Research Article

Received: 16 December 2009 Revised: 8 April 2010 Accepted: 8 April 2010 Published online in Wiley Online Library: 2 August 2010

(wileyonlinelibrary.com) DOI 10.1002/ps.2006

Characterization of *de novo* transcriptome for waterhemp (*Amaranthus tuberculatus*) using GS-FLX 454 pyrosequencing and its application for studies of herbicide target-site genes

Chance W Riggins,[a] Yanhui Peng,[b] C Neal Stewart Jr,[b] and Patrick J Tranel[a*]

Abstract

BACKGROUND: Waterhemp is a model for weed genomics research in part because it possesses many interesting biological characteristics, rapidly evolves resistance to herbicides and has a solid foundation of previous genetics work. To develop further the genomics resources for waterhemp, the transcriptome was sequenced using Roche GS-FLX 454 pyrosequencing technology.

RESULTS: Pyrosequencing produced 483 225 raw reads, which, after quality control and assembly, yielded 44 469 unigenes (contigs + singletons). A total of 49% of these unigenes displayed highly significant similarities to *Arabidopsis* proteins and were subsequently grouped into gene ontology categories. Blast searches against public and custom databases helped in identifying and obtaining preliminary sequence data for all of the major target-site genes for which waterhemp has documented resistance. Moreover, sequence data for two other herbicide targets [4-hydroxyphenylpyruvate dioxygenase (HPPD) and glutamine synthetase], where resistance has not yet been reported in any plant, were also investigated in waterhemp and six related weedy *Amaranthus* species.

CONCLUSION: These results demonstrate the enormous value of 454 sequencing for gene discovery and polymorphism detection in a major weed species and its relatives. Furthermore, the merging of the 454 transcriptome data with results from a previous whole genome 454 sequencing experiment has made it possible to establish a valuable genomic resource for weed science research.

Supporting information may be found in the online version of this article.

Keywords: waterhemp; *Amaranthus tuberculatus*; transcriptome; GS-FLX 454 pyrosequencing; herbicide resistance; target-site genes; 4-hydroxyphenylpyruvate dioxygenase (HPPD)

1 INTRODUCTION

The genus *Amaranthus* L. consists of approximately 70 herbaceous species that are distributed primarily throughout the warm temperate and tropical regions of the world.[1] Most *Amaranthus* species are monoecious annuals, but there is a small group of nine dioecious species, all of which are native to North America.[1,2] Some amaranths have a long documented history of human use and possess economic importance as cultivated pseudocereals, vegetable crops or as ornamentals. In more recent times, several other members of the genus have acquired an economic impact of their own, not as useful plants, but rather as aggressive weeds that decrease crop yields and quality in many agricultural areas of the world.[3,4] Examples of such troublesome weeds include redroot pigweed (*A. retroflexus* L.), smooth pigweed (*A. hybridus* L.), Palmer amaranth (*A. palmeri* S Watson) and waterhemp [*A. tuberculatus* (Moq.) Sauer]. All of these species possess fast and competitive growth rates, high fecundity, discontinuous dormancy, long distance seed dispersal, the ability to form interspecific hybrids and resistance to multiple herbicides.[4–6] The ability to rapidly evolve resistance is one of the reasons why *Amaranthus* species are the most studied weeds, as evidenced by high citation counts in the current weed science literature,[7] and why it is imperative to develop new genomic tools for basic and applied studies in all areas of weed science.

The availability of genomic resources for weed research is small compared with those currently existing for many crops and model plants such as *Arabidopsis*, but new molecular and genomic techniques are beginning to be applied to studies of

* Correspondence to: Patrick J Tranel, Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. E-mail: tranel@illinois.edu

a Department of Crop Sciences, University of Illinois, Urbana, IL, USA

b Department of Plant Sciences, University of Tennessee, Knoxville, TN, USA

weed biology, ecology and herbicide resistance.[7,8] One of the new genomic techniques being utilized by weed scientists is GS-FLX 454 pyrosequencing. This technology allows the rapid generation of sequence data, with considerable savings of costs and time over traditional Sanger-based DNA sequencing methods.[9] With the application of 454 pyrosequencing, hundreds of thousands of high-quality sequence reads can be produced *de novo* from whole genome or transcriptome templates, thus enabling immediate inroads to genetic studies of organisms for which little or no sequence data exist.

As one of the most troublesome weeds throughout the corn and soybean belt of the midwestern United States, waterhemp has received much attention in the last two decades.[10] The species is dioecious and can readily hybridize with other dioecious and monoecious *Amaranthus* species,[11] which raises concern regarding the acquisition or transfer of herbicide resistance and other weedy traits. This scenario is all the more significant when considering that biotypes of waterhemp have evolved resistance to four herbicide groups: those that target acetolactate synthase (ALS; e.g. sulfonylureas), photosystem II (e.g. triazines), protoporphyrinogen oxidase (PPO; e.g. diphenylethers), and EPSP synthase (e.g. glyphosate).[12] Given the increasing threat that waterhemp poses to farmers across North America, it is not surprising that the weed science community has promoted this species as a model candidate for weed genomics research.[7]

A preliminary whole genome pyrosequencing experiment of waterhemp has already produced beneficial results that include identification of genes involved in herbicide resistance and the identification of new molecular markers for studies in population genetics and phylogenetics.[13] This initial run yielded close to 43 Mbp of sequence data and one or more hits to nine of 11 common herbicide target-site genes, which represents a respectable return for a plant with a moderately sized haploid genome of 675 Mbp.[14] Hits to additional candidate genes involved in non-target herbicide resistance, as well as other biological traits of interest to weed scientists, clearly demonstrate the value of 454 sequencing technology for gene discovery. Information from this resource can also be used proactively to gain insights into molecular targets of herbicides before initial reports of resistance are made. Having sequences of these genes beforehand will provide investigators with the necessary tools to study the mechanisms of resistance, as well as monitor its spread once it occurs. For instance, 4-hydroxyphenylpyruvate dioxygenase (HPPD) and glutamine synthetase inhibitors are two classes of herbicides that are emerging as effective alternatives in cases where plants have developed resistance to other herbicides. Because herbicides that inhibit HPPD and glutamine synthetase are not widely applied, no documented cases of naturally evolved resistance have yet been reported in any plant species.[12] However, history shows that resistance to these herbicides will likely occur as their use increases in the future.

In continuation of their efforts to answer the question as to what makes waterhemp such a successful weed, the authors sequenced the transcriptome of plants derived from a population in Adams County, Illinois, using 454 pyrosequencing technology. The main goal was to obtain a deeper coverage of coding sequences in the waterhemp genome that could facilitate a variety of functional and comparative studies of weediness traits. Ancillary goals included the search for herbicide resistance genes underrepresented or absent from the previous whole genome pyrosequencing run. Here, a global characterization of the waterhemp transcriptome as derived from 454 pyrosequencing is presented, and one of the potential applications of this resource in herbicide resistance research is highlighted by identifying and sequencing several important target-site genes in weedy amaranths.

2 MATERIALS AND METHODS

2.1 Plant materials for transcriptome sequencing

As the strategy for transcriptome sequencing was to maximize coverage, source material was pooled from different individuals, sexes, tissues, life stages and plants exposed separately to one of four stress factors (three herbicide treatments and cold stress). Seeds of plants from Adams Co., Illinois, designated as the ACR biotype, served as the source material for transcriptome sequencing. To obtain seedling tissue, seeds were sterilized in 15% bleach for 15 min and then triple rinsed with autoclaved ddH_2O. Afterwards, seeds were suspended in 0.15% (w/v) sterilized agarose and stored at 4 °C for 1 month to break dormancy. Twenty Petri plates were poured (25 mL $plate^{-1}$) with an artificial medium consisting of 10 g L^{-1} sucrose, 1× MS salts (Murashige and Skoog Basal Salt Mixture 4.3 g L^{-1}; Sigma-Aldrich, St Louis, MO) and 10 g L^{-1} Bacto agar. Twenty test tubes were filled with 3 mL of top agar (8 g L^{-1} Bacto agar, 10 g L^{-1} sucrose, 1× MS salts), and ∼20 ACR seeds were taken from the 0.15% agarose and placed in each of the top agar tubes. The top agar was poured over all 20 plates, which were then sealed with parafilm and placed in a growth chamber at 23 °C with a 12 h day length for ∼2 weeks. After 2 weeks the seedlings were carefully removed from the plates, immediately frozen in liquid nitrogen and stored at −80 °C.

For plants grown for flowering tissue and stress treatments, seeds were planted in a commercial potting mix and grown in a greenhouse for 10 days, after which time they were transplanted into flats with 48-well inserts (one seedling per well) containing a 3 : 1 : 1 : 1 mixture of commercial potting mix : soil : peat : sand. The greenhouse was maintained at 28/25 °C day/night with a 16 h photoperiod provided by supplementary lighting from mercury halide and sodium vapor lamps. After 1 week, seedlings were transplanted to 11.4 cm square pots also containing the aforementioned potting mix. The plants were then fertilized as needed with 13-13-13 Osmocote (Scotts Co., Marysville, OH). A week after transplanting to pots, the plants (∼8–10 cm in height) were randomly assorted into groups for either herbicide treatments, storage at 4 °C or obtaining flowering tissues.

Two plants each were exposed to one of three herbicide treatments: lactofen (Cobra®; Valent USA Corp., Walnut Creek, CA) at 110 g AI ha^{-1}; glyphosate (Roundup®; Monsanto Co., St Louis, MO) at 850 g AI ha^{-1}; atrazine (AAtrex®; Syngenta Crop Protection, Greensboro, NC) at 500 g AI ha^{-1}. At 24 h after spraying, the top 4 cm of each plant was harvested and placed in liquid nitrogen, and then stored at −80 °C. Two additional plants were stored in a 4 °C refrigerator with ambient light exposure for 1 week to simulate cold stress. Following this time, the top 4 cm of each plant was harvested, placed in liquid nitrogen and then stored at −80 °C. Tissues from the four different stress treatments were pooled prior to RNA isolation.

Mature flowering material also was obtained from greenhouse-grown plants. Plants were fertilized as necessary with Osmocote slow-release fertilizer and grown under the conditions described above. At flowering, two males were harvested (flower spikes about 4 cm long, multiple spikes per plant), and two females were harvested with young flowers (also about 4 cm from each of multiple flower spikes). Two weeks later, two females that had been exposed to pollen for about 2 weeks were harvested. For these

Table 1. Source information for *Amaranthus* species included in this study. Notations in parentheses correspond to accession labels used by Wassom and Tranel[15]

Taxon	Source information
A. albus L. Tumble pigweed	MH36, Whitman Co., WA (1a, b) MH38, Stoneville, MS (5a, b) PT70, Champaign, IL (4a, b)
A. hybridus L. Smooth pigweed	MH154, Hillsborough, NC (smooth 5a, b) PT12, Wayne Co., IL (smooth 12) MH165, Wooster, OH
A. palmeri S Watson Palmer amaranth	MH253, Dona Ana, NM MH254, Brazos, TX (palmer 8a, b) MH247, Riverside, CA
A. powellii S Watson Powell amaranth	MH234, Moses Lake, WA (powell 8a, b) MH242, Harrow, Ont., Canada (powell 10) MH237, Seneca, NY (powell 12a, b)
A. retroflexus L. Redroot pigweed	MH84, Dona Ana, NM (redroot 2) PT25, Christian Co., IL (redroot 3a, b) PT67, Champaign Co., IL (redroot 5a, b)
A. spinosus L. Spiny amaranth	MH267, Guanica, Puerto Rico (spiny 4a, b) MH203, Brazos, TX (spiny 2a, b) MH205, Tensar Parish, LSU
A. tuberculatus (Moq.) Sauer Waterhemp	MH2, Fayette Co., IL (waterhemp 2) MH320, Fulton Co., OH (waterhemp 4) PT43, Stark Co., IL (waterhemp 15a, b) ILFS1, Macon Co., IL

females, flowers containing both mature and immature seeds were collected. All flowers were immediately placed in liquid nitrogen and then stored at −80 °C. All flowering tissues were pooled prior to RNA isolation.

2.2 Source material for comparative analyses of herbicide genes

The selection of *Amaranthus* species and populations for analyses of resistance genes was largely guided by Wassom and Tranel.[15] In cases where the same populations of species cited in the previous study could not be sampled, seeds from additional populations from the authors' collections served as source material. In total, three populations of seven different species were represented (Table 1). Plants were grown from seeds under greenhouse conditions previously stated. Seeds from each population were initially sown in 800 mL containers with a 3 : 1 : 1 : 1 mixture of commercial potting mix : soil : peat : sand. When seedlings exhibited true leaves, the plants were thinned and transplanted into new containers of the same size to ensure ample material for DNA extraction and subsequent analyses. Leaf material for DNA extraction was harvested from mature plants (to verify identification) and flash frozen in liquid nitrogen prior to extraction using the modified CTAB method.[16]

2.3 454 pyrosequencing, data processing and annotation

Total RNA was isolated from each of the three sets of tissues (seedlings, stressed plants, reproductive growth) using PureLink Plant RNA Reagent (Invitrogen Corp., Carlsbad, CA). Poly A$^+$ mRNA then was isolated from a pool of equal quantities of the three total RNA preparations and used for cDNA synthesis. The cDNAs were then normalized and prepared for 454 pyrosequencing. Steps from mRNA isolation through to pyrosequencing were performed by the WM Keck Center for Comparative and Functional Genomics at the University of Illinois as described previously.[17] A preliminary titration run was followed by two bulk runs. The first bulk run was dedicated to waterhemp cDNAs, whereas in the second run only half the plate was allocated to waterhemp cDNAs.

Initial quality control of the waterhemp transcriptome sequencing was performed at the machine level before base calling. The raw 454 sequences and their quality scores were trimmed using Lucy[18] under default settings to remove the low-quality sequences, end regions with a significant level of ambiguous sequence and short reads ($<$50 bp). In addition, poly-A/T tails, SMART™ adaptors for cDNA synthesis, primers and contaminating vectors were removed from raw 454 sequences with EGassembler (http://egassembler.hgc.jp/) using default settings based on similarities to the NCBI's vector library. Finally, 442 772 high-quality clean sequences were assembled using CAP3[19] and EGassembler[20] with default settings to generate unique sequences.

Unigenes were used as queries to search three protein databases: all *Arabidopsis* proteins from GenBank (130 814 protein sequences; ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/), UniProtKB/Swiss-Prot annotated protein database (353 658 protein sequences; http://www.uniprot.org/downloads) and all green plant proteins from GenBank (677 422 protein sequences; ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/). This procedure was carried out using the NCBI Standalone BlastX program (ftp://ftp.ncbi.nih.gov/blast/). The best five protein hits for each query were parsed through a custom pipeline to create annotated tables, which included available information such as taxonomy, keywords, protein function, accession number and/or gene ontology (GO) terms.

To facilitate faster searches and generate more streamlined output of the waterhemp transcriptome, additional custom databases were created with the Blastall program (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.20/). One file consisted of all Amaranthaceae-specific sequences downloaded from GenBank plus TIGR [*Beta vulgaris* L. and *Suaeda* expressed sequence tags (ESTs)] and contained a total of 79 263 cDNA sequences. A second file was generated from EST libraries from *Helianthus annuus* L. lines obtained from the Compositae Genome Project (http://compgenomics.ucdavis.edu/). This file contained 31 605 unigene sequences. These databases were used for further Blast searches of the full transcriptome dataset to compare results of prior searches against the *Arabidopsis* and other protein databases. Unassigned contiguous sequences (contigs) and singletons from these initial searches were parsed out and subjected to separate BlastX searches against custom databases in an attempt to identify additional genes. Output files were filtered (*E*-values of 10^{-3} or

Table 2. Primers used for amplification and sequencing of herbicide resistance genes in *Amaranthus*

Gene	Primer name	Primer sequence (5′–3′)
HPPD	hppdF2	AAGTGGTCTTGGTGGGTTTG
	hppdF5	GCTGCTGATGTATTGAGTGAGG
	hppdF6	AGGAAGAGAAGTGGTCTTGGTG
	hppdF7	GGTTGGATCATGCTGTAGGG
	hppdR3	ACCACCCTTTTGGTACATCTTG
	hppdR4	AAGTTTCCCTTTCCAAATCCTC
	hppdR5	CGACGGCATAAACTCAAACC
	hppdR7	CACACTCCTTCATCTGCTCC
EPSPS	WH-EPSF1	GGACCACCCAGGGAATCATC
	WH-EPSF3	AGGTTCTTCATCCGAGGTGGTC
	WH-EPSF4	CTCTTGCAGTTGTTGCCTTGTATG
	WH-EPSR1	TCCTCAACTGTTGCCCCAAG
	WH-EPSR2	GACGGGAACATCAGCACAGG
	WH-EPSR4	ATCCGTTCGGTTTCCTTCACTC
Glutamine synthetase	GS1-F4	GTCACGACCAACACGAACTG
	GS1-R1	GTGAAGGGTGACTGGAATGG
	GS2-F3	TAATCGCCAGGAGAGGATAGTG
	GS2-F8	CATACTATTGTGGTGCTGGTGCTGAC
	GS2-R1	ACCCTGAGGACCAGGATAGG
	GS2-R6	TACCACGTATGGGTCCATGTTTGAAG
ALS	ALSf1	Foes *et al.*[22]
	ALSr1	Foes *et al.*[22]
	ALSf2	Foes *et al.*[22]
	ALSr2	Foes *et al.*[22]

less) to include only comparisons with relatively high expectation values.

2.4 Identification and sequencing of herbicide target-site genes

Five herbicide target-site genes were searched for in the transcriptome data: ALS, HPPD, EPSPS, GS and PPO. Contigs and singletons with relatively high-probability hits (% identity; *E*-value $< e^{-50}$) to target-site genes were parsed, translated in all possible frames and added to multiple sequence alignments in MEGA4[21] containing complete target-site gene sequences from other plant species. This procedure helped to verify the position and correct the translation frame of the waterhemp sequences.

Based on the multiple sequence alignments of the waterhemp nucleotide sequences with known target-site genes from other plants, primers for HPPD, EPSPS and GS were designed using Batch-Primer3 and subsequently checked for self-complementarity, hairpins and dimers using IDT OligoAnalyzer. Primers for amplifying regions A and B of the ALS gene were previously described by Foes *et al.*[22] All primers used in this study are listed in Table 2.

All PCR amplifications were performed using MJ Research thermal cyclers in 25 µL volumes with 1× buffer (GoTaq® Flexi Buffer; Promega Corp., Madison, WI), 2.5 mM $MgCl_2$, 200 µM dNTPs, 0.4 µM each primer, 1.0 µL total genomic DNA (10–50 ng) and 1.25 units of GoTaq® polymerase. A touchdown PCR protocol was used for amplifications of the HPPD, EPSPS and GS genes with the parameters: initial denaturation of 95 °C for 2 min, followed by 35 cycles with 1 min denaturation at 95 °C and 1 min extension at 72 °C, and a final extension of 72 °C for 2 min. The annealing temperature was 72 °C for the first cycle, and then decreased by 1 °C each subsequent cycle until it reached 60 °C where it was held constant for the remaining cycles. PCR for ALS featured an initial denaturation of 94 °C for 2 min, followed by 38 cycles of 94 °C for 30 s, 57 °C for 30 s and 72 °C for 45 s, and final extension of 72 °C for 4 min. PCR products were visualized on a 1% agarose gel containing 0.5 µg mL^{-1} ethidium bromide and cleaned using an E.Z.N.A.™ Cycle-Pure Kit (Omega Bio-Tek, Inc., Norcross, GA) following the manufacturer's instructions.

Purified PCR products were directly sequenced using an ABI Prism BigDye Terminator Kit v.3.1 (Applied Biosystems, Foster City, CA) and run on an ABI 3730XL capillary sequencer at the WM Keck Center for Comparative and Functional Genomics at the University of Illinois. Sequencing reactions were prepared in 13–16 µL volumes and contained 1.8 µL of ddH_2O, 5.2 µL of 12.5% (v/v) glycerol, 2.0 µL of 5× cheating buffer, 2.0 µL of 10 µM primer, 1.0 µL of BigDye Terminator v.3.1 and 1.0–4.0 µL of PCR product. Cycle sequencing conditions started at 96 °C for 1 min, followed by 30 cycles of 96 °C for 30 s, 50 °C for 15 s and 60 °C for 4 min, and final extension of 60 °C for 4 min.

Forward and reverse sequences were manually edited and assembled into contigs using the Alignment Explorer in MEGA4. Alignment of coding regions and exon–intron boundaries were determined by comparison with published cDNA and genomic sequences of *Arabidopsis* (http://www.arabidopsis.org) and other plant species available in GenBank. Single nucleotide polymorphisms were observed in HPPD and ALS sequences of several accessions and were coded with IUPAC codes.

2.5 Phylogenetic analyses

Phylogenetic relationships among sampled *Amaranthus* species were examined on the basis of the HPPD and ALS sequence data using distance methods implemented in MEGA4. Pairwise distances were computed for both datasets using nucleotide synonymous and non-synonymous substitution models available in MEGA4 to allow for comparisons among the resulting trees. Under the maximum composite likelihood model, sequences are compared nucleotide by nucleotide, and the sum of log-likelihoods for all pairwise distances in a distance matrix is maximized and then used in tree reconstruction. Alternatively, the Nei–Gojobori method was used to estimate synonymous and non-synonymous substitution parameters for each codon position. Neighbor-joining (NJ) trees were generated from distance matrices and bootstrapped with 1000 replicates for both datasets in MEGA4.

3 RESULTS AND DISCUSSION

3.1 Waterhemp transcriptome sequencing and assembly

Two sequencing runs plus the titration run yielded a total of 483 225 raw reads. The raw sequence files are available from the NCBI Sequence Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/as SRA012013. The average length of each read was 231 bp (Table 3), and the total data size was 111.8 Mb. Length frequency distribution of raw reads is provided in Fig. 1. Initial quality filtering of the 454 reads was performed at the machine level before base calling. These sequences were subsequently trimmed as described in Section 2. Approximately 92% of the sequences (442 772) were determined to be suitable for assembly and generated 44 469 unique sequences. A total of 22 035 contig sequences were assembled, ranging in length from 80 to 2893 bp and with an average sequence length of 434 bp (Fig. 2). An additional 22 434 high-quality sequences ranging in length from 80 to 526 bp

Table 3. Summary of numbers of reads and nucleotides by 454 sequencing runs

	Run A	Run B	Titration run	Total count
Number of raw reads	303 367	168 175	11 683	483 225
Mean length	231 bp	233 bp	223 bp	231 bp
Number of nucleotides	70 019 092	39 134 553	2 607 326	111 760 971

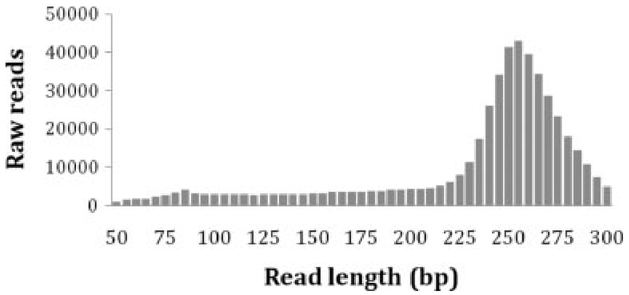

Figure 1. Length frequency distribution of waterhemp GS-FLX 454 sequencing raw reads.

(mean length 210 bp) remained as singletons (Table 4). These results are comparable with those obtained from other *de novo* plant transcriptome assemblies including *Artemisia annua* L.,[23] *Medicago*,[24] California poppy[25] and avocado.[25]

After parsing all contig and singleton hits to the *Arabidopsis* protein database, a total of 8727 (40%) contigs and 13 899 (62%) singletons remained without initial hits. These sequences were then used for BlastN and TBlastX searches (E-value $\leq 10^{-3}$) against custom Amaranthaceae and *Helianthus* databases. Of the 8727 contigs, 1584 received high-quality hits to Amaranthaceae-specific cDNAs. When searched against the *Helianthus* database, 705 hits were recorded, with 525 of these being unique hits not found in previous Blast searches. Of the 13 899 singletons without hits to *Arabidopsis*, 1783 were recorded with high-quality hits to the Amaranthaceae database. Although complete annotation of these new sequence hits was not performed, this demonstrates one type of methodology for pursuing more specific, taxon-guided searches of genes of interest.

3.2 Quality and performance of the 454 assembly

To test the quality and performance of the 454 assembly, contigs were aligned against themselves and the singletons with a NCBI BlastN program. A total of 5772 contigs (25.9%) had best Blast hits (bit score >45, E-value <0.0001) with >95% identity with other contigs and singletons, but in no case did these alignments extend over the entire length of either the Blast subjects or queries. These perfect-match alignments averaged 89 bp and 58 bp in length for contig versus contigs and contigs versus singleton hits respectively. Also, the average coverage of the match alignments was 17.9% and 9.3% of the length of the queried contigs in the cases of contigs versus contigs and contigs versus singletons respectively. Of the contigs, 3067 had BlastX hits (bit score >45) against the all green plant proteins database, and only 392 (1.8%) of those BlastN-paired contigs had the same best BlastX hits in the protein database (Table 5). In recognition that conserved motifs of different genes may exist widely in the genome, and that alternative splice variants of single genes are likely occurrences,[26] the assembly appropriately partitioned these gene regions to produce high identity but short coverage alignments into different contigs.

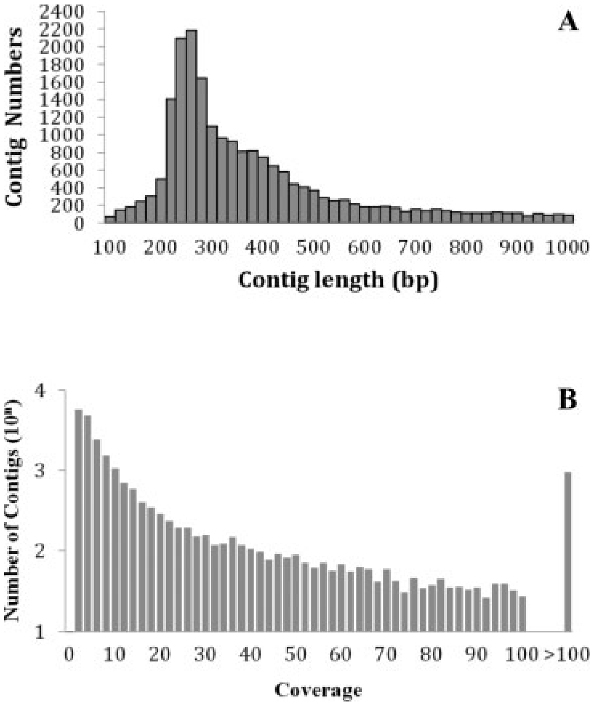

Figure 2. Characteristics of assembled waterhemp GS-FL 454 contigs: (A) length frequency distribution of assembled contigs; (B) average coverage frequency distribution of assembled contigs.

Table 4. Summary of 454 sequencing data trimming, assembling and annotation

	Number of sequences	Number of nucleotides
Raw reads	483 225 (100%)	111 760 971 (100%)
After quality control and trimming	442 772 (91.6%)	102 835 448 (92.0%)
Sequences assembled to contigs	420 338 (87.0%)	98 124 800 (87.8%)
Assembled contigs	22 035	9 561 765
Singletons	22 434 (4.6%)	4 710 648 (4.2%)
Annotated unique sequences	21 845	8 888 445

3.3 Functional annotation of waterhemp unigenes

The 44 469 unigenes were queried against three annotated protein databases (see Section 2.3) in order to assign putative functional roles. Gene descriptions and/or GO terms were assigned to 21 845 (49%) unigenes on the basis of the 'best hit' BlastX search (E-value <0.0001, bit score >45) in the UniPortKB protein database and in accordance with standards provided by the Gene Ontology Consortium (http://www.geneontology.org). The annotated sequences were classified into three general categories associated with cellular, molecular and biological functionalities (Fig. S1).

Table 5. Summary statistics for Blast of assembled GS-FLX 454 contigs against themselves and all singletons. All Blast results refer to hits with bit scores greater than or equal to 45, an *E*-value of <0.0001 and alignments with greater than or equal to 95%

Total assembled waterhemp 454 contigs	22 035
Number of contigs that had best Blast hits with other contigs and singletons	5772 (25.9%)
Average length of alignments as contigs versus contigs	89 bp
Average length of alignments as contigs versus singletons	58 bp
Average coverage of alignments to queried contig as contigs versus contigs	17.9%
Average coverage of alignments to queried contig as contigs versus singletons	9.3%
Number of BlastN-paired contigs that had BlastX hits against protein database	3067 (13.9%)
Number of BlastN-paired contigs that had the same best BlastX hits	392 (1.8%)

A diversity of GO subcategories is represented in the waterhemp unigene set (Fig. S1). This observation is not surprising, as the template for transcriptome sequencing consisted of pooled samples from different plant tissues and stress treatments. Cold stress, for example, is known to have a dramatic impact on the plant transcriptome.[27] Similarly, exposure to herbicides profoundly affects essential metabolic activities in plant cells. Given the inherent complexity of the pooled template, it was not unexpected that many of the unigenes could not be assigned to a particular functional role. However, there were still relatively large percentages of unigenes that did fall within defined categories.

Under the cellular component, approximately 22% of unigenes could be localized to plastids, with more than half associated with chloroplasts. Enzymatic activities involving transferases, hydrolases and kinases rank among the top categories under molecular function. These enzymes fulfill many essential functions in plant cells, and certain types of these enzymes have even been implicated in cold acclimation[27] and in herbicide detoxification.[28] A total of 12% of unigene annotations were grouped under stress/stimuli response, and 11% were associated with transport activity. This latter category includes ABC transporters, which are a diverse gene family involved in the active transport of ligands across membranes.[29] In fact, ABC transporters may serve as one mechanism that confers resistance to plants from glyphosate.[28]

As can be seen from the broad categories shown in Fig. S1, the single 454 transcriptome run provided numerous leads for identifying genes involved in a variety of primary metabolic roles, as well as possible roles related to cold stress and herbicide resistance. To evaluate further the utility of these data for understanding herbicide resistance in waterhemp, the authors focused specifically on the task of identifying and extracting sequence information for a select group of herbicide target-site genes.

3.4 Detection of herbicide target-site genes: comparison between whole genome and transcriptome approaches

The effectiveness of the 454 transcriptome sequencing for identifying genes involved in herbicide resistance in waterhemp can be measured by comparing the results obtained here with the results produced from a preliminary 454 experiment using total genomic DNA as the template.[13] Hits to nine herbicide target-site genes were recorded from the whole genome run, with acetyl-CoA carboxylase receiving the most number of hits (8). Conversely, transcriptome sequencing expanded the number of total hits to 11 target-site genes and contributed to a greater overall percentage of coverage (Table 6). Coverage in this case refers both to the probability of having a gene represented in the assembled sequence output and the amount of coding sequence present. Blast searches of transcriptome data against the whole genomic data did not result in any overlapping coding regions for the herbicide genes examined in this study. Interestingly, there were no hits to protoporphyrinogen IX oxidase II (PPX2; EC 1.3.3.4) in any Blast searches against the transcriptome. To confirm the apparent absence of hits to PPX2, published sequence data of this gene were used to perform additional BlastN and TBlastX searches against the transcriptome, but without any success.

As the percentage of coding regions in most genomes is relatively low compared with the abundance of non-coding sequences,[30,31] it is not surprising that fewer hits to the herbicide genes of interest were observed from the whole genome sequences. Of the genomic hits, a larger proportion of non-coding sequences (introns) was recovered for herbicide genes, as verified upon alignment with fully annotated reference sequences. Conversely, transcriptome reads are derived exclusively from the mRNA pool of expressed genes and include only coding sequences such as exons and 5′ and 3′ flanking untranslated (UTR) regions. Therefore, to maximize coverage and to increase the likelihood of identifying herbicide target genes, or any other gene of interest, sampling the transcriptome rather than the whole genome is more cost effective. In the following section, the results of experiments aimed at developing PCR-based assays for molecular studies of four herbicide target-site genes in *Amaranthus* are discussed.

3.5 Amplification and sequence characteristics of select target-site genes

3.5.1 *4-hydroxyphenylpyruvate dioxygenase (HPPD)*

In plants, the enzyme 4-hydroxyphenylpyruvate dioxygenase (HPPD; EC 1.13.11.27) is involved in the biosynthesis of phenylquinones and in the catabolism of the aromatic amino acids phenylalanine and tyrosine.[32] HPPD activity appears to be localized in the cytosol where it catalyses the formation of homogentisate (HGA), a key precursor of plastoquinones and tocopherols in higher-plant chloroplasts.[33,34] HPPD is a relatively new target for bleaching herbicides,[32] so called because inhibition of this enzyme disrupts the biosynthesis of carotenoids and results in bleaching (or loss of chlorophyll) of foliage.

Little sequence information has been published on plant HPPDs, but examination of the *Arabidopsis* and *Oryza* genomes suggests that HPPD is encoded by a single gene with two exons and one intron.[35] Observations based on multiple sequence alignment of 28 plant HPPDs (data not shown) support this simple gene structure. Exon–intron partitions in the cDNAs/ESTs were identified by comparing these sequences with genomic HPPD sequences of *Arabidopsis* (AT1G06570; http://www.arabidopsis.org) and *Oryza* (AP008208), *Sorghum* (Sb02g008850) and *Populus* (fgenesh4_pm.C_LG_II000277) downloaded from http://www.gramene.org. Length variations of the single intron among these four sequences range from 107 bp in *Arabidopsis* to 758 bp in *Oryza*. Putative HPPD sequences were also translated to identify active site residues and other conserved structural motifs,[33] which helped to confirm the nucleotide alignments.

Table 6. Comparison of the number of hits (contigs + singletons) to herbicide target-site genes from whole genome and transcriptome 454 sequencing of waterhemp

Herbicide target	Whole genome[a]	Transcriptome	Percentage coverage[b]
Acetolactate synthase	0	2	100
D1 protein (plastidic gene)	1	2	100
Tubulin	5	33	100
Protoporphyrinogen oxidase	4	8	90
Phytoene desaturase	3	1	80
Glutamine synthetase	2	7	80
1-Deoxy-D-xylylose-5-phosphate synthase	3	1	55
4-Hydroxyphenylpyruvate dioxygenase	0	2	45
Acetyl-CoA carboxylase	8	8	40
Dihydropteroate synthase	2	2	40
5-Enolpyruvylshikimate-3-phosphate synthase	2	3	40

[a] Number of hits taken from Lee *et al.*[13] Note that whole genome hits include non-coding as well as coding regions of the gene.
[b] Percentage coverage of coding sequence from combined genome and transcriptome 454 runs.

Two contigs, with a combined length of 760 bp, showed high-similarity Blast matches (>82% identity) to other plant HPPDs and were easily aligned by eye in the multiple sequence alignment. All active site residues, as well as the exon–intron boundaries, were represented by both waterhemp contigs. This observation was important for designing primers for genomic sequencing, especially without knowing about intron size of HPPD sequences of *Amaranthus*. Consequently, several forward and reverse primers were designed with the purpose of amplifying short exon-only and longer intron-spanning amplicons.

Results of PCR experiments using different forward and reverse primer combinations led to the selection of eight primers that were consistently effective in most species of *Amaranthus* examined (Table 1). Primer combinations hppdF2/hppdR3, hppdF5/hppdR3, hppdF6/hppdR3, hppdF6/hppdR4, hppdF7/hppdR5 and hppdF7/hppdR7 all produced amplicons. Two primer combinations, hppdF2/hppdR3 and hppdF6/hppdR3, were designed to include the intron, and both resulted in similar large amplicons of approximately 2000–2500 bp. Primers hppdF2/hppdR3 produced apparent single bands in all species except *A. spinosus* and two accessions of *A. albus*, which failed to amplify. Also, two accessions of *A. palmeri* appeared to be heterozygous, as they yielded two amplicons of different size. Primers hppdF6/hppdR3 yielded similar band patterns to those produced with hppdF2/hppdR3, but failed to give amplicons for all accessions of *A. hybridus* and one accession of *A. tuberculatus*. Again, two accessions of *A. albus* and all accessions of *A. spinosus* failed to amplify. These results suggest there might be a very large intron in the HPPD gene of *Amaranthus* species.

Primers designed to amplify the exon-only region of the HPPD gene worked well in all accessions and produced amplicons of the expected size. The forward primer hppdF7, combined with either hppdR5 or hppdR7 reverse primers, yielded amplicons of similar size between 300 and 400 bp respectively. Sequencing the amplicon of hppdF7/hppdR7 gave a 395 bp product, which was then used in the phylogenetic analysis. The coding segment of this gene featured 353 (89%) conserved, 22 (5.5%) variable and 18 (4.5%) informative sites, which was sufficient for inferring phylogenetic relationships among taxa (Fig. 3). All substitutions except one from *A. spinosus* were synonymous. Four plants showed apparent heterozygosity in HPPD, which included two accessions of *A. tuberculatus* (MH2 and MH43) that were distinguished by three and two single nucleotide polymorphisms (SNPs), and one individual of *A. palmeri* (MH254) that differed from other accessions of this species by three substitutions. Both of these species are dioecious and apparent diploids,[14] so heterozygosity is reasoned to be from different allelic forms rather than from multiple homeologous genes. In contrast, a total of eight SNPs in

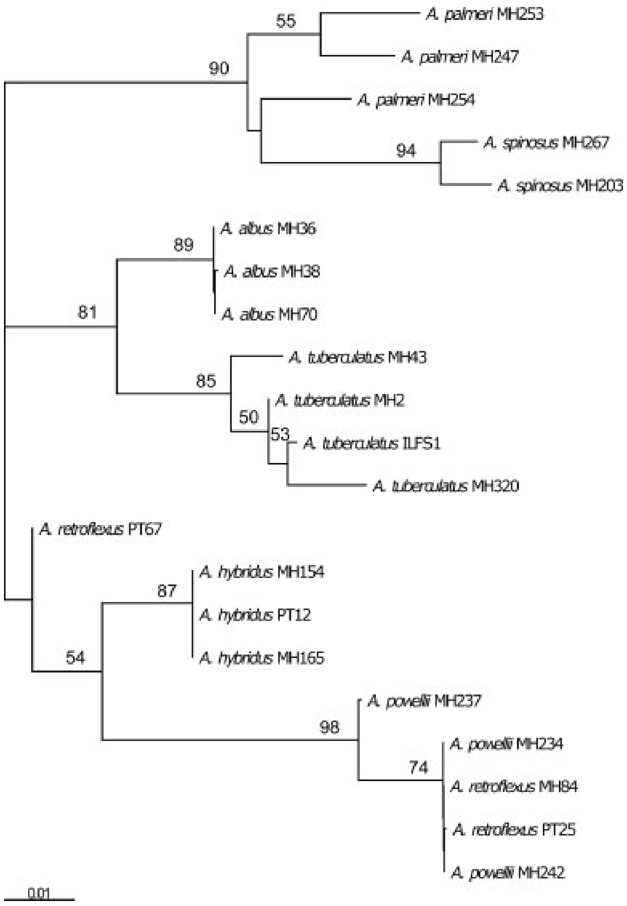

Figure 3. Neighbor-joining tree constructed from HPPD sequence data of *Amaranthus* species. Species annotations are referenced in Table 1. Bootstrap values are provided above branches.

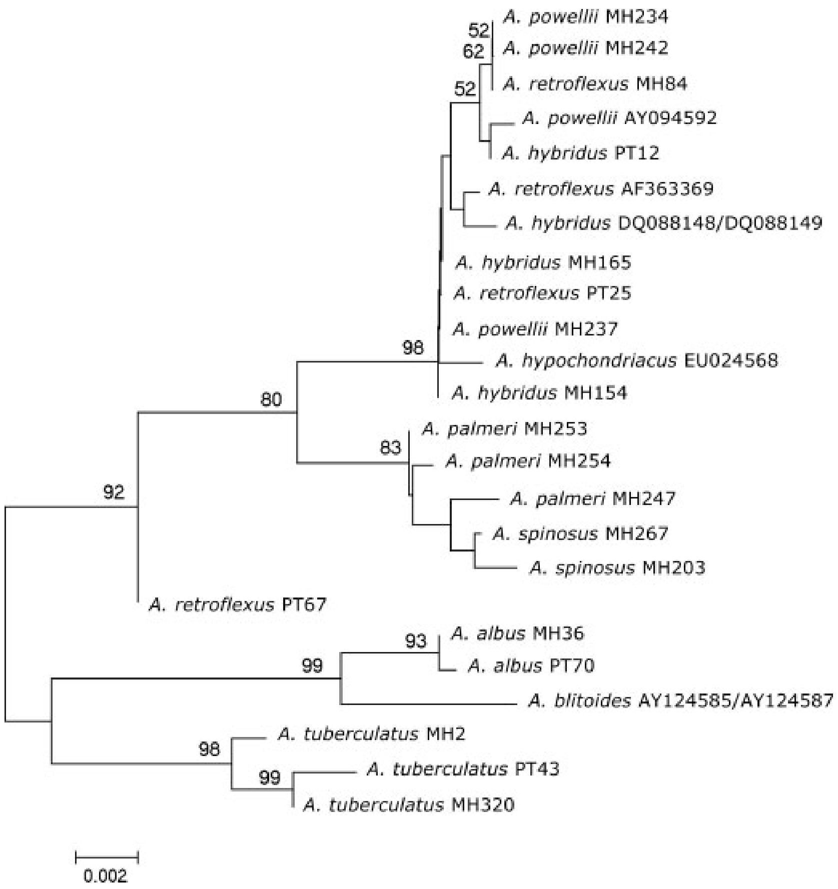

Figure 4. Neighbor-joining tree showing relationships of *Amaranthus* ALS (region A + region B) sequences. Species are annotated with accession labels (Table 1) or GenBank numbers. Clade support is assessed using bootstrap values.

one accession of the monoecious *A. retroflexus* (PT67) set it apart from other populations of this species, thus indicating that this specimen may be of hybrid origin.

3.5.2 *Acetolactate synthase (ALS)*

There is more information about ALS sequence characteristics, mutations and mechanisms of resistance in *Amaranthus* compared with any of the other target-site genes investigated in this study.[36] For these reasons, and the fact that this gene displays monogenic inheritance in *Amaranthus* species,[37] the authors chose to include ALS in this study for comparative purposes and to determine its potential for inferring evolutionary relationships among the sampled taxa. Only one contig in the transcriptome data yielded a hit to this gene, but, at 1987 nucleotides in length, it provided complete coverage of the ALS coding frame.

In addition to the accessions of taxa sequenced for this study, the multiple sequence alignment also included ALS sequences from five *Amaranthus* species previously deposited in GenBank (Fig. 4). These additional sequences were added to take advantage of their availability and to gain some insight into their placements within the ALS gene tree. As ALS primers have been previously described[22] and are routinely used in the authors' laboratory, no new primers were generated from the transcriptome data. Primers ALSf1–ALSr1 were used to amplify region A of ALS, whereas primers ALSf2–ALSr2 amplified region B (Table 2). Regions A and B of the ALS gene were previously identified as regions possessing polymorphisms leading to resistance.[22] These primer combinations worked without fail in all accessions and produced the expected amplicons of approximately 400 bp each. All six accessions of the two dioecious species sampled here were heterozygous for ALS. When the number and location of single nucleotide polymorphisms within ALS of all heterozygous plants were compared, more polymorphisms were observed in region B than in region A (Table 7). In region A, 24 variable sites were observed, three of which resulted in non-synonymous changes. In region B, only 11 variable sites were found, with one producing a non-synonymous change. For distance analysis, amplicons for ALS regions A and B were analyzed separately and concatenated for each sequenced accession. No incongruities in species relationships were observed between the two methods.

Table 7. Number of single nucleotide polymorphisms observed in ALS regions A and B of sampled *Amaranthus* species

Accession	ALS region A	ALS region B
A. tuberculatus MH2	1	–
A. tuberculatus MH43	1	5
A. tuberculatus MH320	1	1
A. palmeri MH247	1	6
A. palmeri MH253	4	4
A. palmeri MH254	2	4

3.5.3 *5-enolpyruvylshikimate-3-phosphate synthase (EPSPS)*

The enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS; EC 2.5.1.19) is the target for glyphosate, which is one of the most widely used herbicides in the world.[38] Inhibition of EPSPS interrupts the production of essential aromatic amino acids derived from the shikimate pathway. Three contigs of 478, 376 and 319 bp displayed high-similarity hits (>94% identity) to other *Amaranthus* EPSPS sequences. Identifying the positions of these non-overlapping contigs within the reading frame of the gene was facilitated by alignments with published complete coding cDNAs from *A. tuberculatus* (GenBank number FJ869880, FJ869881) and *A. palmeri* (FJ861242, FJ861243). The alignment also revealed a total of six polymorphic sites in the contig data. Adding genomic EPSPS sequences of *Arabidopsis* (AT1G48860) further assisted in locating exon–intron boundaries within the waterhemp contigs.

In *Arabidopsis*, splice variants of this gene consist of 7–9 exons, but, as the intron length and structure is unknown in *Amaranthus*, primers were anchored in adjacent exons with expectations of amplifying reasonably sized (<500–800 bp) amplicons. Primer combinations F3-R4, F1-R1 and F4-R2 yielded similar amplicons in most accessions, but all were well over 1500 bp in length, indicating large introns. Sequencing the products of F3-R4, however, proved difficult and only resulted in about 150 bp of coding sequence.

3.5.4 *Glutamine synthetase (GS)*

Glutamine synthetase (EC 6.3.1.2) is the target for the non-selective post-emergence herbicide glufosinate. Inhibition of GS disrupts the synthesis of glutamine from glutamate and ammonia, which ultimately interrupts photosynthesis, leading to membrane damage.[39] GS genes have been isolated from a number of plant species, as observed from the sequences deposited in GenBank. Genes that are responsible for GS activity in plants studied to date are part of a multigene family that encode cytosolic and chloroplastic forms.[40,41] Genes for GS subunits are also among the best conserved sequences in plants.[42] Unlike cytosolic GS sequences, the chloroplastic GS sequence appears to exist as a single copy gene in most plants.[43] The phylogenetic utility of the chloroplastic GS has also been investigated and, in fact, was shown to possess comparable levels of variability with the more commonly used nuclear ribosomal ITS region.[43]

Blast searches of the waterhemp transcriptome revealed seven contigs with high-fidelity hits (>81% identity) to known GS genes. Length variation of these contigs ranged from 202 to 1311 bp and included hits to both cytosolic and chloroplastic forms. Twelve GS sequences (six cytosolic and five chloroplastic) were downloaded from GenBank to create a multiple sequence alignment with which to align the waterhemp contigs and design primers. The alignment included cytosolic and chloroplastic GS sequences from *Beta vulgaris* (AY026353, AF343667) and *Spinacia oleracea* L. (EU057984, EF143582), both from the same family as waterhemp, which helped to assign the contigs to a particular form. This determination, however, was not especially straightforward, as overall sequence similarity between the two GS forms in the same plant is greater than 70%. However, certain 'chloroplastic diagnostic' sites[43] have been used to distinguish these isoforms from the cytosolic sequences and were relied upon in aligning the contigs.

Of the primers developed to amplify genomic GS sequences in *Amaranthus* species, three combinations worked consistently in all species tested (Table 2). The primers GS1-F4 and GS1-R1 resulted in a single band of approximately 500 bp in all accessions. After sequencing, about half of this amplicon was observed to be coding sequence. Sequencing also revealed the presence of a minor component of similar size in several accessions, thus indicating either multiple GS copies or perhaps non-specificity of this primer set for both cytosolic and chloroplastic sequences. Two other primer combinations, one with GS2-F3/GS2-R1 and a second with GS2-F8/GS2-R6, also resulted in apparent single-band amplicons for each accession, but these products were considerably larger in size (∼2000 bp) and were not sequenced. Future studies aimed at developing primers specific for chloroplastic GS in *Amaranthus*, based on the recommendations of Emshwiller and Doyle,[43] should be conducted to confirm whether this gene could be a useful molecular marker at lower taxonomic levels.

3.6 Phylogenetic analyses of *Amaranthus* species

The authors chose to investigate the phylogenetic utility of the HPPD and ALS genes for resolving relationships at low taxonomic levels in *Amaranthus*, based on the results of transcriptome analysis and sequencing experiments. These results, along with indications from the literature, suggested that both sequences were single or low-copy nuclear-inherited genes. If so, then these genes could serve as potentially useful new phylogenetic markers to test current hypotheses of species relationships and the origins of dioecy in the genus *Amaranthus*. Only a few molecular systematic studies have included species from the genus *Amaranthus*,[15,44–46] so a robust hypothesis of phylogenetic relationships of all species within the genus remains to be analyzed. Similarly, the relationships between the dioecious amaranths, which are endemic to North America, and their monoecious counterparts are not well understood.

Current hypotheses of species relationships among seven amaranth species were tested by sampling multiple accessions from populations previously investigated.[15] That study used AFLP markers and expanded sampling to determine relationships and the potential for interspecies hybridization among weedy dioecious and monoecious amaranths. Results of the AFLP study showed that the two dioecious species, waterhemp and Palmer amaranth, are genetically distinct and have their closest relationships among other monoecious species. The present results, based on HPPD and ALS distance data, agree with those previously reported[15] and suggest an independent evolution of dioecy from monoecious ancestors. Palmer amaranth appears to have its closest affinity with spiny amaranth, which is a relationship also supported by studies of pollen morphology[47] and genome sizes.[14] The other dioecious species, waterhemp, forms a distinct lineage but shows genetic affinities to tumble amaranth. The group composed of smooth pigweed, redroot pigweed and Powell amaranth agrees with AFLP data. These three species have very similar morphologies, and their close relationship as a complex has been previously discussed.[6,48] The phylogenetic results agree with a previous hypothesis[49] suggesting an independent origin of the dioecious habit from monoecious ancestors. This hypothesis, however, remains to be tested with complete sampling of all dioecious species.

4 CONCLUSIONS

In summary, a *de novo* characterization of the waterhemp transcriptome derived from 454 pyrosequencing has been presented, and the utility of this approach for identifying target-site genes involved in herbicide resistance has been demonstrated. A comparison of these results with those obtained from a previous whole genome 454 pyrosequencing experiment illustrates the benefits of transcriptome sequencing for expanding gene coverage and promoting efficient gene detection. To highlight this point, important herbicide target-site genes in waterhemp were identified, and four of these, HPPD, GS, EPSPS and ALS, were selected for further investigation. In addition to EPSPS and ALS, for which sequence data were previously available in waterhemp, partial sequence data for HPPD and glutamine synthetase have now been obtained. Having knowledge of the gene sequences for these two enzymes is important as it provides the necessary molecular tools to investigate resistance to herbicides that inhibit HPPD and GS if, and when, it occurs in the future. Another outcome from this study draws attention to the potential utility of the HPPD gene as a new nuclear marker for phylogenetic studies. Variation in this gene was sufficient for resolving relationships among the *Amaranthus* species sampled here and is certainly worth pursuing for use in other taxa.

SUPPORTING INFORMATION

Supporting information may be found in the online version of this article.

ACKNOWLEDGEMENTS

The authors thank Michael Bell for tissue collection, the WM Keck Center for Comparative and Functional Genomics at the University of Illinois for advice, library construction and sequencing and the Illinois Soybean Association for partial funding of this research. They also thank the anonymous reviewers for their helpful comments.

REFERENCES

1 Mosyakin SL and Robertson KR, *Amaranthus* L., in *Flora of North America North of Mexico, Vol. 4*, ed. by *Flora of North America Editorial Committee*. Oxford University Press, New York, NY, pp. 410–435 (2003).

2 Sauer JD, Revision of the dioecious amaranths. *Madroño* **13**:5–46 (1955).

3 Holm L, Doll J, Holm E, Pancho J and Herberger J, *World Weeds: Natural Histories and Distribution*. John Wiley & Sons, Inc., Toronto, ON (1997).

4 Steckel LE, The dioecious *Amaranthus* spp.: here to stay. *Weed Technol* **21**:567–570 (2007).

5 Basu C, Halfhill MD, Mueller TC and Stewart CN Jr, Weed genomics: new tools to understand weed biology. *Trends Plant Sci* **9**:391–398 (2004).

6 Costea M, Weaver SE and Tardif FJ, The biology of Canadian weeds. 130. *Amaranthus retroflexus* L., *A. powellii* S. Watson and *A. hybridus* L. *Can J Plant Sci* **84**:631–668 (2004).

7 Stewart CN, Jr, Tranel PJ, Horvath DP, Anderson JV, Rieseberg LH, Westwood JH, *et al*, Evolution of weediness and invasiveness: charting the course for weed genomics. *Weed Sci* **57**:451–462 (2009).

8 Tranel PJ and Horvath DP, Molecular biology and genomics: new tools for weed science. *Bioscience* **59**:207–215 (2009).

9 Hudson M, Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Resour* **8**:3–17 (2008).

10 Costea M, Weaver SE and Tardif FJ, The biology of invasive alien plants in Canada. 3. *Amaranthus tuberculatus* (Moq.) Sauer var. *rudis* (Sauer) Costea & Tardif. *Can J Plant Sci* **85**:507–522 (2005).

11 Jeschke MR, Tranel PJ and Rayburn AL, DNA content analysis of smooth pigweed (*Amaranthus hybridus*) and tall waterhemp (*A. tuberculatus*): implications for hybrid detection. *Weed Sci* **51**:1–3 (2003).

12 Heap I, *The International Survey of Herbicide Resistant Weeds.* [Online]. Available: www.weedscience.com [9 November 2009].

13 Lee RM, Thimmapuram J, Thinglum KA, Gong G, Hernandez AG, Wright CL, *et al*, Sampling the waterhemp (*Amaranthus tuberculatus*) genome using pyrosequencing technology. *Weed Sci* **57**:463–469 (2009).

14 Rayburn AL, McCloskey R, Tatum TC, Bollero GA, Jeschke MR and Tranel PJ, Genome size analysis of weedy *Amaranthus* species. *Crop Sci* **45**:2557–2562 (2005).

15 Wassom JJ and Tranel PJ, Amplified fragment length polymorphism-based genetic relationships among weedy *Amaranthus* species. *J Heredity* **96**:410–416 (2005).

16 Doyle JJ and Doyle JL, Isolation of plant DNA from fresh tissue. *Focus* **12**:13–15 (1990).

17 Dassanayake M, Haas JS, Bohnert HJ and Cheeseman JM, Shedding light on an extremophile lifestyle through transcriptomics. *New Phytol* **183**:764–775 (2009).

18 Chou HH and Holmes MH, DNA sequencing quality trimming and vector removal. *Bioinformatics* **17**:1093–1104 (2001).

19 Huang X and Madan A, CAP3: a DNA sequence assembly program. *Genome Res* **9**:868–877 (1999).

20 Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, *et al*, EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic Acids Res* **34**:W459–462 (2006).

21 Tamura K, Dudley J, Nei M and Kumar S, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**:1596–1599 (2007).

22 Foes MJ, Liu L, Tranel PJ, Wax LM and Stoller EW, A biotype of common waterhemp (*Amaranthus rudis*) resistant to triazine and ALS herbicides. *Weed Sci* **46**:514–520 (1998).

23 Wang W, Wang Y, Zhang Q, Qi Y and Guo D, Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10**:465–474 (2009).

24 Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao Y and Town CD, Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* **7**:272–281 (2006).

25 Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, *et al*, Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**:347–365 (2009).

26 Barbazuk WB, Fu Y and McGinnis KM, Genome-wide analyses of alternative splicing in plants: opportunities and challenges. *Genome Res* **9**:1381–1392 (2008).

27 Chinnusamy V, Zhu J and Zhu JK, Cold stress regulation of gene expression in plants. *Trends Plant Sci* **12**:444–451 (2007).

28 Yuan JS, Tranel PJ and Stewart CN, Jr, Non-target-site herbicide resistance: a family business. *Trends Plant Sci* **12**:6–13 (2007).

29 Rea PA, Plant ATP-binding cassette transporters. *Annu Rev Plant Biol* **58**:347–375 (2007).

30 Wheat CW, Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* (publ. online 18 Oct. 2008); DOI: 10.1007/s10709-008-9326-y (2008).

31 Imelfort M and Edwards D, *de novo* sequencing of plant genomes using second-generation technologies. *Brief Bioinform* **10**:609–618 (2009).

32 Matringe M, Sailland A, Pelissier B, Rolland A and Zink O, *p*-Hydroxyphenyl-pyruvate dioxygenase inhibitor-resistant plants. *Pest Manag Sci* **61**:269–276 (2005).

33 Fritze IM, Linden L, Freigang J, Auerbach G, Huber R and Steinbacher S, The crystal structures of *Zea mays* and *Arabidopsis* 4-hydroxyphenylpyruvate dioxygenase. *Plant Physiol* **134**:1388–1400 (2004).

34 Moran GR, Minireview – 4-hydroxyphenylpyruvate dioxygenase. *Arch Biochem Biophys* **433**:117–128 (2005).

35 Xiao Y, Di P, Chen J, Liu Y, Chen W and Zhang L, Characterization and expression profiling of 4-hydroxyphenylpyruvate dioxygenase gene (Smhppd) from *Salvia miltiorrhiza* hairy root cultures. *Mol Biol Rep* **36**:2019–2029 (2009).

36 Tranel PJ and Wright TR, Resistance of weeds to ALS-inhibiting herbicides: what have we learned? *Weed Sci* **50**:700–712 (2002).

37 Trucco F, Jeschke MR, Rayburn AL and Tranel PJ, Promiscuity in weedy amaranths: high frequency of female tall waterhemp (*Amaranthus tuberculatus*) × smooth pigweed (*A. hybridus*) hybridization under field conditions. *Weed Sci* **53**:46–54 (2005).

38 Baylis AD, Why glyphosate is a global herbicide: strengths, weaknesses and prospects. *Pest Manag Sci* **56**:299–308 (2000).

39 Tan S, Evans R and Singh B, Herbicidal inhibitors of amino acid biosynthesis and herbicide-tolerant crops. *Amino Acids* **30**:195–204 (2006).

40 Doyle JJ, Evolution of higher-plant glutamine synthetase genes: tissue specificity as a criterion for predicting orthology. *Mol Biol Evol* **8**:366–377 (1991).

41 Pornprom T, Prodmatee N and Chatchawankanphanich O, Glutamine synthetase mutation conferring target-site-based resistance to glufosinate in soybean cell selections. *Pest Manag Sci* **65**:216–222 (2008).

42 Perez-Vicent R, Dorado G and Maldonado JM, Cross-species amplification of glutamine synthetase cDNA by polymerase chain reaction with degenerate primers. *Phys Plant* **98**:705–713 (1996).

43 Emshwiller E and Doyle JJ, Chloroplast-expressed glutamine synthetase (ncpGS): potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). *Mol Phylogenet Evol* **12**:310–319 (1999).

44 Song BH, Chen ZD, Wang XQ and Li FZ, Sequence analysis of the ITS region of nuclear ribosomal DNA (nrDNA) in Chinese *Amaranthus* and its systematic utility. *Acta Bot Sin* **42**:1184–1189 (2000).

45 Müller K and Borsch T, Phylogenetics of Amaranthaceae based on matK/trnK sequence data: evidence from parsimony, likelihood, and Bayesian analyses. *Ann Mo Bot Gard* **92**:66–102 (2005).

46 Sage RF, Sage TL, Pearcy RW and Borsch T, The taxonomic distribution of C4 photosynthesis in Amaranthaceae sensu stricto. *Am J Bot* **94**:1992–2003 (2007).

47 Franssen AS, Skinner DZ, Al-Khatib K and Horak MJ, Pollen morphological differences in *Amaranthus* species and interspecific hybrids. *Weed Sci* **49**:732–737 (2001).

48 Costea M, Sanders A and Waines G, Preliminary results toward a revision of the *Amaranthus hybridus* species complex (Amaranthaceae). *Sida* **19**:931–974 (2001).

49 Mosyakin SL, On the origin of dioecious amaranths (*Amaranthus* L., Amaranthaceae Juss.). *Ukr Bot J* **62**:3–9 (2005).