Research Article

Received: 3 March 2010 Revised: 7 June 2010 Accepted: 8 June 2010 Published online in Wiley Online Library: 16 August 2010

(wileyonlinelibrary.com) DOI 10.1002/ps.2004

Characterization of the horseweed (*Conyza canadensis*) transcriptome using GS-FLX 454 pyrosequencing and its application for expression analysis of candidate non-target herbicide resistance genes

Yanhui Peng,[a] Laura LG Abercrombie,[a] Joshua S Yuan,[b,c] Chance W Riggins,[d] R Douglas Sammons,[e] Patrick J Tranel[d] and C Neal Stewart, Jr[a]*

Abstract

BACKGROUND: The *de novo* transcriptome sequencing of a weedy plant using GS-FLX 454 technologies is reported. Horseweed (*Conyza canadensis* L.) was the first broadleaf weed to evolve glyphosate resistance in agriculture, and also is the most widely distributed glyphosate-resistant weed in the United States and the world. However, available sequence data for this species are scant. The transcriptomic sequence should be useful for gene discovery, and to help elucidate the non-target-based glyphosate resistance mechanism and the genomic basis of weediness.

RESULTS: Sequencing experiments yielded 411 962 raw reads, an average read length of 233 bp and a total dataset of 95.8 Mb (NCBI accession number SRA010952). After trimming and quality control, 379 152 high-quality sequences were retained and assembled into contigs. The assembly resulted in 31 783 unique transcripts, including 16 102 contigs and 15 681 singletons. The average coverage depth for each contig and each nucleotide position was 22-fold and 12-fold respectively. A total of 16 306 unique sequences were annotated by searching a custom plant protein database. The utility of the transcriptome data was demonstrated by further exploration of ABC transporters, which were previously hypothesized to play a role in non-target glyphosate resistance. Real-time RT-PCR primers were designed from the transcriptome data, which made it possible to assess expression patterns of 17 ABC transporters from resistant and susceptible horseweed accessions from Tennessee, with and without glyphosate treatment.

CONCLUSION: These results show that GS-FLX 454 sequencing is a powerful and cost-effective platform for the development of functional genomic tools for a weed species.
© 2010 Society of Chemical Industry

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

Keywords: horseweed; *Conyza canadensis*; glyphosate resistance; high-throughput transcriptome sequencing; GS-FLX 454 DNA sequencing; ABC transporters

1 INTRODUCTION

Glyphosate has become the world's most widely used herbicide for controlling weeds for a number of reasons, including its high efficacy and low cost, and because it is environmentally benign. Using glyphosate along with no-till cropping systems is considered to be a superior economic and environmental choice compared with other systems.[1,2] The widespread use of glyphosate, however, has exerted selection pressure on various species of weeds. In fact, agricultural weeds are becoming more difficult to control as they continue rapidly to evolve herbicide resistance.[3] Horseweed (*Conyza canadensis* L.), which is in the Asteraceae family, was the first broadleaf weed to evolve glyphosate resistance,[4] first occurring in Delaware in 2000. Resistant biotypes are found in 20 US

* *Correspondence to: C Neal Stewart, Jr, Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996-4561, USA. E-mail: nealstewart@utk.edu*

a *Department of Plant Sciences, University of Tennessee, Knoxville, TN USA*

b *Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA*

c *Institute of Plant Genomics and Microbiology, Texas A&M University, College Station, TX, USA*

d *Department of Crop Sciences, University of Illinois, Urbana, IL, USA*

e *Monsanto Company, St Louis, MO, USA*

states and several countries on four continents. The authors have recently performed a phylogeographic study providing evidence that horseweed has evolved glyphosate resistance independently in many locations in the United States.[5] Resistant biotypes seem abruptly to appear and then spread within populations. This within-population spread of resistance is enabled by high seed production (each mature plant can produce more than 200 000 wind-dispersed seeds) coupled with glyphosate treatment which kills non-adapted genotypes.[6]

Horseweed has several attractive features making it amenable for genomics research.[7,8] From the authors' own flow cytometry experiments it is estimated that horseweed has a genome size of about 335 Mb (unpublished data), which is approximately 2.5 times the size of *Arabidopsis thaliana* (L.) Heynh. In fact, horseweed has the smallest known genome of all agricultural weeds.[8] It is self-fertile, has high homozygosity and is relatively easy to maintain in low-light growth rooms until plants bolt, when they quickly outgrow light rack spacing. Horseweed is a true diploid ($2n = 18$), which simplifies sequence analysis compared with polyploid weed species. In addition, a plant transformation and regeneration method has been developed[9] that allows for overexpression or knockdown analysis of potential gene targets.

The likely mechanism for glyphosate resistance in horseweed does not involve the target EPSPS enzyme-coding genes (EPSPS1, GenBank accession number AY545 666; EPSPS2, accession number AY545 667).[5,10] Moreover, the non-target mechanism appears to have a relatively simple Mendelian inheritance, indicating that a single locus is responsible for conferring resistance.[9,11] In addition to the high fecundity and glyphosate selection pressure, the spread of resistant biotypes within and among populations is likely enhanced by autogamy and simple inheritance of dominant-to-semi-dominant non-target-site glyphosate resistance in this species.

Horseweed, like most agricultural weeds, has received little attention from genomics researchers.[8] However, 2019 expressed sequence tags (ESTs) were recently identified and analyzed.[5] In addition, a list of genes upregulated by glyphosate in a resistant biotype compared with an isogenic susceptible genotype were discovered by the use of heterologous microarrays.[5] Many more horseweed gene sequences must be discovered to validate and expand microarray results, to enable gene discovery and cloning candidate genes, and to provide baseline data for further functional genomics analysis. Therefore, transcriptome sequencing was performed using a high-throughput method to obtain the first large-scale genomic information in horseweed.

For non-model plants, such as weeds, the traditional method of obtaining genome or transcriptome data has been through library construction, repeated rounds of normalization/subtraction, followed by traditional Sanger sequencing, which was a lengthy and expensive process in spite of incremental improvements in Sanger sequencing technology. Sequencing of eukaryotic genomes or transcriptomes has remained beyond the typical grant-funded investigator. Recently, the Sanger method has been partially supplanted by several 'next-generation' sequencing technologies that offer dramatic increases in cost-effective sequence throughput.[12,13] One extensively used method relies on Roche GS-FLX 454 technology, which has had a tremendous impact on genomic research for increasing sequencing depth and coverage while reducing time, labour and cost.[13,14] GS-FLX 454, the first next-generation sequencing technology, was released onto the market in October 2005[12] and has been the most widely published next-generation technology, with more than 600 peer-reviewed research publications to date. GS-FLX 454 has been applied to standard sequencing applications, such as *de novo* genome and transcriptome sequencing, resequencing[15–20] and for novel applications previously unexplored by Sanger sequencing.[21] The 454 technology avoids expensive cloning-based library construction by taking advantage of a highly efficient *in vitro* DNA amplification method known as emulsion PCR.[12] Followed by pyrosequencing,[22,23] the GS-FLX 454 system is capable of generating on average 100 Mb of 250 base reads per 7.5 h run (http://www.454.com). Compared with similar expenditure by Sanger sequencing, GS-FLX 454 yields redundant coverage for many more genes. Also, even though the read length is shorter than Sanger sequencing, the lower error level (<0.5%) associated with 454 technology is beneficial for sufficient coverage depth to allow assembly of overlapping reads.[21] It produces less concern about assemblage for transcriptomes. Compared with genomes, transcriptomes are smaller and typically contain much less repetitive DNA. However, the 454 technique is not perfect. As pyrosequencing relies on the magnitude of light emitted to determine the number of repetitive bases, erroneous base calls can occur with homopolymers. Also, 454 is less expensive and faster on a per base basis, but a single 454 run is expensive compared with Sanger, and thus 454 is not suitable for sequencing targeted fragments from small DNA samples. Although next-generation technologies allow genome sequencing to become more efficient, the sequencing of complex genomes remains expensive, often prohibitively so. Therefore, the authors chose to perform transcriptome sequencing of horseweed by using the 454 platform to acquire candidate sequences for functional genomics analysis. Here, the *de novo* assembly of horseweed transcriptome data and the annotation of expressed genes are reported, and the utility of these data for gene expression analysis is demonstrated. Resultant sequence data are publicly available from GenBank (accession number SRA010952).

2 MATERIALS AND METHODS

2.1 Sample preparation for 454 sequencing

Horseweed plants were grown in potting media in a greenhouse at the University of Tennessee, Knoxville, TN, under a 16 : 8 h light : dark photoperiod at ambient temperatures (25 ± 2 °C). Plants were watered and fertilized as necessary with Osmocote slow-release fertilizer. Young leaves and meristematic tissue were harvested at the rosette stage from plants that were approximately 3 months old and 6–8 cm in diameter. Total RNA was isolated and pooled into one sample from the following three sample types represented by six plants each: untreated (water-sprayed) TN-susceptible horseweed (from Knoxville, TN) and TN-resistant biotype from western Tennessee (Lauderdale County, TN), and the latter biotype after 24 h glyphosate-sprayed with the field rate of Round-up Weathermax (0.84 kg AI ha^{-1}; Monsanto, St Louis, MO).[24] RNA extraction was done using TriReagent according to the manufacturer's protocol (MRC, Cincinnati, OH). The pooled sample was used to generate double-stranded cDNA using SMART™ cDNA Library Construction Kit (Clontech, Mountain View, CA). Normalization was performed using TRIMMER cDNA Normalization Kit (Evrogen, Moscow, Russia) to decrease the prevalence of abundant transcripts before sequencing. The cDNA sample was then fractionated into smaller pieces (300–500 bp). The ends of these fragments were subsequently polished by treating with DNA polymerase to fill in or remove any unpaired bases. The short A and B adaptors were then ligated on to each resulting fragment, which

 © 2010 Society of Chemical Industry

provided priming sequences for both emulsion PCR amplification and pyrosequencing, forming the basis of the single-stranded template library. Pyrosequencing using a Roche GS-FLX sequencer was performed by the WM Keck Center for Comparative and Functional Genomics at the University of Illinois, as described previously.[12,25] A preliminary titration run was followed by two bulk runs. The first bulk run was dedicated to horseweed cDNAs, whereas in the second run only half the plate was allocated to horseweed cDNAs. Raw reads of 454 data were submitted to the GenBank Short Read Archive (SRA) database (accession number SRA010952). All contig sequences present in this study are provided in Supporting Information File 6.

2.2 454 sequencing data trimming, assembling and annotation

The raw 454 sequences were processed according to standard protocols used in the authors' waterhemp [*Amaranthus tuberculatus* (Moq.) JD Sauer] paper,[26] and the waterhemp transcriptome data were used as a comparison in this study. Lucy[27] and EGassembler (http://egassembler.hgc.jp/) were used to remove the low-quality sequences, end regions that were rich in ambiguous nucleotides, very short reads (<50 bp), poly (A/T) tails, SMART™ adaptors for cDNA synthesis, primers and potential contaminating vector sequences. The returned high-quality clean sequences were assembled using CAP3[28] and EGassembler. All unique sequences (contigs and singletons) were annotated by similarity search (NCBI Standalone Blast program, ftp://ftp.ncbi.nih.gov/blast/) of three protein databases: *Arabidopsis* all proteins database (AGIallAA.gz, 130 814 protein sequences; ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/other_datasets/12-18-07/), UniProtKB/Swiss-Prot annotated protein database (353 658 protein sequences, http://www.uniprot.org/downloads) and a custom protein database including all green plants proteins from GenBank (677 422 protein sequences, ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/). The best five protein hits for each query sequence were parsed out to create annotated tables, which included available information such as taxonomy, keywords, protein function, accession number and/or gene ontology (GO) terms. The potential micro-RNAs in non-annotated sequences were scanned by searching the miRBase database (ftp://mirbase.org/pub/mirbase/CURRENT/). To help determine which sequences were of non-plant origin, contiguous sequences (contigs) and singletons that had no hits found in the custom plant proteins database were further searched using the Uniport database (Release 15.14).

2.3 Expression analysis of selected ABC transporter genes using real-time RT-PCR

Plants were grown and harvested and total RNA was extracted as described above. Four combinations of plant biotypes and treatments were made: TN-S and TN-R biotypes that were glyphosate treated and untreated were compared for gene expression differences. Young leaves of six individual plants were used for total RNA extraction for each biotype–treatment combination. Therefore, the four combinations were represented by one sample each. The residual genomic DNA in the total RNA extract was removed by several treatments with RNase-free DNase I (Invitrogen, Carlsbad, CA). First-strand cDNA was synthesized using 2 μg of total RNA, 0.5 μg of oligo$(dT)_{18}$ and SuperScript® III reverse transcriptase according to the manufacturer's instructions (Invitrogen), employing an Eppendorf MasterCycler (Eppendorf, Hamburg, Germany). The cDNAs were diluted to 100 μL with sterile water, of which 2 μL was used per real-time PCR sample. Real-time PCR was carried out in an ABI-7000 thermal cycling system using a real-time PCR Power Mix Kit (ABI, Foster City, CA). The reaction mixture (25 μL) contained 2 μL of first strand cDNA, 0.5 μM of each of the forward and reverse primers and appropriate amounts of other components as recommended by the manufacturer (ABI). The ABI-7000 thermal cycler was programmed as follows: 2 min at 95 °C for predenaturation, 40 cycles of 15 s at 94 °C, 15 s at 55 °C and 20 s at 72 °C. Data were collected during the extension step. The cDNA samples were tested by using three independent repetitions in the same condition. For control reactions, either no sample was added or RNA alone was added without reverse transcription to test if the RNA sample was contaminated with genomic DNA. An actin-like housekeeping gene (contig9305, 916 bp) was used as a reference gene. The absolute expression level of this actin-like gene was relatively invariant (average ±0.31 Ct value, within 10% variation) using equal amounts of cDNA samples from glyphosate-treated plants in this study. Furthermore, abiotic stresses (salt, drought and cold; data not shown) did not perturb its expression. The relative expression of target genes to the actin control was calculated using the efficiency-adjusted ΔΔCt method as described by Yuan *et al.*[29] The oligonucleotide primers (Supporting Information File 1) were designed with the Primer Express 2.0 software (ABI). To test the suitability of these primer sets, the specificity and identity of the RT-PCR products were monitored by a melting curve analysis (65–99 °C, 5 °C s^{-1}) of the reaction products, which can distinguish the gene-specific PCR products from the non-specific PCR products. All primers were synthesized by Integrated DNA Technologies (IDT, Iowa City, IA).

3 RESULTS AND DISCUSSION

3.1 Roche GS-FLX sequencing and assembly

Normalized cDNA was used to reduce oversampling of high-abundance transcripts and obtain sufficient coverage of low-abundance transcripts. Two sequencing runs (1.5 plates) plus a titration run yielded a total of 411 962 raw reads. The average length of each read was 233 bp (Supporting Information File 2), with 79.2% distributed between 200 bp and 300 bp, and the total data size was 95.8 Mb (Fig. 1). The sequence yield was somewhat lower compared with genomic DNA 454 sequencing, but was higher than other *de novo* transcriptome sequencing projects for non-model plant species.[30,31] The difference resulted from shorter DNA fragments from the transcriptome preparation or other input effects compared with those data from genome studies. Compared with Sanger EST library sequencing methods, cDNA molecules needed to be fractionated into smaller pieces and size scanned rather than fully cloned into vectors. Shotgun 454 sequences are located evenly across the cDNA of a given gene,[32] which resulted in multiple fragments per single gene, requiring further analysis to assess their relationships.

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 2.2. A total of 94% of sequences (379 152) passed the quality-control filter for assembly into unique sequences. A total of 363 471 high-quality clean sequences resulted in 7.05 Mb representing 16 102 contigs. After assembly, 55% of the contigs (8817) were longer than 300 bp, and 19.5% of them (3145) were longer than 600 bp (Fig. 2). Of these contigs, 15 681 high-quality clean sequences (3.8%) remained as singletons (coverage depth = 1), with the data size totaling

 © 2010 Society of Chemical Industry

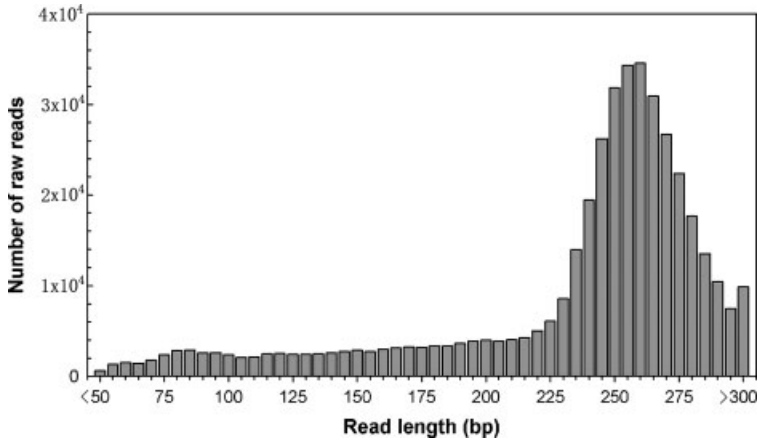

Figure 1. Frequency distribution of horseweed GS-FLX 454 sequence raw read lengths.

Table 1. Summary of 454 sequencing, data trimming, assemblage and annotation

	Sequences (number; % of total)	Nucleotides (number; % of total)
Raw reads	411 962 (100%)	95 822 385 (100%)
After quality control and trimming	379 152 (92.04%)	88 013 743 (91.85%)
Sequences assembled to contigs	363 471 (88.23%)	84 692 110 (88.38%)
Assembled contigs	16 102	7 054 041 (7.36%)
Singletons	15 681 (3.81%)	3 321 633 (3.47%)
Annotated unique sequences	16 306	6 478 624

Table 2. Summary Blast data for assembled 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 horseweed FLX-454 contigs	16 102
Number of contigs that had best Blast hits with other contigs and singletons	4405 (27.1%)
Average length of alignments as contigs versus contigs	92 bp
Average length of alignments as contigs versus singletons	73 bp
Average coverage of alignments to queried contig as contigs versus contigs	18.5%
Average coverage of alignments to queried contig as contigs versus singletons	11.5%
Number of Blastn-paired contigs that had Blastx hits against protein database	2768 (17.2%)
Number of Blastn-paired contigs that had identical best Blastx hits	285 (1.8%)

3.3 Mb. This resulted in 10.35 Mb of new horseweed transcriptome sequencing data representing 31 783 unique sequences (Table 1). A further 2016 quality-trimmed unique transcripts (average length 689 bp; total size 1.39 Mb) obtained from horseweed cDNA libraries using traditional Sanger sequencing techniques[5] were used to gauge the quality of the 454 sequencing and assembly.

The most challenging aspect of *de novo* assembly is obtaining abundant coverage of sequences. In this study, 95.9% of the high-quality trimmed sequences were assembled into contigs with an average length of 438 bp. However, this average length was still shorter than the average length of Sanger ESTs. Given that the average coverage depth for each contig and each nucleotide position was ~22-fold and ~12-fold respectively, this high coverage depth of contigs ensured that the 454 sequences were likely more accurate than traditional Sanger sequences which rely on a single or very few reads.

3.2 Quality and performance of the 454 assembly

To test the quality and performance of the sequence assembly, contigs were aligned against themselves and the singletons using the NCBI Blastn program. A total of 4405 contigs (27.1%) had best Blast hits (i.e. had significantly similar sequences based on a bit score of >45 and an *E*-value of <0.0001 produced by the Blastn program) 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 92 bp and 73 bp in length for contig versus contigs and contigs versus singleton hits respectively. Also, the average coverage of the match alignments were 18.5% and 11.5% of the length of the queried contigs in the case of contigs versus contigs and contigs versus singletons respectively. A total of 2768 of these contigs had Blastx hits (bit score >45) against the all green plants protein database, and only 285 (1.8%) of those Blastn-paired contigs had the same best Blastx hits in the protein database (Table 2). Considering that conserved motifs in different genes widely exist in the genome and different transcripts resulting from alternative splicing of single genes occur frequently,[33,34] the present assembly appropriately partitioned these gene regions, which produced high identity but short coverage alignments into different contigs.

To estimate the error rate of 454 sequencing and the quality of assembly, 2016 high-quality trimmed Sanger-sequenced ESTs[5] were aligned with the 454 contigs and singletons. Of these, 1540 (76.4%) had strong Blast hits to 454 sequences (Table 3). Nucleotide alignments of Sanger versus 454 sequences were 95.8% identical for all alignments, 97.3% for those alignments involving 454 contigs and 99.3% for those alignments with a bit score of >100. The average number of gaps for alignments involving 454 singletons was 0.22 and 7 per 1000 aligned bases. The average number of gaps for alignments involving 454 contigs

 © 2010 Society of Chemical Industry

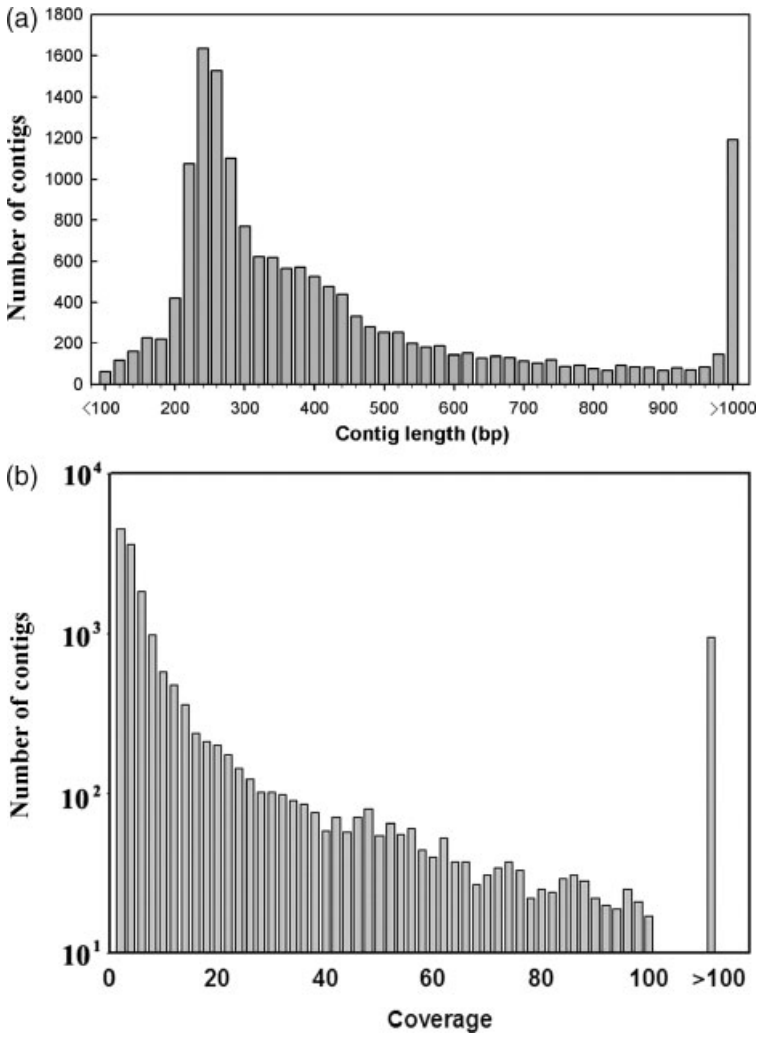

Figure 2. Characteristics of assembled horseweed GS-FLX 454 contigs: (a) length frequency distribution of assembled contigs; (b) average coverage frequency distribution of assembled contigs.

was 0.04 and 1.4 per 1000 aligned bases, which was less than that for 454 singletons. This comparison might overestimate the real 454 sequencing error rates, as they include base mismatches caused by polymorphisms, possible gaps created by alternative splicing and alignments with end regions of Sanger sequences, which are known to have decreased accuracy. The horseweed genotypes between Sanger and 454 sequencing were not the same. However, these results indicated that a sufficient coverage depth could efficiently reduce the error rate in 454 sequencing, and it is reasonable to suggest that it could be more accurate than traditional Sanger sequences on the basis of depth of coverage.

3.3 Functional annotation of 454 unique sequences

One objective of this project was to assign hypothetical protein sequence and function to each EST. All unique sequences (contigs and singletons) were used as queries to search annotated protein databases and were assigned a gene description and/or a GO term (Supporting Information Files 3 and 4). A number of factors, especially the *E*-value, affect the reliability of results when searching databases for similarities. The *E*-value is the probability, due to chance, that there is another alignment with a similarity greater than the given bit score. In short, a lower *E*-value set translates to higher confidence in the search results. A total of 10 698 contigs had hits to the protein database with the *E*-value threshold set at <0.1, which was 1438 ($\sim$16%) more than that with the *E*-value threshold set at <0.0001 (Supporting Information File 5). The authors sought to enlarge the database to allow maximal functional searching for gene discovery in this *de novo* transcriptome sequencing project. The number of contigs that had hits to different protein databases was counted on the basis of the 'best five hits' of Blastx search (*E*-value <0.0001, score bits >45). The greatest yield of protein counts was obtained when searching the all green plant protein database, which hit 629 more contigs than searching the *Arabidopsis* protein database; about 20% more putative proteins were identified. Thus, a total 16 306 unique sequences were annotated. Of these, 13 708 (84.1%) were associated with Biological Process GO classification and were divided into 14 subgroups, 12 404 (76.1%) were associated with 'cellular components' and were further divided into 16 subgroups and 7364 (45.2%) were associated with 'molecular function' and were divided into 15 subgroups (Fig. 3).

Only 39.8% of singletons found hits in the present custom protein database and could be annotated, while 61.5% of the contigs could be annotated, possibly the result of low coverage

 © 2010 Society of Chemical Industry

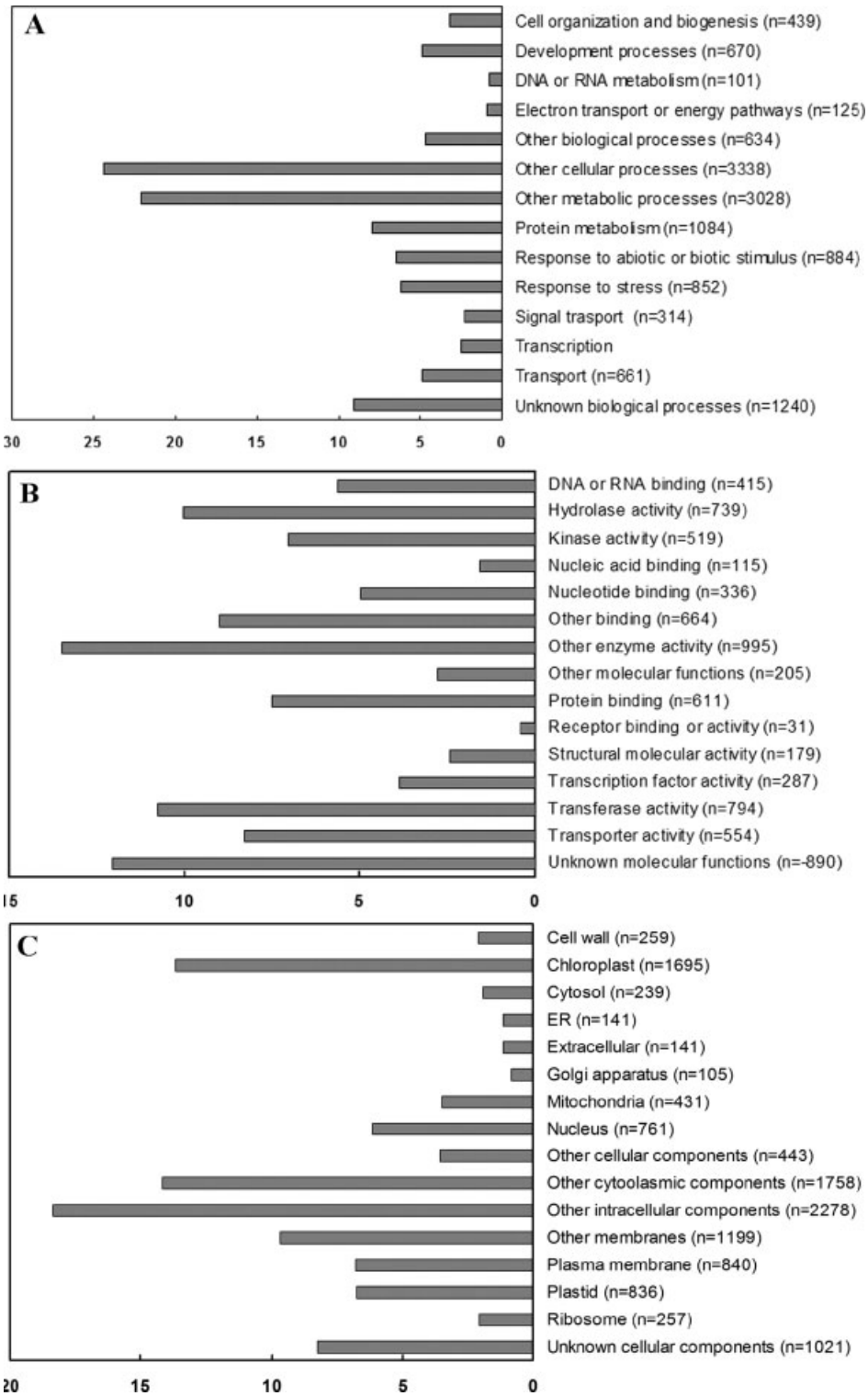

Figure 3. Summary of GO annotation of 454 unique sequences. Annotated sequences were classified into (A) 'biological process', (B) 'molecular Function' and (C) 'cellular component' groups and 45 subgroups.

depth and the short average length of the singletons. The average length of annotated contigs was 526 bp with a 30.3 average coverage, while non-annotated contigs averaged 297 bp with only 9.6 coverage. Similarly, the average length of annotated singletons was 30 bp longer than that of non-annotated singletons. In 15 477 non-annotated unique sequences, ~2.8% of these (431) had hits in the plant micro-RNA database, and ~0.9% of them (134) had hits with non-plant proteins (Table 4). The remainder might be partial sequences that locate to non-coding regions (5′ and 3′ untranslated regions of genes) or some horseweed-specific genes that have no orthologs in these databases. However, the inability to annotate these sequences is a problem that could be common

 © 2010 Society of Chemical Industry

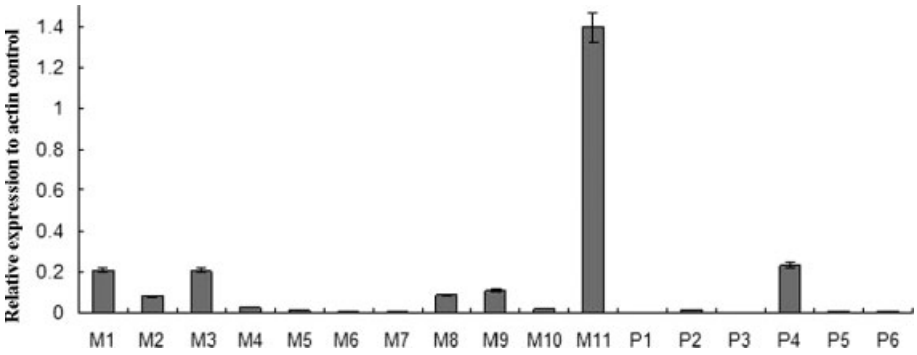

Figure 4. Expression levels of 17 ABC transporter genes in young leaves of glyphosate-treated TN-R biotype horseweed plants relative to an internal control actin gene using real-time RT-PCR. Data are presented as mean ± SE of three technical replicates for each biotype–treatment combination (one pooled sample each).

Table 3. Summary Blast data for assembled Sanger sequences against FLX-454 contigs and singletons. All Blast results refer to hits with bit scores greater than or equal to 45. Alignment lengths refer to nucleotides

Number of Sanger sequences	2016
Number of Sanger sequences with at least one Blast hit against 454 sequences	1540
Percentage Sanger sequences with a Blast hit against 454 sequence	76.4%
Mean percentage identity of Sanger versus all 454 contig Blast hit alignments	95.8%
Mean percentage identity of Sanger versus all 454 Blast hit alignments	97.3%
Mean percentage identity of Sanger versus 454 Blast hit alignment (bit score >100)	99.3%
Mean number of gaps within Sanger versus all 454 contig Blast hit alignments	0.04
Median number of gaps within Sanger versus all 454 contig Blast hit alignments	0
Mean number of gaps within Sanger versus all 454 singleton Blast hit alignments	0.22
Median number of gaps within Sanger versus 454 singleton Blast hit alignments	0

Table 4. Summary data for horseweed-unique annotated and non-annotated sequences

Average length of annotated contigs	526 bp
Average coverage of annotated contigs	30.3-fold
Average length of non-annotated contigs	297 bp
Average coverage of non-annotated contigs	9.6-fold
Average length of annotated singletons	230 bp
Average length of non-annotated singletons	199 bp
Number of non-annotated unique sequences having hits to micro-RNAs	431
Number of non-annotated unique sequences having hits to non-plant proteins	135

Table 5. Comparison of the number of hits (contigs + singletons) to herbicide target-site genes and gene families and non-target gene families from transcriptome 454 sequencing of horseweed and waterhemp

	Horseweed	Waterhemp
Herbicide target gene family		
Acetolactate synthase	6	2
D1 protein (plastidic gene)	4	2
Tubulin	29	33
Protoporphyrinogen oxidase	2	8
Phytoene desaturase	5	1
Glutamine synthetase	9	7
1-Deoxy-D-xylylose-5-phosphate synthase	6	1
4-Hydroxyphenylpyruvate dioxygenase	1	2
Acetyl-CoA carboxylase	6	8
Dihydropteroate synthase	1	2
5-Enolpyruvylshikimate-3-phosphate synthase	2	3
Non-target gene family		
Glutathione *S*-transferase	7	22
Cytochrome P450 monooxygenases	125	191
Glycosyltransferases	76	84
ABC transporter genes	151	192

for many non-model plants. The authors are planning to use either 454 FLX-Titanium or Solexa (Illumina) to sequence unnormalized cDNA (RNAseq) samples to fill in gaps and acquire additional target genes involved in glyphosate resistance. At that time, transcriptome data from resistant and susceptible isogenic lines with and without glyphosate treatment will be compared. More data might also help address the annotation problem. However, if genes involved in weediness are truly novel to weeds, then a comprehensive functional genomics research program would be required to elucidate the genes and their functions.[7,8]

The effectiveness of this horseweed 454 transcriptome sequencing for identifying gene candidates involved in herbicide resistance was estimated by comparing with waterhemp data.[26] Eleven herbicide target-site genes/gene families and four non-target gene families were identified from unique horseweed and waterhemp sequences (Table 5). In ten gene families, more resistance-gene candidates were identified in waterhemp than in horseweed. In the remaining five gene families, more resistance-gene candidates were identified in horseweed than in waterhemp. In short, some 430 unique sequences identified in this study might be involved in the evolution of herbicide resistance. This finding demonstrates the enormous value of 454 transcriptome sequencing for gene discovery in an important weedy plant with scant sequence data. In the following section, the utility of the horseweed transcriptome

 © 2010 Society of Chemical Industry

data is illustrated by exploring a non-target glyphosate resistance hypothesis in this species.[5] Specifically, the transcriptome data enabled expression analysis of ABC transporter genes based on real-time RT-PCR experiments.

3.4 Expression analysis of ABC transporter-like genes

ABC transporters are transmembrane proteins that utilize the energy of ATP hydrolysis to transport a wide variety of substrates across extra- and intracellular membranes, including metabolic products, lipids and sterols and drugs.[35,36] A number of ABC transporter genes were shown to be upregulated in previous microarray analysis by the present authors, suggesting that one or more might contribute to the glyphosate resistance in TN-R horseweed plants.[5] One model for non-target resistance is glyphosate sequestration into vacuoles via active transport of glyphosate by ABC transporters;[5,10,37,38] therefore, overexpression of ABC transporters could account for the resistance. In fact, some gene families that might be involved in glyphosate resistance in horseweed[39] were found in the dataset, which included ABC transporters, glutathione *S*-transferases (GSTs), glycosyltransferases and P450s (Table 5). In the case of ABC transporters, 67 unique sequences belonging to the subfamilies of multidrug resistance protein/multidrug resistance-associated protein (MRP) and pleiotropic drug resistance (PDR) were identified, mainly based on the annotation information of these sequences. Members of these subfamilies were shown to be upregulated at a high frequency by glyphosate in a previous heterologous microarray study.[5] However, multiple unique sequences may represent segments of one unigene, as there is no reference genome to assemble short reads into scaffolds. Therefore, the identification of 67 unique sequences does not mean that 67 different genes were identified in these subfamilies. Subsequently, a preliminary gene-by-gene transcription analysis was performed on 17 of the longer of these ABC-transporter genes (M1 to M11 from the MRP-like subfamily; P1 to P6 from the PDR-like subfamily). Only a few of these 17 genes could determine the closest relatives of the *Arabidopsis* genes, such as M11 to At3g13080. As a whole genome sequence dataset does not yet exist, it is still not clear how many unigenes these sequences represent, a common problem for many non-model organisms.[20,30,31] The most abundant transcript of these 17 ABC transporters, M11 (contig9470, 2120 bp of determined sequence), was found in glyphosate-treated TN-R horseweed plants and was 1.4 times higher than the expression of the actin housekeeping gene that was used as an internal control (Fig. 4). This AtMRP3-like ABC transporter was the highest upregulated gene, with an expression fold-change of 29.6, from a previous heterologous microarray study in horseweed.[5] Also, the identity of M11 with the 70 bp *Arabidopsis* probe sequence was ~90%. All other ABC transporter genes had much lower absolute abundance: M1, M2, M3, M8, M9 and P4 were among those with moderate abundance levels, while the others can be classified as low-abundance transcripts, but were still detectable by real-time RT-PCR (Fig. 4).

Compared with TN-S plants, M2 and P1 had lower expression levels in TN-R horseweed plants. M1 and M8 had the same expression level in both biotypes, while the remainder of ABC transporters had higher expression levels in TN-R horseweed plants. The responses of these ABC transporter genes to 24 h glyphosate treatment varied as shown in Fig. 5. M1, M2, M8, M9, M10, M11, P4 and P5 were shown to be upregulated in both TN-S and TN-R biotypes. However, M1 and M2 had higher expression levels in TN-S plants. M9, M10 and M11 had higher expression levels in TN-R plants. The expression levels of M8, P4 and P5 were comparable between the two biotypes. M3, M6, M7 and P3 were shown to be upregulated in TN-S horseweed, whereas there was almost no response in TN-R plants. M5 and P6 were shown to be upregulated in TN-S horseweed but downregulated in TN-R plants. P1 was downregulated in TN-S horseweed, whereas there was little response in TN-R plants. M4 and P2 had almost no response in both TN-S and TN-R biotype horseweed plants (Fig. 5). M6, M7, M10, M11 and P3 are more likely to be involved in the glyphosate resistance because their expression levels are always higher in resistant lines than in susceptible lines, and these are regarded as good preliminary target genes for further functional genomics studies.

The most interesting results were the strong responses of M10 and M11 transcription to glyphosate. M10 had a very low expression level, $\sim 6 \times 10^{-5}$ in TN-S and $\sim 1.2 \times 10^{-3}$ in TN-R plants,

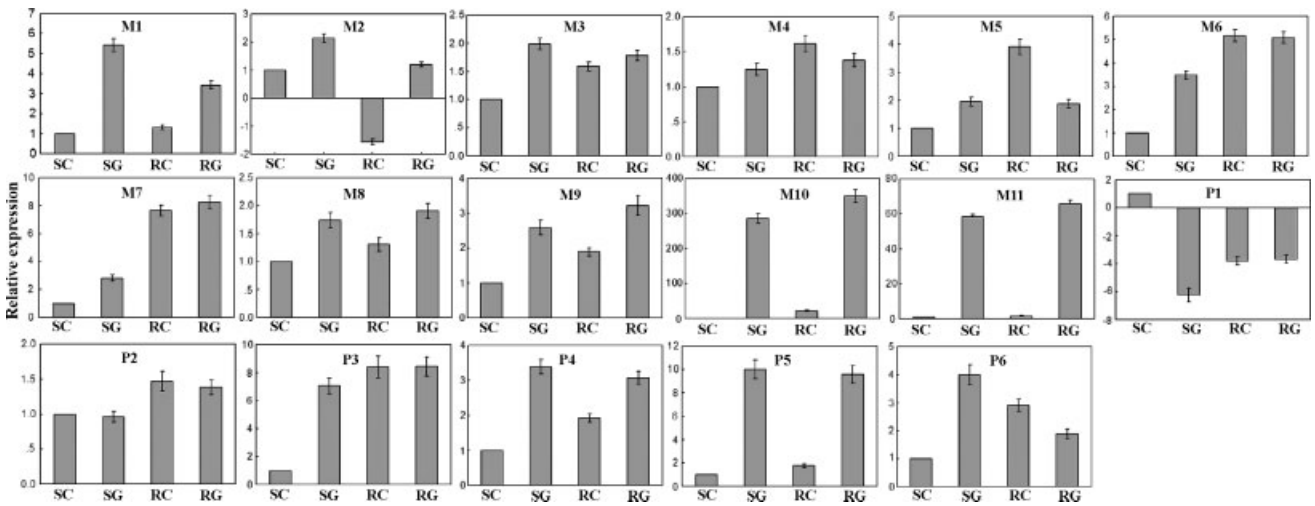

Figure 5. Relative expression profiles (compared with the expression level in TN-S control plants, SC) of 17 ABC transporter genes in young horseweed leaves from the following plant–treatment combinations: Tennessee susceptible–glyphosate sprayed, SG; Tennessee resistant–untreated control, RC; Tennessee resistant–glyphosate sprayed, RG. Data are presented as mean ± SE of three independent real-time RT-PCR analyses. Each RNA sample was isolated from leaves of six individual plants grown under the same conditions for each biotype and treatment and pooled to give one sample each.

 © 2010 Society of Chemical Industry

compared with the actin control gene. M10 was upregulated nearly 300-fold in treated TN-S plants, and 16-fold in treated TN-R plants, compared with their untreated controls, but TN-R plants had the highest expression level (Fig. 4). M11 was upregulated 60-fold and 45-fold in treated TN-S plants and TN-R plants respectively. Therefore, the promoter of this gene could be used to construct an efficient glyphosate reporter. Both M10 and M11 had a higher expression level in treated TN-R plants than in treated TN-S plants. There are several features of M11 that are intriguing with regards to a potential non-target glyphosate resistance candidate. These features include its high levels of absolute transcription, upregulation by glyphosate, which is also highest in resistant plants, and its putative tonoplast localization. Its ortholog in *Arabidopsis* is tonoplast targeted.[40] Thus, M11 could play a very important role in glyphosate transport into vacuoles, thereby resulting in the glyphosate resistance in TN-R horseweed. However, there is no direct evidence that M11 is able to transport glyphosate. It could not eliminate the possibility that the strong induction of the MRP3 homolog might be caused by secondary effects of the herbicide. One of the next steps is to clone the full-length M11, as well as other candidates, and perform functional analyses using overexpression analysis in susceptible and/or knockdown analysis in resistant horseweed biotypes. Transcriptome sequencing has crucially enabled translational research in understanding herbicide resistance mechanisms.

3.5 The significance of transcriptomics in weed science

Potential contributions of transcriptomics research in weeds include better understanding of weediness characteristics, the identification of new molecular targets for improved weed control and the gene discovery for transgenic crop improvement. However, the lack of good models and data hinders advances in genomics in weed biology. Although *Arabidopsis* is fully sequenced and commonly referred as 'the weed', it has few weediness characteristics and does not cause any significant economic loss in crops. Therefore, it is a poor model for weed genomics.[7,8,41] Given the diversity in weedy species, no single species can encompass all weedy traits. Also, among weediness features, herbicide resistance is arguably the most critical trait affecting current and long-term control of weeds in agriculture. Several candidates have been suggested as potential weedy models,[7,8] including herbicide-resistant horseweed and pigweeds, such as waterhemp. A 454 genomic DNA sequencing experiment on waterhemp produced 160 000 reads with an average read length of about 270 bp, yielding a total of about 43 Mb.[42] Subsequently, a 454 transcriptome sequencing project on waterhemp from the same laboratory generated 483 225 raw reads with an average read length of 231 bp, yielding a total of 111.8 Mb,[26] which is comparable with the present study. Horseweed is one of the most attractive weeds for whole-genome sequencing because it has the smallest genome among 25 surveyed weeds most prevalent in the weed science literature.[8] A GS-FLX genomic test run (half-plate) with titanium reagents produced ~600 000 raw reads with an average read length of 403 bp, for a yield total of 248.7 Mb (Y. Peng and C. N. Stewart Jr., unpublished data), which included essentially the whole chloroplast genome (~150 kb). Other weed genomics research has been reported recently. For example, 23 000 unique leafy spurge EST sequences[43] and about 9000 unique cassava EST sequences[44] were obtained through traditional Sanger sequencing. More recently, the Roche GS-FLX 454 sequencing platform has been employed to sequence normalized cDNAs from ten native and ten invasive yellow star-thistle (*Centaurea solstitialis* L.) genotypes (Dlugosch K, Barker M, Lai Z and Rieseberg L, private communication) in the Compositae Genome Project (http://compgenomics.ucdavis.edu/compositae_index.php). An average of 89 000 reads, ~200 bp long, and 32 000 unigenes were obtained per genotype. Compared with these examples, large-scale sequencing of waterhemp and horseweed using the 454 platform yielded abundant data in a short period of time. This suggests that the development of weedy genomic research is largely being enabled by powerful next-generation sequencing platforms.

4 CONCLUSIONS

Young leaves and meristematic tissues from bulked samples of two horseweed biotypes, TN-S and TN-R, with and without glyphosate treatment, were used to carry out a large-scale transcriptome sequencing project using GS-FLX 454 sequencing, *de novo* assemblage and functional annotation of the sequence data. These data were also used to design specific primers and measure the expression of potential candidate genes that might be involved in conferring non-target glyphosate resistance in horseweed. Moreover, the data are sufficient to allow the design of microarray oligonucleotide probes and SNP discovery. The data also enable full-length cDNA cloning of non-target candidate genes via a RACE protocol for the next step in functional genomics research. Because of its very small genome size, horseweed could serve not only as a useful species for identifying non-target herbicide resistance mechanisms[10,39] but also as a good model for weed genomics.[7,8] Sufficient data provided by this large-scale sequencing, coupled with further application of multigenomic tools, will improve understanding of the genetic basis of weediness characteristics and the evolution mechanisms of herbicide resistance in weeds. In the long term, this research should be helpful in weed management and control.

ACKNOWLEDGEMENTS

The authors thank the WM Keck Center for Comparative and Functional Genomics (University of Illinois) for carrying out the 454 sequencing. They also thank the Complex Computation Lab (Iowa State University) for assistance with sequence assembling. The research was funded by the University of Tennessee AgResearch, the Ivan Racheff Endowment, a special USDA-CSREES grant to CNS and by Monsanto.

SUPPORTING INFORMATION

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

REFERENCES

1 Dill GM, Cajacob CA and Padgette SR, Glyphosate-resistant crops: adoption, use and future consideration. *Pest Manag Sci* **64**:326–331 (2008).
2 Duke SO and Powles SB, Glyphosate: a once-in-a-century herbicide. *Pest Manag Sci* **64**:319–325 (2008).
3 Heap I, *The International Survey of Herbicide Resistant Weeds.* [Online]. Available: www.weedscience.com [14 July 2010].
4 Van Gessel MJ, Glyphosate-resistant horseweed from Delaware. *Weed Sci* **49**:703–705 (2001).
5 Yuan JS, Abercrombie LG, Cao Y, Halfhill MD, Zhou X, Peng Y *et al*, Functional genomics analysis of glyphosate resistance in *Conyza canadensis* (horseweed). *Weed Sci* **58**:109–117 (2010).

 © 2010 Society of Chemical Industry

6 Weaver SE, The biology of Canadian weeds. 115. *Conyza canadensis*. *Can. J. Plant Sci* **81**:867–875 (2001).
7 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).
8 Stewart CN, 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).
9 Halfhill MD, Good LL, Basu C, Burris J, Main CL, Mueller TC *et al*, Transformation and segregation of GFP fluorescence and glyphosate resistance in horseweed (*Conyza canadensis*) hybrids. *Plant Cell Rep* **26**:303–311 (2007).
10 Feng PCC, Tran M, Chiu T, Sammons RD, Heck GR and Cajacob CA, Investigations into glyphosate resistant horseweed (*Conyza canadensis*): retention, uptake, translocation, and metabolism. *Weed Sci* **52**:498–505 (2004).
11 Zelaya IA, Owen MDK and VanGessel MJ, Inheritance of evolved glyphosate resistance in *Conyza canadensis* (L.) Cronq. *Theor Appl Genet* **110**:58–57 (2004).
12 Margulies M, Egholm E, Altman WE, Attiya S, Bader JS, Bemben LA *et al*, Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376–380 (2005).
13 Morozova O and Marra MA, Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**:255–264 (2008).
14 Rothberg JM and Leamon JH, The development and impact of 454 sequencing. *Nat Biotechnol* **26**:1117–1124 (2008).
15 Emrich SJ, Barbazuk WB, Li L and Schnable PS, Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17**:69–73 (2007).
16 Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF *et al*, Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**:420–426 (2007).
17 Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I *et al*, High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. *Nat Genet* **40**:987–993 (2008).
18 Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I *et al*, Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**:1636–1647 (2008).
19 Mao C, Evans C, Jensen RV and Sobral BW, Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system. *BMC Microbiol* **8**:72 (2008).
20 Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J *et al*, De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. *BMC Genomics* **11**:262 (2010).
21 Droege M and Hill B, The genome sequencer FLX™ system – longer reads, more applications, straight forward bioinformatics and more complete datasets. *J Biotech* **136**:3–10 (2008).
22 Nyren P, Pettersson B and Uhlen M, Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal Biochem* **208**:171–175 (1993).
23 Ronaghi M, Karamohamed S, Pettersson B, Uhlen M and Nyren P, Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**:84–89 (1996).
24 Mueller TC, Massey JH, Hayes RM, Main CL and Stewart CN, Jr, Shikimate accumulates in both glyphosate-sensitive and glyphosate-resistant horseweed (*Conyza canadensis* L. Cronq.). *J Agric Food Chem* **51**:680–684 (2003).
25 Dassanayake M, Haas JS, Bohnert HJ and Cheeseman JM, Shedding light on an extremophile lifestyle through transcriptomics. *New Phytol* **183**:764–775 (2009).
26 Riggins CW, Peng Y, Stewart CN, Jr and Tranel PJ, Characterization of waterhemp transcriptome using 454 pyrosequencing and its application for studies of herbicide target-site genes. *Pest Manag Sci* (accepted).
27 Chou HH and Holmes MH, DNA sequencing quality trimming and vector removal. *Bioinformatics* **17**:1093–1104 (2001).
28 Huang X and Madan A, CAP3: a DNA sequence assembly program. *Genome Res* **9**:868–877 (1999).
29 Yuan JS, Wang D and Stewart CN, Statistical methods for efficiency adjusted real-time PCR analysis. *Biotechnol J* **3**:112–123 (2008).
30 Alagna F, D'Agostino N, Torchia L, Servili M, Rao R, Pietrella M *et al*, Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. *BMC Genomics* **10**:399–xxx (2009).
31 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–xxx (2009).
32 Weber AP, Weber KL, Carr K, Wilkerson C and Ohlrogge JB, Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144**:32–42 (2007).
33 Black DL, Mechanisms of alternative pre-messenger RNA splicing. *Ann Rev Biochem* **72**:291–336 (2003).
34 Yuan Y, Chung JD, Fu X, Johnson VE, Ranjan P, Booth SL *et al*, Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in *Populus* and *Arabidopsis*. *Proc Natl Acad Sci USA* **106**:22020–22025 (2009).
35 Rea PA, Plant ATP-binding cassette transporters. *Ann Rev Plant Biol* **58**:347–375 (2007).
36 Verrier PJ, Bird D, Burla B, Dassa E, Forestier C, Geisler M *et al*, Plant ABC proteins – a unified nomenclature and updated inventory. *Trends Plant Sci* **13**:151–159 (2008).
37 Shaner DL, The role of translocation as a mechanism of resistance to glyphosate. *Weed Sci* **57**:118–123 (2009).
38 Ge X, d'Avignon DA, Ackerman JJH and Sammons RD, Rapid vacuolar sequestration: the horseweed glyphosate resistance mechanism. *Pest Manag Sci* **66**:345–348 (2010).
39 Yuan JS, Tranel PJ and Stewart CN, Jr, Non-target site herbicide resistance: a family business. *Trends Plant Sci* **12**:6–13 (2007).
40 Dunkley TP, Hester S, Shadforth IP, Runions J, Weimar T, Hanton SL *et al*, Mapping the Arabidopsis organelle proteome. *Proc Natl Acad Sci USA* **103**:6518–6523 (2006).
41 Gressel J, Arabidopsis is not a weed, and mostly not a good model for weed genomics; there is no good model for weed genomics, in *Weedy and Invasive Plant Genomics*, ed. by Stewart CN, Jr. Wiley-Blackwell, Ames, IA, pp. 25–32 (2009).
42 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).
43 Anderson JV, Horvath DP, Chao WS, Foley ME, Hernandez AG, Thimmapuram J *et al*, Characterization of an EST database for the perennial weed leafy spurge: an important resource for weed biology research. *Weed Sci* **55**:193–203 (2007).
44 Lokko Y, Anderson JV, Rudd S, Raji AAJ, Horvath D, Mikel MA *et al*, Characterization of an 18,166 EST dataset for cassava (*Manihot esculenta* Crantz) enriched for drought-responsive genes. *Plant Cell Rep* **26**:1605–1618 (2007).

© 2010 Society of Chemical Industry