DOI Link: https://doi.org/10.59879/4Ra1m
Vol.91, Issue 1, Part.6, June 2025, PP. 47-70

Correlation of Y-DNA Haplogroups and Language Families

Fehér, Tibor✉[1] - Németh, Endre[1]
[1]Hungarian Natural History Museum ✉jafetie@gmail.com
Received April 2025 Accepted May 2025 Published June 2025

Abstract
The distinction between the origin of a language and the origin of a people has been obvious among researchers already since the 19th century. The linguistic and genetic characteristics of cultural groups may behave very differently. However, despite the difficulties listed, it is by no means impossible to study genetic and linguistic features in a common framework. Historical genetics is a powerful tool for analysing migration routes, disjunction and admixtures of biological groups and their demographic expansions or compressions. Taking into account the above observations, we analyse whether the occurrence of a language family and subfamily correlates with certain Y-chromosome subgroups. We conclude that simple mathematical Pearson correlation between language families and deep Y-DNA haplogroup assignments can significantly contribute to the resolution of problems debated among linguists, especially in resolving the upstream phylogeny of Indo-European languages between Proto-Indo-European and well-established subfamilies. Our data also reinforces former observations regarding the genetic origin of proto-Uralic speaking groups and suggests new aspects for understanding the Y-DNA of proto-Altaic speakers. Our study proves that analysing correlations between language families and Y-DNA subhaplogroups is valid but should be based on high resolution Y-DNA data with deep subgroups younger than 5000 years, to obtain meaningful results.

Keywords: Phylogenetics, Language Families, Indo-European, Correlation, Y-DNA

Introduction

The distinction between the origin of a language and the origin of a people has been obvious among researchers already since the 19th century (Budenz, 1880; Szinnyei 1884). The linguistic and genetic characteristics of cultural groups may behave very differently even in the same interactions or as a single unit with eligible external impact.

For example, in case of language-cultural interactions, linguistic attributes may play a significantly more dominant role than genetic ones. The higher-prestige languages tend to dissolve lower-prestige linguistic phenomena without substantially changing the structure of the higher-prestige language. It is not at all uncommon, that the assimilated language leaves little or even undetectable trace in the host language. In contrast, population genetics can detect the trace of the assimilated population with a smaller size in the host population even if no historical or linguistic traces of this assimilation are subsequently available. Naturally, this detection is possible only if the original gene pool of the merging and the hosting population was at least slightly different.

More specifically, the demographic changes have an exponential characteristic. This is true for the reproduction, spread, and change of the gene pool of peoples, as well. Even at a very low rate of migration, significant changes can take place even over centuries in the overall structure of the gene pool of the host population (Cavalli-Sforza et al. 1994).

However, despite the difficulties listed, it is by no means impossible to study genetic and linguistic features in a common framework. The situation is similar to that in thermodynamics for ideal gasses, where the behaviour of the extensive and intensive state variables is sharply different in the physical interactions, but it is still possible to discuss these variables in a unified model.

At this point it is important to summarize the main strengths and goals of historical genetics research. Historical genetics is a powerful tool for analyzing migration routes, disjunction and admixtures of biological groups and their demographic expansions or compressions. With the tools of population genetics and archaeogenetics, a more or less precise chronology and geographical localization of these events is also possible. Some branches of linguistics and history address very similar questions. What migration routes did certain ethnic groups take? Where and when did languages belonging to the same language family diverge from each other? Where and when did peoples who spoke a common language live, what technological innovation resulted in population growth or in a decline due to war events or epidemics? Since these phenomena can also be located in space and time, genetic and cultural phenomena and events can be compared with each other.

In order to discuss historical genetic and linguistic issues in a single framework, the key question is which genetic characteristics are used and in what depth of its resolution. If we analyze the distribution of a genetic subgroup that arose long before the split of a language family, the chances of finding any connection between languages and genetic characteristics are minimal. In the same way, a young genetic subgroup that arose after the language family disintegration cannot provide information about the structure and distribution of the language family. That means the key to the success of your framework is choosing subgroups with appropriate time windows.

Y-chromosome, mtDNA and autosomal markers can all be suitable for studying a relationship between linguistic and genetic characteristics. In the present study, we chose the Y-chromosome primarily because this attribute contains the right amount of data representing the right time depth. Secondly, because at the dawn of history, when language families were formed, the structure of societies and marriage customs were different than nowadays, and there are several signs that clan organizations and paternal lines played an important role in the organization of societies.

It is important to note that the strength of specific Y-chromosome markers is well-known for historical genetic researchers. For example, specific paternal lineage markers were used to investigate the issue of Roma migration through ten European Romani populations. It was found that the genetic make-up of the Romani populations is rather different, but there are some special subgroups which link together the studied European Roma populations (Martinez-Cruz et al. 2015). One of these special markers is the haplogroup H, which is virtually absent from the non-Roma population in Europe but is highly prevalent in India.

In one important aspect, however, the common framework of historical-comparative linguistics and historical genetics certainly differs from the thermodynamic analogy. Namely, it is not possible to combine linguistic and genetic markers into a single universal model. The modelling of linguistic and demographic processes breaks down into different cases depending on certain characteristics of cultural groups. The most important of these characteristics are the following:

1.) the number of languages involved in the interaction,
2.) the relative prestige of the languages in relation to each other,
3.) the relative size of the same language speaking groups and the duration of the interaction.

For the sake of simplicity, we are focusing on populations living in a neighbouring or similar geographic area, but with different languages and genomes. We assume that the geographic boundaries of a community defined on linguistic grounds can be drawn with precision and do not consider the possibility of multilingualism. We also assume that the propensity to have

children is the same in both ideal populations, and that the degree of mixing between populations is proportional to the size of the population.

Without being exhaustive, the main types of interaction are the following.

1. When two populations of roughly the same size and prestige come into contact for a short period of time, the linguistic boundaries do not change, but in areas close to the border, admixtures begin.
2. When two populations of roughly the same size and prestige come into contact for a long period of time, the linguistic boundaries do not change, but the gene pool of the two populations become similar.
3. When two populations of roughly the same size, but with different prestige come into contact for a short period of time, the linguistic boundaries are shifted due to the admixture into the earlier territory of the lower prestige speaking population.
4. When two populations of roughly the same size, but with different prestige come into contact for a long period of time, the lower prestige language disappears and will be replaced by the language with the higher prestige.
5. When a population comes into contact with a higher prestige and larger size population then the assimilation processes take place relatively quickly.
6. If a relatively small but highly prestigious population conquers a large majority, the outcome is not clear. It is possible that the language of the majority population will be taken over by the conquerors over time, but it is also possible that after a sufficiently long period of time the majority will take over the language of the minority conquerors.
7. The lingua franca model, when a high-prestige language becomes a mediating language, which can also relatively quickly become the single language in the earlier colorful linguistic universe of the region.

Despite the static nature and simplicity of our ideal model, two conclusions can be drawn from the types of interactions listed.

- The higher proportions of the original gene pool are preserved better among lower prestige speaking populations living in less accessible or less attractive geographic areas.
- High-prestige language-speaking populations living in easily accessible and attractive areas gradually dissolve in the original gene pool in neighbouring populations, becoming more similar to their environment.

Taking into account the above observations, it is still a legitimate question whether the occurrence of a language subfamily correlates with certain Y-chromosome subgroups. At first glance, such an investigation does not promise too much success. Especially, the most common correlation coefficient, the Pearson correlation coefficient (Pearson 1895), which is sensitive mainly to a linear relationship between two variables seems to be hopeless in detecting relationship between genetic and linguistic data (see Materials and Methods). In the present study, we examine which haplogroups are correlated in the subfamilies of the Indo-European

language family. Best to our knowledge, such a transdisciplinary approach is a novum in linguistics and historical genetics.

Various relationships have been proposed by Indo-European researchers between Proto-Indo-European (PIE) and the well-established, accepted subgroups. These often contradict each other and have different levels of acceptance in Indo-European Studies. It is not the aim of this paper to summarize these discussions, but we would show several examples (Figure 1 to 5) in order to show the complex nature of groupings below PIE but above accepted proto-languages (Anatolian, Tocharian, Germanic, Celtic, Italic, Balto-Slavic, Albanian, Greek, Armenian, Indo-Iranian).

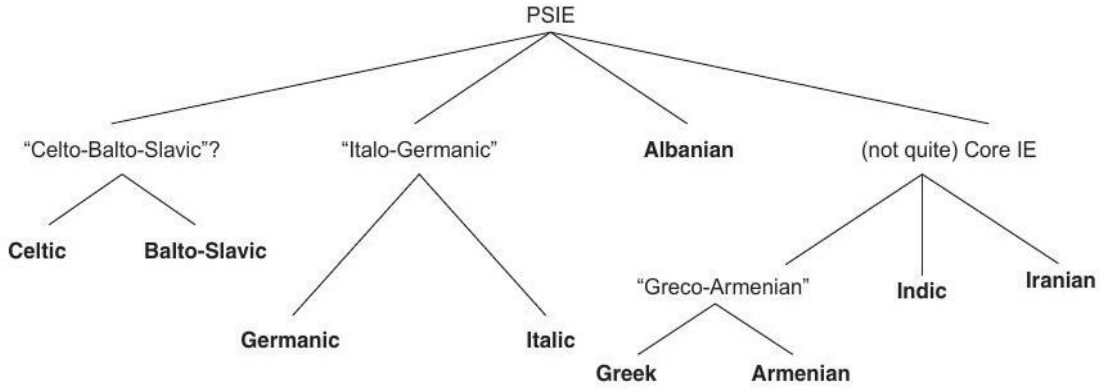


Figure 1 Phylogenetic tree of Indo-European according to Geisler and List (in press), based on the Dyen database
Note: Here and in subsequent figures, boldface indicates the ten "benchmark" groupings. Intermediate nodes are marked solely for convenience of reference. Terms in quotes are not used in the conventional classification, as these nodes are not postulated in the conventional classification. The branch length and angles have no meaning. "PSIE" stands for "Proto-Surviving-Indo-European", the ancestor of all Indo-European languages except the Anatolian and Tocharian branches, which have no surviving languages today.

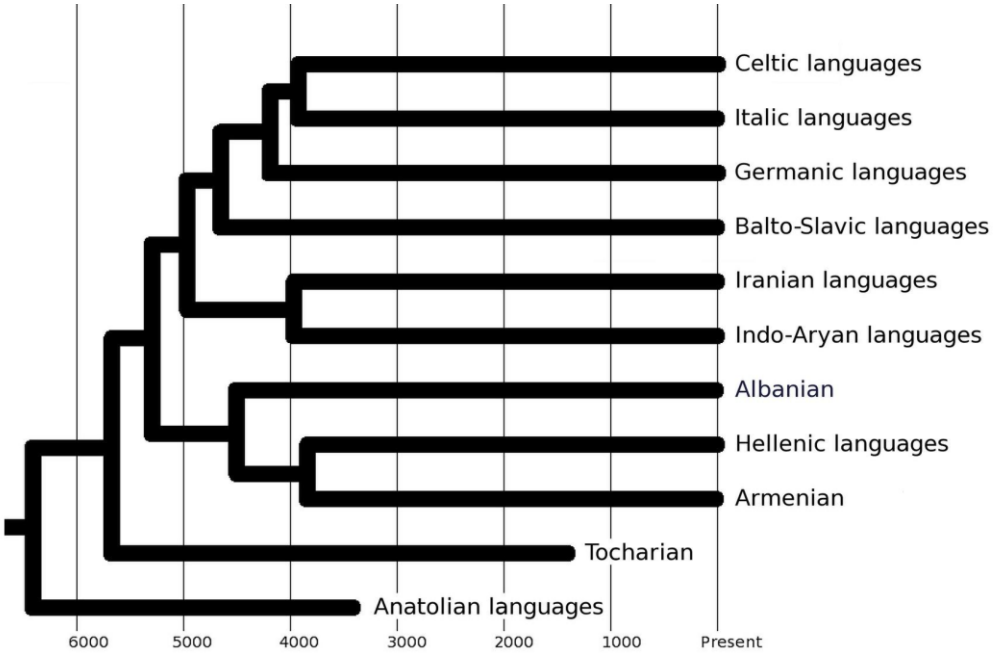


Figure 2: Indo-European language typology by Chang 2015

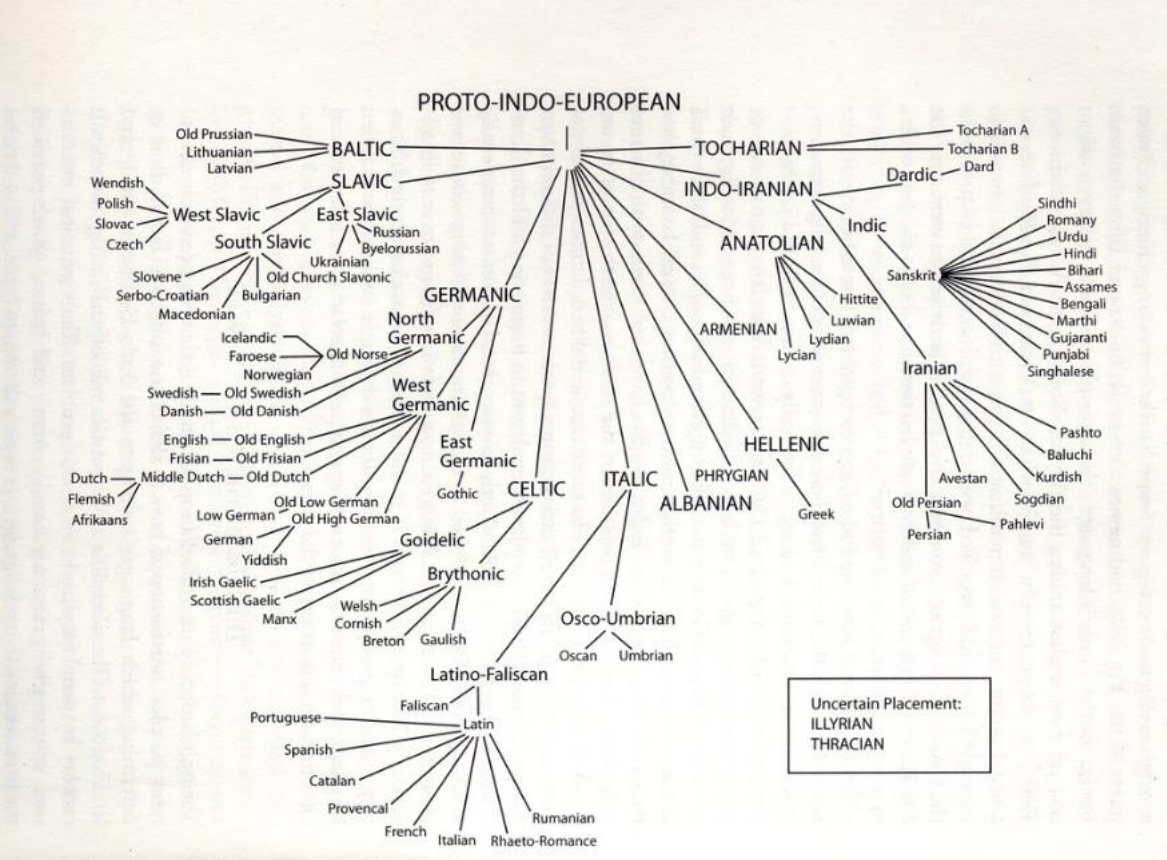


Figure 3: Indo-European language typology by Ali Eminov

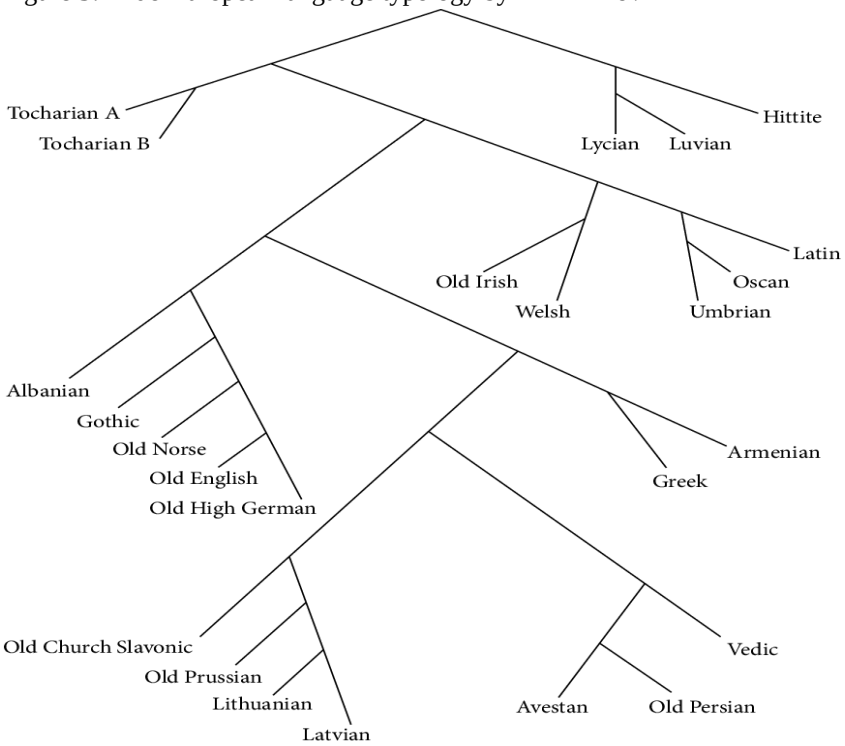


Figure 4: Indo-European language typology by Drinka 2013

ISSN: 0036-5483

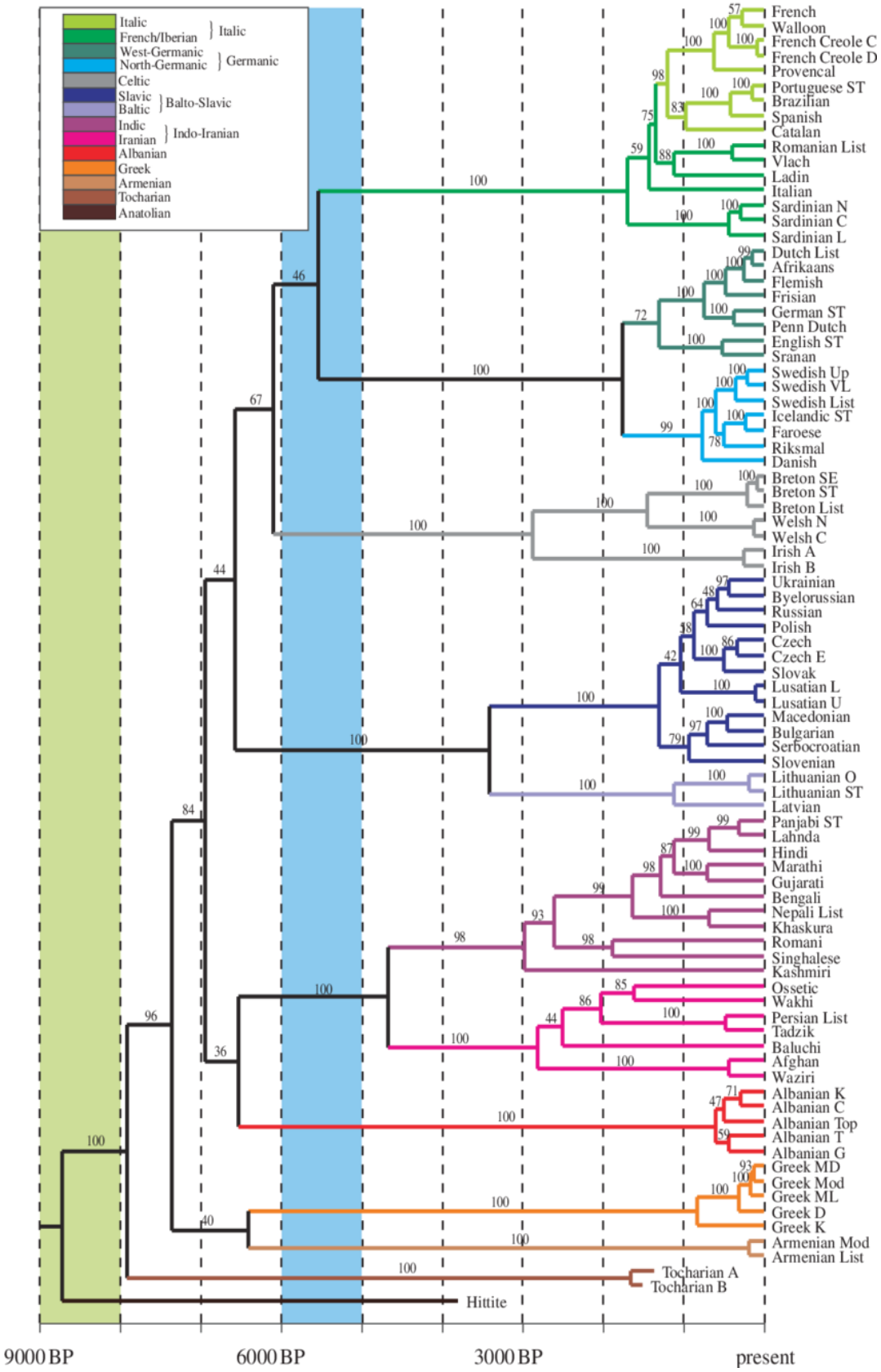


Figure 5: Indo-European language typology by Gray and Atkinson based on the glottochronological approach

While recent ancient DNA research has proven that Indo-European languages (probably excluding the oldest, Anatolian branch) were spread by R1b-M269 and R1a-M417 males from the Yamnaya culture on the Pontic Steppe (Allentoft et al. 2015, Haak et al. 2015, Mathieson et al. 2015, Olalde et al. 2018, Lazaridis et al. 2022, and even more recently Lazaridis et al 2024 and Nikitin et al. 2024), there was no comprehensive analysis on the various IE subgroups and their Y-DNA connections. Clarification on the Y-DNA could also help in establishing an accurate downstream tree from PIE, resolving issues of genetic relatedness vs. later areal or substrate effects among subfamilies (e.g. Germanic is closer to Balto-Slavic or Italo-Celtic; Italic is closer to Celtic or Germanic; placement of Albanian varies extremely from Indo-Iranian through Greek to Germanic etc.). In order to reflect the historically attested language shift in Romance-speaking areas, we split the Romance language family into two categories: Old Italic covering the territories which were Italic-speaking at the beginning of the Iron Age; and Gallo-Romance which were Celtic or Celtiberian-speaking before Roman conquest and later Romanized linguistically with relatively limited gene flow. Extinct Anatolian and Tocharian branches were not included. Baltic and Slavic as well as Iranian and Indo-Aryan were included as different categories, which resulted in 11 IE subcategories in total.

In order to test the correlation of various sub-haplogroups with Indo-European linguistic families, as well as to clarify the wider West Eurasian linguistic and genetic relationships, we included 11 non-IE language families into the analysis, using them as “outliers” and “control groups”. These language families are the following:

Berber, an Afro-Asiatic subfamily spoken from Morocco to Libya and in the Sahara Desert by Tuareg people.

Semitic, an Afro-Asiatic subfamily spoken originally in the Near East and Arabia by Jews, Arabs, Phoenicians, Assyrians, etc., later spread to Ethiopia (Amharic) and North Africa (Carthaginians and later Muslim Arabs).

Niger-Congo, the dominant language family of Sub-Saharan Africa spoken from Senegal to South Africa.

Turkic, covering languages from Anatolian Turkey and Karachay-Balkars in the North Caucasus through Central Asia to Yakut (Sakha) people in Central Siberia, whose classification with Mongolic and Tungusic languages into the larger Altaic family is disputed.

Finno-Ugric, including Finnish, Estonian and Hungarian in Eastern Europe as well as Ob Ugrians (Khanty and Mansi) in Western Siberia.

Dravidian, a language family in South India (Tamil, Kannada, Telugu, Malayalam) and Southwest Pakistan (Brahui), presumed to be dominant in South Asia before the arrival of Indo-Aryan languages ancestral to Hindi/Urdu, Punjabi, Bengali etc. around 1500 BCE.

Sino-Tibetan, the language family covering Chinese, Tibetan, Burmese and various smaller languages.

Amerind, not being a language family but rather an umbrella term for Native American language families excluding the northernmost Na-Dené and Eskimo-Aleut groups.

Northwest Caucasian, a small language family covering Abkhaz, Adyghe, Circassian and Kabard peoples in the Northwestern Caucasus area, with disputed connection to Northeast Caucasian languages.

Vainakh, a subgroup of Northeast Caucasian (Dagestani) languages covering Chechen and Ingush (Dagestani languages were not included due to the lack of high resolution data).

Kartvelian, the language family covering various Georgian dialects in the Southwest Caucasus, and no genetic relationship shown to any of the North Caucasian families.

Materials and Methods

Y-DNA data for 65 populations (rows) belonging to 11 Indo-European and 11 non-Indo-European linguistic categories (columns) were divided into 48 Y-DNA haplogroup categories (columns). The sources of the samples are indicated in Supplementary Table S1.

We examined the correlation between linguistic and genetic characteristics with Pearson correlation. First let's see more precisely what value range the Pearson correlation can range in and what types of relationships the correlation values draw attention to. The values of Pearson correlation coefficient vary from –1 to +1. Values of correlation coefficient close to –1 or +1 indicate a strong linear association between two variables, and those close to zero indicate a low association between variables. Positive values of correlation coefficient indicate a tendency of one variable to increase or decrease together with another variable. Negative values of correlation coefficient indicate a tendency that the increase of values of one variable is associated with the decrease of values of the other variable and vice versa.

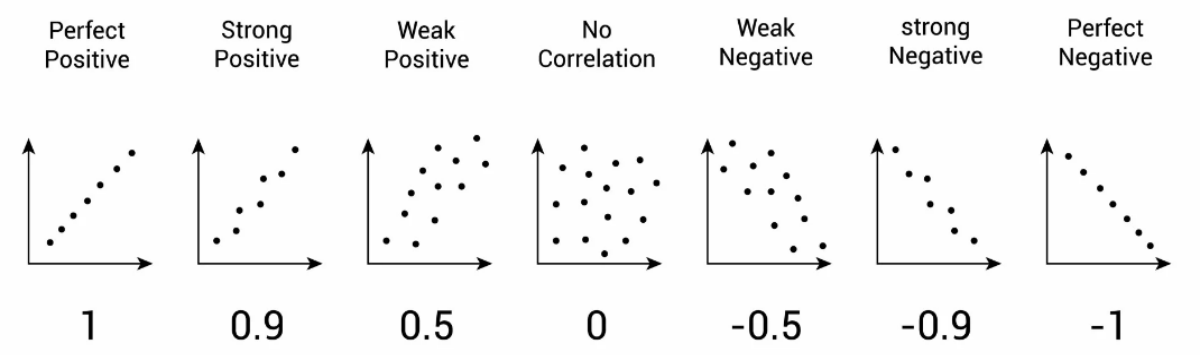


Figure 6: Correlation types based on values

It is worthwhile to examine the genetic composition of language communities based on Pearson's correlation for two reasons. First, in the case of choosing disjoint language categories, the classification of languages into a subfamily is a binary relationship. The English language is not Latin, nor Celtic, but a purely Germanic language by origin. Secondly, if the early spread of the language, say thanks to a technological development, involves a significant increase in the population, then the genetic heritage of the core population can be detected in the communities speaking the successor languages. Therefore we can also examine with the help of correlation whether a subgroup occurs in a language subfamily while it is completely or virtually absent from other subfamilies.

Results

Correlation results are shown in Supplementary Table S1. For better overview and understanding, in this section we list the relevant correlations for each language group.

Table 1: External (outside source) population groups

Language Family	Correlating Haplogroup	Value	Relationship[1]
Berbers (NW Africa)	Hg E1b-L19>M81	0,988	Very Strong
Berbers (NW Africa)	Hg E1b-M35*	0,705	Strong
Berbers (NW Africa)	Hg H2	0,538	Medium
Semitic (Near East)	Hg J1-P58	0,944	Very Strong
Semitic (Near East)	Hg E1b-Z780>M123	0,541	Medium
Semitic (Near East)	Hg T	0,435	Measurable
Semitic (Near East)	Hg J2a*(xM67)	0,431	Measurable
Niger-Congo (Africa)	Hg E1b-M2	0,998	Very Strong
Niger-Congo (Africa)	Hg E*(xE1b)	0,907	Very Strong
Niger-Congo (Africa)	Hg A00-B	0,649	Medium
Dravidian (South India)	Hg H1	0,821	Strong
Dravidian (South India)	Hg L	0,493	Measurable
Sino-Tibetan (East Asia)	Hg O	0,929	Very Strong
Sino-Tibetan (East Asia)	Hg D	0,404	Measurable
“Amerind” (Mex.& Peru)	Hg Q-Z780 & Q-M3	0,945	Very Strong
“Amerind” (Mex.& Peru)	Hg E1b-V12	0,428	Measurable

Hgs E1b-V12 and E1b-V22 had no real source population included in the study (Northeast Africa), which resulted in V12 and V22 appearing with measurable correlation in various non-Afro-Asiatic speaking population groups as a signature of gene flow from the Middle East and North Africa regions (Greece, Italy, colonial Americas).

[1] 0,34-0,49 measurable; 0,50-0,65 medium; 0,66-0,89 strong; 0,90- very strong

Table 2: Non-Indo-European languages in Western Eurasia

Language Family	Correlating Haplogroup	Value	Relationship
Uralic (Finno-Ugric)	Hg N(xL1025)	0,913	Very Strong
Vainakh (NE Caucasian)	Hg J2a-M67	0,916	Very Strong
Vainakh (NE Caucasian)	Hg J1*(xP58)	0,714	Strong
Kartvelian (Georgian)	Hg G2	0,412	Measurable
NW Caucasian	Hg G2	0,454	Measurable
Turkic	Hg C2-M217	0,608	Medium
Turkic	Hg R1b*(xM269)	0,592	Medium
Turkic	Hg Q*(xL527,Z780,M3)	0,443	Measurable
Turkic	Hg R1a-Z93	0,418	Measurable
Turkic	Hg D	0,365	Measurable

In the case of Kartvelian and NW Caucasian, it is clear that a further breakup of Hg G2 would have been needed to produce a stronger correlation result with each language group.

Table 3: Indo-European results

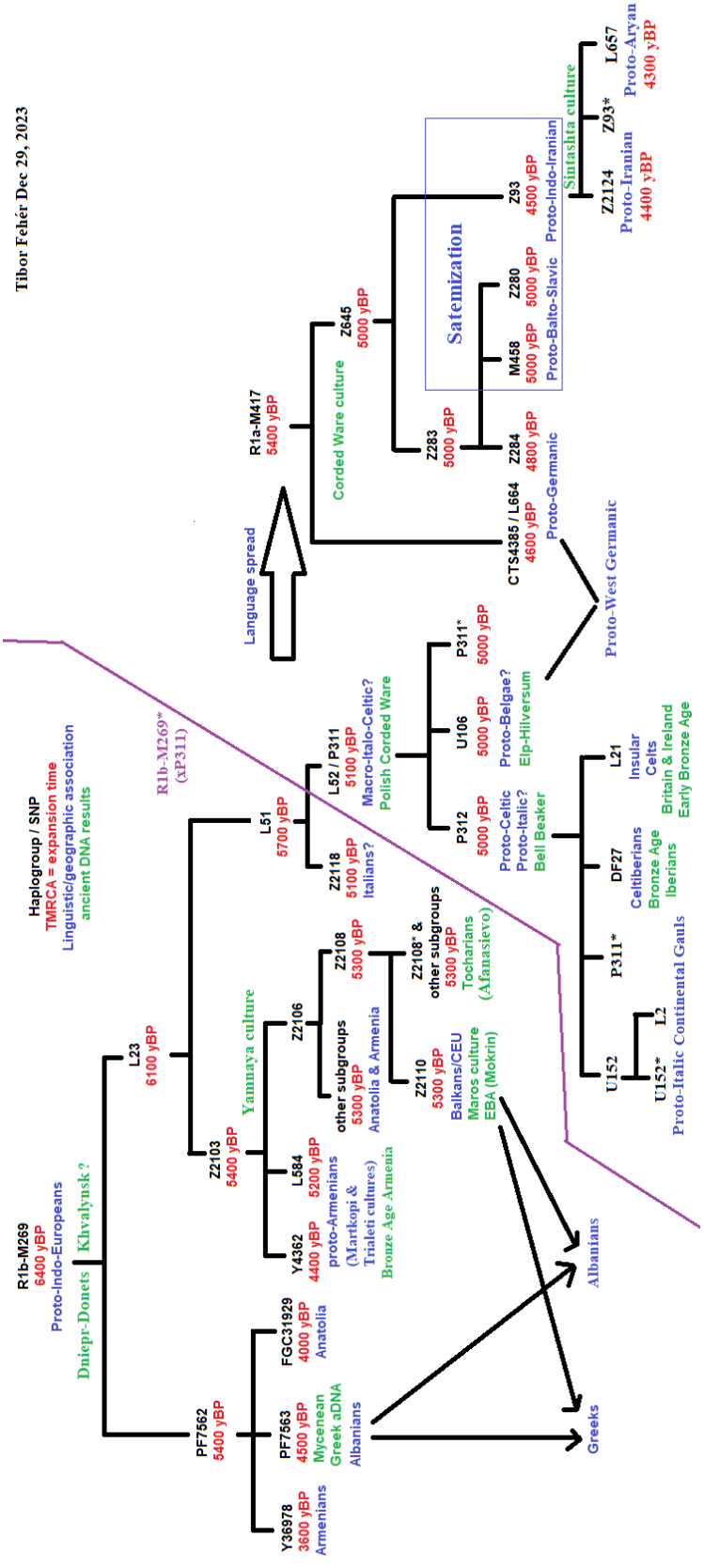
Language Family	Correlating Haplogroup	Value	Relationship
Insular Celtic	Hg R1b-L21	0,939	Very Strong
Insular Celtic	Hg I2-M223	0,452	Measurable
Gallo-Romance	Hg R1b-P312*(xL21,U152)	0,813	Strong
Gallo-Romance	Hg R1b-L2	0,487	Measurable
Old Italic	Hg R1b-U152*(xL2)	0,623	Medium
Old Italic	Hg E1b-V12	0,492	Measurable
Old Italic	Hg I2*(xM223,Y3120)	0,427	Measurable
Old Italic	Hg R1b-L2	0,331	Measurable
Germanic	Hg I1	0,781	Strong
Germanic	Hg R1a-L664	0,708	Strong
Germanic	Hg Q-L527 & Q-L804	0,703	Strong

Germanic	Hg R1b-U106	0,662	Strong
Germanic	Hg R1a-Z282*(xZ280,M458)	0,628	Medium
Germanic	Hg I2-M223	0,581	Medium
Baltic	Hg N-L1025	0,962	Very Strong
Baltic	Hg R1a-Z280	0,507	Medium
Slavic	Hg R1a-M458	0,830	Strong
Slavic	Hg I2-Y3120	0,751	Strong
Slavic	Hg R1a-Z280	0,711	Strong
Albanian	Hg J2b-M241>L283	0,831	Strong
Albanian	Hg E1b-L618>V13	0,782	Strong
Albanian	Hg R1b-M269*(xP311)	0,511	Medium
Greek	Hg J2b*(xM241)	0,809	Strong
Greek	Hg E1b-V12	0,482	Measurable
Greek	Hg E1b-L618>V13	0,391	Measurable
Greek	Hg E1b-Z780>M123	0,353	Measurable
Greek	Hg R1b-M269*(xP311)	0,293	Weak (see below)
Armenian	Hg R1b-M269*(xP311)	0,506	Medium
Iranian	Hg R1a-Z93	0,364	Measurable
Iranian	Hg G2	0,341	Measurable
Indo-Aryan	Hg R2	0,712	Strong
Indo-Aryan	Hg C1 (CxM217)	0,594	Medium
Indo-Aryan	Hg H1	0,392	Measurable
Indo-Aryan	Hg L-M20	0,340	Measurable
Indo-Aryan	Hg R1a-Z93	0,315	Weak (see below)

The Greek branch is the most difficult to define genetically, one strong correlation was shown with the relatively rare J2b1-M205 subgroup, which appears mostly in various Near Eastern populations and could be a pre-Indo-European substrate specific for Greece. Based on the results, we draw the tree of Indo-European relationships.

ISSN: 0036-5483

Figure 7: Correlation of Y-DNA subbranches and Indo-European subfamilies



Discussion

Haplogroups with two or three relatively weak correlation might result from two different processes. I2-M223 is an example in North-Western Europe, which correlates with the Germanic (0,581) and the Celtic (0,452) language families, but most likely represents the pre-IE Mesolithic-Neolithic populations of the area where these languages later spread (see Cassidy et al. 2020). We can see a similar correlation value (0,427) between I2* (xY3120,M223) and Old Italic languages due to Neolithic pre-IE substrate effects (especially in Sardinia, where I2a-PF4188 started a demographic growth 3600 years ago, several centuries after the arrival of IE-speaking Bell Beakers).

On the other hand, Haplogroup G2 shows a different pattern. It correlates with NW Caucasian languages (0,454) and Kartvelian languages (0,412) and to some extent with Iranian (0,341) but importantly with no other IE language group. The NWC and Kartvelian connection here is clearly strong and original when we look at absolute frequencies, therefore the lack of a strong correlation result can be attributed to the low resolution of the haplogroup included in the study. Had we shown more elaborate sub-haplogroups (like G-L1264, G-L1266 or G-Z6552), the correlation result would have been clearer.

Similar unresolved relationship of common origin is shown by R1a-Z280 for Balts and Slavs as well as by R1a-Z93 for Iranians and Indo-Aryans, whose grouping together into the Balto-Slavic and Indo-Iranian subgroups would result in a stronger correlation value. For example, Balto-Slavic grouping together gives 0,893 for R1a-Z280, while separately produces 0,711 for Slavic and 0,507 for Baltic. In a similar manner, R1a-Z93 for Indo-Aryan classified together increases to 0,490 from 0,31-0,36 when classified separately.

Several Indo-European subfamilies showed clear correlation with specific Y-DNA subclades. Baltic showed 0,962 with N-L1025 and **Insular Celtic** showed 0,939 with R1b-L21. Slavic showed strong correlation with R1a-M458 (0,830) and R1a-Z280 (0,710), both being subgroups of R1a-Z283. It is clear from the data that proto-**Balto-Slavic** corresponds to R1a-Z280, which is the core shared by every branch of Balto-Slavs while in different geographic directions they absorbed different “secondary dominant” elements, like N-L1025 from Baltic Finns distinguishing the Baltic subclade; R1a-M458 being most common among West Slavs and I2-Y3120 dominating South Slavs. East Slavic core contains the most of R1a-Z280 but also some of all other admixed groups which is in line with a proto-Balto-Slavic homeland in the present-day Belarus-Northern Ukraine (Pripyat-Polissia) area (see more detailed analysis of the Slavic homeland and genetic results in Kushniarevich et al. 2015). The original source of I2-Y3120 is rather awkward as earlier cousins of this clade are from Western Europe, so the first Y3120 might have arrived to the proto-Slavic homeland with eastward Hallstatt-La Tene

migrations in the 5th-4th centuries BCE. However, this is purely hypothetical without aDNA evidence.

We tried to distinguish **Old Italic** and **Gallo-Romance** in the Italic subgroup of IE languages as the process of Romanization of waste originally Celtic-speaking areas is known from historical sources. At first trial, no haplogroup showed significant correlation with Old Italic apart from the I2* substrate mentioned above, while Gallo-Romance areas showed significant correlation with Hg R1b-P312*(xU152, L21) and with R1b-U152.The classification of Toscana was problematic, as it was neither Italic nor Gallic but Etruscan-speaking in antiquity. However, due to Toscana being south of the La Spezia-Rimini line and having early Steppe-derived proto-Italic genetic admixture before Etruscan times (Posth et al. 2021), we have put Toscana into the Old Italic group. In order to understand the Celtic vs. Italic split better, we have split the R1b-U152 group into an R1b-U152*(xL2) and an R1b-L2 subgroup, considering the near-total dominance of L2 ancient DNA among U152 derived samples outside Italy. While L2 was also numerous inside Italy, at least other U152 subclades (Z192, Z36, Z56) occur there. The effort lead to success, as the R1b-U152*(xL2) group showed a medium (nearly strong) correlation value with Old Italic speaking areas (0,623) while showing very limited but positive correlation with (Western) Gallo-Romance areas (0,233). On the other hand, R1b-L2 showed higher correlation with Gallo-Romance (0,487) than with Old Italic (0,331). Taking into account absolute numbers, U152* accounts for 15-27% in Northern Italy and Tuscany and 6-7% in Southern Italy today, while 1,5-4% in historically Roman Empire areas, which could fit the estimate of Goldhill (2006) about Roman citizens making up max. 10% of the population in Gaul and Iberia. However, R1b-L2 and not U152* dominates the ancient DNA in Iron Age Etruscan areas (Posth et al. 2021), which complicates drawing a conclusion. Further ancient DNA data and even higher resolution testing could lead to better distinction between various U152 subclades concerning Italic vs. Continental Celtic origin.

In case we would define an **Italo-Celtic** superfamily under PIE, it shows correlation with all branches of R1b-P311 (U152 0,567; L21 0,532; P312* 0,529; U106 0,343), and above 0,5 with all P312 branches, which supports the assumption of the R1b-P312 branch representing Proto-Italo-Celtic (see also Fehér 2021). Based on the R1b-U106 ancient DNA data in the Low Countries Bronze Age (see Patterson et al. 2022 and Olalde et al. 2018), well before the Germanization of the area, we might consider R1b-U106 being a distant branch of Italo-Celtic (Belgic?) or an undocumented Bronze Age language later fully absorbed by Celtic and Germanic speakers.

The **Germanic** branch of Indo-European is the most complex one with strong pre-IE substrate effects. R1a-L664 and R1a-Z283* (mostly meaning R1a-Z284) could have carried the language northwest-wards, with the new culture absorbing various hunter-gatherer (I1, Q-L804, Q-L527) and Neolithic (I2-M223) populations in Southern Scandinavia or the German-

Polish Plain. R1b-U106 or some subgroups of it likely joined the spread of Germanic later, in the Iron Age period (see above). The very high amount and diverse background of originally non-IE substrates in Germanic languages might reflect an integrative aspect of Germanic society, its ability to integrate a larger number of genetically different people which is still underway globally especially in English-speaking countries.

R1b-M269* (xP311), mostly covering the Z2103 subgroup (commonly referred as Eastern R1b) showed medium correlation with **Albanian** (0,511), **Armenian** (0,506) and weak with **Greek** (0,293). A deeper subgroup analysis might also help here to reach higher correlation values with various subgroups. R1b-Z2103>Y4362 and >L584 are clearly linked to the spread of proto-Armenians through present-day Georgia to the Armenian Highlands in the Early Bronze Age with Martkopi and Trialeti cultures (see Lazaridis et al. 2022). Contemporary Albanians unite prehistoric Illyrians (J2b2-M241>L283) and Thracians (E1b-V13) genetically while their original IE component is most likely R1b-M269*(xP311), including both R1b-PF7562 (L23-) and R1b-Z2103>Z2110. The former group also appears in Mycenaean Bronze Age Greeks (see Skourtanioti et al. 2023), while most of the Albanian and Greek Z2110 samples belong to the Z2705 subclade which is easily recognizable due to its DYS393=13 and DYS385=11-11 off-modal values. The lack of any R1b or R1a subgroup clearly correlating with Greek and the rarity of R1b and absence of R1a in Middle Bronze Age Greek ancient DNA suggests that the Greek language has spread much more through elite dominance than mass movement of people, or with movements of early absorbed pre-IE substrate groups belonging to various Hg E1b-M35 and Hg J2 subgroups. For example, J2a-M319 was found to be the dominant lineage of Minoans in Crete (Skourtanioti et al. 2023). The similarity of supposedly language-bearing R1b-PF7562 and Z2110 clades as well as the common correlation with E1b-V13 between Albanians and Greeks might indicate an earlier common origin of these clades, as described by Hyllested and Joseph (2022):

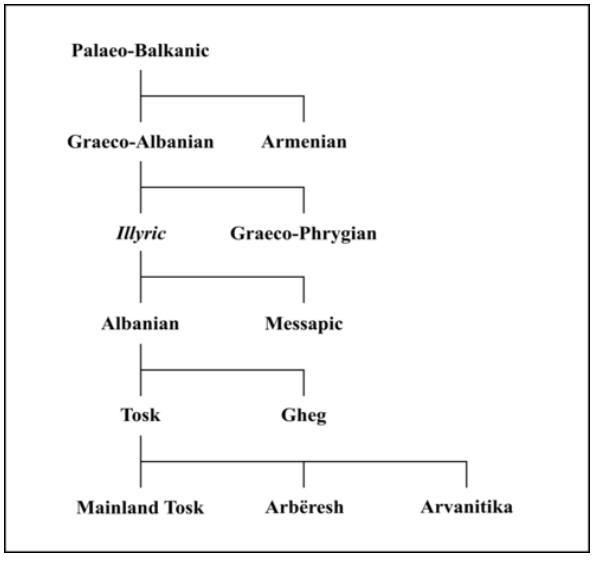


Figure 8: Genetic relationship of Balkans languages based on Hyllested and Joseph

As for the non-IE language families, **Berber** languages have a very strong correlation with Hg E-M81, whose demographic expansion starts around 2700 yBP. The M35* correlation could cover Hg E-V1039 with expansion time 4900 yBP and E-V65>PF2159 with 3200 yBP. While some M35* lines and Hg H2 might represent the Neolithic Capsian culture (8000-2700 BCE) in Northwest Africa. **Semitic** language speakers show very strong correlation with J1-P58 and medium correlation with E1b-Z780>M123 and T-M70. Taking into account the strong correlation of J1*(xP58) with Northeast Caucasian languages, even without including Cushitic and Chadic languages into our analysis, the numbers support the E-M35 and T-M70 groups being the marker of Afro-Asiatic speakers, while J1-P58 entering the gene pool in the Near East later from possibly Hurro-Urartuan sources from the North, which also might be the basis for the traditional southward migration story of Abraham, common forefather of Cohen Jews and Arabs (J1-YSC0000234).

Niger-Congo languages strongly correlate with Hg E*(xE1b) as well as E-M2, or in general with E*(xM35), which is dominant in Sub-Saharan Africa; and also shows medium correlation with Hg A00-B, which is more common among Pygmy, Nilotic and Khoisan ethnic groups not included in this study.

The strongest correlation with the spread of **Turkic languages** was shown by Hg C2-M217, which is also the dominant Y-DNA signature of Mongolian and Tungusic-speaking populations. This phenomenon thus supports the widely debated existence of the Altaic language family. The R1b*(xM269) correlation with Turkic comes down mostly to R1b-M478/M73. Interestingly, the origin of this group is known from the Kunda and Narva cultures of the Eastern Baltic area between 6000 and 8000 yBP, representing Eastern Hunter Gatherers (EHG) and only later appear in Central Asia among Botai herdsmen around 5500-5200 yBP, representing a West to East gene flow on the Eurasian Steppe predating the Indo-European migrations (Tocharians and later Indo-Iranians). This group was likely absorbed by the Altaic-speaking proto-Turks in Southern Siberia/Central Asia in a similar manner like they did with R1a-Z93 Indo-Iranians, R1b-Z2109 Tocharians and N Uralic-speakers. The relatively limited (below 0,5) value for Turkic correlation with Hg Q points out also that proto-Turkic likely appeared on the Eurasian Steppe only with the emerging Xiongnu containing the first C2-M217 aDNA from the 3rd century BCE (see Jeong et al. 2020), and thus no pre-Xiongnu Steppe or South Siberian population could be considered Turkic-speaking. This would also enable the equation of the early Xiongnu language with late proto-Turkic.

Uralic languages have been connected with the spread of Haplogroup N (see Illumäe et al. 2016) and corresponding autosomal admixture (Zeng et al. 2023) including solving the genetically problematic Hungarian question by the present authors both among contemporary Hungarians (see Post et al. 2019) and ancient Magyars (see Fóthi et al. 2020). The N-Uralic relationship is clearly shown by the correlation value above 0,9. The subgroup N-L1025 was

absorbed by Balto-Slavic speakers around 2800-2700 yBP in the East Baltic area and became the marker of Baltic speakers with very limited occurrence among Slavic tribes.

The correlation of **Dravidian language**-speakers in South India with Hg H1-M2826 is likely a result of Paleolithic substrate effect due to the presence and diversity of Hg H1 in India going back more than 40 thousand years. Language spread is more likely connected to secondary correlation with Hg L-M357/L1307 (TMRCA 7800 yBP) and L-M27 (TMRCA: 8200 yBP) from the Near East with farmers through present-day Iran and Turkmenistan towards the Indus Valley (see for example Jeitun culture aDNA NEO310 from Allentoft et al. 2023). South Asia's most common R2 subgroup, R2-L295 has similar TMRCA (7700 yBP) similarly to the most common J2 subgroup, J2b2-Z2432 (8400 yBP) and these two might have been part of the same movement, even though this is not shown by the correlation value and remains purely hypothetical.

As expected, **Sino-Tibetan languages** show very strong correlation with Hg O, but also some with Hg D, which is most frequent in Tibet and Japan. This paper did not aim to resolve the correlation of various South-East Asian language families with Hg O and D sub-haplogroups. To that end, a separate study would be needed with deep Hg O, D, C and N resolutions which are not relevant for Western Eurasia.

Conclusion

We can conclude that simple mathematical correlation between language families and deep Y-DNA haplogroup assignments can significantly contribute to the resolution of problems debated among linguists. We have to admit that the genetic nature of the Anatolian branch is still unresolved. In case of the sub-branching of Indo-European languages except Anatolian, we can conclude that PIE first branched into three well-defined groups, 1) a South-Eastern branch covering the original Yamnaya territory and containing the Armenian, Paleo-Balkans and Tocharian branches; 2) a Western branch in present-day Southeast Poland representing a specific branch of Corded Ware Pottery spreading towards the Rhine and later spreading Bell Beaker culture (Italo-Celtic and potentially an extinct branch in the Lower Rhine area called here "Belgic"); 3) a Northern branch moving to the forested areas and spreading the typical Corded Ware culture (proto-Germanics crossing the Baltic with Battle Axe culture; proto-Balto-Slavs staying in the Upper Dnieper area and proto-Indo-Iranians moving eastwards through Fatyanovo-Balanovo and Sintashta cultures).

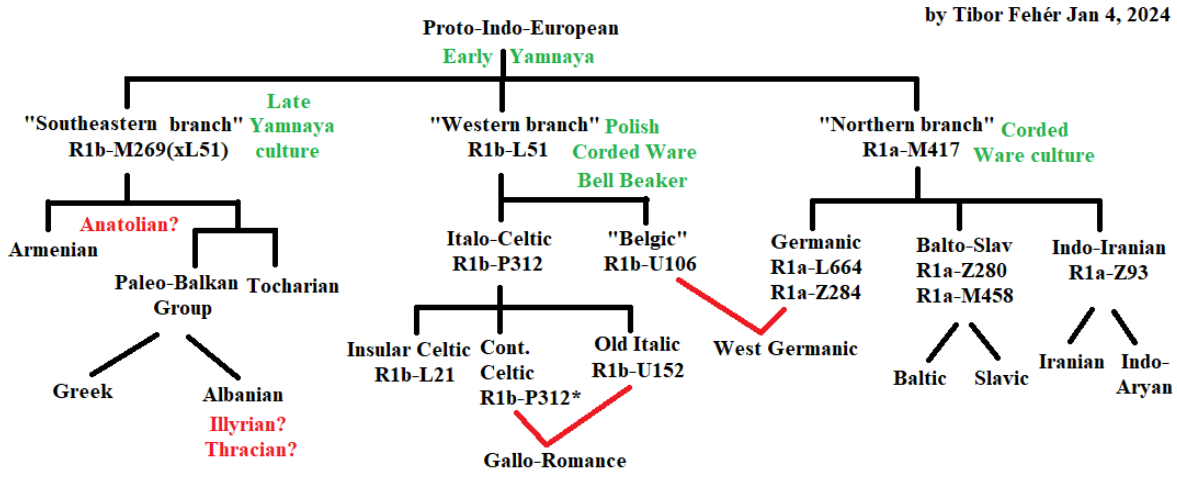


Figure 9: simplified connections among Indo-European branches as revealed by Y-DNA

The similarities between the Germanic and Italo-Celtic branches in this model results from a strong substrate effect in the Iron Age, when proto-Germans absorbed the R1b-U106 group speaking a language close to the Italo-Celtic branch. The split of Greek and Albanian, and their relationships with the Illyrian and Thracian languages cannot be clearly solved, especially due to the lack of deep resolution Greek data. Illyrian (J2b-L283) and Thracian (E1b-V13) both have strong non-IE substrates which are present in both Albanian and Greek. Further research, especially ancient DNA data is needed to define the place and genetic setup of the proto-Anatolian branch.

The study also supports many existing relationships between non-European language families as well as the strong connection between Y-DNA Haplogroup N and Uralic-speaking peoples. Our data also points on a proto-Altaic nucleus spoken by Y-DNA Haplogroup C-carrying people in Northeast Asia, and a later assimilation of the Hg N, Q and R1a elements by Turkic-speaking peoples. Further research is needed on this matter, but Y-DNA data suggests that there were no Altaic languages west of the Lake Baykal-Mongolia area before the late Xiongnu era, which would also open up the question which kind of non-Indo-European, non-Altaic language was spoken by early Xiongnu people.

Our study proves that analysing correlations between language families and Y-DNA subhaplogroups is valid, but should be based on high resolution Y-DNA data with deep subgroups younger than 5000 years, to obtain meaningful results.

Conflict of interest

The authors declare no conflict of interest.

Funding statement

The authors did not receive support from any organization for the submitted work.

References

Allentoft ME, Sikora M, Sjögren KG, et al. Population genomics of Bronze Age Eurasia. Nature. 2015 Jun 11;522(7555):167-72. doi: 10.1038/nature14507. PMID: 26062507.

Allentoft ME, Sikora M, Refoyo-Martínez A, et al. Population genomics of postglacial western eurasia. BioRxiv preprint December 21, 2023.

Bekada A, Fregel R, Cabrera VM, Larruga JM, Pestano J, Benhamamouch S, González AM. Introducing the Algerian mitochondrial DNA and Y-chromosome profiles into the North African landscape. PLoS One. 2013;8(2):e56775. doi: 10.1371/journal.pone.0056775. Epub 2013 Feb 19. PMID: 23431392; PMCID: PMC3576335.

Berniell-Lee, G, Calafell, F, Bosch E, et al., Genetic and Demographic Implications of the Bantu Expansion: Insights from Human Paternal Lineages, *Molecular Biology and Evolution*, Volume 26, Issue 7, July 2009, Pages 1581–1589, https://doi.org/10.1093/molbev/msp069

Boattini A, Martinez-Cruz B, Sarno S, Harmant C, Useli A, Sanz P, et al. (2013) Uniparental Markers in Italy Reveal a Sex-Biased Genetic Structure and Different Historical Strata. PLoS ONE 8(5): e65441. https://doi.org/10.1371/journal.pone.0065441

Budenz, J. 1880. O. Donner: Die gegenseitige verwandtschaft der finnisch-ugrischen sprachen. Nyelvtudományi Közlemények 16. 120–144.

Cassidy, L.M., Maoldúin, R.Ó., Kador, T. *et al.* A dynastic elite in monumental Neolithic society. *Nature* 582, 384–388 (2020). https://doi.org/10.1038/s41586-020-2378-6

Charles B. Chang, A novelty effect in phonetic drift of the native language, Journal of Phonetics, Volume 41, Issue 6, 2013, Pages 520-533, ISSN 0095-4470, https://doi.org/10.1016/j.wocn.2013.09.006.

Di Cristofaro J, Pennarun E, Mazières S, Myres NM, Lin AA, Temori SA, et al. (2013) Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. PLoS ONE 8(10): e76748. https://doi.org/10.1371/journal.pone.0076748

Drinka, Bridget. Phylogenetic and areal models of Indo-European relatedness: The role of contact in reconstruction. Journal of Language Contact. 6. 379-410. (2013). 10.1163/19552629-00602009.

Fehér, T. Celtic and Italic from the West – the Genetic Evidence. 2021, Academia Letters.

Fehér, T. High Resolution Paternal Genetic History of Ireland and its Implications for Demographic History. 2023, EMANIA - Bulletin of the Navan Research Group, No. 26.

Fehér, T., Németh, E., Vándor, A. *et al.* Y-SNP L1034: limited genetic link between Mansi and Hungarian-speaking populations. *Mol Genet Genomics* 290, 377–386 (2015). https://doi.org/10.1007/s00438-014-0925-2

Fóthi, E., Gonzalez, A., Fehér, T. *et al.* Genetic analysis of male Hungarian Conquerors: European and Asian paternal lineages of the conquering Hungarian tribes. *Archaeol Anthropol Sci* 12, 31 (2020). https://doi.org/10.1007/s12520-019-00996-0

Geisler, H. & J.-M. List 'Beautiful trees on unstable ground. Notes on the data problem in lexicostatistics', in H. Hettrich (ed) Die Ausbreitung des Indogermanischen. Thesen aus Sprachwissenschaft, Archäologie und Genetik (2013) (Wiesbaden: Reichert).
Goldhill, Simon (2006). *Being Greek Under Rome: Cultural Identity, the Second Sophistic and the Development of Empire*. Cambridge: Cambridge University Press.

Gray, R., Atkinson, Q. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435–439 (2003). https://doi.org/10.1038/nature02029

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

Hammer MF, Behar DM, Karafet TM, Mendez FL, Hallmark B, Erez T, Zhivotovsky LA, Rosset S, Skorecki K. Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. Hum Genet. 2009 Nov;126(5):707-17. doi: 10.1007/s00439-009-0727-5. Epub 2009 Aug 8. PMID: 19669163; PMCID: PMC2771134.

Hyllested A, Joseph BD. Albanian. In: Olander T, ed. *The Indo-European Language Family: A Phylogenetic Perspective*. Cambridge University Press; 2022:223-245.

Ilumäe AM, Reidla M, Chukhryaeva M, et al. Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families. Am J Hum Genet. 2016 Jul 7;99(1):163-73. doi: 10.1016/j.ajhg.2016.05.025. PMID: 27392075; PMCID: PMC5005449.

Jeong C, Wang K, Wilkin S, et al. A Dynamic 6,000-Year Genetic History of Eurasia's Eastern Steppe. Cell. 2020 Nov 12;183(4):890-904.e29. doi: 10.1016/j.cell.2020.10.015. Epub 2020 Nov 5. PMID: 33157037; PMCID: PMC7664836.

King RJ, Di Cristofaro J, Kouvatsi A, et al. The coming of the Greeks to Provence and Corsica: Y-chromosome models of archaic Greek colonization of the western Mediterranean. BMC Evol Biol. 2011 Mar 14;11:69. doi: 10.1186/1471-2148-11-69. PMID: 21401952; PMCID: PMC3068964.

Kushniarevich A, Utevska O, Chuhryaeva M, et al. Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. PLoS One. 2015 Sep 2;10(9):e0135820. doi: 10.1371/journal.pone.0135820. PMID: 26332464; PMCID: PMC4558026.

Lazaridis, I, Alpaslan-Roodenberg, S, Acar, A *et al.* The genetic history of the Southern Arc: A bridge between West Asia and Europe. *Science* 377, eabm4247 (2022). DOI:10.1126/science.abm4247

Mathieson I, Lazaridis I, Rohland N. et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015 Dec 24;528(7583):499-503. doi: 10.1038/nature16152. Epub 2015 Nov 23. PMID: 26595274; PMCID: PMC4918750.

Myres NM, Rootsi S, Lin AA, et al. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. Eur J Hum Genet. 2011 Jan;19(1):95-101. doi: 10.1038/ejhg.2010.146. Epub 2010 Aug 25. PMID: 20736979; PMCID: PMC3039512.

Olalde, I., Brace, S., Allentoft, M. *et al.* The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555, 190–196 (2018). https://doi.org/10.1038/nature25738

Pearson, Karl (20 June 1895). "Notes on regression and inheritance in the case of two parents". Proceedings of the Royal Society of London. 58:

Post, H., Németh, E., Klima, L. *et al.* Y-chromosomal connection between Hungarians and geographically distant populations of the Ural Mountain region and West Siberia. *Sci Rep* 9, 7786 (2019). https://doi.org/10.1038/s41598-019-44272-6

Posth, C *et al.* Zaro, V, Spyrou, M et al. The origin and legacy of the Etruscans through a 2000-year archeogenomic time transect.*Sci. Adv.*7,eabi7673(2021).DOI:10.1126/sciadv.abi7673

Ramos-Luis E, Blanco-Verea A, Brión M, Van Huffel V, Sánchez-Diz P, Carracedo A. Y-chromosomal DNA analysis in French male lineages. Forensic Sci Int Genet. 2014 Mar;9:162-8. doi: 10.1016/j.fsigen.2013.12.008. Epub 2013 Dec 29. PMID: 24528594.

Sarno, S., Tofanelli, S., De Fanti, S. *et al.* Shared language, diverging genetic histories: high-resolution analysis of Y-chromosome variability in Calabrian and Sicilian Arbereshe. *Eur J Hum Genet* 24, 600–606 (2016). https://doi.org/10.1038/ejhg.2015.138

Sengupta S, Zhivotovsky LA, King R, et al. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet. 2006 Feb;78(2):202-21. doi: 10.1086/499411. Epub 2005 Dec 16. PMID: 16400607; PMCID: PMC1380230.

Underhill, P., Poznik, G., Rootsi, S. *et al.* The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *Eur J Hum Genet* 23, 124–131 (2015). https://doi.org/10.1038/ejhg.2014.50

Skourtanioti, E., Ringbauer, H., Gnecchi Ruscone, G.A. *et al.* Ancient DNA reveals admixture history and endogamy in the prehistoric Aegean. *Nat Ecol Evol* 7, 290–303 (2023). https://doi.org/10.1038/s41559-022-01952-3

Szinnyei, J. 1884. Az ugor népek. In *Budenz-album. Budenz József XXV éves nyelvészeti müködése emlékére kiadják tanítványai,* 26–32. Budapest: Akadémiai könyvkereskedés (Knoll Károly és társa).

Voskarides, K., Mazières, S., Hadjipanagi, D. *et al.* Y-chromosome phylogeographic analysis of the Greek-Cypriot population reveals elements consistent with Neolithic and Bronze Age settlements. *Investig Genet* 7, 1 (2016). https://doi.org/10.1186/s13323-016-0032-8

Zeng TC, Vyazov LA, Kim A *et al*. Postglacial genomes from foragers across Northern Eurasia reveal prehistoric mobility associated with the spread of the Uralic and Yeniseian languages. bioRxiv 2023.10.01.560332; doi: https://doi.org/10.1101/2023.10.01.560332

YFull tree (constantly updated): https://www.yfull.com/tree/ (data obtained as of Dec 23, 2023)

Kazakh DNA project:
https://www.familytreedna.com/public/alash/default.aspx?section=yresults

Indo-European typology by Ali Eminov: https://www.flickr.com/photos/aliarda/5222418678