ISSN 0006-3509, Biophysics, 2021, Vol. 66, No. 3, pp. 395–400. © Pleiades Publishing, Inc., 2021.
Russian Text © The Author(s), 2021, published in Biofizika, 2021, Vol. 66, No. 3, pp. 466–472.

MOLECULAR BIOPHYSICS

The Search for Agroislands in the Chickpea Genome

A. B. Sokolkova[a], S. V. Bulyntsev[b], P. L. Chang[c], N. Carrasquila-Garcia[c], D. R. Cook[c], E. von Wettberg[d], M. A. Vishnyakova[e], S. V. Nuzhdin[a,f], and M. G. Samsonova[a,*]

[a] *Peter the Great St. Petersburg Polytechnic University, St. Petersburg, 195251 Russia*
[b] *Kuban Experimental Station, Vavilov All-Russian Institute of Plant Genetic Resources, Botanika village, Krasnodar krai, 352183 Russia*
[c] *Department of Plant Pathology, University of California, Davis, CA 95616, United States*
[d] *University of Vermont, Burlington, Vermont, VT 05405, United States*
[e] *Vavilov All-Russian Institute of Plant Genetic Resources, St. Petersburg, 190000 Russia*
[f] *University of Southern California, Los Angeles, CA 90089, United States*
**e-mail: m.samsonova@spbstu.ru*

Received December 7, 2020; revised December 7, 2020; accepted December 25, 2020

Abstract—The concentration of genes that control domestication traits in so-called "agroislands" complicates breeding. Therefore, information about where such islands are located in the genome and with what traits of domestication they are associated is of great importance for further research aimed at improving modern varieties. In the search for such information in the genomes of chickpea landraces, the BayPass program was used to assess the association of single nucleotide polymorphisms with population-specific covariates chosen as complexes of domestication syndrome traits in cultivars of the same breeding purpose. The regions of the genome with which these covariates are associated are located on chromosomes 1, 5, 6, 7, and 8. These regions are potential agroislands, as indirectly confirmed by earlier data.

Keywords: chickpea (*Cicer arietinum* L.), genotyping by sequencing, single nucleotide polymorphisms, genomic analysis of associations, adaptation

DOI: 10.1134/S0006350921030192

Plants were first domesticated ca. 10000 years BP; later, plant domestication was independently repeated in many sites around the world. This process involved the growth and selection of plants that had larger edible parts and were easy to harvest, as food surpluses could be produced with these traits. This set of traits is generally termed *domestication syndrome*. Such sets often overlap in crops grown for similar purposes. As an example, the domestication syndrome in grain legumes includes larger seeds, less pod shattering, shorter dormancy, and altered phenological traits. All these are quantitative traits, controlled by multiple genes.

Thus, even in the nascence of breeding, apparently unintended, the intense selection for certain phenotypes was bound to cause a bottleneck effect, i.e., narrowing of the genetic diversity in populations under domestication owing to the selection of only part of the alleles from the gene pool of wild species [1]. According to current idea, domestication occurred gradually, and domesticated and wild forms could exchange genes for a long period of time [2]. This process caused homogenization and slowed the divergence between domesticated and wild forms promoted by selection [3, 4].

The most probable result of these two oppositely directed processes may be the concentration of genes that control domestication traits on so-called domestication islands, or agroislands, which are similar to speciation islands in mosquitoes [5]. This hypothesis tacitly implies that genes of the domestication syndrome have better chances to be preserved in the population in spite of gene exchange and recombination when they are in linkage disequilibrium and are located in one or few sites in the genomes. The pronounced colocation of quantitative trait loci, affecting several traits in several genome regions, which was found in many cultivated plants, strongly supports this viewpoint [6].

The chickpea, a grain legume used most widely in Asian countries as a source of protein, is no exception. Our studies indicated that some chickpea haplotype blocks are enriched in single nucleotide polymorphisms (SNPs) associated with traits of the domestication syndrome [7]. Although the aggregation of domestication genes in agroislands is the product of breeding, it often hampers further crop improvement,

Abbreviations: SNP—single-nucleotide polymorphism; SMS—Synthetic Morphology Score; BF—Bayes factor.

especially in cases when traits are negatively correlated. Therefore, information on the location of such islands in the genome and their association with certain sets of domestication traits is crucial for futher research aimed at the improvement of modern cultivars.

Information on such regions can be obtained with the BayPass program [8]. It assesses the association of SNPs with population-specific covariates chosen as complexes of domestication traits that serve a certain agricultural purpose. Here, we applied BayPass to the analysis of six chickpea populations.

MATERIALS AND METHODS

Genetic material and genomic data. Genetic material was isolated from landraces of the VIR collection collected from major original regions of chickpea growing, regions of its secondary diversification (Ethiopia, Lebanon, Morocco, Turkey, and India), and wider regions of Central Asia and Mediterranea. The study involved 407 accessions, which were divided into six groups according to their sampling regions: ETHI from Ethiopia, IND from India, LEB from Lebanon, MOR from Morocco, TUR from Turkey, and C_Asia from Central Asia.

The genotyping by sequencing and selection of SNPs used in the study is reported in [7]. The sequencing data are available from the National Center of Biotechnologies database, accession code BioProject PRJNA388691. SNPs were sought with the Genome Analysis Tool Kit (GATK) program [9], and further filtering was done with VCFtools software [10]. A total of 2579 SNPs identified in 407 old local chickpea varieties were chosen for the analysis of associations between phenotypic traits and SNPs.

Analysis of associations between phenotypic traits and SNPs was conducted with BayPass [8]. This software can identify SNPs associated with phenotypic traits at the population level. Sixteen quantitative traits assessed at the Kuban experimental station of VIR in 2016 (Table 1) were analyzed. The phenotyping performed in 2016 is described in [7]. Mean values for each phenotypic trait were calculated separately for each of the six geographic groups. As recommended by the authors of BayPass [8], correlations among phenotypic traits were preliminarily analyzed. Spearman's correlation coefficients were calculated with the *rcorr* function from the Hmisc library in R [11]. According to the results, each of 11 traits was assigned to one of four groups of tightly correlated variables. The mean values of each group were replaced by their Synthetic Morphology Scores (SMSs), which were the first principal component of the data (Table 2). The other five phenotypic variables were not tighly correlated with other variables; hence, their mean values were used in the analysis without transformations. Thus, the BayPass program analyzed associations between SNPs and nine phenotypic trait covariates. The analysis was carried out on the assumption of variable independence. Each SNP was assessed by the importance sampling algorithm [8], which calculated the Bayes factor (BF) value, empirical P value, and corresponding regression coefficients. The significance of associations was determined from the empirical P value (>4), and the association strength between SNPs and the variables, by the factor assessment scale according to Jeffrey's prior [12]: "very convincing proof" at $15 < BF < 20$ and "decisive proof" at $BF > 20$. Significantly associated SNPs were annotated with the Legume information system database [13]. Manhattan plots were constructed in R with the CMplot library [14].

RESULTS

Associations between SNPs and the phenotypic covariates were analyzed with regard to the population structure by using 16 quantitative phenotypic traits. The traits were assessed at the Kuban experimental station of VIR in 2016 (Table 1). To estimate associations at the population level, 407 accessions were divided into six groups according to the geographic location of sampling localities: ETHI from Ethiopia, IND from India, LEB from Lebanon, MOR from Morocco, TUR from Turkey, and C_Asia from Central Asia (Fig. 1a).

Correlation analysis was performed to recognize groups of tightly correlated phenotypic traits. The results are shown in Fig. 1b. Eleven traits were divided into four groups of tightly correlated variables. As recommended by the BayPass authors, nine covariates were constructed for each of the six geographical groups: four SMSs (Tables 1 and 2) and five averaged uncorrelated covariates.

The synthetic morphology score SMS1 was obtained from data on two successive vegetation periods linked by a strong negative correlation: the number of flowering days and the number of days from the end of flowering to the start of ripening (Table 2). The SMS2, SMS3, and SMS4 scores correspond to phenotypic traits that describe the weight parameters of plants, pods, and seeds; pod sizes; pods per plant; and seeds per plant (Table 2).

Thus, associations between SNPs and nine phenotypic trait covariates were analyzed with BayPass [8]. The following thresholds were taken to consider an association significant: empirical Bayesian P value > 4 and, according to Jeffrey's prior [12], $BF > 15$.

The analysis with BayPass [8] detected 14 SNPs significantly associated with phenotypic covariates (Fig. 2, Table 3): seven SNPs on chromosome 1, two on chromosome 2, and one on each of chromosomes 4, 5, 6, 7, and 8. Linkage disequilibrium blocks (haplotype blocks), containing 2579 SNPs, were sought with HaploView [15] in our earlier work [7].

Table 1. The phenotypic traits measured in the Kuban experimental station in 2016 and chosen for association analysis

Phenotypic trait	Unit of measure
Flowering duration	days
Height of the lowest pod attachment	cm
Time from sowing to the start of sprouting	days
Ripening duration	days
Time from sprouting to flowering start	days
Time from the end of flowering to the start of ripening	days
1000 seed weight	g
Pod width	mm
Pod length	mm
Number of seeds per plant	pieces
Number of pods per plant	pieces
Seed weight per plant	g
Pod weight	g
Plant height	cm
Plant weight without pods	g
Whole plant weight with pods	g

Table 2. The synthetic morphology scores of phenotypic traits

Synthetic morphology score (SMS)	Phenotypic traits and correlation coefficients between them
SMS1	Flowering duration–days from the end of flowering to the start of ripening ($r = -0.76$*)
SMS2	1000 seed weight–pod width ($r = 0.74$*); 1000 seed weight–pod length ($r = 0.76$*); Pod width–pod length ($r = 0.86$*)
SMS3	Seeds per plant–pods per plant ($r = 0.85$*); Seeds per plant–seed weight per plant ($r = 0.74$*); Seeds per plant–pod weight ($r = 0.60$*); Pods per plant–seed weight per plant ($r = 0.82$*); Pods per plant–pod weight ($r = 0.77$*); Seed weight per plant–pod weight ($r = 0.88$*)
SMS4	Green weight without pods–whole plant weight with pods ($r = 0.92$*)

* $p < 0.0001$.

Each haplotype block was regarded as a set of SNPs located therein and was used for annotation of significantly associated markers. As well, SNPs were annotated by using the Legume information system database [13]. Two SNPs identified in the analysis fall to the sequences of known genes and three occur in linkage disequilibrium regions (Table 3). Two SNPs, on chromosomes 1 and 5, are significantly associated with two covariates: SMS3 and SMS4. An SNP on chromosome 8 (8: 10314452) is significantly associated with as many as four phenotypic covariates: SMS1, SMS3, SMS4, and the plant height covariate (Table 3). This SNP was identified in our previous genome-wide association study. It was significantly associated with weight parameters of plants in the phenotyping conducted in the Kuban region in 2016 [7] and with flowering time in the Kuban region in 2017 [16]. As well, as reported in [17], the search for SNPs significantly associated with bioclimatic covariates in sampling localities with BayPass [8] revealed a joint association of two bioclimatic variables, including temperature parameters, with the SNP 8: 10314452. This SNP is also mapped at approximately 25 kb from the SNP associated with plant weight in [18].

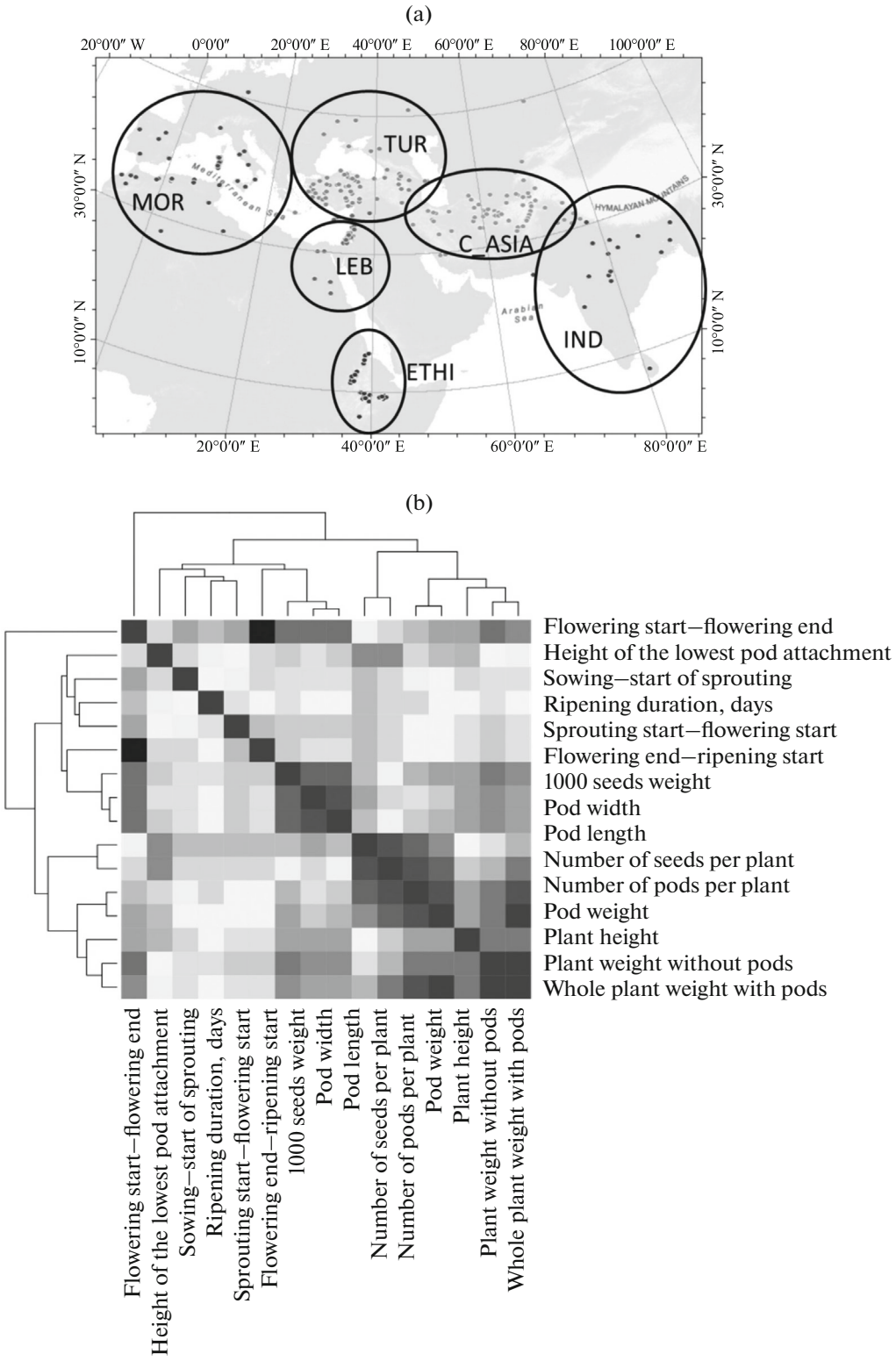

Fig. 1. (a) The geographic origin of accessions and their attribution to one of the six geographic groups. (b) The result of the analysis of phenotypic trait correlations.

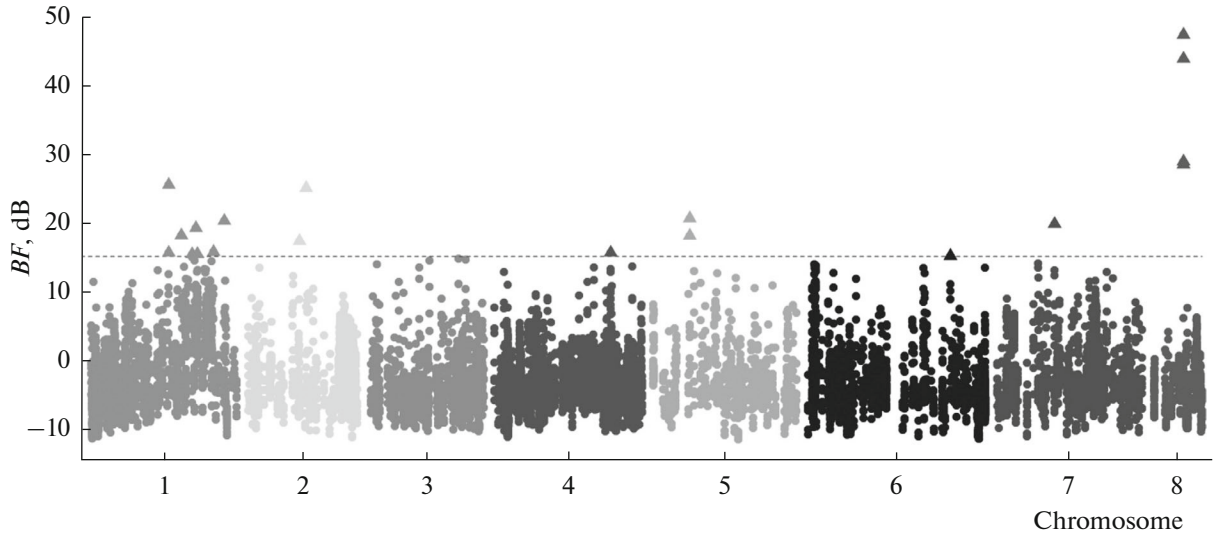

Fig. 2. Association analysis with BayPass. Triangles indicate SNPs with Bayes factor (*BF*) > 15. When one position is associated with a series of phenotypic covariates with different *BF* values, all associations are shown.

Table 3. SNPs detected by the BayPass program [4] in the analysis of associations between SNPs and nine phenotypic trait covariates

Position	Chromosome	Trait	Gene	Linkage disequilibrium block (haplotype block)
25737022	1	SMS3	Ca_18581	–
		SMS4		–
29930356		Mean height of the lowest pod attachment	–	–
33512355		Mean time from sprouting to flowering start	–	–
34754596		SMS4	–	Ca1_Block_36
35261112		SMS4	–	Ca1_Block_37
40608784		SMS4	–	–
44150305		Mean height of the lowest pod attachment	–	–
17432752	2	Mean height of the lowest pod attachment	Ca_16006	–
19604640		Mean height of the lowest pod attachment	–	–
38580554	4	Mean time from sprouting to flowering start	–	Ca4_Block_61
12253018	5	SMS3	–	–
		SMS4	–	–
47657340	6	SMS4	–	–
19377547	7	SMS4	–	–
10314452	8	SMS1	–	–
		SMS3	–	–
		SMS4	–	–
		Mean plant height	–	–

DISCUSSION

The gravitation of genes that control domestication traits to so-called domestication islands, or agroislands, impedes breeding; therefore, information on the location of such islands in the genome and their association with domestication traits is important for further studies aimed at the improvement of modern cultivars. Such information can be obtained with the BayPass program [2], which assesses the associations of SNPs with population-specific covariates. Such covariates can be chosen as complexes of the domestication syndrome traits selected for the same breeding purpose. We applied BayPass to six populations of old local chickpea varieties from the VIR collection. Domestication-related phenotypic traits were divided into five groups of different breeding purposes. They describe phenology, seed weight, seed number, and plant parameters. The genome regions associated with these covariates are mapped on chromosomes 1, 5, 6, 7, and 8. These regions are presumptive agroislands, as proven by the overlap of some of them with haplotype blocks enriched in SNPs. It should be noted that some of the regions also match regions associated with single traits, as found in the genome-wide association studies conducted in 2016 and 2017 [7, 16]. This is an additional argument for their being agroislands.

FUNDING

This work was supported by the Russian Science Foundation, project 16-16-00007.

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interests. The authors declare that they have no conflict of interest.

Statement on the welfare of animals. This article does not contain any studies involving animals performed by any of the authors.

REFERENCES

1. B. L. Gross and K. M. Olsen, Trends Plant Sci. **15**, 529 (2010).
2. T. Pyhajarvi, M. B. Hufford, S. Mezmouk, and J. Ross-Ibarra, Genome Biol. Evol. **5**, 1594 (2013).
3. L. A. F. Frantz, et al., Nat. Genet. **47**, 1141 (2015).
4. S. Abbo and A. Gopher, Trends Plant Sci. **22** (6), 491 (2017). https://doi.org/10.1016/j.tplants.2017.03.010
5. T. L. Turner, M. W. Hahn, and S. V. Nuzhdin, PLoS Biol. **3**, 1572 (2005).
6. R. Bohar, A. Chitkineni, and R. K. Varshney, Biotechniques **69** (3), 158 (2020). https://doi.org/10.2144/btn-2020-0066
7. A. Sokolkova, et al., Int. J. Mol. Sci. **21**, 3952 (2020).
8. M. Gautier, Genetics **201**, 1555 (2015).
9. A. McKenna, M. Hanna, E. Banks, et al., Genome Res. **20**, 1297 (2010).
10. P. Danecek, A. Auton, G. Abecasis, et al., Bioinformatics **27**, 2156 (2011).
11. F. E. Harrell, Jr, Hmisc: Harrell Miscellaneous. R package version 4.1-1 (2018). https://CRAN.R-project.org/package=Hmisc. Accessed October 20, 2020.
12. H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford Univ. Press, Oxford, 1961).
13. S. Dash, J. D. Campbell, E. K. Cannon, et al., Nucleic Acids Res. **44**, D1181 (2016).
14. CMplot: Circle Manhattan Plot. https://github.com/YinLiLin/RCMplot. Accessed November 25, 2020.
15. J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, Bioinformatics **21**, 263 (2005).
16. A. B. Sokolkova, S. V. Bulyntsev, P. L. Chang, et al., Biophysics (Moscow) **66** (1), 32 (2021).
17. A. B. Sokolkova, P. L. Chang, N. Carrasquila-Garcia, et al., Biophysics (Moscow) **65** (2), 237 (2020).
18. R. K. Varshney, M. Thudi, M. Roorkiwal, et al., Nat. Genet. **51**, 857 (2019). https://doi.org/10.1038/s41588-019-0401-3

Translated by Victor Gulevich