INTRODUCTION

Crop domestication has long served as a model for studying the genetics of adaptation, due to the dramatic phenotypic changes that have arisen during a short evolutionary timeframe and the ability to directly compare crop species with their extant wild relatives. Understanding the genetic basis of the phenotypic transformations that define crop species is valuable not only for breeding new varieties, but also for exploring the principles of evolutionary theory (Doebely et al., 2006; Flood & Hancock, 2017; Gross & Olsen, 2010; Meyer & Purugganan, 2013; Olsen & Wendel, 2013). The recent availability of large genome resequencing data sets for multiple crop species has greatly facilitated population genomics analyses. These approaches hold great promise and have already provided key...
insights for understanding the genetic basis of plant domestication traits (Varshney et al., 2019; Wang, Li, et al., 2018; Wang, Mauleon, et al., 2018; Wu et al., 2020; Zhou et al., 2015).

Genome-wide scans of positive selection are a valuable tool to characterize the genomic impact of selection for domestication or improvement-related traits in crop species (Hufford et al., 2012; Zhou et al., 2015). However, most selection scans have been designed and tested exclusively under the assumption of the classical “hard sweep” model (Kaplan et al., 1989; Smith & Haigh, 1974). Under this model, adaptation occurs from a single mutation (Hermisson & Pennings, 2005). It leaves detectable signatures in the targeted genomic region, including a trough of low nucleotide diversity around the selected locus (Kim & Stephan, 2002), a skew in the frequency spectrum of single nucleotide polymorphisms (SNPs) towards an excess of rare variants (Tajima, 1989), the presence of a single extended haplotype (Sabeti et al., 2002), and elevated differentiation between adaptively-diverged populations (Akey et al., 2002). Studies aimed at detecting hard sweeps constitute the paradigm for genomic scans of adaptation (Nielsen, 2005; Siol et al., 2010; Vitti et al., 2013).

The alternative to the hard sweep model, the “soft sweep”, has gained increasing attention as a potentially common and under-recognized pattern of selection (Barrett & Schluter, 2008; Messer & Petrov, 2013; Pennings & Hermisson, 2006). Under the soft sweep model, multiple adaptive alleles for the same trait are selected within a species, either from standing genetic variation or from multiple independent adaptive substitutions. Substantial theoretical work (Ferrer-Admetlla et al., 2014; Harris & DeGiorgio, 2020; Harris et al., 2018; Hermisson & Pennings, 2005; Innan & Kim, 2004; Wilson et al., 2014) and increasing empirical studies in various organisms, from viruses to insects and mammals, have highlighted its potential importance and prevalence (Avalos et al., 2017; Garud et al., 2015, 2021; Schrider & Kern, 2017).

Despite their probable importance in adaptation, soft sweeps have received little attention in crop species. The rare exceptions include studies focused on individual domestication genes, including the Rht-B1 gene affecting plant height in wheat (Raquin et al., 2008); Pgtb1 reducing branching in pearl millet (Remigereau et al., 2011); a MYB transcription factor controlling seed colour in amaranth (Stetter et al., 2020); and the Dt1 gene controlling growth habit (determinacy) in soybean, which we characterized in a previous study (Zhong et al., 2017). To our knowledge, no studies have systematically tested on a genome-wide scale for the role of soft sweeps in the process of domestication or improvement of any crop species; thus, the prevalence and genomic impact of soft sweeps is mostly unknown for crop domestication traits.

Most statistical methods for detecting selection were expressly designed to detect hard sweeps. The most commonly applied methods include the following: $\alpha$ ratios to search for genomic regions with a significant reduction of nucleotide diversity in the population undergoing selection; Tajima’s $D$ to identify loci with an excess of rare SNPs; $F_{ST}$ to detect SNPs with strong interpopulation differentiation; and extended haplotype homozygosity (EHH, often calculated as the integral of extended haplotype homozygosity, or iHH) to detect haplotypes with extended linkage disequilibrium (LD). These statistics are expected to have limited power to detect soft sweeps, because sweeps from multiple adaptive haplotypes produce distinct patterns of genetic variation in the vicinity of the adaptive site compared to that expected under a hard sweep. Similarly, incomplete hard sweeps, where a favoured mutation has not reached a high frequency across the entire study population, may not be detected by these tests (although EHH-based methods can detect incomplete sweeps if they are still young).

Fortunately, a few recently developed test statistics facilitate the simultaneous detection of both hard and soft sweeps (Garud et al., 2015; Schrider & Kern, 2016). In the series of H (homozygosity) statistics developed by Garud et al. (2015), $H_1$ is the sum of the squares of frequencies of each haplotype in a sample, and $H_2$ is calculated similarly to $H_1$ but excludes the most abundant haplotype; as such, hard sweeps are expected to have a higher $H_1$ and lower $H_2$ value than soft sweeps. $H_2$ equals the sum of the squares of frequencies after combining the first and second most abundant haplotypes into a single group. In this way, $H_2$ should have similar power to detect both hard and soft sweeps if the frequency sum of the first two most abundant haplotypes in soft sweeps approximates the frequency of the most abundant haplotype in a hard sweep. The $H_2/H_1$ ratio can then be used to differentiate between hard and soft sweeps. However, $H_2$ values rapidly decline as the softness of a sweep (i.e., the number of independent haplotypes being selected simultaneously) increases. The selection coefficient ($s$) and the time since the end of the sweep also influence the power of the $H_2$ test; $H_2$ values decrease as $s$ decreases and sweep signatures decay in the generations following the sweep.

For crop species, selection during domestication often coincides with major changes in population size caused by founder events, bottlenecks and extensive population expansion (Gaut et al., 2015; Guo et al., 2010; Hyten et al., 2006). These demographic events can strongly influence the power and ability of statistics to detect both hard and soft sweeps (Siol et al., 2010). Therefore, it is important to distinguish the effects of population history when attempting to infer signatures of selection in crop genomes. Many sophisticated statistical models have been developed to infer the demographic history of populations from genomic data (Beichman et al., 2018; Schraiber & Akey, 2015). In particular, models based on sequentially Markovian coalescent (SMC) methods, including the pairwise sequentially Markovian coalescent model (PSMC) (Li & Durbin, 2011), multiple sequentially Markovian coalescent model (MSMC) (Schiffels & Durbin, 2014), and SMC++ (Terhorst et al., 2017), have been successfully applied in crop species for demographic inference (Cubry et al., 2018; Meyer et al., 2016; Sun et al., 2020).

Soybean (Glycine max [L.] Merr.), an economically important food crop that is valued for its high levels of seed oil and protein, was domesticated from its wild progenitor Glycine soja (Sieb. and Zucc.) in China more than 3000 years ago (Hymowitz, 1970). Extensive genetic resources have become available for this species (Fang et al., 2017; Lam et al., 2010; Liu et al., 2020; Valliyodan et al., 2016; Zhou et al., 2015) following the release of the soybean reference
genome (Schmutz et al., 2010). Mainly via classical forward genetic approaches, several genes associated with soybean domestication have been isolated and functionally verified at the molecular level. These confirmed domestication-associated loci can serve as “positive controls” in selection scans — where signatures of hard or soft selective sweeps should be detectable by test statistics. Here, we focus on six such genes; these include five hard sweep targets, where a single causative mutation is uniquely present in cultivated soybean, and one gene that we previously documented to be a soft sweep target with multiple selectively-favoured mutations (Dt1) (Zhong et al., 2017) (Table 1). For the traits controlled by these domestication-related genes, we consider “domestication traits” in the narrow sense to be those that were favoured early in the domestication process and that distinguish members of a crop species from its wild ancestor, and ‘improvement traits’ to be those that were favoured at later stages of crop cultivation and are present in a subset of crop landraces or varieties (Olsen & Wendel, 2013). It should be noted that since improvement traits are not universally selected in the crop, a hard sweep for an improvement trait will create the signature of an incomplete hard sweep when considering the crop species in its entirety.

Among these six positive control genes, five were targets of hard sweeps during soybean domestication or improvement (Table 1). The domestication gene B1 controls the powdery bloom on the seed coat of wild soybean (Zhang et al., 2018). A nonsynonymous mutation in B1 accounts for the transition from the “bloom” to “no-bloom” phenotypes during domestication and also appears to increase seed oil content in domesticated soybeans. GmHs1-1 is a domestication gene for seed-coat impermeability; the transition from impermeability to permeability in domesticated soybean was caused by artificial selection on a nonsynonymous mutation in this gene (Sun et al., 2015). The improvement gene G functions in controlling seed coat colour and seed dormancy in some varieties (Wang, Li, et al., 2018); a point mutation led to an alternative splicing site and generated a premature stop codon, thereby reducing seed dormancy. The improvement gene Pdh1 controls pod dehiscence in some varieties under dry conditions (Funatsuki et al., 2014). Loss-of-function by a premature stop codon in Pdh1 confers the shattering-resistant phenotype under low humidity. Similarly, loss-of-function by a premature stop codon in Tof12 controls earlier flowering and maturity in domesticated soybean (Lu et al., 2020). Among these hard sweep genes, the improvement genes G and Pdh1 are considered to represent incomplete hard sweeps, because the favoured alleles were not universally fixed in domesticated soybean but instead account for phenotypic differences among varieties.

The soft sweep positive control gene, Dt1, was the first domestication-related gene to be molecularly characterized in soybean; it controls the difference in stem growth architecture between wild soybean and many cultivated varieties (indeterminate vs. determinate growth) (Table 1). Four nonsynonymous mutations that independently emerged in different cultivar subgroups control the transition from indeterminacy to determinacy (Liu et al., 2010; Tian et al., 2010). As an illustrative study of a domestication-related soft
sweep, we previously performed a detailed evolutionary analysis of the four Dt1 determinate-growth haplotypes and identified four independent selective sweeps (Zhong et al., 2017).

In this study, we aimed to understand the roles of hard and soft selective sweeps during soybean domestication and improvement. Using previously-published whole genome resequencing data of 302 wild and cultivated accessions (Zhou et al., 2015), we attempted to: (1) detect selection candidates by genome scans through five popular test statistics based on different selection signatures ($\alpha$, Tajima's $D$, $F_{ST}$, $iHH$, and $H12$); (2) estimate the relative abundances in the genome of hard and soft sweeps in wild and cultivated populations with $H12$, in combination with $H2/H1$; (3) characterize the genomic signatures of selection for the six known domestication/improvement genes ($B1$, $GmHs1$, $G$, $Pdh1$, $Tof12$, and $Dt1$) and use them as positive controls to evaluate the performance of different statistics in selection scans; and (4) identify a set of candidate genes that share common selection features with positive controls, and further explore their potential biological functions by comparison with previously-identified QTLs and GWAS loci.

2 | MATERIALS AND METHODS

2.1 | Data source

Resequencing data of 302 accessions, including 62 wild genotypes, 130 landraces, and 110 improved cultivars, representing germplasm before and after the domestication and improvement stages, were obtained from previously published data (see table S1 in Zhou et al., 2015). The SNPs for the 302 genotypes stored in variant call format (vcf) that were used for selection analyses were downloaded from SoyBase (https://www.soybase.org/); the variants were identified by both GATK and SAMtools, and the SNPs with minor allele frequencies (MAF) lower than 1% were discarded. The reference genome sequence and its gene annotation files (wm82.a1.v1.1) required for SNP annotation were accessed from Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html). Raw Illumina short reads for 18 genotypes (9 wild and 9 landraces) used for demographic analyses were downloaded from the Sequence Read Archive (SRA) of NCBI, and the wild reference genome “soyW01” was obtained from Liu et al. (2020). The physical positions for markers associated with various QTLs and GWAS traits were obtained from SoyBase.

2.2 | Population structure and demographic analyses

Since inferences on selection can be confounded by undetected population structure, we assessed the population structure of the wild and cultivated soybean accessions used in analyses. We first thinned the raw SNP data set to mitigate effects of linkage disequilibrium; each SNP with an $R^2$ value >0.9 with any other SNP within a 50-SNP sliding window (advanced by increments of 10 SNPs) was removed using the “--indep-pairwise” option of PLINK v.1.9 (Purcell et al., 2007). Based on the pruned SNPs set, population structure was inferred for the 62 wild genotypes, 130 landrace accessions, and 110 improved cultivars using ADMIXTURE v.1.3.0 (Alexander et al., 2009). The number of ancestral population clusters ($K$) was set to vary from 1 to 6, and the optimal $K$ was determined with the lowest value of cross-validation (CV) error. Genetic differentiation ($F_{ST}$) for pairwise subgroups within the three soybean groups were calculated using “--weir-fst-pop” flag in VCFtools (Danecek et al., 2011) (Table S1). Because pairwise genetic differentiation between subpopulations within domesticated soybean was determined to be low (see Results), landrace and improved samples were each treated as a single population in subsequent analyses; for wild soybean, a representative and robust subpopulation, “Wild_pop1” (composed of 17 accessions that consistently form a group at K values ranging from 2 to 6), was used in demographic simulation and selection analyses where uncontrolled population structure would be a confounding factor (described below).

To obtain the timing of demographic events and the corresponding effective population size ($N_e$) parameters required for construction of demographic null models, we employed a sequentially Markovian coalescent-based model, PSMC (Li & Durbin, 2011), to infer the demographic history for wild and domesticated soybean based on 18 soybean accessions (9 wild genotypes and 9 landraces, representing all genetic subgroups) with sequencing depth >15x (Table S2). The generation was set at one per year (g 1), and the mutation rate was assumed as $7 \times 10^{-9}$ per site per generation based on the rate in Arabidopsis (Ossowski et al., 2010). To test whether the inferred demographic parameters were consistent with the empirical data, the summary statistics of $\theta$ and Tajima's $D$ were compared between the real data set and an ensemble of neutral coalescent simulations generated under those parameters by Hudson’s ms simulator (Hudson, 2002). For further details and the full ms command lines, see Supporting Information Note 1.

2.3 | Methods for detecting selection

Five statistics based on different selection characteristics were used to identify selection candidates across all gene units (defined as the open reading frame and its upstream and downstream 1 kb). To assess the reduction of nucleotide diversity associated with domestication, we first calculated the $\pi$ value (Nei, 1987) for each site in the three soybean groups (wild, landrace, improved) using VCFtools (Danecek et al., 2011) with the “--site-pi” option. Then, the $\pi$ values for each gene unit and ratios of $\pi$ values in the landrace or improved group to that in the wild group were calculated based on gene coordinates using a custom script. The top 5% of genes with significantly low $\pi_{landrace}/\pi_{wild}$ values were considered candidate genes of positive selection. Tajima’s $D$ (Tajima, 1989) was calculated for each gene based on its $\theta_w$ (Watterson, 1975) and $\pi$ value in wild soybean, landraces, and improved cultivar populations individually. The top 5% of genes with significantly negative values in the cultivated but not wild...
population were taken as candidates. To assess genetic differentiation between groups, we calculated Wright’s (1951) $F_{ST}$ between the wild species and the other two groups for each SNP of all genes using VCFtools with "--weir-fst-pop" option. Those genes containing more than 3 SNPs that fell into the 2.5% right tail of the empirical distribution were deemed selective targets. To identify genes containing nonsynonymous SNPs or other protein-coding variation, we annotated all of the SNPs with the software ANNOVAR (Wang et al., 2010).

Using a custom script, we calculated a series of haplotype homozygosity-based statistics, $H_1$, $H_2$ & $H_12$ (Garud et al., 2015), for all genes in the three focal groups (wild soybean, domesticated landraces, and improved cultivars). The genes located in the 5% right tail of $H12$ empirical distribution were regarded as selection candidates, and we then used $H2/H1$ for differentiating hard from soft sweeps. Those with $H2/H1 ≥ 0.059$ among the candidates identified by $H12$ were considered soft-sweep genes, while the rest were considered hard sweep genes. The numerical threshold value for distinguishing soft-sweep candidates was determined as follows: assuming there are $n$ haplotypes in a given gene and each haplotype has a frequency of $q_i, i = 1, 2, …, n$, where $q_1 ≥ q_2 ≥ … ≥ q_i ≥ … ≥ q_n$, and $\sum q_i = 1$; if we assume that the frequency of the second most frequent haplotype is equal or greater than one-fourth of the most frequent haplotype, that is, $q_2 ≥ (1/4)q_1$; then, $H2/H1 = \sum_i q_i^2 / \sum_i q_i^2 ≥ q_2^2/(q_1^2 + q_2^2) ≥ 1/1 + 17 = 0.059$

To assess the effects of rare SNPs on the analyses, $H$ statistics were calculated using the whole genome data set and then repeated for data sets where rare SNPs with MAF $< 3%$ or MAF $< 5%$ were removed (see Supporting Information Note 2, 3). Integral of extended haplotype homozygosity (iHH) (Sabeti et al., 2002) based on linkage disequilibrium (LD) was calculated for each haplotype of all genes with frequency $> 5%$ in the three populations with a custom script. More specifically, we first took a haplotype of a given gene as a core haplotype and calculated its homozygosity; then, we extended the haplotype by adding 20 SNPs and calculated its homozygosity again (iEHH), repeating the process for both directions until the EHH values decayed to 0.05. The integral of EHH was finally obtained by summing up all the EHH values and multiplying by 20. The genes containing haplotype(s) located within the 5% right tail of any of the empirical distributions were taken as candidate genes for positive selection. Missing genotypes in haplotype-related analyses involved in iHH and H12/H1/H2 were imputed using Beagle v.5.1 with default parameters (Browning et al., 2018).

### 2.4 | Genome scans for selection candidates based on positive control genes

We used the shared population characteristics of the positive control genes to scan the genome for candidate targets of selection during domestication and improvement. Specifically, we first identified variants which are nonsynonymous, and simultaneously, for which homozygous mutations are specific to cultivated soybeans. Then, to identify (nearly) complete hard sweep genes (like $B1$ and GmHs1-1), we further screened out SNPs showing significantly elevated differentiation between wild soybean and all cultivated accessions ($F_{ST_WL} > 0.579; F_{ST_WI} > 0.676$); To find more recently targeted, improvement-related hard-sweep genes (like $G$, $Pdh1$, and Tof12), we identified the SNPs with significantly elevated $F_{ST}$ values between wild soybean and improved varieties ($F_{ST_WI} > 0.676$) and where the iHH values for the corresponding haplotypes were also above the significance threshold in landraces or improved varieties ($\text{ln}(\text{iHH}_L) > 6.066$; $\text{ln}(\text{iHH}_I) > 6.229$). For the latter improvement-related gene set, we examined their physical locations to assess whether they overlap with known soybean domestication/improvement QTLs.

### 3 | RESULTS

#### 3.1 | Population structure and demographic inference for wild and domesticated soybean

Population structure and demographic history were first examined for wild and domesticated soybean to exclude their confounding effects on selection inferences. ADMIXTURE analyses revealed genetic substructure, with optimal $K$-values of 5, 4, and 4 for wild, landrace, and improved varieties, respectively (Figure 1a, Figure S1).

We calculated genetic differentiation ($F_{ST}$) for pairwise subgroups within the three soybean groups at their optimal $K$-values (Table S1). Both the $F_{ST}$ values between pairwise groups and $N_e$ patterns (Orozco-Terwengel, 2016) suggested little population structure in domesticated soybean but clear substructure within the wild progenitor species. The average $F_{ST}$ value was relatively small for subgroups within landrases ($0.105 ± 0.0005$) and improved varieties ($0.099 ± 0.0007$), while it was two-fold greater for those within the wild progenitor ($0.217 ± 0.0113$). Given the derivative relationship of improved cultivars from landraces and the almost negligible differentiation between these groups ($F_{ST} = 0.038$), we focused solely on landrace genomes to represent the domesticated taxon and used those accessions for PSMC demographic analysis (Figure 1b).

We observed consistent patterns of effective population size ($N_e$) for all of the landrace pseudogenomes created by accessions from different subgroups. However, contrasting $N_e$ patterns were present in the wild progenitors, where the pseudogenomes generated by two individuals from the same subgroup have similar downward trajectory over time while the others created by accessions from different groups have an upward trend. Additionally, we found that the domesticated soybean lineage experienced a lasting bottleneck that started before domestication (Figure 1b); this is similar to patterns observed in African rice (Meyer et al., 2016) and grape (Zhou et al., 2017) and could reflect predomestication population management by hunter-gatherers or other factors including climate (Cubry et al., 2018; Purugganan, 2019). Given the occurrence of population structure in the wild species, subsequent analyses that would be sensitive to uncontrolled population structure used a single genetic
subgroup of wild soybean, “Wild_pop1”, to represent the progenitor species. The reliability of the inferred demographic parameters for landraces and this one wild subpopulation was demonstrated by the close distribution of $\pi$ and Tajima’s $D$ statistics between the genomic data sets and coalescent simulations under those parameters (Figure 1c).

3.2 Selection-detection methods and selection candidates during domestication and improvement

Five widely used selection tests were calculated within or between soybean populations for all 53,927 annotated genes across the 20 chromosomes of the soybean genome: $\pi$ ratios, measuring the reduction of nucleotide diversity; Tajima’s $D$, testing for skewed frequency spectrum; $F_{ST}$, assessing elevated population differentiation; iHH, based on extended linkage disequilibrium (LD); and H12 based on extended haplotype homozygosity and skewed frequency spectrum. Their density distribution plots are presented in Figure 2.

Locus-specific reductions in crop diversity (evidenced as a differentially low $\pi_{\text{cultivated}}/\pi_{\text{wild}}$ ratio) are the most commonly used feature to detect genomic targets of selection in crop domestication. We considered the 5% of genes with lowest ratios from the distribution of $\pi_{\text{cultivated}}/\pi_{\text{wild}}$ values as candidate targets of selection; this yielded 2696 genes in total across landrace and improved populations. However, the density distribution of $\pi$ ratios did not reveal obvious low-diversity tails for either of the crop-wild comparisons (landrace/wild and improved/wild); this distributional skew was most evident as an excess of low ratios in improved cultivars (Figure 2a). Improved varieties are descended from a bottlenecked subset of landraces, and the downward skew in $\pi$ ratios for this group illustrates the confounding effects of genome-wide bottlenecks on the power of the $\pi$ ratio test to reveal targets of selection, and the challenges in establishing informative significance thresholds.

For Tajima’s $D$, the top 5% of genes with significantly negative Tajima’s $D$ values in cultivated soybean were taken as candidate targets of selection; this yielded 2697 genes in both landrace and improved populations. However, as with $\pi$ ratios, demographic
effects appear to have shaped Tajima's $D$ values, which are shifted in the positive direction in both landraces and improved varieties (Figure 2b). This pattern probably reflects the domestication bottleneck during soybean domestication/improvement, as strong population bottlenecks can skew Tajima's $D$ in a positive direction (Tajima, 1989).

$F_{ST}$ was calculated for all SNPs of all genes between wild soybean and each cultivated group in turn (Figure 2c). We chose 2.5% of SNPs in the right tail of the $F_{ST}$ distribution as candidate SNPs. To control for false positives, we considered only genes containing four or more candidate SNPs as selective targets; this yielded 5101 and 5250 genes for landraces and improved cultivars, respectively. Since most causative mutations in crop domestication genes involve protein-coding variants (Meyer & Purugganan, 2013), we also applied a further filter, considering only those genes with nonsynonymous SNPs or other protein-coding variation; this yielded 1856 and 1833 genes in landraces and improved cultivars, respectively.

For the iHH analysis, the iHH statistic was calculated for each haplotype detected at >5% frequency in each gene. We observed 151,803 and 131,179 haplotypes in landraces and improved lines, respectively, and the corresponding genes of the top 5% of haplotypes were taken as selection targets (Figure 2d). By this method, 6374 and 5542 genes were identified in landraces and improved cultivars, respectively.

One challenge of applying the H12 statistic to detect selective sweeps is determining the size of the analysis window, which dramatically affects the haplotype frequency spectrum and the H12 density distribution (Garud et al., 2015). If the window size is too short, it is difficult to distinguish effects of selection from demography, resulting in a high false-positive rate; but if the size is too long, the statistical power is weakened, and many true positives will be missed (Hoban et al., 2016). We used seven incrementally larger window sizes, spanning the gene alone, and the gene plus flanking 50, 100, 150, 200, 250, or 300 SNPs, to examine the distributions of H12 density in landraces and improved soybean populations (Figure 2e). We then selected the distribution where a tail become apparent as a target from short to long window size. By this criterion, the windows of gene ± 100 SNPs and gene ± 125 SNPs was chosen for the landrace and improved soybean populations, respectively.
2,698, and 2700 genes were accordingly identified as candidates by a cutoff point of the 95% quantile.

The pattern of candidate locus sharing among the above five selection tests is presented in Figure S2. About 33% of the genes in the genome were found in the union set, suggesting the pooled candidates include a considerable proportion of false-positive genes. Among these genes, those detected by at least two tests account for 12.2% of genes in the genome (6600), and a set of 213 loci (0.39% of all genes) was shared by all five tests (Table S3).

### 3.3 Selection modes between domesticated and wild populations

The same H12 analysis for landraces and improved populations above was also applied to the focal wild subpopulation (Wild_pop1). H2/H1, in combination with H12, is expected to be able to differentiate hard from soft sweeps. Among the candidates identified by H12, the higher the H2/H1 value, the more likely a locus is to be a soft sweep target. Soft sweep candidates were considered those whose second most frequent haplotype had a frequency larger than one-fourth that of the most frequent haplotype. This criterion yielded an H2/H1 threshold value of 0.059 for calling hard versus soft sweeps, where H2/H1 values ≥0.059 were categorized as soft sweep targets (see Materials and Methods). For the seven window sizes used in the analysis, the proportion of inferred soft-sweep candidate genes generally increased with increasing window size, as would be expected (Figure 3a). Notably, regardless of window size, a contrasting pattern of H2/H1 distributions was evident between wild and cultivated soybean. Specifically, the proportion of inferred hard-sweep genes (from 49.2% to 98.1%) is obviously greater than soft-sweep genes (from 1.9% to 50.8%) in both of the cultivated populations (landraces, improved varieties). In contrast, the inferred soft-sweep genes predominate in the wild subpopulation (from 42.5% to 100%) (Figure 3a).

As a further analysis, we constructed a demographic null model based on the $N_e$ parameters inferred above to examine whether the selection candidates in the wild (G ± 25SNPs) and domesticated (G ± 100SNPs) populations are likely to be confounded by effects of their demographic history (see Supporting Information Note 1).
H12 values for the selection candidates in both the wild subpopulation (H12 > 0.308) and landraces (H12 > 0.305) are significantly greater than that from the neutral demographic model (greater than the H12 value, corresponding to the 99.5% quantile) (Figure 3b); this pattern suggests that those candidate genes show real selection signals rather than artefacts of demography. Additionally, this pattern
was unbiased when the calculation was performed based on the whole genome or when the removal of rare SNPs was considered (Figure 3c; see Supporting Information Note 2, 3). These results potentially suggest a basic difference between domesticated and wild soybean in the prevalence of hard versus soft selective sweeps, with hard sweeps dominating under domestication but not in the wild species.

3.4 | Genomic signatures at known domestication and improvement loci

For Dt1 and five other functionally-verified soybean domestication genes, we assessed selection signatures and their significance in the context of the whole genome. These six genes represent different selection stages (domestication or improvement) and selection modes (hard or soft sweeps) (Table 1). Based on the phenotypes they control and the high frequency of causative mutations in cultivated soybean, B1 and GmHs1-1 are both considered hard-sweep genes that were targets of selection early in soybean domestication. As such, we would expect significant genetic differentiation between wild soybean and each cultivated population, while the performance for other statistics will depend on their robustness to the time since the end of the sweep and demographic effects. In $\pi$ ratio analyses, neither gene showed a significant selection signature in landrace/wild comparisons, although GmHs1-1 did show a significant deviation in the improved/wild comparison (Figure 4a). For both genes, Tajima’s $D$ values for landraces were marginally or significantly negative, consistent with signatures of positive selection during the initial domestication process (Figure 4b). For $F_{ST}$ analyses, B1 showed significant differentiation in both landrace/wild and improved/wild comparisons, as predicted (Figure 4c). The causative site of GmHs1-1 also presents a significantly high $F_{ST}$ value in both comparisons according to the allele frequency among three populations published in the original study ($F_{ST} = 0.78$ and 1.00 for wild/landrace and wild/improved comparisons, respectively, calculated with data of Sun et al., 2015). This SNP was not detected in the resequencing data analysed in the present study, possibly as an artefact of low sequence quality in the region. In H12 and iHH analyses, neither B1 nor GmHs1-1 showed significant signatures in landrace/wild comparisons (Figure 4d and e), indicating the sweeps for these two genes are not young enough to be detected by LD-based statistics focused on the major haplotypes. Unexpectedly, we discovered a new minor haplotype of B1 (labelled H02), which is derived from the major causative haplotype (H01) and differs from H01 by a single nonsynonymous SNP (Table S4 and Figure S3); this haplotype showed a significant selection signature in both cultivated populations by the iHH analysis, and its frequency continued to increase during modern cultivar improvement (Figure 4e). This pattern suggests that a second selection episode, favouring a modified haplotype of previous causative allele, has been ongoing at the B1 locus during soybean improvement.

In contrast to the early-selected B1 and GmHs1-1 domestication genes, the G, Pdh1, and Tof12 loci probably represent more recent hard-sweep genes that were targets of selection during crop improvement, since intrapopulation differences are present among landraces. The sweeps at these loci can be considered near complete in improved varieties, where their causative mutations have been nearly fixed. As such, the power of the $\pi$ ratio and Tajima’s $D$ statistics will be stronger in the improved population than landraces, and we would expect significant genetic differentiation in the wild/improved comparison and extended LD detected in landraces or improved populations in the iHH and H12 analyses. The observations for Pdh1 and Tof12 were completely in line with this expectation. Tof12 was detected by all the five tests in the improved varieties, and Pdh1 was detected by $F_{ST}$, H12, and iHH tests; although Pdh1 did not reach the significance threshold in the $\pi$ ratio and Tajima’s $D$ tests, it was much closer to the threshold of improved varieties than landraces (Figure 4a–d).

Unlike Pdh1 and Tof12, the genomic signatures of G appeared to be similar to the expected signatures of domestication genes. G was significantly detected in both landrace/wild and improved/wild comparisons by $F_{ST}$ but showed significant signatures of selection only in landraces in $\pi$ ratio, H12, and iHH analyses (Figure 4a–d). A reasonable explanation for its lack of selection signature in improved varieties is that the period of time when G was a target of selection occurred earlier than the other two improvement genes, allowing more time for the genomic signature to decay. Additionally, it is worth noting that G was categorized as a soft-sweep gene due to its H2/H1 value above the threshold in landraces (Figure 4d). Further examination showed that two major haplotypes were present in cultivated soybean at the G locus (Table S4 and Figure S3); the causative haplotype (H01) was derived from another cultivated haplotype (H02), and H01 also differs from H02 by a single mutation (the causative site). Meanwhile, H02 has a very high iHH value (slightly lower than threshold and H01) in landraces (Figure 4e), indicating H02 was subject to artificial selection before a new round of selection on H01. Thus, the G gene provides another case for stepwise selection during soybean domestication.

Dt1 provides a case example of an improvement gene that underwent a soft sweep from multiple independent mutations (Zhong et al., 2017). It is not surprising to see that signatures of selection were not detected at Dt1 by $\pi$ ratio, Tajima’s $D$, or $F_{ST}$ analyses, as these popular statistics are implemented with the assumption of hard sweeps (Figure 4a–c). Notably, however, Dt1 also did not show any significant signatures by the H12 & H2/H1 statistics, which were designed to detect both hard and soft sweeps (Figure 4d). For the iHH method, we anticipated detecting the four favoured haplotypes we previously documented, each of which is associated with an independent selective sweep around a causative mutation (Zhong et al., 2017). However, none of the four selected haplotypes stood out in the genome-wide background in landraces (Figure 4e), although we confirmed the existence of the four previously-documented selective sweeps in the genome resequencing data set (Figure S4). We found that only one determinate haplotype was detected, and only...
in the improved population (Figure 4e), which probably reflects later selection for overall improved crop performance in the course of modern breeding.

A summary of the performance of the selection tests on the six domestication and improvement genes is presented in Table 2. Among the five methods, $F_{ST}$ is the most robust for detecting hard sweeps, including those at $B1$, GmHs1-1, Tof12, G, and $Pdh1$ (the latter two occurring in improved populations). The LD-based $iHH$ method appears to be powerful for detecting recent or ongoing selection events, as is the case with G, $Pdh1$, and Tof12. However, no methods were found to be efficient for identifying soft sweeps on multiple haplotypes, as occurs at Dt1.

3.5 Genome-wide identification of selection candidates based on positive control genes

Since demographic history can obscure signatures of selection in the genome, we used signatures from the six known selection targets to inform identification of other candidate loci with similar signatures. For the positive control genes, all causative mutations are nonsynonymous substitutions (Table 1) and all causative homozygotes are specific to cultivated soybean (Table S4). Using these characteristics as initial gene filtering criteria, we identified a total of 11,273 genes across the genome. Next, we took two strategies to identify complete hard sweep genes and recent/ongoing selective sweeps (Figure 5).

To identify candidate genes with complete hard sweeps during domestication, like $B1$ and GmHs1-1, we searched for nonsynonymous SNPs whose $F_{ST}$ values are significantly elevated between wild soybean and all cultivated accessions (both landraces and improved varieties). This yielded 326 candidate genes. To find more recently targeted, improvement-related hard-sweep genes like G, $Pdh1$, and Tof12, we applied two criteria: significantly elevated $F_{ST}$ values between wild soybean and improved varieties, and also $iHH$ values above the significance threshold in landraces or improved varieties. This double filtering yielded 66 candidate improvement genes. Among the 66 genes, 44 were also identified in the set of 326 loci identified by the first method. Thus, a total of 348 unique genes were identified as associated with hard sweeps at some stage of soybean domestication. Notably, these 348 selection candidates account for just 0.6% of genes in the soybean genome but include all five of the hard sweep positive control genes (Figure 5), suggesting that these represent a high-quality candidate gene data set.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$\pi$ ratio</th>
<th>Tajima’s D</th>
<th>H12</th>
<th>$F_{ST}$</th>
<th>$iHH$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L/W</td>
<td>I/W</td>
<td>L</td>
<td>I</td>
<td>L/W</td>
</tr>
<tr>
<td>$B1$</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>GmHs1-1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>$^a$</td>
<td>√</td>
</tr>
<tr>
<td>$G$</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>$Pdh1$</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Tof12</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Dt1</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

Abbreviations: I, improved cultivars; L, landraces; W, wild soybean.

$^a$Calculated from the original study by Sun et al. (2015) since the causative SNP was not present in resequencing data.

**Figure 5** Sankey diagram illustrating the filtering scheme for genome-wide identification of selection candidates based on the genomic signatures of positive control genes. The numbers in parentheses indicate the gene number after the corresponding gene filtering criteria and are proportional to the height of the bars [Colour figure can be viewed at wileyonlinelibrary.com]
We were particularly interested in the 66 recent hard-sweep candidates detected by both $F_{ST}$ and iHH because of the analytical complementarity between these two methods (genetic differentiation and extended linkage disequilibrium). We examined the chromosomal locations of these candidates and found that they are clustered in 30 locations on 20 chromosomes (Figure S5); this pattern suggests that some of the detected selection signatures are probably due to genetic hitchhiking effects caused by linkage with selected loci, and that not all 66 genes have been individual targets of selection. To explore the potential biological relevance of these candidate genes for soybean domestication and improvement, we searched for colocalization with genetic markers of previously identified soybean QTLs and GWAS traits (located within 500 kb of each candidate gene). Twenty-five of the 30 candidate gene clusters have corresponding QTLs in varying numbers, and many of these QTLs are associated with essential domestication/improvement traits, including seed features (including seed yield, morphology, size, and oil content), plant height, and flowering time (Figure S5). The 13 candidate genes that overlapped in the sets of 213 and 348 loci are clustered in six locations on 6 chromosomes, which are mainly associated with seed features and reproductive period of soybean (Figures S5, S6, and Table S5). Given that three essential genes for domestication and improvement (G for dormancy, Pdh1 for seed shattering, and Tof12 for flowering and maturity) were detected in this small set of 66 candidate genes, we have reason to have confidence in the power of our selection-detecting strategies and the potential importance of these genes in soybean domestication and improvement.

4 | DISCUSSION

While increasingly recognized for their role in adaptive responses in wild species, soft sweeps have not been systematically examined in the context of domestication. This knowledge gap prompted the current study, where we sought to characterize the prevalence of hard and soft sweeps in domesticated and wild soybean, and to evaluate our detection abilities when applying widely used selection tests. Our genome scans of domesticated G. max and a representative subpopulation of its wild ancestor, G. soja, suggest that there is a qualitative difference between them in the prevalence of hard versus soft sweeps: whereas we find extensive signatures of soft sweeps in the wild species, hard sweeps appear to predominate in the domesticate (Figure 3a). For both wild and domesticated soybean, our selective sweep inferences appear robust to demographic models (Figure 3b). To the extent that these findings for domesticated soybean are generalizable to other crop systems (see caveats below), they suggest that human-mediated selection during domestication may occur primarily via hard sweeps. Consistent with this result, our genome scans using statistics designed to detect hard sweeps yielded a promising set of 66 candidate loci, many of which overlap previously-described QTLs and molecularly-confirmed genes for soybean domestication and improvement traits (Figure S5). Below we discuss the implications and limitations of our findings for soybean and other domestication systems.

4.1 | The power and limitations of various statistics on selection detection

Genome scans to identify statistical outlier loci are a widely used method for detecting targets of positive selection in population genomics studies. Based on our observations, $F_{ST}$ is the most robust method for detecting domestication-related hard sweeps. For all five of our hard sweep positive control genes, clusters of SNPs comprising the causative sites and linked neutral variants could be easily detected at a 99% cutoff threshold using this metric (Figure 4c). This result corresponds well with a previous study of dog domestication, in which an $F_{ST}$-based statistic, hapFLK, was scanned across 25 dog breeds and successfully identified all 12 positive controls (Schlamp et al., 2016). We believe that the statistics utilizing genetic differentiation are likely to be the most powerful strategy in detecting domestication-mediated selection. However, while this method is powerful, the false-positive rate of $F_{ST}$ scans should not be overlooked. In our analyses, more than half the genes in the genome contained at least one outlier SNP at a 95% cutoff threshold. This false positive rate can be effectively reduced by the following measures: elevating the significance threshold (in our case to 99%), focusing on genes with multiple outlier SNPs, and focusing on the subset of those genes that contain at least one nonsynonymous SNP. The third criterion may be especially important in plant domestication studies: all of the causative SNPs in our six positive control genes involve protein-coding changes (including premature stops), and nonsense and missense mutations of this nature are reported to be the predominant causative changes in crop domestication genes (Meyer & Purugganan, 2013).

The LD-based method iHH provides unique power to detect recent or ongoing selective sweeps, whether complete or not. Three selection events, involving the causative haplotypes of the improvement genes G, Pdh1, and Tof12, and a newly discovered minor haplotype of B1 (H02), were successfully detected by the iHH in landraces (Figure 4e). While about 9,000 genes were identified by iHH in our soybean genome scan analysis, we can prioritize loci by focusing on haplotypes carrying protein-coding variation as with the $F_{ST}$ outlier analysis.

Among the other metrics used in our study, $\pi$ ratios (to identify crop-specific reductions in diversity) and Tajima’s D are well known to be sensitive to demographic history (Siel et al., 2010). Population bottlenecks and expansions are an inherent feature of most domestication events, and they can reduce nucleotide diversity and/or skew the site frequency spectrum in similar ways to selection. Notably, among our six positive control selection targets, neither of the narrow-sense domestication genes (B1 and GmHs1-1) showed significantly skewed $\pi$ ratios in the landrace/wild comparison, and only one of the three improvement genes (Tof12) showed the expected skew in the improved/wild comparison. Additionally, the
timing of the selective event will affect the ability to detect it using these site frequency spectrum-based measures, as the signal fades over time. This effect was especially evident in our Tajima’s D analyses, where positive control genes that were targets of hard sweeps during the initial domestication process (B1 and GmHs1-1) have (marginally) significant Tajima’s D values in landraces while the more recent hard-sweep genes (G, Pdh1, and Tof12) approach the significance threshold in the improved population (Figure 4b).

The failure of H12 and H2/H1 to detect the previously-documented soft sweep at Dt1 can offer some important insights into the complexities and limitations inherent in soft sweep detection (Figure 4d). The “softness” of a sweep corresponds to the number of haplotypes independently selected, and the ability of H12 and H2/H1 to detect sweeps is maximized if there are only one or two selected haplotypes (the former corresponding to a hard sweep) (Garud et al., 2015). In the case of Dt1, the softness created by four independently-selected haplotypes substantially impedes detection ability. The Dt1 sweep is further softened by the fact that the ancestral haplotype (confering indeterminate growth) is still preferred in some varieties and habitats, so that diversifying selection keeps the soft sweep incomplete. This has the effect of sharply reducing the H12 value. The complexity of having multiple haplotypes favoured by selection also has an impact on the iHH test. We observed that none of the four causative haplotypes in Dt1 were successfully detected in landraces by iHH (Figure 4e). Since three of the four haplotypes appear to have originated within cultivated soybean (Zhong et al., 2017), it seems unlikely that the haplotypes are too old to be detectable based on haplotype length. A more reasonable possibility is that with so many independently-selected haplotypes, the weakness of selection on any given haplotype reduces the power of H12. Collectively, the higher degree of softness, the intermediate ending frequency, and the weaker selection coefficient together contributed to the escape of Dt1 from detection by H12.

4.2 Selection modes in cultivated and natural environments

Our H12 & H2/H1 analyses of wild and cultivated soybean suggest that hard sweeps predominate over soft sweeps in domesticated species (Figure 3a). While the generalizability of this finding must await additional soft sweep scans in other crop species, there are several reasons why this pattern would be expected. From a theoretical perspective, the probabilities of generating soft sweeps, either from standing genetic variation or from multiple mutational origins, are governed by the population mutation parameter ($\Theta = 4N_e \mu$). All else being equal, the larger the value of $\Theta$, the larger the probabilities of yielding soft sweeps (Hermisson & Pennings, 2017; Jensen, 2014; Messer & Petrov, 2013; Orr & Betancourt, 2001). Soft sweeps are therefore expected to occur more frequently in populations with larger $N_e$. Since crop species typically experience population and genetic bottlenecks during the domestication process, their smaller $N_e$ would reduce the soft sweep probability relative to the wild progenitor. On top of this, the demographic events during domestication and improvement are expected to accelerate the hardening process of soft sweeps, since a soft sweep pattern could be erased by bottlenecks and genetic drift eliminating all but the highest frequency adaptive allele (Hermisson & Pennings, 2017).

Differences in selective pressures between wild and domesticated environments are also likely to contribute to differences in soft sweep prevalence. Wild species naturally experience greater spatial and temporal environmental heterogeneity than crop species growing in managed agricultural habitats. In addition, selection during domestication (in the narrow sense) may be more consistent and unidirectional than natural selection in wild species (Milla et al., 2015). Both of these factors would differentially favour the fixation of a single fittest haplotype in crop species, creating hard sweeps in place of soft sweeps, whereas multiple haplotypes are more likely to be favoured in the variable environmental conditions experienced by a wild species. Consistent with this prediction, several instances of soft sweeps have been documented in the maize wild progenitor teosinte (Fustier et al., 2017), and signatures of soft sweeps were found to be abundant in the Drosophila melanogaster genome (Garud et al., 2015). Soft sweeps have also been reported to be the dominant mode of adaptation in the human genome (Schrider & Kern, 2017).

It should be noted that the inference of selection mode by H12 & H2/H1 analyses comes with an important caveat: assessment of the abundance of hard versus soft sweeps relies on a complete and accurate detection of both forms of selective sweep, but the H12 method has limited ability for detecting sweeps at loci with multiple alleles selected, those with weak selection strength, or those where selection occurred in the distant past (Garud et al., 2015). This limitation is evidenced by the poor performance of H12 with two hard-sweep genes associated with early domestication selection (GmHs1-1, B1), and with the soft-sweep gene Dt1 (Figure 4d). The detectable proportion of hard versus soft sweeps within species will therefore be inevitably biased to some extent by this limitation of H12 (Figure 3a). In the context of the present study, this limitation means that while the patterns we observe strongly suggest a greater prevalence of soft sweeps in wild soybean than in the domesticated species, the lack of evidence for soft sweeps in the crop should not be taken as evidence that they have not occurred at all (indeed, the well studied Dt1 gene demonstrates otherwise), but rather that soft sweeps have had a less statistically-detectable impact on the genome of the domesticated species than the wild progenitor.

4.3 Candidate genes for soybean domestication and improvement

The large statistical toolkit that has been designed specifically for detecting hard sweeps — which, by our results are the predominant selection mode in domesticated soybean — has allowed us to identify pools of promising candidate genes for this crop (Tables S3, S5). The apparent success of our approach derives from the application
of analytically complementary test statistics (elevated population differentiation and extended linkage disequilibrium as well as incorporation of information on protein-coding variation (nonsynonymous SNPs); the latter criterion may be of particular value for crop species given the apparent role of protein-coding variation in domestication and improvement traits (Meyer & Purugganan, 2013). It is also evident from our analyses that while application of selection tests individually is prone to high recovery of false positives, the indiscriminate stacking of multiple tests is equally counterproductive. Indeed, only 13 genes overlapped between the 213 loci identified by all five selection tests and the 348 genes identified based on the common signatures of positive hard-sweep genes (Figure S6). Overstacking is especially ill advised when combining analyses that are inherently incompatible, as their intersection will probably miss many true positive genes. For example, \( x \) ratios give priority to genes with little or no nucleotide variation in the selected population, whereas such genes cannot be detected by Tajima’s \( D \).

To avoid such logical incompatibility, we suggest a future focus in soybean on the 66 genes with both highly differentiated nonsynonymous SNP(s) and widely extended haplotype homozygosity for the corresponding haplotype characterized by the SNP(s) (Figure 5; Table S5). These candidates include three functionally verified improvement genes (\( Pdh1 \), and \( Tof12 \)), and they overlap with many previously identified QTLs for soybean domestication/improvement traits (Figure S5). The genes in this candidate set that have not yet been molecularly characterized are prime candidates for functional studies to assess their contribution to the genetic basis of soybean domestication and improvement traits, and they may potentially provide important genetic targets for future soybean breeding. More generally, this strategy of scanning for selection candidates based on the characteristics of positive control genes could also prove useful when applied to other crop species.

ACKNOWLEDGEMENTS
This work was supported by the National Natural Science Foundation of China (grant nos. 31801050, 31960433) and China Scholarship Council (grant no. 201806820031). We thank PhD candidate Huayi Zeng (Washington University in St. Louis) for his help in coding skills. We appreciate all members of Zhu and Olsen laboratories for their helpful comments.

CONFLICT OF INTEREST
The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS
Limei Zhong, Youlin Zhu, and Kenneth M. Olsen conceived the research, Limei Zhong analysed all the data and wrote the manuscript draft, Youlin Zhu and Kenneth M. Olsen supervised the project and revised the manuscript.

DATA AVAILABILITY STATEMENT
The whole-genome resequencing data for 302 soybean accessions was previously published in Zhou et al. (2015). The raw Illumina short reads were downloaded from the Sequence Read Archive (SRA) of NCBI. The genotype metrics for the 302 soybean accessions were downloaded from SoyBase (https://soybase.org/data/v2/Glycine/max/diversity/Wm82.gnm1.div.Zhou_Jiang_2015/). Custom scripts used in the manuscript are available in https://github.com/limeizhong/hard-soft-sweeps. Further details for the usage of software and custom scripts involved in the manuscript are provided in Supporting Information Note 4.

REFERENCES

ORCID
Limei Zhong https://orcid.org/0000-0002-7472-5245
Kenneth M. Olsen https://orcid.org/0000-0002-8338-3638
683 common bean genotypes identifies yield component trait associations across a north–south cline. Nature Genetics, 52(1), 118–125. https://doi.org/10.1038/s41588-019-0546-0


