frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 29 June 2022
doi: 10.3389/fpls.2022.909073

Genome-Guided Analysis of Seven Weed Species Reveals Conserved Sequence and Structural Features of Key Gene Targets for Herbicide Development

Sarah Shah[1,2], Thierry Lonhienne[2], Cody-Ellen Murray[3], Yibi Chen[1,2], Katherine E. Dougan[1,2], Yu Shang Low[2], Craig M. Williams[2], Gerhard Schenk[2,4,5], Gimme H. Walter[3], Luke W. Guddat[2] and Cheong Xin Chan[1,2]**

[1] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia, [2] School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia, [3] School of Biological Sciences, The University of Queensland, Brisbane, QLD, Australia, [4] Sustainable Minerals Institute, The University of Queensland, Brisbane, QLD, Australia, [5] Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Parimalan Rangan,
National Bureau of Plant Genetic Resources (ICAR), India

Reviewed by:
Sathishkumar Ramalingam,
Bharathiar University, India
Muthusamy Ramakrishnan,
Nanjing Forestry University, China

***Correspondence:**
Luke W. Guddat
l.guddat@uq.edu.au
Cheong Xin Chan
c.chan1@uq.edu.au

Specialty section:
This article was submitted to Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science

Received: 31 March 2022
Accepted: 08 June 2022
Published: 29 June 2022

Citation:
Shah S, Lonhienne T, Murray C-E, Chen Y, Dougan KE, Low YS, Williams CM, Schenk G, Walter GH, Guddat LW and Chan CX (2022) Genome-Guided Analysis of Seven Weed Species Reveals Conserved Sequence and Structural Features of Key Gene Targets for Herbicide Development. Front. Plant Sci. 13:909073. doi: 10.3389/fpls.2022.909073

Herbicides are commonly deployed as the front-line treatment to control infestations of weeds in native ecosystems and among crop plants in agriculture. However, the prevalence of herbicide resistance in many species is a major global challenge. The specificity and effectiveness of herbicides acting on diverse weed species are tightly linked to targeted proteins. The conservation and variance at these sites among different weed species remain largely unexplored. Using novel genome data in a genome-guided approach, 12 common herbicide-target genes and their coded proteins were identified from seven species of Weeds of National Significance in Australia: *Alternanthera philoxeroides* (alligator weed), *Lycium ferocissimum* (African boxthorn), *Senecio madagascariensis* (fireweed), *Lantana camara* (lantana), *Parthenium hysterophorus* (parthenium), *Cryptostegia grandiflora* (rubber vine), and *Eichhornia crassipes* (water hyacinth). Gene and protein sequences targeted by the acetolactate synthase (ALS) inhibitors and glyphosate were recovered. Compared to structurally resolved homologous proteins as reference, high sequence conservation was observed at the herbicide-target sites in the ALS (target for ALS inhibitors), and in 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase (target for glyphosate). Although the sequences are largely conserved in the seven phylogenetically diverse species, mutations observed in the ALS proteins of fireweed and parthenium suggest resistance of these weeds to ALS-inhibiting and other herbicides. These protein sites remain as attractive targets for the development of novel inhibitors and herbicides. This notion is reinforced by the results from the phylogenetic analysis of the 12 proteins, which reveal a largely consistent vertical inheritance in their evolutionary histories. These results demonstrate the utility of high-throughput genome sequencing to rapidly identify and characterize gene targets by computational methods, bypassing the experimental characterization of individual genes. Data generated from this study provide a useful reference for future investigations in herbicide discovery and development.

Keywords: genetics, herbicide resistance, weed genomics, sequence analysis, gene targets, protein targets, acetolactate synthase, EPSP synthase

INTRODUCTION

The management and control of weed species remain a major global problem, largely due to increased frequencies of herbicide resistance and lack of new herbicides entering the market. Weeds threaten the environment by damaging the natural landscapes and waterways, leading to displacement of native species, degradation of the ecosystem, and reduction in biodiversity. In Australia, over 30 Weeds of National Significance (Atlas of Living Australia, 2022) have been identified to date, due to the negative impacts they have on the environment, economy, and society. These weeds were selected based on their invasiveness and ability to spread and persist in natural and agricultural environments, and the severity of the risks they pose. Legislation has been implemented to contain their spread by mandating reporting and removal. In 2018, the financial cost of weed control across natural agriculture and non-agriculture sectors (e.g., management of environments and indigenous lands) was estimated between $4.5 and $6.0 billion per annum in Australia (McLeod, 2018).

Application of herbicides is one of the most common approaches for managing weeds. Other approaches include manual/mechanical removal and biological control using insects and plant pathogens. Increasingly, combinations of these approaches are used in an integrated pest management framework. However, the use of herbicides, which is often more cost-effective and less labor-intensive, remains highly popular. Two of the most widely used commercial herbicides are the acetolactate synthase [ALS, also known as acetohydroxyacid synthase (AHAS)] inhibitors, and glyphosate (Funke et al., 2006; Garcia et al., 2017). ALS is the first enzyme in the branched-chain amino acid (BCAA) biosynthesis pathway that yields valine, isoleucine and leucine (Garcia et al., 2017), whereas glyphosate inhibits 5-enolpyruvylshikimate-3-phosphate (EPSP) synthase that is central to the biosynthesis pathway of aromatic amino acids (Funke et al., 2006). It is most commonly assumed that ALS inhibitors function by preventing substrate from gaining access to the active site of the enzyme. However, it has also been proposed that inhibition of ALS could also result in the accumulation of 2-keto-butyrate, a substrate for ALS known to be toxic to plants (Dayan and Duke, 2020).

A sound understanding of the specificity and effectiveness of herbicides at targeting weed species is critical for selecting or designing target-specific herbicides with optimized potency. The risk of weeds developing resistance should also be minimized along with potential negative impact on non-target species and the broader environment. An intimate knowledge of metabolic pathways linked to herbicide action will guide targeted prediction of herbicide efficiency for specific weeds. Genome and proteome data from weed species present a powerful resource for identifying inherent variations in gene and protein sequences (e.g., membrane transporters and cytochrome P_{450} enzymes) that may influence and impinge upon the effectiveness of specific herbicides, and for identifying genes or proteins conferring herbicide resistance (Ravet et al., 2018; Fonseca et al., 2020; Garcia et al., 2020). To date, however, little genome-scale data are available from invasive or problematic weed species, presenting a significant knowledge gap. Problematic weed species are highly diverse and classified in various Orders, distinct from the model weed plant of *Arabidopsis thaliana*, thus observations based on analysis on *Arabidopsis* may not be generalizable to other species.

In this study, we adopted a genome-guided approach to assess 12 common herbicide-target genes and their coded proteins among seven invasive weed species in Australia. Novel genome data generated from these species enabled the intersection of sequence and protein structural information to elucidate target sites for glyphosate and ALS inhibitors, and for potential development of other herbicides.

MATERIALS AND METHODS

Weed Species and Collection of Leaf Material

Seven species of Weeds of National Significance were chosen for this study (**Figure 1** and **Table 1**), based on their particularly pernicious role in the destruction of the Australian environment and their impact on agricultural yields. For example, lantana forms dense thickets that smother native vegetation (Gooden et al., 2009) and is potentially poisonous to livestock (Sharma et al., 2007; Ntalo et al., 2022), whereas alligator weed impacts both aquatic and terrestrial environments and can displace native flora, especially along river banks (Sainty et al., 1997). Leaf material for each weed species was collected from field infestations in Queensland or plants in culture (**Table 1**), and transported to the University of Queensland (Saint Lucia campus, Brisbane) for subsequent analysis.

Each field sample was identified by reviewing morphology and habit of the plant against the diagnostic features outlined in the Lucid Key Server for Environmental Weeds of Australia.[1] Leaf samples were cut from plants, placed inside paper seed

[1] https://keyserver.lucidcentral.org/weeds

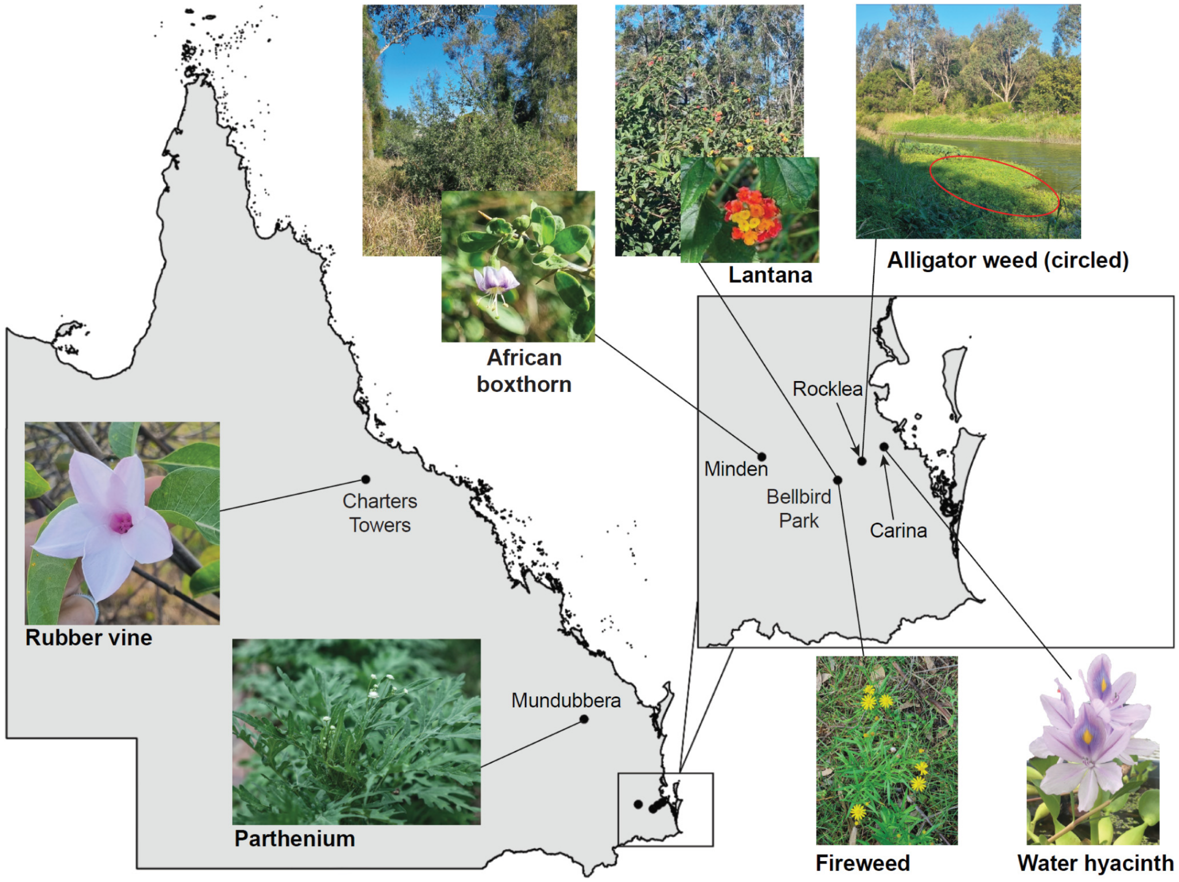


FIGURE 1 | Locations in Queensland, Australia, where plant material or seeds (to grow plants for leaf samples) were collected from the Weeds of National Significance investigated in this study. Clockwise from left: Rubber vine (*Crytostegia grandiflora*) in Charters Towers, pictured in flower; African Boxthorn (*Lycium ferocissimum*) infestation from Minden, with flower inset; lantana (*Lantana camara*) infestation at Bellbird Park, with flower inset; alligator weed (*Alternanthera philoxeroides*) infestation (and other weeds) in Stable Swamp Creek, Rocklea; water hyacinth (*Eichhornia crassipes*, also known as *Pontederia crassipes*) from Bulimba Creek, Carina (photo by Jennifer Evans and used with permission); fireweed (*Senecio madagascariensis*) from Bellbird Park, in flower; and parthenium (*Parthenium hysterophorus*) from Mundubbera.

TABLE 1 | Seven weed species used in this study.

Common name	Species name	Order	Leaf collection site
Alligator weed	*Alternanthera philoxeroides*	Caryophyllales	Rocklea, Queensland
African boxthorn	*Lycium ferocissimum*	Solanales	Minden, Queensland
Fireweed	*Senecio madagascariensis*	Asterales	Bellbird Park, Queensland
Lantana	*Lantana camara*	Lamiales	Bellbird Park, Queensland
Parthenium	*Parthenium hysterophorus*	Asterales	Lab cultures (Queensland Department of Agriculture and Fisheries, Ecoscience Precinct, Brisbane); grown from seeds collected at Mundubbera, Queensland
Rubber vine	*Cryptostegia grandiflora*	Gentianales	Lab cultures (Univeristy of Queensland), grown from seeds acquired from Tropical Weeds Research Centre, Charters Towers, Queensland
Water hyacinth	*Eichhornia crassipes* (also known as *Pontederia crassipes*)	Commelinales	Carina, Queensland

packets, and kept cool using ice bricks inside polystyrene boxes until delivery. The leaves were cleaned with reverse-osmosis water, snap-frozen in liquid nitrogen and stored at −80°C until DNA extraction.

Generation of Genome Data, *de novo* Genome Assembly, and Gene Prediction

For each weed species, genomic DNA was extracted from the leaf tissue using the ISOLATE II Plant DNA Kit (Meridian Bioscience, Cincinnati, OH, United States) following the manufacturer's manual. Integrity of the extracted DNA was confirmed by agarose gel electrophoresis (**Supplementary Figure 1**), and DNA quality was assessed using Nanodrop 2000 Spectrophotometer (ThermoFisher Scientific, Waltham, MA, United States). All seven DNA samples (2–4 μg each; 20–60 mg/mL) were sent to the Australian Genome Research Facility (Melbourne, Australia) for preparation of sequencing libraries (Illumina TruSeq; average insert size 445 bp) and sequencing using the NovaSeq 6000 platform (2 × 150 bp paired-end, multiplexed in a single S4 flow cell); base calling was conducted using llumina bcl2fastq version 2.20.0.422. In total, 928.2 Gbp of sequence data were generated from the seven species (**Supplementary Table 1**). All sequence data were processed using fastp v0.20.0 (Chen et al., 2018) at default setting to remove adapter sequences, poly-G fragments (a known artifact in NovaSeq 6000 data), and remaining sequence reads of length < 50bp; this yielded 895.5 Gbp of filtered data (**Supplementary Table 1**). For each dataset, a haploid genome size was estimated based on the distribution of *k*-mer coverage ($k = 21$) generated using Jellyfish v2.3.0 (Marçais and Kingsford, 2011) adapting the *in silico* approach from earlier studies (Lin et al., 2015; González-Pech et al., 2021). Briefly, a haploid genome size is estimated by dividing the number of all observed *k*-mers by the fold-coverage of *k*-mers at the first peak of the frequency distribution. All sequence data generated from this study are available on NCBI Sequence Read Archive (BioProject accession PRJEB47218).

For each species, using the corresponding filtered reads, MaSuRCA v3.3.3 (Zimin et al., 2013) was used to assemble the genome *de novo* into contiguous sequences (i.e., genome scaffolds). Sequences that exhibit unusual G + C content and genome-scaffold lengths were deemed putative contaminant sequences and were removed following Chen et al. (2020), yielding the final genome assembly (i.e., the representative, primary genome assembly). Benchmark Universal Single-Copy Orthologs (BUSCO) v4.1.2 (Waterhouse et al., 2018) were used to assess the completeness of each assembled genome, based on recovery of core conserved genes in the viridiplantae_odb10 dataset (at *genome* mode). The assembled genome of alligator weed is highly fragmented (1.16 million scaffolds, N50 length 3.3 Kbp) with low BUSCO recovery (47.0%; **Supplementary Table 2**). Different tools are expected to yield different genome assemblies in which some sequence regions are better resolved in one compared to another. To maximize recovery of genome data from alligator weed, the data were also assembled *de novo* using another common tool, CLC Genomics Workbench v10 (Qiagen); this CLC assembly is considered as secondary genomic evidence for alligator weed. For each primary genome assembly from the seven weed species, MAKER v2.31.10 (Cantarel et al., 2008; Holt and Yandell, 2011) was used to predict protein-coding genes, guided by the manually curated protein sequences in SwissProt release 02032020. All assembled genomes, and the predicted gene and protein sequences are available at https://doi.org/10.48610/c22c10c.

Identification of 18S Ribosomal RNA Sequences for Species Validation

From each genome dataset, full-length 18S rRNA sequences were identified using hmmscan v3.2.1 (Eddy, 2011) with a hidden Markov model profile created using reference 18S rRNA sequences from 1,897 angiosperm species (Maia et al., 2014), and a BLASTn v2.10.1+ (Altschul et al., 1990) search using this reference as query. The 18S rRNA sequences were recovered from the pimary genome assemblies of all weed species except for alligator weed, in which the more-complete 18S rRNA sequence was recovered from the CLC assembly.

Identification and Analysis of Herbicide-Targeting Genes

We focused on 12 key genes that are known or are potential targets for commercial herbicides (i.e., glyphosate and ALS

TABLE 2 | The 12 genes analyzed in this study.

Gene name	Gene product (protein)	Herbicide targeting the gene or the implicated molecular function
ALS (CSU)	Catalytic subunit (CSU) or large subunit of acetolactate synthase (ALS)	Group B (inhibition of ALS), first gene in BCAA pathway
ALS (RSU)	Regulatory subunit (RSU) or small subunit of acetolactate synthase (ALS)	Group B (inhibition of ALS), first gene in BCAA pathway
TD	Threonine dehydratase (deaminase)	BCAA pathway
KARI	Ketol-acid reductoisomerase	BCAA pathway
DHAD	Dihydroxy-acid dehydratase	BCAA pathway
BCAT-2	Branched-chain-amino-acid aminotransferase (isoform 2 sequences are the most commonly available)	BCAA pathway
IPMS	2-Isopropylmalate synthase 1	BCAA pathway
IPMI	Isopropylmalate isomerase (dehydratase)	BCAA pathway
IMDH	3-Isopropylmalate dehydrogenase	BCAA pathway
aroA	5-enolpyruvylshikimate-3-phosphate (EPSP) synthase	Group G (inhibition of EPSP synthase)
accA	Acetyl-coA carboxylase subunit alpha	Group A (inhibition of ACCase)
psbA	Photosystem II protein D1	Group C (inhibition of photosynthesis PS II)

The relevant herbicides are grouped based on mode of action following the Herbicide-Resistance Action Committee classification system (Beffa et al., 2019).

inhibitors) (**Table 2**). Of these, the catalytic subunit (CSU) of ALS is the target for ALS inhibitors. The gene encoding EPSP synthase is the target for glyphosate, whereas genes encoding acetyl-CoA carboxylase (ACCase) and D1 protein of photosystem II are targets for other herbicides. In all cases the presence of these genes is essential for survival of plants.

To recover the full-length sequence of each gene from the genome data generated for each weed species, we first searched among the protein sequences predicted from the genome data, using the corresponding plant protein sequences as queries in BLASTp searches. To curate these query sequences, relevant plant protein sequences that are categorized as InterPro families, i.e., ALS CSU (IPR012846), ALS regulatory subunit (RSU; IPR004789), threonine dehydratase (TD; IPR005787), ketol-acid reductoisomerase (KARI; IPR013023), and branched-chain-amino-acid aminotransferase (BCAT-2; IPR005786), were downloaded from the InterPro database on October 15, 2020. For the other proteins not classified as InterPro families, the sequences were retrieved from UniProt and NCBI GenBank. The complete list of query sequences is shown in **Supplementary Table 3**. If full-length protein was not recovered this way, their coding sequences were searched among the assembled genome sequences using known protein sequences as query (tBLASTn). If no full-length gene or protein sequence were recovered in these two steps, a gene-targeted assembly approach was adopted to recover the full-length coding sequence using MegaGTA v0.1-alpha (Li et al., 2017). Briefly, for each weed species, a multiple sequence alignment and the associated Hidden Markov model profile (i.e., the alignment profile) for the target gene were first created using homologous sequences from taxa within the same Order (**Supplementary Table 4**), which were sourced from InterPro release 80.0 and the associated databases within. The alignment and the associated profile were then used to recover the target protein-coding sequence in full length. Given that conservation of gene functions are manifested in the coded protein sequences, we used the predicted protein sequences from each genome in subsequent analysis.

Validation of ALS CSU From African Boxthorn Using Polymerase Chain Reaction

As a proof of concept of our targeted approach using whole-genome data, we validated the ALS CSU coding sequence from African boxthorn using a PCR-based primer-walking approach. Primers (**Supplementary Table 5**) were designed based on highly conserved regions in the middle of ALS_CSU coding sequences. During the first phase, a PCR reaction was performed on the genomic DNA using degenerate primers, A1F and A1R, from which the amplified product together with the primers B1F and B1R were used in a second PCR reaction. The final amplified product was verified using agarose (1% w/v) gel electrophoresis, extracted using the QIAquick gel extraction kit (Qiagen, United States) following the manufacturer's procedure, and sequenced using the Sanger method at the Australian Genome Research Facility (Brisbane, Australia). This sequence was then used to design primers targeting up-/down-stream regions, as part of the primer-walking strategy; these primers include those targeting sequence coding for the N-terminal (C1R, C2R) and C-terminal (D1F) protein regions, including degenerate primers (C1F, D1R) designed based on other similar ALS_CSU sequences from other closely related plants. During the second phase, the sequence coding for the N-terminal region was identified by performing a PCR reaction on the genomic DNA using primers C1F and C1R, from which the amplified product together with primers C1F and C2R were used in a second PCR reaction. An independent PCR reaction was performed on the genomic DNA using primers D1F and D1R to verify the sequence coding for the C-terminal region. The sequence coding for the mitochondrial transit peptide is highly variable, and thus cannot be determined using this approach.

Each PCR reaction mix (50 μL) contained 1 × Thermopol® buffer, 0.2 mM dNTPs, forward primer (1 μM), reverse primer (1 μM), template DNA (∼1000 ng), and *Taq* DNA polymerase (3 U). If genomic DNA was the template, 4 μL (∼30 ng) of the purified DNA was used. For each PCR reaction, initial denaturation (95°C, 1 min) was followed by 35 cycles of denaturation (95°C, 20 s), annealing [primer-specific temperature (**Supplementary Table 5**), 30 s], and elongation (70°C, 45 s; 1 min for the 35th elongation step). The reaction was kept at 10°C prior to subsequent analysis.

Assessment of Conserved Amino Acids and Target Sites for ALS and EPSP Synthase

For each of ALS RSU, ALS CSU, and EPSP synthase, we compared protein sequences from the seven weed species against the structurally resolved protein sequence of a related species as reference. For ALS, we used the *At*RSU (UniProt Q93YZ7; PDB 6VZ8) and *At*CSU (UniProt P17597; PDB 1YBH) from *Arabidopsis thaliana*. For EPSP synthase, the protein from *Escherichia coli* (UniProt P0A6D3; PDB 1G6S) was used; no protein structures for plant EPSP synthase have been resolved. For each set of protein sequences, multiple sequence alignment was performed using Cluster Omega (Sievers et al., 2011). To assess interaction of a protein with other molecules, we used PoseView[2] (Stierand et al., 2006) to generate two-dimensional representations of the molecular complex formed by a protein interacting with substrate, cofactor, and herbicide molecules.

Inference of Phylogenetic Tree

For each gene, the query sequences used in the initial BLASTp searches (**Supplementary Table 3**) and the gene sequences recovered from the weed species were combined to reconstruct a homologous protein-sequence set; this yielded 12 homologous sets. For each set, multiple sequence alignment was performed using MAFFT v7.471 (Katoh and Standley, 2013) in L-INS-i mode. For the analysis of the 18S rRNA sequences, multiple sequence alignment was performed using MUSCLE v3.8.31 (Edgar, 2004). Aligned sequence positions

[2] https://www.biosolveit.de/free-to-academics/

that are not parsimoniously informative were removed with trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009) using the *automated1* setting. Using the processed alignment as input, a maximum-likelihood phylogenetic tree was inferred using IQ-TREE v2.0.3 (Nguyen et al., 2015) with the most appropriate substitution model chosen using ModelFinder (Kalyaanamoorthy et al., 2017), and node support assessed using an ultrafast bootstrap approximation (Hoang et al., 2018) based on 2,000 sample replicates.

RESULTS AND DISCUSSION

Genome Data From Seven Weed Species

We generated new genome data from seven invasive weed species in Australia to guide our identification of gene targets for common herbicides, focusing on glyphosate and ALS inhibitors. **Supplementary Table 2** shows the statistics for each assembled draft genome and the associated number of predicted protein-coding genes. We further assessed the completeness of these genome data based on the recovery of core conserved single-copy genes in Viridiplantae (i.e., the group encompassing all plants and green algae) using the standard benchmark tool for genome or gene-repertoire completeness, BUSCO (Waterhouse et al., 2018); in this context, a high recovery of these genes (e.g., >90%) indicates that the genome data are near complete.

All draft genomes are highly fragmented with a high number of scaffolds (>80,000) and short N50 lengths (<21 Kbp). For instance, the fragmented genome assembly for alligator weed (1,160,435 scaffolds, N50 length 3.30 Kbp) is likely due to its large genome size, i.e., only 1.75 Gbp was recovered in the assembly compared to the estimated size of 4.97 Gbp; this is also confirmed by the low (47.0%) genome completeness based on BUSCO analysis. In contrast, the most contiguous assembly is the rubber vine genome (82,815 scaffolds, N50 length 20.26 Kbp), with the estimated genome size of 0.48 Gbp (0.38 Gbp recovered in the assembly) and BUSCO recovery at 88.5%. Based on the common measures of contiguity of assembled genome sequences, the quality of these assemblies is insufficient for a comprehensive comparative genomic analysis, as the predicted protein-coding genes are likely fragmented, and the gene repertoire incomplete. As these sequence data were generated *de novo* with no good reference available, ploidy number and genome size cannot be readily identified from the data without further experiments. However, these data remain suitable for identifying specific genes in each weed species that are targeted by glyphosate and ALS inhibitors; these genes are expected to be present in all plant genomes. Moreover, not considering the assembled genome of alligator weed, all the other six assembled genomes have an average BUSCO completeness of 79.5% (between 62.9% and 90.6%; **Supplementary Table 2**), suggesting that most protein-coding genes encoded in these weed genomes are present in these albeit fragmented, *de novo* assembled sequences.

Given the fragmented genome assemblies, we took additional precaution to ensure that the data were truly representative of each weed species. Our comparison of 18S rRNA sequences from the seven genomes against those from other plant species revealed robust groupings (node support > 90%) of each weed species with their closely related species or genus, within the expected Order. For instance, fireweed and parthenium grouped within Asterales, while water hyacinth, the only monocot among the seven weed species, grouped within Order Commelinales as expected, close to other monocot Orders such as Asparagales (**Supplementary Figure 2**). This relationship was also supported at the protein level by the individual trees of two subunits of the ALS protein, i.e., the catalytic subunit (CSU; **Supplementary Figure 3**), and the regulatory subunit (RSU; **Supplementary Figure 4**). These results provide first-line molecular evidence for the identity of the seven weed species from which the genome data were generated.

Genome-Guided Identification of Genes Targeted by Herbicides

From the genome data, we adopted an integrated approach (Materials and Methods) to identify, in each weed species, 12 key genes that are current or potential targets for common herbicides, particularly glyphosate and ALS inhibitors, including seven involved in the BCAA pathway (**Table 2** and **Figure 2**; see Section "Materials and Methods"). The identified gene and protein sequences from each genome are available as **Supplementary Data Sheet 1**. The *ALS* gene is a current target for commercial herbicides, and for this we focus on

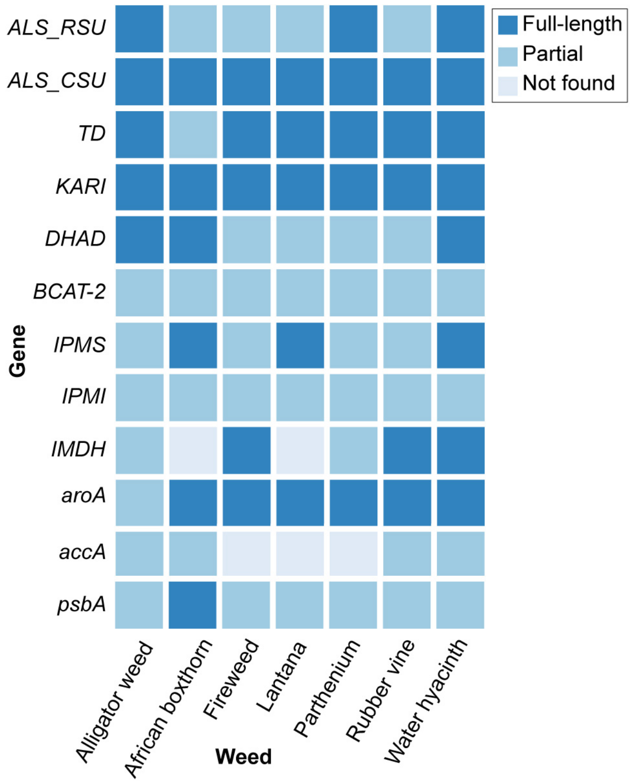


FIGURE 2 | The 12 key genes identified in the genomes of seven weed species.

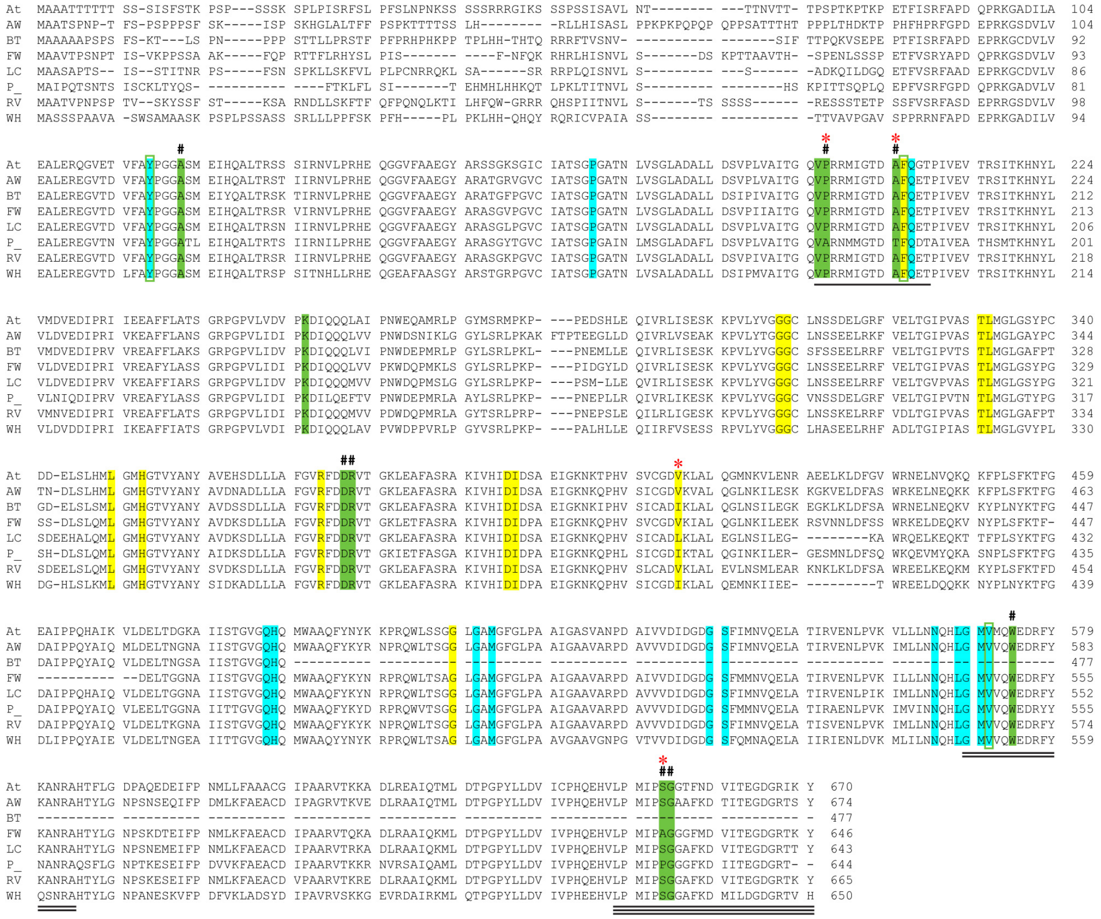
the catalytic subunit (CSU) of ALS. However, other potential binding sites can be found in the regulatory subunits (RSU) (Lonhienne et al., 2020). Inhibitors of other enzymes in the BCAA pathway, e.g., ketol-acid reductoisomerase (KARI) (Biou et al., 1997) and dehydroacid dehydratase (DHAD) (Yan et al., 2018) are also known to exhibit good herbicidal activity, thus detailed knowledge of the genes encoding these enzymes or regulators is useful for future efforts in herbicide discovery.

All 12 genes were recovered in the seven genomes, except for the *accA* (recovered in four) and *IMDH* (recovered in five) (**Figure 2**). This observation was likely due to the fragmented nature of the draft genome sequences, which also affects genome quality and completeness (**Supplementary Table 2**), not because of loss of these genes in these genomes. Recovery of most of these genes from the draft genomes demonstrates the utility of our high-throughput genome-sequencing approach for identifying gene targets in leiu of the more-tedious targeted gene sequencing approach (e.g., *via* amplification of specific gene regions).

Catalytic Subunit of Acetolactate Synthase

The CSU of ALS contains the target site for the largest number of commercial herbicides (Garcia et al., 2017). The holoenzyme of ALS also contains an RSU that binds to the CSU (Lonhienne et al., 2020). The RSU upregulates the activity of the enzyme by ~6-fold. This enhancement of the activity can be offset in the presence of the BCAAs (Lonhienne et al., 2020). The cryo-EM structure of the complex suggests that the presence of the RSU could affect the affinity of herbicide binding to the CSU, and could therefore be of relevance in explaining herbicidal activity for specific species.

We identified the complete sequences of the CSUs for all seven weed species (**Figure 2**); the coding sequence for the CSU in African boxthorn was further validated using PCR in a primer-walking approach (see Section "Materials and Methods"). The CSU is the target of five commercial herbicide families (see examples below) that bind to the catalytic pocket, thus blocking active site access to substrates (i.e., pyruvate and 2-ketobutyrate) (McCourt and Duggleby, 2006).



```
At  MAAATTTTTT SS-SISFSTK PSP---SSSK SPLPISRFSL PFSLNPNKSS SSSRRRGIKS SSPSSISAVL NT-------- ---TTNVTT- TPSPTKPTKP ETFISRFAPD QPRKGADILA  104
AW  MAATSPNP-T FS-PSPPPSF SK-----ISP SKHGLALTFF PSPKTTTTSS LH-------- -RLLHISASL PPKPKPQPQP QPPSATTTHT PPPLTHDKTP PHFHPRFGPD EPRKGSDVLV  104
BT  MAAAAAPSPS FS-KT---LS PN-----PPP STTLLPRSTF PFPRHPHKPP TPLHH-THTQ RRRFTVSNV- ---------- -------SIF TTPQKVSEPE PTFISRFAPD EPRKGCDVLV  92
FW  MAAVTPSNPT IS-VKPPSSA AK-----FQP RTTFLRHYSL PIS------- ----F-NFQK RHRLHISNVL S-------DS KPTTAAVTH- -SPENLSSSP ETFVSRYAPD QPRKGSDVLV  93
LC  MAASAPTS-- IS--STITNR PS-----FSN SPKLLSKFVL PLPCNRRQKL SA------SR RRPLQISNVL S--------- ---S------ -ADKQILDGQ ETFVSRFAAD EPRKGCDVLV  86
P_  MAIPQTSNTS ISCKLTYQS- ---------- ----FTKLFL SI-------T EHMHLHHKQT LPKLTITNVL S--------- --------HS KPITTSQPLQ EPFVSRFGPD QPRKGADVLV  81
RV  MAATVPNPSP TV--SKYSSF ST-----KSA RNDLLSKFTF QFPQNQLKTI LHFQW-GRRR QHSPIITNVL S-------TS SSSS------ -RESSSTETP SSFVSRFASD EPRRGSDVLV  98
WH  MASSSPAAVA -SWSAMAASK PSPLPSSASS SRLLLPPFSK PFH-----PL PKLHH-QHQY RQRICVPAIA SS-------- ---------- -TTVAVPGAV SPPRRNFAPD EPRKGADILV  94

                                                                                              *        *
                      #                                                                       #        #
At  EALERQGVET VFAYPGGASM EIHQALTRSS SIRNVLPRHE QGGVFAAEGY ARSSGKSGIC IATSGPGATN LVSGLADALL DSVPLVAITG QVPRRMIGTD AFQGTPIVEV TRSITKHNYL  224
AW  EALEREGVTD VFAYPGGASM EIHQALTRST IIRNVLPRHE QGGVFAAEGY ARATGRVGVC IATSGPGATN LVSGLADALL DSVPLVAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  224
BT  EALEREGVTD VFAYPGGASM EIYQALTRSK TIRNVLPRHE QGGVFAAEGY ARATGFPGVC IATSGPGATN LVSGLADALL DSIPIVAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  212
FW  EALEREGVTD VFAYPGGASM EIHQALTRSR VIRNVLPRHE QGGVFAAEGY ARASGVPGVC IATSGPGATN LVSGLADALL DSVPIIAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  213
LC  EALEREGVTD VFAYPGGASM EIHQALTRSN IIRNVLPRHE QGGVFAAEGY ARASGLPGVC IATSGPGATN LVSGLADALL DSVPLVAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  206
P_  EALEREGVTN VFAYPGGATL EIHQALTRTS IIRNILPRHE QGGVFAAEGY ARASGYTGVC IATSGPGAIN LMSGLADAFL DSVPLVAITG QVARNMMGTD TFQDTAIVEA THSMTKHNYL  201
RV  EALEREGVTD VFAYPGGASM EIHQALTRSR IIRNVLPRHE QGGVFAAEGY ARASGKPGVC IATSGPGATN LVSGLADALL DSVPIVAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  218
WH  EALEREGVTD LFAYPGGASM EIHQALTRSP SITNHLLRHE QGEAFAASGY ARSTGRPGVC IATSGPGATN LVSALADALL DSIPMVAITG QVPRRMIGTD AFQETPIVEV TRSITKHNYL  214

At  VMDVEDIPRI IEEAFFLATS GRPGPVLVDV PKDIQQQLAI PNWEQAMRLP GYMSRMPKP- ---PEDSHLE QIVRLISESK KPVLYVGGGC LNSSDELGRF VELTGIPVAS TLMGLGSYPC  340
AW  VLDVEDIPRI VKEAFFLANS GRPGPVLIDI PKDIQQQLVV PNWDSNIKLG GYLSRLPKAK FTPTEEGLLD QIVRLVSEAK KPVLYTGGGC LNSSEELRKF VELTGIPVAS TLMGLGAYPC  344
BT  VMDVEDIPRV VREAFFLAKS GRPGPVLIDV PKDIQQQLVI PNWDEPMRLP GYLSRLPKL- ---PNEMLLE QIVRLISESK KPVLYVGGGC SFSSEELRRF VELTGIPVTS TLMGLGAFPT  328
FW  VLDVEDIPRI VREAFYLASS GRPGPVLIDI PKDIQQQLVV PKWDQPMKLS GYLSRLPKP- ---PIDGYLD QIVRLISESK KPVLYVGGGC LNSSEELRRF VELTGIPVAS TLMGLGSYPG  329
LC  VLDVEDIPRV VKEAFFIARS GRPGPVLIDV PKDIQQQMVV PNWDQPMSLG GYLSRLPKP- ---PSM-LLE QVIRLISESK KPVLYVGGGC LNSSEELRRF VELTGVPVAS TLMGLGSYPG  321
P_  VLNIQDIPRV VREAFYLASS GRPGPVLIDI PKDILQEFTV PNWDEPMRLA AYLSRLPKP- ---PNEPLLR QIVRLIKESK KPVLYVGGGC VNSSDELGRF VELTGIPVTN TLMGLGTYPG  317
RV  VMNVEDIPRI VREAFFLATS GRPGPVLIDI PKDIQQQMVV PDWDQPMRLA GYTSRLPRP- ---PNEPSLE QILRLIGESK KPVLYVGGGC LNSSKELRRF VDLTGIPVAS TLMGLGAFPT  334
WH  VLDVDDIPRI IKEAFFIATS GRPGPVLVDI PKDIQQQLAV PVWDPPVRLP GYVSRLPKP- ---PALHLLE QIIRFVSESS RPVLYVGGGC LHASEELRHF ADLTGIPIAS TLMGLGVYPL  330

                                                                                             *
                                       ##
At  DD-ELSLHML GMHGTVYANY AVEHSDLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDSA EIGKNKTPHV SVCGDVKLAL QGMNKVLENR AEELKLDFGV WRNELNVQKQ KFPLSFKTFG  459
AW  TN-DLSLHML GMHGTVYANY AVDNADLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDSA EIGKNKQPHV SICGDVKVAL QGLNKILESK KGKVELDFAS WRKELNEQKK KFPLSFKTFG  463
BT  GD-ELSLSML GMHGTVYANY AVDSSDLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDSA EIGKNKIPHV SICADIKLAL QGLNSILEGK EGKLKLDFSA WRNELNEQKV KYPLNYKTFG  447
FW  SS-DLSLQML GMHGTVYANY AVDKSDLLLA FGVRFDDRVT GKLETFASRA KIVHIDIDPA EIGKNKQPHV SVCGDVKIAL QGLNKILEEK RSVNNLDFSS WRKELDEQKV NYPLSFKTF-  447
LC  SDEEHALQML GMHGTVYANY AVDKSDLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDSA EIGKNKQPHV SICADLKLAL EGLNSILEG- --------KA WRQELKEQKT TFPLSYKTFG  432
P_  SH-DLSLQML GMHGTVYANY AIDKSDLLLA FGVRFDDRVT GKIETFASGA KIVHIDIDPA EIGKNKQPHL SICGDIKTAL QGINKILER- GESMNLDFSQ WKQEVMYQKA SNPLSFKTFG  435
RV  SDEELSLQML GMHGTVYANY SVDKSDLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDSA EIGKNKQPHV SLCADVKLAL EVLNSMLEAR KNKLKLDFSA WREELKEQKM KYPLSFKTFD  454
WH  DG-HLSLKML GMHGTVYANY SIDKADLLLA FGVRFDDRVT GKLEAFASRA KIVHIDIDPA EIGKNKQPHV SICGDIKLAL QEMNKIIEE- ---------T WREELDQQKK NYPLNYKTFG  439

                                                                                                                      #
At  EAIPPQHAIK VLDELTDGKA IISTGVGQHQ MWAAQFYNYK KPRQWLSSGG LGAMGFGLPA AIGASVANPD AIVVDIDGDG SFIMNVQELA TIRVENLPVK VLLLNNQHLG MVMQWEDRFY  579
AW  DAIPPQYAIQ MLDELTNGNA IISTGVGQHQ MWAAQFYKYR NPRQWLTSGG LGAMGFGLPA AIGAAVARPD AVVVDIDGDG SFIMNVQELA TIRVENLPVK IMLLNNQHLG MVVQWEDRFY  583
BT  DAIPPQYAIQ VLDELTNGSA IISTGVGQHQ ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  477
FW  ---------- --DELTGGNA IISTGVGQHQ MWAAQFYKYN RPRQWLTSAG LGAMGFGLPA AIGAAVARPD AIVVDIDGDG SFMMNVQELA TIRVENLPVK ILLLNNQHLG MVVQWEDRFY  555
LC  DAIPPQHAIQ VLDELTGGNA IISTGVGQHQ MWAAQFYKYN RPRQWLTSGG LGAMGFGLPA AIGAAVARPD AVVVDIDGDG SFIMNVQELA TIRVENLPIK IMLLNNQHLG MVVQWEDRFY  552
P_  DAIPPQYAIQ VLEELTGGNA IITTGVGQHQ MWAAQFYKYD RPRQWVTSGG LGAMGFGLPA AIGAAVARPD AVVVDIDGDG SFMMNVQELA TIRAENLPVK IMVINNQHLG MVVQWEDRYY  555
RV  DAIPPQYAIQ VLDELTKGNA IISTGVGQHQ MWAAQFYKYK NPRQWLTSGG LGAMGFGLPA AIGAAVARPD AVVVDIDGDG SFIMNVQELA TISVENLPVK IMLLNNQHLG MVVQWEDRFY  574
WH  DLIPPQYAIE VLDELTNGEA IITTGVGQHQ MWAAQYYNYK RPRQWLTSAG LGAMGFGLPA AVGAAVGNPG VTVVDIDGDG SFQMNAQELA IIRIENLDVK MLILNNQHLG MVVQWEDRFY  559

                                                                                    *
                                                                                    ##
At  KANRAHTFLG DPAQEDEIFP NMLLFAAACG IPAARVTKKA DLREAIQTML DTPGPYLLDV ICPHQEHVLP MIPSGGTFND VITEGDGRIK Y  670
AW  KANRAHTYLG NPSNSEQIFP DMLKFAEACD IPAGRVTKVE DLRAAIQTML DTPGPYLLDV IVPHQEHVLP MIPSGAAFKD TITEGDGRTS Y  674
BT  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- -  477
FW  KANRAHTYLG NPSKDTEIFP NMLKFAEACD IPAARVTQKA DLRAAIQKML DTPGPYLLDV IVPHQEHVLP MIPAGGGFMD VITEGDGRTK Y  646
LC  KANRAHTYLG NPSNEMEIFP NMLKFAEACD IPAARVTRKA DLRAAIQKML DTPGPYLLDV IVPHQEHVLP MIPSGGAFKD VITEGDGRTT Y  643
P_  NANRAQSFLG NPTKESEIFP DVVKFAEACD IPAARVTKKR NVRSAIQAML DTPGPYLLDV IVPHQEHVLP MIPPGGGFKD IITEGDGRT- -  644
RV  KANRAHTYLG NPSKESEIFP NMLKFAEACD VPAARVTKRE DLRAAIQKML DTPGPYLLDV IVPHQEHVLP MIPSGGAFKD VITEGDGRTK Y  665
WH  QSNRAHTYLG NPANESKVFP DFVKLADSYD IPAVRVSKKG EVRDAIRKML QTPGPYLLDV IVPHEEHVLP MIPSGGAFKD MILDGDGRTV H  650
```

FIGURE 3 | Alignment of *At*CSU and the CSU sequences from the seven weed species. The catalytic loop, the mobile loop, and the C-terminal arm are underlined (1, 2, and 3 lines, respectively). The residues that interact with ThDP (highlighted in cyan), FAD (yellow), and the herbicide chlorimuron ethyl (CE; green, including three positions bounded in green boxes) are highlighted. A hash (#) above the alignment represents the mutation sites responsible for resistance to herbicides. A red asterisk (*) above the alignment designates residues that are not fully conserved across the eight sequences. Legend: *A. thaliana* (At), alligator weed (AW), African boxthorn (BT), fireweed (FW), lantana (LC), parthenium (P), rubber vine (RV), and water hyacinth (WH).

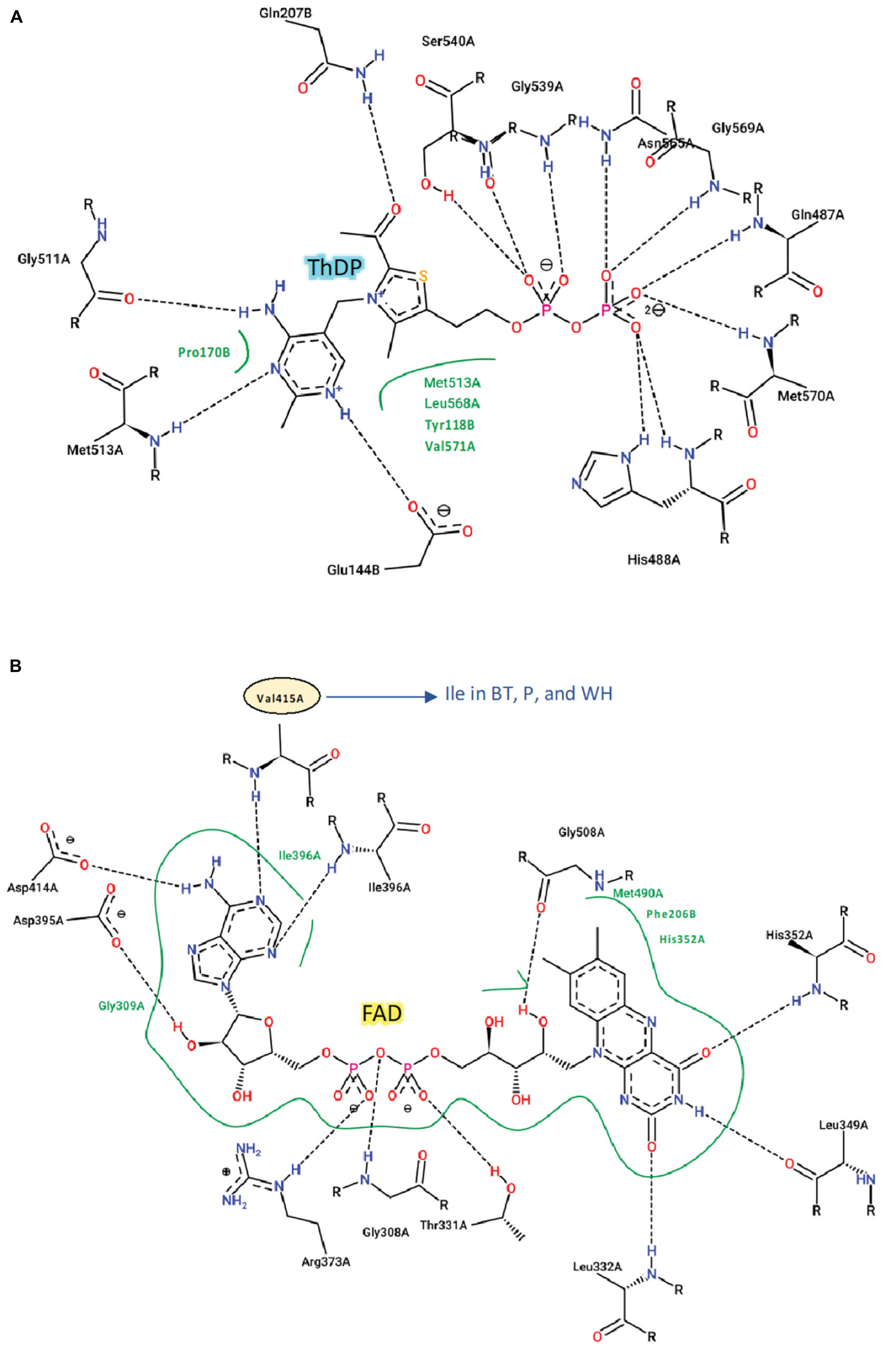


FIGURE 4 | Interactions of ThDP, FAD with *At*CSU. Two-dimensional representation of the interactions of **(A)** ThDP and **(B)** FAD with *At*CSU. Residues labeled in black and green form hydrogen bonds and hydrophobic interactions, respectively. BT, African boxthorn; P, parthenium; WH, water hyacinth.

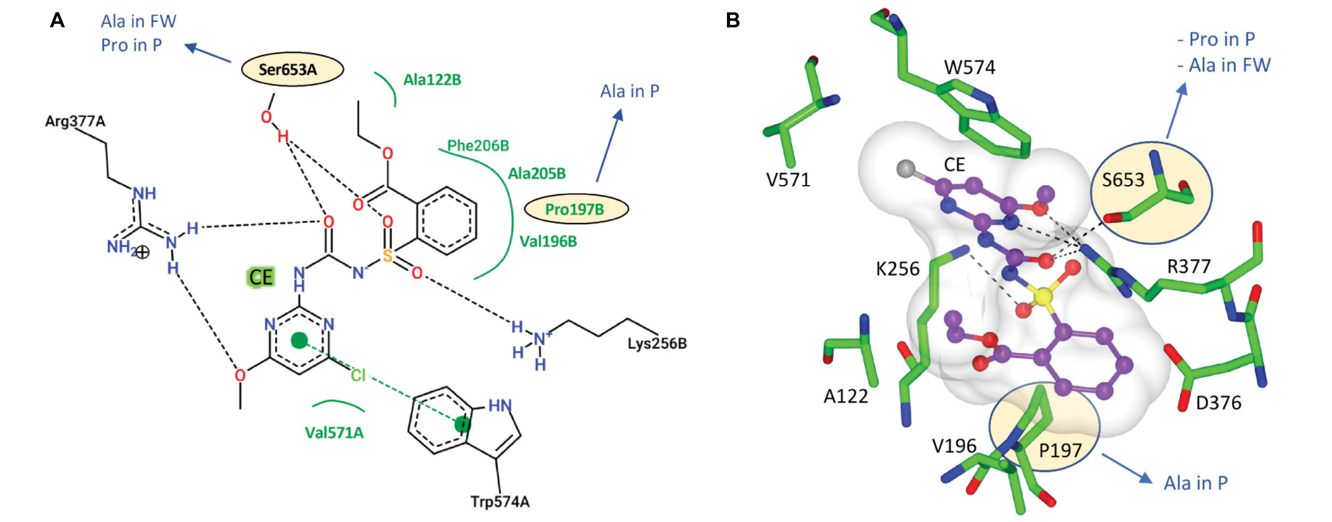


FIGURE 5 | Interactions of CE with *At*CSU. **(A)** Two-dimensional representation of CE interacting with *At*CSU. Residues labeled in black and green form hydrogen bonds and hydrophobic interactions, respectively. **(B)** Three-dimensional representation of CE interacting with *At*CSU (PDB ID 1YBH). The residues with observed variations in the seven weeds are circled. FW, fireweed; P, parthenium.

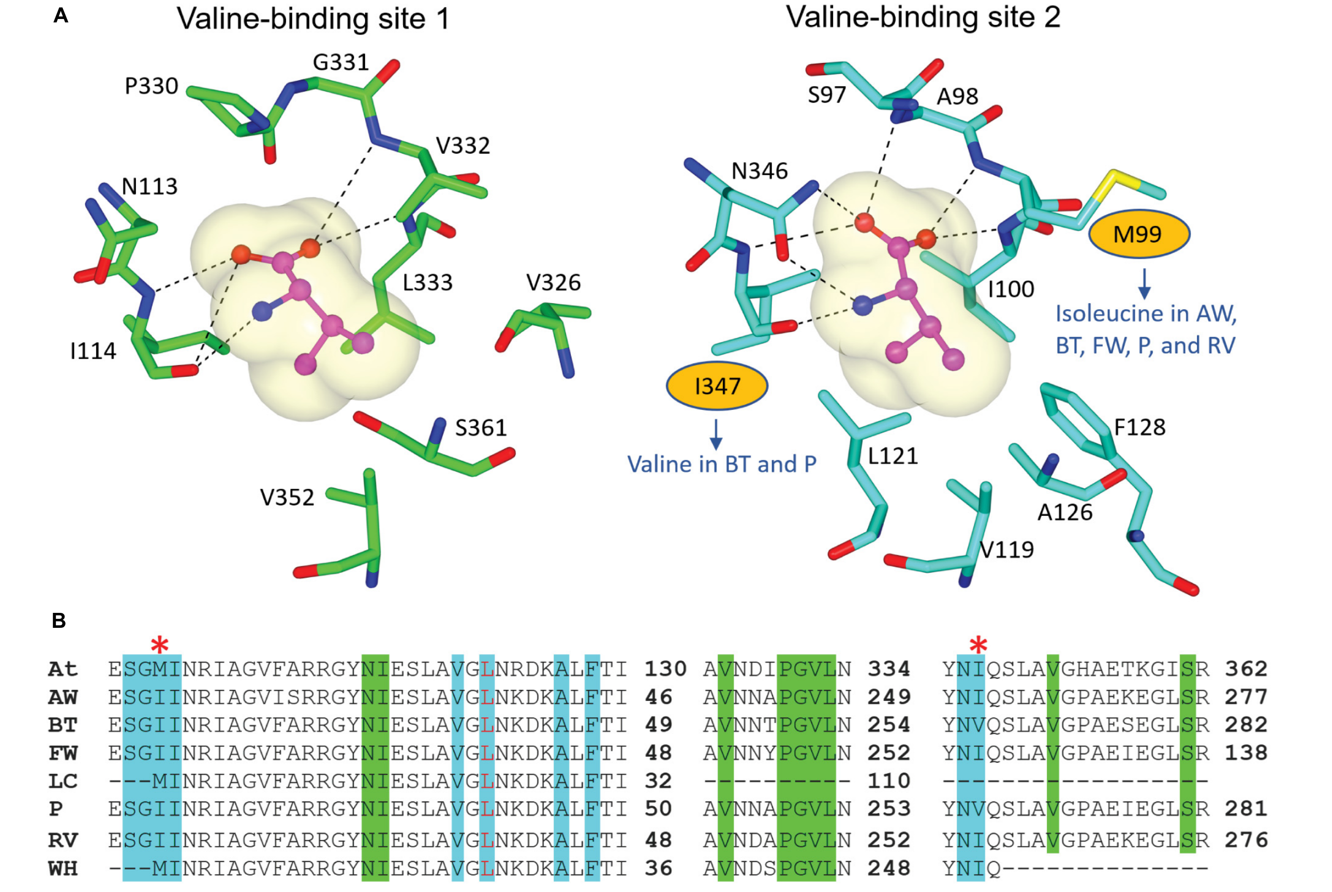

A
Valine-binding site 1
P330
G331
V332
N113
L333
V326
I114
S361
V352
Valine-binding site 2
S97
A98
N346
M99
I100
Isoleucine in AW, BT, FW, P, and RV
I347
Valine in BT and P
L121
F128
A126
V119
B
*
*
At ESGMINRIAGVFARRGYNIESLAVGLNRDKALFTI 130 AVNDIPGVLN 334 YNIQSLAVGHAETKGISR 362
AW ESGIINRIAGVISRRGYNIESLAVGLNKDKALFTI 46 AVNNAPGVLN 249 YNIQSLAVGPAEKEGLSR 277
BT ESGIINRIAGVFARRGYNIESLAVGLNKDKALFTI 49 AVNNTPGVLN 254 YNVQSLAVGPAESEGLSR 282
FW ESGIINRIAGVFARRGYNIESLAVGLNKDKALFTI 48 AVNNYPGVLN 252 YNIQSLAVGPAEIEGLSR 138
LC ---MINRIAGVFARRGYNIESLAVGLNKDKALFTI 32 ---------- 110 ------------------
P ESGIINRIAGVFARRGYNIESLAVGLNKDKALFTI 50 AVNNAPGVLN 253 YNVQSLAVGPAEIEGLSR 281
RV ESGIINRIAGVFARRGYNIESLAVGLNKDKALFTI 48 AVNDAPGVLN 252 YNIQSLAVGPAEKEGLSR 276
WH ---MINRIAGVFARRGYNIESLAVGLNKDKALFTI 36 AVNDSPGVLN 248 YNIQ--------------

FIGURE 6 | Residues involved in BCAA binding in the RSU. **(A)** Valine-binding sites in *At*RSU (PDB ID 6VZ8). The residues where there are variations in the weed sequences are circled. **(B)** Alignment of the RSU sequences of the seven weeds and *At*RSU. The residues that interact with valine in binding site 1 and 2 are highlighted in green and cyan, respectively. A red asterisk (*) above the alignment designates residues that are not fully conserved across the eight sequences. At, *A. thaliana*; AW, alligator weed; BT, African boxthorn; FW, fireweed; LC, lantana; P, parthenium; RV, rubber vine; WH, water hyacinth.

The CSU sequences were analyzed for the conservation of residues that are essential for the binding of herbicides and for ALS activity (i.e., the catalytic loop and the binding sites for the cofactors ThDP and FAD) (**Figure 3**). The *Arabidopsis thaliana* CSU (*At*CSU) was chosen as the model to analyze relevant structural features because it is the only plant CSU for which experimentally verified structural information (i.e., crystal structures) are available (McCourt et al., 2006; Garcia et al., 2017).

The catalytic loop [also known as the Q-loop (Lonhienne et al., 2018)] is highly conserved across all species (**Figure 3**), with F206 and Q207 (based on numbering of *At*CSU; this numbering is used hereinafter) being invariant. These residues directly interact with the substrate and have critical roles in catalysis (Lonhienne et al., 2017). Residues that interact with ThDP are also fully conserved (**Figure 4A**), and those that interact with FAD are highly conserved, with only one residue that is variable; Val415 in *At*CSU is replaced by isoleucine in African boxthorn, partenium, and water hyacinth (**Figure 4B**). Presumably, this modification has no effect on the binding of FAD as the interaction involves only the nitrogen atom of the backbone amide group of Val415. In general, the high conservation of all these sites strongly suggests that all the weed CSU sequences assembled and identified here are fully functional (**Figure 2**).

The CSU is the target for 58 commercial herbicides (Garcia et al., 2017). These herbicides fall into five different chemical classes, sulfonylureas (SUs), imidazolinones (IMIs), triazolopyrimidines (TPs), pyrimidinylbenzoates (PYBs), and the sulfonylamino-carbonyl-triazolinones (SCTs) (Garcia et al., 2017). Since the introduction of these herbicides in the 1980s, eight mutation sites that confer resistance have been reported (Heap, 2022). All mutations are in the catalytic pocket in regions that are in contact with the inhibitor (**Figure 3**). We chose chlorimuron ethyl (CE, a sulfonylurea) for investigation, as it is one of the most common commercial herbicides utilized in the field, and a three-dimensional structure has been solved in complex with *At*CSU (PDB identifier: 1YBH). The sequence alignment of the CSUs shows that amongst the 13 residues that interact with CE, which includes the eight residues for which mutations have been reported (Heap, 2022), three are variable (**Figure 3**). The first, corresponding to Pro197 in *At*, has been replaced by alanine in parthenium (**Figure 5A**). The second, corresponding to Ser653 in *At*, has been replaced by proline and alanine, in parthenium and fireweed, respectively (**Figure 5B**). The third corresponds to Ala205, a residue which to date has not been reported as a resistance mutation site. These results suggest that in all seven weed species, residues involved in herbicide binding are generally highly conserved. For five species (alligator weed, African boxthorn, lantana, rubber vine, and water hyacinth), there is no variation suggesting that the efficiency of herbicides should not be affected should alternate weed resistance mechanisms prevent the herbicide reaching the binding site. The CSUs coded in the genomes of fireweed and parthenium vary at S653 (**Figure 5B**), which is one of the major resistance sites that appear in weeds subjected to ALS-inhibiting herbicides (Heap, 2022). The S653 mutation affects the efficacy of the herbicides of the PYBs, IMIs, and SCTs families (Heap, 2022), thus these herbicides are likely ineffective for controlling parthenium and fireweed. The additional variation in parthenium, where P197 replaced by an alanine residue is also a site where mutation provides resistance, suggests that the ALS-inhibiting herbicides may not be suitable for application against parthenium.

Regulatory Subunit of Acetolactate Synthase

We recovered full-length RSU sequences from the genomes of alligator weed, parthenium, and water hyacinth, and partial sequences from those of the other four species (**Figure 2**). The RSU plays an important role in regulating the activity of ALS. Binding of RSU to the CSU increases ALS activity, but also allows

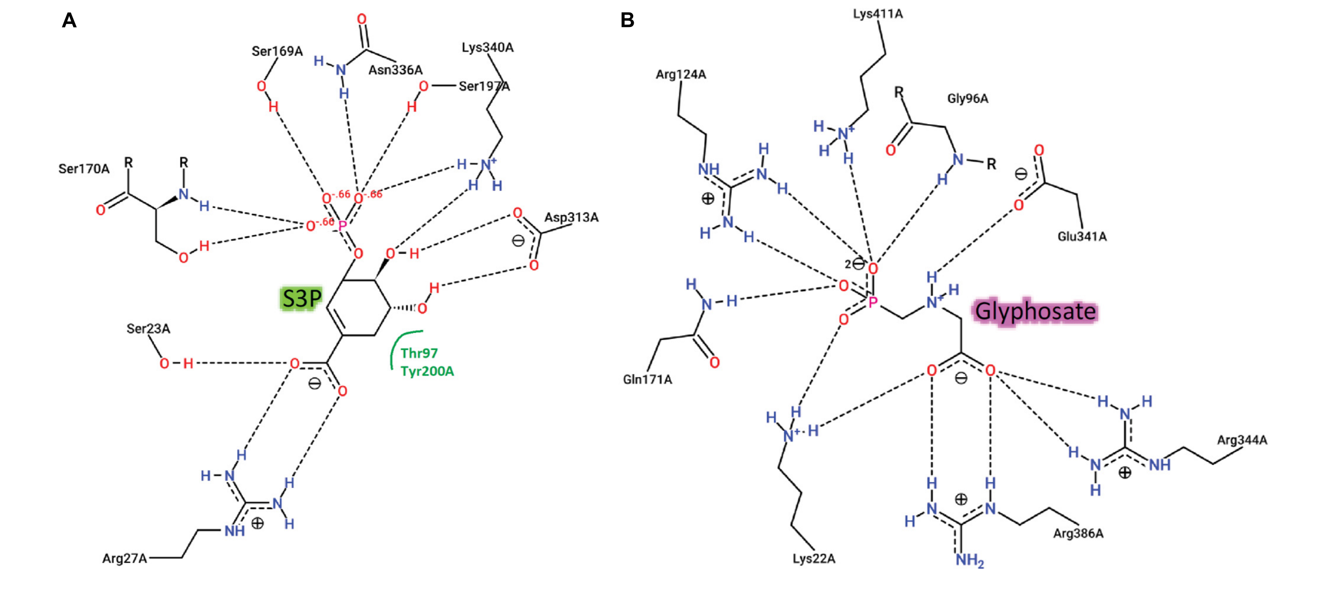


FIGURE 7 | Two-dimensional representations of the interactions of *E. coli* EPSPS with **(A)** the substrate, S3P and **(B)** the herbicide, glyphosate.

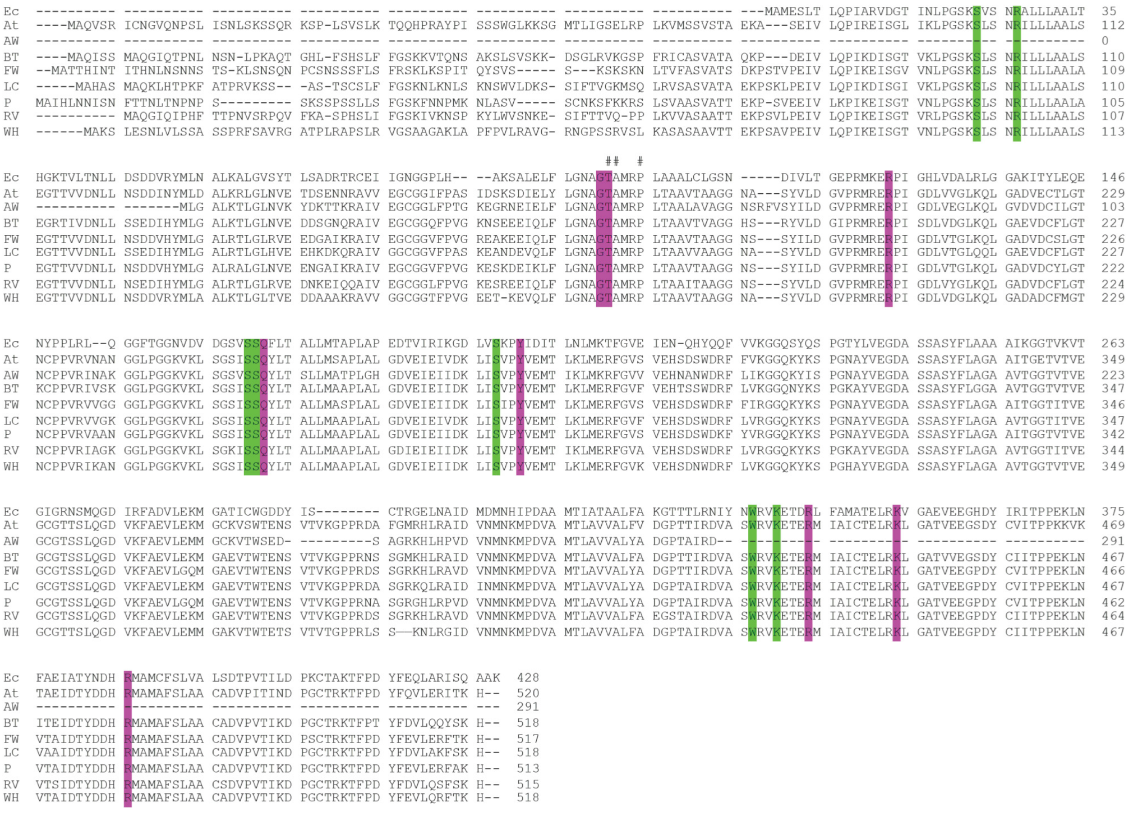
feedback inhibition by the BCCAs (Lonhienne et al., 2020). All recovered RSU sequences have the regions implicated in BCAA binding. There are two different binding sites for the BCAAs in the plant enzyme, both can bind valine, but their specificity for the individual BCAAs is likely to be different (Lonhienne et al., 2020). The alignment of eight RSU sequences (from seven weeds and *A. thaliana*) shows complete conservation of the residues at binding site 1 (**Figure 6A**), and two variable residues (M99 and I347) at binding site 2 (**Figure 6B**). Presumably, the variations at binding site 2 do not affect its functionality, because these residues interact with valine through backbone carbonyl and amide groups. The high degree of conservation across all seven weed species suggests that these sites are attractive for the continued development of novel inhibitors with potential as new herbicide leads.

EPSP Synthase, the Target of Glyphosate

We recovered full-length protein sequences of EPSP synthase in all genomes except for the genome of alligator weed, in which a partial sequence was recovered (*aroA* in **Figure 2**). EPSP synthase, an enzyme in the shikimate pathway involved in the synthesis of aromatic amino acids, is the target for glyphosate (Funke et al., 2006). EPSP synthase catalyzes the transfer of the enolpyruvyl moiety of phosphoenolpyruvate (PEP) to shikimate-3-phosphate (S3P), generating 5-enolpyruvyl-shikimate-3-phosphate (EPSP). Glyphosate competes with PEP, mimicking an intermediate state of the ternary enzyme-substrates complex (Funke et al., 2006). Currently, no crystal structure of a plant EPSP synthase is available, although the structure of *Escherichia coli* EPSP synthase in complex with S3P and glyphosate [PDB identifier: 1G6S (Funke et al., 2006)] has been solved, which identifies residues that interact with S3P and glyphosate (**Figure 7**). The *E. coli* EPSP synthase shared high sequence similarity with its counterpart protein in plants (**Figure 8**), thus presenting an excellent model for comparison. An alignment of the EPSP synthase sequences from *At* and the seven weed species shows that all residues involved in the binding of S3P and glyphosate are highly conserved (**Figure 8**). Three mutation sites that confer resistance have been reported (Heap, 2022). These three residues are also totally conserved (**Figure 8**). These results demonstrate that glyphosate should bind potently to the EPSP synthase from these weeds.

Vertical Inheritance of Genes Targeted by Herbicides

For each of the 12 proteins, we further assessed their evolutionary relationship among plant taxa. Similar to what we observed for the CSU (**Supplementary Figure 3**) and RSU



```
Ec  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---MAMESLT LQPIARVDGT INLPGSKSVS NRALLLAALT  35
At  ----MAQVSR ICNGVQNPSL ISNLSKSSQR KSP-LSVSLK TQQHPRAYPI SSSWGLKKSG MTLIGSELRP LKVMSSVSTA EKA---SEIV LQPIREISGL IKLPGSKSLS NRILLLAALS  112
AW  ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------  0
BT  ----MAQISS MAQGIQTPNL NSN-LPKAQT GHL-FSHSLF FGSKKVTQNS AKSLSVSKK- DSGLRVKGSP FRICASVATA QKP---DEIV LQPIKDISGT VKLPGSKSLS NRILLLAALS  110
FW  --MATTHINT ITHNLNSNNS TS-KLSNSQN PCSNSSSFLS FRSKLKSPIT QYSVS----- S---KSKSKN LTVFASVATS DKPSTVPEIV LQPIKEISGV VNLPGSKSLS NRILLLAALA  109
LC  -----MAHAS MAQKLHTPKF ATPRVKSS-- -AS-TSCSLF FGSKNLKNLS KNSWVLDKS- SIFTVGKMSQ LRVSASVATA EKPSKLPEIV LQPIKDISGI VKLPGSKSLS NRILLLAALS  110
P   MAIHLNNISN FTTNLTNPNP S--------- SKSSPSSLLS FGSKFNNPMK NLASV----- SCNKSFKKRS LSVAASVATT EKP-SVEEIV LKPIKEISGT VNLPGSKSLS NRILLLAALA  105
RV  ---------- MAQGIQIPHF TTPNVSRPQV FKA-SPHSLI FGSKIVKNSP KYLWVSNKE- SIFTTVQ-PP LKVVASAATT EKPSVLPEIV LQPIKEISGT VRLPGSKSLS NRILLLAALS  107
WH  ------MAKS LESNLVLSSA SSPRFSAVRG ATPLRAPSLR VGSAAGAKLA PFPVLRAVG- RNGPSSRVSL KASASAAVTT EKPSAVPEIV LQPIKEISGT VNLPGSKSLS NRILLLAALS  113

                                                                          ##   #
Ec  HGKTVLTNLL DSDDVRYMLN ALKALGVSYT LSADRTRCEI IGNGGPLH-- --AKSALELF LGNAGTAMRP LAAALCLGSN -----DIVLT GEPRMKERPI GHLVDALRLG GAKITYLEQE  146
At  EGTTVVDNLL NSDDINYMLD ALKRLGLNVE TDSENNRAVV EGCGGIFPAS IDSKSDIELY LGNAGTAMRP LTAAVTAAGG NA---SYVLD GVPRMRERPI GDLVVGLKQL GADVECTLGT  229
AW  ---------- -------MLG ALKTLGLNVK YDKTTKRAIV EGCGGLFPTG KEGRNEIELF LGNAGTAMRP LTAALAVAGG NSRFVSYILD GVPRMRERPI GDLVEGLKQL GVDVDCILGT  103
BT  EGRTIVDNLL SSEDIHYMLG ALKTLGLNVE DDSGNQRAIV EGCGGQFPVG KNSEEEIQLF LGNAGTAMRP LTAAVTVAGG HS---RYVLD GIPRMRERPI SDLVDGLKQL GAEVDCFLGT  227
FW  EGTTVVDNLL NSDDVHYMLG ALRTLGLRVE EDGAIKRAIV EGCGGVFPVG REAKEEIQLF LGNAGTAMRP LTAAVTAAGG NS---SYILD GVPRMRERPI GDLVTGLKQL GADVDCSLGT  226
LC  EGTTVVDNLL SSEDIHYMLG ALRTLGLHVE EHKDKQRAIV GGCGGVFPAS KEANDEVQLF LGNAGTAMRP LTAAVVAAGG NA---SYVLD GVPRMRERPI GDLVTGLQQL GAEVDCFLGT  227
P   EGTTVVDNLL NSDDVHYMLG ALRALGLNVE ENGAIKRAIV EGCGGVFPVG KESKDEIKLF LGNAGTAMRP LTAAVTAAGG NS---SYILD GVPRMRERPI GDLVTGLKQL GADVDCYLGT  222
RV  EGTTVVDNLL NSEDIHYMLG ALRTLGLRVE DNKEIQQAIV EGCGGLFPVG KESREEIQLF LGNAGTAMRP LTAAITAAGG NS---SYVLD GVPRMRERPI GDLVYGLKQL GADVDCFLGT  224
WH  EGTTVVDNLL NSDDVRYMLA ALKTLGLTVE DDAAAKRAVV GGCGGTFPVG EET-KEVQLF LGNAGTAMRP LTAAVTAAGG NA---SYVLD GVPRMRERPI GDLVDGLKQL GADADCFMGT  229

Ec  NYPPLRL--Q GGFTGGNVDV DGSVSSQFLT ALLMTAPLAP EDTVIRIKGD LVSKPYIDIT LNLMKTFGVE IEN-QHYQQF VVKGGQSYQS PGTYLVEGDA SSASYFLAAA AIKGGTVKVT  263
At  NCPPVRVNAN GGLPGGKVKL SGSISSQYLT ALLMSAPLAL GDVEIEIVDK LISVPYVEMT LKLMERFGVS VEHSDSWDRF FVKGGQKYKS PGNAYVEGDA SSASYFLAGA AITGETVTVE  349
AW  NCPPVRINAK GGLPGGKVKL SGSVSSQYLT SLLMATPLGH GDVEIEIIDK LISVPYVEMT IKLMKRFGVV VEHNANWDRF LIKGGQKYIS PGNAYVEGDA SSASYFLAGA AVTGGTVTVE  223
BT  KCPPVRIVSK GGLPGGKVKL SGSISSQYLT ALLMAAPLAL GDVEIEIIDK LISVPYVEMT LKLMERFGVF VEHTSSWDRF LVKGGQNYKS PGKAYVEGDA SSASYFLGGA AVTGGTVTVE  347
FW  NCPPVRVVGG GGLPGGKVKL SGSISSQYLT ALLMASPLAL GDVEIEIIDK LISIPYVEMT LKLMERFGVS VEHSDSWDRF FIRGGQKYKS PGNAYVEGDA SSASYFLAGA AITGGTITVE  346
LC  NCPPVRVVGK GGLPGGKVKL SGSISSQYLT ALLMAAPLAL GDVEIEIVDK LISVPYVEMT LKLMERFGVF VEHSDSWDRF LVRGGQKYKS PGNAYVEGDA SSASYFLAGA AVTGGTITVE  347
P   NCPPVRVAAN GGLPGGKVKL SGSISSQYLT ALLMAAPLAL GDVEIEIIDK LISVPYVEMT LKLMERFGVS VEHSDSWDKF YVRGGQKYKS PGNAYVEGDA SSASYFLAGA AITGGTVTVE  342
RV  NCPPVRIAGK GGLPGGKVKL SGKISSQYLT ALLMAAPLAL GDIEIEIIDK LISVPYVEMT IKLMERFGVA VEHSDSWDRF LVRGGQKYKS PGKAYVEGDA SSASYFLAGA AVTGGTITVE  344
WH  NCPPVRIKAN GGLPGGKVKL SGSISSQYLT ALLMAAPLAL GDVEIEIIDK LISVPYVEMT LKLMERFGVK VEHSDNWDRF LVKGGQKYKS PGHAYVEGDA SSASYFLAGA AVTGGTVTVE  349

Ec  GIGRNSMQGD IRFADVLEKM GATICWGDDY IS-------- CTRGELNAID MDMNHIPDAA MTIATAALFA KGTTTLRNIY NWRVKETDRL FAMATELRKV GAEVEEGHDY IRITPPEKLN  375
At  GCGTTSLQGD VKFAEVLEKM GCKVSWTENS VTVKGPPRDA FGMRHLRAID VNMNKMPDVA MTLAVVALFV DGPTTIRDVA SWRVKETERM IAICTELRKL GATVEEGSDY CVITPPKKVK  469
AW  GCGTSSLQGD VKFAEVLEMM GCKVTWSED- ---------S AGRKHLHPVD VNMNKMPDVA MTLAVVALYA DGPTAIRD-- ---------- ---------- ---------- ----------  291
BT  GCGTSSLQGD VKFAEVLEKM GAEVTWTENS VTVKGPPRNS SGMKHLRAID VNMNKMPDVA MTLAVVALFA DGPTTIRDVA SWRVKETERM IAICTELRKL GATVVEGSDY CIITPPEKLN  467
FW  GCGTSSLQGD VKFAEVLGQM GAEVTWTENS VTVKGPPRDS SGRKHLRAVD VNMNKMPDVA MTLAVVALYA DGPTTIRDVA SWRVKETERM IAICTELRKL GATVEEGPDY CVITPPEKLN  466
LC  GCGTSSLQGD VKFAEVLEKM GAEVTWTENS VTVKGPPRDA SGRKQLRAID INMNKMPDVA MTLAVVALYA DGPTAIRDVA SWRVKETERM IAICTELRKL GATVEEGPDY CVITPPEKLN  467
P   GCGTSSLQGD VKFAEVLGQM GAEVTWTENS VTVKGPPRNA SGRGHLRPVD VNMNKMPDVA MTLAVVALYA DGPTAIRDVA SWRVKETERM IAICTELRKL GATVEEGPDY CVITPPEKLN  462
RV  GCGTSSLQGD VKFAEVLEKM GAEVTWTENS VTVKGPPRDS SGRKHLRAVD VNMNKMPDVA MTLAVVALFA EGSTAIRDVA SWRVKETERM IAICTELRKL GATVEEGPDY CIITPPEKLN  464
WH  GCGTTSLQGD VKFAEVLEMM GAKVTWTETS VTVTGPPRLS S——KNLRGID VNMNKMPDVA MTLAVVALFA DGPTAIRDVA SWRVKETERM IAICTELRKL GATVEEGPDY CIITPPEKLN  467

Ec  FAEIATYNDH RMAMCFSLVA LSDTPVTILD PKCTAKTFPD YFEQLARISQ AAK  428
At  TAEIDTYDDH RMAMAFSLAA CADVPITIND PGCTRKTFPD YFQVLERITK H--  520
AW  ---------- ---------- ---------- ---------- ---------- ---  291
BT  ITEIDTYDDH RMAMAFSLAA CADVPVTIKD PGCTRKTFPT YFDVLQQYSK H--  518
FW  VTAIDTYDDH RMAMAFSLAA CADVPVTIKD PSCTRKTFPD YFEVLERFTK H--  517
LC  VAAIDTYDDH RMAMAFSLAA CADVPVTIKD PGCTRKTFPD YFDVLAKFSK H--  518
P   VTAIDTYDDH RMAMAFSLAA CADVPVTIKD PGCTRKTFPD YFEVLERFAK H--  513
RV  VTSIDTYDDH RMAMAFSLAA CSDVPVTIKD PGCTRKTFPD YFDVLQSFSK H--  515
WH  VTAIDTYDDH RMAMAFSLAA CADVPVTIKD PGCTRKTFPD YFEVLQRFTK H--  518
```

FIGURE 8 | Alignment of the EPSP synthase sequences of the seven weeds, *A. thaliana* and *E. coli*.The residues that interact with the S3P and glyphosate are highlighted in green and purple, respectively. A hash (#) above the alignment represents the mutation sites responsible for resistance to glyphosate (Heap, 2022). Ec, *E. coli*; At, *A. thaliana*; AW, alligator weed; BT, African boxthorn; FW, fireweed; LC, lantana; P, parthenium; RV, rubber vine; WH, water hyacinth.

(**Supplementary Figure 4**) of ALS, each phylogenetic tree of the seven proteins implicated in the BCAA pathway (TD, KARI, DHAD, BCAT-2, IPMS, IPMI, and IMDH; **Supplementary Figures 5–11**) suggests vertical inheritance of the genes, in which all seven weed species are positioned within monophyletic clades of their corresponding Orders among seed plants (spermatophytes) (The Angiosperm Phylogeny Group, 2009). We observed a similar trend in the protein trees for AroA (**Supplementary Figure 12**) and accA (**Supplementary Figure 13**), but not in the tree of psbA (**Supplementary Figure 14**), in which the various Orders are not grouped in monophyletic clades. In contrast to the other 11 genes that are encoded in the nuclear genome of plants, gene *psbA* is encoded in the chloroplast genome. Our tree (**Supplementary Figure 14**) supports the notion that psbA is not suitable as a phylogenetic marker for plants, in contrast to the established chloroplast marker gene *rbcL* (Hollingsworth et al., 2009).

CONCLUDING REMARKS

Our results demonstrate the use of whole-genome sequencing data from invasive weeds to quickly screen for potential gene and protein targets for herbicides, without the need for high-quality assembled genomes that are commonly cost-prohibitive in plants mainly because of their large genome sizes and the complication of polyploidy. Based on comparative sequence analysis, we identified conserved amino acids and target sites for the widely used herbicides (e.g., ALS inhibitors and glyphosate), and that most of these protein targets are evolutionarily conserved in plants through vertical inheritance. Compared to the other examined weed species, fireweed and parthenium, for which mutations were identified at the target sites of ALS protein, are likely more resistant to ALS-inhibiting and other herbicides. Although this hypothesis remains to be tested experimentally, given the high sequence conservation we observed among the seven phylogenetically diverse species, the proteins we identified present excellent models for further research without the need for expressing and purifying the proteins individually.

While we focused on targets of two major herbicides in this study, our approach and the data we generated can be readily applied to identify targets of *de novo* herbicides including novel leads that are under development, and other proteins relevant to herbicide tolerance (Werck-Reichhart et al., 2000; Dimaano and Iwakami, 2021). These data present a useful resource of sequence information that can be exploited for future efforts in herbicide discovery and development. They serve as a potential guide for selecting or designing herbicides, particularly in sustaining crop improvement.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/bioproject/, PRJEB47218. All assembled genomes, and the predicted gene and protein sequences are available at https://doi.org/10.48610/c22c10c.

AUTHOR CONTRIBUTIONS

TL, GS, LG, and CXC conceived the study and designed the research. SS, YC, and KD performed bioinformatic analysis of genome data, gene-target identification, and phylogenetic analyses. TL and YL performed computational analysis of protein-sequence conservation. C-EM and TL contributed to collection of biological material and data generation for this study. SS and TL prepared the first draft of the manuscript. C-EM, CW, GS, GW, LG, and CXC contributed to writing and iterative revisions of the manuscript. All authors reviewed and approved the final manuscript.

FUNDING

The Australian Government, through the Department of Agriculture, Water, and the Environment (CT-06), was gratefully acknowledged for funding. SS and YC were supported by Research Training Program scholarships. KD was supported by Australian Research Council grant DP190102474 awarded to CXC and Debashish Bhattacharya.

ACKNOWLEDGMENTS

We thank Jason Callander (Queensland Department of Agriculture and Fisheries) and Kelli Pukallus (Queensland Department of Agriculture and Fisheries and Tropical Weeds Research Centre) for providing some of the plant material.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.909073/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/s0022-2836(05)80360-2

Atlas of Living Australia (2022). *Species List: Weeds of National Significance (WoNS) [Online]*. Available: https://lists.ala.org.au/speciesListItem/list/dr823 (accessed June 7, 2022).

Beffa, R., Menne, H., and Köcher, H. (2019). "Herbicide Resistance Action Committee (HRAC): herbicide classification, resistance evolution, survey, and resistance mitigation activities," in *Modern Crop Protection Compounds*, eds P. Jeschke, M. Witschel, W. Krämer, and U. Schirmer (Weinheim: Wiley-VCH Verlag GmbH & Co), 5–32.

Biou, V., Dumas, R., Cohen-Addad, C., Douce, R., Job, D., and Pebay-Peyroula, E. (1997). The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 Å resolution. *EMBO J.* 16, 3405–3415. doi: 10.1093/emboj/16.12.3405

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., et al. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* 18, 188–196. doi: 10.1101/gr.6743907

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Chen, Y., González-Pech, R. A., Stephens, T. G., Bhattacharya, D., and Chan, C. X. (2020). Evidence that inconsistent gene prediction can mislead analysis of dinoflagellate genomes. *J. Phycol.* 56, 6–10. doi: 10.1111/jpy.12947

Dayan, F. E., and Duke, S. O. (2020). Discovery for new herbicide sites of action by quantification of plant primary metabolite and enzyme pools. *Engineering* 6, 509–514. doi: 10.1016/j.eng.2020.03.004

Dimaano, N. G., and Iwakami, S. (2021). Cytochrome P450-mediated herbicide metabolism in plants: current understanding and prospects. *Pest Manage. Sci.* 77, 22–32. doi: 10.1002/ps.6040

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comp. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fonseca, E. C. M., da Costa, K. S., Lameira, J., Alves, C. N., and Lima, A. H. (2020). Investigation of the target-site resistance of EPSP synthase mutants P106T and T102I/P106S against glyphosate. *RSC Adv.* 10, 44352–44360. doi: 10.1039/d0ra09061a

Funke, T., Han, H., Healy-Fried, M. L., Fischer, M., and Schonbrunn, E. (2006). Molecular basis for the herbicide resistance of Roundup Ready crops. *Proc. Natl. Acad. Sci. U.S.A.* 103, 13010–13015. doi: 10.1073/pnas.0603638103

Garcia, M. D., Nouwens, A., Lonhienne, T. G., and Guddat, L. W. (2017). Comprehensive understanding of acetohydroxyacid synthase inhibition by different herbicide families. *Proc. Natl. Acad. Sci. U.S.A.* 114, E1091–E1100. doi: 10.1073/pnas.1616142114

Garcia, M. J., Palma-Bautista, C., Vazquez-Garcia, J. G., Rojano-Delgado, A. M., Osuna, M. D., Torra, J., et al. (2020). Multiple mutations in the *EPSPS* and *ALS* genes of *Amaranthus hybridus* underlie resistance to glyphosate and ALS inhibitors. *Sci. Rep.* 10:17681. doi: 10.1038/s41598-020-74430-0

González-Pech, R. A., Stephens, T. G., Chen, Y., Mohamed, A. R., Cheng, Y., Shah, S., et al. (2021). Comparison of 15 dinoflagellate genomes reveals extensive sequence and structural divergence in family Symbiodiniaceae and genus *Symbiodinium*. *BMC Biol.* 19:73. doi: 10.1186/s12915-021-00994-6

Gooden, B., French, K., and Turner, P. J. (2009). Invasion and management of a woody plant, *Lantana camara* L., alters vegetation diversity within wet sclerophyll forest in southeastern Australia. *For. Ecol. Manage.* 257, 960–967. doi: 10.1016/j.foreco.2008.10.040

Heap, I. (2022). *The International Herbicide-Resistant Weed Database [Online]*. Available online at: https://www.weedscience.org/ (accessed June 7, 2022).

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35, 518–522. doi: 10.1093/molbev/msx281

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., et al. (2009). A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U.S.A.* 106, 12794–12797. doi: 10.1073/pnas.0905845106

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491. doi: 10.1186/1471-2105-12-491

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Li, D., Huang, Y., Leung, C. M., Luo, R., Ting, H. F., and Lam, T. W. (2017). MegaGTA: a sensitive and accurate metagenomic gene-targeted assembler using iterative de Bruijn graphs. *BMC Bioinformatics* 18:408. doi: 10.1186/s12859-017-1825-3

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., et al. (2015). The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* 350, 691–694. doi: 10.1126/science.aad0408

Lonhienne, T., Garcia, M. D., Noble, C., Harmer, J., Fraser, J. A., Williams, C. M., et al. (2017). High resolution crystal structures of the acetohydroxyacid synthase-pyruvate complex provide new insights into its catalytic mechanism. *ChemistrySelect* 2, 11981–11988. doi: 10.1002/slct.201702128

Lonhienne, T., Garcia, M. D., Pierens, G., Mobli, M., Nouwens, A., and Guddat, L. W. (2018). Structural insights into the mechanism of inhibition of AHAS by herbicides. *Proc. Natl. Acad. Sci. U.S.A.* 115, E1945–E1954. doi: 10.1073/pnas.1714392115

Lonhienne, T., Low, Y. S., Garcia, M. D., Croll, T., Gao, Y., Wang, Q., et al. (2020). Structures of fungal and plant acetohydroxyacid synthases. *Nature* 586, 317–321. doi: 10.1038/s41586-020-2514-3

Maia, V. H., Gitzendanner, M. A., Soltis, P. S., Wong, G. K.-S., and Soltis, D. E. (2014). Angiosperm phylogeny based on 18S/26S rDNA sequence data: constructing a large data set using next-generation sequence data. *Int. J. Plant Sci.* 175, 613–650. doi: 10.1086/676675

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

McCourt, J. A., and Duggleby, R. G. (2006). Acetohydroxyacid synthase and its role in the biosynthetic pathway for branched-chain amino acids. *Amino Acids* 31, 173–210. doi: 10.1007/s00726-005-0297-3

McCourt, J. A., Pang, S. S., King-Scott, J., Guddat, L. W., and Duggleby, R. G. (2006). Herbicide-binding sites revealed in the structure of plant acetohydroxyacid synthase. *Proc. Natl. Acad. Sci. U.S.A.* 103, 569–573. doi: 10.1073/pnas.0508701103

McLeod, R. (2018). *Annual Costs of Weeds in Australia*. Canberra, ACT: eSYS Development Pty Ltd, Centre for Invasive Species Solutions.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Ntalo, M., Ravhuhali, K. E., Moyo, B., Hawu, O., and Msiza, N. H. (2022). *Lantana camara*: poisonous species and a potential browse species for goats in southern Africa—a review. *Sustainability* 14:751. doi: 10.3390/su14020751

Ravet, K., Patterson, E. L., Krahmer, H., Hamouzova, K., Fan, L., Jasieniuk, M., et al. (2018). The power and potential of genomics in weed biology and management. *Pest Manage. Sci.* 74, 2216–2225. doi: 10.1002/ps.5048

Sainty, G., McCorkelle, G., and Julien, M. (1997). Control and spread of Alligator Weed *Alternanthera philoxeroides* (Mart.) Griseb., in Australia: lessons for other regions. *Wetl. Ecol. Manage.* 5, 195–201. doi: 10.1023/A:1008248921849

Sharma, O. P., Sharma, S., Pattabhi, V., Mahato, S. B., and Sharma, P. D. (2007). A review of the hepatotoxic plant *Lantana camara*. *Crit. Rev. Toxicol.* 37, 313–352. doi: 10.1080/10408440601177863

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. doi: 10.1038/msb.2011.75

Stierand, K., Maass, P. C., and Rarey, M. (2006). Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. *Bioinformatics* 22, 1710–1716. doi: 10.1093/bioinformatics/btl150

The Angiosperm Phylogeny Group (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* 161, 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* 35, 543–548. doi: 10.1093/molbev/msx319

Werck-Reichhart, D., Hehn, A., and Didierjean, L. (2000). Cytochromes P450 for engineering herbicide tolerance. *Trends Plant Sci.* 5, 116–123. doi: 10.1016/s1360-1385(00)01567-3

Yan, Y., Liu, Q., Zang, X., Yuan, S., Bat-Erdene, U., Nguyen, C., et al. (2018). Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. *Nature* 559, 415–418. doi: 10.1038/s41586-018-0319-4

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.