Bioinformatics, 2022, 0–0
doi: 10.1093/*bioinformatics*/xxxxx
Advance Access Publication Date: 17 March 2023
Application Notes

Application Notes

SNPulse – plotting minor allele frequency of a SNP overtime

Pham Xuan Huy Nguyen[1,*]

[1]Department of Biology, Box 118, 211 00, Lund University, Sweden.

*To whom correspondence should be addressed.

Abstract

Motivation: Minor allele frequency (MAF) is the frequency of the less common allele at a specific genetic locus in a population. It provides important information about the genetic variation in a population. Therefore, visualising its information is key to provide insights into the evolutionary dynamics of a population and help identify important genetic changes that have occurred over time.
Results: SNPulse allows user to upload PLINK files and set parameters, retrieval its information and produce a plot of MAF and the count of allele overtime of the input SNP.
Availability: https://github.com/npxhuy/popgen_binp29
Contact: npxhuy@gmail.com

1 Introduction

Minor allele frequency (MAF) refers to the frequency of the second most common allele in a given population. MAF is an important factor in population genetics studies since it provides information to differentiate between common and rare variants in a population (Chanock et al, 2014). In population genetics, the MAF is commonly used to describe genetic diversity within populations and to compare genetic variation between populations. It is often used to infer evolutionary history, as allele frequencies can change over time as a result of genetic drift, selection, or migration. Therefore, understanding MAF can provide insight into the evolutionary forces shaping populations and species.

Several studies have highlighted the importance of considering MAF in genetic research. Rare variants with a MAF of less than 5% accounted for a significant proportion of genetic variation in human populations, and that a comprehensive understanding of MAF was necessary for accurately characterizing human genetic diversity (Abecasis et al, 2010). Similarly, a study on the genetics of height found that genetic variants with a low MAF were more likely to be associated with height, suggesting that rare variants can contribute significantly to complex traits (Marouli et al, 2017).

As genome sequencing and population genetics studies become more widespread, several web tools were made to analyse MAF such as SNPStats, HaploReg, or Vizome. However, the mentioned tools do not accept PLink file format, which is currently a popular file format used in genetic association studies. In addition, SNPStats will only do analysis on the SNPs and will not produce any plot, while Vizome is built with many other complex tools. SNPulse is much simpler than many of these tools, which was built with a morden, simple and direct design. SNPulse also accepts Plink file format, that previously had to be converted into another format in order to use in other tools. It also allows the user to modify the SNP and time step and produce desired plot.

2 Methods

SNPulse was written in R studio (v 4.2.1), and can be executed within the R studio environment. The required inputs include Plink files, a *.txt* file, and two other parameters including time step and SNP's name. Note that the Plink files usually include a *.bed* file, which has the information of genotypes of the individuals in binary format. This *.bed* file must be converted to *.ped* in order for SNPulse to read.

For each SNP's name and time step the user input, the tool follows the same steps (Fig. 1). The *.txt* file contains individuals ID and their mean date in two separated columns. Time vector is an integer positive number. The *.bim* and *.fam* are the Plink files. The *.ped* file is generated from *.ped, .bim* and *.fam* files using *plink* the command line program. The tutorial for the usage of *plink* command program is written in the *Readme* tab of the application. (Fig. 2). Using the mentioned files, the application will generate a dataset that will be used for plotting.

3 Results

The web application consists of two tabs, *Home* and *Readme*. In the *Readme* tab contains the information of the input files. Additionally, it also shows how to generate the *.ped* file from the Plink files (Fig. 2). In the *Home* tab, there are two separated panels. The side panel allows the user to upload their desired Plink files and parameters.

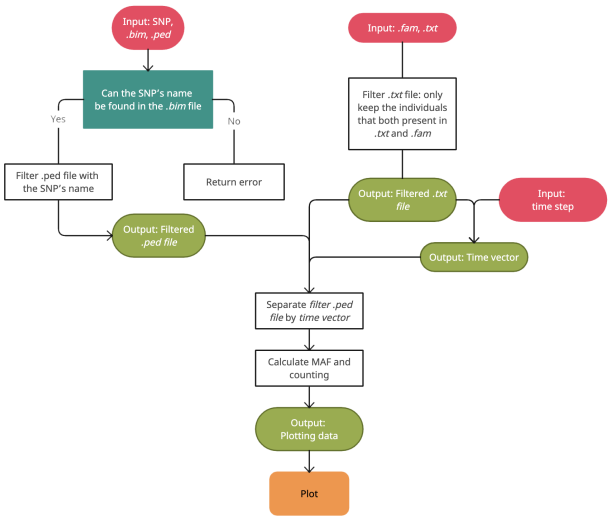

Fig. 1. Pipeline. Pipeline used in R studio.

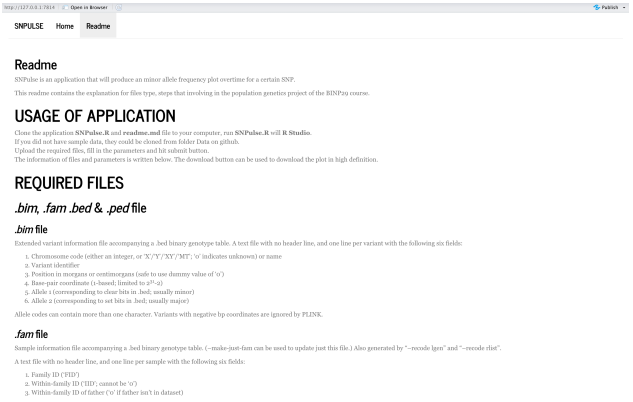

Fig. 2. Web application. Readme tab

After retrieving the information and making plotting data, the plot will be plotted on the main panel. The produced plot is a combined plot between a point-to-point graph, which plots the MAF overtime, and a stacked bar plot, which plots the count of major and minor allele overtime. The main panel also features a download button, which will download the produced plot as *.png* file whose name is the SNP's name (Fig. 3).

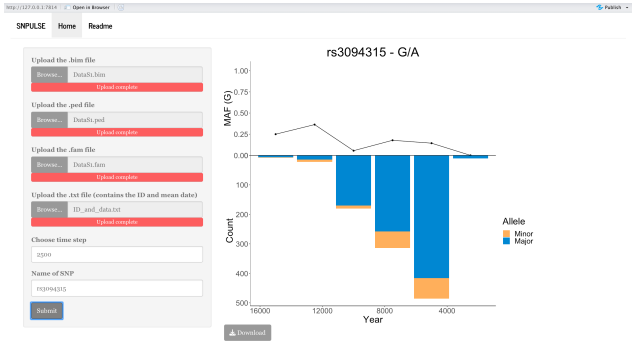

Fig. 3. Web application. Interface and sample outcome.

User can modify the name of SNP or time step and new data will be retrieved and plotted automatically without having to click the submit button again.

4 Discussion

SNPulse is a helpful tool to plot minor allele frequency overtime. It is useful for someone who wants to observe the trend of allele frequency to understand or predict how populations evolve. Besides, it also allows the user to download the plot with high definition quality, which can be convenient for someone who wants to use the plot in their report or study.

However, there are still some limitations in this application. The moderated reading speed makes the plot generated speed depend on the size of the input files, specifically *.ped* file. Users can rapidly change the time step parameter once the plot has been drawn and the application can reproduce the plot simultaneously in real time. However, it can take a few minutes to produce another plot when modifying the SNP's name.

Another limitation is that the user has to input the *.txt* file in the right format: the first column contains the ID and the second column contains the mean date. The application will not read the *.txt* file properly, hence will not produce the data for plotting. A tutorial of how to extract those information to a *.txt* file from an *.xlsx* is in the *Readme* tab, which can help the user to overcome this limitation.

In summary, this application is helpful for plotting minor allele frequency overtime. Although having few limitations, further development and optimisation on the code can accelerate the plotting speed as well as expanding its function.

Acknowledgement

I am thankful for Ms. Sara Behnamian for sharing her data that I used as example data set to run my application, Mr. Eran Elhaik for valuable comment on application and for Lund university for their financial support.

References

Chanock, S. J., & Ostrander, E. A. (2014). Discovery and Characterization of Cancer Genetic Susceptibility Alleles. *Elsevier EBooks*, 309-321.e3. https://doi.org/10.1016/b978-1-4557-2865-7.00022-9

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R. A., Hurles, M. E., & McVean, G. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. https://doi.org/10.1038/nature09534

Marouli, E., Graff, M., Medina-Gomez, C. et al. Rare and low-frequency coding variants alter human adult height. *Nature* 542, 186–190 (2017). https://doi.org/10.1038/nature21039