PLOS ONE

RESEARCH ARTICLE

PopSc: Computing Toolkit for Basic Statistics of Molecular Population Genetics Simultaneously Implemented in Web-Based Calculator, Python and R

Shi-Yi Chen[1☯]*, Feilong Deng[1☯], Ying Huang[2], Cao Li[1], Linhai Liu[1], Xianbo Jia[1], Song-Jia Lai[1]*

1 Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, 611130, China, **2** College of Veterinary Medicine, Sichuan Agricultural University, Chengdu, 611130, China

☯ These authors contributed equally to this work.
* sychensau@gmail.com (SYC); laisj5794@163.com (SJL)

Citation: Chen S-Y, Deng F, Huang Y, Li C, Liu L, Jia X, et al. (2016) PopSc: Computing Toolkit for Basic Statistics of Molecular Population Genetics Simultaneously Implemented in Web-Based Calculator, Python and R. PLoS ONE 11(10): e0165434. doi:10.1371/journal.pone.0165434

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: September 2, 2016

Accepted: October 11, 2016

Published: October 28, 2016

Data Availability Statement: All relevant data are available from the Figshare repository at the following: https://dx.doi.org/10.6084/m9.figshare.4037427.v1.

Funding: SYC is grateful for the financial supports from Department of Science and Technology of Sichuan Province, China ("Three Districts Talent Program" and "Fumin Project") and National Natural Sciences Foundation of China (31172197). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Although various computer tools have been elaborately developed to calculate a series of statistics in molecular population genetics for both small- and large-scale DNA data, there is no efficient and easy-to-use toolkit available yet for exclusively focusing on the steps of mathematical calculation. Here, we present PopSc, a bioinformatic toolkit for calculating 45 basic statistics in molecular population genetics, which could be categorized into three classes, including (i) genetic diversity of DNA sequences, (ii) statistical tests for neutral evolution, and (iii) measures of genetic differentiation among populations. In contrast to the existing computer tools, PopSc was designed to directly accept the intermediate metadata, such as allele frequencies, rather than the raw DNA sequences or genotyping results. PopSc is first implemented as the web-based calculator with user-friendly interface, which greatly facilitates the teaching of population genetics in class and also promotes the convenient and straightforward calculation of statistics in research. Additionally, we also provide the Python library and R package of PopSc, which can be flexibly integrated into other advanced bioinformatic packages of population genetics analysis.

Introduction

Theoretical framework of population genetics had been significantly promoted due to the accumulated evidence in molecular biology, which resulted in an interdisciplinary Molecular Population Genetics more than two decades ago [1]. A primary task of molecular population genetics is to investigate the distribution and dynamic change of allele frequencies at population level in relation to a series of evolutionary processes and demographic events. Accordingly, both various mathematical models and a large number of statistical tests have been developed

Competing Interests: The authors have declared that no competing interests exist.

to comprehensively address these issues. A widely known example is that Japanese scientist Fumio Tajima proposed a simple and sophisticated statistic of Tajima *D* value for testing hypothesis of neutral evolution on basis of the observed DNA polymorphisms in 1989 [2], which has still been actively cited in enormous publications [3].

Prior to the advent of high-throughput sequencing techniques, the common molecular data available for studies of population genetics always only involve a short DNA fragment directly sequenced with tens of thousands of base pairs in length or several hundreds of polymorphism loci individually genotyped. By aiming at these small-scale and standard molecular data, kinds of computer programs, such as the actively cited DnaSP [4] and Arlequin [5], had been elaborately explored for calculating a number of statistics in population genetics with different strengths and weaknesses [6]. Recently, to better address the large-scale molecular data, such as genome-wide polymorphisms by massive resequencing, many high-throughput bioinformatic packages have been proposed for much more efficiently calculating the basic statistics, especially written in R programming language [7–9].

Despite the fact that there are various tools available for population genetics analysis on both small- and large-scale molecular data, we also found two limitations in this field deserving to be addressed. First, it is very inconvenient for using these existing tools, because all of them require strict format of input data, when you intend to compute certain statistics directly starting with the ready-made metadata of variants (such as allele frequencies) instead of raw DNA sequences. So, we believe that the easy-to-use online calculator would be very helpful for conducting some trivial tasks in teaching as well as research of molecular population genetics. Second, there is no toolkit available yet for exclusively performing the calculation step of statistics, which, in expectation, should be able to be directly and flexibly employed and incorporated into an advanced bioinformatic package. Such toolkit, therefore, could facilitate the development of analysis packages for bioinformatists without good background in population genetics, and also for biologists only holding the elementary skills in bioinformatics. In the present study, we specially addressed the two issues.

Design of PopSc and Statistics

To address the issues as mentioned above, PopSc was designed to exclusively perform the calculation step of statistics of interest in molecular population genetics. Therefore, the initial input data into PopSc are these descriptive metadata, such as the allele or haplotype frequencies, number of segregating sites, counts of different mutation types, and mismatch distribution, etc., which must be generated from upstream raw DNA sequences or genotyping data in advance. After completing the computational duties, PopSc will directly pass the numerical output of statistic on to users. Such philosophy of design will promote the convenient calculation without requiring tedious input of raw DNA sequences and also guarantee the possibility of direct integration of PopSc toolkit into other advanced packages of bioinformatic analysis.

Generally guided by the popularity in scientific literatures, we comprehensively collected a total of 45 basic statistics in molecular population genetics into PopSc (Table 1). All of them could be roughly categorized into three classes, including (i) genetic diversity of DNA sequences, (ii) statistical tests for neutral evolution, and (iii) measures of genetic differentiation among populations. The required input data of PopSc for computing these statistics would be slightly different dependent on which one is selected, and each of them was clearly defined in the reference manuals. To facilitate practical uses, the calculation of each statistic is performed by one independent function. The mathematical formulas of these statistics were employed in intact from initial publications, meanwhile, all calculated results of PopSc were also carefully

Table 1. Summaries of the included statistics into PopSc.

Classes	Statistics
Genetic diversity: the basic statistics about genetic diversity among a set of nucleotide sequences.	Heterozygosity *H*, Haplotype diversity *Hd* [10]; Nucleotide diversity π [11]; Average nucleotide differences *k* [12]; Polymorphism information content, *PIC* [13]
Neutrality test: the classical statistical tests for DNA neutral evolution based on the mutation frequencies, haplotype distribution, and mismatch distribution, respectively.	Tajima's *D* [2]; Fu and Li's *D*, *F*, *D**, *F** [14]; Strobeck's *S*, Fu's *W*, F_s, Watterson's *W* [15, 16]; Fay and Wu's *H*, H_n, Zeng's *E* [17, 18]; Ramos-Onsins's R_2, R_3, R_4, R_{2E}, R_{3E}, R_{4E}, *Ch*, *Che*, *ku* [19]; Raggedness index *rg* [20]; Kelly's Z_{ns}, Z_A [21, 22]
Population structure: the statistics for measuring genetic differentiation among populations.	Wright's F_{ST}, $\bar{F}_{ST}$, F_{IS} [23]; Nei's G_{ST}, D_{ST}, J_T, J_S, R_{ST} [24]; Hedrick G'_{ST} [25]; Jost *D* [26]; Weir and Cockerham's θ_U, θ_{RH}, θ_w, f_U, f_{RH}, f_w [27]

doi:10.1371/journal.pone.0165434.t001

checked by either referring to values as being outputted by prevalent tools, such as DnaSP [4], or according to artificial verification step by step.

Implementation and Availability

Because the web browser is independent of operating system and requires no installation, PopSc was first implemented as online calculator (Fig 1). The web-based calculator of PopSc provides the convenient and straightforward calculation of statistics without requiring *ab initio* input of DNA sequences or other types of raw molecular data. Mathematical formulas and the related documentation for each statistic are also clearly shown on the respective web pages,

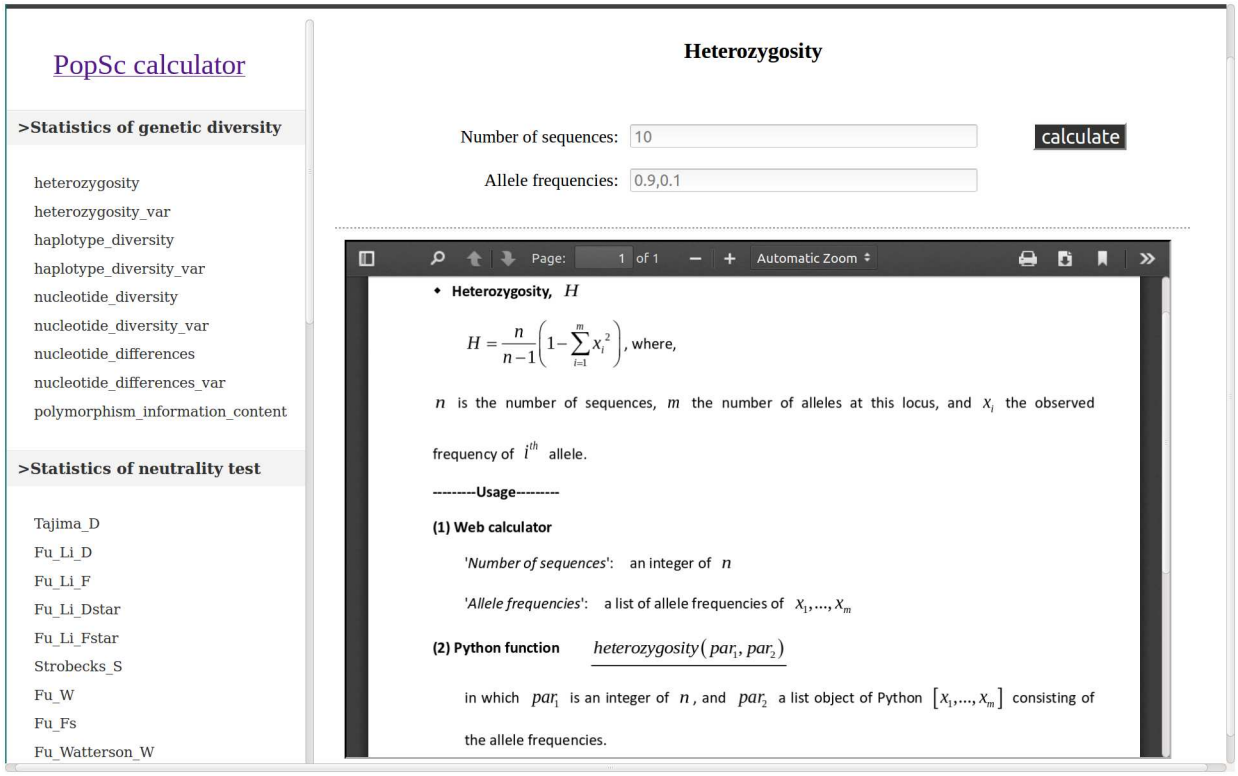

Fig 1. Screenshot of PopSc online calculator.

doi:10.1371/journal.pone.0165434.g001

which would be very helpful for teaching the molecular population genetics in class; and such teaching purpose had also been successfully addressed by the very popular tool of GENALEX, which is an add-in program for Microsoft Excel [28].

During the past years, programming languages of Python and R are becoming more and more popular in bioinformatic analyses. Therefore, PopSc was further independently implemented and provided as the Python library and R package, which can be easily integrated into other advanced bioinformatic packages in molecular population genetics. The Python library and R package of PopSc are deposited in official repositories (http://pypi.python.org and http://cran.r-project.org, respectively) and could be installed by standard commands. The web-based calculator, source codes and reference manuals of PopSc could also be freely available at: http://chenshiyi.com/popsc.html.

Of course, PopSc is not an upgrade or even a replacer to the exiting computer tools because it just performs the step of mathematical calculation for each statistic. Therefore, PopSc only accepts the descriptive metadata, which must be independently prepared from raw DNA sequences or genotyping data with the aid of the custom scripts or other computer tools, such as MEGA [29]. PopSc certainly remains open for the additional implementation of other statistics in molecular population genetics.

Author Contributions

Conceptualization: SYC.

Funding acquisition: SYC SJL.

Software: SYC FD YH.

Validation: CL LL XJ SJL.

Writing – original draft: SYC SJL.

Writing – review & editing: SYC.

References

1. Nei M. Molecular population genetics and evolution: North-Holland Publishing Company; 1975.
2. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123(3):585–95. PMID: 2513255
3. Vitti JJ, Grossman SR, Sabeti PC. Detecting natural selection in genomic data. Annu Rev Genet. 2013; 47:97–120. doi: 10.1146/annurev-genet-111212-133526 PMID: 24274750
4. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009; 25(11):1451–2. doi: 10.1093/bioinformatics/btp187 PMID: 19346325
5. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2005; 1:47–50.
6. Excoffier L, Heckel G. Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet. 2006; 7(10):745–58. doi: 10.1038/nrg1904 PMID: 16924258
7. Keenan K, McGinnity P, Cross TF, Crozier WW, Prodöhl PA. diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. Methods Ecol Evol. 2013; 4(8):782–8.
8. Adamack AT, Gruber B. PopGenReport: simplifying basic population genetic analyses in R. Methods Ecol Evol. 2014; 5(4):384–7.
9. Pfeifer B, Wittelsbürger U, Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol. 2014; 31(7):1929–36. doi: 10.1093/molbev/msu136 PMID: 24739305
10. Nei M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978; 89(3):583–90. PMID: 17248844

11. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA. 1979; 76(10):5269–73. PMID: 291943
12. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983; 105 (2):437–60. PMID: 6628982
13. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980; 32(3):314–31. PMID: 6247908
14. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993; 133(3):693–709. PMID: 8454210
15. Fu YX. New statistical tests of neutrality for DNA samples from a population. Genetics. 1996; 143 (1):557–70. PMID: 8722804
16. Fu YX. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics. 1997; 147(2):915–25. PMID: 9335623
17. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000; 155(3):1405–13. PMID: 10880498
18. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006; 174(3):1431–9. doi: 10.1534/genetics.106.061432 PMID: 16951063
19. Ramos-Onsins SE, Rozas J. Statistical properties of new neutrality tests against population growth. Mol Biol Evol. 2002; 19(12):2092–100. PMID: 12446801
20. Harpending HC. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol. 1994; 66(4):591–600. PMID: 8088750
21. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997; 146(3):1197–206. PMID: 9215920
22. Rozas J, Gullaud M, Blandin G, Aguadé M. DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics. 2001; 158(3):1147–55. PMID: 11454763
23. Wright S. The genetical structure of populations. Ann Eugen. 1951; 15(4):323–54. PMID: 24540312
24. Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA. 1973; 70 (12):3321–3. PMID: 4519626
25. Hedrick PW. A standardized genetic differentiation measure. Evolution. 2005; 59(8):1633–8. PMID: 16329237
26. Jost LO. G_{ST} and its relatives do not measure differentiation. Mol Ecol. 2008; 17(18):4015–26. PMID: 19238703
27. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984; 38(6):1358–70.
28. Peakall RO, Smouse PE. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes. 2006; 6(1):288–95.
29. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013; 30(12):2725–9. doi: 10.1093/molbev/mst197 PMID: 24132122