More rule than exception: Parallel evidence of ancient migrations in grammars and genomes of Finno-Ugric speakers

Patrícia Santos[1, 2], Gloria Gonzalez-Fortes[1], Emiliano Trucchi[3], Andrea Ceolin[4], Guido Cordoni[5], Cristina Guardiano[6], Giuseppe Longobardi[7], Guido Barbujani[1]*

[1] Dipartimento di Scienze della Vita e Biotecnologie, Università di Ferrara, Italy
[2] Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche 5199 (UMR5199 PACEA), Université de Bordeaux, France
[3] Department of Life and Environmental Sciences, Marche Polytechnic University, Italy
[4] Department of Linguistics, University of Pennsylvania, USA
[5] School of Veterinary Medicine, University of Surrey, UK
[6] Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia, Italy
[7] Department of Language and Linguistic Science, University of York, UK
* Correspondence: g.barbujani@unife.it

Abstract: To reconstruct aspects of human demographic history, linguistics and genetics complement each other, reciprocally suggesting testable hypotheses on population relationships and interactions. Relying on a linguistic comparative method exclusively based on syntactic data, here we focus on the complex relation of genes and languages among Finno-Ugric (FU) speakers, in comparison to their Indo-European (IE) and Altaic (AL) neighbors. Syntactic analysis supports three distinct clusters corresponding to these three Eurasian families; yet, the outliers of the FU group show linguistic convergence with their geographical neighbors. By analyzing genome-wide data in both ancient and contemporary populations, we uncovered remarkably matching patterns, with north-western FU speakers linguistically and genetically closer in parallel degrees to their IE-speaking neighbors, and eastern FU speakers to AL-speakers. Therefore, our study indicates plausible secondary convergence in the syntax of languages of different families, providing evidence that such interference effects were accompanied, and possibly caused, by recognizable processes at the population level. In particular, based on the comparison of modern and ancient genomes, our analysis identified the Pontic-Caspian steppes as the possible origin of the demographic processes that led to the expansion of the FU into Europe.

Keywords: genomes; languages; genetic and linguistic distances; human migrations; culture

1. Introduction

Darwin proposed that linguistic diversity along human history tends to be correlated with the biological differentiation of populations [1]. Indeed, factors isolating populations from each other, such as geographical distance and barriers to migration, are likely to promote both biological and cultural divergence, whereas factors favouring contacts have the opposite effect at both levels [2–5]. In fact, despite élite dominance and other processes of horizontal language transmission creating local mismatches [6], parallel genetic and linguistic changes have often appeared as the rule rather than the exception [2,7–12]. This implies that linguistic diversity may offer a set of testable hypotheses about the demographic processes shaping genetic diversity, and vice versa.

Until recently, though, comparative studies of genes and languages suffered from serious limitations, simply because of the data available. On the one hand, only seldom were whole genomes considered in these comparisons. On the other, classical etymological comparison of vocabulary items, still normally used to measure language differences even in modern quantitative studies (see e.g. [10,13,14]), work well within a language family; but words cannot be used for broader comparisons: for, by definition, across different language families there are no recognizable common etymologies (i.e. lexical cognates; see Ref. [15] for an important attempt to remedy some of these problems). However, the Parametric Comparison Method (PCM) [16–19]), which explores the phylogenetic information contained in the generative rules of syntax [4,20,21], has in principle overcome the limitations of vocabulary-based taxonomic methods, paving the ground for comparisons across language families. Through parameters, i.e. abstract and universally definable syntactic polymorphisms, the PCM quantifies language differences/similarities across languages into a synthetic measure. Such similarities may reveal both vertically inherited ancient identities but also horizontally (secondarily) exchanged properties (see Ref. [22]).

In this study, through a multidisciplinary approach comparing grammars and genomes, we contribute to a better understanding of population diversity, both cultural and biological, in Western/Central Eurasia. We shall focus on Altaic- (hereafter: AL) Finno-Ugric- (FU) and Indo-European- (IE) speaking populations, with a special emphasis on FU speakers. The reason is that FU appears as a possible exception to the general gene-language correspondence, one worth some deeper investigation. Indeed, its monophyletic unity was acknowledged linguistically already in the 18th century [23] and remains virtually undisputed (with the possible caveats in Ref. 24), but FU-speaking populations fail to be identified as a genetic group. In particular, the westernmost FU-speaking populations in Central and Northern Europe have been shown to display peculiar properties in a study of their gene-syntax-geography relations [12].

As for Indo-European, despite a long tradition of studies, it is still debated whether early IE languages came into Europe from the Pontic-Caspian steppes (and spread west in the Bronze Age [24,25]) or from Anatolia (and spread with the dispersal of early Neolithic farmers [14,26]). Thus, we compared the syntax and the genomes of several AL- FU- and IE-speaking populations with the available genome-wide data, both contemporary and ancient, in the area of interest [27]. Of particular interest was one Bronze-Age population from the Pontic-Caspian steppe, the Yamnaya, the likely source of the Bronze-Age migration leading to a Westwards diffusion of DNA of Central Asian origin and, according to some authors, of IE languages in Europe [28–30]. By contrast, a recent analysis of Asian genomes suggested that the spread of IE languages in South Asia and Anatolia may have little, if anything, to do, with migration from the Pontic-Caspian steppes [31]. An analogous uncertainty surrounds the homeland of early Uralic-speakers, whether in the river Volga basin [25] or further East, in Siberia [32].

Our multidisciplinary approach comparing grammars and genomes will ultimately help us better frame the evolution of this cultural and biological diversity in Western/Central Eurasia, reinforcing the idea of widespread congruence between the two types of variables [36].

2. Materials and Methods

Genomic dataset

The dataset analyzed in this study comprises the high-coverage sequenced genomes of 45 individuals from 17 populations from Eurasia (Supplementary Table S1). The samples were collected from Pagani et al. (2016) [33] and downloaded from the public database ENA (European Nucleotide Archive). For the sake of equal representation, a random subset of three individuals per population was chosen for populations with a larger sample size, to perform all the analyses.

Ancient and modern Genome-Wide SNP array data from Ref. [34] were used to estimate Outgroup *f3*-statistic and *qpAdm* analysis (Supplementary Tables S2 and S3, respectively).

Dataset preparation

Samples from Ref. [33] were in Complete Genomics MasterVar format files (reads mapped against the human genome reference hg19/GRCh37). The MasterVar file was converted into a Variant Call Format (VCF) by the cgatool mkvf (version 1.8.0.1) from Complete Genomics. The VCF file created only contains SNP variants called with a high confidence (>40 dB). All the VCF files from the different individuals were merged using BCFtools (version v1.6-36) merge with the option '-m none' to output the multiallelic sites in different lines. All duplicated variants were excluded from the data. The VCF files were phased using SHAPEIT2 (version v2.r837) using the 1000 Genomes phase 3 haplotypes as a reference panel, as recommended. Heterozygous sites not present in the 1000 Genomes data were left unphased. In the end, genotypes were obtained for 11,931,455 autosomal SNPs.

Principal Component Analysis

A general description of genetic variation was obtained by principal component analysis (PCA). QTLtools [35] (version v1.1) was used on scaled and centered genotype data on relatively independent (50 Kb distance) and non-rare variants (minor allele frequency = 0.05).

Genomic distances

Weir and Cockerham's genomic distances between populations were calculated by the 4P software [36] (version 1.0). Genomic regions that may be under selection were masked using bedtools subtract (version v2.26) and variants with a missing call rate exceeding 10% were excluded, resulting in a total of 9,881,752 autosomal SNPs.

Linguistic dataset

For the analysis of linguistic data by PCM [4,16,19,21,22]), we used the 94 binary parameters defining properties of nominal structures for 69 modern Eurasian languages recently employed in [22]. The original dataset of 69 languages has been reduced to a subset of 34 IE, FU and AL languages, to improve resolution on the 17 populations for which genetic data are available and their neighbors [17].

Data were available for the main FU subfamilies (Ugric, Volgaic, Permic, Balto-Finnic), with the exception of Lapp (Saami). For three languages, Mari (Cheremiss), Udmurt (Votyak) and Khanty (Ostyak), we encoded two diastratic variants. The two Even (Tungusic, AL) varieties are instead diatopic. For details see Ref. [22]. The relevant IE languages belong to three subfamilies, namely Indo-Iranian, Germanic and Slavic (see Supplementary Figure S1).

Linguistic distances and phylogenies

As in [22], a matrix of Jaccard-Tanimoto distances was derived from the data matrix, visualized in a heatmap. By means of a Principal Coordinate Analysis (PCoA, also called MDS, Multidimensional Scaling), calculated using the software PAST [37], we visualized the syntactic relationships between languages. We also represented the syntactic data in tree form through a UPGMA tree by a dedicated bootstrapped algorithm, in combination with the software PHYLIP [38] and Mesquite [39], and 2. through a character-based Bayesian tree, using BEAST v1.83 [40].

ChromoPainter and fineSTRUCTURE

ChromoPainter [41] (version v2) is a method to quantify distances between individual genomes. This method uses sampled chromosomes as "donors" and match (or "paint") other chromosomes to the donors' DNA, thus creating a cluster based on who shares blocks of SNPs. Each individual is "painted" as a combination of all the other sequences. In the heatmap, each square represents the number of DNA segments that each row (recipient) copies from each column (donor).

We used ChromoPainter output to cluster individuals into genetically homogeneous groups using fineSTRUCTURE [41] (version 2.1.3), a powerful approach for inference of fine-scale population structure from haplotype data. Each individual is presented as a matrix of non-recombining genomic chunks received from a set of multiple donors. Clusters of individuals are then inferred from the patterns of similarities among these copying matrices, by a Bayesian approach, and the tree is finally plotted using FigTree (version 1.4.2).

Outgroup f3-statistics

We performed an *f3* analysis using the *qp3Pop* package in ADMIXTOOLS (version 412). The outgroup *f3*-statistic (X, Y; Outgroup) is a function of shared branch length between two genomes, say X and Y, in the absence of admixture with the outgroup. Y is extracted from a set of individuals, among whom we look for the the most closely related to the individual under exam (X). Throughout the analysis we used the African Yoruba as an outgroup that we assumed to diverge from population X before all the other populations being analyzed. In this analysis, high values of *f3* indicate that X and Y are genetically closer.

The modern samples from Pagani et al. (2016) [33] used in this study were merged with the Yamnaya, Anatolian, Sintashta and Nganasan individuals from Ref. 38 and used as source populations. Variants with a missing call rate exceeding 10% were excluded, resulting in 249,286 SNPs suitable for the analysis.

Modelling admixture

Using *qpAdm* package in ADMIXTOOLS (version 412) we estimated the proportions of ancestry in a *Test* population deriving from a mixture of three reference populations by leveraging shared genetic drift with a set of outgroup populations. The reference populations used were: Yamnaya, Anatolia and Nganasan (used here as a proxy for the genetically still undescribed Siberian population). As outgroup populations we used: Han, Mbuti, Karitiana, Ulchi and Mixe. The detail: YES parameter was set, which reports a normally distributed Z-score for the goodness of fit of the model (estimated with a Block Jackknife).

3. Results

3.1. Linguistic analyses

3.1.1. Syntactic comparison

The PCoA inferred from syntactic data (Figure 1) shows a first, neat division between all the IE languages, with positive values of the first component (accounting for 78% of variation), while FU and AL are all found in the left, negative, area of the graph. In that area, the second PC (accounting for 11% of variation) separates FU from AL. In sum, each group appears to form a well-defined cluster. While the clouds corresponding to IE and AL are compact, although with individual outliers (Indo-Iranian and Buryat, respectively) the FU languages appear more scattered. Finnish and even more so Estonian fall particularly close to the IE languages. Such a resemblance between the Balto-Finnic group of FU and IE emerges even more neatly in the Bayesian phylogenetic analysis (Supplementary Figure S2), where the Balto-Finnic node joins the IE cluster rather than the FU one. The second important aspect that emerges from the PCoA is a split between IE and the other two groups, which might in turn hint at some closer FU-AL relatedness.

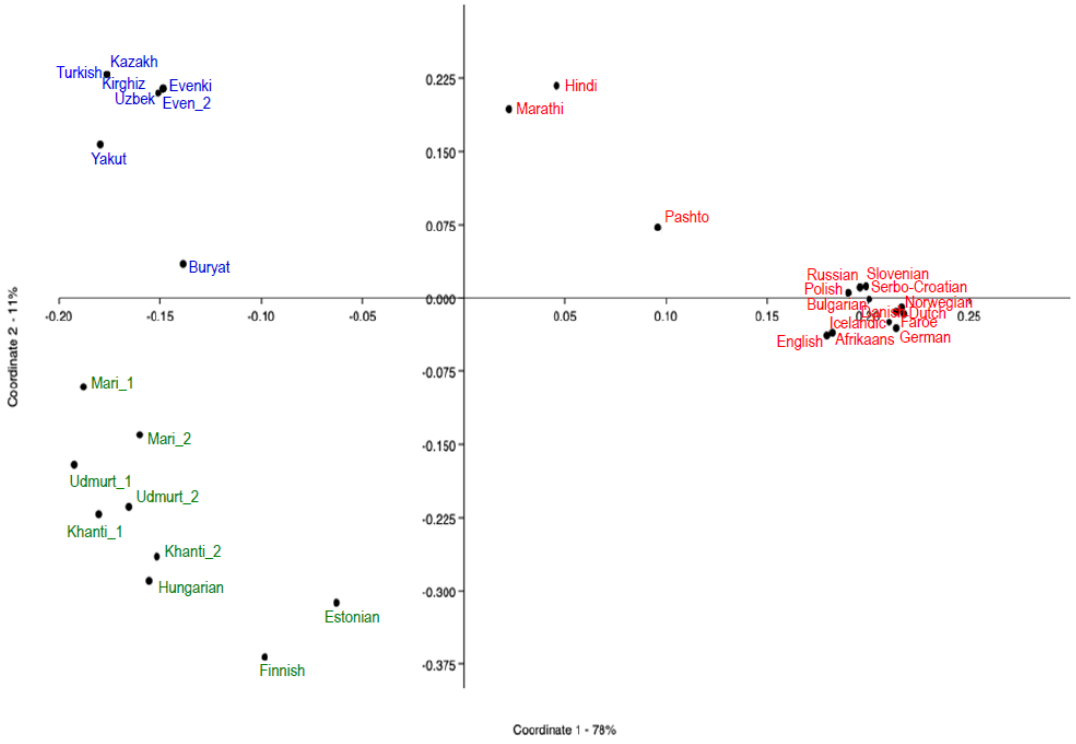


Figure 1. PCoA from the syntactic distances in 34 Eurasian populations. Language groups coded as follow: Finno-Ugric (green), Altaic (Blue), Indo-European (red).

In the UPGMA tree (Figure 2), languages from the same family, IE, FU and AL neatly cluster together without exception; FU languages form a monophyletic cluster within which the Balto-Finnic (Finnish and Estonian) and Ugric (Hungarian and Khanty) families are well identified, with the addition of a node comprising geographically closer Udmurt (Permic) and Mari (Volgaic) [12,17,42,43].

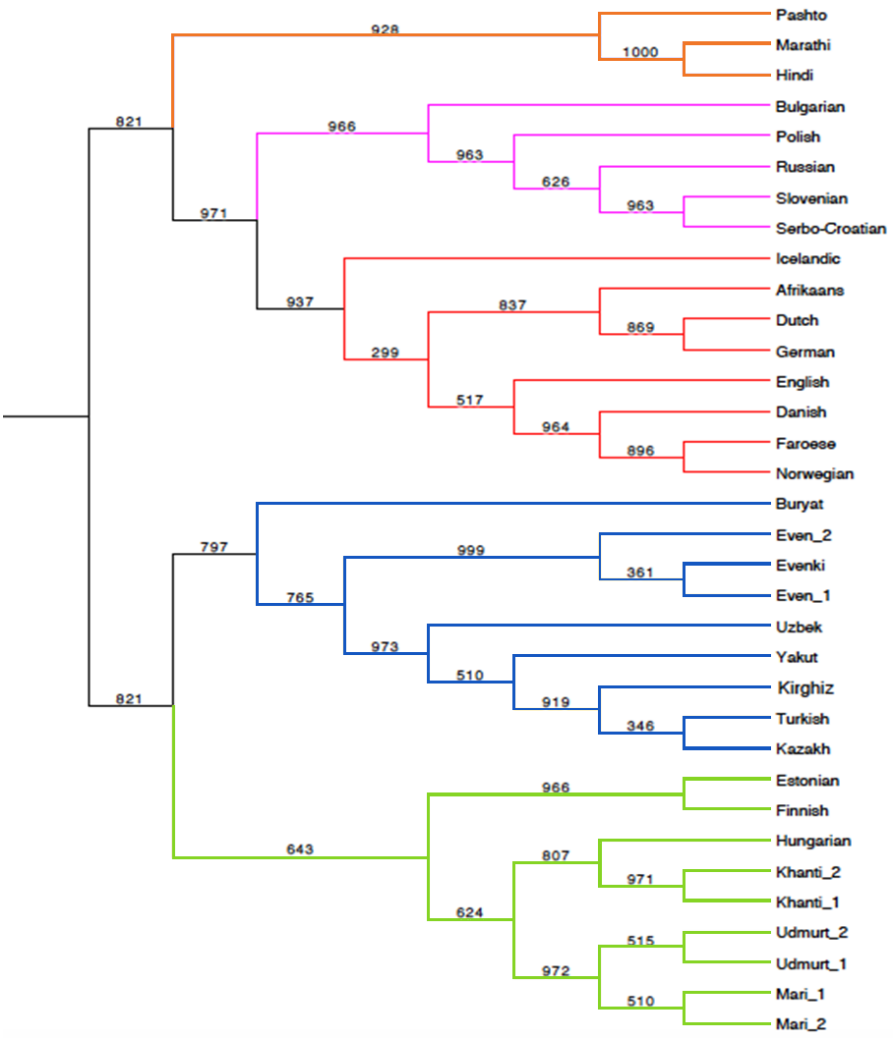


Figure 2. UPGMA tree inferred from the Jaccard syntactic distances. Bootstrap values, base =1000, at the nodes. Orange=Indo-Iranian IE, pink=Slavic IE, red=Germanic IE, Blue=AL, Green=FU.

The outlying positions of Balto-Finnic (Finnish and Estonians), Indo-Iranian (Pashto, Marathi and Hindi) and Mongol (Buryat) within the three groups is also visualized in the Heatmap of the syntactic distances (Supplementary Figure S3). Interestingly, the minimum syntactic distances between the European FU speakers Finnish and Estonians are with the steppe populations Udmurt and Mari, while Hungarians are closer to Khanty, an ethnic group from West Siberia, which however track its origins to the South Ural steppe [44]. Thus, all three extant Uralic languages in Europe seem to be linguistically connected to the Russian steppes.

3.2. Genetic comparison

3.2.1. Population structuring in Eurasia

We selected 17 populations, 7 speaking IE, 6 FU and 4 AL languages, for which whole-genome data were available (Figure 3a; Supplementary Table S1). The first principal component (Figure 3b) mostly reflects geography and separates eastern from western Eurasian populations, whereas the second component separates western Eurasians along a north-south cline. The AL-speaking populations fall

into a single cluster along the first PC axis. The European IE-speaking populations form a cluster along the PC2 axis, separated from the Iranians, the latter belonging to the Asian group of IE languages.

Conversely, the FU-speaking populations are scattered along the PC1 axis. Estonians fall within the IE diversity at the negative end of the X-axis, while Finns occupy an intermediate position between the IE-speakers and the FU-speaking Udmurt and Mari people, i.e. the modern inhabitants of the Pontic steppes (Figure 3b).

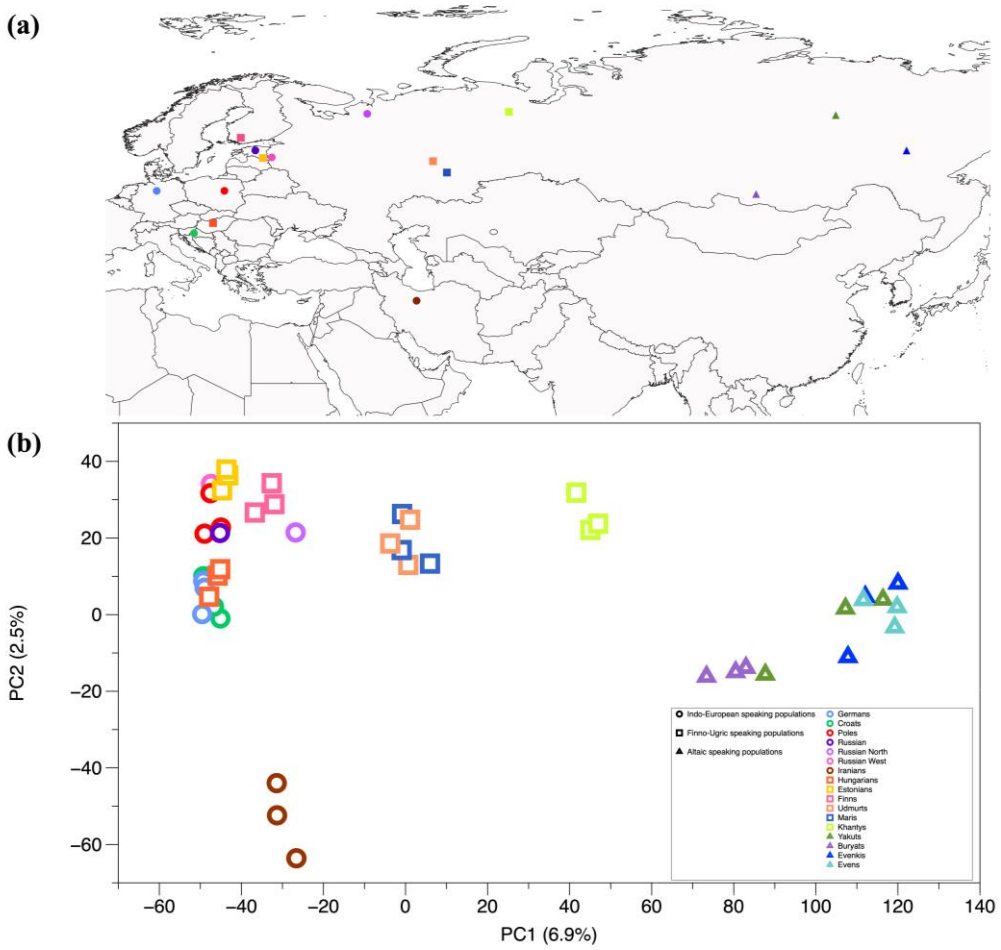


Figure 3. Geographical locations and Principal Component Analysis (PCA) of genomic variation. Populations speaking an IE, FU and AL language are represented by circles, squares and triangles, respectively. **(a)** Geographical locations of the samples in this study. **(b)** Projection on two dimensions of the main components (PCA) of genomic variation in IE, FU and AL speaking populations.

3.2.2. Genetic distances between populations

Next, we calculated genetic distances (*Fst*) between pairs of populations (Figure 4). All AL and IE speaking populations are genetically closer to other populations of their language family than to populations belonging to a different family. Instead, that is not the case for the FU-speakers; all of Estonians, Finns and Hungarians are genetically closer to their respective European neighbors speaking IE. Also, among the Eastern populations, the Mari and Udmurt seem genetically more similar to the Europeans than to the AL-speakers. Exceptions are the easternmost and Trans-Uralic Khanty (Ostyaks), which seem equally close to Mari, Udmurt and most of the AL speakers. This observation can be

reconciled with historical data, which place the origins of the Khanty people in the Russian steppes followed by a northward migration into western Siberia about 500 AD [44].

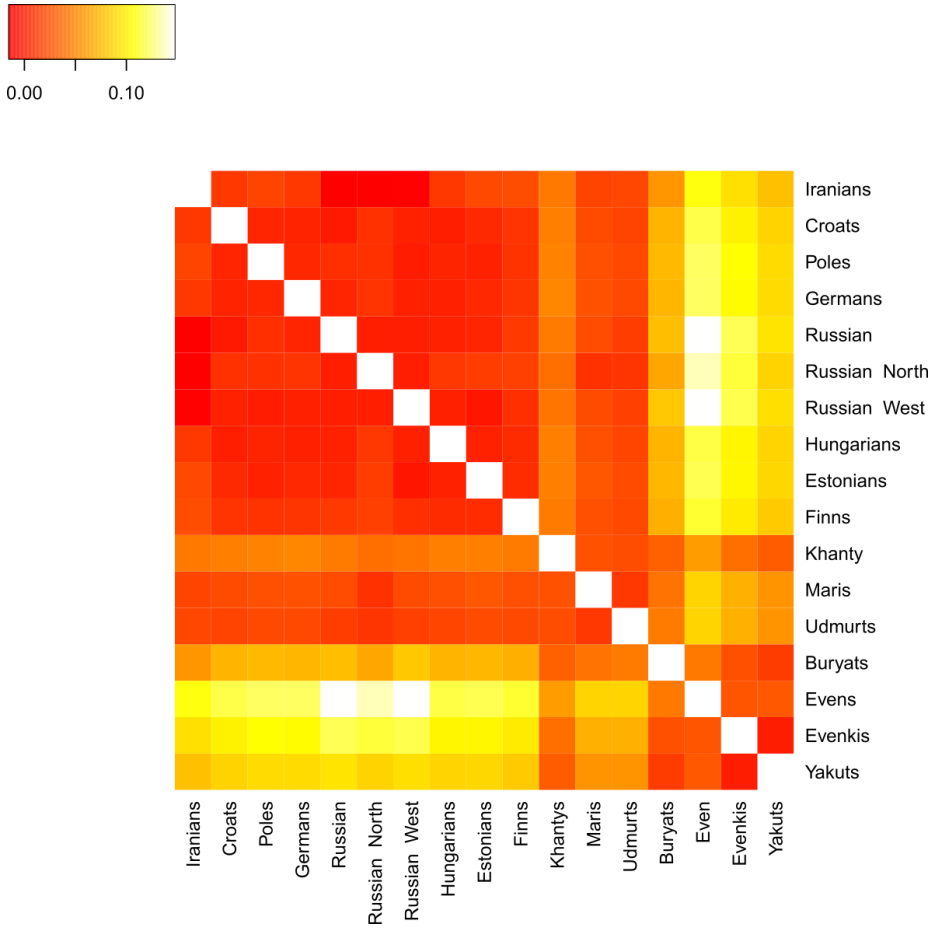


Figure 4. Pairwise genetic distances between Eurasian populations. Darker colors indicate that populations are genetically closer, whereas lighter colors indicate that populations are genetically distant.

3.2.3. Shared haplotypes

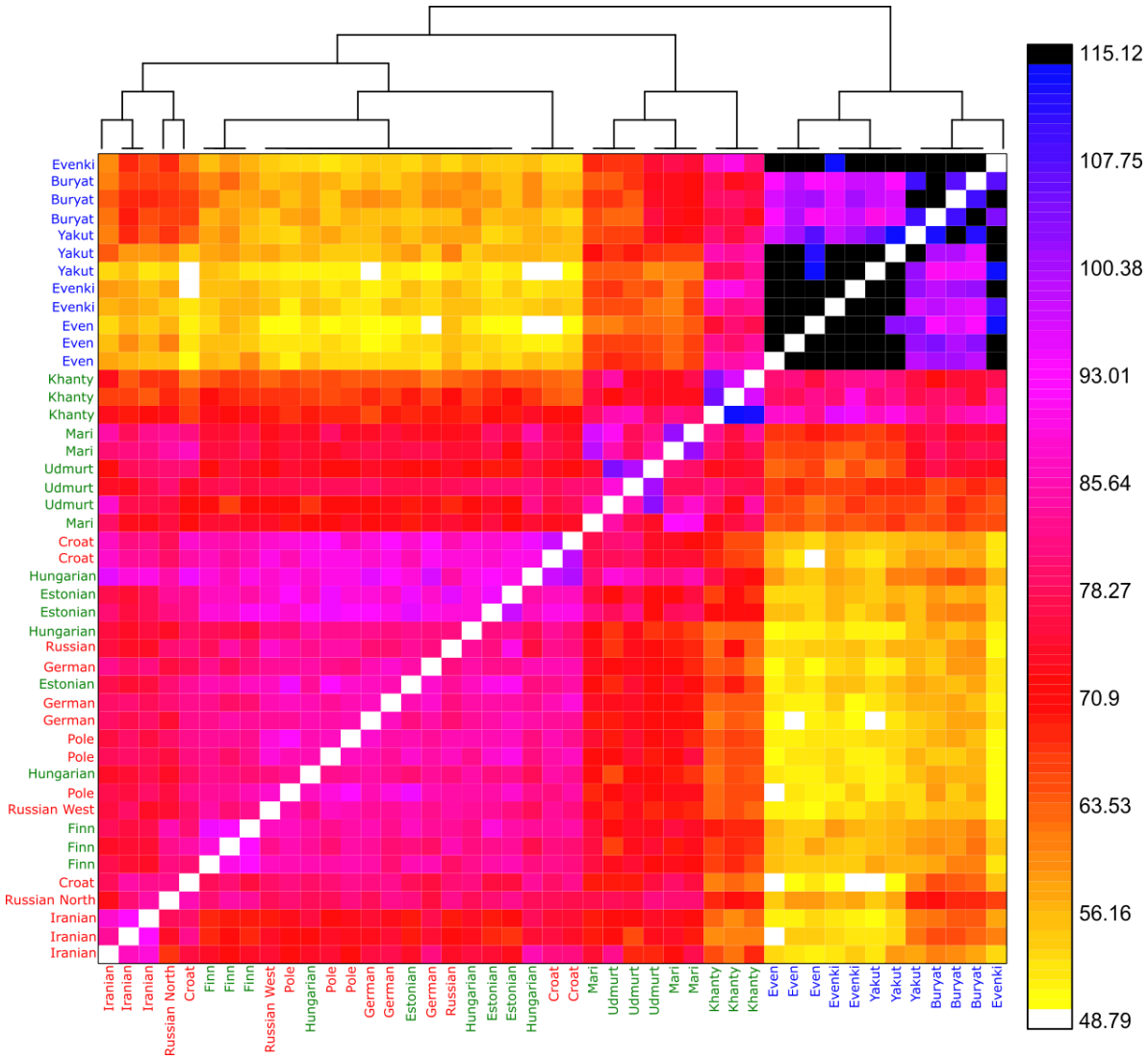
In the analysis of genetic distances, each single-nucleotide polymorphism is independently considered, regardless of its association with other polymorphisms. To analyze the patterns of population resemblance in finer detail, we thus moved to the haplotype level, using ChromoPainter and fineSTRUCTURE (Figure 5). This approach does not depend on prior information on sample groupings and operates instead with data-driven natural groups defined by patterns of haplotype sharing.

This approach also led us to identify three main genetic groups, broadly corresponding to the three main language families. However, as already observed in the *Fst* analysis, there were exceptions. The Western FU-speaking populations (Estonians, Finns and Hungarians) seem to mainly share co-ancestry with the other Europeans, regardless of the language spoken. Conversely, among the Eastern FU-speakers, Udmurt, Mari and Khanty there is a high level of haplotype sharing. Also, this analysis revealed for the first time some co-ancestry of Finns (and partly Estonians and Hungarians) with AL-speakers of Siberian origin.

The evolutionary tree inferred from these data (fineSTRUCTURE cluster analysis; Figure 5b) shows two deep splits, the first isolating all AL speakers, and the second separating Eastern FU speakers from

a group composed by Western FU and IE speakers. All this could even point to different ancestries for the UR-speaking populations, with phenomena of horizontal language diffusion leading them to a shared linguistic identity. But lexical analyses and, in a more modulated fashion, even the syntactic ones support an original FU linguistic unity around the Russian steppes, later fragmented by northward (Khanthy) and westward (Finns and Estonians) migrations and contacts. To better understand these results, we resorted to ancient DNA.

(a)



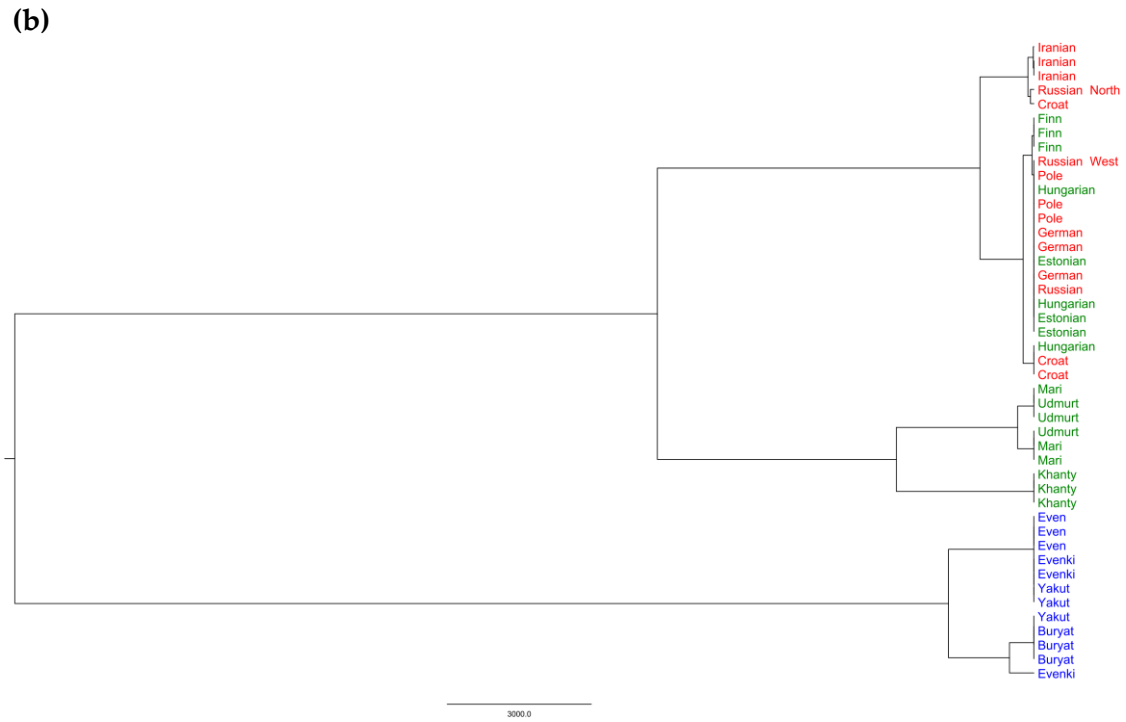


Figure 5. Estimates of shared ancestry between Eurasian individuals. **(a)** Co-ancestry heatmap. Each of the 51 individuals is represented as a row, where each pixel represents the level of co-ancestry (higher for darker colors) shared with each of the other individuals. **(b)** fineSTRUCTURE cluster analysis obtained from the co-ancestry matrix. Red=IE; Green=FU; Blue=AL.

3.2.4. Affinities between modern and ancient populations

Our genetic analysis showed Udmurt and Mari to be the Asian populations more closely related to the Europeans (Figure 4 and Figure 5a). Also, at the linguistic level, they share syntax traits with all three Uralic speakers in Europe (Finns, Estonians and Hungarian, Supplementary Figure 3). We hypothesized this observation may be related to shared ancestry with Yamnaya, an ancient pastoralist population that lived in the current Udmurt and Mari territories, around the Pontic-Caspian steppes, and that expanded into Central and Western Europe in the third millennium BCE, contributing a Caucasian genomic component that nowadays is widespread in Europeans [28,30]. We tested for genetic continuity from the ancient Steppe populations, Yamnaya (~4700 yBP) and the more recent Sintashta (~3900 yBP) on the one hand [28,29], to current Udmurt and Mari on the other. An ancient Anatolian sample [45] was also included in our tests, potentially accounting for the genetic legacy of early farmers from the Near East.

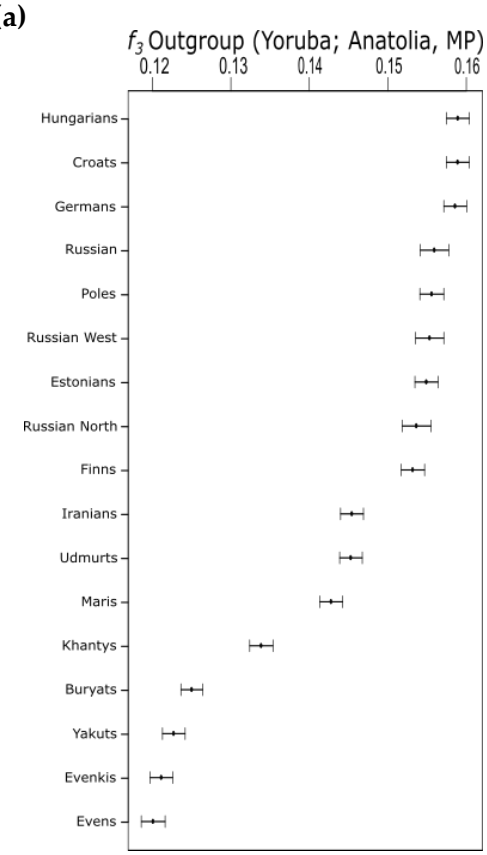
We formulated outgroup *f*3-statistics of the form *f*3(AP, MP; Yoruba), where *AP* was represented in turn by each of the three ancient populations, and MP was each of the modern samples in our dataset (Figure 6 and Supplementary Figure S4). In general, we found all ancient samples to share more genetic drift with modern Europeans and Russians than with non-European populations. Among the Eastern populations, Udmurt and Mari are the ones sharing the most genetic drift with Yamnaya and Sintashta; on the other hand, the Iranians (IE) are the Asian sample closest to the Anatolian farmers, in agreement with recent findings [30]. Also, within the European populations the *f*3 values show opposite trends for the Anatolian and the Yamnaya/Sintashta, the former sharing more genetic drift with southern and central Europeans (Croats and Germans) and the latter being closer to Northeast Europeans, including

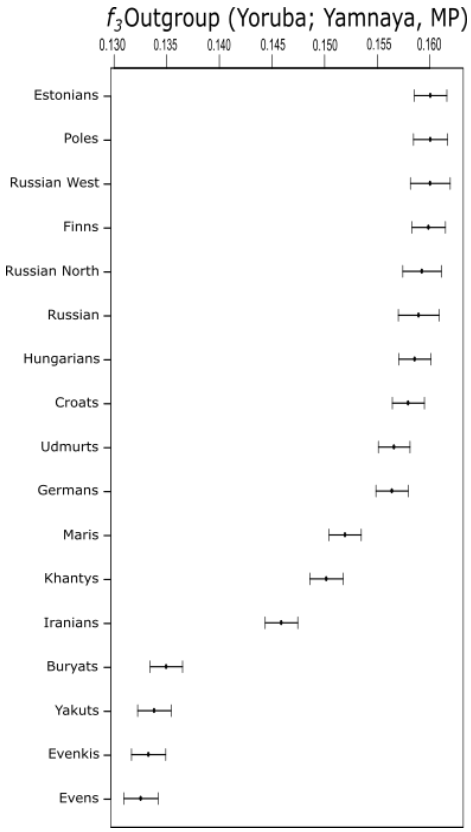
the FU-speaking Estonians and Finns, once again in general agreement with previous findings (e.g., Ref. [28]). It is interesting to notice the peculiar behavior of the Hungarians. They appear much closer to the ancient Anatolians than to the Yamnaya, which is common among southern European populations; however, they are the modern Europeans sharing most genetic drift with the Sintashta. This may be indicative of a relatively more recent genetic contact between them and the Steppe populations, i.e. after the process leading to the spread of the Yamnaya component into Europe.

Contrary to what could be expected, the modern FU inhabitants of the Russian steppes, Mari and Udmurt, appear more distant from Yamnaya than Estonians and Finns. One possible explanation would be the presence, in their genomes, of a Siberian-related component, known to be widespread in contemporary Central and North Asian populations [42,46–48]. We tested for its presence in our samples by modelling Nganasan, a population from the Taymyr Peninsula, as a proxy of the carriers of this Siberian component (as also in Refs. [26] and [31]). We did find support for the presence of such a Siberian component among Mari and Udmurt; the outgroup *f*3 statistics of the form (Nganasan, MP/AP; Yoruba) showed that Udmurt and Mari are indeed closer to Nganasan than Yamnaya, which shared similar *f*3 values with other European population with regards to Nganasan. Figure 6b shows a clear trend; the Nganasans share more genetic drift with all AL speakers, followed by Udmurt and Mari, and then by European populations, no matter if FU- or IE speakers.

To further test whether the peculiar genetic position of the Udmurt and Mari is really associated with the higher presence of a Siberian genetic component in their genome, we ran a *qpAdm* analysis (Figure 7 and Supplementary Table S4). All the FU-speaking populations were successfully modelled as a mixture of Yamnaya, Anatolian and Nganasan-related ancestry, with the exception of the Khanty, who seem to have no Anatolian ancestry. In particular, the Mari and Udmurt genomes appear to contain a large component that can be related with a Siberian genetic ancestry, confirming our expectations. Furthermore, this Siberian ancestry is present, at low though non-negligible percentages, in the Western FU-speaking Finns (but less saliently in Estonians).

(a)





(b)

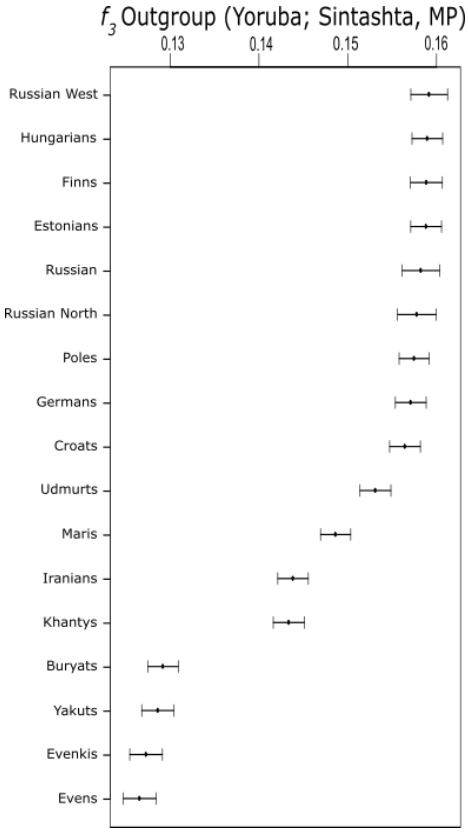


Figure 6. Outgroup *f3*-statistic analysis. Shared genetic drift between ancient and modern (MP) populations. **(a)** Shared genetic drift between Anatolian, Yamnaya, Sintashta, **(b)** Nganasan and modern/ancient populations.

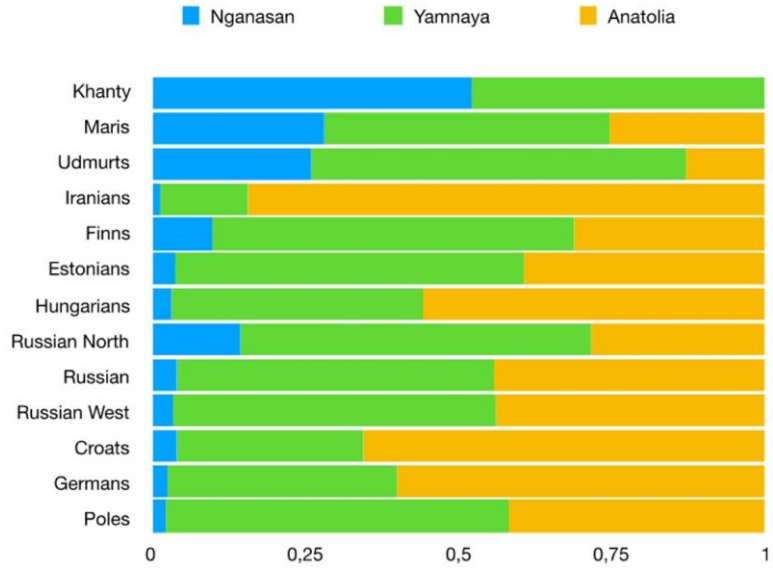


Figure 7. Admixture proportions from three sources estimated using *qpAdm.* Sources used were Nganasan, Yamnaya and Anatolia (percentages and chi-square values are shown in the Supplementary Table S4).

4. Discussion

Syntactic diversity

Syntax distinguishes IE, FU and AL languages quite well, although IE and AL have single outliers (Indo-Iranian and Buryat, respectively). Conversely, the FU family turns out to be less compact, in spite of the greater geographic spanning and population size of IE, and of the weaker lexical evidence purportedly supporting AL (See Ref. 18 for the state of the debate). The whole family appears scattered and in some structural contiguity with their eastern and western neighbors.

A previous study comparing Uralic lexical data, including the FU speakers, had suggested that some secondary contact played a role in the divergence of these languages [49]. The scattered pattern is now more clearly observed and measurable through the cross-family application of our syntactic analysis. The outlying position of Estonian and Finnish among FU languages is evident (see Figures 1 and 2). As for the other Western FU language, Hungarian, qualitative analysis shows that the language shares some parameter values with IE, as opposed to the rest of FU. Yet, such similarities do not emerge in the trees, possibly reflecting the much later arrival in Europe of the Hungarian language [43].

The very distribution of similarities and differences in the syntactic parameters suggests that the pronounced scattering of FU languages is likely to be secondary, i.e. due to cultural contacts. Indeed, there is no evidence of potential convergence of Khanty, Udmurt, Mari with IE. In addition, the main syntactic changes detaching Finnish and especially Estonian on the one hand, and Hungarian on the

other, from the other FU languages are: A. different from each other; B. unidirectional, i.e. of a kind that is often acquired but hardly reversed; C. shared with neighboring IE languages at the time of the contacts with the respective FU languages [22]. This tends to exclude that these properties are ancestral (proto-Uralic) and have been lost by the more eastern varieties because of recent convergence with Asian languages. The similarities of eastern varieties with AL languages are, instead, more ambiguous as to whether they may be shared inheritance or a secondary effect.

In sum, our syntactic phylogenetic analysis supports the original wisdom that FU has been a monophyletic cluster, and is well compatible with the traditional view that the western FU languages have reached Europe from the East at some ancient point; but syntax also detects and measures the pattern of secondary similarities with neighboring languages.

Genome diversity

The genetic analysis shows that the three main groups identified by the linguistic analysis are also biologically differentiated; however, while IE and AL samples form distinct genetic clusters, both in the PCA and ChromoPainter analyses, a peculiar pattern emerges within the FU language family. While the Khantys show affinities with a well-differentiated cluster comprising all AL speakers, the other FU speakers appear to be part of a broad group, including all IE-speaking individuals. In particular, the Western FU-speakers, namely Finns, Estonians and Hungarians, are genetically closer to IE populations in Europe than to the Asian UR-speaking populations. Estonians and Finns also share more ancestry with each other than with the Hungarians. This genetic similarity can reflect: (i) a different source of steppe ancestry in the Hungarians (more closely related with the Sintashta) than in Finns and Estonians (genetically closer to the Yamnaya) (Figure 6a); and/or (ii) a lower contribution of Siberian ancestors to the Hungarian genomes than to the Estonians and especially the Finns (Figure 6b).

Comparison of genetic and linguistic results

Judging whether or not linguistic and genetic data mirror each other may be partly a matter of taste. However, there is little doubt that the syntactic and genomic findings of this study match and corroborate each other. In five out of six cases, linguistic and genetic evidence were consistent (Table 1), the exception being the third one. In this field, however, exceptions are as interesting as the rules, as they call our attention to phenomena that need be further investigated. By looking into the syntactic features of Western FU languages, and into their speakers' genomes, we could recognize peculiar processes affecting the demographic history of people speaking Estonian, Finnish and Hungarian.

Table 1. Synopsis of the main results of this study. Note that ancient Siberian ancestry is (here and elsewhere: Refs. 29, 34) approximated by a modern population, Nganasans.

	Syntax	Modern genomes	Ancient genomes
1	AL languages form a cluster	AL speakers form a cluster	Higher Siberian component in AL than in all the other populations
2	Indo-Iranian languages distinct from European IE languages	Indo-Iranian speakers distinct from other IE speakers	Higher Anatolian component in Indo-Iranian speakers than in other IE speakers
3	FU languages separated from IE and AL	In the tree, FU speakers and IE speakers fall in the same cluster	Yamnaya and Anatolian components similar in Western FU speakers and their European IE-speaking neighbours
4	Estonian closer to IE and more distant than Finnish from other FU languages	Estonians closer to IE speakers than Finns	Siberian component lower in Estonians than in Finns
5	Mari, Khanty and Udmurt closer to AL than to IE languages	Mari, Khanty and Udmurt speakers more distant from IE speakers than Finns, Estonians and Hungarians	Higher Nganasan component in Mari, Khanty and Udmurt speakers than in any other population
6	Easternmost FU Khanty least distant from easternmost Yakut of all AL languages	Khanty speakers halfway between the Mari/Udmurt speakers and eastern AL populations	Khanty speakers have the Siberian and Yamnaya component, but no Anatolian one

Indeed, the Bayesian syntactic tree matches the strong similarity between IE and Balto-Finnic revealed by the genomic tree and PCA, but, on the whole, syntax supports the FU unity to a stronger extent than genetics, and neatly recognizes the Ugric group (Hungarian and Khanty). On the contrary, at the genetic level the FU-speaking populations cluster according to geography, with Hungarian speakers close to Central Europeans, Khanty speakers close to their Eastern AL-speaking neighbors, and the steppe-dwelling Mari and Udmurt speakers in an intermediate position. This result suggests that syntax can also capture secondary demographic events (e.g. population admixture), which genetics can identify only if they have entailed substantial demographic change.

In particular, syntax shows more limited secondary effects on Hungarian from its IE geographic neighbors, and preserves well its historical similarity with Khanty. As we shall discuss later in this paper, historical data suggest that the establishment of Hungarian in Central Europe was the product of an episode of élite dominance, i.e. a deep change of language with limited demographic impact [12]. As a consequence, the genomes of modern speakers of Hungarian were affected only marginally by the phenomena that radically modified their language.

Demographic scenarios reconciling linguistic and genetic evidence

The general picture emerging from our combined analysis of linguistic and genetic data is one in which speakers of the AL, IE and FU languages have long formed three separated groups, evolved independently, and then had contacts leading to various degrees of linguistic and biological exchange (the data cannot exclude that some even much more ancient unity -or close contact- may have involved two of the groups, proto-Uralic and proto-Altaic speakers).

While little is known about the ancient demographic history of AL populations, genomic data are now available from pre-historic peoples of Western Eurasia and the Near East. Analysis of genomes from pre-historic inhabitants of the Russian Steppes [29] identified a Westward migration of people from the Pontic steppes that contributed a Yamnaya ancestral component today widespread in Europe, with the highest prevalence among Estonians and Finns. Although linguistic data were not considered in that study, the authors linked the westward Yamnaya migration with the expansion of the IE languages. Later on, several studies have related the presence of the Estonian and Finnish languages in Europe to a northward migration of people of Siberian ancestry [42]. But the issue is far from settled, as this Siberian influx seems to be too recent to explain the presence of the FU in the Baltic area [48].

Our multidisciplinary analysis seems to point to a different scenario, potentially better reconciling linguistic and genetic evidence. The linguistic similarity of the FU languages Mari and Udmurt with the Balto-Finnic ones, as well as the genetic relationships of these modern populations (Mari and Udmurt in the steppes and Estonians and Finns in north-western Europe) with the ancient Yamnaya (Figure 6a), suggests instead a scenario involving a demographic and linguistic expansion of people with steppe-related ancestry into the Baltic area, consistent with the hypothesis of an expansion of the FU languages from the Volga-river basin [32,43,50].

A possible link between the expansion of the Yamnaya ancestry and the FU languages into northeast Europe also fits nicely with the estimated dates of the FU diversification, between 5000-3000 yBP [43], which coincides in time with the first appearance of the Yamnaya genomic component in ancient European populations [28,29,34]. Particularly in the Baltic region, analysis of ancient DNA have dated the first contacts between Yamnaya migrants and the local communities around the early Bronze Age (5000-4500 yBP), involving Baltic hunter-gatherers with no Neolithic ancestry [51]. This contact pre-dates shortly the first diversification of the Finno-Volgaic branch of FU, ca. 4500-3500 yBP [43], which includes Mari, Finnish and Estonian.

The reconstruction of the lexical history of the FU languages suggests a later northward expansion of the FU languages from the Baltic area to southern Finland around 2000-1600 yBP [43]. This expansion was accompanied by the separation of the northern (Finnish) from the southern (Estonian) group [43]. It is this last diversification that overlaps in time with the first appearance of the Siberian component in ancient remains around the Baltic area [48]. Therefore, ancient DNA, linguistic and archaeological studies agree in suggesting that people related to the Corded-ware culture moved from the coastal Baltic areas into south Finland around 2000-1600 yBP, where they first came into contact with the ancestors of modern Saami (Lapps) [43,47,48,52,53], some ~1000 years after the spread of the extant FU languages to the area [43,48]. Traces of this contact, and of the limited admixture that must have followed, are still detectable in the genomes of Finns [54] (also in Figure 7).

The syntactic similarities/differences we were able to quantify between Estonian and Finnish and IE and the other FU languages (Figure 1) seem in agreement with this migration model, although by

themselves they are not informative about the direction of such a cultural exchange, whether South-to-North or North-to-South.

Finally, our genomic analysis gives evidence of a second event of expansion of the FU-languages from the Russian steppes into Europe without involving a Siberian mediation. That is the case of the FU-speakers from Hungary. There is historical evidence that at the beginning of the medieval era, the language spoken in nowadays Hungary was still Late Latin (at least as an official language), later subject to the effects of Slavic, Germanic and Avar invasions [55]. The main linguistic shift can be approximately dated around 895-905 AD, when people coming from the East conquered Hungary, imposing their own language belonging to the Ugric family [55,56]. Ancient-DNA studies of the invaders have shown that they were genetically close to the Sintashta of the steppes, and apparently unrelated with Siberian ancestors [57], in fine agreement with our genetic analysis. Therefore, the presence of a FU language in Europe is not necessarily correlated with the presence of a Siberian component in the DNA of its speakers.

Speculations on the diffusion of IE into Europe

Linguists and archaeologists have long discussed the timing and modes of spread of IE languages in Europe. Gimbutas [24] associated it with the westward spread of the Kurgan culture, from the Pontic steppes during the Bronze age, whereas Renfrew [26] saw it as a consequence of the Neolithic farmers' demic diffusion from Anatolia (see also Refs. [14] and [58]). These alternatives (hereafter referred to as the Steppe and the Anatolian hypothesis, respectively) are paralleled at the genetic level, by studies supporting population dispersal of Yamnaya-related populations in the Early Bronze Age [28,29], or of Anatolia-related populations during the Neolithic transition [2,3,7,31,59,60].

The genomic similarity between the Yamnaya and the first FU speakers of Europe may be difficult to reconcile with the view that the Yamnaya were also the first who introduced IE languages in Europe, as suggested by studies of genomic, not linguistic, data [28,29]. One possibility, supported by a study of Iberia [61], is that the arrival in Europe of the Steppe genomic component did not necessarily entail the same linguistic changes in all areas. In the absence of adequate data to formally test this hypothesis, we still may speculate that the small, but non-negligible, ancestry component associated with the Anatolian Neolithic [31] among the Yamnaya may reflect previous Northward gene flow from the Near East into the Pontic steppes. If so, it would be possible to reconcile genetic evidence for the Neolithic demic diffusion from the Near East, linguistic evidence on a Near East centre of IE diffusion [14,26,31,58,62], and data suggesting a role of Yamnaya people in spreading both IE [28,29,63], and FU (this study) languages, by imagining the existence of some linguistic diversity within the Yamnaya-like populations and concluding that IE languages have entered Europe in two moments and by two routes. The first one would correspond to the main Neolithic expansion, Northwest into Southern and then Central Europe, but also North, towards the Pontic Steppes. The linguistic impact of this migration would have not been the same for all people in the Pontic steppes; some would retain their original FU languages, some would acquire an IE language. The former would then mostly move towards the Baltic and Finnish area, whereas the latter would correspond to the IE-speaking populations dispersing in Europe in the Bronze Age [29,64], giving rise to the Bell Beaker and Corded Ware cultures.

5. Conclusions

Full inference of complex processes requires the study of broader datasets than available for the present study. Nonetheless, this study exemplifies how appropriate quantitative and qualitative tools allow one to measure cross-family language variation, offering a novel insight into human prehistory and generating testable hypothesis for large-scale genomic analyses. Of course, we must warn about the risk of over-interpreting correlations between languages and genes, especially in the absence of accurate dates of linguistic diversification and expansion, which are not yet well established.

But, on the whole, our analysis, based on linguistic features that can be compared across families and are stable in time, suggests that Darwin's prediction of a general correspondence between biological evolution and language transmission is still generally valid, and that exceptions to this rule are both limited (more than it may appear from simply relying on traditional and non-quantitative language taxonomies, as e.g. in [68]) and extremely useful for a detailed reconstruction of human past.

Supplementary Figure S1. Approximate geographical location of the 34 languages considered.
Supplementary Figure S2. Bayesian phylogeny (BEAST) from the syntactic dataset.
Supplementary Figure S3. Heatmap from the syntactic distances. Dark red represents maximum distance, dark blue minimum distance.
Supplementary Figure S4. Outgroup *f3*-statistics analysis.
Supplementary Table S1. Whole-genome samples collected for the populations under study.
Supplementary Table S2. Ancient DNA samples used in this study.
Supplementary Table S3. Human Origins data on present-day humans used in this study.
Supplementary Table S4. Statistics of the *qpAdm* models.

Author Contributions: Conceptualization, G.L., C.G. and G.B.; methodology, P.S., G.G.F., G.L., C.G. and G.B.; software, P.S. and G.C..; formal analysis, P.S., G.G.F., E.T., A.C. and G.C.; investigation, P.S., G.G.F., E.T., A.C. and G.C.; genetic data curation, P.S.; linguistic data A.C., C.G. and G.L.; writing—original draft preparation, P.S., G.L and G.B..; writing—review and editing, G.G.F, C.G., G.L. and G.B.; visualization, P.S.; supervision, G.L. and G.B.; funding acquisition, G.L., C.G. and G.B.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ERC Adv.Gr. 295733 Darwin's Last Challenge (*LanGeLin*) 2012-2018 (PI Giuseppe Longobardi, Co-I Guido Barbujani) and the MIUR PRIN 2017K3NHHY *Models of language variation and change: new evidence from language contact* (C. Guardiano).

Acknowledgments: We thank Andrea Benazzo for his bioinformatics support. Part of the analyses were carried out while P.S. was Hugo Reyes-Centeno's guest at the DFG Center for Advanced Studies "Words, Bones, Genes, Tools: Tracking Linguistic, Cultural and Biological Trajectories of the Human Past", at the University of Tübingen. We are also indebted to Ándras Bárány, Judit Gervain, Anders Holmberg, István Kenesei, Paul Kiparsky, Katalin Kiss, Márton Sóskuthy for help with Finno-Ugric evidence and analyses, and to Monica-Alexandrina Irimia and Nina Radkevic for assistance in collecting more language data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*; London: John Murray, 1859;

2. Sokal, R.R. Genetic, geographic, and linguistic distances in Europe. *Proc. Natl. Acad. Sci.* **1988**, *85*, 1722–1726, doi:10.1073/pnas.85.5.1722.

3. Barbujani, G.; Pilastro, A. Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 4670–4673, doi:10.1073/pnas.90.10.4670.

4. Longobardi, G. Methods in parametric linguistics and cognitive history. *Linguist. Var. Yearb.* **2003**, *3*, 101–138, doi:10.1075/livy.3.06lon.

5. Creanza, N.; Ruhlen, M.; Pemberton, T.J.; Rosenberg, N.A.; Feldman, M.W.; Ramachandran, S. A comparison of worldwide phonemic and genetic variation in human populations. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 1265–72, doi:10.1073/pnas.1424033112.

6. Renfrew, C. Archaeology, Genetics and Linguistic Diversity. *R. Anthropol. Inst. Gt. Britain Irel.* **1992**, *27*, 445–478.

7. Cavalli-Sforza, L.L.; Piazza, A.; Menozzi, P.; Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 6002–6, doi:10.1073/pnas.85.16.6002.

8. Poloni, E.S.; Passarino, G.; Santachiara-Benerecetti, A.S.; Langaney, A.; Excoffier, L.; Poloni, E. *Human Genetic Affinities for Y-Chromosome P49a,f/Taql Haplotypes Show Strong Correspondence with Linguistics*; 1997; Vol. 61;.

9. Belle, E.M.S.; Barbujani, G. Worldwide analysis of multiple microsatellites: Language diversity has a detectable influence on DNA diversity. *Am. J. Phys. Anthropol.* **2007**, *133*, 1137–1146, doi:10.1002/ajpa.20622.

10. Gray, R.D.; Drummond, A.J.; Greenhill, S.J. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science (80-.).* **2009**, *323*, 479–483, doi:10.1126/science.1166858.

11. Henn, B.M.; Botigué, L.R.; Gravel, S.; Wang, W.; Brisbin, A.; Byrnes, J.K.; Fadhlaoui-Zid, K.; Zalloua, P.A.; Moreno-Estrada, A.; Bertranpetit, J.; et al. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* **2012**, *8*, doi:10.1371/journal.pgen.1002397.

12. Longobardi, G.; Ghirotto, S.; Guardiano, C.; Tassi, F.; Benazzo, A.; Ceolin, A.; Barbujani, G. Across language families: Genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.* **2015**, *157*, 630–640, doi:10.1002/ajpa.22758.

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

14. Gray, R.D.; Atkinson, Q.D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **2003**, *426*, 435–439, doi:10.1038/nature02029.

15. Jäger, G. Support for linguistic macrofamilies from weighted sequence alignment. *Proc. Natl. Acad. Sci.* **2015**, *112*, 12752 LP – 12757, doi:10.1073/pnas.1500331112.

16. Longobardi, G.; Guardiano, C. Evidence for syntax as a signal of historical relatedness. *Lingua* **2009**, *119*, 1679–1706, doi:10.1016/j.lingua.2008.09.012.

17. Longobardi, G.; Guardiano, C. Toward a syntactic phylogeny of modern Indo-European languages. *J. Hist. Linguist.* **2013**, *3*, 122–152, doi:10.1075/jhl.3.1.07lon.

18. Ceolin, A. Significance testing of the Altaic family. *Diachronica* **2019**, *36*, 299–336, doi:10.1075/dia.17007.ceo.

19. Longobardi, G.; Guardiano, C. Phylogenetic reconstruction in syntax. In *the Parametric Comparison Method*; Ledgeway, A., Roberts, I., Eds.; Cambridge University Press, 2017; pp. 241-272 BT-The Cambridge Handbook of Historical ISBN 9781107049604.

20. Nichols, J. *Linguistic diversity in space and time*; Chicago: The University of Chicago Press, 1992;

21. Guardiano, C.; Longobardi, G. *Parametric comparison and language taxonomy. In: Battlori M, Picallo C, Roca F, editors. Grammaticalization and parametric variation*; Oxford: Oxford University Press, 2005;

22. Ceolin, A.; Guardiano, C.; Irimia, M.-A.; Longobardi, G. Formal syntax and deep history. *Front. Psychol.* **2020**, *11*, 2384.

23. Gyarmathi, S. *Affinitas linguae hungaricae cum linguis fennicae originis grammatice demonstrata*; Les éditions chapitre.com, 1799;

24. Gimbutas, M. The Three Waves of Kurgan People into Old Europe, 4500–2500 BC. *Arch. suisses d'anthropologie générale* **1979**, *43*, 113–137.

25. D., A. *The horse, the wheel and language. How Bronze-Age riders from the Eurasian steppes shaped the modern world*; 2007;

26. Renfrew, C. *Archaeology and Language: The Puzzle of Indo-European Origins*; 1987;

27. Heggarty, P. *Indo-European and the Ancient DNA Revolution*; Max Planck Institute for Evolutionary Anthropology, Leipzig, December 2-3, 2013 (JIES Monograph Series No 65), 2018;

28. Haak, W.; Lazaridis, I.; Patterson, N.; Rohland, N.; Mallick, S.; Llamas, B.; Brandt, G.; Nordenfelt, S.; Harney, E.; Stewardson, K.; et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **2015**, *522*, 207–211, doi:10.1038/nature14317.

29. Allentoft, M.E.; Sikora, M.; Sjögren, K.-G.; Rasmussen, S.; Rasmussen, M.; Stenderup, J.; Damgaard, P.B.; Schroeder, H.; Ahlström, T.; Vinner, L.; et al. Population genomics of Bronze Age Eurasia. *Nature* **2015**, *522*, 167–172, doi:10.1038/nature14507.

30. Narasimhan, V.M.; Patterson, N.; Moorjani, P.; Rohland, N.; Bernardos, R.; Mallick, S.; Lazaridis, I.; Nakatsuka, N.; Olalde, I.; Lipson, M.; et al. The formation of human populations in South and Central Asia. *Science (80-.).* **2019**, *365*, eaat7487, doi:10.1126/science.aat7487.

31. de Barros Damgaard, P.; Martiniano, R.; Kamm, J.; Moreno-Mayar, J.V.; Kroonen, G.; Peyrot, M.; Barjamovic, G.; Rasmussen, S.; Zacho, C.; Baimukhanov, N.; et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **2018**, *360*, doi:10.1126/science.aar7711.

32. Janhunen, J. Proto-Uralic–what, where, and when? *Mémoires la Société Finno-Ougrienne* **2009**, *258*, 57–78.

33. Pagani, L.; Lawson, J.; Jagoda, E.; Mörseburg, A.; Clemente, F.; Hudjashov, G.; DeGiorgio, M.; Eriksson, A.; Saag, L.; Wall, J.; et al. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **2016**, doi:10.1038/nature19792.

34. Mathieson, I.; Lazaridis, I.; Rohland, N.; Mallick, S.; Patterson, N.; Roodenberg, S.A.; Harney, E.; Stewardson, K.; Fernandes, D.; Novak, M.; et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **2015**, *528*, 499–503, doi:10.1038/nature16152.

35. Delaneau, O.; Ongen, H.; Brown, A.A.; Fort, A.; Panousis, N.I.; Dermitzakis, E.T. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **2017**, *8*, 1–7, doi:10.1038/ncomms15452.

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

37. Hammer, O.; Harper, D.; Ryan, P. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontol. Electron.* **2001**, *4*, 1–9.

38. Felsenstein, J.; Felenstein, J. *Inferring phylogenies*; Sunderland, MA: Sinauer associates, 2004;

39. Maddison, W.P.; Maddison, D.R. Mesquite: a modular system for evolutionary analysis. **2004**, *version 1.*

40. Bouckaert, R.; Vaughan, T.G.; Barido-Sottani, J.; Duchêne, S.; Fourment, M.; Gavryushkina, A.; Heled, J.; Jones, G.; Kühnert, D.; De Maio, N.; et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **2019**, *15*, e1006650.

41. Lawson, D.J.; Hellenthal, G.; Myers, S.; Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **2012**, *8*, 11–17, doi:10.1371/journal.pgen.1002453.

42. Tambets, K.; Yunusbayev, B.; Hudjashov, G.; Ilumäe, A.M.; Rootsi, S.; Honkola, T.; Vesakoski, O.; Atkinson, Q.; Skoglund, P.; Kushniarevich, A.; et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol.* **2018**, *19*, doi:10.1186/s13059-018-1522-1.

43. Honkola, T.; Vesakoski, O.; Korhonen, K.; Lehtinen, J.; Syrjänen, K.; Wahlberg, N. Cultural and climatic changes shape the evolutionary history of the Uralic languages. *J. Evol. Biol.* **2013**, *26*, 1244–1253, doi:10.1111/jeb.12107.

44. Pimenoff, V.N.; Comas, D.; Palo, J.U.; Vershubsky, G.; Kozlov, A.; Sajantila, A. Northwest Siberian Khanty and Mansi in the junction of West and East Eurasian gene pools as revealed by uniparental markers. *Eur. J. Hum. Genet.* **2008**, *16*, 1254–1264, doi:10.1038/ejhg.2008.101.

45. Lazaridis, I.; Nadel, D.; Rollefson, G.; Merrett, D.C.; Rohland, N.; Mallick, S.; Fernandes, D.; Novak, M.; Gamarra, B.; Sirak, K.; et al. Genomic insights into the origin of farming in the ancient Near East. *Nature* **2016**, *536*, 419–424, doi:10.1038/nature19310.

46. Jeong, C.; Balanovsky, O.; Lukianova, E.; Kahbatkyzy, N.; Flegontov, P.; Zaporozhchenko, V.; Immel, A.; Wang, C.C.; Ixan, O.; Khussainova, E.; et al. The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* **2019**, *3*, 966–976, doi:10.1038/s41559-019-0878-2.

47. Saag, L.; Laneman, M.; Varul, L.; Malve, M.; Valk, H.; Razzak, M.A.; Shirobokov, I.G.; Khartanovich, V.I.; Mikhaylova, E.R.; Kushniarevich, A.; et al. The Arrival of Siberian Ancestry Connecting the Eastern Baltic to Uralic Speakers further East. *Curr. Biol.* **2019**, *29*, 1701-1711.e16, doi:10.1016/j.cub.2019.04.026.

48. Lamnidis, T.C.; Majander, K.; Jeong, C.; Salmela, E.; Wessman, A.; Moiseyev, V.; Khartanovich, V.; Balanovsky, O.; Ongyerth, M.; Weihmann, A.; et al. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat. Commun.* **2018**, *9*, doi:10.1038/s41467-018-07483-5.

49. Lehtinen, J.; Honkola, T.; Korhonen, K.; Syrjänen, K.; Wahlberg, N.; Vesakoski, O. Behind Family Trees: Secondary Connections in Uralic Language Networks. *Lang. Dyn. Chang.* 2014, *4*, 189–221.

50. Kallio, P. Suomen kantakielen absoluuttista kronologiaa. *Virittäjä* **2006**, *110*, 2–25.

51. Jones, E.R.; Zarina, G.; Moiseyev, V.; Lightfoot, E.; Nigst, P.R.; Manica, A.; Pinhasi, R.; Bradley, D.G. The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. *Curr. Biol.* **2017**, *27*, 576–582, doi:10.1016/J.CUB.2016.12.060.

52. Miettinen, T. *Suomenlahden ulkosaarten esihistoriaa. In: Suomenlahden ulkosaaret: Lavansaari, Seiskari, Suursaari, Tytärsaari (R. Hamari, M. Korhonen, Miettinen Timo & I. Talve, eds)*; Suomalaisen Kirjallisuuden Seura, Helsinki, 1996;

53. Kivikoski, E. Suomen Historia I: Suomen Esihistoria. *Werner-Söderström Oy, Porvoo* **1961**.

54. Palo, J.U.; Ulmanen, I.; Lukka, M.; Ellonen, P.; Sajantila, A. Genetic markers and population history: Finland revisited. *Eur. J. Hum. Genet.* **2009**, *17*, 1336—1346, doi:10.1038/ejhg.2009.53.

55. Csányi, B.; Bogácsi-Szabó, E.; Tömöry, G.; Czibula, Á.; Priskin, K.; Csõsz, A.; Mende, B.; Langó, P.; Csete, K.; Zsolnai, A.; et al. Y-Chromosome Analysis of Ancient Hungarian and Two Modern Hungarian-Speaking Populations from the Carpathian Basin. *Ann. Hum. Genet.* **2008**, *72*, 519–534, doi:10.1111/j.1469-1809.2008.00440.x.

56. Cavalli-Sforza, L.L. *Geni, popoli e lingue*; 1997;

57. Neparáczki, E.; Kocsy, K.; Tóth, G.E.; Maróti, Z.; Kalmár, T.; Bihari, P.; Nagy, I.; Pálfi, G.; Molnár, E.; Raskó, I.; et al. Revising mtDNA haplotypes of the ancient Hungarian conquerors with next generation sequencing. *PLoS One* **2017**, *12*, doi:10.1371/journal.pone.0174886.

58. Bouckaert, R.; Lemey, P.; Dunn, M.; Greenhill, S.J.; Alekseyenko, A. V.; Drummond, A.J.; Gray, R.D.; Suchard, M.A.; Atkinson, Q.D. Mapping the origins and expansion of the Indo-European language family. *Science (80-.).* **2012**, *337*, 957–960, doi:10.1126/science.1219669.

59. Menozzi, P.; Piazza, A.; Cavalli-Sforza, L. Synthetic Maps of Human Gene Frequencies in Europeans. **1978**, *201*, 786–792.

60. Sokal, R.R.; Oden, N.L.; Legendre, P.; Fortin, M.-J.; Kim, J.; Thomson, B.A.; Vaudor, A.; Harding, R.M.; Barbujani, G. Genetics and Language in European Populations 1990, 157–175.

61. Olalde, I.; Mallick, S.; Patterson, N.; Rohland, N.; Villalba-Mouco, V.; Silva, M.; Dulias, K.; Edwards, C.J.; Gandini, F.; Pala, M.; et al. The genomic history of the Iberian Peninsula over the past 8000 years. *Science (80-.).* **2019**, *363*, 1230–1234, doi:10.1126/science.aav4040.

62. Chikhi, L.; Nichols, R.A.; Barbujani, G.; Beaumont, M.A. Y genetic data support the Neolithic demic diffusion model. *PNAS* **2002**, *99*, doi:10.3389/fpsyg.2018.00483.

63. Ning, C.; Wang, C.-C.; Gao, S.; Yang, Y.; Zhang, X.; Wu, X.; Zhang, F.; Nie, Z.; Tang, Y.; Robbeets, M.; et al. Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan. *Curr. Biol.* **2019**, *29*, 2526-2532.e4, doi:10.1016/j.cub.2019.06.044.

64. Tassi, F.; Vai, S.; Ghirotto, S.; Lari, M.; Modi, A.; Pilli, E.; Brunelli, A.; Susca, R.R.; Budnik, A.; Labuda, D.; et al. Genome diversity in the Neolithic Globular Amphorae culture and the spread of Indo-European languages. *Proc. R. Soc. B Biol. Sci.* **2017**, *284*, 20171540, doi:10.1098/rspb.2017.1540.