Received: 3 February 2016 | Revised: 21 January 2017 | Accepted: 10 March 2017

DOI: 10.1111/mec.14120

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

All roads lead to weediness: Patterns of genomic divergence reveal extensive recurrent weedy rice origins from South Asian *Oryza*

Zhongyun Huang[1] | Nelson D. Young[1] | Michael Reagon[2] | Katie E. Hyma[1] | Kenneth M. Olsen[3] | Yulin Jia[4] | Ana L. Caicedo[1]

[1]Department of Biology, University of Massachusetts, Amherst, MA, USA

[2]Department of Biology, Ohio State University Lima, Lima, OH, USA

[3]Department of Biology, Washington University, St. Louis, MO, USA

[4]Dale Bumpers National Rice Research Center, USDA-ARS, Stuttgart, AR, USA

Correspondence
Ana L. Caicedo, Department of Biology, University of Massachusetts, Amherst, MA, USA.
Email: caicedo@bio.umass.edu

Present address
Katie E. Hyma, LeClairRyan, Rochester, NY, USA

Funding information
Directorate for Biological Sciences, Grant/Award Number: IOS-1032023

Abstract

Weedy rice (*Oryza* spp.), a weedy relative of cultivated rice (*O. sativa*), infests and persists in cultivated rice fields worldwide. Many weedy rice populations have evolved similar adaptive traits, considered part of the 'agricultural weed syndrome', making this an ideal model to study the genetic basis of parallel evolution. Understanding parallel evolution hinges on accurate knowledge of the genetic background and origins of existing weedy rice groups. Using population structure analyses of South Asian and US weedy rice, we show that weeds in South Asia have highly heterogeneous genetic backgrounds, with ancestry contributions both from cultivated varieties (*aus* and *indica*) and wild rice. Moreover, the two main groups of weedy rice in the USA, which are also related to *aus* and *indica* cultivars, constitute a separate origin from that of Asian weeds. Weedy rice populations in South Asia largely converge on presence of red pericarps and awns and on ease of shattering. Genomewide divergence scans between weed groups from the USA and South Asia, and their crop relatives are enriched for loci involved in metabolic processes. Some candidate genes related to iconic weedy traits and competitiveness are highly divergent between some weed-crop pairs, but are not shared among all weed-crop comparisons. Our results show that weedy rice is an extreme example of recurrent evolution, and suggest that most populations are evolving their weedy traits through different genetic mechanisms.

KEYWORDS
adaptation, agricultural weeds, *Oryza*, parallel evolution, recurrent evolution, weedy traits

1 | INTRODUCTION

Agricultural weeds offer striking examples of rapid evolution and adaptation. Defined as unwanted plants growing in the agricultural environment (Monaco, Weller & Ashton, 2002)), agricultural weeds are responsible for a ~30% annual reduction in crop productivity worldwide (Oerke, 2006). Weedy plants' cost to society is a direct result of their ability to continuously infest and persist in crop fields. This rapid adaptation is thought to occur through the evolution of a suite of traits known as the 'agricultural weed syndrome' (Vigueira, Olsen & Caicedo, 2013). Understanding how agricultural weeds arise and evolve can help us design methods to prevent their adaptation to crop fields.

Although adaptive traits can vary among weed species, traits considered part of the agricultural weed syndrome are common to many weed groups. Examples of such traits include rapid growth, efficient seed dispersal and seed dormancy (Baker,1965). The evolution of similar traits in different groups is known as parallel or

convergent evolution (Arendt & Reznick, 2008). Recently, much interest has centred on determining the extent to which phenotypes involved in parallel evolution have similar or different genetic bases (e.g. Elmer & Meyer, 2011; Hoekstra, Hirschmann, Bundey, Insel & Crossland, 2006; Nachman, Hoekstra & D'Agostino, 2003; Protas et al., 2006). The repeated evolution of weed syndrome traits in agricultural weeds makes these ideal systems in which to address questions about the genetic basis of parallel evolution (Vigueira et al., 2013).

Weedy or red rice (*Oryza* spp.), an aggressive, interfertile weed of cultivated rice, displays signs of parallel phenotypic evolution (Ziska et al., 2015). Weedy rice infests cultivated rice (*O. sativa*) fields worldwide (FAO, 2002). However, morphological and molecular-based studies have increasingly suggested that the origin of weedy rice populations may vary across sites. For example, surveys of simple sequence repeat (SSR) markers and genomewide single nucleotide polymorphism (SNP) suggest that weedy rice groups found in the USA are most closely related to the *aus* and *indica* cultivated varieties, which are native to Asia (Gealy, Agrama & Eizenga, 2009; Londo & Schaal, 2007; Reagon et al., 2010), whereas studies based on isozymes and SSR markers suggest that weedy rice in Bhutan and northeastern China is related to genetically distinct *japonica* cultivated rice varieties (Cao et al., 2006; Ishikawa et al., 2005). Recent SSR data have further revealed genetic contributions from wild rice populations (*O. rufipogon*) to weedy rice backgrounds in Malaysia (Song, Chuah, Tam & Olsen, 2014) and Thailand (Pusadee, Schaal, Rerkasem & Jamjod, 2012).

Although surveys of weedy rice origins have been ongoing for many years, few have made use of high density genomewide variation. Additionally, many studies have failed to survey a broad enough range of *Oryza* germplasm to examine all possible origins. For example, early studies of US weedy rice did not include cultivated *aus* varieties, so their involvement in weed origins was only recently discovered (Londo & Schaal, 2007; Reagon et al., 2010). Similarly, wild germplasm has not been routinely included in many studies (Ziska et al., 2015). Moreover, the case of US weedy rice demonstrates that weed strains do not necessarily evolve from local *Oryza* groups; neither *indica* nor *aus* rice varieties are cultivated in the USA, indicating that US weeds are of exotic origin (Reagon et al., 2010).

Despite these limitations, the diversity of genetic backgrounds detected thus far for weedy rice populations is consistent with multiple independent evolutionary origins. Many characterized weedy rice populations reportedly have some traits consistent with the agricultural weed syndrome. Weedy rice traits include seed dormancy, a high proportion of seed dispersal (shattering), presence of a red pericarp, asynchronous maturity and enhanced growth (Chauhan, 2013; Rathore, Singh & Kumar, 2013). However, the extent to which independently evolved weedy rice populations worldwide share these typical traits remains an open question. Additionally, the extent to which similar genetic mechanisms have been involved in the evolution of convergent weedy rice traits has recently become an exciting area of inquiry (Qi et al., 2015; Thurber, Jia, Jia & Caicedo, 2013).

The extent of parallel evolution in weedy rice at the genetic and phenotypic levels cannot be understood without first elucidating the separate phylogenetic origins of weedy rice populations around the world. In this study, we focus on genomic and phenotypic characterization of weedy rice from South Asia, an area with great *Oryza* diversity. The wild ancestor of cultivated Asian rice, the *O. rufipogon*/*O. nivara* complex, grows natively in this region, and South Asia is believed to be the domestication site of the *indica* lineage of cultivated Asian rice, which comprises the *indica* and *aus* varieties (Civáň, Craig, Cox & Brown, 2015; Garris, Tai, Coburn, Kresovich & McCouch, 2005; Huang et al., 2012; Londo, Chiang, Hung, Chiang & Schaal, 2006; Zhu, Zheng, Luo, Gaut & Ge, 2007). In fact, South Asia harbours the greatest diversity of *indica* cultivars and is the only geographic area, where *aus* cultivars are grown (Khush, 1997). This is significant, because it raises the possibility that weedy *Oryza* from the USA and South Asia may be related. The second major lineage of cultivated rice, the *japonica*, composed of the *aromatic*, *tropical* and *temperate japonica* varieties, is believed to have been domesticated in China (Londo et al., 2006), but some *japonica* cultivars are also grown in South Asia.

With the technological shift from hand transplanting of paddy-grown seedlings to direct-seeded rice cultivation in recent years, weedy rice has emerged as a severe agricultural threat in South Asia (Chauhan, 2013). Because of the diversity present, the evolutionary dynamics of weedy, wild and cultivated *Oryza* could be more complex compared to other regions. Although weeds categorized as *Oryza* spp. have been reported in several countries in South Asia (Moody 1989), no regional characterization of weedy rice genetic diversity has been carried out, nor have the origins of weeds in this entire region been explored.

Using genome-scale genotyping on samples of South Asian wild, weedy and cultivated rice and leveraging previous data produced for US weedy rice (Burgos et al., 2014), here we attempt to answer the following questions: (1) From which *Oryza* groups has South Asian weedy rice arisen?, (2) How are US and South Asian weedy rice related?, and (3) Which loci have contributed to weedy rice evolution in South Asia and are these shared among different weedy groups?

2 | MATERIALS AND METHODS

2.1 | Plant material and DNA extraction

In this study, we focus on weeds in the South Asia geographic area, including Bangladesh, Myanmar, India, Nepal, Pakistan and Sri Lanka. We obtained seed for 59 South Asian samples classified as weedy rice from the International Rice Research Institute (IRRI), spanning collections made from 1963 to 1999 (Table S1). These samples were identified as weeds because all were unwanted noncrop *Oryza* growing within cultivated rice fields. Due to the homogeneity of rice cultivars, weedy *Oryza* are readily recognized. We further obtained seed for 77 cultivated *O. sativa* accessions that capture the diversity of cultivated varieties in South Asia and close neighbouring

countries, as well as 29 samples of the wild ancestral species to cultivated Asian rice (*O. rufipogon*/*O. nivara*) and four out-group samples (*O. meridionalis* and *O. barthii*) from the USDA Genetic Stocks *Oryza* Collection (GSOR) or IRRI (Table S1). Wild rice plants are distinct from weedy or cultivated rice, as they grow in natural habitats rather than within cultivated fields. We also included genotype information for six accessions from South-East Asia (K. M. Olsen, unpublished), to test for possible contributions of outside groups to South Asian weeds. Lastly, we included genotype data for 17 US weedy rice samples from the black hull awned (BHA) and straw hull awnless (SH) groups reported in Burgos et al. (2014). Our total number of samples was 186.

One individual per accession was grown at the University of Massachusetts Amherst. Approximately 100 mg of green leaf tissue was collected from each sample. A Retsch Mixer Mill MM400 with 3.2-mm stainless steel beads (BioSpec Products) was used for tissue grinding, and DNA extractions were performed with Qiagen DNeasy Miniprep Kits (Qiagen, MD, USA). DNA was quantified with a Qubit2.0 Fluorometer following the instructions in the Qubit dsDNA HS Assay Kit.

2.2 | GBS library preparation and sequence analysis

Genotyping by sequencing (GBS; Elshire et al., 2011) was performed at the Cornell University Institute of Biotechnology to detect genomewide polymorphisms. DNA samples were digested with the enzyme *Ape*KI, and the fragments were ligated with individual barcoded and common adapters. DNA fragments were pooled for PCR amplification, and 100-base pair (bp) fragments were single-end sequenced on an Illumina HiSeq 2000 platform. Initial data processing was also performed at Cornell with the standard Tassel pipeline (Bradbury et al., 2007). Reads were aligned to the MSU6 rice genome using Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009). Sites containing more than two SNP variants were eliminated. The minimum minor allele frequency was set to 1%. GBS quality results are included in Figure S1. Further filtering was performed in-house to remove SNP with >10% missing data and individuals with >95% missing data. SNP adjacent to mononucleotide repeats of five bp or more was also removed. We obtained a total of 51934 SNP, which were fairly evenly distributed among chromosomes (Figure S1). Raw reads were submitted to the NCBI Short Read Archive (SRA) under experiment (SRX576894).

2.3 | Population structure and phylogenetic analyses

High-quality SNP was analysed for population structure using STRUCTURE (version 2.3.3, Hubisz, Falush, Stephens & Pritchard, 2009). Due to data set size limitations of the program (Falush, Stephens & Pritchard, 2003; Pritchard, Stephens & Donnelly, 2000), we randomly selected approximately 10,000 SNP for each STRUCTURE analysis with an approximate 15,000 bp spacing. As cultivated and weedy *Oryza* are highly self-fertilizing, and even wild *Oryza* species have a tendency to self (Oka, 1974), we recoded heterozygous calls as 'N' and ran all simulations with the data coded as haploid. STRUCTURE was given no prior information on ancestral populations and was run with a model with admixture and no correlated allele frequencies. *K* values were varied from 1 to 15, and three replicates were run per *K* using a 100,000 burn-in period and 500,000 subsequent replications. The best *K* was detected based on Evanno, Regnaut and Goudet (2005) method. For comparison, we also analysed our complete SNP data set with the Bayesian clustering analysis FASTSTRUCTURE (version 1.0, Raj, Stephens & Pritchard, 2014) with no prior grouping. FASTSTRUCTURE runs were conducted for *K* from 1 to 15, and the optimal number of clusters was determined using the chooseK.py program in FASTSTRUCTURE.

SmartPCA from EIGENSOFT (Patterson, Price & Reich, 2006; Price et al., 2006) was applied to investigate the genetic divergence among individuals using the full set of SNP. The four out-group accessions were excluded from the principal component analysis (PCA), due to their outlier status.

Basic population genetics statistics for each *Oryza* subgroup determined by STRUCTURE and PCA results were calculated with ARLEQUIN (ver 3.5.2.2, Excoffier & Lischer, 2010) using the full set of SNP. AMOVA, molecular diversity indices and pairwise F_{ST} were computed for each subgroup.

RAXML (Randomized Axelerated Maximum Likelihood) version 8 (Stamatakis, 2014) was used to infer the phylogeny of the complete 186 *Oryza* accessions based on the full set of SNP. We used the RAXML HPC2 on XSEDE tool carried by the CIPRES portal http://www.phylo.org/, with a GTRGAMMA model and a bootstrap of 100. As our input sequences are concatenated SNP, we used the ascertainment bias correction (ASC) setting. The highest likelihood tree was plotted using ITOL v3 (Letunic & Bork, 2016).

2.4 | Phenotypic characterization of *Oryza* plants

All accessions were phenotyped for seed morphology. We classified hull colour as black or straw, seed pericarp colour as red or white, and awns as present or absent (Table S1). A subset of accessions chosen to represent *Oryza* groups identified by STRUCTURE and PCA was further phenotyped for five growth traits believed to be diagnostic or adaptive in weedy rice. The subset included 50 weedy rice, 14 *aus*, 14 *indica* and 16 *O. rufipogon*/*O. nivara* accessions. Three replicates of each accession were grown in a randomized design distributed across two Conviron PGW36 growth chambers, under 11-hour day length with 25°C temperature, until 30 days after flowering.

Days to flowering was measured as the number of days from germination to the time the first panicle emerges 50% from the sheath (Reagon, Thurber, Olsen, Jia & Caicedo, 2011). Height was measured at first flowering as the distance from soil surface to panicle base. Tiller number was also recorded at flowering. Emergence growth rate was calculated as plant height at 10 days divided by ten. Seed shattering was measured as breaking tensile strength (BTS) of seeds 30 days after flowering using the method described in Thurber et al. (2010); lower BTS values correspond to stronger

shattering. We randomly chose three seeds on three different panicles of the same plant, and the averages from nine seeds were used for each individual.

We used PCA to summarize the phenotypic divergence among the *Oryza* groups. The five phenotypic traits described above were decomposed into two primary axes of variation and plotted to show differentiation. All calculations were performed with the R package PCAMETHODS (Stacklies, Redestig, Scholz, Walther & Selbig, 2007), using the correction for missing data option. Raw data were mean centred and scaled based on unit variance before running the PCA.

To test for trait differences among groups, we first tested for chamber effects on each of the five traits. The parametric trait (height) was analysed by two-way ANOVA taking into account both chamber and *Oryza* group factors. For nonparametric traits with chamber effects, measurements from one chamber were analysed with Kruskal–Wallis tests; nonparametric traits without chamber effects were analysed for both chambers with Kruskal–Wallis tests.

2.5 | Coalescent analysis on the demographic history of weedy rice

To further investigate how demographic processes may have influenced the evolution of *aus*-like weedy rice, we used an approximate Bayesian computation (ABC) approach implemented in the program DIYABC v. 2.10 (Cornuet et al., 2014). We compared three demographic scenarios: (1) South Asian *aus*-like weedy rice evolving directly from within the *aus* cultivated rice group, (2) *aus*-like weeds evolving from wild *O. rufipogon/O. nivara*, and (3) *aus*-like weedy rice evolving from a hybridization event between *aus* and *O. rufipogon/O. nivara*. To reduce computing time, all analyses were based on a subset of 1,571 SNP (out of an original 51934 SNPs) that had a minimum minor allele frequency of 5% and no missing data. We considered our data set as haploid due to the high selfing rate in rice. Priors for timing of divergence between *aus* and *O. rufipogon/O.nivara* were based on previous estimates for rice domestication, and ranged from 10 to 15,000 years. We set the divergence time between weedy rice and its prospective progenitor population to be less than or equal to that of the timing of domestication and with a prior which ranged from 10 to 10,000 years, assumed to follow a uniform distribution. Demographic scenario selection and parameter estimates were based on a total of three million simulations (one million per scenario) as suggested by DIYABC instructions. Posterior probabilities of the three scenarios were calculated by direct estimation and logistic regression considering between 500 and 30,000 data sets that were closest to the observed values. Model selection was based on summary statistics transformed by linear discriminant analysis (LDA). Based on the demographic scenario with the highest posterior probability, we estimated the posterior distribution of all demographic parameters.

2.6 | F_{ST} scan and outlier detection

We performed population genomic scans to identify SNP-specific high F_{ST} outliers using both BAYESCAN v 2.1 (Foll & Gaggiotti, 2008) and LOSITAN (Antao, Lopes, Lopes, Beja-Pereira & Luikart, 2008), in order to compare the results obtained with these two distinct methods. LOSITAN uses the island model as a null distribution of F_{ST}, while BAYESCAN assumes that population have diverged independently from a common ancestor. Loci with minor allele frequency of <5% were removed from the data set. Sites with heterozygote calls occurring in more than 20% of the accessions were also removed to limit effects of possibly misaligned paralogous loci. F_{ST} scans were carried out separately for each weed group and its putative cultivated ancestor, and for the *aus* and *indica* cultivated groups.

For BAYESCAN, the 'snp' option was applied to recognize the matrix of SNP genotypes as input data. The analyses were run using default settings that included 20 pilot runs of 5,000 steps each, followed by 50,000 burn-in and 5,000 sampling steps with a thinning interval of 10. The prior odds parameters were set to the default of 10. False discovery rate (FDR) was set to 0.1 with the PLOT_BAYESCAN R function for outlier detection. For LOSITAN, 50,000 simulations were run on the same data set with default parameters. Both the 'neutral mean F_{ST}' and 'force mean F_{ST}' options were used. Loci outside the 95% confidence interval and those with F_{ST} = 1 were considered outliers.

High F_{ST} outlier SNP was considered candidates for positive selection under population divergence. We identified all genes containing outlier SNP based on the MSU6 reference genome annotation. To identify functional terms over-represented among the list of candidate genes, we performed gene ontology (GO) term enrichment with AGRIGO (Du, Zhou, Ling, Zhang & Su, 2010), using the *Oryza sativa* MSU6.1 nonTE genome as background. Significance was evaluated using a hypergeometric statistical test, with a Hochberg FDR multiple correction and a significant cut-off of 0.05; the minimum number of mapping entries was set to two.

3 | RESULTS

3.1 | The origins of South Asian weedy rice

We obtained 50,557 high-quality GBS SNP using the 165 Asian (South and South-East Asia) and out-group *Oryza* samples. To investigate population structure, we first used this entire set of SNP in a PCA (Figure 1b, Table S2). Two tight, but slightly overlapping clusters of *indica* and *aus* cultivated rice groups are evident, consistent with the close evolutionary relationship between these cultivars (e.g. McNally et al., 2009). The *japonica* cluster is well differentiated from *indica* and *aus*, but is more diffuse, likely because it contains accessions belonging to three cultivar groups within the *japonica* lineage: *tropical japonica*, *temperate japonica* and *aromatic*. In contrast to cultivated *Oryza*, wild rice accessions do not form a cluster and are scattered along the axes of both principal components, consistent with the higher levels of genetic diversity in this group (Caicedo et al., 2007; Huang et al., 2012). Some overlap with the various cultivated *O. sativa* groups is reflective of the status of this species complex as ancestral to domesticated rice. Interestingly, South Asian weedy rice is similarly scattered along both principal component axes. Many accessions overlap with the *indica* and *aus* groups and with various

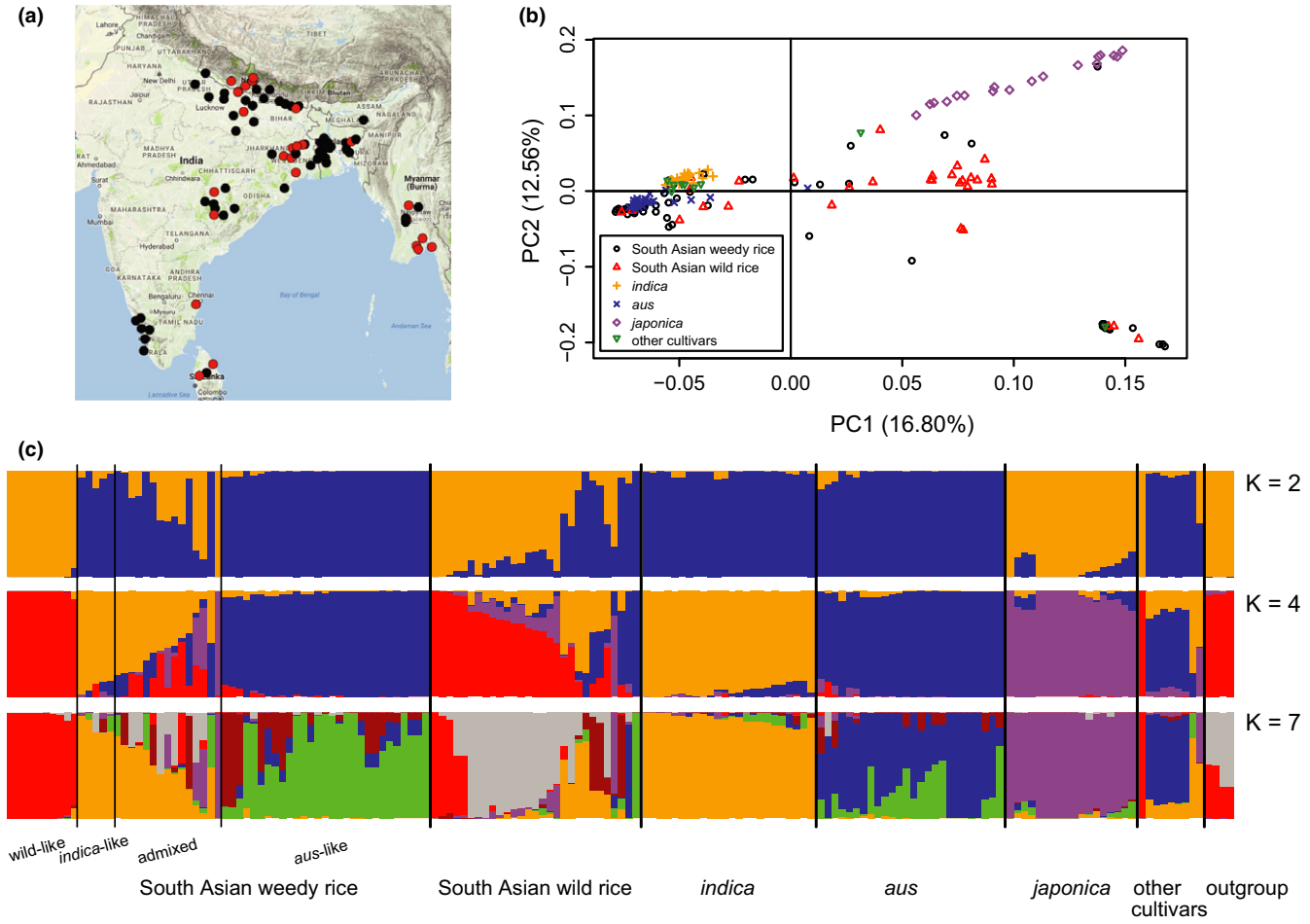

FIGURE 1 (a) Map of South Asia showing geographic collection localities for the weedy rice (black dots) and wild rice (red dots) used in this study. (b and c) Population structure of *Oryza* accessions from South and South-East Asia (59 weedy rice, 29 wild rice, 24 *indica*, 26 *aus*, 18 *japonica*, nine other cultivars and four out-group). (b) Principal component analysis excluding four out-group accessions. Principal component 1 (PC1) explains 16.80% variance and PC2 explains 12.56% variance. Cultivar identities are based on previous information and STRUCTURE $K = 4$ results. (c) Estimated population structure based on 10295 SNP. Each individual is represented by a coloured bar, with coloured partitions reflecting the relative proportion of genetic membership in a given cluster. Results are shown for $K = 2$, $K = 4$ and $K = 7$ clusters

wild rice accessions, but several weedy samples do not cluster with defined groups. The PCA result suggests that South Asian weeds have great heterogeneity in genetic background.

To further investigate relationships between South Asian weedy rice and other *Oryza* groups, we carried out STRUCTURE analysis on the same panel using a subset of 10,295 SNP. A clear peak in ΔK (Evanno et al., 2005) occurs at $K = 2$ populations (Table S3) and in this model the *indica* and *aus* groups are differentiated from *japonica*, a subset of wild rice, and the out-group species (Figure 1c, Table S4). Two types of South Asian weedy rice are also evident.

Because the Evanno method can underestimate K when there is hierarchical population structure (e.g. Vigouroux et al., 2008; Waples & Gaggiotti, 2006), we also examined population models at the two other ΔK peaks: $K = 4$ and $K = 7$ (Figure 1c, Table S3). The $K = 4$ model is generally consistent with PCA results (Figure 1b,c, Table S4). The three cultivated *O. sativa* groups, *indica*, *aus* and *japonica*, comprise three mostly distinct populations, and these groups are also largely differentiated from wild rice, which has a more heterogeneous genetic background. South Asian weedy rice comprises a mix of possible ancestries, with primary contributions from *aus*, *indica* and wild rice. While some weedy individuals have admixed backgrounds, three distinct subgroups of weedy rice can be differentiated based on our population genetics results: we have designated these as the *aus*-like, *indica*-like and wild-like weedy groups. The $K = 7$ results, which also correspond to the highest likelihood model, mirror those of $K = 4$, with greater heterogeneity in wild rice and in *aus* and wild rice ancestries in weeds but no identification of further groupings (Figure 1c, Table S4). Our STRUCTURE results were also consistent with results using FASTSTRUCTURE and the full set of SNP (Figure S2).

Both PCA and STRUCTURE detected *aus*-like (29), *indica*-like (5) and wild-like (10) groups among the South Asian weedy rice samples. Hereafter in the manuscript we define each of these weed groups as comprising individuals with at least 80% ancestry from each given

ancestral population in the $K = 4$ STRUCTURE results (Table S4). For wild-like weeds, this implies a minimum of 80% ancestry from the 'red' wild rice group (Figure 1c). Remaining weeds are classified as admixed (15). F_{ST} measures supported relationships between weedy groups and putative ancestors (Table S5). Basic population genetic statistics revealed similar levels of diversity in weedy groups and putative ancestral cultivar groups, suggesting that South Asian weeds may not have gone through overly strong bottlenecks (Table S6).

Most weedy rice samples in our study were collected from four distinct geographic regions: south India and Sri Lanka, central India, north India and Nepal, and northeast India and Bangladesh (Figure 1a, Table S1), consistent with the extent of rice agriculture in these regions. Additional collections came from Myanmar and Pakistan. Taking into consideration genetic similarity, it is evident that weedy rice samples tend to cluster within geographic regions of South Asia (Figure S3). A Fisher's exact test detected high correlation between weed population structure and the four main geographic regions ($p = .00028$; Table S7). *Aus*-like weeds are the most common group, but they are excluded from south India and Sri Lanka. *Indica*-like weeds only occur in the two northern regions. Wild-like weeds are confined to south India and Sri Lanka as well as north India and Nepal. The largest diversity of weed types occurs in north India and Nepal (Tables S1 and S5).

3.2 | Phenotypic characterization of South Asian weedy rice

To examine phenotypic trends across *Oryza* groups, we defined weedy groups as *aus*-like, *indica*-like and wild-like using the genetic structure criteria outlined above, and also limited each cultivar group to individuals with at least 80% ancestry from that group in the $K = 4$ STRUCTURE results. We excluded *japonica* from analyses, due to its very limited contribution to South Asian weeds. Due to great genetic heterogeneity, all wild rice samples were grouped together.

Among seed traits, red pericarp colour is one commonly associated with weedy rice (Ziska et al., 2015) and is a trait common in the wild ancestor of cultivated rice but rare in cultivated rice (Sweeney, Thomson, Pfeil & McCouch, 2006). As expected, most wild rice accessions in our study have red pericarps, while white pericarps dominate the *indica* cultivated variety (Table 1). Notably, however, many *aus* and *indica* cultivars from South Asia do have red pericarps despite their domesticated status. Red pericarps are dominant in all groups of South Asian weedy rice, but especially so in the *aus*-like and wild-like groups (Table 1).

Black hull colour, another trait common to wild rice (Zhu et al., 2011) is common in our wild *Oryza* group and its weedy relatives. Although *aus* accessions tend to be straw hulled, *aus*-like weeds are commonly black hulled (Table 1). In contrast, both *indica* cultivars and *indica*-like weedy rice tend to have straw coloured hulls. Awns are rare in cultivated groups, particularly *indica*, and common in wild rice. However, there is a high incidence of awns across weedy groups, with awns particularly dominant in wild-like and *aus*-like weeds (Table 1). In general, compared to cultivated groups, weedy rice groups have a high occurrence of red pericarp and awn presence. Wild-like weeds and wild rice share similar percentage of red pericarp, black hull and awn presence.

We characterized a subset of our accessions for various growth-related traits. While growth under artificial chamber conditions is unlikely to be identical to that in a native environment, we used this as a means to explore whether weedy rice populations differed phenotypically from their related groups (Table S8). In a PCA of the growth trait data, there is extensive evident scatter in all groups (Figure 2; Figure S4; Table S9). The 50% concentration ellipse suggests much phenotypic overlap between the *aus* and *indica* cultivar groups, and no overlap between these groups and wild rice. Weedy rice overlaps both with cultivars and wild rice, indicating phenotypic resemblance to the three *Oryza* groups. Geographically, weedy rice accessions from south India and Sri Lanka tend to cluster together along PC1, as well as central India (Figure S4). Other weed groups seem to have greater heterogeneity in growth traits.

In general, Asian weedy rice displays a moderate emergence growth rate and tiller number compared to cultivated and wild rice, a range of heights at flowering, a low to moderate number of days to flower and a high level of seed shattering (Table 2). Significant differences for some growth traits are evident among weed groups, and between weedy groups and their closest relatives. Wild-like weeds, in particular, shattered significantly more and flowered significantly earlier than at least one weed group in our conditions (Table 2, Table S8). Wild-like weedy rice also flowers significantly earlier than wild rice. For crop-like weeds, both *aus*-like and *indica*-like weedy rice shatter significantly more than their cultivar relatives, making this the phenotype where crop-like weeds most obviously diverged from their putative ancestors (Table 2). *Aus*-like weeds also flower significantly later than *aus* cultivars.

Despite the relative ease of seed shattering in weed groups, surprisingly low levels of shattering were observed in some weed samples (e.g. arr54, arr29; Table S8). This is unexpected, as easy shattering is a trait that is often considered diagnostic of weedy rice. In weeds classified as admixed based on STRUCTURE analyses, this could be due to introgression from crops. No obvious differences in genomic background based on STRUCTURE were observed between low-shattering and high-shattering samples belonging to *aus*-like or *indica*-like weed groups, however. Rare low-shattering individuals could represent accessions that have lost the high-shattering trait due to introgression with crops undetected by our analysis due to the overall similarity between crop and weed genomic backgrounds, or could be due to favouring of low-shattering genotypes when the weed seed is harvested with the crop and a portion of this seed is used for next year's cultivation.

3.3 | The relationship between US and South Asian weedy rice

Previous research has shown that two main genetically differentiated types of weedy rice occur in the United States, which are also largely distinguishable in morphological traits. Known as SH, for their

TABLE 1 Seed morphology characteristics observed in the South Asian *Oryza* groups defined by population structure analyses

Oryza groups	No. of accessions	Pericarp colour		Hull colour		Awn presence	
		Red	White	Black	Straw	Present	Absent
Weedy rice	59						
Aus-like	29	29 (100%)	0 (0%)	24 (83%)	5 (17%)	23 (79%)	6 (21%)
Indica-like	5	3 (60%)	2 (40%)	0 (0%)	5 (100%)	2 (40%)	3 (60%)
Wild-like	10	9 (90%)	1 (10%)	9 (90%)	1 (10%)	10 (100%)	0 (0%)
Admixed	15	9 (64%)	5 (36%)	8 (53%)	7 (47%)	9 (60%)	6 (40%)
Aus	26	12 (46%)	13 (50%)	3 (12%)	22 (85%)	5 (19%)	20 (77%)
Indica	24	5 (21%)	19 (79%)	5 (21%)	19 (79%)	0 (0%)	23 (96%)
Wild rice	29	28 (97%)	1 (3%)	23 (79%)	6 (21%)	24 (83%)	5 (17%)

Oryza group designation is based on at least 80% ancestry from each given ancestral group in the $K = 4$ STRUCTURE results (Table S4). Numbers represent the counts of accessions with the phenotype described in the column. Numbers in parentheses are percentages of accessions with the phenotype described.

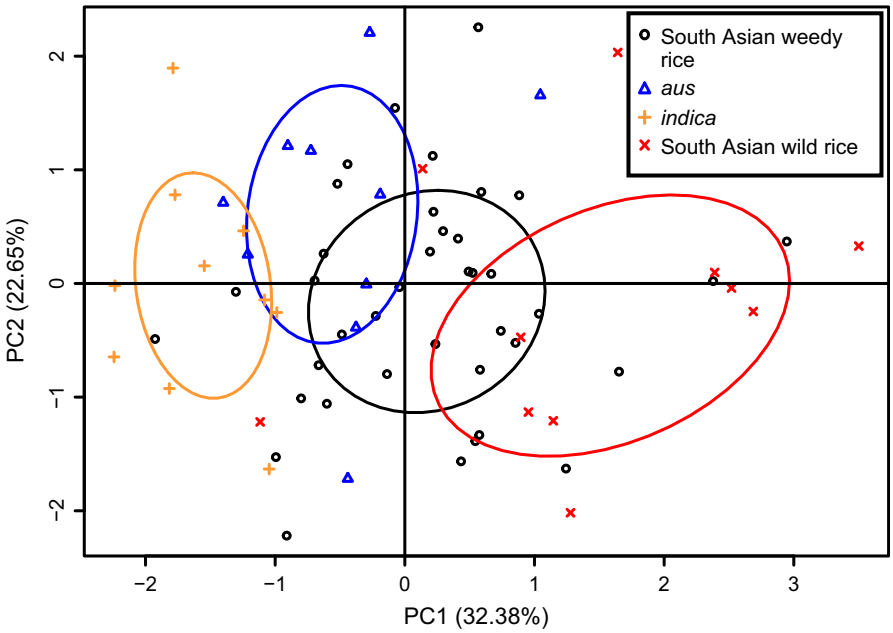

FIGURE 2 Principal component analysis for five growth-related traits (days to flowering, height at flowering, tiller number, emergence growth rate and seed shattering). A panel of 76 accessions including 39 South Asian weedy rice, 10 *aus*, 10 *indica*, three *japonica*, three other cultivars and 11 wild rice was used for phenotyping. Data points are labelled based on *Oryza* groups, and the 50% confidence ellipses for each group are shown

straw hull awnless seed morphology, and BHA, for their predominantly black hull awned seeds, these weedy groups seem to have arisen from *indica* and *aus* cultivated ancestors, respectively (Londo & Schaal, 2007; Reagon et al., 2010). However, neither of these domesticated rice varieties has ever been commercially cultivated in the USA, while both are typically grown in South Asia. Additionally, the occurrence of weedy rice populations in South Asia with genetic resemblance to *aus* and *indica* cultivated groups gives rise to the question of whether US weeds could be derived directly from South Asian weeds.

We carried out a PCA on 45,249 SNP from South Asian weedy rice (59) and US weedy rice of both the BHA (11) and SH (6) groups (Figure S5a, Table S10). The US weed groups are clearly differentiated along PC2, consistent with previous studies (Reagon et al., 2010). South Asian weeds are more diverse, but some lie close to US weeds (Figure S5a). A similar pattern is seen in the STRUCTURE analyses (Figure S5b, Tables S11 and S12). In the highest ΔK model, $K = 5$ populations, some Asian weedy rice groups share genomic background with BHA and SH, but Asian weeds also show more heterogeneous background than SH or BHA groups.

To further investigate relationships among all *Oryza* populations, we carried out a STRUCTURE analysis with all 186 weedy, wild and cultivated accessions from Asia and the USA. In the highest ΔK model with $K = 4$ populations (Figure S6, Tables S13 and S14), BHA weeds, *aus* cultivars and some Asian weeds clearly share ancestry, while SH, *indica* and a separate set of Asian weeds also share ancestry. Remaining Asian weeds seem more similar to wild rice. These results were also confirmed by FASTSTRUCTURE (Figure S2).

We also constructed a maximum-likelihood tree with the full set of SNP (51934) from all samples (Figure 3). As expected, BHA US weeds are nested within the clade that includes *aus* cultivars and related Asian weedy rice, and SH weeds are nested within a clade that includes *indica* cultivars and Asian weedy rice. Strikingly, both US weed groups are monophyletic and do not group with any Asian

TABLE 2 Average values for growth-related traits in the South Asian *Oryza* groups defined by population structure analyses

	No. of accessions	Emergence growth rate at 10 days (cm/day)[a]	Height at flowering (cm)[b]	Days to flower (day)[c]	Tiller number[d]	Shattering (g)[e]
p-Value (Kruskal–Wallis)		.095	NA	**1.10×10^{-5}**	**.018**	**2.08×10^{-12}**
p-Value (ANOVA)		NA	**.016**	NA	NA	NA
Aus-like weed	16	2.40 (0.73)	44.97 (11.81)^{ab}	127.77 (20.18)^{a}	8.56 (4.09)^{c}	4.61 (9.54)^{c}
Indica-like weed	5	2.40 (0.90)	43.72 (6.53)^{ab}	119.88 (8.63)^{ab}	10.13 (6.40)^{abc}	17.46 (16.71)^{b}
Wild-like weed	10	2.53 (0.80)	36.93 (10.97)^{b}	102.53 (17.60)^{c}	11.42 (4.90)^{ab}	4.20 (13.86)^{c}
Aus	10	2.61 (0.62)	48.92 (16.52)^{a}	112.22 (12.19)^{bc}	8.50 (2.59)^{c}	20.72 (18.04)^{ab}
Indica	10	2.73 (0.94)	42.91 (14.59)^{ab}	104.00 (14.53)^{bc}	6.60 (2.67)^{c}	29.69 (15.75)^{a}
Wild rice	11	1.87 (0.92)	44.40 (12.66)^{ab}	139.9 (43.27)^{a}	13.11 (5.93)^{a}	2.47 (5.37)^{c}

Numbers in parentheses represent standard deviations. Significant *p*-values are in bold.
Letters beside each measurement indicate significant differences between groups determined by Tukey HSD for normally distributed data and BH *p*-value adjustment method for non-normally distributed data.
[a]Chamber effect detected. Only measurements from chamber 1 were used in analyses. *p*-Value of group effect is reported.
[b]Chamber effect detected. The group factor *p*-value from a two-way ANOVA test is reported. The *p*-value for chamber factor is .00025, and the chamber × group *p*-value is .031.
[c]Chamber effect detected. Only measurements from chamber 1 were used in analyses. *p*-value of group effect is reported.
[d]Chamber effect detected. Only measurements from chamber 1 were used in analyses. *p*-value of group effect is reported.
[e]No chamber effect detected for this trait. *p*-Value of group effect is reported.

weeds. Instead, their sister taxa are cultivars, suggesting that both US weed groups stem from single colonization events in the USA, and both represent direct de-domestication events from cultivated ancestors. Curiously, of the two closest *indica* cultivars to SH weeds, one is from South-East Asia, suggesting the possibility that SH origins could be from South Asia or South-East Asia.

In contrast to US weeds, the relationship patterns seen for Asian weedy rice are more varied. Neither *indica*-like weeds nor *aus*-like weeds are monophyletic, which suggests they could have arisen more than once. All *indica*-like weeds are nested within clades of *indica* cultivars, suggesting de-domestication origins for these weeds (Figure 3). Some admixed weedy accessions with a high proportion of *indica* ancestry (e.g. arr82, arr27, arr37; Table S4) appear basal to the *indica* clade, suggesting origins from wild ancestors that gave rise to *indica*, or hybridization with wild rice. In contrast to *indica*-like weeds, no *aus*-like weed accessions nests within the main *aus* cultivar clade (Figure 3). Instead, most *aus*-like weeds form a sister clade to *aus* cultivars and US weedy rice, and some are basal to the entire *aus* and *aus*-like clade.

Because the origin of *aus*-like weeds is not immediately obvious from the tree topology and observed bootstrap support, we carried out coalescent analyses to determine if *aus*-like weeds arose through de-domestication from *aus* cultivars, descent from wild ancestors prior to domestication of *aus*, or hybridization between wild and cultivated rice (Figure S7). Results strongly supported a scenario of de-domestication, in which the majority of *aus*-like weeds arose from within the cultivated *aus* group (logistic regression: 0.7796) (Figure S7, Table S15). The estimate for divergence times, while recent compared to other studies on *aus* domestication (~6,000 years) (Choi et al., 2017), support weedy rice divergence after the divergence of *aus* from *O. rufipogon* (Table S16).

As suggested by the population structure analyses, wild-like weedy rice samples from South Asia occur within a clade that also contains *O. rufipogon* and *O. nivara*, and are clearly more closely related to these than to the out-group species (Figure 3). Thus wild-like weedy rice likely descends directly from wild populations.

3.4 | F_{ST} outlier scans for potential weedy trait-related loci under selection

We conducted F_{ST} outlier scans to detect loci that are highly differentiated between the various weed groups and their putative ancestral/closest relative groups. We excluded wild-like weedy rice and focused on weed-crop comparisons from both the USA and Asia, for two reasons. First, many of the traits favoured during domestication are traits that seem to have been reversed during weed evolution (Reagon et al., 2011; Thurber et al., 2010; Ziska et al., 2015); thus, we expect clear signals of positive selection on genes underlying such traits in weed strains that are descended from or related to cultivated rice. Second, the four weed-crop comparisons include weed groups that originated separately from similar cultivated ancestors (e.g. BHA and *aus*-like weeds; SH and *indica*-like weeds), as well as weed groups that have adapted to the same geographic areas (e.g. BHA and SH; *aus*-like and *indica*-like); this provides a framework for examining the extent of parallel genetic evolution in each case. In order to identify loci and functional terms exclusive to weed evolution rather than divergence between any two *O. sativa* populations, we also conducted F_{ST} outlier scans on *aus* vs *indica* cultivars.

In all cases, LOSITAN yielded a greater number of high F_{ST} SNP outliers than BAYESCAN (Table 3). No clear pattern was observed between methods for outlier numbers among different comparisons. For all comparisons, we identified the set of outlier SNP shared between LOSITAN and BAYESCAN (hereafter overlap outliers), and the set composed of SNP that appear in the LOSITAN or BAYESCAN results (hereafter union outliers). In general, a high proportion of outlier

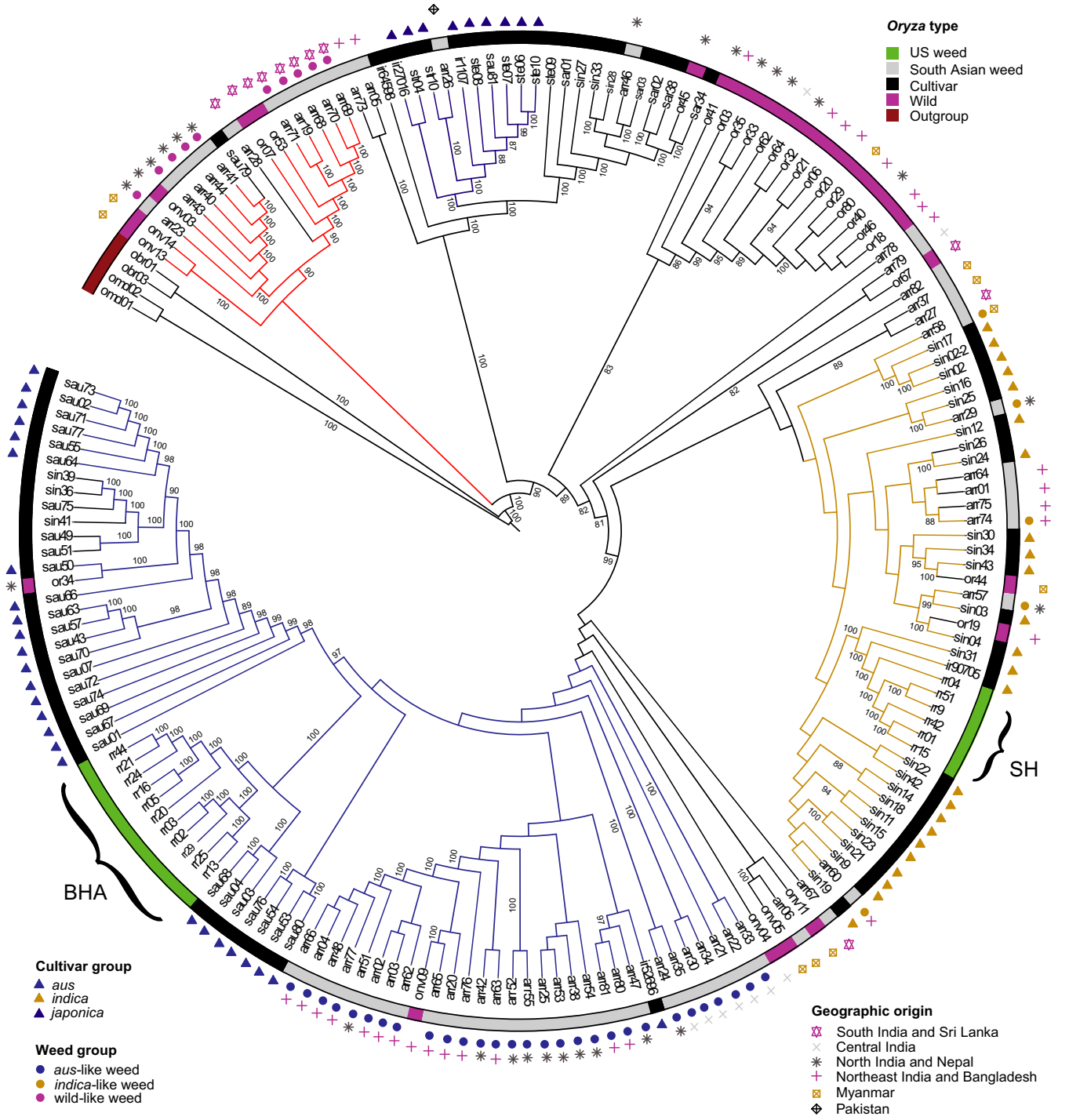

FIGURE 3 Maximum-likelihood tree of 186 *Oryza* accessions (59 South Asian weedy rice, 29 wild rice, 77 cultivated rice, 11 BHA US weedy rice, six SH US weedy rice and four out-group) based on all SNP obtained through GBS. Outer colour ring represents different *Oryza* groups, as indicated in the figure key. Coloured triangles beside accession IDs represent cultivar groups, and coloured filled circles represent weed groups. The geographic origins of weedy rice accessions are labelled with a set of coloured shapes as indicated in the figure key. Branches with bootstrap values >80% are labelled. Branch colours within clades correspond to the predominant ancestry colour (>80%) in the STRUCTURE $K = 4$ results shown in Figure 1c; branches leading to accessions with admixed ancestry are shown in black

SNP was found to be located within gene coding regions, consistent with the proportion (~70%) among all SNP tested (Table 3). We thus confined ourselves to gene coding regions and identified the genes containing outlier SNP. These genes were considered to be possible candidates for evolution under positive selection during weed-crop divergence or crop variety divergence.

We first focused on genes in the overlap lists (Table S17). The number of SNP outliers overlapping between the two methods was

TABLE 3 Summary of high F_{ST} SNP outliers from BAYESCAN and LOSITAN analyses

Comparisons between *Oryza* groups[a]	Total SNP tested	BAYESCAN F_{ST}	LOSITAN F_{ST}[b]	No. of outliers detected by BAYESCAN	No. of outliers detected by LOSITAN	Overlap outliers[c]	Union outliers[d]
Aus-like (29) vs *aus* (25)	16370 (11231)	0.19	0.21	115 (87)	882 (577)	33 (22)	964 (636)
BHA (11) vs *aus* (25)	18086 (12094)	0.35	0.34	45 (31)	2788 (1917)	17 (11)	2816 (1937)
Indica-like (5) vs *indica* (23)	15480 (10627)	0.06	0.06	14 (13)	2797 (1910)	14 (13)	2797 (1910)
SH (6) vs *indica* (23)	15077 (10312)	0.39	0.42	0 (0)	3974 (2773)	0 (0)	3974 (2773)
Aus (25) vs *indica* (23)	22819 (15752)	0.39	0.49	52 (41)	466 (323)	45 (32)	473 (329)

Numbers in parentheses represent outlier SNP that lie within gene coding regions.
[a]Sample sizes are indicated in parentheses.
[b]Values correspond to the dataset F_{ST} values reported by LOSITAN.
[c]Numbers of outliers that overlap between the BAYESCAN and LOSITAN results.
[d]Numbers of outliers that form the union of both BAYESCAN and LOSITAN results.

small, and no shared outliers were detected for the SH-*indica* comparison. For the remaining weed-crop comparisons, no candidate genes were shared among the overlap lists. We looked at possible gene function in each list by considering rice genome annotations, biological process gene ontology (GO) terms, and the function of *Arabidopsis thaliana* orthologous genes (Table S17). In general, no gene function or lower level GO term is shared exclusively among weed-crop comparisons and not present in the crop–crop comparisons. However, there are a number of genes potentially affecting pollen germination and tube growth that occur across all comparisons (Table S17). Other processes shared by more than one pair of comparisons include hormone (particularly gibberellin) pathways, and biotic and abiotic stress responses. An interesting trend is that in most comparisons, series of candidate genes are located closely in the genome. This suggests that some outlier SNPs lie in areas likely that have undergone selective sweeps, compounding the already extensive levels of linkage disequilibrium (LD) that exist in rice (LD breakdown in different cultivar groups can range from 75 to 500 Kb (Mather et al., 2007)), and making it more difficult to identify the gene targeted by selection. In all, however, the overlap lists results do not reveal any trends exclusive to weed evolution.

Because BAYESCAN and LOSITAN use different methods to detect outliers, we then focused on genes containing outlier SNP from the union lists. Nine candidate genes were shared among all four weed-crop comparisons and were not present in the crop–crop comparison (Table 4). No functional trend was observed for these shared genes. We also examined union lists for each comparison for over-represented biological process (BP) GO terms. The *aus-indica* comparison had the smallest list of union outliers, and there were no significantly over-represented biological process GO terms. In contrast, between 13 and 61 significantly over-represented terms were detected for each crop-weed comparisons (Table S18). The top ranked significant BP GO terms were strikingly similar among comparisons, with many terms related to metabolic processes, and the lowest level shared GO term among all being protein amino acid phosphorylation (Table 5). Many of these metabolism-related GO terms were also present in *aus-indica* F_{ST} outliers, although none were significantly over-represented as with the crop-weed comparisons.

TABLE 4 Candidate genes highly divergent across all four weed-crop comparisons that are not outliers in the crop–crop comparison

MSU locus	Annotation	GO term
LOC_Os03 g03920	Ubiquitin family domain containing protein, expressed	Molecular function
LOC_Os03 g12180	MA3 domain containing protein, expressed	Biological process
LOC_Os03 g12440	Zinc-binding protein, putative	Biological process
LOC_Os06 g05750	Transferase family domain containing protein, expressed	Metabolic process
LOC_Os06 g17220	UDP-glycosyltransferase, putative, expressed	Metabolic process
LOC_Os07 g01710	Phytosulfokine receptor precursor, putative, expressed	Response to stress, signal transduction, protein modification process
LOC_Os08 g18060	Expressed protein	NA
LOC_Os11 g14180	Expressed protein	NA
LOC_Os12 g35030	Plus-3 domain containing protein, expressed	Flower development, biosynthetic process, nucleobase, nucleoside, nucleotide and nucleic acid metabolic process

As a separate approach, we looked for GO term over-representation in the list of shared outlier genes in weed-crop comparisons grouped by ancestry or geographic region, excluding genes present in the *aus-indica* comparison (Table S19). Shared genes (105) between the two weed-crop comparisons with *aus* ancestry (*aus*-like vs *aus* and BHA vs *aus*), had no significantly over-represented GO terms. Shared genes between the two weed-crop comparisons with *indica* were over-represented for numerous metabolic process related terms, much like for individual weed-crop comparisons. When considering outlier genes shared between weed-crop comparisons occupying the same geographic region, metabolic process

TABLE 5 Significantly enriched biological process GO terms shared among all four weed-crop comparisons

Term	Number of loci			
	Aus-like vs *aus*	BHA vs *aus*	*Indica*-like vs *indica*	SH vs *indica*
Metabolic process	142	408	483	570
Primary metabolic process	113	345	400	475
Macromolecule metabolic process	95	290	333	390
Cellular macromolecule metabolic process	81	253	302	338
Macromolecule modification	37	113	129	139
Protein modification process	36	113	126	136
Post-translational protein modification	34	108	117	127
Phosphorus metabolic process	33	105	110	127
Phosphate metabolic process	33	105	110	127
Protein amino acid phosphorylation	32	99	101	114
Phosphorylation	32	103	106	121

related terms were again significantly over-represented (Table S19). However, an additional category of GO terms related to immunity was over-represented among outlier genes shared by US weed-crop lists; these were driven by the presence of two shared outlier genes (LOC_Os07g11510 and LOC_Os07g11410) which have been implicated as seed allergenic proteins (Wang, Yang, Zhao, Li & Zhang, 2014; Table S19).

3.5 | Divergence in genes from previously characterized weedy syndrome related pathways

As evidenced by our phenotypic survey (Tables 1 and 2), certain traits differentiate Asian weed groups from their cultivated relatives. Likewise, US weedy rice groups have been documented as having greater seed shattering, higher incidence of red pericarps and divergent flowering times compared to their ancestral groups (Gross et al., 2010; Reagon et al., 2011; Thurber, Reagon, Olsen, Jia & Caicedo, 2014; Thurber et al., 2010). Because traits that differentiate cultivated from weedy rice are often similar to those that differentiate cultivated from wild rice, candidate genes associated with some of these traits have been discovered in rice domestication studies. Following extensive literature surveys, we thus examined if genes from pathways potentially involved in weedy syndrome phenotypes contained SNP detected as outliers by BAYESCAN or LOSITAN in any of our weed-crop comparisons. We focused on genes involved in iconic weedy rice traits including seed shattering, pericarp and hull pigmentation, seed dormancy, and flowering time variability, and genes involved in growth and competitiveness traits including tillering, starch and cellulose synthesis, and chlorophyll synthesis. For all candidate genes examined, absence of association with an outlier SNP can be due to either absence of genotyped SNP in the locus, or lack of differentiation in SNP genotyped at that locus (Table 6, Table S20). Below we discuss only traits for which outlier SNP were detected. We also verified that no gene identified as a crop-weed outlier was an outlier in the crop–crop comparison.

Genes encoding at least three of the enzymes in the anthocyanidin and proanthocyanidin synthesis pathways (Furukawa et al., 2007; Gu et al., 2011), which produce the pigments leading to the red pericarp phenotype, contained outlier SNP in various weed-crop comparisons (Table 6; Table S20). In particular, genes coding for anthocyanidin glucosyltransferases (GT) are associated with highly diverged SNP in all four weed-crop comparisons. Notably, more anthocyanin synthesis pathway genes contain outlier SNP in weed groups related to *indica* than in those related to *aus* ancestors (Table 6), consistent with the lower incidence of red pericarps in the *indica* group (Table 1). Our analysis did not detect outlier SNP in the classic pericarp pigmentation gene, *Rc* (Furukawa et al., 2007) due to the lack of genotyped SNP; however, in the SH-*indica* comparison, the two markers closest to *Rc*, (S6_6058740 and S6_6086630) are both highly divergent SNP, and in the BHA-*aus* comparison, the closest SNP downstream of *Rc* (S6_6204793) is also an outlier.

Of the three major genes known to influence seed shattering in rice, only *qSH1* (Konishi et al., 2006) was associated with outlier SNP in weed-crop comparisons, and only in *indica* derived weeds (Table 6; Table S20). However, several genes in the flowering time pathway contain high divergence SNP in three weed-crop comparisons (Table 6; Table S20), consistent with flowering time divergence in various weed-crop comparisons (Table 1; Thurber et al., 2014). Interestingly, these genes are in relatively downstream positions in the pathway (Higgins, Bailey & Laurie, 2010). We investigated why the gene *Hd1*, a regulator of *Hd3a* and major contributor to flowering time diversity in rice (Takahashi, Teshima, Yokoi, Innan & Shimamoto, 2009), was not detected as containing outlier SNP. In most comparisons, the closest genotyped SNP was distant from *Hd1* (greater than 18Kb), decreasing the likelihood of detecting an association. For the BHA-*aus* comparison, a SNP 8Kb away from the gene was genotyped but was not identified as an outlier, consistent with lack of divergence at *Hd1* genes previously reported for these groups (Thurber et al., 2014).

Of three known genes in the tillering pathway, only *MOC1* was associated with outlier SNP in SH-*indica* comparisons (Table 6; Table S20). The cellulose synthesis pathway, which affects cell wall metabolism and has been reported to be associated with plant architecture-related traits such as height, leaf morphology and the brittle culm phenotype (Ding et al., 2015; Tanaka et al., 2003), contained four genes with outlier SNP. These cellulose synthase genes were divergent in BHA vs *aus* and *indica*-like vs *indica* comparisons (Table 6). The starch synthesis pathway, which has been under selection in domesticated rice for cooking qualities (Waters, Henry, Reinke & Fitzgerald, 2006) and has likely experienced relaxation of

TABLE 6 Weedy trait candidate genes for which crop-weed divergent SNP were observed

Trait	Candidate genes or products of candidate genes	RGAP ID	Comparisons with genotyped SNP[a]	Comparisons in which SNP was an outlier[b]	Supporting SNP
Pericarp colour	F3H (flavanone-3-hydroxylase)	LOC_Os01 g25010	AlvA,BvA,IlvI,SvI	IlvI	S1_14101586
		LOC_Os08 g37456	BvA, IlvI,SvI	IlvI	S8_23723995
		LOC_Os02 g52840	BvA,IlvI,SvI	SvI	S2_32306220
	GT(anthocyanidin glucosyltransferase)	LOC_Os06 g17250	AlvA, BvA	**AlvA**	S6_9991506
		LOC_Os06 g18790	AlvA,BvA,IlvI,SvI	AlvA	S6_10659095
		LOC_Os05 g45180	AlvA, BvA	BvA	S5_26163683
		LOC_Os07 g05420	IlvI	IlvI	S7_2488859
		LOC_Os05 g45200	AlvA, IlvI,SvI	SvI	S5_26173788
	LDOX (leucoanthocyanidin dioxygenase)	LOC_Os01 g27490	AlvA,IlvI,SvI	SvI	S1_15346903
		LOC_Os03 g18030	AlvA,BvA,IlvI,SvI	SvI	S3_10042362
Shattering	*qSH1*	LOC_Os01 g62920	AlvA,IlvI,SvI	IlvI,SvI	S1_36448657, S1_36448657
Flowering	*Hd3a*	LOC_Os06 g06320	IlvI	IlvI	S6_2940098
	RFT1	LOC_Os06 g06300	AlvA, BvA, IlvI,SvI	**AlvA**, IlvI	S6_2926114, S6_2926161
	RCN1	LOC_Os11 g05470	AlvA,BvA,IlvI,SvI	SvI	S11_2448979
	OsMADS56	LOC_Os10 g39130	AlvA,BvA,IlvI,SvI	AlvA,SvI	S10_20795711, S10_20801678
	OsMADS14	LOC_Os03 g54160	IlvL,SvL	SvI	S3_31033610
Tillering	*MOC1*	LOC_Os06 g40780	IlvI,SvI	SvI	S6_24314050
Cellulose synthesis	CSLH3 – cellulose synthase-like family H	LOC_Os04 g35030	AlvA, BvA	**BvA**	S4_21121242
	CSLD5 – cellulose synthase-like family D	LOC_Os06 g22980	AlvA,BvA	BvA	S6_13415496
	CSLC3 – cellulose synthase-like family C	LOC_Os08 g15420	AlvA, BvA,IlvI,SvI	BvA,IlvI	S8_9385705, S8_9385697
	CSLF6 – cellulose synthase-like family F	LOC_Os08 g06380	BvA, IlvI,SvI	IlvI	S8_3548272
Starch synthesis	Soluble starch synthase	LOC_Os04 g53310	AlvA,BvA,IlvI.SvI	AlvA	S4_31565911

[a]AlvA represents *aus*-like vs *aus*, BvA represents BHA vs *aus*, IlvI represents *indica*-like vs *indica*, and SvI represents SH vs *indica*.
[b]Comparisons for which a divergent gene was supported by both BAYESCAN and LOSITAN methods are highlighted in bold.

selection in weedy rice, yielded one gene with a highly divergent SNP between *aus*-like weeds and *aus* (Table 6).

4 | DISCUSSION

4.1 | Multiple independent origins for South Asian weedy rice from local wild and cultivated genetic backgrounds

For several years, the different strains of weedy rice infesting cultivated rice fields worldwide have been suspected of having separate evolutionary origins, but only recently have systematic studies begun to be undertaken around the world (Ziska et al., 2015). The region of South Asia encompassing sub-Himalayan countries had not previously been well examined for weedy rice origins. Our results clearly show that multiple weedy rice groups occur in South Asia and that their genetic backgrounds are consistent with a close relationship to diverse local *Oryza* groups (Figure 1b,c). These local groups correspond to the *aus* and *indica* cultivars, which arose in South Asia (Civáň et al., 2015), and to wild rice (*O. rufipogon*/*O. nivara*), which grows natively in the region. There is suggestive evidence that both *aus*-like and *indica*-like South Asian weedy rice have arisen more than once, as neither group is monophyletic (Figure 3). Interestingly, only in the case of *indica*-like weeds do samples nest within clades containing cultivars, the signature of a de-domestication event. However, coalescent analyses support de-domestication from *aus* cultivars as the origin of *aus*-like weeds in South Asia (Figure S7). A possible explanation for lack of nesting within the *aus*

clade could be current *aus* cultivars representing only a portion of the diversity of the originally domesticated *aus* population. In finding a close relationship between weedy rice and cultivated rice groups our results complement those of weeds from the USA (Reagon et al., 2010) and other areas (e.g. Song et al., 2014).

A significant portion of South Asian weeds are most closely related to wild rice in the region (Figures 1b,c and 3). Ancestry of weedy rice from wild rice relatives has often been suggested (Wet & Harlan, 1975), but it has not been well documented, in part because many regions where weedy rice is an agricultural problem do not harbour local wild germplasm (e.g. USA; Gealy et al., 2009) or because wild samples have not been included in analyses (e.g. Ishikawa et al., 2005; Chung & Park, 2010; Zhang, Dai, Wu, Song & Qiang, 2012; Sun et al., 2013; Qiu et al., 2014). However, wild rice contributions have been detected in weedy rice populations in Malaysia and Thailand (Prathepha, 2009; Pusadee et al., 2012; Song et al., 2014). Together with our results, this shows that the wild ancestor species of cultivated rice can serve as a source of weedy rice in some regions of Asia.

4.2 | Weedy rice is an extreme example of recurrent evolution at a global level

The genetic heterogeneity of weedy rice from South Asia is consistent with at least three, and likely more, independent evolutionary origins in this limited geographic area. Additionally, our study revealed that US weed origins constitute yet two more independent evolutionary events (Figure 3). US weedy rice groups are most closely related to *indica* and *aus* cultivars (Reagon et al., 2010), but previous sampling could not discern between direct US weed origins from South Asian cultivars, and from Asian weeds that in turn were related to Asian cultivars. Our results are compelling in supporting an origin of each US weedy rice group directly from cultivars through de-domestication. Although SH US weeds could also have arisen from *indica* outside of South Asia (Figure 3), taken together, our results suggest more than four separate evolutionary events giving rise to the organisms we refer to as weedy rice in a single world region. Despite their separate origins, all these groups have adapted to the same environment—cultivated rice fields—and all function as agricultural weeds.

The plethora of weedy rice origins from *Oryza* groups native to South Asia suggests that weedy rice is an extreme example of recurrent evolution. Weedy rice studies in other world regions furthermore suggest that this recurrent evolution is occurring at a global scale. Although not all rice-growing regions have been equally surveyed, and relationships among world weedy rice groups will require characterization with common markers, the evidence so far indicates additional independent origins in other regions. This includes *indica* and *japonica* origins in Korea (Chung & Park, 2010) and China (Zhang et al., 2012), and local elite cultivars and wild origins in Malaysia (Song et al., 2014). Remarkably, rice cultivars seem to be the most common source of weedy rice, raising concerns of how agricultural practices contribute to the rise of economically devastating weeds.

Many other crop species also have weedy relatives that infest agricultural environments. These include weedy radish (Klinger, Arriola & Ellstrand, 1992; Snow, Uthus & Culley, 2001), johnsongrass and shattercane, which are weedy types of sorghum (Anderson, Nissen, Martin & Roeth, 1998; Arriola & Ellstrand, 1996; Paterson, Schertz, Lin, Liu & Chang, 1995), weedy beets (Ford-Lloyd & Hawkes, 1986), weedy finger millet (Samarajeewa, Horiuchi & Oba, 2006), and weedy sunflowers (Whitney, Randell, Rieseberg, Elle & Whitlock, 2006). Whether extensive repeated evolution is common in other crop-related weed groups is currently unknown. There is some evidence that weedy sunflowers may have evolved multiple times from wild ancestors (Kane & Rieseberg, 2008). However, most agricultural weed species have not been extensively studied with respect to their origins.

The impressive scale at which recurrent evolution is occurring for weedy rice is perhaps comparable to another famous system, the three-spined stickleback (*Gasterosteus aculeatus*) (Jones et al., 2012), which has independently adapted to many different freshwater habitats from marine habitats. However in the case of weedy rice, a more varied set of ancestors serves as source populations (both cultivated and wild *Oryza*). The extreme repeated evolution of weedy rice at a global scale presents an unprecedented opportunity to examine convergence and parallelism at the phenotypic and genetic levels.

4.3 | Phenotypic convergence for weediness most likely does not involve the same genetic mechanisms

Several traits have often been considered typical of weedy rice. These include red pericarp, seed dormancy, seed shattering, and traits related to flowering time and competitive growth. As more systematic surveys of weedy rice around the world have accumulated, however, it has become clear that there is variation for what are thought of as weediness traits; for example, some weedy populations in temperate areas do not have seed dormancy (Xia, Xia, Ellstrand, Yang & Lu, 2011). This is of importance, as understanding what minimal set of shared traits are necessary for a plant to be weedy can have an impact on management strategies and on identification of plant groups most likely to give rise to agricultural weeds.

In our survey of a set of traits, red pericarp and easy seed shattering were the most highly convergent traits among weedy groups (Tables 1 and 2; Table S8). Although not present in every individual, weedy rice in South Asia has a high proportion of red pericarp despite *aus*, *indica* or wild rice ancestry (Table 1). The prevalence of red pericarps among weed groups suggests that proanthocyanidins in the pericarp may confer an advantage to weeds, perhaps through deterrence against pathogens and predators or increased seed dormancy (Shirley, 1998; Gu et al., 2011). Consistent with the high incidence of red pericarps in South Asian and US weeds, genes in the anthocyanidin and proanthocyanidin synthesis pathway were detected as F_{ST} outliers in all four pairs of weed-crop comparisons (Table 6).

Seed shattering, which leads to efficient seed dispersal, has long been considered a trait that increases reproductive fitness in weedy rice. South Asian weeds with diverse ancestry tend to have a greater seed shattering compared to cultivated varieties (Table 2). Other weedy rice phenotype studies also report high shattering in the USA (Thurber et al., 2010) and in Japan (Akasaka, Konishi, Izawa & Ushiki, 2011). Despite the convergence in shattering among South Asian and US weed groups, we did not detect outlier SNP in known shattering candidate genes in all weed-crop comparisons (Table 6). This is consistent with reports that the domestication *sh4* gene does not influence the shattering phenotype in US weeds (Thurber et al., 2010), and that shattering has likely re-evolved the through distinct genetic mechanisms in US weed groups (Qi et al., 2015; Thurber et al., 2013).

As might be expected, a shared genetic background seems to have a large impact on the degree of convergence of weedy traits. Both BHA and *aus*-like weeds are predominantly black hulled and awned, while SH and *indica*-like weeds are predominantly straw hulled and awnless (Table 1; Reagon et al., 2010), suggesting that weedy rice groups that originated from similar ancestors are likely to harbour similar seed morphology traits. Remarkably, the convergence between *aus*-related weed groups holds even though *aus* cultivars are neither predominantly black hulled nor awned (Table 1). This suggests that these traits might convey an advantage to weedy rice, and they are favoured when the standing variation of the ancestor makes evolution of the traits possible. Despite this convergence among *aus*-related groups, we did not detect F_{ST} outliers among awn and hull colour candidate genes examined.

While not convergent, flowering time in weedy rice is an interesting trait, as it shows significant variance among South Asian weed groups: wild-like weeds flower earliest while *aus*-like weeds flower latest (Table 2). Variation in flowering time among weedy groups colonizing the same environment has also been previously reported for US weeds (Reagon et al., 2011; Thurber et al., 2014). Moreover, as in US weeds, we also detected divergence in flowering time between weeds and their related groups. Although five candidate flowering genes were highly divergent in three of our four weed-crop comparisons, very rarely were the same genes detected as outliers among more than one comparison. Different allelic combinations among multiple genes in the rice flowering pathway may contribute to the high variance in flowering time strategies in weedy rice populations, making this an intriguing example of emphatically nonparallel phenotypic evolution in a recurrently evolving group.

Our phenotype-agnostic genome divergence scans suggest a lack of convergent evolution at the level of individual genes in weeds evolving from cultivated backgrounds. At a higher functional level, however, F_{ST} outlier lists from all four weed-crop comparisons contained an excess of genes associated with various metabolic processes. This suggests that weedy rice evolution in general may be characterized by an increase in divergence in genes involved in various primary metabolic processes, with an emphasis on protein phosphorylation (Table 5). While this could implicate genes involved in signal transduction in the divergence of crops from weeds, it provides no information about the pathways that may be involved.

Strikingly, pathogen defence genes, which are often among the most quickly evolving genes in genomes, did not seem to overly contribute to highly diverged genes between weed and crop groups (Bishop, Dean & Mitchell-Olds, 2000). There was also no strong signal of shared genetic mechanisms among weed groups adapted to the same world region (US weeds or Asian weeds), nor among weeds evolving from similar genetic backgrounds (e.g. *aus* derived weeds, *indica* derived weeds). Taken together, our results and those of others suggest that despite being a weed group that can evolve repeatedly in time and space, weedy rice groups converge phenotypically on only a few key traits. Moreover, at the genetic level, very few genes are consistently being recruited for weed evolution, and genetic mechanisms involved in weed evolution are only shared among weed groups at very broad functional levels. Thus, there seem to be multiple genetic paths to evolve weedy rice and possibly only a few constraints on the phenotypes that can contribute to successful weeds. The remarkable ease with which weedy rice can recurrently evolve at a global level makes the management of the noxious weedy rice more complicated.

ACKNOWLEDGEMENTS

We thank L. Tyler, S. Lee, G. Delulio and Z. He for technical assistance. We thank the Olsen laboratory group and B. K. Song for helpful comments on the manuscript. We are grateful to the University of Massachusetts greenhouse staff, S. Perera and J. Rodriguez Hernandez for plant care support. This work was supported by a grant from the US National Science Foundation Plant Genome Research Program (IOS-1032023) to A.L.C., K.M.O. and Y.J.

DATA ACCESSIBILITY

Raw genotyping-by-sequencing data has been deposited at the NCBI Short Read Archive (experiment SRX576894). SNP data have been deposited at DRYAD doi: 10.5061/dryad.8p9j6. Phenotype data are included in the supplementary data files.

AUTHOR CONTRIBUTIONS

A.L.C., K.M.O. and Y.J. designed the study. Z.H. performed the research. N.D.Y. helped in data analysis. K.E.H. contributed materials. A.L.C. and Z.H. wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

Akasaka, M., Konishi, S., Izawa, T., & Ushiki, J. (2011). Histological and genetic characteristics associated with the seed-shattering habit of weedy rice (*Oryza sativa* L.) from Okayama, Japan. *Breeding Science*, *61*, 168–173.

Anderson, D. D., Nissen, S. J., Martin, A. R., & Roeth, F. W. (1998). Mechanism of primisulfuron resistance in a shattercane (*Sorghum bicolor*) biotype. *Weed Science*, *46*, 158–162.

Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., & Luikart, G. (2008). LOSITAN: A workbench to detect molecular adaptation based on a F_{ST}-outlier method. *BMC Bioinformatics*, *9*, 323.

Arendt, J., & Reznick, D. (2008). Convergence and parallelism reconsidered: What have we learned about the genetics of adaptation? *Trends in Ecology & Evolution*, *23*, 26–32.

Arriola, P. E., & Ellstrand, N. C. (1996). Crop-to-weed gene flow in the genus sorghum (Poaceae): Spontaneous interspecific hybridization between johnsongrass, sorghum halepense, and crop sorghum, S. Bicolor. *American Journal of Botany*, *83*, 1153–1159.

Baker, H. G. (1965). Characteristics and modes of origin of weeds. In: H. G. S. Baker (Ed.), *The genetic of colonizating species* (pp. 147–169). New York: Academic Press.

Bishop, J. G., Dean, A. M., & Mitchell-Olds, T. (2000). Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 5322–5327.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*, 2633–2635.

Burgos, N. R., Singh, V., Tseng, T. M., Black, H., Young, N. D., Huang, Z., … Caicedo, A. L. (2014). The impact of herbicide-resistant rice technology on phenotypic diversity and population structure of United States weedy rice. *Plant Physiology*, *166*, 1208–1220.

Caicedo, A. L., Williamson, S. H., Hernandez, R. D., Boyko, A., Fledel-Alon, A., York, T. L., … Purugganan, M. D. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genetics*, *3*, e163.

Cao, Q., Lu, B.-R., Xia, H., Rong, J., Sala, F., Spada, A., & Grassi, F. (2006). Genetic diversity and origin of weedy rice (*Oryza sativa* f. spontanea) populations found in north-eastern China revealed by simple sequence repeat (SSR) markers. *Annals of Botany*, *98*, 1241–1252.

Chauhan, B. S. (2013). Strategies to manage weedy rice in Asia. *Crop Protection*, *48*, 51–56.

Choi, J. Y., Platts, A. E., Fuller, D. Q., Hsing, Y.-I., Wing, R. A., & Purugganan, M. D. (2017). The rice paradox: Multiple origins but single domestication in Asian rice. *Molecular Biology and Evolution*, *34*(4), 969–979.

Chung, J.-W., & Park, Y.-J. (2010). Population structure analysis reveals the maintenance of isolated sub-populations of weedy rice. *Weed Research*, *50*, 606–620.

Civáň, P., Craig, H., Cox, C. J., & Brown, T. A. (2015). Three geographically separate domestications of Asian rice. *Nature Plants*, *1*, 15164.

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., … Estoup, A. (2014). DIYABC v2.0: A software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism. DNA sequence and microsatellite data. *Bioinformatics*, *30*(8), 1187–1189.

Ding, Z., Lin, Z., Li, Q., Wu, H., Xiang, C., & Wang, J. (2015). DNL1, encodes cellulose synthase-like D4, is a major QTL for plant height and leaf width in rice (*Oryza sativa* L.). *Biochemical and Biophysical Research Communications*, *457*, 133–140.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: A GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, *38*, W64–W70.

Elmer, K. R., & Meyer, A. (2011). Adaptation in the age of ecological genomics: Insights from parallelism and convergence. *Trends in Ecology & Evolution*, *26*, 298–306.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE*, *6*, e19379.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*, 2611–2620.

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, *10*, 564–567.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

FAO (2002). *FAO rice information*, Vol. *3*. Rome: Produced by Agriculture and Consumer Protection, Food and Agriculture Organization of the United Nations (FAO).

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, *180*, 977–993.

Ford-Lloyd, B. V., & Hawkes, J. G. (1986). Weed beets : Their origin and classification. *Acta Horticulture*, 399–402.

Furukawa, T., Maekawa, M., Oki, T., Suda, I., Iida, S., Shimada, H., … Kadowaki, K. (2007). The Rc and Rd genes are involved in proanthocyanidin synthesis in rice pericarp. *The Plant Journal*, *49*, 91–102.

Garris, A. J., Tai, T. H., Coburn, J., Kresovich, S., & McCouch, S. (2005). Genetic structure and diversity in *Oryza sativa* L. *Genetics*, *169*, 1631–1638.

Gealy, D. R., Agrama, H. A., & Eizenga, G. C. (2009). Exploring genetic and spatial structure of U.S. weedy red rice (*Oryza sativa*) in Relation to rice relatives worldwide. *Weed Science*, *57*, 627–643.

Gross, B. L., Reagon, M., Hsu, S.-C., Caicedo, A. L., Jia, Y., & Olsen, K. M. (2010). Seeing red: The origin of grain pigmentation in US weedy rice. *Molecular Ecology*, *19*, 3380–3393.

Gu, X. Y., Foley, M. E., Horvath, D. P., Anderson, J. V., Feng, J., Zhang, L., … Chen, Z. (2011). Association between seed dormancy and pericarp color is controlled by a pleiotropic gene that regulates abscisic acid and flavonoid synthesis in weedy red rice. *Genetics*, *189*(4), 1515–1524.

Higgins, J. A., Bailey, P. C., & Laurie, D. A. (2010). Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses. *PLoS ONE*, *5*, e10065.

Hoekstra, H. E., Hirschmann, R. J., Bundey, R. A., Insel, P. A., & Crossland, J. P. (2006). A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, *313*, 101–104.

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., … Han, B. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature*, *490*, 497–501.

Hubisz, M. J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9*, 1322–1332.

Ishikawa, R., Toki, N., Imai, K., Sato, Y. I., Yamagishi, H., Shimamoto, Y., … Sato, T. (2005). Origin of weedy rice grown in Bhutan and the force of genetic diversity. *Genetic Resources and Crop Evolution*, *52*, 395–403.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*, 55–61.

Kane, N. C., & Rieseberg, L. H. (2008). Genetics and evolution of weedy Helianthus annuus populations: Adaptation of an agricultural weed. *Molecular Ecology*, *17*, 384–394.

Khush, G. S. (1997). Origin, dispersal, cultivation and variation of rice. In T. Sasaki, & G. Moore (Eds.), *Oryza: From molecule to plant* (pp. 25–34). Dordrecht: Springer Netherlands.

Klinger, T., Arriola, P. E., & Ellstrand, N. C. (1992). Crop-weed hybridization in radish (*Raphanus sativus*): Effects of distance and population size. *American Journal of Botany*, *79*, 1431–1435.

Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., & Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. *Science*, *312*, 1392–1396.

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, *44*, W242–W245.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760.

Londo, J. P., Chiang, Y.-C., Hung, K.-H., Chiang, T.-Y., & Schaal, B. A. (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 9578–9583.

Londo, J. P., & Schaal, B. A. (2007). Origins and population genetics of weedy red rice in the USA. *Molecular Ecology*, *16*, 4523–4535.

Mather, K. A., Caicedo, A. L., Polato, N. R., Olsen, K. M., McCouch, S., & Purugganan, M. D. (2007). The Extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics*, *177*, 2223–2232.

McNally, K. L., Childs, K. L., Bohnert, R., Davidson, R. M., Zhao, K., Ulat, V. J., … Leach, J. E. (2009). Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 12273–12278.

Monaco, T. J., Weller, S. C., & Ashton, F. M. (2002). *Weed science: Principles and practices*, 4th edn. New York: Wiley-Blackwell.

Moody, K. (1989). *Weeds reported in rice in South and Southeast Asia*. Los Baños, Philippines: International Rice Research Institute.

Nachman, M. W., Hoekstra, H. E., & D'Agostino, S. L. (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 5268–5273.

Oerke, E.-C. (2006). Crop losses to pests. *Journal of Agricultural Science*, *144*, 31–43.

Oka, H.-I. (1974). Experimental studies on the origin of cultivated rice. *Genetics*, *78*, 475–486.

Paterson, A. H., Schertz, K. F., Lin, Y. R., Liu, S. C., & Chang, Y. L. (1995). The weediness of wild plants: Molecular analysis of genes influencing dispersal and persistence of johnsongrass, *Sorghum halepense* (L.) Pers. *Proceedings of the National Academy of Sciences of the United States of America*, *92*, 6127–6131.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*, e190.

Prathepha, P. (2009). Seed morphological traits and genotypic diversity of weedy rice (*Oryza sativa* f. spontanea) populations found in the Thai Hom Mali rice fields of north-eastern Thailand. *Weed Biology and Management*, *9*, 1–9.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*, 904–909.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Protas, M. E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W. R., … Tabin, C. J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics*, *38*, 107–111.

Pusadee, T., Schaal, B. A., Rerkasem, B., & Jamjod, S. (2012). Population structure of the primary gene pool of *Oryza sativa* in Thailand. *Genetic Resources and Crop Evolution*, *60*, 335–353.

Qi, X., Liu, Y., Vigueira, C. C., Young, N. D., Caicedo, A. L., Jia, Y., … Olsen, K. M. (2015). More than one way to evolve a weed: Parallel evolution of US weedy rice through independent genetic mechanisms. *Molecular Ecology*, *24*, 3329–3344.

Qiu, J., Zhu, J., Fu, F., Ye, C.-Y., Wang, W., Mao, L., … Fan, L. (2014). Genome re-sequencing suggested a weedy rice origin from domesticated Indica-Japonica hybridization: A case study from southern China. *Planta*, *240*, 1353–1363.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*, 573–589.

Rathore, M., Singh, R., & Kumar, B. (2013). Weedy rice: An emerging threat to rice cultivation and options for its management. *Current Science*, *105*, 1067–1072.

Reagon, M., Thurber, C. S., Gross, B. L., Olsen, K. M., Jia, Y., & Caicedo, A. L. (2010). Genomic patterns of nucleotide diversity in divergent populations of U.S. weedy rice. *BMC Evolutionary Biology*, *10*, 180.

Reagon, M., Thurber, C. S., Olsen, K. M., Jia, Y., & Caicedo, A. L. (2011). The long and the short of it: SD1 polymorphism and the evolution of growth trait divergence in U.S. weedy rice. *Molecular Ecology*, *20*, 3743–3756.

Samarajeewa, K. B. D. P., Horiuchi, T., & Oba, S. (2006). Finger millet (*Eleucine corocana* L. Gaertn.) as a cover crop on weed control, growth and yield of soybean under different tillage systems. *Soil and Tillage Research*, *90*, 93–99.

Shirley, B. W. (1998). Flavonoids in seeds and grains: Physiological function, agronomic importance and the genetics of biosynthesis. *Seed Science Research*, *8*, 415–422.

Snow, A. A., Uthus, K. L., & Culley, T. M. (2001). Fitness of hybrids between weedy and cultivated radish: Implications for weed evolution. *Ecological Applications*, *11*, 934–943.

Song, B.-K., Chuah, T.-S., Tam, S. M., & Olsen, K. M. (2014). Malaysian weedy rice shows its true stripes: Wild Oryza and elite rice cultivars shape agricultural weed evolution in Southeast Asia. *Molecular Ecology*, *23*, 5003–5017.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., & Selbig, J. (2007). pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, *23*, 1164–1167.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Sun, J., Qian, Q., Ma, D. R., Xu, Z. J., Liu, D., Du, H. B., & Chen, W. F. (2013). Introgression and selection shaping the genome and adaptive loci of weedy rice in northern China. *New Phytologist*, *197*(1), 290–299.

Sweeney, M. T., Thomson, M. J., Pfeil, B. E., & McCouch, S. (2006). Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell Online*, *18*, 283–294.

Takahashi, Y., Teshima, K. M., Yokoi, S., Innan, H., & Shimamoto, K. (2009). Variations in Hd1 proteins, Hd3a promoters, and Ehd1 expression levels contribute to diversity of flowering time in cultivated rice. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(11), 4555–4560.

Tanaka, K., Murata, K., Yamazaki, M., Onosato, K., Miyao, A., & Hirochika, H. (2003). Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiology*, *133*, 73–83.

Thurber, C. S., Jia, M. H., Jia, Y., & Caicedo, A. L. (2013). Similar traits, different genes? Examining convergent evolution in related weedy rice populations. *Molecular Ecology*, *22*, 685–698.

Thurber, C. S., Reagon, M., Gross, B. L., Olsen, K. M., Jia, Y., & Caicedo, A. L. (2010). Molecular evolution of shattering loci in U.S. weedy rice. *Molecular Ecology*, *19*, 3271–3284.

Thurber, C. S., Reagon, M., Olsen, K. M., Jia, Y., & Caicedo, A. L. (2014). The evolution of flowering strategies in US weedy rice. *American Journal of Botany*, *101*, 1737–1747.

Vigouroux, Y., Glaubitz, J. C., Matsuoka, Y., Goodman, M. M., Sánchez, J., & Doebley, J. (2008). Population structure and genetic diversity of new world maize races assessed by DNA microsatellites. *American Journal of Botany*, *95*, 1240–1253.

Vigueira, C. C., Olsen, K. M., & Caicedo, A. L. (2013). The red queen in the corn: Agricultural weeds as models of rapid adaptive evolution. *Heredity*, *110*, 303–311.

Wang, J., Yang, L., Zhao, X., Li, J., & Zhang, D. (2014). Characterization and phylogenetic analysis of allergenic tryp_alpha_amyl protein family in plants. *Journal of Agriculture and Food Chemistry*, *62*, 270–278.

Waples, R. S., & Gaggiotti, O. (2006). INVITED REVIEW: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, *15*, 1419–1439.

Waters, D. L. E., Henry, R. J., Reinke, R. F., & Fitzgerald, M. A. (2006). Gelatinization temperature of rice explained by polymorphisms in starch synthase. *Plant Biotechnology Journal*, *4*, 115–122.

Wet, J. M. J. D., & Harlan, J. R. (1975). Weeds and domesticates: Evolution in the man-made habitat. *Economic Botany*, *29*, 99–108.

Whitney, K. D., Randell, R. A., Rieseberg, L. H., Elle, A. E. E., & Whitlock, E. M. C. (2006). Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. *American Naturalist*, *167*, 794–807.

Xia, H.-B., Xia, H., Ellstrand, N. C., Yang, C., & Lu, B.-R. (2011). Rapid evolutionary divergence and ecotypic diversification of germination behavior in weedy rice populations. *New Phytologist*, *191*, 1119–1127.

Zhang, L., Dai, W., Wu, C., Song, X., & Qiang, S. (2012). Genetic diversity and origin of Japonica- and Indica-like rice biotypes of weedy rice in the Guangdong and Liaoning provinces of China. *Genetic Resources and Crop Evolution*, *59*, 399–410.

Zhu, B.-F., Si, L., Wang, Z., Zhou, Y., Zhu, J., Shangguan, Y., … Han, B. (2011). Genetic control of a transition from black to straw-white seed hull in rice domestication. *Plant Physiology*, *155*, 1301–1311.

Zhu, Q., Zheng, X., Luo, J., Gaut, B. S., & Ge, S. (2007). Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of rice. *Molecular Biology and Evolution*, *24*, 875–888.

Ziska, L. H., Gealy, D. R., Burgos, N., Caicedo, A. L., Gressel, J., Lawton-Rauh, A. L., … Merotto, A. (2015). Chapter three – weedy (red) rice: An emerging constraint to global rice production. *Advances in Agronomy 129*, 181–228.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Huang Z, Young ND, Reagon M, et al. All roads lead to weediness: Patterns of genomic divergence reveal extensive recurrent weedy rice origins from South Asian *Oryza*. *Mol Ecol*. 2017;26:3151–3167. https://doi.org/10.1111/mec.14120