Brief Communication

sstar: A Python package for detecting archaic introgression from population genetic data with *S**

Xin Huang[1, 2, *], Martin Kuhlwilm[1, 2, *]

[1] Department of Evolutionary Anthropology, University of Vienna, Djerassiplatz 1, 1030 Vienna, Austria

[2] Human Evolution and Archaeological Sciences (HEAS), University of Vienna; Austria.

* Corresponding authors: xin.huang@univie.ac.at; martin.kuhlwilm@univie.ac.at

Abstract

*S** is a widely used statistic for detecting archaic admixture from population genetic data. To apply *S**, several previous studies used the freezing-archer repository, which is a compilation of Python 2 and R scripts originally developed to detect Neanderthal and Denisovan introgression into modern humans. When studying other species or populations, however, freezing-archer is not directly applicable. Here, we implemented a new Python 3 package named sstar, with better performance than freezing-archer. Reliability was also improved after fixing bugs in freezing-archer. Compared with freezing-archer, sstar is more user-friendly, versatile, and well-documented.

Keywords: Introgression; Archaic admixture; *S**; Python

Admixture between populations is a topic of great interest (Fontsere et al. 2019), especially since we know that it does affect our own species. For example, we have direct evidence for genetic exchange with other hominins like Neandertals (Green et al. 2010). To detect archaic admixture from population genetic data, a statistic named *S** was introduced to search for patterns of variation and linkage expected in the case of introgression (Plagnol and Wall 2006). This statistic has been applied in subsequent studies in modern humans (Huerta-Sanchez et al. 2014; Vernot and Akey 2014; Vernot et al. 2016; Xu et al. 2017; Jacobs et al. 2019) as well as other organisms (Cong et al. 2016; Kuhlwilm et al. 2019).

Several methods also have been developed to detect ancient admixture. Some of these depend on knowledge of the source genome to infer introgressed fragments (Sankararaman et al. 2014; Peter 2020), while others identify highly discordant regions in the genome (Skov et al. 2018; Yuan et al. 2021) similar to *S**. These methods are often available as public packages that can be easily applied to new datasets. Although *S** has been a fundamental method to expand our knowledge on archaic admixture and was widely used in a range of landmark studies, there is no user-friendly and versatile package available.

A previous implementation of *S** is freezing-archer (Vernot and Akey 2014; Vernot et al. 2016). It was originally developed with Python 2, which is no longer supported (from 2020). Moreover, freezing-archer was specifically designed with human demographic models and used for detecting introgressed fragments from Neanderthals and Denisovans into the Papuan population (Vernot et al. 2016). If users want to apply any part of the analysis with other demographic models from humans or non-human species, they must carefully read and understand the source codes of freezing-archer before manually changing the parameters inside the codes. To improve the efficiency, robustness and reproducibility when using *S** for detecting introgression, we implemented sstar, based on the algorithm and pipeline in freezing-archer, using Python 3. sstar can efficiently calculate *S** with population genomic data, and also provides a more user-friendly pipeline with snakemake (Mölder et al. 2021) for downstream analysis. sstar also fixed several bugs (Table S1) discovered in freezing-archer and supports multi-processing for accelerating calculation.

The whole pipeline in sstar is illustrated in Figure 1A. We define the population without introgressed fragments as the reference population, the population which received introgressed fragments as the target population, and the population which donated introgressed fragments as the source population (Figure S1). Besides, we assume genotype data are diploid, biallelic, phased and not missing in all populations. We also remove variants with derived alleles that are fixed in both the reference and target populations. Users can calculate *S** for sliding windows across genomes without source populations by defining the length of a window and step size. If users want to determine whether *S** is statistically significant, they can use simulated data from a demographic model without introgression and build up a generalized linear model with different *S**, quantiles of *S**, numbers of mutations, and local recombination rates to predict the expected *S**. If a genome from the source population is available, users can calculate a source match percentage between a haplotype from the target population and the source genome. Users can also obtain a *p*-value for this source match percentage. To estimate *p*-values for source match percentages, users need to provide a panel of samples from the reference population for

estimating empirical null distributions for source match percentages. If genomes from multiple source populations are available, users can estimate posterior probabilities for determining which source population a fragment is more likely from.

We compared sstar to freezing-archer with data used in Vernot et al. (2016). The obtained results are in principle identical in both cases (Figure S2 & S3). We found freezing-archer may overestimate *p*-values before fixing bugs (Figure S4). Therefore, the results in Vernot and Akey (2014) and Vernot et al. (2016) may be conservative. We also assessed the performance of sstar. sstar outperforms freezing-archer in the whole pipeline with a ~6-fold to ~10-fold increase in processing speed using a single thread (Figure 1B, Table S2 & S3). We conclude that this package can be used to calculate the *S** statistic reliably and efficiently and perform the downstream analysis to identify introgressed fragments in the genome and assign them to a source population.

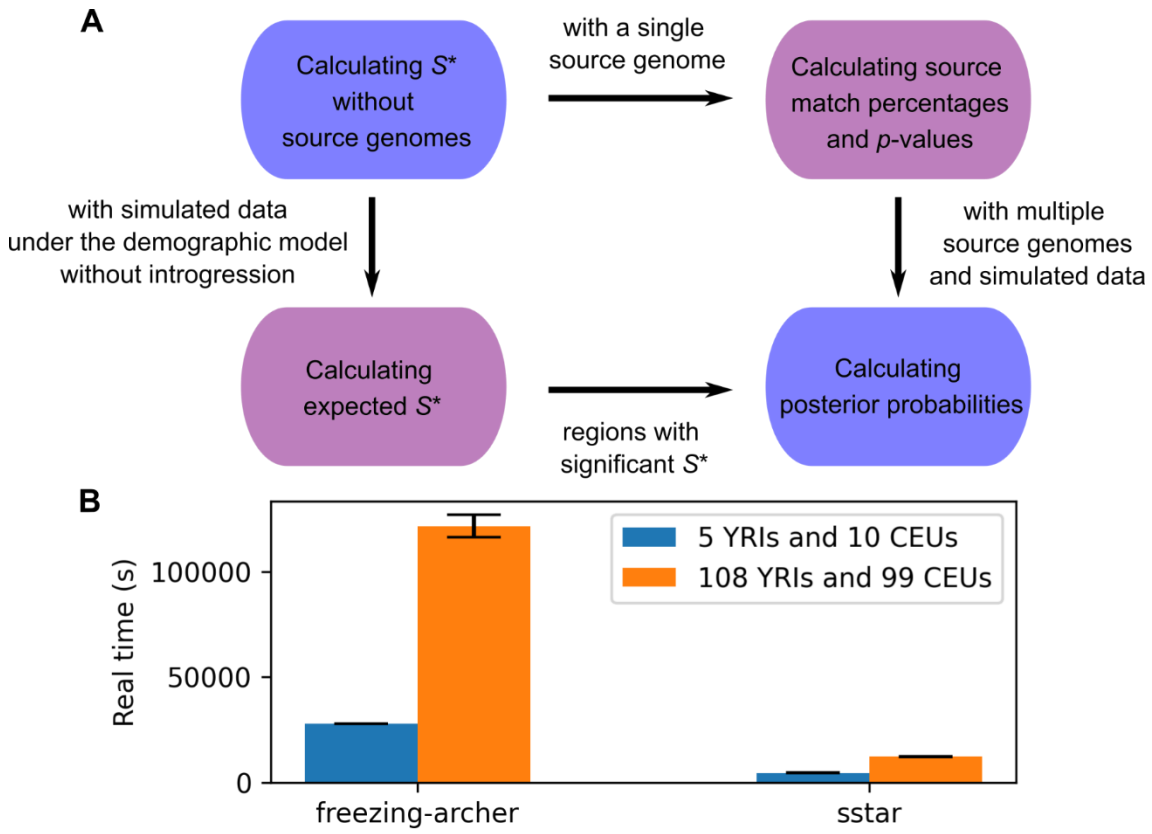

Figure 1 Pipeline implemented in sstar and its performance. (A) The whole workflow. *S** scores can be estimated without source genomes. Expected *S** scores can be obtained with simulated data and regions with significant *S** can be determined. If a single source genome is available, source match percentages and *p*-values can be estimated. Here, source match percentages

measure the similarity between haplotypes from the target population and a genome from the source population, and *p*-values indicate the significance of source match percentages. If multiple source genomes are available, posterior probabilities can be estimated for determining which source population fragments originated from. (B) Performance comparison between freezing-archer and sstar with two different datasets using the genotype data of the human chromosome 21 from the 1000 Genomes Project. The first dataset contained five YRI individuals as the reference population and ten CEU individuals as the target population. The second dataset consisted of 108 YRI individuals as the reference population and 99 CEU individuals as the target population.

Data availability

Source codes can be found in https://github.com/xin-huang/sstar and the manual can be found in https://sstar.readthedocs.io/en/latest/. Other supplementary data can be found in https://figshare.com/ndownloader/files/34320419.

Acknowledgments

We thank Benjamin Vernot for helpful discussion on freezing-archer. This project has been funded by the Vienna Science and Technology Fund (WWTF) and the City of Vienna through project VRG20-001.

Author contribution

M.K. and X.H. designed the study. X.H. implemented the package. M.K. tested the package. X.H. and M.K. analysed the data and wrote the manuscript.

Competing interests

The authors declare no conflict of interests.

References

Cong Q, Shen J, Warren AD, Borek D, Otwinowski Z, Grishin N V. 2016. Speciation in

Cloudless Sulphurs Gleaned from Complete Genomes. *Genome Biol. Evol.* 8: 915–931.

Fontsere C, de Manuel M, Marques-Bonet T, Kuhlwilm M. 2019. Admixture in Mammals and How to Understand Its Functional Implications. *BioEssays* 41: 1900123.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* 328: 710–722.

Huerta-Sanchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, et al. 2014. altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512: 194–197.

Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, Mondal M, Pagani L, Ricaut F-X, Stoneking M, et al. 2019. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell* 177: 1010-1021.e32.

Kuhlwilm M, Han S, Sousa VC, Excoffier L, Marques-Bonet T. 2019. Ancient admixture from an extinct ape lineage into bonobos. *Nat. Ecol. Evol*. 3: 957–965

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. 2021. Sustainable data analysis with Snakemake. *F1000Res* 10: 33.

Peter BM. 2020. 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. *bioRxiv*. https://www.biorxiv.org/node/1196895.abstract, last accessed March 8, 2022.

Plagnol V, Wall JD. 2006. Possible Ancestral Structure in Human Populations. *PLoS Genet.* 2: e105.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* 507: 354–357.

Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, Durbin R. 2018. Detecting archaic introgression using an unadmixed outgroup. *PLOS Genet.* 14: e1007641.

Vernot B, Akey JM. 2014. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science* 343: 1017–1021.

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S,

McCoy RC, Norton H, et al. 2016. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352: 235–239.

Wall JD, Lohmueller KE, Plagnol V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* 26: 1823–1827.

Xu D, Pavlidis P, Taskent RO, Alachiotis N, Flanagan C, DeGiorgio M, Blekhman R, Ruhl S, Gokcumen O. 2017. Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. *Mol. Biol. Evol.* 34: 2704–2715.

Yuan K, Ni X, Liu C, Pan Y, Deng L, Zhang R, Gao Y, Ge X, Liu J, Ma X, et al. 2021. Refining models of archaic admixture in Eurasia with ArchaicSeeker 2.0. *Nat. Commun.* 12: 6232.