






Genetic diversity and GWAS of agronomic traits using an ICARDA lentil (*Lens culinaris* Medik.) Reference Plus collection

Karthika Rajendran¹ , Clarice J. Coyne^{2*} , Ping Zheng³, Gopesh Saha⁴ , Dorrie Main³, Nurul Amin⁴ , Yu Ma³, Ted Kisha², Kirstin E. Bett⁵ , Rebecca J. McGee⁶ and Shiv Kumar¹

¹Biodiversity and Crop Improvement Program, International Center for Agricultural Research in the Dry Areas, Rabat, Morocco, ²Plant Germplasm Introduction and Testing Research Unit, USDA, ARS, Pullman, WA, USA, ³Department of Horticulture, Washington State University, Pullman, WA, USA, ⁴Department of Crops and Soils, Washington State University, Pullman, WA, USA, ⁵Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK, S7N 5A8, Canada and ⁶Grain Legume Genetics and Physiology Research Unit, USDA, ARS, Pullman, WA, USA

Received 31 July 2020, revised 4 February 2021; Accepted 5 February 2021 – First published online 7 July 2021

Abstract

Genotyping of lentil plant genetic resources holds the promise to increase the identification and utilization of useful genetic diversity for crop improvement. The International Center for Agriculture Research in the Dry Areas (ICARDA) lentil reference set plus collection of 176 accessions was genotyped using genotyping-by-sequencing (GBS) and 22,555 SNPs were identified. The population structure was investigated using Bayesian analysis (STRUCTURE, $k=3$) and principal component analysis. The two methods are in concordance. Genome-wide association analysis (GWAS) using the filtered SNP set and ICARDA historical phenotypic data discovered putative markers for several agronomic traits including days to first flower, seeds per pod, seed weight and days to maturity. The genetic and genomic resources developed and utilized in this study are available to the research community interested in exploring the ICARDA reference set plus collection using GWAS.

Keywords: GBS, genotyping by sequencing, GWAS, *Lens culinaris*, lentil, SNPs

Introduction

Lentil (*Lens culinaris* Medik.) is an important protein crop. It is a diploid ($2n = 2x = 14$) and possesses a large genome

(~4 Gbp) (Arumuganathan and Earle, 1991). It ranks fifth, after dry beans, chickpea, dry peas and cowpea, for pulses production in the world (FAOSTAT, 2017). The current global lentil production is estimated at 6.33 million metric tons and falls short of the global demand which is expected to increase soon due to rapid population growth and plant protein market (Reda, 2015). In order to bridge the demand-supply gap, efforts are required to accelerate the genetic gain, which is abysmally low mainly due to the narrow genetic base of cultivated lentil. This presents a serious barrier towards developing cultivars for future needs (Lombardi *et al.*, 2014; Khazaei *et al.*, 2016). Integration of genomic tools with conventional breeding approaches

*Corresponding author. E-mail: clarice.coyne@usda.gov

†Present address: VIT School of Agricultural Innovations and Advanced Learning (VAIAL), Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India.

‡ Present address: Brotherton Seed Co., 451 S Milwaukee Ave, Moses Lake, WA, USA.

§ Present address: Department of Crop and Soil Sciences, Breeder Seed Production Center, Bangladesh Agricultural Research Institute, Debiganj, Panchagarh, Bangladesh.

would help to alleviate bottlenecks by improving selection efficiency and speed up the breeding process in developing improved cultivars.

Recent efforts on increasing the utilization of plant genetic resources (PGRs) has focused on leveraging genomic tools to unlock the genetic potential of *ex situ* collections held in national and international gene banks all over the world (McCouch *et al.*, 2012, 2013; Mascher *et al.*, 2019). The linkage between genomic characterization and PGR on a global scale can assist with the future challenges to agricultural production such as climate change (Zimmerer and De Haan, 2017). Even without significant genotypic information for most crops, the USDA germplasm distributions doubled from 2006 to 2012 (Heisey and Rubenstein, 2015). For plant scientists, and especially plant breeders, access to new positive alleles is paramount for gradual and sustainable genetic gains over the breeding cycles. This requires the utilizing genomic-based tools specifically for genomic-assisted breeding (Varshney *et al.*, 2015), genomic selection (Annicchiarico *et al.*, 2017) and breeding-assisted genomics, the recent paradigm switch suggested by Poland (2015).

Recent technological advances are facilitating the expansion of genomic resources for food crops, particularly for pulse crops, in recent years (Varshney, 2016). It is mainly due to the notable reduced costs in sequencing and a surge in bioinformatics tool development (Varshney *et al.*, 2020). Many pulse genomes that have been sequenced include pea, lentil, common bean, kabuli chickpea, desi chickpea, cowpea and pigeonpea (Varshney *et al.*, 2012, 2013; Jain *et al.*, 2013; Schmutz *et al.*, 2014; Ogutcu *et al.*, 2018; Kreplak *et al.*, 2019; Lonardi *et al.*, 2019, respectively). Currently, genotyping by sequencing (GBS) is increasingly popular among pulse breeders to screen germplasm quickly and inexpensively (e.g. Guindon *et al.*, 2019; Ma *et al.*, 2020). As a high throughput approach, GBS in lentil has facilitated the discovery of genome-wide (SNPs), development of high-density linkage maps and assessment of the genetic diversity in the germplasm collection (Temel *et al.*, 2015; Wong *et al.*, 2015; Khazaei *et al.*, 2017a, b; Ma *et al.*, 2020).

The International Centre for Agricultural Research in the Dry Areas (ICARDA) has a global mandate for the genetic improvement of lentil. The ICARDA lentil reference set (Kumar *et al.*, 2015), representing the major production and geographical (51 countries) regions, was phenotyped for economically important traits, but was genotyped with only microsatellites. The objectives of this project were to (1) construct a public available lentil SNP genotype set for internationally available lentil PGRs, (2) explore the population structure and diversity, and (3) assess the genotyped collection for possible marker identification (allelic contribution/function) for agronomic traits using genome-wide association study

(GWAS) by data mining historical data collected by ICARDA.

Materials and methods

Plant material and field data

In this study, the ICARDA Reference Plus collection of 176 lentil lines (130 Generation Challenge Program (GCP)) reference set (Furman, 2006; Kumar *et al.*, 2015), plus 39 abiotic stress-tolerant lines and seven recombinant inbred lines parents were selected based on phenotypic diversity from the world lentil germplasm collection held by ICARDA (online Supplementary Table S1). The field data presented in online Supplementary Table S2 were historic data collected by ICARDA (e.g. Migicovsky *et al.*, 2016; González *et al.*, 2018a). Plant materials were grown using an α -lattice design with two replicates at two ICARDA experiment stations: (1) Tel Hadya, Syria and (2) Terbol, Lebanon, from 2007 to 2011. During the crop growing period, all crop management practices typical for the area were followed. Lines were phenotyped for days to first flower (number of days from sowing to the appearance of the first flower); plant height (average height of five plants from the ground to the tip of the foliage at maturity); seeds per pod (average number of seeds in 10 randomly chosen dry pods); days to maturity (number of days from sowing until 90% of the pods were golden brown); biomass yield of each plot (weight of dried mature plants in a plot); seed yield (seed yield of a plot after threshing, expressed as kg/ha); straw yield (calculated as the difference between biomass yield and seed yield); harvest index (calculated as the ratio of seed to biomass yield); and hundred-seed weight (average weight of two samples of 100 randomly chosen seeds in g). Phenotypic values were also combined across years and averaged in cases of replication for a particular accession (online Supplementary Table S2; Supplementary Fig. S1).

Genotyping

DNA was extracted, using a DNeasy Plant Kit (QIAGEN, Valencia, CA), from a single plant per accession grown in the greenhouse at the USDA-ARS Western Regional Plant Introduction Station in Pullman, WA in 2013. DNA was quantified using a spectrophotometer (Nano-Drop Technologies, Wilmington, DE, USA).

The two-enzyme genotyping-by-sequencing procedure of Poland *et al.* (2012) was followed using the modifications of Wong *et al.* (2015). Briefly, 200 ng of genomic DNA was double-digested with *Pst*I and *Msp*I and ligated to two adapters, of which one contained a barcode sequence. Samples were pooled, PCR amplified and cleaned

up using a column (Qiagen QIAquick PCR Purification Kit). Four libraries of 48 bar-coded samples were sequenced in four lanes using an Illumina HiSeq2000 by the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley.

Genotypic data analysis

The sequencing data were processed to remove low-quality data and analysed using 'Stacks' software (Catchen *et al.*, 2013). Unfiltered Fastq sequence Illumina data were assigned to individual samples via the barcode sequence using 'Stacks' software (Catchen *et al.*, 2011, 2013). The RAD-Tags algorithm was used to examine raw reads from Illumina sequencing runs by checking that the barcode and the RAD cut-site are intact, and at the same time de-multiplex the data. The default parameters of the 'Stack' software were used to call the SNPs. 'Stacks' uses short-read sequence data to identify and genotype loci in a set of individuals and in-house scripts generated a VCF format genotype data file.

A second analysis was conducted using FreeBayes software (Garrison and Marth, 2012) to call the variants using the lentil reference genome, Lc1.2 of 'CDC Redberry' (Ramsay *et al.*, 2016; Bett, 2020). Analyses of pipeline details are presented in Yu *et al.* (2017). Briefly, the SNPs were identified with the lentil reference genome version Lc1.2 using BamTools (Barnett *et al.*, 2011) and the FreeBayes variant caller (<https://github.com/ekg/freebayes>) (Garrison, 2012; Garrison and Marth, 2012). The 50% missing data were used as a minimum to keep the variant in the final filtered VCF file to give future users of the SNP data flexibility on filtering without resorting to reanalysing the raw data (Glaubitz *et al.*, 2014; O'Leary *et al.*, 2018).

Structure and diversity

The genetic structure of the collection was determined using the software STRUCTURE (Pritchard *et al.*, 2000). PGDSpider software was used to convert the *.VCF file to the STRUCTURE software format (Lischer and Excoffier, 2012). Parameter set included a 10,000 burn-in period, 10,000 Markov Chain Monte Carlo (MCMC) replications, admixture model, run for each subpopulation (K) value ranging from 1 to 7. The best K value was determined by plotting the rate of change in the log probability of data (ΔK) against the successive K values (Evanno *et al.*, 2005) implemented in STRUCTURE HARVESTER (Earl and von Holdt, 2012). The K value was considered to be optimum while ΔK reaches the maximum. A tree was constructed in NTSys-pc using Prevosti's Distance substituting the probability of assignment to each population at $K = 3$ for allele frequency (Rohlf 2009). The distance matrices were used

to produce a dendrogram based on clustering using the unweighted pair-group method with arithmetic averages (UMGMA) in the SAHN module of NTSYS-PC program version 2.02 k (Rohlf, 2009). A tree view was created using the distance matrix and UPGMA (Sneath and Sokal, 1973) clustering method modules in STRUCTURE. Genetic diversity between subpopulations as determined by the STRUCTURE software was calculated using Analysis of Molecular Variance (AMOVA) which calculates PhiPT (Excoffier *et al.*, 1992). Phi-statistics is a modified version of Wright's F that refers to the relative contributions of between-subpopulation separation to the overall genetic variation in the whole sample. The variance components are used to calculate phi-statistics which are analogous to Wright's F -statistics, $\Phi_{ST} = (\sigma^2 a + \sigma^2 b) / \sigma^2 T$ (Schneider *et al.*, 2000). AMOVA was calculated using the 'Distance' and 'AMOVA' functions in GenAlEx 6.5 (Peakall and Smouse, 2006, 2012). Principal component analysis (PCA) was conducted using the 'PCA' module in TASSEL using the SNP data and graphed using SigmaPlot Version 13.0 (Systat Software, San Jose, CA, USA) (Bradbury *et al.*, 2007).

GWAS

The GWAS was conducted using phenotypic data means collected from 2007 to 2011 (online Supplementary Table S2) held in the lentil germplasm database of ICARDA using the GLM-PCA batch commands in the software TASSEL 5.2.29 (Bradbury *et al.*, 2007). SNPs were filtered to a maximum of 15% of the lines missing the SNP call with a minimum allele frequency of 0.05 (Glaubitz *et al.*, 2014; O'Leary *et al.*, 2018). Marker-trait associations (MTAs) were analysed for hundred-seed weight, days to flower, plant height, days to maturity, seeds per pod, biomass yield, seed yield, and harvest index using a generalized linear model and a population stratification (structure) correction based on principal component (3) analysis (PCA) (Price *et al.*, 2006). The significance of associations between SNPs and traits was based on the threshold $P < 1.57 \times 10^{-4}$, a modified Bonferroni correction calculated by dividing 1 by the total number of SNPs (6373) in the analysis (Li *et al.*, 2016).

Results

The non-reference (*de novo*) pipeline identified 11,225 SNPs in the 176 accessions originating from 51 countries. The SNP dataset from the *de novo* analysis was further filtered allowing for 15% missing data which left 1021 SNPs, i.e. SNP calls were available in 85% of the accessions. Reanalysing the variants using FreeBayes and the lentil reference genome version Lc1.2 (Ramsay *et al.*, 2016; Ogutcen *et al.*, 2018) increased the SNPs identified to

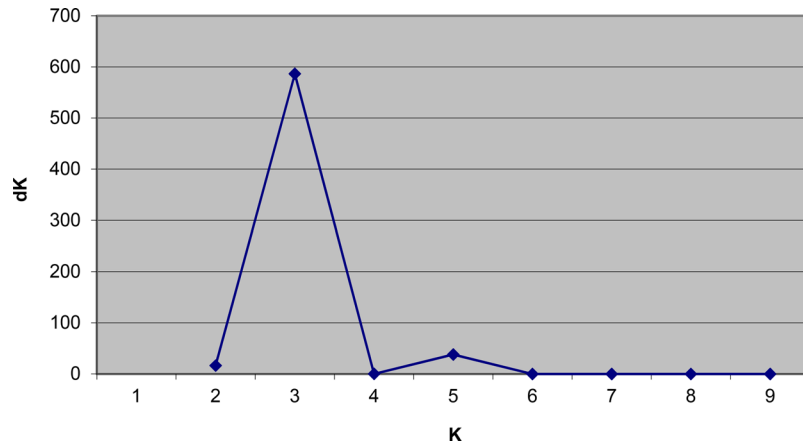


Fig. 1. The subpopulations of $K = 3$ as determined by the *ad hoc* statistic ΔK based on the rate of change in the log probability of data between successive $K = 1-7$ (Evanno *et al.*, 2005).

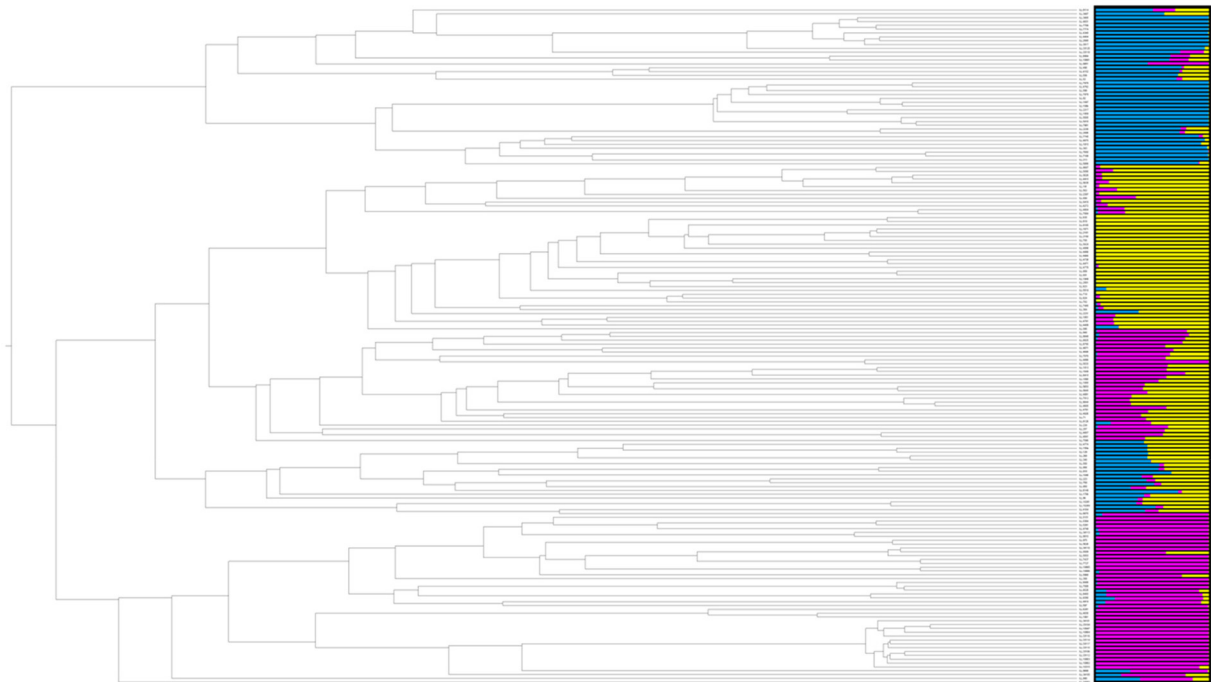


Fig. 2. Dendrogram based on UPGMA and the subpopulations ($K = 3$) calculated using the Bayesian clustering method of the software STRUCTURE based on SNP data for 176 lentil accessions (Pritchard *et al.*, 2000).

22,555. These SNPs were filtered allowing for the same 15% missing data and increased the SNPs sixfold to 6373 versus 1021 *de novo* called variants. Finally, allowing for no missing data, the number of SNPs was 4195 using the reference-based FreeBayes analysis versus zero SNPs for the *de novo* Stack pipeline.

The collection of 176 accessions was analysed for subpopulation structure using two methods: a Bayesian clustering method (online Supplementary Table S3) and PCA based on the SNP genotypes. In the Bayesian approach, first

proposed by Pritchard *et al.* (2000), the Evanno method used to determine the number of subpopulations supports $k = 3$ with far lower support for $k = 5$ (Evanno *et al.*, 2005). Both Bayesian and PCA methods indicated three subpopulations (Figs. 1 and 2). Further, the UPGMA tree is also in agreement with the STRUCTURE software clustering subpopulations (Fig. 3; online Supplementary Figs. S1, S2; Supplementary Table S3). One cluster of the dendrogram contained most of the admixture accessions (Fig. 1). Genetic diversity among and between subpopulations was

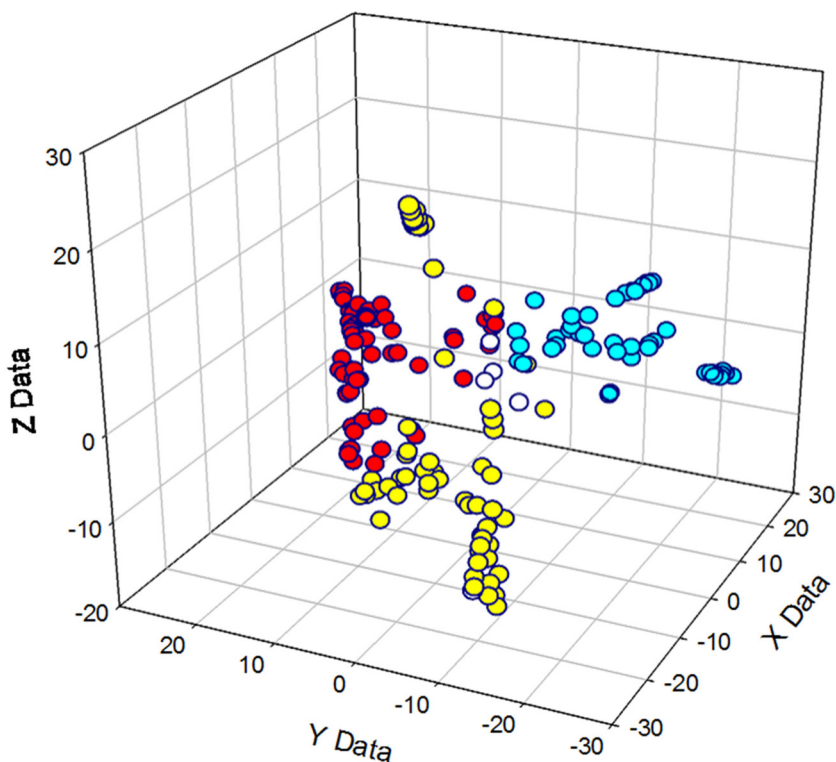


Fig. 3. Principal component analysis of 176 lentil accessions at $K = 3$ based on SNP genotyping. Colours correspond to greater than 50% association with a subpopulation and colours correspond to Fig. 3. White accessions are admixtures. Further views available in online Supplementary Material (Fig. S2).

calculated using AMOVA (Excoffier *et al.*, 1992). The partition of total genetic diversity within the three ($k = 3$) subpopulations was 74% and among the three subpopulations was 26%. Φ_{PT} value was 0.256, $P > 0.001$.

Frequency histograms of the phenotypic data are presented in online Supplementary Fig. S1. The MTAs were identified across five chromosomes for four traits: days to first flower, days to maturity, seeds per pod, and 100-seed weight (Table 1; online Supplementary Fig. S3). The range of variance explained by the MTAs for the four traits ranged from $R^2 = 0.10$ to 0.17. The associations are moderate (Table 1). For days to flower, two MTAs were found on chromosome 3. The most MTAs were found for days to maturity, eight MTAs on chromosomes 2, 3, 5, 6 and 7. Two MTAs were found for seeds per pod on chromosomes 2 and 7. For 100-seed weight, one MTA was identified on chromosome 2.

Discussion

Genotyping of PGRs has spurred discoveries of the genetic control of important agronomic traits in several crops. For example, SNP genotyping of the entire USDA + 18 K soybean collection has been used in 124 research studies (Song *et al.*, 2015). Genotyping-by-sequencing has been

a useful approach for genotyping and analysis of lentil PGRs. Wong *et al.* (2015) were the first to report the use of the two-enzyme GBS approach of Poland *et al.* (2012) in lentil and identified 266,356 genome-wide SNPs before filtering. After filtering to 20% missing SNP data, the final dataset was 32,019 SNPs. In comparison, filtering the ICARDA Lentil Reference Plus SNP set to 20% missing data in the present study resulted in fewer SNPs but still a useful set of 11,171. Recently, a single enzyme GBS of lentil resulted in the discovery of 6693 SNPs (Pavan *et al.*, 2019). While fair genome coverage was gained through these GBS experimental designs, a lentil exome capture resource has now been developed (Ogutcen *et al.*, 2018). This approach will enhance the coverage of the transcribed genes and perhaps improve the success in identifying causal genes and alleles. However, the cost is significantly higher for the lentil exome capture sequencing approach compared to GBS. Therefore, GBS is still a good option for SNP discovery and diversity studies in lentil.

The ICARDA Lentil Reference Plus collection subpopulation ($K = 3$) result is similar to a larger lentil genetic diversity study of 352 accessions by Khazaei *et al.* (2016). However, the populations they reported ($K = 3$) separated into three geographic regions: South Asia, Mediterranean and northern temperate. In contrast, the ICARDA Reference Plus

Table 1. Significant SNP markers identified using genome-wide associations for four traits based on ICARDA's historical phenotypic data collected from 2007 to 2011

Trait	Marker	Chromosome	Base pair position on Lc v1.2	P value	R ²
Days to flower	SLCCHR3_146944147	Chr 3	146944147	5.22 × 10 ⁻⁵	0.11
Days to flower	SLCCHR3_116163111	Chr 3	116163111	1.06 × 10 ⁻⁴	0.12
Days to maturity	SLCCHR2_80372417	Chr 2	80372417	1.70 × 10 ⁻⁵	0.17
Days to maturity	SLCCHR3_74144140	Chr 3	74144140	1.20 × 10 ⁻⁴	0.15
Days to maturity	SLCCHR3_150274854	Chr 3	150274854	1.50 × 10 ⁻⁴	0.15
Days to maturity	SLCCHR5_229889997	Chr 5	229889997	9.58 × 10 ⁻⁵	0.17
Days to maturity	SLCCHR6_51257354	Chr 6	51257354	5.54 × 10 ⁻⁵	0.16
Days to maturity	SLCCHR6_148579031	Chr 6	148579031	1.29 × 10 ⁻⁴	0.12
Days to maturity	SLCCHR7_64647913	Chr 7	64647913	3.92 × 10 ⁻⁵	0.16
Days to maturity	SLCCHR7_236098668	Chr 7	236098668	1.42 × 10 ⁻⁴	0.15
Seeds per pod	SLCCHR2_270863768	Chr 2	270863768	9.44 × 10 ⁻⁵	0.13
Seeds per pod	SLCCHR7_166950176	Chr 7	166950176	1.00 × 10 ⁻⁴	0.10
100 seed weight	SLCCHR2_141782657	Chr 2	141782657	3.73 × 10 ⁻⁵	0.13

collection's subpopulations did not stratify geographically. One aspect to consider is the ICARDA Reference Plus collection includes 20% breeding lines involving parents from different geography and elite lines thus listed as unknown origin in the ICARDA database. Lombardi *et al.* (2014) also reported three subpopulations studying 505 cultivars and landraces but found weaker geographic clustering outside of a breeding program cluster. Idrissi *et al.* (2018) reported subpopulations of two studying 74 Mediterranean lentil landraces stratifying geographically detecting a northern gene pool composed of Turkish, Italian and Greek landraces, and a southern gene pool composed of Moroccan landraces. A larger study of the genetic diversity of the Mediterranean *ex situ* lentil collection ($n = 349$) held at the Italian National Research Council also partitioned into two subpopulations with lower support for three and five subpopulations using the Evanno method (Pavan *et al.*, 2019).

Two other diversity studies have been published on the ICARDA lentil collection using genetic markers. The work of Hamwiah *et al.* (2009) detected two clusters in the ICARDA core collection using 26 microsatellites and UPGMA and PCA analyses. Their collection of 109 accessions (52% wild *Lens* spp.) overlaps our set by only five accessions. An earlier study of an ICARDA set (308 accessions) that included the *Lens* wild relatives (175 accessions) used 22 expressed sequence tags (ESTs) and found eight subpopulations ($K = 8$) using the same Evanno method to select the best fitting model (Alo *et al.*, 2011). *Lens culinaris* ssp. *culinaris* accessions clustered into two subpopulations, the other six subpopulations were single taxon wild accessions. The Alo *et al.* (2011) ICARDA set overlaps our study by nine accessions.

For GWAS, precise phenotypic quantitative trait values are required. The data available for our test were the historical data from ICARDA collected over 5 years, not an experiment *per se*. Further, quantitative trait data, particularly days to flower, days to maturity and seed weight are known to have a genotype by environment interaction (Abbo *et al.*, 1992; Singh *et al.*, 2009; Kahriman *et al.*, 2015). Also, our GWAS experiment might have been affected by the use of phenotypic means from different years. Nonetheless, interesting single-SNP defined regions were identified for four traits: days to first flower, days to maturity, seeds per pod and seed weight. These results will be useful for future meta-analyses as more lentil GWAS studies are published for agronomic traits. Days to first flower was highly significantly correlated with days to maturity and the chromosome 3 MTAs for these two traits are in the same region. One hundred seed weight was significantly negatively correlated with seeds per pod and had no correlation to days to maturity. Comparison with earlier QTL studies is limited by the lack previously of a lentil consensus linkage map so numbering of linkage groups was not consistent. An early genetic study using isozymes, RAPDs and RFLPs, found four linkage groups with factors controlling seed weight in three wide crosses with contrasting seed sizes (*L. culinaris* × *L. orientalis*) (Abbo *et al.*, 1992). A major QTL for seed weight was identified using SSRs explaining 48.4% of the variance (Verma *et al.*, 2015). Three QTL for seed weight were reported on two linkage groups with one QTL explaining 34–50% of the variance (Jha *et al.*, 2017). Our GWAS MTA on chromosome 2 explains less of the variance (12.9%) than these other studies. However, another lentil GWAS study recently found one

significant MTA lentil seed weight using SSRs which explained a similar amount of the variance (Singh *et al.*, 2019).

Days to flowering has been an important selection criterion for lentil breeders (Erskine *et al.*, 1990). Sarker *et al.* (1999) originally reported and named a lentil early flowering gene (*sn*). Weller *et al.* (2012) establish the importance of the *Hr* locus (orthologue of *Early Flowering 3*) on photoperiod response of flowering in lentil located on chromosome 3 (Bett, unpublished data). The one significant MTA for days to flowering identified in the ICARDA Reference Plus collection was located also on chromosome 3 in our study. Once the lentil reference genome is published, it can be determined if this MTA maps close to the *Hr* locus. A bi-parental mapping population reported a major flowering time QTL explained 60% of the variance (Kahriman *et al.*, 2015). Three other QTL for days to flower have been reported on a bi-parental SNP-based map (Fedoruk *et al.*, 2013). A single major locus for days to flower was recently reported in a wide cross between *L. culinaris* × *L. odemensis* (Polanco *et al.*, 2019). A lentil GWAS study on flowering time in 324 *L. culinaris* lines using 255,714 SNP markers identified three MTAs (two on chromosome 2, one on chromosome 5) using the CDC Redberry Lc1.2 assembly (Ramsay *et al.*, 2016; Neupane, 2019). Flowering time MTAs were also reported for days to flower and days to maturity using GWAS with unmapped SSR and EST markers by Kumar *et al.* (2018a, b).

The data mining of historical genebank phenotypic data for GWAS is relatively new and mostly untested. Nguyen and Norton (2020) recently reviewed this approach for GWAS and genomic selection. Two examples of this wealth of data on barley and wheat were published by Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) (González *et al.*, 2018a; Philipp *et al.*, 2018). González *et al.* (2018b) published a strategy to utilize historical phenotypic data collected during seed regeneration to assemble large mapping populations of accessions to discover the genetic effects. Their proposed strategy is not crop-specific and can be used as a guide for the phenotypic evaluation of basically any collection with quality phenotypic data. Utilizing phenotypic historic data from *ex situ* genebanks was thought to 'elevate them to bio-digital resource centres'. A successful application of the historical data approach was used for a GWAS study confirming the association between fruit colour and the MYB1 locus in apple (Migicovsky *et al.*, 2016).

Incorporating this genotype data of the ICARDA Reference Plus collection into genebank databases will bring the world's plant science research community closer to 'unlocking' genetic diversity within these collections (Tanksley and McCouch, 1997). Linking the genotypic data to *ex situ* PGR accessions has been limited based on current genebank database software schema (Postman

et al., 2010; van Treuren and van Hintum, 2014). Finkers *et al.* (2015) proposed using semantic web technology. A USDA effort (www.breedinginsight.org) was undertaken to link genomic data directly to GRIN Global databases. The most advanced effort to link genotypes to PGR accessions has been developed, the database 'Germinate' (Raubach *et al.*, 2021). This version integrates both the phenotypic and genotypic data with the PGR accession. Currently, this lentil accession genotypic data is available for download from the PulseDB database and in the future will be linked to the ICARDA genebank database. Seed of the ICARDA Reference Plus collection is available for requestors directly from ICARDA (<https://www.icarda.org/>).

Data availability

The lentil SNP data set in vcf format file as well as corresponding FASTA sequences are available for downloading on the Pulse Crops Database (<https://www.pulsedb.org/>). All raw sequence data are available through the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with BioProject number: PRJNA639210 (<http://www.ncbi.nlm.nih.gov/bioproject/639210>).

Supplementary material

The supplementary material for this article can be found at <https://doi.org/10.1017/S147926212100006X>

Acknowledgements

The authors thank the USAID for a linkage grant (KR, SK, RJM, CJC), CRP-Grain Legumes (KR, SK) and the Northern Pulse Growers Association (DM, RJM, CJC) and for funding and support from USDA ARS Project Nos. 5348-21000-017-00D (CJC), #5348-21000-024-00D (RJM). The authors further thank the 'Lentil Genome Sequencing (LenGen) Project' and its Project Leaders (KE Bett and DR Cook), and the researcher(s) responsible for generating the data. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

References

- Abbo S, Ladizinsky G and Weeden NF (1992) Genetic analysis and linkage study of seed weight in lentil. *Euphytica* 58: 259–266.
- Alo F, Furman BJ, Akhunov E, Dvorak J and Gepts P (2011) Leveraging genomic resources of model species for the assessment of diversity and phylogeny in wild and domesticated lentil. *Journal of Heredity* 102: 315–329.
- Annicchiarico P, Nazzicari N, Pecetti L, Romani M, Ferrari B, Wei Y and Brummer EC (2017) GBS-based genomic selection for

- pea grain yield under severe terminal drought. *The Plant Genome* 10: 2.
- Arumuganathan K and Earle ED (1991) Nuclear DNA content of some important plant species. *Plant Molecular Biology* 9: 208–218.
- Barnett DW, Garrison EK, Quinlan AR, Strömberg MP and Marth GT (2011) Bamtools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics (Oxford, England)* 27: 1691–1692.
- Bett KE (2020) Lentil Genome Sequencing (LenGen) Project. Available at <https://knowpulse.usask.ca/study/2653517> (accessed 17 June 2020).
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y and Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)* 23: 2633–2635.
- Catchen JM, Amores A, Hohenlohe P, Cresko W and Postlethwait JH (2011) Stacks: building and genotyping loci *de novo* from short-read sequences. *G3: Genes, Genomes, Genetics* 1: 171–182.
- Catchen J, Hohenlohe PA, Bassham S, Amores A and Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* 22: 3124–3140.
- Earl DA and von Holdt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4: 359–361.
- Erskine W, Ellis RH, Summerfield RJ, Roberts EH and Hussain A (1990) Characterization of responses to temperature and photoperiod for time to flowering in a world lentil collection. *Theoretical and Applied Genetics* 80: 193–199.
- Evanno G, Regnaut S and Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14: 2611–2620.
- Excoffier L, Smouse PE and Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131: 479–491.
- FAOSTAT (2017) Available at: <http://www.fao.org/faostat/en/#data/QC/visualize> (accessed 20 September 2017).
- Fedoruk MJ, Vandenberg A and Bett KE (2013) Quantitative trait loci analysis of seed quality characteristics in lentil using single nucleotide polymorphism markers. *The Plant Genome* 6. doi: 10.3835/plantgenome2013.05.0012.
- Finkers R, Chibon PY, van Treuren R, Visser R and Hintum TV (2015) Genebanks and genomics: how to interconnect data from both communities? *Plant Genetic Resources* 13: 90–93.
- Furman BJ (2006) Methodology to establish a composite collection: case study in lentil. *Plant Genetic Resources* 4: 2–12.
- Garrison E (2012) FreeBayes source repository. Available at <https://github.com/ekg/freebayes>.
- Garrison E and Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv: 1207.3907*.
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q and Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9: e90346.
- González MY, Weise S, Zhao Y, Philipp N, Arend D, Börner A, Oppermann M, Graner A, Reif JC and Schulthess AW (2018a) Unbalanced historical phenotypic data from seed regeneration of a barley *ex situ* collection. *Scientific Data* 5: 1–10.
- González MY, Philipp N, Schulthess AW, Weise S, Zhao Y, Börner A, Oppermann M, Graner A and Reif JC (2018b) Unlocking historical phenotypic data from an *ex situ* collection to enhance the informed utilization of genetic resources of barley (*Hordeum* sp.). *Theoretical and Applied Genetics* 131: 2009–2019.
- Guindon MF, Martin E, Cravero V, Gali KK, Warkentin TD and Cointry E (2019) Linkage map development by GBS, SSR, and SRAP techniques and yield-related QTLs in pea. *Molecular Breeding* 39: 54.
- Hamwieh A, Udupa SM, Sarker A, Jung C and Baum M (2009) Development of new microsatellite markers and their application in the analysis of genetic diversity in lentils. *Breed Science* 59: 77–86.
- Heisey P and Rubenstein DR (2015) Using crop genetic resources to help agriculture adapt to climate change: economics and policy. *USDA-ERS Economic Information Bulletin* 139: 1–23.
- Idrissi O, Piergiovanni A, Toklu F, Houasli C, Udupa S, De Keyser E, Van Damme P and De Riek, J (2018) Molecular variance and population structure of lentil (*Lens culinaris* Medik.) landraces from Mediterranean countries as revealed by simple sequence repeat DNA markers: implications for conservation and use. *Plant Genetic Resources* 16: 249–259.
- Jain M, Misra G, Patel RK, Priya P, Jhanwar S, Khan AW, Shah N, Singh VK, Garg R, Jeena G and Yadav M (2013) A draft genome sequence of the pulse crop chickpea (*Cicer arietinum* L.). *The Plant Journal* 74: 715–729.
- Jha R, Bohra A, Jha UC, Rana M, Chahota RK, Kumar S and Sharma TR (2017) Analysis of an intraspecific RIL population uncovers genomic segments harbouring multiple QTL for seed relevant traits in lentil (*Lens culinaris* L.). *Physiology and Molecular Biology of Plants* 23: 675–684.
- Kahriman A, Temel HY, Aydogan A and Tanyolac MB (2015) Major quantitative trait loci for flowering time in lentil. *Turkish Journal of Agriculture and Forestry* 39: 588–595.
- Khazaei H, Caron CT, Fedoruk M, Diapari M, Vandenberg A, Coyne CJ, McGee R and Bett KE (2016) Genetic diversity of cultivated lentil (*Lens culinaris* Medik.) and its relation to the world's agro-ecological zones. *Frontiers in Plant Science* 7: 1093.
- Khazaei H, Fedoruk M, Caron CT, Vandenberg A and Bett KE, (2017a) Single nucleotide polymorphism markers associated with seed quality characteristics of cultivated lentil. *The Plant Genome* 11: 170051.
- Khazaei H, Podder R, Caron CT, Kundu SS, Diapari M, Vandenberg A and Bett KE (2017b) Marker-trait association analysis of iron and zinc concentration in lentil (*Lens culinaris* Medik.) seeds. *The Plant Genome* 10. doi: 10.3835/plantgenome2017.02.0007.
- Kreplak J, Madoui M-A, Cápál P, Novák P, Labadie K, Aubert G, Bayer P, Kishore KG, Symes RA, Main D, Klein A, Bérard A, Fukova I, Fournier C, d'Agata L, Belsler C, Berrabah W, Šimková H, Lee HT, Kougbeadjo A, Térézol M, Huneau C, Turo CJ, Mohellibi N, Neumann P, Falque M, Gallardo-Guerrero K, McGee R, Tar'an B, Bendahmane A, Aury J-M, Batley J, Le Paslier MC, Ellis THN, Warkentin T, Coyne CJ, Salse J, Edwards D, Lichtenzweig J, Macas J, Doležel J, Wincker P and Burstin J (2019) A reference genome for pea provides insight into legume evolution. *Nature Genetics* 51: 1411–1422.
- Kumar S, Rajendran K, Kumar J, Hamwieh A and Baum M (2015) Current knowledge in lentil genomics and its application for crop improvement. *Frontiers in Plant Science* 6: 78.
- Kumar J, Gupta S, Biradar RS, Gupta P, Dube S and Singh NP (2018a) Association of functional markers with flowering time in lentil. *Journal of Applied Genetics* 59: 9–21.

- Kumar J, Gupta S, Gupta DS and Singh NP (2018b) Identification of QTLs for agronomic traits using association mapping in lentil. *Euphytica* 214: 75.
- Li F, Chen B, Xu K, Gao G, Yan G, Qiao J, Li J, Li H, Li L, Xiao X and Zhang T (2016) A genome-wide association study of plant height and primary branch number in rapeseed (*Brassica napus*). *Plant Science* 242: 169–177.
- Lischer HEL and Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics (Oxford, England)* 28: 298–299.
- Lombardi M, Materne M, Cogan NOI, Rodda M, Daetwyler HD, Slater AT, Forster JW and Kaur S (2014) Assessment of genetic variation within a global collection of lentil (*Lens culinaris* Medik.) cultivars and landraces using SNP markers. *BMC Genetics* 15: 150.
- Lonardi S, Muñoz-Amatrián M, Liang Q, Shu S, Wanamaker SI, Lo, S, Tanskanen J, Schulman AH, Zhu T, Luo MC, Alhakami H, Ounit R, Hasan AM, Verdier J, Roberts PA, Santos JPR, Ndeve A, Doležel J, Vrána J, Hokin SA, Farmer AD, Cannon SB and Close TJ (2019) The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *The Plant Journal* 98: 767–782.
- Ma Y, Marzougui A, Coyne CJ, Sankaran S, Main D, Porter LD, Mugabe D, Smitchger JA, Zhang C, Amin M and Rasheed N (2020) Dissecting the genetic architecture of Aphanomyces root rot resistance in lentil by QTL mapping and Genome-Wide Association Study. *International Journal of Molecular Sciences* 21: 2129.
- Mascher M, Schreiber M, Scholz U, Graner A, Reif JC and Stein N (2019) Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. *Nature Genetics* 51: 1076–1081.
- McCouch SR, McNally KL, Wang W and Hamilton RS (2012) Genomics of gene banks: a case study in rice. *American Journal of Botany* 99: 407–423.
- McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G, Dempewolf H, Dingkuhn M, Feuillet C, Gepts P, Grattapaglia D, Guarino L, Jackson S, Knapp S, Langridge P, Lawton-Rauh A, Lijua Q, Lusty C, Michael T, Myles S, Naito K, Nelson RL, Pontarollo R, Richards CM, Rieseberg L, Ross-Ibarra J, Rounsley S, Hamilton RS, Schurr U, Stein N, Tomooka N, van der Knaap E, van Tassel D, Toll J, Valls J, Varshney RK, Ward J, Waugh R, Wenzl P and Zamir D (2013) Agriculture: feeding the future. *Nature* 499: 23–24.
- Migicovsky Z, Gardner KM, Money D, Sawler J, Bloom JS, Moffett P, Chao CT, Schwaninger H, Fazio G, Zhong GY and Myles S (2016) Genome to phenome mapping in apple using historical data. *The Plant Genome* 9: 1–15.
- Neupane S (2019) Flowering time response of diverse lentil (*Lens culinaris* Medik.) germplasm grown in multiple environments. MS Thesis, University of Saskatchewan.
- Nguyen GN and Norton SL (2020) Genebank phenomics: a strategic approach to enhance value and utilization of crop germplasm. *Plants* 9: 817.
- Ogutten E, Ramsey L, von Wettberg EJB and Bett K (2018) Capturing variation in *Lens*: development and utility of an exome capture array for lentil. *Applications in Plant Science* 6: e01165.
- O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM and Portnoy DS (2018) These aren't the loci you're looking for: principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* 27: 3193–3206.
- Pavan S, Bardaro N, Fanelli V, Marcotrigiano AR, Mangini G, Taranto F, Catalano D, Montemurro C, De Giovanni C, Lotti C and Ricciardi L (2019) Genotyping by sequencing of cultivated lentil (*Lens culinaris* Medik.) highlights population structure in the Mediterranean gene pool associated with geographic patterns and phenotypic variables. *Frontiers in Genetics* 10: 872.
- Peakall R and Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6: 288–295.
- Peakall R and Smouse PE (2012) Genalex 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. *Bioinformatics (Oxford, England)* 28: 2537–2539.
- Philipp N, Weise S, Oppermann M, Börner A, Graner A, Keilwagen J, Kilian B, Zhao Y, Reif JC and Schulthess AW (2018) Leveraging the use of historical data gathered during seed regeneration of an *ex situ* genebank collection of wheat. *Frontiers in Plant Science* 9: 609.
- Polanco C, de Miera LES, González AI, García P, Fratini R, Vaquero F, Vences FJ and de la Vega MP (2019) Construction of a high-density interspecific (*Lens culinaris* x *L. odemensis*) genetic map based on functional markers for mapping morphological and agronomical traits, and QTLs affecting resistance to *Ascochyta* in lentil. *PLoS ONE* 14: e0214409.
- Poland JA (2015) Breeding-assisted genomics. *Current Opinion in Plant Biology* 24: 119–124.
- Poland JA, Brown PJ, Sorrells ME and Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7: e32253.
- Postman J, Hummer K, Bretting P, Kinard G, Bohning M, Emberland G, Sinnott Q, Mackay M, Cyr P, Millard M, Gardner C, Weaver B, Ayala-Silva T, Franko T, Mackay M and Guarino L (2010) GRIN-Global: An international project to develop a global plant genebank information management system. *Acta Horticulturae* 859: 49–56.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904–909.
- Pritchard JK, Stephens M and Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
- Ramsay L, Chan C, Sharpe AG, Cook DR, Penmetsa RV, Chang P, Coyne C, McGee R, Main D, Edwards D, Kaur S, Vandenberg A and Bett KE (2016) *Lens culinaris* CDC Redberry genome assembly v1.2. Retrieved from <https://knowpulse.usask.ca/genome-assembly/Lc1.2>.
- Raubach S, Kilian B, Dreher K, Amri A, Bassi FM, Boukar O, Cook D, Cruickshank A, Fatokun C, El Haddad N, Humphries A, Jordan D, Kehel Z, Kumar S, Laborasa SJ, Loi NH, Mace E, McCouch S, McNally K, Marshall DF, Mikwa EO, Milne I, Odeny DA, Plazas M, Prohens J, Rieseberg LH, Schafleitner R, Sharma S, Stephen G, Tin HQ, Abou Togola A, Emily Warchefsky E, Peter Werner P and Shaw PD (2021) From bits to bites: advancement of the germinate platform to support genetic resources collections and pre-breeding informatics for crop wild relatives. *Crop Science* 62: 1–29.
- Reda A (2015) Lentil (*Lens culinaris* Medik.) current status and future prospect of production in Ethiopia. *Advances in Plants & Agriculture Research* 2: 00040.
- Rohlf FJ (2009) *Numeric Taxonomy and Multivariate Analysis System (NTSYSpc)*, version 2.21c. Setauket, New York: Exeter Software.
- Sarker A, Erskine W, Sharma B and Tyagi MC (1999) Inheritance and linkage relationship of days to flower and morphological

- loci in lentil (*Lens culinaris* Medikus subsp. *culinaris*). *Journal of Heredity* 90: 270–275.
- Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, Torres-Torres M, Geffroy V, Moghaddam SM, Gao D, Abernathy B, Barry K, Blair M, Brick MA, Chovatia M, Gepts P, Goodstein DM, Gonzales M, Hellsten U, Hyten DL, Jia G, Kelly JD, Kudrna D, Lee R, Richard MMS, Miklas PN, Osorno JM, Rodrigues J, Thareau V, Urrea CA, Wang M, Yu Y, Zhang M, Wing RA, Cregan PB, Rokhsar DS and Jackson SA (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics* 46: 707–713.
- Schneider S, Roessli D and Excoffier L (2000) *Arlequin Ver 2000: A Software for Population Genetics Data Analysis*. Geneva, Switzerland: Genetics and Biometry Laboratory, University of Geneva.
- Singh S, Singh I, Gil RK, Kumar S and Sarker A (2009) Genetic studies for yield and component characters in large seeded exotic lines of lentil. *Journal of Food Legumes* 22: 229–232.
- Singh A, Dikshit HK, Mishra GP, Aski M and Kumar S (2019) Association mapping for grain diameter and weight in lentil using SSR markers. *Plant Gene* 20: 100204.
- Sneath PHA and Sokal RR (1973) *Numerical Taxonomy*. San Francisco: W. H. Freeman and Company.
- Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL and Cregan PB (2015) Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes, Genomes, Genetics* 5: 1999–2006.
- Tanksley SD and McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science (New York, N.Y.)* 277: 1063–1066.
- Temel HY, Göl D, Akkale HBK, Kahriman A and Tanyolac MB (2015) Single nucleotide polymorphism discovery through Illumina-based transcriptome sequencing and mapping in lentil. *Turkish Journal of Agriculture and Forestry* 39: 470–488.
- van Treuren R and van Hintum TJ (2014) Next-generation genebanking: plant genetic resources management and utilization in the sequencing era. *Plant Genetic Resources* 12: 298–307.
- Varshney RK (2016) Exciting journey of 10 years from genomes to fields and markets: some success stories of genomics-assisted breeding in chickpea, pigeonpea and groundnut. *Plant Science* 242: 98–107.
- Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM and Farmer AD (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 30: 83.
- Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B and Millan T (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology* 31: 240–246.
- Varshney RK, Singh VK, Hickey JM, Xun X, Marshall DF, Wang J, Edwards D and Ribaut JM (2015) Analytical and decision support tools for genomics-assisted breeding. *Trends in Plant Science* 21: 354–363.
- Varshney RK, Sinha P, Singh VK, Kumar A, Zhang Q and Bennetzen JL (2020) 5Gs for crop genetic improvement. *Current Opinion in Plant Biology* 56: 190–196.
- Verma P, Goyal R, Chahota RK, Sharma TR, Abidin MZ and Bhatia S (2015) Construction of a genetic linkage map and identification of QTLs for seed weight and seed size traits in lentil (*Lens culinaris* Medik.). *PLoS ONE* 10: e0139666.
- Weller JL, Liew LC, Hecht VF, Rajandran V, Laurie RE, Ridge S, Wenden B, Vander Schoor JK, Jaminon O, Blassiau C and Dalmais M (2012) A conserved molecular basis for photoperiod adaptation in two temperate legumes. *Proceedings of the National Academy of Sciences* 109: 21158–21163.
- Wong MML, Verma NG, Ramsay L, Yuan HY, Caron C, Diapari M, Vandenberg A and Bett KE (2015) Classification and characterization of species within the genus *Lens* using genotyping-by-sequencing (GBS). *PLoS ONE* 10: e0122025.
- Yu LX, Zheng P, Bhamidimarri S, Liu XP and Main D (2017) The impact of genotyping-by-sequencing pipelines on SNP discovery and identification of markers associated with verticillium wilt resistance in autotetraploid alfalfa (*Medicago sativa* L.). *Frontiers in Plant Science* 8: 89.
- Zimmerer KS and De Haan S (2017) Agrobiodiversity and a sustainable food future. *Nature Plants* 3: 17047.