Plant Science 285 (2019) 122–131

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Multi-trait analysis of domestication genes in *Cicer arietinum* – *Cicer reticulatum* hybrids with a multidimensional approach: Modeling wide crosses for crop improvement

Min-Gyoung Shin[a], Sergey V. Bulyntsev[c], Peter L. Chang[b,d], Lijalem Balcha Korbu[d,e], Noelia Carrasquila-Garcia[d], Margarita A. Vishnyakova[c], Maria G. Samsonova[f], Douglas R. Cook[d], Sergey V. Nuzhdin[b,f,*]

[a] University of Southern California, Program Quantitative and Computational Biology, Dornsife College of Letters Arts & Science, Los Angeles, CA 90089, USA
[b] University of Southern California, Program Molecular & Computational Biology, Dornsife College of Letters Arts & Science, Los Angeles, CA 90089, USA
[c] Federal Research Center The NI Vavilov All Russian Institute of Plant Genetic Resources, St Petersburg, Russia
[d] University of California Davis, Department of Plant Pathology, Davis, CA 95616, USA
[e] Ethiopian Institute of Agricultural Research, Debre Zeit, Ethiopia
[f] Peter the Great St Petersburg Polytechnich University, Department of Applied Mathematics, St Petersburg, Russia

ARTICLE INFO

Keywords:
Chickpea wild – domestic introgression
Nested association mapping
QTL mapping
GWAS
Bayesian network
Random forest
Machine learning

ABSTRACT

Domestication and subsequent breeding have eroded genetic diversity in the modern chickpea crop by ~100-fold. Corresponding reductions to trait variation create the need, and an opportunity, to identify and harness the genetic capacity of wild species for crop improvement. Here we analyze trait segregation in a series of wild x cultivated hybrid populations to delineate the genetic underpinnings of domestication traits. Two species of wild chickpea, *C. reticulatum* and *C. echinospermum*, were crossed with the elite, early flowering *C. arietinum* cultivar ICCV96029. KASP genotyping of F2 parents with an FT-linked molecular marker enabled selection of 284 F3 families with reduced phenological variation: 255 F3 families of *C. arietinum x* reticulatum (AR) derived from 17 diverse wild parents and 29 F3 families of *C. arietinum x* echinospermum (AE) from 3 wild parents. The combined 284 lineages were genotyped using a genotyping-by-sequencing strategy and phenotyped for agronomic traits. 50 QTLs in 11 traits were detected from AR and 35 QTLs in 10 traits from the combined data. Using hierarchical clustering to assign traits to six correlated groups and mixed model based multi-trait mapping, four pleiotropic loci were identified. Bayesian analysis further identified four inter-trait relationships controlling the duration of vegetative growth and seed maturation, for which the underlying pleiotropic genes were mapped. A random forest approach was used to explore the most extreme trait differences between AR and AE progenies, identifying traits most characteristic of wild species origin. Knowledge of the genomic basis of traits that segregate in wild-cultivated hybrid populations will facilitate chickpea improvement by linking genetic and phenotypic variation in a quantitative genetic framework.

1. Introduction

Modern agriculture must meet the nutritional demands of a growing human population using increasingly limited land and water resources. Ideally, cultivation will intensify while energy-intensive inputs, such as water and nitrogen, will decrease. Grain legumes, including chickpea, are the primary source of nutritional nitrogen for approximately 30% of the world's human population [1]. Legumes were not, however, beneficiaries of the Green Revolution. Grain legumes in particular were underinvested and often relegated to marginal lands where abiotic stress, shortened growing seasons, and poor soils conspire to limit yield [2]. Nevertheless, grain legumes remain a vital asset in the effort to

Abbreviations: AR, *C. arietinum* × *reticulatum*; AE, *C. arietinum* × *echinospermum*

* Corresponding author at: University of Southern California, Program Molecular & Computational Biology, Dornsife College of Letters Arts & Science, Los Angeles, CA 90089 USA.

E-mail addresses: mingyous@usc.edu (M.-G. Shin), s_bulyntsev@mail.ru (S.V. Bulyntsev), peterc@usc.edu (P.L. Chang), lijupeace@gmail.com (L.B. Korbu), noecarras@ucdavis.edu (N. Carrasquila-Garcia), m.vishnyakova.vir@gmail.com (M.A. Vishnyakova), m.samsonova@spbstu.ru (M.G. Samsonova), drcook@ucdavis.edu (D.R. Cook), snuzhdin@usc.edu (S.V. Nuzhdin).

https://doi.org/10.1016/j.plantsci.2019.04.018
Received 18 December 2018; Received in revised form 17 April 2019; Accepted 20 April 2019
Available online 25 April 2019
0168-9452/ © 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

meet the demand for food in impoverished, often food-insecure countries. Paradoxically, in the case of chickpea, domestication and modern breeding have constrained the capacity for crop improvement by eroding ancestral diversity [3–5].

Germplasm collections with abundant phenotypic and genetic variation are essential to adapting crops to current and future agricultural challenges [5–8]. Crop wild progenitors are likely to be of particular value because they possess unexplored variation ('genomic gems') that could accelerate breeding gains beyond what is possible from the domesticated gene pool alone. Despite this fact, the systematic, large-scale use of wild germplasm has been limited to only a few crop species [5].

During domestication and breeding, humans selected cultivated plants for superior agronomic performance. Within the cultivated gene pool, the assembled agronomic traits are themselves augmented by the complexity of genetic context, including pleiotropy and linkage, creating what are in essence distinct sub-species. By contrast, superior traits present in wild germplasm typically exist in the context of inferior agronomic performance, despite their presumed ecological utility. Such features complicate trait selection from wide cross populations and have the practical impact of focusing breeding efforts on few loci of large effect size. Mining of large-effect loci is an obvious target and has been pursued to great impact [9]. Quantitative genetics tools can access minor effect loci with large cumulative effects, which for example is an implicit goal of genomic selection [9,10], but to our knowledge, such tools are not in use with wild germplasm. More nuanced but also of great importance, wild populations offer unique opportunities to reduce the burden of minor effect deleterious alleles that are often elevated in domesticated species [5], by identification and introgression of wild genome backgrounds that lack undesirable pleiotropies on crop phenotypes. To do so, one must estimate the effects of introgressions not only on the primary yield-related traits, but – ideally – on the full range of agricultural phenotypes. Thus, wild x cultivated hybrids potentially have multiple uses in crop improvement: identifying desirable traits of wild origin for introgression, reducing the load of deleterious alleles, and re-cycling useful cultivated traits previously selected during domestication.

While wild x cultivated populations provide the means to bridge germplasm collections, maximizing their impact requires capturing the full wealth of wild adaptations, which in turn depends on comprehensive knowledge of wild species' diversity. Ideally one would leverage population genomic and origin environment data to select wild parents of diverse origins. Such populations would enable the identification of fixed differences between wild and cultivated species (including, but not limited to, domestication traits) as well as those that continue to segregate among wild populations.

In chickpea, wild germplasm has seen limited use for crop improvement. In particular, systematic use of wild *Cicer* germplasm is hampered by the lack of adequately diverse inter-specific mapping populations and also by the scarcity of validated genetic markers and few known genes/QTLs for important agronomic traits [6]. Our recent comprehensive sampling and genomic analyses of chickpea's wild progenitor populations create an opportunity to reverse this situation [5]. The wild progenitors of chickpea, *Cicer reticulatum* and *C. echinospermum*, are confined to southeastern Turkey, near the origin of the Tigris and Euphrates rivers in a region historically known as Mesopotamia. We used ecological principles to guide collection across the full range of habitats in which wild chickpea occurs [5]. Sequencing of ~1000 wild accessions, either as field-collected DNA or seed, was sufficient to describe the nature of wild species' diversity. Twenty wild accessions representing the genetic and environmental breadth of the wild collection were crossed into cultivated accessions as the first step towards trait introgression.

We report on a subset of 284 F3 phenology-normalized lineages obtained from crosses with a single early flowering cultivated lineage (ICCV 96,029). Twenty-three core phenotypes were scored, including the broad list of traits and protocols first instituted almost one century ago by Nikolay Vavilov and colleagues at the All Russian NI Vavilov Institute for Plant Genetic Resources, Saint Petersburg [4]. Together, these traits survey aspects of vegetative growth, plant architecture, seed and pod traits, disease susceptibility, nodulation, and flower color.

Because the present data involves numerous phenotypes for which genetic control may involve genetic interactions, we utilized multi-dimensional analyses to disentangle relationships between correlated traits. Specifically, we applied two types of linear mixed models: a single-trait model and a multi-trait model, while a Bayesian network analysis was performed to find influential relationships between phenotypes. A Bayesian network is a directed acyclic graph in which each node contains quantitative probability information [11]. A Bayesian framework analysis has been used previously to understand causal markers that drive disease resistance and metabolic pathways [11–14]. Lastly, random forests, a machine learning technique, was utilized to explore traits that differ most between progeny obtained from the two wild species, *C. reticulatum* and *C. echinospermum*. The results presented here will increase the immediate utility of these wild x cultivate population for genomic breeding, by linking genetic variation to phenotypic variation through quantitative genetic and genomic approaches.

2. Materials and methods

2.1. Germline development

von Wettberg et al. [5] have described the construction of the hybrid germplasm analyzed here. Briefly, all parental accessions were sequenced to at least 30-fold coverage. One full set of 2521 F2 progeny, representing 20 wild parents crossed into the early flowering parent ICCV96029, was genotyped to facilitate trait-marker discovery. To correct for the confounding effect of segregating phenology, F2 plants were genotyped for the FT locus using KASP genotyping, because ICCV96029′s early flowering habit is substantially explained by variants tightly linked to FT [5]. A subset of 284 F2-derived lines homozygous the ICCV96029 early flowering locus was selected. The genotyped F2 plants were selfed to produce F3 progeny and at least 20 seeds per lineage were grown in the field. As the frequency of alleles averaged among these progeny correspond to the allelic state known from F2, phenotyping of F3 is akin to estimating breeding values for genotyped F2 individuals. Note that while additive effects are well-recovered in this approach, the dominance deviations would not be; accordingly we don't attempt to estimate them.

2.2. DNA analysis

Genomic DNA was isolated using the Qiagen DNeasy 96 format Plant Mini Kit (Valencia, CA, USA). DNA was digested with restriction enzymes HindIII and NlaIII and ligated with adapters. The 'barcode' adapter ligates to HindIII allowing sample pooling. The 'common' adapter ligates to NlaIII. Products were selected for size and amplified through 14 rounds of PCR. 100 base paired-end reads were generated on an Illumina HiSeq4000 at the University of California at Davis Genome Center DNA Technologies Core. Illumina data is available online at the National Center for Biotechnology Information under the BioProject umbrella PRJNA507624. Illumina reads were mapped to the *C. arietinum* CDCFrontier reference genome [15] using BWA MEM [16] under default mapping parameters. Polymorphisms were called using the GATK pipeline, which considers indel realignment and base quality score recalibration, and calls variants across all samples simultaneously through the HaplotypeCaller program in GATK. Variants were filtered using standard hard filtering parameters according to GATK Best Practices recommendation. More precisely, GBS data were filtered to only retain SNP calls with Mapping Quality (MQ) > 37 and Quality by Depth (QD) > 24. Both metrics take into consideration the quality of the mapping and genotype calls to ensure that only those with highest confidence were used. The SNPs were also filtered to retain those with

MQRankSum < |2.0|, which ensures that there is no difference in the Mapping Quality scores for alleles at a given locus. This filtering removed nearly 60% of variant sites reported by GATK and only retained those that passed all three criteria. The resulting SNP were filtered using a minor allele frequency (MAF) threshold of > 5% and genotype call-rate > 90%. 14,201 SNPs remained in the *C. arietinum x C. reticulatum* (AR) crosses, while in the combined data of *C. arietinum x C. reticulatum* (AR) and *C. arietinum x C. echinospermum* (AE) 4713 SNPs were retained. Considering the small sample size of AE, AE specific association analysis was not performed.

2.3. Field cultivation and phenotyping

During the growing season of 2017, 284 hybrid lines were phenotyped at the Kuban branch of the VIR (Fed Res Center The NI Vavilov All Russian Institute of Plant Genet Resources). The Kuban station is situated at the step zone of near Kuban flatlands, 80 km away from the Kavkaz foothills. The typical soils in these regions are Kavkaz blackland soils, with a fertile layer of 140–150 cm and slightly alkaline pH. The humus horizons depth is typically between 130–170 cm, with humus content approaching 3.6–4.6%. The climate of the station is characterized by suboptimal rainfall, and high fluctuations of all climatic parameters. Temperatures are typical of moderate-continental sites, with hot summers. Mid-temperature of the coldest month (January) is −2.6C, and of the hottest month (July) 23C. Total yearly rainfall is 565 mm. These climatic conditions are beneficial for the chickpea cultivation. The station has been engaged in experimental cultivation of over 1000 chickpea varieties since 1930.

The sowing of all hybrid lines was carried out on the same day on May 2, 2017 using hand drills according to the scheme of 70 × 5 x 100 cm. Each line was sown in a row 100 cm long, where each row of 20 seeds was sown to a depth of 5 cm. The distance between the rows was 70 cm, and between the seeds in a row was 50 cm. The onset of germinations was on May 10, 2017, while the appearance of complete shoots differed among individuals and was between May 12, 2017–June 20, 2017. During the growing season, 4-fold manual weeding was conducted. Harvesting of chickpea seeds was carried out manually from July 25, 2017 to August 20, 2017, as the pods matured.

During the plant's vegetative period, 23 phenological, morphological, agronomical, and biological descriptors were analyzed. At the time of seed collections, full structural analyses were executed for 5 randomly chosen plants per accession. The following phenotypes were recorded in the field throughout the growing season: 'start of germination – sowing' (days), 'start of flowering - start of germination' (days), 'full flowering - start of germination' (days), 'full flowering - start of flowering' (days), 'start of maturation - start of flowering' (days), 'start of maturation - start of germination' (days), 'full maturation - start of germination' (days), and 'full maturation – start of maturation' (days). Plant architecture was recorded as: growth habit (score), plant height (cm), location of the lowest pod (cm), pod per plant, pod width (mm), pod length (mm), number of seeds per pod, 1000 seed weight (g), weight of seeds per plot (g), pod dehiscence (score), presence of nodules on roots, number of pods per peduncle. Seed and flower color were recorded by visual observation. Growth habit was measured based on five criteria: prostrate, semi-prostrate, spreading, semi-erect or erect. Disease incidence of Ascochyta susceptibility (score) was recorded for at least 5 plants per accession.

2.4. Linkage map

The linkage map was built based on 1950 F2 lines of *C. arietinum x C. reticulatum*. Before the construction of the linkage map, SNP genotypes were imputed and error-corrected using an in-house script. We first estimated recombination events; when multiple SNPs were present in the same recombination bin, one representative SNP was assigned as the marker to represent the entire bin. Recombination break points were determined based on the dissimilarity value of genotypes flanking SNP k which is defined as follow:

$$disimilarity(k) = abs\left(\sum_{i \in Sl} genotype(i) - \sum_{i \in Sr} genotype(i)\right)$$

where Sl is the set of left flanking SNPs of SNP k and Sr is the set of right flanking SNPs. Genotype(i) denotes the genotype of SNPi which was converted into a numeric value −1, 0, 1 according to its heterozygosity. The size of the window flanking each SNP was determined based on the average number of recombination events across all individuals and chromosomes with the given target window size. We aimed to obtain one recombination event per chromosome on average, considering that the utilized samples were from F2 plants.

JoinMap V4.0 [17] was used to build the linkage map. SNPs located in the same chromosome were run separately using the independence LOD option to calculate groups. Based on the location of SNPs in the CDCFrontier genome, the fixed order of markers was assigned. The Maximum likelihood mapping function was used to estimate map distances.

2.5. LD decay and population structure

The genome-wide mean Linkage Disequilibrium (LD) was measured using R-square between SNPs. The average of R-square values for SNP pairs was calculated in intervals of 50Kb. The calculation was carried out using PLINK v1.07 [18].

Bayesian Markov Chain Monte Carlo model (MCMC) implemented in STRUCTURE was used to find population structure of genetic data [19]. Burn-in iteration and Markov chain Monte Carlo (MCMC) replications were set to 10,000 and 25,000, respectively. The number of subpopulations was set from 2 to 21 (K) in five independent runs. The optimal K value was determined based on the log probability of data [LnP(D)]. Principal Component Analysis (PCA) was performed using SNPRelate [20].

2.6. Association analysis

A linear mixed model was used to find markers associated with 23 traits. The linear mixed model we used incorporates the realized relationship matrix (RRM), which is equivalent to including markers as covariates to control confounding factors induced by relatedness of samples [21]. In addition to the RRM, we added family information as covariate in order to remove cross-specific confounding effects. The model can be written as follows:

$$y = X\beta + g + f + e$$

where y is an n× 1 vector of the phenotypes, X is an n× m matrix of the fixed effects, and β is an m× 1 vector of the coefficients of the fixed effect. Additionally, g is the random effect of the mixed model, with $g \sim N(0, \sigma_g^2 K)$ where K is the relationship matrix that reflects the genetic similarity between samples, f is a fixed effect which represents the family affiliation of each individual, and e is an n× n matrix of the residual effects. The model was implemented using Fast-LMM [22].

We also explored marker associations using a multi-trait linear mixed model. To avoid convergence issues due to the similarity of genotypes and covariates, residuals were calculated using a linear model incorporating the family information as fixed covariate and marker associations were investigated using the residuals. The model was implemented using GEMMA [23]. For both models, we used Bonferroni-corrected p-values with a threshold of 0.05 to determine the significance.

2.7. Bayesian network analysis

To investigate genetic interactions among traits, we utilized a

Table 1
Mean and standard deviation of traits in the combined data, AR, and AE.

Trait	mean (all)	sd (all)	mean (AR)	sd (AR)	mean (AE)	sd (AE)	p-value*
start of germination – sowing	11.5	3	11.6	3	10.5	2.3	2.49E-02
start of flowering – start of germination	27.4	3.4	27.4	3.5	27.4	2.4	8.99E-01
full bloom – start of germination	34.9	8.6	35.1	8.8	32.4	6.5	4.45E-02
full bloom – start of flowering	7.4	7.8	7.7	8	5	5.8	2.93E-02
start of maturation – start of flowering	45.1	5.9	44.5	5.8	50.1	4.2	7.91E-08
start of maturation – start of germination	72.8	6.3	72.2	6.3	78.2	3.3	2.96E-11
full maturation – start of germination	82	7.2	81	6.8	90.3	5.2	6.45E-11
full maturation – start of maturation	9.2	5	8.8	5	12.2	4.1	2.40E-04
growth habit	2.9	1.5	2.8	1.5	3.7	1	4.40E-05
plant height	48.1	9.1	47.1	8.5	57.4	8.5	4.36E-07
location of the lowest pod	8.1	3.4	7.7	3	11.9	4.5	2.48E-05
pod per plant	162.9	86.5	164.1	86.3	152	89.1	5.00E-01
pod width	0.9	0.2	0.9	0.2	0.9	0.1	8.16E-04
pod length	2	0.3	2.1	0.2	2	0.3	6.99E-02
number of seeds per pod	1.4	0.5	1.4	0.5	1.4	0.5	8.43E-01
1000 seed weight	167.7	37.7	163.7	36.1	211	26.3	1.53E-07
weight of seeds per plot	101	58	103.7	58.6	76.9	46.4	7.77E-03
double pod	1.1	0.3	1.1	0.3	1	0.2	4.26E-02
Ascochyta susceptibility	0.8	1	0.8	1	0.7	0.7	6.54E-01
presence of nodules on roots	1	0.2	1	0.2	1.1	0.4	1.05E-01
pod dehiscence	2	0.2	2	0.2	1.9	0.3	6.51E-01
color of flowers	3.6	0.6	3.6	0.7	3.9	0.4	1.86E-04
color of seeds	3.3	0.6	3.3	0.6	3.6	0.6	4.25E-03

* p-value from *t*-test.

Bayesian network analysis. The analysis was conducted using the R package bnlearn [24]. Traits were discretized using a quantile-based discretization. To find trait relationships that are consistent across different algorithms, six different types of learning algorithms were applied: hc, gs, iamb, inter.iamb, mmhc, and rsmax2. Stable trait relationships were identified by means of bootstrap resampling using a minimum frequency of 85% as the significance threshold [25]. The final set of trait relationships was determined by requiring consistency across the six models.

2.8. Random forest

To explore traits that differ most between the AR and AE progenies, we measured trait importance. We applied the random forest, a popular machine learning method, to assess trait importance. The trait importance measure corresponds to the importance index of the random forest, which is defined as follow:

$$\text{Importance index} = \frac{1}{ntree}\sum_{t} \tilde{errOOB}_t^j - errOOB_t$$

where t is a tree that corresponds to a single decision tree, ntree is the number of trees, errOOB_t is the prediction error of tree t, and $\tilde{\text{errOOB}}_t^j$ is the prediction error of tree t when a predictor *j* is perturbed [26]. The ranking of the importance index can be utilized to prioritize predictors for a prediction. In the sense of predicting AR and AE, the importance index ranks the ability of traits to distinguish crosses based on the parental origin (i.e., *C. reticulatum* versus *C. echinospermum*). To measure the importance index, we first grouped correlated traits using a correlation coefficient 0.5 and only utilized one representative trait from each group to prevent spurious signal [26]. To balance the sample size of two crosses, the SMOTE [27] algorithm in the R package DMwR was used. To build the random forest, we used the R package randomForest. The number of trees and number of predictors in each tree was set to 500 and 4, respectively.

3. Results

3.1. LD decay and population structure

To assess the genome-wide Linkage Disequilibrium (LD), LD decay was calculated using R-square between SNPs. The average of R-square values for SNP pairs was calculated

in intervals of 50 Kb. Based on our result, the LD reached its half of the maximum value at 1.9 Mb, which is substantially longer than previously reported LD decay measures that were less than 800 Kb [28,29] (Fig. S5). The observed long decay was as expected since our data was generated from short cycle of breeding and selfing.

Population structure was assessed using the Bayesian Markov Chain Monte Carlo model (MCMC) implemented in STRUCTURE [19]. The optimal K was determined to be 14 after investigating the log probability of data [LnP(D)]. The output of STRUCTURE showed that progeny materials share both cultivar and wild parental genetic characteristics (Fig. S4). The subpopulation pattern of progeny clearly followed the characteristic of their wild parent. Principal Component Analysis (PCA) showed the similar pattern, exhibiting progeny materials being intermediate genetic sources of their cultivar and wild parents (Fig. S3).

3.2. Analysis of phenotypes

Representatives of multiple, diverse genetic populations of two species of wild *Cicer*, *C reticulatum* (R) and *C. echinospermum* (E), were crossed with the elite chickpea cultivar *C. arietinum* (A) ICCV96029 [5] to generate F2 populations. 284 derived F3 lines were selected to survey populations from each of twenty diverse wild parents: 17 *C. reticulatum* (AR) parents (255 lines) and three *C. echinospermum* (AE) accessions (29 lines). Among 23 traits, we observed broad differences between AR and AE with maturation time related traits showing the greatest difference (Table 1), especially time to seed maturity. AE progeny tended to have more erect growth, greater plant height, and higher location of the lowest pod. 1000 seed weight was also higher in AE. Color of seeds was also significantly different between the two crosses, with seeds of AE generally darker than seeds from AR.

Several traits were highly correlated. ‘full flowering - start of germination’ and ‘full flowering - start of flowering’ showed the highest positive correlation (correlation coefficient 0.9; Fig. 1). Pod width and pod length were also highly positively correlated (correlation coefficient 0.5; Fig. 1). Maturation time related traits clustered in two separate groups, with a correlation coefficient threshold 0.5. Highly negative correlations were found between the following pairs of traits: the

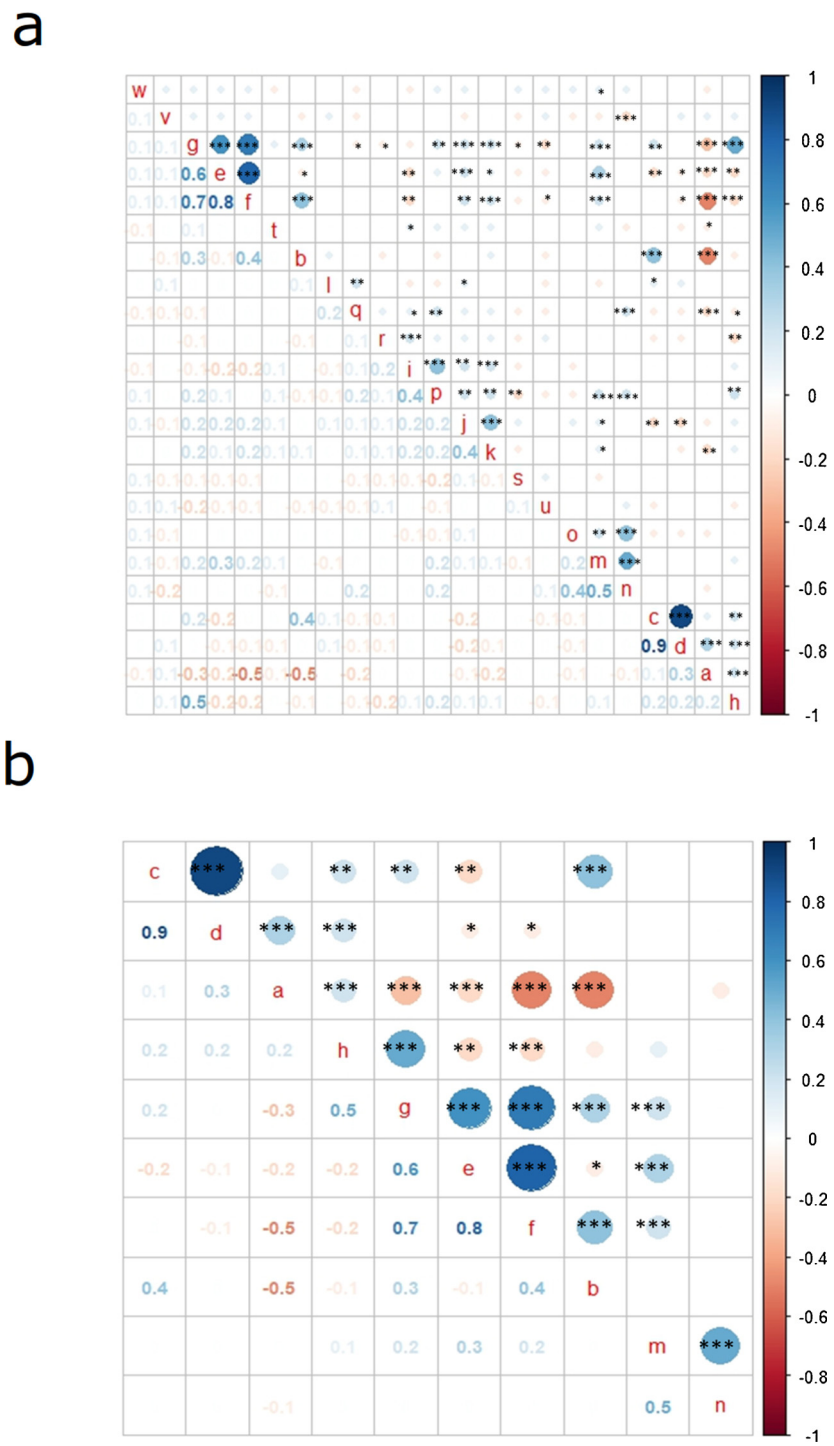

Fig. 1. Correlation analysis results of (A) 23 traits and (B) traits that exhibited strong correlations (correlation coefficient > .5). Correlation significance is denoted as * < 0.05, ** < 0.01, *** < 0.001. The corresponding trait of the each alphabet are: a) start of germination - sowing, b) start of flowering - start of germination, c) full flowering - start of germination, d) full flowering - start of flowering, e) start of maturation - start of flowering, f) start of maturation - start of germination, g) full maturation - start of germination, h) full maturation - start of maturation, i) growth habit, j) plant height, k) location of the first pod, l) number of pods per peduncle, m) pod width, n) pod length, o) number of seeds per pod, p) 1000 seed weight, q) seed weight per plot, r) double pod, s) Ascochyta susceptibility, t) presence of nodules on roots, u) dehiscence of pods, v) color of flowers, w) color of seeds.

pair 'start of germination – sowing' and 'start of maturation - start of germination', and the pair 'start of germination – sowing' and 'start of flowering - start of germination'.

3.3. Association analysis

We conducted association analyses to investigate QTLs that are associated with 23 traits of 284 F3 plants. Genotype data represent F2 parents, while phenotypes were averaged among 20 F3 sibling plants for each F2 lineage. The analysis was carried out on the combined data of AR and AE, and then separately for AR. AR specific analysis was performed with 14,201 markers, while for the combined data analysis 4713 pan-collection markers (markers that exist in both datasets) were selected. AE was not analyzed separately due to its small sample size. To remove confounding signals from the wild parents and family effects from each cross, we applied a mixed-linear model with a relationship matrix as a random effect and the family information as fixed cofactors. Each associated marker was remapped to the linkage map built based on 1950 F2 lines of *C. arietinum x C. reticulatum* (Table A.1 in Supplementary material). Trait-associated markers that were mapped to the same linkage map bin (defined by 1 cM distance) were considered as a single QTL. Using the Bonferroni multiple test correction, we detected 50 QTLs in 11 traits of AR and 35 QTLs in 10 traits of the combined data (Fig. 2, Table A.2 in Supplementary material). To

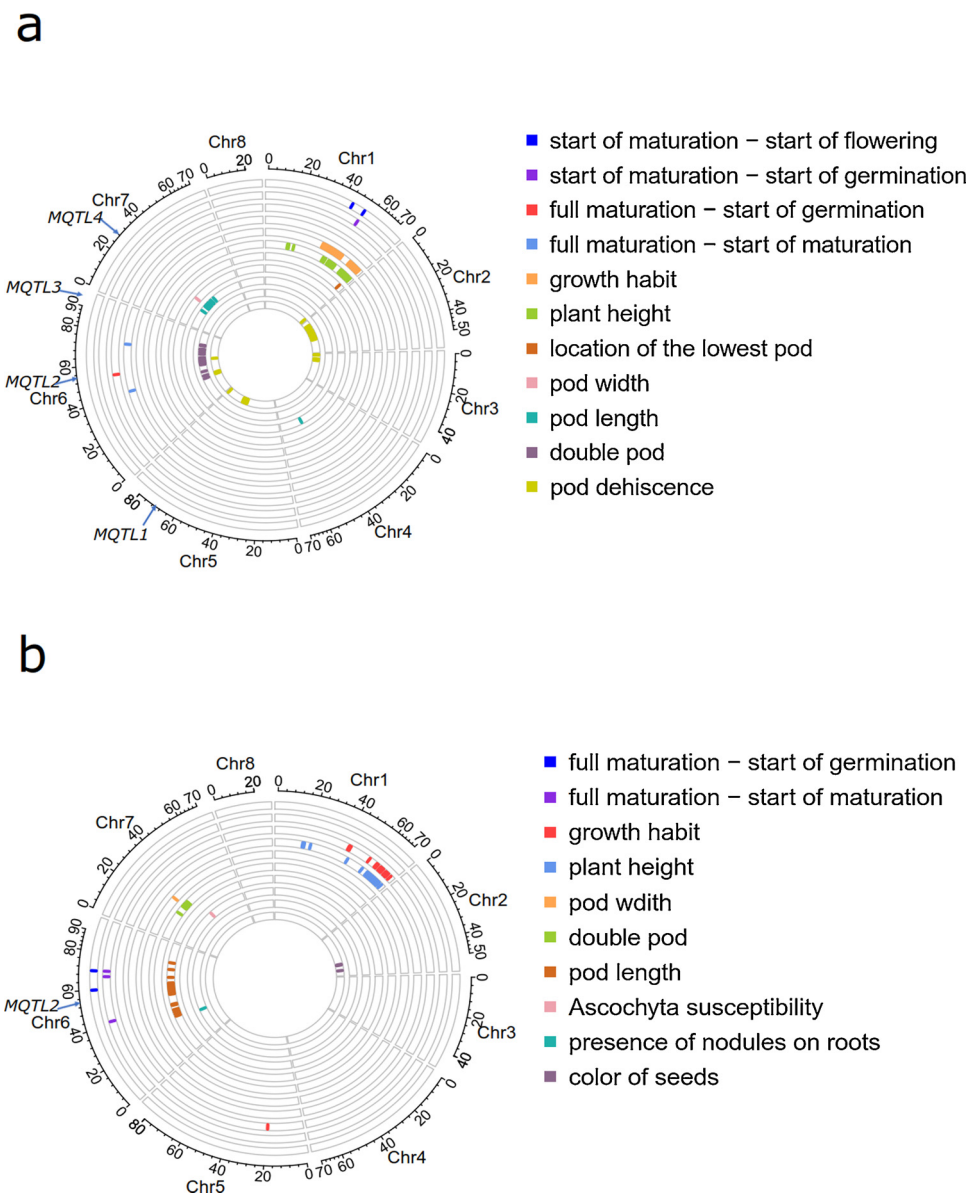

Fig. 2. Association hits on each chromosome. (A) *C. reticulatum* and (B) All data combined.

Table 2
Groups of traits with correlation coefficients bigger than 0.5.

Group A	full flowering – start of germination, full flowering – start of flowering
Group B	full maturation – start of germination, full maturation – start of maturation
Group C	start of maturation – start of flowering, full maturation – start of germination, start of maturation – start of germination
Group D	pod width, pod length
Group E	start of germination – sowing, start of maturation – start of germination
Group F	start of germination – Sowing, start of flowering – start of germination

investigate the variance explained by QTLs, R^2 of the most significant marker in each QTL was assessed. QTL markers associated with plant height exhibited comparatively high R^2 values above 0.5 (Table A.2 in Supplementary material). In addition, QTL markers associated with seed maturation related traits showed moderately high R^2 values above 0.3. Among the observed QTLs, those associated with Ascochyta susceptibility, presence of nodules on roots, and color of seeds were discovered only in the combined data set. The growth habit QTL on chromosome 5 was uniquely found in the combined data, whereas other QTLs were present on chromosome 1 in both data sets. Previously reported growth habit SNP Ca_Kabuli_Chr1 6262577 was found to be within a 1 cM window with the marker 1: 6481879 and the marker 1:6660037 that showed moderate level of p-value in AR and AR + AE. repectively (p-value: 0.00088 and 0.0002) [30].

3.4. Multi-trait association analysis

To explore QTLs with pleiotropic effects, we implemented a multi-trait association analysis. Traits were grouped based on correlation coefficients using hierarchical clustering and traits with a correlation coefficient higher than 0.5 were assigned to the same group (Fig. 1, Table 2). As a result, six groups were determined and they were

Table 3
QTLs showed significance in the multi-trait association analysis.

Data type	Group	QTL name	Marker	Chr	cM	p-value	Gene
AR	Group B	MQTL1	5:38343696	5	68	8.12E-01	LOC101492219
AR	Group B	MQTL2	6:30959263	6	54	4.26E-02	LOC101490414
AR	Group B	MQTL3	6:57185489	6	95	5.15E-02	receptor-like protein kinase FERONIA-like
AR	Group C	MQTL1	5:38343696	5	68	1.34E-01	LOC101492219
AR	Group C	MQTL2	6:30959263	6	54	3.15E-01	LOC101490414
AR	Group D	MQTL4	7:11287194	7	30	8.05E-02	AP2-like ethylene-responsive transcription factor
Combined	Group B	MQTL2	6:30959263	6	54	4.24E-01	LOC101490414

Table 4
Influential relationship of traits. Left: influencing trait and right: influenced trait.

Data type	From	To
C. arietinum x C. reticulatum, combined	start of maturation – start of flowering	start of maturation – start of germination
C. arietinum x C. reticulatum, *C.* combined	start of maturation – start of germination	full maturation – start of germination
combined	full maturation – start of maturation	full maturation – start of germination
C. arietinum x C. reticulatum	start of germination – sowing	start of maturation – start of germination
combined	plant height	location of the lowest pod
C. arietinum x C. reticulatum, combined	growth habit	1000 seed weight

analyzed using the multi-trait mixed model. In AR, four QTLs were found in group B, C, and D (Fig. 2a, Table 3). There were three QTLs associated with group B, and among them, two QTLs (MQTL1 and MQTL2) overlapped with QTLs found in group C, suggesting their global role in the duration of seed maturation. In the combined data, a single QTL, MQTL2, was found associated with group B (Fig. 2b, Table3).

By comparing these results to the single-trait association results, we investigated multi-trait QTLs that also appeared as single-trait QTLs or were located close to single-trait QTLs. MQTL2 associated with group B in both data sets was one of the QTLs associated with full maturation - start of germination. Similarly, MQTL4 associated with group D was one of the QTLs associated with pod length. On the other hand, MQTL1, which was associated with group B and C, did not overlap with the single-trait association results. In summary, the multi-trait analysis efficiently captured pleiotropic QTLs in two groups among a total of six groups of highly correlated traits.

3.5. Influential relationships of traits

To investigate the influential relationship between traits, we implemented a Bayesian network analysis using phenotypes of AR and AE. To guarantee the stability of the Bayesian model, six types of Bayesian relationships were measured and only relationships that appeared across all six models were reported. Directional arcs of Bayesian networks can be interpreted as causal on the condition that all traits that are interrelated were measured. According to the analysis, there were four influential relationships found, all related to the duration of either vegetative growth or seed maturation. We observed relationships involving the timing of germination and other phenology-related traits, including relationships between 'start of germination – sowing':'start of maturation – start of germination' uniquely in AR and between' full maturation – start of maturation': 'full maturation - start of germination' in the combined data. Similarly, there was a relationship between plant height and location of the lowest pod. Less obvious is the observed relationship of 1000 seed weight and 1000 seed weight. Rubio et al [31] found that chickpeas with erect growth habit have greater seed size, implicating genetic or physiological relationship between these two traits [31]. Under the assumption that our data capture all of the representative correlated traits, the relationships found by Bayesian network analysis can be considered as dependencies between traits.

Based on the GWAS results, one can infer whether the deduced influential relationships are governed by common QTLs or not. In the maturation time related relationships, the pair 'start of maturation - start of flowering':'start of maturation - start of germination' shared the common QTL located on chromosome 1 (Fig. 2). The result may indicate that these paired traits share common regulatory loci or pathways. On the other hand, the QTLs of 'start of maturation - start of germination' and 'full maturation - start of germination' were located on chromosome 1 and 6, respectively. With respect to the pair, 'growth habit':'1000 seed weight', the QTL associated with growth habit did not contribute measurably to 1000 seed weight. Likewise, 'start of germination -sowing' was not associated with a QTL despite the fact that its counterpart trait, 'start of maturation - start of germination' had a QTL on chromosome 1.

Comparison of the multi-trait association results with the Bayesian network results make it possible to illustrate which relationships were affected by pleiotropic QTLs. Based on the multi-trait association results, two seed maturation related relationships, 'start of maturation - start of flowering':'start of maturation - start of germination', and 'start of maturation - start of germination':'full maturation - start of germination', were governed by two pleiotropic QTLs, MQTL1 and MQTL2 (Tables 2–4). Furthermore, the other seed maturation relationships, 'full maturation – start of maturation':'full maturation - start of germination', were governed by three QTLs, MQTL1, MQTL2 and MQTL3. On the other hand, plant height, seed weight and growth habit related relationships showed no evidence of pleiotropic QTLs.

3.6. Trait importance index analysis using machine learning

To assess the degree to which each trait differs in two crosses, trait importance was measured using the importance index of the random forest method fitted to predict the wild parent origin of each cross, i.e., AR versus AE. Before the analysis, sample size was balanced using the algorithm SMOTE [27]. To prevent biased results that can derive from correlation of traits, traits with a correlation coefficient higher than 0.5 were assigned to the same group and only one representative trait from each group was used for the analysis (Table 2). The results showed that seed maturation time, seed weight, and plant height-related traits most effectively classify the two types of crosses (Fig. 3). Combining the trait importance results with the Bayesian network results, we found that the influenced traits have higher importance than the influencing traits. More specifically, location of the lowest pod showed higher importance than plant height, while 1000 seed weight had higher importance than growth habit. The combination of these two analyses suggests that, for future breeding schemes, the choice of the wild parent should consider

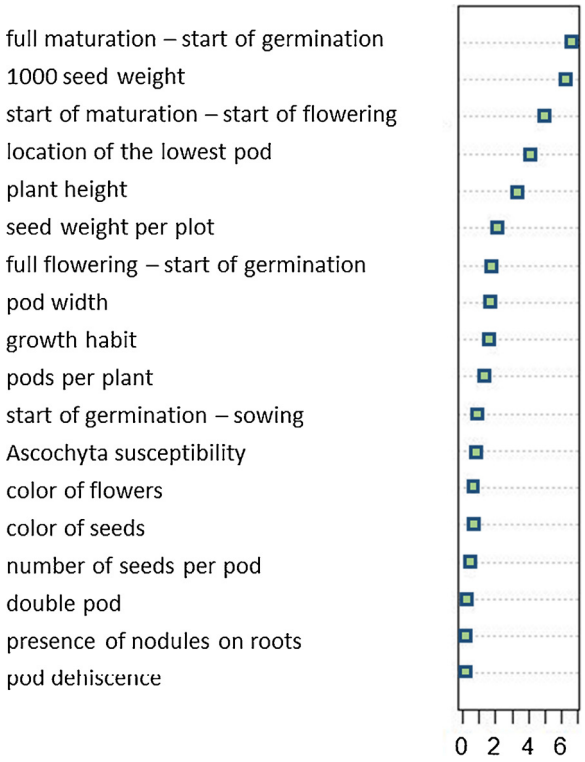

Fig. 3. Trait importance index sorted by the importance of each trait. x-axis corresponds to the decreased accuracy of the cross type prediction which corresponds to the importance of each trait.

not only the traits that differ most, which corresponds to a higher trait importance, but also the traits that influence other traits irrespective of their difference between crosses.

4. Discussion

4.1. By the integration of single and multiple trait association analysis, genomic regions involved in important traits of chickpea were discovered

For millennia, agriculturalists and breeders have focused on selecting organisms with the desirable phenotypes. Strong selection on a few traits during domestication by definition caused a reduction in adaptive variation due to genetic bottlenecks, the strength of which vary depending on the crop, but are particularly pronounced in chickpea [5]. Counteracting these forces requires tapping diverse germplasm, including wild species and more primitive landraces that can be sources of variants useful for improving elite varieties. Systematic breeding with wild materials is uncommon, in part because the desirable characteristics of elite cultivated genotypes can be greatly reduced in wide cross populations. The burden of disentangling traits in wide cross populations is not only labor intensive, but often desirable traits are missed entirely due to genetic interactions. Modern genomics and quantitative genetics can resolve these complexities and reduce the barrier to use of wild species in crop improvement. Here we conducted an analysis of the genetic underpinning of individual domestication traits and the genetic basis of multi-trait complexes. The resulting inferences may be of value to optimize wide-cross breeding by allowing a priori weights for different genes and variants to be integrated into marker-assisted and genomic-selection programs.

We explored QTLs of 23 chickpea traits by applying both single and multiple trait association analysis. As depicted in the correlation analysis, some pairs of traits were strongly correlated. In several cases the associated QTLs overlapped, even when traits were not strongly correlated. For example, plant height and growth habit were poorly correlated (correlation coefficient 0.2), which is consistent with a previous study reporting overall low correlation coefficients of these traits [32]. Nevertheless, breeders have an interest combining these characters, because tall plants with reduced canopy density are desirable for machine harvesting. Interestingly, despite their low correlation, the two traits occur in moderately distant QTL region (18 cM in AR and 10 cM in the combined data), suggesting that targeting these regions will improve plant height and growth habit simultaneously.

Most of the QTLs found in the combined data were also found in AR alone. QTLs specific to the combined AR + AE data include Ascochyta susceptibility, presence of nodules and color of seeds. Among possible explanations for combined data-specific QTLs are: i) the subset of markers in the combined data was not present in AR because the MAF of the markers was lower in AR, ii) the combined data had better power to detect causal markers due to the increased sample size, or iii) the effect of the causal markers in AE is greater. QTLs associated with Ascochyta susceptibility and presence of nodules on roots corresponded to case i). In the case of the QTLs associated with color of seeds, two linked SNP showed a slightly higher seed color increase as a function of genotype, suggesting the combined data-specific signal originates from iii). (Figs. S1 and S2). Future studies with larger sample sizes and perhaps with specific AR and AE parents are required to test whether these are species specific QTLs (i.e., uniquely derived from *C. reticulatum* vs *C. echinospermum*), or artifacts of the current genetic materials.

Multi-trait association analysis efficiently identifies signals caused by pleiotropic loci. Our analysis of multi-trait associations found loci in common with single-trait association analysis, and also loci that were not detected in the single-trait analysis. MQTL2 and MQTL4 are the cases where an association was found by both association analysis methods. On the other hand, MQTL1 was located in a region not identified in single-trait analysis. This can be explained by the fact that the multi-trait model has higher power, and thus signal detection is more efficient than in the single-trait model. Differential discovery of genetic associations is expected to be especially high for loci that are measurably pleiotropic, but that make relatively small contributions to single traits.

4.2. Bayesian network analysis provides more information on trait relationships than correlation analysis can do

To investigate the genetic regulation of multiple trait interactions, we used Bayesian network analysis. Bayesian network analysis found trait interrelationships that could not be detected based on trait correlation alone. Using six Bayesian network methods to ensure the consistency of the relationships, we found six pairs of traits with influential relationships. Four of the relationships were involved the regulation of seed maturation time, while the remainder involved characteristics of plants, pods and seeds.

In the case of maturation time influences, all interactions involved germination-related traits. Delayed germination exposes seedlings to differing environments, i.e., day length and potentially temperature vary with time, which may influence the rate of plant development. Thus at least some of the observed signal may be caused by germination environment, rather that direct genetic interactions. Interestingly, however, the results of GWAS implicate a genetic component, at least in the case of the pair ‘start of maturation - start of flowering’:’start of maturation - start of germination’, which share a common QTL located on chromosome 1 (Fig. 2).

By better understanding the influential relationships between agronomic traits, we can predict the outcomes of changing certain traits and prevent or induce undesirable or desirable changes in other traits. For instance, according to Bayesian network results, regulating the height of the chickpea plant may change the lowest pod position. While the result may sound obvious – taller plants have more upright growth, while upright growth raises the height of the lowest pods–, quantifying

the nature of such relationships and ultimately determining the nature of their genetic control, including pleiotropy and linkage, has broad implications especially for crops such as chickpea where a quantitative genetic framework is largely absent.

4.3. The combination of QTL mapping, Bayesian networks, and trait importance can contribute to in-depth strategy for future breeding

Introgressing wild germplasm into the cultivated genepool provides an opportunity to overcome obvious limitations in selective breeding of cultivated plants caused by a narrow genetic basis. When multiple traits are considered, the choice of the parental type is expected to affect the relationship between the target traits. For example, Guo et al., (2017) investigated timing and quality traits in two types of Petunia crosses: P. axillaris x P. exserta (AEx) and P. integrifolia x P. axillaris (IA). The authors found that the correlation of the development rate and time to flower was different between the two crosses depending on temperature. Moreover, the percentage of traits with significant correlations was different: 63.8% and 43.2% in AEx and IA respectively. This example illustrates the importance of prioritizing wild materials based on the characteristics of the target traits. This priority can be achieved by assessing how much each trait differs between cross types.

In the current evaluation, we used the random forest method to measure the importance of each trait, identifying traits that best differentiate populations based on their wild species origin (*C. reticulatum* vs *C. echinospermum*). The integration of trait importance index, influential relationships and GWAS results gives insight into how wild species might be used systematically in crop improvement. For instance, 1000 seed weight was influenced by growth habit in Bayesian network analysis. Seed weight is an important agricultural trait and thus introgressing QTL associated with seed weight is a desirable choice for breeding. However, according to our GWAS results, no corresponding QTL region was found. The reason of absence of signal can be explained by the polygenic characteristic of seed weight of legumes of which QTL detection requires large number of samples [33,34]. We did however identify multiple QTL associated with growth habit. Bayesian network analysis suggests a potential solution: introgression of the growth habit QTL is predicted to increase seed weight, despite the presence seed weight-associated QTL. Following the trait importance index, growth habit might be discarded as a breeding objective. However, understanding that growth habit can influence seed size might lead to a re-prioritization, permitting the use of growth habit-associated QTL to pursue increased seed size. We suggest that the insights of this study, based on integration of multiple traits and different analytic methods, can contribute to improving cultivar chickpeas in future breeding systems.

5. Conclusions

This study investigated agronomic traits of F3 progeny derived from the elite, *C. arietinum* cultivar ICCV96029 and two species of wild chickpea, *C. reticulatum* and *C. echinospermum*, were crossed with the elite, early flowering *C. arietinum* cultivar ICCV96029. The application of a single-trait genome-wide association analysis revealed 50 QTLs in 11 traits for AR and 35 QTLs in 10 traits for the combined data. By performing a multi-trait association analysis, we found four pleiotropic QTLs. Moreover, we detected phenotypic-level inter relationships of the duration of vegetative growth and seed maturation, seed weight, growth habit, plant height and the location of the lowest pod utilizing a Bayesian analysis. Combining the results from the two types of association approaches and the Bayesian analysis, we detected QTLs that govern traits that showed inter relationships. Lastly, a random forest approach was applied to find the traits differ most between the AR and AE progenies. By combining multiple traits and analytic methods, this study provides an insight for the efficient scheme to target agronomic QTLs for future breedings.

Availability of data and material

The phenotype dataset that underlies the conclusions of this study is attached as Supplementary Table 3. The Illumina sequences of two interspecific chickpea inbred lines have been deposited in NCBI under the BioProject PRJNA507624. The in-house script used for the GBS data binning can be found at https://github.com/wioxio/GBS_binning/blob/master/GBS_binning.r.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

CRediT authorship contribution statement

Min-Gyoung Shin: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. **Sergey V. Bulyntsev:** Investigation, Resources. **Peter L. Chang:** Formal analysis. **Lijalem Balcha Korbu:** Investigation. **Noelia Carrasquila-Garcia:** Investigation. **Margarita A. Vishnyakova:** Investigation. **Maria G. Samsonova:** Funding acquisition, Investigation, Resources. **Douglas R. Cook:** Funding acquisition, Investigation, Project administration, Resources, Writing - review & editing. **Sergey V. Nuzhdin:** Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgements

This work was supported by the US National Science Foundation Plant Genome Program NSF PGRP 1339346 to D.R.C.; by a cooperative agreement from the United States Agency for International Development under the Feed the Future Program AID-OAA-A-14–00008 to D.R.C. and S.N.; by a gift from Mars Incorporated to D.R.C.; by a grant from the Government of Norway through the Global Crop Diversity Trust CWR14NOR2 3.3 07 to D.R.C. The work on field cultivation, phenotyping, and field data analysis as well as co-execution of data analysis and co-writing the manuscript was supported by the Russian Scientific Fund project no. 16-16-00007 to M.V., S.B, S.N. and M.G.S

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.plantsci.2019.04.018.

References

[1] A. Afshin, R. Micha, S. Khatibzadeh, D. Mozaffarian, Consumption of nuts and legumes and risk of incident ischemic heart disease, stroke, and diabetes: a systematic review and meta-analysis, Am. J. Clin. Nutr. 100 (2014) 278–288.
[2] Peña C.D. La, J. Pueyo, T. Coba, D. Peña, J.J. Pueyo, Legumes in the reclamation of marginal soils, from cultivar and inoculant selection to transgenic approaches, Agron. Sustain. Dev. 32 (2012) 65–91.
[3] J. Berger, S. Abbo, N.C. Turner, Ecogeography of annual wild Cicer species: the poor state of the world collection, Crop Sci. 43 (2003) 1076–1090.
[4] E. Plekhanova, M.A. Vishnyakova, S. Bulyntsev, P.L. Chang, N. Carrasquilla-garcia, K. Negash, E. Wettberg, Von, N. Noujdina, Genomic and phenotypic analysis of Vavilov's historic landraces reveals the impact of environment and genomic islands of agronomic traits, Sci. Rep. 7 (2017) 4816.

[5] E.J.B. von Wettberg, P.L. Chang, F. Başdemir, N. Carrasquila-Garcia, L.B. Korbu, S.M. Moenga, G. Bedada, A. Greenlon, K.S. Moriuchi, V. Singh, et al., Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation, Nat. Commun. 9 (2018) 649.
[6] S. Gupta, T. Kumar, S. Verma, C. Bharadwaj, S. Bhatia, Development of gene-based markers for use in construction of the chickpea (Cicer arietinum L.) genetic linkage map and identification of QTLs associated with seed weight and plant height, Mol. Biol. Rep. 42 (2015) 1571–1580.
[7] V.K. Singh, A.W. Khan, D. Jaganathan, M. Thudi, M. Roorkiwal, H. Takagi, V. Garg, V. Kumar, A. Chitikineni, P.M. Gaur, et al., QTL-seq for rapid identification of candidate genes for 100-seed weight and root/total plant dry weight ratio under rainfed conditions in chickpea, Plant Biotechnol. J. 14 (2016) 2110–2119.
[8] Y. Li, P. Ruperao, J. Batley, D. Edwards, T. Khan, T.D. Colmer, J. Pang, K.H.M. Siddique, T. Sutton, Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data, Front. Plant Sci. 9 (2018) 190.
[9] Z.A. Desta, R. Ortiz, Genomic selection: genome-wide prediction in plant improvement, Trends Plant Sci. 19 (2014) 592–601.
[10] X. Wang, Y. Xu, Z. Hu, C. Xu, Genomic selection methods for crop improvement: current status and prospects, Crop J. 6 (2018) 330–340.
[11] C.E. Schlosberg, W. Duan, N.L. Saccone, Application of Bayesian network structure learning to identify causal variant SNPs from resequencing data, BMC Proc. 5 (2011) S109.
[12] J. Hou, G. Stacey, J. Cheng, Exploring soybean metabolic pathways based on probabilistic graphical model and knowledge-based methods, EURASIP J. Bioinform. Syst. Biol. 2015 (2015) 5.
[13] Y. Jin, Y. Su, X. Zhou, S. Huang, T. Alzheimer, N. Initiative, Heterogeneous multimodal biomarkers analysis for Alzheimer's disease via Bayesian network, EURASIP J. Bioinform. Syst. Biol. 2016 (2016) 12.
[14] T.Y. Curtis, V. Bo, A. Tucker, N.G. Halford, Construction of a network describing asparagine metabolism in plants and its application to the identification of genes affecting asparagine metabolism in wheat under drought and nutritional stress, Food Energy Secur. 7 (2018) e00126.
[15] R.K. Varshney, C. Song, R.K. Saxena, S. Azam, S. Yu, A.G. Sharpe, S. Cannon, J. Baek, B.D. Rosen, B. Tar'an, et al., Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement, Nat. Biotechnol. 31 (2013) 240.
[16] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595.
[17] B.V.W. Kyazma, Join Map 4: Software for the Calculation of Genetic Linkage Maps in Experimental Populations, [WWW document] URL http://dendrome.ucdavis.edu/resources/tooldocs/joinmap/JM4manual.pdf . (Accessed 15 September 2012) 59 (2006).
[18] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.
[19] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[20] X. Zheng, D. Levine, J. Shen, S.M. Gogarten, C. Laurie, B.S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data, Bioinformatics 28 (2012) 3326–3328.
[21] Christoph Lippert, Gerald Quon, Eun Yong Kang, M. Carl, Jennifer Kadie, D.H. Listgarten, The benefits of selecting phenotype-specific variants for applications of mixed models in genomics, Sci. Rep. 3 (2013) 1815.
[22] C. Lippert, J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, D. Heckerman, Fast linear mixed models for genome-wide association studies, Nat. Methods 8 (2011) 833–837.
[23] X. Zhou, M. Stephens, Efficient multivariate linear mixed model algorithms for genome- wide association studies, Nat. Methods 11 (2014) 407–409.
[24] M. Scutari, Learning bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (2010) 1–22.
[25] Radhakrishnan Nagarajan, Marco Scutari, S. Lèbre, Bayesian Networks in R with Applications in Systems Biology, Springer-Verlag, New York, 2013.
[26] R. Genuer, J. Poggi, C. Tuleau-malot, Variable selection using random forests, Pattern Recognit. Lett. 31 (2010) 2225–2236.
[27] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, (2002), pp. 321–357.
[28] A. Kujur, D. Bajaj, H.D. Upadhyaya, S. Das, R. Ranjan, T. Shree, M.S. Saxena, S. Badoni, V. Kumar, S. Tripathi, et al., Employing genome-wide SNP discovery and genotyping strategy to extrapolate the natural allelic diversity and domestication patterns in chickpea, Front. Plant Sci. 6 (2015) 162.
[29] D. Bajaj, S. Das, S. Badoni, V. Kumar, M. Singh, K.C. Bansal, A.K. Tyagi, S.K. Parida, Genome-wide high-throughput SNP discovery and genotyping for understanding natural (functional) allelic diversity and domestication patterns in wild chickpea, Sci. Rep. 5 (2015) 12468.
[30] H.D. Upadhyaya, D. Bajaj, R. Srivastava, A. Daware, U. Basu, S. Tripathi, C. Bharadwaj, A.K. Tyagi, S.K. Parida, Genetic dissection of plant growth habit in chickpea, Funct. Integr. Genom. 17 (2017) 711–723.
[31] J. Rubio, F. Flores, M.T. Moreno, J.I. Cubero, J. Gil, Effects of the erect / bushy habit, single/double pod and late/early flowering genes on yield and seed size and their stability in chickpea, Field Crops Res. 90 (2004) 255–262.
[32] H.D. Upadhyaya, R. Ortiz, P.J. Bramel, S. Singh, Phenotypic diversity for morphological and agronomic characteristics in chickpea core collection, Euphytica 123 (2002) 333–342.
[33] R. Hovav, K.C. Upadhyaya, A. Beharav, S. Abbo, Major flowering time gene and polygene effects on chickpea seed weight, Plant Breed. 122 (2003) 539–541.
[34] T. Isemura, A. Kaga, S. Tabata, P. Somta, P. Srinives, T. Shimizu, et al., Construction of a genetic linkage map and genetic analysis of domestication related traits in mungbean (Vigna radiata), PLoS One 7 (2012) e41304.