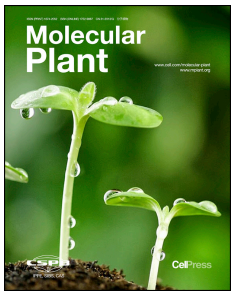
Accepted Manuscript



A Genome-wide Scenario of Terpene Pathways in Self-pollinated *Artemisia annua* L

Dong-Ming Ma, Zhilong Wang, Liangjiang Wang, Fatima Alejos-Gonzales, Ming-An Sun, De-Yu Xie

PII: S1674-2052(15)00298-1
DOI: 10.1016/j.molp.2015.07.004
Reference: MOLP 156

To appear in: *MOLECULAR PLANT*
Accepted Date: 8 July 2015

Please cite this article as: **Ma D.-M., Wang Z., Wang L., Alejos-Gonzales F., Sun M.-A., and Xie D.-Y.** (2015). A Genome-wide Scenario of Terpene Pathways in Self-pollinated *Artemisia annua* L. Mol. Plant. doi: 10.1016/j.molp.2015.07.004.

Running title: Terpene Biosynthesis in Self-pollinated *Artemisia annua*

Dong-Ming Ma[1], Zhilong Wang[1], Liangjiang Wang[2], Fatima Alejos-Gonzales[1], Ming-An Sun[3], and De-Yu Xie[1]

1, Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC 27695, USA

2, Department of Genetics and Biochemistry, Clemson University, Clemson, SC 29634, USA

3, Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Corresponding author:

De-Yu Xie
Department of Plant and Microbial Biology
North Carolina State University
Raleigh, NC 27695, USA
Email: dxie@ncsu.edu
Tel: 919-515-2129
Fax: 919-515-3436

Current address for Dr. Zhilong Wang:

Dr. Zhilong Wang
Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization
Hunan Agricultural University
Changsha Hunan 410128, China
Phone: +86 0731-84638423
Fax: +86 0731-84638423
E-mail: zhilongwang@126.com

Manuscript Title: A Genome-wide Scenario of Terpene Pathways in Self-pollinated *Artemisia annua* L

Summary:

A genome-wide gene-to-terpene landscape is characterized to show the network of terpene biosynthetic pathways differentially expressed in tissues of self-pollinated *Artemisia annua* plants.

Abstract

Scenarios of genes to metabolites in *Artemisia annua* remain uninvestigated. Here, we report the use of an integrated approach combining metabolomics, transcriptomics and gene function analyses to characterize gene-to-terpene and terpene pathway scenarios in a self-pollinating variety of this species. Eighty-eight metabolites including twenty-two sesquiterpenes (e.g., artemisinin), twenty-six monoterpenes, three triterpenes and thirty-eight other non-polar metabolites were identified from fourteen tissues. These metabolites were differentially produced by leaves and flowers at lower to higher positions. Sequences from cDNA libraries of six tissues were assembled into 18,871 contigs, and genome-wide gene expression profiles in tissues were strongly associated with developmental stages and spatial specificities. Sequence mining identified forty-seven genes that mapped to the artemisinin, non-amorphadiene sesquiterpene, monoterpene, triterpene, 2-C-methyl-D-erythritol 4-phosphate (MEP) and mevalonate pathways. Pearson correlation analysis resulted in network integration that characterized significant correlations of gene-to-gene expression patterns and gene expression-to-metabolite levels in six tissues simultaneously. More importantly, manipulations of amorpha-4,11-diene synthase (ADS) gene expression not only controlled the activity of this pathway toward artemisinin, artemisinic acid and arteannuin b but also altered non-amorphadiene sesquiterpene and genome-wide volatile profiles. Such gene-to-terpene landscapes associated with different tissues are fundamental to the metabolic engineering of artemisinin.

Keywords: *Artemisia annua*, artemisinin, terpene, metabolic profiling, pyrosequencing

Introduction

Artemisinin is an unusual endoperoxide sesquiterpene lactone (Liu et al., 1979) with a unique oxygen bridge structure (Fig. 1) and is the active antimalarial metabolite in *Artemisia annua* (Klayman, 1985; Liu et al., 1979; Milhous and Weina, 2010; Tu et al., 1982). Artemisinin-based combination therapy (ACT) is currently the first-line treatment of malarial diseases (Banek et al., 2014; Gogtay et al., 2013; Jelinek, 2013; Shunmay et al., 2008; Taylor, 2013), and numerous efforts have been made over many years to understand artemisinin biosynthesis to increase artemisinin production for cost-effective ACT. The biochemical steps from farnesyl diphosphate (FDS or farnesyl pyrophosphate, FPP) to artemisinic acid and dihydroartemisinic acid (Fig. 1) have been characterized by gene cloning and characterization (Brown, 2010). Additionally, genes encoding amorpha-4,11-diene synthase (ADS), a cytochrome P450 monooxygenase (CYP71AV1), a cytochrome P450 reductase (CPR), a Δ-11(13)-double bond reductase (DBR2), an aldehyde dehydrogenase 1 (ALDH1), a cytochrome b5 monooxygenase (CYB5) and an alcohol dehydrogenase (ADH1) have been biochemically and transgenically characterized as being involved in the biosynthesis of artemisinic acid and dihydroartemisinic acid (Fig. 1) (Bouwmeester et al., 1999; Chang et al., 2000; Paddon et al., 2013; Ro et al., 2006; Teoh et al., 2009; Teoh et al., 2006; Zhang et al., 2008). Although the steps from dihydroartemisinic acid to artemisinin (Fig. 1) remain unknown, the currently favored hypothesis involves non-enzymatic spontaneous autoxidation reactions in the presence of high levels of singlet oxygen (Brown, 2010; Sy and Brown, 2002).

Increasing artemisinin production is one of the primary research goals in combating malaria. Unfortunately, increasing artemisinin yields in plants has been challenging, given that artemisinin production is controlled by different facts. Indeed, germplasm investigations have demonstrated that the artemisinin content is low, and dramatically variable from 0.003 to 0.2% (g/g, dry weight) in different varieties and from 0 to 0.39% (DW) in individual plants of the same variety (Charles et al., 1990; Paul et al., 2014). Considering that *A. annua* is a cross-pollinating species, the most difficult breeding challenge is an inability of progeny to produce artemisinin due to segregation (Alejos-Gonzalez et al., 2011; Delabays et al., 2001; Graham et al., 2010; Simonnet et al., 2008). Thus, although a few varieties have been bred to produce high levels of artemisinin, e.g., from 1 to 2.4% (DW) (Brisibe et al., 2012; Cockram et al., 2012; Delabays et al., 2001; Graham et al., 2010; Larson et al., 2013; Simonnet et al., 2008), efforts are still necessary to overcome yield instability. In fact, germplasm selection has become particularly important for developing new ACT treatments. In the 1970s, crude extracts of artemisinin from *A. annua* were used to treat malaria in China (Zhong-Yi-Yan-Jiu-Yue-Zhong-Yao-Yan-Jiu-Suo, 1978), and recent experiments have further demonstrated that crude extracts of dry plants could be used as a new pACT product (*A. annua* ACT) (Weathers et al., 2014a; Weathers et al., 2014b) to overcome the resistance of the parasite to artemisinin monotherapy (Elfawal et al., 2015). In addition, numerous studies have shown that artemisinin production is controlled by multiple factors such as glandular trichomes (Duke and Paul, 1993; Lommen et al., 2005; Tellez et al., 1999), NaCl, jasmonic acid and salicylic acid (Kjaer et al., 2012). Hence, optimization of these factors can improve artemisinin production.

Much time and effort have been invested in metabolic engineering and synthetic biology to produce artemisinin or its precursors. *ADS*, *CYP71AV1* and *CPR* (Fig. 1) were

115 overexpressed in *A. annua* and tobacco plants, and promising levels of artemisinic acids,
116 artemisinin or new compound were produced in transgenic plants (Alam and Abdin,
117 2011; Chen et al., 2013; Farhi et al., 2011; Jiang et al., 2010; Shen et al., 2012; Tang et
118 al., 2014; van Herpen et al., 2010; Xiang et al., 2012; Zhang et al., 2010). Such efforts
119 have demonstrated that metabolic engineering is a promising approach to increasing the
120 production of artemisinin and its precursors in plants (Farhi et al., 2013; Liu et al., 2011).
121 In comparison, artemisinin synthetic biology has made exciting progress. Engineered
122 yeast strains have been developed to produce artemisinic acid by fermentation (Paddon et
123 al., 2013; Ro et al., 2006; Teoh et al., 2006), and the combination of synthetic biology
124 and semi-synthesis recently reached an industrial level to complement the demand of
125 plant-based artemisinin (Paddon et al., 2013; Turconi et al., 2014). In particular, the
126 improvement of semi-synthesis using artemisinic acid as a substrate resulted in an
127 increase in the yield of artemisinin to approximately 60 tons (Turconi et al., 2014). This
128 success will help ameliorate the limitation of the insufficient artemisinin supply from
129 plants.

130

131 Regardless, relatively fewer investigations have been performed to understand the effects
132 of the *in planta* activities of the terpene biosynthetic network on artemisinin biosynthesis.
133 Numerous monoterpenes and sesquiterpenes have been isolated from heterozygous *A.*
134 *annua* plants grown in different environments (Bhakuni et al., 2001; Goel et al., 2007;
135 Holm et al., 1997; Pu et al., 2013; Tellez et al., 1999; Wei et al., 1992). All data thus far
136 reveal that the *A. annua* terpene network is complicated, though it remains unclear
137 whether the terpenes isolated are produced by all tissues or only specific tissues. We
138 hypothesize that the complexities of the genome-wide terpene pathways toward different
139 terpene molecules are spatially and developmentally dependent and that the activities of
140 those pathways can essentially affect artemisinin production in both specific tissues and

the entire plant. To overcome progeny segregation, promote the metabolic engineering of artemisinin and understand the metabolic networks of terpenes, we recently developed a self-pollinating population of *A. annua* (Alejos-Gonzalez et al., 2011). When the plants were grown in a growth chamber, the artemisinin contents in the leaves and inflorescences of F2 plants at lower to higher positions ranged from 0.01%-0.1% (g/g, DW), and these levels were reproducible among progeny and individual plants. Additionally, these self-pollinating plants displayed an approximately 100% regeneration capacity through tissue culture (Alejos-Gonzalez et al., 2013). In the present study, we used an integrated approach of pyrosequencing and metabolomics to characterize genome-wide terpene biosynthesis in leaves at nine positions and in flowers at five stages. Terpene and gene expression profiles were integrated to characterize the scenarios of genome-wide gene-to-terpene pathways. *ADS* was selected to assess the effects of its down-regulation and overexpression on profiles of artemisinin, other amorphadiene and non-amorphadiene sesquiterpenes and volatiles. The resulting data provide new fundamental gene-to-terpene landscapes for enhancing the metabolic engineering of artemisinin.

Results

GC-MS-based Non-polar Metabolite Profiles

Eighty-eight metabolites (Tables 1 and 2, S-Table 3) were identified from 14 tissues using GC-MS analysis via two temperature programs (TP-1 and TP-2) of the column oven. These metabolites included 26 monoterpenes (monoterpenoids) (Table 1 and

S-Fig. 2), 21 sesquiterpenes (sesquiterpenoids) (Table 2 and S-Fig. 3), two triterpenes, one diterpene and another 38 metabolites (including acids, an alcohol, an aldehyde and alkenes) (S-Table 3). Most of metabolites could be detected by GC-MS using both TP1 and TP2. In comparison, TP-1 provided a better profile for sesquiterpenes than TP-2. TP-1 (but not TP-2) could detect artemisinic acid and β-pinene, whereas TP-2 (but not TP-1) could detect α-pinene (Table 1 and S-Fig. 2). Therefore, the metabolites separated by TP-1 were primarily used for principle component analysis (PCA), clustering analysis, gene-to-metabolite mapping and integration.

The profiles of metabolites were different among tissues (S-Fig. 1 and S-Table 4). The resulting 3-dimensional PCA plot, which was created using 87 metabolites (separated using TP-1) from 14 tissues, indicated that the patterns of non-polar metabolite profiles are affected by the tissue and position in the plant. The accumulation patterns of metabolites from nine groups of leaf tissues were clearly grouped together but separate from those patterns of metabolites from flower tissues (Fig. 2). By comparing leaves from the 7^{th} (L7) through 17^{th} (L17) nodes, the metabolite profiles in neighboring leaves, e.g., L11 and L12, were found to be ordinately close together. In addition, the accumulation patterns of the 87 metabolites in head inflorescences, including head inflorescence bud (HB) tissue and fully open head (FOH) inflorescences, were ordinately distant from the patterns of the nine groups of leaves. The accumulation patterns of these metabolites in 50% opening head (50%OH) inflorescences and inflorescence branch 15 and 16 (IB15 and IB16) tissues were ordinately grouped together and localized between the leaf and FOH and HB samples.

A heatmap coupled with clustering analysis was generated to further understand metabolic differentiation in the 14 tissues (Fig. 3). The nine groups of leaves appeared in one clade, which was separated from the FOH, 50%OH, IB15 and IB16 samples. The L7 and L10, L11 and L12, and L16 and L17 samples were clustered together. The heatmap and clustering data supported the PCA results.

Monoterpene Profiles

Twenty-six monoterpenes (Table 1; S-Fig. 2) were identified from 14 tissues (separated by both TP-1 and TP-2). Based on the recent classification of 134 *A. annua* monoterpenes (Brown, 2010), these 26 molecules comprise three acyclic, ten monocyclic, twelve bicyclic and one tricyclic monoterpenes (Table 1). The detection of these metabolites was also associated with various tissues (Table 1; S-Table 3), and only 1,8-cineole was detected in all 14 tissues. The levels of 1,8-cineole exhibited an increasing trend from L7 through L17 (Fig. 4 A) and were higher in the five flower tissues than in the nine leaf tissues, with a maximum value in FOH tissue. The 25 other metabolites were detected only in certain tissues at specific positions (Table 1). For example, β-pinene was only detected in L13 through L17 and the five flower tissues but not in L7-L12. The levels of β-pinene were higher in the five flower tissues than in L14-L17 leaf tissues, with a maximum value in IB15 tissue. β-Terpinene was only detected in IB15 and IB16, terpinen-4-ol only in HB and FOH tissues, and 1, 2-dimethyl-3-(1-methylethenyl)-cyclopentane only in L10 and L11.

Sesquiterpene Profiles

GC-MS analysis identified twenty-one sesquiterpenes (Table 2; S-Fig. 3) from 14 tissues, comprising one acyclic, six monocyclic, eleven bicyclic and four tricyclic structures. These metabolites could be categorized into two amorphadiene-derived (artemisinic acid and arteannuin b) and 19 non-amorphadiene-derived sesquiterpenes (Table 2). Based on their presence or absence and abundance, the accumulation patterns of the 21 metabolites were characterized by tissue specificity (Table 2; S-Table 4). Arteannuin b, caryophyllene, α-copaene and germacrene D were detected in all 14 tissues, with variable accumulation patterns (Fig. 4B and C, S-Fig. 4 A, Table 2), and β-farnesene was detected in 13 tissues. By contrast, the 16 other sesquiterpenes were only detected in certain tissues, e.g., only in certain positional leaves, at certain flower stages, or at certain leaf and flower positions (Table 2).

The peak area values of the 21 sesquiterpenes (S-Table 3) were recorded to reflect their relative abundance in tissues and to characterize their tissue specificity. Five sesquiterpenes were selected for a comparison of levels in 14 tissues. The content of artemisinic acid was significantly higher in the five flower tissues than in the nine leaf tissues. Of the former, artemisinic acid levels were higher in HB, FOH and 50%OH than in IB15 and IB16 (Fig. 4 B and S-Table 3). Artemisinic acid was detected in L11-L17 but not in L7 and L10, and the content increased from L11 to L17. The levels of arteannuin b in IB15 and HB were higher than the levels in the other 12 tissues (Fig. 4 B and Table 2). In the nine different leaf tissues, the arteannuin b level gradually increased from L7 to L13, was similar in L13 L14 and L15, and then increased in L16 and L17. The levels of caryophyllene were higher in the five flower tissues than in the nine leaf tissues, with the lowest value in L7 (Fig. 4 C); of the former, its content was

238 higher in FOH and 50%OH than in IB15 and IB16 (Fig. 4 C). The contents of
239 α-copaene gradually increased from L7 to L17, were similar among L17, IB15, IB16 and
240 HB and then increased in FOH and 50%OH (Fig. 4 C). β-Farnesene was detected in 13
241 tissues but not in L7, and the content was significantly higher in the five flower tissues
242 than in the nine leaf samples; for the latter, its content was the highest in L10, followed
243 by a slight decreasing trend from L12 through L17. In addition, the levels of
244 germacrene D were higher in HB, FOH and 50%OH than in the other tissues (S-Fig. 4).
245 Of the nine leaf tissues, the germacrene D level gradually increased from L7 to L17.

246

247 **Triterpenes**

248

249 β-Amyrin and squalene were detected from 12 tissues (Fig. 4D). The levels of squalene
250 were notably higher in L7, L10 and L11 than in the other six leaf tissues. In the five
251 flower tissues, the squalene levels were higher in HB than in IB15 and IB16, and the
252 levels of β-amyrin were higher in L15 and L17 than in the other tissues.

253

254 **Artemisinin Accumulation Profile**

255

256 The artemisinin content in 13 tissues was estimated using LC-ESI-MS analysis. The
257 highest artemisinin level was observed in HB (Fig. 4 E). As reported previously in F2
258 progeny (Alejos-Gonzalez et al., 2011), the artemisinin content increased from L10 to
259 L13 and was similar in L13, L14 and L15. In addition, the artemisinin contents in IB16
260 and FOH were close to those in L14 and L15. The highest artemisinin level was

261 approximately 0.03% (g/g, fresh weight, which is equivalent to 0.3% in dry weight) in
262 HB of F3 plants.

263

264 **Transcriptome Sequencing**

265

266 Pyrosequencing of six different tissues generated 1,133,865 high-quality reads, including
267 277,706 from L7, 230,260 from L15, 270,435 from IB15, 132,229 from HB, 98,985 from
268 FOH and 123,519 from 50%OH. All sequences were assembled to obtain 18,871
269 contigs. Approximately 95% of the contigs could be functionally annotated using
270 sequence similarity analysis (E-value $\leq 1\times10^{-5}$) with the UniProt database (Magrane and
271 UniProt-Consortium, 2011). For each contig, the reads per kilobase per million
272 (RPKM) (Magrane and UniProt-Consortium, 2011) were calculated to indicate the levels
273 of gene expression in six tissues. In addition, all contigs from the six tissues were used
274 to develop an HCL Euclidean heatmap. The resulting heatmap (Fig. 5) showed that L7
275 and L15 were clustered together but differentially separated from FOH, HB and 50%OH.
276 In addition, the 50%OH transcriptome was between those of HB and FOH, indicating a
277 global transcriptomic shifting during reproductive tissue development.

278

279 **Genes Involved in Artemisinin Biosynthesis**

280

281 Five known pathway genes, namely, *ADS*, *AV1*, *CPR*, *DBR2* and *ALDH1* (Fig. 1), which
282 were cloned from trichomes and other tissues (Bouwmeester et al., 1999; Chang et al.,
283 2000; Chen et al., 2013; Covello et al., 2007; Mercke et al., 2000; Paddon et al., 2013; Ro
284 et al., 2006; Teoh et al., 2009; Teoh et al., 2006; Xiang et al., 2012; Zhang et al., 2010),

285 were identified in the 18,871 contigs. In addition, two new cDNAs reported to encode
286 an alcohol dehydrogenase (ADH1) and a cytochrome b5 (CYB5) were recently cloned
287 from *A. annua* (Paddon et al., 2013). Twenty-nine *ADH1* and 15 *CYB5* homolog
288 contigs were identified from the sequences from the six tissues.

289

290 One contig was obtained for each of the *ADS*, *ALDH1* and *DBR2* sequences. The *ADS*
291 cDNA sequence was obtained from four tissues, and its RPKM value was the highest in
292 HB, followed by FOH, 50%OH and then IB15 (Fig. 6 A). Although the *ADS* cDNA
293 could not be sequenced using the L7 and L15 samples, the RT-PCR results showed that it
294 was expressed in these two and the seven other leaf tissues (L10-L14 and L16-17) (S-Fig.
295 5 and S-Fig. 6). RT-PCR also demonstrated that the expression levels of *ADS* were
296 higher in the five flower tissues than in the nine leaf tissues (S-Fig. 5 and S-Fig. 6). The
297 *DBR2* cDNA sequence was obtained from the six tissues, and its RPKM value was the
298 highest in L7, followed by L15 and HB, IB15, 50%OH and then FOH (Fig. 6 A). In
299 contrast with the order of the RPKM values, RT-PCR analysis showed that the expression
300 levels of *DBR2* were higher in the five flower tissues than in the nine leaf tissues (S-Fig.
301 5 and S-Fig. 6). The *ALDH1* cDNA sequence was obtained from four tissues; its
302 RPKM value was the highest in HB, followed by 50%OH, FOH and then IB15 (Fig. 6
303 A). The RT-PCR results showed that *ALDH1* was expressed in all 14 tissues (S-Fig. 5
304 and S-Fig. 6); its expression levels were relatively similar in the five flower tissues, L14,
305 L15, L16 and L17 but higher in these tissues than in L13, L12 L11, L10 and L7.

306

307 Two contigs each were obtained for *CYP71AV1* and *CPR*. The *CYP71AV1* cDNA
308 sequence was obtained from five tissues but not from L7. Sequence alignment showed
309 that the two *CYP71AV1* contigs are identical to the reported sequence of *CYP71AV1* (Ro

et al., 2006; Teoh et al., 2006*)*. Its RPKM value was the highest in 50%OH, followed by HB, IB15, FOH and then L15. The RT-PCR results showed that the expression levels of *CYP71AV1* in the five flower tissues were much higher than in the leaf tissues (S-Fig. 5 and S-Fig. 6), which supported the order of RPKM values. Two *CPR* contigs, namely, *CPR1* and CPR2, were identified as two homologs by sequence alignment (Supplementary Results). The *CPR1* sequence was obtained from the six tissues, and its RPKM value was the highest in L7, followed by L15, 50%OH, FOH, IB15 and HB (Fig 6 A). RT-PCR analysis showed that all 14 tissues expressed *CPR1* at a relatively similar level (S-Fig. 5 and S-Fig. 6). The *CPR2* sequence was obtained from four flower tissues but not L7 and L15 tissues, with the RPKM value of *CPR2* the highest in 50%OH, followed by FOH, HB and then IB15 (Fig. 6 A). RT-PCR analysis showed that *CPR2* was highly expressed in 50%OH, FOH, IB15 and HB but was barely detectable in leaf tissues (S-Fig. 7), supporting the order of its RPKM values.

In addition, two *FDS* contigs, *AaFDS1* and *AaFDS2*, were assembled from the sequences of the six tissues, and sequence alignment showed that these two contigs are two homologs (Supplementary Results). The RPKM value of *AaFDS1* was the highest in 50%OH, followed by HB, FOH, IB15, L15 and then L7, whereas the RPKM value of *AaFDS2* was the highest in 50%OH, followed by FOH, HB, L15, IB15 and then L7. RT-PCR analysis showed that *AaFDS2* was highly expressed in all 14 tissues (S-Fig. 5 and S-Fig. 6).

Genes Involved in Non-amorphadiene Sesquiterpene and Triterpene Biosynthesis

Gene annotation revealed five homologs (Fig. 6 B) associated with non-amorphadiene sesquiterpene biosynthesis. Three contigs, *AaGAS*, *AaCPS* and *AaECS*, are homologs of reported germacrene A synthase, β-caryophyllene synthase and 8-epi-cedrol synthase genes, respectively. *AaGAS* sequences were obtained from 5 tissues; the highest RPKM values were in HB, followed by 50%OH, IB15, L7 and L15 (Fig. 6 B). RT-PCR showed that its expression was higher in the five flower tissues than in the nine leaf tissues (S-Fig. 5 and S-Fig. 6), with its expression being variable in the latter. *AaCPS* sequences were obtained from five tissues but not from L15, and were assembled into two contigs. Sequence alignment identified the contigs as two fragments of a reported *AaCPS* (gb|AF472361.1|). The RPKM value of *AaCPS* was the highest in HB, followed by 50%OH, FOH, B15 and L7; RT-PCR analysis showed that the expression levels of *AaCPS* were higher in the five flower tissues than in the nine leaf tissues (S-Fig. 5 and S-Fig. 6), supporting the order of the RPKM values. An *AaECS* sequence was obtained from the six tissues, with its RPKM value being highest in L15, followed by 50%OH, IB15, FOH, L7 and HB. RT-PCR analysis showed that the expression levels of *AaECS* were higher in the nine leaf tissues than in the five flower tissues (S-Fig. 5 and S-Fig. 6). Two additional contigs were annotated as unknown sesquiterpene synthase (SQTS) genes: unknown *SQTS1* and *2*, the sequences of which were obtained from the six tissues. Sequence alignment identified them as homologous to a reported terpene synthase cDNA (gb|JQ837261.1|.) Therefore, they belong to one gene, namely, unknown *SQTS*. Further sequence alignments showed that this unknown *SQTS* shares 72% and 74% identity to *AaECS* and *AaGAS* (Supplementary results), respectively. Its RPKM value was the highest in 50%OH, followed by HB, FOH, IB15, L15 and L7 (Fig. 6 B).

One contig for a squalene synthase (AaSQS) gene was assembled from the sequences of the six tissues, and its RPKM value was the highest in L7, followed by L15, 50%OH, HB, IB15 and FOH (Fig. 6 B). RT-PCR analysis showed its expression in all 14 tissues (S-Fig. 5 and S-Fig. 6).

In addition, although the beta-farnesene synthase (BFS) sequence was not obtained from the six tissues, the RT-PCR results indicated that this gene was indeed expressed in all 14 tissues (S-Fig. 5 and S-Fig. 6). β-Amyrin synthase (AaBAS) is the first cyclase leading to triterpenes; although the *AaBAS* cDNA sequence could not be obtained from the sequences of the six tissues, RT-PCR analysis demonstrated its expression in all 14 tissues (S-Fig. 5 and S-Fig. 6).

Genes Involved in Monoterpene Biosynthesis

One contig each for *A. annua* limonene synthase (*AaLS*), 1,8-cineole synthase (*AaCIN*) and β-pinene synthase (*AaBPS*) genes was identified among 18,871 contigs. The sequence of the *AaLS* cDNA was obtained from two tissues, and its RPKM value was higher in L7 than in L15 (Fig. 6 C). RT-PCR analysis showed that *AaLS* was expressed not only in L7 and L15 but also in the other 12 tissues (S-Fig. 5, S-Fig. 6). *AaBPS* was sequenced from three tissues, and its RPKM value was higher in IB15 and 50%OH than in IB15 (Fig. 6 C). RT-PCR analysis showed that *AaBPS* was expressed in all 14 tissues, with higher levels in the five flower tissues than in the nine leaf tissues (S-Fig. 5 and S-Fig. 6). The sequence of *AaCIN* was obtained from five tissues, and its RPKM value was the highest in 50%OH, followed by HB, IB15, L7 and FOH (Fig. 6 C).

Five contigs for linalool synthase (*AaLAS*), namely, *AaLAS1*, *AaLAS2*, *AaLAS3*, *AaLAS4* and *AaLAS5*, were assembled from the sequences of the six tissues (Fig. 6 C). Sequence alignment showed that these five contigs constitute five homologs (Supplementary Results). The sequences of *AaLAS1* and *AaLAS2* were obtained from four tissues. The RPKM value of *AaLAS1* was the highest in 50%OH, followed by FOH, IB15 and HB, and that of *AaLAS2* was the highest in HB, followed by 50%OH, IB15 and then FOH. The sequence of *AaLAS3* was obtained from five tissues, with the highest RPKM value in L7, followed by L15, IB15, FOH and 50%OH. *AaLAS4* and *AaLAS5* sequences were obtained from the six tissues. The RPKM value of *AaLAS4* was the highest in L7, followed by L15, IB15, 50%OH, FOH and HB, and that of *AaLAS5* was the highest in 50%OH, followed by FOH, HB, IB15, L15 and L7.

Genes Involved in the MEP Pathway

Six *DXS* contigs, *AaDXS1*, *AaDXS2*, *AaDXS3*, *AaDXS4*, *AaDXS5* and *AaDXS6*, were assembled from the sequences of the six tissues. Sequence alignment showed that the six contigs represent four homologs (Supplementary Results). The *AaDXS1*, *-2* and *-3* contigs are three fragments of the same homolog; thus, *AaDXS1* was used to represent this homolog. The RPKM value of *AaDXS1* was the highest in 50%OH, followed by FOH, IB15, HB, L7 and L15 (S-Fig. 8 A). RT-PCR analysis showed that *AaDXS1* was expressed in all 14 tissues (S-Fig. 5 and S-Fig. 6). The RPKM value of *AaDXS4* was the highest in 50%OH, followed by IB15, L15 and HB, FOH and L7, and that of *AaDXS5* was the highest in HB, followed by FOH, L15, 50%OH, IB15 and L7. The cDNA

407 sequences of *AaDXS6* were obtained from three tissues, with a higher RPKM value in L7
408 than in IB15 and 50%OH.

409

410 Two contigs each for *DXR* (*AaDXR1* and *AaDXR2*) and *HDR* (*AaHDR1* and *AaHDR2*)
411 were assembled based on the sequences from the six tissues. Sequence alignment
412 showed that the two *DXR* contigs are homologs and that the two *HDR* contigs are also
413 homologs (Supplementary Results). The RPKM value of *AaDXR1* was the highest in
414 L7, followed by HB, IB15, L15, 50%OH and FOH (S-Fig. 8 A and B). RT-PCR
415 analysis showed that *AaDXR1* was expressed at a similar level in all 14 tissues (S-Fig. 5
416 and S-Fig. 6). The RPKM value of *AaDXR2* was the highest in 50%OH, followed by
417 IB15, L15, L7, HB and FOH. The RPKM value of *AaHDR1* was the highest in FOH,
418 followed by 50%OH, L7, IB15, HB and L15, and RT-PCR analysis showed its
419 expression in all 14 tissues (S-Fig. 5 and S-Fig. 6). The RPKM value of *AaHDR2* was
420 the highest in 50%OH, followed by IB15, FOH, L15, HB and L7.

421

422 One contig each for *MCT*, *CMK*, *MDS* and *HDS* was obtained from the sequences of the
423 six tissues, namely, *AaMCT*, *AaCMK*, *AaMDS* and *AaHDS*, respectively (S-Fig. 8 B).
424 The RPKM value of *AaMCT* was the highest in 50%OH, followed by FOH, L15, L7, HB
425 and IB15, that of *AaCMK* was the highest in FOH, followed by 50%OH, IB15, HB, L7
426 and L15, that of *AaMDS* was the highest in L7, followed by IB15, FOH, L15 and HB,
427 and that of *AaHDS* was the highest in L7, followed by L15, HB, FOH, 50%OH and IB15.

428

429 Two contigs for *IPPI*, *AaIPPI1* and *AaIPPI2*, were assembled from the sequences of the
430 six tissues (S-Fig. 8 B). Sequence alignment showed that the two contigs are two

homologs (Supplementary Results). The RPKM value of *AaIPPI1* was the highest in 50%OH, followed by HB, OF, IB15, L15 and L7, and RT-PCR analysis showed that *AaIPPI1* was expressed at a similar level in all 14 tissues (S-Fig. 5 and S-Fig. 6). The RPKM value of *AaIPPI2* was the highest in HB, followed by L7, IB15, FOH, L17 and 50%OH.

Genes Involved in the MVA Pathway

Four *AACT* contigs, *AaAACT1*, *AaAACT2*, *AaAACT3* and *AaAACT4*, were assembled from the sequences of the six tissues (S-Fig. 9 A). Sequence alignment showed that these four contigs are four homologs (Supplementary results). The RPKM value of *AaAACT1* was the highest in FOH, followed by L7, L17, HB and IB15, that of *AaAACT2* was the highest in HB, followed by 50%OH, L15, L7, FOH and IB15, and that of *AaAACT3* was the highest in HB, followed by 50%OH, IB15, L15, L7 and FOH, with the values for the last three tissues being similar. The RPKM value of *AaAACT4* was the highest in L7, followed by IB15, L15, 50%OH, FOH and HB.

Four *HMGS* contigs, *AaHMGS1*, *AaHMGS2*, *AaHMGS3* and *AaHMGS4*, were assembled from the sequences of the six tissues, and these four contigs represent two different homologs (Supplementary Results). As *AaHMGS1* and *AaHMGS3* are two fragments of the same homolog, *AaHMGS1* was used to represent this homolog. The RPKM value of *AaHMGS1* was the highest in 50%OH, followed by HB, L7, FOH and IB15 (S-Fig. 9 A). Similarly, as *AaHMGS2* and *AaHMGS4* are two fragments of the same homolog,

454 *AaHMGS2* was used to represent this homolog. The RPKM value of *AaHMGS2* was the
455 highest in IB15, followed by L7, HB, FOH and 50%OH (S-Fig. 9 A).

456

457 Three *HMGR* contigs, *AaHMGR1*, *AaHMGR2* and *AaHMGR3*, were assembled from the
458 sequences of the six tissues, and these three contigs constitute three homologs
459 (Supplementary Results). The RPKM value of *AaHMGR1* was the highest in HB,
460 followed by IB15, 50%OH, L15, FOH and L7 (S-Fig. 9 B). RT-PCR analysis showed
461 higher expression levels in the five flower tissues than in the nine leaf tissues (S-Fig. 5
462 and S-Fig. 6), supporting the expression pattern indicated by the RPKM values. The
463 RPKM value of *AaHMGR2* was the highest in HB, 50%OH, FOH, IB15, L7 and L15,
464 and that of *AaHMGR3* was the highest in HB, 50%OH, FOH, IB15, FOH and L7.

465

466 Two *PMK* contigs (*AaPMK1* and *AaPMK2*) and one contig each for *AaMVK* and
467 *AaMPDC* were assembled from the sequences of the six tissues (S-Fig. 9 B). Sequence
468 alignment showed that the *AaPMK1* and *AaPMK2* contigs are two fragments of the same
469 *PMK* homolog, which is thus described as *AaPMK*. The RPKM value of *AaPMK* was
470 the highest in HB, followed by 50%OH, FOH, L7, L15 and IB15 (S-Fig. 9 B). The
471 RPKM value of *AaMVK* was the highest in 50%OH, followed by FOH, IB15, HB, L15
472 and L7 and that of *AaMPDC* was the highest in 50%OH, followed by HB, L7, IB15, L15
473 and FOH.

474

475 **Integration of Gene Expression Levels and Terpenoid Profiles**

476

Pearson partial correlation analysis resulted in significant correlation networks ($r>0.9$ or $r<-0.9$, $p<0.05$) for 34 (out of 47) genes and 22 (out of 51) terpenoid metabolites (Fig. 7) in L7, L15, IB15, HB, FOH and 50%OH. The correlation networks were characterized by two groups, a big large and a small one, with gene-to-gene RPKM value profiles and RPKM value-to-metabolite accumulation patterns being either positively or negatively correlated in the six tissues. The large network showed positive or negative RPKM value correlations for 33 genes (nine MVA, six MEP, six artemisinin, six monoterpene and two non-amorphadiene pathway genes as well as two *AaFDS*, one *AaIPPI* and one triterpene genes). In addition, the large network was characterized by four patterns, each with one, one, 25 and six genes. For the 25 genes, 23 were positively correlated with each other, whereas two MVA pathway genes, *AaMVK* and *AaAACT4*, were negatively correlated with the others. The expression levels of six other genes were positively correlated each other, and *AaHDR1 and DBR2* were separated from the other genes by metabolites. It was interesting that *AaHMGR1* formed a single group. The networks also showed positive correlations between the RPKM values of certain genes and the metabolite accumulation patterns. The 22 terpenoids included nine monoterpenoid, eight non-amorphadiene, three triterpene and two amorphadiene metabolites. *AaFDS2* positively correlated with four sesquiterpenes, β-farnesene, caryophyllene, humulene and cis-muurola-4(14), 5-diene. The profile of artemisinic acid (arteannuic acid) was closely and positively correlated to *ADS* and *ALDH1*, and *unknown-SQTS (1* and *2)* was positively correlated with spiro[5.5]undec-2-ene ($C_{15}H_{24}$). Of five *AaLAS* homologs, only *AaLAS5* was positively correlated with linalool. In addition, negative correlations between genes and metabolites were found in the large network; for example, *CPR1* was negatively correlated with arteannuin b.

Effects of Overexpression and Down-regulation of *ADS* on Artemisinin, Artemisinic Acid and Arteannuin b Contents

One binary construct (Fig. 8 A) was developed to overexpress *ADS*. Multiple transgenic lines were obtained via *Agrobacterium*-mediated transformation of F6 homozygous plants (Fig. 8 B), and five T0 lines (ADS-GFP-1, -2, -3, -4 and -8) were selected to examine the effects of *ADS* overexpression on artemisinin synthesis. RT-PCR showed that the expression levels of *ADS* were increased in these five lines (Fig. 8 C), and artemisinin levels were significantly increased in leaf samples from three of the lines (Fig. 8 E and F). In addition, artemisinic acid and arteannuin b productions were significantly increased in two lines (Fig. 9 A).

In addition, one *ADS* RNAi construct (Fig. 8 A) developed to down-regulate gene expression was introduced into F6 progeny plants to obtain multiple transgenic RNAi lines (Fig. 8 B). Three lines (ADS-RNAi-2, -7 and -13) were selected for metabolic characterization. The expression levels of *ADS* were significantly reduced in all three lines (Fig. 8 D) as were the contents of artemisinin (Fig. 8 E and F), artemisinic acid and arteannuin b (Fig. 9 A).

Effects of Overexpression and Down-regulation of *ADS* on Non-amorphadiene Sesquiterpene and Metabolite Profiles

Six non-amorphadiene sesquiterpenes in transgenic vs. wild-type plants were detected by GC-MS analysis, and their contents were found to be differentially altered by *ADS*

526 over-expression and RNAi. β-Farnesene, caryophyllene and copaene were apparently
527 reduced in the three selected overexpressing plants (Fig. 9 B). In the three selected
528 RNAi plants, caryophyllene and copaene were apparently increased, whereas the content
529 of β-farnesene was decreased.

530

531 In addition, 76 other volatile metabolites (of the 88 in Supplementary table 2) were also
532 identified in transgenic vs. wild-type plants by GC-MS analysis. A PCA generated
533 using these 76 metabolites and the eight sesquiterpene metabolites described above
534 showed that *ADS* overexpression and RNAi altered genome-wide metabolite profiles
535 (Fig. 9 C). Metabolite profiles from three overexpressing, two wild-type and three
536 RNAi lines were separately grouped together. The metabolite profiles of the wild-type
537 plants were ordinately located in the middle of the overexpressing and RNAi lines.
538 These results indicated that *ADS* overexpression and RNAi altered non-polar metabolite
539 profiles in opposite manners.

540

541 **Discussion**

542

543 In our investigation, both genome-wide terpene profiling and analyses of individual
544 metabolites revealed a spatial and tissue-specificity trend of total volatile and terpene
545 scenarios in self-pollinating *A. annua*. PCA and clustering analyses showed a clear
546 ordinate separation between leaf and flower samples (Figs. 2 and 3), and the data
547 indicated that leaf position and flowering stage affect these metabolite profiles. In
548 addition, the number and molecular diversity of terpenes increased from lower to higher
549 leaf positions to flower tissues (S-Table 4). These data support a general theory that

550 volatile production is associated with tissue and development (Dudareva et al., 2013).
551 In addition, metabolite profiles varied among the five flower tissues. An ordinate
552 separation occurred between 50%OH and FOH or HB, and this phenomenon most likely
553 results from the shift of metabolite biosynthesis during the transition from HB formation
554 to blooming. This type of metabolic shift has been observed in the field, with
555 recommendations for harvesting plants during the flowering stage to obtain a better yield
556 of artemisinin and other essential oil compositions (Chalchat et al., 1994; Ferreira et al.,
557 1995a; Ferreira et al., 1995b). IB15 and IB16 were clustered together (Fig. 3), most
558 likely due to their similarity in tissue compositions, including inflorescence heads and
559 branch parts. Variation was also observed in the same type of tissues. For example,
560 the three biological samples of FOH were slightly separated from each other in the PCA
561 analysis. This separation most likely results from the effects of the uneven opening
562 status of ray and disc flowers in head inflorescences; when the head inflorescences were
563 fully opened, the opening stages of ray and disc flowers on the heads varied per head and
564 among heads.

565

566 Pyrosequencing of six cDNA libraries provided new information for enhancing the
567 understanding of artemisinin biosynthesis. The first committed step toward artemisinin
568 was unknown until ADS protein isolation (Bouwmeester et al., 1999) and cloning and
569 functional characterization of its cDNA (Chang et al., 2000; Mercke et al., 2000; Wallaart
570 et al., 2001). Trichomes were then targeted for EST generation and pyrosequencing
571 (Covello et al., 2007; Graham et al., 2010; Wang et al., 2009). These previous
572 sequencing investigations provided an appropriate platform for cloning *CYP71AV1*, *CPR*,
573 *DBR2* and *ALDH1*, which were sequentially demonstrated to catalyze the steps from
574 amorpha-4,11-diene to artemisinic acid and dihydroartemisinic acid *in vitro* (Ro et al.,
575 2006; Teoh et al., 2009; Teoh et al., 2006; Zhang et al., 2009; Zhang et al., 2008).

576 Recently, two new genes, namely, *ADH1* and *CYB5*, were cloned and integrated into
577 engineered yeasts to significantly scale up artemisinic acid production (Paddon et al.,
578 2013). Although such progress has greatly enhanced the understanding of artemisinin
579 biosynthesis, it has been highly challenging to obtain transgenic *A. annua* plants that can
580 be applied for artemisinin production. One remaining question is whether other genes
581 are involved in this pathway. In the present work, two *CPR* (*CPR1* and *CPR2*)
582 homologs were obtained from sequence assembly. CYP71AV1 and CPR have been
583 demonstrated to form redox partners in the conversion of amorpha-4, 11-diene to
584 artemisinic acid (Paddon et al., 2013; Ro et al., 2006; Teoh et al., 2006). The *CPR1*
585 sequence was obtained from six tissues, whereas the *CPR2* sequence was identified from
586 four flower tissues (Fig. 6A). RT-PCR also demonstrated the strong expression of
587 *CPR2* in flower tissues (S-Fig. 7). It is noteworthy that the contents of artemisinic acid
588 are higher in flower tissues than in leaves (Fig. 4 B), and the results suggest that *CPR2* is
589 likely involved in the formation of potential redox complexes associated with artemisinic
590 acid in flower tissues. In addition, CYB5 has been recently reported to partner with
591 CYP71AV1 and CPR (Paddon et al., 2013). We identified 15 CYB5 homolog contigs
592 based on our sequencing (data not shown), which will be pursued in future experiments.
593 Taken together, our sequence data suggest that additional genes are involved in the
594 biosynthesis of artemisinin.

595

596 The genome-wide terpene scenario is informative for understanding the competitive
597 pathways that affect artemisinin production in plants. Although numerous
598 sesquiterpenes and monoterpenes have been detected from tissues of heterozygous *A.*
599 *annua* plants grown under different field conditions (Ahmad and Misra, 1994; Bilia et al.,
600 2014; Goel et al., 2007; Holm et al., 1997; Woerdenbag et al., 1994; Woerdenbag et al.,
601 1993), no genome-wide terpene landscapes for a given growth environment have been

602 characterized in efforts to understand pathway networks that compete with artemisinin
603 production. In our investigation, artemisinin and 21 other sesquiterpenes (Table 2 and
604 S-Fig. 3) were identified from different tissues. Of them, arteannuin b, artemisinin,
605 caryophyllene, germacrene D and α-copaene were detected in 14 analyzed tissues (Table
606 2), artemisinic acid and β-farnesene were detected in 13 tissues, and the other 15
607 sesquiterpenes were differentially accumulated in 14 tissues. All of these sesquiterpenes
608 are synthesized from farnesyl diphosphate FPP) (Fig. 1), which is also the substrate of
609 triterpenes; therefore, the availability of FPP is the limiting factor for artemisinin
610 production. ADS cyclizes FPP to amorphadiene, which is the first step toward
611 artemisinic acid, arteannuin b and artemisinin. The sequence assembly obtained one
612 *ADS* contig, and the *ADS* RPKM values in HB, 50%OH, FOH and IB15 tissues were
613 approximately 324, 160, 176 and 89, respectively. Additionally, *AsGAS*, *AaCPS*,
614 *AaECS* and an unknown *SQTS* were identified as controlling steps toward
615 non-amorphadiene metabolites. The summed *AsGAS*, *AaCPS*, *AaECS* and unknown
616 *SQTS* RPKM values were approximately 518, 691, 357 and 91 in HB, 50%OH, FOH and
617 IB15, respectively. Moreover, the metabolite profiles indicate the existence of
618 additional sesquiterpene synthase genes involve in the formation of 19 other
619 non-amorphadiene sesquiterpenes. Accordingly, these data indicate that the
620 transcriptional activity of *ADS* essentially controls the production of artemisinin, a
621 hypothesis that was supported by *ADS* overexpression and RNAi in F6 plants. Indeed,
622 the overexpression of *ADS* increased the contents of artemisinin, artemisinic acid and
623 arteannuin b in transgenic lines (Fig. 8 E and F) but decreased the levels of
624 caryophyllene, germacrene D, copaene and β-farnesene. In contrast, the
625 down-regulation of *ADS* by RNAi decreased the contents of artemisinin, artemisinic acid
626 and arteannuin b but increased the contents of caryophyllene and copaene (Fig. 8 E and
627 F).

628

629 The genome-wide gene-to-terpene scenarios demonstrate the association of terpenoid
630 formation with gene expression during plant development. First, the gene expression
631 profiles are associated with artemisinin production during plant development.
632 Numerous reports have shown that artemisinin accumulation reaches a peak value prior
633 to the blooming stage (Chan et al., 1995; Ferreira, 2008; Liersch et al., 1986).
634 Unfortunately, the mechanism underlying this phenomenon remains unclear. In our
635 investigation, 14 tissues associated with different development stages were collected to
636 analyze artemisinin production and gene expression. The peak value of artemisinin in
637 HB (Fig. 4 E) was associated with the highest RPKM values for *ADS* and *ALDH1* (Fig. 6
638 A), and the RPKM values for *CYP71AV1* and *DBR2* in HB were the second highest
639 among the six sequenced tissues (Fig. 6 A). In addition to HB, other flower tissues
640 produced higher levels of artemisinic acids and artemisinin than did leaves. This
641 phenotype also resulted from the expression of *ADS*, *CYP71AV1*, *DBR2*, *ALDH1* and
642 *CPR*. The RPKM values for *ADS*, *CYP71AV1*, *ALDH1* and two *CPR* contigs were
643 higher in flowering tissues than in leaves (Fig. 6 A). RT-PCR data also supported this
644 result (S-Fig. 5 and S-Fig. 6). In addition, other individual sesquiterpenes were found to
645 be associated with plant development (Tables 1 and 2, S-Table 4). The diversity of
646 sesquiterpene molecules shows an apparent increasing trend from lower leaf positions to
647 flowers (S-Table 4). The number of detected sesquiterpenes was greater in flowering
648 tissues than in leaves (S-Table 4), and the most diverse sesquiterpene species were
649 detected in HB and FOH tissues. This metabolic phenomenon was closely associated
650 with gene expression. For example, the RPKM values of *AaGAS*, *AaCPS* and the
651 unknown *SQTS* were higher in flower tissues than in leaves (Fig. 6 B), and RT-PCR
652 supported this result (S-Fig. 5 and S-Fig. 6). Second, an integration analysis showed
653 significant correlations of the expression levels of the genes themselves and the

accumulation patterns of metabolites in the six tissues. The large network showed significant expression correlations of 34 genes in the six tissues (Fig. 7). Of these, twenty-three genes were positively correlated each other, and six other genes were positively correlated each other (Fig. 7). These correlations indicate that their expression profiles are simultaneously associated with six tissues. In addition, Pearson partial analysis significantly correlated twenty-two terpenoids from the six tissues in the networks. More importantly, the networks correlated gene expression levels with the formation certain metabolites in the six tissues. For example, *AaFDS2* was significantly correlated with β-farnesene, caryophyllene, humulene and cis-muurola-4(14),5-diene in these tissues. Third, based on this integration, we propose that development-associated terpene biosynthesis scenarios are closely related to glandular trichome formation on leaves and flowers. We previously estimated that the density of trichomes on leaf surfaces increased from L1 through 17 in F2 plants (Alejos-Gonzalez et al., 2011). The density of glandular trichomes was also higher on the surfaces of head inflorescence and flowers than on the surfaces of leaves, and this developmental feature is inherited by the progeny. We observed that glandular trichomes were the main site of the formation of essential oil drops that are released into the environment (S-Fig. 4 B). Essential oil drops are mainly composed of volatile monoterpenes and sesquiterpenes. Glandular trichomes are main site of artemisinin biosynthesis (Covello et al., 2007; Duke and Paul, 1993; Ferreira and Janick, 1995; Lommen et al., 2005). These data suggest the significance of glandular trichomes involved in the formation of terpenes in self-pollinated *A. annua*. In the immediate future, we will take advantage of these self-pollinated plants and gene-to-metabolite scenarios to focus on the manipulation of glandular trichome development and terpene biosynthesis.

Materials and Methods

Plant Growth

We previously reported the self-pollinating *A. annua* plants used in this study, which form an appropriate platform for understanding the biosynthesis of artemisinin and other terpenes (Alejos-Gonzalez et al., 2011). Seeds of F3 self-pollination progeny were used to grow plants in this investigation (Supplementary Materials and Methods).

Collection of Plant Materials

Samples were collected from 65- and 70-day old F3 progeny of self-pollinating plants grown under a short photoperiod (9/15 hrs). Tissue selection was based on our previous characterization of trichomes and artemisinin profiles (Alejos-Gonzalez et al., 2011). In brief, 65-day old plants were starting to develop first and second raceme inflorescences, whereas the 70-day old plants had flowered for approximately 5 days. The leaf samples were collected from the 7^{th}, 9^{th}, 10^{th}, 11^{th}, 12^{th}, 13^{th}, 14^{th}, 15^{th}, 16^{th} and 17^{th} nodes of the 65-day old plants. The flowering samples were collected from the inflorescence branches (including branch and head inflorescence), fully open head inflorescence and head inflorescence buds of the 70-day old plants. All sampling details are summarized in S-Table 1. After being excised from the plants, the tissues were immediately frozen in liquid nitrogen and then stored at -80°C for 454 sequencing and metabolic profiling.

Extraction of Non-polar Metabolites

Hexane was used to extract non-polar metabolites from 14 different tissues (S-Table 1). Frozen samples were ground into fine powder in liquid nitrogen. One hundred milligrams of powder from each sample was suspended in 0.5 ml of LS-MS grade hexane (100%) in a 2 ml Eppendorf tube. The samples were vigorously vortexed for 1 min and then sonicated for 10 min. The tubes were centrifuged at 10,000 rpm for 5 min, and the resulting supernatants were pipetted into new 2 ml tubes. The remaining pellet in each tube was extracted with a second 0.5 ml aliquot of hexane. The extraction steps were identical to those steps used in the first extraction. The two hexane supernatants were pooled together to obtain 1 ml of extract (in a 1.5 ml tube), followed by an additional centrifugation step at 13,000 rpm for 5 min. The resulting clean supernatant was pipetted into a new 1.5 ml tube. Two hundred microliters of hexane extract was pipetted into a 400 µl insert contained in 2 ml vials for GC-MS analysis. Three biological replicates were performed for each tissue position, and three technical replicates were carried out for each biological replicate.

GC-MS analysis

Metabolite analysis was performed using a gas chromatograph 6890 coupled with 5975C MSD (Agilent Technologies, USA). An RTX-5 capillary column (30 m×0.25 mm×0.25 µm) was used to separate metabolites. The splitless mode was used in the inlet. The injection temperature was set at 250°C. Two gradient temperature programs (TPs) of the column oven were optimized for metabolite analysis. The first (TP-1) was initially set at 80°C for 2 min, then ramped to 260°C at a constant rate of 10°C/min, and held at 260°C for 25 min. The second (TP-2) was initially set at 60°C and then ramped to 260°C at a constant rate of 10°C/min, and held at 260°C for 25 min. Pure helium was

used as the carrier gas, with a flow rate of 1 ml/min. A positive electron impact ion source (70 EV) was used to ionize compounds, and mass fragments were scanned in the range of 40-800 (m/z), with a 4 min of solvent delay. Total ion chromatographs were recorded for all biological and technical replicates of the 14 tissues. All metabolite peaks were recorded and stored in the format of Chemistation data files. A mixture of even-numbered chain n-alkenes (C10–C40) was purchased from Restek (Florida, USA, cat#31266), and these standards were used to estimate the retention index (RI) values, which were used to deconvolute the metabolite peaks.

Metabolite Peak Deconvolution and Statistical Analysis

Most of the peaks detected by GC-MS analysis were deconvoluted using Agilent MassHunter Mass Profiler (MHMP) and Mass Profiler Professional (MPP) software. The MHMP software was developed to operate using compound data from different samples in ".cef" (compound exchange file) format files. All compound data were produced by MassHunter Qualitative Analysis software, and all data underwent both qualitative and quantitative evaluations. Two steps were performed to identify the metabolites. First, the ChemiStation mass data were translated into MassHunter data files using Agilent MassHunter GC/MS Translator software (version B.05.02). Second, unknown peaks were deconvoluted by mass analysis using the Mass Hunter Qualitative Analysis software (Version B.06.00). Compound searches and identification in the NIST 11 library were also completed using the same software. After this devolution, each deconvoluted metabolite peak data was exported as a ".cef" file.

All deconvoluted metabolites were used as the data matrix for PCA, multivariate analysis and other statistical analyses. The ".cef" data files of the metabolites were imported into

753 MPP software (version B.12.5, Agilent), which was developed by Agilent for PCA,
754 multivariate analysis, heat map and other statistical analyses (included in the software).
755 For each metabolite, the presence or absence as well as its abundance in tissues were used
756 to generate plots and heatmaps using MPP software. This software has been reported to
757 be effective for characterizing metabolic profiles (Bannur et al., 2014; Chan et al., 2013;
758 Chen et al., 2012; Hu et al., 2013) and can analyze the absolute abundance and retention
759 time, align data, normalize all data and set up the baseline to the median of all sample
760 data. In brief, the absolute abundance and retention time tolerance (Barnes et al., 2013)
761 across the sample sets were filtered using a minimum of 5000 counts and 0.05 min.
762 Then, all the data were aligned, normalized and baselined to the median level of all
763 samples for each experiment. The fold changes obtained for all metabolites were log 2
764 values obtained from the transformation of original accounts from the mean value for
765 each group compared with the minimum peak abundance of each identified compound.
766 Hierarchical clustering (condition tree), coupled with the heat map and PCA, were
767 generated to demonstrate the relationships among the samples with regard to metabolite
768 profiles.

769

770 **HPLC-MS Analysis of Artemisinin**

771

772 To estimate artemisinin levels using HPLC-MS analysis on a 2010 eV LC/UV/ESI/MS
773 instrument (Shimadzu), the samples (S-Table 1) were analyzed following our previously
774 reported protocol (Alejos-Gonzalez et al., 2013; Alejos-Gonzalez et al., 2011). The
775 steps and methods are included in Supplementary Materials and Methods.

776

RNA Isolation, cDNA Library Construction and 454 Sequencing

Total RNA was isolated from 14 different tissues (S-Table 1) using an RNeasy Midi Kit (Qiagen) (http://www.qiagen.com/). One gram of sample was homogenized into fine powder in liquid nitrogen. The RNA extraction followed the manufacturer's protocol. RNA samples were treated with RNase-free DNase I (Promega, USA) to eliminate any genomic DNA contamination. Each step followed the manufacturer's protocol. Messenger RNA (mRNA) was purified from total RNA samples using an Oligotex mRNA Mini Kit (Qiagen, catalog#: 70022). Each step followed the protocol from Qiagen.

The synthesis, amplification and normalization of cDNA were completed using three different kits. Each step of the experiments followed the manufacturers' protocols. The SMART cDNA synthesis protocol (Clontech, USA) was employed to prepare the construction of the *A. annua* cDNA library. Five micrograms of mRNA purified as described above was used as the template to synthesize the first-strand cDNA using the Advantage-2 PCR kit (Clontech, USA), following the manufacturer's protocol. Based on the instructions, PCR was optimized at 15 cycles to amplify double-stranded cDNA, which was then purified using a Purelink® PCR Purification Kit (Invitrogen, USA). The purified double-stranded cDNA samples were normalized using a cDNA nominalization kit (catalog #: NK003, Evrogen JSC, 16/10 Miklukho-Maklaya str., Moscow, Russia) (http://www.evrogen.com/).

Sequencing was completed using a 454 GS FLZ Titanium Sequencer at the Genomic Sciences Laboratory at North Carolina State University (http://research.ncsu.edu/gsl/). All cDNA samples were processed by exactly following the Roche's rapid library preparation kit instructions (Roche, IN, USA). Six cDNA libraries were run on one lane on the FLX Titanium platform following the manufacturer's protocol.

Transcript Assembly and Annotation

Raw sequence data were processed by trimming adaptor sequences and removing low-quality or empty reads. The clean reads obtained from all the samples were assembled and clustered using TIGR Gene Indices clustering tools software (Pertea et al., 2003). The resulting contigs were functionally annotated by searching the UniProt database (Magrane and UniProt-Consortium, 2011) using the standalone BLASTX program (Altschul et al., 1997) with an E-value $\leq 1\times10^{-5}$.

Estimation of Gene Expression Levels Using Transcriptomes

The level of gene expression in each sample was estimated by the number of reads that were assembled into a contig. The expression value was normalized for the length of the contig and for the total number of reads from a cDNA library by calculating RPKM (reads per kilobase per million). The gene expression data were then used for a hierarchical clustering analysis of the samples and contigs using the TM4 MeV software package (Saeed et al., 2003). The Euclidean metric was used as the distance measure.

823

824 **Integration of Gene Expression and Terpenoid Molecule Profiles**

825

826 Correlation between the expression levels of 47 genes and the profiles of 51 terpenoids in
827 L7, L15, IB15, HB, FOH and 50%OH were carried out using the program R 2.10.1. For
828 the unknown *SQTS*, two contigs (1 and 2) were used to test whether they could be
829 integrated consistently. RPKM values for genes and the peak values of metabolites
830 were used as a matrix for Pearson partial correlation analysis. The resulting correlation
831 networks were obtained and visualized using Cytoscape software (Cytoscape 2.6.3)
832 (Smoot et al., 2011).

833

834 **Semi-quantitative RT-PCR**

835

836 Semi-quantitative RT-PCR was performed to characterize the expression profiles of 19
837 genes. The methods are included in Supplementary Materials and Methods.

838

839 **Cloning of *ADS*, Genetic Transformation and Metabolite Analysis**

840

841 *ADS* cDNA was cloned from the leaves of F6 progeny plants. Binary vectors for *ADS*
842 overexpression and RNAi were constructed to understand *ADS*'s involvement in the
843 biosynthesis of artemisinin *in planta*. The gene cloning, genetic transformation and
844 sample collection methods are included in Supplementary Materials and Methods.

GC-MS based profiling using TP-1, metabolite annotation and PCA analysis was as described above.

The methods and procedures are included in Supplementary Materials and Methods.

Student's T Test

Student's test (P values < 0.05) was used to evaluate the statistical significance of artemisinin and selected terpenes in the samples.

Acknowledgement

This project was funded by the North Carolina Biotechnology Center, North Carolina.

Supplementary Results (9 figures, 4 tables and sequence alignments)

860 **References**

861

862 Ahmad, A., and Misra, L.N. (1994). Terpenoids from *Artemisia annua* and constituents
863 of its essential oil. Phytochemistry 37:183-186.

864 Alam, P., and Abdin, M.Z. (2011). Over-expression of HMG-CoA reductase and
865 amorpha-4,11-diene synthase genes in *Artemisia annua* L. and its influence on
866 artemisinin content. Plant Cell Rep 30:1919-1928.

867 Alejos-Gonzalez, F., Perkins, K., Winston, M.I., and Xie, D.-Y. (2013). Efficient somatic
868 embryogenesis and organogenesis of self-pollination *Artemisia annua* progeny
869 and artemisinin formation in regenerated plants. AJPS 4:2206-2217

870 Alejos-Gonzalez, F., Qu, G.S., Zhou, L.L., Saravitz, C.H., Shurtleff, J.L., and Xie, D.Y.
871 (2011). Characterization of development and artemisinin biosynthesis in
872 self-pollinated *Artemisia annua* plants. Planta 234:685-697.

873 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and
874 Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
875 protein database search programs. Nucleic Acids Res 25:3389-3402.

876 Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and
877 Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
878 protein database search programs. Nucleic Acids Res 25:3389-3402.

879 Banek, K., Lalani, M., Staedke, S.G., and Chandramohan, D. (2014). Adherence to
880 artemisinin-based combination therapy for the treatment of malaria: a systematic
881 review of the evidence. Malar J 13.

882 Bannur, Z., Teh, L.K., Hennesy, T., Rosli, W.R.W., Mohamad, N., Nasir, A., Ankathil,
883 R., Zakaria, Z.A., Baba, A., and Salleh, M.Z. (2014). The differential metabolite
884 profiles of acute lymphoblastic leukaemic patients treated with 6-mercaptopurine
885 using untargeted metabolomics approach. Clin Biochem 47:427-431.

886 Barnes, B.B., Wilson, M.B., Carr, P.W., Vitha, M.F., Broeckling, C.D., Heuberger, A.L.,
887 Prenni, J., Janis, G.C., Corcoran, H., Snow, N.H., et al. (2013). "Retention
888 projection" enables reliable use of shared gas chromatographic retention data
889 across laboratories, instruments, and methods. Anal Chem 85:11650-11657.

890 Bhakuni, R.S., Jain, D.C., Sharma, R.P., and Kumar, S. (2001). Secondary metabolites of
891 *Artemisia annua* and their biological activity. Curr Sci 80:35-48.

892 Bilia, A.R., Santomauro, F., Sacco, C., Bergonzi, M.C., and Donato, R. (2014). Essential
893 oil of *Artemisia annua* L.: An extraordinary component with numerous
894 antimicrobial properties. Evid-based Complement Altern Med 2014:1-7.

895 Bouwmeester, H.J., Wallaart, T.E., Janssen, M.H.A., van Loo, B., Jansen, B.J.M.,
896 Posthumus, M.A., Schmidt, C.O., De Kraker, J.-W., Konig, W.A., and Franssen,
897 M.C.R. (1999). Amorpha-4,11-diene synthase catalyses the first probable step in
898 artemisinin biosynthesis. Phytochemistry 52:843-854.

899 Brisibe, E.A., Udensi, O., Chukwurah, P.N., de Magalhaes, P.M., Figueira, G.M., and
900 Ferreira, J.F.S. (2012). Adaptation and agronomic performance of *Artemisia*
901 *annua* L. under lowland humid tropical conditions. Ind Crop Prod 39:190-197.

902 Brown, G.D. (2010). The biosynthesis of artemisinin (Qinghaosu) and the
903 phytochemistry of *Artemisia annua* L. (Qinghao). Molecules 15:7603-7698.

904 Chalchat, J.C., Garry, R.P., and Lamy, J. (1994). Influence of harvest time on yield and
905 composition of *Artemisia annua* oil produced in France. J Essent Oil Res 6
906 261-268.

907 Chan, D., Fussell, R.J., Hetmanski, M.T., Sinclair, C.J., Kay, J.F., Grant, A., and
908 Sharman, M. (2013). Investigation of the fate of trifluralin in shrimp. J Agric
909 Food Chem 61:2371-2377.

910 Chan, K.L., Teo, C.K.H., Jinadasa, S., and Yuen, K.H. (1995). Selection of high
911 artemisinin yielding *Artemisia annua* Planta Med 61:285-287.

912 Chang, Y.-J., Song, S.-H., Park, S.-H., and Kim, S.-U. (2000). Amorpha-4,11-diene
913 synthase of *Artemisia annua*: cDNA isolation and bacterial expression of a
914 terpene synthase Involved in artemisinin biosynthesis. Arch Biochem Biophys
915 383:178-184.

916 Charles, D.J., Simon, J.E., Wood, K.V., and Heinstein, P. (1990). Germplasm variation in
917 artemisinin content of Artemisia annua using an alternative method of artemisinin
918 analysis from crude plant extracts. J Nat Prod 53:157-160.

919 Chen, Q.Y., Park, H.C., Goligorsky, M.S., Chander, P., Fischer, S.M., and Gross, S.S.
920 (2012). Untargeted plasma metabolite profiling reveals the broad systemic
921 consequences of xanthine oxidoreductase inactivation in mice. PLoS ONE
922 7:e37149.

923 Chen, Y.F., Shen, Q., Wang, Y.Y., Wang, T., Wu, S.Y., Zhang, L., Lu, X., Zhang, F.Y.,
924 Jiang, W.M., Qiu, B., et al. (2013). The stacked over-expression of FPS,
925 CYP71AV1 and CPR genes leads to the increase of artemisinin level in *Artemisia*
926 *annua* L. Plant Biotechnol Rep 7:287-295.

927 Cockram, J., Hill, C., Burns, C., Arroo, R.R.J., Woolley, J.G., Flockart, I., Robinson, T.,
928 Atkinson, C.J., Davies, M.J., Dungey, N., et al. (2012). Screening a diverse
929 collection of *Artemisia annua* germplasm accessions for the antimalarial
930 compound, artemisinin. Plant Genet Resour-Charact Util 10:152-154.

931 Covello, P.S., Teoh, K.H., Polichuk, D.R., Reed, D.W., and Nowak, G. (2007).
932 Functional genomics and the biosynthesis of artemisinin. Phytochemistry
933 68:1864-1971.

934 Delabays, N., Simonnet, X., and Gaudin, M. (2001). The genetics of artemisinin content
935 in *Artemisia annua* L. and the breeding of high yielding cultivars. Curr Med
936 Chem 8:1795-1801.

937 Dudareva, N., Klempien, A., Muhlemann, J.K., and Kaplan, I. (2013). Biosynthesis,
938 function and metabolic engineering of plant volatile organic compounds. New
939 Phytol 198:16-32.

940 Duke, S.O., and Paul, R.N. (1993). Development and fine structure of the glandular
941 trichomes of *Artemisia annua* L. Int J Plant Sci 154:107-118.

942 Elfawal, M.A., Towler, M.J., Reich, N.G., Weathers, P.J., and Rich, S.M. (2015). Dried
943 whole-plant *Artemisia annua* slows evolution of malaria drug resistance and
944 overcomes resistance to artemisinin. PNAS 112:821-826.

945 Farhi, M., Kozin, M., Duchin, S., and Vainstein, A. (2013). Metabolic engineering of
946 plants for artemisinin synthesis. In: Biotechnology and Genetic Engineering
947 Reviews--Harding, S.E., and Adams, G.G., eds. London: Taylor & Francis Ltd.
948 135-148.

949 Farhi, M., Marhevka, E., Ben-Ari, J., Algamas-Dimantov, A., Liang, Z., Zeevi, V.,
950 Edelbaum, O., Spitzer-Rimon, B., Abeliovich, H., Schwartz, B., et al. (2011).
951 Generation of the potent anti-malarial drug artemisinin in tobacco. Nat Biotech
952 29:1072-1074.

953 Ferreira, J.F.S. (2008). Seasonal and post-harvest accumulation of artemisinin,
954 artemisinic acid, and dihydroartemisinic acid in three accessions of *Artemisia*
955 *annua* cultivated in West Virginia, USA Planta Med 74.:O1.

Ferreira, J.F.S., and Janick, J. (1995). Floral morphology of *Artemisia annua* with special reference to trichomes. Int J Plant Sci 156:807-815.

Ferreira, J.F.S., Simon, J.E., and Janick, J. (1995a). Developmental studies of *Artemisia annua*: Flowering and artemisinin production under greenhouse and field conditions Planta Med 61:167-170.

Ferreira, J.S., Simon, J.E., and Janick, J. (1995b). Relationship of artemisinin content of tissue-cultured, greenhouse-grown, and field-grown plants of *Artemisia annua.* Planta Med. 61:351-355.

Goel, D., Singh, V., Ali, M., Mallavarupu, G., and Kumar, S. (2007). Essential oils of petal, leaf and stem of the antimalarial plant *Artemisia annua*. J Nat Med 61:187-191.

Gogtay, N., Kannan, S., Thatte, U.M., Olliaro, P.L., and Sinclair, D. (2013). Artemisinin-based combination therapy for treating uncomplicated *Plasmodium vivax* malaria. Cochrane Database Syst Rev.

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.

Holm, Y., Laakso, I., Hiltunen, R., and Galambosi, B. (1997). Variation in the essential oil composition of *Artemisia annua L.* of different origin cultivated in Finland. Flav Fragr J 12:241-246.

Hu, J.B., Jiang, F.S., Gu, H.C., Ding, Z.S., Yao, L., Fan, Y.S., and Ding, X.H. (2013). Metabolomics study on the effects of Jieduquyuziyin prescription on systemic lupus erythematosus mice by LC-Q-TOF/MS. Chromatographia 76:791-800.

Jelinek, T. (2013). Artemisinin based combination therapy in travel medicine. Travel Med Infect Dis 11:23-28.

Jiang, L.X., Liu, H., Wang, J.B., Tan, F.R., Zhao, K., Wu, X., Zhu, H., Tang, K.X., and Tang, X.M. (2010). Characterization and comparison of three transgenic *Artemisia annua* varieties and wild-type variety in environmental release trial. J Med Plants Res 4:2719-2728.

Kjaer, A., Grevsen, K., and Jensen, M. (2012). Effect of external stress on density and size of glandular trichomes in full-grown *Artemisia annua*, the source of anti-malarial artemisinin. Aob Plants.

Klayman, D.L. (1985). Qinghaosu (Artemisinin) - An antimalarial drug from China. Science 228:1049-1055.

Larson, T.R., Branigan, C., Harvey, D., Penfield, T., Bowles, D., and Graham, I.A. (2013). A survey of artemisinic and dihydroartemisinic acid contents in glasshouse and global field-grown populations of the artemisinin-producing plant *Artemisia annua* L. Ind Crop Prod 45:1-6.

Liersch, R., Soicke, H., Stehr, C., and Tüllner, H.-U. (1986). Formation of artemisinin in *Artemisia annua* during one vegetation period. Planta Med 52:387-390.

Liu, B.Y., Wang, H., Du, Z.G., Li, G.F., and Ye, H.C. (2011). Metabolic engineering of artemisinin biosynthesis in *Artemisia annua* L. Plant Cell Rep 30:689-694.

Liu, J.-M., Ni, M.-Y., Fan, J.-F., Tu, Y.-Y., Wu, Z.-H., Wu, Y.-L., and Chou, W.-S. (1979). Structure and reaction of arteannuin. Acta Chim Sin 37:129-143.

Lommen, W.J., Schenk, E., Bouwmeester, H.J., and Verstappen, F.W. (2005). Trichome dynamics and artemisinin accumulation during development and senescence of *Artemisia annua* leaves. Planta Med 71:336.

Magrane, M., and UniProt-Consortium. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database-the Journal of Biological Databases and Curation.

Mercke, P., Bengtsson, M., Bouwmeester, H.J., Posthumus, M.A., and Brodelius, P.E. (2000). Molecular cloning, expression, and characterization of amorpha-4,11-diene synthase, a key enzyme of artemisinin biosynthesis in *Artemisia annua L.* Arch Biochem Biophys 381:173-180.

Milhous, W.K., and Weina, P.J. (2010). The Botanical Solution for Malaria. Science 327:279-280.

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-+.

1016 Paul, S., Khanuja, S.P.S., and Gupta, M.M. (2014). Breeding strategy for genetic
1017 improvement up to four generations in relation to artemisinin with canopy and
1018 other secondary metabolites in *Artemisia annua* L. Ind Crop Prod 56:67-73.

1019 Pertea, G., Huang, X.Q., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y.,
1020 White, J., Cheung, F., Parvizi, B., et al. (2003). TIGR Gene Indices clustering
1021 tools (TGICL): a software system for fast clustering of large EST datasets.
1022 Bioinformatics 19:651-652.

1023 Pu, G.B., Ma, D.M., Wang, H., Ye, H.C., and Liu, B.Y. (2013). Expression and
1024 localization of amorpha-4,11-diene synthase in *Artemisia annua* L. Plant Mol Biol
1025 Rep 31:32-37.

1026 Ro, D.-K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L., Ndungu, J.M., Ho,
1027 K.A., Eachus, R.A., Ham, T.S., Kirby, J., et al. (2006). Production of the
1028 antimalarial drug precursor artemisinic acid in engineered yeast. Nature
1029 440:940-943.

1030 Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa,
1031 M., Currier, T., Thiagarajan, M., et al. (2003). TM4: A free, open-source system
1032 for microarray data management and analysis. Biotechniques 34:374-+.

1033 Shen, Q., Chen, Y.F., Wang, T., Wu, S.Y., Lu, X., Zhang, L., Zhang, F.Y., Jiang, W.M.,
1034 Wang, G.F., and Tang, K.X. (2012). Overexpression of the cytochrome P450
1035 monooxygenase (cyp71av1) and cytochrome P450 reductase (cpr) genes
1036 increased artemisinin content in *Artemisia annua* (Asteraceae). Genet Mol Res
1037 11:3298-3309.

1038 Shunmay, Y., Van Damme, W., Socheat, D., White, N.J., and Mills, A. (2008). Access to
1039 artemisinin combination therapy for malaria in remote areas of Cambodia.
1040 Malaria J 7:1-14, 14P.

1041 Simonnet, X., Quennoz, M., and Carlen, C. (2008). New *Artemisia annua* hybrids with
1042 high artemisinin content. In: Proceedings of the International Symposium on
1043 Asian Plants with Unique Horticultural Potential--Zhang, D., Lee, J.M., and Tao,
1044 R., eds. Leuven 1: International Society Horticultural Science. 371-373.

1045 Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8:
1046 new features for data integration and network visualization. Bioinformatics
1047 27:431-432.

1048 Sy, L.K., and Brown, G.D. (2002). The mechanism of the spontaneous autoxidation of
1049 dihydroartemisinic acid. Tetrahedron 58:897-908.

1050 Tang, K.X., Shen, Q., Yan, T.X., and Fu, X.Q. (2014). Transgenic approach to increase
1051 artemisinin content in *Artemisia annua* L. Plant Cell Rep 33:605-615.

1052 Taylor, S. (2013). Artemisia annua, Artemisinin, ACTs and Malaria Control in Africa:
1053 Tradition, Science and Public Policy. Soc Hist Med 26:799-801.

1054 Tellez, M.R., Canel, C., Rimando, A.M., and Duke, S.O. (1999). Differential
1055 accumulation of isoprenoids in glanded and glandless *Artemisia annua L.*
1056 Phytochemistry 52:1035-1040.

1057 Teoh, K.H., Polichuk, D.R., Reed, D.W., and Covello, P.S. (2009). Molecular cloning of
1058 an aldehyde dehydrogenase implicated in artemisinin biosynthesis in *Artemisia*
1059 *annua*. Botany 87:635-642.

1060 Teoh, K.H., Polichuk, D.R., Reed, D.W., Nowak, G., and Covello, P.S. (2006). *Artemisia*
1061 *annua L.* (*Asteraceae*) trichome-specific cDNAs reveal CYP71AV1, a
1062 cytochrome P450 with a key role in the biosynthesis of the antimalarial
1063 sesquiterpene lactone artemisinin. FEBS Lett 580:1411-1416.

1064 Tu, Y.-y., Ni, M.-y., Zhong, Y.-r., Li, L.-n., Cui, S.-l., Zhang, M.-q., Wang, X.-z., Ji, Z.,
1065 and Xiao-tian, L. (1982). Studies on the constituents of *Artemisia annua* Part II
1066 Planta Med ; 44:143-114.

1067 Turconi, J., Griolet, F., Guevel, R., Oddon, G., Villa, R., Geatti, A., Hvala, M., Rossen,
1068 K., Goller, R., and Burgard, A. (2014). Semisynthetic artemisinin, the chemical
1069 path to industrial production. Org Process Res Dev 18:417-422.

1070 van Herpen, T., Cankar, K., Nogueira, M., Bosch, D., Bouwmeester, H.J., and
1071 Beekwilder, J. (2010). Nicotiana benthamiana as a production platform for
1072 artemisinin precursors. PLoS ONE 5.

1073 Wallaart, T.E., Bouwmeester, H.J., Hille, J., Poppinga, L., and Maijers, N.C.A. (2001).
1074 Amorpha-4,11-diene synthase: cloning and functional expression of a key enzyme
1075 in the biosynthetic pathway of the novel antimalarial drug artemisinin. Planta
1076 212:460-465.

1077 Wang, W., Wang, Y., Zhang, Q., Qi, Y., and Guo, D. (2009). Global characterization of
1078 *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing.
1079 BMC Genomics 10:465.

Weathers, P.J., Elfawal, M.A., Towler, M.J., Acquaah-Mensah, G.K., and Rich, S.M. (2014a). Pharmacokinetics of artemisinin delivered by oral consumption of *Artemisia annua* dried leaves in healthy vs. *Plasmodium chabaudi*-infected mice. J Ethnopharmacol 153:732-736.

Weathers, P.J., Jordan, N.J., Lasin, P., and Towler, M.J. (2014b). Simulated digestion of dried leaves of *Artemisia annua* consumed as a treatment (pACT) for malaria. J Ethnopharmacol 151:858-863.

Wei, Z.-x., Pan, J.-p., and Li, Y. (1992). Artemisinin G: a sesquiterpene from *Artemisia annua* Planta Med 58:300

Woerdenbag, H., Pras, N., Chan, N.G., Bang, B.T., Bos, R., van Uden, W., Van Y, P., Van Boi, N., Batterman, S., and Lugt, C. (1994). Artemisinin, related sesquiterpenes, and essential oil in *Artemisia annua* during a vegetation period in Vietnam. Planta Med 60:272-274.

Woerdenbag, H.J., Pras, N., Chan, N.G., Bang, B.T., Bos, R., van Uden, W., van Y, P., Van Boi, N., Batterman, S., and Lugt, C.B. (1993). Artemisinin, related sesquiterpenes, and essential oil in *Artemisia annua* during a vegetation period in Vietnam Planta Med 59:A699.

Xiang, L., Zeng, L.X., Yuan, Y., Chen, M., Wang, F., Liu, X.Q., Zeng, L.J., Lan, X.Z., and Liao, Z.H. (2012). Enhancement of artemisinin biosynthesis by overexpressing dxr, cyp71av1 and cpr in the plants of *Artemisia annua* L. Plant Omics 5:503-507.

Zhang, Y., Nowak, G., Reed, D.W., and Covello, P.S. (2010). The production of artemisinin precursors in tobacco. Plant Biotechnol J DOI: 10.1111/j.1467-7652.2010.00556.x:In press.

Zhang, Y., Teoh, K.H., Reed, D.W., and Covello, P.S. (2009). Molecular cloning and characterization of Dbr1, a 2-alkenal reductase from *Artemisia annua*. Botany 87:643-649.

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 Δ-11(13) reductase and its role in glandular trichome-dependent biosynthesis of artemisinin in *Artemisia annua*. J Biol Chem 283:21501-21508.

1111 Zhong-Yi-Yan-Jiu-Yue-Zhong-Yao-Yan-Jiu-Suo. (1978). Qing Hao Kang Nue Yan
1112 Jiu:1971-1978 (Antimalarial studies of *Artemisia annua* L:1971-1978): Zhong Yi
1113 Yan Jiu Yue Zhong Yao Yan Jiu Suo.

1114

1115

1116

Figure legends

Figure 1. The compartmented global MVA and MEP pathways toward terpene biosynthesis in *Artemisia annua*.

The following enzymes are involved in the MVA pathway: AACT, acetoacetyl-CoA thiolase; HMGS, 3-hydroxy-3-methylglutaryl-coenzyme A synthase; HMGR, 3-hydroxy-3-methylglutaryl coenzyme A reductase; MK, mevalonate kinase (MVK); PMK, 5-phosphatemevalonate kinase; MPDC, mevalonate-5-phosphate decarboxylase; IPPI, isopentenyl pyrophosphate isomerase. The following enzymes are involved in the MEP pathway: DXS, 1-deoxy-D-xylulose-5-phosphate synthase; DXR,1-deoxy-D-xylulose 5-phosphate reductase; MCT, 2-C-methyl-D-erythritol-4-phosphate cytidyltransferase; CMK, cytidyl (4-diphospho)-2-C-methyl-D-erythritol kinase; MDS, 2-C-methyl-D-erythritol-2,4-cyclodiphohate synthase (MCS); HDS, 1-hydroxyl-2-methyl-2-(E)-butenyl-diphosphate synthase; HDR, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase; IPPI, isopentenyl pyrophosphate isomerase. The following enzymes are involved in the pathways toward monoterpene and diterpene formation: LAS, linalool synthase; LS, limonene synthase; CIN, cineole synthase; BPS, beta-pinene synthase; GGPPS, geranylgeranyl pyrophosphate synthase. The following enzymes are involved in the artemisinin biosynthetic pathway: FDS, farnesyl diphosphate; ADS, amorpha-4,11-diene synthase; CPR, cytochrome P450 reductase; CYP71AV1, cytochrome P450 monooxygenase; DBR2, artemisinic aldehyde Δ-11(13) double bond reductase; ALDH1, aldehyde dehydrogenase 1; CYB5 and ADH1, cytochrome b5 monooxygenase and alcohol dehydrogenase (however, the steps toward artemisinic acid have not been clearly demonstrated). The following enzymes are involved in the biosynthetic pathways toward other sesquiterpenes and triterpenes: BFS,

1143 beta-farnesene synthase; CPS, beta-caryophyllene synthase; GAS, germacrene A
1144 synthase; ECS, 8-epi-cedrol synthase; SQS, squalene synthase; BAS, beta-amyrin
1145 synthase. Question marks indicate that these steps toward artemisinin production are
1146 unclear.

1147

1148 Figure 2. Principle component analysis for metabolites from fourteen tissues.
1149 Eighty-seven metabolites (S-Table 2) identified by GC-MS analysis and their abundance
1150 were used as the data matrix for PCA. Sample abbreviations: L7-L17, the 7^{th} to 17^{th}
1151 leaves; IB15 and IB16, the inflorescence branches at the 15^{th} and 16^{th} nodes; 50%OH:
1152 50% full open and 50% head inflorescence buds; HB: head inflorescence buds; FOH:
1153 fully open head inflorescence.

1154

1155 Figure 3. A heat map coupled with clustering analysis for eighty-seven metabolites
1156 displaying differential profiles in fourteen tissues. Eighty-seven metabolites identified
1157 by GC-MS and their abundance were used as the data matrix for heat map and clustering
1158 analysis. Sample abbreviations: L7-L17, the 7^{th} to 17^{th} leaves; IB15 and IB16, the
1159 inflorescence branches at the 15^{th} and 16^{th} nodes; 50%OH, 50% full open and 50%
1160 head inflorescence buds; HB, head inflorescence buds; FOH, fully open head
1161 inflorescence. "*": monoterpenes; "**": sesquiterpenes; "***": triterpenes.

1162

1163 Figure 4. Accumulation patterns of artemisinin and of eight other terpenes in different
1164 tissues. The artemisinin contents were estimated using a standard curve established
1165 with an authentic artemisinin standard. The levels of eight other terpenes were
1166 estimated either using their standard curve or using an equivalent value calculated by an

1167 equivalent terpene standard curve. A, contents of 1,8-cineole and β-pinene; B, contents
1168 of artemisinic acid and arteannuin B; C, contents of β-copaene, caryophyllene and
1169 β-farnesene; D, contents of squalene and β-amyrin; and E, artemisinin content patterns.
1170 Sample abbreviations: L7-L17, the 7^{th} to 17^{th} leaves; IB15 and IB16, the inflorescence
1171 branches at the 15^{th} and 16^{th} nodes; 50%OH, 50% full open and 50% head inflorescence
1172 buds; HB, head inflorescence buds; FOH, fully open head inflorescence.

1173

1174 Figure 5. An HCL Euclidean heat map showing genome-wide gene expression
1175 differentiation among two positional groups of leaves (L7 and L15, the 7^{th} and 15^{th}
1176 leaves), one positional group of inflorescence branches (IB15, inflorescence branch at the
1177 15^{th} node), full open head inflorescence (FOH), head inflorescence buds (HB) and 50%
1178 full open and 50% head inflorescence buds (50%OH).

1179

1180 Figure 6. Reads per kilobase per millions (RPKM) of 454 sequencing for genes involved
1181 in the biosynthesis of artemisinin and other terpenes in six different tissues. The y-axis
1182 values are expressed using a logarithmic scale. The numbers labeled on the top of each
1183 bar are true RPKM values. A, two farnesyl diphosphate synthase (*FDS*) genes and
1184 seven genes involved in artemisinin biosynthesis (abbreviations: *ADS*, amorpha-4,
1185 11-diene synthase; *CPR*, cytochrome P450 reductase; *CYP71AV1*, cytochrome P450
1186 monooxygenase; *DBR2*, artemisinic aldehyde Δ11(13) reductase; *ALDH1*, aldehyde
1187 dehydrogenase 1). B, Four genes involved in other sesquiterpene and triterpene
1188 biosynthesis (abbreviations: "*Aa*" prefix in each gene, *A. annua*; *AaCPS*,
1189 beta-caryophyllene synthase; *AaGAS*, germacrene A synthase; *AaECS*, 8-epi-cedrol
1190 synthase; Unknown sesquiterpene synthase (SQTS); *AaSQS*, squalene synthase). C,
1191 Four genes involved in monoterpene biosynthesis (abbreviations: *AaLAS*, linalool

1192 synthase; *AaLS*, limonene synthase; *AaCIN*, cineole synthase; *AaBPS*, beta-pinene
1193 synthase). Sample abbreviations: L7 and L15, the 7th and 15th leaves; IB15,
1194 inflorescence branch at the 15th node; FOH, full open head inflorescence; HB, head
1195 inflorescence buds; 50%OH, 50% full open and 50% head inflorescence buds.

1196

1197 Figure 7. A large and a small significant correlation networks generated by the
1198 integration of expression levels (RKPM values) of 34 genes and the accumulation levels
1199 of 22 terpenoids using Pearson partial correlation analysis ($r>0.9$ or $r<-0.9$, $p<0.05$).
1200 Abbreviations, Cyclo-H3M6(M): cyclohexene, 3-methyl-6-(1-methylethylidene)-;
1201 Cyclo-1M4M(M): cyclohexanol, 1-methyl-4-(1-methylethenyl)-; B[3.1.0]H4M1M(M):
1202 bicyclo[3.1.0]hexan-3-ol, 4-methylene-1-(1-methylethyl)-, (1. α.,3. α .,5. α.);
1203 Spiro[5.5]3TM(S): spiro[5.5]undec-2-ene, 3,7,7-trimethyl-11-methylene-, (-)-;
1204 CM-4(14),5-D(S): cis-muurola-4(14),5-diene; TriDol (M):
1205 tricyclo[4.2.1.1(2,5)]decan-9-ol.

1206

1207 Figure 8. Effects of the overexpression and down-regulation of *ADS* on artemisinin levels
1208 in leaves at positions 7 and 8. A, cassettes for the overexpression and RNAi of *ADS*; B,
1209 transgenic vs. WT plants; C, RT-PCR showing the overexpression of *ADS* in 4 transgenic
1210 lines; D, RT-PCR showing *ADS* downregulation in 3 transgenic lines; E, selective ion
1211 chromatographs of artemisinin (indicated by an arrow) showing a reduction in one *ADS*
1212 RNAi transgenic line and an increase in one overexpressing line compared to WT plants;
1213 F, significant increase and decrease of artemisinin in overexpressing and RNAi
1214 transgenic plants, respectively. Bars labeled with different lowercase letters (a, b, c, d,
1215 or e) in C, D and E indicate a significant difference (P values < 0.05), and those labeled

with the same lowercase letters indicate an insignificant difference. AR: artemisinin, WT: wild-type control plants (F6 progeny).

Figure 9. Effects of the overexpression and down-regulation of *ADS* on *A. annua* metabolite profiles. A, contents of artemisinic acid and arteannuin b (the content of arteannuin b is indicated as an artemisinic acid equivalent); B, contents of six other sesquiterpenoids (five, except for farnesene, is indicated as farnesene equivalent); C, a PCA plot generated using 84 volatile metabolites including eight metabolites in A and B.

Table Legend

Table 1. Monoterpenoids detected in tissues of self-pollinated F3 progeny plants by GC-MS-based profiling. Abbreviations: L7, 10, 11, 12, 13, 14, 15, 16 and 17, leaf from the 7^{th} node to 17^{th} node, respectively, on the primary stem; HB, head inflorescence buds; FOH, fully open head inflorescences; 50%OH, 50% of open head inflorescences; IB15 and IB16: inflorescence branches 15 and 16, respectively (*:α-pinene was detected when the column oven's temperature was initially set at 60ºC instead of 80ºC for 2 min).

Table 2. Sesquiterpenoids detected in the tissues of self-pollinated F3 progeny plants by GC-MS-based profiling. Abbreviations: L7, 10, 11, 12, 13, 14, 15, 16 and 17, leaf from the 7^{th} node to 17^{th} node, respectively, on the primary stem; HB, head inflorescence buds; FOH, fully open head inflorescences; 50%OH, 50% of open head inflorescences; IB15 and IB16, inflorescence branches 15 and 16, respectively.

Supplementary Materials and Methods

Supplementary Figure Legends:

S-Figure 1. Total ion chromatographs of metabolites from leaves, fully open head inflorescence (FOH) and head inflorescence buds (HB). These profiles were separated by the first column temperature program (TP-1). The twelve identified terpene molecules are labeled with their generic names. The peaks detected at the same retention time as squalene and beta-amyrin in the FOH and HB samples are an unknown metabolite and 1-docosene, respectively.

S-Figure 2. Structures of the 26 monoterpenes detected in *A. annua* tissues.

S-Figure 3. Structures of the 21 sesquiterpenoids detected in *A. annua* tissues.

S-Figure 4. A, accumulation profiles of germacrene D in 14 A. annua tissues; B, two glandular trichomes, one with an essential oil drop and the other after the release of an essential oil drop.

S-Figure 5. RT-PCR images showing the expression patterns of nineteen representative genes in seven selected tissues. An "*Aa*" prefix in each gene denotes *A. annua*. Four MEP pathway genes, *AaDXS1*: 5-deoxy-xylulose-5-phosphate synthase gene contig 1, *AaDXR1*: 5-deoxy-xylulose-5-phosphate reductase contig 1, *AaHDR1*: 4-hydroxy-3-methylbut-2-enyl diphosphate reductase contig 1, *AaIPPI1*: isopentenyl pyrophosphate isomerase contig 1. Two monoterpene synthase genes, *AaLS*: limonene

synthase, *AaBPS*: beta-pinene synthase. Two triterpene synthase genes, *AaBAS*: beta-amyrin synthase, *AaSQS*: squalene synthase. Four sesquiterpene synthase genes, *AaCPS*: beta-caryophyllene synthase, *AaGAS1*: germacrene A synthase contig 1, *AaECS1*: 8-epi-cedrol synthase contig 1, *AaBFS*: beta-farnesene synthase. One MVA pathway gene, *AaHMGR*: 3-hyroxy-3-methylglutary coenzyme reductase. *AaFDS2*: farnesyl diphosphate synthase gene contig 2. Five artemisinin pathway genes, *ADS*: amorpha-4,11-diene synthase, *CPR1*: cytochrome P450 reductase contig 1, *CYP71AV1*: cytochrome P450 monooxygenase, *DBR2*: artemisinic aldehyde Δ11(13) reductase, *ALDH1*: aldehyde dehydrogenase 1. Sample abbreviations, L7, L10 and L15: the 7th, 10th and 15th leaves, respectively, IB15: inflorescence branch at the 15th node, FOH: full open head inflorescence, HB: head inflorescence buds, 50%OH: 50% full open and 50% head inflorescence buds.

S-Figure 6. RT-PCR images showing the expression patterns of nineteen representative genes in seven selected tissues. An "*Aa*" prefix in each gene denotes *A. annua*. Four MEP pathway genes: *AaDXS1*, 5-deoxy-xylulose-5-phosphate synthase contig 1; *AaDXR1*, 5-deoxy-xylulose-5phosphate reductase contig 1; *AaHDR1*, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase contig 1; *AaIPPI1*, isopentenyl pyrophosphate isomerase contig 1. Two monoterpene synthase genes: *AaLS*, limonene synthase; *AaBPS*, beta-pinene synthase. Two triterpene synthase genes: *AaBAS*, beta-amyrin synthase; *AaSQS*, squalene synthase. Four sesquiterpene synthase genes: *AaCPS*, beta-caryophyllene synthase; *AaGAS1*, germacrene A synthase contig 1; *AaECS1*,*8*-epi-cedrol synthase contig 1; *AaBFS*, beta-farnesene synthase. One MVA pathway gene: *AaHMGR*, 3-hyroxy-3-methylglutary coenzyme reductase. *AaFDS2*, farnesyl diphosphate synthase contig 2. Five artemisinin pathway genes: *ADS*, amorpha-4,11-diene synthase; *CPR1*, cytochrome P450 reductase contig 1; *CYP71AV1*,

cytochrome P450 monooxygenase; *DBR2*, artemisinic aldehyde Δ11(13) reductase; *ALDH1*, aldehyde dehydrogenase 1. Sample abbreviations: L11-17, the 11^{th}-17^{th} leaves; IB16, inflorescence branch at the 16^{th} node.

S-Figure 7. A RT-PCR image showing the expression patterns of *CPR2* in 14 selected tissues.

S-Figure 8. Reads per kilobase per millions (RPKM) of 454 sequencing for nine genes in the 2-C-methyl-D-erythritol-4-phosphate (MEP) pathway toward isopentenyl diphosphate and dimethylallyl diphosphate in six different tissues. The y-axis values are expressed using a logarithmic scale. The numbers labeled on the top of each bar are true RPKM values. An "*Aa*" prefix in each gene denotes *A. annua*. A, six contigs of *AaDXS* (5-deoxy-D-xylulose-5-phosphate synthase) and two contigs of *AaDXR* (5-deoxy-D-xylulose-5-phosphate reductase); B, six genes: *AaMCT*, 2-C-methyl-D-erythritol-4-phosphate cytidyltransferase; *AaCMK*, cytidyl (4-diphospho)-2-C-methyl-D-erythritol kinase; *Aa*MDS, 2-C-methyl-D-erythritol-2,4-cyclodiphohate synthase (*AaMCS*); *AaHDS*, 1-hydroxyl-2-methyl-2-(E)-butenyl-diphosphate synthase; *AaHDR*, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase; *AaIPPI*, isopentenyl pyrophosphate isomerase. Sample abbreviations: L7 and L15, the leaves at the 7^{th} and 15^{th} nodes, respectively; IB15, inflorescence branch at the 15^{th} node; FOH, full open head inflorescence; HB, head inflorescence buds; 50%OH 50% full open and 50% head inflorescence buds.

S-Figure 9. Reads per kilobase per millions (RPKM) of 454 sequencing for seven genes in the mevalonic acid pathway (MVP) in six different tissues. The y-axis values are expressed using a logarithmic scale. The numbers labeled on the top of each bar are true RPKM values. An "*Aa*" prefix in each gene denotes *A. annua*. A, four contigs of *AaAACT* (acetoacetyl coenzyme A thiolase) and four contigs of *AaHMGS* (3-hydroxy-3-methylglutaryl coenzyme A synthase); B, five genes: *AaMVK*, mevalonate kinase (*MK*); *AaPMK*, 5-phosphatemevalonate kinase; *AaMVAPP*, 5-phosphomevalonate decarboxylase; *AaHMGR*, 3-hydroxy-3-methylglutaryl coenzyme A reductase. Sample abbreviations: L7 and L15, the leaves at the 7^{th} and 15^{th} nodes, respectively; IB15, inflorescence branch at the 15^{th} node; FOH, full open head inflorescence; HB, head inflorescence buds; 50%OH 50% full open and 50% head inflorescence buds.

Table 1. Monoterpenoids detected in tissues of self-pollinated F3 progeny plants by GC-MS-based profiling. Abbreviations: L7, 10, 11, 12, 13, 14, 15, 16 and 17, leaf from the 7th node to 17th node, respectively, on the primary stem; HB, head inflorescence buds; FOH, fully open head inflorescences; 50%OH, 50% of open head inflorescences; IB15 and IB16: inflorescence branches 15 and 16, respectively (*:α-pinene was detected when the column oven's temperature was initially set at 60ºC instead of 80ºC for 2 min).

Compound names	**Chemical Formulas**	**Tissues**	**Types and #**
1,3,6-Heptatriene, 2,5,6-trimethyl-	C10H16	L14, FOH	Acyclic (3)
1,6-Octadien-3-ol,3,7-dimethyl-; Linalool	C10H18O	HB, FOH, 50%OH	
Citronellol	C10H20O	L7-16, IB16, HB, FOH, 50%OH	
β-Phellandrene	C10H16	L14, IB15, HB, FOH, 50%OH	Monocyclic (10)
β-Terpinene	C10H16	IB15, IB16	
Terpinen-4-ol	C10H18O	HB, FOH	
Cyclohexene, 3-methyl-6-(1-methylethylidene)-	C10H16	L12, HB, FOH, 50%OH	
Carveol	C10H16O	HB, FOH	
Cyclohexanol, 1-methyl-4-(1-methylethenyl)-	C10H18O	FOH, 50%OH	
Cyclohexanone, 2-methyl-5-(1-methylethenyl)-	C10H16O	L16, FOH	
(1S-(1α,2α,4β))-1-isopropenyl-4-methyl-1,2-cyclohexanediol	C10H18O2	IB15, HB	
α-Campholenal	C10H16O	L13, L16, L17, IB15, IB16, HB, FOH, 50%OH	
Cyclopentane, 1,2-dimethyl-3-(1-methylethenyl)-	C10H18	L10, L11	

α-pinene*	C10H16	L12-17, IB15, IB16, HB, FOH, 50%OH	Bicyclic (12)
β-pinene	C10H16	L13-17, IB15, IB16, HB, FOH, 50%OH	
1,8-Cineole	C10H18O	All tissues	
L-camphor	C10H16O	L17, IB16	
trans-Pinocarveol	C10H16O	L12-17, IB15, IB16, HB, FOH, 50%OH	
Pinocarveol	C10H16O	L14, HB, FOH	
Pinocarvone	C10H14O	L13-17, IB15, IB16, HB, FOH, 50%OH	
(-)-Myrtenol	C10H16O	L14, IB15, IB16, HB, 50%OH	
7-Propylidene-bicyclo[4.1.0]heptane	C10H16	HB, FOH, 50%OH	
Bicyclo[3.1.0]hexan-3-ol, 4-methylene-1-(1-methylethyl)-, (1.α.,3.α.,5.α.)-;	C10H16O	IB15, HB, FOH, 50%OH	
1,7,7-Trimethylbicyclo[2.2.1]hept-5-en-2-ol	C10H16O	L14, HB, FOH	
1(2H)-Naphthalenone,octahydro-,trans-	C10H16O	L14, IB16	
Tricyclo[4.2.1.1(2,5)]decan-9-ol, stereoisomer	C10H16O	HB	Tricyclic (1)

1332

1333

1334

1335 Table 2. Sesquiterpenoids detected in the tissues of self-pollinated F3 progeny plants by
1336 GC-MS-based profiling. Abbreviations: L7, 10, 11, 12, 13, 14, 15, 16 and 17, leaf from the 7th
1337 node to 17th node, respectively, on the primary stem; HB, head inflorescence buds; FOH, fully
1338 open head inflorescences; 50%OH, 50% of open head inflorescences; IB15 and IB16,
1339 inflorescence branches 15 and 16, respectively.

Compound names	**Chemical Formulas**	**Tissues**	**Types and #**
β-Famesene	C15H24	L10-17, IB15, IB16, HB, FOH, 50%OH	Acyclic (1)
Germacrene B	C15H24	HB, 50%OH	Monocyclic (6)
Germacrene D	C15H24	All tissues	
Humulene	C15H24	HB, FOH, 50%OH	
Germacra-1(10),4,11(13)-trien-12-oic acid, 6-alpha-hydroxy-,γ-lactone,(E,E)-	C15H20O2	IB15, HB, FOH, 50%OH	
Cyclohexane, 1-ethenyl-1-methyl-2-(1-methylethenyl)-4-(1-methylethylidene)-	C15H24	L13, L14, L15	
Caryophyllene	C15H24	All tissues	Bicyclic (11)
Arteannuic acid	C15H22O2	L11-17, IB15, IB16, HB, FOH, 50%OH	
cis-muurola-3,5-diene	C15H24	IB15, HB	
cis-muurola-4(14),5-diene	C15H24	IB15, HB, FOH, 50%OH	
Azulene, 1,2,3,5,6,7,8,8a-octahydro-1,4-dimethyl-7-(1-methylethenyl)-, [1S-(1.α.,7.α., 8.a.β)]-	C15H24	HB,FOH	

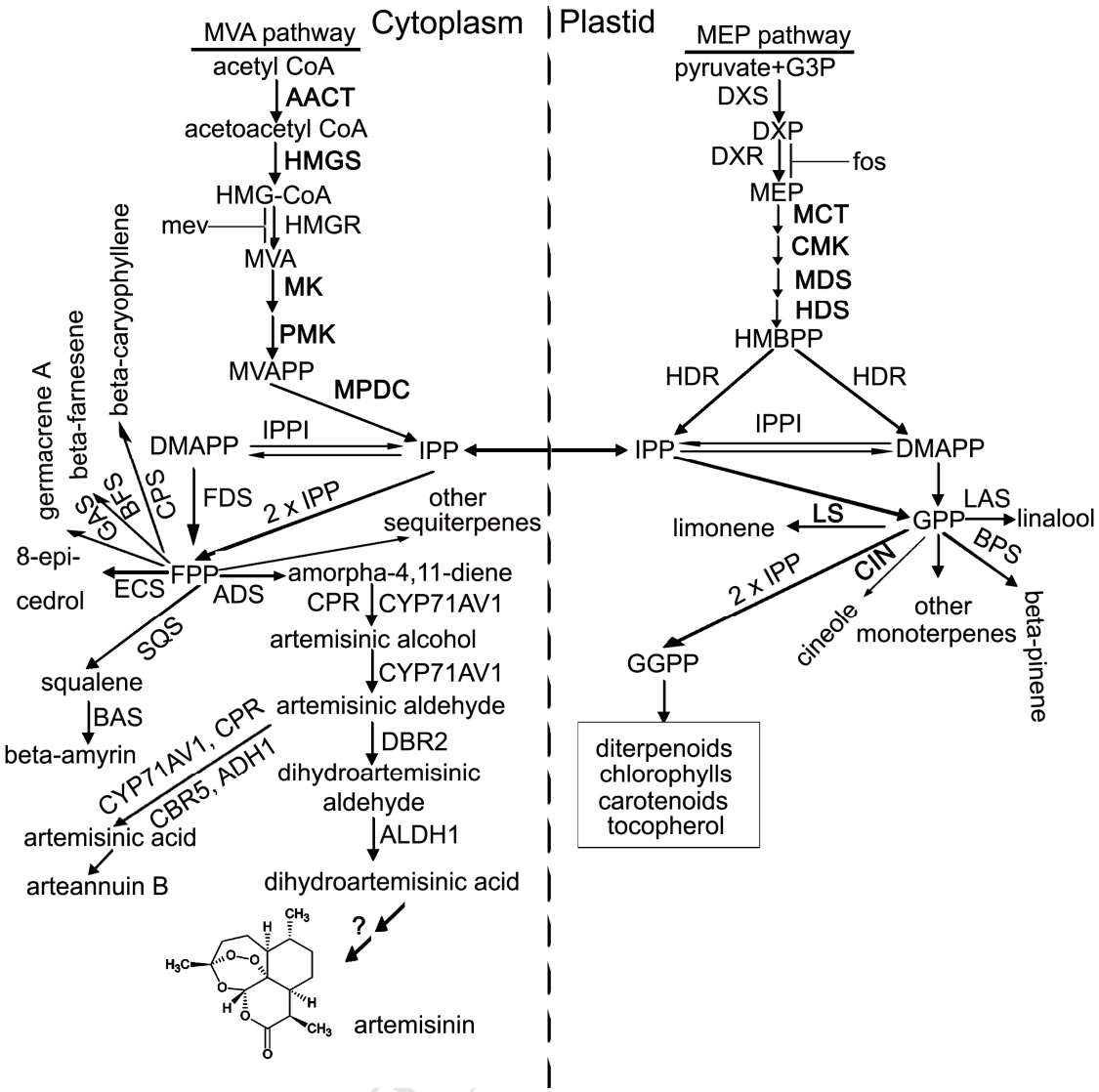
Bicyclo[5.3.0]decane, 2-methylene-5-(1-methylvinyl)-8-methyl-	C15H24	L13, IB16, HB, FOH	
Bicyclo[4.3.0]nonane, 7-methylene-2,4,4-trimethyl-2-vinyl-	C15H24	FOH	
Bicyclo[5.2.0]nonane, 4-methylene-2,8,8-trimethyl-2-vinyl-	C15H24	IB15, IB16, HB	
Bicyclo[7.2.0]undec-4-ene, 4,11,11-trimethyl-8-methylene-	C15H24	L10-17	
Spiro[4.5]dec-7-ene, 1,8-dimethyl-4-(1-methylethenyl)-, [1S-(1.α.,4.β.,5.α.)]-	C15H24	IB15, HB, FOH	
Spiro[5.5]undec-2-ene, 3,7,7-trimethyl-11-methylene-, (-)-	C15H24	IB15, IB16, HB, FOH, 50%OH	
1H-Cyclopenta[1,3]cyclopropa[1,2]benzene, octahydro-7-methyl-3-methylene-4-(1-methylethyl)-,[3aS-(3a.α.,3b.β.,4.β.,7.α.,7aS*)]-; β-Cubebene	C15H24	L14, IB15, FOH	Tricyclic (4)
Arteannuin b	C15H20O3	All tissues	
1H-3a,7-Methanoazulene, octahydro-1,4,9,9-tetramethyl-	C15H26	L13, IB15, IB16, HB, FOH, 50%OH	
α-copaene	C15H24	All tissues	

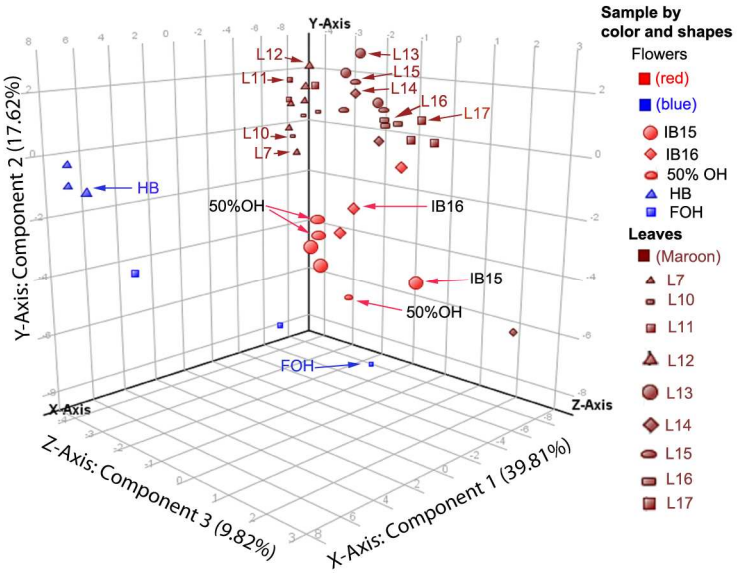
1340

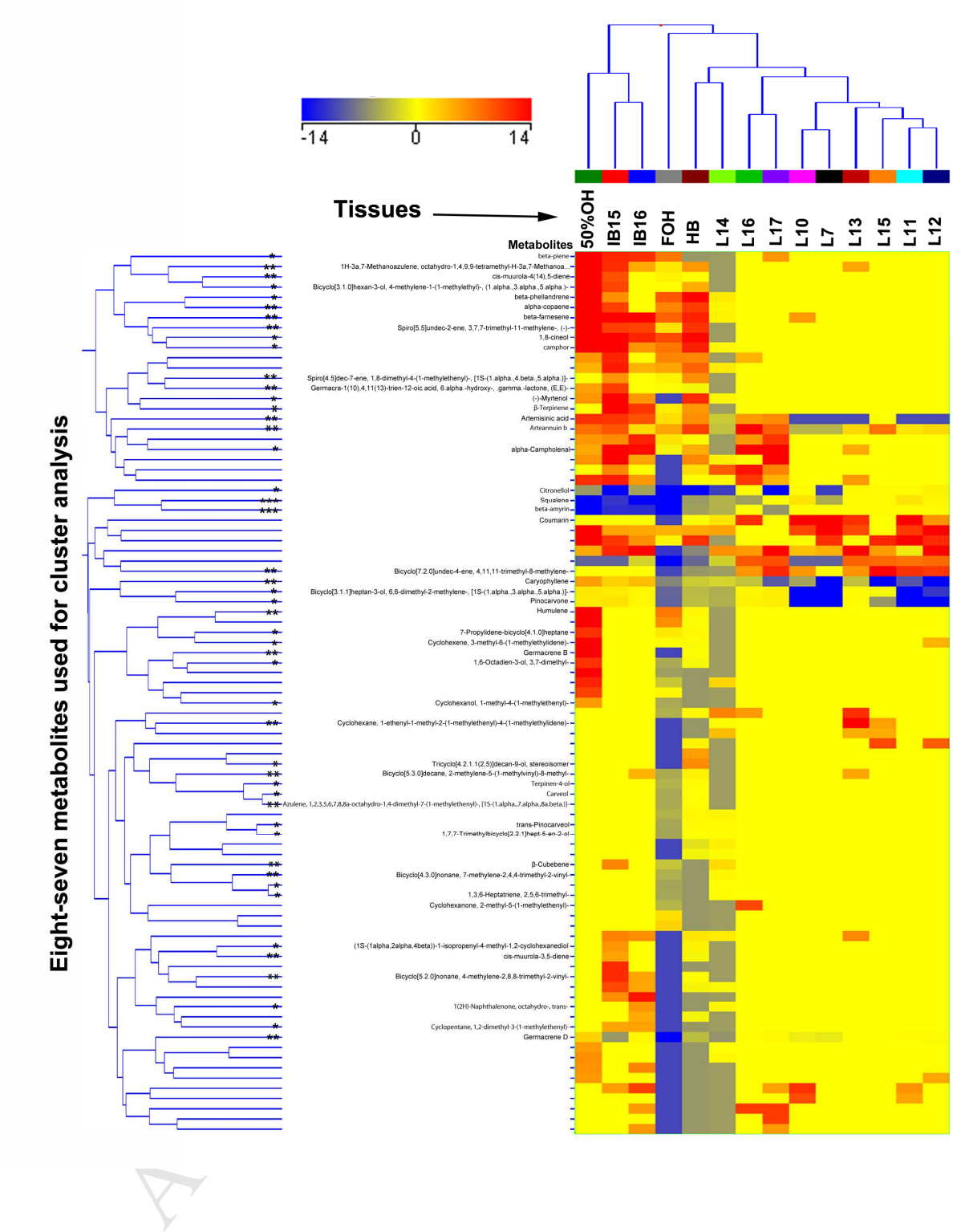
1341

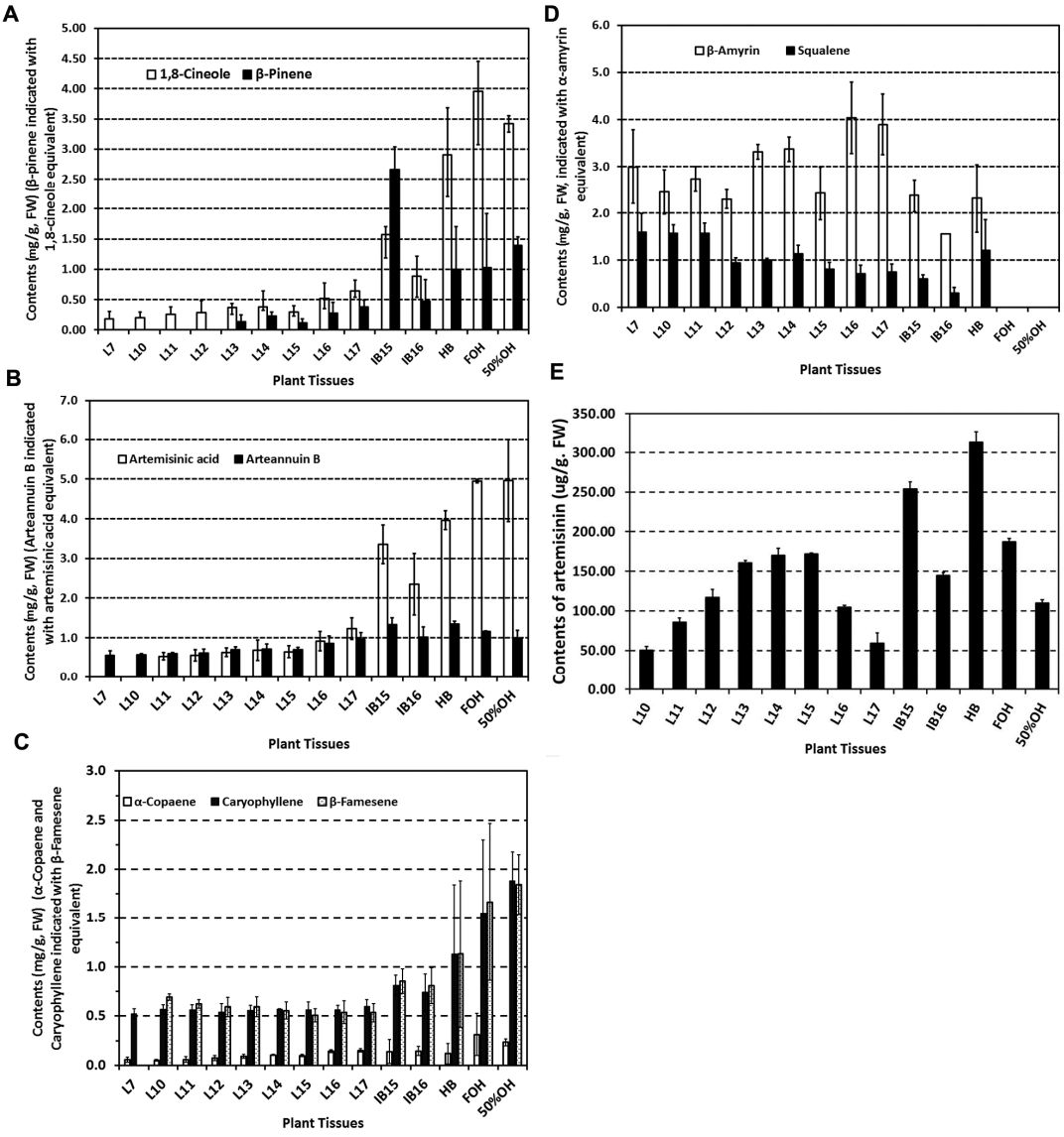
1342

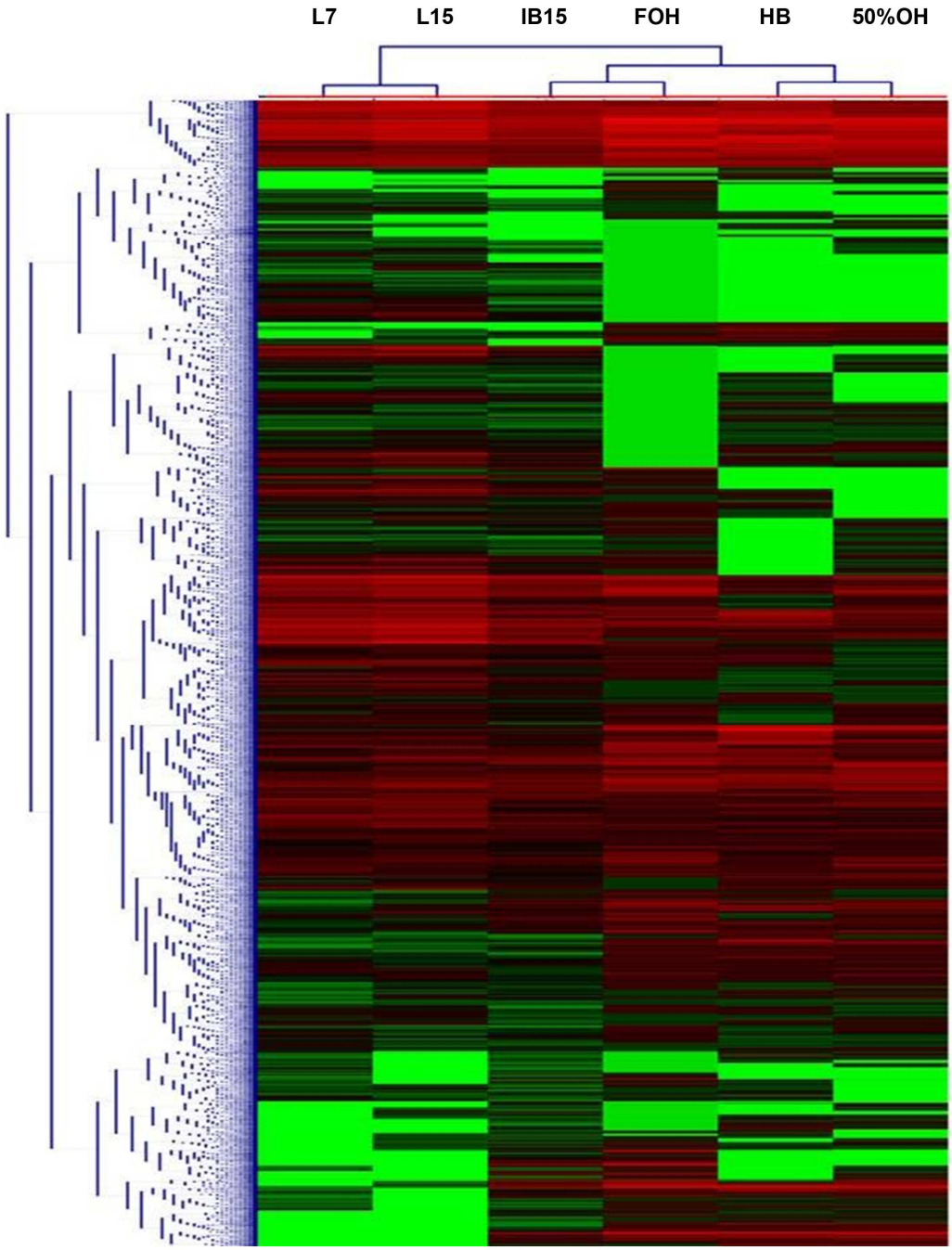
1343


MVA pathway
Cytoplasm
Plastid
MEP pathway
acetyl CoA
AACT
acetoacetyl CoA
HMGS
HMG-CoA
mev
HMGR
MVA
MK
PMK
MVAPP
MPDC
DMAPP
IPPI
IPP
germacrene A
beta-farnesene
beta-caryophyllene
GAS
BFS
CPS
FDS
2 x IPP
other sequiterpenes
8-epi-cedrol
ECS
FPP
ADS
amorpha-4,11-diene
CPR
CYP71AV1
artemisinic alcohol
CYP71AV1
artemisinic aldehyde
DBR2
dihydroartemisinic aldehyde
ALDH1
dihydroartemisinic acid
SQS
squalene
BAS
beta-amyrin
CYP71AV1, CPR
CBR5, ADH1
artemisinic acid
arteannuin B
?
artemisinin
pyruvate+G3P
DXS
DXP
DXR
fos
MEP
MCT
CMK
MDS
HDS
HMBPP
HDR
HDR
IPPI
IPP
DMAPP
LAS
LS
limonene
GPP
linalool
BPS
CIN
2 x IPP
cineole
other monoterpenes
beta-pinene
GGPP
diterpenoids
chlorophylls
carotenoids
tocopherol

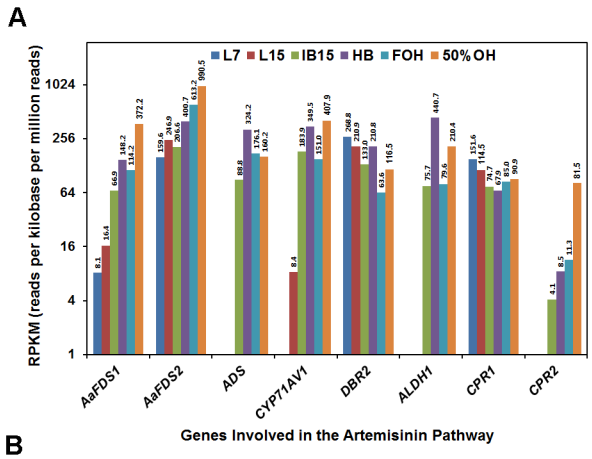


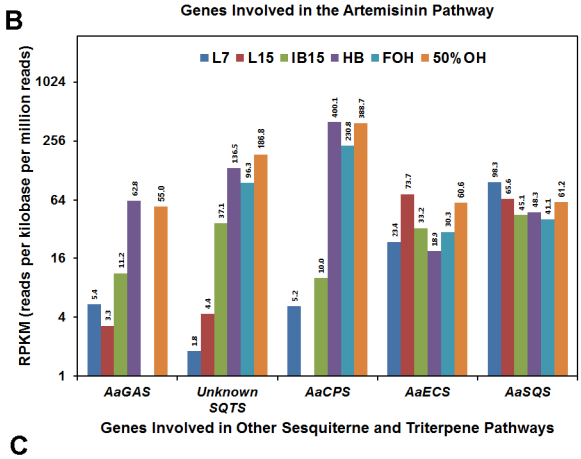

-14
0
14
Tissues
50%OH
IB15
IB16
FOH
HB
L14
L16
L17
L10
L7
L13
L15
L11
L12
Metabolites
Eight-seven metabolites used for cluster analysis
beta-piene
1H-3a,7-Methanoazulene, octahydro-1,4,9,9-tetramethyl-H-3a,7-Methanoa...
cis-muurola-4(14),5-diene
Bicyclo[3.1.0]hexan-3-ol, 4-methylene-1-(1-methylethyl)-, (1.alpha.,3.alpha.,5.alpha.)-
beta-phellandrene
alpha-copaene
beta-farnesene
Spiro[5.5]undec-2-ene, 3,7,7-trimethyl-11-methylene-, (-)-
1,8-cineol
camphor
Spiro[4.5]dec-7-ene, 1,8-dimethyl-4-(1-methylethenyl)-, [1S-(1.alpha.,4.beta.,5.alpha.)]-
Germacra-1(10),4,11(13)-trien-12-oic acid, 6.alpha.-hydroxy-, .gamma.-lactone, (E,E)-
(-)-Myrtenol
β-Terpinene
Artemisinic acid
Arteannuin b
alpha-Campholenal
Citronellol
Squalene
beta-amyrin
Coumarin
Bicyclo[7.2.0]undec-4-ene, 4,11,11-trimethyl-8-methylene-
Caryophyllene
Bicyclo[3.1.1]heptan-3-ol, 6,6-dimethyl-2-methylene-, [1S-(1.alpha.,3.alpha.,5.alpha.)]-
Pinocarvone
Humulene
7-Propylidene-bicyclo[4.1.0]heptane
Cyclohexene, 3-methyl-6-(1-methylethylidene)-
Germacrene B
1,6-Octadien-3-ol, 3,7-dimethyl-
Cyclohexanol, 1-methyl-4-(1-methylethenyl)-
Cyclohexane, 1-ethenyl-1-methyl-2-(1-methylethenyl)-4-(1-methylethylidene)-
Tricyclo[4.2.1.1(2,5)]decan-9-ol, stereoisomer
Bicyclo[5.3.0]decane, 2-methylene-5-(1-methylvinyl)-8-methyl-
Terpinen-4-ol
Carveol
Azulene, 1,2,3,5,6,7,8,8a-octahydro-1,4-dimethyl-7-(1-methylethenyl)-, [1S-(1.alpha.,7.alpha.,8a.beta.)]-
trans-Pinocarveol
1,7,7-Trimethylbicyclo[2.2.1]hept-5-en-2-ol
β-Cubebene
Bicyclo[4.3.0]nonane, 7-methylene-2,4,4-trimethyl-2-vinyl-
1,3,6-Heptatriene, 2,5,6-trimethyl-
Cyclohexanone, 2-methyl-5-(1-methylethenyl)-
(1S-(1alpha,2alpha,4beta))-1-isopropenyl-4-methyl-1,2-cyclohexanediol
cis-muurola-3,5-diene
Bicyclo[5.2.0]nonane, 4-methylene-2,8,8-trimethyl-2-vinyl-
1(2H)-Naphthalenone, octahydro-, trans-
Cyclopentane, 1,2-dimethyl-3-(1-methylethenyl)-
Germacrene D

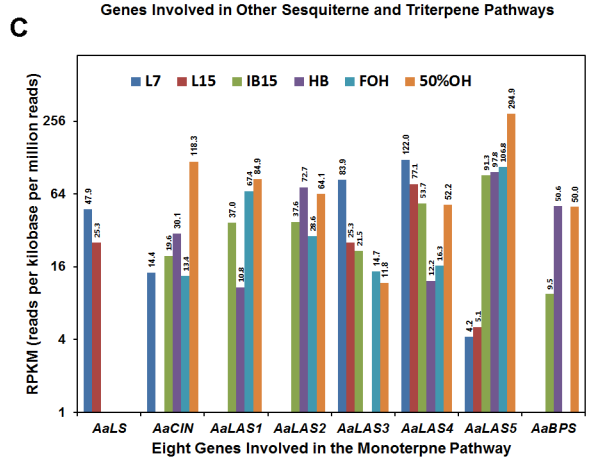

A
1,8-Cineole
β-Pinene
Contents (mg/g. FW) (β-pinene indicated with 1,8-cineole equivalent)
Plant Tissues
B
Artemisinic acid
Arteannuin B
Contents (mg/g. FW) (Arteannuin B indicated with artemisinic acid equivalent)
Plant Tissues
C
α-Copaene
Caryophyllene
β-Famesene
Contents (mg/g. FW) (α-Copaene and Caryophyllene indicated with β-Famesene equivalent)
Plant Tissues
D
β-Amyrin
Squalene
Contents (mg/g. FW, indicated with α-amyrin equivalent)
Plant Tissues
E
Contents of artemisinin (ug/g. FW)
Plant Tissues
L7
L10
L11
L12
L13
L14
L15
L16
L17
IB15
IB16
HB
FOH
50%OH

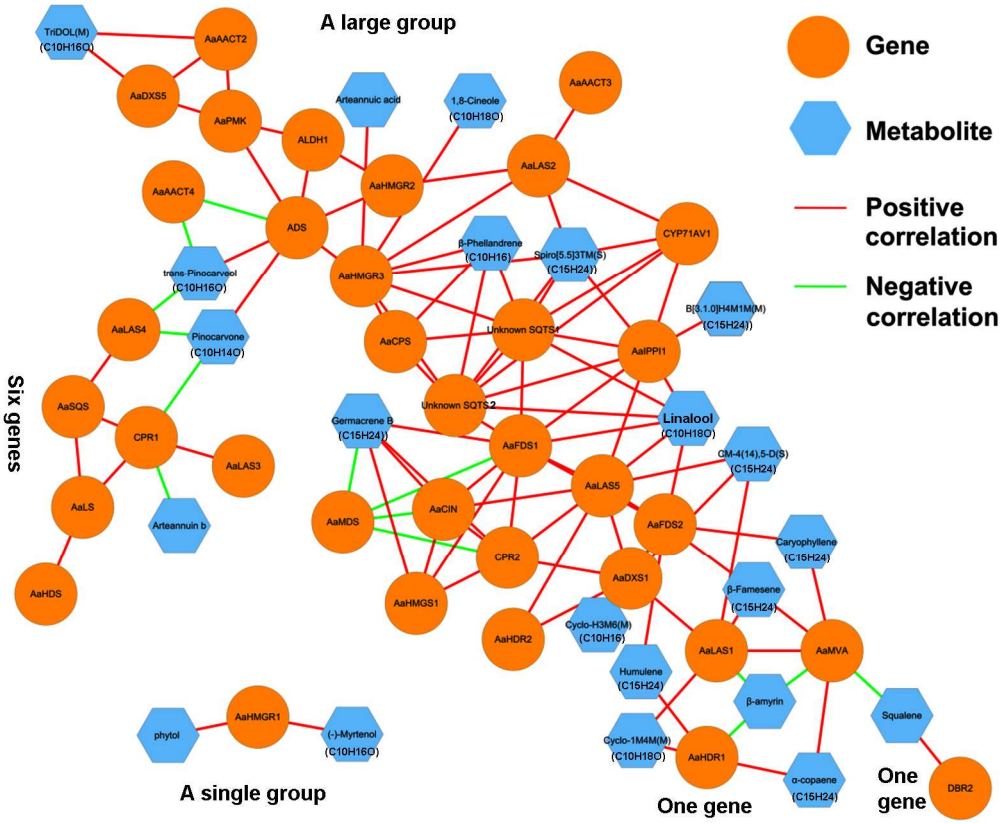


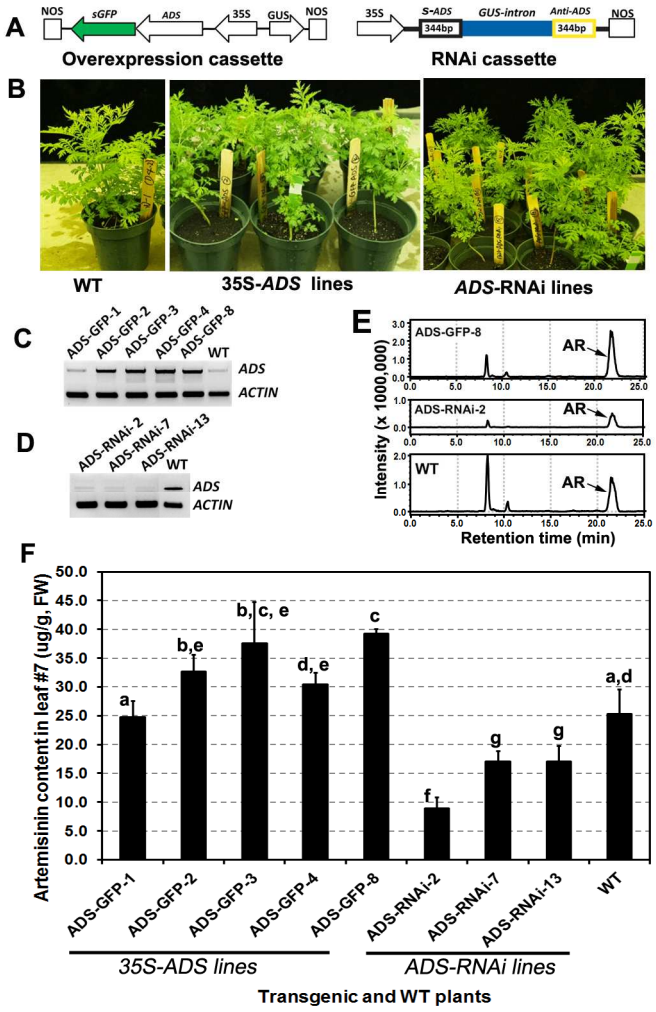
HCL Euclidean Heat Map

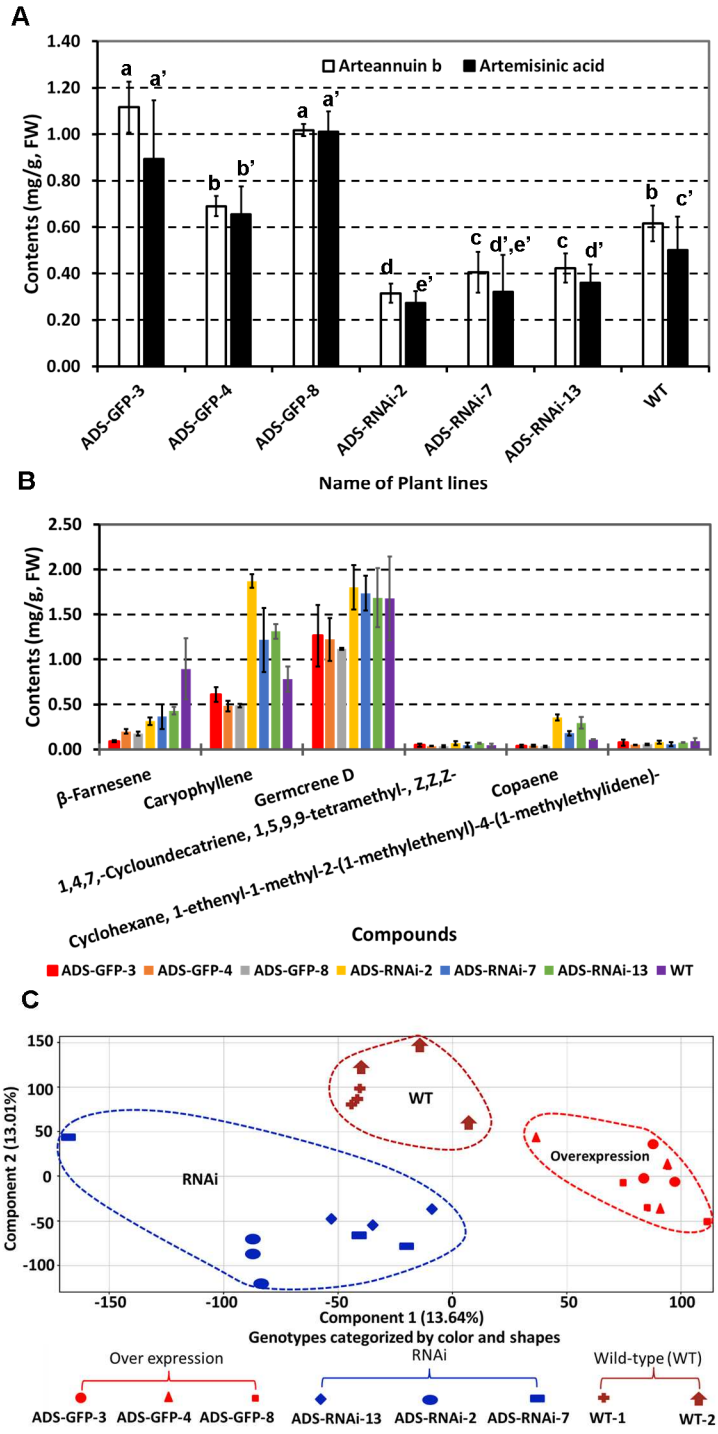

A
L7 L15 IB15 HB FOH 50%OH
RPKM (reads per kilobase per million reads)
1024
256
64
16
4
1
AaFDS1
AaFDS2
ADS
CYP71AV1
DBR2
ALDH1
CPR1
CPR2
Genes Involved in the Artemisinin Pathway


B
L7 L15 IB15 HB FOH 50%OH
RPKM (reads per kilobase per million reads)
1024
256
64
16
4
1
AaGAS
Unknown SQTS
AaCPS
AaECS
AaSQS
Genes Involved in Other Sesquiterne and Triterpene Pathways


C
L7 L15 IB15 HB FOH 50%OH
RPKM (reads per kilobase per million reads)
256
64
16
4
1
AaLS
AaCIN
AaLAS1
AaLAS2
AaLAS3
AaLAS4
AaLAS5
AaBPS
Eight Genes Involved in the Monoterpne Pathway


A large group
Gene
Metabolite
Positive correlation
Negative correlation
Six genes
A single group
One gene
One gene
TriDOL(M) (C10H16O)
AaAACT2
AaDXS5
AaPMK
ALDH1
Arteannuic acid
1,8-Cineole (C10H18O)
AaAACT3
AaLAS2
AaHMGR2
AaAACT4
ADS
β-Phellandrene (C10H16)
Spiro[5.5]3TM(5) (C15H24)
CYP71AV1
trans-Pinocarveol (C10H16O)
AaHMGR3
B[3.1.0]H4M1M(M) (C15H24)
AaLAS4
Pinocarvone (C10H14O)
AaCPS
Unknown SQTS1
AaIPPI1
AaSQS
Unknown SQTS2
Linalool (C10H18O)
Germacrene B (C15H24)
CPR1
AaFDS1
GM-4(14),5-D(S) (C15H24)
AaLAS3
AaLAS5
AaLS
AaCIN
AaFDS2
Caryophyllene (C15H24)
Arteannuin b
AaMDS
CPR2
AaDXS1
β-Famesene (C15H24)
AaHDS
AaHMGS1
Cyclo-H3M6(M) (C10H16)
AaHDR2
AaLAS1
AaMVA
Humulene (C15H24)
β-amyrin
Squalene
phytol
AaHMGR1
(-)-Myrtenol (C10H16O)
Cyclo-1M4M(M) (C10H18O)
AaHDR1
α-copaene (C15H24)
DBR2


A
NOS
sGFP
ADS
35S
GUS
NOS
Overexpression cassette
35S
S-ADS
344bp
GUS-intron
Anti-ADS
344bp
NOS
RNAi cassette
B
WT
35S-ADS lines
ADS-RNAi lines
C
ADS-GFP-1
ADS-GFP-2
ADS-GFP-3
ADS-GFP-4
ADS-GFP-8
WT
ADS
ACTIN
D
ADS-RNAi-2
ADS-RNAi-7
ADS-RNAi-13
WT
ADS
ACTIN
E
ADS-GFP-8
AR
ADS-RNAi-2
AR
WT
AR
Intensity (x 1000,000)
Retention time (min)
F
Artemisinin content in leaf #7 (ug/g, FW)
a
b, e
b, c, e
d, e
c
f
g
g
a, d
ADS-GFP-1
ADS-GFP-2
ADS-GFP-3
ADS-GFP-4
ADS-GFP-8
ADS-RNAi-2
ADS-RNAi-7
ADS-RNAi-13
WT
35S-ADS lines
ADS-RNAi lines
Transgenic and WT plants


A
Arteannuin b
Artemisinic acid
Contents (mg/g, FW)
ADS-GFP-3
ADS-GFP-4
ADS-GFP-8
ADS-RNAi-2
ADS-RNAi-7
ADS-RNAi-13
WT
Name of Plant lines
B
Contents (mg/g, FW)
β-Farnesene
Caryophyllene
Germcrene D
1,4,7,-Cycloundecatriene, 1,5,9,9-tetramethyl-, Z,Z,Z-
Copaene
Cyclohexane, 1-ethenyl-1-methyl-2-(1-methylethenyl)-4-(1-methylethylidene)-
Compounds
ADS-GFP-3 ADS-GFP-4 ADS-GFP-8 ADS-RNAi-2 ADS-RNAi-7 ADS-RNAi-13 WT
C
Component 2 (13.01%)
WT
RNAi
Overexpression
Component 1 (13.64%)
Genotypes categorized by color and shapes
Over expression
RNAi
Wild-type (WT)
ADS-GFP-3 ADS-GFP-4 ADS-GFP-8
ADS-RNAi-13 ADS-RNAi-2 ADS-RNAi-7
WT-1 WT-2