sstar: A Python Package for Detecting Archaic Introgression from Population Genetic Data with S*

Xin Huang [1,2]* Patricia Kruisz,[3] and 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, Djerassiplatz 1, 1030 Vienna, Austria

[3]Department of Bio Data Science, Faculty of Engineering, University of Applied Sciences Wiener Neustadt, Biotech Campus Tulln, Konrad Lorenz-Straße 10, 3430 Tulln, Austria

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

Associate editor: Daniel Falush

Abstract

S* is a widely used statistic for detecting archaic admixture from population genetic data. Previous studies used freezing-archer to apply S*, which is only directly applicable to the specific case of Neanderthal and Denisovan introgression in Papuans. Here, we implemented sstar for a more general purpose. Compared with several tools, including SPrime, SkovHMM, and ArchaicSeeker2.0, for detecting introgressed fragments with simulations, our results suggest that sstar is robust to differences in demographic models, including ghost introgression and two-source introgression. We believe sstar will be a useful tool for detecting introgressed fragments in various scenarios and in non-human species.

Key words: **introgression, archaic admixture, S*, Python.**

Admixture between populations is a topic of great interest (Fontsere et al. 2019), especially in hominins (Peter 2020). 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 (Wall et al. 2009; 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). Although the S* statistic is a powerful approach for detecting introgressed fragments without source genomes, there is no user-friendly and versatile package available. A previous implementation of S* is freezing-archer, which was specifically designed with human demographic models and used for detecting introgressed fragments from Neanderthals and Denisovans into Papuans (Vernot et al. 2016). Users must carefully read and understand the source codes of freezing-archer before manually changing the parameters inside the code. To improve the efficiency, robustness and reproducibility when using S* for detecting introgression, we implemented sstar.

The whole pipeline is illustrated in figure 1A. We define the population without introgressed fragments as the reference population, the population that received introgressed fragments as the target population, and the population that donated introgressed fragments as the source population (supplementary fig. S1, Supplementary Material online). We assume genotype data are diploid, biallelic and not missing in all the individuals of a dataset. We remove variants with derived alleles that are fixed in both the reference and target populations. Users can calculate S* for sliding windows across genomes by defining the window length and step size. To assess significance of S* scores, users can simulate data under a demographic model without introgression and build a generalized additive model (GAM) with different S* scores, quantiles of S*, numbers of mutations, and local recombination rates to predict the expected S* scores, as described previously (Vernot et al. 2016). If a genome from a potential source population is available, users can calculate the source match rate between an individual from the target population and an individual from the source population. If genomes from two different source populations are available, the origin of candidate introgressed fragments can be determined by comparing the source match rates with different source populations.

We evaluated the performance of sstar with precision-recall curves because precision-recall curves may be more informative than receiver operating characteristic curves on imbalanced data sets (Saito and Rehmsmeier 2015). We simulated data with msprime 1.0 (Kelleher et al. 2016; Baumdicker et al. 2022) for different demographies and sample sizes. Two models tested ghost introgression: a Human-Neanderthal model (Gower et al. 2021) and a Bonobo-Ghost model (Kuhlwilm et al. 2019). Two further models tested two-source introgression: a Human-Neanderthal-Denisovan model (Malaspinas et al. 2016; Jacobs et al. 2019) and a Chimpanzee-Ghost-Bonobo model.

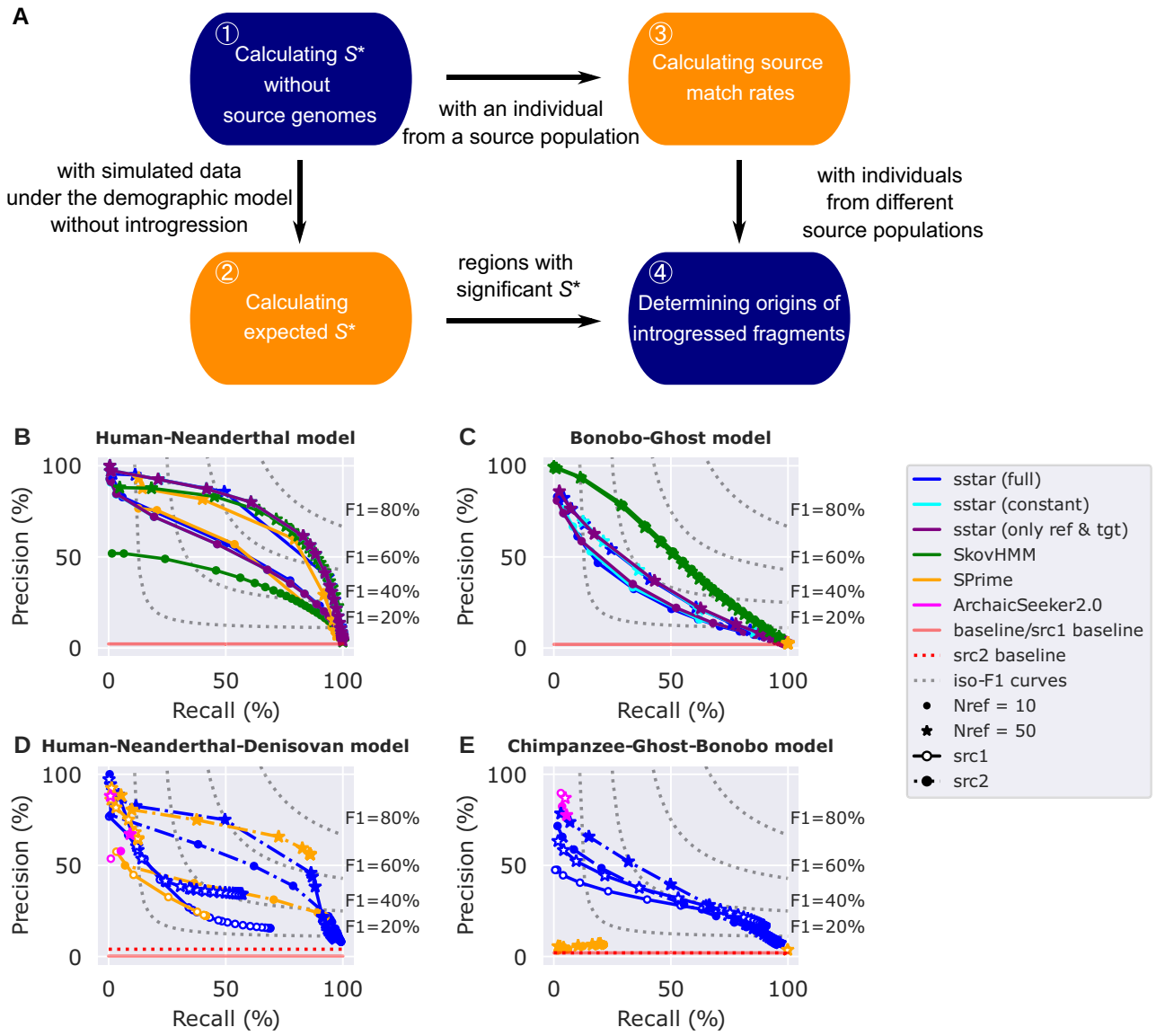

FIG. 1. The sstar workflow and its performance in different demographic models and sample sizes with SPrime, SkovHMM, and ArchaicSeeker2.0. Different points represent precision and recall estimated with different cutoffs (supplementary tables S1–S6, Supplementary Material online). For ArchaicSeeker2.0, we only used the best results because this tool does not provide options to define candidate introgressed fragments with different cutoffs. Nref is the diploid sample size of the reference population (10 or 50). sstar (full) are results inferred with GAMs using simulated data from full demographic models without introgression (supplementary figs. S6–S9, Supplementary Material online). sstar (constant) are results inferred with GAMs using simulated data from constant effective population size models without introgression (supplementary figs. S10 and S11, Supplementary Material online). sstar (only ref and tgt) are results inferred with GAMs using simulated data from models with only the reference and target populations, these populations are also constant in size (supplementary figs. S12 and S13, Supplementary Material online). src1 represents the performance for identifying the introgressed fragments from the source population 1. src2 represents the performance for identifying the introgressed fragments from the source population 2. A baseline is the performance of a random classifier, where the precision is equal to the true proportion of the introgressed fragments. An F_1 score is the harmonic mean of a given pair of precision and recall, dotted hyperbolic curves represent F_1 isometrics. (*A*) The sstar workflow. (*B*) The precision-recall curves of sstar, SPrime, and SkovHMM for detecting introgression without source genomes under a Human-Neanderthal model(Gower et al. 2021, supplementary fig. S2 and table S7, Supplementary Material online). (*C*) The precision-recall curves of sstar, SPrime, and SkovHMM for detecting introgression without source genomes under a Bonobo-Ghost model(Kuhlwilm et al. 2019, supplementary fig. S3 and table S8, Supplementary Material online). (*D*) The precision-recall curves of sstar, SPrime, and ArchaicSeeker2.0 for detecting introgression with source genomes from two-source populations under a Human-Neanderthal-Denisovan model from stdpopsim(Adrion et al. 2020, supplementary fig. S4 and table S9, Supplementary Material online). The src1 population is the Neanderthal population. The src2 population is the Denisovan population. (*E*) The precision-recall curves of sstar, SPrime, and ArchaicSeeker2.0 for detecting introgression with source genomes from two-source populations under a hypothetical Chimpanzee-Ghost-Bonobo model(supplementary fig. S5 and table S10, Supplementary Material online) modified from Kuhlwilm et al. (2019). The src1 population is the Ghost population, and the src2 population is the Bonobo population, both introgressing into the Central Chimpanzee population.

For ghost introgression, we compared sstar with SPrime (Browning et al. 2018), another tool using an S*-like approach, and SkovHMM (Skov et al. 2018), a tool based on hidden Markov models (HMMs). In the Human-Neanderthal model, our results show that sstar and SPrime performed better than SkovHMM, when sample size was small (ten reference individuals, fig. 1*B*). In the Bonobo-Ghost model, SPrime performed poorly (fig. 1*C*), assigning the whole simulated sequence as a single introgressed fragment, while sstar and SkovHMM still detected introgressed fragments.

One key step in sstar is calculating the expected S* scores with simulated data from demographic models without introgression, requiring detailed knowledge on population history (supplementary figs. S6–S9, Supplementary Material online). Using approximate models (supplementary figs. S10–S13, Supplementary Material online), our results suggest that sstar still performed similarly to those using the full history (fig. 1*B* and *C*). For two-source introgression, we compared sstar with SPrime, and ArchaicSeeker2.0 (Yuan et al. 2021; Zhang et al. 2022), another HMM-based tool. Both sstar and SPrime performed better when identifying Denisovan fragments than identifying Neanderthal fragments (fig. 1*D*). This may be due to the Denisovan introgression event in Papuans being more recent and its admixture proportion being larger than for the Neanderthal introgression. More ancient events like in the Chimpanzee-Ghost-Bonobo model cannot be well determined by SPrime, while sstar still retained power (fig. 1*E*).

We conclude that sstar is robust for detecting introgressed fragments. Since no single tool could perform well in all scenarios, users should choose appropriate tools based on their data. We believe sstar will be useful in various scenarios, especially considering small samples, and non-human data sets.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Benjamin Vernot for discussions on freezing-archer; Andrew Kern and Peter Ralph for implementing Snakemake pipelines; Graham Gower for plotting demographic models with demesdraw and help from the PopSim Consortium; and Benjamin M. Peter and an anonymous reviewer. This project has been funded by the Vienna Science and Technology Fund (WWTF) and the City of Vienna through project VRG20-001.

Author Contributions

M.K. and X.H. designed the study, analysed the data, and wrote the manuscript. X.H. implemented sstar. M.K. tested sstar. X.H., P.K., and M.K. implemented the benchmarking pipelines.

Data Availability

Source codes for sstar can be found in https://github.com/admixVIE/sstar (last accessed on August 17, 2022) and the manual can be found in https://sstar.readthedocs.io/en/latest/ (last accessed August 17, 2022). Codes for replicating the benchmarking can be found in https://github.com/admixVIE/sstar-analysis (last accessed August 17, 2022). Computational tools installed through conda can be found in https://github.com/admixVIE/sstar-analysis/blob/main/environment.yml (last accessed August 17, 2022). Tools listed below cannot be installed through conda but can be found in the websites from their authors: ArchaicSeeker2.0 (https://github.com/Shuhua-Group/ArchaicSeeker2.0, last accessed August 17, 2022), ms program ([Hudson 2002]; https://home.uchicago.edu/~rhudson1/source/mksamples.html, last accessed August 17, 2022), SkovHMM (https://github.com/LauritsSkov/Introgression-detection, last accessed August 17, 2022), SPrime (https://github.com/browning-lab/sprime, last accessed August 17, 2022), and SPrime pipeline ([Zhou and Browning 2021]; https://github.com/YingZhou001/sprimepipeline, last accessed August 17, 2022). Demographic models in Demes YAML format (Gower et al. 2022) can be found in https://github.com/admixVIE/sstar-analysis/tree/main/config/simulation/models (last accessed August 17, 2022).

References

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, *et al.* 2020. A community-maintained standard library of population genetic models. *Elife* **9**:e54967.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, *et al.* 2022. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* **220**:iyab229.

Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. 2018. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* **173**:53–61.e9.

Cong Q, Shen J, Warren AD, Borek D, Otwinowski Z, Grishin NV. 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.

Gower G, Picazo PI, Fumagalli M, Racimo F. 2021. Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife* **10**:e64669.

Gower G, Ragsdale A, Bisschop G, Gutenkunst R, Hartfield M, Noskova E, Schiffels S, Schiffels S, Struck T, Kelleher J, *et al.* 2022. Demes: a standard format for demographic models.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**:337–338.

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.

Kelleher J, Etheridge AM, McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. **12**:e1004842.

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.

Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, *et al*. 2016. A genomic history of Aboriginal Australia. *Nature* **538**: 207–214.

Peter BM. 2020. 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. bioRxiv:2020.03.13. 990523.

Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet*. **2**:e105.

Saito T, Rehmsmeier M. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**:e0118432.

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.

Zhang R, Yuan K, Xu S. 2022. Detecting archaic introgression and modeling multiple-wave admixture with ArchaicSeeker 2.0. *STAR Protoc*. **3**:101314.

Zhou Y, Browning SR. 2021. Protocol for detecting introgressed archaic variants with SPrime. *STAR Protoc*. **2**:100550.