Oleg Balanovsky
Vavilov Institute for General Genetics (Moscow)

Olga Utevska
V. N. Karazin National University (Kharkov)

Elena Balanovska
Research Centre for Medical Genetics (Moscow)

Genetics of Indo-European populations: the past, the future*

We describe our experience of comparing genetic and linguistic data in relation to the Indo-European problem. Our recent comparison of the genetic variation with lexicostatistical data on North Caucasian populations identified the parallel evolution of genes and languages; one can say that history of the populations was reflected in the linguistic and the genetic mirrors. For other linguistic families one can also expect this similarity, though it could be blurred by elite dominance and other events affecting gene and lexical pools differently. Indeed, for Indo-European populations of Europe, in contrast with the Caucasus case, the partial correlation indicates a more important role of geography ($r = 0.32$) rather than language ($r = 0.21$) in structuring the gene pool; though high pair correlation ($r = 0.67$) between genetics and linguistics distances allows using the lexicostatistical data as good predictors of genetic similarity between populations. The similarity between genetics and linguistics was identified for both Y-chromosomal data (populations are clustered according to their language) and mitochondrial DNA (populations are clustered according to their language group). In general, we believe that there is no single genetic marker definitively linked with the expansion of Indo-European populations. Instead, we are starting a new research project aiming to identify a set of markers partially linked with separate Indo-European groups, thus allowing partial reconstructions of the multi-layer mosaic of Indo-European movements.

Keywords: gene pool, Indo-European populations, Y-chromosome, correlation between genetic and linguistic variation.

The Indo-European problem from the genetic point of view

Genetics and linguistics are very different branches of science and humanities; the only overlap between them is *population*. Any language exists within a population which speaks it across generations; any gene pool exists within a population which reproduces it across generations. However, the nature of language and the nature of the gene are so different that no direct comparison of language structure and genetic structure makes sense. The only possibility for comparisons lies in the population *history*, because both linguistics and genetics try to reconstruct this history from language structure or genetic structure.

One should not expect that population histories told by genetics and linguistics coincide; one might expect that both genetics and linguistics tell something truthful about population history. Keeping this in mind, we briefly describe our experience of comparing genetic and linguistic data, particularly in the case of Indo-European *populations*.

* This work has been supported by The Presidium of RAS programs: “Molecular and cell biology”, “Fundamental sciences for the medicine”, “Gene pool dynamics”, and RFBR grants 13-04-01711, 12-04-31732.

Discrepancy between genetic and linguistic data

Hungarians are a good example of the crucial discrepancy between linguistic and genetic data. Linguistically they belong to the Ugric group, which allows to hypothesize migration from the Trans-Ural region (where other Ugric-speaking populations could be found) to the Middle-Danube (or Great Hungarian) plain [Szıj, 2005]. Historical records about the Magyar invasion which gave rise to the Hungarian state strongly support this hypothesis. Genetic (and other anthropological) data show, however, that the Hungarian gene pool has almost nothing in common with the Trans-Uralic gene pool, but is very similar to the gene pools of its closest neighbor populations of the Balkans and East Europe [Czeizel et al., 1991; Semino et al., 2000; Bogacsi-Szabo et al., 2005; Tomory et al., 2007; Csanyi et al., 2008]. So, linguistics and history tell that the migration took place, while genetics tells that it did not. Shall one conclude that either genetics or linguistics tells a completely wrong story? In this case — certainly not. It is generally accepted that Magyars (Hungarians) were strong enough to mix with the previous population of this region and make them speak the Hungarian language, but they were not *numerous* enough to make a recordable contribution to their gene pool. This is the well known pattern of *elite dominance model*: invasion with language change, but without a change in the gene pool. Thus, genetics provided some information about the relative numbers of migrants, which was missing in the data of other sciences. Combining linguistic, historical and genetic data, one can reconstruct the population history in more details than when lacking even one of these sources. To conclude, the discrepancy between linguistics and genetics might yield useful information on *population history*.

Coevolution of genes and languages

Without a doubt, coinciding results of linguistic and genetic studies could tell even more about population history; finding such examples is always pleasant for researchers. Our study on North Caucasian populations [Balanovsky et al., 2011] provided the best fit published to date. We studied the Y-chromosomal variation among 10 ethnic groups (*populations*) speaking North Caucasian languages and compared this genetic variation with lexicostatistical data on these languages. The 11th population studied was Iranic-speaking Ossets. The linguistic part of this study was performed by A. V. Dybo and O. A. Mudrak, while the genetic team included a number of researchers, with major contributions from O. P. Balanovsky and Kh. D. Dibirova.

To clarify the genetic terminology used here, the *haplotype* defines concrete Y-chromosomal lineage and *haplogroup* signifies a large group of haplotypes that have a common origin. Haplogroups are therefore like branches on the family tree of humankind, while haplotypes are leaves. One haplotype originates from another due to mutation. Our study was performed on both levels: the level of haplogroups and the level of haplotypes.

At the level of *haplogroups*, four independent methods were used for comparing genetic and linguistic data.

First, the dendrogram that shows the interrelations of gene pools was compared with the dendrogram that represents language splits. Both dendrograms virtually coincided.

Second, genetic boundaries were identified, subdividing the meta-population of the Caucasus into regional gene pools. These genetic boundaries coincided with linguistic boundaries (between Dagestan and Nakh speakers; between Nakh and Iranian speakers; between Iranian and Abkhaz-Adyghe speakers).

Table 1. The correlation between genetic, linguistic and geographic distances between populations (data on the Y-chromosome).

Type of the correlation	Correlated parameters	North Caucasian populations	Indo-European populations (from Europe)
Pair correlation	Genetics and linguistics	0.64	0.67
Pair correlation	Genetics and geography	0.60	0.70
Partial correlation	Genetics and linguistics (geography held constant)	0.34	0.21
Partial correlation	Genetics and geography (language held constant)	0.21	0.32

These two methods revealed an excellent correlation between genetics and linguistics. But correlation does not necessarily mean a causal link: it may also mean that both systems depend on a third one. This third underlying factor could be the geography. To explore this possibility, genetic distances, linguistic distances and geographic distances between the same set of Caucasian populations were computed, and correlation between these distances was calculated [Balanovsky et al., 2011]. Table 1 shows that the correlation between genetics and geography ($r = 0.60$) was almost as high as the correlation between genetics and linguistics ($r = 0.64$). When partial correlations were computed (a statistical method to exclude influence of the third factor), the correlation between genetics and linguistics became noticeably higher ($r = 0.34$) than the correlation between genetics and geography ($r = 0.21$). This indicates that the linguistic structure itself correlates with the genetic structure, rather than that both simply depend on the geographic structure.

The fourth method to be applied was an estimation of the genetic variation between North Caucasian populations, grouped in two different ways. The linguistic grouping meant subdividing the populations into Dagestan, Nakh, Iranian, and Abkhaz-Adyghe groups. The geographic grouping meant subdividing the same populations into West Caucasian, Central Caucasian and East Caucasian groups. The genetic variation between linguistic groups (0.27) was twice as high as the genetic variation between geographic groups (0.15). One should conclude that linguistic relationship is a more important factor than geographic vicinity for structuring the gene pool of North Caucasian populations.

At the level of *haplotypes* we found many haplotype clusters present in one population but absent or rare in all the other populations. These population-specific clusters were dated using the molecular clock approach. These dates estimate the time when the given population split from the related populations. The crucial point is that glottochronology also provides dates of language splits, which are the same as the splits of populations of speakers. Therefore, we have this unique possibility to compare genetic dates of population events and the linguistic dates of the same events. These dates mostly coincided, as described in details in [Balanovsky et al., 2011]. Thus, we identified a parallel evolution of genes and languages in the Caucasus.

To explain this coevolution we suggested the following model. The Caucasian populations originated from a common root (proto-population) that split into daughter populations which went on to occupy different parts of the Caucasus; there they later split into "granddaughter" populations, and so on. These *population* events also caused the split of languages, so that the tree of population splits became the tree of the North Caucasian linguistic family. These population events also allowed each population to accumulate its own specific clusters of haplotypes. In other words, the model implies that population history was reflected in two

mirrors — the linguistic and the genetic one. Because (i) this population history in a mountainous region was not too strongly blurred by migrations and (ii) both "mirrors" were based on an extensive dataset and analyzed by adequate methods, both reflections coincided.

The important conclusion is that in other regions of the world and other linguistic families one can also expect a similarity between genetic and linguistic data. However, even providing this similarity exists in nature, to see it in research data three important conditions should be met: (i) genetic analysis is done properly; (ii) linguistic analysis is done properly; (iii) the population history did not include elite dominance or other events that are visible in one "mirror" but not visible in another.

This is why we have included here this brief description of the study of North Caucasian populations here: for the Indo-European case, it provides the basic model and comparison point. However, one could hardly expect that Indo-European populations followed this model with the same precision as those in the North Caucasus.

Correlation of genes and languages

We do not expect that the history of Indo-Europeans followed the same clear model as that of the North Caucasians. It is therefore even more interesting to apply the same methodology to the IE case. So far, we have performed only one, but the most important kind of analysis — the correlation analysis of genetic, linguistic and geographic distances between the Indo-European populations of Europe. (We did not include Indo-Iranian populations because the Indian gene pool is much too different from the European one). This kind of analysis had already been performed earlier, in 2000 [Rosser et al., 2000], where it was found that both correlations are about $r = 0.3$. Twelve years later we repeated this analysis using a dataset that was ten times as large (Table 1). We found correlations that were twice as high (0.67 between genetics and linguistics and 0.70 between genetics and geography). In contrast with the case of the Caucasus, the partial correlation indicates a more important role of geography (genetics and geography $r = 0.32$, while genetics and linguistics only $r = 0.21$). However, the high pair correlation with linguistics ($r = 0.67$) allows to use the statistical data as good predictors of genetic similarity between populations.

Of course, the single correlation coefficient does not tell much about the Indo-European homeland or their migrations. To study them from the genetic point of view, we should take a look at the overall gene pool structure.

Indo-European gene pool: the obvious geographic patterns, the hidden language parallels

Genetic studies of the Eurasian populations resulted in a general agreement on the main patterns of the gene pool. It became clear that populations at the extremes of the Indo-European area have little in common genetically (like Western Europe *vs* India, or Scandinavia *vs* Armenia). Moreover, in many cases IE-speaking populations are genetically similar to their geographic non-IE neighbors; for example, French and Spaniards are genetically similar to Basque [Martínez-Cruz et al., 2012; Behar et al., 2012], Russians to Finnish speakers [Balanovsky et al., 2008], and Indian IE speakers to Dravidian populations [Kivisild et al., 2003]. Therefore, one might suppose that the elite dominance model had worked many times throughout the history of Indo-European populations.

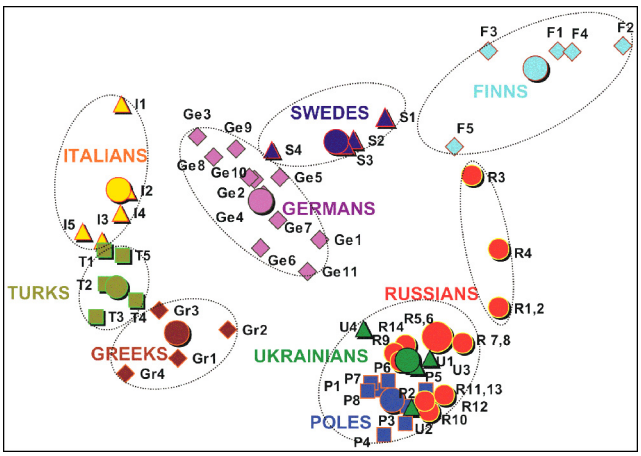

Figure 1 (adapted from Balanovsky et al., 2008). Genetic relationships between European populations (data on Y-chromosome). Populations of different ethnic groups are marked by signs of different color and shape. It can be seen that populations cluster together according to their language.

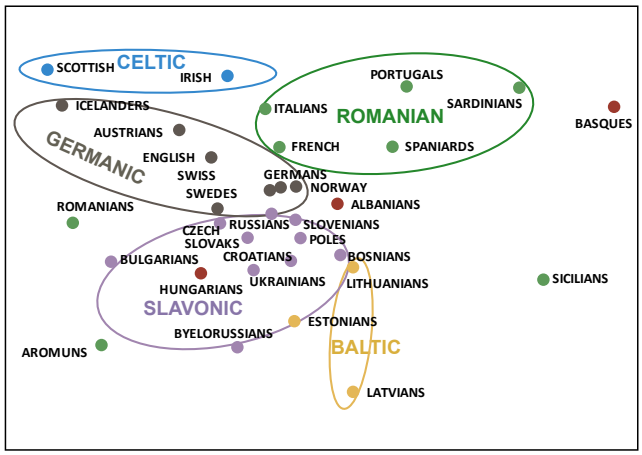

Figure 2. Genetic relationships between European populations (data on mitochondrial DNA). It can be seen that populations tend to cluster together according to their language group.

On the other side, on the smaller geographic (and linguistic) scale a similarity between genetics and linguistics begins to appear. For example, Indian populations are genetically closer to the populations of Iran than to any other. Another example is that West and East Slavic populations form a genetic continuum across their large area (Figure 1), which coincides with their linguistic similarity while contradicting the large geographic distance between then. All these examples were based on the Y-chromosome which generally provides the clearest pattern. However, data on mitochondrial DNA led to similar conclusions. For example, European ethnic groups form *genetic clouds* according to their linguistic grouping (Figure 2), though the pattern is less clear compared with the Y-chromosomal results. All of these draw a very complicated picture of genetic-linguistic interrelations among IE populations and demand more detailed studies; one of the possible future studies is described below.

Genetic data on Neolithization of Europe

The most elaborated theory, explaining the spread of Indo-European languages across Europe and the formation of the European gene pool, links both events with Neolithization. The main pattern in the European gene pool is gradual change from the southeast (Anatolia, then the Balkans) to Northwest Europe [Cavalli-Sforza et al., 1994]. This pan-European genetic trend was first shown on classical genetic markers and was confirmed multiple times by other genetic systems. This geographic trend demonstrates a wonderful correlation with the archeological map of the gradual distribution of farming across Europe. This allowed to develop the "demic diffusion model", which states that farming populations (growing in numbers much faster than hunter-gathering groups) spread from Anatolia, and that each generation of farmers migrated further until they reached the geographic limits of the European subcontinent. Each generation mixed with autochthonous hunter-gatherers, and the initial Near Eastern gene pool gradually dissolved. This (geographically gradual) dissolution resulted in the gene pool gradient that was found in European populations [Ammerman & Cavalli-Sforza, 1984; Cavalli-Sforza et al., 1994]. Many researchers believe that these farmers were Indo-Europeans,

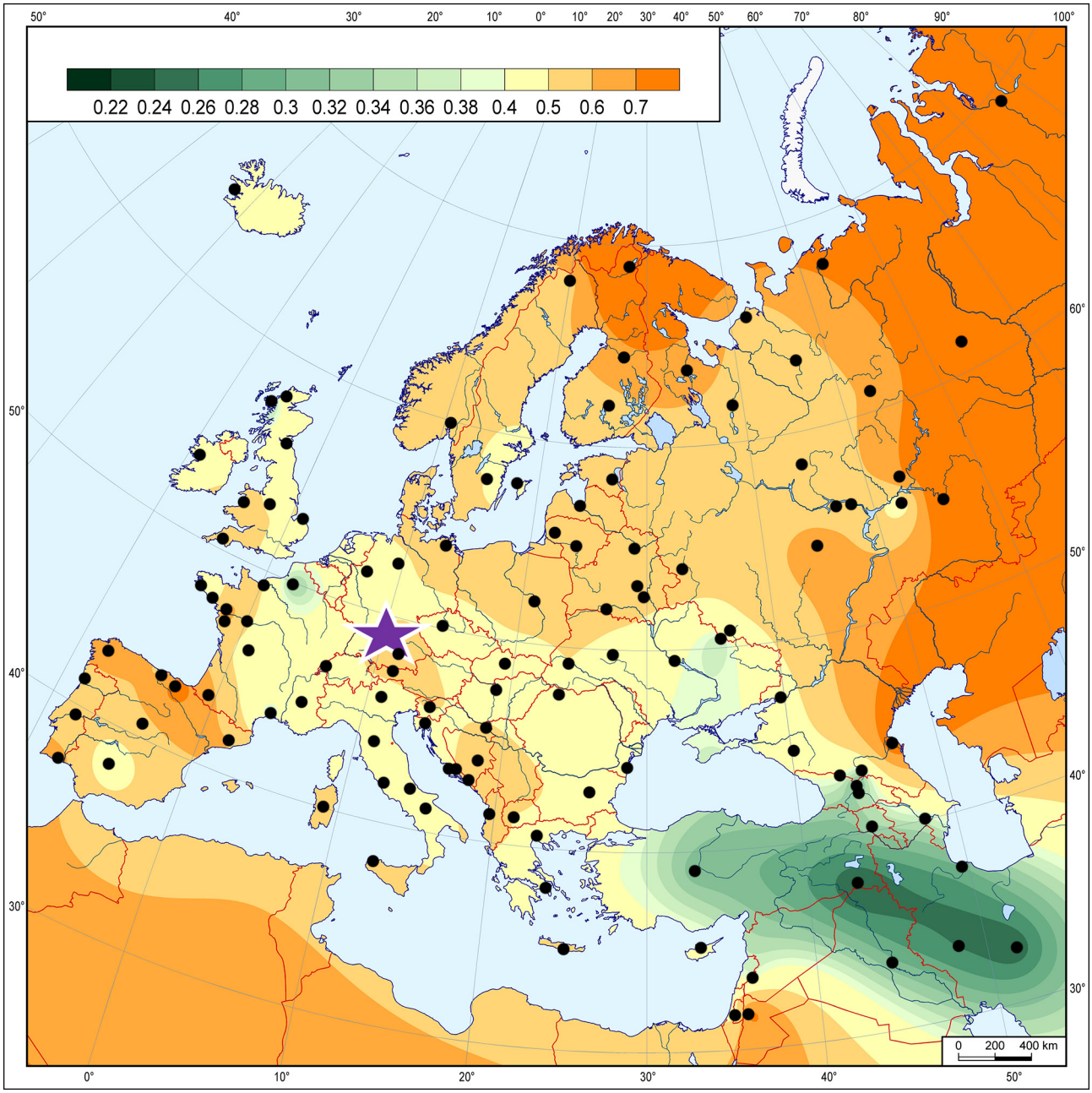

Figure 3 (adapted from Haak et al., 2010). Map of genetic distances between the first Neolithic population of Europe and present day gene pools (data on mitochondrial DNA). A genetic distance map plots genetic distances from a single selected population (reference population) to all populations of the mapped area. It is the researcher's choice to select the reference population. This map plots genetic distances from the first widespread Neolithic culture in Europe (Linear Band Ceramic) to the present day populations of Europe and Near East. The location of the studied ancient population is shown by the asterisk.

and that, therefore, the spread of farming, the spread of Indo-European languages and the formation of the present-day European gene pool were three consequences of the same population history.

This elegant theory was predominant among geneticists in the 1970s and 1980s. But in the ensuing two decades it was generally rejected, since new data on mitochondrial DNA demonstrated that the European gene pool has a Paleolithic rather than Neolithic age [Richards et al., 1996; Comas et al., 1997; Torroni et al., 1998]. Of course, the "out-of-Anatolia" trend in the

European gene pool was still a stable fact, but it was reinterpreted as the result of a Paleolithic rather than Neolithic migration into Europe (both migration waves entered Europe through Anatolia and the Balkans). The spread of farming also remained a stable fact, but it was reinterpreted as a result of "cultural diffusion" (spread of farming technology without the accompanying spread of farmers themselves) rather than demic diffusion.

Recently, our study on ancient mitochondrial DNA [Haak et al., 2010] provided direct data on ancient DNA and its comparison with data on modern DNA by means of a geographical map of genetic distances. The data on ancient DNA were from the early Neolithic European site, while the modern DNA data covered the entire Europe and Near East (Figure 3). It was found that the gene pool of the early Neolithic farmers was drastically different from the modern European one, but showed close affinities with the modern (and probably ancient) Near Eastern gene pool. One may conclude that the direct migration of farmers from Anatolia to Central Europe did indeed take place (as stated by Ammerman and Cavalli-Sforza), but that their gene pool was subsequently dissolved among autochthonous European populations. This is, therefore, a compromise between "demic diffusion" and "cultural diffusion" models. Of course, we do not know whether this pattern of Neolithization of Europe was indeed linked with the spread of Indo-Europeans; however, we do know that no one has so far suggested a better theory, and there are no reasons to abandon it.

The "genetic Indo-European" marker: myth or reality?

Since the 1990s, human population genetics mainly used two genetic systems: mitochondrial DNA and Y-chromosome. Within these systems, a number of haplogroups was discovered, many of which have clear patterns of distribution across the Earth. It became very popular to link the spread of every haplogroup with certain population events (like back migration to Africa [Cruciani at al., 2002], Mesolithic recolonization of Europe from Mediterranean refugia [Torroni et al., 2001; Rootsi et al., 2004], spread of Islam [Eaaswarkhanth et al., 2010], and many others).

It is more hard, however, to find a good "candidate" haplogroup that would mark the Indo-European expansion. It was stated many times (mainly on Internet forums, but also in some research papers) that Y-chromosomal haplogroup R1a could be the "Indo-European marker". Using both published data from available literature and our own unpublished data (523 populations worldwide altogether), we have constructed the gene geographical map of the distribution of this haplogroup across Eurasia (Figure 4). The map shows that (in agreement with the possible link with Indo-European movements) this haplogroup is widespread in Central and East Europe, in West Central Asia (where the genetic legacy of Iranic-speaking Scythians has survived) and in North India (particularly in the upper castes). However, the low frequency of this haplogroup in West Europe is in disagreement with the possible link with Indo-Europeans. The frequency in West Europe, Armenia and Anatolia (typical Indo-European areas) is as low as in Mongolia, which certainly was not a part of the Indo-European area. The highest frequency of this haplogroup is found in East Slavic populations, which stimulated some nationalistic activists to go as far as to claim the origin of all Indo-European populations from Russians, and insist that present day Russian people carrying the R1a haplogroup are the most direct descendants of "Proto-Indo-Europeans". This marginal "theory" could hardly be called science. But the haplogroup R1a is indeed very interesting as a possible marker of at least some episodes of the history of IE populations, such as their substitution by Turkic speakers in West Central Asia.

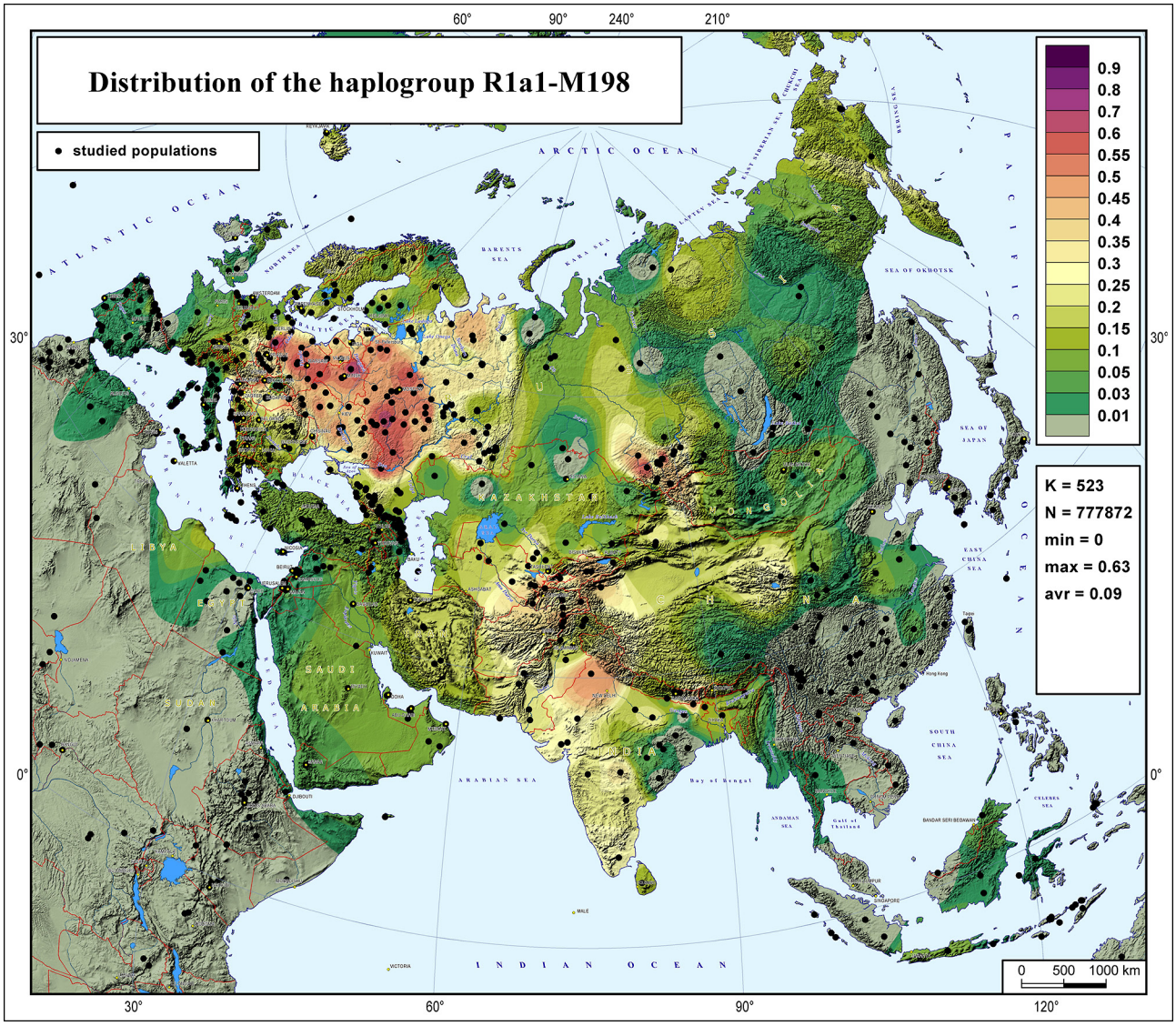

Figure 4. Map of distribution of the haplogroup R1a-M198 (data on Y-chromosome).

The main problem of such an interpretation lies in genetic data on the date and place of origin of this haplogroup. The date is too old, and the place is too far to the East to fit any hypothesis on the IE homeland that is currently being discussed by linguists and archeologists. The *date* is problematic (like every genetic date) and could easily change in the future. But the most reliable method to estimate the *place* of the haplogroup origin (the gradient of genetic diversity within the haplogroup) shows that haplogroup R1a initially spread *from* India rather than in the opposite direction [Underhill et al., 2010; Sharma et al., 2009; Sengupta et al., 2006].

There are other haplogroups that could mark the spread of some Indo-European branches. For example, we found the haplogroup G1-M285 to be widespread in the Kazakh clan of the *Argyns*, who could be descendants of IE-speaking Saks, assimilated by groups of Turkic speakers in the 1st millennium AD. The same haplogroup is also spread in some Iranic-speaking and Armenian-speaking populations, which might indicate either a common origin or intensive contacts between these populations (Figure 5).

In general, we believe that the "Indo-European marker" does not exist, simply because the first population to speak Proto-Indo-European must have possessed a spectrum of haplogroups which were shared (or identical) with its sister and neighbor populations that spoke other languages. It is unlikely that a mutation occurred exactly at the time when the first

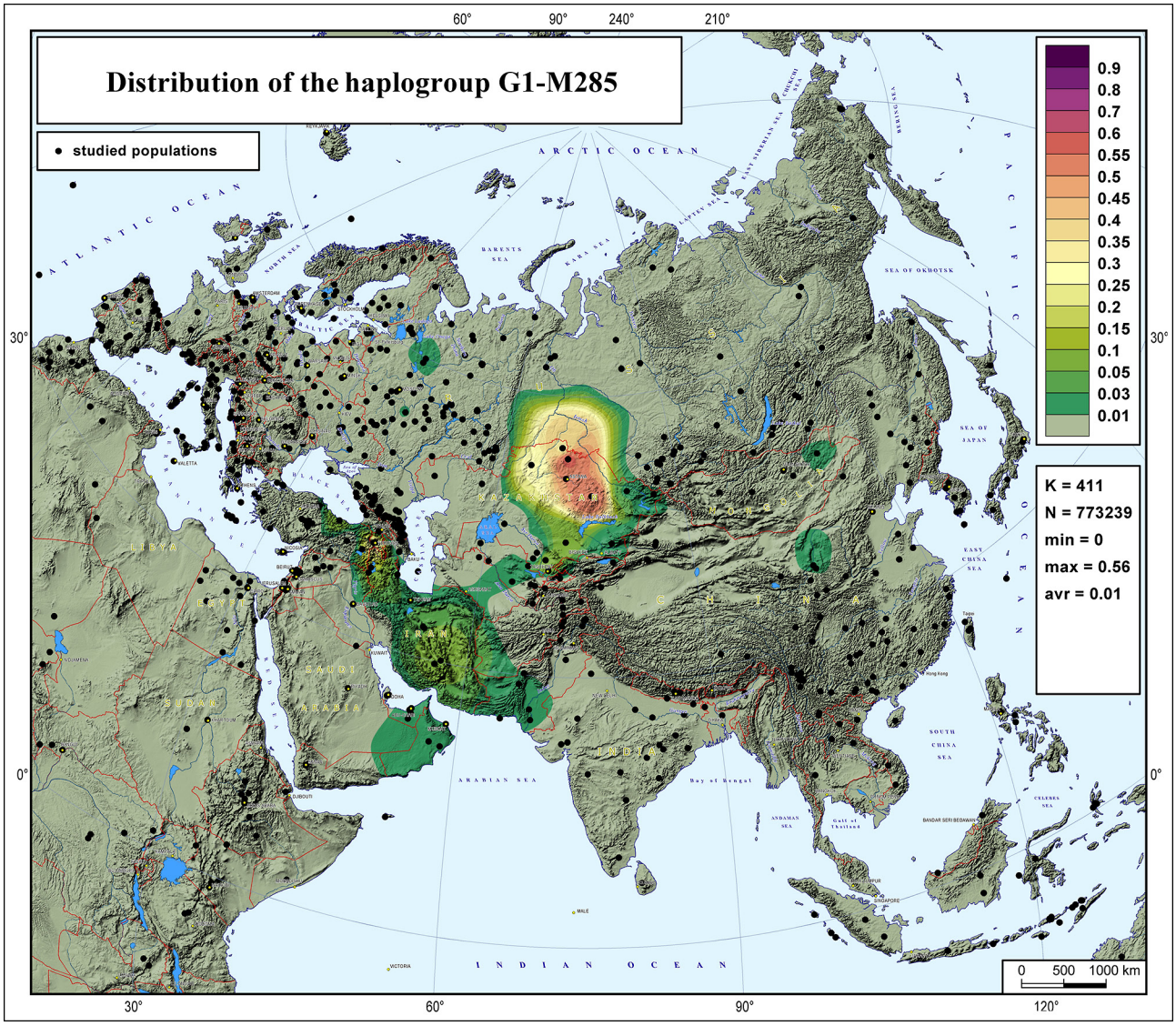

Figure 5. Map of distribution of the haplogroup G1-M285 (data on Y-chromosome).

Proto-Indo-European phrase was spoken. The mutations that occurred earlier must have been shared between IE and neighboring non-IE populations. The mutations that occurred later must have been specific for only a subset of IE populations. And this, in turn, defines the possibilities for future studies.

Possibilities for future studies

At first the progress of population genetics was mainly due to the development of new concepts, approaches, and methods. Although experimental data were, of course, an important component of this discipline, the driving force was the theory. Leading researchers of the first period (Sergey Chetverikov, Theodosius Dobzhansky, Alexander Serebrovsky, Ronald Fisher, Samuel Wright and others) and, later, Masatoshi Nei arrived at their discoveries through intellectual, rather than experimental, means. In the "post-Lysenko renaissance" of human population genetics in Russia, linked with the name of Yuri Rychkov (and, simultaneously, the seminal work of Luca Cavalli-Sforza and his colleagues in Europe), the research was already based on creating databases of human genetic variation; however, the analysis of this

experimental data, its interpretation and conclusions were still based on the development of new concepts, methods, and ideas.

The present day situation in human population genetics is completely different. The progress in experimental methods of DNA analysis in the last 10–20 years (from RFLP analysis to STR analysis to direct sequencing to next generation sequencing, with the "third generation sequencing" expected in the very near future) was so fast that, unfortunately, scientific progress became a technically driven rather than intellectually driven process. Every 3–5 years a new type of genetic systems becomes available. And each kind of these markers has features that look very promising. For example, the non-recombinative nature of mtDNA allowed the possibility of genetic dating; Y-chromosome allowed to trace paternal lines that allow for possible comparison with genealogies and clan ancestry legends; multiple SNP panels allowed the full coverage of the genome which seems to solve the problem of biased representation of analyzed loci.

However, rapid change of the genetic systems also created "scientific fashion". A new type of markers would become popular simply because it was fashionable, rather than due to any really important advantages. Even worse is the fact that studies, dealing with previous type(s) of markers, often meet problems with being published in prestigious journals, only because these markers are "out of fashion" rather than because of the studies themselves. This creates a paradoxical situation in which the most widely cited papers that make global conclusions are often based on weak datasets and small sample sizes. This is because publishing in a prestigious journal is possible when the marker is still "in fashion"; but at that particular time the accumulated dataset is still relatively small. And after many researchers have worked with the marker and accumulated a large dataset from all over the world, the marker is already "out of fashion", so that the papers based on these datasets cannot reach a high citation level.

It seems that, in order to reach its full potential in contributing towards solving the Indo-European problem, human population genetics needs to combine the data on all types of markers which are (and were) widely used in population genetic studies. Moreover, it is necessary to switch back to the "theoretical" style of research, since the problem itself is more complicated than simply tracing historical migration. Many geneticists in different countries could work in these directions. Below we provide an overview of the concrete project which our team plans to perform in the nearest three years. It is based on two (rather than "all") genetic systems. It also includes only one new and two older analytical methods. And it certainly does not pretend to "employ the full potential" of genetics in contributing towards Indo-European studies. However, we hope that it might represent an important step in this direction.

Although there is no single genetic marker that could be definitively linked with all Indo-Europeans, there could be sets of genetic markers, partially linked with some Indo-European branches. The more genetic markers we take into consideration, the higher are the chances that such "sub-Indo-European" markers might be identified. The present day genetic techniques allow this approach. For medical genetic purposes, hundreds of thousands of polymorphic markers were found in the human genome. The method of "genetic chips" provides the possibility to check for presence or absence of these numerous markers in the individual DNA sample. Using this technique we plan to perform the following five-step study.

Step 1: Six population pairs will be genotyped by 130000 genetic markers; each population pair includes the IE population and its non-IE neighbor. These pairs are: Russians and Karelians, Ukrainians and Nogais; Ossets and Adyghe; Armenians and Georgians; Tajiks and Turkmens; Pomiri and Kirghiz. Three additional pairs are also planned in collaboration with foreign laboratories: French and Basque; Iranians and Syrians, Brahmins and Dravidian populations.

Step 2: The genetic markers that are typical for IE populations but not for their non-IE neighbors will be identified. We do not expect to find markers that will be present in all IE populations and absent in all non-IE groups. We do expect to find markers that are *frequent* in groups of IE populations but rare in their non-IE "couples". For example, we hope to find markers that are present in Tajiks, Pomiri and Ossets (Iranic group) but absent among their non-IE neighbors.

Step 3: Maps of spatial distribution of these markers will be created, and statistical analysis of their frequencies will be performed. The results will be interpreted in terms of the population history of Indo-European branches under study, including their migrations, admixture and differentiation.

Step 4: The formal correlation analysis will be performed between matrices of genetic, lexicostatistical and geographical distances between Indo-European populations. Although the "pan-IE" analysis might be non-informative, we expect novel results from analysis at the level of different branches of the IE family, particularly Balto-Slavic and Iranian ones.

Step 5: Out of all the types of genetic markers, Y-chromosomal haplogroups have the highest level of inter-population differentiation and, therefore, have the maximum power to distinguish between populations. This is particularly true when large haplogroups, spread over vast areas, are subdivided into sub-haplogroups with geographically restricted areas. We plan to follow this approach with the aim to better trace migrations and reconstruct at least some parts of the multi-layer mosaic of Indo-European movements.

Literature

Ammerman A. J., Cavalli-Sforza L. L. *Neolithic Transition and the Genetics of Populations in Europe.* Princeton, N. J.: Princeton University Press. 1984.

Balanovsky O., Rootsi S., Pshenichnov A., Kivisild T., Churnosov M., Evseeva I., Pocheshkhova E., Boldyreva M., Yankovsky N., Balanovska E., Villems R. Two sources of the Russian patrilineal heritage in their Eurasian context. *American Journal of Human Genetics*. 2008. Vol. 82(1). P. 236–250.

Balanovsky O., Dibirova Kh., Dybo A., Mudrak O., Frolova S., Pocheshkhova E., Haber M., Platt D., Schurr T., Haak W., Kuznetsova M., Radzhabov M., Balaganskaya O., Druzhinina E., Zakharova T., Hernanz D., Zalloua P., Koshel S., Ruhlen M., Renfrew C., Wells R. S., Tyler–Smith C., Balanovska E. & The Genographic Consortium. Parallel Evolution of Genes and Languages in the Caucasus Region. *Molecular Biology and Evolution*. 2011. Vol. 28(10). P. 2905–2920.

Behar D. M., Harmant C., Manry J., van Oven M., Haak W., Martinez-Cruz B., Salaberria J., Oyharçabal B., Bauduer F., Comas D., Quintana-Murci L. & The Genographic Consortium. The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. *American Journal of Human Genetics*. 2012. Vol. 90(3). P. 486–493.

Bogacsi-Szabo E., Kalmar T., Csanyi B., Tomory G., Czibula A., Priskin K., Horvath F., Downes C. S., Rasko I. Mitochondrial DNA of ancient Cumanians: culturally Asian steppe nomadic immigrants with substantially more western Eurasian mitochondrial DNA lineages. *Human Biology*. 2005. Vol. 77. P. 639–662.

Cavalli-Sforza L. L., Menozzi P., Piazza A. *The History and Geography of Human Genes.* Princeton: Princeton University Press. 1994.

Comas D., Calafell F., Mateu E., Pérez-Lezaun A., Bosch E., Bertranpetit J. Mitochondrial DNA variation and the origin of the Europeans. *Human Genetics*. 1997. Vol. 99(4). P. 443–449.

Cruciani F., Santolamazza P., Shen P., Macaulay V., Moral P., Olckers A., Modiano D., Destro-Bisol G. et al. An Asia to Sub-Saharan Africa back migration is supported by high-resolution analysis of human Y chromosome haplotypes. *American Journal of Human Genetics*. 2002. Vol. 70. P. 1197–1214.

Csanyi B., Bogacsi-Szabo E., Tomory Gy., Czibula A., Priskin K., Csosz A., Mende B., Lango P., Csete K., Zsolnai A., Conant E. K., Downes C. S., Rasko I. Y-Chromosome Analysis of Ancient Hungarian and Two

Modern Hungarian-Speaking Populations from the Carpathian Basin. *Annals of Human Genetics*. 2008. Vol. 72. P. 519–534.

Czeizel A. E., Benkman H. G., Goedde H. W. *Genetics of the Hungarian populations*. Berlin: Springer-Verlag. 1991.

Eaaswarkhanth M., Haque I., Ravesh Z., Romero I. G., Meganathan P. R., Dubey B., Khan F. A., Chaubey G., Kivisild T., Tyler-Smith C., Singh L., Thangaraj K. Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *European Journal of Human Genetics*. 2010. Vol. 18(3). P. 354–363.

Haak W., Balanovsky O., Sanchez J. J., Koshel S., Zaporozhchenko V., Adler C. J., Der Sarkissian C. S., Brandt G., Schwarz C., Nicklisch N., Dresely V., Fritsch B., Balanovska E., Villems R., Meller H., Alt K. W., Cooper A., Members of the Genographic Consortium. Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biology*. 2010. Nov 9; 8(11): e1000536.

Kivisild T., Rootsi S., Metspalu M., Mastana S., Kaldma K., Parik J., Metspalu E., Adojaan M., Tolk H. V., Stepanov V., Gölge M., Usanga E., Papiha S. S., Cinnioğlu C., King R., Cavalli-Sforza L., Underhill P. A., Villems R. The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *American Journal of Human Genetics*. 2003. Vol. 72(2). P. 313–332.

Martínez-Cruz B., Harmant C., Platt D. E., Haak W., Manry J., Ramos-Luis E., Soria-Hernanz D. F., Bauduer F., Salaberria J., Oyharçabal B., Quintana-Murci L., Comas D., Genographic Consortium. Evidence of pre-roman tribal genetic structure in basques from uniparentally inherited markers. *Molecular Biology and Evolution*. 2012. Vol. 29(9). P. 2211–2222.

Richards M., Côrte-Real H., Forster P., Macaulay V., Wilkinson-Herbots H., Demaine A., Papiha S., Hedges R., Bandelt H. J., Sykes B. Paleolithic and neolithic lineages in the European mitochondrial gene pool. *American Journal of Human Genetics*. 1996. Vol. 59(1). P. 185–203.

Rootsi S., Magri C., Kivisild T., Benuzzi G., Help H., Bermisheva M., Kutuev I., Barać L., Pericić M., Balanovsky O., Pshenichnov A., Dion D., Grobei M., Zhivotovsky L. A., Battaglia V., Achilli A., Al-Zahery N., Parik J., King R., Cinnioğlu C., Khusnutdinova E., Rudan P., Balanovska E., Scheffrahn W., Simonescu M., Brehm A., Goncalves R., Rosa A., Moisan J. P., Chaventre A., Ferak V., Füredi S., Oefner P. J., Shen P., Beckman L., Mikerezi I., Terzić R., Primorac D., Cambon-Thomsen A., Krumina A., Torroni A., Underhill P. A., Santachiara-Benerecetti A. S., Villems R., Semino O. Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *American Journal of Human Genetics*. 2004. Vol. 75(1). P. 128–137.

Rosser Z. H., Zerjal T., Hurles M. E., Adojaan M., Alavantic D., Amorim A., Amos W., Armenteros M., Arroyo E., Barbujani G., Beckman G., Beckman L., Bertranpetit J., Bosch E., Bradley D. G., Brede G., Cooper G., Côrte-Real H. B., de Knijff P., Decorte R., Dubrova Y. E., Evgrafov O., Gilissen A., Glisic S., Gölge M., Hill E. W., Jeziorowska A., Kalaydjieva L., Kayser M., Kivisild T., Kravchenko S. A., Krumina A., Kucinskas V., Lavinha J., Livshits L. A., Malaspina P., Maria S., McElreavey K., Meitinger T. A., Mikelsaar A. V., Mitchell R. J., Nafa K., Nicholson J., Nørby S., Pandya A., Parik J., Patsalis P. C., Pereira L., Peterlin B., Pielberg G., Prata M. J., Previdere C., Roewer L., Rootsi S., Rubinsztein D. C., Saillard J., Santos F. R., Stefanescu G., Sykes B. C., Tolun A., Villems R., Tyler-Smith C., Jobling M. A. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American Journal of Human Genetics*. 2000. V. 67. P. 1526–1543.

Semino O., Passarino G., Quintana-Murci L., Liu A., Béres J., Czeizel A., Santachiara-Benerecetti A. S. MtDNA and Y-chromosome polymorphisms in Hungary: inferences from the palaeolithic, neolithic and Uralic influences on the modern Hungarian gene pool. *European Journal of Human Genetics*. 2000. Vol. 8(5). P. 339–346.

Sengupta S., Zhivotovsky L. A., King R., Mehdi S. Q., Edmonds C. A., Chow C. E., Lin A. A., Mitra M., Sil S. K., Ramesh A., Usha Rani M. V., Thakur C. M., Cavalli-Sforza L. L., Majumder P. P., Underhill P. A. 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. *American Journal of Human Genetics*. 2006. Vol. 78(2). P. 202–221.

Sharma S., Rai E., Sharma P., Jena M., Singh S., Darvishi K., Bhat A. K., Bhanwer A. J., Tiwari P. K., Bamezai R. N.. The Indian origin of paternal haplogroup R1a1(*) substantiates the autochthonous origin of Brahmins and the caste system. *Journal of Human Genetics*. 2009. Vol. 54 (1). P. 47–55.

Szij E. Research on the prehistory of the Hungarians and Finno-Ugric studies. In: Mende B. G. (ed.) *Research on the prehistory of the Hungarians: A review.* Hungarian Academy of Sciences Archaeological Institute. 2005. VAH 18: 115–156.

Tomory G., Csanyi B., Bogacsi-Szabo E., Kalmar T., Czibula A., Csosz A., Priskin K., Mende B., Lango P., Downes C. S., Rasko I. Comparison of maternal lineage and biogeographic analysis of ancient and modern Hungarian populations. *American Journal of Physical Anthropology*. 2007. Vol. 134. P. 354–68.

Torroni A., Bandelt H. J., Macaulay V., Richards M., Cruciani F., Rengo C., Martinez-Cabrera V., Villems R., Kivisild T., Metspalu E., Parik J., Tolk H. V., Tambets K., Forster P., Karger B., Francalacci P., Rudan P., Janicijevic B., Rickards O., Savontaus M. L., Huoponen K., Laitinen V., Koivumäki S., Sykes B., Hickey E., Novelletto A., Moral P., Sellitto D., Coppa A., Al-Zaheri N., Santachiara-Benerecetti A. S., Semino O., Scozzari R. A signal, from human mtDNA, of postglacial recolonization in Europe. *American Journal of Human Genetics*. 2001. Vol. 69(4). P. 844–852.

Torroni A., Bandelt H. J., D'Urbano L., Lahermo P., Moral P., Sellitto D., Rengo C., Forster P., Savontaus M. L., Bonné-Tamir B., Scozzari R. mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *American Journal of Human Genetics*. 2001. 1998. Vol. 62(5). P. 1137–1152.

Underhill P. A., Myres N. M., Rootsi S., Metspalu M., Zhivotovsky L. A., King R. J., Lin A. A., Chow C. E., Semino O., Battaglia V., Kutuev I., Järve M., Chaubey G., Ayub Q., Mohyuddin A., Mehdi S. Q., Sengupta S., Rogaev E. I., Khusnutdinova E. K., Pshenichnov A., Balanovsky O., Balanovska E., Jeran N., Augustin D. H., Baldovic M., Herrera R. J., Thangaraj K., Singh V., Singh L., Majumder P., Rudan P., Primorac D., Villems R., Kivisild T. Separating the post–Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European Journal of Human Genetics*. 2010. Vol. 18(4). P. 479–484.

О. П. Балановский, О. М. Утевская, Е. В. Балановская. Молекулярно-генетические исследования индоевропейских популяций: прошлое и будущее.

Представлен опыт сравнения генетических и лингвистических данных в связи с индоевропейской проблематикой. Наше сравнение генетического разнообразия и лексикостатистических данных по северокавказским популяциям выявило параллелизм в эволюции генов и языков; можно сказать, что популяционная история отражается в лингвистическом и генетическом зеркалах. Для других лингвистических семей можно ожидать такого же сходства, хотя оно и может быть искажено событиями «доминирования элиты» и другими факторами, по-разному влияющими на генный фонд и на лексический фонд. И действительно, для индоевропейских популяций Европы, в отличие от Кавказа, частные корреляции выявили большую роль географического ($r = 0.32$), чем лингвистического фактора ($r = 0.21$) в структурировании генофонда; но при этом бóльшáя парная корреляция ($r = 0.67$) между генетическими и лингвистическими расстояниями позволяет использовать лексикостатистические данные для прогноза генетического сходства между популяциями. Сходство генетических и лингвистических данных выявлено как по Y-хромосоме (популяции кластеризуются по языку), так и по митохондриальной ДНК (популяции кластеризуются по принадлежности к языковой группе). В целом, мы считаем, что не существует одного генетического маркера, вполне связанного с расселением индоевропейцев. Вместо этого, мы начинаем новый проект, направленный на выявление групп маркеров, частично связанных с отдельными группами индоевропейцев, что позволит реконструировать некоторые части многослойной мозаики индоевропейских миграций.

Ключевые слова: генофонд, индоевропейские популяции, Y-хромосома.