The Journal of Open Source Software

SeleDiff: A fast and scalable tool for estimating and testing selection differences between populations

Xin Huang[1, 2], **Li Jin**[1, 3], **and Yungang He**[4]

1 Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai, 200031, China **2** Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China **3** State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, 200433, China **4** Institutes of Biomedical Sciences, Shanghai Medical College, Fudan University, Shanghai, 200032, China

DOI: 10.21105/joss.01545

Software

- Review
- Repository
- Archive

Submitted: 29 June 2019
Published: 18 July 2019

Summary

Detecting and quantifying selection is a classical task in population genetics. Over the last two decades, many studies detected selection signals in genomes. Few studies, however, quantified differences in selective pressures between populations, due to the lack of efficient tools. Here we implemented an open-source software package, SeleDiff, with an established probabilistic method to estimate and test differences in selective pressures between populations. Extensive simulation revealed that SeleDiff is robust in various demographic models, as well as fast and scalable for analyzing large-scale genomic datasets. Thus, SeleDiff is helpful for analyzing selection as genomic datasets grow, and is available at https://github.com/xin-huang/SeleDiff.

Introduction

Analyzing natural selection is critically important in population genetics (Haldane, 1990). In the past 20 years, researchers have learned extensively about selection signals in genomic data (Vitti, Grossman, & Sabeti, 2013), but a deeper understanding of selection strength has remained elusive (Thurman & Barrett, 2016). This is particularly due to difficulties in estimating selective pressures using empirical data. In addition, as the amount of genomic data has dramatically increased, researchers require more efficient software for analyzing large-scale genomic datasets. To meet these computational demands, we introduced and evaluated SeleDiff, a fast and scalable tool for quantifying differences in selective pressures between populations.

Results

SeleDiff implements a probabilistic method from our previous study (He et al., 2015). In this approach, we introduced logarithm odds ratios of allele frequencies to measure differences in selective pressures. For a bi-allelic locus in the population i, let $p_i(t)$ and $q_i(t)$ denote the derived and ancestral allele frequencies at time t. We define the absolute fitness of the derived and ancestral alleles as w_D and w_A. The relative fitness becomes

$$e^s = \frac{w_D}{w_A},$$

where s is the (genic) selection coefficient. The selection (coefficient) difference between populations i and j is

$$d_{ij} = s_i - s_j = \frac{1}{t}\left[\ln \frac{p_i(t)/q_i(t)}{p_j(t)/q_j(t)} + \Omega\right] = \frac{1}{t}(\ln \mathrm{OR} + \Omega),$$

where OR stands for odds ratio; Ω approximately follows a normal distribution with a mean of zero and reflects the uncertainty of allele frequencies caused by factors other than selection; t is the divergence time from populations i and j to their most recent common ancestor. Thus, the expectation and variance of d_{ij} are

$$\begin{aligned} \mathrm{E}(d_{ij}) &= \mathrm{E}(s_i - s_j) = \frac{1}{t}\ln \mathrm{OR} \\ \mathrm{var}(d_{ij}) &= \frac{1}{t^2}[\mathrm{var}(\mathrm{OR}) + \mathrm{var}(\Omega)] \end{aligned}.$$

Given a dataset with n loci, we can estimate $\mathrm{var}(\Omega)$ as

$$\mathrm{median}\left\{\frac{\ln^2 \mathrm{OR}_k}{0.455} - \mathrm{var}(\ln \mathrm{OR}_k)\right\}, 1 \le k \le n,$$

where $\mathrm{var}(\ln \mathrm{OR}) \approx 1/[N_i \hat{p}_i(t)] + 1/[N_i \hat{q}_i(t)] + 1/[N_j \hat{p}_j(t)] + 1/[N_j \hat{q}_j(t)]$. Here, N_i and N_j are the sample sizes of populations i and j. We add 0.5 to allele counts less than 5 for continuity correction. To test the selection differences in a locus, we proposed a statistic:

$$\delta = \frac{[\mathrm{E}(d_{ij})]^2}{\mathrm{var}(d_{ij})},$$

where δ follows a central χ^2-distribution with one degree of freedom in the absence of selection differences.

We evaluated SeleDiff in different demographic models (Figure 1) simulated by SLiM 2 (Haller & Messer, 2017). In Models 1–70, we assume larger selection coefficients in Population 1 than in Population 2 (Figure 1A–E). Without migration, SeleDiff accurately estimates selection differences ranging from 0 to 0.002/generation in scenarios with different population sizes (Figure 2A, Models 1–9). The estimated differences (Figure 2A, Models 10–17) are slightly smaller in scenarios with low initial frequencies ($\le$ 0.02) of the selective allele or long divergence times ($\ge$ 5000 generations), because alleles with low initial frequencies are easily lost regardless of their selection coefficients, and alleles with small selection coefficients can reach high frequencies with long enough time. SeleDiff is affected little by time-varied population sizes (Figure 2A, Models 18–37), except for extremely severe bottlenecks in populations under less selective pressures (Figure 2A, Model 23). In Models 38–46 (Figure 2A), populations diverge into subpopulations, and selection stops in one of these subpopulations. If we ignore their structures, then the estimated differences diminish because SeleDiff treats all the individuals in a group homogenously. Therefore, it is important to select samples carefully and interpret results cautiously. In models with moderate migration rates (0.00001–0.0001/generation), the estimated differences are only slightly smaller than the given values, whereas strong migration reduces differences between populations (Figure 2B, Models 47–70), a well-known phenomenon in population genetics (Crow & Kimura, 2009). SeleDiff also works well in complex models (Figure 2B, Model 1a–6d) involving multiple demographic events from human evolution (Gravel et al., 2011). Thus, SeleDiff is robust in various demographic models, and

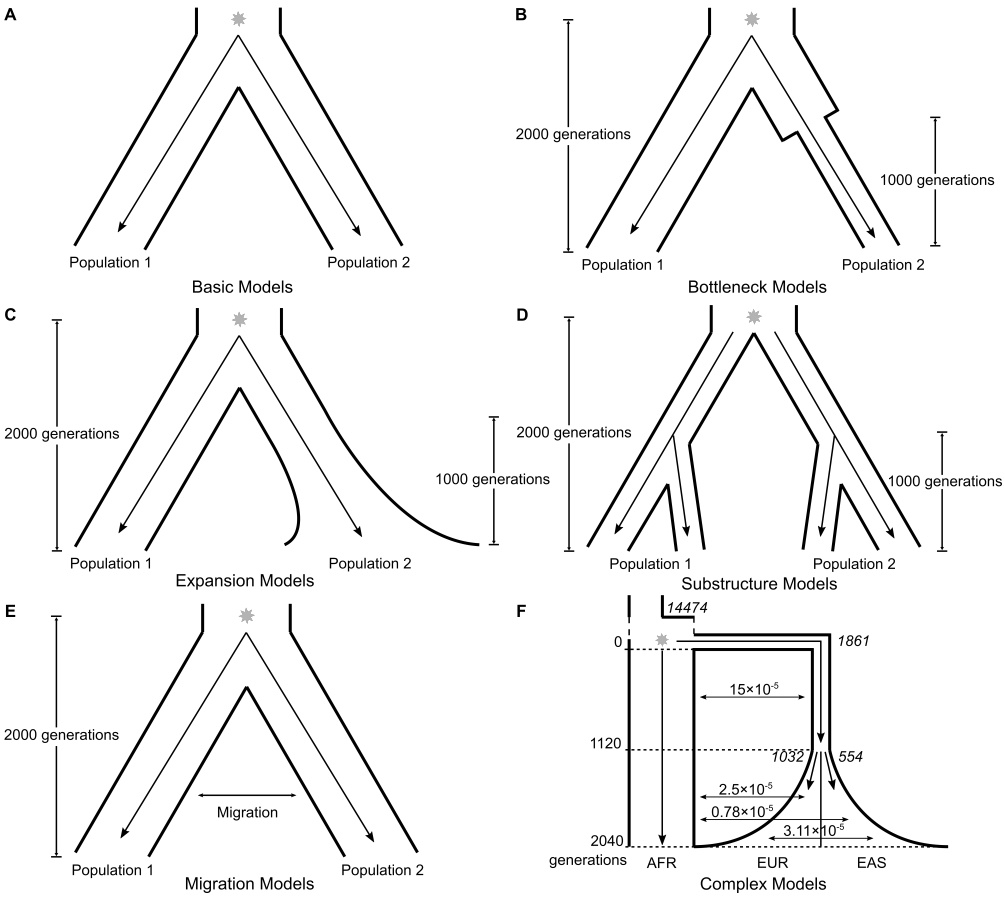

Figure 1: The demographic models in simulation.

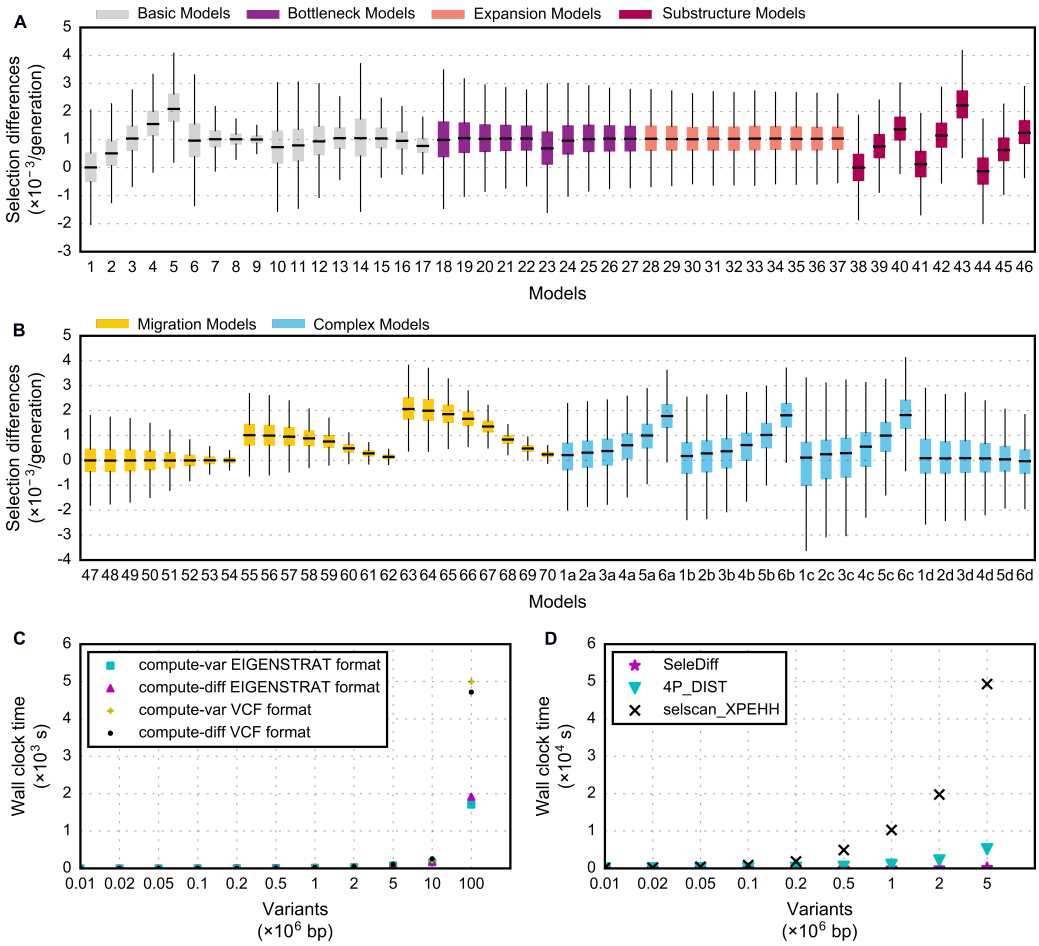

Figure 2: Accuracy and speed of SeleDiff.

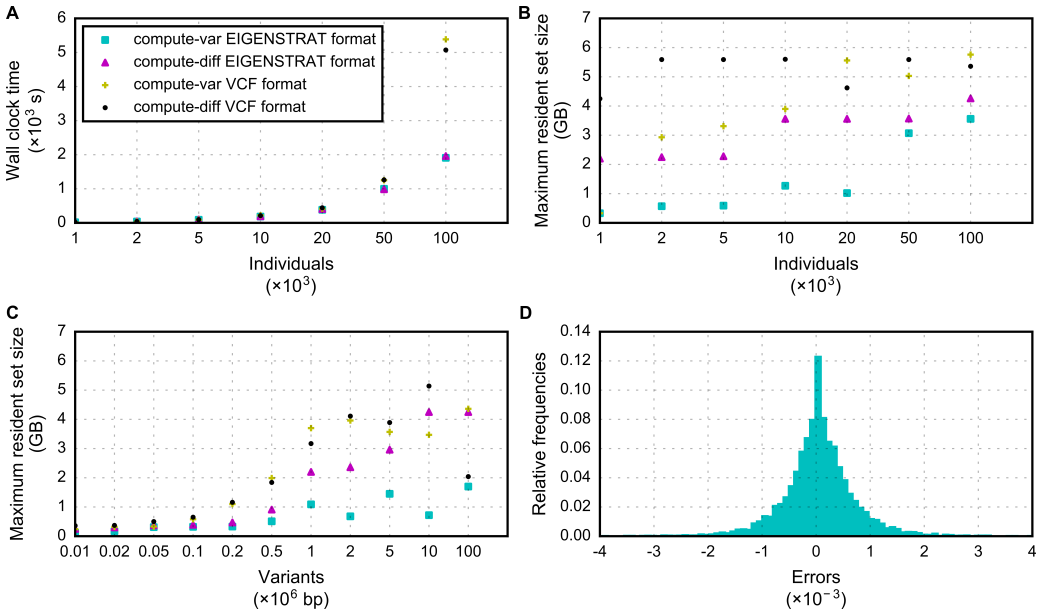

Figure 3: Performance of SeleDiff.

indicates the lower bounds of differences in selective pressures when migration or substructure exists.

Finally, we compared the performance of SeleDiff with other cross-population methods in two recent programs—4P and selscan—for genome-wide selection scans (Benazzo, Panziera, & Bertorelle, 2014; Szpeich & Hernandez, 2014). All the programs were executed with a single thread. SeleDiff can analyze a dataset containing 10^8 base pairs of variants in less than 1 hour (Figure 2C) with less than 4 gigabytes of random-access memory (Figure 3), and is much faster than the other two programs (Figure 2D). To enhance the scalability of SeleDiff, we integrated it with a newly developed online algorithm— t-digest (Dunning & Friedman, 2014). T-digest allows SeleDiff to estimate var (Ω) from genome-wide data with only a small amount of memory (Figure 3). In summary, SeleDiff can help researchers detect and quantify natural selection from massive genomes in this era of big data.

Acknowledgements

This work was supported by grants from National Natural Science Foundation of China (91331109 and 91731310 to Y.H.; 31271338 and 3133038 to L.J.). L.J. was also supported by the Shanghai Leading Academic Discipline Project (B111) and the Center for Evolutionary Biology at Fudan University. The authors declare no conflict of interest.

References

Benazzo, A., Panziera, A., & Bertorelle, G. (2014). 4P: Fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol Evol*, *5*, 172–175. doi:10.1002/ece3.1261

Crow, J. F., & Kimura, M. (2009). *An Introduction to Population Genetics Theory*. The Blackburn Press.

Dunning, T., & Friedman, E. (2014). *Practical Machine Learning: A New Look at Anomaly Detection*. O'Reilly Media, Inc.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., et al. (2011). Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci USA*, *108*, 11983–11988. doi:10.1073/pnas.1019276108

Haldane, J. B. S. (1990). A mathematical theory of natural and artificial selection—I. *Bltn Mathcal Biology*, *52*, 209–240. doi:10.1007/BF02459574

Haller, B. C., & Messer, P. W. (2017). SLiM 2: Flexible, interactive forward genetic simulations. *Mol Biol Evol*, *34*, 230–240. doi:10.1093/molbev/msw211

He, Y., Wang, M., Huang, X., Li, R., Xu, H., Xu, S., & Jin, L. (2015). A probabilistic method for testing and estimating selection differences between populations. *Genome Res*, *25*, 1903–1909. doi:10.1101/gr.192336.115

Szpeich, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*, *31*, 2824–2827. doi:10.1093/molbev/msu211

Thurman, T. J., & Barrett, R. D. (2016). The genetic consequences of selection in natural populations. *Mol Ecol*, *25*, 1429–1448. doi:10.1111/mec.13559

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu Rev Genet*, *47*, 97–120. doi:10.1146/annurev-genet-111212-133526