frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 November 2024
DOI 10.3389/fgene.2024.1493270

OPEN ACCESS

EDITED BY
Hovirag Lancioni,
University of Perugia, Italy

REVIEWED BY
Sanatan Tudu,
Wildlife Institute of India, India
Vikash Mishra,
Chimera Transplant Research Foundation, India

*CORRESPONDENCE
Alina Urnikytė,
alina.urnikyte@mf.vu.lt

RECEIVED 08 September 2024
ACCEPTED 28 October 2024
PUBLISHED 20 November 2024

CITATION
Daniūtė G, Pranckėnienė L, Pakerys J, Kloviņš J, Kučinskas V and Urnikytė A (2024) Populations of Latvia and Lithuania in the context of some Indo-European and non-Indo-European speaking populations of Europe and India: insights from genetic structure analysis. *Front. Genet.* 15:1493270. doi: 10.3389/fgene.2024.1493270

Populations of Latvia and Lithuania in the context of some Indo-European and non-Indo-European speaking populations of Europe and India: insights from genetic structure analysis

Gintė Daniūtė[1], Laura Pranckėnienė[2], Jurgis Pakerys[3], Jānis Kloviņš[4], Vaidutis Kučinskas[2] and Alina Urnikytė[2]*

[1]Department of Human and Medical Genetics, Biomedical Science Institute, Faculty of Medicine, Vilnius University, Vilnius, Lithuania, [2]Faculty of Medicine Population Genomics Laboratory, Translational health research Institute, Faculty of Medicine, Vilnius University, Vilnius, Lithuania, [3]Department of Baltic Studies, Institute for the Languages and Cultures of the Baltic, Vilnius University, Vilnius, Lithuania, [4]Latvian Biomedical Research and Study Centre, Riga, Latvia

The aim of this study was to investigate the relationship among Lithuanian, Latvian, Indian, and some other populations through a genome-wide data analysis of single nucleotide polymorphisms (SNPs). Limited data of Baltic populations were mostly compared with geographically closer modern and ancient populations in the past, but no previous investigation has explored their genetic relationships with distant populations, like the ones of India, in detail. To address this, we collected and merged genome-wide SNP data from diverse publicly available sources to create a comprehensive dataset with a substantial sample size especially from Lithuanians and Latvians. Principal component analysis (PCA) and admixture analysis methods were employed to assess the genetic structure and relationship among the populations under investigation. Additionally, we estimated an effective population size (*Ne*) and divergence time to shed light on potential past events between the Baltic and Indian populations. To gain a broader perspective, we also incorporated ancient and modern populations from different continents into our analyses. Our findings revealed that the Balts, unsurprisingly, have a closer genetic affinity with individuals from Indian population who speak Indo-European languages, compared to other Indian linguistic groups (such as speakers of Dravidian, Austroasiatic, and Sino-Tibetan languages). However, when compared to other populations from the European continent, which also speak Indo-European and some Uralic languages, the Balts did not exhibit a stronger resemblance to Indo-European-speaking Indians. In conclusion, this study provides an overview of the genetic relationship and structure of the populations investigated, along with insights into their divergence times.

KEYWORDS

Lithuanian population, Latvian population, Indian population, genetics, population structure

Introduction

The origins, spread, and timing of the Indo-European language family, which gave rise to the main languages of the Baltic countries and to many languages of India, have been subjects of debate. One of the main proposals, the "Steppe" hypothesis, suggests that the spread of Indo-Europeans occurred from the Pontic Steppe ca. 6,500 years BP (Gimbutas 1970; Mallory, 1989; Anthony, 2007) The alternative "Anatolian" hypothesis proposes that the expansion originated from Anatolia region and occurred much earlier, around 9,500–8,500 years BP (Renfrew, 1987; Bellwood, 2005). In addition to that, one of the recent studies proposed a hybrid model: here, the primary homeland is seen south of the Caucasus (as in the "Anatolian" hypothesis) while the later secondary homeland is located in the steppe region (in line with the "Steppe" hypothesis) (Heggarty et al., 2023). As for development of the languages that are of primary importance for our study, it is widely accepted that the Baltic languages grew out of the Balto-Slavic branch of the Proto-Indo-European, while the Indo-European languages of India are surely known to have evolved from the Indo-Iranian branch. The common linguistic features of the Balto-Slavic and the Indo-Iranian languages were inherited from the stages of diversification of the Proto-Indo-European and the two branches do not seem to have shared an exclusive common ancestor (Pronk, 2022; Kümmel, 2022). In some studies, however, a potential ancestor of the Indo-Iranian and the Balto-Slavic protolanguages is proposed and is referred to as Indo-(Iranian-Balto-)Slavic node (Ringe et al., 2002; Nakhleh et al., 2005; Olander, 2019); see the latest critical evaluation of this proposal in (Heggarty et al., 2023). Despite the well-researched linguistic relationship, there have been no detailed studies investigating the genetic relationship between geographically distant Baltic and Indian populations and the present study aims to fill this gap. The results of this study could potentially lead to a wider-scale examination of the genetic relationship between the Balto-Slavic and the Indo-Iranian speaking populations.

The territory of the present-day Baltic countries, Lithuania and Latvia, was first inhabited during the Final Paleolithic ca. 11,000–10000 years BP by the hunter-gatherers (HG) who were migrating as the climate was warming up after the last glaciation (Rimantienė, 1996; Zagorska 2001; Ostrauskas, 2008; Jochim, 2011). HG populations were dominant in the area up until ca. 5,000–4500 BP and left a lasting impact on the genetic profile of inhabitants of Lithuania and Latvia which are known to carry the largest share of west European HG ancestry in Europe (Urnikyte et al., 2019; Reščenko et al., 2023). An event of major importance for further formation of the Baltic populations was the arrival of Indo-European speaking populations which ca. 5,000–4,500 years BP reached Europe through migration waves from the Pontic Steppe (Allentoft et al., 2015; Haak et al., 2015). The migrant populations merged with the local ones and subsequently gave rise to the Baltic-speaking populations via the intermediate stage of Balto-Slavic unity (Pronk, 2022). Throughout the history, the Baltic populations were influenced by neighboring non-Indo-European speaking (Finnic) as well as Indo-European-speaking (Germanic, Slavic) populations. Despite certain admixture, the contemporary Lithuanians and Latvians are considered relatively homogeneous populations, with the genetic diversity of Baltic tribes diminishing over the past millennium due to the geographical isolation of these lands (Kasperavičiūtė et al., 2004; Pliss et al., 2015; Ruzgaitė et al., 2015; Urnikyte et al., 2019). In contrast, India is known for its high genetic, cultural and linguistic diversity (Majumder, 1998). Indo-European migrations have also played a significant role in the formation of the Indian population (Diamond and Bellwood, 2003; Pugach and Stoneking, 2015). Indian populations are divided into tribal and caste groups, with castes comprising most of the population (Basu et al., 2003). There are approximately 400 tribal groups in India, speaking hundreds of languages, which belong to the four major language families: the Indo-European (in northern India), the Dravidian (in southern India), the Sino-Tibetan (in northeastern India), and the Austroasiatic (in fragmented areas of eastern and central India) (Majumder and Basu, 2015).

Recent studies have shown that Lithuanian and Latvian populations are genetically close to their neighbors: Estonians (speaking a Finnic language of the Uralic language family), and Belarusians and Poles (speaking East and West Slavic languages of the Indo-European family respectively) (Urnikyte et al., 2021). Most Indian groups have inherited of their ancestry from the Ancestral North Indians (ANI) related to Middle Easterners, Caucasians, Central Asians, and Europeans, and Ancestral South Indians (ASI) who are distantly related to West Eurasians (Reich et al., 2009; Metspalu et al., 2011; Moorjani et al., 2013). Furthermore, individuals from higher castes within Indian population have been found to be genetically more closely related to Europeans than Asians (Heggarty et al., 2023; Bamshad et al., 2001; Thanseem et al., 2006; Bose et al., 2021).

Materials and methods

Samples

The baltic context

We made use of samples from 416 unrelated adult participants from two major ethnolinguistic groups (Aukštaitija (N = 218) and Žemaitija (N = 198)) in Lithuania. All participants indicated a minimum of three generations of Lithuanian ethnicity. The data used were from LITGEN project, already used in previous paper (Urnikyte et al., 2019). To get a diverse dataset of the Baltic populations, we merged Lithuanian samples with Latvian population samples. The 287 Latvian sample data is published by the Latvian Biomedical Research and Study center (LVBMC) (Reščenko et al., 2023). Participants reflect all the regions and the major ethnic groups of Latvia (Courland (N = 61), Semigallia (N = 24), Vidzeme (N = 147), Latgale (N = 55)) (see Supplementary Table S1.1). After merging we ended up with a total of 703 samples, and 153,244 SNPs. All the study participants signed written informed consent in compliance with the Declaration of Helsinki.

The Indian context

The genome wide SNP data of 456 Indians was obtained from the following publicly available sources: 102 samples from Telugu (South Asian Ancestry) ethnic group living in United Kingdom and 103 samples from Gujarati (South Asian Ancestry) ethnic group

currently residing in Houston, Texas (The 1000 Genomes Project Consortium et al., 2015); 140 samples from 29 ethnic groups from Metspalu et al., 2011 study (Metspalu et al., 2011); 22 samples from 6 ethnic groups from Tätte et al. (2019) study (Tätte et al., 2019); 44 samples from 3 ethnic groups from Pathak et al. (2018) study (Pathak et al., 2018); 19 samples from Parsi ethnoreligious group from Chaubey et al. (2017) study (Chaubey et al., 2017); 26 samples from 10 ethnic groups from Chaubey et al. (2010) study (Supplementary Table S1.2). In total, we generated a combined dataset of 343,295 SNPs from a total of 1,159 subjects from Lithuania (N = 416), Latvia (N = 287), and India (N = 456).

For further analysis, to see a broader view in the context of other world populations, the generated pooled dataset was merged with 36 samples from Estonian (Tambets et al., 2018) and with 17 populations from 1000 Genomes project (The 1000 Genomes Project Consortium et al., 2015) (Supplementary Table S1.3). In total this dataset consisted of 2,883 individuals and 343,295 SNP.

Lithuanian, Latvian, and Indian dataset was also merged with a dataset of ancient and recent samples from Eurasia. Besides Lithuanians, Latvians and Indians, this dataset consisted of 406 ancient and recent samples from Lazaridis et al. (2016) study (Lazaridis et al., 2016), 17 recent samples from Behar et al. (2013) study (Behar et al., 2013), 54 ancient samples from Mathieson et al., (2018) study (Mathieson et al., 2018) and 11 ancient samples from Mittnik et al. (2018) study (Mittnik et al., 2018) (Supplementary Table S1.4). In total this dataset consisted of 1,647 individuals and 13,522 SNPs.

Datasets were merged and processed using *plink* v1.90, *VCFtools* v.01.16 and *bcftools* v.1.14 programs (Purcell et al., 2007; Danecek et al., 2011; Danecek et al., 2021).

Methods

Principal component analysis, admixture, and outgroup f3-statistics

Principal component analysis (PCA) was performed with SmartPCA package from EIGENSOFT v7.2.1 (Patterson et al., 2006) on the pruned SNPs. SNPs in linkage disequilibrium were removed with the indep-pairwise option of PLINK (v1.07) using a window size of 50 SNPs, a step size of 5, and a r2 threshold of 0.5.

Inbreeding coefficients (F) among individuals from Lithuanian, Latvian and Indian populations was measured using plink v1.90 program (Purcell et al., 2007). Negative F values were changed to zeros as they probably represent sampling errors. Kinship was measured using KING v.2.2.5 (Manichaikul et al., 2010). PCA outliers, individuals having F than that expected for second cousin mating offspring ($F \geq 0.0156$) and second-degree relatives (kinship coefficient >0.0884) were omitted from all further analyses (in total 17 samples), except for Indian samples (Supplementary Table S2.1). PCA results were plotted using R v4.3.1 software ggplot2 package (Wickham, 2016).

Assessment of each individual genetic ancestry was performed with ADMIXTURE v1.3.0 (Alexander et al., 2009). In analysis we included both present-day and ancient samples with K from 2 to 12 with 10 iterations. The results were plotted using the PONG tool (Behr et al., 2016).

Outgroup f3-statistics were computed using qp3Pop program from ADMIXTOOLS V4.1 (Patterson et al., 2012). Mbuti was considered as outgroup in the analysis and calculated the shared drift between each putative ancient group and all the modern groups in the dataset in the form (Mbuti; Ancient, Modern).

Effective population size and divergence time

The long-term *Ne* and divergence time for the populations under study were estimated using R Package NeON (Neon, 2015) based on the genetic distance between SNPs. For the analysis we used binary PLINK files and updated genetic map information of the markers to calculate the *Ne* over time. First, the squared correlation coefficient of linkage disequilibrium (*r2LD*) within predefined recombination distance categories between markers is estimated. We used a function that generates 250 overlapping recombination distance categories with a step of 0.001 centiMorgan (cM) from 0.005 to 0.25. Then, *Ne* with a confidence interval 95% for each recombination distance category applying the formula $Ne \approx 1/(4c) * [(1/r2)-2]$ is calculated (c is the distance between markers in Morgan). The long-term *Ne* was estimated as the harmonic mean of the effective population size along the generations in the past for each population (Wright, 1931). Knowing the values of *Ne* and having the matrix of the calculated pairwise F_{ST} values with 4P software (Benazzo et al., 2015), we could estimate the time of divergence between populations using the *Tdverg* function of the NeON R package. Divergence time between pairs of study populations in generations was estimated as follows: $T = ln(1 - F_{ST})/ln(1-1/2Ne)$, where T represents divergence time. A generation is assumed to be 25 years long. The evolutionary history based on estimated divergence times was inferred using the UPGMA method implemented in MEGA X software v.10.2.5 (Kumar et al., 2018).

Results

Principal component analysis results on Baltic and Indian populations

Principal component analysis (PCA) was performed on genome wide 343,295 SNP data of 1,142 individuals, which were distinguished by the country of origin (Lithuania, Latvia, and India). We revealed 2 Latvian and 5 Indian populations outliers (Supplementary Figure S2.1), which were subsequently removed from further analyses.

The Indian population was classified based on the families of languages they mostly speak (Indo-European, Dravidian, Austroasiatic, and Sino-Tibetan). This classification was chosen because it provides information about geographical location of the samples (Supplementary Table S1.2). Samples from the Lithuanian and Latvian populations were distinguished solely by their country of origin, as they exhibited a high level of homogeneity and formed a single cluster. This approach was adopted in order to accurately represent the diversity within the Indian population and to ensure that meaningful conclusions can be drawn from the analysis (Supplementary Figure S2.1).

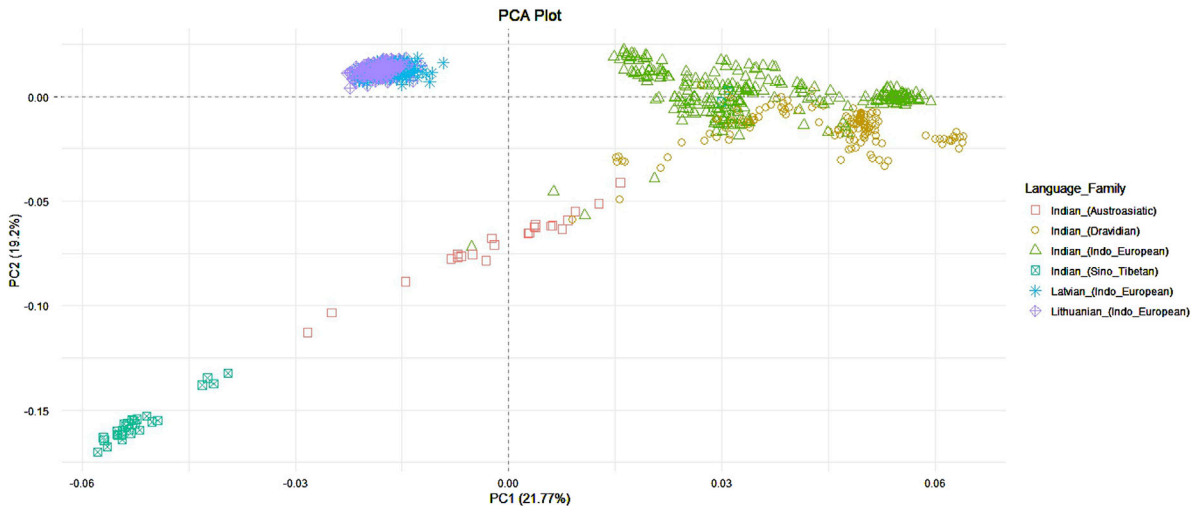

FIGURE 1
PCA among Lithuanian, Latvian, and Indian language families.

The first two principal components explained 45,49% of genetic variance among populations and showed that individuals from Lithuanian and Latvian populations form one cluster, while individuals from the Indian population form a distinct group. Also, as quite expected, the Indian population had a much bigger data dispersion in the plot when compared to the Balts. This shows a higher genetic diversity of the Indian population due to a complex interplay of historical migrations, long-term population structure, and social factors like the caste system and linguistic diversity. The samples of Indian speakers of Indo-European languages showed a closer allocation to the Balts compared to other linguistic groups of Indians. The most distant from the Balts are samples which belong to the speakers of Sino-Tibetan languages (Figure 1).

Principal component analysis results on baltic, indian and other world populations

Further PCA was performed on the Lithuanian, Latvian, Indian dataset merged with other world populations from 1000 Genomes project. In total it included 2,859 individuals and 343,295 SNP. Individuals from the Indian population were again distinguished based on the family of their main language to get a more accurate view and see which linguistic group from this large diverse population has a greater resemblance to the Balts. The first principal component explained 57,2% of genetic variance among populations, the second principal component – 26,97% (Figure 2). The first principal component distinguishes ESN, GWD, LWK, YRI and MSL populations (Africa) from other analyzed populations. The second principal component distinguishes KHV, JPT, CHB, CHS populations (East Asia) from populations of Europe and South Asia. Lithuania and Latvia together form a distinct cluster but are closer to the Indian populations, GIH, PJL, STU, BEB (South Asia) and Estonians (Europe). Close to these clusters another group is found of TSI, FIN, CEU and GBR populations (Europe). The Indian population has the biggest data dispersion in the plot when compared to other populations. Indo-European and Dravidian speaking individuals cluster closer to the Baltic populations than Sino-Tibetan or Austroasiatic speaking individuals.

Principal component analysis results on baltic, indian and ancient and modern populations from eurasia

Principal component analysis was conducted on the dataset by merging the samples from the Baltic and Indian populations with those from other ancient and modern populations across the Eurasian continent (as shown in Supplementary Figure S1.4). The purpose of this analysis was to determine whether ancient populations, particularly those from Anatolian and Steppe regions, exhibit a closer genetic relationship to Indo-European speaking populations. This investigation also aimed to evaluate the hypotheses of the homeland of the Indo-European languages discussed in the Introduction. Some modern populations from the European continent were included in the analysis with the objective to investigate whether the speakers of the Baltic languages share a stronger genetic connection with the Indian population compared to other neighboring European populations. This dataset consisted of 1,628 samples and 13,522 SNPs. Indian ethnolinguistic groups which were assigned to Austroasiatic or Sino-Tibetan languages families were removed from analysis since they have been shown as clearly distant from the Baltic populations in the previous analysis (Figure 1). After filtering, the dataset consisted of 1,569 samples and 10,208 SNPs. First principal component analysis revealed 1 Levantian Neolithic and 2 Iranian Neolithic outliers which were removed from the further analysis (Supplementary Figure S2.4). Ancient samples were assigned to populations as in Supplementary section table 1.4., and individuals from modern populations were distinguished based on the language branches and subbranches, the Indo-European family is represented by Baltic (Lithuanian and Latvian populations), East Slavic (Belarusian, Russian, Ukrainian populations), West Slavic (Polish population), North Germanic (Icelandic, Norwegian populations), West Germanic (English and Scottish populations), the Uralic family

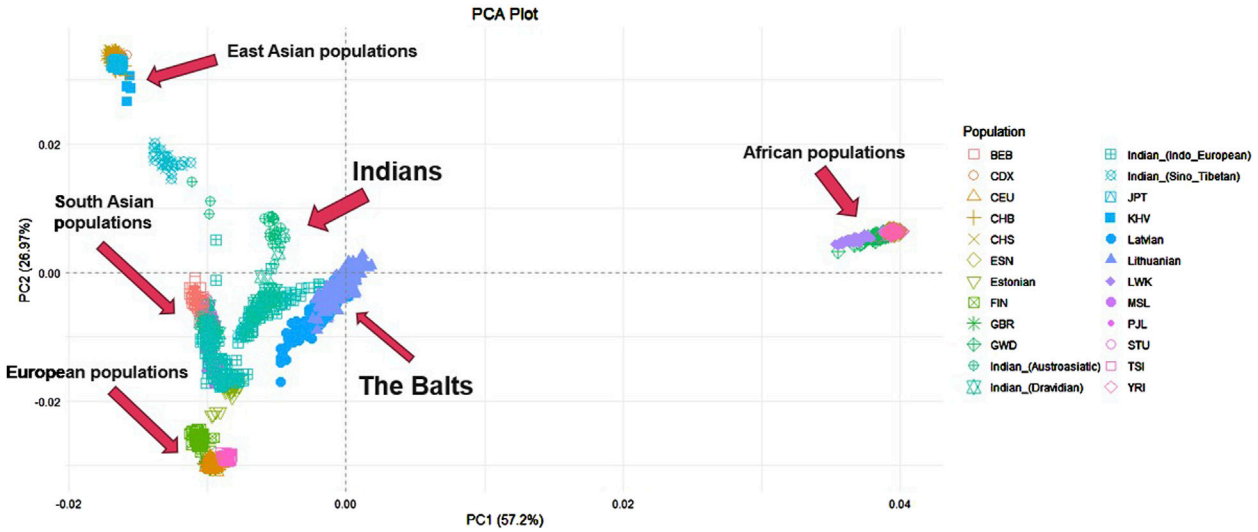

FIGURE 2
PCA of Lithuanian, Latvian, Indian and other world populations. YRI - Yoruba in Ibadan, Nigeria, LWK - Luhya in Webuye, Kenya, GWD - Gambian in Western Divisions in the Gambia, MSL - Mende in Sierra Leone, ESN - Esan in Nigeria, CEU - Utah Residents with Northern and Western European Ancestry, TSI - Toscani in Italia, FIN - Finnish in Finland, GBR - British in England and Scotland, CHB - Han Chinese in Bejing, China, JPT - Japanese in Tokyo, Japan, CHS - Southern Han Chinese, China, CDX - Chinese Dai in Xishuangbanna, China, KHV - Kinh in Ho Chi Minh City, Vietnam, PJL - Punjabi from Lahore, Pakistan, BEB - Bengali from Bangladesh, STU - Sri Lankan Tamil from the UK.

is represented by Finnic (Estonian and Finnish populations), and Mordvin (Mordovian population). The first principal component explained 2.6% of genetic variance and distinguished the Indian populations from the analyzed modern Eurasian populations. However, Indo-European speaking Indians clustered closer to the Eurasian populations compared to the Dravidian speaking Indians. Balts display a tight cluster partly overlapping the West Slavic, and North Germanic groups with close proximity towards one of the extremes of the Finnic population clusters (Figure 3). Compared to ancient samples, modern Balts and Finns positioned as the closest present-day European populations to ancient hunter-gatherer groups (Latvian HG, Western HG, Eastern HG, Scandinavian HG), neolithic ancient samples from Lithuania, Estonia and Latvia, also Ukrainian Mesolithic and neolithic samples, and showed high proximity to the Steppe and late Neolithic Bronze Age European samples (Figure 3). The Balts did not show a closer genetic relationship with Indians compared to other Eurasian populations (Figure 3).

Admixture analysis

The dataset of ancient and modern samples from Eurasian continent was used, with previously mentioned outliers and Indian populations speaking Austroasiatic or Sino-Tibetan languages removed. This dataset consisted of 1,569 samples and 10,208 SNPs. We included 17 modern populations (Lithuanian, Latvian, Belarusian, Ukrainian, Russian, Polish, English, Scottish, Orcadian, Icelandic, Norwegian, Iranian, Indian (Indo-European), Indian (Dravidian), Mordovian, Estonian, Finnish (Uralic)) and a group of 34 ancient populations. The best CV value of admixture analysis was determined when populations were distributed into 4 clusters (Supplementary Table S2.2).

Admixture analysis revealed that the Baltic population is characterized by three main genetic components reflecting Hunter Gatherers ancestry (in green), other components found in the European Neolithic/Anatolian Neolithic (in purple) and in the Iranian, Lithuanian Mesolithic, Lithuanian EMN Narva (red). Indians are also characterized by three genetic components–blue found maximized in Lithuanian and Estonian EMN Narva, Lithuanian Late Neolithic, and Estonian Middle Neolithic; red found in the Iranian, and purple which is observed only in one part of Indo-European speaking Indians and found in the European Neolithic/Anatolian Neolithic samples. All modern European populations are mainly characterized by three color compounds as well as the Balts. Some of them, however, have a small part of purple color (West Slavs). Out of modern European populations, the Polish population has the biggest part of the blue color component, which is characteristic to Indians. From the ancient populations, Lithuanian Early-Middle Neolithic Narva samples have the biggest amount of blue color (Supplementary Figures S3, S4). The full admixture analysis is provided in Supplementary Figure S2.8, S2.9.

Subsequently we used outgroup-f3 statistics (Patterson et al., 2012) to measure shared ancestry between Baltic and ancient populations from different geographic regions. Among the modern populations of the Lazaridis et al., (2016) dataset (Lazaridis et al., 2016), Balts showed the highest f3 values when testing Western HG, Scandinavian and Eastern HG, Latvian HG, Latvia middle Neolithic, Lithuania early middle Neolithic and late Neolithic, Ukraine Mesolithic and Neolithic, Steppe Eneolithic, and EMBA steppe pastoralists, but not for the Lithuania Mesolithic, and European Neolithic farmer component (Figure 4). No significant allele sharing was detected between the Balts and the Anatolian Neolithic farmers.

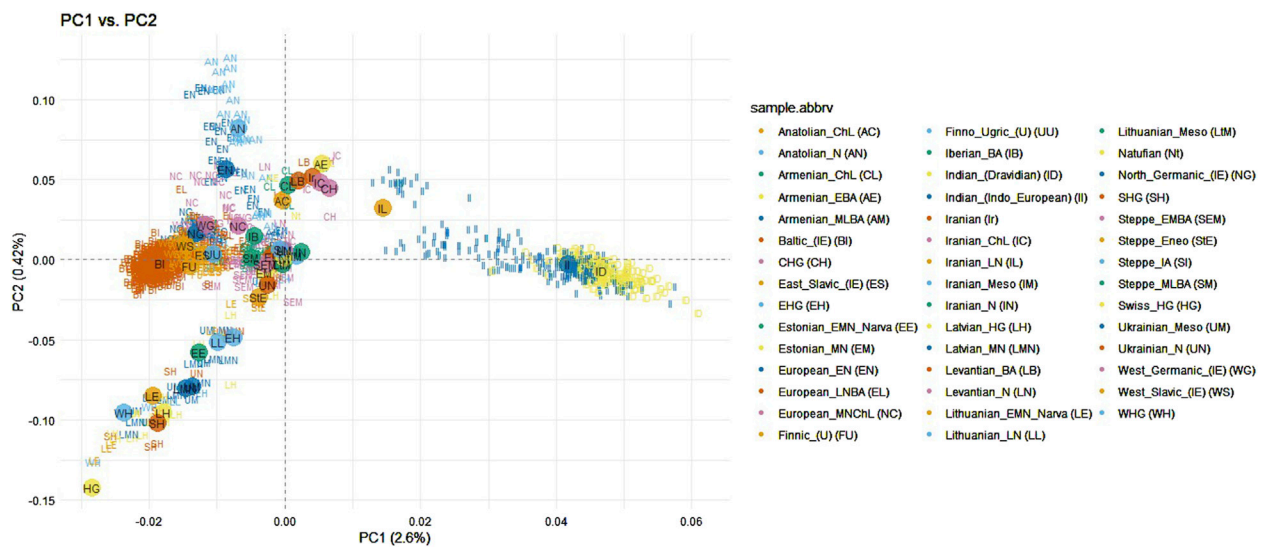

FIGURE 3
PCA of Indian (Indo-European and Dravidian speaking), ancient and modern Eurasian populations when modern populations were distinguished by language subfamilies. Abbreviations of ancient populations are explained in Supplementary section table 1.4.

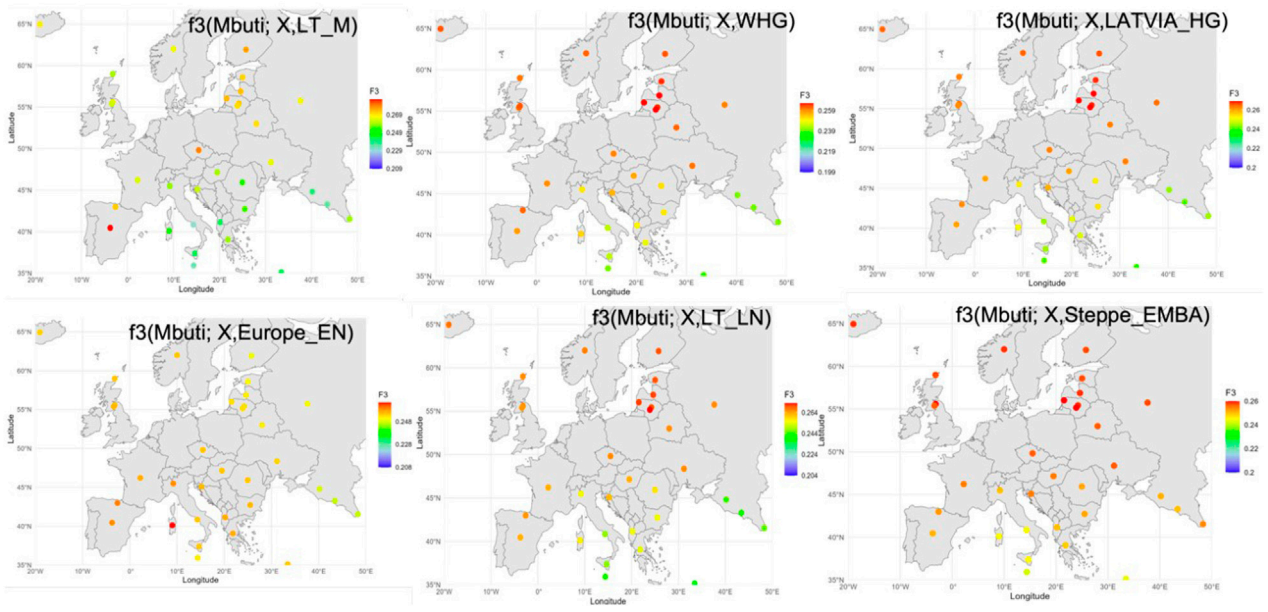

FIGURE 4
Geographical distribution of outgroup f3-statistics showing shared genetic drift top values between antient and modern samples analysed. Each analysed modern population is displayed in its corresponding location on the map. The darker the red color, the higher the pairwise f3 statistics between modern population and the ancient population group, representing higher allele sharing.

Ne and divergence time

To reconstruct the evolutionary relationships among the Baltic, Indian (speaking Indo-European and Dravidian languages), and modern populations from Europe, effective population size (*Ne*) of each population (Supplementary Table S2, S3; Supplementary Figures S2.10, S2.11) and divergence times (Supplementary Table S2.5) among them were estimated. The estimated long-term *Ne*, calculated as the harmonic mean, was about 5,000 for Balts, with confidence intervals (CI) [4,672; 5,258]. Comparing the Ne values for populations analyzed we observed a variation from ≈10,000 in the Indian populations to 3,987 in Finnish population (Supplementary Figure S2.3). The constructed phylogenetic tree displays two main long branches which distinguish Indians from other populations in the European

continent. This indicates an earlier separation that occurred many years ago. Divergence time analysis showed that Indian population separated from other European continent populations approximately 11,569 years ago on average (Supplementary Table S2.5). The branch representing populations from the European continent further divides into two smaller main branches, which can be partly explained by the language families and their branches and partly by the geographic proximity of the populations. One branch comprises Indo-European speakers of the Baltic (Lithuanian and Latvian), the West Slavic (Polish), and the East Slavic (Russian, Belarusian, Ukrainian), which belong to the Balto-Slavic branch, but it also includes speakers of Uralic languages from the Finnic (Estonian and Finnish) and the Mordvin (Mordovian) branches. The speakers of different subbranches and even families are notably intermixed here due to the areal proximity: Polish (West Slavic) shares a branch with Ukrainian (East Slavic), Lithuanian and Latvian (Indo-European) share a branch with Estonian (Uralic), and Russian (Indo-European) shares a branch with Mordovian (Uralic). The other branch forms a group of Germanic-speaking populations, where speakers of West Germanic (English, Scottish, Orcadian) and North Germanic (Icelandic, Norwegian) are intermixed. Within the first branch group, the Finnish population is initially separated from the remaining populations. Subsequently, the Mordovian and Russian populations are distinguished from the rest. Further smaller branches then separate into a group consisting of Estonian, Latvian, and Lithuanian populations and another group comprising Polish, Belarusian, and Ukrainian populations. The most recent separation occurred between Lithuanian and Latvian populations and between Polish and Ukrainian populations. The Scottish population splits from the rest in the Germanic branch first. Following that, the Orcadian and Norwegian populations are distinguished, and the most recent separation occurred between English and Icelandic populations (Figure 5).

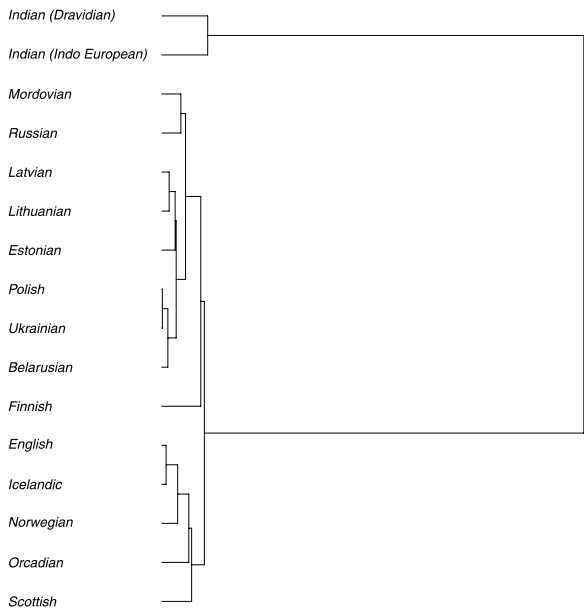

FIGURE 5
Phylogenetic tree based on the divergence time (in generations) among analyzed populations.

Discussion

The results of our present study, using increased number of the Baltic samples, confirm previous findings that the Lithuanian and Latvian populations exhibit a high level of homogeneity (Kasperavičiūtė et al., 2004; Pliss et al., 2015; Ruzgaitė et al., 2015; Urnikyte et al., 2019). The Indo-European speaking Indian population clusters closer to Lithuanian, Latvian, and other Indo-European and Uralic speaking European populations in our sample compared to the Dravidian, Austroasiatic, or Sino-Tibetan speaking Indian populations. This supports previous findings indicating that Indo-European speaking Indians exhibit a closer genetic relationship to Europeans than Dravidian, Austroasiatic, or Sino-Tibetan speakers (Bamshad et al., 2001; Thanseem et al., 2006).

Furthermore, in all PCA plots, the Indian population displays greater data dispersion compared to other populations. This observation helps explain the higher genetic diversity observed within the Indian population (Majumder, 1998; Pugach and Stoneking, 2015). When considering other world populations, the Indian population appears to be genetically closer to European populations of our samples, including Lithuania and Latvia, than to African or East Asian populations. This finding is supported by shared genetic drift with Europeans (Bose et al., 2021). However, when conducting PCA analysis on the Baltic, Indians, and samples from ancient and recent times across the Eurasian continent, Lithuanians and Latvians did not exhibit a higher genetic relationship with Indo-European speaking Indians compared to other populations. This demonstrates that the Baltic speaking populations (within the Balto-Slavic branch) and the Indo-European-speaking Indian populations (of the Indo-Iranian branch) did not share an exclusive ancestor, and this is in line with the linguistic interpretation that does not support the tentative Indo-Slavic node in the evolution of the Indo-European language family, as discussed in the Introduction. We may also add that the results of PCA analysis demonstrate that the Indo-European speaking populations from the European continent in our (limited) sample remain equally genetically similar to the Indo-European speaking Indians.

PCA and f3-statistics analysis on ancient populations revealed that the Balts and other neighboring modern populations from the European continent display a closer genetic relationship to populations from the ancient Steppe region (from the Early-Middle and Middle-Late Bronze Age) compared to the populations from the ancient Anatolian region. These findings align with the steppe hypothesis, which suggests that the major spread of Indo-Europeans to Europe was from the Pontic Steppe (Gimbutas, 1970; Mallory, 1989; Anthony, 2007) or the hybrid origin model which presupposes a secondary Indo-European homeland in the steppe region (Heggarty et al., 2023), rather than the Anatolian hypothesis alone (Renfrew, 1987; Bellwood, 2005).

Furthermore, an analysis of the population's genetic structure using the admixture method revealed that the genetic composition of the Indian population differs from that of other populations across the Eurasian continent, both ancient and modern. However, once again, the Indo-European speaking part of the Indian

population showed more similar genetic composition to other European populations compared to Dravidians.

The divergence time analysis indicates that Indian population separated from other European continent populations approximately 11,569 years ago on average. They separated from Lithuanians 12,673 years ago and from Latvians 12,969 years ago. The most recent divergence for the Indian population, in the context of European populations, occurred with the Mordovian population approximately 11,195 years ago. This period may be associated with the retreat of the last glacier, as well as the separation and dispersal of Indo-European languages across the Eurasian continent.

In conclusion, these results largely confirm previous findings reported by other authors in the field of population genetic studies or, at the very least, do not conflict with them. The Indo-European speaking Indian population exhibits a higher similarity to Europeans in both genetic structure analysis methods (PCA and admixture), as previously discussed in literature (Bamshad et al., 2001; Thanseem et al., 2006; Bose et al., 2021). However, as already mentioned above, the applied methods did not reveal that the Balts population is more genetically similar to Indians than other Indo-European speaking European populations. This work represents the first attempt in the field of population genetic studies to investigate the genetic relationship between the Baltic and the Indian populations using autosomal markers. Additionally, the study included a large number of samples from the investigated Baltic and Indian populations, ensuring diversity within the datasets.

The study does have several limitations that should be considered. First, while the sample size was large, it may not fully capture the entire genetic diversity of the Baltic and Indian populations, potentially limiting the generalizability of the findings. Second, the study focused on autosomal markers, which, while informative, may not provide a complete picture of genetic relationships. Other genetic markers, such as mitochondrial DNA or Y-chromosome data, could offer additional insights.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ebi.ac.uk/ena/browser/view/PRJEB25864; PRJEB25864. Additional data information is available in Supplementary Table S1.

Ethics statement

The studies involving humans were approved by Vilnius Regional Biomedical Research Ethics Committee (hereinafter - VRBTEK). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

GD: Investigation, Writing–original draft, Writing–review and editing. LP: Formal Analysis, Investigation, Methodology, Supervision, Validation, Writing–original draft, Writing–review and editing. JP: Writing–original draft, Writing–review and editing. JK: Data curation, Investigation, Writing–review and editing. VK: Supervision, Validation, Writing–review and editing. AU: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1493270/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19 (9), 1655–1664. doi:10.1101/gr.094052.109

Allentoft, M., Sikora, M., Sjögren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., et al. (2015). Population genomics of Bronze Age eurasia. *Nature* 522, 167–172. doi:10.1038/nature14507

Anthony, D. W. (2007). *The horse, the wheel, and language: how bronze-age riders from the Eurasian steppes shaped the modern world*. Princeton, NJ: Princeton University Press.

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., et al. (2001). Genetic evidence on the origins of Indian caste populations. *Genome Res.* 11 (6), 994–1004. doi:10.1101/gr.gr-1733rr

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., et al. (2003). Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* 13 (10), 2277–2290. doi:10.1101/gr.1413403

Behar, D. M., Metspalu, M., Baran, Y., Kopelman, N. M., Yunusbayev, B., Gladstein, A., et al. (2013). No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. *Hum. Biol.* 85 (6), 859–900. doi:10.3378/027.085.0604

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). Pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32 (18), 2817–2823. doi:10.1093/bioinformatics/btw327

Bellwood, P. (2005). *First farmers: the origins of agricultural societies*. London: Blackwell.

Benazzo, A., Panziera, A., and Bertorelle, G. (2015). 4P: fast computing of population genetics statistics from large DNA polymorphism panels. *Ecol. Evol.* 5 (1), 172–175. doi:10.1002/ece3.1261

Bose, A., Platt, D. E., Parida, L., Drineas, P., and Paschou, P. (2021). Integrating linguistics, social structure, and geography to model genetic diversity within India. *Mol. Biol. Evol.* 38 (5), 1809–1819. doi:10.1093/molbev/msaa321

Chaubey, G., Ayub, Q., Rai, N., Prakash, S., Mushrif-Tripathy, V., Mezzavilla, M., et al. (2017). "Like sugar in milk": reconstructing the genetic history of the Parsi population. *Genome Biol.* 18 (1), 110. doi:10.1186/s13059-017-1244-9

Chaubey, G., Metspalu, M., Choi, Y., Mägi, R., Romero, I. G., Soares, P., et al. (2010). 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

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27 (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10 (2), giab008. doi:10.1093/gigascience/giab008

Diamond, J., and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science.* 300 (5619), 597–603. doi:10.1126/science.1078208

Gimbutas, M. (1970). "Proto-Indo-European culture: the Kurgan culture during the fifth, fourth and third millennia BC," in *Indo-European and Indo-Europeans: papers presented at the third Indo-European conference at the university of Pennsylvania*. Editors G. Cardona, H. M. Hoenigswald, and A. Senn (Philadelphia: University of Pennsylvania Press).

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., et al. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522, 207–211. doi:10.1038/nature14317

Heggarty, P., Anderson, C., Scarborough, M., King, B., Bouckaert, R., Jocz, L., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science.* 6656. doi:10.1126/science.abg0818

Jochim, M. (2011). "The Upper Paleolithic," in *European prehistory: A survey*. Editor S. Milisauskas (New York: Springer), 67–124.

Kasperavičiūtė, D., Kučinskas, V., and Stoneking, M. (2004). Y chromosome and mitochondrial DNA variation in Lithuanians. *Ann. Hum. Genet.* 68 (5), 438–452. doi:10.1046/j.1529-8817.2003.00119.x

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., and Mega, X. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35 (6), 1547–1549. doi:10.1093/molbev/msy096

Kümmel, M. J. (2022). "Indo-Iranian," in *The Indo-European language family*. Editor T. Olander (Cambridge University Press), 246–268.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* 536 (7617), 419–424. doi:10.1038/nature19310

Majumder, P. P. (1998). People of India: biological diversity and affinities. *Evol. Anthropol. Issues. Evol. Anthropol. Issues, News, Rev.* 6 (3), 100–110. doi:10.1002/(sici)1520-6505(1998)6:3<100::aid-evan4>3.0.co;2-i

Majumder, P. P., and Basu, A. (2015). A genomic view of the peopling and population structure of India. *Cold Spring Harb. Perspect. Biol.* 7 (4), a008540. doi:10.1101/cshperspect.a008540

Mallory, J. P. (1989). *In search of the Indo-Europeans: Language, archaeology, and myth*. New York, NY: Thames and Hudson.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26 (22), 2867–2873. doi:10.1093/bioinformatics/btq559

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., et al. (2018). The genomic history of southeastern Europe. *Nature* 555 (7695), 197–203. doi:10.1038/nature25778

Metspalu, M., Romero, I. G., Yunusbayev, B., Chaubey, G., Mallick, C. B., Hudjashov, G., 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

Mittnik, A., Wang, C.-C., Pfrengle, S., Daubaras, M., Zariņa, G., Hallgren, F., et al. (2018). The genetic prehistory of the Baltic Sea region. *Nat. Commun.* 9 (1), 442. doi:10.1038/s41467-018-02825-9

Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P. R., Govindaraj, 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

Nakhleh, L., Ringe, D. A., and Warnow, T. (2005). Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Lang. Balt.* 81 (2), 382–420. doi:10.1353/lan.2005.0078

Neon, M. M. (2015). An R package to estimate human effective population size and divergence time from patterns of linkage disequilibrium between SNPS. *J. Comput. Sci. Syst. Biol.* 8 (1). doi:10.4172/jcsb.1000168

Olander, T. (2019). Indo-European cladistic nomenclature. *Indoger Forsch* 124 (1), 231–244. doi:10.1515/if-2019-0008

Ostrauskas, T. (2008). "Vėlyvasis paleolitas," in *Akmens amžius ir ankstyvasis metalų laikotarpis*. Editor A. Girininkas (Vilnius: Baltos lankos), 11–47.

Pathak, A. K., Kadian, A., Kushniarevich, A., Montinaro, F., Mondal, M., Ongaro, L., 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

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., et al. (2012). Ancient admixture in human history. *Genetics* 192 (3), 1065–1093. doi:10.1534/genetics.112.145037

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Pliss, L., Timša, L., Rootsi, S., Tambets, K., Pelnena, I., Zole, E., et al. (2015). Y-chromosomal lineages of Latvians in the context of the genetic variation of the eastern-baltic region. *Ann. Hum. Genet.* 79 (6), 418–430. doi:10.1111/ahg.12130

Pronk T. (2022). "Balto-Slavic," in *The Indo-European language family*. Editor T. Olander (Cambridge University Press), 269–292.

Pugach, I., and Stoneking, M. (2015). Genome-wide insights into the genetic history of human populations. *Investig. Genet.* 6 (1), 6. doi:10.1186/s13323-015-0024-0

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

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

Renfrew, C. (1987). *Archaeology and language: the puzzle of Indo-European origins*. (New York: Cambridge University Press).

Reščenko, R., Brīvība, M., Atava, I., Rovīte, V., Pečulis, R., Silamiķelis, I., et al. (2023). Whole-genome sequencing of 502 individuals from Latvia: the first step towards a population-specific reference of genetic variation. *Int. J. Mol. Sci.* 24 (20), 15345. doi:10.3390/ijms242015345

Rimantienė, R. (1996). *Akmens amžius lietuvoje*. (Vilnius: Žiburio leidykla).

Ringe, D., Warnow, T., and Taylor, A. (2002). Indo-European and computational cladistics. *Trans. Philol. Soc.* 100, 59–129. doi:10.1111/1467-968x.00091

Ruzgaitė, G., Čaplinskienė, M., Baranovienė, R., Jankauskienė, J., Kukienė, J., Savanevskytė, K., et al. (2015). Forensic application of Y-chromosomal STR analysis in Lithuanian population. *Biologija* 61 (2). doi:10.6001/biologija.v61i2.3140

Tambets, K., Yunusbayev, B., Hudjashov, G., Ilumäe, A. M., Rootsi, S., Honkola, T., et al. (2018). Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol.* 19 (1), 139. doi:10.1186/s13059-018-1522-1

Tätte, K., Pagani, L., Pathak, A. K., Kõks, S., Ho Duy, B., Ho, X. D., et al. (2019). The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci. Rep.* 9 (1), 3818. doi:10.1038/s41598-019-40399-8

Thanseem, I., Thangaraj, K., Chaubey, G., Singh, V. K., Bhaskar, L. V. K. S., Reddy, B. M., et al. (2006). Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* 7 (1), 42. doi:10.1186/1471-2156-7-42

The 1000 Genomes Project Consortium Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., et al. (2015). A global reference for human genetic variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Urnikyte, A., Flores-Bello, A., Mondal, M., Molyte, A., Comas, D., Calafell, F., et al. (2019). Patterns of genetic structure and adaptive positive selection in the Lithuanian population from high-density SNP Data. *Sci. Rep.* 9 (1), 9163. doi:10.1038/s41598-019-45746-3

Urnikyte, A., Molyte, A., and Kučinskas, V. (2021). Genome-wide landscape of north-Eastern European populations: a view from Lithuania. *Genes* 12 (11), 1730. doi:10.3390/genes12111730

Wickham, H. (2016). *GGPLOT2: elegant graphics for data analysis*. Cham: Springer International Publishing.

Wright, S. (1931). Evolution in mendelian populations. *Genetics* 16 (2), 97–159. doi:10.1093/genetics/16.2.97

Zagorska, I. (2001). "Vēlā paleolīta beigas 8500.-7600. g. pr. Kr," in *Latvijas senākā vēsture, 9. g. t. pr. Kr. - 1200 g.* Editors Ē. Mugurēvičs, and A. Vasks (Rīga: Latvijas vēstures institūta apgāds), 22–39.