Improving Selection Detection with Population Branch Statistic on Admixed Populations

Burak Yelmen[1,2,*], Davide Marnetto[1], Ludovica Molinaro[1,2], Rodrigo Flores[1], Mayukh Mondal[1,†], and Luca Pagani[1,3,*,†]

[1]Institute of Genomics, University of Tartu, Estonia

[2]Institute of Molecular and Cell Biology, University of Tartu, Estonia

[3]Department of Biology, University of Padova, Italy

†These senior authors contributed equally to this work.

*Corresponding authors: E-mails: burakyelmen@gmail.com; lp.lucapagani@gmail.com

Accepted: 22 February 2021

Abstract

Detecting natural selection signals in admixed populations can be problematic since the source of the signal typically dates back prior to the admixture event. On one hand, it is now possible to study various source populations before a particular admixture thanks to the developments in ancient DNA (aDNA) in the last decade. However, aDNA availability is limited to certain geographical regions and the sample sizes and quality of the data might not be sufficient for selection analysis in many cases. In this study, we explore possible ways to improve detection of pre-admixture signals in admixed populations using a local ancestry inference approach. We used masked haplotypes for population branch statistic (PBS) and full haplotypes constructed following our approach from Yelmen et al. (2019) for cross-population extended haplotype homozygosity (XP-EHH), utilizing forward simulations to test the power of our analysis. The PBS results on simulated data showed that using masked haplotypes obtained from ancestry deconvolution instead of the admixed population might improve detection quality. On the other hand, XP-EHH results using the admixed population were better compared with the local ancestry method. We additionally report correlation for XP-EHH scores between source and admixed populations, suggesting that haplotype-based approaches must be used cautiously for recently admixed populations. Additionally, we performed PBS on real South Asian populations masked with local ancestry deconvolution and report here the first possible selection signals on the autochthonous South Asian component of contemporary South Asian populations.

Key words: natural selection, PBS, XP-EHH, local ancestry inference, admixed populations, South Asia.

Significance

Detecting natural selection in recently admixed populations can be difficult due to the obscurity of the source of selection signals. In this study, we used local ancestry inference methods to obtain ancestry assigned haplotypes out of simulated admixed populations and reported improvement in detecting selection signals before the admixture event using these haplotypes instead of the admixed ones. We additionally demonstrated that methods utilizing haplotype structure to detect selection must be used cautiously for recently admixed populations. Finally, we applied our approach to real South Asian genomes to report first possible selection signals for autochthonous South Asian populations.

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Introduction

Investigating signatures of selection in admixed populations is challenging due to the fact that independent signals from source populations may be obscured by each other (Huerta-Sánchez et al. 2014; Galaverni et al. 2017; Pierron et al. 2018). Although recent advancements in ancient DNA (aDNA) grants us with the opportunity to independently study source populations prior to the admixture event, detecting selection signals in populations with admixture background is still widely an unexplored and challenging endeavor, especially in areas where aDNA preservation is limited. When conventional methods like population branch statistics (PBS; Yi et al. 2010) or cross-population extended haplotype homozygosity (XP-EHH; Sabeti et al. 2007) are utilized to detect possible signals on admixed populations, it is not easy to resolve whether the candidate signals are due to selection acting after the admixture event or before and in the latter case, on which source population the selection event took place. Here, we particularly concentrated on detecting selection signals before admixture. We postulate that applying local ancestry inference as a preliminary step and then searching for signatures of selection within each set of source haplotypes may greatly decrease false positives and may help assign the observed sweeps to the correct ancestral population. In a previous study, we used a local ancestry deconvolution-based approach to create surrogates of the two main ancestral components of contemporary South Asian human genomes to shed light into the demographic history of these highly admixed human populations (Yelmen et al. 2019) and to address patterns of selection after admixture. Here, we test our idea through forward simulations and investigate the possible use of these surrogates for frequency based (PBS) and haplotype based (XP-EHH) selection tests in humans or any other diploid organism.

Results

Comparison of PBS Scores Based on Forward Simulations

We used forward simulations to create an admixed human population by mixing available genomes from European (French) and East Asian (Han) individuals for 200 generations (assuming 30 years per generations, which will be around six thousand years ago). Notably, European-East Asian population splits may mimic the split between the North and South genomic components found within contemporary South Asian populations (see Yelmen et al. for a full description of South Asian demography). For this reason, we decided to respectively label "N" and "S" the European and East Asian ancestries retrieved from this simulation test, to keep consistency with the subsequent real case scenario applied on South Asian human genomes. Utilizing both PCAdmix (Brisbin et al. 2012) and ELAI (Guan 2014) and using German and Japanese genomes (two populations deemed to be genetically close to French and Han, respectively) as reference populations for local ancestry deconvolution, we formed MASK_S (German-assigned chunks masked out) and MASK_N (Japanese-assigned chunks masked out) populations from the simulated data. Then, we performed PBS on both (masked and naive) and compared with the results obtained from the original source populations (French and Han), which were used as a true dataset, set as our standard (see Materials and Methods). Overall SNP by SNP correlation showed significantly higher correlation between the PBS scores of original source populations and masked populations compared to the unmasked admixed populations (Supplementary figures S1 and S2*a*). However, when we only analyzed SNPs with PBS scores above the 99th percentile (with SNPs selected based on Han and French source population PBS scores), the ELAI approach retained higher correlation whereas PCAdmix performed poorly (figure 1, Supplementary figure S2*b*). The poor performance of PCAdmix was due to false negatives, as can be seen from the low PBS values accumulating along the *x* axis. Additionally, both masking approaches performed better compared with the naive populations when we concentrated only on the top 50 scoring SNPs except for MASK_N_ELAI (Supplementary table S1*a*).

Aside from SNP by SNP comparison, we also compared window-based mean PBS scores as widely used in selection studies (Huerta-Sánchez et al. 2013). Overall, precision and true positive rate indicators were higher in masked populations compared with the naive population (table 1), showing that our approach helps retrieving a lower fraction of false signals compared with simply studying the admixed genomes. Although setting signal threshold to 99.9% reduced the hits significantly which, in return, made the comparison difficult (Supplementary table S2), 99.5% threshold results mostly showed improvement using masked populations.

Comparison of XP-EHH Scores Based on Forward Simulations

Since XP-EHH cannot be performed with missing chunks, masked populations could not be utilized in a straightforward manner. We, therefore, created ancestral random breeders (ARBs, see methods for more details), full genomes with no gaps made by combining together masked haplotypes from multiple individuals, as described in Yelmen et al. (2019), resulting in ARB_N and ARB_S individuals. This approach assigns ancestries to the chunks of admixed genomes and combines matching ancestries to create whole genomes, bypassing the lack of a proper imputation panel when either of the ancestry is no longer available in unadmixed form. Therefore, they can be seen as "Frankensteins" of stitched haplotypes of the same ancestry. Similar to the PBS analysis, we performed XP-EHH on ARB (filled in after local ancestry deconvolution) and naive (prior to any type of processing) genomes, and compared these with the results from French

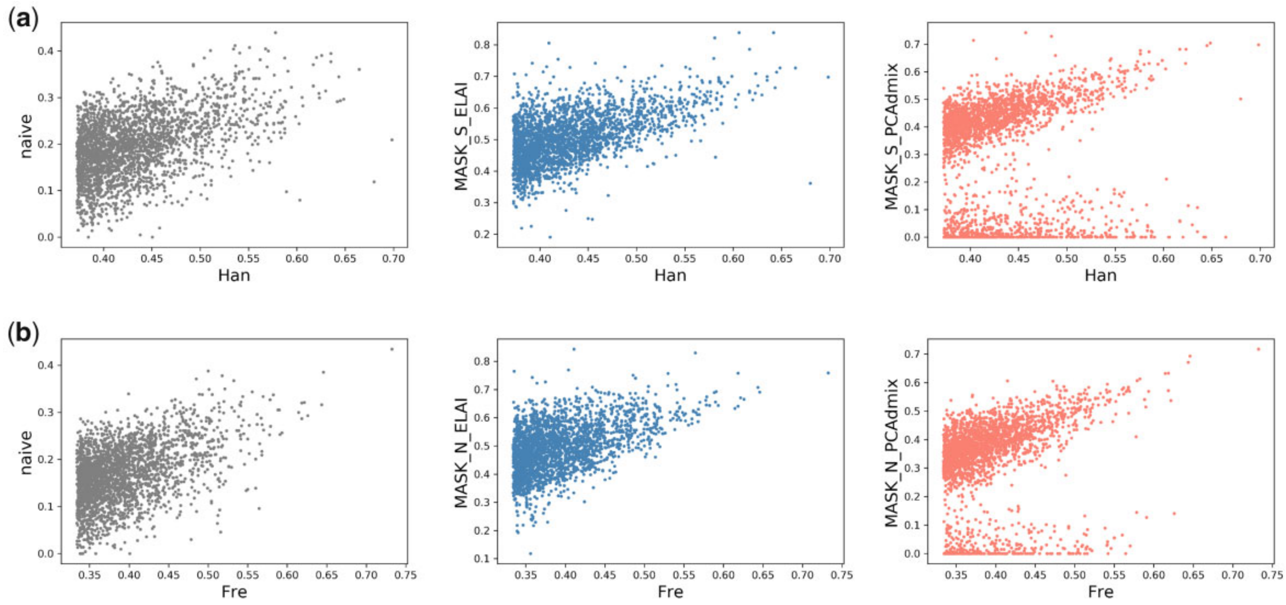

Fig. 1.—SNP by SNP PBS comparison for SNPs with PBS values above 99% threshold (SNPs selected based on Han and French source population scores) using Spearman's correlation. (*a*) Han vs naive (correlation coefficient: 0.409, 95% confidence interval: 0.377–0.441, *p*-value <2.2e−16, *n* = 2,529), Han vs MASK_S_ELAI (correlation coefficient: 0.510, 95 percent confidence interval: 0.479–0.542, *p*-value <2.2e−16, *n* = 2,529), Han versus MASK_S_PCAdmix (correlation coefficient: 0.210, 95% confidence interval: 0.169–0.252, *p*-value <2.2e−16, *n* = 2,529). (*b*) French vs naive (correlation coefficient: 0.400, 95% confidence interval: 0.367–0.436, *p*-value <2.2e−16, *n* = 2,530), French versus MASK_N_ELAI (correlation coefficient: 0.452, 95% confidence interval: 0.422–0.483, *p*-value <2.2e−16, *n* = 2,530), French versus MASK_N_PCAdmix (correlation coefficient: 0.394, 95% confidence interval: 0.354–0.432, *p*-value <2.2e−16, *n* = 2,530).

Table 1

Comparison of Possible Selection Signals Based on Mean PBS Scores (Above 99.5% Noted as Positive) for 50 kb Windows with TP, FP, FN, TPR (Measuring the Fraction of Correctly identified positives), and FDR, Measuring Expected Fraction of FPs) Indicators for Each Tested Population (Admixed Naive along with PCAdmix and ELAI-Masked Populations, see Materials & Methods for details) Compared with True Source Population

	TP	FP	FN	TPR	FDR
Naive	25	39	29	0.46	0.61
MASK_S_ELAI	27	23	27	0.50	0.46
MASK_S_PCAdmix	25	18	29	0.46	0.42
Naive	50	33	60	0.45	0.40
MASK_N_ELAI	41	21	69	0.37	0.34
MASK_N_PCAdmix	68	38	42	0.62	0.36

and Han source populations (see Materials and Methods). This time, ARB populations performed worse in comparison to admixed naive population for position by position XP-EHH scores (Supplementary fig. S3), although both ARB and naive populations performed very poorly when we checked top 50 scoring SNPs in comparison to the source populations (Supplementary table S1b).

PBS on Real South Asian Populations

Given that applying XP-EHH on admixed populations proved to be imprecise using either ARB or admixed genomes, we then resorted to PBS as our only viable approach to detect signatures of selection in a case study: contemporary South Asian human populations. Genomic composition of South Asians can be characterized in a broad perspective as an admixture of West Eurasian and South Asian components (Reich et al. 2009; Chaubey et al. 2011; Metspalu et al. 2011; Moorjani et al. 2013; Basu et al. 2016; Lazaridis et al. 2016; Damgaard et al. 2018; Pathak et al. 2018; Yelmen et al. 2019). Even though recent studies shed more light into this composition (Narasimhan et al. 2019), there is still no available unadmixed aDNA attributed to the South Asian component for a selection scan to be carried out. In our previous study, we checked for local admixture imbalance between ancestral components of South Asians and detected some possible selection signals pointing to selective pressures acting after the admixture event (Yelmen et al. 2019). In this study, we instead focused on events that took place prior to the admixture of the two genomic components, and reported the first time possible selection signals for the South Asian component of contemporary South Asians (fig. 2). We additionally report the genes present within 50-kb windows with possible positive signals (Supplementary table 3) and top 20 scoring SNPs with related genes (Supplementary table S4). We stress that the PBS analysis could also detect postadmixture selection, for example, by searching for overlap between the signals found on each ancestry independently; however, postadmixture imbalance (Yelmen et al. 2019) remains the strategy of choice

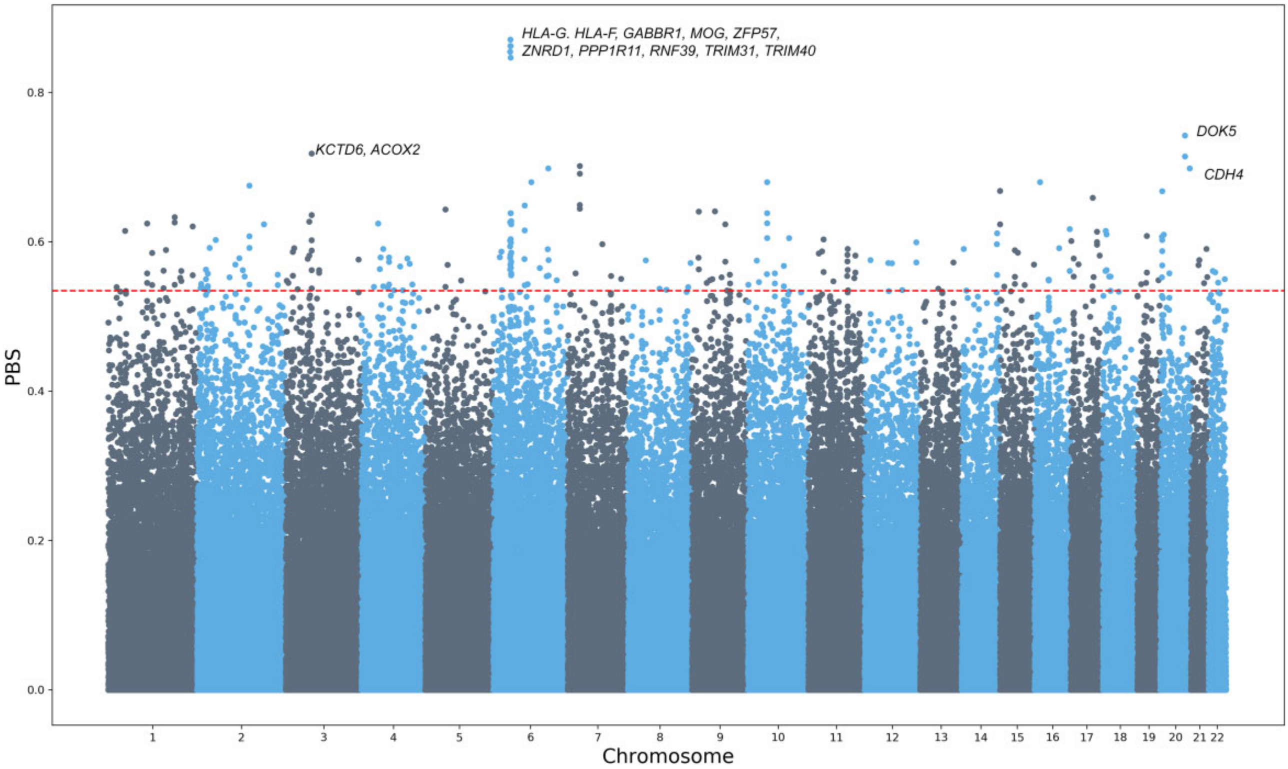

Fig. 2.—PBS on masked South Asian genomes (MASK_S) with the dashed line marking 99.9% threshold. Genes within 50-kb range of the highest peaks are annotated.

given the short evolutionary time elapsed after the admixture event.

Discussion

In our study, we explored the possibility of improving selection tests on admixed populations using local ancestry deconvolution approaches. Following our previous work (Yelmen et al. 2019), we created surrogates of components from a simulated admixed population both in the form of masked and reconstructed genomes. Knowing the true source populations which we merged to create the admixture, we were able to compare selection scores and observed an increased performance in the site-specific PBS using masked populations instead of the admixed one. However, it is important to note that even with this improvement, the scores are far from matching the ones obtained from the true source population in our simulations (Supplementary Table 1). Additionally, although we calculated the true positive rate and precision for window-based comparison, due to the vast number of true negatives, it was not possible to obtain other meaningful indicators such as accuracy or false positive rate.

On the other hand, haplotype aware XP-EHH analyses revealed that both reconstructed and admixed genomes performed poorly compared with the benchmark (Supplementary table 1b), with the naive approach being slightly better than our proposed solution in the case of haplotype-based methods. This might be an expected outcome since combining chunks from different individual genomes disrupts haplotype integrity (Yelmen et al. 2019), although it should still be seen as a strategy to be preferred over simple imputation in the absence of a suitable reference panel. Nonetheless, we found some regions above the threshold level (>2) in the admixed dataset, though there we never simulated any selection (after admixture). This suggests that the selection signal found in an admixed population using XP-EHH (or similar selection tests) does not necessarily advocate for the selection to have happened in that admixed population. It might be a remnant effect coming from source populations where the variant was already selected. Moreover, some signals detected in the admixed populations were not present in the source populations (Supplementary fig. 4). These findings suggest that XP-EHH and haplotype-aware methods are probably not a good option for selection analysis in admixed populations.

In addition to the analyses on simulated data, we also performed PBS on real masked genomes and reported the first possible selection signals for the South Asian component of contemporary South Asian populations (fig. 1, Supplementary tables S2 and S3). Some possible signals included high-scoring SNPs from chromosome 6 within 50 kb of *HLA-G* and *HLA-F* and *TRIM31* and *TRIM40* genes related to immune system (Ishitani et al. 2003; Rajagopalan and Long 2012; Fu et al. 2017; Liu et al. 2017; Zhao et al. 2017; Lin and Yan 2019) along with *KCTD6* related to sweet taste signaling pathway (Liu et al. 2013), *ACOX2* related to branched fatty acid processing (Bjørklund et al. 2015; Vilarinho et al. 2016), *DOK5* coding for adapter proteins involved in signal transduction (Favre et al. 2003), and *CDH4* that is thought to be involved in brain segmentation and neuronal growth (Babb et al. 2005). These highlight for the first time putative selection signals that took place in the autochthonous South Asian population, as yet unsampled in its unadmixed form. However, it is important to note that signals related to *HLA* genes should be interpreted very cautiously due to difficulties related to variant calling and genotyping for that region. Furthermore, in our previous work (Yelmen et al. 2019), we found this region to be under unbalanced admixture between "North" and "South" South Asian components, with "North" haplotypes to appear preferred over "South" haplotypes.

In conclusion, our work shows that applying selection scans on admixed populations of any diploid organism with no prior deconvolution yields several off target results, while a preliminary local ancestry deconvolution step may help improve the detection of true signals, in the case of PBS. On the other hand, more research is required to assess the effect of different local ancestry methods on detecting true selection signals since we detected varying results between PCAdmix and ELAI applications. Our results also show that haplotype-aware methods to detect selection may be severely impaired in presence of recent admixture between highly divergent populations. Overall, our results should inform future studies in the field to be cautious when reporting selection scans from human or other organisms where recent admixture has been detected, with particular reference to those populations where at least one of the admixing sources is no longer available in its unadmixed form due to complete assimilation or extinction (in case of wild species) of that particular group.

Materials and Methods

Samples and Simulations

We used SNP array data of French, Han, Japanese, Yoruba (Li et al. 2008), and German (Yunusbayev et al. 2015) populations along with masked South Asian populations (Brahmin Gujarati, Gujarati, Khatri, Maratha, Pallan, Chamar, Dharkar, Kanjar, Gujjar, and Ror) (Metspalu et al. 2011; Basu et al. 2016; Pathak et al. 2018) with same sample sizes as described in Yelmen et al. 2019 (see Supplementary table S5). A simulated admixed population consisting of 500 samples was created via admix-simu (Williams 2016) using French and Han as admixing groups. First, we only kept positions whose minor allele frequency is more than 1% (using vcftools –maf command), remove all the indels and kept only SNPs (using vcftools –remove-indels) and kept only the first biallelic position in case of multiple allele (using bcftools –norm -d all). After filtering, we used admix-simu to simulate the admixed populations, which were admixed 200 generations ago with 50/50 contributions from the two sources.

Local Ancestry Inference

Local ancestry inference was performed with PCAdmix (Brisbin et al. 2012) setting window size to ten SNPs after default LD pruning ($r^2 > 0.8$, based on a built-in window size), a value suggested by the software's developers as the smallest size for reliable resolution. Additionally, we only used windows assigned to ancestral proxies with posterior probabilities (provided in the PCAdmix .fbk output) higher than 0.95, based on simulation tests described in Yelmen et al. 2019. For the simulated admixture events we used Germans and Japanese as reference groups, given that French and Han were used as mixing populations. For the South Asian real populations, we used French and Paniya as proxies for "North" South Asian (N) and "South" South Asian (S) ancestry donors, respectively, following Yelmen et al. 2019. For ELAI (Guan 2014), we used -mg 100 to define the admixture event that happened 100 generations ago. We find that using -mg 200 gives spurious results, although the true admixture in this case was 200 generations ago. Further we used -s 20 for having 20 steps for EM. We ran 10 times independently for ELAI and then averaged it out, and used the same source populations described for the PCAdmix approach.

Haplotype Masking

For each local ancestry method, we kept the regions assigned to either ancestry with at least 95% confidence and marked all the other regions as unknown or unassigned. Based on ancestry assigned genomic chunks, two sets (PCAdmix and ELAI) of 500 MASK_N and 500 MASK_S individuals were created to be used in PBS analysis. These haploid individuals retained information only for the confidently assigned sites and showed gaps or "NA" for the rest of their genome.

Ancestral Random Breeders

Additionally, to overcome issues introduced by the missing data and by the lack of suitable imputation panels for the "South" South Asian component, we generated 20 ARBs for each ancestry, using ELAI as the local ancestry method, to be used in XP-EHH analysis. ARBs were created for each N and S ancestry, by taking all the MASK_N or MASK_S individuals and replenishing the masked-out haplotypes by randomly picking (with replacement) a nonmasked haplotype from another donor within the same ancestry assignation at that locus. This process hence created a number of ARB haploid individuals which feature the genetic makeup of the original individual used as a scaffold, and, where not available, a random set of haplotypes from a given ancestry drawn from that individual's population. The reconstructed ARB population can therefore be seen as a set of random breeders, to be considered as the best available proxy to the actual ancestry source within the studied population. To control for fluctuations in allele frequencies introduced by this drawing approach, we worked only on a small number of recipient ARBs (20 individuals, from the simulated populations or from the whole set of available South Asian genomes) by retaining only positions for which a minimum number of available donor haplotypes were available in a given ancestry status to minimize the donation of the same haplotype to multiple ARBs, and by maximizing the length of the donated haplotype and its affinity to the surrounding sequences of a given receiving ARB, to minimize the number of ancestry switches artificially introduced by the ARB making process. This method has been initially described in Yelmen et al. (2019) to which we refer for further details and testing.

Population Branch Statistic

We used scikit-allel package (Miles et al. 2020) for calculating PBS score for each available position using allel.pbs function with window_size = 1 and window_step = 1. Two different sets of analyses were performed for simulated data. PBS for MASK_N_ELAI, MASK_N_PCAdmix, naive (as an admixture of French and Han), and French group was performed using Yoruba and Japanese outgroups [PBS(N, Japanese, Yoruba)]. PBS for MASK_S_ELAI, MASK_S_PCAdmix, naive (as an admixture of French and Han) and Han group was performed using Yoruba and German outgroups [PBS(S, German, Yoruba)]. For window-based analysis, genomes were divided into 50-kb regions and mean PBS was calculated for each region. We defined 99.9% and 99.5% thresholds as possible selection signals for each region.

Cross -Population Extended Haplotype Homozygosity

We used scikit-allel package (Miles et al. 2020) for calculating XP-EHH score for each available position using default parameters with allel.xpehh function. Two different sets of analyses were performed for simulated data. XP-EHH for ARB_N, naive (as an admixture of French and Han), and French group was performed using Japanese outgroup. XP-EHH for ARB_S, naive (as an admixture of French and Han), and Han group was performed using German outgroup.

Computing Correlation Coefficients and Sensitivity/ Specificity Measures

We calculated Spearman's correlation coefficient with R version 3.6.3 (Development Core Team R 2020) to assess the SNP by SNP correlation of PBS and XP-EHH scores. We used bootstrapping with 1,000 replicates to acquire 95 percent confidence intervals. For 50 kb window analyses, we calculated the true positive rate (TPR) and false discovery rate (FDR) as

$$TPR = \frac{TP}{TP + FN}$$

$$FDR = \frac{FP}{FP + TP}$$

where TP is true positives, FN is false negatives, and FP is false positives.

Data Availability

All data used in this article are publicly available from the original publications and on https://evolbio.ut.ee link. All data used for the analyses were obtained from literature, no new sample data were collected.

Acknowledgements

This work was supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.16-0024, MOBTT53: L.P., L.M., D.M., B.Y.; Project No. MOBEC008: M.M.); Horizon 2020 research and innovation programme under grant no. 810645: MM; the Estonian Research Council grant PUT (PRG243): L.P. Thanks to High Performance Computing Center of the University of Tartu for providing computational resources.

Literature Cited

Babb SG, et al. 2005. Zebrafish R-cadherin (Cdh4) controls visual system development and differentiation. Dev Dyn. doi:10.1002/dvdy.20431.

Basu A, Sarkar-Roy N, Majumder PP. 2016. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci Usa. 113(6):1594–1599. doi:10.1073/pnas.1513197113.

Bjørklund SS, et al. 2015. Expression of an estrogen-regulated variant transcript of the peroxisomal branched chain fatty acid oxidase ACOX2 in breast carcinomas. BMC Cancer. 15(1):1–13. doi:10.1186/s12885-015-1510-8.

Brisbin A, et al. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol. 84(4):343–364. doi:10.3378/027.084.0401.

Chaubey G, et al. 2011. Population genetic structure in indian austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Mol Biol Evol. 28(2):1013–1024. doi:10.1093/molbev/msq288.

Damgaard Peter de Barros, et al. 2018. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. Science (80-). 360(6396):eaar7711.

Development Core Team R. 2020. R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: https://www.r-project.org/. Accessed February 1, 2021.

Favre C, et al. 2003. DOK4 and DOK5: new dok-related genes expressed in human T cells. Genes Immun. 4(1):40–45. doi:10.1038/sj.gene.6363891.

Fu B, et al. 2017. Natural Killer Cells Promote Fetal Development through the Secretion of Growth-Promoting Factors. Immunity. 47(6):1100–1113.e6. doi:10.1016/j.immuni.2017.11.018.

Galaverni M, et al. 2017. Disentangling timing of admixture, patterns of introgression, and phenotypic indicators in a hybridizing Wolf population. Mol Biol Evol. 34(9):2324–2339. doi:10.1093/molbev/msx169.

Guan Y. 2014. Detecting structure of haplotypes and local ancestry. Genetics. 196(3):625–642. doi:10.1534/genetics.113.160697.

Huerta-Sánchez E, et al. 2013. Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. Mol Biol Evol. 30(8):1877–1888. doi:10.1093/molbev/mst089.

Huerta-Sánchez E, et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 512(7513):194–197. doi:10.1038/nature13408.

Ishitani A, et al. 2003. Protein Expression and peptide binding suggest unique and interacting functional roles for HLA-E, F, and G in maternal-placental immune recognition. J Immunol. 171(3):1376–1384.

Lazaridis I, et al. 2016. Genomic insights into the origin of farming in the ancient Near East. Nature. 536(7617):419–424. doi:10.1038/nature19310.

Li JZ, et al. 2008. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science. 319(5866):1100–1104.

Lin A, Yan WH. 2019. The emerging roles of human leukocyte antigen-F in immune modulation and viral infection. Front Immunol. 10:964. doi:10.3389/fimmu.2019.00964.

Liu B, et al. 2017. The ubiquitin E3 ligase TRIM31 promotes aggregation and activation of the signaling adaptor MAVS through Lys63-linked polyubiquitination. Nat Immunol. 18(2):214–224. doi:10.1038/ni.3641.

Liu Z, Xiang Y, Sun G. 2013. The KCTD family of proteins: structure, function, disease relevance. Cell Biosci. 3(1):45.doi:10.1186/2045-3701-3-45.

Metspalu M, et al. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet. 89(6):731–744. doi:10.1016/j.ajhg.2011.11.010.

Miles A, et al. 2020. cggh/scikit-allel: v1.3.2. Available from: 10.5281/zenodo.3976233. Accessed December 1, 2020.

Moorjani P, et al. 2013. Genetic evidence for recent population mixture in India. Am J Hum Genet. 93(3):422–438. doi:10.1016/j.ajhg.2013.07.006.

Narasimhan VM, et al. 2019. The formation of human populations in South and Central Asia. Science. 365(6457):eaat7487. doi:10.1126/science.aat7487.

Pathak AK, et al. 2018. The Genetic Ancestry of Modern Indus Valley Populations from Northwest India. Am J Hum Genet. 103(6):918–929. doi:10.1016/j.ajhg.2018.10.022.

Pierron D, et al. 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat Commun. 9(1):1–9. doi:10.1038/s41467-018-03342-5.

Rajagopalan S, Long EO. 2012. Cellular senescence induced by CD158d reprograms natural killer cells to promote vascular remodeling. Proc Natl Acad Sci U S A. 109(50):20596–20601. doi:10.1073/pnas.1208248109.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461(7263):489–494. doi:10.1038/nature08365.

Sabeti PC, The International HapMap Consortium, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature. 449(7164):913–918. doi:10.1038/nature06250.

Vilarinho S, et al. 2016. ACOX2 deficiency: a disorder of bile acid synthesis with transaminase elevation, liver fibrosis, ataxia, and cognitive impairment. Proc Natl Acad Sci U S A. doi:10.1073/pnas.1613228113.

Williams A. 2016. admix-simu: admix-simu: program to simulate admixture between multiple populations. Available from: 10.5281/zenodo.45517. Accessed November 1, 2020.

Yelmen B, et al. 2019. Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations. Mol Biol Evol. 36(8):1628–1642. doi:10.1093/molbev/msz037.

Yi X, et al. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science. 329(5987):75–78.

Yunusbayev B, et al. 2015. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. PLoS Genet. 11(4):1–24.

Zhao C, et al. 2017. The E3 Ubiquitin Ligase TRIM40 Attenuates Antiviral Immune Responses by Targeting MDA5 and RIG-I. Cell Rep. 21(6):1613–1623. doi:10.1016/j.celrep.2017.10.020.

Associate editor: Dr. Tanja Slotte