The Genome of *Artemisia annua* Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis

Qian Shen[1,7], Lida Zhang[1,7], Zhihua Liao[2,7], Shengyue Wang[3], Tingxiang Yan[1], Pu Shi[1], Meng Liu[1], Xueqing Fu[1], Qifang Pan[1], Yuliang Wang[1], Zongyou Lv[1], Xu Lu[1,6], Fangyuan Zhang[1,2], Weimin Jiang[1], Yanan Ma[1], Minghui Chen[1], Xiaolong Hao[1], Ling Li[1], Yueli Tang[1], Gang Lv[3], Yan Zhou[3], Xiaofen Sun[1], Peter E. Brodelius[5], Jocelyn K.C. Rose[4] and Kexuan Tang[1,*]

[1]Joint International Research Laboratory of Metabolic & Developmental Sciences, Key Laboratory of Urban Agriculture (South) Ministry of Agriculture, Plant Biotechnology Research Center, Fudan-SJTU-Nottingham Plant Biotechnology R&D Center, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]SWU-TAAHC Medicinal Plant Joint R&D Centre, School of Life Sciences, Southwest University, Chongqing 400715, China

[3]Chinese National Human Genome Center at Shanghai, Shanghai 201203, China

[4]Plant Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY 14853, USA

[5]Department of Chemistry and Biomedical Sciences, Linnaeus University, 39182 Kalmar, Sweden

[6]State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 211198, China

[7]These authors contributed equally to this article.

***Correspondence: Kexuan Tang (kxtang@sjtu.edu.cn)**

https://doi.org/10.1016/j.molp.2018.03.015

ABSTRACT

***Artemisia annua*, commonly known as sweet wormwood or Qinghao, is a shrub native to China and has long been used for medicinal purposes. *A. annua* is now cultivated globally as the only natural source of a potent anti-malarial compound, artemisinin. Here, we report a high-quality draft assembly of the 1.74-gigabase genome of *A. annua*, which is highly heterozygous, rich in repetitive sequences, and contains 63 226 protein-coding genes, one of the largest numbers among the sequenced plant species. We found that, as one of a few sequenced genomes in the Asteraceae, the *A. annua* genome contains a large number of genes specific to this large angiosperm clade. Notably, the expansion and functional diversification of genes encoding enzymes involved in terpene biosynthesis are consistent with the evolution of the artemisinin biosynthetic pathway. We further revealed by transcriptome profiling that *A. annua* has evolved the sophisticated transcriptional regulatory networks underlying artemisinin biosynthesis. Based on comprehensive genomic and transcriptomic analyses we generated transgenic *A. annua* lines producing high levels of artemisinin, which are now ready for large-scale production and thereby will help meet the challenge of increasing global demand of artemisinin.**

Keywords: *Artemisia annua*, artemisinin, genome, evolution, transcriptome, metabolic engineering

Shen Q., Zhang L., Liao Z., Wang S., Yan T., Shi P., Liu M., Fu X., Pan Q., Wang Y., Lv Z., Lu X., Zhang F., Jiang W., Ma Y., Chen M., Hao X., Li L., Tang Y., Lv G., Zhou Y., Sun X., Brodelius P.E., Rose J.K.C., and Tang K. (2018). The Genome of *Artemisia annua* Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis. Mol. Plant. **11**, 776–788.

INTRODUCTION

Malaria is a global health problem: in 2016 alone, there were an estimated 216 million new cases of malaria, 445,000 deaths, and nearly 1 one billion people living in areas with a high risk of the disease (World Health Organization, 2017). Artemisinin, an endoperoxide sesquiterpene lactone, is an effective anti-malarial compound that is synthesized in the glandular trichomes of the Chinese medicinal plant *Artemisia annua*. Due to her discovery of the anti-malaria function of artemisinin, which has

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

saved millions of lives, the Chinese scientist Youyou Tu received a Nobel Prize in Physiology or Medicine in 2015. Artemisinin-based combination therapies (ACTs) are recommended by the WHO (World Health Organization, 2017) for treatment of uncomplicated malaria caused by the *Plasmodium falciparum* parasite (Duffy and Mutabingwa, 2006).

Currently, the supply of ACTs is reliant on the agricultural production of artemisinin. However, plant-based production sometimes cannot meet the global demand due to the low amount of artemisinin produced in *A. annua* leaves (0.1%–1.0% of dry weight). Alternatively, a semi-synthetic system can be used for the production of artemisinin, in which yeast are engineered to synthesize its precursor, artemisinic acid (Ro et al., 2006; Paddon et al., 2013). However, the semi-synthetic production of artemisinin is expensive and thus cannot replace its agricultural production at present (Peplow, 2016). The use of the whole *Artemisia* plant as a malaria therapy was found to be more effective than a comparable dose of pure artemisinin, and was shown to be able to overcome resistance to pure artemisinin in a rodent malaria model and human clinical trial (Elfawal et al., 2015; Daddy et al., 2017). Moreover, parasite resistance to artemisinin has recently been confirmed by the WHO in the Greater Mekong subregion (World Health Organization, 2017). Besides its anti-malarial activity, many other therapeutic effects of artemisinin on diseases such as cancer (Efferth, 2006; Tin et al., 2012), tuberculosis (Zheng et al., 2017), and diabetes (Li et al., 2017) have been reported. Hence, artemisinin is a potential multi-functional compound and is of high medicinal value. There is a considerable interest in increasing the artemisinin content of *A. annua* and an urgent need to identify other potential anti-malarial compounds.

Artemisinin is synthesized from isopentenyl pyrophosphate via farnesyl pyrophosphate, which is converted to amorpha-4,11-diene by the action of amorpha-4,11-diene synthase (ADS) (Chang et al., 2000). The next reaction is the three-step oxidation of amorpha-4,11-diene to artemisinic acid, via artemisinic alcohol and artemisinic aldehyde, through the action of a cytochrome P450 monooxygenase (CYP71AV1) (Teoh et al., 2006). Artemisinic aldehyde is converted by artemisinic aldehyde Ä11(13) reductase (a double-bond reductase, DBR2) into dihydroartemisinic aldehyde (Zhang et al., 2008). ALDH1 (aldehyde dehydrogenase 1) (Teoh et al., 2009) then catalyzes the oxidation of artemisinic aldehyde and dihydroartemisinic aldehyde to produce artemisinic acid and dihydroartemisinic acid, respectively (Supplemental Figure 1). The conversion of dihydroartemisinic acid to artemisinin occurs via a light-induced non-enzymatic photochemical oxidation process (Czechowski et al., 2016).

In the last two decades, metabolic engineering has been demonstrated to be a useful approach to increase artemisinin content in *A. annua*. Previous studies employed several metabolic engineering strategies to enhance artemisinin production, including overexpression of artemisinin biosynthetic pathway genes (Banyai et al., 2010; Nafis et al., 2011; Lu et al., 2013a; Ma et al., 2015), overexpression of transcription factors (TFs) that can enhance the expression of artemisinin biosynthetic genes (Ma et al., 2009; Yu et al., 2012; Lu et al., 2013b; Zhang et al., 2015; Shen et al., 2016), and overexpression of the ADP-FPS fusion gene to stimulate substrate channeling (Han et al., 2016). However, in these reports either *A. annua* cultivars producing low levels of artemisinin (0.02%–0.4% of dry weight) were used as transgenic recipients, or the improvement of artemisinin content was not as efficient when high-artemisinin-producing cultivars (0.8%–1.0% of dry weight) were used as transgenic recipients. Moreover, most of previous studies focused only on modifying the upstream or downstream parts of the artemisinin biosynthetic pathway, and thus were unable to effectively boost the entire metabolic flux toward artemisinin biosynthesis. This may be one of the reasons for the failure to obtain transgenic *A. annua* lines with high artemisinin content. Recently, a synthetic biology strategy for artemisinin biosynthesis, in which the complete biosynthetic pathway of artemisinic acid, the precursor of artemisinin, is introduced into tobacco plants, has also been reported (Fuentes et al., 2016; Malhotra et al., 2016). However, whether this strategy allows efficient artemisinin production in other plant species needs to be further studied.

A. annua is a member of the Asteraceae, the largest family of flowering plants, which comprises more than 23 000 species, including many with considerable medicinal, ornamental, and economic importance (Vidic et al., 2016). A major impediment to the exploitation of these resources in basic and breeding sciences has been the absence of reference genome sequences; to date, only the sunflower genome has been released (Kane et al., 2011; Renaut et al., 2013; Badouin et al., 2017), although some transcriptome and metabolomics datasets are available (Graham et al., 2010; Soetaert et al., 2013; Ma et al., 2015). Here, we report a high-quality draft genome sequence of *A. annua* obtained by an integrative approach combining short-read sequencing (Illumina and Roche 454) and long-read sequencing (PacBio RSII), which is highly effective in assembling complex genomes such as that of *A. annua*. We also performed transcriptomic analyses, and identified some novel genes involved in the biosynthesis of artemisinin or other terpenoids based on the genomic and transcriptomic data. We demonstrate that the availability of a reference genome sequence coupled with RNA sequencing (RNA-seq) is helpful for metabolic engineering of important secondary metabolites, leading to improvement in artemisinin production.

RESULTS

Genome Sequencing, Assembly, and Annotation

We generated a high-quality draft genome sequence of a high-artemisinin-producing *A. annua* cultivar, Huhao 1, a highly heterozygous diploid (2*n* = 2*x* = 18 chromosomes) (Xie et al., 1995; Torrell and Vallès, 2001). According to standard 17-mer curves, the heterozygosity of the *A. annua* genome is 1.0%–1.5% (Supplemental Figure 2). The DNA sequencing reads were obtained using Illumina, Roche 454, and PacBio sequencing technologies. A total of 442 gigabases (Gb) of high-quality Illumina reads, 3.1 Gb of Roche 454 reads, and 22 Gb of PacBio long reads (Supplemental Tables 1–3) were generated, resulting in ∼260× coverage of the *A. annua* genome. The total genome assembly amounted to 1.74 Gb, consisting of 39 579 scaffolds, with a scaffold N50 of 104.86 kb and a contig N50 of 18.95 kb (Table 1). The completeness of the genome assembly was evaluated by using the Core Eukaryotic Genes Mapping

Chromosome number (2*n*)	18[a]
Estimate of genome size	1.74 Gb
Number of scaffolds	39 579
N50 of scaffolds	104 858 bp
Number of contigs	197 282
N50 of contigs	18 950 bp
GC content	31.5%
Number of protein-coding genes	63 230
Average gene density (per 100 kb)	4.78
Mean gene length	3803
Mean coding sequence length	994
Total size of coding regions	85.29 Mb
Fraction of coding regions	4.76%
Fraction of protein-coding genes	18.19%
Repetitive elements share in genome	61.57%

Table 1. *A. annua* Genome Assembly and Annotation Statistics.
[a]Xie et al., 1995; Torrell and Vallès, 2001.

Approach (CEGMA) (Parra et al., 2007) and Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simao et al., 2015). CEGMA results indicated that 98.0% of core eukaryotic genes were contained in our assembly (243 out of 248 core eukaryotic genes) (Supplemental Table 4). Based on BUSCO analysis 89.2% of plant sets were identified as complete (1284 out of 1440 BUSCOs) (Supplemental Figure 3). Completeness of the assembled genome sequence was also estimated by mapping a multiple-sourced eval dataset, including 166 full-length genes and 86 708 unigenes from the NCBI database, 19 168 transcripts from Roche 454 sequencing, and 79 190 Trinity-assembled transcripts, to the assembled genome. All of the 166 full-length gene sequences completely mapped to the assembly. Of 19 168 Roche 454 transcripts, 91.8% and 94.7% were successfully mapped to the genome with the ''hard'' and ''soft'' criteria, respectively. Of the 12 616 (66 574) Trinity-assembled transcripts, 85.1% (78.9%) were successfully mapped to the genome with >90% coverage and 89.2% (84.8%) mapped with >70% coverage, while 75.8% and 86.6% of the unigenes were mapped with the ''hard'' and ''soft'' criteria, respectively (Supplemental Table 5). We identified mitochondrial and chloroplast contigs from the sequenced genome data by performing a BLAST search against plant mitochondrial and chloroplast genome sequences from the NCBI Organelle Genome Resources using alignment length $\geq$10 kb as a threshold. We found nine and ten contigs putatively associated with the mitochondrial and chloroplast genomes, respectively (Supplemental Tables 6 and 7).

By combining homology-based and *de novo* approaches, we found that 61.57% of the assembled *A. annua* genome consists of repetitive elements; 60.10% of the genome was predicted to consist of interspersed repeats and transposable elements, while 1.47% was predicted to be tandem repeats (Supplemental Table 8). Of the repetitive elements, 32.66% could not be classified into any known gene family, and the most abundant characterized elements were LTRs (long terminal repeats), which accounted for 22.69% of the genome. Next, protein-coding gene models were constructed using a pipeline combining *de novo*, homology-based, and EST-aided prediction, resulting in 63 226 protein-coding gene models. The predicted gene structures were further refined using a transcriptome sequencing dataset. Prediction of alternatively spliced forms of these protein-coding genes identified 6758 splice variants for 3064 genes. Of the predicted gene models, 54 317 (85.9%) were supported by transcriptome data from at least one of nine organs/tissues (young leaf, old leaf, bud, flower, stem, seed, root, epidermis, and mixed trichome cells), indicating the high accuracy of the gene predictions (Supplemental Figure 4). Among 41 518 genes supported by transcriptome data with transcripts per kilobase per million (TPM) >1, more than 28 870 were expressed in buds, whereas only about 20 000 expressed genes were detected in trichomes (Supplemental Figure 5). A total of 41 884 genes (66.2%) were assigned to gene ontology (GO) categories, of which 26 752 fell within the ''metabolic processes'' category (Figure 1A).

The *A. annua* genome we report here is one of a few sequenced Asteraceae genomes. Orthologous clustering of the predicted *A. annua* proteins with those of 16 other representative green plants revealed that 7522 *A. annua* genes distributed among 2899 families are common to these species. A total of 36 435 *A. annua* genes clustered with those from at least one of the 16 genomes. Another 2496 genes, from 871 families, grouped only with genes from *Helianthus annuus*. These genes are referred to as Asteraceae-specific genes. Further comparison of *A. annua*, *H. annuus*, *A. thaliana*, and *S. lycopersicum* proteins revealed 8401 clusters of genes distributed among all four eudicot genomes and 3739 gene families that were unique to *A. annua* (Figure 1B). These *A. annua* specific gene families were enriched in genes involved in the GO-defined biological processes such as SCF-dependent proteasomal ubiquitin-dependent protein catabolism, shade avoidance, stomatal movement, and nitrate metabolism, as well as those associated with the transferase activity and kinase activity (Supplemental Table 9).

Evolution and Gene Family Expansion Analysis

A phylogenetic tree was constructed based on a concatenated sequence alignment of the 67 single-copy genes shared by the *Artemisia* genus and 16 other green plant species (Figure 1C). In this phylogenetic tree, *A. annua*, as expected, clustered with *H. annuus*, another Asteraceae species, and these two species were most closely related to the Solanaceae family. Although the 67 single-copy genes already provided phylogenetic signals that allowed for phylogeny construction of the 16 green plant species, accurate dating of the divergence times between Asteraceae and other families still requires a larger gene set. We identified the gene families that have expanded or contracted in the *Artemisia* lineage. In total, 7286 gene families were expanded in *A. annua*, whereas 3950 gene families were contracted (Figure 1C). Gene family expansion in *A. annua* has resulted in a notably large number of genes, making *A. annua* to be one of the genome-sequenced plant species with the largest number of genes compared with the other 50 plant species with sequenced genomes listed by Michael and Jackson (2013). Notably, 2717 TFs were identified in *A. annua* (Supplemental Table 10), which is substantially more than the number in most sequenced plant genomes, and three TF

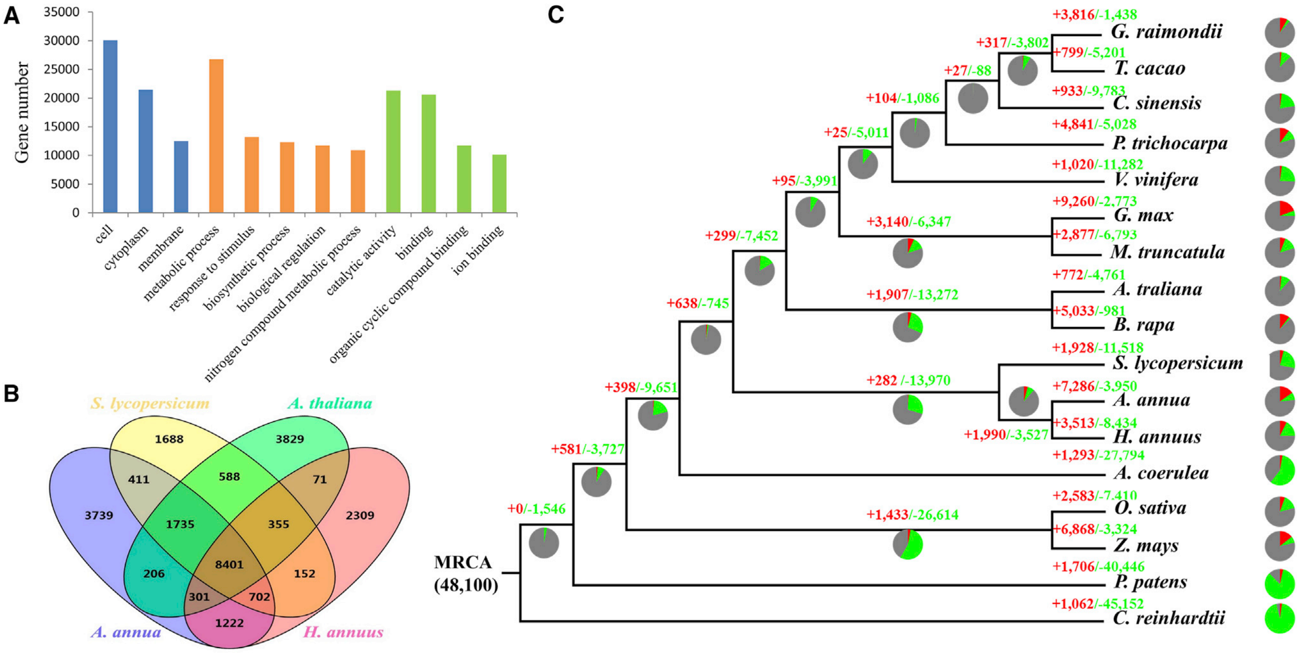

Figure 1. Characterization of *A. annua* Genome Evolution and Gene Family Expansion.

(A) Gene Ontology (GO) functional classification.

(B) Protein clusters shared between *A. annua* and other more phylogenetically distant plant species. *A. annua* is a representative of the *Artemisia* genus, *Helianthus annuus* is a representative of the *Helianthus* genus, *Solanum lycopersicum* is a representative of the *Solanum* genus, and *Arabidopsis thaliana* is a representative of the *Arabidopsis* genus.

(C) The phylogenetic tree was constructed from a concatenated alignment of 67 single-copy genes from 17 plant species. Gene family expansion is indicated in red, and gene family contraction is indicated in green; the corresponding proportions among total changes are shown using the same colors as in the pie charts. MRCA, most recent common ancestor.

families, C3H, FAR1 and Nin-like, were particularly larger than their counterparts in the other plant species.

The diverse array of secondary metabolites that are synthesized by *A. annua* provides protection against pests, pathogens, and herbivores across its natural range of habitats (Aftab et al., 2014). An important class of these defense compounds is terpenoids, including artemisinin, which are synthesized and stored in glandular trichomes. In this regard, a striking feature of the *A. annua* genome is the existence of unusually large gene families related to terpene biosynthesis and the expansion of several prenyltransferase families (Figure 2A). The terpene synthase (TPS) family genes are generally divided into seven clades, whereas in most plants the majority of TPS genes fall into one or two clades, indicating lineage-specific expansion. We found 122 putative TPS genes in the *A. annua* genome, most of which belonged to the TPS-a and TPS-b subfamilies. Phylogenetic analysis of putative full-length TPSs from 11 sequenced plant genomes revealed that *A. annua* and *H. annuus* TPSs were clustered together in each subclade. The presence of Asteraceae-specific members of the TPS-a and TPS-b subfamilies indicates lineage-specific expansion of the TPS family (Figure 2B). The positions of TPS genes from the model flowering plant *A. thaliana* on the branches of the TPS-a and TPS-b clades indicate that almost all of the Asteraceae TPS genes arose by gene duplications that occurred after the divergence of the Asteraceae lineage from the *Solanum lycopersicum* and *A. thaliana* lineages (Chen et al., 2011).

Genomic and Transcriptomic Analyses of the Artemisinin Biosynthetic Pathway and Its Regulation

Although the artemisinin biosynthetic pathway has been extensively studied and some key components and enzyme-catalytic steps for artemisinin biosynthesis have been well characterized, by combining genomic and transcriptomic analyses we obtained some new insights about this biosynthetic pathway and their transcriptional regulation in *A. annua*. Analysis of the *A. annua* genome revealed that the copy numbers of *DXR*, *FPS*, *ADS*, and *ALDH1* were in line with both previous transcriptomic sequencing reports (Figure 3A) (Graham et al., 2010; Ma et al., 2015). However, as for gene *HMGR* (3-hydroxy-3-methylglutaryl coenzyme A reductase), two genes (AA201470 and AA271980) in the *A. annua* genome were annotated as *HMGR* and showed 97% nucleotide sequence identity to the previously reported *A. annua HMGR1* (GenBank: AF142473.1) and *Chrysanthemum* × *morifolium HMGR2* (GenBank: KT809341.1) genes, respectively, whereas Graham et al. (2010) and Ma et al. (2015) reported four contigs and three contigs for *HMGR*, respectively. Moreover, through genomic sequence alignment we showed that these two *HMGR* genes shared less than 50% nucleotide sequence identity. Therefore, we consider them to be two different genes rather than two alleles. Similarly, two copies of *CYP71AV1* and *DBR2* were found in the *A. annua* genome (Figure 3A). The two *CYP71AV1* paralogs have 97% nucleotide sequence identity (Supplemental Figure 6); one paralog (AA566140) was located on scaffold QH_S14321 (37.6 kb)and based on RNA-seq analysis was highly expressed in trichomes, and in buds and young leaves

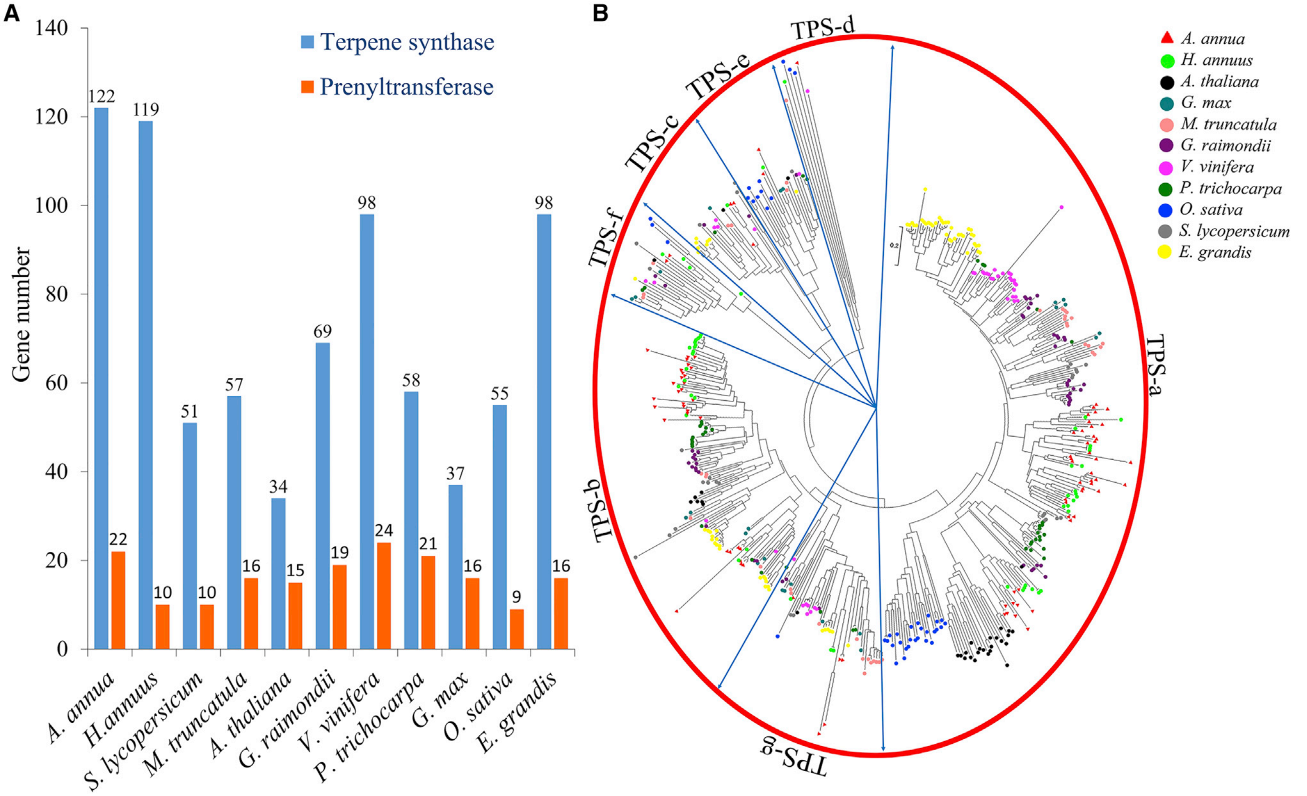

Figure 2. Interspecific Phylogenetic Analysis and Classification of Terpene Synthase Genes from *A. annua* and 10 Other Sequenced Plant Genomes.

(A) Comparison of terpene synthase and prenyltransferase genes in *A. annua* and 10 other sequenced plant genomes. BLAST analysis of the genomes was performed using the terpene synthase and prenyltransferase conserved domain by HMM.

(B) Phylogenetic tree of terpene synthases (TPS). Putative full-length TPS proteins (>400 amino acids in length) identified in 11 sequenced plant genomes, including 88 from *A. annua*, 51 from *H. annuus*, 30 from *S. lycopersicum*, 40 from *M. truncatula*, 33 from *A. thaliana*, 54 from *G. raimondii*, 43 from *V. vinifera*, 44 from *P. trichocarpa*, 24 from *G. max*, 41 from *O. sativa*, and 69 from *E. grandis*, were subjected to phylogenetic analysis. The scale bar (0.2) shows the number of amino acid substitutions per site.

that have a large number of trichomes, while the other paralog (AA502080) was located on scaffold QH_S10223 (58.7 kb) and showed basal expression in all the organs/tissues examined (Figure 3A). PCR analysis using genomic DNA further confirmed that there are two copies of *CYP71AV1* in *A. annua*. The two *DBR2* (AA049700 and AA049710) paralogs also have extremely high nucleotide sequence identity, but unlike *CYP71AV1*, the two *DBR2* paralogs were tandem repeated on scaffold QH_S00290 (316.65 kb) and showed a similar expression profile. The two *DBR2* paralogs were both highly expressed in trichomes, and in buds and young leaves (Figure 3A).

The biosynthesis of plant secondary metabolites is usually species, organ, or tissue specific (Goossens, 2014), as exemplified by artemisinin, which is only synthesized and stored in the glandular trichomes on plant surfaces. The expression of biosynthetic pathway genes is known to be regulated by a number of TFs (Vom Endt et al., 2002), and most of the known artemisinin metabolism-related genes were found to be expressed at higher levels in young leaves and buds, or in bud trichomes (Figure 3A). Several TFs in *A. annua* such as the basic helix-loop-helix family gene *AaMYC2* (Shen et al., 2016), the WRKY family gene *AaGSW1* (Chen et al., 2017), and the HD-ZIP family gene *AaHD1* (Yan et al., 2017) have already been demonstrated to regulate artemisinin biosynthesis. To investigate the transcriptional regulatory network underlying artemisinin biosynthesis, we performed hierarchical clustering (HCL) heatmap analysis coupled with Pearson correlation analysis using MultiExperiment Viewer (MeV4.9.0) software as previously described (Eisen et al., 1998). We identified several groups of TFs with expression patterns that were significantly correlated with those of *ADS*, *CYP71AV1*, *DBR2*, and *ALDH1*, and as such might be involved in regulating artemisinin biosynthesis (Figure 3B), including some TFs that have already been shown to be involved in artemisinin biosynthesis, such as AaORA, a member of the AP2/ERF family (Lu et al., 2013b) (Supplemental Figures 10–14). In addition, we identified an MYB family TF, AaMYB2 (AA147030), a member of the R2R3-MYB clade, which closely clustered with artemisinin biosynthesis-specific genes based on expression pattern, similar to AaORA (Figure 3C). MYB TFs participate in a range of plant developmental processes and the biosynthesis of secondary metabolites. However, little is known about whether this TF family is associated with artemisinin biosynthesis (Matías-Hernández et al., 2017).

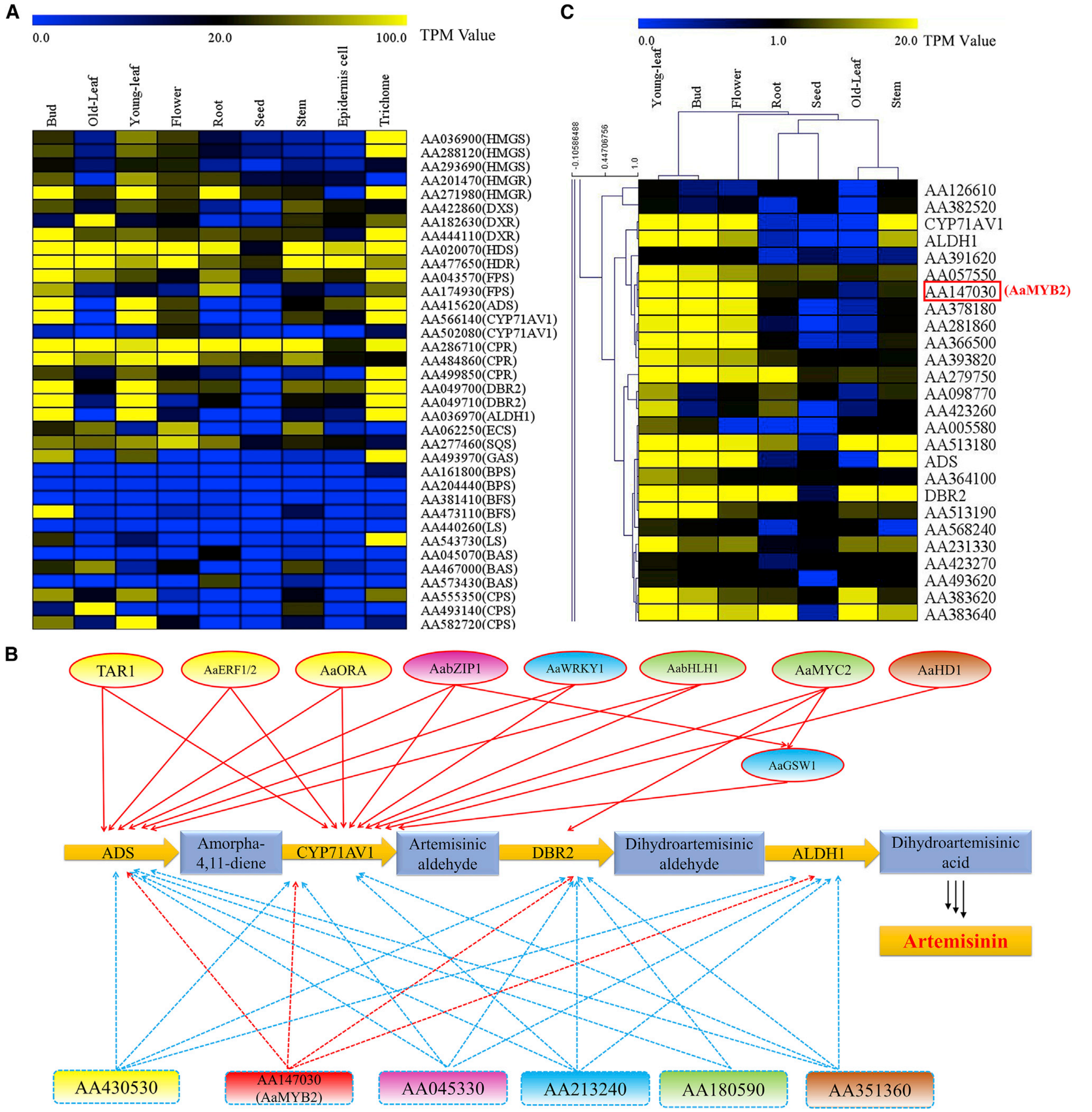

Figure 3. Artemisinin-related Gene Expression in Selected Tissues and the Artemisinin Biosynthesis Transcriptional Regulatory Network.

(A) Copy number and expression levels of genes functioning in the artemisinin biosynthetic pathway and its competitive pathway.

(B) Hierarchical clustering heatmap analysis coupled with Pearson correlation analysis of the MYB transcription factor family (part of the MYB family in *A. annua*) and artemisinin biosynthetic pathway genes (*ADS*, *CYP71AV1*, *DBR2*, and *ALDH1*) using MeV 4.9 software. AaMYB2 is marked by a red box.

(C) Transcription factors bind the promoters of artemisinin biosynthetic pathway genes such as *ADS*, *CYP71AV1*, *DBR2*, and *ALDH1*. The activation of gene expression results in an increase in artemisinin content in *A. annua*. ADS, amorpha-4,11-diene synthase; CYP71AV1, cytochrome P450 monooxygenase; CPR, cytochrome P450 reductase; DBR2, artemisinic aldehyde Ä11(13) reductase; ALDH1, aldehyde dehydrogenase 1. Red solid arrows represent transcription factors reported previously and demonstrated to bind to the promoters of pathway genes (TAR1, Tan et al., 2015; AaERF1/2, Yu et al., 2012; AaORA, Lu et al., 2013b; AabZIP1, Zhang et al., 2015; AaWRKY1, Ma et al., 2009; AabHLH1, Ji et al., 2014; AaMYC2, Shen et al., 2016; AaHD1, Yan et al., 2017; AaGSW1, Chen et al., 2017). Red dotted arrows represent the induction of pathway gene expression by the predicted AaMYB2 transcription factor. Blue dotted arrows represent putative transcription factors identified by hierarchical clustering heatmap analysis coupled with Pearson correlation analysis that may regulate artemisinin biosynthetic pathway genes. Different colors indicate different transcription factor families.

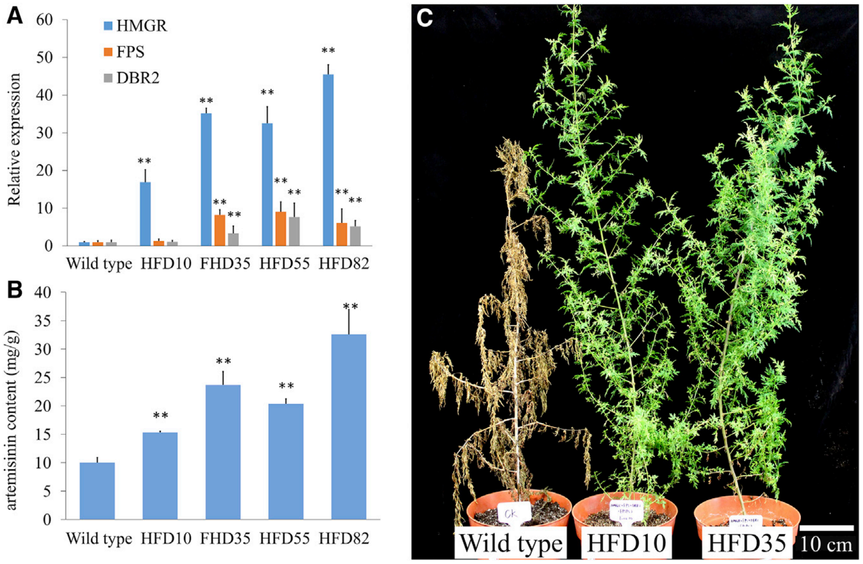

Figure 4. Genome and RNA-Seq Guided Plant Metabolic Engineering to Increase Artemisinin Content in *A. annua*.

(A) Relative expression levels of *HMGR*, *FPS*, and *DBR2* in wild-type and four independent transgenic *A. annua* overexpression lines determined by qPCR.

(B) HPLC analysis of artemisinin content in wild-type and four independent *HMGR*, *FPS*, and *DBR2*-overexpressing transgenic *A. annua* lines.

(C) Three-month-old transgenic *A. annua* lines overexpressing *HMGR*, *FPS*, and *DBR2* (middle and right) and wild-type (left) after 3 g/l glyphosate ammonium treatment. Data represent the means ± SE from three replicates. The *â-actin* gene (GenBank: EU531837.1) was used as the control for gene expression normalization.

Statistical significance was determined by two-independent-samples Æ test (**$P < 0.01$). Asterisks indicate the differences between wild-type and transgenic lines.

Metabolic Engineering of Artemisinin Biosynthesis

The information obtained by genome sequencing and RNA-seq has potential value in enhancing artemisinin content through metabolic engineering. Overexpressing a single artemisinin biosynthetic pathway gene (such as *FPS* and *HMGR*) has been partially successful, but has not resulted in substantial increases in artemisinin accumulation (Banyai et al., 2010; Nafis et al., 2011). HMGR is responsible for the conversion of HMG-coenzyme A into mevalonic acid, the precursor of isopentenyl diphosphate, and its expression level was low in trichomes when compared with that of *ADS*, *CYP71AV1*, and so forth. FPS, the enzyme responsible for generating farnesyl pyrophosphate, the immediate substrate for the biosynthesis of sesquiterpenes including artemisinin, is a key enzyme located at a branching point of artemisinin biosynthesis. DBR2, the enzyme responsible for converting artemisinic aldehyde to dihydroartemisinic aldehyde, is another crucial enzyme in the formation of artemisinin, sitting in the downstream position in artemisinin biosynthesis. To increase artemisinin content, we generated transgenic *A. annua* lines simultaneously overexpressing *HMGR* (AA201470), *FPS*, and *DBR2* (Figure 4A and Supplemental Figure 7). We found that many transgenic lines showed high artemisinin content in the leaves, among which line HFD82 had the highest artemisinin content (3.2%, dry weight) as determined by high-performance liquid chromatography (HPLC) analysis (Figure 4B and Supplemental Figure 8). The artemisinin contents of these lines were further confirmed by liquid chromatography–mass spectrometry analysis (Supplemental Figure 9). In addition to having high artemisinin content, because the *EPSPS* gene encoding 5-enolpyruvylshikimate-3-phosphate synthase that confers resistance to glyphosate was used as a selection marker during plant transformation, the transgenic lines overexpressing *HMGR*, *FPS*, and *DBR2* exhibited resistance to the herbicide glyphosate (Figure 4C and Supplemental Figure 7) and thus can potentially be used for large-scale cultivation.

We also produced transgenic *A. annua* lines with constitutive overexpression of *AaMYB2*, in which the transcript levels of *AaMYB2* showed a 1.2- to 7.4-fold increase compared with those in the wild-type plants. As expected, the transcript levels of genes in the artemisinin biosynthetic pathway, including *ADS*, *CYP71AV1*, *DBR2*, and *ALDH1*, showed substantial increases in *AaMYB2*-overexpressing lines (Figure 5A). Compared with the wild-type plants, the artemisinin and dihydroartemisinic acid contents were 51%–103% and 83%–144% higher, respectively, in the *AaMYB2*-overexpressing lines (Figure 5B), which showed no morphological or developmental abnormalities (Figure 5C). These results indicate that AaMYB2 is a positive regulator of artemisinin biosynthesis.

DISCUSSION

The *de novo* assembly of large genomes with a high degree of heterozygosity remains a particular challenge (Schatz et al., 2012; Kajitani et al., 2014). For example, self-incompatibility is the most common mode of outcrossing in the Asteraceae (Hiscock, 2000), which might result in high heterozygosity. It is both technically challenging and time-consuming to generate haploid, pure lines with low heterozygosity. The *A. annua* genome is highly heterozygous (heterozygosity is 1.0%–1.5%, shown in Supplemental Figure 2), with a high density of LTRs and low GC content. Such characteristics are barriers to the assembly of diploid genome sequences generated using only short-read sequencing (Roche 454 or Illumina), whereas the PacBio RSII sequencing platform generates long reads, thereby faciliting the sequence assembly and enhancing the assembly quality. The integrated PacBio strategy used here was highly effective in assembling the complex *A. annua* genome.

Analysis of the *A. annua* genome revealed that 871 gene families might be unique to the Asteraceae family. Due to its morphological complexity and the large number of species, systematic analysis of this family has been challenging for taxonomists (Judd et al., 2007). Our phylogenetic analysis, based on sequence alignment of 67 single-copy genes, supports the traditional phylogeny (Judd et al., 2007). For example, of the species tested, *A. annua* and *H. annuus* showed the closest evolutionary

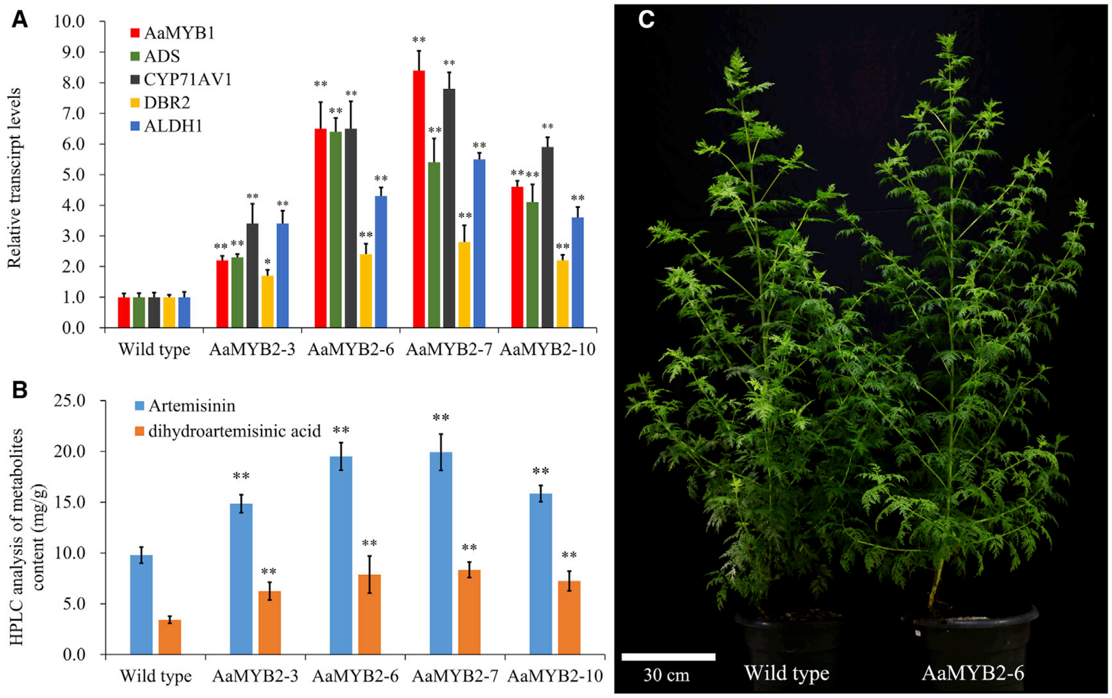

Figure 5. The AaMYB2 Transcription Factor Positively Regulates Artemisinin Biosynthesis in *A. annua*.

(A) Relative expression levels of *AaMYB2* and artemisinin biosynthetic pathway genes (*ADS*, *CYP71AV1*, *DBR2*, and *ALDH1*) in wild-type and four independent *AaMYB2*-overexpressing transgenic *A. annua* lines determined by qPCR.

(B) HPLC analysis of artemisinin and dihydroartemisinic acid contents in wild-type and four independent *AaMYB2*-overexpressing transgenic *A. annua* lines. Data represent the means ± SE from three replicates. The *â-actin* gene was used as the control for normalization.

(C) No morphological difference is observed between wild-type and an *AaMYB2*-overexpressing transgenic *A. annua* line.

Statistical significance was determined by two-independent-samples *Æ* test (**P* < 0.05; ***P* < 0.01). Asterisks indicate the differences between wild-type and *AaMYB2*-overexpressing lines.

relationship to *S. lycopersicum* (*Solanum* family), which is consistent with a previous tomato genome sequence report (Sato et al., 2012), indicating the usefulness of the *A. annua* genome in phylogenetic studies.

The *A. annua* genomic and associated transcriptomic datasets provide not only new insights into the artemisinin biosynthetic pathway and its regulation but also valuable tools for artemisinin metabolic engineering. Previous studies manipulating artemisinin biosynthesis in *A. annua* mainly focused on either upstream (Nafis et al., 2011) or downstream (Yuan et al., 2015) of the artemisinin biosynthetic pathway, resulting in increased but still limited artemisinin accumulation in transgenic *A. annua* lines. Multiple enzymatic steps are involved in artemisinin biosynthesis, implying that there may be more than one enzymatic step limiting the metabolic flux into artemisinin biosynthesis. In this study, by simultaneously overexpressing multiple genes functioning in the upstream (*HMGR*), midstream (*FPS*), and downstream (*DBR2*) of the artemisinin biosynthetic pathway, which may effectively boost the entire pathway, we obtained transgenic *A. annua* lines with significantly increased artemisinin content. Thus, an efficient strategy to increase the production of plant secondary metabolites might be to simultaneously enhance the expression of genes encoding enzymes functioning in different steps in a biosynthetic pathway or to overexpress the TFs such as AaMYB2 that regulate the expression of multiple genes in a biosynthetic pathway. *AaMYB2* closely clustered with artemisinin biosynthesis-specific genes based on expression pattern, and we found that overexpression of this gene could significantly enhance artemisinin and dihydroartemisinic acid content in transgenic *A. annua* lines. This also implies that the other TF genes that clustered with the artemisinin biosynthetic pathway genes may have similar functions, and as such represent candidate genes for metabolically engineered enhancement of artemisinin levels.

In short, our study adds abundant valuable information to the limited genomic resources of Asteraceae, one of the biggest plant families with diverse specialized metabolites. The *A. annua* genome and transcriptome data we provide here should be valuable for both fundamental biological research and applied breeding programs. The transgenic *A. annua* lines with high artemisinin content generated in this study should be a useful aid in enhancing the global supply of artemisinin from plant sources.

METHODS

Sample Preparation, DNA Extraction, and RNA Extraction

The *A. annua* cultivar used for sequencing, "Huhao 1," is a high artemisinin producer variety originating from Youyang, the major and traditional *A. annua* growing area, and was developed at Shanghai Jiao Tong University after several years of selection. *A. annua* plants were grown in the university greenhouse and fresh young leaves were collected from 4-month-old plants. External contaminants were removed by washing with Milli-Q water three times, then leaves were frozen in liquid nitrogen and stored at −80°C until DNA extraction. Genomic DNA (gDNA) was extracted using the CTAB method (Stewart and Via, 1993) with phenol–chloroform followed by RNase A and proteinase K treatments to remove RNA and protein contamination.

Total RNA was extracted from seven organs/tissues (young leaf, old leaf, bud, flower, stem, seed, and root) collected from five independent plants, using the Column Plant RNAout kit (TIANDZ, China). The trichome cells and epidermal cells were collected by laser capture microdissection (LCM) as described previously (Matas et al., 2011; Olofsson et al., 2012). The mRNA from seven tissues was enriched using an Illumina TruSeq RNA Library Prep Kit, which involves poly(A) pull-down using oligo-dT attached magnetic beads, following the manufacturer's standard protocol. Total RNA from LCM samples was extracted using the RNeasy Micro Kit (Qiagen, Germany), and mRNA was amplified in a two-round amplification using the TargetAmp kit (Epicenter Biotechnologies, USA) following the manufacturer's instructions. The DNA and RNA quantity and size distribution were verified using a Bioanalyzer (Agilent Technologies, USA).

Library Construction and DNA Sequencing

A shotgun library with an insert size of 300- to 800-bp was prepared from 1 ìg of gDNA using the Roche GS DNA Library Preparation Kit following the manufacturer's standard protocol (Roche, Switzerland), and 3-Gb sequence data were produced by a Roche 454 GS FLX. The paired-end (PE) libraries with different inserts (170, 300, and 800 bp) were prepared using a standard protocol (Illumina, USA). The three PE libraries were sequenced (2 × 150- bp reads) on a HiSeq 2500 (Illumina), generating 169 Gb of sequence data. The 500-bp library was sequenced (2 × 300- bp reads) on a MiSeq (Illumina) and 56 Gb of sequence data were generated. Mate-pair-end (MP) libraries with varied insert sizes (3, 5, 8, 10, and 20 kb) were constructed using the Cre-Loxp Inverse PCR Paired-End (CLIP-PE) method (Peng et al., 2012) and sequenced in the 2 × 150-bp format on a HiSeq 2500 (Illumina), generating 217 Gb of sequence data. Four flowcells were used on the HiSeq 2500 and another four on MiSeq. Samples were not multiplexed with those from other species.

For the PacBio sequencing library construction and sequencing, 5 ìg of gDNA, extracted from the fresh young leaves of five plants, was sheared using a Covaris g-TUBE (Covaris, USA) followed by purification via binding to pre-washed AMPure XP beads (Beckman Coulter, USA). Sheared gDNA was end-repaired using the PacBio DNA Template Prep Kit 2.0 (Pacific Biosciences, USA) and then ligated with blunt adapters, followed by exonuclease incubation to remove all unligated adapters and DNA. The final "SMRT bells" were annealed with primers and bound to the proprietary polymerase using the PacBio DNA/Polymerase Binding Kit P4 (Pacific Biosciences) to form the "Binding Complex". After dilution, the library was loaded onto the PacBio *RS*II sequencer with DNA Sequencing Kit 2.0 (Pacific Biosciences) and an SMRT Cell 8 Pac V3 for sequencing. A primary filtering analysis was performed on the sequencer, and the secondary analysis was performed using the SMRT analysis pipeline version 2.1.0 (Pacific Biosciences).

Genome Assembly and Evaluation

The 1.5% heterozygosity along with the high repetitive content (62%) of the *A. annua* genome poses a challenge for assembly with next-generation sequencing short reads. Extremely fragmental contigs with N50 lengths of 100–200 bp were obtained when we used the Illumina PE and MP sequencing method and assembly recipe. To alleviate the effects of the complex genome and short reads, we sequenced a 2 × 300 MiSeq pair-end library with an insert size of 500 bp and merged read pairs to obtain long reads using FLASH (Magoè and Salzberg, 2011). Although the volume of the Roche 454 data is small compared with Illumina data, the Roche 454 data were fed to Newbler (v2.9, Roche) at an early stage of the assembly process to construct the core contigs. *A. annua* Roche 454 data were solely sequenced with 12 flowcells, resulting in a total of 3.1 Gb with a mean read length of 585 bp. The samples were not multiplexed with those of other species. Newbler (v2.9, Roche) was run with default parameters except that "large Genome" and "heterozygote Mode" were set to true.

Considering the complexity of the *A. annua* genome, we sequenced five different mate-pair libraries with insert sizes of 3, 5, 8, 10, and 20 kb. All mate-pair reads were processed by Trimmomatic to remove the adapters and low-quality sequences before genome assembly (Bolger et al., 2014). After evaluation, cleaning, and some quality trimming, these mate-pair reads were mapped to core contigs to extract contig pairing information. SSPACE (v3.0) (Boetzer et al., 2011) was used to collect, calculate, and summarize the information and generate the genome scaffolds.

To obtain the final genome assembly with better assembly continuity, we used a 2-step gap-closing process to close sequence gaps within the scaffolds. First, GapCloser (Paulino et al., 2015) was used to close small-sized gaps with Illumina pair-end reads. Second, for the remaining larger-sized gaps, PacBio long reads, which have longer single sequence lengths, were used to cross and fill some of these gaps. PBJelly2 (English et al., 2012) was applied in this process to ensure the quality of the consensus sequence from local assembly around the gaps. The assembled genome sequences were searched again for vector and *Escherichia coli* genome sequences to remove contaminant sequences.

We mapped all contigs onto each other using Lastz and, when two contigs were more than 95% identical over more than 90% of the shorter contig, we discarded the shorter sequence to filter out homologous contigs. Based on this criterion, 173 highly homologous contigs were removed. Since the 1.5% heterozygosity of the *A. annua* genome was not high enough to require the assembly of two separate genome sets, we did not screen out secondary haploid scaffolds from the assembly and retained this information for subsequent analysis.

The completeness of the final genome assembly was evaluated by using CEGMA (Parra et al., 2007) and BUSCO (Simao et al., 2015). We performed a BUSCO testing run for sequenced plant genomes using version 3 run_BUSCO.py software with the embryophyta protein set (run_BUSCO.py -i plant_species.fa -o plant_species -l embryophyta_odb9/-m proteins). We also collected a multiple-sourced eval dataset and mapped these sequences to the assembled genome to evaluate the completeness.

Estimation of the Genome Size

The genome size (G) of *A. annua* was estimated based on the occurrence distribution from 17-mer resampling of sequencing reads, calculated with the equation $G = N_{17mer}/D_{17mer}$, where N_{17mer} is the number of total 17-mers and D_{17mer} is average sequence depth. For N_{17mer}, low-frequency 17-mers, which possibly resulted from sequencing errors (depth <3), were excluded from the calculation to alleviate the deviation from sequencing error (Marçais and Kingsford, 2011). The total genome assembly was 1.74 Gb. The total K-mer (K = 18) number was 154 863 960 403, and the volume peak was 88. The genome size can be estimated as (total K-mer number)/(volume peak), which is 1 759 817 732 bp.

RNA-Seq

Two micrograms of *A. annua* mRNA were enriched using the Illumina TruSeq RNA Library Prep Kit (Illumina), which uses oligo(dT)-attached magnetic beads to pull down poly(A) RNA. The *A. annua* mRNA was then used to synthesize cDNA using a non-strand-specific SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, USA), and the library was constructed using the Roche shotgun-library-preparation method. In total, 625 Mb of transcriptome sequence was generated using a Roche 454 GS FLX instrument. The mRNA from nine organs/tissues (young leaf, old leaf, bud, flower, stem, seed, root, epidermal cells, and trichome cells) of *A. annua* were converted into RNA-seq libraries and sequenced using an Illumina HiSeq2500, and 237 577 070 pairs of 100-bp reads were generated.

Low-quality raw read sequences were removed using Trimmomatic (version 0.30) (Bolger et al., 2014). Cleaned reads were then mapped to the genome of *A. annua* using TopHat2 software (Kim et al., 2013). Reads counts were calculated with HTseq (Anders et al., 2014) using BAM results from TopHat2, and TPM values were then calculated for every gene in the tissues analyzed.

Gene Prediction

The repeat-masked *A. annua* genome sequence was used for gene predictions. Three predictors, Fgenesh (Solovyev et al., 2006), GeneMark (Lomsadze et al., 2014), and Augustus (Keller et al., 2011), were used for *ab initio* gene prediction with default parameters, and for Fgenesh and Augustus, *A. thaliana* was selected as the species template. For each locus, predicted genes from Fgenesh, GeneMark, and Augustus were annotated based on BLAST searches against the NCBI nr, UniProt, and KEGG protein databases (Breuza et al., 2016; Kanehisa et al., 2016; Rigden et al., 2016). The predicted gene with the highest score from BLAST was considered the best prediction and selected as the gene model of the locus. All steps were performed using in-house perl scripts. The selected gene predictions were improved using PASA with two cycles of annotation comparison (Haas et al., 2008). Sequence improvement included adding untranslated regions, exon boundary adjustments, merging genes, and adding alternative transcripts.

In total 205 982 transcripts were constructed from about ∼237 million pairs of PE Illumina (Hiseq2500) RNA-seq reads from nine organ/tissue libraries. In total 19 168 transcript assemblies were constructed from ∼0.8 million Roche 454 RNA-seq sequences using Trinity with default parameters: Trinity release 2014-07-17, Trinity.pl –seqType fq –JM 500G –full_cleanup –output Artemisia_output –left All_paired_1 fq –right All_paired_2 fq –single All_unpaired.fq –CPU 24 (Grabherr et al., 2011). An additional 86 708 *A. annua* unigenes were downloaded from the NCBI public database.

We evaluated the possibility that the high number of gene models in our assembly were false positives by using *A. annua* predicted proteins as queries in searches against the NCBI nr protein databases. Among the 63 226 predicted *A. annua* proteins, 59 190 proteins were similar to the nr proteins (E-value <1e−5), and 54 586 proteins were similar to the nr proteins (E-value <1e−10). Furthermore, the predicted *A. annua* genes were validated by BLASTN (2.2.26 version) searches against the sunflower genes (https://sunflowergenome.org/), the *A. annua* transcripts assembled from RNA-seq sequences, and the downloaded *A. annua* unigenes. Of the 63 226 predicted genes, 55 198 genes (87.3%) were similar to sunflower genes (E-value <1e−10), and 55 091 genes (87.1%) were supported by the unigenes or transcripts (E-value <1e−50, overall length coverage >50% of the predicted gene).

The 63 226 predicted genes encode 66 920 transcripts. The lengths of the transcripts (mRNA) range from 153 to 16 465 bp, and the average length is 1324 bp. For the predicted genes, the average number of exons per gene is 5.2, the average exon length is 219 bp, and the average intron length is 655 bp.

Gene Family Analysis

OrthoMCL (Li et al., 2003) was used to define gene families across 17 plant genomes. Besides the predicted proteins from *A. annua*, we used the following sources to obtain proteins from 16 other species: *Aquilegia coerulea*, *Brassica rapa*, *Chlamydomonas reinhardtii*, *Citrus sinensis*, *Glycine max*, *Gossypium raimondii*, *Helianthus annuus*, *Physcomitrella patens*, *Populus trichocarpa*, *Theobroma cacao*, *Vitis vinifera*, and *Zea mays* (version 10.1, http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV10.1); *Arabidopsis thaliana* (version 10, https://www.arabidopsis.org/); *Medicago truncatula* (version 4.0, http://medicago.jcvi.org/medicago/display.php?pageName=General&section=Download); *Solanum lycopersicum* (version 2.4, http://solgenomics.net/organism/solanum_lycopersicum/genome); and *Oryza sativa* (version 7.0, http://rice.plantbiology.msu.edu/). All protein sets were filtered for a minimum protein length of 50 amino acid residues. An all-against-all comparison was performed using BLASTP with the threshold E-value <1e−5. For defining gene families, the Markov cluster algorithm was used to cluster the BLASTP results into groups of homologous proteins at an inflation factor of 1.5.

The output of OrthoMCL was parsed to identify gene families. A concatenated alignment of single-copy genes was performed using ClustalW (Larkin et al., 2007) with default settings. The phylogenetic tree was constructed using maximum parsimony with the MEGA program (version 6.0) with 1000 bootstrap replicates (Tamura et al., 2013). Computational analysis of changes in gene family size in 17 plant genomes was done using CAFE (De Bie et al., 2006).

Identification of Transcription Factors

We refined the TF prediction pipeline by updating the hidden Markov model (HMM) profiles used to identify TFs. The HMM profiles were downloaded from Pfam (version 27.0) for most signature domains. The predicted *A. annua* proteome and eight other sequenced plant proteomes were scanned using HMMER suite (http://hmmer.janelia.org/) based on pfam profiles to identify TFs, after which TFs were assigned to different families according to assignment rules (Jin et al., 2014).

Identification of Terpene Synthase Genes

The protein sequences were scanned using pfamscan based on the HMMER suite. The TPSs were identified by screening with the HMM profiles of the PFAM motifs PF01397 (N-terminal TPS domain) and PF03936 (TPS, metal binding domain). The requirement for the presence of both domains was strict.

Phylogeny Reconstruction of TPS Proteins

Putative full-length TPSs (>400 amino acids in length) identified in 11 sequenced plant genomes, including 88 from *A. annua*, 51 from *H. annuus*, 30 from *S. lycopersicum*, 40 from *M. truncatula*, 33 from *A. thaliana*, 54 from *G. raimondii*, 43 from *V. vinifera*, 44 from *P. trichocarpa*, 24 from *G. max*, 41 from *O. sativa*, and 69 from *E. grandis*, were subjected to phylogenetic analysis. To perform phylogeny reconstruction, we performed multiple protein sequence alignments of the TPS homologs using ClustalW (Larkin et al., 2007) with the default settings. The neighbor-joining trees were constructed using the MEGA program (version 6.0) with 1000 bootstrap replicates (Tamura et al., 2013).

Hierarchical Clustering Analysis Using RNA-Seq Data

Hierarchical clustering analysis based on RNA-seq data was performed using MultiExperiment Viewer (MeV4.9.0) software to predict potential TFs involved in the biosynthesis of artemisinin (Saeed et al., 2003). Sample clustering was carried out using the HCL method, and the evolutionary distances were computed with Poisson correction and average linkage clustering (Eisen et al., 1998).

Construction of Transforming Vectors and Transformation of *A. annua*

The full-length open reading frames (ORFs) of *HMGR* (AA201470), *FPS*, and *DBR2* from *A. annua* were cloned from young leaf cDNA using gene-specific primers and then ligated into the modified pCAMBIA1305.1 vector with a glyphosate selection marker gene, and with expression driven by the CaMV35S promoter (Supplemental Figure 7, Supplemental Table 11). The full-length ORF of *AaMYB2* was cloned from *A. annua* young leaf cDNA using AaMYB2-F and AaMYB2-R primers (Supplemental Table 11) and then ligated into the pHB$^+$ vector under the control of the CaMV35S promoter.

The resulting constructs were introduced into *Agrobacterium tumefaciens* strain EHA105, and *A. annua* transgenic plants were generated as described previously (Zhang et al., 2009).

Gene Expression and HPLC Analyses

The relative transcript levels of all *A. annua* genes tested were measured by qPCR. The relative transcript levels were normalized to the transcript abundance of *A. annua â-actin* (GenBank: EU531837.1). For qPCR analysis, mRNA was extracted using the RNAprep pure Plant Kit (Tiangen, China) and reverse transcribed into cDNA using the PrimeScript RT Master Mix (TaKaRa, China). PCR amplification was performed in a Roche LightCycler 96 (Roche) using the SYBR Green qPCR Master Mix (Tiangen) according to the manufacturer's instructions. The thermal profile for SYBR Green qPCR was 95°C for 2 min, followed by 40 cycles of 95°C for 20 s, 54°C for 20 s, and 72°C for 20 s. Samples for HPLC analysis were prepared as described previously (Lu et al., 2013b). Artemisinin standard was purchased from Sigma and dihydroartemisinic acid standard was obtained from Guangzhou Honsea Sunshine Bio Science and Technology (Sunshine Bio, China).

Data and Seeds Access

For genome-sequencing data, the NCBI accession number is PKPP00000000, and for RNA-seq raw data, the SRA accession number is SRP129502.

The seeds of the "Huhao 1" cultivar and the transgenic overexpression lines were deposited in our university seed bank, and they are freely accessible for research only. The seed deposit information is as follows: ID Huhao 1-01, type *A. annua* Huhao 1 cultivar; ID Huhao 1-02, type HFD transgenic lines; ID Huhao 1-03, type AaMYB2 transgenic lines. Contact person: Dr. Guirong Wang, School of Agriculture and Biology, Shanghai Jiao Tong University, Room 1-413, Building of Agriculture & Biology, 800 Dongchuan Road, Shanghai 200240, China, Tel: +86-21-34206144, Email: grwang@sjtu.edu.cn.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the China National High Technology Research and Development Program (2011AA100605), China National Transgenic Plant Research and Commercialization Project (2016ZX08002-001), and Shanghai Jiao Tong University Agri-Engineering Program (AF1500028).

AUTHOR CONTRIBUTIONS

Q.S., L.Z., Z.L., and K.T. designed experiments; Q.S., L.Z., S.W., G.L., and Y.Z. prepared libraries and generated sequence data; T.Y., P.S., M.L., X.F., Z.L., X.L., F.Z., and W.J. performed gene cloning, plasmid construction, and plant transformation; Q.P., Y.W., Y.M., M.C., X.H., L.L., Y.T., and X.S. contributed molecular biology analysis and metabolic analysis; L.Z., S.W., G.L., Y.Z., and P.E.B. performed bioinformatics analyses; Q.S., L.Z., Z.L., J.K.C.R., and K.T. wrote the manuscript.

ACKNOWLEDGMENTS

We are very grateful to Dr. Liangsheng Zhang (Fujian Agriculture and Forestry University) for providing help and kind suggestions. We express appreciation to the Instrumental Analysis Center of Shanghai Jiao Tong University for the kind help from Dr. Lei Feng in liquid chromatography–mass spectrometry analysis. J.K.C.R. is supported by an affiliation with the Atkinson Center for a Sustainable Future. No conflict of interest declared.

Received: July 27, 2017
Revised: March 12, 2018
Accepted: March 25, 2018
Published: April 24, 2018

REFERENCES

Aftab, T., Ferreira, J.F., Khan, M.M.A., and Naeem, M. (2014). *Artemisia annua*—Pharmacology and Biotechnology (Berlin Heidelberg: Springer).

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics **31**:166–169.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Brière, C., Owens, G.L., Carrère, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature **546**:148–152.

Banyai, W., Kirdmanee, C., Mii, M., and Supaibulwatana, K. (2010). Overexpression of farnesyl pyrophosphate synthase (*FPS*) gene affected artemisinin content and growth of *Artemisia annua* L. Plant Cell Tissue Organ Cult. **103**:255–265.

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

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**:2114–2120.

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M.L., Magrane, M., Tognolli, M., Bridge, A., Baratin, D., and Redaschi, N.; UniProt Consortium (2016). The UniProtKB guide to the human proteome. Database (Oxford) **2016**. https://doi.org/10.1093/database/bav120.

Chang, Y.J., Song, S.H., Park, S.H., and Kim, S.U. (2000). Amorpha-4, 11-diene synthase of *Artemisia annua*: cDNA isolation and bacterial expression of a terpene synthase involved in artemisinin biosynthesis. Arch. Biochem. Biophys. **383**:178–184.

Chen, F., Tholl, D., Bohlmann, J., and Pichersky, E. (2011). The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. **66**:212–229.

Chen, M., Yan, T., Shen, Q., Lu, X., Pan, Q., Huang, Y., Tang, Y., Fu, X., Liu, M., Jiang, W., et al. (2017). GLANDULAR TRICHOME-SPECIFIC WRKY 1 promotes artemisinin biosynthesis in *Artemisia annua*. New Phytol. **214**:304–316.

Czechowski, T., Larson, T.R., Catania, T.M., Harvey, D., Brown, G.D., and Graham, I.A. (2016). *Artemisia annua* mutant impaired in artemisinin synthesis demonstrates importance of nonenzymatic conversion in terpenoid metabolism. Proc. Natl. Acad. Sci. USA **113**:15150–15155.

Daddy, N.B., Kalisya, L.M., Bagire, P.G., Watt, R.L., Towler, M.J., and Weathers, P.J. (2017). *Artemisia annua* dried leaf tablets treated malaria resistant to ACT and i.v. artesunate: case reports. Phytomedicine **32**:37–40.

De Bie, T., Cristianini, N., Demuth, J.P., and Hahn, M.W. (2006). CAFE: a computational tool for the study of gene family evolution. Bioinformatics **22**:1269–1271.

Duffy, P.E., and Mutabingwa, T.K. (2006). Artemisinin combination therapies. Lancet **367**:2037–2039.

Efferth, T. (2006). Molecular pharmacology and pharmacogenomics of artemisinin and its derivatives in cancer cells. Curr. Drug Targets **7**:407–421.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**:14863–14868.

Elfawal, M.A., Towler, M.J., Reich, N.G., Weathers, P.J., and Rich, S.M. (2015). Dried whole-plant *Artemisia annua* slows evolution of malaria drug resistance and overcomes resistance to artemisinin. Proc. Natl. Acad. Sci. USA **112**:821–826.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., and Worley, K.C. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One **7**:e47768.

Fuentes, P., Zhou, F., Erban, A., Karcher, D., Kopka, J., and Bock, R. (2016). A new synthetic biology approach allows transfer of an entire metabolic pathway from a medicinal plant to a biomass crop. Elife **5**. https://doi.org/10.7554/eLife.13664.

Goossens, A. (2014). It is easy to get huge candidate gene lists for plant metabolism now, but how to get beyond? Mol. Plant **8**:2–5.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**:644–652.

Graham, I.A., Besser, K., Blumer, S., Branigan, C.A., Czechowski, T., Elias, L., Guterman, I., Harvey, D., Isaac, P.G., Khan, A.M., et al. (2010). The genetic map of *Artemisia annua* L. identifies loci affecting yield of the antimalarial drug artemisinin. Science **327**:328–331.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. **9**:R7.

Han, J., Wang, H., Kanagarajan, S., Hao, M., Lundgren, A., and Brodelius, P.E. (2016). Promoting artemisinin biosynthesis in *Artemisia annua* plants by substrate channeling. Mol. Plant **9**:946–948.

Hiscock, S.J. (2000). Self-incompatibility in *Senecio squalidus* L. (Asteraceae). Ann. Bot. **85**:181–190.

Ji, Y., Xiao, J., Shen, Y., Ma, D., Li, Z., Pu, G., Li, X., Huang, L., Liu, B., and Ye, H. (2014). Cloning and characterization of AabHLH1, a bHLH transcription factor that positively regulates artemisinin biosynthesis in *Artemisia annua*. Plant Cell Physiol. **55**:1592–1604.

Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res. **42**:D1182–D1187.

Judd, W., Campbell, C., Kellogg, E., and Stevens, P. (2007). Plant Systematics: A Phylogenetic Approach (Sunderland, MA: Sinauer Associates Press).

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. **24**:1384–1395.

Kane, N., Gill, N., King, M., Bowers, J., Berges, H., Gouzy, J., Bachlava, E., Langlade, N., Lai, Z., and Stewart, M. (2011). Progress towards a reference genome for sunflower. Botany **89**:429–437.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. **44**:D457–D462.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics **27**:757–763.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. **14**:R36.

Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., and Lopez, R. (2007). Clustal W and clustal X version 2.0. Bioinformatics **23**:2947–2948.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

Li, J., Casteels, T., Frogne, T., Ingvorsen, C., Honore, C., Courtney, M., Huber, K.V., Schmitner, N., Kimmel, R.A., and Romanov, R.A. (2017). Artemisinins Target GABAA receptor signaling and impair a cell identity. Cell **168**:86–100.e15.

Lomsadze, A., Burns, P.D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. **42**:e119.

Lu, X., Shen, Q., Zhang, L., Zhang, F., Jiang, W., Lv, Z., Yan, T., Fu, X., Wang, G., and Tang, K. (2013a). Promotion of artemisinin biosynthesis in transgenic *Artemisia annua* by overexpressing *ADS*, *CYP71AV1* and *CPR* genes. Ind. Crop. Prod. **49**:380–385.

Lu, X., Zhang, L., Zhang, F., Jiang, W., Shen, Q., Zhang, L., Lv, Z., Wang, G., and Tang, K. (2013b). AaORA, a trichome-specific AP2/ERF transcription factor of *Artemisia annua*, is a positive regulator in the artemisinin biosynthetic pathway and in disease resistance to *Botrytis cinerea*. New Phytol. **198**:1191–1202.

Ma, D., Pu, G., Lei, C., Ma, L., Wang, H., Guo, Y., Chen, J., Du, Z., Wang, H., Li, G., et al. (2009). Isolation and characterization of AaWRKY1, an *Artemisia annua* transcription factor that regulates the amorpha-4,11-diene synthase gene, a key gene of artemisinin biosynthesis. Plant Cell Physiol. **50**:2146–2161.

Ma, D., Wang, Z., Wang, L., Alejos-Gonzales, F., Sun, M., and Xie, D. (2015). A genome-wide scenario of terpene pathways in self-pollinated *Artemisia annua*. Mol. Plant **8**:1580–1598.

Magoè, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27**:2957–2963.

Malhotra, K., Subramaniyan, M., Rawat, K., Kalamuddin, M., Qureshi, M.I., Malhotra, P., Mohmmed, A., Cornish, K., Daniell, H., and Kumar, S. (2016). Compartmentalized metabolic engineering for artemisinin biosynthesis and effective malaria treatment by oral delivery of plant cells. Mol. Plant **9**:1464–1477.

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764–770.

Matas, A.J., Yeats, T.H., Buda, G.J., Zheng, Y., Chatterjee, S., Tohge, T., Ponnala, L., Adato, A., Aharoni, A., Stark, R., et al. (2011). Tissue- and cell-type specific transcriptome profiling of expanding tomato fruit provides insights into metabolic and regulatory specialization and cuticle formation. Plant Cell **23**:3893–3910.

Matías-Hernández, L., Jiang, W., Yang, K., Tang, K., Brodelius, P.E., and Pelaz, S. (2017). AaMYB1, and its orthologue AtMYB61, affect terpene metabolism and trichome development in *Artemisia annua* and *Arabidopsis thaliana*. Plant J. **90**:520–534.

Michael, T.P., and Jackson, S. (2013). The first 50 plant genomes. Plant Genome **6**. https://doi.org/10.3835/plantgenome2013.03.0001in.

Nafis, T., Akmal, M., Ram, M., Alam, P., Ahlawat, S., Mohd, A., and Abdin, M.Z. (2011). Enhancement of artemisinin content by constitutive expression of the HMG-CoA reductase gene in high-yielding strain of *Artemisia annua* L. Plant Biotechnol. Rep. **5**:53–60.

Olofsson, L., Lundgren, A., and Brodelius, P.E. (2012). Trichome isolation with and without fixation using laser microdissection and pressure catapulting followed by RNA amplification: expression of genes of terpene metabolism in apical and sub-apical trichome cells of *Artemisia annua* L. Plant Sci. **183**:9–13.

Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D., et al. (2013). High-level semi-synthetic production of the potent antimalarial artemisinin. Nature **496**:528–532.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061–1067.

Paulino, D., Warren, R.L., Vandervalk, B.P., Raymond, A., Jackman, S.D., and Birol, I. (2015). Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics **16**:1–12.

Peng, Z., Zhao, Z., Nath, N., Froula, J.L., Clum, A., Zhang, T., Cheng, J.-f., Copeland, A.C., Pennacchio, L.A., and Chen, F. (2012). Generation of long insert pairs using a Cre-LoxP Inverse PCR approach. PLoS One **7**:e29437.

Peplow, M. (2016). Synthetic biology's first malaria drug meets market resistance. Nature **530**:389–390.

Renaut, S., Grassa, C., Yeaman, S., Moyers, B., Lai, Z., Kane, N., Bowers, J., Burke, J., and Rieseberg, L. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat. Commun. **4**:1827.

Rigden, D.J., Fernández-Suárez, X.M., and Galperin, M.Y. (2016). The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. Nucleic Acids Res. **44**:D1–D6.

Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho, K.A., Eachus, R.A., Ham, T.S., Kirby, J., et al. (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature **440**:940–943.

Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., and Thiagarajan, M. (2003). TM4: a free, open-source system for microarray data management and analysis. Biotechniques **34**:374–378.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485**:635–641.

Schatz, M.C., Witkowski, J., and McCombie, W.R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biol. **13**:243.

Shen, Q., Lu, X., Yan, T., Fu, X., Lv, Z., Zhang, F., Pan, Q., Wang, G., Sun, X., and Tang, K. (2016). The jasmonate-responsive AaMYC2 transcription factor positively regulates artemisinin biosynthesis in *Artemisia annua*. New Phytol. **210**:1269–1281.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics **31**:3210–3212.

Soetaert, S.S., Van Neste, C.M., Vandewoestyne, M.L., Head, S.R., Goossens, A., Van Nieuwerburgh, F.C., and Deforce, D.L. (2013). Differential transcriptome analysis of glandular and filamentous trichomes in *Artemisia annua*. BMC Plant Biol. **13**:220.

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. **7**:S10.

Stewart, C.N., Jr., and Via, L.E. (1993). A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques **14**:748–749.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. **30**:2725–2729.

Tan, H., Xiao, L., Gao, S., Li, Q., Chen, J., Xiao, Y., Ji, Q., Chen, R., Chen, W., and Zhang, L. (2015). TRICHOME AND ARTEMISININ REGULATOR 1 is required for trichome development and artemisinin biosynthesis in *Artemisia annua* L. Mol. Plant **8**:1396–1411.

Teoh, K.H., Polichuk, D.R., Reed, D.W., Nowak, G., and Covello, P.S. (2006). *Artemisia annua* L. (Asteraceae) trichome-specific cDNAs reveal CYP71AV1, a cytochrome P450 with a key role in the biosynthesis of the antimalarial sesquiterpene lactone artemisinin. FEBS Lett. **580**:1411–1416.

Teoh, K.H., Polichuk, D.R., Reed, D.W., and Covello, P.S. (2009). Molecular cloning of an aldehyde dehydrogenase implicated in artemisinin biosynthesis in *Artemisia annua*. Botany **87**:635–642.

Tin, A.S., Sundar, S.N., Tran, K.Q., Park, A.H., Poindexter, K.M., and Firestone, G.L. (2012). Antiproliferative effects of artemisinin on human breast cancer cells requires the downregulated expression of the E2F1 transcription factor and loss of E2F1-target cell cycle genes. Anticancer Drugs **23**:370–379.

Torrell, M., and Vallès, J. (2001). Genome size in 21 *Artemisia* L. species (Asteraceae, Anthemideae): systematic, evolutionary, and ecological implications. Genome **44**:231–238.

Vidic, D., Æavar Zeljkoviæ, S., Dizdar, M., and Maksimoviæ, M. (2016). Essential oil composition and antioxidant activity of four Asteraceae species from Bosnia. J. Essent. Oil Res. **28**:445–457.

Vom Endt, D., Kijne, J.W., and Memelink, J. (2002). Transcription factors controlling plant secondary metabolism: what regulates the regulators? Phytochemistry **61**:107–114.

World Health Organization. (2017). World Malaria Report 2017. http://www.who.int/malaria/publications/world-malaria-report-2017/report/en/.

Xie, D., Kang, N., and Li, G. (1995). Studies on the karyotype of *Artemisia annua*. Chin. Bull. Bot. **12S**:71–72.

Yan, T., Chen, M., Shen, Q., Li, L., Fu, X., Pan, Q., Tang, Y., Shi, P., Lv, Z., Jiang, W., et al. (2017). HOMEODOMAIN PROTEIN 1 is required for jasmonate-mediated glandular trichome initiation in *Artemisia annua*. New Phytol. **213**:1145–1155.

Yu, Z.X., Li, J.X., Yang, C.Q., Hu, W.L., Wang, L.J., and Chen, X.Y. (2012). The jasmonate-responsive AP2/ERF transcription factors AaERF1 and AaERF2 positively regulate artemisinin biosynthesis in *Artemisia annua* L. Mol. Plant **5**:353–365.

Yuan, Y., Liu, W., Zhang, Q., Xiang, L., Liu, X., Chen, M., Lin, Z., Wang, Q., and Liao, Z. (2015). Overexpression of artemisinic aldehyde Ä11(13) reductase gene-enhanced artemisinin and its relative metabolite biosynthesis in transgenic *Artemisia annua* L. Biotechnol. Appl. Biochem. **62**:17–23.

Zhang, Y., Teoh, K.H., Reed, D.W., Maes, L., Goossens, A., Olson, D.J.H., Ross, A.R.S., and Covello, P.S. (2008). The molecular cloning of artemisinic aldehyde Ä11(13) reductase and its role in glandular trichome-dependent biosynthesis of artemisinin in *Artemisia annua*. J. Biol. Chem. **283**:21501–21508.

Zhang, L., Jing, F., Li, F., Li, M., Wang, Y., Wang, G., Sun, X., and Tang, K. (2009). Development of transgenic *Artemisia annua* (Chinese wormwood) plants with an enhanced content of artemisinin, an effective anti-malarial drug, by hairpin-RNA-mediated gene silencing. Biotechnol. Appl. Biochem. **52**:199–207.

Zhang, F., Fu, X., Lv, Z., Lu, X., Shen, Q., Zhang, L., Zhu, M., Wang, G., Sun, X., Liao, Z., et al. (2015). A basic leucine zipper transcription factor, AabZIP1, connects abscisic acid signaling with artemisinin biosynthesis in *Artemisia annua*. Mol. Plant **8**:163–175.

Zheng, H., Colvin, C.J., Johnson, B.K., Kirchhoff, P.D., Wilson, M., Jorgensen-Muga, K., Larsen, S.D., and Abramovitch, R.B. (2017). Inhibitors of *Mycobacterium tuberculosis* DosRST signaling and persistence. Nat. Chem. Biol. **13**:218–225.