De Novo Genome Assembly of the Economically Important Weed Horseweed Using Integrated Data from Multiple Sequencing Platforms


Department of Plant Science, University of Tennessee, Knoxville, Tennessee 37996 (Y.P., T.L., M.N.-R., C.N.S.); Department of Biology, Indiana University, Bloomington, Indiana 47405 (Z.L., L.H.R.); Department of Plant Sciences (M.O., M.J.) and Genome Center (H.O., R.W.K.), University of California, Davis, California 95616; Monsanto, Inc., St. Louis, Missouri 63130 (R.D.S.); and Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4 (L.H.R.)

Horseweed (Conyza canadensis), a member of the Compositae (Asteraceae) family, was the first broadleaf weed to evolve resistance to glyphosate. Horseweed, one of the most problematic weeds in the world, is a true diploid (2n = 2x = 18), with the smallest genome of any known agricultural weed (335 Mb). Thus, it is an appropriate candidate to help us understand the genetic and genomic bases of weediness. We undertook a draft de novo genome assembly of horseweed by combining data from multiple sequencing platforms (454 GS-FLX, Illumina HiSeq 2000, and PacBio RS) using various libraries with different insertion sizes (approximately 350 bp, 600 bp, 3 kb, and 10 kb) of a Tennessee-accessed, glyphosate-resistant horseweed biotype. From 116.3 Gb (approximately 350× coverage) of data, the genome was assembled into 13,966 scaffolds with 50% of the assembly = 33,561 bp. The assembly covered 92.3% of the genome, including the complete chloroplast genome (approximately 153 kb) and a nearly complete mitochondrial genome (approximately 450 kb in 120 scaffolds). The nuclear genome is composed of 44,592 protein-coding genes. Genome resequencing of seven additional horseweed biotypes was performed. These sequence data were assembled and used to analyze genome variation. Simple sequence repeat and single-nucleotide polymorphisms were surveyed. Genomic patterns were detected that associated with glyphosate-resistant or -susceptible biotypes. The draft genome will be useful to better understand weediness and the evolution of herbicide resistance and to devise new management strategies. The genome will also be useful as another reference genome in the Compositae. To our knowledge, this article represents the first published draft genome of an agricultural weed.

In the past few years, genomics approaches have revolutionized plant biology. Complete or draft genome data are currently available for tens of plant species (http://www.phytozome.net). From the model plant Arabidopsis (Arabidopsis thaliana; Arabidopsis Genome Initiative, 2000) to important crop plants such as rice (Oryza sativa; Goff et al., 2002; Yu et al., 2002; International Rice Genome Sequencing Project, 2005), soybean (Glycine max; Schmutz et al., 2010), maize (Zea mays; Schnable et al., 2009), and chickpea (Cicer arietinum; Varshney et al., 2013) to economic woody plants such as wine grape (Vitis vinifera; Jaillon et al., 2007) and poplar (Populus trichocarpa; Tuskan et al., 2006), powerful complete genome data sets and tools allow the unprecedented ability to explore the genetics and genomics of plant form and function. Furthermore, genome sequences of additional plant species will soon be available from in-progress large-scale sequencing projects (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html).

One critical group of plants that has been largely overlooked in the genomics revolution consists of economically significant agricultural weeds (Stewart, 2009; Stewart et al., 2009). Weeds cause about $36 billion annual damage in the United States alone (Pimentel et al., 2000). The cost is even higher if one includes weeds in pastures, golf courses, aquatic environments, etc. Unchecked, weeds effectively outcompete crops for resources and decrease the yield of food, feed, and fiber. Despite their profound economic significance, weedy plant genomics data are very scarce relative to other economically important plants.

Weeds experienced surreptitious accidental domestication (Warwick and Stewart, 2005) and adaptation to changing agricultural environments to become among the world’s most successful plants. They are persistent,
adaptable, stress tolerant, competitive, and capable of extreme levels of reproduction (e.g. some species can produce one million seeds per plant). Herbicide application has been the dominant weed-control measure in recent decades, but increasingly, the evolution of herbicide resistance is challenging this practice. More than 400 biotypes of weeds have evolved resistance to one or more of all the major groups of herbicides (Heap, 2014), among which resistance to glyphosate is currently of greatest concern. Horseweed (Conyza canadensis), which is in the Compositae (Asteraceae) family, was the first eudicot weedy species to evolve glyphosate resistance and is also one of the most widely distributed glyphosate-resistant species in the world (VanGessel, 2001; Heap, 2014). Horseweed is a true diploid (2n = 18), with a self-fertilizing mating system and the smallest genome of any agricultural weed (335 Mb; Stewart et al., 2009). We sought to describe the reference genome of horseweed in order to enable genomic approaches to elucidate the mechanism of nontarget herbicide resistance and to gain a better understanding of the evolution and spread of herbicide resistance across agricultural landscapes. Weed genomics is critical to understand weed biology, and understanding weed biology is critical in weed management. To date, transcriptomes using next-generation sequencing of a few weedy species have been produced to aid the discovery of genes that contribute to herbicide resistance and other weediness traits, with most progress occurring in Conyza and Amaranthus spp. and Lolium rigidum (Lee et al., 2009; Peng et al., 2010; Riggins et al., 2010; Lai et al., 2012; Gaines et al., 2014).

Horseweed shares many other weediness features with the world’s most damaging weeds. These features include a relatively short vegetative phase, indeterminate flowering, self-compatibility, high seed output, long-distance seed dispersal, competitiveness with crop plants, a deep root system, discontinuous dormancy, environmental plasticity, and allelopathy (Basu et al., 2004; Chao et al., 2005). Also, horseweed is amenable to genetic transformation and regeneration (Halfhill et al., 2007), which facilitates overexpression or knockdown analysis of potential gene targets. Autogamous, a single horseweed plant may produce well over 200,000 seeds. The small seeds have pappi that enable wind and water dispersal (Bhowmik and Bekech, 1993; Weaver 2001) to sometimes very long distances (DeVlaming and Proctor, 1968; Weaver and McWilliams, 1980; Andersen, 1993; Dauer et al., 2006; Shields et al., 2006). Moreover, horseweed seeds are not dormant and can germinate immediately after maturity in the fall, spring, or midsummer (Bhowmik and Bekech, 1993; Weaver, 2001). Thus, horseweed is a bona fide weed with model plant qualities.

The Compositae is the largest and most diverse plant family, with over 24,000 described species. Despite the evolutionary success and economic importance of plants in this family, a draft whole-genome publication for any member of the family or for any plants from closely related families is lacking. However, genomes for lettuce (Lactuca sativa; https://1gr.genomecenter.ucdavis.edu), sunflower (Helianthus annuus; http://www.sunflowergenome.org), and several other members of the family are available, but de novo genome assembly can often be arduous in many cases when genomes are large and complex (Kane et al., 2011; Truco et al., 2013). A draft sequence of a streamlined Compositae genome would greatly advance the genomics and genetics of economically, ecologically, and evolutionarily important plants within this important plant family.

In this study, in order to further increase sequence resources for horseweed, we performed de novo genome sequencing and assembly of the Tennessee glyphosate-resistant horseweed biotype TN-R using multiple sequencing platforms, including 454 GS-FLX, Illumina HiSeq 2000 and PacBio RS. Using a modified strategy that has proven effective in vertebrates (Li et al., 2010a), date palm (Phoenix dactylifera; Al-Dous et al., 2011), and flax (Linum usitatissimum; Wang et al., 2012), the draft genome of horseweed was assembled into scaffolds. Genomes and transcriptomes of this and another seven horseweed biotypes were also sequenced, and the data were used to investigate genome variation and to facilitate molecular marker development and gene discovery. The primary goal of this project was to build substantial genomic resources for horseweed, which will subsequently be useful to elucidate the genomic basis of weediness traits.

RESULTS
De Novo Sequencing and Assembly

The reference genome source was DNA extracted from six individual plants of a single glyphosate-resistant biotype that was originally collected in 2002 from a soybean field in Jackson, TN (TN-R; Mueller et al., 2003). Comprehensive genome shotgun sequences were obtained using seven libraries with insert sizes that ranged from 350 bp to 10 kb and three sequencing platforms (454 GS-FLX, Illumina HiSeq 2000, and PacBio RS; Table I). Two plates of 454 GS-FLX runs on three independent libraries (approximately 600 bp) produced 2,238,807 raw reads with an average read length of approximately 390 bp, for a yield total of approximately 860 Mb data. Two lanes of HiSeq 2000 runs (2 × 100 bp) of two short paired-end libraries produced 786,389,990 raw reads, for a yield total of approximately 77 Gb data. One lane of a HiSeq 2000 run (2 × 100 bp) of one 3-kb mate-paired library produced 381,844,926 raw reads, for a yield total of approximately 37 Gb data. Ten SMRT cells of PacBio RS runs produced 513,084 sequences with an average read length of 3.1 kb, which yielded 1.44 Gb of total data (Table II). All the sequence data together represented approximately 350× coverage of the horseweed genome.

De novo genome assembly is often not straightforward (Baker, 2012), in that many factors impact the integrity and accuracy of the draft assembly. To integrate data from multiple sequencing platforms, we used a de novo assembly strategy (Supplemental Fig.
Table I. Summary of sequencing read yield results from multiple platforms and mate-paired libraries of various sizes (approximately 350 bp, 600 bp, 3 kb, and 10 kb)

<table>
<thead>
<tr>
<th>Platform</th>
<th>Libraries</th>
<th>Read Length</th>
<th>Filtered Reads</th>
<th>Data (Approximate Coverage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 GS-FLX</td>
<td>600 bp (3)</td>
<td>390 bp</td>
<td>2,238,807</td>
<td>860 Mb (2.5×)</td>
</tr>
<tr>
<td>Illumina PE</td>
<td>350 bp (2)</td>
<td>2 × 100 bp</td>
<td>786,389,990</td>
<td>77 Gb (230×)</td>
</tr>
<tr>
<td>Illumina MP</td>
<td>3 kb (1)</td>
<td>2 × 100 bp</td>
<td>381,844,926</td>
<td>37 Gb (110×)</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>10 kb (1)</td>
<td>3.1 kb</td>
<td>513,084</td>
<td>1.4 Gb (4.5×)</td>
</tr>
<tr>
<td>Total</td>
<td>Seven libraries</td>
<td>0.1–10 kb</td>
<td>1,170,986,807</td>
<td>116.3 Gb (350×)</td>
</tr>
</tbody>
</table>

Table II. Summary of the de novo genome assembly of horseweed

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Contigs</th>
<th>Scaffolds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence number</td>
<td>20,075</td>
<td>13,966</td>
</tr>
<tr>
<td>N50 (bp; n)</td>
<td>8,698; 14,359</td>
<td>13,506; 10,318</td>
</tr>
<tr>
<td>N70 (bp; n)</td>
<td>14,317; 8,886</td>
<td>23,243; 6,219</td>
</tr>
<tr>
<td>N90 (bp; n)</td>
<td>20,764; 5,122</td>
<td>33,561; 3,575</td>
</tr>
<tr>
<td>Average length (bp)</td>
<td>16,258</td>
<td>26,546</td>
</tr>
<tr>
<td>Maximum single sequence (bp)</td>
<td>102,072</td>
<td>182,395</td>
</tr>
<tr>
<td>Base pairs (Mb)</td>
<td>311.27</td>
<td>344.88</td>
</tr>
</tbody>
</table>

base ratio, which was higher than sequenced woody eudicot genomes but equivalent to other eudicot genomes and much lower than monocot genomes (Fig. 1C; Supplemental Table S2). The G/C content of coding regions was 38.6%, which was higher than that of the entire genome (Fig. 1D). The assembled draft of horseweed genome data has been submitted to the National Center for Biotechnology Information with accession number SUB535309.

To inspect the de novo assembly accuracy, multiple independent sources of horseweed DNA and complementary DNA sequences were used to align with the assembled genome (Table III). These sequences included EST reads from Sanger sequencing, transcriptome shotgun reads from 454 and Illumina sequencing, and partial genomic reads, which were used for the genome assembly (Zhou et al., 2009; Peng et al., 2010; Yuan et al., 2010; sample identifiers WSYE and NUSE from the One-Thousand Plant Transcriptomes project [www.onekp.com]). Over 90% of sequences could be aligned to the genome. The mapping and BLAST varied among sequence reads. However, in the aligned sequences, 91.2% or more had over 90% coverage and identity. The non-perfect alignments indicated that EST reads could be used to further improve assembly accuracy and gene prediction.

Identification of Repetitive Elements and Annotation of Protein-Coding Genes

By using the plant repeat database (RepeatMasker libraries 20140131) and RepeatMasker 4.0.5, a total of 233,521 loci (19.5 Mb) were identified to be various nucleotide repeat elements. These sequences represented 6.25% of the assembled genome (Table IV) and were composed of 172,825 simple sequence repeats (SSRs),
28,494 low-complexity elements, 23,425 different types of retroelements, 5,424 small RNA structures, 2,283 DNA transposons, and 675 unclassified elements. Retroelements were the largest mobile elements found in the genome (2.61%), of which most were long terminal repeat-type retroelements (2.54%). Small RNAs contributed 0.61% to the genome, whereas identified DNA transposons represented just 0.16% of the genome.

Masked SSR loci were analyzed using the Simple Sequence Repeat Identification Tool (http://archive.gramene.org/db/markers/ssrtool) with a threshold minimum of six repeats. The results included 51,892 loci (Supplemental Table S3) containing 44,589 dimer (85.9%), 6,864 trimer (13.2%), 299 tetramer (0.58%), 60 pentamer (0.12%), and 80 hexamer (0.15%) motifs. The majority of these loci contained six to nine repeats (Supplemental Fig. S7). Considering the position in scaffold, repeat number, and length, 1,316 loci were identified to have appropriate sequences to develop PCR primers, which could be used as SSR markers (Okada et al., 2013).

Using previous EST and transcriptome data as a starting point (Zhou et al., 2009; Peng et al., 2010; Yuan et al., 2010), 44,592 horseweed gene models were predicted. A total of 25,439 unique transcripts were returned after removing redundancies. The longest gene (using coding sequence only and without tallying introns) was composed of 17,120 bp. It was annotated to encode a midasin-like protein. The mean and median transcript lengths were 2,210 and 1,830 bp, respectively. Over 95% of the transcripts were longer than 715 bp (Fig. 2A). The predicted gene sequences were further compared with the UniProt and National Center for Biotechnology Information nonredundant protein databases for assigning biological information. Using a threshold e value of $10^{-4}$ or less, 79.6% of these

<table>
<thead>
<tr>
<th>Sequence Type</th>
<th>Sequences</th>
<th>Aligned</th>
<th>Alignment Coverage</th>
<th>Alignment Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger ESTs (more than 600 bp)</td>
<td>5,482</td>
<td>5,012</td>
<td>91.2</td>
<td>93.7</td>
</tr>
<tr>
<td>454 GS-FLX RNAseq (more than 200 bp)</td>
<td>145,796</td>
<td>132,579</td>
<td>90.9</td>
<td>92.6</td>
</tr>
<tr>
<td>Illumina RNAseq (100 bp)</td>
<td>3,987,575</td>
<td>3,591,820</td>
<td>90.1</td>
<td>91.6</td>
</tr>
<tr>
<td>454 GS-FLX genomic reads (more than 390 bp)</td>
<td>499,955</td>
<td>450,754</td>
<td>90.2</td>
<td>95</td>
</tr>
<tr>
<td>Illumina genomic reads (100 bp)</td>
<td>9,315,014</td>
<td>8,476,173</td>
<td>91.0</td>
<td>95</td>
</tr>
</tbody>
</table>
returned hits at e values of $10^{-10}$ or greater (Fig. 2C). Within the annotated genes, 18,406 (20,248 of 25,439) had at least one hit to the database. The majority of the alignments had an identity of 50% (20,248 of 25,439) had at least one hit to the database. Moreover, most of the hits had only one high-scoring segment pair with the aligned genes (Fig. 2, B and D). Overall, the majority of the predicted genes in the assembled horseweed genome could be aligned with hits in a known protein database with high identity and specificity, which indicated that most of these genes were likely protein-coding sequences.

Gene Ontology (GO) annotation resulted in 14,897 GO-annotated genes. The largest gene category with molecular function was kinase activity (1,780), following by nucleotide binding (1,740), transporter activity (1,031), and another 15 subgroups with 50 or more terms (Supplemental Fig. S8A). Among genes annotated to various biological processes, the largest subgroup was cellular component (4,582), followed by metabolic process (3,799), and 27 additional groups with 50 or more terms (Supplemental Fig. S8B). The potential cellular localization of the genes was predicted by assigning 17,878 cellular component GO terms to various queries (Supplemental Fig. S8C). Genes that encoded proteins with enzyme activity were further divided into six groups (transferases, hydrolases, oxidoreductases, ligases, lyases, and isomerases; Supplemental Fig. S8D).

Fisher’s exact test showed 10 GO terms that were significantly overrepresented in horseweed compared with Arabidopsis (Fig. 3). Enrichment analysis using the GO terms as well as related gene families included vacuolar transport, photosynthetic acclimation, detoxification of nitrogenous compounds, response to UV-B light, chloroplast thylakoid lumen, protein targeting to chloroplasts, protein peptidyl-prolyl isomerization, peptidyl-prolyl cis-trans-isomerase activity, hydrolase activity (acting on carbon-nitrogen), glycolysis, and NAD(P)H dehydrogenase complex assembly. Four transporter subgroup GO terms were overrepresented in the horseweed genome at the $P < 0.001$ level: drug transmembrane transporters (GO:0015238), xanthine transmembrane transporters (GO:0042907), uracil transmembrane transporters (GO:0015210), and allantoin uptake transporters (GO:0005274; Supplemental Fig. S9). However, these data are not sufficient to draw conclusions about biological function and the evolution of herbicide resistance; further gene expression analysis and other follow-on experiments need to be performed.

Gene families that are commonly associated with nontarget herbicide resistance include cytochrome P450s, glutathione S-transferases (GSTs), glycosyltransferases (GTs), and ATP-binding cassette (ABC) transporters (Yuan at al., 2007). Relative to Arabidopsis, the horseweed genome has more members in each of these families (Table V). There were 323 unique cytochrome P450 genes at 401 loci in horseweed, which represents a 26% increase over Arabidopsis. Also, 155 ABC transporter genes at 213 loci were found in horseweed, which is 14% more than in Arabidopsis. Horseweed had 6% more GTs than Arabidopsis. There were 54 horseweed GSTs, which is one more gene than is found in Arabidopsis (Table V).

### Table IV. Repeat element analysis in the horseweed genome

For elements, most repeats were fragmented by insertions or deletions and have been counted as one element.

<table>
<thead>
<tr>
<th>Repeat Elements</th>
<th>Elements</th>
<th>Length</th>
<th>Percentage of Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retroelements</td>
<td>23,425</td>
<td>8,509,514</td>
<td>2.61</td>
</tr>
<tr>
<td>Short interspersed elements</td>
<td>29</td>
<td>1,433</td>
<td>0.00</td>
</tr>
<tr>
<td>Long interspersed elements</td>
<td>902</td>
<td>219,700</td>
<td>0.07</td>
</tr>
<tr>
<td>Long terminal repeat elements</td>
<td>22,889</td>
<td>8,288,381</td>
<td>2.54</td>
</tr>
<tr>
<td>Yeast transposon</td>
<td>15,599</td>
<td>4,999,694</td>
<td>1.54</td>
</tr>
<tr>
<td>family1/Copia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gypsy/Dictyostelium</td>
<td>7,110</td>
<td>3,276,883</td>
<td>1.00</td>
</tr>
<tr>
<td>intermediate repeat</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sequence1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA transposons</td>
<td>2,283</td>
<td>531,765</td>
<td>0.16</td>
</tr>
<tr>
<td>Hobo-Activator</td>
<td>337</td>
<td>61,039</td>
<td>0.02</td>
</tr>
<tr>
<td>Tcl-1-B6s0-Pogo</td>
<td>719</td>
<td>166,937</td>
<td>0.05</td>
</tr>
<tr>
<td>Enhancer supressor mutator</td>
<td>2</td>
<td>644</td>
<td>0.00</td>
</tr>
<tr>
<td>Tourist/Harbinger</td>
<td>418</td>
<td>137,376</td>
<td>0.04</td>
</tr>
<tr>
<td>Unclassified</td>
<td>675</td>
<td>278,725</td>
<td>0.09</td>
</tr>
<tr>
<td>Small RNAs</td>
<td>5,424</td>
<td>2,003,521</td>
<td>0.61</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>172,825</td>
<td>7,669,927</td>
<td>2.35</td>
</tr>
<tr>
<td>Low complexity</td>
<td>28,494</td>
<td>1,400,133</td>
<td>0.43</td>
</tr>
<tr>
<td>Total</td>
<td>233,521</td>
<td>20,393,585</td>
<td>6.25</td>
</tr>
</tbody>
</table>

*Plastid and Mitochondrial Genomes*

Plastid genome sequences from 454 GS-FLX reads were isolated by searching against known plastid genome databases and then subjected to de novo assembly. The assembled plastid genomes were further mapped with accurate Illumina reads to fix homopolymer errors. The entire chloroplast genome (approximately 153 kb; Supplemental Fig. S10) and most of the mitochondrial genome (approximately 450 kb in scaffolds) were obtained (Supplemental Table S4; Supplemental Data Sets S1 and S2). The horseweed chloroplast genome contained two inverted repeats (24,936 bp each), a large single-copy fragment (84,634 bp), and a small single-copy fragment (18,063 bp). The orientation of the two inverted repeats was determined by PCR. A total of 95 protein-coding genes (88 unique), 39 tRNA genes (28 unique), and eight ribosomal RNA genes (four unique) were annotated (Supplemental Data Set S1). These genes comprised 59.7% of the chloroplast genome. The remainder, the noncoding portion of the chloroplast genome, was composed of introns, intergenic spacers, and pseudogenes. The largest gene was theycf2 gene, which was 6,085 bp. The smallest gene was a 34-bp tRNA gene. The G/C content in the chloroplast genome was 37.2%. The chloroplast genome was nearly identical to that of lettuce.
and sunflower, except in the orientation of the two inverted repeat elements (Supplemental Fig. S11). The lettuce chloroplast has reverse inverted repeat elements compared with horseweed and sunflower.

The horseweed mitochondrial genome reads from 454 GS-FLX, Illumina paired-end, and true-mate pairs were parsed and assembled into 123 scaffolds (N50 = 10,057 bp), with the help of plant mitochondrial databases. The scaffold sizes ranged from 315 to 43,498 bp, for a total of 453,334 bp in the mitochondrial genome (Supplemental Fig. S12; Supplemental Data Set S2), which was moderately sized compared with other plant mitochondrial genomes.

**Genome Resequencing and Genomic Variation Analysis**

We resequenced four population pairs that were either glyphosate resistant (R) or susceptible (S) from the United States. The representative populations used for resequencing were from California (CA-R versus CA-S), Delaware (DE-R versus DE-S), Indiana (IN-R versus IN-S), and Tennessee (TN-R versus TN-S). Individual seedlings surviving a glyphosate application were used for eventual DNA donors, wherein DNA was pooled for sequencing (Table VI; Supplemental Fig. S13). Within two HiSeq 2000 flow cell lanes, a total 55.6 Gb (170× coverage) of paired-end reads was produced, and each of the seven additional genomes was assembled using TN-R as the reference genome (Table VI). Interpopulation genome variation included multiple-nucleotide variation, which was defined as 4 bp or less, with equal numbers of nucleotides at each locus, single-nucleotide variation, insertions/deletions, and replacement (Table VII). When a minimum cutoff value (20 times or greater coverage at the variant loci, 20% or greater frequency) was applied to avoid calling errors, the frequency of detected variants in the genome varied among biotypes, ranging from 0.75 to 1.59 counts per kb (mean of 1.01; Table VII). Compared with the TN-R reference genome, CA-R had the fewest variants, followed by, in ascending order, IN-R, CA-S, IN-S, DE-R, TN-S, and DE-S. IN-R, IN-S, and TN-S had fewer variants relative to each other compared with the rest of the other biotypes. Compared with their
partners (DE-S and CA-S, respectively), DE-R and CA-R had the fewest variants. Similarly, IN-R had fewer variants than IN-S (Supplemental Table S5). The minimum average quality score for the position of each variant was 30 (one error in 1,000; Fig. 4A), which ensured that the results were quite reliable. To determine whether a variant was homozygous, we applied a threshold of 90% probability and found that IN-R and CA-R had the highest frequency of heterozygosity (nearly 80%), whereas only 50% of variants were heterozygous in the DE-R, DE-S, and TN-S biotypes (Fig. 4B). Most of these sequence variations among samples was composed of single-nucleotide variations/single-nucleotide polymorphisms (SNPs) in all eight biotypes (Fig. 4C). Fisher's exact test showed that a total of 2,010 specific variants were found in all four glyphosate-resistant biotypes, which were distributed among 1,370 loci. Only 539 specific variants were found in all four glyphosate-susceptible biotypes, and these were located at 425 loci (Fig. 4D). Since the physical map of the horseweed genome was not available, the location of each specific variant in the sorted scaffold arrays instead of in the genome was listed (Supplemental Fig. S14). A genome-wide association study requires SNP information from the genomes of many control and test individuals. Specific SNPs in glyphosate-resistant and glyphosate-susceptible genomes were located in hotspots, which were defined as the ones with low P values ($P \leq 0.01$), and might be associated with the phenotypes of glyphosate resistance (Supplemental Fig. S14). However, these findings are not definitive, given the low sample sizes of biotypes analyzed.

Phylogenetic analysis of whole-genome SNP variation provided a phylogeographic hypothesis regarding the evolution of glyphosate resistance and its spread (Fig. 5). Biotypes from proximate locations generally shared the same clades, such as the Delaware pairs and the Indiana pairs, which is consistent with our previous studies (Yuan et al., 2010; Okada et al., 2013).

Table V. Analysis of gene families (cytochrome P450s, GSTs, ABC transporters, and GTs) that are potentially involved in nontarget glyphosate resistance in horseweed compared with those tallied for Arabidopsis

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>Arabidopsis</th>
<th>Horseweed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytochrome P450s</td>
<td>256</td>
<td>323 (401 Loci)</td>
</tr>
<tr>
<td>GSTs</td>
<td>53</td>
<td>54 (54 Loci)</td>
</tr>
<tr>
<td>ABC transporters</td>
<td>136</td>
<td>155 (213 Loci)</td>
</tr>
<tr>
<td>GTs</td>
<td>361</td>
<td>381 (457 Loci)</td>
</tr>
</tbody>
</table>
Although not in the same clade, the TN-R branch was proximate to TN-S. The position of CA-S on the phylogram was basal, which might be of biological importance with regard to gene flow in the species. The phylogenetic pattern suggests that glyphosate resistance in horseweed has evolved independently multiple times (Fig. 5).

5-Enolpyruvylshikimate-3-Phosphate Synthase Genes

In the horseweed genome, there are two 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) gene copies (EPSPS1 and EPSPS2); thus, no gene amplification was observed. Eleven EPSPS1 and 32 EPSPS2 variants were detected (Supplemental Fig. S15). However, no EPSPS mutation cosegregated with glyphosate resistance in our data. Furthermore, none of the observed EPSPS sequence variants have been previously reported to be associated with resistance. There also were no significant differences in EPSPS gene expression level among these eight biotypes in response to glyphosate treatment, based on an RNAseq study (Y. Peng, Y. Sang, S. Allen, and C.N. Stewart, unpublished data). Therefore, we conclude that glyphosate resistance in horseweed is conferred by a non-target-site mechanism.

DISCUSSION

To our knowledge, horseweed is the first economically important weed to have its genome sequenced and assembled. The de novo strategy using whole-genome shotgun sequencing in this study was similar to those used in several recent reports (Al-Dous et al., 2011; Wang et al., 2012; Varshney et al., 2013), except that we also included third-generation sequencing reads in our study. The 92.3% coverage of 13,996 scaffolds should be considered to be a draft genome comparable to rice, soybean, maize, chickpea, grape, poplar, and date palm (Goff et al., 2002; Yu et al., 2002; Tuskan et al., 2006; Jaillon et al., 2007; Schnable et al., 2009; Schmutz et al., 2010; Al-Dous et al., 2011; Varshney et al., 2013). Further improvements in assembly accuracy and backfilling will require the use of either long-insertion mate-pair reads or physical maps (International Rice Genome Sequencing Project, 2005; Wang et al., 2012; Al-Mssallem et al., 2013).

The draft horseweed genome has proven to be immediately useful. We have successfully cloned 10 target promoters of interest to study glyphosate resistance and response (data not shown). Moreover, 12 SSR loci have already been tested and successfully used in a recent horseweed population genetics study (Okada et al., 2013). Also, the genome sequences have been used to annotate the explosive composition B (hexahydro-1,3,5-trinitro-1,3,5-triazine and 2,4,6-trinitrotoluene) response genes in Baccharis halimifolia (Ali et al., 2014). Furthermore, we observed that the horseweed chloroplast genome is nearly identical to those of sunflower and lettuce (https://lgr.genomecenter.ucdavis.edu). Higher plants have various mitochondrial genome architectures with respect to size and construction compared with the chloroplast genome, which is more conserved (Tian et al., 2006). Brassica carinata has a small mitochondrial genome (232,241 bp), whereas Cucurbita pepo has a much larger chloroplast genome (982,832 bp; Supplemental Table S4).

Next-generation sequencing technologies have been widely used in whole-genome sequencing and resequencing, which has led to the development of rapid genome-wide SNP detection applications in model and crop plant species for exploring within-species diversity, genotyping by sequencing, construction of haplotype maps, and performing genome-wide association studies of interesting traits (Craig et al., 2008; Huang

Table VII. Summary of genomic variation among the sequenced horseweed biotypes

<table>
<thead>
<tr>
<th>Biotype</th>
<th>MNV</th>
<th>SNV</th>
<th>Deletion</th>
<th>Insertion</th>
<th>Replacement</th>
<th>Heterozygous</th>
<th>Homozygous</th>
<th>Total No.</th>
<th>Size (bp)</th>
<th>Size per Variant (kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-R</td>
<td>7,069</td>
<td>188,525</td>
<td>35,475</td>
<td>3,664</td>
<td>621</td>
<td>184,596</td>
<td>50,758</td>
<td>235,354</td>
<td>247,452</td>
<td>1.32</td>
</tr>
<tr>
<td>CA-S</td>
<td>13,253</td>
<td>279,016</td>
<td>10,329</td>
<td>5,407</td>
<td>1,035</td>
<td>235,416</td>
<td>73,624</td>
<td>309,040</td>
<td>329,025</td>
<td>0.99</td>
</tr>
<tr>
<td>DE-S</td>
<td>11,215</td>
<td>352,016</td>
<td>33,265</td>
<td>10,712</td>
<td>1,210</td>
<td>228,343</td>
<td>180,075</td>
<td>408,418</td>
<td>432,327</td>
<td>0.76</td>
</tr>
<tr>
<td>DE-R</td>
<td>13,024</td>
<td>419,703</td>
<td>41,670</td>
<td>11,843</td>
<td>1,255</td>
<td>234,044</td>
<td>253,451</td>
<td>487,495</td>
<td>516,307</td>
<td>0.63</td>
</tr>
<tr>
<td>IN-S</td>
<td>9,890</td>
<td>273,165</td>
<td>16,479</td>
<td>6,049</td>
<td>832</td>
<td>248,444</td>
<td>58,371</td>
<td>306,415</td>
<td>323,209</td>
<td>1.01</td>
</tr>
<tr>
<td>IN-R</td>
<td>10,394</td>
<td>298,866</td>
<td>16,720</td>
<td>7,360</td>
<td>965</td>
<td>270,989</td>
<td>66,728</td>
<td>337,717</td>
<td>356,007</td>
<td>0.92</td>
</tr>
<tr>
<td>TN-R</td>
<td>6,650</td>
<td>174,915</td>
<td>48,417</td>
<td>3,646</td>
<td>617</td>
<td>155,282</td>
<td>78,963</td>
<td>234,245</td>
<td>246,577</td>
<td>1.33</td>
</tr>
<tr>
<td>TN-S</td>
<td>9,725</td>
<td>330,880</td>
<td>58,920</td>
<td>10,432</td>
<td>1,043</td>
<td>205,585</td>
<td>205,335</td>
<td>410,920</td>
<td>433,873</td>
<td>0.75</td>
</tr>
<tr>
<td>Total</td>
<td>81,220</td>
<td>2,317,086</td>
<td>264,687</td>
<td>59,113</td>
<td>7,578</td>
<td>1,762,299</td>
<td>967,305</td>
<td>2,729,604</td>
<td>2,884,777</td>
<td>0.96*</td>
</tr>
</tbody>
</table>

Table VI. Summary of genome resequencing of seven additional horseweed biotypes using the Illumina HiSeq 2000 platform

<table>
<thead>
<tr>
<th>Population Identifier</th>
<th>Accession Location</th>
<th>Phenotype</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-S</td>
<td>Fresno, CA</td>
<td>S</td>
<td>36×</td>
</tr>
<tr>
<td>CA-R</td>
<td>Fresno, CA</td>
<td>R</td>
<td>19×</td>
</tr>
<tr>
<td>DE-S</td>
<td>Georgetown, DE</td>
<td>S</td>
<td>24×</td>
</tr>
<tr>
<td>DE-R</td>
<td>Georgetown, DE</td>
<td>R</td>
<td>28×</td>
</tr>
<tr>
<td>IN-S</td>
<td>Knox County, IN</td>
<td>S</td>
<td>16×</td>
</tr>
<tr>
<td>IN-R</td>
<td>Knox County, IN</td>
<td>R</td>
<td>16×</td>
</tr>
<tr>
<td>TN-S</td>
<td>Jackson, TN</td>
<td>S</td>
<td>30×</td>
</tr>
</tbody>
</table>
With the reported horseweed genome data, 51,892 SSR markers and over 2.7 million SNPs, as well as chloroplast DNA markers, were identified. Although glyphosate-susceptible and glyphosate-resistant plants, obviously, had opposite phenotypic responses to glyphosate treatment, 99.93% of their genomes were identical. We found many SNPs in glyphosate-resistant or glyphosate-susceptible genomes that were located in hot pots, which were defined as those with low $P$ values ($P \leq 0.01$), and might be associated with glyphosate resistance. However, many SNPs will be silent at the protein sequence level and not of interest for further study in terms of biological functions. Other mutations involved in gene expression regulation (e.g. within promoter regions) are of potential interest for elucidating resistance and other traits. Interestingly, horseweed exhibited considerable morphological variability among accessions (Supplemental Fig. S13); thus; integrating genomic information with resistance phenomena could be useful in weed science not only in studying the rapid evolution of herbicide resistance but also other traits.

The evolution of glyphosate resistance is a critical process that, in most cases, has been elucidated with regard to genomics and molecular mechanisms. For over a decade, we have known that glyphosate resistance in horseweed is a semidominant trait that is determined by a single simple Mendelian locus (Zelaya et al., 2004). The major focus of weed scientists to combat resistance has been to explore more herbicide control strategies while essentially ignoring the genomic and evolutionary bases of resistance. With few genomic resources for weeds and little expertise to utilize the available resources among the weed science community (Stewart, 2009), discovering the molecular mechanisms underlying nontarget resistance has been extremely difficult. However, now that the genomics era has found its way to weed science, we can begin to answer fundamental questions about what makes weeds so weedy and capable of adapting to control measures and, thereby, design approaches that reduce the further evolution of herbicide resistance in weeds (Basu et al., 2004; Stewart et al., 2009). The currently known mechanisms of glyphosate resistance in weedy plants include EPSPS target-site mutations (Baerson et al., 2002; Collavo and Sattin, 2012; Sammons et al., 2009, 2010; Atwell et al., 2010; Todesco et al., 2010; Wu et al., 2010; Kump et al., 2011; Deschamps et al., 2012).

Figure 4. Summary of genome variants and distribution among biotypes. A, Total variant counts in different biotypes. B, Percentage of heterozygous and homozygous variants among biotypes. C, Distribution of variant types. Del, Deletion; Ins, insertion; MNV, multiple-nucleotide variation; Rep, replacement; SNV, single-nucleotide variation. D, Number of specific variants in glyphosate-resistant (GR) and glyphosate-susceptible (GS) populations. [See online article for color version of this figure.]
An important goal of effective weed management is to stop or slow the evolution and spread of herbicide resistance in weeds. Thus, it is vital to understand whether glyphosate resistance originated once and spread from a single source horseweed population or originated multiple independent times within distinct populations. If resistance originated once and spread, say, via seeds, then the molecular mechanisms would be expected to be the same or similar among populations because of identical descent. In that case, the prevention of seed dispersal and movement by machinery or other means would be an effective resistance management strategy. Alternatively, if resistance originated multiple times and spread from multiple sources, reduction in both seed dispersal and selection pressure will be needed. Management would also benefit from understanding evolutionary dynamics. In the case of horseweed, the evolution of glyphosate resistance occurred independently in multiple locations and might be caused by more than one non-target resistance gene (Yuan et al., 2010; Okada et al., 2013; this study). Moreover, approximate Bayesian computation (ABC) analyses of microsatellite marker variation indicated that resistant populations underwent expansion after greatly increased glyphosate use in California in the 1990s, but many years before it was detected, strongly suggesting that diversity in weed-control practices prior to herbicide regulation probably kept resistance frequencies low (Okada et al., 2013). Data from this study also seem not to support the long-range rapid spread of resistance from a single source. At one time, the application of a combination of herbicides simultaneously or in sequence was considered to be an effective resistance management strategy. However, resistance to multiple herbicides is becoming increasingly common (Ashworth et al., 2014; Heap, 2014), and this approach is no longer consistently effective. Herbicides with new modes of action or more diversified control measures are desperately needed. Understanding the genomic mechanisms underlying non-target resistance would allow the development of novel approaches, such as allelopathy and the expanded use of safeners.

For the first time, we have genomics data to begin to address the genetic basis of weediness. Do weeds have unique genes that endow weediness or merely variants of genes common to all plants? Most researchers have postulated that the latter situation is most plausible (Basu et al., 2004; Stewart et al., 2009). Indeed, our finding lends credence to the gene family expansion hypothesis, given that non-target resistance candidate gene families, ABC transporters, and cytochrome P450 genes, and to a lesser extent GST and GT genes, were overrepresented in horseweed relative to other published plant genomes. Weeds, at least some of them, might have an inordinate adeptness for rapid evolution that is manifested by herbicide resistance; gene family expansion and/or changes in gene expression might be why weeds are weeds (Yuan et al., 2007; Shaner, 2009; Ge et al., 2010; Peng et al., 2010; Gaines et al., 2014; Sammons and Gaines, 2014). Other genes of interest with enriched GO terms, such as vacuole transport, photosynthetic acclimation, and detoxification of nitrogenous compounds, were found (Fig. 3). Finally, horseweed seems to have relatively more protein-coding genes (44,592; also supported by EST data) in comparison with other sequenced plant genomes while maintaining the smallest genome of all economically important weeds (Sterck et al., 2007; Stewart et al., 2009; Paterson et al., 2010). However, de novo genome assembly (contigs or scaffolds) tends to underestimate repeat numbers and overestimate gene numbers, because repeats are compacted on each other and genes are somewhat fragmentary and get annotated as two unique genes in some cases (Chaisson and Pevzner, 2008; Baker, 2012; Seabury et al., 2014).
Repeat numbers in the horseweed genome are lower than in most other sequenced plant genomes, in part because horseweed has a relatively compact genome. However, read-mapping analyses revealed that the read depth of repeated regions was much higher than for nonrepeats. This finding implies that repeat numbers are underestimated in the horseweed genome and that a de novo repeat-searching model will be needed for further analysis.

The small genome size of horseweed and other economically important agricultural weeds might have a practical and evolutionary significance. Twenty-three of the 25 most-studied weedy plant species have 1C genome sizes of less than 5,000 Mb, and 13 of these have genome sizes of less than 2,000 Mb (Stewart et al., 2009). Recently, a meta-analysis was performed in which invasiveness among plant species was negatively associated with genome size and positively associated with chromosome number (Pandit et al., 2014). This study analyzed 890 species among 62 genera. Even though it presents one data set, horseweed appears to be, potentially, an archetypical weed in that it has a very streamlined genome but with a large number of genes, which could give it the capacity for rapid evolution in changing environments. Thus, beyond its use for practical agricultural research, the horseweed genome might shed light on standing questions of invasion, evolution, and changing climates (Caplat et al., 2013). The bane of farmers could be the harbinger of genomic enlightenment.

MATERIALS AND METHODS

Plants and Phenotyping

Seeds sampled from horseweed (Conyza canadensis) populations were germinated and grown in potting medium (3B potting medium, 10-cm-diameter pot, one plant per pot; Supplemental Fig. S13) in a greenhouse under a 16-/8-h light/dark photoperiod at ambient temperatures (25°C ± 2°C). Twenty-four 3-month-old plants from each population were treated with glyphosate at the rosette stage when plants were 6 to 8 cm in diameter, and glyphosate-treated plants were considered to be resistant if they were alive at the end of 3 weeks (Yuan et al., 2010).

DNA Isolation and Genome Sequencing Using Multiple Platforms

Genomic DNA was extracted from six individual plants of the glyphosate-resistant biotype that was collected originally from a soybean field in Jackson, TN (TN-R), using Plant DNeasy kits (Qiagen). Whole-genome shotgun sequencing was performed using the Illumina HiSeq 2000, 454 GS-FLX Titanium, and PacBio RS sequencing platforms. A total of three sequencing libraries were constructed with insert sizes of approximately 600 bp for the 454 GS-FLX Titanium system. Two paired-end sequencing libraries with insert sizes of approximately 350 bp and one mate-pair library with insert sizes of approximately 3,000 bp was constructed for the Illumina HiSeq 2000 system. One library with insert sizes of approximately 10,000 bp was constructed for the PacBio RS sequencing system.

De Novo Genome Assembly

The de novo genome assembly strategy is shown in Supplemental Figure S1. The 454 GS-FLX reads were assembled with Newbler. The Illumina paired-end sequence reads were divided by sequencing depth (approximately 50×) using error-corrected PacBio long reads as guidance (Koren et al., 2012), and subsets were assembled using SOAPdenovo and CLC Genomic Workbench. NGen was used to combine the primary assembled contigs. CLC Genomic Workbench was used to find the true mate pairs by mapping with the total contigs. NGen was used for the final scaffold construction and gap closure. All the parameter settings were described previously (Varshney et al., 2012; Wang et al., 2012). To check the completeness of the assembly, we mapped ESTs and transcriptome data to the genome assembly using BLASTN (with e value cutoff of 10−5).

Assembly of Chloroplast and Mitochondrial Genomes

The sequencing reads were screened against custom plant chloroplast and mitochondria genome databases. The subset sequences for chloroplast and mitochondria were assembled respectively. The complete chloroplast genome was annotated using DOGMA (Wyman et al., 2004). To remove nonmitochondrial assembly, the mitochondrial contigs were further screened against plant mitochondrial genome databases using BLASTN with e value cutoff of 10−5.

Gene and Repeat Annotation

Gene prediction was performed using homology-based and transcript-based methods. Previous horseweed ESTs and transcriptome data were aligned to the genome assembly using BLAT (bit score > 10; coverage ≥ 0.98) to generate spliced alignments (Kent, 2002), which were linked according to their overlap using PASA (Haas et al., 2003). Plant proteins of other species were also mapped to the genome using TBlastN with e value cutoff of 10−5. Predicted transcripts were assigned biological information by searching the UniProt protein database using BLASTX with e value cutoff of 10−5. RepeatMasker version 4.0.5 (http://www.repeatmasker.org/) was used to search putative transposable element regions of the horseweed genome against an updated set of RepeatMasker libraries (20140131). To develop SSR markers for further population genetic studies, the masked SSR loci were further analyzed using the Simple-Repeat Sequence Identification Tool (http://archive.gramene.org/db/markers/sortool) with a threshold minimum of six repeats.

GO Annotation and Enrichment Analysis

Results of BLASTX of horseweed transcripts were further mapped using the GO database to assign GO terms and carry out GO annotation using the Blast2Go program integrated in CLC Genomic Workbench (http://www. blast2go.com). Enrichment analysis was performed by using the entire set of available Arabidopsis (Arabidopsis thaliana) transcripts as a reference. The P value of Fisher’s exact test was set to P ≤ 0.001 to reduce noise and output the most specific terms.

Genome Resequencing and Variation Analysis

DNA samples of another seven horseweed biotypes were isolated using the method described above. Paired-end DNA libraries (approximately 350-bp insertion) were constructed for each biotype, bar coded, and subjected to Illumina sequencing using the HiSeq 2000 platform. After trimming and quality-control steps, the remaining data for each population were assembled into genomes with the TN-R assembly as reference using CLC Genomic Workbench 7.03. The probabilistic variant detection model was chosen to report the variants. A minimum cutoff value (20 times or greater coverage at the variant loci, 20% or greater frequency) was applied to avoid issues of sequencing and assembly errors in variant calling. Fisher’s exact test was carried out to detect specific variants in glyphosate-susceptible or glyphosate-resistant groups. The cutoff sample frequency was 90% or greater, which meant that only variants present in all test samples, but not in any reference samples, would be reported. To carry out phylogenetic analysis based on whole-genome single-nucleotide variation profiles among sequenced horseweed biotypes, the shrunk-genomes method of the program PhyloSNP (Faison et al., 2014) with a position delta of zero surrounding each SNP was chosen. The algorithm of a quantitative method was used, in which the created matrix of presence/absence and the position of each SNP from the shrunk-genome alignment can be used directly to generate a distance matrix among the biotypes. Finally, the tree was built by using the neighbor-joining method in Phylip 3.695.
Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Data flow of de novo genome assembly.

Supplemental Figure S2. Factors effect on the quality of de novo genome assembly.

Supplemental Figure S3. Length and quality distribution of PacBio reads.

Supplemental Figure S4. PacBio reads rescue.

Supplemental Figure S5. The benefit of including PacBio long reads.

Supplemental Figure S6. True and false mate-paired reads.

Supplemental Figure S7. Repeats distribution of all SSR loci.

Supplemental Figure S8. GO annotation.

Supplemental Figure S9. Enriched GO terms in transporter subgroup.

Supplemental Figure S10. Map of horseweed chloroplast genome.

Supplemental Figure S11. Synaptic analysis of the horseweed chloroplast genome.

Supplemental Figure S12. Horseweed mitochondrial genome.

Supplemental Figure S13. Morphological variation among horseweed populations.

Supplemental Figure S14. Distribution of specific SNPs.

Supplemental Figure S15. Alignment of EPSPS protein sequences from different horseweed biotypes.

Supplemental Table S1. Mapping of mate-paired reads.

Supplemental Table S2. GC content of plant genomes.

Supplemental Table S3. Summary of SSR markers.

Supplemental Table S4. Summary of plant mitochondrial genomes.

Supplemental Table S5. Variants within genomes among horseweed biotypes.

Supplemental Data Set S1. Annotation of horseweed chloroplast genome.

Supplemental Data Set S2. Horseweed mitochondrial genome sequences.

ACKNOWLEDGMENTS

We thank Sara Allen, Laura Abercrombie, Qiong Jia, Guanglin Li, and Nolan Kane for assistance with material sampling, sequencing, and data collection. We also thank the providers of the original sources of germplasm, including Anir Shrestha, David Grantz, Vince Davis, Robert M. Hayes, and Barbara Scott.

Received July 29, 2014; accepted September 9, 2014; published September 10, 2014.

LITERATURE CITED


Downloaded from on July 31, 2017 - Published by www.plantphysiol.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.


Copyright © 2014 American Society of Plant Biologists. All rights reserved.


CORRECTIONS

Vol. 166: 1241–1254, 2014


On p. 1243 of this article, the assembled draft of horseweed genome data submitted to the National Center for Biotechnology Information is incorrectly listed as having the accession number SUB535309.

The correct accession number for the Whole Genome Shotgun project deposited at DDBJ/EMBL/GenBank is JSWR0000000.