Letter to the Editor

Revisiting Adaptive Introgression at the HLA Genes in Lithuanian Genomes with Machine Learning

Josef Hackl[1], Xin Huang[1, 2, *]

1 Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria

2 Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria

* Corresponding author: xin.huang@univie.ac.at

Keywords: Adaptive introgression; Balancing selection; HLA; Population genetics; Machine learning

Dear Editor,

We are writing to discuss the article titled 'Disentangling Archaic Introgression and Genomic Signatures of Selection at Human Immunity Genes,' published by Urnikyte et al (2023). This study employed an *ad-hoc* approach, first applying the machine learning tool, ArchIE (Durvasula and Sankararaman 2019), to detect introgression candidates, followed by the use of the iHS statistic (Voight et al. 2006) to identify candidates under positive selection. According to the authors, the *HLA-C* gene displays both introgression and positive selection signals, suggesting it as a candidate for adaptive introgression in Lithuanians. However, this approach is problematic due to the varying effectiveness of the methods employed (Zhang et al. 2023) and the confounding effects of introgression on methods used to detect selection (Racimo et al. 2015). More specifically, adaptive introgression can be confounded by balancing selection (Fijarczyk and Babik 2015), and the human leukocyte antigen (HLA) genes are well known examples for long-term balancing selection (Andrés et al. 2009; Gelabert et al. 2024). Considering this, we reanalyzed the Lithuanian genomic data using a recently developed machine learning approach, MaLAdapt (Zhang et al. 2023), which is specifically designed to detect adaptive introgression through supervised learning. Our results suggest that the HLA genes are not candidates for adaptive introgression.

We downloaded the Lithuanian genomes from Urnikyte et al. (2023), as well as the chromosome 6 variants of the Altai Neanderthal (Prüfer et al. 2014) from http://cdna.eva.mpg.de/neandertal/Vindija/ and the chromosome 6 variants of modern humans identified by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) from https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Since the HLA genes are located on human chromosome 6, our analysis focused exclusively on this

chromosome. We used BEAGLE 5.4 (version: 06Aug24.a91) to phase and impute missing genotypes in these genomes, utilizing all European populations (CEU, FIN, GBR, IBS, TSI) from the 1000 Genomes Project as the reference panel. In BEAGLE 5.4 (Browning et al. 2018; Browning et al. 2021), we set the argument ne = 10,000 for effective population size, with other parameters left as default. Next, we extracted biallelic single nucleotide polymorphisms (SNPs) from the imputed data and merged them with biallelic SNPs from the YRI population in the 1000 Genomes Project, as the YRI population is considered a reference population that did not experience introgression (Huang et al. 2022). We applied MaLAdapt to extract various input features using sliding windows of 50,000 base pairs with a step size of 10,000 base pairs from this merged dataset. Finally, we used the pretrained model provided by MaLAdapt to predict the probability of adaptive introgression for each 50 kb window from these input features. We estimated the probability of adaptive introgression for each HLA gene by either averaging or taking the maximum of the probabilities from the 50 kb windows that overlap with the gene. Our results (Figure 1) show that the probabilities of adaptive introgression for the HLA genes fall within the middle range of the empirical distribution, which indicates that they are not outliers. Since adaptive introgression is rare and typically identified through outliers, these findings suggest that the HLA genes are not candidates for adaptive introgression.

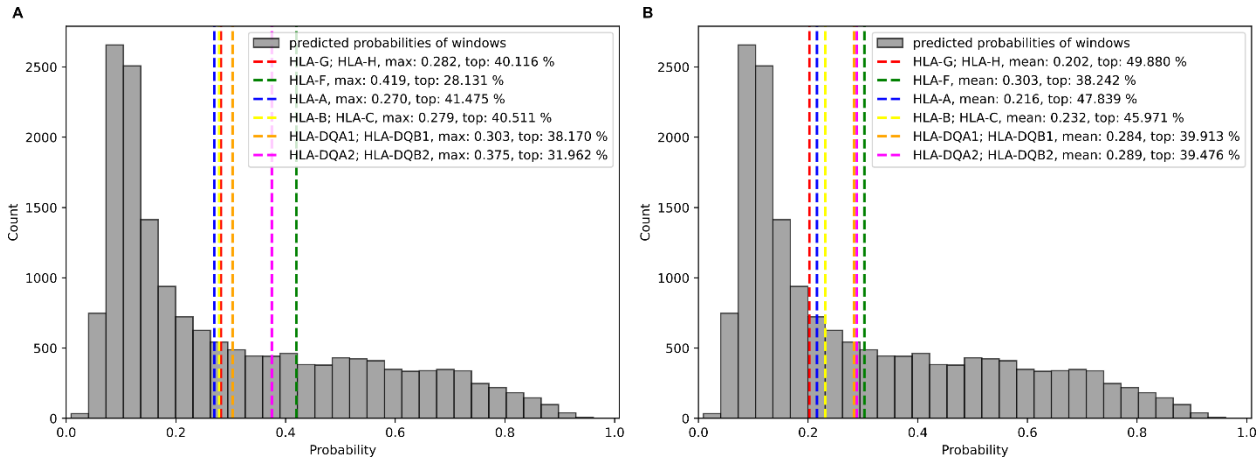

Figure 1. Distribution of adaptive introgression probabilities on chromosome 6 in Lithuanian genomes predicted by MaLAdapt using the pretrained model MaLAdapt_25_-sweep-all_model. A, the adaptive introgression probabilities for the HLA genes by taking the maximum of the probabilities from the 50 kb windows that overlap with the genes; B, the adaptive introgression probabilities for the HLA genes by averaging the probabilities from the 50 kb windows that overlap with the genes. The grey bars represent the distribution of adaptive introgression probabilities for all 50 kb windows across chromosome 6 in the Lithuanian genomes. The dashed lines indicate the adaptive introgression probabilities for the HLA genes. We downloaded the pretrained model from

https://drive.google.com/drive/folders/10r8e5WbhcgAIjC0DVmIe4saVYODRgFCO?usp=share_link on June 5, 2024, and used the ranges of the HLA genes as outlined in Table 1 of Urnikyte et al. (2023). The hg19 coordinates are as follows: *HLA-G; HLA-H*: 29798610–29897944; *HLA-F*: 29698426–29746527; *HLA-A*: 29898105–29947740; *HLA-B; HLA-C*: 31198205–31348022; *HLA-DQA1; HLA-DQB1*: 32598042–32644388; *HLA-DQA2; HLA-DQB2*: 32698044–32748039; *HLA-DOB*: 32748045–32797488.

We also applied the B1 statistic from BetaScan (Siewert and Voight 2017) to identify candidates for long-term balancing selection on chromosome 6 in the Lithuanian genomes. For the scan, only SNPs with minor allele frequencies greater than 0.05 in the imputed data were used. SNPs located in regions defined by the RepeatMasker table, simple repeats table, and segmental duplication table from the UCSC Table Browser (hg19 coordinates, last accessed in October 2024) were removed. Additionally, SNPs with *p*-values less than 10^{-3} from exact Hardy-Weinberg equilibrium tests in each population, performed using PLINK 1.9 (Chang et al. 2015), were excluded. As per Siewert and Voight (2017), only SNPs with folded allele frequencies greater than 0.15 were used as cores for calculating the B1 scores. All other parameters in BetaScan were kept at their default values. Our results show a peak in B1 scores within the HLA genes, particularly in *HLA-B; HLA-C* (highest B1 score of 36.350388), *HLA-DQA1; HLA-DQB1* (highest B1 score of 71.446036), and *HLA-DQA2; HLA-DQB2* (highest B1 score of 43.007245), suggesting they are candidates for long-term balancing selection. This is consistent with previous studies (DeGiorgio et al. 2014; Bitarello et al. 2018) and was also noted by Urnikyte et al. (2023).

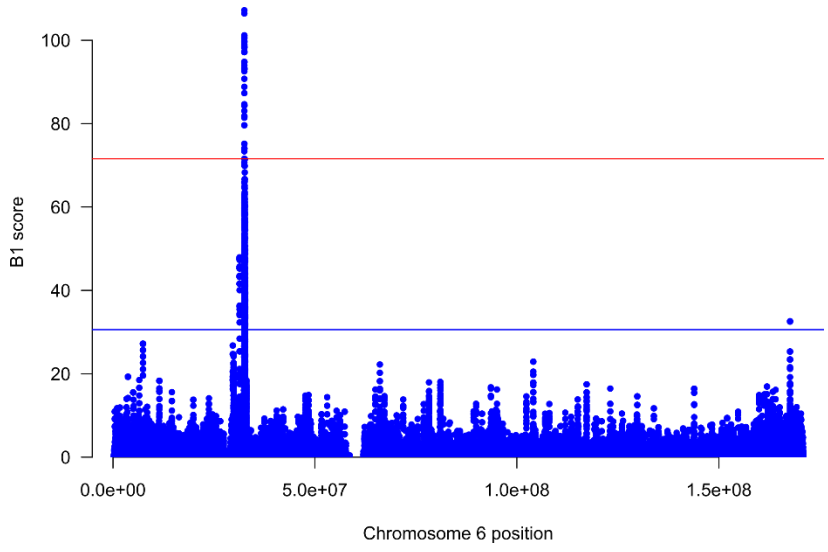

Figure 2. Manhattan plot of B1 scores on chromosome 6 in Lithuanian genomes. The red horizontal line (B1 = 71.572528) represents the top 0.05%, and the blue horizontal line (B1 = 30.589559) represents the top 1%. This plot was created using the qqman package (Turner 2018).

It remains controversial whether the HLA genes are under balancing selection or if they experienced adaptive introgression from archaic humans (Ding et al. 2014; Yasukochi and Ohashi 2017). Recently developed machine learning-based methods for detecting adaptive introgression, such as genomatnn and MaLAdapt, have not reported signals of adaptive introgression in the HLA genes using populations from the 1000 Genomes Project (Gower et al. 2021; Zhang et al. 2023). However, the HLA regions have been consistently identified as candidates for balancing selection in both modern and ancient human populations across various recent studies employing different approaches (Siewert and Voight 2017; Bitarello et al. 2018; Gelabert et al. 2024). Moreover, a recent method, based on the ancestral recombination graph, strongly supports balancing selection in the HLA regions with trans-species polymorphism (Deng et al. 2024). Considering that polymorphisms maintained by balancing selection are typically shared across populations or species (Hedrick 2007; Bitarello et al. 2023), it is likely that the HLA genes in Lithuanian genomes are maintained by balancing selection shared among various human populations. Although a study by Abi-Rached et al. (2011) suggested that the *HLA-B* locus was under adaptive introgression, they used a simulator that assumed neutrality at this locus, even though the classical class I *HLA* loci are well-known examples of balancing selection (Yasukochi and Ohashi 2017).

We would like to point out a similar issue with the machine learning tool, ArchIE, used by Urnikyte et al. (2023), which relies on the ms simulator based on the Wright–Fisher neutral model (Hudson 2002) to generate training data. Since ms cannot simulate data under natural selection, this raises concerns about how ArchIE performs when analyzing data that includes natural selection. Simulation misspecification can impact the performance of supervised learning tools (Mo and Siepel 2023). Therefore, it is critical to document the specific demographic model used for generating the simulated data, which was not reported in Urnikyte et al. (2023). If Urnikyte et al. (2023) used the demographic model hard-coded in ArchIE, an additional issue arises: the ArchIE code trains on data simulated from a four-population model, whereas a three-population model was reported (Huang 2024). This discrepancy may also affect the performance of ArchIE. Furthermore, a recent study (Ray et al. 2024) highlighted the importance of balancing the training data to achieve results similar to those originally reported by ArchIE, as introgression is rare and only a small proportion of the training data contains introgressed fragments. It remains unclear whether Urnikyte et al. (2023) balanced their training data—ensuring a similar amount of non-introgressed and introgressed fragments—before applying ArchIE to detect archaic introgressed fragments in Lithuanian genomes. Hence, it is important to thoroughly document the details when applying machine learning approaches (Walsh et al. 2021), as factors like data preprocessing, training data, and hyperparameters can significantly impact the final performance of these models. Using version control tools with code hosting platforms like GitHub, model hosting platforms like Hugging Face, and reproducible workflow

management systems like Snakemake (Mölder et al. 2021) helps document the details and ensures that computational steps can be easily reproduced or modified by others (Huang 2024).

As interest in applying machine learning, particularly deep learning, to population genetics and evolutionary biology continues to grow (Huang et al. 2024), it is crucial for researchers to understand the underlying principles. For instance, since ArchIE is trained using simulated data that does not account for natural selection, its performance on data with natural selection should be carefully examined when applied in such contexts. Additionally, it is essential to develop robust machine learning applications that allow users to easily comprehend and adapt them to their own data (Huang 2024). One limitation of our study is that we used the pretrained model provided by MaLAdapt, which was trained on a specific human demographic model (Zhang et al. 2023). Currently, retraining MaLAdapt for a specific dataset is challenging due to its implementation. Reduced performance of MaLAdapt has been observed in other species (Romieu et al. 2024), likely due to demographic model misspecification. However, since the pretrained model was trained on a human population setting that includes Eurasians, to which the Lithuanians belongs, we expect the reduction in moder performance in our analysis to be minimal.

Acknowledgements

J.H. and X.H. thank Martin Kuhlwilm for discussions and comments on the manuscript; and the Life Science Compute Cluster at the University of Vienna for providing computing resources.

Competing interests

J.H. and X.H. declare no conflict of interests.

Author contributions

X.H. designed the study. J.H. and X.H. analyzed the data and wrote the manuscript.

Data availability

The Snakemake workflow for reproducing the analysis can be found in https://github.com/xin-huang/Lithuanian-archaic-introgression, last accessed October 10, 2024.

References

Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., Kimani, J., Carringtong, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S.G.E., Maisers, M., Guethlein, L.A., Tavoularis, S., Little, A., Green, R.E.,

Norman, P.J., Parham, P., 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**, 89–94.

Andrés, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustanmante, C.D., Clark, A.G., Nielsen, R., 2009. Targets of balancing selection in the human genome. *Mol Biol Evol* **12**, 2755–2764.

Bitarello, B.D., de Filippo, C., Teixeira, J.C., Schmidt, J.M., Kleinert, P., Meyer, D., Andrés, A.M., 2018. Signatures of long-term balancing selection in human genomes. *Genome Biol Evol* **10**, 939–955.

Bitarello, B.D., Brandt, D.Y.C., Meyer, D., Andrés, A.M., 2023. Inferring balancing selection from genome-scale data. *Genome Biol Evol* **15**, evad032.

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A one-penny imputed genome from next generation reference panels. *Am J Hum Genet* **103**, 338–348.

Browning, B.L., Tian, X., Zhou, Y., Browning, S.R., 2021. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108**, 1880–1890.

Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047-8.

DeGiorgio, M., Lohmueller, K.E., Nielsen, R., 2014. A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet* **10**, e1004561.

Ding, Q., Hu, Y., Jin, L., 2014. Non-Neanderthal origin of the HLA-DPB1*0401. *J Biol Chem* **289**, 10252.

Deng, Y., Nielsen, R., Song, Y.S., 2024. Robust and accurate Bayesian inference of genome-wide genealogies for large samples. bioRixv. https://doi.org/10.1101/2024.03.16.585351, last accessed October 10, 2024.

Durvasula, A., Sankararaman, S., 2019. A statistical model for reference-free inference of archaic local ancestry. *PLoS Genet* **15**, e1008175.

Fijarczyk, A., Babik, W., 2015. Detecting balancing selection in genomes: Limits and prospects. *Mol Ecol* **14**, 3529–3545.

Gelabert, P., Bickle, P., Hofmann, D., Teschler-Nicola, M., Anders, A., Huang, X., Hämmerle, M., Olalde, I., Fournier, R., Ringbauer, H., Akbari, A., Cheronet, O., Lazaridis, I., Broomandkhoshbacht, N., Fernandes, D.M., Buttinger, K., Callan, K., Candilio, F., Bravo, G., Curtis, E., Ferry, M., Keating, D., Freilich, S., Kearns, A., Harney, É., Lawson, A.M., Mandl, K., Michel, M., Oberreiter, V., Zagorc, B., Oppenheimer, J., Sawyer, S., Schattke, C., Ozdogan, K.T., Qiu, L., Workman, J.N., Zalzala, F., Mallick, S., Mah, M., Micco, A., Pieler, F., Pavuk, J., Šefčáková, A., Lazar, C., Vasic, R., Starovic, A., Djuric, M., Škrivanko, M.K., Šlaus, M., Bedić, Ž., Novotny, F., Szabó, L.D., Cserpák-Laczi, O.,

Hága, T., Hajdú, Z., Mirea, P., Nagy, E.G., Virág, Z.M., Horváth, A.M., Horváth, L.A., Biró, K.T., Domboróczki, L., Szeniczey, T., Jakucs, J., Szelekovszky, M., Zoltán, F., Sztáncsuj, S., Tóth, K., Csengeri, P., Pap, I., Patay, R., Putica, A., Vasov, B., Havasi, B., Sebők, K., Raczky, P., Lovász, G., Tvrdý, Z., Rohland, N., Novak, M., Ruttkay, M., Krošláková, M., Bátora, J., Cheben, I., Boric, D., Dani, J., Kuhlwilm, M., Palamara, P.F., Hajdu, T., Pinhasi, R., Reich, D., 2024. Social and genetic diversity in the first farmers of Central Europe. *Nature Hum Behav*.

Gower, G., Picazo, P.I., Fumagalli, M., Racimo, F., 2021. Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife* **10**, e64669.

Hedrick, P.W., 2007. Balancing selection. *Curr Biol* **17**, R230–R231.

Huang, X., Kruisz, P., Kuhlwilm, M., 2022. sstar: A Python package for detecting archaic introgression from population genetic data with *S**. *Mol Biol Evol* **39**, msac212.

Huang, X., 2024. Developing machine learning applications for population genetic inference: Ensuring precise terminology and robust implementation. EcoEvoRixv. https://doi.org/10.32942/X2N90M, last accessed October 10, 2024.

Huang, X., Rymbekova, A., Dolgova, O., Lao, O., Kuhlwilm, M., 2024. Harnessing deep learning for population genetic inference. *Nat Rev Genet* **25**, 61–78.

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

Mo, Z., Siepel, A., 2023. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLoS Genet* **19**, e1011032.

Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. *F1000 Res* **10**, 33.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., Li, H., Mallick, S., Dannemann, M., Fu, Q., Kircher, M., Kuhlwilm, M., Lachmann, M., Meyer, M., Ongyerth, M., Siebauer, M., Theunert, C., Tandon, A., Moorjani, P., Pickrell, J., Mullikin, J.C., Vohr, S.H., Green, R.E., Hellmann, I., Johnson, P.L.F., Blanche, H., Cann, H., Kitzman, J.O., Shendure, J., Eichler, E.E., Lein, E.S., Bakken, T.E., Golovanova, L.V., Doronichev, V.B., Shunkov, M.V., Derevianko, A.P., Viola, B., Slatkin, M., Reich, D., Kelso, J., Pääbo, S., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49.

Racimo, F., Sankararaman, S., Nielsen, R., Huerta-Sánchez, E., 2015. Evidence for archaic adaptive introgression in humans. *Nat Rev Genet* **16**, 359–371.

Ray, D.D., Flagel, L., Schrider, D.R., 2024. IntroUNET: Identifying introgressed alleles via sematic segmentation. *PLoS Genet* **20**, e1010657.

Romieu, J., Camarata, G., Crochet, P.A., de Navascués, M., Leblois, R., Rousset, F., 2024. Performance evaluation of adaptive introgression classification methods. bioRixv. https://doi.org/10.1101/2024.06.12.598278, last accessed October 10, 2024.

Siewert, K.M., Voight, B.F., 2017. Detecting long-term balancing selection using allele frequency correlation. *Mol Biol Evol* **34**, 2996–3005.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**, 68–74.

Turner, S.D., 2018. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J Open Source Softw* **3**, 731.

Urnikyte, A., Masiulyte, A., Pranckeniene, L., Kučinskas, V., 2023. Disentangling archaic introgression and genomic signatures of selection at human immunity genes. *Infect Genet Evol* **116**, 105528.

Voight, B.F., Kudaravalli, S., Wen, X., Pritchard, J.K., 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72.

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., ELIXIR Machine Learning Focus Group, Harrow, J., Psomopoulos, F.E., Tosatto, S.C.E., 2021. DOME: Recommendations for supervised machine learning validation in biology. *Nat Methods* **18**, 1122–1127.

Yasukochi, Y., Ohashi, J., 2017. Elucidating the origin of *HLA-B*73* allelic lineage: Did modern humans benefit by archaic introgression? *Immunogenetics* **69**, 63–67.

Zhang, X., Kim, B., Singh, A., Sankararaman, S., Durvasula, A., Lohmueller, K.E., 2023. *MaLAdapt* reveals novel targets of adaptive introgression from Neanderthals and Denisovans in worldwide human populations. *Mol Biol Evol* **40**, msad001.