nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0401-3

Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits

Rajeev K. Varshney[1*], Mahendar Thudi[1], Manish Roorkiwal[1], Weiming He[2], Hari D. Upadhyaya[1], Wei Yang[2], Prasad Bajaj[1], Philippe Cubry[3], Abhishek Rathore[1], Jianbo Jian[2], Dadakhalandar Doddamani[1], Aamir W. Khan[1,4], Vanika Garg[1,5], Annapurna Chitikineni[1], Dawen Xu[2], Pooran M. Gaur[1], Narendra P. Singh[6], Sushil K. Chaturvedi[6,22], Gangarao V. P. R. Nadigatla[7], Lakshmanan Krishnamurthy[1], G. P. Dixit[6], Asnake Fikre[8,23], Paul K. Kimurto[9], Sheshshayee M. Sreeman[10], Chellapilla Bharadwaj[11], Shailesh Tripathi[11], Jun Wang[2,12], Suk-Ha Lee[13], David Edwards[4], Kavi Kishor Bilhan Polavarapu[5], R. Varma Penmetsa[14], José Crossa[15], Henry T. Nguyen[16], Kadambot H. M. Siddique[4], Timothy D. Colmer[4], Tim Sutton[17,18], Eric von Wettberg[19], Yves Vigouroux[3], Xun Xu[2,20*] and Xin Liu[2,21*]

We report a map of 4.97 million single-nucleotide polymorphisms of the chickpea from whole-genome resequencing of 429 lines sampled from 45 countries. We identified 122 candidate regions with 204 genes under selection during chickpea breeding. Our data suggest the Eastern Mediterranean as the primary center of origin and migration route of chickpea from the Mediterranean/Fertile Crescent to Central Asia, and probably in parallel from Central Asia to East Africa (Ethiopia) and South Asia (India). Genome-wide association studies identified 262 markers and several candidate genes for 13 traits. Our study establishes a foundation for large-scale characterization of germplasm and population genomics, and a resource for trait dissection, accelerating genetic gains in future chickpea breeding.

Changes in global climate pose immense challenges for plant breeders to maintain and further enhance yield in varying environments[1]. Globally, more than 2 billion people experience micronutrient deficiency, per The World Health Organization estimates[2] (https://ourworldindata.org/micronutrient-deficiency). The development and adoption of improved crop varieties with higher yield and nutrition is expected to reduce the number of malnourished people across the world, especially in South Asia and Sub-Saharan Africa[3].

Chickpea (*Cicer arietinum* L.) is an important source of protein for millions of people in developing countries. It is also a rich source of β-carotene and minerals including phosphorus, calcium, magnesium, iron and zinc. In addition, chickpea crops add 60–103 kg ha^{-1} nitrogen to the soil through symbiotic nitrogen fixation. Drought and heat are among the major abiotic stresses that can cause more than 70% yield loss in chickpea. The productivity of chickpea, a cool-season legume crop, is expected to be further reduced by the predicted increase in global temperature due to global warming. A lack of genetic diversity in chickpea, stemming potentially from a series of bottlenecks in its evolutionary past, has long been thought to exacerbate the challenge posed by these abiotic and biotic stresses[4]. If so, one would expect both domestication and modern selection to have eroded genetic variation, leaving wild relatives as the most diverse germplasm reservoir and landraces as a resource of intermediate diversity. A few studies with a limited number of markers have enabled identification of genomic regions associated with abiotic and biotic tolerance traits[5,6], showing that some variation for tolerance to these stresses is available and that germplasm resources could be harnessed to effectively meet these challenges. More than 90,000 chickpea accessions have been conserved in genebanks globally[7]. The diversity of the crop has been poorly exploited, owing to limited availability of high-density marker information and detailed

[1]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India. [2]BGI-Shenzhen, Shenzhen, China. [3]Institut de Recherche pour le Développement (IRD), University of Montpellier, Montpellier, France. [4]The University of Western Australia (UWA), Crawley, Western Australia, Australia. [5]Osmania University, Hyderabad, India. [6]Indian Council of Agricultural Research (ICAR), Indian Institute of Pulses Research (IIPR), Kanpur, India. [7]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Nairobi, Kenya. [8]Ethiopian Institute of Agricultural Research (EIAR), Debre Ziet, Ethiopia. [9]Egerton University, Njoro, Kenya. [10]University of Agricultural Sciences–Bangalore, Bangalore, India. [11]ICAR–Indian Agricultural Research Institute (IARI), New Delhi, India. [12]iCarbonX, Shenzhen, China. [13]Seoul National University, Seoul, South Korea. [14]University of California–Davis, Davis, CA, USA. [15]International Maize and Wheat Improvement Center (CIMMYT), Mexico, Mexico. [16]University of Missouri, National Center for Soybean Biotechnology, Columbia, SC, USA. [17]South Australian Research and Development Institute, Adelaide, South Australia, Australia. [18]University of Adelaide, Glen Osmond, South Australia, Australia. [19]University of Vermont, Burlington, VT, USA. [20]China National Gene Bank (CNGB), Shenzhen, China. [21]State Key Laboratory of Agricultural Genomics, Shenzhen, China. [22]Present address: Rani Lakshmi Bai Central Agricultural University, Jhansi, India. [23]Present address: International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Addis Ababa, Ethiopia. *e-mail: r.k.varshney@cgiar.org; xuxun@genomics.cn; liuxin@genomics.cn

phenotypic information for key adaptive traits. Whole-genome resequencing (WGRS) has proved useful for understanding the extent and patterns of genetic variation, population structure, linkage disequilibrium and unused genetic potential for crop improvement in some crop species[8–15]. The advent of next-generation sequencing (NGS) technology has drastically reduced the cost for sequencing[16] and enabled whole-genome-level analysis to identify alleles gained or lost during domestication, diversification and adaptation. This knowledge could be used to develop climate-change-resilient varieties. By coupling whole-genome information with detailed study of phenotypic variation, it is possible to harness accessions with low frequency variants that may contribute to key phenotypes such as abiotic and biotic stress tolerance or yield components.

This study uses the power of NGS technology to harness the germplasm wealth available in genebanks and provides insights into naturally occurring genetic variation, population structure, domestication and selection in 429 chickpea genotypes that encompass the diversity of cultivated chickpea. We assess the extent to which the 300 accessions of the reference set[17] show a decline in diversity from wild relatives to landraces and then breeding lines, as well as the extent of diversification of landraces into major market classes and into geographically distinct forms that may reflect different patterns of cultivation and use in divergent regions with long histories of chickpea use such as the Fertile Crescent, South Asia and the East African highlands. We propose a new migration route of chickpea from the Fertile Crescent in the Eastern Mediterranean to South Asia, where >90% chickpea cultivation currently occurs, as well as to other regions. In addition, we also established marker-trait associations (MTAs) for drought and heat tolerance related traits that can be used in marker-assisted breeding to develop new chickpea varieties with enhanced yield and climate resilience.

Results

Germplasm sequencing, genome-wide variations, population structure and linkage disequilibrium decay. We undertook WGRS of the chickpea reference set (300 genotypes) and analyzed the data along with WGRS data on 100 chickpea released varieties[18] and 29 lines from the chickpea genome paper[19]. Thus, in total, 429 chickpea genotypes were used to understand genome diversity, population structure, crop domestication and post-domestication diversification (Supplementary Table 1 and Supplementary Fig. 1). In brief, we analyzed 2.57 terabase pairs (Tbp) of raw data comprising 28.36 billion reads with an average of 10.22× coverage or 6 gigabase pairs (Gbp) of raw data per sample. Aligning the cleaned reads to the chickpea reference genome assembly of CDC Frontier[19] resulted in 10.21× vertical and 95.33% horizontal genome coverage, while unique mapping provided 6.84× mean depth and 88.06% average genome coverage (Supplementary Table 1). The coverage of resequencing data is comparable to earlier studies in pigeonpea[10], pearl millet[11], maize[13], rice[14], soybean[8,15] and chickpea[18,19].

Using the mapped re-sequence data, we identified genome-wide variations including 4,972,803 single-nucleotide polymorphisms (SNPs), 596,100 small insertions or deletions (indels), 4,931 copy number variations (CNVs) and 60,742 presence absence variations (PAVs) across 429 lines (Table 1). Of the 4.97 million SNPs, most (85%) were present in intergenic regions and an average of 4% SNPs were located in coding sequence. We also analyzed the reference genotype (CDC Frontier sequenced at ~11.9×) by using the same SNP-calling procedure and identified 107,375 heterozygous SNPs and 20,544 homozygous SNPs. The homozygous SNPs are more likely to be variant calling errors; hence the error rate should be less than 1%, given that the sequenced genome length is ~532 megabase pairs (Mbp). The ratio of non-synonymous to synonymous SNPs varied between 0.86 (ICC 20194) and 1.56 (PBA HatTrick) with an average of 1.20 (Supplementary Table 2), which is comparable to that observed in sorghum (1.0)[20], pigeonpea (1.18)[10], rice (1.29)[14] and soybean (1.36)[15]. In comparing different chickpea seed types, the abundance of SNPs, indels and PAVs was higher in desi genotypes as compared to kabuli genotypes while CNV abundance was lower (Fig. 1 and Supplementary Table 3). Similarly, in comparing the genotypes on the basis of biological status, a higher number of variations was observed in landraces compared to elite cultivars and breeding lines (Fig. 1 and Supplementary Tables 4–9). The genome-wide variations identified in this study were more abundant than in previous studies[18,19], probably due to large number of genotypes used in the present study. Wild species genotypes had more unique SNPs; however, 523,260 common SNPs were identified in landraces, breeding lines, elite cultivars and wild chickpea genotypes (Fig. 2a,b). In the present study, we identified 4.7% (95) non-frameshift or neutral indels (Fig. 2c) that block protein synthesis of genes involved in transcription factor activity and DNA binding activity (Supplementary Fig. 2). Among desi seed types, ICC 15618, a heat tolerant genotype had the maximum number of CNVs, when compared to CDC Frontier (Supplementary Table 10). Of 1,202 CNVs in coding regions, 86.18 % (1,036 CNVs) had predicted function, whereas 89.26% of PAVs (6,606) had predicted functions (Supplementary Tables 11 and 12). Gene ontology annotation of CNVs indicated that these genes are involved in regulation of multi-cellular organismal processes (Supplementary Fig. 3). PAV genes are involved in regulation of cellular processes, response to stimulus and reproduction (Supplementary Fig. 4).

We determined three sub-populations ($K=3$) using the Admixture model in STRUCTURE[21]. Allelic admixture in some genotypes is evident among different sub-populations (Supplementary Fig. 5), probably due to the breeding history among cultivated chickpea genotypes, as reported earlier[22]. We also explored relationships among 429 chickpea genotypes on the basis of 4.97 million SNPs using principal coordinate and phylogenetic analyses. More than 25% of genetic variance was explained by principal component 1 and principal component 2 (Supplementary Fig. 6). The seven accessions of wild species form an out-group from the cultivated genotypes. Of the cultivated chickpea genotypes, Pusa 1103 ((Pusa 256 × *Cicer reticulatum*) × Pusa 362; an elite variety developed at Indian Agricultural Research Institute (IARI), New Delhi, India, tolerant to drought and soil borne diseases) and ICC 9636 (landrace originating from Afghanistan) grouped away from the other cultivated genotypes. Further, among cultivated genotypes, desi and kabuli genotypes formed separate clusters with little admixture. Phylogenetic analysis also revealed four clusters, which further supports the presence of four sub-populations (Fig. 3). Among the four clusters identified, Cluster I (170 genotypes) is the largest group followed by Cluster III (110 genotypes), Cluster II (84 genotypes) and Cluster IV (58 genotypes). Clusters I and III are dominated by elite cultivars while Clusters II and IV have been dominated by landraces (Supplementary Table 13 and Fig. 3). Clustering of two breeding lines (CDC Vanguard and ICC 14402) and two elite cultivars (Dohad Yellow and ICC 96970) within Cluster II, along with 80 landraces, may be due to the presence of one or more of the landraces in their genetic background. On the basis of market class, all the clusters are dominated by desi genotypes with the exception of Cluster II, which is dominated by kabuli genotypes and Cluster IV is the smallest cluster with 65% landraces (Fig. 3) and the kabuli market class. Nevertheless, all four clusters were interspersed with landraces, breeding lines and elite cultivars, potentially reflecting breeding for different environments or market types. On the basis of geographical distribution of lines, there was no clear demarcation of different clusters. Clusters I, II, III and IV comprised 170, 84, 110 and 58 genotypes from 27, 16, 27 and 18 countries, respectively. This indicates substantial historical movement of germplasm, wider use of diverse pedigrees in developing breeding lines than previously appreciated or multiple origins of ecotypes adapted to different climatic contexts (for example, temperate, Mediterranean

Table 1 | Genome-wide variations identified in 429 chickpea genotypes

Groups[a]	SNPs					Indels				CNVs	PAVs
	Total	Intron	Intergenic	Exon	Others	Total	Intron	Intergenic	Exon		
All genotypes (429)	4,972,803	512,627	4,239,339	194,844	25,993	596,100	95,117	495,387	5,596	4,931	60,742
Market type (412)	4,956,853	511,569	4,224,651	194,690	25,943	595,650	95,080	494,983	5,587	-	-
Desi (272)	3,405,151	324,925	2921861	137,948	20,417	222,285	25,563	194,565	2,157	3,603	44,094
Kabuli (128)	2,730,493	254,237	2,347,305	117,620	11,331	172,111	19,376	151,198	1,537	3,860	40,089
Pea shaped (12)	1,207,509	108,848	1,043,051	49,657	5,953	74,206	9,550	63,977	679	2,039	25,407
Biological status (419)	4,968,032	512,253	4,235,043	194,766	25,970	595,038	95,020	494,439	5,579	-	-
Landraces (268)	3,345,197	316,831	2,873,798	134,377	20,191	227,247	25,840	199,049	2,358	3,180	40,694
Elite lines (100)	2,422,703	222,882	2,085,841	98,460	15,520	121,833	14,282	106,516	1,035	2,503	33,250
Breeding lines (44)	1,932,979	175,029	1,662,291	86,434	9,225	137,622	15,325	121,215	1,082	3,659	35,837
Wild (7)	3,897,195	432,632	3,284,577	170,854	9,132	429,611	79,175	346,790	3,646	2,736	38,905
C. echinospermum (3)	3,228,018	357,233	2,725,269	138,057	7,459	269,358	50,443	216,537	2,378	2,111	27,321
C. reticulatum (4)	2,608,835	313,884	2,164,815	126,366	3,770	258,187	50,325	205,587	2,275	2,073	27,896

[a]Numbers of genotypes in each group are given in parentheses.

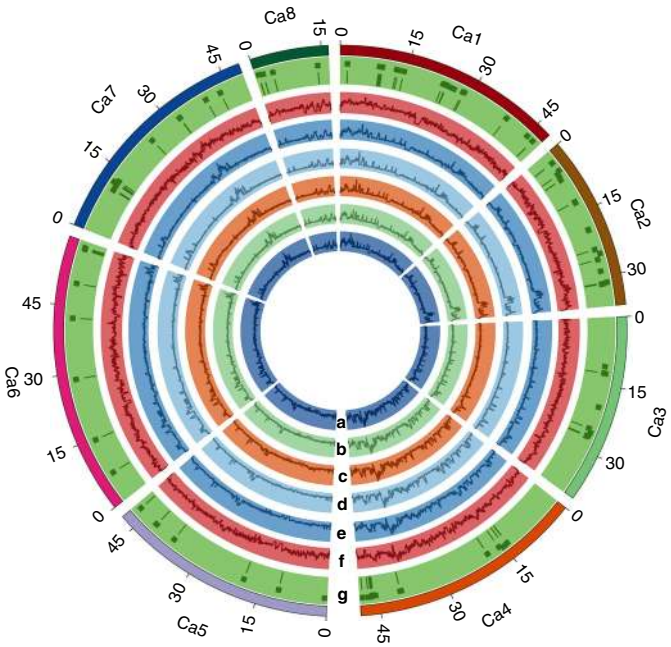

Fig. 1 | A circos diagram illustrating the genome-wide variations among 429 chickpea lines. a–**g**, Eight pseudomolecules traverse from in to out. **a**, SNP density in desi genotypes; **b**, SNP density in kabuli genotypes; **c**, SNP density in cultivars; **d**, SNP density in breeding lines; **e**, SNP density in landraces; **f**, SNP density in wild lines; **g**, the candidate genomic regions underwent selection during crop breeding and post-domestication diversification. The rectangles represent the genomic regions and lines represent the genes within these regions. A very high number of SNPs were observed in wild lines as compared to landraces, cultivars and breeding lines. A total of 122 candidate genomic regions harboring 204 genes were identified, of which the maximum regions were present on pseudomolecule Ca1 followed by Ca2.

or sub-tropical)[22,23]. Grouping of drought tolerant and heat tolerant lines in different clusters indicated optimal variability for these traits in the germplasm studied, which can be deployed for chickpea improvement (Supplementary Table 1).

To understand the linkage disequilibrium patterns between different chromosomes of desi, kabuli and all chickpeas, we calculated r^2 between pairs of SNPs by using Haploview[24]. The linkage disequilibrium of overall samples dropped to half of its maximum at 180 kilobases (kb) ($r^2 = 0.2$). The linkage disequilibrium decay among desi (190 kb) and kabuli genotypes (210 kb) did not significantly differ. However, on the basis of biological status, linkage disequilibrium decay in breeding lines was slower (~320 kb) compared to landraces (~180 kb) and elite cultivars (~190 kb) (Supplementary Fig. 7a,b). Chromosome-wise linkage disequilibrium decay varied from ~100 to ~425 kb among different groups of chickpea (Supplementary Fig. 8). In general, the linkage disequilibrium decay observed in the present study was similar to cultivated soybean (150 kb)[8] and slower than cereals; for example, rice (<10 kb for *O. rufipogon* and *O. nivara*, 65 kb for *indica*)[14], sorghum (19.7 kb and 10.3 kb for the improved inbreds and landraces)[20] and maize (<1 kb)[25]. The bigger linkage disequilibrium blocks in breeding lines may be due to selection for positive alleles during breeding programs and the self-pollinated nature of the crop.

Genomic regions affected by selection during and after domestication. To understand the diversity patterns, we estimated the diversity parameters in the whole population and sub-populations. We observed broad variation in nucleotide diversity among individual pseudomolecules (Ca1 to Ca8) of landraces, ranging from 0.47 per kb (Ca5) to 1.62 per kb (Ca4) and with an average nucleotide diversity of 0.81 per kb at the whole-genome level (Supplementary Table 14). The distribution of nucleotide diversity per kb indicates high allelic diversity in wild chickpea genotypes compared to landraces despite a very small number of wild genotypes being used in this study (Fig. 4a and Supplementary Fig. 9). The overall nucleotide diversity was much lower than reported in *Medicago truncatula*[26] (4.3 per kb), wild soybean (3.0 per kb) and cultivated soybean[15] (1.9 per kb). A significant reduction in diversity was observed from wild genotypes (3.80 per kb) to landraces (0.86 per kb) and breeding lines (0.84 per kb) (Table 2, Supplementary Table 14 and Fig. 4b), suggesting that about 80% of genetic diversity captured in this study has been lost during chickpea domestication[27]. We observed negative Tajima's *D* values on all eight pseudomolecules in landraces and elite cultivars, indicating an excess of low frequency polymorphisms relative to expectation (Supplementary Table 14). These low values are consistent with population size expansion (for example,

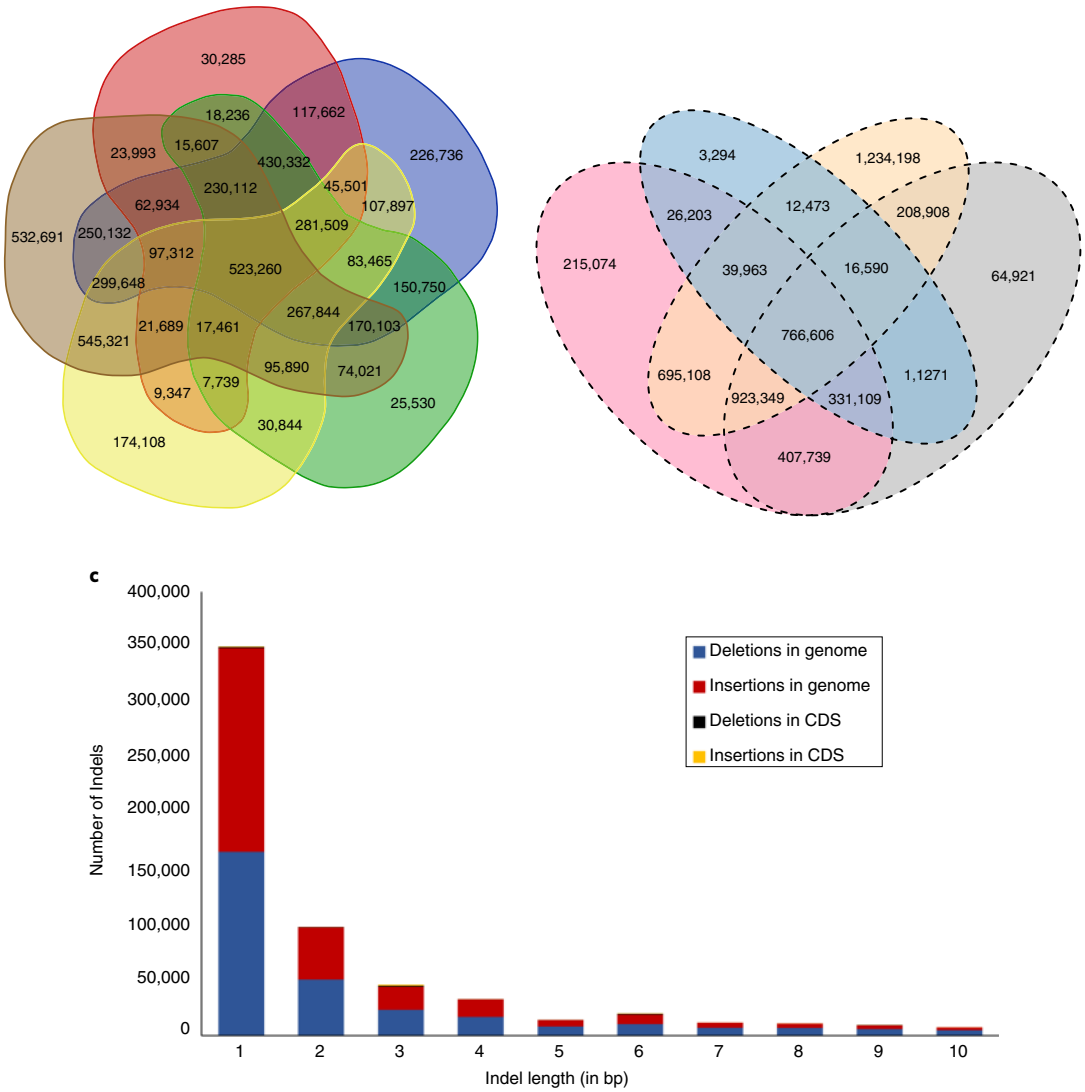

Fig. 2 | Genome-wide variations, population structure and genetic diversity in 429 chickpea genotypes. a, Venn diagram representing the number of unique and shared SNPs on the basis of biological status. A total of 523,260 SNPs were common among different chickpea genotype groups including breeding lines, landraces, elite cultivars and wild genotypes (*C. reticulatum* and *C. echinospermum*). **b**, Venn diagram representing the number of unique and shared SNPs on the basis of seed type. A total of 766,606 SNPs were common among desi, kabuli, pea and wild chickpea genotypes. **c**, Distribution of small insertions and deletions in genomic and coding regions.

after a bottleneck or a selective sweep) and/or purifying selection, also indicating a possible strong selection during domestication and post-domestication diversification[28,29].

To identify possible regions and genes under selection during crop improvement and post-domestication diversification, where we have the greatest sampling of genotypes, we calculated reduction of diversity (ROD) and the population difference (*F* index, F_{ST}) in 20 kb non-overlapping windows along the genome[30]. Regions with substantial lower diversity level in the breeding lines compared to the landraces (high ROD, top 5% in the whole genome), and substantial high population differences (high F_{ST}, top 5% in the whole genome) between the two germplasm types were identified as possible candidate regions under selection during more recent crop breeding. We found consistent regions using 20 kb and 100 kb windows, possibly because of the relatively low number of genomic regions that underwent post-domestication selection in chickpea or the larger genome size of chickpea. In total, we identified 122 candidate regions with 204 genes under selection during crop breeding and post-domestication-diversification (Fig. 1, Supplementary Figs. 10–13 and Supplementary Tables 15 and 16). Then we carried out gene ontology annotations to investigate possible functions of those genes. The gene onotology terms of the biological process category revealed that among 204 candidate genes under selection during crop breeding, were mostly related to response to stress, DNA repair, protein kinase activity, seed development, germination and flower development, suggesting selection for key biotic and abiotic stress resistance and phenological related traits (Supplementary Table 16). Among 204 genes in the candidate regions, we identified 12 unique genes with non-synonymous SNPs, which had large allele frequency differences between landrace and breeding populations (≥20% allele frequency difference; Supplementary Table 17). Gene annotations, however, were available for only ten genes. One of these genes, Ca_13939 has associated function with disease resistance (Supplementary Table 18). In addition, we also identified 169 non-synonymous deleterious mutations (SIFT (sorting intolerant from tolerated) score <0.05) in 88 genes using SIFT analysis

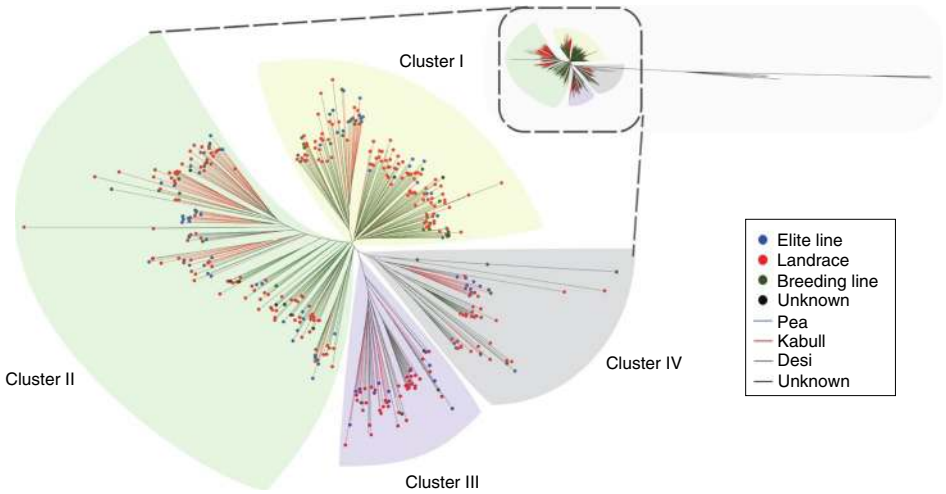

Fig. 3 | Population diversity in 429 chickpea genotypes. Phylogenetic tree constructed using SNPs identified. Wild accessions completely separated from cultivated chickpea. Among cultivated four clusters identified with no clear pattern on the basis of biological status and seed type. Clusters I, II and IV are interspersed with landraces, breeding lines and elite cultivars as well as from different geographies.

(Supplementary Table 19). These genes are attractive candidates for further investigation for their contribution to phenotypic changes during post-domestication improvement.

Genomic regions that underwent domestication and breeding selection possess lower diversity in the same regions in wild species (compared to landraces) and landraces (compared to breeding lines), respectively. These genomic regions were investigated to identify gene gain and loss events among different groups on the basis of seed type (desi and kabuli) and biological status (landraces, breeding lines and elite cultivars) (Supplementary Table 20). Kabuli lines had fewer lost genes compared to desi (Supplementary Fig. 14). Loss of genes was also greater in breeding lines. In total, 350 and 144 genes were lost among different chickpea groups on the basis of biological status and seed type, respectively (Supplementary Tables 21 and 22). Of the 85 gene ontology terms in the biological process category, 11 were related to defense responses to biotic stresses, and 5 were related to abiotic stress response, seed development, germination and flower development suggestive of domestication-syndrome traits. We observed >100 genes lost in the majority of wild genotypes (Supplementary Fig. 14).

Center of origin, migration route and diversity. Population difference (*F* index, F_{ST}) and diversity indices (π and ω) were estimated across different geographical groups (Fertile Crescent, South Asia, Central Asia, East Africa, Mediterranean and Americas). Pair-wise genome-wide F_{ST} values for populations from the Fertile Crescent and Mediterranean were the lowest (0.0156), indicating these populations are very close to each other (Supplementary Table 23). These two regions (in south-west Asia, that is, the Fertile Crescent and Mediterranean) have also been designated the primary centers of origin of chickpea by Vavilov[31]. Detailed analysis indicated the lowest F_{ST} values for individuals from the Mediterranean with Central Asia (0.0198) followed by populations of East Africa (0.0231), Americas (0.0237) and South Asia (0.0306) (Supplementary Table 23). Although F_{ST} alone may not entirely reflect migration history[32], the higher F_{ST} values between both the Mediterranean and Fertile Crescent and South Asia are consistent with archeological evidence. Similarly, the high F_{ST} value between the Fertile Crescent and Ethiopia is also in the line with archeological evidence of an ancient introduction of chickpea to Ethiopia. That said, the relatively low F_{ST} value between the Mediterranean and Ethiopia could indicate more recent introduction of chickpeas to Ethiopia by international partners. The relatively low F_{ST} value among the Fertile Crescent and Central Asia could indicate ongoing gene flow or the introduction of kabuli genotypes into the Fertile Crescent, consistent with linguistic evidence tying these market types to the city of Kabul in modern Afghanistan. Our results could also indicate possible migration of chickpea to the New World (Americas) directly from Central Asia or East Africa, rather than solely from Iberia or the Mediterranean during the colonial period. On the basis of these results, we speculate that New World chickpea growers may have been able to source seeds more widely from Spain alone. Furthermore, these results confirm Vavilov's hypothesis[31] of Ethiopia (East Africa) being the secondary center of diversity.

To understand the diversity patterns in different geographic regions, two diversity parameters (π, ω) were estimated in the above mentioned six geographic regions (Supplementary Table 24). The highest genetic diversity was observed in the population of South Asia ($\pi = 1.05$, $\omega = 1.16$) and the lowest in the population of Americas ($\pi = 0.85$, $\omega = 0.70$). These results were expected since South Asia is the major (>90%) chickpea-growing area at present, and the Americas were the last region where the chickpea was introduced[33]. Lower values of π and ω in the Fertile Crescent and Mediterranean region, however, were unexpected, which may be explained by the relatively low number of samples and a lack of statistical framework in this study. An alternative explanation was proposed by Abbo et al.[4], which is that chickpea cultivation was abandoned in the Fertile Crescent for an almost 2,000-year period due to threats from Ascochyta blight, and then chickpea was re-introduced to the region from Central or South Asia as a spring-sown rather than a winter-sown crop.

Genome-wide association study (GWAS) for agronomic traits. We used 3.65 million SNPs and phenotyping data for 20 drought and heat tolerance related traits collected over one to six locations (Patancheru, Kanpur and Bangalore in India, Nairobi and Nakuru in Kenya and Debre Zeit in Ethiopia) for one to six seasons on 272 genotypes for identifying markers associated with key agronomic traits. We used different statistical models including generalized linear model (GLM)/compressed mixed linear model (CMLM), fixed and random model circulating probability unification (FarmCPU) and efficient mixed model association expedited (EMMAX) and identified 262, 624 and 938 MTAs, respectively (Supplementary Table 25). The MTAs obtained using EMMAX look spurious as 683 MTAs were associated with yield per plant out of 938 total MTAs identified. Although the *P* values of MTAs obtained from

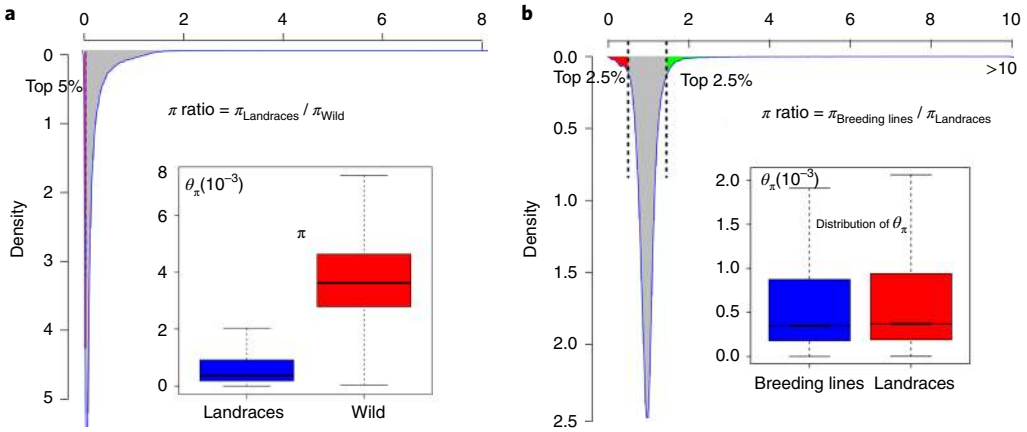

Fig. 4 | Selection sweeps and reduction of diversity a, Nucleotide diversity (θ_π per kb) indicates that wild chickpea genotypes possess high allelic diversity compared to landraces despite a very small number of wild lines being used in this study. **b**, Candidate selection sweep regions are with a 2.5% significance level of reduction of diversity (ROD) are shown in green and 0.25% in red. A significant reduction in diversity was observed from wild genotypes to landraces and breeding lines.

Table 2 | Diversity levels among different groups on the basis of biological status and seed type

	Biological status[a]					Seed type	
	Wild (7)	Landraces (268)	Breeding lines (44)	Elite cultivars (100)	Cultivated (412)	Desi (272)	Kabuli (128)
θ_π (kb)	3.80	0.86	0.84	0.74	0.84	0.81	0.81
θ_ω (kb)	2.79	0.87	0.86	0.74	0.90	0.88	0.85

[a]Number of genotypes in each group are given in parenthesis.

FarmCPU were high, and both false positives and false negatives were reduced by testing multiple markers simultaneously by considering selected associated markers as covariates, the analysis is constrained due to removal of significant SNPs in linkage disequilibrium, which confirms the existence of a truly associated locus. Furthermore, the majority of SNPs associated with traits were in genes with unknown functions (Supplementary Tables 26 and 27). Hence, we have focused on results from CMLM method for in-depth analysis to understand the molecular mechanism of drought and heat tolerance. In brief, the CMLM-based GWAS identified 262 MTAs with 203 unique SNPs. Further, a total of 173 MTAs fall in the 'robust' category and of these MTAs, 51 MTAs were 'consistent' and 5 MTAs were 'stable' (see Methods and Supplementary Table 28). On the basis of SNP annotation, 48 SNPs were present in 47 unique genes with known function (Supplementary Table 29). We have discussed association of some traits with genes of known function in the Supplementary Note.

Discussion

Crop domestication, post-domestication diversification and recent breeding efforts have selected traits to meet human needs and resulted in narrow genetic diversity in cultivated gene pools of most annual crops including chickpea. Efforts to increase diversity by using exotic germplasm, including wild *Cicer* species, have been able to enhance the genetic diversity to some extent[34]. However, there is still a large yield gap that needs to be filled by using wild species or under-used landraces for crossing and bring superior alleles into advanced germplasm. This study reports the sequencing and analysis of a large germplasm collection, not only for chickpea but for legume crops.

Our resequencing of the 300 genotypes reference set, which encompasses 78% of the variation in the larger 3,000 genotype composite collection, provided 4.97 million SNPs and a comprehensive hapmap for chickpea. Higher variation was observed in desi genotypes, as compared to kabuli genotypes, while CNV abundance was lower. Similarly, a higher level of variation was observed in landraces, as compared to elite cultivars and breeding lines. Wild species genotypes had more unique SNPs. This study, therefore, provides a resource (both alleles/haplotypes as well as lines) for re-structuring breeding programs; for example, development of new and knowledge-based crosses to enhance diversity in the elite gene pool. Population structure identified allelic admixture in some genotypes that can be attributed to breeding history where desi and kabuli genotypes have been inter-crossed frequently. All four clusters in our cluster analysis were interspersed with landraces, breeding lines and elite cultivars. This is a reflection of breeding for different environments or market types. However, no clear clustering pattern was observed on the basis of the geographic origin of the genotypes, which indicates extensive movement of the germplasm across regions and also extensive use of diverse lines in breeding programs. Cluster analysis has also identified useful variation in this germplasm for drought and heat stress tolerance, two important traits for developing climate-smart ready varieties. Furthermore, as expected, breeding lines were found to contain large linkage disequilibrium blocks, which underline a selection of bigger genomic regions for positive alleles and their subsequent fixation in the breeding programs. These observations highlight the importance of breaking these large linkage blocks by using new genetic crosses such as multi-parent advanced generation inter-crosses, so that hitchhiking by mildly deleterious alleles can be removed from these linkage disequilibrium blocks.

A four-fold reduction in diversity was observed from wild genotypes to landraces, highlighting the loss of about 80% of genetic diversity. This study has focused on the identification of allele(s) and genomic region(s) impacted during domestication and post-domestication diversification and identified 122 candidate

domestication regions and 204 genes that underwent selection, which can be further explored to understand physiological processes important to bring changes in phenotype of interest. Reproductive success in chickpea depends on time to flowering, as it is cultivated following rainy seasons in arid and semi-arid environments that are prone to heat and moisture stress, particularly during flowering and pod set. Furthermore, chickpea is grown both as a spring-sown crop in Mediterranean and temperate regions such as the Fertile Crescent and Central Asia and as a winter-sown crop in sub-tropical monsoonal regions such as South Asia and East Africa. Due to these significant post-domestication shifts in climate, genes involved in vernalization and flowering are likely candidates underlying these differences, as has been shown in wheat[35] as well as chickpea[36]. Notably, Ca_13671, a gene that encodes a Vernalization 1 (*VRN1*) ortholog, and Ca_13939, which encodes an ortholog of a gene for disease resistance, are among key genes with signatures of selection.

In terms of center of origin of chickpea, Vavilov suggested southwest Asia (Fertile Crescent) and the Mediterranean as possible primary centers of origin with South Asia and Ethiopia as secondary centers. A number of migration routes have been proposed in the past for introduction of chickpea in different geographical regions. Our study, based on F_{ST} analyses, suggests a migration route from the Mediterranean/Fertile Crescent to South Asia (India) and then perhaps East Africa and Central Asia in parallel. Regarding migration of chickpea to India from East Africa versus Central Asia, our data are in accordance with the linguistic indications that the large-seeded, cream-colored chickpeas reached India only two centuries ago, apparently through Afghanistan, as its Hindi name is kabuli chana (chickpea) in allusion to the Afghanistan capital Kabul[37]. These same genotypes may have been introduced back to the primary center of origin as well. Our study speculates about possible introduction of chickpea to the New World (Americas) directly from Central Asia or East Africa rather than the Mediterranean basin alone.

With rising temperatures and increasing climatic fluctuations due to global warming, identifying adaptive genetic variation present in the cultivated gene pool and understanding its mode of action will be an essential breeding action to address oncoming climate change[38]. High-density sequencing and phenotyping of the chickpea reference set, a diverse germplasm covering different agro-climatic zones and germplasm types (landraces and elite varieties) enables the identification of superior line(s) and gene(s). For instance, ICC 14778, a desi landrace reported to have stable heat as well as drought tolerance can be used to breed for tolerance to both the stresses. GWAS exploits the recombination events to identify the candidate gene(s)/marker(s) associated with trait of interest to dissect the molecular mechanism for stress response[39]. In self-fertilized species such as chickpea, linkage disequilibrium generally decays with lower rate of around 200 kb, suggesting that it will be difficult to perform GWAS with single-gene resolution. The current study deployed ~3.65 million SNPs, about 2,000 times more markers than the previous study[6], to undertake the high-resolution GWAS. This study reports 262 MTAs and candidate genes (such as *TIC*, *REF6*, aspartic protease, *cc-NBS-LRR*, *RGA3*) for both drought and heat tolerance. Among different drought tolerance mechanisms, drought escape mechanism has been considered the most important[40]; we identified significant MTAs or candidate genes and their haplotypes for early phenology such as *REF6*, which could enable selection of lines that escape drought as well as heat stress, such that the markers identified in this study can be simultaneously used for enhancing tolerance to both heat and drought. Further, the MTAs identified in the present study can be used to develop 'steep, cheap and deep' root ideotypes that have been recently proposed as adaptive for drought-prone soil conditions[41].

In summary, this study has established a foundation for large-scale characterization of germplasm, population genetics and crop breeding. The comprehensive chickpea hapmap with 4.97 million SNPs developed in this study is a valuable resource for undertaking high-resolution GWAS and better imputation of low-coverage sequencing data for facilitating large-scale germplasm characterization at lower cost. We have achieved a better understanding regarding population structure of germplasm, domestication and post-domestication divergence as well as the center of origin and migration routes of chickpea to different geographical regions. This study should enable breeders to enhance the use of diverse germplasm and candidate genes in developing improved (climate-change-ready) cultivars that hopefully will contribute significantly to the increased productivity and sustainability of agricultural development in developing countries.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-019-0401-3.

Received: 7 February 2018; Accepted: 21 March 2019;

References

1. Mba, C., Guimaraes, E. P. & Ghosh, K. Re-orienting crop improvement for the changing climatic conditions of the 21st century. *Agri. & Food Sec.* **1**, 7 (2012).
2. Ritchie, H. & Roser, M. Micronutrient deficiency. *Our World in Data* https://ourworldindata.org/micronutrient-deficiency (2017).
3. Atlin, G. N., Cairns, J. E. & Das, B. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. *Glob. Food Sec.* **12**, 31–37 (2017).
4. Abbo, S., Berger, J. & Turner, N. C. Viewpoint: evolution of cultivated chickpea: four bottlenecks limit diversity and constrain adaptation. *Funct. Plant Biol.* **30**, 1081–1087 (2003).
5. Varshney, R. K. et al. Genetic dissection of drought tolerance in chickpea (*Cicer arietinum* L.). *Theor. Appl. Genet.* **127**, 445–462 (2014).
6. Thudi, M. et al. Understanding the genetic architecture of drought and heat tolerance in chickpea through genome-wide and candidate gene-based association mapping. *PLoS ONE* **9**, e96758 (2014).
7. Upadhyaya, H. D. et al. Genomic tools and germplasm diversity for chickpea improvement. *Plant Genet. Resour.* **9**, 45–58 (2011).
8. 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).
9. 3K RGP. The 3,000 rice genomes project. *GigaScience* **3**, 7 (2014).
10. Varshney, R. K. et al. Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* **49**, 1082–1088 (2017).
11. Varshney, R. K. et al. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* **35**, 969–976 (2017).
12. Fang, L. et al. Genomic analyses in cotton identify signatures of selection and loci associated with fiber quality and yield traits. *Nat. Genet.* **49**, 1089–1098 (2017).
13. Romero Navarro, J. A. et al. A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat. Genet.* **49**, 476–480 (2017).
14. Xu, X. et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30**, 105–111 (2012).
15. 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).
16. Varshney, R. K., Nayak, S. N., May, G. D. & Jackson, S. A. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530 (2009).
17. Upadhyaya, H. D. et al. Genetic structure, diversity, and allelic richness in composite collection and reference set in chickpea (*Cicer arietinum L.*). *BMC Plant Biol.* **8**, 106 (2008).
18. Thudi, M. et al. Recent breeding programs enhanced genetic diversity in both desi and kabuli varieties of chickpea (*Cicer arietinum* L.). *Sci. Rep.* **6**, 38636 (2016).

19. Varshney, R. K. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* **31**, 240–246 (2013).
20. Mace, E. S. et al. Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat. Commun.* **4**, 2320 (2013).
21. Pritchard, J. K., Stephens, M. & Donnelly, P. J. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
22. Penmetsa, R. V. et al. Multiple post-domestication origins of kabuli chickpea through allelic variation in a diversification-associated transcription factor. *New Phytol.* **211**, 1440–1451 (2016).
23. Roorkiwal, M. et al. Exploring germplasm diversity to understand the domestication process in *Cicer* spp. using SNP and DArT markers. *PLoS ONE* **9**, e102016 (2014).
24. Barrett, J. C. et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
25. Gore, M. A. et al. A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
26. Branca, A. et al. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. *Proc. Natl Acad. Sci. USA* **108**, E864–E870 (2011).
27. von Wettberg, E. J. et al. Ecology and community genomics of an important crop wild relative as a prelude to agricultural innovation. *Nat. Commun.* **9**, 649 (2018).
28. Eldon, B., Birkner, M., Blath, J. & Freund, F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics* **199**, 841–856 (2015).
29. Ferretti, L., Ledda, A., Wiehe, T., Achaz, G. & Ramos-Onsins, S. E. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. *Genetics* **207**, 229–240 (2017).
30. Hudson, R. R. et al. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).
31. Vavilov, N. I. Centres of origin of cultivated plants. *Bull. Appl. Bot. Genet. Plant Breed.* **16**, 1–248 (1926).
32. Whitlock, M. C. & McCauley, D. E. Indirect measures of gene flow and migration: FST does not=1/(4Nm+1). *Heredity* **82**, 117–125 (1999).
33. Redden, R. J. & Berger, J. D. History and origin of chickpea. in *Chickpea Breeding and Management* (eds Yadav, S. S. et al.) 1–13 (C.A.B. International, 2007).
34. McCouch, S. et al. Agriculture: feeding the future. *Nature* **499**, 23–24 (2014).
35. Kamran, A. et al. The effect of VRN1 genes on important agronomic traits in high-yielding Canadian soft white spring wheat. *Plant Breed.* **133**, 321–326 (2014).
36. Samineni, S. et al. Vernalization response in chickpea is controlled by a major QTL. *Euphytica* **207**, 453–461 (2016).
37. van der Maesen, L. J. G. in *The Chickpea* (eds Saxena, M. C. & Singh, R. B.) 11–34 (C.A.B. International, 1987).
38. Hoffmann, A. A. & Sgrò, C. M. Climate change and evolutionary adaptation. *Nature* **470**, 479–485 (2011).
39. Rafalski, J. A. Association genetics in crop improvement. *Curr. Opin. Plant Biol.* **13**, 174–180 (2010).
40. Kashiwagi, J. et al. Traits of relevance to improve yield under terminal drought stress in chickpea (*C. arietinum* L.). *Field Crops Res.* **145**, 88–95 (2013).
41. Lynch, J. P. Steep, cheap and deep: an ideotype to optimize water and N acquisition by maize root systems. *Ann. Bot.* **112**, 347–357 (2013).

Acknowledgements

R.K.V. acknowledges the funding support from CGIAR Generation Challenge Programme, Department of Science and Technology Government of India under the Australia-India Strategic Research Fund, Ministry of Agriculture and Farmers Welfare, Government of India and Bill & Melinda Gates Foundation, USA. Shenzhen Municipal Government of China (grant no. JCYJ20150831201643396 and no. JCYJ20170817145512476 under the Basic Research Program) and the Guangdong Provincial Key Laboratory of Genome Read and Write (grant no. 2017B030301011) are acknowledged to provide support to X.X. and X.L. This work has been undertaken as part of the CGIAR Research Program on Grain Legumes and Dryland Cereals. ICRISAT is a member of the CGIAR Consortium.

Author contributions

R.K.V. conceived and designed the experiments. R.K.V., X.X. and X.L. coordinated sequencing and genome analysis. W.H., W.Y., J.J., H.D.U., N.P.S., S.K.C., G.V.P.R.N., L.K., A.F., K.K.B.P., P.M.G. and S.M.S. performed the experiments. W.H., P.B., A.R., D.D., V.G., A.W.K., H.D.U., J.C. and Y.V. performed statistical analysis. R.K.V., M.T., M.R., A.C., P.C., C.B., S.T., J.W., S.-H.L., D.E., K.K.B.P., R.V.P., J.C., H.T.N., K.H.M.S., T.D.C., T.S., E.v.W., Y.V., X.X. and X.L. analyzed the data. R.K.V., H.D.U., A.C., P.M.G., N.P.S., S.K.C., G.P.D., D.X., J.W., X.X. and X.L. contributed to the reagents, materials and analysis tools. R.K.V., M.T., M.R., W.H. and X.L. wrote the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0401-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.K.V., X.X. or X.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Plant material. A set of 300 genotypes from the chickpea reference set were resequenced on Illumina HiSeq 2000 using the WGRS approach (Supplementary Table 1). The reference set consists of 267 landraces, 13 advanced lines and cultivars, 7 wild *Cicer* accessions and 13 accessions with unknown biological status. The reference set captures 78% of allelic diversity present in the global composite collection of chickpea germplasm, making it a more tractable germplasm set than the larger core collection[17]. In addition, WGRS data of 100 elite chickpea cultivars released between 1948 and 2012 (ref. [18]) and 29 lines from the chickpea genome paper[19] were analyzed along with the data generated in the present study. Stringent filtering steps as described earlier[18] were adopted for obtaining clean data.

Variant detection. Genome-wide variations such as SNPs, indels, CNVs and PAVs were identified in sequence data for 429 lines. For calling SNPs, the clean reads were mapped on to the reference genome of chickpea genotype CDC Frontier using SOAP2 (ref. [42]). In brief, the major parameters of the mapping step were '-m 300 -x 600 -s 35 -l 32 -v 5 -p 4'. We then used SOAPsnp3 to calculate the likelihood of all possible genotypes for each sample (major parameters: '-i sample.soap –d ref.fa -u –M ref.mat'). Then the likelihoods of genotypes of all the samples were combined to calculate the maximum likelihood estimation of the allele frequency in the population[43]. To filter out low-quality variants, the loci with sequencing depth higher than 10,000, lower than 400, mapping times higher than 1.5 or a quality score lower than 20 were filtered out. The loci with estimated allele frequency not equal to 0 or 1 were determined as SNPs. After obtaining the SNPs, we also determined the genotype of each individual at the SNP locus by assigning the most likely genotype from SOAPsnp3 result of each sample. For the later analysis, we further filtered the SNPs with half or more individuals not covered (with missing genotype), because those SNPs might not be informative for analyses such as population analysis. Small insertions and deletions (1–10 bp), referred to as indels, were identified using SOAPindel[44]. In brief, we first mapped the clean reads of each sample against the reference with gaps allowed (major parameters settings to: '-g 10 -m 300 -x 600 -s 35 -l 32 -v 5 -p 4'). Then we applied SOAPindel to identify indels in each sample from the gap allowed mapping results (major parameters of SOAPindel settings to: '-p 0.01 -c 3 -h0.5 -k 5 -m 1'). Finally, we combined indels of each sample according to the locus and length of the indels, to obtain their total number for different groups.

We identified CNVs as described previously[18]. Assuming a Poisson's distribution of sequencing depth, the genome regions were divided into initial windows where the sequencing depth did not significantly differ. The mean depth of each window was then calculated and compared to nearby windows. Initial windows were thus further merged if there were no significant depth differences in the nearby initial windows. Merging of the window process was repeated once more, thus the edges and the copy number of each window were decided. As we detected lost genes later, we only retained CNVs that had more copies than the reference genome (copy number >1).

We filtered the identified indels to obtain the PAVs. For a deletion region, if the average sequencing depth was less than 10% of the genome-wide average sequencing depth, we determined this sample to have absent variation in this region. For an insertion region, if the average sequencing depth was more than 50% of the genome-wide average sequencing depth, we determined this sample to have present variation in this region. For the sub-populations or groups, if more than three samples consistently had PAV at one region, this sub-population or group was determined to have this PAV.

Population structure and genetic relationships. We conducted population structure analysis using STRUCTURE software[20]. We ran 10,000 iterations, and the number of clusters (K) was set to 2–7. We used the final SNP dataset with 4.9 million SNPs (population SNPs with missing genotypes less than 50%) to conduct the principal component analysis (PCA). We performed the PCA according to previously described procedure[45]. The eigenvector decomposition of the transformed genotype data was performed using the R function eigen, and the significance of the eigenvectors was determined with a Tracey–Widom test, implemented in the program twstats, provided by the EIGENSOFT software[45,46]. We used the final SNP dataset to construct the phylogenetic tree. All SNPs were used to calculate the genetic distances between different accessions following the procedure previously described[14]. Then, the neighbor-joining method in the software PHYLIP[47] was used to construct the phylogenetic tree according to the distance matrix. Finally, MEGA4 (ref. [48]) was used to display the phylogenetic tree. The ANGSD[49] program was used to estimate Tajima within different groups. Tajima was calculated over a non-overlapping window of 100 kb using folded site frequency spectrum (SFS).

Linkage disequilibrium analysis. Before calculating linkage disequilibrium, marker data was filtered for minor allele frequency (MAF) and missing percentage. All markers below 0.05 MAF and having more than 80% missing data were discarded from linkage disequilibrium analyses. Linkage disequilibrium decay on each pseudomolecule and across all pseudomolecules was calculated using 3.6 million SNPs distributed on Ca1-Ca8 employing Haploview[24]. To measure linkage disequilibrium levels in different populations (such as desi, kabuli, wild, cultivated, breeding lines, elite cultivars and so on), we calculated the correlation coefficient (r^2) of alleles by setting major parameters to '-maxdistance 1000 -dprime -minGeno 0.6 -minMAF 0.1 -hwcutoff 0.000'.

Domestication, selection and gene loss analysis. ROD was determined as described earlier[14], while π, ω and Tajima's D were computed as described earlier[19]. Lost genes in genomes of different populations were identified by merging all the deletions to continuous regions and then extracting the genes in those combined regions. The genes were considered lost if the coverage was less than 10% and the sequencing depth was lower than 10% of the average depth.

Marker-trait association analyses. Association mapping was done using GLM/CMLM, EMMAX and FarmCPU statistical methods[50–53]. TASSEL 5.0 software was used for GLM/CMLM and EMMAX statistical analysis. SNPs with allele frequencies <5% were filtered out for association analysis. For the FarmCPU analysis, the first four axes of the PCA estimated with GAPIT R package were used as covariates. For GLM/CMLM analysis, model selection was based on quantile–quantile plots. 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. To correct for population structure, PCA was performed. The final number of principal components that appropriately explain population structure was determined from scree plots[46]. The critical P values for assessing the significance of markers were calculated on the basis of a false discovery rate separately for each trait. A false discovery rate cut-off of 0.05 was used for determining significance. In the case of EMMAX and FarmCPU analyses, a P value threshold was estimated for SNP significance by randomly permutated genotype and phenotype, ten times. MTAs with more than 10% phenotypic variation explained were considered as 'robust', MTAs identified for more than one location were considered as 'stable' and MTAs identified across more than 1 year/season were defined as 'consistent'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study have been deposited in the NCBI under accession code SRA: SRP096939; BioProject: PRJNA362278, and these data are also available in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession code CNP0000370.

References

42. Li, R. et al. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
43. Yi, X. et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
44. Li, S. et al. SOAPindel: Efficient identification of Indels from short paired reads. *Genome Res.* **23**, 195–200 (2013).
45. Patterson, N., Price, A. L. & Reich, D. Population structure and eigen analysis. *PLoS Genet.* **2**, 2074–2093 (2006).
46. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
47. Felsenstein, J. Phylip: phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
48. Tamura, K. et al. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
49. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
50. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
51. Lipka, A. E. et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
52. Tang, Y. et al. GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. *Plant Genome* **9**, 2 (2016).
53. Liu, X., Huang, M., Fan, B., Buckler, E. S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**, e1005767 (2016).

nature research

Corresponding author(s): Rajeev K. Varshney, Xun Xu, Liu Xin

Last updated by author(s): Feb 25, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software used
Data analysis	R-packages, SOAP2, SOAPsnp3, SOAPIndel, FRAPPE, EIGENSOFT, ANGSD, Haploview, TASSEL 4.0, STRUCTURE2.3.4, PHYLIP, MEGA4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The resequencing data of chickpea Reference set is made available at NCBI. Accession code: SRA: SRP096939; BioProject: PRJNA362278

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Chickpea reference set with 300 genetically most diverse accessions from genebank representing 78% of diversity of total collection were used for analysis. These 300 lines are an ideal set of germplasm for allele mining, association genetics, and in applied breeding for the development of broad-based elite breeding lines/cultivars with superior yield.
Data exclusions	Genotyping data was filtered using various well established criteria including % of missing, minor allele frequency and others. These exclusion have been defined for each analysis in the Methods section. This is pre-established criteria commonly used by all the studies.
Replication	Phenotyping experiments were conducted in replication and confirmed the reproducibility of data. Genotyping data analysis was randomly replicated and confirm that results are reproducible.
Randomization	Distribution of samples in different experimental group is defined in the Methods section. In brief samples were defined based on seed type (desi, kabuli and intermediate) and biological status (breeding lines, landraces, cultivars and wilds) of the samples. Role of co-variate is not relevant here as we made grouping based on the genotype status using seed type and biological status.
Blinding	There is no blind data collection. Our study include phenotyping data on the lines where we record the data on agronomic traits on a particular lines for use in GWAS and other studies, so having blind data is not possible.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging