nature genetics

Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits

Rajeev K Varshney[1,2], Rachit K Saxena[1], Hari D Upadhyaya[1], Aamir W Khan[1], Yue Yu[3], Changhoon Kim[4], Abhishek Rathore[1], Dongseon Kim[4], Jihun Kim[4], Shaun An[3], Vinay Kumar[1], Ghanta Anuradha[5], Kalinati Narasimhan Yamini[5], Wei Zhang[3], Sonnappa Muniswamy[6], Jong-So Kim[4], R Varma Penmetsa[7], Eric von Wettberg[8] & Swapan K Datta[9]

Pigeonpea (*Cajanus cajan*), a tropical grain legume with low input requirements, is expected to continue to have an important role in supplying food and nutritional security in developing countries in Asia, Africa and the tropical Americas. From whole-genome resequencing of 292 *Cajanus* accessions encompassing breeding lines, landraces and wild species, we characterize genome-wide variation. On the basis of a scan for selective sweeps, we find several genomic regions that were likely targets of domestication and breeding. Using genome-wide association analysis, we identify associations between several candidate genes and agronomically important traits. Candidate genes for these traits in pigeonpea have sequence similarity to genes functionally characterized in other plants for flowering time control, seed development and pod dehiscence. Our findings will allow acceleration of genetic gains for key traits to improve yield and sustainability in pigeonpea.

Along with cereals, legumes provide balanced nutrition to human diets in the form of essential amino acids, minerals, vitamins, fiber and resistant starch. In addition, legumes enhance and sustain soil health through symbiotic nitrogen fixation. Yield levels, however, have remained stagnant during the last six decades in the majority of legume crops like pigeonpea (*C. cajan*), a tropical grain legume grown on 5 million hectares. Narrow genetic diversity in the elite gene pool coupled with the limited genomic resources available until recently has been a major bottleneck for using modern breeding approaches for crop improvement. It is therefore essential to catalog available genome-wide sequence variations in germplasm, particularly in landraces and other non-adapted lines, to develop markers for breeding traits.

The distribution of wild relatives of the crop, archaeological remains, linguistic evidence and the extensive usage of pigeonpea in daily cuisines together support India as the center of origin of the pigeonpea crop[1,2]. Cultivated pigeonpea originated from its wild progenitor *Cajanus cajanifolius* in central India around 3,500 years ago[3]. The completion of a draft genome sequence[4] and the availability of high-throughput sequencing technologies[5] make it possible to rapidly detect genomic variation among germplasm of *Cajanus* breeding lines, landraces and wild species. A reference set of 300 pigeonpea accessions from eight geographical regions (South Asia, South America, Mesoamerica, Oceania–Pacific, sub-Saharan Africa, Europe, East Asia and Southwest Asia) was assembled[6] (**Fig. 1a**). This reference set encompasses 95% of the total genetic diversity present in a larger composite collection of 1,000 accessions spanning the wide geographical distribution of pigeonpea germplasm[6] (**Fig. 1** and **Supplementary Table 1**). To gain a better understanding of the patterns of genome-wide variation and the genetic basis of agronomic traits in *Cajanus* spp., we resequenced 300 accessions from the pigeonpea reference set, which includes modern cultivars, traditional landraces and wild species accessions (**Supplementary Table 1**). By conducting a genome-wide association study (GWAS), we identified genomic regions affected by domestication and breeding and genetic loci associated with phenotypic variation for agronomically important traits of relevance to pigeonpea breeding.

RESULTS

Resequencing of *Cajanus* accessions

High-quality whole-genome resequencing (WGRS) data were generated for 292 accessions from the reference set, including 117 breeding lines, 166 landraces, 2 others and 7 accessions from three wild species, *Cajanus cajanifolius*, *Cajanus scarabaeoides* and *Cajanus platycarpus* (**Supplementary Table 1**). We generated a total of 21.7 billion paired-end

[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. [2]School of Agriculture and Environment and Institute of Agriculture, University of Western Australia, Crawley, Western Australia, Australia. [3]Shenzhen Millennium Genomics, Inc., Shenzhen, China. [4]MACROGEN, Inc., Seoul, Republic of Korea. [5]Institute of Biotechnology, Professor Jayshankar Telangana State Agricultural University (PJTSAU), Hyderabad, India. [6]Agricultural Research Station (ARS)–Gulbarga, University of Agricultural Sciences (UAS), Karnataka, India. [7]Department of Plant Sciences, University of California–Davis, Davis, California, USA. [8]Biological Sciences and International Center for Tropical Botany, Florida International University, Miami, Florida, USA. [9]Visva-Bharati, Shantiniketan, India. Correspondence should be addressed to R.K.V. (r.k.varshney@cgiar.org).

Received 2 December 2016; accepted 25 April 2017; published online 22 May 2017; doi:10.1038/ng.3872

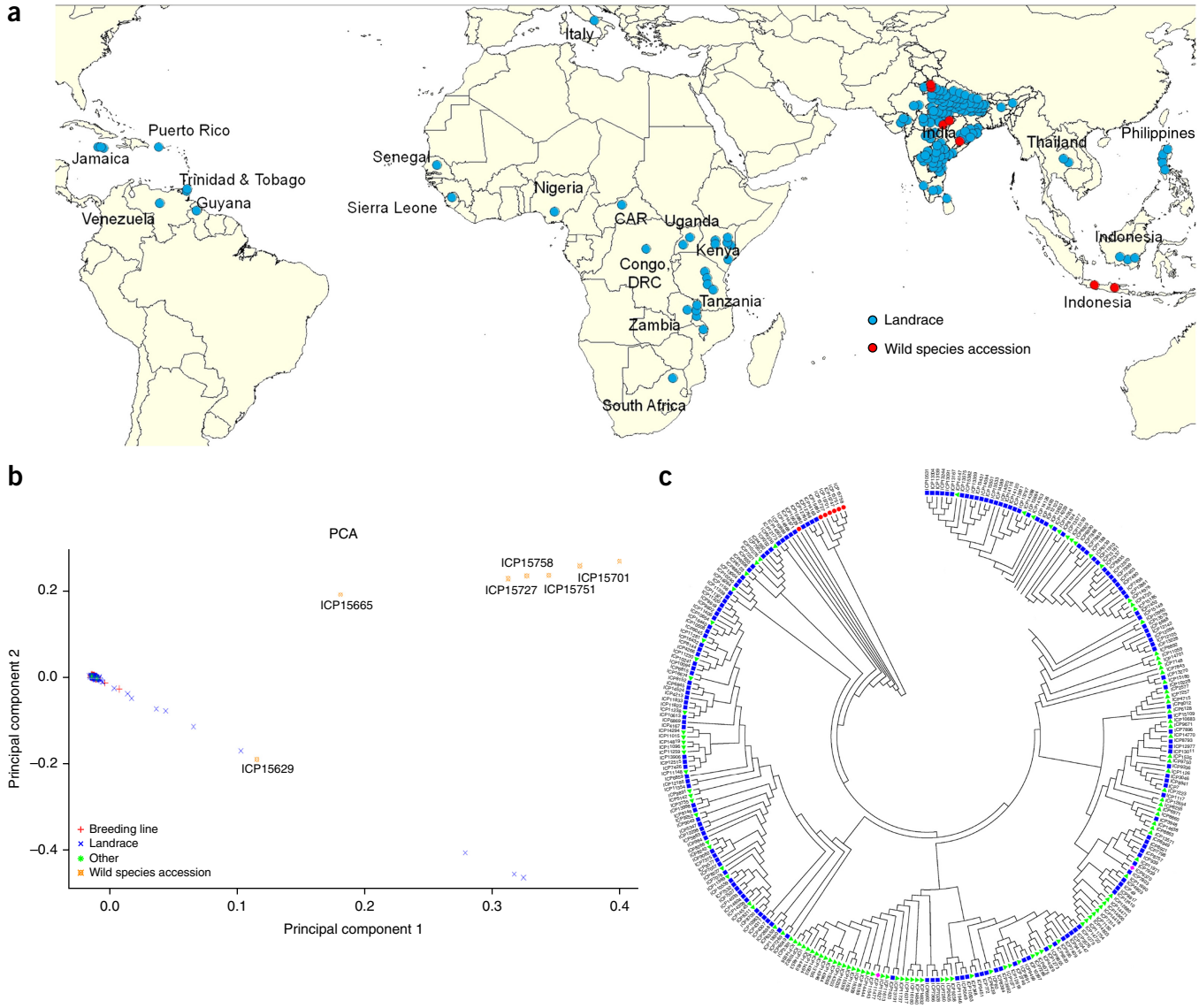

Figure 1 An overview on diversity in 292 *Cajanus* accessions. (**a**) Geographical distribution of the collection sites for 166 landraces (represented in blue color) and seven wild species accessions (represented in red color). Information on pigeonpea growing zones across the globe is taken from the public domain, while the map was drawn using licensed ESRI 2004 software. CAR, Central African Republic; DRC, Democratic Republic of the Congo. (**b**) PCA of 292 *Cajanus* accessions including 117 breeding lines, 166 landraces, 7 accessions from three wild relative species and 2 accessions with no information using SNPs detected in whole-genome resequencing data. (**c**) Neighbor-joining tree analysis of 292 *Cajanus* accessions (166 landraces represented in blue, 117 breeding lines in green, 2 others in pink and 7 wild species accessions in red) using SNPs detected in whole-genome resequencing data.

reads of 101 bp in length (2.19 Tb of sequence) that were mapped to the reference genome of pigeonpea cultivar 'Asha' (ICPL87119)[4] using BWA[7]. We obtained sequencing depths that ranged from 5× to 12× and genome coverage of approximately 93% (**Supplementary Table 2**). Mapping results were used to identify small-scale variation evident as SNPs and indels and larger-scale variants including copy number variations (CNVs) and presence-and-absence variations (PAVs) (**Table 1**).

Patterns of variation across *Cajanus* accessions

The WGRS data provided a total count of 17.2 million variants across the 292 *Cajanus* accessions (**Fig. 2** and **Supplementary Table 3**). These included 15.1 million SNPs, 0.9 million small insertions and 1.2 million small deletions (indels of 1–5 bp in length). A total of 3.02 million SNPs were found in genes, and variants occurred approximately twice as often in noncoding regions (65.9%; 1.99 million) as in coding regions (34.1%; 1.03 million) (**Supplementary Table 3**). Within coding regions, nonsynonymous SNPs (0.5 million) were moderately more frequent than synonymous SNPs (0.4 million). This 1.18 ratio of nonsynonymous-to-synonymous substitutions in pigeonpea is intermediate to the ratios observed in *Arabidopsis thaliana* (0.83)[8] and soybean (1.61)[9].

To identify genomic regions affected by demography and selection during domestication, we examined WGRS data for genomic regions with SNP frequencies that deviated from the whole-genome average (**Supplementary Fig. 1** and **Supplementary Table 4**). We scanned the entire genome with non-overlapping windows of 10 kb in length separately in subsets of accessions comprising modern breeding lines, landraces

Table 1 Summary of whole-genome variations identified in breeding lines, landraces and wild species accessions

Sequence variants		Breeding lines	Landraces	Wild species
SNPs	Total	2,783,600	5,583,628	10,795,005
	Intergenic	2,336,405	470,0294	8,118,234
	Intron	210,341	438,338	1,707,486
	Exon	168,585	311,975	807,274
Indels	Total	493,999	935,816	1,375,995
	Insertions	204,603	400,250	627,000
	Deletions	289,396	535,566	748,995
SVs	Total	317	265	250
	CNVs	282	228	173
	PAVs	35	37	77

SVs, structural variants; CNVs, copy number variants; PAVs, presence-and-absence variants.

and wild species accessions (**Supplementary Table 5**), as these groups reflect different phases of crop evolutionary history. A number of regions were identified as having variation consistent with neutral selection nonsynonymous mutation (*Ka*)/synonymous mutation (*Ks*) ratio = 1, purifying selection (*Ka*/*Ks* < 1) and positive selection (*Ka*/*Ks* > 1) in all three groups of accessions. The presence of higher nonsynonymous-to-synonymous substitution ratios at the whole-genome level in cultivated pigeonpea (in breeding lines and landraces) as compared to wild species accessions suggests that cultivated pigeonpea has accumulated a higher ratio of deleterious to non-deleterious mutations. A combination of positive and purifying selection, as well as neutral evolution, may have shaped the cultivated pigeonpea genome during the domestication process. To increase our chances of capturing all possible genomic variations influencing the selection process, we also scanned the genome with non-overlapping windows of 1 Mb in length (**Supplementary Fig. 1** and **Supplementary Table 4**). A total of 19 genomic regions with a nonsynonymous-to-synonymous substitution ratio greater than 2.5 were identified (**Supplementary Fig. 1** and **Supplementary Table 6**). These 19 genomic regions covered a total of 1,749 genes whose gene ontology–based annotation suggests an involvement in regulatory processes, recognition of external signals, etc. (**Supplementary Fig. 2** and **Supplementary Table 6**), which is consistent with findings for candidate regions involved in selective sweeps in *Arabidopsis*[8] and soybean[9].

SNPs among the subgroups of breeding lines (2.7 million), landraces (5.5 million) and wild species accessions (10.7 million) accounted for 17.9%, 36.1% and 69.8%, respectively, of all SNPs (**Supplementary Fig. 3**). The greatest number of unique SNPs was identified in wild species accessions (9.1 million), followed by landraces (2.7 million) and breeding lines (0.5 million) (**Supplementary Fig. 3**), a pattern consistent with progressive bottlenecks on genetic diversity during domestication and subsequent breeding. A total of 0.7 million SNPs were common to all three groups, representing low levels of standing variation from wild species percolating through domestication and subsequent landrace diversification into breeding material. More SNPs (~2 million) were common to breeding lines and landraces than to landraces and wild species accessions (1.4 million) and breeding lines and wild species accessions (0.8 million), consistent with the closer relationship of landraces and breeding lines (**Supplementary Fig. 3**). Pairwise genome-wide fixation index (F_{ST}) values based on SNPs have provided results similar to those described above with respect to the closer relationship of breeding lines and landraces (F_{ST} = 0.007) as compared to breeding lines and wild species (F_{ST} = 0.264) and landraces and wild species (F_{ST} = 0.263).

We also examined WGRS data for larger structural variations (SVs) ≥1,000 bp in length (**Table 1**). To ascertain how SVs were shaped by major evolutionary transitions, from wild species to domesticated accessions and from domesticated accessions to intensive modern breeding lines, we pooled individuals from the same subgroups (breeding lines, landraces and wild species accessions) separately. This aggregation deepened coverage, yielding 1,036×, 1,523× and 75× coverage for breeding lines, landraces and wild species accessions, respectively. We identified a total of 282, 228 and 173 CNVs and 35, 37 and 77 PAVs in breeding lines, landraces and wild species accessions, respectively. The size of SVs (including CNVs and PAVs) varied from 0.002 Mb to 13.3 Mb in breeding lines, from 0.001 Mb to 0.2 Mb in landraces and from 0.001 Mb to 2 Mb in wild species accessions (**Supplementary Figs. 4–6** and **Supplementary Tables 7–9**). We note that, although wild species material likely corresponds to the group with the greatest number of variants owing to the likely few million years over which the species diverged from one another, our power to detect SVs in the wild species was far more limited, owing to fewer samples, lower sequencing coverage of their pooled aggregate and the unavailability of a wild species reference genome for alignment. Despite the potential undercounting of SVs among the wild species, the SVs we identified in this group are useful for understanding levels of variation.

Phylogenetic relationships

We used sequence variation data to obtain a broad view of the genetic relationships among the *Cajanus* accessions. Principal-component analysis (PCA) reflected limited genetic diversity in cultivated accessions, comprising breeding lines and landraces, and a more diverse spread of wild species accessions (**Fig. 1b**). Analyses based on pairwise dissimilarity using neighbor joining also identified two distinct groups (**Fig. 1c**). Of the two major groups shown in **Figure 1c**, group I contained six wild species accessions (ICP15758, ICP15751, ICP15747, ICP15701 and ICP15727 from *C. scarabaeoides* and ICP15665 from *C. platycarpus*) and three landraces (ICP14163, ICP12766 and ICP12765) and group II included the remaining landraces, all breeding lines and one wild species accession (ICP15629) from *C. cajanifolius*. These results also assigned *C. cajanifolius* as the closest wild species to cultivated pigeonpea and support the notion of it being the most likely progenitor species[2]. The subgroups within group II, which predominantly (but not exclusively) contained cultivated material from both breeding lines and landraces, could have been derived from diverse landraces having been used as parents by breeders in developing new breeding lines or from some landraces or early domesticates, such as ICP15629, having sufficient agronomic suitability to be used for crop production without further intercrossing and selection by breeders. Thus, it seems that the true genetic history of the breeding lines and landraces in pigeonpea is more obscure than the more classical view (as in maize, for example) in which landraces represent early domesticated forms and breeding lines represent outputs from subsequent deliberate breeding and selection. Despite the ambiguity within cultivated materials, however, several minor groups showed some degree of separation between breeding lines and landraces. Further analysis with STRUCTURE[10,11] identified numerous accessions showing admixture, highlighting the prevalence of genetic mixtures of breeding lines and landraces in pigeonpea breeding (**Supplementary Fig. 7**).

Sterility mosaic disease (SMD), Fusarium wilt (FW) and photoperiod insensitivity are among the most important target traits for contemporary breeding in pigeonpea. Among a total of seven SMD-resistant accessions, five breeding lines (ICP11230, ICP11238,

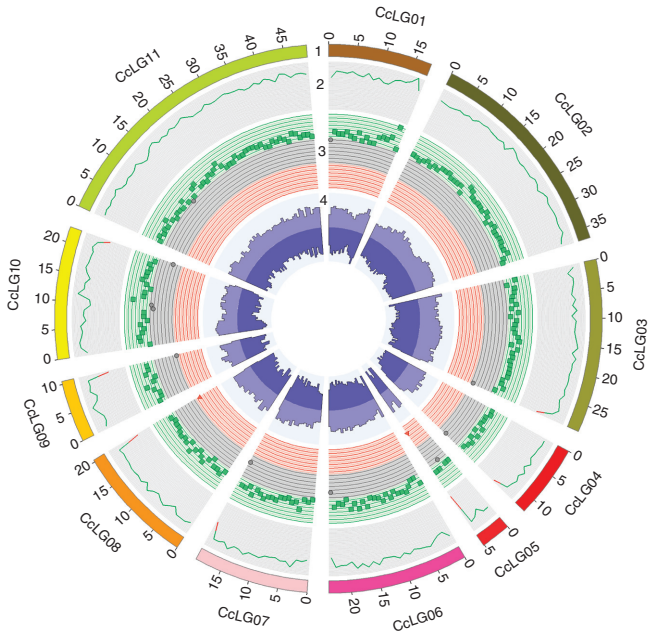

Figure 2 A Circos image representing variations identified across 292 *Cajanus* accessions. (1) 11 pseudomolecules (CcLG01 to CcLG11) in different colors with scale of Mb genome. (2) Sequence coverage. The range of the coverage plot axis is 0 to 40,000 (green, >20,000; red, <20,000). (3) Green squares, dark shaded circles and red triangles represent SNP regions in breeding lines, landraces and wild relatives, respectively. The range of the SNP axis is 0 to 36,000 (green square, >24,000; dark shaded circle, 12,000–24,000; red triangle, <12,000). (4) Indels (the range of the indel axis is 0 to 3,500; light purple, insertions; dark purple, deletions).

ICP11015, ICP11096 and ICP11148) were closely clustered together in a subgroup within group II, while one breeding line (ICP11059) along with one landrace (ICP7436) was distantly placed in another subgroup. Interestingly, four FW-resistant accessions (ICP14819, ICP3633, ICP8863 and ICP10240) were scattered in different subgroups within group II. Of four photoperiod-insensitive accessions, three (ICP14903, ICP14936 and ICP14944) were found in close proximity to one another as compared to the fourth accession (ICP14900). These observations suggest recurrent use of common source material for SMD resistance and photoperiod insensitivity, but different sources for developing FW resistance in breeding (**Fig. 1c**).

To assess the geographical distribution and to validate the center of origin with the present data set, we calculated the genetic distances among populations using *F* statistics (**Supplementary Tables 10–12**). Landraces and wild species accessions were classified by their continent, country and state of origin. On the continent scale, pairwise F_{ST} values correlated with geographical distances between populations when the center of origin was considered to be South Asia (**Supplementary Table 10**). The F_{ST} value (0.102) for populations from South Asia and sub-Saharan Africa was lowest as compared to those for South Asian and South American populations (0.126) and South Asian and Mesoamerican populations (0.167). These F_{ST} values suggest that the dispersal route of pigeonpea was from South Asia to sub-Saharan Africa and, finally, to South America and Mesoamerica. It has been proposed that pigeonpea migrated to the New World through the slave trade from Africa and South Asia[12]. On the country scale, we found that populations from Uganda and Tanzania were either derived from or closer to accessions from India than populations from Kenya and Malawi were (**Supplementary Table 11**). Interestingly, the F_{ST} values suggest that pigeonpea might have migrated from India to the East African country of Uganda, then from Uganda to Tanzania and Kenya, and from Kenya to Malawi. Alternatively, or in combination with historical migration of cultivated pigeonpea, the genetic relationships of pigeonpea accessions from India and East African countries may reflect local selection in East Africa for adaptation to local climates or consumer preferences such as seed size, appearance and taste.

When F_{ST} values were calculated according to states in India (**Supplementary Table 12**), our data set validated the central Indian state of Madhya Pradesh as the likely center of origin of pigeonpea. We saw populations from Rajasthan, Odisha, Uttar Pradesh and Maharashtra closely clustered with the Madhya Pradesh population in comparison to populations from Bihar, Karnataka, Andhra Pradesh (including the newly formed Telangana), Gujarat and Tamil Nadu. While calculating pairwise F_{ST} values for different states and Madhya Pradesh, the highest F_{ST} value was 0.060 with the South Indian state of Tamil Nadu. This is not surprising, as this corresponds to the most geographically distant population pair sampled. Still, high F_{ST} values were detected even over short distances. For instance, the F_{ST} value of 0.028 for Madhya Pradesh and Gujarat was the second highest, but these are two of the geographically closest states. We also observed very low F_{ST} values for certain pairs of populations, for example, Andhra Pradesh and Rajasthan, Andhra Pradesh and Maharashtra, Andhra Pradesh and Gujarat, Bihar and Rajasthan, Bihar and Uttar Pradesh, Karnataka and Rajasthan, Maharashtra and Rajasthan, and Odisha and Rajasthan. Given the lack of prominent geographical boundaries between Indian states, as well as a history of shared territories, shifting administrative boundaries and migration of human populations, we had limited capacity to dissect the historical migration of pigeonpea from Madhya Pradesh to other states within India.

Impact of domestication and breeding on genetic diversity

It has been found that low levels of genetic diversity in annual crops like grain legumes arise because of human selection during domestication and breeding for favorable alleles and lead to small effective population sizes[8,9,13]. Loci selected during domestication and modern breeding should have the lowest variability across the genome. To infer the effects of domestication and breeding, we performed comparative analyses in two ways, by comparing wild species accessions with landraces and by comparing landraces with breeding lines. To detect selective sweeps driven by domestication and breeding, we calculated the reduction of diversity (ROD), based on the ratio of diversity in non-overlapping windows of 10 kb in length along the entire genome. From the comparison of wild species accessions with landraces and landraces with breeding lines, a total of 2,945 and 1,323 genomic regions, respectively, were identified with higher ROD values (**Supplementary Table 13**). These regions of reduced diversity in landraces and breeding lines might have experienced selective sweeps during domestication and breeding, respectively. Further, 1,722 and 671 genomic regions were identified as regions with maximum diversity loss (ROD = 1) during domestication and breeding, respectively. Maximum-ROD regions were also analyzed for pairwise F_{ST}. F_{ST} values for ROD regions reached up to 0.081 (breeding lines versus landraces) and 0.947 (landraces versus wild species) (**Supplementary Table 14**). These F_{ST} values provide evidence that the identified ROD regions have been subjected to selection pressure during domestication and breeding. Interestingly, we also identified 666 and 1,643 genomic regions with low genetic variation consistent

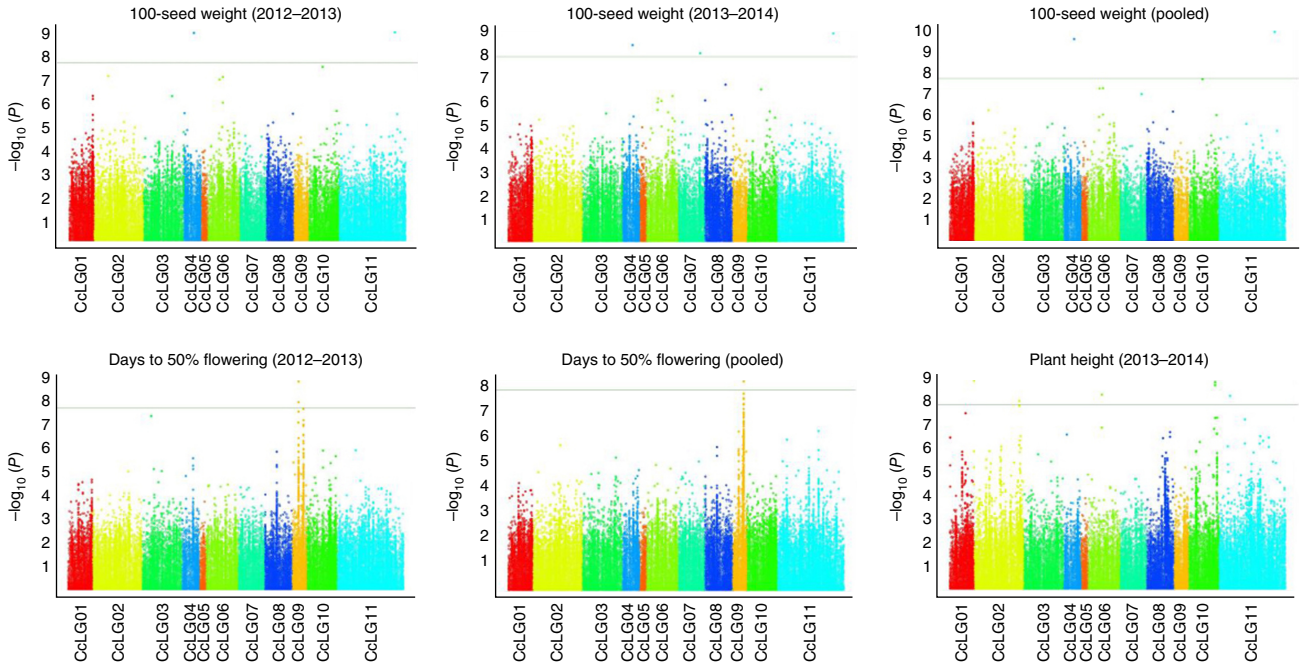

Figure 3 Significant marker–trait associations for 100-seed weight, days to 50% flowering and plant height. MTAs were detected through genome-wide association analysis using SUPER GWAS in the pigeonpea genome. The *y* axis in each graph represent $-\log_{10}P$ for the *P* value of the MTAs, while linkage groups are indicated on the *x* axis.

with positive selection during domestication and breeding, respectively. To assess the role of SVs in domestication and breeding, we searched the ROD regions for the presence of CNVs and PAVs. In total, there were 69 potentially selected SVs (68 CNVs and 1 PAV) detected as targets of selection (**Supplementary Tables 7** and **9**). Annotation of genes in genomic intervals affected by domestication and breeding (**Supplementary Table 15**) is presented in the **Supplementary Note**.

Genome-wide associations with agronomic traits

The genome-wide linkage disequilibrium (LD) decay rate, determined using 446,568 high-quality SNPs for cultivated pigeonpea, was 70 kb on average (**Supplementary Fig. 8** and **Supplementary Note**). The scale of LD across different pseudomolecules (CcLGs) was not uniform, possibly owing to recombination rate, natural selection, mutation, gene conversion or other causes (**Supplementary Figs. 9–12**). WGRS data from 286 accessions were analyzed with phenotyping data for eight agronomic traits collected over 2 years, 2012–2013 and 2013–2014 (**Supplementary Tables 16** and **17**), using the SUPER GWAS method[14] for estimating marker–trait associations (MTAs) (Online Methods). In total, 241 MTAs were identified with $P < 0.05$ for eight traits with data for two cropping seasons (**Supplementary Table 18**). The significance for MTAs was determined using a false discovery rate (FDR)-adjusted *P*-value threshold of $P < 0.05$. FDR-adjusted *P* values were calculated using R software. Of a total of 241 MTAs, 53 were detected for year 1 (cropping season 2012–2013), 90 were detected for year 2 (cropping season 2013–2014) and the remaining 98 were detected for the pooled data. A total of 37 MTAs were detected across different data sets. For instance, of the 26 MTAs identified across the data sets for 100-seed weight, 17 were identified in both cropping seasons as well as in the pooled data. The nine remaining MTAs were present in only one cropping season or only in the pooled data (**Fig. 3**).

We observed that many of the MTAs identified in a given year showed relatively weak or no associations in the other year. At least four traits, namely days to 50% flowering, plant height, number of primary branches per plant and number of secondary branches per plant, had high sensitivity to temperature and photoperiod length. For instance, we identified 29 and 1 MTA for days to 50% flowering and number of pods per plant, respectively, in the cropping season 2012–2013, yet detected only weak or no associations ($P > 0.05$) in the cropping season 2013–2014. For days to 50% flowering, 66 MTAs were detected in the pooled data, although only 20 of these were shared with the 2012–2013 cropping season and the pooled data set (**Supplementary Table 18**). Interestingly, 86 of a total of 95 MTAs for days to 50% flowering were detected on CcLG09. Several studies in the human genome have shown that closely linked SNPs tend to be in strong LD with each other, especially for alleles that increases the risk of complex inherited diseases[15–17]. The strong LD among associated SNPs is assumed to be true for days to 50% flowering in pigeonpea. On the other hand, one and two MTAs were detected for number of primary branches per plant and number of seeds per plant, respectively, only with the trait phenotyping data from the 2013–2014 season. In the case of the number of secondary branches per plant, one MTA was detected in the 2012–2013 data set and two MTAs were detected in the 2013–2014 data set, with no MTA common to the two seasons. In the case of plant height, 65 MTAs were detected in the 2013–2014 cropping season data set and 2 MTAs were detected in the pooled data set; no MTA could be identified for the 2012–2013 cropping season. For days to 75% maturity, eight MTAs were detected only with the pooled data set. GWAS analysis with the available data set suggests that many quantitative trait loci are associated with adaptation to environment in pigeonpea.

A further detailed analysis was used to count the number of favorable alleles in each accession for significant MTAs identified for 100-seed weight, days to 50% flowering and plant height (**Supplementary Table 19**).

100-seed weight is an important trait for pigeonpea, as it has a critical role in milling and in determining the price farmers obtain from traders. The acceptable seed weight for dehulled pigeonpea ('dhal') is 10–14 g/100 seeds. Keeping this in mind, the seed weight should be neither too small nor too large for milling. In case of vegetable pigeonpea, a larger seed size is preferred, as it yields a higher price. Hence, MTAs and accessions carrying favorable alleles identified in this study would help in rapidly improving pigeonpea varieties and/or breeding lines with seed weights outside of this desirable range through genomics-assisted breeding with sequence variation identified in this study (**Supplementary Table 20**).

The *Cajanus* genome shows the presence of variable numbers of copies of large, multi-kilobase regions in the genome of different groups of accessions. SVs could in principle account for a substantial proportion of variation in human and plant genomes. Initial discoveries in humans have identified associations of SVs with a range of disorders such as autism[18,19], schizophrenia[20,21] and neuroblastoma[22]. In some plant species, SVs have been reported to affect leaf size[23], fruit shape[24], aluminum tolerance[25] and agronomical traits, including leaf development and disease resistance[26]. In spite of this, the functional impact of most SVs has yet to be clarified[27]. Therefore, we mapped MTAs on identified SVs to explore the possible association with target traits. A total of 183 MTAs for all the target traits were mapped on 63 SVs across the genome in breeding lines (**Supplementary Table 21**). In the case of wild species, 29 MTAs were detected for different traits located in 19 SVs across the genome (**Supplementary Table 21**). Interestingly, 64 of 86 MTAs located on CcLG09 for days to 50% flowering were clustered on six SVs. We also detected 26 MTAs for 100-seed weight corresponding to 21 SVs across the genome in breeding lines. These results indicate that SVs had a crucial role in improving the fitness of breeding lines as compared to their progenitor wild species. Moreover, the abundance of MTAs on CcLG09, along with the long LD blocks and evidence of a 'hitchhiking effect', indicates that this pseudomolecule has been strongly affected by domestication and breeding.

DISCUSSION

The availability of genomic resources such as genome sequence or molecular markers is not in itself enough to improve crop productivity. An effective means to harness the enormous genetic diversity present in the germplasm collections of genebanks for traits of interest to breeders is an acute need[28]. A systematic evaluation and utilization of available germplasm is important for crop improvement through accessing allelic variations affecting important agronomic traits. In the present study, from analysis of resequencing data of wild species, landraces and breeding lines, we identified regions of the pigeonpea genome that have undergone selective sweeps corresponding to domestication and modern breeding. We observe that both favorable alleles for agronomic traits and SVs contribute to reductions in molecular diversity and in specific genomic regions, such as on CcLG09, that have experienced selective sweeps. The extent of genetic bottlenecking in pigeonpea during domestication from wild *Cajanus* is strong but is much more moderate in going from landraces to breeding lines (for example, see **Supplementary Fig. 3**). This weaker bottlenecking from landraces to breeding lines is less severe than the paradigm for domestication, for example, in maize[26]. This may be due to the relatively recent onset of intensive breeding efforts in pigeonpea, an absence of broadly adapted megavarieties that serve as predominant breeding lines (and thus suppress levels of diversity within modern breeding programs) or the focus of breeding of pigeonpea on a range of locally relevant agroecologies and consumer preferences. The limited intensive breeding history in pigeonpea is evident, for example, from the observation that ICP15629, an accession of the progenitor *C. cajanifolius* (**Fig. 1c**), appears to have served as a cultivated landrace. Furthermore, basal clades identified from molecular taxonomy (**Fig. 1c**) comprise not only additional landraces but also breeding lines. In addition, the same analysis also indicates that less basal clades also comprise both landraces and breeding lines (**Fig. 1c**), with only relatively few fine-scale and more terminal clades exclusively comprising either landraces or breeding lines, again supporting the notion of a limited extent of intensive breeding in pigeonpea.

Although genes that are functionally characterized in other species served as candidate genes[29] in our analysis (for example, pigeonpea homologs of *LIGULELESS1*, *SHATTERING1* and *EARLY FLOWERING3* (*ELF3*)), it is formally possible that the targets of selection in pigeonpea breeding may be other nearby loci that, from an absence of functional characterization, are not obvious candidates. Further study will be needed to delineate the causal genes for these and other MTAs, such as those for 100-seed weight and other traits. The data we have obtained from resequencing and GWAS will facilitate future studies of the genetic underpinning of agronomically relevant traits, which may also facilitate comparative studies or serve as candidates for similar traits in other crops. Targeting of an *ELF3* ortholog in pigeonpea during domestication is reminiscent of the role of *ELF3* orthologs in the domestication of garden pea and lentil[30]. In addition to underscoring a major role for this gene and its orthologs in the domestication of both temperate legumes (pea and lentil) and the tropical legume pigeonpea, the reduced photoperiod sensitivity engendered by *ELF3* variation in pigeonpea may also facilitate its spread and cultivation across broader latitudinal ranges.

In conclusion, genome data for the reference set of pigeonpea, inferences drawn from their analysis and MTAs for agronomically relevant traits collectively provide valuable resources to accelerate genetic gains in pigeonpea crop improvement programs to the benefit of smallholder farmers in the developing world who grow this multipurpose food security crop.

URLs. Genome assembly for pigeonpea, https://www.ncbi.nlm.nih.gov/bioproject/72815; MEGA4, http://www.megasoftware.net/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors are thankful to the US Agency for International Development (USAID) for providing financial support to R.K.V. The authors would like to thank A. Gafoor, B. Poornima and P. Bajaj for their support in this work. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes. ICRISAT is a member of the CGIAR Consortium.

AUTHOR CONTRIBUTIONS

R.K.V., R.K.S., Y.Y., C.K., D.K., J.K., S.A., V.K., J.-S.K. and W.Z. contributed to generation of whole-genome resequencing data. H.D.U. and R.K.V. contributed genetic material. H.D.U., G.A., K.N.Y. and S.M. performed phenotyping. R.K.V., R.K.S., H.D.U., A.W.K., C.K., A.R., D.K., J.K., S.A., J.-S.K., R.V.P., E.v.W. and S.K.D. worked on different analyses. R.K.V. and R.K.S. together with C.K., A.R., J.-S.K., R.V.P. and E.v.W. wrote and finalized the manuscript. R.K.V. and R.K.S. directed the project, and R.K.V. conceived and designed the study.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Van der Maesen, L.G.J. in *The Pigeonpea* (eds. Nene, Y.L., Hall, S.D. & Sheilla, V.K.) 15–46 (C.A.B. International, 1990).
2. Saxena, R.K. *et al.* Genetic diversity and demographic history of *Cajanus* spp. illustrated from genome-wide SNPs. *PLoS One* **9**, e88568 (2014).
3. Vavilov, N.I. The origin, variation, immunity, and breeding of cultivated plants. *Chron. Bot.* **13**, 1–366 (1951).
4. Varshney, R.K. *et al.* Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* **30**, 83–89 (2011).
5. Varshney, R.K., Thudi, M., May, G.D. & Jackson, S.A. Legume genomics and breeding. *Plant Breed. Rev.* **33**, 257–304 (2010).
6. Upadhyaya, H.D. *et al.* Phenotyping chickpeas and pigeonpeas for adaptation to drought. *Front. Physiol.* **3**, 179 (2012).
7. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
8. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsisthaliana*. *Science* **317**, 338–342 (2007).
9. Lam, H.M. *et al.* Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* **42**, 1053–1059 (2010).
10. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
11. Tang, H., Peng, J., Wang, P. & Risch, N.J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
12. Van der Maesen, L.J.G. *Cajanus DC and Atylosia W. & A. (Leguminosae) (Agricultural University Wageningen Papers)* (Wageningen Universiteit Project, 1986).
13. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
14. Wang, Q., Tian, F., Pan, Y., Buckler, E.S. & Zhang, Z. A SUPER powerful method for genome wide association study. *PLoS One* **9**, e107684 (2014).
15. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
16. Gudbjartsson, D.F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).
17. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
18. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
19. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
20. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
21. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat. Genet.* **41**, 1223–1227 (2009).
22. Diskin, S.J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987–991 (2009).
23. Horiguchi, G., Gonzalez, N., Beemster, G.T., Inzé, D. & Tsukaya, H. Impact of segmental chromosomal duplications on leaf size in the *grandifolia-D* mutants of *Arabidopsisthaliana*. *Plant J.* **60**, 122–133 (2009).
24. Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J. & van der Knaap, E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527–1530 (2008).
25. Maron, L.G. *et al.* Aluminum tolerance in maize is associated with higher *MATE1* gene copy number. *Proc. Natl. Acad. Sci. USA* **110**, 5241–5246 (2013).
26. Chia, J.M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
27. Saxena, R.K., Edwards, D. & Varshney, R.K. Structural variations in plant genomes. *Brief. Funct. Genomics* **13**, 296–307 (2014).
28. Upadhyaya, H.D. *et al.* Pigeonpea composite collection and identification of germplasm for use in crop improvement programmes. *Plant Genet. Resour.* **9**, 97–108 (2011).
29. Meyer, R.S. & Purugganan, M.D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
30. Weller, J.L. *et al.* A conserved molecular basis for photoperiod adaptation in two temperate legumes. *Proc. Natl. Acad. Sci. USA* **109**, 21158–21163 (2012).

ONLINE METHODS

DNA extraction and sequencing. Single plants from each *Cajanus* accession were used to collect young leaves (3 to 4 weeks after planting; leaves frozen in liquid nitrogen). Total DNA was extracted with the cetyltrimethylammonium bromide (CTAB) method following a standard procedure. At least 10 µg of genomic DNA from each accession was used to construct a sequencing library. Paired-end sequencing libraries with an insert size of approximately 400 bp were sequenced on an Illumina HiSeq 2000 sequencer. Paired-end resequencing reads were mapped using BWA[7] (ver. 0.5.9) with default parameters onto the pigeonpea reference genome[4]. Mapped reads were converted into BAM files using SAMtools (ver. 0.1.18), and duplicate reads were removed with SAMtools. Genome coverage of mapped reads on the reference genome was estimated using GATK (ver. 1.4-11).

Variation calling and annotation. After removing duplicate reads, variants including SNPs and indels were detected using BCFtools (ver. 0.1.17) in SAMtools. Captured variants were annotated with ANNOVAR (ver. 2011Nov28) and SnpEff (ver. 3.2). For these variants, SNPs were counted with VCFtools (ver. 0.1.10) and indels were counted with bedops (ver. 2.4.3) and in-house Perl scripts. Using the *C. cajan* genome assembly (see URLs) in SnpEff, the genetic position of variants was identified and variants at each position were counted.

Identification of CNVs and PAVs. CNVs for the 292 samples were identified with Control-FREEC (ver 7.0). Because pigeonpea is a diploid plant, two copies were considered to be the threshold. Copy number values greater than 2 were considered to represent a 'gain' while those less than 2 were considered to represent a 'loss'. A PAV matrix was generated from the CNV data set using an in-house Perl script.

Population genetics analysis. We filtered variants with quality score >30 from a VCF file. Using PLINK (ver. 1.9), we generated a pruned SNP set that has approximate linkage equilibrium between SNPs. The option used for analysis included a window of 50 SNPs, a shift window of 5 SNPs at each step and $r^2 = 0.6$. Finally, the neighbor-joining tree was constructed using MEGA4 software on the basis of a distance matrix (see URLs). PCA of whole-genome SNPs was performed with EIGENSOFT 4.2, and the first two eigenvectors were plotted in 2D.

Ratios of nonsynonymous and synonymous SNPs. We identified high-variance genomic regions on the basis of patterns of synonymous (S) and nonsynonymous (N) SNPs. Numbers of synonymous and nonsynonymous SNPs were obtained for each pseudomolecule with an interval of 10 kb or 1 Mb and were then averaged across accessions for wild species, landraces and breeding lines separately. An N/S ratio was then calculated across each linkage group for all three accession types and was plotted for comparison. The resulting plots showed that the wild species accessions had minimum variation throughout the genome, with the average N/S ratio equal to ~1, whereas the breeding lines and landraces were highly variable when compared to wild species accessions. For the breeding lines and landraces, we looked for regions where the N/S ratio was less than that for the wild species and above an N/S ratio cutoff of 2.5 for further in-depth analysis.

Reduction of diversity. To detect ROD values in two different combinations of wild species accessions versus landraces and landraces versus breeding lines, genetic diversity (π) was calculated[31] in wild species accessions, landraces and breeding lines. Further, to detect the genomic regions affected by domestication and breeding, we calculated ROD values in 10-kb non-overlapping windows as following

$$\mathrm{ROD}_{\mathrm{lw}} = 1 - \pi_{\mathrm{landraces}}/\pi_{\mathrm{wild}}$$

$$\mathrm{ROD}_{\mathrm{bl}} = 1 - \pi_{\mathrm{breeding\,lines}}/\pi_{\mathrm{landraces}}$$

Geographical differentiation analyses. F_{ST} provides insights into the evolutionary processes that influence the genetic variation within and among populations and is one of the most widely used statistics in population genetics. To calculate F_{ST} between populations grouped on the continent, country and state scale, we used the method of Weir and Cockerham[32] in the R language. At first, we calculated genome-wide diversity for each genotype; the population-specific mean diversities were calculated as the arithmetic mean across the genotypes.

Phenotyping. Germplasm accessions in the pigeonpea reference set were planted in different experimental fields during the rainy seasons of 2012–2013 and 2013–2014 at ICRISAT, Patancheru, India. ICRISAT is located at an altitude of 545 m above sea level, 17°32′ N, 78°16′ E. All accessions were grown in the fields in well-distributed alfisols (red soils) and uniform conditions. Fields were selected with good drainage conditions and free from weeds. Before sowing of seeds, the fields were prepared by deep plowing followed by 2–3 passes of a harrow, leveling and construction of ridges spaced 75 cm apart. We applied a basal dose of diammonium phosphate (DAP) at 100 kg/ha. Subsequently, sowing was done by hand with 3–4 seeds per hill spaced at a distance of 25 cm along the ridge. Irrigation was provided after sowing (in case of low soil moisture) and subsequently to maintain the crop throughout the entire cropping season. Fifteen days after sowing, seedlings were thinned out manually to achieve an equal density of each individual accession. All the field management practices followed during the cropping seasons, including irrigation, weed management, etc., were carried out equally in each year for all accessions. For phenotyping, all accessions were planted in two replications in an α-lattice design. Three individual plants from each accession in every replication were used to collect the trait phenotyping data. The phenotyping procedure and scoring standard followed practices outlined in the genebank manual[33]. The pigeonpea germplasm accessions of the reference set were phenotyped for eight agronomic traits (days to 50% flowering, plant height, primary branches per plant, secondary branch per plant, 100-seed weight, days to 75% maturity, pods per plant and seeds per pod). The data for each year (2012–2013 and 2013–2014) and the pooled data over both years were analyzed using residual maximum likelihood (REML) on the GENSTAT software program. The data were analyzed for grand mean (GM), standard error of differences (SED), least significant difference (LSD), coefficient of variation (CV) and heritability (HERT).

Genome-wide LD and association (GWAS) analyses. After identifying genome-wide SNPs across the populations, SNP loci for which more than 20% of data were missing across accessions and with MAF less than 5% were eliminated from the analysis. As the sample size for the accessions of wild species was far smaller, we did not use wild species accessions for LD decay estimation. LD was measured by the pairwise correlation coefficient (r^2) for each SNP pair in TASSEL 5.0 (ref. 34). The P values for each r^2 estimate were determined with the two-sided Fisher's exact test implemented in TASSEL. Only r^2 values with $P < 0.05$ were included in further analyses. Association analysis for the target traits was carried out using the SUPER (settlement of MLM under progressively exclusive relationship) GWAS method[35] employed using GAPIT[36]. In the SUPER GWAS method, genome-wide SNPs were divided into small bins. Each bin comprised the most significant SNPs. Subsequently, influential bins were selected. Further, a maximum-likelihood method was used to optimize the size and number of bins selected. Kinship was defined for associated markers and to exclude the markers that are in LD with the testing markers. Furthermore, principal components[37] were calculated and used as fixed effects to correct for stratification. Linear model testing was performed by plotting the observed P values from the association test against an expected (cumulative) probability distribution. These quantile–quantile plots indicate the extent to which the analysis produced more significant results than expected by chance. The significance of MTAs was determined using a threshold of FDR-adjusted $P < 0.05$. FDR-adjusted P values were calculated using R.

Data availability. The WGRS data set generated and analyzed in the current study is available from NCBI under BioProject accession PRJNA383013.

31. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).

doi:10.1038/ng.3872

32. Weir, B.S. & Cockerham, C.C. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**, 1358–1370 (1984).
33. Upadhyaya, H.D. & Gowda, C.L.L. *Managing and Enhancing the Use of Germplasm—Strategies and Methodologies* (International Crops Research Institute for the Semi-Arid Tropics, 2009).
34. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
35. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2011).
36. Lipka, A.E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
37. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).