

# Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty

Zhenzhen Yang,<sup>1,2,3</sup> Eric K. Wafula,<sup>2,3</sup> Loren A. Honaas,<sup>1,2,3</sup> Huiting Zhang,<sup>1,2</sup> Malay Das,<sup>‡,4</sup> Monica Fernandez-Aparicio,<sup>§,4,5</sup> Kan Huang,<sup>¶,5</sup> Pradeepa C.G. Bandaranayake,<sup>\*\*6</sup> Biao Wu,<sup>††,6</sup> Joshua P. Der,<sup>‡‡,2,3</sup> Christopher R. Clarke,<sup>4</sup> Paula E. Ralph,<sup>2</sup> Lena Landherr,<sup>2</sup> Naomi S. Altman,<sup>7</sup> Michael P. Timko,<sup>5</sup> John I. Yoder,<sup>6</sup> James H. Westwood,<sup>4</sup> and Claude W. dePamphilis<sup>\*,1,2,3</sup>

<sup>1</sup>Intercollege Graduate Program in Plant Biology, Huck Institutes of the Life Sciences, The Pennsylvania State University

<sup>2</sup>Department of Biology, The Pennsylvania State University

<sup>3</sup>Institute of Molecular Evolutionary Genetics, Huck Institutes of the Life Sciences, The Pennsylvania State University

<sup>4</sup>Department of Plant Pathology, Physiology and Weed Science, Virginia Polytechnic Institute and State University

<sup>5</sup>Department of Biology, University of Virginia

<sup>6</sup>Department of Plant Sciences, University of California, Davis

<sup>7</sup>Department of Statistics and Huck Institutes of the Life Sciences, The Pennsylvania State University

<sup>‡</sup>Present address: Department of Biological Sciences, Presidency University, Kolkata, India

<sup>§</sup>Present address: UMR1347 Agroécologie, Institut National de la Recherche Agronomique (INRA), Dijon, France

<sup>¶</sup>Present address: Rutgers Business School, Rutgers University, New Brunswick-Piscataway, NJ

<sup>\*\*</sup>Present address: Agricultural Biotechnology Center, Faculty of Agriculture, University of Peradeniya, Peradeniya, Sri Lanka

<sup>††</sup>Present address: Molefarming Laboratory, Davis, CA

<sup>‡‡</sup>Present address: Department of Biological Science, California State University Fullerton

\*Corresponding author: E-mail: cwd3@psu.edu.

Associate editor: Michael Purugganan

## Abstract

The origin of novel traits is recognized as an important process underlying many major evolutionary radiations. We studied the genetic basis for the evolution of haustoria, the novel feeding organs of parasitic flowering plants, using comparative transcriptome sequencing in three species of Orobanchaceae. Around 180 genes are upregulated during haustorial development following host attachment in at least two species, and these are enriched in proteases, cell wall modifying enzymes, and extracellular secretion proteins. Additionally, about 100 shared genes are upregulated in response to haustorium inducing factors prior to host attachment. Collectively, we refer to these newly identified genes as putative “parasitism genes.” Most of these parasitism genes are derived from gene duplications in a common ancestor of Orobanchaceae and *Mimulus guttatus*, a related nonparasitic plant. Additionally, the signature of relaxed purifying selection and/or adaptive evolution at specific sites was detected in many haustorial genes, and may play an important role in parasite evolution. Comparative analysis of gene expression patterns in parasitic and nonparasitic angiosperms suggests that parasitism genes are derived primarily from root and floral tissues, but with some genes co-opted from other tissues. Gene duplication, often taking place in a nonparasitic ancestor of Orobanchaceae, followed by regulatory neofunctionalization, was an important process in the origin of parasitic haustoria.

**Key words:** novel traits, transcriptome, protease, parasitism, gene duplication, Orobanchaceae.

## Introduction

Throughout evolutionary history, organisms have evolved a variety of sophisticated novel traits for survival and reproduction. For instance, insects have evolved wings for flying, and plants have evolved different patterns of floral shapes and colors to maximize the attraction of insects and other animals for pollination. The origin of such novel traits has been of longstanding interest to evolutionary biologists, and a wide range of approaches has been used to gain insights into the origin of specific traits. For example, examination of the

sensory functions of cilia, the secretory structure of sponges, an early diverging group of multicellular animals, provided insights into the origin of the sensory system of metazoans (Ludeman et al. 2014). Phylogenetic histories of genes known to be involved in eye development and phototransduction revealed that a greater variety of eye types as found in pancrustacean arthropods, appeared to be associated with a higher rate of gene duplication (Rivera et al. 2010). Mutation and gene duplication have played an important role in generating complex pathways for refined eye

© The Author 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

development (Gehring 2011; Nei 2013), which resulted in many different eye types including the camera eye, the compound eye, and the mirror eye (Salvini-Plawen and Mayer 1961). The complete genome analysis of the basal angiosperm *Amborella trichopoda*, the sister species to all other extant flowering plants, revealed that a whole-genome duplication (WGD) led to the creation of many novel genes and functions associated with floral development and evolution, ultimately contributing to the diversification of flowering plants (*Amborella* Genome Project 2013).

As seen in the above case studies, gene duplication is frequently associated with the evolution of novel functions (Stephens 1951; Nei 1969; Ohno 1970; Kaessmann 2010; Liberles et al. 2010). The most extensively documented proposal for the evolution of novel gene function is the classic gene duplication model proposed by Ohno (1970) and extended by Force, Lynch, and many others (Force et al. 1999; Lynch and Conery 2000; Tirosh and Barkai 2007; Liberles et al. 2010). Following gene duplication, one copy may retain its original function, whereas the other copy diverges, and can have a variety of different fates, including pseudogenization (Lynch and Conery 2000), hypofunctionalization (Duarte et al. 2006), subfunctionalization, or neofunctionalization. Subfunctionalization is due to complementary loss of some of the functional attributes that are initially shared by the new paralogs following duplication, whereas neofunctionalization can occur when one of the paralogs evolves a new expression pattern or sequence attribute and acquires a new function (Force et al. 1999; Lynch and Conery 2000; Tirosh and Barkai 2007; Liberles et al. 2010). Subneofunctionalization was proposed to describe processes that involved both (He and Zhang 2005). Polyploidy duplicates all of the genes in the genome at once (Otto and Whitton 2000), providing ample opportunities for the function of paralogous gene copies to diverge (Crow and Wagner 2006; Nei 2013), especially in plants, where both angiosperms (Jiao et al. 2011; *Amborella* Genome Project 2013) and seed plants (Jiao et al. 2011) have been hypothesized to be ancestrally polyploid, and a large number of more recent polyploidy events have been detected (Schlueter et al. 2004; Cui et al. 2006; Soltis et al. 2009; Jiao et al. 2012; Vanneste et al. 2013). Thus, novel gene creation through single gene duplications, large-scale genome duplication, and neofunctionalization may all play significant roles in the origin of a novel function.

For plants, the evolution of parasitism is one of the most extraordinary examples of evolution of novel traits, as parasitic plants have evolved the ability to form a connection that allows it to feed off plants of other species, allowing some parasites to completely abandon photosynthesis, one of the hallmarks of life for most plants. Parasitism is enabled by specialized feeding structures known as haustoria (Kuijt 1969; Heide-Jørgensen 2013a), which have evolved independently at least 11 times in angiosperm evolution (Barkman et al. 2007; Westwood et al. 2010). Most haustorial parasitic plants invade host roots, whereas some are able to form haustorial connections with stems, and rarely, leaves. Parasitic plants can be classified as hemiparasites (if they retain some photosynthetic capability, and thus are at least

partly autotrophic) or holoparasites (if they are entirely heterotrophic). Furthermore, parasitic plants can be classified by their degree of host dependence. Facultative parasites must retain photosynthetic abilities and are opportunistic parasites that are able to complete their life cycle without attaching to a host (Kuijt 1969; Westwood et al. 2010; Naumann et al. 2013), whereas obligate parasites must form a host attachment in order to complete their life cycle. The Orobanchaceae species that are the focus of this work are unusual in that this is the only family of plants that spans the entire spectrum from facultative to obligate parasitism and from hemiparasites to holoparasites (Westwood et al. 2010), even including a basal nonparasitic lineage, and thus this family is an ideal group for investigating the evolution of parasitism.

Haustoria of parasitic plants transport water, minerals (including nitrogen), and often organic nutrients including host photosynthate. Bulk flow of water and dissolved nutrients from the vascular system of host plants is achieved by parasitic plants by maintaining a lower water potential relative to the host (Smith and Stewart 1990). Holoparasites acquire almost all of their fixed carbon from their hosts through phloem connections (Dorr and Kollmann 1995); sugar accumulated in haustoria ranges from 6- to 8-fold higher than in their hosts (Aber et al. 1983). The ability of the haustorium to efficiently transfer host resources results in substantial yield loss to several economically important crop plants. For example, in sub-Saharan Africa, witchweed (*Striga* spp.) infests over 50 million hectares of arable farmland cultivated with corn and legumes, causing annual yield loss estimated to exceed \$10 billion USD (Scholes and Press 2008). Furthermore, infestations of staple crops can result in substantial or near complete yield loss, exacerbating problems of low food security.

Identifying genes with key roles in parasitism may reveal novel strategies to control weedy agricultural pest species (Aly et al. 2009; Alakonya et al. 2012; Bandaranayake et al. 2012; Westwood et al. 2012; Bandaranayake and Yoder 2013b; Ranjan et al. 2014). Despite decades of research, only a few genes have been previously characterized with specific roles in the parasitic process in Orobanchaceae. One quinone oxidoreductase gene (*TvQR1*) is necessary for haustorium initiation through redox bioactivation of haustorial inducing factors (HIFs) in *Triphysaria*, a facultative root hemiparasite (Bandaranayake et al. 2010; Ngo et al. 2013). Additionally, *TvPirin* is upregulated by HIFs (or by contact with host roots) and putatively functions as a positive regulator of other genes needed for haustorial development (Bandaranayake et al. 2012). Finally, mannose 6-phosphate reductase (*M6PR*) in *Orobanche*, a root holoparasite, was also shown to be involved in parasite metabolism. Silencing of the parasite *M6PR* gene by RNAi from the host resulted in decreased mannitol concentration in the haustorium tubercle and increased tubercle mortality, thus clarifying the role of mannitol in parasitism (Aly et al. 2009).

Although not in the family of Orobanchaceae, experimental characterization also supports the role of two parasitism genes in *Cuscuta*: One encoding a cysteine protease in the stem parasite, *Cuscuta reflexa* (Bleischwitz et al. 2010), and a

*SHOOT MERISTEMLESS-Like (STM)* gene in *C. pentagona* (Alakonya et al. 2012). *STM* encodes a KNOTTED-like homeobox transcription factor (TF) with a known role in promoting cytokinin biosynthesis in the shoot apical meristem. Silencing of the *STM* gene in *Cuscuta* by the production of small RNA by host plants resulted in reduced haustorial development and increased growth of infected host plants (Alakonya et al. 2012).

Next generation transcriptome sequencing technologies provide great opportunities to use advanced, yet inexpensive, genome-scale tools to gain a better understanding of the molecular biology of parasitic plants (Westwood et al. 2010, 2012; Ranjan et al. 2014). A major focus of the Parasitic Plant Genome Project (PPGP; <http://ppgp.huck.psu.edu/>, last accessed December 18, 2014) is the identification of genes related to haustorial initiation and development through a comparative transcriptomics approach of multiple stages of parasite growth and development in Orobanchaceae, the youngest (estimated at 32 My) and one of the largest and most diverse extant lineages of parasitic plants (Naumann et al. 2013). We focus on three species of parasitic plants in Orobanchaceae that together represent the complete range of parasitic dependence: *Triphysaria versicolor* (a facultative hemiparasite), *Striga hermonthica* (an obligate hemiparasite), and *Phelipanche aegyptiaca* (syn. *Orobanche aegyptiaca*; an obligate holoparasite) (Westwood et al. 2010, 2012). Also included in this study was a single autotrophic lineage, the genus *Lindenbergia*, which is sister to the rest of the Orobanchaceae (Young et al. 1999). Overall, this project yielded nearly 3 billion sequence reads from more than 30 tissue- and stage-specific as well as normalized whole-plant libraries. Several published studies have already utilized the data generated by this project for focused analyses on a range of topics, including gene targets for host-induced gene silencing (Bandaranayake and Yoder 2013b) and herbicide control (Westwood et al. 2012); host-dependent parasite gene expression at the host–parasite interface (Honaas 2013); genes involved in strigolactone (Liu et al. 2014), karrakin (Nelson et al. 2011; Nelson 2013), and ABA pathways (Lechat et al. 2012) related to parasite seed germination; horizontal gene transfer in the parasites (Zhang, Cowled, et al. 2013; Zhang et al. 2014); evolutionary conservation and loss of genes involved in photosynthetic (Wickett et al. 2011) and symbiotic processes (Delaux et al. 2014); and housekeeping genes for quantitative polymerase chain reaction (qPCR) studies (Fernandez-Aparicio et al. 2013) and molecular markers for systematic research (Eizenberg et al. 2012). However, to date, no comprehensive analysis has been published of the entire data set to identify genes that are likely essential to the parasitic lifestyle and the origin of parasitic processes.

The origin of parasitism in plants has been proposed to follow two general mechanisms. The first considers the striking morphological similarity between some parasitic plant haustoria, root nodules, and crown galls; it was thus proposed that parasites may have evolved through endophytic association or horizontal gene transfer of genes from bacteria or other microorganisms that could confer parasitic ability (Atsatt 1973). The second mechanism, termed the


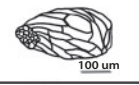
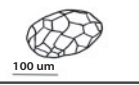

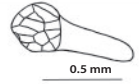
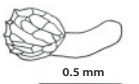

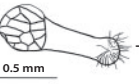

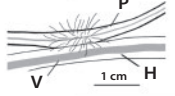
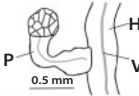
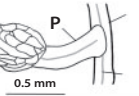
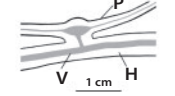
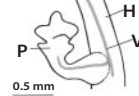
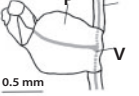
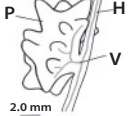
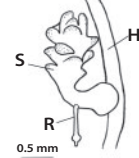
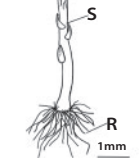

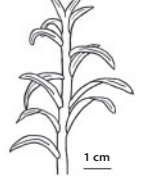




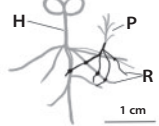
endogenous model (Bandaranayake and Yoder 2013a), was that parasitic functions may have evolved through neofunctionalization from plant genes encoding nonparasitic functions. These mechanisms are not necessarily mutually exclusive, and both may have been important to the evolution of parasitism.

In this study, we have focused on seeking evidence relevant to the endogenous model for the origin of parasitism. We utilized differential expression (DE) analysis and expression clustering to identify upregulated genes associated with haustorium initiation, development, and physiology. Through identification of a “core set of parasitism genes” shared by multiple species of parasites, our results also shed light on the evolutionary mechanism(s) that led to the origin of the haustorium in the Orobanchaceae. As the haustorium is a novel structure at the core of the parasitic process, comparative analysis of the genes and gene expression patterns of both parasitic and nonparasitic plants enabled us to propose its genetic origins.

## Results

### Assembly Statistics

For each of the three parasitic plants in this study (*T. versicolor*, *S. hermonthica*, and *P. aegyptiaca*), we generated 11–14 stage-specific libraries (Westwood et al. 2012), plus additional whole-plant normalized libraries using RNA from all developmental stages in each species (fig. 1). Additionally, a whole-plant normalized library of *Lindenbergia philippensis* was sequenced to represent the nonparasitic sister lineage of the parasites. A grand total of 2,995,494,710 Illumina reads and 3,153,353 Roche 454 GS-FLX reads were generated. Hybrid assemblies combining all sequencing data for each species resulted in unigene numbers ranging from 117,470 in *Striga* to 131,173 in *Triphysaria* (table 1). Average unigene length varied between 581 bp (*Triphysaria*) and 745 bp (*Striga*), whereas average N50 lengths ranged from 789 to 1,183 bp with the N50 unigene counts ranging from 21,356 to 24,729. To evaluate the completeness of our transcriptome sequence data sets, we examined the frequency of capture of three known conserved sets of plant genes in the transcriptome assemblies, namely the universally conserved orthologs (UCOs) (Kozik et al. 2008; Der et al. 2011; Williams et al. 2014), conserved single copy genes from COSII (Fulton et al. 2002; Wu et al. 2006; Williams et al. 2014), and the set of conserved single copy genes in PlantTribes2 (Wall et al. 2008) ([http://fgp.bio.psu.edu/tribedb/10\\_genomes/](http://fgp.bio.psu.edu/tribedb/10_genomes/), last accessed December 18, 2014). The UCO list was obtained from the Compositae genome project ([http://compgenomics.ucdavis.edu/compositae\\_reference.php](http://compgenomics.ucdavis.edu/compositae_reference.php), last accessed December 18, 2014) and the COSII gene list was obtained from SolGenomics (<http://solgenomics.net/documents/markers/cosii.xls>, last accessed December 18, 2014). The single copy gene list containing 970 single copy orthogroups from PlantTribes2.0 ([http://fgp.bio.psu.edu/tribedb/10\\_genomes/](http://fgp.bio.psu.edu/tribedb/10_genomes/), last accessed December 18, 2014) was identified as single copy in the seven angiosperm genomes included in the classification: *Arabidopsis thaliana* Columbia (version 7), *Carica*

	Description	<i>Triphysaria versicolor</i>	<i>Striga hermonthica</i>	<i>Phelipanche aegyptiaca</i>
Stage 0	Imbibed seed			
Stage 1	Roots of germinated seedlings ( <i>Triphysaria</i> ) or germinated seedlings (after exposure to GR24 in <i>Striga</i> and <i>Phelipanche</i> )			
Stage 2	Roots of germinated seedlings ( <i>Triphysaria</i> ) or germinated seedlings ( <i>Striga</i> and <i>Phelipanche</i> ) after exposure to host roots ( <i>Triphysaria</i> and <i>Phelipanche</i> ) or DMBQ ( <i>Striga</i> )			
Stage 3	Haustoria attached to host roots; penetration stages, pre-vascular connection (~ 48hrs)			
Stage 4.1	Haustoria attached to host roots; penetration stages after vascular connection (~72 hrs in <i>Striga</i> and <i>Phelipanche</i> and ~120 hrs in <i>Triphysaria</i> )			
Stage 4.2	Spider stage	N/A	N/A	
Stage 5	Late post-attachment stage from below ground plants 5.1 Pre-emerged shoots (S) from soil 5.2 Roots (R) on pre-emerged shoots	N/A		
Stage 6.1	Vegetative structures; leaves/stems			
Stage 6.2	Reproductive structures; floral buds (up through anthesis)			
Stage 6.3	Roots in late post-attachment		N/A	N/A

**FIG. 1.** An illustration of stages of each parasitic plant used in the PPGP (Westwood et al. 2012) in this study. Drawings are based on original photographs, as shown in Nickrent et al. (1979), Musselman and Hepper (1986), Zhang (1988), and Rumsey and Jury (1991). Additional sequences from the parasite–host interface (Honaas et al. 2013) were also used to study haustorial-specific gene expression (stage 4). H, host; P, parasite; V, vasculature; R, root; S, shoot; N/A, not applicable.

*papaya* (version 1), *Populus trichocarpa* (version 1), *Medicago truncatula* (version 1), *Oryza sativa* (version 5), *Sorghum bicolor* (version 1), and *Vitis vinifera* (version 1). The *Ar. thaliana* proteins from each of the three conserved single copy gene

lists were used as the query in tBLASTn search of each transcriptome assembly. A gene was considered detected if it returned a hit with an *E* value smaller than 1e-10 and at least 30 amino acids long. Results from this analysis shown

**Table 1.** Transcriptome Assembly Statistics in the Postprocessed Combined Assembly for Each Study Species.

Species	Assembly ID	Number of Contigs	Assembly Size (Mb)	Number of N50 Contigs	N50 Contig Length (bp)	Mean Contig Length (bp)
<i>Triphysaria versicolor</i>	TrVeBC3	131,173	76.20	24,729	789	580.91
<i>Striga hermonthica</i>	StHeBC3	117,470	87.53	21,356	1,183	745.17
<i>Phelipanche aegyptiaca</i>	PhAeBC5	129,450	83.80	21,552	1,010	643.48

**Table 2.** Transcriptome Gene Capture Statistics in Three Parasitic Species.

Gene Set	Total	TrVeBC3	TrVeBC3 Proportion (%)	StHeBC3	StHeBC3 Proportion (%)	PhAeBC5	PhAeBC5 Proportion (%)
COSII single copy	220	216	98.18	214	97.27	201	91.36
PlantTribes2.0 single copy	970	949	97.84	952	98.14	869	89.59
UCO	357	357	100.00	356	99.72	354	99.16

in table 2 indicate that gene coverage ranged from at least 90% in *Phelipanche* (PlantTribes single copy analysis) to 100% (UCO analysis) in *Triphysaria* combined assemblies. These results suggested that our assemblies have excellent gene coverage and are very likely to capture the large majority of the expressed genes in a transcriptome. Additionally, to validate the accuracy of the de novo transcriptome assemblies, we used reverse transcription (RT)-PCR to amplify a total of 33 contigs spanning a range of assembly sizes in the three parasitic species. The estimated sizes from the amplified cDNAs agree well with the expected sizes from the de novo transcriptome assemblies ( $R^2$  for the three species of *Triphysaria*, *Striga*, and *Phelipanche* range from 0.973 to 0.999), suggesting a high degree of accuracy in the de novo assemblies (supplementary fig. S1, Supplementary Material online). Seven of the *Triphysaria* sequences were selected for validation sequencing and matched precisely the predicted length of the contig and differed from the reference assembly by at most a few single nucleotide polymorphisms, as would be expected for allelic variants in an outcrossing species. To conclude, we have produced high-quality large scale transcriptome assemblies that serve as a valuable resource for comparative studies of parasitic plant gene content and expression.

### Validation Genes with Known Roles in Orobanchaceae Parasitism

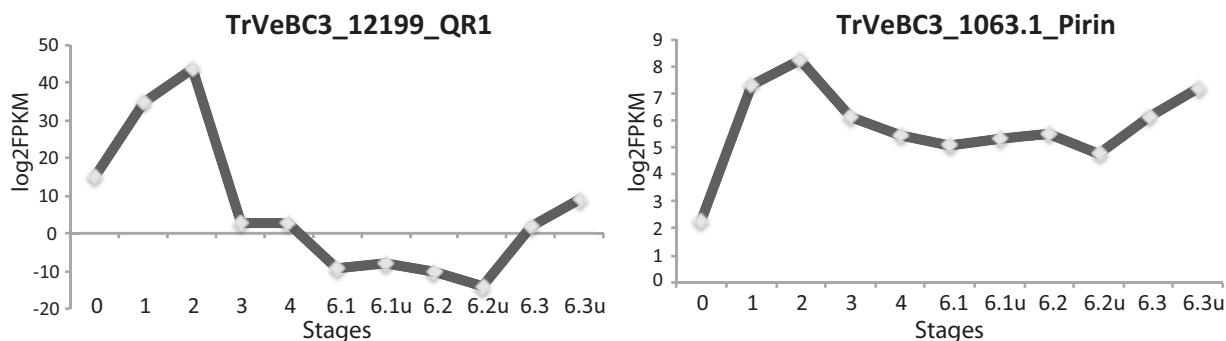
We validated the expression data from RNA sequencing (RNA-seq) using expression profiles of two genes that are known to play a role in parasitism: *TvQR1* (Bandaranayake et al. 2010; Ngo et al. 2013) and *TvPirin* (Matvienko et al. 2001; Bandaranayake et al. 2012; Ngo et al. 2013). Both genes are upregulated in *Triphysaria* roots exposed to the HIF (quinone 2,6 dimethoxy-1,4-benzoquinone; DMBQ) (stage 2) compared with roots without exposure (stage 1). All BLASTn alignments of the *TvQR1* gene with an *E* value cutoff of  $e^{-10}$  or smaller were used to construct the putative full-length transcript representing *TvQR1*. The expression level of *TvQR1* in each stage was calculated as the combined expression of all

the unigene hits within TrVeBC3\_12199 trinity component that contributed to the construction of the full-length reference (TrVeBC3\_12199.1–TrVeBC3\_12199.9). For profiling gene expression we also used three libraries made from host–parasite interfaces of haustoria from the three parasitic plants (Honaas 2013). The interface tissues from the three parasitic plants were targeted by a laser-capture microdissection approach (Honaas 2013; Honaas et al. 2013). Reads from stage 4 interface libraries were mapped onto the combined assemblies to quantify the gene expression in the interface. The RNA-Seq data for the gene *TvQR1* showed high and specific expression for root tissue (stage 1 and stage 2) but low expression levels in other tissues (haustoria, seed, and above ground tissue). When root tissue was treated with DMBQ (stage 2), the expression of *TvQR1* increased relative to roots without any treatment (stage 1) (fig. 2), which is consistent with results obtained in previous studies (Bandaranayake et al. 2010). These results confirm the expected expression for *TvQR1* (fig. 2).

The only significant hit for gene *TvPirin* in the assembly was contig TrVeBC3\_1063.1, which included the full-length coding sequences and 5′- and 3′-untranslated regions. As expected (Matvienko et al. 2001; Bandaranayake et al. 2012), this gene showed the highest expression in stage 1 and stage 2 (root tissue), with the expression in stage 2 (root treated with DMBQ) higher than stage 1 (untreated root) (fig. 2). Postattachment root tissue (6.3) also showed relatively high expression levels for *TvPirin*, suggesting that this gene is highly expressed in roots. Given the consistency in expression validation as well as assembly validation (supplementary fig. S1, Supplementary Material online), our RNA-Seq assemblies should be able to provide good estimates of gene expression in the species within this study.

### Differential Gene Expression and Clustering to Identify Haustorial Genes

A DE analysis for the common stages present in all three parasitic plants (stages 0, 1, 2, 3, 4, 6.1, and 6.2) was performed to identify DE patterns for any pairwise comparison among



**Fig. 2.** Gene expression profiles from RNA-seq data for two previously characterized parasitism genes in *Triphysaria* (QR1, left, and Pirin, right). The two genes were shown to be upregulated in stage 2 relative to stage 1 by RT-PCR, which was confirmed by RNA-Seq data. The organ/stage of the parasite sampled is shown along the x axis. Labels on the x axis refer to stage (see fig. 1) and “u” means that this facultative parasite was growing “unattached” to any host (i.e., 6.1u means unattached stage 6.1). The y axis gives expression values as FPKM on a log<sub>2</sub> scale.

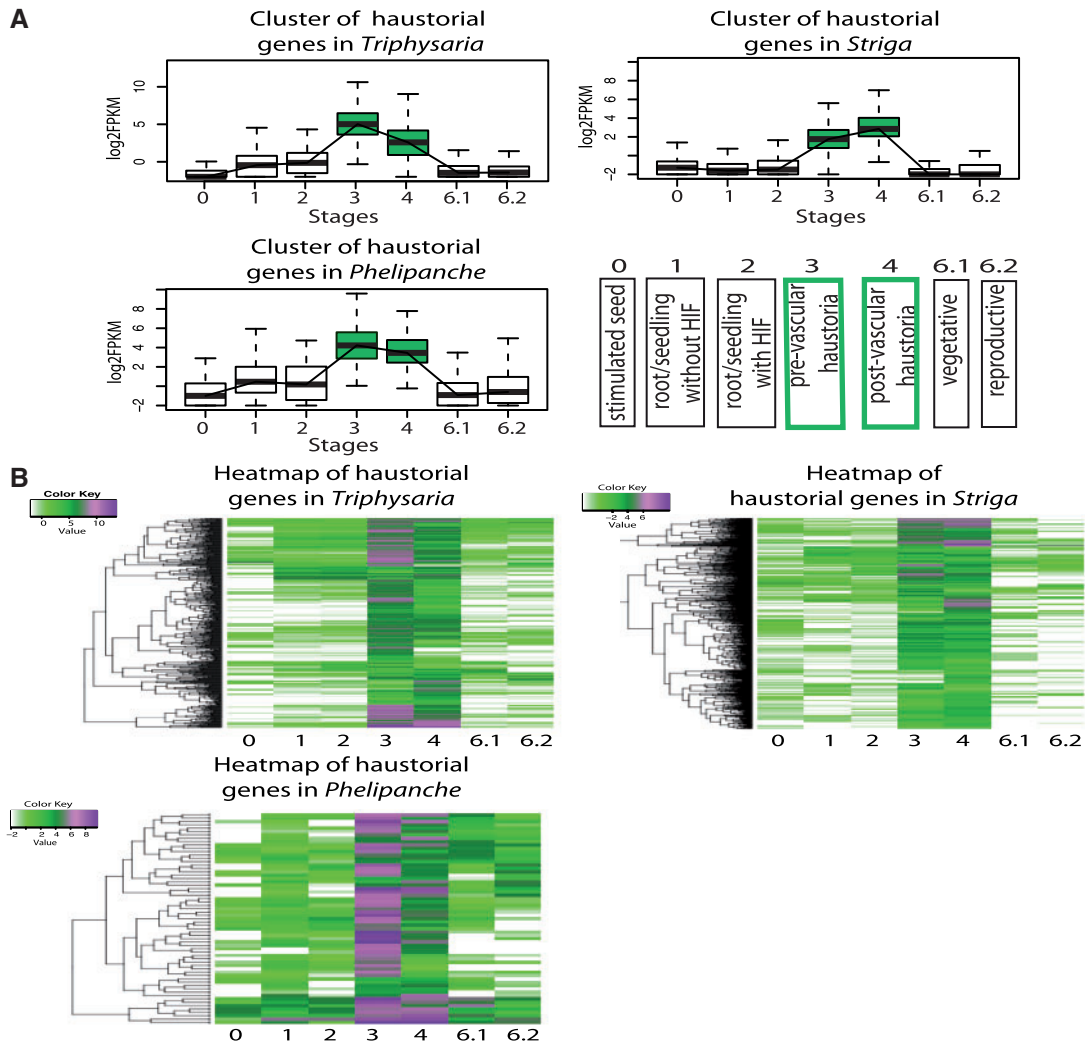
the seven stages. Next, we conducted two clustering analyses using K-means clustering and self-organizing maps (SOM clustering) to identify clusters of coexpressed genes with high expression in postattachment haustorial stages (3 and/or 4) for each parasitic plant. Clusters of coexpressed genes that exhibited significantly higher gene expression in postattachment haustorial stages were extracted from the K-means cluster analysis for each species. We refer to these upregulated genes in postattachment haustorial stages as “haustorial genes.” A boxplot and expression heat map for each cluster of haustorial genes in each species were used to visualize the specific expression patterns for each of the parasitic plants (fig. 3A and B). DE analyses and clustering approaches were performed for expression of both unigenes and a more inclusive putative transcript definition, the “component-orthogroup” (supplementary data S1 and S2, Supplementary Material online). The latter is defined as a representative sequence for a Trinity component containing all associated unigenes (e.g., splice forms, alleles, subassemblies), so long as they are assigned to the same orthogroup (i.e., clusters of homologous genes representing narrowly defined gene lineages) in the gene family classification used by the (*Amborella* Genome Project 2013). The unigenes and component-orthogroups identified by SOM clustering are shown in the supplementary section, Supplementary Material online (supplementary data S3, Supplementary Material online).

Genes that are differentially upregulated in developmentally similar stages of haustorial development, and are evolutionarily conserved across species, are likely to play an important role in parasitism. We examined three species in Orobanchaceae that exhibit varying levels of host dependence and photosynthetic ability. These species serve as divergent biological replicates; shared gene sequences that show conserved upregulation in parasitic structures are likely to be important to the parasitic process. We used orthogroups to associate homologous (and putatively orthologous) genes across the Orobanchaceae species in this study (see Materials and Methods, supplementary data S4 and S5, Supplementary Material online). The final list of candidate haustorial genes was defined as the union of orthogroups represented by upregulated unigenes or component-orthogroups from each species. We also identified unique

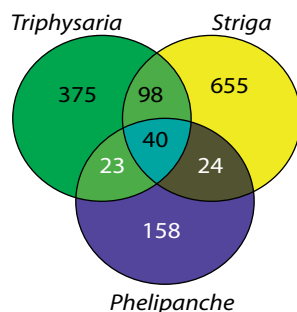
orthogroups and orthogroups present in only two of the three species. Both K-means clustering and SOM clustering were used to identify genes with high and specific expression after attachment to a host (supplementary data S3 and S6, Supplementary Material online). As a result, we identified 185 orthogroups that contained genes (874 unigenes and 488 component-orthogroups) highly expressed in at least two of the species (fig. 4). Forty orthogroups were identified that show their highest level of expression during haustorial development in all three parasitic plants. It is important to note that most of the two-way shared haustorial genes are likely to represent a true set of haustorial genes, because at least 70% of these two-way shared orthogroups also contain genes with increased expression in haustorial stages in the third species (supplementary data S7, Supplementary Material online).

### Shared Haustorial Genes Are Enriched for Proteolysis and Extracellular Region Localization

To determine whether the haustorial genes are enriched for specific molecular processes or biochemical pathways, we took the highly expressed unigenes belonging to an orthogroup shared by at least two species and aligned them with BLASTx to the TAIR database (Lamesch et al. 2012). Best hits in *Arabidopsis* ( $E$  value  $\leq e^{-10}$ ) were used as the input for a DAVID enrichment analysis (Huang et al. 2009) for enriched Pfam domains, Gene Ontology (GO) molecular functions (MFs), biological processes (BPs), and cellular components (CCs) (supplementary data S8, Supplementary Material online). All three parasites share enrichment for the Pfam term serine carboxypeptidase. In addition, the haustorial genes in *Triphysaria* and *Striga* are significantly enriched for eukaryotic aspartyl protease (AP), peroxidase, and leucine-rich repeat N-terminal domains (table 3 and supplementary data S3 and S8, Supplementary Material online). There was a similar high level of enrichment for eukaryotic AP and leucine-rich repeat N-terminal domains in *Phelipanche*, but this was not significant after correcting the  $P$  value for multiple testing. The smaller number of haustorial genes in *Phelipanche* limited the power to detect significantly enriched



**FIG. 3.** Gene expression clustering boxplots (A) and heatmaps (B) of upregulated genes in postattachment haustorial stages 3 and 4 (haustorial genes) in parasitic Orobanchaceae. (A) One cluster of highest expression in post attachment haustorial stages with K-means clustering in each species (511 in *Triphysaria*, 958 in *Striga*, and 126 in *Phelipanche*). Expression in each stage is represented by a boxplot. The upper whisker of the boxplot indicates the highest expression value for features within each cluster; the lower whisker, the lowest expression value; and the middle line, the median expression. The upper and lower edges of the box represent the 75th and 25th percentile, respectively. Expression of genes in the postattachment haustorial stages is highlighted in green. A description of the focal stages is shown on the lower right. (B) Gene expression heat map of component-orthogroups with upregulated expression in postattachment haustorial (stage 3 and/or 4) stages identified by K-means and hierarchical clustering in *Triphysaria*, *Striga*, and *Phelipanche*. The color-intensity in the heat map represents expression value ( $\log_2\text{FPKM}$ ).



**FIG. 4.** Venn diagram illustrating the number of orthogroups with upregulated expression in stage 3 and/or stage 4 (by K-means and SOM) in *Triphysaria*, *Striga*, and *Phelipanche*.

terms. Finally, pectate lyase (PL) was significantly enriched among shared genes in *Triphysaria*.

The largest GO BP category in terms of the number of genes (supplementary data S8, Supplementary Material online) is “proteolysis,” represented by genes encoding AP or serine-type peptidase and subtilase, followed by oxidation–reduction processes, such as peroxidases, and protein phosphorylation, such as kinases. There are also two genes involved in transport activity; one is an oligopeptide transporter and the other is a glutamate-receptor protein. Moreover, six genes involved in cell wall modification were identified, including three genes encoding PL or PL-like proteins, one encoding pectin methylesterase inhibitor (PMEI), one encoding cellulase, and one encoding carbohydrate-binding X8 protein. GO MF terms such as “serine-type peptidase

**Table 3.** Enriched Pfam Domains in the Shared Set of Haustorial Unigenes Identified by Either K-Means or SOM Clustering in *Triphysaria*, *Striga*, and *Phelipanche*.

Pfam Term	<i>Triphysaria</i>		<i>Striga</i>		<i>Phelipanche</i>	
	Fold Change	Bonferroni-Adjusted P Value	Fold Change	Bonferroni-Adjusted P Value	Fold Change	Bonferroni-Adjusted P Value
Serine carboxypeptidase	24.0	5.08E-07*	24.0	4.26E-08*	39.1	3.58E-05*
Eukaryotic AP	39.6	2.64E-06*	40.7	9.94E-08*	41.4	NS
Peroxidase	22.4	7.56E-11*	14.0	3.80E-05*	NA	NA
Leucine rich repeat N-terminal domain_2	6.1	1.67E-02*	7.4	1.07E-04*	6.7	NS
FAD_binding_4	16.1	3.85E-02*	23.1	6.99E-06*	NA	NA
PL	41.2	2.02E-06*	15.9	NS	NA	NA

NOTE.—Significance levels for category enrichment relative to background are given as Bonferroni-adjusted *P* values. NA means that enrichment information for the particular term is not identified by the test; NS, nonsignificant.

**Table 4.** Enriched GO Cellular Component (GO-CC), Biological Process (GO-BP) and Molecular Function (GO-MF), KEGG Pathway, and Tissue Expression Terms among Shared Set of Haustorial Unigenes Identified by Either K-means or SOM Clustering in *Triphysaria*, *Striga*, and *Phelipanche*.

Term	<i>Triphysaria</i>		<i>Striga</i>		<i>Phelipanche</i>	
	Fold Change	Bonferroni-Adjusted P Value	Fold Change	Bonferroni-Adjusted P Value	Fold Change	Bonferroni-Adjusted P Value
GO-CC: External encapsulating structure	5.3	3.65E-12*	5.1	1.65E-12*	3.6	NS
GO-CC: Cell wall	5.2	1.69E-11*	5.0	7.09E-12*	3.7	NS
GO-CC: Extracellular region	3.5	4.53E-11*	3.3	3.97E-10*	2.8	NS
GO-BP: Proteolysis	3.4	1.92E-06*	4.4	7.51E-14*	4.9	3.88E-05*
GO-BP: Cellular response to hydrogen peroxide	18.0	1.87E-09*	10.2	4.18E-03*	NA	NA
GO-BP: Cellular response to ROS	16.1	7.26E-09*	9.1	8.65E-03*	NA	NA
GO-MF: Serine-type peptidase activity	13.3	4.35E-13*	15.8	9.14E-21*	21.9	3.01E-10*
GO-MF: Aspartic-type endopeptidase activity	15.0	2.59E-06*	18.3	7.31E-11*	15.7	NS
GO-MF: Electron carrier activity	3.6	6.71E-04*	2.8	4.98E-02*	NA	NA
KEGG: Phenylalanine metabolism	14.2	8.68E-10*	10.6	7.95E-06*	NA	NA
KEGG: Methane metabolism	14.0	9.99E-10*	10.5	8.77E-06*	NA	NA
KEGG: Phenylpropanoid biosynthesis	10.9	1.62E-08*	8.2	6.07E-05*	NA	NA
Tissue-specificity: Specifically expressed in root cap cells	49.7	7.06E-01 (NS)	47.2	7.68E-01 (NS)	NA	NA
Tissue-specificity: Expressed in flowers, but not in leaves	33.2	8.40E-01 (NS)	NA	NA	NA	NA

NOTE.—Significance levels for category enrichment relative to background are given as Bonferroni-adjusted *P* values. NA means enrichment information for the particular term was not identified by the test; NS, nonsignificant.

\**P* < 0.05.

activity” and “aspartic-type endopeptidase activity” were significantly enriched (table 4). The KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway terms “phenylalanine metabolism” and “methane metabolism” and “phenylpropanoid biosynthesis” were also significantly enriched in the haustorial upregulated gene set. In addition, other enriched GO terms, such as “cellular response to hydrogen peroxide” and “cellular response to reactive oxygen species (ROS),” support the suggestion by Torres et al. (2006) that ROS may be an important signaling intermediate during the parasitic plant–host plant interaction.

Significant enrichment of genes with the GO CC category “external encapsulating structure,” “cell wall,” and “extracellular regions” were found in both *Triphysaria* and *Striga*

(table 4). A concordant pattern of enrichment was observed for these categories in *Phelipanche*, though, like the domain analysis, the conservative test, corrected for multiple comparisons, did not find these enrichments significant in this species. Future research with experimental evidence is needed to determine whether the extracellular predictions for these candidate haustorial genes hold true or not, and whether the proteins they encode affect the parasite–host interactions.

### Two Examples Illustrating Haustorial Gene Expression Evolution

To further understand the evolutionary origin of some of these haustorial-related genes, we examined patterns of



gene expression in orthologs in three nonparasitic model plant or crop species: Thale cress (*Ar. thaliana*), barrel medic (*M. truncatula*), and tomato (*Solanum lycopersicum*). To illustrate this approach, we present a detailed analysis of the PL and peroxidase gene families, which were identified by the presence of enriched Pfam domains in *Triphysaria* (PL) or in both *Triphysaria* and *Striga* (peroxidase). Following an early, possibly angiosperm-wide duplication in the PL gene family (fig. 5A), a subsequent gene duplication gave rise to two paralogous gene lineages shared by *Mimulus* and members of Orobanchaceae. One paralogous gene in parasitic Orobanchaceae and their orthologs in *Arabidopsis* and tomato show principal expression levels in floral tissues (supplementary data S21, Supplementary Material online; Goda et al. 2004; Sun and van Nocker 2010); the other paralog exhibits abundant expression in the haustorium of parasitic Orobanchaceae (fig. 5A). The more recent gene duplication in a common ancestor of *Mimulus* and Orobanchaceae gave rise to two *Striga* genes, one having peak expression in haustorial tissue and the other in flower (fig. 5A). Conservation of principal gene expression in floral tissues of nonparasites and the maintenance of a floral-expressed ortholog in *Striga* strongly suggests that the haustorial expression of this gene was co-opted from an ancestral gene acting in flowers, and was recruited to haustoria following gene duplication through regulatory neofunctionalization. Alternatively, because the ancestral gene may have been expressed in both floral tissue and root, a two-step process in which the ancestral gene first subfunctionalized and then shifted to haustoria would be an example of subneofunctionalization (He and Zhang 2005).

In contrast to the PL gene, a peroxidase gene family shows a shift of gene expression from roots in all related nonparasitic model species to haustorial tissue of parasitic plants without gene duplication (fig. 5B). The peroxidase gene highly expressed in roots of *Arabidopsis* was characterized to be involved in the production of ROS (Kim et al. 2010), stress response (Llorente et al. 2002), and pathogenic responses (Ascencio-Ibanez et al. 2008).

### Identifying Haustorial Initiation Genes and “Parasitism Genes”

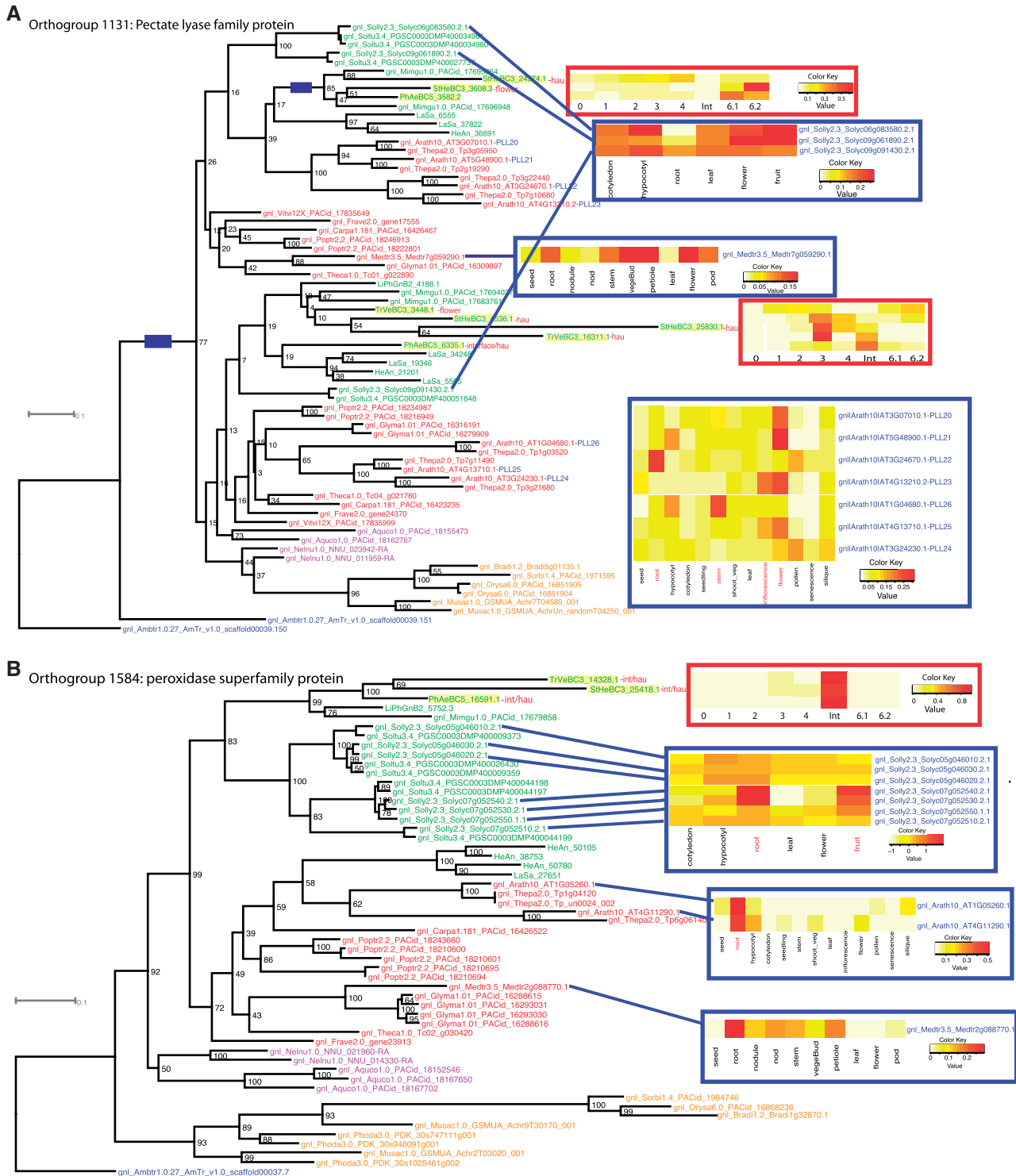
Genes that are upregulated in response to the haustorial initiation factor (HIF) DMBQ might also play an important role in parasitism. For example, *TvQR1*, which is upregulated following DMBQ exposure, encodes a quinone reductase that acts early in the HIF signaling pathway (Bandaranayake et al. 2010). Stage 1 in the tiny-seeded *Striga* and *Phelipanche* (Westwood et al. 2012) consists of whole germinated seedlings, whereas in the much larger-seeded species (*Triphysaria*), stage 1 comprised excised radicles of germinated seedlings. Upon treatment of stage 1 seedlings or roots with DMBQ or a host root, the plants progress to stage 2 (haustorial initiation). DESeq was used to identify “haustorial initiation genes” (HIGs), defined here as unigenes and component-orthogroups that have significantly higher gene expression

in stage 2 compared with stage 1 (supplementary data S9, Supplementary Material online).

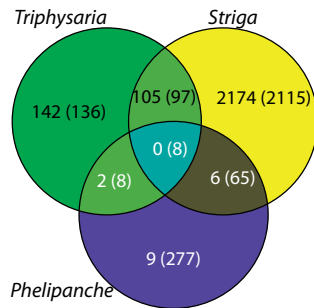
*Triphysaria* HIGs were enriched for GO MF terms such as “magnesium ion binding,” “calcium ion binding,” “calcium-transporting ATPase activity,” “calcium ion transmembrane transporter activity,” and “cation-transporting ATPase activity” (supplementary data S10, Supplementary Material online). The co-occurrence of putative functions of calcium ion transport and ion binding with ATPase activity suggest a possible involvement of a Ca<sup>2+</sup>ATPase (Brini and Carafoli 2011) in the regulation of the haustorium initiation pathway. In *Striga*, however, a distinct set of enriched GO MF terms was detected, including “nucleotide binding,” “ATPase activity,” and “ATP-dependent helicase activity.” This may suggest a different picture associated with HIF exposure between facultative and obligate parasitic plants.

By combining the list of shared upregulated genes in haustoria (fig. 4 and supplementary data S5, Supplementary Material online) with HIGs (fig. 6 and supplementary data S11, Supplementary Material online), we define a joint set of genes in these parasites that we call parasitism genes. Parasitism genes are defined as having enhanced expression in parasitic structures, and likely play a role in parasite biology, as opposed to parasite-specific sequences which are defined only by their joint presence in parasitic and absence from nonparasitic plants (see below). In total, we identify 1,809 parasitism genes in these three parasitic species that are assigned to 298 orthogroups that were shared by at least two species. As most of the parasitism genes shared by two of the parasitic plants also show an upregulated pattern in the third species (supplementary data S7, Supplementary Material online), this set of genes upregulated in parasitic processes constitutes a shared set of parasitism genes in Orobanchaceae. There are almost 300 gene families shared by at least two species that have genes upregulated in parasitic processes.

The number of orthogroups containing HIGs varied widely in the three parasites (2,285 in *Striga*, 17 in *Phelipanche*, and 249 in *Triphysaria*). Although the low number of HIGs in *Phelipanche* could be a result of lower power to detect DE (fewer replicated libraries and a smaller total volume of data; supplementary data S12, Supplementary Material online), we also observed that a large number of genes are highly upregulated in *Phelipanche* stage 1, suggesting that *Phelipanche* may automatically begin haustorial initiation without HIF exposure, which was reported by a previous study where haustorial initiation, as an exception, does not require the application of HIF (Joel and Losner-Goshen 1994a). To examine this possibility, we expanded the list of HIGs by including genes highly expressed in stage 1 (a cluster of upregulated expression in stage 1 relative to any other stages 0, 2, 3, 4, 6.1, and 6.2) of *Phelipanche* and performed another Venn diagram analysis. In this expanded set, there are eight additional orthogroups including HIGs that were shared by all three parasitic plants as well as an additional 65 orthogroups shared by two species (fig. 6—number in parenthesis). An examination of the eight orthogroups revealed genes coding for the following functions: Cytochrome P450, heat



**FIG. 5.** Gene family phylogeny and gene expression profile of two orthogroups showing: (A) Shift of gene expression from flower to haustorium following gene duplication, and (B) shift of gene expression from root to haustorium without gene duplication. Gene duplication events relevant to the origin of parasite genes are shown on the tree with blue rectangular bars. Sequence names are color-coded to represent different lineages: Basal angiosperms (blue), monocots (yellow), basal eudicots (purple), rosids (red), and asterids (green). Parasite genes are highlighted with yellow background and green foreground. The expression of parasite genes and nonparasite genes in *Arabidopsis*, *Medicago*, and tomato are shown using heat maps. Green and red lines are connecting genes from the phylogeny to the heatmap for nonparasitic genes (except in *Arabidopsis* orthogroup 1131 where genes are labeled as PLLs) and parasitic genes. The tissue with the highest expression was labeled in red. The color intensity in heatmaps refers to expression measurements with an RNA-Seq approach in parasites and tomato, and with microarrays in *Arabidopsis* and *Medicago*. Int or int means “interface” tissue of haustoria (~stage 4) (Honaas 2013), and hau means “haustoria”.



**Fig. 6.** HIGs: Orthogroups containing genes upregulated in root or seedlings following HIF exposure (stage 2) compared with germinating seedlings (stage 1) in *Triphysaria*, *Striga*, and *Phelipanche*. The numbers in parenthesis are the corresponding set of orthogroups when including genes that are highly expressed in stage 1 (relative to any other stages of stage 0, 2, 3, 4, 6.1, and 6.2—by K-means clustering) of *Phelipanche*.

shock protein 70, ribosomal protein, peptidase C48, oleosin, ATPase, pyruvate kinase, and integrase. This is consistent with the possibility that *Phelipanche* starts haustorial initiation at an earlier stage (stage 1) than other two hemiparasites. Alternatively, because haustorium development in response to HIFs in *Phelipanche* is not as evident as in *Striga* or *Triphysaria* (Joel and Losner-Goshen 1994a), there may actually be fewer HIGs in *Phelipanche*.

### A Majority of Parasitism Genes Evolved through Gene Duplication

We next explored the possibility that these putative parasitism genes evolved through gene duplication. We utilized an automated tree-building pipeline (Jiao et al. 2011; *Amborella* Genome Project 2013) to construct gene families for 114 orthogroups containing parasitism genes with a manageable number of genes to score for gene duplications (orthogroup number ranging from 1,000 to 9,999). Each orthogroup contained homologs from 22 other plant species used in the construction of the classification (*Amborella* Genome Project 2013), plus the genes from Orobanchaceae identified here. Manual inspection of the gene tree phylogenies was performed to find parasite genes that may have been missing in one or more “whole plant” combination assemblies. We also manually examined each alignment (and resulting tree) for frame shift and translation errors that could result in extremely long (greater than 10× others) branches. These errors were corrected when possible, or the sequence was eliminated from the matrix to avoid spurious topologies. Together, these gene family phylogenies give us a broad view of how parasitism genes evolved.

Of the 114 orthogroups (supplementary data S13, Supplementary Material online) containing parasitism genes, gene duplications were detected in 58 trees at  $\geq 50\%$  bootstrap support and 38 trees at  $\geq 80\%$  bootstrap support value. By mapping the duplication events observed in parasitic plants onto phylogenetic species trees, and examining bootstrap support values for key supporting nodes, we determined when the putative parasite paralogs were duplicated

**Table 5.** Phylogenetic Placement of Gene Duplications Observed in Gene Families with Shared Parasitism Genes.

Duplicated Lineages	Orthogroups with Duplication	
	BS $\geq 80\%$	BS $\geq 50\%$
Parasite-wide	9 (23.68)	13 (22.41)
Orobanchaceae-wide	4 (10.53)	9 (15.52)
Orobanchaceae+ <i>Mimulus</i>	21 (55.26)	28 (48.28)
EuAsterid1-wide	1 (2.63)	1 (1.72)
Asterid-wide	1 (2.63)	1 (1.72)
Core-eudicot-wide	5 (13.165)	8 (13.79)
Eudicot-wide	3 (7.89)	8 (13.79)
<b>Total</b>	<b>38 (100)</b>	<b>58 (100)</b>

NOTE.—Values in parentheses are given in percentage.

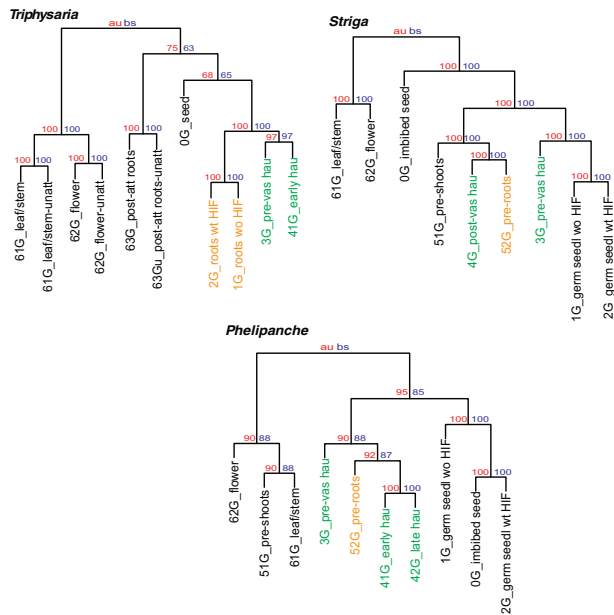
(supplementary data S14, Supplementary Material online). A detailed scheme illustrating various duplication events for when parasitism genes were duplicated is shown in supplementary data S15, Supplementary Material online. The greatest proportion of duplicated gene families supported a gene duplication event that occurred in a common ancestor of *Mimulus* and Orobanchaceae (but not seen in Solanaceae, other asterids, or rosids) (table 5).

As with the parasite genes in general, most of the duplicated parasitism genes detected in this analysis were annotated with terms related to peptidase activity (such as AP, serine carboxypeptidase) and cell wall modification processes (PL, PME1, carbohydrate-binding X8 protein, and glycosyl hydrolase). In addition, three TFs (homeodomain-like TF, ethylene responsive TF, and lateral organ boundaries (LOB) domain-containing protein), genes with transporter activity (cationic amino acid transporter, major facilitator family protein, NOD26-like intrinsic protein, and an oligopeptide transporter), a peroxidase, and a leucine-rich repeat-containing protein were also derived from scorable gene duplications.

### Regulatory Neofunctionalization and Origin of the Haustorium from Root and Flower

#### Tissue Expression Clustering

To obtain a global view of transcriptional profiles throughout parasite growth and development in each species, expression values for each unigene were clustered by tissue and stage expression levels using complete linkage and correlation distances using the pvclust routine in R (Racine 2012) (supplementary data S16, Supplementary Material online). The expression clustering in each of the three parasitic plants (fig. 7) shows that overall gene expression from vegetative and reproductive above-ground tissues is quite different from the below-ground structures. In both *Striga* and *Triphysaria*, above-ground stage 6.1 (vegetative structures) clustered with above-ground stage 6.2 (reproductive structures; floral buds), and were separated from the remaining tissues with 100% bootstrap support. In contrast, pre-emergent shoots occurring underground (stage 5.1) in *Phelipanche* clustered with above-ground shoots (stage 6.1), and the two shoot transcriptomes clustered with floral buds (stage 6.2). It



**FIG. 7.** Overall similarity of transcriptional profiles of all stages in three parasites. Numerical values represent supports as estimated by the approximately unbiased (on left, in red) and bootstrap (on right, in blue) as described in the Materials and Methods. Clustering was performed with complete linkage and correlation distance. Haustoria tissues (from stages 3 and 4) are labeled in green, whereas root tissues (from stages 1, 2, and 5.2) are labeled in orange.

is notable that *Phelipanche* and *Striga* both produce pre-emergent shoots (stage 5.2), but the overall expression patterns of this stage in the two species is somewhat different. The fact that *Striga* pre-emergent shoots (stage 5.1) do not cluster with emergent shoots (stage 6.1), but are more similar to other below-ground stages (haustorium—stage 4 and pre-attachment roots—stage 5.2), may be due to the fact that photosynthetic activity in *Striga* shoots only becomes active after emergence, whereas *Phelipanche* is not capable of photosynthetic activity at all.

Cluster analysis (fig. 7) also shows that in all three species haustorial gene expression is overall most similar to root expression. In *Triphysaria* and *Phelipanche*, expression patterns from both stage 3 and stage 4 haustorial tissues cluster with roots. (Roots from *Triphysaria* were taken from germinated seedlings [stages 1 and 2], whereas roots for *Phelipanche* were from the late postattachment stage, but prior to shoot emergence from the soil [stage 5.2].) In *Striga*, a late stage 4 haustorial tissue is most similar to roots prior to the above-ground emergence of shoots, a scenario similar to *Phelipanche*.

#### Expression of Orthologs of Parasite Genes in Nonparasitic Models

The pattern we see of haustorial expression being most similar to root is consistent with the longstanding hypothesis that the haustorium was derived from a modified root (Kuijt 1969; Musselman and Dickison 1975; Joel 2013). However, it is also possible that individual genes functioning in the haustorium have been recruited from genes normally expressed in other plant organs. To investigate this scenario,

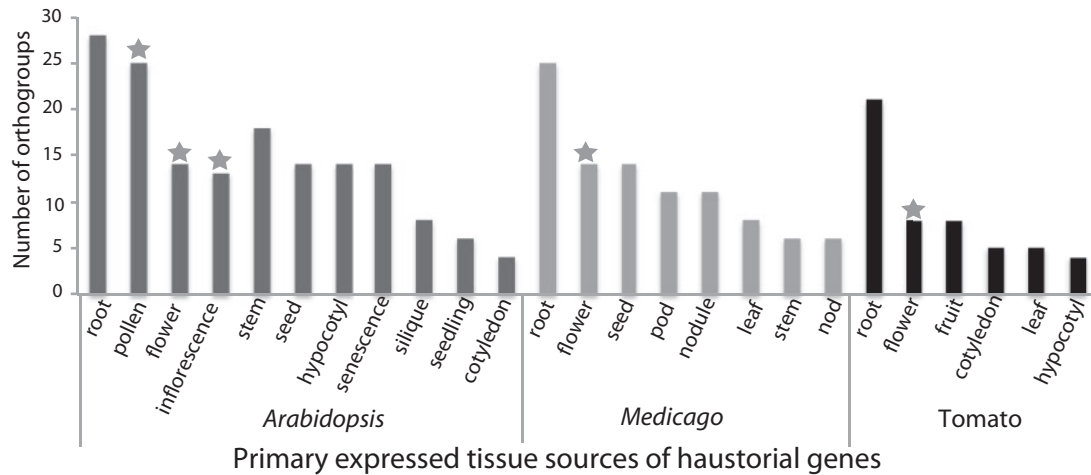
we compared gene expression of candidate parasitism genes with extensive gene expression data from multiple tissues and organs of *Ar. thaliana*, *M. truncatula*, and tomato to examine the evolution of expression patterns across large evolutionary distances. When we trace the haustorial gene expression back to orthologous genes in nonparasitic plants, we found significantly higher organ-specific expression in root and floral tissue, than in leaf, seed, or hypocotyl (fig. 8 and supplementary data S22, Supplementary Material online).

In all three nonparasitic model plant species, the orthologs of haustorial genes are expressed most highly in root and floral (or fruit) tissues, suggesting that these were the major sources of genes recruited to the haustorium. For both *Medicago* and tomato, root is the most frequent source, which is consistent with the similarity between root and haustorial expression in the parasites (fig. 7). In *Arabidopsis*, orthologs of haustorial genes are also commonly upregulated in roots, but even more are upregulated in pollen. It is possible that the slightly different picture obtained from the three nonparasitic plants might be due to differences in tissue sampling in the nonparasitic model species. For example, floral tissue sampling is less extensive in tomato and *Medicago* than in *Arabidopsis*.

#### Parasitism Genes Show Signatures of Adaptive Evolution or Relaxed Constraint in Parasite Lineages

To examine whether altered selection patterns play a role in the evolution of parasitism, we utilized the branch model in PAML for hypothesis testing. We labeled the parasite genes as the foreground and the remaining genes as the background, and then identified genes that show accelerated evolution (dN/dS ratio) in parasitic lineages compared with the background. We focused on parasitism genes identified in our analyses. The branch model implemented in PAML was used to identify orthogroups that show a significantly higher dN/dS ratio in parasitic lineages compared with nonparasitic lineages, whose GO BPs include proteolysis, cell wall modification, oxidation–reduction process, transport, protein glycosylation, cytokinin metabolic process, and ubiquitin-dependent protein catabolic process (supplementary table S1, Supplementary Material online). To examine whether there are sites that have evolved under positive selection, the branch-site model was performed on orthogroups that show an elevated dN/dS ratio in foreground parasite lineages relative to background nonparasitic lineages. Nine orthogroups were found to contain sites under positive selection. They include two orthogroups encoding AP and one orthogroup each encoding serine carboxypeptidase, expansin, glycosyl transferase, pectin methyltransferase inhibitor, PAR1 protein, and C2H2, and C2H2 zinc finger family protein (supplementary data S17, Supplementary Material online), respectively.

We then compared the dN/dS ratio of haustoria-specific genes with nonhaustorial-specific genes. The dN/dS ratio was calculated by selecting orthologous pairs across three

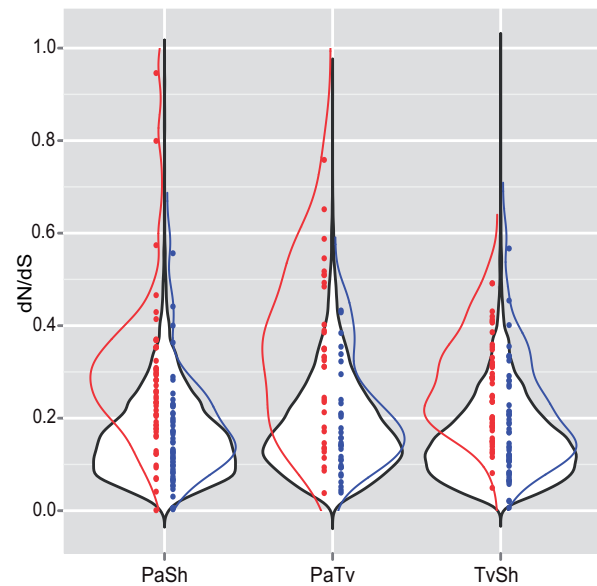


**FIG. 8.** Haustorial genes in parasitic species were recruited from root, flower, and other tissues. Values on the y axis show the number of orthogroups containing haustorial genes as identified from expression analysis of nonparasitic model species *Arabidopsis*, *Medicago*, and tomato. Tissues on the x axis represent the principally expressed tissue for orthologs of upregulated haustorial genes in *Arabidopsis* (light black), *Medicago* (gray), and tomato (black). Floral tissues are highlighted with stars on top of the bars.

different parasitic species by best BLAST hit based on an E-value cutoff of  $e^{-10}$  (PaSh: between *P. aegyptiaca* and *S. hermonthica*, PaTv: between *P. aegyptiaca* and *T. versicolor*, TvSh: between *T. versicolor* and *S. hermonthica*). dN and dS were calculated separately by codeml in PAML (supplementary data S18, Supplementary Material online). The distribution for genome-wide dN/dS values was represented with a symmetric violin plot from each pairwise species comparison. This overall distribution was made by calculating the dN/dS ratio for the orthologous pairs from all unigenes from each species pair. The distributions of all haustorial gene pairs (in red) and a randomly chosen equally sized set of nonhaustorial gene pairs (in blue) were represented by a dotplot and a density plot. In all three cases, the peaks of the density plots for the haustorial genes are above the peak for nonhaustorial genes. The haustorial genes exhibit significantly greater dN/dS ratio compared with the nonhaustorial genes for all three pairwise species comparisons (Wilcoxon rank,  $P$  value < 0.01) (fig. 9).

### The Majority of the Parasite-Specific Sequences Have Unknown Functions

The transcriptome assemblies allowed us to identify parasite-specific sequences (Yoshida et al. 2010), some of which may be associated with parasitic functions. This was done by building a secondary OrthoMCL orthogroup classification of all of the genes and transcripts that were not assigned to the initial orthogroup classification of 22 plant genomes. In addition to the singleton genes from the 22 genomes, we used in the secondary classification all of the unassigned transcripts from the three parasitic species and from *Lindenbergia*, *Lactuca*, and *Helianthus*. This strategy, which includes multiple nonparasitic lineages closely related to the parasites, allows a highly sensitive means of distinguishing sequences found only in the parasitic Orobanchaceae. We identified 84 novel orthogroups that contain sequences from all three parasitic Orobanchaceae species, but lack sequences from any



**FIG. 9.** Haustorial genes show evidence of adaptive selection or relaxed selective constraint. Symmetric violin plots show the genome-wide distributions of dN/dS values for comparisons of *Phelipanche aegyptiaca* (Pa), *Striga hermonthica* (Sh) and *Triphysaria versicolor* (Tv). Red dots represent haustorial genes shared by all three species, whereas blue dots represent randomly selected nonhaustorial genes from the relevant species. The density plots colored in red and blue represent the frequency distributions of the individual dN/dS values as seen in the dot distributions of haustorial and nonhaustorial genes.

nonparasitic plant (supplementary table S2, Supplementary Material online). 178, 180, and 139 unigenes were found in these orthogroups from *Triphysaria*, *Striga*, and *Phelipanche*, respectively. The large majority of these sequences had no significant BLAST alignments to any of the nonparasitic species, whereas a few of the predicted peptide sequences (6, 18, and 13 from *Triphysaria*, *Striga*, and *Phelipanche*, respectively)

had hits to genes of unknown function in the annotation databases. Most of the significant hits in the annotation databases corresponded to sequences that were transposon and retrotransposon related, such as GAG-preintegrase domain protein and retrotransposon gag protein (supplementary table S2, Supplementary Material online). A sequence with homology to a homing intron endonuclease with two LAGLIDADG motifs (Belfort and Roberts 1997) was also among the sequences with annotations. One orthogroup contained sequences with distant homology to genes annotated as mitochondrial aconitase with a putative role in mitochondrial oxidative electron transport (Yan et al. 1997). We also obtained the stage-specific expression profiles for each of parasite-specific genes. Six and four of the parasite-specific unigenes in *Striga* and *Triphysaria* were also among the list of significantly upregulated haustorial genes (supplementary table S2, Supplementary Material online). In addition, three parasite-specific orthogroups contained genes from all three parasite species that exhibited a pattern of increased expression in the haustorial postattachment penetration stages (supplementary table S2, Supplementary Material online).

## Discussion

The PPGP has used large-scale transcriptome sequencing to interrogate multiple stages of parasite growth and development of three related parasitic plants spanning a wide range of parasite ability, enabling an integrated analysis of genes upregulated in parasitic processes in Orobanchaceae. The large comparative framework has made it possible to identify, for the first time, a set of genes we believe are essential or core to parasitism. These candidate parasitism genes will be a valuable resource for future functional studies as we strive to understand the genetic changes that led to the parasitic lifestyle, as well as those that resulted from the transition to heterotrophy. Among the core parasitism genes are specific members of gene families encoding cell wall modifying enzymes (cellulase, PLs, glycosyl hydrolases, and pectin methyl-esterase), and peroxidase enzymes, proteins that are known to be involved in the parasite invasion process (Singh A and Singh M 1993; Antonova and TerBorg 1996; Losner-Goshen et al. 1998; Pérez-de-Luque 2013). We also identified a variety of genes (encoding proteases, transporters, regulatory proteins [TFs and receptor protein kinases], and others, including many genes of unknown function) that are coexpressed in parasitic stages and may be important in haustorial development and function. Because homologs of these haustorially expressed genes encode proteins also functioning in nonparasitic plants, this supports the endogenous mechanism for the origin of parasitism (Bandaranayake and Yoder 2013a). For this collection of genes, the evolution of regulatory sequences resulting in novel expression in the haustorium was likely essential to the evolution of parasitic functions. By expanding our analyses to orthologous genes in nonparasitic model plants, we have gained insights into the evolution of parasitism and the source of genes that shifted expression to parasite tissues, and presumably function there. Although some genes have gained haustorial expression in the absence

of detectable gene duplication, a majority of the parasitism genes originated following a gene duplication event.

## Cell Wall Degradation Enzymes and the Haustorium

Enzymatic degradation of the host cell walls has been suggested to be important in the penetration of the parasite across the host root cortex as it attempts to reach and connect with host vascular tissues (Kuijt 1977). Baird and Riopel (1984) observed that the intrusive cells at the tip of the penetration peg of the haustorium contained a densely staining cytoplasm, indicative of high levels of cell wall hydrolytic activity. Additionally, early histological studies observed that during penetration of the host endodermis by the haustoria of *P. aegyptiaca*, there was the dissolution of the middle lamella between host cell walls (Joel and Losner-Goshen 1994b) and degradation of the cutin of the Casparian strips (Joel et al. 1998). Cell wall degrading enzymes such as cellulase, polygalacturonase, and xylanase were found in the tubercles of *P. aegyptiaca* (Singh A and Singh M 1993), suggesting that they play a role in the penetration process necessary for establishing haustorial connections with the host vasculature (Pérez-de-Luque 2013). In the enriched set of upregulated prehaustorial and haustorial genes in dodder (*C. pentagona*), many genes encoding cell wall modifying enzymes were found including PLs, pectin methyl-esterase, cellulases, and expansins (Ranjan et al. 2014).

In our study, four glycosyl hydrolase and five PL genes were upregulated in haustorial tissues in at least two species of Orobanchaceae. GO enrichment analysis of the CC terms identified cell wall and extracellular localization annotation terms as being significantly enriched among the upregulated haustorial genes, suggesting that proteins encoded by these genes tend to be secreted where they could impact cell wall integrity of the parasite or the host. Consistent with this idea was the evidence of disintegration of the middle lamella in *S. gesnerioides* attacking cowpea (Reiss and Bailey 1998). Glycosyl hydrolases were shown to have a role in hydrolysis and degradation of structural or storage polysaccharides, including cellulose and hemicellulose (Henrissat et al. 1995). Pectic enzymes have long been recognized as important proteins in cell wall loosening or disassembly. Almost all cell wall penetration processes, including pollen tube growth and bacterial or fungal pathogenic invasion, involve the modification of pectins that are integral for cell wall stability. For instance, studies reported the role of pectic enzymes in degrading the plant cell wall during the invasion process by bacterial or fungal pathogens (Delorenzo et al. 1991; Volpi et al. 2011). Similarly, a recent study identified a PL to be required for root infection by rhizobia during nodulation (Xie et al. 2012). Additional evidence also supports a direct role of PLs in loosening of the cell wall in fruit ripening (Marin-Rodriguez et al. 2002). Thus, it is likely that PLs play a role in loosening the host cell wall for invasion by parasitic plants.

Immunological detection of pectin methyl-esterase (PME) at the penetration site and demethylated pectins at the cell walls adjacent to the intrusive cells of *Orobanche* (Losner-Goshen et al. 1998) implies a role for pectin degradation in

the haustorial penetration process. Three orthogroups (orthogroup 2875, 6176, and 19181) encoding putative PMEIs were identified to contain genes that were specific to haustorial tissue in at least two species of Orobanchaceae. It would be interesting to determine whether PMEIs and PMEs have distinct roles in parasite–host interactions. A previous analysis revealed another gene involved in cell wall modification, a beta-expansin, which was highly expressed in the haustorial interface between *T. versicolor* and its grass family host (Honaas et al. 2013).

### Proteases, Transporters, and the Haustorium

Genes involved in proteolysis, largely proteases, and proteinases, account for a great proportion of transcripts upregulated in the haustorium and shared by at least two of the parasite species that we investigated. The upregulated haustorial genes identified in this study include four genes encoding subtilisin-like serine protease similar to those required for virulence in bacterial pathogens (Kennan et al. 2010). In addition, a subtilisin-like protein from soybean was reported to activate defense-related genes (Pearce et al. 2010). In nonparasitic plants, serine proteases often play a role in various processes including protein degradation/processing, hypersensitive response, and signal transduction (Antao and Malcata 2005), but what roles they take on in Orobanchaceae parasites is not yet clear.

In addition to serine protease, the eukaryotic APs are also enriched among upregulated haustorial genes. APs, a large gene family with members present in all living organisms, play central roles in protein degradation, processing, and maturation (Chen et al. 2009). Plant APs are expressed in various organs including seed, root, grain, leaf, and flower (Chen et al. 2009). Also they play a role in seed germination, where they degrade seed-storage proteins to provide amino acids to growing plants (Higgins 1984). Other studies identified APs of blood-feeding malaria parasites to play a role in degrading hemoglobin proteins to amino acids for nutrition (Brinkworth et al. 2001). In addition, a marked expansion within the AP gene family was found in the xylem-feeding hemiparasites, *Triphysaria* and *S. hermonthica* (Dorr 1997; Neumann et al. 1999), but not in the phloem-feeding obligate parasite, *Phelipanche* (Aly et al. 2011), which has a lower rate of nutrient uptake from the xylem stream, suggesting that these proteins may play a pivotal role in nutrient mobilization only in the hemiparasites.

Our analyses indicate that four APs show peak expression in haustorial tissue, whereas their orthologs in *Arabidopsis* and tomato (or paralogs in the parasitic plants) have peak expression in root and flower (supplementary fig. S2, Supplementary Material online). This provides another line of evidence for gene recruitment to haustorial function, and suggests that this has occurred through regulatory neofunctionalization. A recent study reported that a rice AP plays an indispensable role in pollen tube germination and growth, and that loss of function results in reduced male fertility (Huang et al. 2013). In addition, a secretome analysis (Kall et al. 2007) of predicted peptides for upregulated haustorial genes

revealed that haustorial genes are more likely than nonhaustorial genes to be extracellularly localized or contain a signal peptide structure (supplementary data S19, Supplementary Material online). The fact that upregulated haustorial genes are enriched for proteases with signal peptides suggests that the evolution of parasitism may be associated with an expansion of the suite of secreted proteases to aid parasite attack and/or feeding.

Analysis of selective constraints showed that proteases with expression specific to haustoria in the parasites show a greater dN/dS as compared with their orthologous genes in nonparasitic plants. The greater dN/dS ratio for these upregulated haustorial proteases suggests either a relaxation of purifying selection or adaptive evolution of these protease-encoding genes associated with the evolution of parasitism. The fact that particular sites were indicated as evolving adaptively, especially in the functional domains, provides support for the latter hypothesis (supplementary data S17, Supplementary Material online).

We also found five genes encoding transporters upregulated in the haustorium that are shared by at least two species: One ABC transporter, two oligopeptide transporters, one zinc transporter, and one glutamate transporter. Upregulated genes encoding transporters including sugar transporter, amino acid transporter, and ammonium transporter were also identified to be enriched in haustorial tissue of the parasite *Cuscuta* (Ranjan et al. 2014). Interestingly, *S. hermonthica* infection has been shown to increase amino acid levels in xylem sap of its *Sorghum* host, with glutamate being the predominant form of translocated nitrogen (Pageau et al. 2003). The assimilation of host <sup>15</sup>N-labeled nitrate into the parasite (Pageau et al. 2003) provided evidence for the potential role of a glutamate transporter in nitrogen translocation between the host and parasite.

### Gene Duplication and Regulatory Neofunctionalization—Origin of Parasitism

The identification of haustorium genes allowed us to gain new insights into the evolution of parasitism. We have used phylogenetic analysis to show that a majority of the genes with a putative role in parasite functions arose by gene duplication. Most of the duplications occurred in a nonparasitic common ancestor of the parasitic Orobanchaceae species and *Mimulus* (a nonparasitic plant in the related nonparasitic family Phrymaceae) (Schaferhoff et al. 2010; Refulio-Rodriguez and Olmstead 2014). This suggests that either multiple independent gene duplications, or a WGD event occurring before the divergence of *Mimulus* and Orobanchaceae (Wickett et al. 2011), may have resulted in the diversification of genes important to haustorial development, and ultimately contributed to the rise of parasitic plants. In contrast, relatively few of the parasitism genes arose through duplications occurring in a more recent common ancestor of just the parasites or Orobanchaceae. The fact that the gene duplications that produced parasitism genes occurred in a nonparasitic ancestor, well before the origin of the parasitic Orobanchaceae about 32 Ma (Naumann et al. 2013), is consistent with the idea that

gene duplication does not immediately give rise to novel functions, as described recently by the WGD Radiation Lag-Time Model (Schranz et al. 2012).

We mapped expression data from multiple stages of parasite development onto parasite genes and found that under most circumstances, the two copies derived from a gene duplication event show different expression profiles. In addition, we interrogated the expression profiles of orthologous genes from tomato, *Medicago*, and *Arabidopsis*. By comparing the parasite duplicate's expression with orthologous gene expression in these related nonparasitic plants, we gained insight into how parasitism genes evolved, both in gene sequence and in gene expression. The fact that these parasitism genes shift their expression from root or flower in related nonparasitic plants to haustoria of parasitic plants, through gene duplication or otherwise (fig. 7 and supplementary fig. S2, Supplementary Material online), supports inferences of neofunctionalization in the evolution of parasitism (Conant and Wolfe 2008; Innan and Kondrashov 2010). While investigating the evolution of parasitism genes, we also found that a number of them play a role in symbiotic nodulation in nonparasites such as genes encoding ethylene responsive factor (ERF) TF (Vernie et al. 2008), oligopeptide transporter (Nogales et al. 2009), peroxidase, and pectinesterase inhibitor (Young et al. 2011; Zouari et al. 2014), suggesting possible parallels between the evolution of parasitism and that of mutualism, both of which involve invasion of host tissues. Similar evidence that *P. aegyptiaca* induced upregulation of genes involved in nodulation in *Lotus japonicus* supports this idea (Hiraoka et al. 2009).

### Origin of the Haustorium Involves Co-Option of Root and/or Flower Genes

The results of this study shed light on the origin of haustoria in Orobanchaceae, where molecular phylogenetic studies have identified the nonparasitic *Lindenbergia* as sister to the parasitic Orobanchaceae and are thus consistent with a single origin of haustorial parasitism in Orobanchaceae (dePamphilis et al. 1997; Olmstead et al. 2001; Schneeweiss et al. 2004; Bennett and Mathews 2006; Angiosperm Phylogeny Group 2009; McNeal et al. 2013). Two lines of evidence—global gene expression data in the parasites, and expression specificity in related nonparasitic plants—suggest that gene expression patterns in haustorial tissues are most similar to those of root. The second largest number of haustorial genes shows floral-specific expression patterns in nonparasitic models. These observations suggest a possible mechanism of parasitism through neofunctionalization, where genes with a role in root and floral biology in nonparasite species were co-opted to haustorial function in parasite species.

The root is a likely source for processes useful to subterranean haustorial structures. Both haustoria and roots operate underground, are physically adjacent, and haustoria are derived from apex of the primary root, sometimes from lateral root extensions (Heide-Jørgensen 2013b). Additionally, both haustoria and root are highly specialized organs for nutrient

uptake and transfer. The recruitment of many haustorial genes from those normally expressed in floral tissue such as pollen is more surprising, but the idea that haustorial growth is similar to the intrusive growth of pollen tubes was explored recently (Thorogood and Hiscock 2010; Pérez-de-Luque 2013). The authors propose that neighboring host cells recognize the parasite as alien without reacting against the “invasion,” similar to the way that plants recognize intrusive growth of pollen tubes (Thorogood and Hiscock 2010; Pérez-de-Luque 2013). Specifically we found that genes, like PLs that are used in polarized pollen tube growth to rapidly invade stylar tissue (Krichevsky et al. 2007), are expressed during the invasion of host tissue by the growing parasitic haustorium. One possible explanation is that the penetration peg of the haustorium, which grows rapidly into the host tissue, may have co-opted genes from the polarized, invasive growth found in the pollen tubes of flowers (Sampedro and Cosgrove 2005; Krichevsky et al. 2007; Honaas et al. 2013). It is also likely that some pectic enzymes involved in loosening the pollen tube cell walls so that they can elongate into the female reproductive tissues (Taniguchi et al. 1995) are recruited in the penetration and growth of the haustoria toward the host vascular tissue.

### Parasite-Specific Genes—Mobile Elements

Of the 84 parasite-specific orthogroups detected in our analysis, most contained sequences with no known function. However, almost all of the remaining sequences with an annotated Pfam domain have significant BLAST alignments with genes encoding proteins involved in the transfer of mobile elements, including retrotransposon gag protein, GAG-preintegrase domain, and a LAGLIDADG homing endonuclease (supplementary table S2, Supplementary Material online). Retrotransposon gag proteins are associated with the transposition of retrotransposons to telomere-associated structures in *Drosophila* (Rashkova et al. 2002), whereas GAG-preintegrase domain proteins are associated with chromosomal rearrangements by retrovirus insertion activity (Houzet et al. 2003). Interestingly, some LAGLIDADG endonucleases (Belfort and Roberts 1997) encoded by self-splicing group I introns are implicated in the highly mobile transfer and insertion of copies of the intron to specific target sequences. Intron homing by the LAGLIDADG endonuclease activity is implicated in the widespread horizontal gene transfer of the self-splicing group I intron in plant mitochondrial *cox1* genes (Vaughn et al. 1995; Cho et al. 1998; Barkman et al. 2007; Sanchez-Puerta et al. 2008, 2011). Thus, both transposable elements and homing introns by the retrotransposon gag proteins, GAG-preintegrase domain proteins, and LAGLIDADG homing endonuclease are involved in horizontal gene transfer (Daniels et al. 1990; Rodelsperger and Sommer 2011), supporting the possibility of a mechanistic link between parasitism in plants and at least some horizontal gene transfers (Barkman et al. 2007; Xi et al. 2013; Zhang, Fernandez-Aparicio, et al. 2013).



## Conclusion

In this article, we have shown that parasitic plants have evolutionarily recruited many genes for haustorial development and host penetration from genes that were involved in other processes in related nonparasitic plants, primarily root or flower development. These candidate parasitism genes are being functionally characterized to determine whether they are essential to parasite function and survival. The observation that genes with similar GO classifications (cell wall modification process and transporters) are also upregulated in the haustoria of *Cuscuta* (Ranjan et al. 2014) increases the likelihood that these genes do play important roles in haustorial function. In Orobanchaceae, genes recruited from root or pollen tube development show evidence of potentially adaptive changes in elevated dN/dS ratios and sites with excess nonsynonymous changes in parasitic lineages. The study of parasitic haustoria in Orobanchaceae indicates that two modes of regulatory neofunctionalization—either following gene duplication or in unduplicated orthogroup lineages—have provided the mechanism through which a novel structure has evolved.

## Materials and Methods

### Tissues, Libraries, and Sequence Data

Multiple stages of parasite development from the species *T. versicolor*, *S. hermonthica*, and *P. aegyptiaca* within Orobanchaceae were interrogated by transcriptome sequencing. Detailed descriptions of the stages ranging from seed and seedling, through haustorial development to above ground tissues such as leaf, stem, and flowering, are illustrated by Westwood et al. (2012) and in figure 1. Tissues from each stage were subjected to RNA extraction and library preparation, followed by subsequent Illumina paired-end sequencing (Honaas et al. 2013). Methods for Illumina and 454 paired-end mRNA-Seq library construction and sequencing are as described in Wickett et al. (2011). Additional Illumina transcriptome sequences were also obtained from a single normalized library for each species using pooled RNAs of all stages. Finally, a normalized whole plant library was also constructed and Illumina sequenced, as above, for *L. phillypensis*, representing the nonparasitic sister group to the parasitic Orobanchaceae (dePamphilis et al. 1997; Young et al. 1999; Olmstead et al. 2001; Schneeweiss et al. 2004; Angiosperm Phylogeny Group 2009; McNeal et al. 2013).

### Assembly, Cleaning, and Annotation (Including Gene Family Classification)

Duplicate reads in the Illumina sequence data were removed with CLC Assembly Cell version 3.2.0 (<http://www.clcbio.com/products/clc-assembly-cell/>, last accessed December 18, 2014). Adapters and bases with a quality score lower than Q20 were trimmed from the ends of the reads, and these reads were retained only if at least half of the sequence had quality  $\geq$ Q20. Raw Roche 454 sequence files in Standard Flowgram Format were converted to FASTA and associated quality files along with clipping of sequence adapters and low-

quality bases using sff\_extract version 0.2.10 ([http://bioinf.comav.upv.es/sff\\_extract/](http://bioinf.comav.upv.es/sff_extract/), last accessed December 18, 2014).

De novo assemblies of Illumina reads from each species were performed using Trinity release 2011-10-29 (Grabherr et al. 2011) and de novo hybrid assemblies of combined 454 and Illumina reads were performed using CLC Assembly Cell version 3.2.0 with default parameters. Assembled transcripts from both assemblies were combined by assigning hybrid CLC transcripts to Trinity components that yielded the best bit-score with BLASTN ( $E$  value =  $1e-10$ ). The resulting combined assemblies were filtered by removing contigs without coding regions (Iseli et al. 1999) as well as redundant transcripts (Edgar 2010). The assemblies for parasite species (*Triphysaria*, *Striga*, and *Phelipanche*) were then cleaned to remove contaminant sequences using a three-step process: 1) The transcripts were screened against the NCBI nonredundant protein database using BLASTX ( $E$  value =  $1e-5$ ) to remove nonplant transcripts, 2) the transcripts were then screened with BLASTN ( $E$  value =  $1e-10$ ) against a collection of publicly available genomes and expressed sequence tag data sets from the experimental host plants (to identify and remove host transcripts), and 3) after performing this screen on each of the parasite species, BLASTN ( $E$  value =  $1e-10$ ) of host candidate sequences was screened against the Orobanchaceae species (*Lindenbergia*, *Triphysaria*, *Striga*, and *Phelipanche*) databases (not including the parasite species being cleaned) to retain the transcripts that were much better matches to other Orobanchaceae family members than to the host plant. Open reading frames (ORFs) and protein sequences were predicted from the reconstructed transcript assemblies with ESTScan version 2.0 (Iseli et al. 1999). We experimented with different reference sequences to guide the ESTScan predictions, and other protein prediction programs such as GeneWise (Birney et al. 2004). Finally we chose ESTScan with an *Arabidopsis* reference to obtain the best balance of length and protein number in the resulting protein set.

The predicted protein sequences were used for BLASTP ( $E$  value =  $1e-5$ ) searches against Swissprot, TAIR10, and trEMBL databases to assign putative functional annotations in the form of human readable descriptions using the automated assignment of human readable descriptions (AHRD) pipeline (<https://github.com/groupschoof/AHRD>, last accessed December 18, 2014). AHRD uses similarity searches and lexical analysis for automatic AHRD to protein sequences. These translated transcripts were also annotated with Pfam domains using InterProScan version 4.8 (McDowall and Hunter 2011), and identified domains were directly translated into GO terms.

### Defining a Component-Orthogroup from Each De Novo Assembly

Very large transcriptome sequence data sets, including those produced by this project, result in complex de novo assemblies including many splice variants and distinct alleles (Grabherr et al. 2011). Due to the assembly complexity, we combined expression information for unigenes that were

assigned to the same Trinity component and mapped to the same orthogroup (Wickett et al. 2011). We call these unigenes from the same component and orthogroup a “component-orthogroup.”

### Read Mapping and Expression Normalization

High-quality nonredundant Illumina reads from individual stage-specific samples were independently mapped on each parasite’s postprocessed transcripts using CLC Genomic Workbench version 6.0.4 (parameters: mismatch cost = 2, insertion cost = 3, deletion cost = 3, length fraction = 0.5, similarity = 0.8, min insert size = 100, and max insert size = 300). Transcript abundance was then estimated using the CLC Genomic Workbench RNA-Seq program with unique reads counted for their matching transcripts, and nonspecifically mapped reads allocated on a proportional basis relative to the number of uniquely mapped reads. The numbers of reads mapped per library were normalized by the fragments per kilobase per million mapped reads (FPKM) (Mortazavi et al. 2008) method that corrects for biases in the total transcript size and normalizes for the total number of read sequences obtained in each sample library. The read counts and FPKM values of transcripts for each Trinity component classified as an orthogroup were summed up to obtain each component-orthogroup’s expression in each library.

### PV-Clustering of Global Transcriptional Profiles

To get an overall picture of the global gene expression profile, stages were clustered by the pvclust (Suzuki and Shimodaira 2006) command in R using the expression of all unigenes in each stage. Pvclust not only clusters stages but also infers confidence support with approximately unbiased multiscale resampling and bootstrap resampling (bs) values, obtained by resampling genes from the total population of unigenes.

### Identification of Differentially Expressed Genes for Candidate Parasite Gene Assignment

We first used a DE analysis to limit the number of genes for candidate parasitism gene identification. DE analysis was performed within each species using DESeq package in R (Anders and Huber 2010), which utilizes the negative binomial distribution to model the read count data for variance estimation. Only the stages that were shared among the three parasite species were used in this analysis (stage 0, 1, 2, 3, 4, 6.1, and 6.2). DE analysis using pairwise comparison was undertaken to identify genes that were differentially expressed in at least two stages.

### K-Means Clustering to Identify Clusters with High Expression in Haustoria

DE analysis resulted in a list of genes with varying expression patterns across the seven shared stages. K-means clustering was used to identify clusters of coexpressed features with high expression in haustoria tissue. The expression of the DE unigenes measured by log<sub>2</sub>FPKM in stages 0, 1, 2, 3, 4, 6.1, and 6.2 constituted an expression matrix as the input for clustering analysis. To determine the optimum number of clusters

needed to identify a single cluster with high expression in haustorium tissue, an R script was developed to generate a series of pdf files to represent each cluster’s expression profile for a total number of specified clusters ranging from 2 to 30. Each cluster contained a set of coexpressed genes whose expression in each stage was reflected in a boxplot with the median expression of the coexpressed genes connected by a line. The criteria used to find the appropriate number of clusters were the smallest number of clusters that enabled the visualization of a haustorial-specific (stage 3 and or stage 4) cluster. This means that for a smaller cluster number, we cannot identify the haustoria-specific pattern; a large cluster number may split the haustoria-specific cluster into several clusters, but is unnecessary in terms of gene identification.

### Hierarchical Clustering to Identify Putative Parasite Features within Each Species

After the appropriate number of clusters was chosen, hierarchical clustering was used to identify genes with expression patterns of interest; that cluster was extracted with the cutree function in R. Each cluster’s expression profile was determined by plotting a heat map using the heatmap.2 function from the gplots package in R.

### Self Organizing Maps (SOM Clustering) to Identify Patterns of Haustorial-Specific Expression

SOM clustering was used to maximize the identification of genes that show a high and specific expression pattern. The analysis was performed on the web server with the unsupervised learning SOM clustering of GenePattern developed by the Broad Institute (Reich et al. 2006). SOM clustering clearly reveals the overall pattern from the genome-wide gene expression data by reducing dimensionality of the original data. SOM clustering of GenePattern involved three steps: Data preprocessing, SOM clustering, and SOMClusterViewer. The expression matrix (log<sub>2</sub>FPKM) of the differentially expressed features was used as the input file for preprocessing, in which a row normalization and gene filtering were performed. The threshold and filter were performed with default parameters (floor: -3; ceiling: 18; min fold change: 1.5; min delta: 5). SOM clustering was performed by manually selecting the cluster-range. The default was chosen at the beginning and changed until the pattern with haustorial high and specific expression was revealed. The SOMClusterViewer displayed the expression profile of each cluster identified by SOM clustering. Finally for all three data sets from the three species, the cluster range was chosen at 6–8, which identified one or more clusters with high and specific expression in haustorial tissue.

### Enrichment Analysis—Parasite Genes Versus Whole Plant Background

The identification of a list of parasite genes allowed us to ask what biological functions are enriched compared with the whole plant background within each species. For each parasite gene, we obtained its BLASTx best hit in *Arabidopsis* and used these TAIR hits to identify GO assignments (Ashburner et al. 2000) and perform enrichment analysis. *Arabidopsis* was

selected for this analysis because of its relatively complete GO-term annotation. Enrichment analysis was performed with DAVID using Bonferoni adjusted *P* values for multiple tests (Huang et al. 2009) by comparing GO assignments for foreground (putative orthologs of parasite genes in *Arabidopsis*) versus background (all genes from the *Arabidopsis* genome). Enriched components with statistical significance, with annotations including Pfam domains, GO MF, GO BP, GO CC, and KEGG pathway, were identified for a set of shared parasite genes from each species.

### Phylogenies and Ka/Ks Constraint Analysis for Parasite Genes

Transcripts from the Orobanchaceae were assigned into orthogroups defined by 586,228 protein-coding genes of 22 representatives of sequenced land plant genomes (*Amborella* Genome Project 2013) using OrthoMCL. The selected taxa include nine rosids (*Ar. thaliana*, *Thellungiella parvula*, *Ca. papaya*, *Theobroma cacao*, *Po. trichocarpa*, *Fragaria vesca*, *Glycine max*, *M. truncatula*, and *V. vinifera*), three asterids (*So. lycopersicum*, *Sol. tuberosum*, and *Mimulus guttatus*), two basal eudicots (*Nelumbo nucifera* and *Aquilegia coerulea*), five monocots (*O. sativa*, *Brachypodium distachyon*, *Sor. bicolor*, *Musa acuminata*, and *Phoenix dactylifera*), one basal angiosperm (*A. trichopoda*), one lycophyte (*Selaginella moellendorffii*), and one moss (*Physcomitrella patens*). Of the plants with sequenced genomes, *Mimulus*, an emerging asterid model plant of family Phrymaceae, is the most closely related to Orobanchaceae (Schaferhoff et al. 2010; Refulio-Rodriguez and Olmstead 2014). Candidate orthogroups for unigenes from transcriptome assemblies of *Lindenbergia*, *Triphysaria*, *Striga*, *Phelipanche*, and two Asteraceae species, *Lactuca sativa* and *Helianthus annuus*, were identified by retaining BLASTP (McGinnis and Madden 2004) hits with *E* value  $\leq 1e-5$  for predicted peptide searches against the orthogroup-classified proteomes from those 22 sequenced plant genomes. HMM (Eddy 2011) searches of the translated transcripts were then performed on constructed candidate HMM orthogroup classification profiles, and orthogroups yielding the best bitscore were assigned to the transcripts. Once unigenes were found that had high and DE in haustoria, phylogenies were estimated for their corresponding orthogroups. Amino acid alignments of sequences within these orthogroups (including any translated transcripts that were assigned to the orthogroup as described above) were generated with MAFFT (Katoh and Standley 2013) and the corresponding DNA sequences were forced onto the amino acid alignments using a custom perl script. DNA alignments were then trimmed with trimAL (Capella-Gutierrez et al. 2009) to remove sites with less than 10% of the taxa. Orthogroup alignments were required to contain transcripts with alignment coverage of at least 50%. Otherwise, the failing transcripts were removed from the orthogroup amino acids and DNA alignments, and the alignments were regenerated. This process was iterated until all of the sequences covered at least 50% of the alignment. Finally, maximum likelihood (ML) phylogenetic trees of DNA alignments

for orthogroups containing parasite sequence(s) were generated using RAxML (Stamatakis 2006) with the GTRGAMMA model. To evaluate the reliability of the branches on the tree, 100 pseudosamples for the alignment were generated to estimate branch support using the bootstrap method (Felsenstein 1985).

### Scoring Gene Duplications

Gene duplication events were scored by referring to each rooted gene tree. Genes from *Physcomitrella* and/or *Selaginella* were used as outgroups to root each tree; when these outgroups were not present in the orthogroup, *Amborella* and/or monocots were used. Because our interest in this article was focused on when parasitism genes evolved, we limited our analysis to gene trees that contain parasitism genes. Possible topologies showing gene duplications giving rise to parasitism genes are illustrated in [supplementary data S15, Supplementary Material](#) online. Gene duplications were scored if a parasitism gene from one or more of the three parasitic species (*Triphysaria*, *Striga*, and *Phelipanche*) was present in each duplicated clade, and if bootstrap values for key nodes met defined criteria. In addition, a sequence from at least one taxon had to be present in each duplicated clade. To illustrate, a *Mimulus+Orobanchaceae*-wide duplication (including nonparasitic *Lindenbergia* and three parasites—*Triphysaria*, *Striga*, and *Phelipanche*), and shown as (((M1O1)bootstrap1, (M2O2)bootstrap2)bootstrap3), is required to meet the following criteria: 1) Each clade defined by the nodes M1O1 and M2O2 contains at least one gene from the parasite taxa; 2) at least one taxon of *Mimulus* or Orobanchaceae has to be present in both clades defined by M1O1 or M2O2; and 3) bootstrap2 and at least one of either bootstrap1 or bootstrap2 must be greater than or equal to the bootstrap stringency cutoffs of 50% or 80%. An example of a scored gene tree with a supported gene duplication is given in [supplementary figure S3, Supplementary Material](#) online.

### Selective Constraint Analysis

To perform the constraint analysis to infer adaptive or purifying selection, PAML (Yang 1997, 2007) software based on ML was utilized for hypothesis testing. The branch model in PAML was used to test whether the foreground branch of interest has significantly different dN/dS ratio ( $\omega$ ) compared with the background  $\omega$ . The codeml tool in PAML was used to perform such analyses. To estimate significance of one particular hypothesis, a likelihood ratio test was used. The one-ratio model and branch model in codeml of PAML were used to test whether the branch model fits the model significantly better than the one-ratio model. When the one-ratio model is correct, the distribution of the likelihood-ratio test statistic follows a chi-square distribution with the degrees of freedom being equal to the number of additional parameters in the branch model test. The test statistics was calculated by taking twice the difference between log likelihood-values from the two tests. These models were fitted to examine which model was more appropriate for the data. The branch-site model was further used when the

branch-model identified significant differences between the foreground and background lineages. The branch-site model was also used to identify sites under positive selection for the indicated foreground lineages. To perform the branch-site model, the codeml file was set to model = 2, NSsites = 2. The null model was set to fix  $\omega$  at 1, whereas the alternative model was set to estimate  $\omega$ . Sites identified by PAML with a probability greater than 0.95 by Bayes Empirical Bayes analysis were examined further through looking at the peptide alignment.

### Expression of Haustorial Orthologs in Nonparasitic Species

We utilized the existing expression profile data for growth and developmental stages in *Ar. thaliana*, *M. truncatula*, and *Sol. lycopersicum*, which we refer to as *Arabidopsis*, *Medicago*, and tomato in this analysis, respectively. The expression profiles for the *Arabidopsis* and *Medicago* genes were extracted from the PLEXdb database (Dash et al. 2012) that contains microarray data (*Arabidopsis*, AT40: Expression Atlas of *Arabidopsis* Development [AtGenExpress]; *Medicago*, ME1: The *M. truncatula* Gene Expression Atlas), whereas the data for tomato were from the digital expression (RNA-Seq) experiment (D007: Transcriptase analysis of various tissues in wild species *Sol. pimpinellifolium*, LA1589) in the Tomato Functional Genomics Database (Fei et al. 2011). First, parasite genes with upregulated haustorial expression were identified, and their putative orthologs were obtained as the best blast hits within the same orthogroup in *Arabidopsis*, *Medicago*, and tomato among the sequences. To find the expression information of the genes in *Arabidopsis* and *Medicago* used in the gene family analysis, we used the microarray expression information of probes for these genes by BLASTn. For tomato, expression for each gene was retrieved directly from the RNA-Seq database. To make the expression easily comparable across species, the expression values for similar tissues were averaged (supplementary data S20, Supplementary Material online). In the *Arabidopsis* gene expression atlas, all vegetative\_leaf and rosette\_leaf were combined as “leaf,” the sepal, petal, stamen, and carpel were combined as “flower,” different root tissues (root7, root\_MS1, and root\_GM) were combined as “root.” In the *Medicago* gene expression atlas, expression data for tissue responding to nod factors (Nod 4d, Nod 10d, and Nod 14d) were combined as “nod,” root and root-0d as “root,” and seed 10d, seed 12d, seed 16d, seed 20d, seed 36d as “seed.” In the tomato RNA-Seq data, newly developed leaves and mature green leaflets were combined and labeled as “leaf,” flower buds 10 days before anthesis or younger and flowers at anthesis (0 DPA) as “flower,” and fruit at 10 DPA, 20 DPA, 33 DPA as “fruit.” We scored the expression for upregulated haustorial orthogroups based on the principally expressed tissue. To do this, we divided the expression of a gene in a given tissue by its summed expression across all tissues to obtain a normalized expression in each tissue. Then, we identified genes as “principally expressed” in a tissue if its normalized expression was  $\geq 2$ -fold higher in that tissue than in any other. If similarly high expression was found in two tissues, both tissues were scored. Genes with broad expression in

more than three tissues were not scored. In addition to this binary scoring of the principally expressed tissue for haustorial genes, we also scored each tissue quantitatively using the average of each tissue’s expression across all haustorial gene orthologs. We excluded genes whose highest expressions across all tissues were in the lower 25th percentile. We then averaged expression of each tissue across all genes and ranked each tissue based on the expression. All orthogroups were subject to this step and finally each tissue type that supported a possible haustorial origin was scored by the number of upregulated haustorial orthogroups and the average expression across all orthogroups.

### Supplementary Material

Supplementary figures S1–S3, tables S1 and S2, and data S1–S23 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>) and the PGP website through the link provided ([http://ppgp.huck.psu.edu/5\\_Our\\_own\\_link\\_Supplementary\\_Data\\_Final\\_submission\\_2014.zip](http://ppgp.huck.psu.edu/5_Our_own_link_Supplementary_Data_Final_submission_2014.zip), last accessed December 18, 2014).

### Acknowledgments

Sequence data are archived at National Center for Biotechnology Information BioProject ID SRP001053, and at <http://ppgp.huck.psu.edu>. This research was supported by awards DBI-0701748 and IOS-1238057 to J.H.W., C.W.D., M.P.T., and J.H.Y. from the NSF’s Plant Genome Research Program, with the additional support for Y.Z., L.A.H., and H.Z. from the Plant Biology and Biology Department graduate programs at Penn State, the National Institute of Food and Agriculture Project No. 131997 to J.H.W., and from NSF IOS-1213059 for M.P.T. and K.H. M.F.-A. was supported by an International Outgoing European Marie Curie postdoctoral fellowship (PIOF-GA-2009-252538). The authors thank Yongde Bao for Illumina sequencing at the University of Virginia, Lynn Tomsho and Stephan Schuster for 454 and Illumina sequencing at Penn State University, and UC Davis Genome Center for generating normalized whole plant libraries and Illumina sequence data. They also thank Verlyn Stromberg for technical assistance, Yu Zhang for providing constructive ideas in candidate gene identification, Ningtao Wang and Iman Farasat for developing the initial R code for the cluster analyses, and Lenwood Heath and Yeting Zhang for helpful suggestions on the manuscript. Authors’ contribution is as follows: Conceived of project: J.H.W., C.W.D., J.I.Y., and M.P.T.; cultivated plants and generated staged tissue samples: M.D., M.F.A., P.C.G.B., K.H., L.A.H., and P.E.R.; RNAs and libraries: M.D., M.F.A., P.C.G.B., K.H., L.A.H., L.L., P.E.R., and B.W.; RT-PCR: Z.Y., H.Z., and C.R.C.; conceived of paper: C.W.D. and Z.Y.; designed and performed data analyses and results visualization: Z.Y. (primary), E.K.W. (primary), J.P.D., H.Z., N.S.A., and C.W.D. (primary); wrote the paper: Z.Y. (primary), C.W.D. (primary), and E.K.W.; additional text, editing, and comments: J.H.W., J.I.Y., M.P.T., N.S.A., C.R.C., P.E.R., L.A.H., M.F.A., K.H., and P.C.G.B.; Read and approved the text: all authors.

## References

- Aber M, Fer A, Salle G. 1983. Transfer of organic substances from the host plant *Vicia faba* to the parasite *Orobanche renata* Forsk. *Z Pflanzenphysiol.* 112:297–308.
- Alakonya A, Kumar R, Koenig D, Kimura S, Townsley B, Runo S, Garces HM, Kang J, Yanez A, David-Schwartz R, et al. 2012. Interspecific RNA interference of *SHOOT MERISTEMLESS-Like* disrupts *Cuscuta pentagona* plant parasitism. *Plant Cell* 24:3153–3166.
- Aly R, Cholak H, Joel DM, Leibman D, Steinitz B, Zelcer A, Naglis A, Yarden O, Gal-On A. 2009. Gene silencing of mannose 6-phosphate reductase in the parasitic weed *Orobanche aegyptiaca* through the production of homologous dsRNA sequences in the host plant. *Plant Biotechnol J.* 7:487–498.
- Aly R, Hamamouch N, Abu-Nassar J, et al. 2011. Movement of protein and macromolecules between host plants and the parasitic weed *Phelipanche aegyptiaca* Pers. *Plant Cell Rep.* 30:2233–2241.
- Amborella Genome Project. 2013. The *Amborella* genome and the evolution of flowering plants. *Science* 342:1241089.
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.
- Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc.* 161:105–121.
- Antao CM, Malcata FX. 2005. Plant serine proteases: biochemical, physiological and molecular features. *Plant Physiol Biochem.* 43:637–650.
- Antonova TS, TerBorg SJ. 1996. The role of peroxidase in the resistance of sunflower against *Orobanche cumana* in Russia. *Weed Res.* 36: 113–121.
- Ascencio-Ibanez JT, Sozzani R, Lee TJ, Chu TM, Wolfinger RD, Cella R, Hanley-Bowdoin L. 2008. Global analysis of *Arabidopsis* gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. *Plant Physiol.* 148: 436–454.
- Ashburner M, Ball CA, Blake JA, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 25:25–29.
- Atsatt PR. 1973. Parasitic flowering plants: how did they evolve? *Am Nat.* 107:502–510.
- Baird WV, Riopel JL. 1984. Experimental studies of haustorium initiation and early development in *Agalinis purpurea* (L) Raf (Scrophulariaceae). *Am J Bot.* 71:803–814.
- Bandaranayake PCC, Filappova T, Tomilov A, Tomilova NB, Jamison-McClung D, Ngo Q, Inoue K, Yoder JI. 2010. A single-electron reducing quinone oxidoreductase is necessary to induce haustorium development in the root parasitic plant *Triphysaria*. *Plant Cell* 22: 1404–1419.
- Bandaranayake PCC, Tomilov A, Tomilova NB, Ngo QA, Wickett N, dePamphilis CW, Yoder JI. 2012. The *TvPirin* gene is necessary for haustorium development in the parasitic plant *Triphysaria versicolor*. *Plant Physiol.* 158:1046–1053.
- Bandaranayake PCC, Yoder J. 2013a. Evolutionary origins. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobancheaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 69–70.
- Bandaranayake PCC, Yoder JI. 2013b. Trans-specific gene silencing of acetyl-CoA carboxylase in a root-parasitic plant. *Mol Plant Microbe Interact.* 26:575–584.
- Barkman TJ, McNeal JR, Lim SH, Coat G, Croom HB, Young ND, dePamphilis CW. 2007. Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evol Biol.* 7:248.
- Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25:3379–3388.
- Bennett JR, Mathews S. 2006. Phylogeny of the parasitic plant family Orobancheaceae inferred from phytochrome A. *Am J Bot.* 93: 1039–1051.
- Birney E, Clamp M, Durbin R. 2004. GeneWise and genomewise. *Genome Res.* 14:988–995.
- Bleichschwitz M, Albert M, Fuchsbauer HL, Kaldenhoff R. 2010. Significance of Cuscutain, a cysteine protease from *Cuscuta reflexa*, in host-parasite interactions. *BMC Plant Biol.* 10:227.
- Brini M, Carafoli E. 2011. The plasma membrane Ca<sup>2+</sup> ATPase and the plasma membrane sodium calcium exchanger cooperate in the regulation of cell calcium. *Cold Spring Harb Perspect Biol.* 3:a004168.
- Brinkworth RI, Prociw P, Loukas A, Brindley PJ. 2001. Hemoglobin-degrading aspartic proteases of blood-feeding parasites—substrate specificity revealed by homology models. *J Biol Chem.* 276: 38844–38851.
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.
- Chen J, Ouyang Y, Wang L, Xie W, Zhang Q. 2009. Aspartic proteases gene family in rice: gene structure and expression, predicted protein features and phylogenetic relation. *Gene* 442:108–118.
- Cho Y, Qiu YL, Kuhlman P, Palmer JD. 1998. Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci U S A.* 95: 14244–14249.
- Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9:938–950.
- Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? *Mol Biol Evol.* 23: 887–892.
- Cui L, Wall PK, Leebens-Mack JH, et al. 2006. Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16:738–749.
- Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. 1990. Evidence for horizontal transmission of the P-transposable element between *Drosophila* species. *Genetics* 124:339–355.
- Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA. 2012. PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res.* 40:D1194–D1201.
- Delaux PM, Varala K, Edger PP, Coruzzi GM, Pires JC, Ane JM. 2014. Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* 10:e1004487.
- Delorenzo G, Cervone F, Hahn MG, Darvill A, Albersheim P. 1991. Bacterial endopectate lyase—evidence that plant-cell wall pH prevents tissue maceration and increases the half-life of elicitor-active oligogalacturonides. *Physiol Mol Plant Pathol.* 39:335–344.
- dePamphilis CW, Young ND, Wolfe AD. 1997. Evolution of plastid gene *rps2* in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. *Proc Natl Acad Sci U S A.* 94:7367–7372.
- Der JP, Barker MS, Wickett NJ, dePamphilis CW, Wolf PG. 2011. *De novo* characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. *BMC Genomics* 12:99.
- Dorr I. 1997. How *Striga* parasitizes its host: a TEM and SEM study. *Ann Bot.* 79:463–472.
- Dorr I, Kollmann R. 1995. Symplasmic sieve element continuity between *Orobanche* and its host. *Bot Acta.* 108:47–55.
- Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Mol Biol Evol.* 23: 469–478.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol.* 7: e1002195.
- Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461.
- Eizenberg H, Aly R, Cohen Y. 2012. Technologies for smart chemical control of broomrape (*Orobanche* spp. and *Phelipanche* spp.). *Weed Sci.* 60:316–323.
- Fei ZJ, Joung JG, Tang XM, et al. 2011. Tomato Functional Genomics Database: a comprehensive resource and analysis package for tomato functional genomics. *Nucleic Acids Res.* 39: D1156–D1163.
- Felsenstein J. 1985. Confidence-limits on phylogenies—an approach using the bootstrap. *Evolution* 39:783–791.

- Fernandez-Aparicio M, Huang K, Wafula EK, Honaas LA, Wickett NJ, Timko MP, dePamphilis CW, Yoder JJ, Westwood JH. 2013. Application of qRT-PCR and RNA-Seq analysis for the identification of housekeeping genes useful for normalization of gene expression values during *Striga hermonthica* development. *Mol Biol Rep.* 40: 3395–3407.
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.
- Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14: 1457–1467.
- Gehring WJ. 2011. Chance and necessity in eye evolution. *Genome Biol Evol.* 3:1053–1066.
- Goda H, Sawa S, Asami T, Fujioka S, Shimada Y, Yoshida S. 2004. Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in *Arabidopsis*. *Plant Physiol.* 134:1555–1573.
- Grabherr MG, Haas BJ, Yassour M, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.
- He XL, Zhang JZ. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164.
- Heide-Jørgensen HS. 2013a. Introduction: the parasitic syndrome in higher plants. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 1.
- Heide-Jørgensen HS. 2013b. The haustorium. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 4.
- Henrissat B, Callebaut I, Fabrega S, Lehn P, Morion JP, Davies G. 1995. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci U S A.* 92:7090–7094.
- Higgins TJV. 1984. Synthesis and regulation of major proteins in seeds. *Annu Rev Plant Phys.* 35:191–221.
- Hiraoka Y, Ueda H, Sugimoto Y. 2009. Molecular responses of *Lotus japonicus* to parasitism by the compatible species *Orobanche aegyptiaca* and the incompatible species *Striga hermonthica*. *J Exp Bot.* 60: 641–650.
- Honaas LA. 2013. Tissue specific *de novo* transcriptomics in the parasitic Orobanchaceae [dissertation]. University Park, Pennsylvania: The Pennsylvania State University.
- Honaas LA, Wafula EK, Yang Z, et al. 2013. Functional genomics of a generalist parasitic plant: laser microdissection of host-parasite interface reveals host-specific patterns of parasite gene expression. *BMC Plant Biol.* 13:9.
- Houzet L, Battini JL, Bernard E, Thibert V, Mougél M. 2003. A new retroelement constituted by a natural alternatively spliced RNA of murine replication-competent retroviruses. *EMBO J.* 22: 4866–4875.
- Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.
- Huang JY, Zhao XB, Cheng K, Jiang YH, Ouyang YD, Xu CG, Li XH, Xiao JH, Zhang QF. 2013. OsAP65, a rice aspartic protease, is essential for male fertility and plays a role in pollen germination and pollen tube growth. *J Exp Bot.* 64:3351–3360.
- Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11:97–108.
- Iseli C, Jongeneel CV, Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol.* 138–148.
- Jiao Y, Leebens-Mack J, Ayyampalayam S, et al. 2012. A genome triplification associated with early diversification of the core eudicots. *Genome Biol.* 13:R3.
- Jiao Y, Wickett NJ, Ayyampalayam S, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97–100.
- Joel DM. 2013. Functional structure of the mature haustorium. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 54–55.
- Joel DM, Losner-Goshen D. 1994a. The attachment organ of the parasitic angiosperms *Orobanche cumana* and *O. aegyptiaca* and its development. *Can J Bot.* 72:564–574.
- Joel DM, Losner-Goshen D, editors 1994b. In: Proceedings of the Third International Workshop on *Orobanche* and Related *Striga* Research. Biology and management of *Orobanche*, Amsterdam, The Netherlands, November 8–12, 1993. Amsterdam: Royal Tropical Institute.
- Joel DM, Losner-Goshen D, Goldman-Guez T, Portnoy VH, editors 1998. In: Proceedings of the 4th International Workshop on *Orobanche*. Current problems in *Orobanche* research, September 23–26, 1998, Bulgaria. Albena (Bulgaria): Institute for Wheat and Sunflower Dobroudja.
- Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.
- Kall L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* 35:W429–W432.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.
- Kennan RM, Wong W, Dhungyel OP, et al. 2010. The subtilisin-like protease AprV2 is required for virulence and uses a novel disulphide-tethered exosite to bind substrates. *PLoS Pathog.* 6: e1001210.
- Kim MJ, Ciani S, Schachtman DP. 2010. A peroxidase contributes to ROS production during *Arabidopsis* root response to potassium deficiency. *Mol Plant.* 3:420–427.
- Kozik A, Matvienko M, Kozik I, Vvan Leeuwen H, Van DDeynze A, Michelmore RM, editors 2008. Plant and Animal Genome Conference. 12–16, January 2008, San Diego, CA, USA.
- Krichevsky A, Kozlovsky SV, Tian GW, Chen MH, Zaltsman A, Citovsky V. 2007. How pollen tubes grow. *Dev Biol.* 303:405–420.
- Kuijt J. 1969. The biology of parasitic flowering plants. Berkeley (CA): University of California Press.
- Kuijt J. 1977. Haustoria of phanerogamic parasites *Annu Rev Phytopathol.*, Vol. 15. p 91–118.
- Lamesch P, Berardini TZ, Li D, et al. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40:D1202–D1210.
- Lechat MM, Pouvreau JB, Peron T, et al. 2012. *PrCYP707A1*, an ABA catabolic gene, is a key component of *Phelipanche ramosa* seed germination in response to the strigolactone analogue GR24. *J Exp Bot.* 63:5311–5322.
- Liberles DA, Kolesov G, Dittmar K. 2010. Understanding gene duplication through biochemistry and population genetics. In: Dittmar K, Liberles D, editors. Evolution after gene duplication. Hoboken (NJ): John Wiley & Sons. p. 10–11.
- Liu Q, Zhang Y, Matusova R, et al. 2014. *Striga hermonthica* MAX2 restores branching but not the very low fluence response in the *Arabidopsis thaliana* max2 mutant. *New Phytol.* 202:531–541.
- Llorente F, Lopez-Cobollo RM, Catala R, Martinez-Zapater JM, Salinas J. 2002. A novel cold-inducible gene from *Arabidopsis*, RCI3, encodes a peroxidase that constitutes a component for stress tolerance. *Plant J.* 32:13–24.
- Losner-Goshen D, Portnoy VH, Mayer AM, Joel DM. 1998. Pectolytic activity by the haustorium of the parasitic plant *Orobanche* L. (Orobanchaceae) in host roots. *Ann Bot.* 81:319–326.
- Ludeman DA, Farrar N, Riesgo A, Paps J, Leys SP. 2014. Evolutionary origins of sensation in metazoans: functional evidence for a new sensory organ in sponges. *BMC Evol Biol.* 14:3.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
- Marin-Rodriguez MC, Orchard J, Seymour GB. 2002. Pectate lyases, cell wall degradation and fruit softening. *J Exp Bot.* 53:2115–2119.

- Matvienko M, Torres MJ, Yoder JI. 2001. Transcriptional responses in the hemiparasitic plant *Triphysaria versicolor* to host plant signals. *Plant Physiol.* 127:272–282.
- McDowall J, Hunter S. 2011. InterPro protein classification. *Methods Mol Biol.* 694:37–47.
- McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32: W20–W25.
- McNeal JR, Bennett JR, Wolfe AD, Mathews S. 2013. Phylogeny and origins of holoparasitism in Orobanchaceae. *Am J Bot.* 100:971–983.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 5:621–628.
- Musselman LJ, Dickison WC. 1975. The structure and development of the haustorium in parasitic Scrophulariaceae. *Bot J Linn Soc.* 70: 183–212.
- Musselman LJ, Hepper FN. 1986. The witchweeds (*Striga*, Scrophulariaceae) of the Sudan Republic. *Kew Bull.* 41:205–221.
- Naumann J, Salomo K, Der JP, et al. 2013. Single-copy nuclear genes place haustorial Hydnoraceae within piperales and reveal a cretaceous origin of multiple parasitic angiosperm lineages. *PLoS One* 8: e79204.
- Nei M. 1969. Gene duplication and nucleotide substitution in evolution. *Nature* 221:40–42.
- Nei M, editor. 2013. Mutation-driven evolution. Oxford: Oxford University Press. p. 156–157.
- Nelson DC. 2013. Are karrikin signaling mechanisms relevant to strigolactone perception?. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 221–230.
- Nelson DC, Scaffidi A, Dun EA, Waters MT, Flematti GR, Dixon KW, Beveridge CA, Ghisalberti EL, Smith SM. 2011. F-box protein MAX2 has dual roles in karrikin and strigolactone signaling in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A.* 108: 8897–8902.
- Neumann U, Vian B, Weber HC, Salle G. 1999. Interface between haustoria of parasitic members of the Scrophulariaceae and their hosts: a histochemical and immunocytochemical approach. *Protoplasma* 207:84–97.
- Ngo QA, Albrecht H, Tsuchimatsu T, Grossniklaus U. 2013. The differentially regulated genes *TvQR1* and *TvPirin* of the parasitic plant *Triphysaria* exhibit distinctive natural allelic diversity. *BMC Plant Biol.* 13:28.
- Nickrent DL, Musselman LJ, Riopel JL, Eplee RE. 1979. Haustorial initiation and non-host penetration in witchweed (*Striga asiatica*). *Ann Bot.* 43:233–236.
- Nogales J, Munoz S, Olivares J, Sanjuan J. 2009. Genetic characterization of oligopeptide uptake systems in *Sinorhizobium meliloti*. *FEMS Microbiol Lett.* 293:177–187.
- Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
- Olmstead RG, dePamphilis CW, Wolfe AD, Young ND, Elisons WJ, Reeves PA. 2001. Disintegration of the Scrophulariaceae. *Am J Bot.* 88:348–361.
- Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet.* 34:401–437.
- Pageau K, Simier P, Le Bizec B, Robins RJ, Fer A. 2003. Characterization of nitrogen relationships between *Sorghum bicolor* and the root-hemi-parasitic angiosperm *Striga hermonthica* (Del.) Benth. using (KNO<sub>3</sub>)-N-15 as isotopic tracer. *J Exp Bot.* 54:789–799.
- Pearce G, Yamaguchi Y, Barona G, Ryan CA. 2010. A subtilisin-like protein from soybean contains an embedded, cryptic signal that activates defense-related genes. *Proc Natl Acad Sci U S A.* 107: 14921–14925.
- Pérez-de-Luque A. 2013. Haustorium invasion into host tissues. In: Joel DM, Gressel J, Musselman LJ, editors. Parasitic Orobanchaceae—parasitic mechanisms and control strategies. Heidelberg (Germany): Springer. p. 82–83.
- Racine JS. 2012. RStudio: a platform-independent IDE for R and Sweave. *J Appl Econ.* 27:167–172.
- Ranjan A, Ichihashi Y, Farhi M, Zumstein K, Townsley B, David-Schwartz R, Sinha NR. 2014. *De novo* assembly and characterization of the transcriptome of the parasitic weed *Cuscuta pentagona* identifies genes associated with plant parasitism. *Plant Physiol.* 166: 1186–1199.
- Rashkova S, Karam SE, Kellum R, Pardue ML. 2002. Gag proteins of the two *Drosophila* telomeric retrotransposons are targeted to chromosome ends. *J Cell Biol.* 159:397–402.
- Refulio-Rodriguez NF, Olmstead RG. 2014. Phylogeny of Lamiidae. *Am J Bot.* 101:287–299.
- Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. 2006. GenePattern 2.0. *Nat Genet.* 38:500–501.
- Reiss GC, Bailey JA. 1998. *Striga gesnerioides* parasitising cowpea: development of infection structures and mechanisms of penetration. *Ann Bot.* 81:431–440.
- Rivera AS, Pankey MS, Plachetzki DC, Villacorta C, Syme AE, Serb JM, Omilian AR, Oakley TH. 2010. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC Evol Biol.* 10:123.
- Rodelsperger C, Sommer RJ. 2011. Computational archaeology of the *Pristionchus pacificus* genome reveals evidence of horizontal gene transfers from insects. *BMC Evol Biol.* 11:239.
- Rumsey FJ, Jury SL. 1991. An account of *Orobanche* L. in Britain and Ireland. *Watsonia* 18:257–295.
- Salvini-Plawen L, Mayer E. 1961. On the evolution of photoreceptors and eyes. *Evol Biol.* 10:63–207.
- Sampedro J, Cosgrove DJ. 2005. The expansin superfamily. *Genome Biol.* 6:242.
- Sanchez-Puerta MV, Abbona CC, Zhuo S, Tepe EJ, Bohs L, Olmstead RG, Palmer JD. 2011. Multiple recent horizontal transfers of the *cox1* intron in Solanaceae and extended co-conversion of flanking exons. *BMC Evol Biol.* 11:277.
- Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD. 2008. Frequent, phylogenetically local horizontal transfer of the *cox1* group I intron in flowering plant mitochondria. *Mol Biol Evol.* 25: 1762–1777.
- Schaferhoff B, Fleischmann A, Fischer E, Albach DC, Borsch T, Heubl G, Muller KF. 2010. Towards resolving Lamiales relationships: insights from rapidly evolving chloroplast sequences. *BMC Evol Biol.* 10:352.
- Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* 47:868–876.
- Schneeweiss GM, Colwell A, Park JM, Jang CG, Stuessy TF. 2004. Phylogeny of holoparasitic *Orobanche* (Orobanchaceae) inferred from nuclear ITS sequences. *Mol Phylogenet Evol.* 30:465–478.
- Scholes JD, Press MC. 2008. *Striga* infestation of cereal crops—an unsolved problem in resource limited agriculture. *Curr Opin Plant Biol.* 11:180–186.
- Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr Opin Plant Biol.* 15:147–153.
- Singh A, Singh M. 1993. Cell-wall degrading enzymes in *Orobanche aegyptiaca* and its host *Brassica campestris*. *Physiol Plant.* 89:177–181.
- Smith S, Stewart GR. 1990. Effect of potassium levels on the stomatal behavior of the hemi-parasite *Striga hermonthica*. *Plant Physiol.* 94: 1472–1476.
- Soltis DE, Albert VA, Leebens-Mack J, et al. 2009. Polyploidy and angiosperm diversification. *Am J Bot.* 96:336–348.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Stephens SG. 1951. Possible significances of duplication in evolution. *Adv Genet.* 4:247–265.
- Sun L, van Nocker S. 2010. Analysis of promoter activity of members of the PECTATE LYASE-LIKE (PLL) gene family in cell separation in *Arabidopsis*. *BMC Plant Biol.* 10:152.
- Suzuki R, Shimodaira H. 2006. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22: 1540–1542.

- Taniguchi Y, Ono A, Sawatani M, Nanba M, Kohno K, Usui H, Kurimoto M, Matuhasi T. 1995. Cry j I, a major allergen of Japanese cedar pollen, has pectate lyase enzyme activity. *Allergy* 50:90–93.
- Thorogood CJ, Hiscock SJ. 2010. Compatibility interactions at the cellular level provide the basis for host specificity in the parasitic plant *Orobanchae*. *New Phytol.* 186:572–575.
- Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neo-functionalization of yeast duplicates. *Genome Biol.* 8:R50.
- Torres MA, Jones JD, Dangl JL. 2006. Reactive oxygen species signaling in response to pathogens. *Plant Physiol.* 141:373–378.
- Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions revisited. *Mol Biol Evol.* 30:177–190.
- Vaughn JC, Mason MT, Sper-Whitis GL, Kuhlman P, Palmer JD. 1995. Fungal origin by horizontal transfer of a plant mitochondrial group I intron in the chimeric *Coxl* gene of *Peperomia*. *J Mol Evol.* 41: 563–572.
- Vernie T, Moreau S, de Billy F, et al. 2008. EFD Is an ERF transcription factor involved in the control of nodule number and differentiation in *Medicago truncatula*. *Plant Cell* 20:2696–2713.
- Volpi C, Janni M, Lionetti V, Bellincampi D, Favaron F, D'Ovidio R. 2011. The ectopic expression of a pectin methyl esterase inhibitor increases pectin methyl esterification and limits fungal diseases in wheat. *Mol Plant Microbe Interact.* 24:1012–1019.
- Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW. 2008. PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.* 36:D970–D976.
- Westwood JH, dePamphilis CW, Das M, Fernandez-Aparicio M, Honaas LA, Timko MP, Wafula EK, Wickett NJ, Yoder JI. 2012. The Parasitic Plant Genome Project: new tools for understanding the biology of *Orobanche* and *Striga*. *Weed Sci.* 60:295–306.
- Westwood JH, Yoder JI, Timko MP, dePamphilis CW. 2010. The evolution of parasitism in plants. *Trends Plant Sci.* 15:227–235.
- Wickett NJ, Honaas LA, Wafula EK, et al. 2011. Transcriptomes of the parasitic plant family Orobanchaceae reveal surprising conservation of chlorophyll synthesis. *Curr Biol.* 21:2098–2104.
- Williams JS, Der JP, dePamphilis CW, Kao TH. 2014. Transcriptome analysis reveals the same 17 S-Locus F-box genes in two haplotypes of the self-incompatibility locus of *Petunia inflata*. *Plant Cell* 26: 2873–2888.
- Wu F, Mueller LA, Crouzillat D, Petiard V, Tanksley SD. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* 174:1407–1420.
- Xi Z, Wang Y, Bradley RK, Sugumaran M, Marx CJ, Rest JS, Davis CC. 2013. Massive mitochondrial gene transfer in a parasitic flowering plant clade. *PLoS Genet.* 9:e1003265.
- Xie F, Murray JD, Kim J, Heckmann AB, Edwards A, Oldroyd GED, Downie A. 2012. Legume pectate lyase required for root infection by rhizobia. *Proc Natl Acad Sci U S A.* 109:633–638.
- Yan LJ, Levine RL, Sohal RS. 1997. Oxidative damage during aging targets mitochondrial aconitase. *Proc Natl Acad Sci U S A.* 94: 11168–11172.
- Yang ZH. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
- Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
- Yoshida S, Ishida JK, Kamal NM, Ali AM, Namba S, 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.
- Young ND, Debelle F, Oldroyd GED, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature* 480:520–524.
- Young ND, Steiner KE, dePamphilis CW. 1999. The evolution of parasitism in Scrophulariaceae/Orobanchaceae: plastid gene sequences refute an evolutionary transition series. *Ann Mo Bot Gard.* 86: 876–893.
- Zhang DL, Qi JF, Yue JP, et al. 2014. Root parasitic plant *Orobanche aegyptiaca* and shoot parasitic plant *Cuscuta australis* obtained Brassicaceae-specific strictosidine synthase-like genes by horizontal gene transfer. *BMC Plant Biol.* 14
- Zhang GJ, Cowled C, Shi ZL, et al. 2013. Comparative analysis of *Bat* genomes provides insight into the evolution of flight and immunity. *Science* 339:456–460.
- Zhang Y, Fernandez-Aparicio M, Wafula EK, et al. 2013. Evolution of a horizontally acquired legume gene, albumin 1, in the parasitic plant *Phelipanche aegyptiaca* and related species. *BMC Evol Biol.* 13:48.
- Zhang ZY. 1988. Taxonomy of the Chinese *Orobanche* and its relationships with related genera. *Acta Phytotax Sin.* 26:394–403.
- Zouari I, Salvioli A, Chialva M, Novero M, Miozzi L, Tenore GC, Bagnaresi P, Bonfante P. 2014. From root to fruit: RNA-Seq analysis shows that arbuscular mycorrhizal symbiosis may affect tomato fruit metabolism. *BMC Genomics* 15.