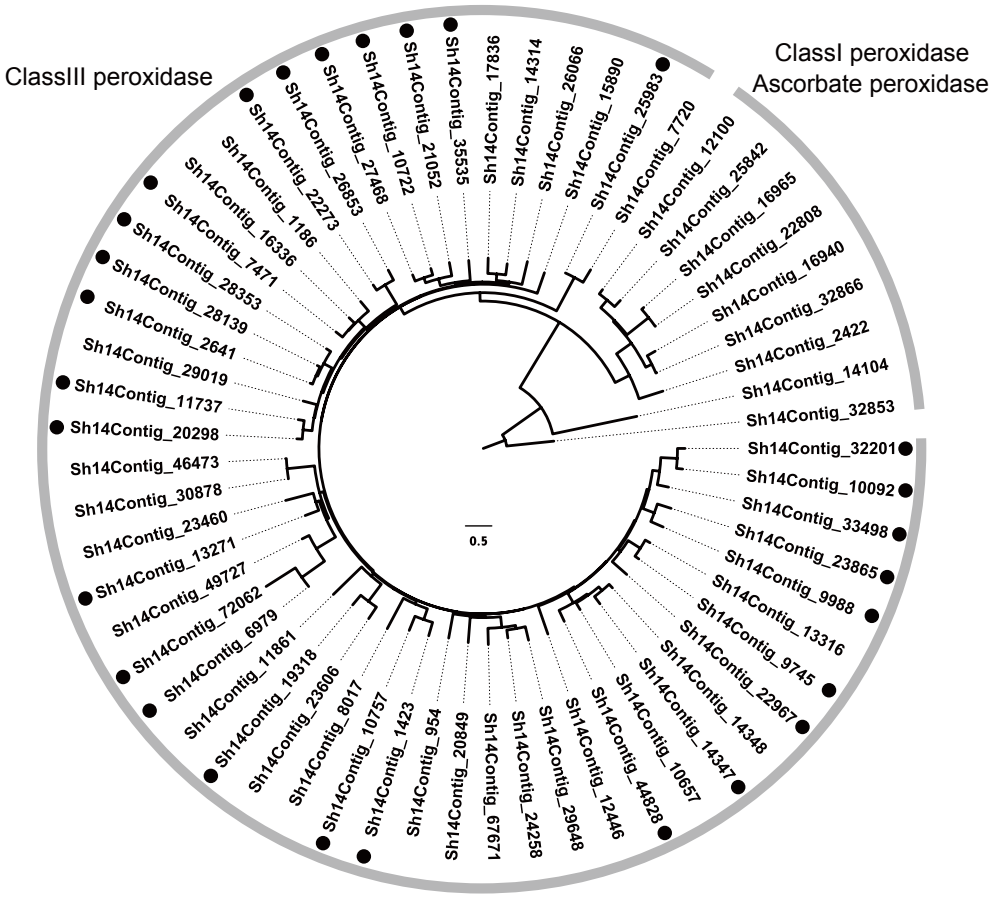


Figure E.7. Phylogenetic tree of AA2 family proteins.
Phylogenetic tree of *S. hermonthica* AA2 family proteins. The AA2 family proteins are classified into two clades, class I and class III peroxidases. All upregulated *S. hermonthica* proteins belong to class III peroxidase. Upregulated contigs are highlighted with dots.

E.9 Lateral root development genes

Because of developmental and morphological similarities between haustorium of parasitic plants and lateral roots, we proposed a hypothesis that parasitic plants might have recruited a lateral root developmental program to form a haustorium. We determined *S. hermonthica* orthologues related to lateral root development in *Arabidopsis* from the literature[S141]. Out of the 24 lateral root developmental genes in *Arabidopsis*, we identified 17 orthologues in the *S. hermonthica* transcriptome, corresponding to 50 contigs (Data S1Q). The orthologous genes in *S. asiatica* genome were searched by *InParanoid*4.1[S142] and tBLASTn search and all genes were found. The expression of the orthologues of the lateral root related genes was confirmed by RT-qPCR in seedling, 3-d and 7-d post infection stages using SaRPS2 as an internal control (Figure S3). Similar to *S. hermonthica* RNA-seq results, the *S. asiatica* LRD genes were upregulated during host penetration stages.

F. Horizontal gene transfer

F.1 Horizontally-transferred genes

For searching horizontally transferred genes in *S. asiatica* genome from grass host species, the *S. asiatica* annotation was subjected to BLASTp search with threshold e-value 1e-10 against a database of combined predicted proteins from the genome of 28 different plant species, including *Striga* host plants, rice, sorghum, foxtail millet and maize. The *S. asiatica* proteins that have at least one hit to grass species in their top 20 hits are selected, and modified Alien Index (AI) values[S143] were calculated as below formula. Modified AI = log((Best E-value for dicots) + 1e-200) - log((Best E-value for grasses) + 1e-200). The genes that have modified AI>30 and genes that do not have dicot hit are selected for further analysis. Maximum-likelihood phylogenetic trees were drawn by RAxML program with blast hit homolog genes from 28-species database as well as non-redundant (nr) database. Manual investigation of phylogenetic trees found 34 positive HGT candidate genes, which can be assigned into 20 orthogroups by orthoMCL analysis (Data S1R). These candidate genes are located in scaffolds with moderate coverage rates after mapping of the genome short reads, suggesting these genes are located in scaffold encoding nuclear genes (Data S1R). A few HGT candidates are closely located in the genome, and therefore the genomic regions were compared using CoGE with GEvo function. The gene scaffold555T52903, a homolog of the *ShContig9483* gene which were previously reported as HGT[S144], and scaffold555T52910, homologue of Alanyl-tRNA synthetase (Figure 7) are located in 30 kb region in the *S. asiatica* genome. The genomic region shows similarities to *Panicum hallii* chromosome 3, *Setalia italica* scaffold 3 and weak similarity to *S. bicolor* scaffold 3, suggesting that the conserved region among grass species were transferred into the parasite genome. The sequence similarities were observed in intron and untranslated regions, but not intergenic regions. It may suggest that transfer of gene-coding region or alternatively loss of conservation in intergenic region possibly due to selection pressures. Two other genomic regions are found to contain multiple HGT candidate genes (Figures F.1 and F.2). In addition, the genes similar to Pong transposon are frequently found as HGT genes, suggesting transposon transfer between host and parasites (Figure F.3).

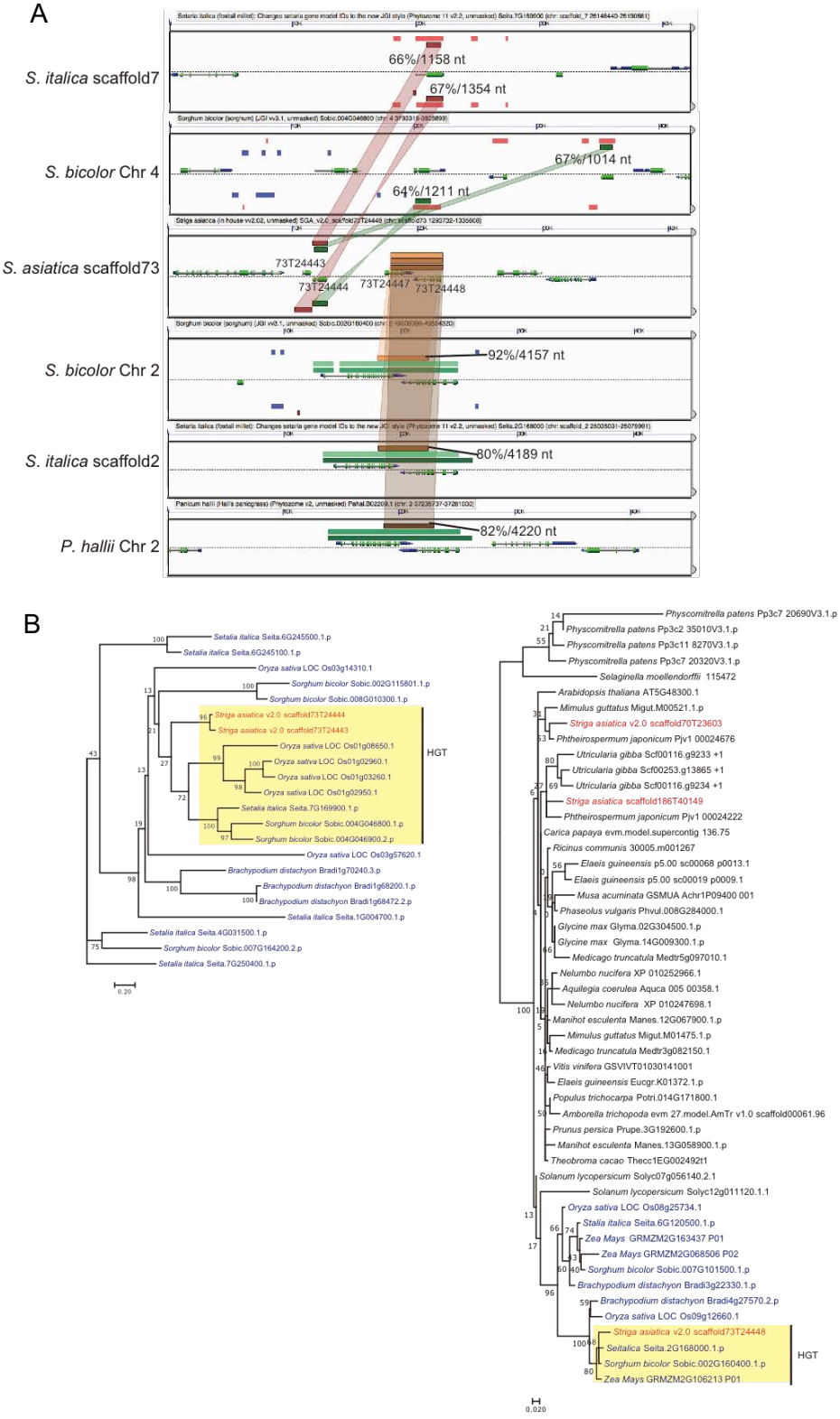


Figure F.1. HGT of genomic region scaffold73
A. Genomic comparison among *S. asiatica* scaffold 73, *S. italica* scaffold2 and 7, S. bicolor Chr2 and 4 and P. hallii Chr2. Regions with similarity detected by BlastZ (score 10000<) are highlighted by colors. **B.** Phylogenetic trees of HGT genes in scaffold73. The nodes including HGT events are highlighted with yellow.

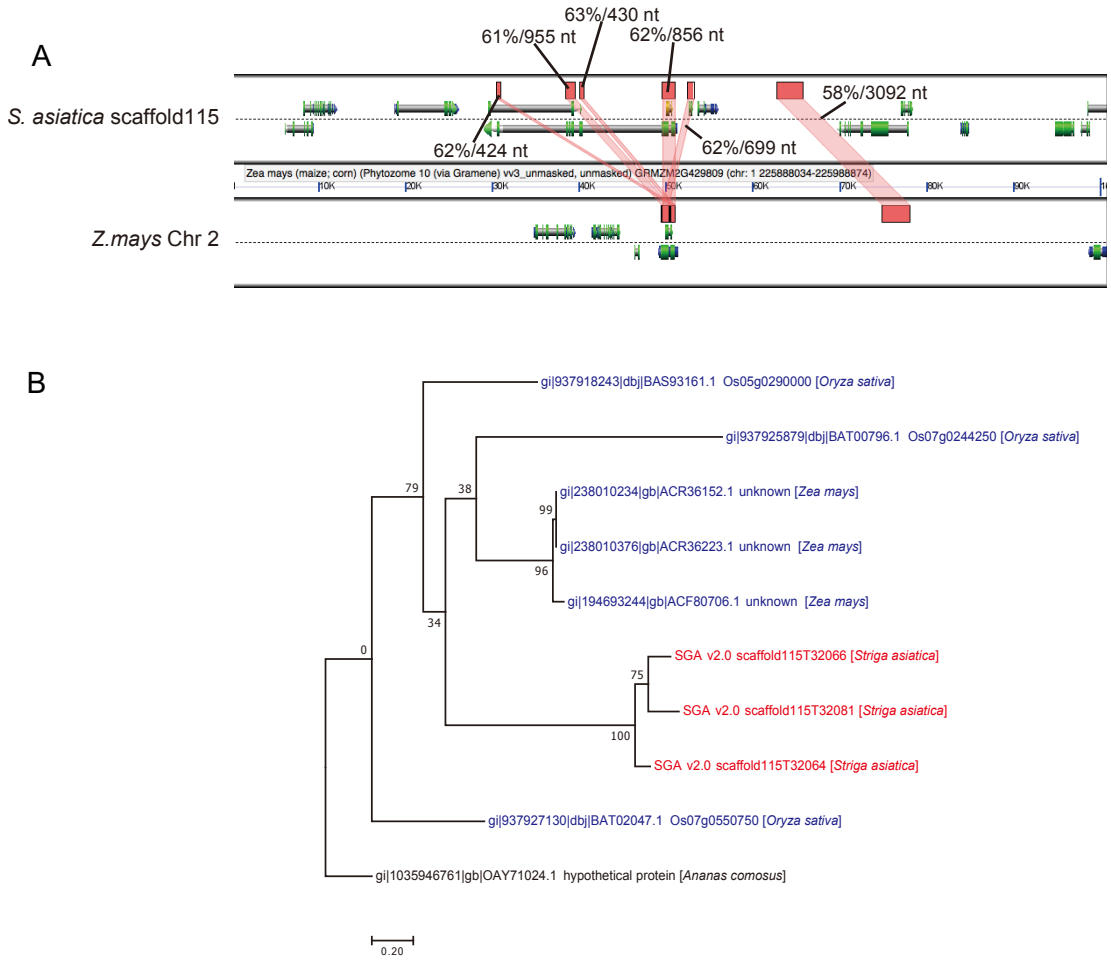


Figure F.2. Horizontal transfer of genomic region.
A. Genomic comparison between *S. asiatica* scaffold115 and *Z. mays* Chr2. **B.** The phylogenetic tree of HGT genes on saffold115.

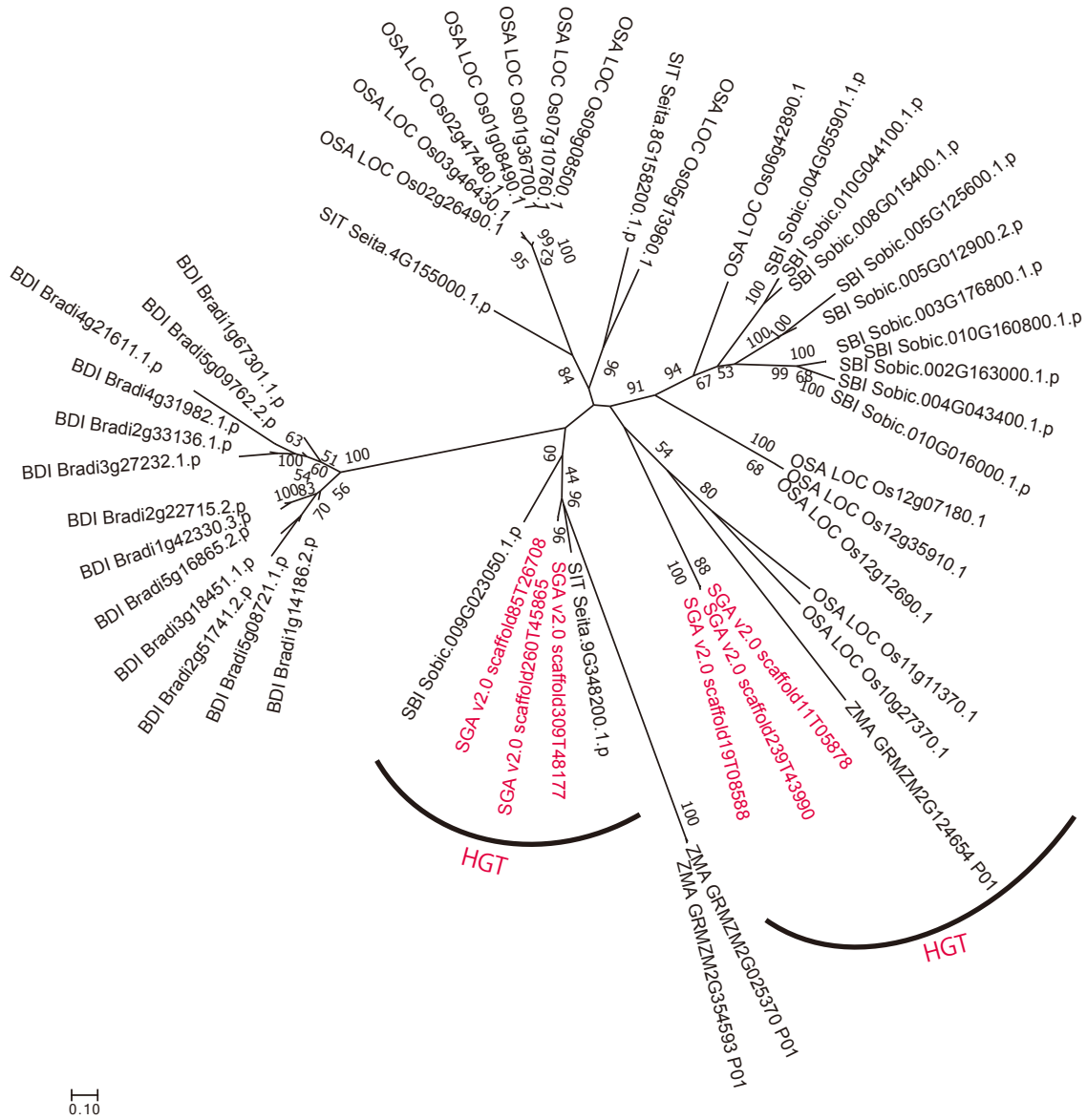


Figure F.3. HGT of a transposon family.
Phylogenetic tree of transposons that are obtained from grass species, having similarity to Pong transposon.

F.2 Horizontally transferred retrotransposons

The LTR retrotransposons and their *Gypsy* and *Copia* superfamilies are ubiquitous in fungi, plants, and animals and therefore appear to predate their divergence ~1500 MYA[S145]. Given the effects of genetic isolation and the mutagenic nature of retrotransposon replication[S146–S148], cladistic analysis generally separates retrotransposon families along organismal species lines, consonant with the vertical passage of these elements from the last common ancestor. When inter-species trees of retrotransposons contain branches in a phylogenetically inconsistent position, the horizontal transfer of an ancestral

element from one phylogenetic branch to another is often posited. Horizontal transfers of plant retrotransposons have been suggested in several specific cases[13, 135–138], but elsewhere the evidence has been equivocal[S153].

We carried out exhaustive alignments and phylogenetic analyses of reverse transcriptase (*rt*) domains, including 35,690 from *Copia* and 54,973 from *Gypsy* elements, retrieved from the genome sequences of *S. asiatica* and those of the monocots *Sorghum bicolor*, *Zea mays*, *Oryza sativa* ssp. *japonica* and ssp. *indica*, *O. rufipogon*, and *O. glaberrima* and the eudicots *Glycine max*, *S. tuberosum*, and *Vitis vinifera*. The *rt* sequences from the *Copia* and *Gypsy* superfamilies were analysed separately, producing 221 and 151 clusters respectively, of which 12 and 3 contained *S. asiatica rt* sequences mixed with those of other genomes. The *rt* domains of candidate elements were further characterised by exonerate-search[S154] using known *rt* sequences from GypsyDB[S155]. Resulting *rt* fragments were clustered by homology search against each other (BLASTn -evalue 1e-20) and subsequently clustered by silix-software[142] (silix -i 0.60 -r 0.70). The resulting clusters were aligned with the clustal-omega[143] and prank-ms[144] multiple aligners. Phylogenetic trees were constructed by FastTree (fasttree –nt –gtr –gamma)[S156]. Ages of LTR retrotransposons containing both LTRs were made as previously[S157]; a clock of 1.3×10^{-8} changes nt^{-1} $year^{-1}$ was used.

As alternatives to horizontal transfer, incomplete sampling or spotty evolutionary retention of retrotransposon groups can be invoked[S153]; in both cases, ancient, conserved, and widespread lineages that passed vertically could appear to be phylogenetically disjunct single representatives. However, regarding sampling, *rt* sequences are fairly easy to identify and are well represented in plant genome assemblies. Given the high number of *rt* sequences sampled, clustered, and aligned, both spotty retention and unusually high conservation for the several cases of apparent horizontal transfer would be needed to discount the examples given. Interestingly, these and previously reported horizontal transfers all involve elements of the superfamily *Copia* (Figure 7E, Figure S4). One possible explanation is that extant *Gypsy* elements, as reported here for *S. asiatica*, are older than those of *Copia*; *Gypsy* transfers may therefore have been lost from the genome already.

G. Supplemental References

S1. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956–60.

S2. Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics *27*, 764–70.

S3. Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116.

S4. The potato genome sequencing consortium (2011). Genome sequence and analysis of the tuber crop potato. Nature *475*, 189–195.

S5. Hamilton, J.P., and Robin Buell, C. (2012). Advances in plant genome sequencing. Plant J. *70*, 177–190.

S6. Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–63.

S7. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.* (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res *20*, 265–272.

S8. Boetzer, M., Henkel, C. V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

S9. Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., *et al.* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. 1384–1395.

S10. Hernandez, D., François, P., Farinelli, L., Østerås, M., and Schrenzel, J. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Res. *18*, 802–809.

S11. Schmutz, J., Mcclean, P.E., Mamidi, S., Wu, G.A., Cannon, S.B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., *et al.* (2014). A reference genome for common bean and genome-wide analysis of dual domestications. Nat. Publ. Gr. *46*, 707–713.

S12. Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548.

S13. Hawkins, J.S., Proulx, S.R., Rapp, R. a, and Wendel, J.F. (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proc. Natl. Acad. Sci. U. S. A. *106*, 17811–17816.

S14. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.* (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

S15. Thomas, J., and Pritham, E.J. (2015). Helitrons , the Eukaryotic Rolling-circle Transposable Elements. 1–32.

S16. Heitkam, T., Holtgräwe, D., Dohm, J.C., Minoche, A.E., Himmelbauer, H., Weisshaar, B., and Schmidt, T. (2014). Profiling of extensively diversified plant LINEs reveals distinct plant-specific subclades. Plant J. *79*, 385–397.

S17. Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., and Schulman, A.H. (2004). Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics *166*, 1437–1450.

S18. Gao, D., Li, Y., Kim, K. Do, Abernathy, B., and Jackson, S.A. (2016). Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome Biol., 1–17.

S19. Initiative, T.I.B. (2010). Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature *463*, 763–768.

S20. Schulman, A.H. (2013). Retrotransposon replication in plants. Curr. Opin. Virol. *3*, 604–614.
S21. Baidouri, M. El, and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol. Evol. *5*, 954–965.
S22. Timmis, J.N., Ayliff, M.A., Huang, C.Y., and Martin, W. (2004). Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. *5*, 123–135.
S23. Kapitonov, V. V, and Jurka, J. (2007). A universal classification of eukaryotic transposable elements implemented ...: Univerzity Karlovy - UKAŽ. *181*, 2006–2007.
S24. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 4–9.
S25. Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.
S26. 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.* (2013). MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. Plant Physiol. *164*, 513–524.
S27. Frith, M.C., Hamada, M., and Horton, P. (2010). Parameters for accurate genome alignment. BMC Bioinformatics *11*, 80.
S28. Han, Y., and Wessler, S.R. (2010). MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, 1–8.
S29. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.
S30. Steinbiss, S., Willhoeft, U., Gremme, G., and Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. *37*, 7002–7013.
S31. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., and Wu, C.H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics *31*, 926–932.
S32. Yoshida, S., Ishida, J.K., Kamal, N.M., Ali, A.M., Namba, S., and Shirasu, K. (2010). A full-length enriched cDNA library and expressed sequence tag analysis of the parasitic weed, Striga hermonthica. BMC Plant Biol. *10*, 55.
S33. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
S34. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D. a, Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–52.
S35. Yandell, M., and Holt, C. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.
S36. Westwood, J.H., dePamphilis, C.W., Das, M., Fernández-Aparicio, M., Honaas, L.A., Timko, M.P., Wafula, E.K., Wickett, N.J., and Yoder, J.I. (2012). The Parasitic Plant Genome Project: New Tools for Understanding the Biology of Orobanche and Striga. Weed Sci. *60*, 295–306. Available at: https://www.cambridge.org/core/product/identifier/S0043174500021330/type/journal_article.
S37. Yang, Z., Wafula, E.K., Honaas, L.A., Zhang, H., Das, M., Fernandez-aparicio, M., Huang, K., Bandaranayake, P.C.G., Wu, B., Der, J.P., *et al.* (2014). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. Mol. Biol. Evol.
S38. Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., *et al.* (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. *40*, 1178–1186.
S39. Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., *et al.* (2017). UniProt: The universal protein

knowledgebase. Nucleic Acids Res. *45*, D158–D169.
S40. Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.
S41. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: A b initio prediction of alternative transcripts. Nucleic Acids Res. *34*, 435–439.
S42. Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy to use annotation pipeline designed for emerging model organism genomes. Genome Res. *18*, 188–96.
S43. Ramírez-Sánchez, O., Pérez-Rodríguez, P., Delaye, L., and Tiessen, A. (2016). Plant Proteins Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins. Genomics, Proteomics Bioinforma. *14*, 357–370.
S44. Eilbeck, K., Moore, B., Holt, C., and Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics *10*, 1–15.
S45. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., *et al.* (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. Nucleic Acids Res. *40*, 1202–1210.
S46. McDowall, J., and Hunter, S. (2011). InterPro Protein Classification BT - Bioinformatics for Comparative Proteomics. In, C. H. Wu and C. Chen, eds. (Totowa, NJ: Humana Press), pp. 37–47.
S47. Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. *16*, 1–14.
S48. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.
S49. Wall, P.K., Leebens-Mack, J., Müller, K.F., Field, D., Altman, N.S., and Depamphilis, C.W. (2008). PlantTribes: A gene and gene family resource for comparative genomics in plants. Nucleic Acids Res. *36*, 970–976.
S50. Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., *et al.* (2011). Ancestral polyploidy in seed plants and angiosperms. Nature *473*, 97–100.
S51. Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H., Jin, H., Marler, B., Guo, H., *et al.* (2012). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. *40*, 1–14.
S52. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
S53. McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. J. Stat. Softw. *4*, 1–14.
S54. Tang, H., and Lyons, E. (2012). Unleashing the Genome of Brassica Rapa. Front. Plant Sci. *3*, 1–12.
S55. Haas, B.J., Delcher, A.L., Wortman, J.R., and Salzberg, S.L. (2004). DAGchainer: A tool for mining segmental genome duplications and synteny. Bioinformatics *20*, 3643–3646.
S56. Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., *et al.* (2013). Architecture and evolution of a minute plant genome. Nature *498*, 94–8.
S57. Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., Zheng, C., Alberti, A., Anthony, F., Aprea, G., *et al.* (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science *345*, 1181–1184.
S58. Pedersen, B.S., Tang, H., and Freeling, M. (2011). Gobe: An interactive, web-based tool for comparative genomic visualization. Bioinformatics *27*, 1015–1016.
S59. Freeling, M., Woodhouse, M.R., Subramaniam, S., Turco, G., Lisch, D., and Schnable, J.C. (2012). Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr. Opin. Plant Biol. *15*, 131–9.

S60. Searcy, D.G., and MacInnis, A.J. (1970). Measurements by DNA renaturation of the genetic basis of parasitic reduction. Evolution (N. Y). *24*, 796–806.
S61. Searcy, D.G. (1970). Measurements by DNA Hybridization in vitro of the Genetic Basis of Parasitic Reduction. Evolution (N. Y). *24*, 207–219.
S62. Ames, R.M., Money, D., Ghatge, V.P., Whelan, S., and Lovell, S.C. (2012). Determining the evolutionary history of gene families. Bioinformatics *28*, 48–55.
S63. 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*, 4–11. Available at: http://dx.doi.org/10.1038/s41467-018-04721-8.
S64. dePamphilis, C.W. (1995). Genes and genomes. In Parasitic Plants, M. Press and J. Graves, eds. (Chapman and Hall).
S65. Huang, D.W., Sherman, B.T., and Lempicki, R. a (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.
S66. Zimmermann, P., Bleuler, S., Laule, O., Martin, F., Ivanov, N. V, Campanoni, P., Oishi, K., Lugon-moulin, N., Wyss, M., Hruz, T., *et al.* (2014). ExpressionData - A public resource of high quality curated datasets representing gene expression across anatomy , development and experimental conditions. BioData Min. *7*, 18.
S67. Severin, A.J., Woody, J.L., Bolon, Y.-T., Joseph, B., Diers, B.W., Farmer, A.D., Muehlbauer, G.J., Nelson, R.T., Grant, D., Specht, J.E., *et al.* (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol. *10*, 160.
S68. Westwood, J. (2013). The physiology of the established parasite-host association. In Parasitic Orobanchaceae, D. M. Joel, J. Gressel, and L. J. Musselman, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg).
S69. 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*.
S70. Ravin, N. V., Gruzdev, E. V., Beletsky, A. V., Mazur, A.M., Prokhortchouk, E.B., Filyushin, M.A., Kochieva, E.Z., Kadnikov, V. V., Mardanov, A. V., and Skryabin, K.G. (2016). The loss of photosynthetic pathways in the plastid and nuclear genomes of the non-photosynthetic mycoheterotrophic eudicot Monotropa hypopitys. BMC Plant Biol. *16*.
S71. Rogers, W.E., and Nelson, R.R. (1962). Penetration and nutrition of *Striga asiatica*. Phytopathology *52*, 1064–1070.
S72. Těšitel, J., Plavcová, L., and Cameron, D.D. (2010). Interactions between hemiparasitic plants and their hosts: The importance of organic carbon transfer. Plant Signal. Behav. *5*, 1072–1076.
S73. Wicke, S. (2013). Genomic Evolution in Orobanchaceae. In Parasitic Orobanchaceae: Parasitic Mechanisms and Control Strategies (Berlin, Heidelberg: Springer Berlin Heidelberg).
S74. Wickett, N.J., Honaas, L.A., Wafula, E.K., Das, M., Huang, K., Wu, B., Landherr, L., Timko, M.P., Yoder, J., Westwood, J.H., *et al.* (2011). Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. Curr. Biol. *21*, 2098–104.
S75. Press, M.C., Smith, S., and Stewart, G.R. (1991). Carbon Acquisition and Assimilation in Parasitic Plants. Funct. Ecol. *5*, 278–283.
S76. Tuohy, J., Smith, E.A., and Stewart, G.R. (1986). The parasitic habit: trends in morphological and ultrastructural reductionism.
S77. Shah, N., Smirnoff, N. and Stewart, G.R. (1987). Photosynthesis and stomatal characteristics of Striga hermonthica in relation to its parasitic habit. Physiol. Plant. *69*, 699–703.
S78. Press, M.C., Tuohy, J.M., and Stewart, G.R. (1987). Gas exchange characteristics of the sorghum-striga host-parasite association. Plant Physiol. *84*, 814–819.
S79. Smith, S., and Stewart, G.R. (1990). Effect of Potassium Levels on the Stomatal Behavior of the Hemi-Parasite Striga hermonthica. Plant Physiol. *94*, 1472–1476.
S80. Singh, R., Singh, S., Parihar, P., Singh, V.P., and Prasad, S.M. (2015). Retrograde signaling

between plastid and nucleus: A review. J. Plant Physiol. *181*, 55–66.
S81. Ehleringer, J.R., and Marsahall, J.D. (1995). No Title. In Parasitic Plants, M. C. Press and J. D. Graves, eds. (Chapman and Hall, London).
S82. Dorr, I. (1997). HowStriga Parasitizes its Host: a TEM and SEM Study. Ann Bot *79*, 463–472.
S83. Heide-Jorgensen, H.S., and Kuijt, J. (1993). Epidermal derivatives as xylem elements and transfer cells : a study of the host-parasite interface in two species of *Triphysaria* (Scrophulariaceae). Protoplasma *174*, 173–183.
S84. Heide-Jorgensen, H.S., and Kuijt, J. (1995). The Haustorium of the Root Parasite Triphysaria (Scrophulariaceae), with Special Reference to Xylem Bridge Ultrastructure. Am. J. Bot. *82*, 782–797.
S85. Tomilov, A.A., Tomilova, N.B., Wroblewski, T., Michelmore, R., and Yoder, J.I. (2008). Trans-specific gene silencing between host and parasitic plants. Plant J *56*, 389–397.
S86. Aly, R., Hamamouch, N., Abu-Nassar, J., Wolf, S., Joel, D.M., Eizenberg, H., Kaisler, E., Cramer, C., Gal-On, A., and Westwood, J.H. (2011). Movement of protein and macromolecules between host plants and the parasitic weed Phelipanche aegyptiaca Pers. Plant Cell Rep. *30*, 2233–2241.
S87. Kim, G., LeBlanc, M.L., Wafula, E.K., DePamphilis, C.W., and Westwood1, J.H. (2014). Genomic-scale exchange of mRNA between a parasitic plant and its hosts. Science *345*, 808–811.
S88. Pallas, V., and Gómez, G. (2013). Phloem RNA-binding proteins as potential components of the long-distance RNA transport system. Front. Plant Sci. *4*, 1–6.
S89. Hayashi, K. -i. (2012). The interaction and integration of auxin signaling components. Plant Cell Physiol. *53*, 965–975.
S90. Werner, T., Köllmer, I., Bartrina, I., Holst, K., and Schmülling, T. (2006). New insights into the biology of cytokinin degradation. Plant Biol. *8*, 371–381.
S91. Shah, N., Smirnoff, N. and Stewart, G.R., Shah, N., Smirnoff, N., and Stewart, G.R. (1987). Photosynthesis and stomatal characteristics of Striga hermonthica in relation to its parasitic habit. Physiol. Plant. *69*, 699–703.
S92. Fujioka, H., Samejima, H., Suzuki, H., Mizutani, M., Okamoto, M., and Sugimoto, Y. (2019). Aberrant protein phosphatase 2C leads to abscisic acid insensitivity and high transpiration in parasitic Striga. Nat. Plants *5*, 258–262.
S93. Toh, S., Kamiya, Y., Kawakami, N., Nambara, E., McCourt, P., and Tsuchiya, Y. (2012). Thermoinhibition uncovers a role for strigolactones in Arabidopsis seed germination. Plant Cell Physiol *53*, 107–117.
S94. Taylor, a, Martin, J., and Seel, W.E. (1996). Physiology of the parasitic association between maize and witchweed (Striga hermonthica): is ABA involved? J. Exp. Bot. *47*, 1057–1065.
S95. Kanno, Y., Hanada, A., Chiba, Y., Ichikawa, T., Nakazawa, M., Matsui, M., Koshiba, T., Kamiya, Y., and Seo, M. (2012). Identification of an abscisic acid transporter by functional screening using the receptor complex as a sensor. Proc. Natl. Acad. Sci. *109*, 9653–9658.
S96. Park, S., Fung, P., Nishimura, N., Jensen, D.R., Fujii, H., Zhao, Y., Lumba, S., Santiago, J., Rodrigues, A., Chow, T.F., *et al.* (2009). Abscisic Acid Inhibits Type 2C Protein Phosphatases via the PYR/PYL Family of START Proteins. Science *324*, 1068–1069.
S97. Yue Ma, Izabela Szostkiewicz, Arthur Korte, Danièle Moes, Y.Y., and Alexander Christmann, E.G. (2009). Regulators of PP2C phosphatase activity function as abscisic acid sensors. Science *209*, 1064–1069.
S98. Fujita, Y., Yoshida, T., and Yamaguchi-Shinozaki, K. (2013). Pivotal role of the AREB/ABF-SnRK2 pathway in ABRE-mediated transcription in response to osmotic stress in plants. Physiol. Plant. *147*, 15–27.
S99. Hubbard, K.E., Nishimura, N., Hitomi, K., Getzoff, E.D., and Schroeder, J.I. (2010). Early abscisic acid signal transduction mechanisms : newly discovered components and newly emerging questions. Genes Dev *24*, 1695–1708.

S100. Vahisalu, T., Kollist, H., Wang, Y.-F., Nishimura, N., Chan, W.-Y., Valerio, G., Lamminmäki, A., Brosché, M., Moldau, H., Desikan, R., *et al.* (2008). SLAC1 is required for plant guard cell S-type anion channel function in stomatal signalling. Nature *452*, 487–491.
S101. Logan, D.C., and Stewart, G.R. (1991). Role of ethylene in the germination of the hemiparasite *Striga hermonthica*. Plant Physiol. *97*, 1435–1438.
S102. Iverson, R.D., Westbrooks, R.G., Eplee, R.E., and Tasker, A. V. (2011). Overview and status of the witchweed (Striga asiatica) eradication program in the Carolinas. In Invasive plant management issues and challenges in the United States : 2011 overview, A. R. L. and R. G.Westbrooks, ed. (Washington, DC), pp. 51–68.
S103. Merchante, C., Alonso, J.M., and Stepanova, A.N. (2013). Ethylene signaling : simple ligand , complex regulation. Curr. Opin. Plant Biol. *16*, 554–560. Available at: http://dx.doi.org/10.1016/j.pbi.2013.08.001.
S104. Qiao, H., Chang, K.N., Yazaki, J., and Ecker, J.R. (2009). Interplay between ethylene, ETP1/ETP2 F-box proteins, and degradation of EIN2 triggers ethylene responses in Arabidopsis. Genes Dev *23*, 512–521.
S105. Yang, C., Lu, X., Ma, B., Chen, S.-Y., and Zhang, J.-S. (2015). Ethylene signaling in rice and Arabidopsis: conserved and diverged aspects. Mol. Plant *8*, 495–505.
S106. Waters, M.T., Brewer, P.B., Bussell, J.D., Smith, S.M., and Beveridge, C.A. (2012). The Arabidopsis ortholog of rice DWARF27 acts upstream of MAX1 in control of plant development by strigolactones. Plant Physiol *159*, 1073–1085.
S107. Umehara, M., Hanada, A., Yoshida, S., Akiyama, K., Arite, T., Takeda-Kamiya, N., Magome, H., Kamiya, Y., Shirasu, K., Yoneyama, K., *et al.* (2008). Inhibition of shoot branching by new terpenoid plant hormones. Nature *455*, 195–200.
S108. Gomez-Roldan, V., Fermas, S., Brewer, P.B., Puech-Pages, V., Dun, E.A., Pillot, J.P., Letisse, F., Matusova, R., Danoun, S., Portais, J.C., *et al.* (2008). Strigolactone inhibition of shoot branching. Nature *455*, 189–194.
S109. Alder, A., Jamil, M., Marzorati, M., Bruno, M., Vermathen, M., Bigler, P., Ghisla, S., Bouwmeester, H., Beyer, P., and Al-Babili, S. (2012). The path from beta-carotene to carlactone, a strigolactone-like plant hormone. Science (80-.). *335*, 1348–1351.
S110. Zhang, Y., Dijk, A.D.J. Van, Scaffidi, A., Flematti, G.R., Hofmann, M., Charnikhova, T., Verstappen, F., Hepworth, J., Krol, S. Van Der, and Leyser, O. (2014). Rice cytochrome p450 maX1 homologs catalyze distinct steps in strigolactone biosynthesis. Nat. Chem. Biol. *10*, 1028–1033.
S111. Abe, S., Sado, A., Tanaka, K., Kisugi, T., Asami, K., Ota, S., Il, H., and Yoneyama, K. (2014). Carlactone is converted to carlactonoic acid by MAX1 in Arabidopsis and its methyl ester can directly interact with AtD14 in vitro.
S112. Brewer, P.B., Yoneyama, K., Filardo, F., Meyers, E., Scaffidi, A., Frickey, T., Akiyama, K., Seto, Y., Dun, E.A., Cremer, J.E., *et al.* (2016). *LATERAL BRANCHING OXIDOREDUCTASE* acts in the final stages of strigolactone biosynthesis in *Arabidopsis*. Proc. Natl. Acad. Sci. *113*, 6301–6306.
S113. Das, M., Fernández-aparicio, M., Yang, Z., Huang, K., Wickett, N.J., Alford, S., Wafula, E.K., Bouwmeester, H., Timko, M.P., Yoder, J.I., *et al.* (2015). Parasitic plants *Striga* and *Phelipanche* dependent upon exogenous strigolactones for germination have retained genes for strigolactone biosynthesis. Am. J. Plant Sci. *6*, 1151–1166.
S114. Stirnberg, P., Furner, I.J., and Ottoline Leyser, H.M. (2007). MAX2 participates in an SCF complex which acts locally at the node to suppress shoot branching. Plant J. *50*, 80–94.
S115. Arite, T., Umehara, M., Ishikawa, S., Hanada, A., Maekawa, M., Yamaguchi, S., and Kyozuka, J. (2009). *d14*, a strigolactone-insensitive mutant of rice, shows an accelerated outgrowth of tillers. Plant Cell Physiol *50*, 1416–1424.
S116. Waters, M.T., Scaffidi, A., Sun, Y.K., Flematti, G.R., and Smith, S.M. (2014). The karrikin response system of Arabidopsis. Plant J. *79*, 623–631.

S117. Zhou, F., Lin, Q., Zhu, L., Ren, Y., Zhou, K., Shabek, N., Wu, F., Mao, H., Dong, W., Gan, L., *et al.* (2013). D14-SCF(D3)-dependent degradation of D53 regulates strigolactone signalling. Nature *504*, 406–410.
S118. Jiang, L., Liu, X., Xiong, G., Liu, H., Chen, F., Wang, L., Meng, X., Liu, G., Yu, H., Yuan, Y., *et al.* (2013). DWARF 53 acts as a repressor of strigolactone signalling in rice. Nature *504*, 401–405.
S119. Stanga, J.P., Smith, S.M., Briggs, W.R., and Nelson, D.C. (2013). SUPPRESSOR OF MORE AXILLARY GROWTH2 1 controls seed germination and seedling development in Arabidopsis. Plant Physiol. *163*, 318–30.
S120. Soundappan, I., Bennett, T., Morffy, N., Liang, Y., Stanga, J.P., Abbas, A., Leyser, O., and Nelson, D.C. (2015). SMAX1-LIKE/D53 family members enable distinct MAX2-dependent responses to strigolactones and karrikins in Arabidopsis. Plant Cell *27*, 3143–3159.
S121. Yoshida, S., and Shirasu, K. (2009). Multiple layers of incompatibility to the parasitic witchweed, Striga hermonthica. New Phytol *183*, 180–189.
S122. Hirayama, K., and Mori, K. (1999). Synthesis of (+) -Strigol and (+) -Orobanchol , the germination stimulants, and their stereoisomers by employing lipase-catalyzed asymmetric acetylation as the key step. European J. Org. Chem. *1999*, 2211–2217.
S123. Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Res *9*, 868–877.
S124. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–9.
S125. Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics *21*, 3674–3676.
S126. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth *9*, 357–359.
S127. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.
S128. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.
S129. Chitwood, D.H., Kumar, R., Headland, L.R., Ranjan, A., Covington, M.F., Ichihashi, Y., Fulop, D., Jiménez-Gómez, J.M., Peng, J., Maloof, J.N., *et al.* (2013). A quantitative genetic basis for leaf morphology in a set of precisely defined tomato introgression lines. Plant Cell *25*, 2465–81.
S130. Wehrens, R., and Buydens, L.M.C. (2007). Self-and super-organizing maps in R: the Kohonen package. J. Stat. Softw. *21*, 19.
S131. Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.
S132. Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.
S133. Visser, J.H., Inge, D., and Kollmann, R. (1984). The “hyaline body” of the root parasite *Alectra orobanchoides* benth. (Scrophulariaceae)—Its anatomy, ultrastructure and histochemistry. Protoplasma *121*, 146–156.
S134. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. *42*, 490–495.
S135. Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). DbCAN: A web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. *40*, 445–451.
S136. Cosgrove, D.J., and Jarvis, M.C. (2012). Comparative structure and biomechanics of plant primary and secondary cell walls. Front. Plant Sci. *3*, 1–6.
S137. King, B.C., Waxman, K.D., Nenni, N. V, Walker, L.P., Bergstrom, G.C., and Gibson, D.M. (2011). Arsenal of plant cell wall degrading enzymes reflects host preference among plant

pathogenic fungi. Biotechnol. Biofuels *4*, 4. http://www.biotechnologyforbiofuels.com/content/4/1/4.
S138. Jarvis, M.C., Briggs, S.P.H., and Knox, J.P. (2003). Intercellular adhesion and cell separation in plants. Plant, Cell Environ. *26*, 977–989.
S139. Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M., and Henrissat, B. (2013). Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes. Biotechnol. Biofuels *6*, 41.
S140. Saathoff, A.J., Donze, T., Palmer, N. a, Bradshaw, J., Heng-Moss, T., Twigg, P., Tobias, C.M., Lagrimini, M., and Sarath, G. (2013). Towards uncovering the roles of switchgrass peroxidases in plant processes. Front. Plant Sci. *4*, 202.
S141. Lavenus, J., Goh, T., Roberts, I., Guyomarc'h, S., Lucas, M., De Smet, I., Fukaki, H., Beeckman, T., Bennett, M., and Laplaze, L. (2013). Lateral root development in *Arabidopsis*: Fifty shades of auxin. Trends Plant Sci. *18*, 1360–1385.
S142. O'Brien, K.P., Remm, M., and Sonnhammer, E.L.L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. *33*, 476–480.
S143. Gladyshev, E.A., Meselson, M., and Arkhipova, I.R. (2008). Massive Horizontal Gene Transfer in Bdelloid Rotifers. Science (80-.). *320*, 1210–1214.
S144. Miyawaki, K., Tarkowski, P., Matsumoto-Kitano, M., Kato, T., Sato, S., Tarkowska, D., Tabata, S., Sandberg, G., and Kakimoto, T. (2006). Roles of Arabidopsis ATP/ADP isopentenyltransferases and tRNA isopentenyltransferases in cytokinin biosynthesis. Proc. Natl. Acad. Sci. U. S. A. *103*, 16598–16603.
S145. Hedges, S.B. (2002). The origin and evolution of model organisms. Nat Rev Genet *3*, 838–849.
S146. Gabriel, A., Willems, M., Mules, E.H., and Boeke, J.D. (1996). Replication infidelity during a single cycle of Ty1 retrotransposition. Proc. Natl. Acad. Sci. U. S. A. *93*, 7767–7771.
S147. Boutabout, M., Wilhelm, M., and Wilhelm, F.X. (2001). DNA synthesis fidelity by the reverse transcriptase of the yeast retrotransposon Ty1. Nucleic Acids Res. *29*, 2217–2222.
S148. Abram, M.E., Ferris, A.L., Shao, W., Alvord, W.G., and Hughes, S.H. (2010). Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J. Virol. *84*, 9864–9878.
S149. Roulin, A., Piegu, B., Wing, R.A., and Panaud, O. (2008). Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus Oryza. Plant J *53*, 950–959.
S150. Cheng, X., Zhang, D., Cheng, Z., Keller, B., and Ling, H.Q. (2009). A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics *181*, 1183–1193.
S151. Sharma, A., and Presting, G.G. (2014). Evolution of centromeric retrotransposons in grasses. Genome Biol. Evol. *6*, 1335–1352.
S152. Roulin, A., Piegu, B., Fortune, P.M., Sabot, F., D'Hont, A., Manicacci, D., and Panaud, O. (2009). Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon Route66 in Poaceae. BMC Evol. Biol. *9*, 58.
S153. Moisy, C., Schulman, A.H., Kalendar, R., Buchmann, J.P., and Pelsy, F. (2014). The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years. Theor. Appl. Genet., 1–13.
S154. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.
S155. Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., *et al.* (2011). The Gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. *39*, 70–74.
S156. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One *5*.

S157. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat Genet *20*, 43–45.

Click here to access/download

Supplemental Videos and Spreadsheets

DataS1_sy.xlsx

Click here to access/download

Supplemental Videos and Spreadsheets

DataS3.csv

Declaration of Interests form

Click here to access/download

Supplemental Videos and Spreadsheets

DI_form.pdf