DOI: 10.11134/btp.4.2026.3

UDC 575.17:004.9

DEVELOPMENT OF THE CROPPOPSTRUCTURE GENOMIC PLATFORM FOR GENETIC DIVERSITY ANALYSIS AND SUPPORT OF BREEDING DECISIONS

D. Yerbolatov[1], A. Amalova[1], Y. Genievskaya[1], Y.K. Turuspekov[1]

[1] *Institute of Plant Biology and Biotechnology, Almaty, Kazakhstan*
**e-mail: yerlant@yahoo.com*

ABSTRACT

Understanding genetic diversity and population structure is essential for efficiently using crop germplasm collections in breeding, yet a complete analysis requires many separate programs and command-line skills. This study aimed to develop CropPopStructure, a web-based genomic platform that performs a full population genetic analysis of a germplasm collection from a VCF file and a population list, and to validate it on data with a known structure. The platform combines a Nextflow pipeline based on PLINK, VCFtools, and R packages with an R Shiny web interface accessible from a browser. It validates the input data and runs eleven analysis modules: genetic diversity indices, principal component analysis (PCA), discriminant analysis of principal components (DAPC), sNMF ancestry estimation, genomic kinship, a neighbor-joining tree, analysis of molecular variance (AMOVA), pairwise F_{ST}, private SNPs, SNP density, and linkage disequilibrium (LD) decay. HapMap files can be converted to VCF with TASSEL within the platform. The platform was tested on a simulated collection of 98 homozygous lines from four populations genotyped with 35,000 SNPs and generated under a hierarchical coalescent model. PCA separated all four populations (PC1 = 6.32%, PC2 = 1.96% of variance), and the BIC criterion indicated two main genetic groups corresponding to the model's two lineages. AMOVA attributed 21.9 % of molecular variance to differences among populations (Φ_{PT} = 0.219; $p < 0.001$). Pairwise F_{ST} values were 0.037–0.039 between sister populations and 0.124–0.125 between lineages, reproducing the simulated hierarchy. CropPopStructure allows breeders and geneticists without bioinformatics training to obtain reproducible results that support forming core collections, selecting genetically distant parents, and controlling population structure in association studies.

Keywords: population structure; genetic diversity; SNP genotyping; germplasm collection; breeding; Nextflow; Shiny; simulation.

INTRODUCTION

Plant germplasm collections are the main source of starting material for breeding, and their efficient use depends on knowledge of their genetic diversity and structure [1; 2]. Information on how a collection is divided into genetically distinct groups is needed to form core collections, to select parental pairs, and to plan crosses [3]. Unaccounted population structure is also a major source of false-positive associations in genome-wide association studies (GWAS) [4; 5; 6].

Advances in genotyping, such as genotyping-by-sequencing (GBS) and high-density SNP arrays, have made it routine to genotype hundreds of accessions with tens of thousands of markers [7; 8; 9]. Analytical methods have expanded alongside data volume. Model-based methods estimate each accession's ancestry proportions [10; 11; 12]. Multivariate methods, primarily PCA and DAPC, reveal structure without assuming Hardy–Weinberg equilibrium [13; 14]. Analysis of molecular variance (AMOVA) and fixation indices (F_{ST}) quantify population differentiation [15; 16].

Each method is implemented in a separate program with its own input format, including PLINK, VCFtools, adegenet, and TASSEL [17; 18; 19; 20]. A complete analysis of a collection therefore requires format conversion, sequential execution of about a dozen tools, and reconciliation of sample identifiers between them. For breeders and geneticists without programming experience, this is a serious barrier, and manual execution reduces reproducibility. General-purpose web platforms such as Galaxy lower the entry threshold but do not offer a ready end-to-end workflow for the population structure analysis of plant collections [21].

Workflow management systems such as Nextflow and Snakemake describe an analysis as a reproducible chain of steps with automatic parallelization and resumption after failure [22; 23; 24]. Such pipelines follow the FAIR principles, but on their own they still require work on the command line [25]. Combining a workflow system with a web interface makes the same reproducible analysis available to users who work only in a browser. Simulated data are convenient for testing bioinformatic tools: unlike a real collection, the true structure of a simulated population is known in advance, which makes it possible to evaluate how accurately a method recovers it [26].

This work aimed to develop CropPopStructure, a web-based genomic platform that combines complete population structure analysis of crop germplasm collections in a single interface to support breeding decisions, and to validate it on a simulated collection with a known structure.

MATERIALS AND METHODS

Platform architecture

CropPopStructure consists of two layers: a computational pipeline written in Nextflow (DSL2) and a web interface written in R Shiny [22; 27; 28] (Figure 1). The pipeline includes 19 processes: data preparation in PLINK 1.9 and VCFtools, quality control and construction of a unified data object in R, and eleven analysis modules [18; 29]. All programs are installed into a single conda environment from the Bioconda repository [30]. Modules run independently in parallel, and after a failure, the pipeline resumes from the last successfully completed step.

DOI: 10.11134/btp.4.2026.3

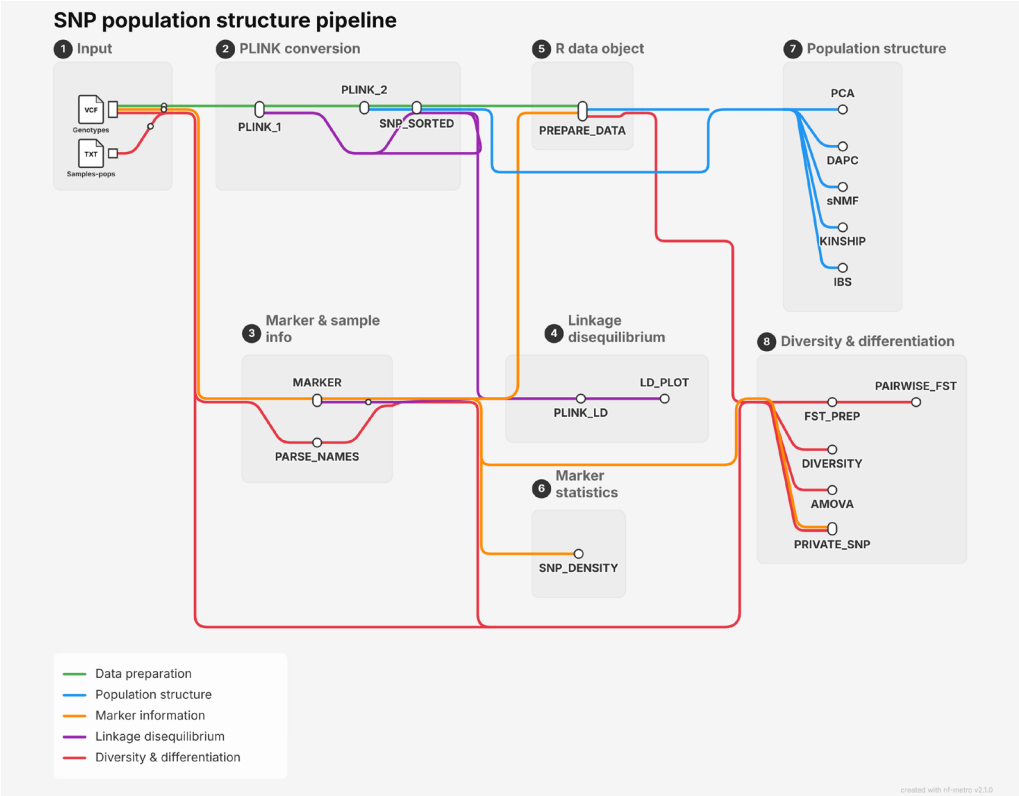


Figure 1. Metro map of the CropPopStructure Nextflow pipeline. Stations are pipeline processes grouped into eight stages; coloured lines show the data flows for data preparation, population structure, marker information, linkage disequilibrium, and diversity and differentiation.

The user uploads a VCF file (uncompressed or vcf.gz) and a two-column text file that assigns each sample to a population. Before the run, the platform checks the input data (Table 1). Each analysis is executed as a background job in a separate folder on the server, so closing the browser does not interrupt the computation. Progress is displayed from the Nextflow trace, and figures and tables appear in the interface as soon as each module finishes. Each analysis receives a permanent link, and you can download all results as a single archive. Figures are saved in TIFF (600 dpi) and PDF formats.

Genotypes in HapMap format can be converted to VCF within the platform using TASSEL 5 [20]. The interface sets analysis parameters and passes them to the scripts through environment variables, so the platform is not tied to a particular crop; it detects chromosome names from the data and sorts them automatically.

Data preparation

The VCF file is converted to PLINK binary format with the --double-id option, which preserves the original sample names. Genotypes are recoded into a numeric 0/1/2 matrix (number of copies of the alternative allele) [17]. Population membership is taken from the user's file and serves as the single source for all modules. At this step, we calculate missing-data rates per sample and per marker, and the number of monomorphic markers.

Analysis modules

Table 2 lists the analysis modules, methods, and their software implementation. All plots are produced with ggplot2 and base R graphics [31].

Simulated collection

To validate the platform, a collection with a known structure was simulated with the coalescent simulator msprime [46; 47]. The model included four populations with a hierarchical history. The ancestral population split into two lineages 1,500 generations ago; 400 generations ago, each lineage split into two sister populations: PopA and PopB, and PopC and PopD. Each population had an effective size of 10,000. Weak symmetric migration at a rate of 5×10^{-5} per generation was al-

Table 1. Input data checks performed before the analysis

Check	Action if violated
VCF contains a #CHROM header line	Run blocked
Every VCF sample is listed in the population file	Run blocked; missing samples listed
No duplicated sample names	Run blocked
SNP identifiers are present and unique	Run blocked; correction command suggested
At least two populations; allowed characters in names	Run blocked
Multi-allelic sites and indels	Warning
Populations with fewer than three samples	Warning

DOI: 10.11134/btp.4.2026.3

Table 2. Analysis modules of the CropPopStructure platform

Module	Method and main parameters	Implementation
Genetic diversity	MAF, H_o, H_e, PIC, Shannon index, F_{IS}, percentage of polymorphic loci; no imputation [32; 33]	custom R functions
PCA	Standardized genotype matrix; missing values replaced by marker mean [13]	R (prcomp)
DAPC	find.clusters (K = 1–20, BIC), k-means on 20 PCs, cross-validation of retained PCs (100 replicates) [14; 34]	adegenet [19; 35]
sNMF	K = 1–15, 10 replicates, α = 100, K chosen by cross-entropy [12]	LEA [36]
Genomic kinship	VanRaden relationship matrix, heatmap	rrBLUP, ComplexHeatmap [37; 38; 39]
NJ tree	1 − IBS distance, neighbor-joining [40]	PLINK, ape [41]
AMOVA	Nei's genetic distance, 999 permutations [15; 42]	StAMPP, pegas [43; 44]
Pairwise F_{ST}	Weighted Weir–Cockerham estimator [16]	VCFtools [18]
Private SNPs	Alternative allele present in only one population	custom R functions
SNP density	Number of markers in 1 Mb windows along chromosomes	custom R functions
LD decay	r^2 for marker pairs within 10 Mb; means and medians in 100 kb bins; threshold $r^2 = 0.2$ [4; 45]	PLINK, R

lowed among all four populations. Mutation and recombination rates were 10^{-8} per base pair per generation. The genome was modeled as seven chromosomes of 30 Mb each (1H–7H, by analogy with barley).

To reproduce the genetics of self-pollinating crops, we modeled each sample as a homozygous line: we simulated a single haplotype and wrote it in diploid form. After simulation, we retained biallelic SNPs with a minor allele frequency of at least 0.05 and randomly sampled 35,000 markers. To mimic real data, 1 % of genotypes were randomly set to missing. The final collection included 98 samples: PopA (25), PopB (25), PopC (24), and PopD (24) (Table 3). The simulated data were analyzed in CropPopStructure with default parameters (K = 3 in DAPC; 999 permutations in AMOVA).

RESULTS

The platform automatically performed the complete analysis of the simulated collection (98 samples × 35,000 SNPs) on the laboratory server, without manual intervention. The results of the modules that directly reflect population structure are presented below.

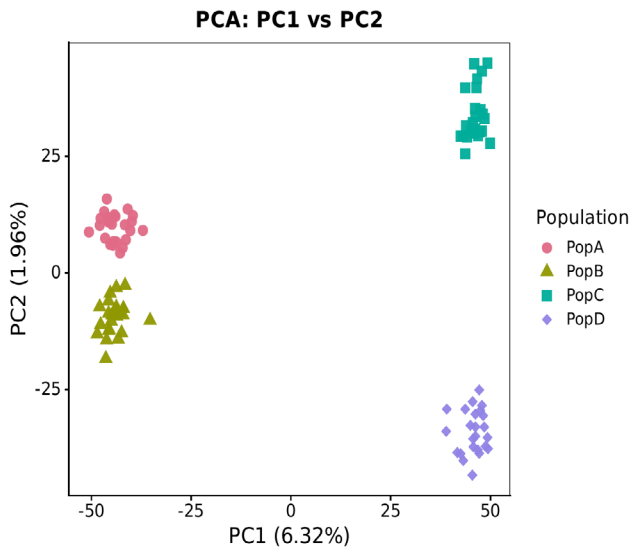


Figure 2. Principal component analysis of the simulated collection (PC1 and PC2). The percentage of explained variance is given in parentheses.

Table 3. Parameters of the simulated collection

Parameter	Value
Number of samples (PopA / PopB / PopC / PopD)	98 (25 / 25 / 24 / 24)
Number of SNPs	35,000
Chromosomes	7 × 30 Mb
Effective population size, N_e	10,000
Split between lineages AB and CD	1,500 generations ago
Split within lineages	400 generations ago
Migration rate	5×10^{-5} per generation
Mutation / recombination rate	10^{-8} / 10^{-8} per bp per generation
Sample type	homozygous lines
Minimum minor allele frequency	0.05
Missing genotypes	1 %

DOI: 10.11134/btp.4.2026.3

Principal component analysis

PCA clearly divided the collection into four non-overlapping groups that matched the simulated populations (Figure 2). The first principal component (6.32% of variance) separated the model's two lineages: PopA and PopB were on the negative side of the axis, and PopC and PopD on the positive side. The second component (1.96 %) separated the sister populations within each lineage. No sample fell into another population's cluster.

Discriminant analysis of principal components

The BIC curve of the find.clusters algorithm reached its minimum at K = 2 (Figure 3). Values at K = 3 and K = 4 exceeded the minimum by only 1–2 units, and at K > 4 the BIC increased monotonically. This agrees with the hierarchical model: the most pronounced level of structure is the two lineages, while differences within lineages are weaker.

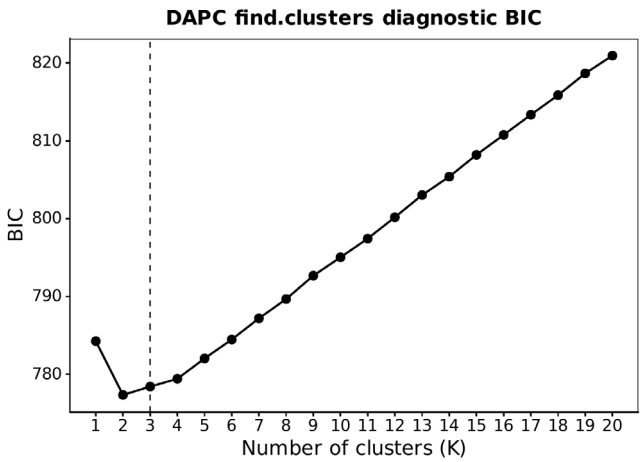


Figure 3. Bayesian information criterion (BIC) values for K = 1–20 clusters. The dashed line marks the K used in the analysis.

With K = 3, DAPC merged PopA and PopB into a single cluster of 50 samples and assigned PopC and PopD to separate clusters of 24 samples each (Figure 4). In the scatter plot, the first discriminant function (LD1) separated the lineages and the second (LD2) separated PopC from PopD. All samples were assigned to their clusters with posterior probabilities close to 1; no admixed samples were detected.

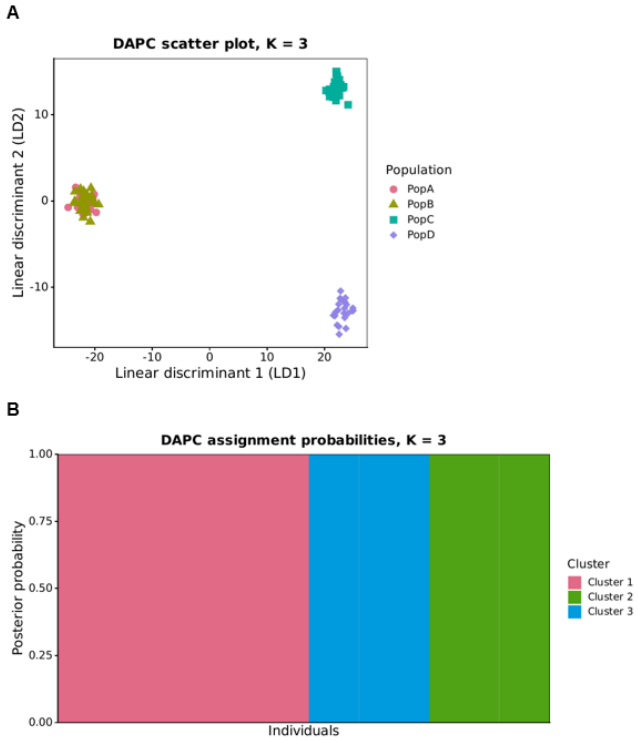


Figure 4. DAPC at K = 3: A – scatter plot of samples on the first two discriminant functions, with colors and shapes indicating the simulated populations; B – posterior probabilities of cluster membership.

Analysis of molecular variance

AMOVA showed that 21.9% of molecular variance was due to differences among populations and 78.1% to variation within populations (Table 4). The differentiation index Φ_{PT} was 0.219, and the differences were highly significant ($p < 0.001$; 999 permutations).

Genetic distances and pairwise F_{ST}

The mean Nei's genetic distance between samples of the same population was 0.318–0.319 (Table 5). It increased to 0.333–0.334 between samples of sister populations and to 0.373–0.374 between samples of different lineages.

Pairwise F_{ST} values (Table 5, Figure 5) followed the same hierarchy. Differentiation between sister populations was moderate (PopA–PopB: 0.037; PopC–PopD: 0.039). Between

Table 4. Analysis of molecular variance (AMOVA)

Source of variation	df	SSD	MSD	Variance component	%
Among populations	3	1.197	0.399	0.0142	21.9
Within populations	94	4.765	0.051	0.0507	78.1
Total	97	5.963	0.061	0.0649	100

$\Phi_{PT} = 0.219$; $p < 0.001$ (999 permutations).

Table 5. Pairwise F_{ST} (above the diagonal) and mean Nei's genetic distances between samples (below and on the diagonal)

	PopA	PopB	PopC	PopD
PopA	0.318	0.037	0.124	0.124
PopB	0.333	0.318	0.125	0.125
PopC	0.373	0.374	0.319	0.039
PopD	0.373	0.373	0.334	0.318

DOI: 10.11134/btp.4.2026.3

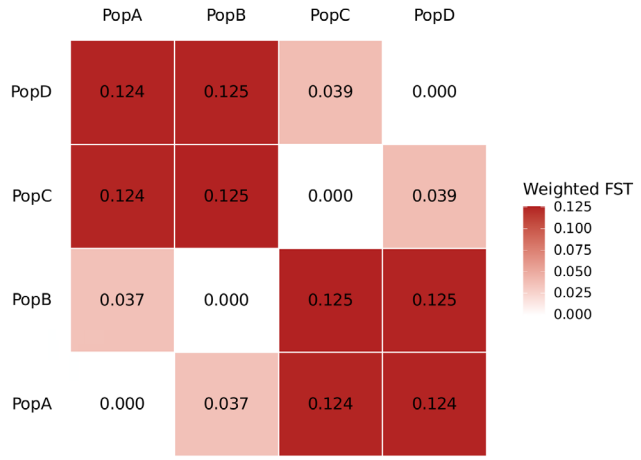


Figure 5. Heatmap of pairwise weighted F_{ST} values among the simulated populations.

populations of different lineages, it was about 3.3 times higher (0.124–0.125). Differences between F_{ST} values within each level did not exceed 0.002, consistent with the symmetric simulation model.

DISCUSSION

CropPopStructure recovered the simulated structure at all levels of the hierarchy. PCA identified all four populations, and the principal components matched the model's order of events: PC1 reflects the older split between lineages, and PC2 the more recent splits within lineages. The low proportion of variance explained by the first components (6.32 and 1.96 %) is typical of data with tens of thousands of markers and does not indicate weak structure [13]. The clear cluster separation in Figure 2 shows that the platform reliably distinguishes even sister populations with F_{ST} of about 0.04.

Pairwise F_{ST} values agreed quantitatively with the model. The older split between lineages produced about three times greater differentiation than the recent splits within lineages. Absolute F_{ST} values depend on the estimator and on the allele frequency spectrum: the Weir–Cockerham estimator accounts for sample size, and the removal of rare alleles (MAF $\geq$ 0.05) shifts the values [16; 48]. Therefore, when comparing collections, the platform reports the calculation method, ensuring uniform implementation [49]. AMOVA (21.9% of variance among populations) and Nei's genetic distances independently confirmed the same pattern [15; 42].

Choosing the number of clusters in DAPC illustrates a general issue with clustering methods. The BIC criterion pointed to K = 2, the upper level of the hierarchical structure, whereas the true number of populations was four. Similar BIC values at K = 2–4 reflect multiple levels of structure rather than a methodological error [14]. The same behavior is known for STRUCTURE and ADMIXTURE: formal criteria for choosing K often indicate the deepest level of subdivision [50; 51]. Clustering results should therefore be interpreted at several values of K, together with PCA, F_{ST} and biological information about the collection [52]. The platform supports this approach: it reports the BIC curve and sNMF cross-entropy for K = 1–15 and allows quick reruns for other K values. Using homozygous lines in the simulation reflects the genetics of self-pollinating crops such as barley, wheat, and chickpea. In such data, observed heterozygosity is near zero, and FIS is near one, which must be considered when interpreting diversity indices [33]. Methods that do not require Hardy–Weinberg equilibrium (PCA, DAPC, sNMF) are particularly suitable for such collections [12; 14]. The platform also supports heterozygous data from cross-pollinating crops.

For breeding programs, the platform's outputs translate directly into decisions. PCA, DAPC, and sNMF-identified groups guide the formation of core collections that capture a gene bank's diversity with a minimal number of accessions [1]. Pairwise F_{ST}, genetic distances, and the kinship matrix help select genetically distant parental pairs and avoid crosses between closely related lines [3; 38]. Principal components and the kinship matrix are standard covariates for controlling population structure and relatedness in GWAS [5; 6], and LD decay determines the marker density required for association mapping [4]. Private SNPs indicate populations that carry unique variation and deserve conservation priority. CropPopStructure's main advantages over using separate programs are reproducibility and accessibility. The analysis chain is defined in Nextflow and all programs are installed from Bioconda, so identical inputs always produce identical results [22; 30]. The web interface removes the need for the command line, and input validation prevents the most common errors: mismatched sample names, empty SNP identifiers, and invalid population names. Unlike general-purpose platforms, CropPopStructure provides a ready end-to-end workflow specifically for the population structure analysis of collections [21]. We validated the platform on a single simulation scenario with relatively strong, symmetric structure; a full evaluation requires scenarios with weak differentiation, unequal sample sizes, admixed individuals, and different proportions of missing data [26; 51]. The memory required for LD calculation grows rapidly with the number of markers. Future work will apply the platform to real crop collections, add containerization for portability between servers, and introduce user access control [53].

CONCLUSIONS

The CropPopStructure web platform combines, in a single browser interface, complete population structure analysis of crop germplasm collections from SNP data, from VCF validation to eleven types of analysis with ready figures and tables. On a simulated collection of 98 homozygous lines from four populations, the platform correctly recovered the simulated hierarchical structure: PCA and pairwise F_{ST} identified all four populations and both levels of differentiation, and AMOVA confirmed significant differences among populations (Φ_{PT} = 0.219; $p < 0.001$). The platform makes population genetic analysis accessible to breeders and geneticists without programming skills, ensures reproducible results, and provides the information needed to form core collections, select parental lines, and control population structure in association studies.

FUNDING

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Program No. BR 24992903).

DOI: 10.11134/btp.4.2026.3

AUTHOR CONTRIBUTIONS

D.Y. – development of the Nextflow pipeline and the Shiny web application, data simulation, analysis of results, writing of the manuscript; A.A. – validation, analysis of the results, manuscript discussion; Y.G. – analysis methodology, development of the R scripts of the analysis modules; Y.T. – validation, supervision, funding acquisition, manuscript writing and discussion. All authors have read and agreed to the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

LITERATURE

1. Glaszmann J.C., Kilian B., Upadhyaya H.D., Varshney R.K. Accessing genetic diversity for crop improvement // Current Opinion in Plant Biology. 2010. Vol. 13, No. 2. P. 167–173.

2. McCouch S., Baute G.J., Bradeen J. et al. Agriculture: Feeding the future // Nature. 2013. Vol. 499. P. 23–24.

3. Varshney R.K., Graner A., Sorrells M.E. Genomics-assisted breeding for crop improvement // Trends in Plant Science. 2005. Vol. 10, No. 12. P. 621–630.

4. Flint-Garcia S.A., Thornsberry J.M., Buckler E.S. Structure of linkage disequilibrium in plants // Annual Review of Plant Biology. 2003. Vol. 54. P. 357–374.

5. Yu J., Pressoir G., Briggs W.H. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness // Nature Genetics. 2006. Vol. 38, No. 2. P. 203–208.

6. Price A.L., Patterson N.J., Plenge R.M. et al. Principal components analysis corrects for stratification in genome-wide association studies // Nature Genetics. 2006. Vol. 38, No. 8. P. 904–909.

7. Elshire R.J., Glaubitz J.C., Sun Q. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species // PLoS ONE. 2011. Vol. 6, No. 5. e19379.

8. Wang S., Wong D., Forrest K. et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array // Plant Biotechnology Journal. 2014. Vol. 12, No. 6. P. 787–796.

9. Comadran J., Kilian B., Russell J. et al. Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley // Nature Genetics. 2012. Vol. 44, No. 12. P. 1388–1392.

10. Pritchard J.K., Stephens M., Donnelly P. Inference of population structure using multilocus genotype data // Genetics. 2000. Vol. 155, No. 2. P. 945–959.

11. Alexander D.H., Novembre J., Lange K. Fast model-based estimation of ancestry in unrelated individuals // Genome Research. 2009. Vol. 19, No. 9. P. 1655–1664.

12. Frichot E., Mathieu F., Trouillon T., Bouchard G., François O. Fast and efficient estimation of individual ancestry coefficients // Genetics. 2014. Vol. 196, No. 4. P. 973–983.

13. Patterson N., Price A.L., Reich D. Population structure and eigenanalysis // PLoS Genetics. 2006. Vol. 2, No. 12. e190.

14. Jombart T., Devillard S., Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations // BMC Genetics. 2010. Vol. 11. P. 94.

15. Excoffier L., Smouse P.E., Quattro J.M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data // Genetics. 1992. Vol. 131, No. 2. P. 479–491.

16. Weir B.S., Cockerham C.C. Estimating F-statistics for the analysis of population structure // Evolution. 1984. Vol. 38, No. 6. P. 1358–1370.

17. Purcell S., Neale B., Todd-Brown K. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses // American Journal of Human Genetics. 2007. Vol. 81, No. 3. P. 559–575.

18. Danecek P., Auton A., Abecasis G. et al. The variant call format and VCFtools // Bioinformatics. 2011. Vol. 27, No. 15. P. 2156–2158.

19. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers // Bioinformatics. 2008. Vol. 24, No. 11. P. 1403–1405.

20. Bradbury P.J., Zhang Z., Kroon D.E. et al. TASSEL: software for association mapping of complex traits in diverse samples // Bioinformatics. 2007. Vol. 23, No. 19. P. 2633–2635.

21. Afgan E., Baker D., Batut B. et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update // Nucleic Acids Research. 2018. Vol. 46, No. W1. P. W537–W544.

22. Di Tommaso P., Chatzou M., Floden E.W. et al. Nextflow enables reproducible computational workflows // Nature Biotechnology. 2017. Vol. 35, No. 4. P. 316–319.

23. Köster J., Rahmann S. Snakemake — a scalable bioinformatics workflow engine // Bioinformatics. 2012. Vol. 28, No. 19. P. 2520–2522.

24. Ewels P.A., Peltzer A., Fillinger S. et al. The nf-core framework for community-curated bioinformatics pipelines // Nature Biotechnology. 2020. Vol. 38, No. 3. P. 276–278.

25. Wilkinson M.D., Dumontier M., Aalbersberg I.J. et al. The FAIR Guiding Principles for scientific data management and stewardship // Scientific Data. 2016. Vol. 3. 160018.

26. Hoban S., Bertorelle G., Gaggiotti O.E. Computer simulations: tools for population and evolutionary genetics // Nature Reviews Genetics. 2012. Vol. 13, No. 2. P. 110–122.

27. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2024. URL: https://www.R-project.org/

28. Chang W., Cheng J., Allaire J.J. et al. shiny: Web Application Framework for R. R package. URL: https://CRAN.R-project.org/package=shiny

29. Chang C.C., Chow C.C., Tellier L.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets // GigaScience. 2015. Vol. 4. P. 7.

30. Grüning B., Dale R., Sjödin A. et al. Bioconda: sustainable and comprehensive software distribution for the life sciences // Nature Methods. 2018. Vol. 15, No. 7. P. 475–476.

DOI: 10.11134/btp.4.2026.3

31. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. New York: Springer, 2016. 260 p.

32. Botstein D., White R.L., Skolnick M., Davis R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms // American Journal of Human Genetics. 1980. Vol. 32, No. 3. P. 314–331.

33. Wright S. The genetical structure of populations // Annals of Eugenics. 1951. Vol. 15, No. 4. P. 323–354.

34. Schwarz G. Estimating the dimension of a model // The Annals of Statistics. 1978. Vol. 6, No. 2. P. 461–464.

35. Jombart T., Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data // Bioinformatics. 2011. Vol. 27, No. 21. P. 3070–3071.

36. Frichot E., François O. LEA: An R package for landscape and ecological association studies // Methods in Ecology and Evolution. 2015. Vol. 6, No. 8. P. 925–929.

37. Endelman J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP // The Plant Genome. 2011. Vol. 4, No. 3. P. 250–255.

38. VanRaden P.M. Efficient methods to compute genomic predictions // Journal of Dairy Science. 2008. Vol. 91, No. 11. P. 4414–4423.

39. Gu Z., Eils R., Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data // Bioinformatics. 2016. Vol. 32, No. 18. P. 2847–2849.

40. Saitou N., Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees // Molecular Biology and Evolution. 1987. Vol. 4, No. 4. P. 406–425.

41. Paradis E., Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R // Bioinformatics. 2019. Vol. 35, No. 3. P. 526–528.

42. Nei M. Genetic distance between populations // The American Naturalist. 1972. Vol. 106, No. 949. P. 283–292.

43. Pembleton L.W., Cogan N.O.I., Forster J.W. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations // Molecular Ecology Resources. 2013. Vol. 13, No. 5. P. 946–952.

44. Paradis E. pegas: an R package for population genetics with an integrated–modular approach // Bioinformatics. 2010. Vol. 26, No. 3. P. 419–420.

45. Hill W.G., Robertson A. Linkage disequilibrium in finite populations // Theoretical and Applied Genetics. 1968. Vol. 38, No. 6. P. 226–231.

46. Kelleher J., Etheridge A.M., McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes // PLoS Computational Biology. 2016. Vol. 12, No. 5. e1004842.

47. Baumdicker F., Bisschop G., Goldstein D. et al. Efficient ancestry and mutation simulation with msprime 1.0 // Genetics. 2022. Vol. 220, No. 3. iyab229.

48. Bhatia G., Patterson N., Sankararaman S., Price A.L. Estimating and interpreting FST: the impact of rare variants // Genome Research. 2013. Vol. 23, No. 9. P. 1514–1521.

49. Meirmans P.G., Hedrick P.W. Assessing population structure: FST and related measures // Molecular Ecology Resources. 2011. Vol. 11, No. 1. P. 5–18.

50. Evanno G., Regnaut S., Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study // Molecular Ecology. 2005. Vol. 14, No. 8. P. 2611–2620.

51. Puechmaille S.J. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem // Molecular Ecology Resources. 2016. Vol. 16, No. 3. P. 608–627.

52. Lawson D.J., van Dorp L., Falush D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots // Nature Communications. 2018. Vol. 9. P. 3258.

53. Kurtzer G.M., Sochat V., Bauer M.W. Singularity: Scientific containers for mobility of compute // PLoS ONE. 2017. Vol. 12, No. 5. e0177459.

DOI: 10.11134/btp.4.2026.3

УДК 575.17:004.9

РАЗРАБОТКА ГЕНОМНОЙ ПЛАТФОРМЫ CROPPOPSTRUCTURE ДЛЯ АНАЛИЗА ГЕНЕТИЧЕСКОГО РАЗНООБРАЗИЯ И ПОДДЕРЖКИ СЕЛЕКЦИОННЫХ РЕШЕНИЙ

Д. Ерболатов[1], А. Амалова[1], Ю. Гениевская[1], Е.К. Туруспеков[1]

[1] Институт биологии и биотехнологии растений, Алматы, Казахстан
**e-mail: yerlant@yahoo.com*

АННОТАЦИЯ

Знание генетического разнообразия и популяционной структуры необходимо для эффективного использования коллекций сельскохозяйственных культур в селекции, однако полный анализ требует запуска многих отдельных программ и навыков работы в командной строке. Цель работы — разработать веб-платформу CropPopStructure, выполняющую полный популяционно-генетический анализ коллекции по VCF-файлу и списку популяций, и проверить её на данных с известной структурой. Платформа объединяет конвейер Nextflow на основе PLINK, VCFtools и R-пакетов с веб-интерфейсом R Shiny. Она проверяет входные данные и выполняет одиннадцать модулей анализа: индексы генетического разнообразия, анализ главных компонент (PCA), дискриминантный анализ главных компонент (DAPC), оценку предковой принадлежности sNMF, геномное родство, NJ-дерево, AMOVA, попарные FST, приватные SNP, плотность SNP и затухание неравновесия по сцеплению. Файлы HapMap конвертируются в VCF средствами TASSEL. Платформа проверена на симулированной коллекции из 98 гомозиготных линий четырёх популяций (35 000 SNP), полученной по иерархической коалесцентной модели. PCA разделил все четыре популяции (PC1 — 6,32 %, PC2 — 1,96 % дисперсии), критерий BIC указал на две основные генетические группы, соответствующие двум ветвям модели. AMOVA отнесла 21,9 % молекулярной изменчивости к различиям между популяциями (ФРТ = 0,219; p < 0,001). Попарные FST составили 0,037–0,039 между сестринскими популяциями и 0,124–0,125 между ветвями, что воспроизводит заданную иерархию. CropPopStructure позволяет селекционерам и генетикам без подготовки в биоинформатике получать воспроизводимые результаты для формирования ядерных коллекций, подбора генетически отдалённых родительских форм и учёта популяционной структуры в ассоциативных исследованиях.

Ключевые слова: популяционная структура; генетическое разнообразие; SNP-генотипирование; коллекция генетических ресурсов; селекция; Nextflow; Shiny; симуляция.

ӘОЖ 575.17:004.9

ГЕНЕТИКАЛЫҚ ӘРТҮРЛІЛІКТІ ТАЛДАУ ЖӘНЕ СЕЛЕКЦИЯЛЫҚ ШЕШІМДЕРДІ ҚОЛДАУ ҮШІН CROPPOPSTRUCTURE ГЕНОМДЫҚ ПЛАТФОРМАСЫН ӘЗІРЛЕУ

Д. Ерболатов[1], А. Амалова[1], Ю. Гениевская[1], Е.К. Туруспеков[1]

[1] Өсімдіктер биологиясы және биотехнологиясы институты, Алматы, Қазақстан
**e-mail: yerlant@yahoo.com*

АҢДАТПА

Ауыл шаруашылығы дақылдары коллекцияларын селекцияда тиімді пайдалану үшін олардың генетикалық әртүрлілігі мен популяциялық құрылымын білу қажет, алайда толық талдау көптеген жеке бағдарламаларды іске қосуды және командалық жолда жұмыс істеу дағдысын талап етеді. Жұмыстың мақсаты — VCF файлы мен популяциялар тізімі бойынша коллекцияның толық популяциялық-генетикалық талдауын орындайтын CropPopStructure веб-платформасын әзірлеу және оны құрылымы белгілі деректерде тексеру. Платформа PLINK, VCFtools және R пакеттеріне негізделген Nextflow конвейерін R Shiny веб-интерфейсімен біріктіреді. Ол кіріс деректерін тексеріп, он бір талдау модулін орындайды: генетикалық әртүрлілік индекстері, негізгі компоненттер талдауы (PCA), негізгі компоненттердің дискриминанттық талдауы (DAPC), sNMF арқылы ата-тектік үлесті бағалау, геномдық туыстық, NJ ағашы, AMOVA, жұптық FST, жеке (private) SNP, SNP тығыздығы және тіркесу теңсіздігінің бәсеңдеуі. HapMap файлдары TASSEL көмегімен VCF форматына түрлендіріледі. Платформа иерархиялық коалесценттік модель бойынша алынған, төрт популяцияның 98 гомозиготалы линиясынан (35 000 SNP) тұратын симуляцияланған коллекцияда тексерілді. PCA төрт популяцияны да ажыратты (PC1 — 6,32 %, PC2 — 1,96 % дисперсия), ал BIC критерийі модельдің екі тармағына сәйкес келетін екі негізгі генетикалық топты көрсетті. AMOVA молекулалық өзгергіштіктің 21,9 %-ын популяциялар арасындағы айырмашылықтарға жатқызды (ФРТ = 0,219; p < 0,001). Жұптық FST мәндері туыс популяциялар арасында 0,037–0,039, ал тармақтар арасында 0,124–0,125 болды, бұл берілген иерархияны дәл қайталайды. CropPopStructure биоинформатикалық дайындығы жоқ селекционерлер мен генетиктерге ядролық коллекцияларды қалыптастыру, генетикалық алшақ ата-аналық формаларды іріктеу және ассоциативтік зерттеулерде популяциялық құрылымды ескеру үшін қайталанатын нәтижелер алуға мүмкіндік береді.

Түйін сөздер: популяциялық құрылым; генетикалық әртүрлілік; SNP-генотиптеу; генетикалық ресурстар коллекциясы; селекция; Nextflow; Shiny; симуляция.