

SNP2STR: A USER FRIENDLY SOFTWARE FOR CREATING INPUT FILES FOR STRUCTURE SOFTWARE FROM SNP DATA*

Almira Konjić^{*1}, Vlad Gheorghe², Fuad Gaši¹, Mirza Uzunović³, Jasmin Grahić¹

Original scientific paper

Abstract

STRUCTURE is a software package that uses molecular data from multiple loci to analyze population structure. To analyze Single Nucleotide Polymorphism (SNP) data using STRUCTURE, a specific input file format is required. However, because SNP studies often involve thousands of loci, these input files are typically too large to edit manually.

To address this challenge, several software tools and functions have been developed. However, in the case of plant data, some of these tools either fail to process the data correctly or generate inaccurate input files.

To overcome these issues, a new software tool - SNP2STR - has been developed. SNP2STR converts SNP data into the input format required by Structure. It generates a file with an optional first row containing marker names, followed by two rows of data per individual, formatted according to the corresponding markers.

The software is open-source and available on GitHub. It also offers potential for future upgrades based on user feedback.

Key words: *molecular data, population structure analysis, genetic clustering, Open-source bioinformatics tools*

INTRODUCTION

In recent decades, plant genetic analysis has undergone a significant transformation by applying molecular markers, catalyzing advancements in molecular breeding in agriculture (Mammadov *et al.*, 2012). Early-generation markers, such as restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), and amplified fragment length polymorphisms (AFLPs), were instrumental in initial genetic studies but were limited by low throughput, poor reproducibility, or complex protocols (Varshney *et al.*, 2007; Mammadov *et al.*, 2012; Mueller *et al.*, 1999). These were followed by second-generation markers like simple sequence repeats (SSRs or microsatellites), which offered improved reproducibility and informativeness

* Paper presented at the 32nd International Scientific-Expert Conference of Agriculture and Food Industry, 1st-2nd December 2022, Sarajevo, Bosnia and Herzegovina

¹ University of Sarajevo - Faculty of Agriculture and Food Sciences, Sarajevo, B&H

Correspondence: a.konjic@ppf.unsa.ba

² Independent Researcher, Berlin, Germany

³ Independent Researcher, Visoko, Bosnia and Herzegovina

(Powell *et al.*, 1996). More recently Single nucleotide polymorphisms (SNPs) have become the preferred choice among these markers due to their abundance, genome-wide distribution, and suitability for high-throughput genotyping. SNP markers enable detailed investigations into genetic diversity, linkage mapping, and marker-assisted selection in various plant species (Ganal *et al.*, 2009). The availability of high-density SNP datasets has also facilitated powerful approaches such as whole-genome scanning (WGS), genome-wide association studies (GWAS), and association genetics, allowing researchers to link genomic variation to phenotypic traits at unprecedented resolution (Zhu *et al.*, 2008).

As SNP genotyping technologies now routinely generate hundreds of thousands of markers, with densities reaching up to 480K per individual, the resulting datasets are large and complex, requiring efficient computational tools for downstream analysis. Several programs have been developed for handling and analyzing such data, including PLINK (Purcell *et al.*, 2007), VCF2PopTree (Subramanian *et al.*, 2019), and R-based tools (R Core Team, 2024). However, population structure analysis, an essential step in assessing genetic subgroups and potential admixture, often relies on programs such as STRUCTURE (Pritchard *et al.*, 2000) and fastSTRUCTURE (Raj *et al.*, 2014). These tools use multi-locus genotype data to infer hidden population structure, estimate ancestry proportions, and assign individuals to genetic clusters (Pritchard *et al.*, 2000, Raj *et al.*, 2014, Falush *et al.*, 2003). This is especially important in studies of plant genetics focused on managing genetic resources, enhancing breeding strategies, and detecting gene flow between cultivars and their wild relatives.

The STRUCTURE software implements a Bayesian model-based clustering method that infers population structure by estimating the likelihood that each individual belongs to one or more of K populations (Pritchard *et al.*, 2000). While robust and widely accepted, STRUCTURE is computationally intensive and becomes impractical for datasets with tens of thousands of SNPs or hundreds of individuals. In response to this limitation, fastSTRUCTURE was developed using a variational Bayesian framework that delivers similar clustering results with significantly reduced runtime. It is more suitable for fast, large-scale SNP data analysis in plants (Raj *et al.*, 2014).

Nonetheless, STRUCTURE remains highly relevant and is frequently used in high-density SNP studies where analytical precision and model flexibility are prioritized, as its complete Bayesian Markov Chain Monte Carlo (MCMC) approach enables a deeper exploration of complex genetic patterns, offering superior performance in detecting subtle population structures, admixture, and scenarios involving small sample sizes, polyploidy, or non-random mating (Porras-Hurtado *et al.*, 2013).

The STRUCTURE software requires specifically formatted input files, which presents a significant practical challenge in modern SNP-based studies. STRUCTURE expects genotype data to be organized in a strict text-based format where two rows and markers representing each individual are consistently encoded. These requirements are difficult to fulfill manually, especially when working with large datasets in PLINK-derived *.ped* and *.map* formats. This preprocessing step often becomes a technical bottleneck in research workflows, particularly when dealing with high-density SNP data.

To address this limitation, we introduce SNP2STR, a user-friendly Python-based tool designed to automatically convert SNP datasets from PLINK's *.ped* and *.map* formats into the input format required by STRUCTURE. SNP2STR efficiently arranges the data into a compatible matrix format, optionally includes marker names in the first row, and ensures that each individual is represented by two correctly formatted rows corresponding to allele calls. By automating this conversion process, SNP2STR removes a significant barrier to the effective use of STRUCTURE, facilitating its application in plant population genetics even when working with extensive SNP datasets.

This paper presents the rationale behind SNP2STR, describes its core functionalities, and demonstrates its utility in enabling efficient and accurate population structure analysis with STRUCTURE in the context of high-density plant SNP data.

MATERIALS AND METHODS

Data

The data used in this study was previously published by Konjić *et al.* (2023) and consists of traditional apple accessions from the Bosnia and Herzegovina apple core collection genotyped with the Axiom® Apple 480K SNP array (Bianco *et al.*, 2016).

Input files

The software was developed to process *.map* and *.ped* input files (Figure 1). In this study we used Axiom data *.map* and *.ped* and files were generated using Axiom Analysis Suite (Affymetrix, Inc., version 4.0.3) from raw *.CEL* files.

1	#Sample	Filename	Genotypes															
2	MD	Amerikanka.CEL	A	A	T	C	T	T	G	G	A	A	C	C	G	G	T	T
3	MD	Ananas.CEL	C	C	T	C	C	G	G	A	A	T	C	G	G	T	C	T
4	MD	Babovaca.CEL	A	A	T	C	T	T	G	G	G	T	T	A	G	T	T	G
5	MD	Bascenka.CEL	A	A	T	C	T	T	G	G	G	T	T	A	A	T	T	G

A

1	MDC000001.124	AX-115533715	0.000	7112
2	MDC000001.124	AX-115533717	0.000	7810
3	MDC000001.124	AX-115533718	0.000	7922
4	MDC000001.124	AX-115533719	0.000	8069
5	MDC000002.328	AX-115533721	0.000	941

B

Figure 1. (A) Example of a space delimited *.ped* file containing mandatory six columns (Family ID, Individual ID, Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown), Phenotype). The remaining columns contain genotypes (B) Example of a *.map* file where each line describes a single marker and has exactly four columns (chromosome, rs# or snp identifier, genetic distance (morgans), base-pair position (bp units))

Software development:

The SNP2STRtool was developed in Python ver. 3.6 (Van Rossum and Drake, 1995) as a command-line application for converting SNP genotype data from PED format to STRUCTURE-compatible format. The software utilizes the pandas library ver. 2.2.3 (McKinney, 2010) for efficient data manipulation and CSV processing of genomic datasets. The tool is implemented as a Python package using setuptools for distribution, with the core processing logic contained in a single module that handles file parsing, genotype conversion, and output formatting. The software includes comprehensive error handling for common data inconsistencies such as varying sequence lengths across samples and mismatched population data. The package was designed following standard Python packaging conventions with a command-line interface.

Software implementation

SNP2STR is a Python package designed to prepare SNP files for analysis within the STRUCTURE software. The script has one mandatory argument: the path to the *.ped* file that needs to be modified. Moreover, the user can optionally specify a *.map* file and a *.txt* file with the list of population identifiers.

The script produces a single output file. This is a *.txt* file with values separated by a single space. The file header, if present, shows the names of chromosomes. Each variety has two rows of sequences, one for each strand. Bases are encoded to align with the required STRUCTURE format.

The user can choose to omit the header in the output file. Note that if no *.map* file is specified, the header will be omitted by default.

To produce the output, SNP2STR reads the *.ped* file line by line. For each variety, it extracts the base sequence and encodes it into numerical format (ATGC → 1234), missing data is represented by -9 (Pritchard *et al.*, 2009). The software then splits the sequence into two strands by taking alternating values. For example, the first row contains the first, third, and fifth values of the *.ped* sequence. The second row contains the second, fourth, and sixth values.

RESULTS AND DISCUSSION

The generated software package is hosted on the Python Package Index (PyPI) and can be installed with the pip tool (Figure 2.)



Figure 2. Example of how to install SNP2STR software with the pip tool

The software is also available on GitHub [<https://github.com/vlad-ds/snp2str>].

The script is operated through a command line interface (Figure 3)

```
Welcome to snp2str!  
usage: snp2str [-h] [--output OUTPUT] [--skip-header] [files [files ...]]  
  
Convert PED files to create input format for structure.  
  
positional arguments:  
  files          List paths for .ped, .map and pop.txt file. If omitted, will search for default files in current folder.  
  
optional arguments:  
  -h, --help      show this help message and exit  
  --output OUTPUT Path for the output file. Defaults to ./output.csv  
  --skip-header    Skip header row in output file.
```

Figure 3. Software interface

Software can be run in python and also from Windows Powershell in Windows OS and terminal in Linux OS.

Performance

In order to test the software a dataset containing 64 apple accessions from the *ex situ* site in Srebrenik, Bosnia and Herzegovina were used. The accessions have been genotyped with the Axiom® Apple480K SNP array (Bianco *et al.*, 2016). Dataset was available in *.map* and *.ped* format which were then used as input files for SNP2STR software (Figure 4).

```
# Basic usage with a PED file  
snp2str path/to/your/file.ped  
  
# Using all optional files  
snp2str path/to/your/file.ped path/to/your/populations.txt path/to/your/file.map  
  
# Specify output path  
snp2str path/to/your/file.ped --output custom_output.csv  
  
# Skip header in output file  
snp2str path/to/your/file.ped --skip-output-header  
  
# Skip first line in input PED file  
snp2str path/to/your/file.ped --skip-input-header
```

Figure 4. Command line interface when running the software

The output file (Figure 5) is written as follows. First, if present in the file, the header is added, then the rows are written. Each population label gets two rows, one for each strand. The first value of each row is the population label. The second value is a population identifier if specified in the *.txt* file, otherwise, this value is omitted. The subsequent values are the base numbers for each chromosome on that strand.

1	Amerikanka.CEL	2	3	1	4	3	2	3	1	3	1	2	3	4	1	2	3	2	1	4	4	2	1	2	3	2	1	2	1	2	3	1	2	1	2	1	4	4	1				
2	Amerikanka.CEL	2	3	1	4	3	2	3	1	3	1	4	3	4	2	3	3	3	4	4	4	1	2	3	2	3	4	1	4	3	3	2	1	2	3	1	4	3	1				
3	Ananas.CEL	4	3	1	2	3	2	2	1	1	-9	2	1	2	1	2	1	3	1	4	2	4	1	2	2	3	4	1	4	1	3	2	1	2	3	1	4	1	2				
4	Ananas.CEL	4	3	1	4	3	4	2	3	3	-9	2	1	4	1	4	3	3	3	4	4	4	3	2	3	2	3	4	1	4	3	3	3	1	4	3	3	4	1	3			
5	Babovaca.CEL	2	3	3	2	1	2	2	1	3	1	2	1	2	1	2	3	2	1	4	2	4	1	2	2	2	1	4	1	2	1	1	2	3	4	1	3	2	1	3			
6	Babovaca.CEL	2	3	3	2	3	2	3	1	3	1	2	3	4	1	2	3	3	3	4	4	4	3	3	3	2	3	4	1	2	3	1	3	3	4	3	3	4	1	3			
7	Bascenka.CEL	2	3	3	2	1	2	3	1	3	1	2	3	2	1	2	3	3	3	4	2	4	1	2	3	2	1	4	1	2	1	1	2	3	4	3	3	2	1	3			
8	Bascenka.CEL	2	3	3	2	1	2	3	1	3	3	2	3	2	1	4	3	3	3	4	2	4	1	3	3	2	1	4	1	2	3	1	2	3	1	2	3	4	3	3	2	1	3
9	Bedrika.CEL	2	2	1	2	1	2	2	1	3	1	2	1	-9	1	2	3	3	1	4	2	2	1	2	2	2	3	4	1	2	1	1	3	3	4	1	1	2	1	2			
10	Bedrika.CEL	2	3	3	2	3	4	3	1	3	3	2	3	-9	1	2	3	3	3	4	2	4	1	2	2	4	3	4	3	4	3	1	3	3	4	3	3	4	1	3			
11	Bihorka.CEL	2	3	1	2	3	2	3	1	3	1	2	3	2	1	2	1	4	2	4	1	2	3	2	1	2	1	2	1	1	2	3	2	1	1	2	1	2	1	2			
12	Bihorka.CEL	2	3	3	4	3	2	3	3	3	3	4	3	4	3	4	3	2	3	4	2	4	1	2	3	2	3	4	3	2	3	1	2	3	2	3	3	4	3	3			
13	Bistrica.CEL	2	3	1	2	3	2	3	1	-9	1	2	1	4	1	2	3	2	1	4	2	2	1	2	2	2	1	4	1	2	1	1	3	3	2	1	1	4	1	1			
14	Bistrica.CEL	2	3	3	2	3	2	3	1	-9	3	2	3	4	1	2	3	3	3	4	4	4	1	2	3	4	3	4	1	4	3	1	3	3	3	4	3	3	4	1			

Figure 5. Input format for STRUCTURE software after conducting snp2str

To our knowledge conversion to structure format can also be done with PGD spider (Lischer & Excoffier 2012) and Rstudio (R Studio Team 2020). However, in trying to convert the .vcf file to STRUCTURE format in PGD spider, it yielded a table with distorted numbers. The cause of this issue is unclear and may be due to format incompatibility between the .vcf file and PGDSpider's input requirements, a software bug, or errors in the conversion settings or process. As far as Rstudio is concerned, the operation would involve a few different steps and a range of skills in the program. These limitations highlight the need for a stand-alone, scriptable solution that eliminates manual reformatting and conversion errors—precisely the gap that SNP2STR is designed to fill.

Even though STRUCTURE software is widely known and popular when it comes to population structure analysis, it takes too long to analyse large datasets containing SNPs (Raj *et al.*, 2014). Therefore, a new software named fastStructure (Raj *et al.*, 2014) was designed specifically for inferring population structure from SNP data. As such, this software accepts outputs derived in PLINK as inputs, but it also accepts files in the 'classical' STRUCTURE format.

However, STRUCTURE software analysis remains highly relevant and is used in high-density SNP studies where analytical precision and model flexibility are prioritized. This makes it especially valuable in detecting subtle population structures, admixture, and scenarios involving small sample sizes, polyploidy, or non-random mating—cases where a more nuanced analysis is essential (Porras-Hurtado *et al.*, 2013). Additionally, STRUCTURE offers advanced modeling capabilities such as the LOCPRIOR model, correlated allele frequencies, and detailed ancestry estimation, features not fully supported by fastSTRUCTURE. These qualities make STRUCTURE the preferred choice in studies requiring publication-standard resolution or methodological benchmarking, even if it necessitates using a subset of SNPs or longer computational time.

In this context, SNP2STR remains a valuable tool for facilitating data preparation, enabling researchers to efficiently utilize the method best suited to their dataset size and research goals. The main advantage of SNP2STR lies in its straightforward interface

and ability to streamline the conversion of SNP datasets into STRUCTURE-compatible formats, reducing manual effort and potential formatting errors

Overall, the value of the newly developed software lies in its simple user interface, which clearly saves time in the analysis process. The tool has been validated on dataset from Axiom® Apple 480 K SNP array (Bianco *et al.*, 2017), and so far no limitations have been observed.

Future developments

Future development efforts are focused on expanding the software's compatibility with additional input formats. Specifically, future updates aim to support direct processing of Variant Call Format (.vcf) files, thereby streamlining the workflow for users working with raw genomic data.

CONCLUSIONS

High-density SNP datasets have revolutionised plant population genetics, but the strict input format required by STRUCTURE remains a technical bottleneck. SNP2STR removes this hurdle by converting .ped and .map files into STRUCTURE-ready matrices in a single command. Testing on a 64-accession apple dataset showed that the script produces fully compliant output in seconds and without user intervention.

REFERENCES

- Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., Poncet, C., Micheletti, D., Kerschbamer, E., Di Pierro, E.A., Larger, S., Pindo, M., Van de Weg, E., Davassi, A., Laurens, F., Velasco, R., Durel, C.E., & Troggio, M. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *Plant Molecular Biology*, 86(1), 62–74. <https://doi.org/10.1111/tpj.13145>
- Falush, D., Stephens, M., & Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567–1587. <https://doi.org/10.1093/genetics/164.4.1567>
- Ganal, M.W., Altmann, T., & Röder, M.S. (2009). SNP identification in crop plants. *Current Opinion in Plant Biology*, 12(2), 211–217. <https://doi.org/10.1016/j.pbi.2008.12.009>
- Konjić, A., Kurtović, M., Grahić, J., Pojskić, N., Kalajdžić, A., & Gaši, F. (2023). Using high-density SNP array to investigate genetic relationships and structure of apple germplasm in Bosnia and Herzegovina. *Horticulturae*, 9(5), 527. <https://doi.org/10.3390/horticulturae9050527>
- Lischer, H.E.L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, 28(2), 298–299. <https://doi.org/10.1093/bioinformatics/btr642>

- Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International Journal of Plant Genomics*, 2012, 1–11. <https://doi.org/10.1155/2012/728398>
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56).
- Mueller, U.G., & Wolfenbarger, L.L. (1999). AFLP genotyping and fingerprinting. *Trends in Ecology & Evolution*, 14(10), 389–394. [https://doi.org/10.1016/S0169-5347\(99\)01659-6](https://doi.org/10.1016/S0169-5347(99)01659-6)
- Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M.V. (2013). An overview of STRUCTURE: Applications, parameter settings, and supporting software. *Frontiers in Genetics*, 4, Article 98. <https://doi.org/10.3389/fgene.2013.00098>
- Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S., & Rafalski, A. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding: New Strategies in Plant Improvement*, 2(3), 225–238.
- Pritchard, J.K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945–959. <https://doi.org/10.1093/genetics/155.2.945>

SNP2STR: SOFTVER ZA KREIRANJE INPUT DATOTEKA ZA STRUCTURE SOFTVER IZ SNP PODATAKA

Sažetak

STRUCTURE je softverski paket koji koristi molekularne podatke iz više lokusa za analizu strukture populacije. Naime, Za analizu podataka zasnovanih na jednonukleotidnim polimorfizmima (engl. *Single Nucleotide Polymorphism* – SNP) pomoću softvera STRUCTURE, potreban je specifičan format ulazne datoteke. Međutim, budući da SNP studije često obuhvataju hiljade lokusa, takve datoteke obično su prevelike za manuelno kreiranje.

Kako bi se prevazišao ovaj izazov, razvijeni su različiti softverski alati i funkcije. Mešutim, kada je riječ o biljnim podacima, neki od tih alata ne uspijevaju ispravno obraditi podatke ili generišu netačne ulazne datoteke.

Kako bi se adresirali navedeni problemi razvijen je novi softverski alat — SNP2STR. SNP2STR vrši pretvaranje SNP podataka u input format koji zahtjeva STRUCTURE softver. Ovaj softver generiše datoteku koja sadrži nazive markera, sa po dva reda po jedinki formatirana u skladu sa pripadajućim markerima.

Softver je open-source i dostupan na GitHub-u. Također nudi potencijal za buduća unapređenja na osnovu povratnih informacija korisnika.

Ključne riječi: molekularni podaci, analiza strukture populacije, genetički clustering, open-source bioinformatički alati