

---

# **Russian Journal of Genetic Genealogy**

**Vol 1, №1, 2010**

ISSN: 1920-2989

**Russian Journal of Genetic Genealogy**

Publisher Lulu inc., 2010

*All rights reserved. No part of this publication may be reproduced, altered in any form or by any means: mechanical, electronic, with photocopying, etc., without the prior written permission of the publisher of the journal, or authors of articles.*

*When citing a reference to this publication is required.*

Editor

*Michael Temosh*

Art editor

*Roman Sychev*

*Nataliya Zyryanova*

Technical editor

*Denis Grigoriev*

Reviewer

*Alexander Kireev*

Contact address

*[rjgg.editor@gmail.com](mailto:rjgg.editor@gmail.com)*

© **RJGG, 2010**

## Contents

European prehistory in mirror of genetics: a contemporary view <i>Oleg Balanovsky</i> .....	5
Geographic distribution and molecular evolution of ancestral Y chromosome haplotypes in the Low Countries <i>Gerhard Mertens, Hugo Goossens</i> .....	17
Whit Athey, developer of Y-haplogroup predictor <i>Denis Grigoriev</i> .....	29
Comment on "Geographic distribution and molecular evolution of ancestral Y chromosome haplotypes in the Low Countries" by Gerhard Mertens and Hugo Goossens <i>Guido Deboeck</i> .....	35
Y-haplogroups of carriers of the Aryan language <i>A.A. Aliev, A.S. Smirnov</i> .....	38
Origin of "Jewish" clusters of E1b1b1 (M35) haplogroup <i>A.A. Aliev</i> .....	47
Modern carriers of haplogroup E1b1b1c1 (M34) are the descendants of the ancient Levantines <i>A.A. Aliev, Bob Del Turco</i> .....	50



# European prehistory in mirror of genetics: a contemporary view\*

Oleg Balanovsky<sup>1</sup>

## Introduction

Numerous branches of knowledge are currently affected by a certain eurocentrism, this being especially the case of the studies focused on human population genetics. These studies were (and remain) developed predominantly in European laboratories and are less popular on other continents. Population geneticists need human populations as subject matter of their research, and in most cases we study those which are in proximity of our homes. That is why Europe is the continent which is by far better studied genetically, and the European genetic landscape is much deeper understood and discussed than any other part of the world. For the same reason the controversies between different schools of thought regarding various aspects of the European gene pool became much more apparent

## The two concepts

The variation of "classical" genetic markers (which became referred to like that when the DNA came to the fore to replace them) was best summarized in the book by Luigi Luca Cavalli-Sforza and his colleagues published in 1994. Unsurprisingly, the largest chapter of this book is the "European" one, which describes how the European genetic landscape had been formed during the Neolithic expansion from the Near East. It was one of the most minutely-elaborat-

ed concepts in population genetics at that time; nonetheless it was almost entirely rejected in the subsequent decade.

The European genetic landscape, as restored based on the analysis of classical markers, shows three principle features: 1) a general homogeneity (the Europeans are genetically very similar to each other, compared to populations of other continents); 2) the presence of only a few outliers (isolated peripheral populations such as Icelanders, Saami, or Sardinians); their peculiarities are the secondary, having arose after these populations were demographically split off and underwent the genetic drift from the main European corpus; 3) Clear geographic patterns of gradual genetic changes.

To identify these geographic patterns Cavalli-Sforza and his colleagues (Menozzi et al., 1978, Cavalli-Sforza et al., 1994) and independently Russian geneticists (Rychkov & Balanovskaya, 1992) developed the method of "synthetic maps". These maps are created by a complex mathematical algorithm but in a simpler way they consist of displaying the geographic distribution of an "ideal" genetic marker, which correlates with geographical patterns of the majority of real markers presenting the data (Menozzi et al., 1978; Rychkov & Balanovskaya, 1992; Balanovskaya & Nurbaev, 1997a). This synthetic map visually demonstrated gradual changes with a remarkable geographical pattern: from Anatolia via the Balkans over the rest of Europe i.e. from the Southeast to the Northwest (Fig 1). This picture was interpreted as a result of the gradual spread of farming (and farmers) across Europe which was known since Gordon Childe (1928) to follow the same trajectory.

This concept was additionally substanti-

Received: June 8 2009; accepted: June 8 2009; published June 30 2009

<sup>1</sup>Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia

\* This study has been originally published in: Balanovsky O.P. (2009) Human genetics and Neolithic dispersal' The East European Plain on the eve of agriculture edited by Pavel M. Dolukhanov, Graeme R. Sarson and Anvar M. Shukurov. BAR -S1964. ISBN 978 1 4073 0447 2. P:235-46



**Figure 1.** Synthetic map, summarizing genetic variation in Europe (from Cavalli-Sforza et al., 1994).

ated in two ways. First, the "isogenes" (lines connecting the same gene pools on the genetic map) have shown a remarkable agreement with isochrones (lines showing the early arrival of agriculture based on archaeological and radiometric evidence). Second, the concept and the mathematical model of the so-called demic diffusion was developed (Ammerman & Cavalli-Sforza, 1984). It implies a slow (generation by generation) migration of farmers which assimilated indigenous populations, and thereby gradually dissolved the initial "farming" gene pool. As a consequence, the geographic trajectory of migration becomes a geographic line of gradual genetic changes: from a "mainly farming" gene pool in Anatolia to a "mainly indigenous" one in the Europe's north-west and north-east (as the most distant from Anatolia).

This elegant, reasonable and sufficiently substantiated concept of the origin and composition of the European gene pool dominated population genetics in the 1980-1990s. Generally, four elements of this concept could be potentially criticised:

- i) the data-set (classical markers);
- ii) the methodology (synthetic maps);
- iii) the logical foundation (attributing south-east-northwest pattern to Neolithisation) or
- iv) controversial results obtained with a use of independent data, methods and logics. Critics used all four elements but with a variable success.

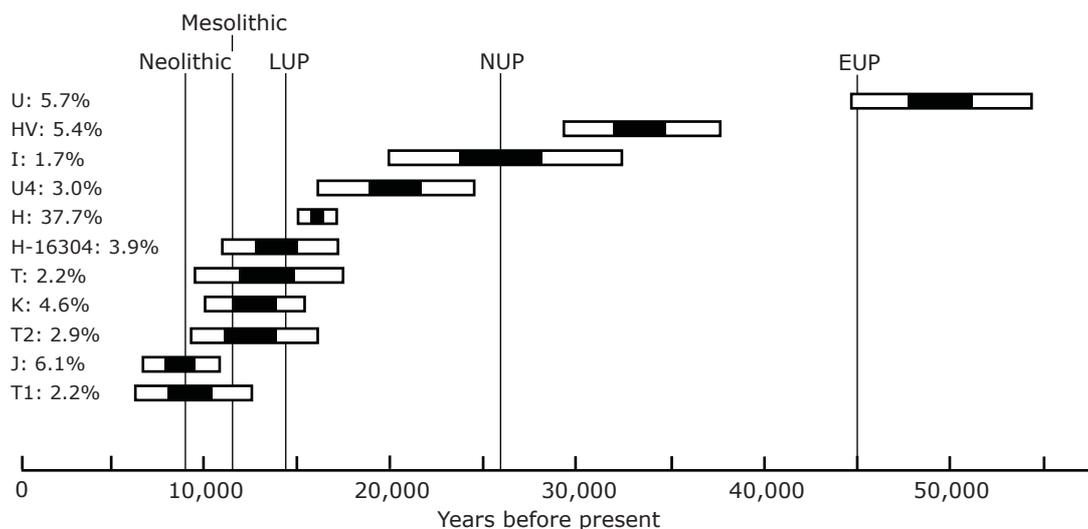
The popular idea that classical markers are "worse" than new DNA markers has never been positively proven and should be considered rather as a scientific fashion. Yet some critics tend to reject the classical markers arguing that they are affected by natural selection, and therefore

their variation would be the result of both historical and biological factors. However, many DNA markers can be equally affected by biological factors and therefore geographic distribution of a genetic marker reflects the history which to some degree was blurred by biological selection affecting this marker. And, second, the biological factors differently affect various markers and therefore disappear when averaging, in the case when numerous markers are considered (Yamazaki & Maryama, 1973; Lewontin & Krakauer, 1975; Balanovskaya & Nurbaev, 1997b).

The method of synthetic maps was attacked by Robert Sokal, who used an alternative method (autocorrelation analysis) for revealing geographical patterns in the genetic data (Sokal, Oden, 1978). Using computer simulation he demonstrated that synthetic maps compiled from interpolated maps produce gradual pattern even from randomly permuted data, hence the obtained patterns are artificial (Sokal et al., 1999). However, our recent simulations (Balanovsky et al., 2008) failed to recognise any difference between synthetic maps from interpolated surfaces (criticized by Sokal and colleagues), on the one hand, and from non-interpolated raw data (considered as control by Sokal and colleagues), on the other. It is equally remarkable that Sokal and colleagues did not express doubt that the gradual genetic pattern from Anatolia is the main feature of the European gene pool, yet they did question the methods applied for to identify this pattern.

Curiously enough, the applied logic conclusion (attributing the observed genetic pattern to the Neolithic expansion as the both followed the same trajectory) had never been criticized to the best of our knowledge, though population geneticists were well aware that a correlation never proves the cause-effect relationship. The interpretation in terms of the Neolithic expansion seemed so obvious, transparent, and natural, that this logical mistake became only apparent when controversial results started emerging from independent data.

The evidence, demonstrating the Palaeolithic time for the origin of the European gene pool was based on mitochondrial DNA (mtDNA). The principal difference between mtDNA and Y chromosomal markers on the one hand, and the classical and autosomal DNA markers on the other, resides in the presence or absence of recombination. Autosomal markers recombine and therefore each marker is inherited independently from all other markers. MtDNA and the main portion of the Y chromosome do not recombine. That is why every occurring mutation



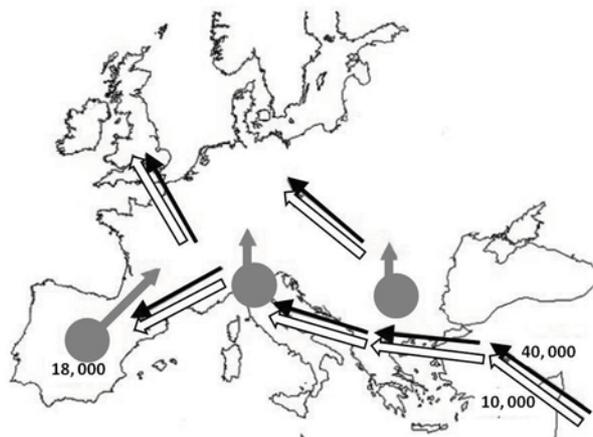
**Figure 2.** Ages of mitochondrial haplogroups in Europe (from Richards et al., 2000).  
EUP - Early Upper Palaeolithic; MUP - Middle Upper Palaeolithic; LUP - Late Upper Palaeolithic.

gets transmitted from generation to generation alongside other mutations, which did occur earlier. In other words, the mutation (a mistake in the genetic text) became forever part of this text and is transmitted together with all the mistakes that had appeared in this text earlier. When comparing different texts (so-called haplotypes) it is possible to trace mutations back in time and reconstruct a "genealogy" of these texts. i.e. to draw their "family tree". This tree is commonly rooted in the most recent common ancestor (a "mitochondrial Eve") and each branch of the tree differs by its particular set of mutations. Next, each twig of a certain branch carries all mutations characteristic to this branch, together with a set of additional "twig-specific" mutations. These branches and twigs are called haplogroups (subhaplogroups) each of them unites a group of closely related haplotypes (which can be compared with a leaves of this tree). Assuming an average rate of mutations one can calculate the age of each haplogroup by multiplying the number of accumulated mutations by the mutation rate.

This methodology, applied to the European mitochondrial pool (Richards et al., 1996), demonstrated that most branches (haplogroups) found in Europe were much older than the Neolithic and most of them fell into the age range of the Upper Palaeolithic. Based on this evidence it was concluded, that European gene pool was formed by the initial peopling of the continent by anatomically modern humans (AMH) during the Upper Palaeolithic, and that it is still present in the most of present-day Europeans. As for the Neolithic expansion, it had, therefore, a limited impact on the European gene pool.

Hence, this new concept imposed the "cultural diffusion" model of Neolithisation in contrast to "demic diffusion" model advanced by Ammerman and Cavalli-Sforza (1984).

The following decade witnessed a heated debate between two camps of geneticists, namely the "cultural diffusionists" and the "demists". Despite the ongoing debate, the methodological limits and benefits of both models are apparent. The source database for demic diffusion model was much richer (hundreds of markers studied in dozens of populations) while Richards and colleagues were restricted to one marker (mtDNA) studied in a limited set of populations. Arguably, as Barbujani and colleagues (1998) pointed out, the Palaeolithic origins of haplogroups found in Europeans do not necessarily imply that these haplogroups were present in Europe since the Palaeolithic. As haplogroups age is calculated based on its diversity, they could have accumulated diversity in other parts of the world arriving into Europe being already diverse. This problem of the pre-existing diversity met elegant solution in the following paper by Richards and colleagues (2000), which became the most recognized study of European genetics. In that paper the founder mtDNA lineages were identified which were deemed as the starting points for the entire European diversity accumulated in situ. Although different criteria for "founding" resulted in slightly different time assessments, all calculations demonstrated the Upper Palaeolithic age for most European clusters of lineages (haplogroups), while haplogroups whose appearance in Europe can be attributed to the Neolithic period make up only a quarter of the total European gene pool (Fig. 2).



**Figure 3.** A scheme of the main demographic processes documented in the archeological record of Europe (from Barbuĵani, Bertorelle, 2001).

Numbers are approximate dates, in years before the present. Black arrows, Paleolithic colonization; grey arrows, Late Palaeolithic recolonization from glacial refugia (grey circles); white arrows, Neolithic demic diffusion.

The opponents of this concept did not miss the opportunity to point out that the deepest time assessment for an *in situ* European haplogroup was paradoxically older than the age of AMH appearance in Europe (Barbuĵani, Bertorelle, 2001). It should be noted that the time estimates are based largely on the calibration points used (mutation rate). But nevertheless the ability to present a time estimate was the strongest among of Richards et al. arguments whereas the contrary concept was based exclusively on the similarity between the genetic pattern and that of the Neolithic spread. This enabled one to pinpoint the main logical weak point in the "Neolithic" concept, namely, that the AMH initial settlement of Europe followed the same geographical trajectory which was later used by expanding Neolithic farmers. When Barbuĵani and Bertorelle (2001) summed up this discussion they admitted, that the gradual "out of Anatolia" geographic pattern, as established by classical and (later) other markers was correct. Yet this pattern could have originated from both, the Palaeolithic and Neolithic migrations, as the both were believed to follow the same Anatolian route (Fig. 3). The lesson learnt was that geographic patterns of genetic variation do not allow distinguish between these scenarios, and one needs non-recombining systems which are essential for time estimations.

Nearly simultaneously with the seminal publication summarising the mtDNA data (Richards et al., 2000), two papers on the second non-recombining system appeared, summing up the paternal perspective, i.e. Y chromosomal

variations in Europe (Semino et al., 2000; Rosser et al., 2000). The both papers were based on extensive datasets. Although written in a different manner they established similar features.

Rosser and colleagues followed a phenomenological approach, describing patterns of Y chromosomal variation. They found very clear geographical clines in the distribution of all haplogroups and statistically calculated that the genetic similarity of populations was affected by their geographic proximity rather than linguistic similarity. In contrast, to that Semino and colleagues following an interpretative approach concluded that the observed geographical patterns could have been caused by the factors of similar geographic distribution in the Palaeolithic epoch. One may easily note, that in doing so they committed the same logical mistake as they interpreted geographical pattern "by association" with the known event of the same spatial pattern. And indeed, having reanalysed Semino's dataset, Chikhi et al (2002) came to the opposite conclusion and interpreted the clines as having been formed during the Neolithic. At that time, time the estimates for Y chromosomal haplogroups were much less informative and reliable than those for mtDNA. The reason for this is that it is hard to distinguish on the Y chromosome the pre-existing diversity (which founder population had brought from its homeland) and one accumulated *in situ*. (Founder analysis, which was the convenient instrument for mtDNA, proved to be too complex to be applied to the Y chromosome).

Since the 1990s, the studies on mitochondrial DNA and Y chromosome diversity became dominant in population genetics, which resulted in a specific "two-system" way of thinking. According to it, the greater part of migratory events was allegedly reflected in the both systems. Over the following years numerous studies were published on Y chromosomal and mtDNA variation in virtually all European countries. Most of them provided missing pieces for the European genetic puzzle but did refrain from making oversimplified and/or general conclusions. Those which did could be roughly classified into two groups: those describing the overall genetic landscape (based on the data of the totality of haplogroups) and deducing a particular genetic event from the distribution of particular haplogroup (the haplogroup-driving approach).

### Mitochondrial landscape of Europe

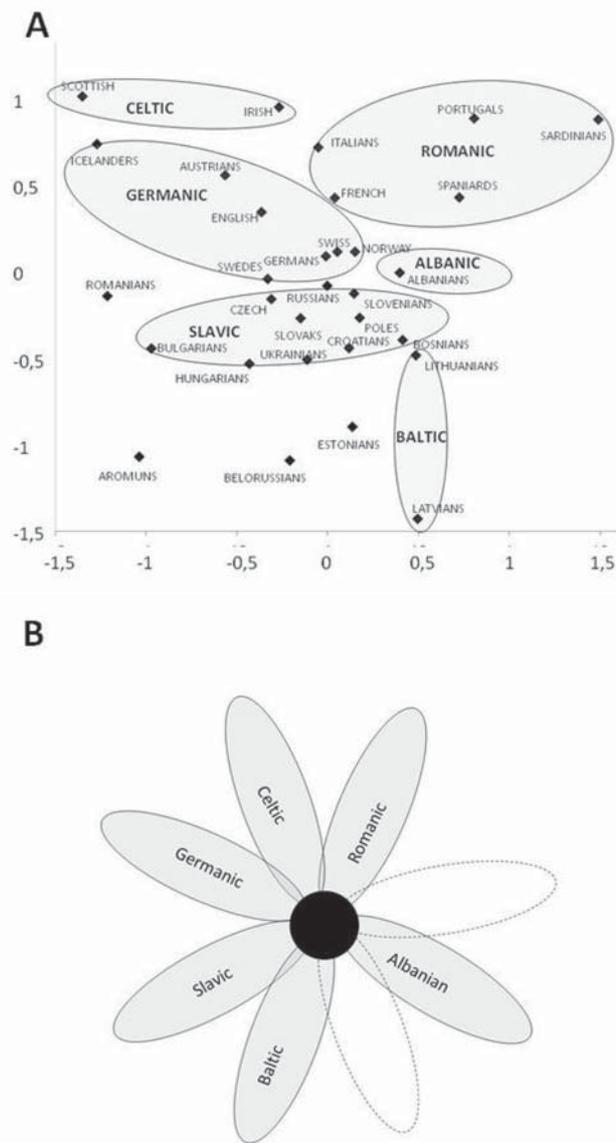
From the perspective of mitochondrial DNA, the European gene pool consists of 7-10 most

frequent haplogroups. All but one of them came from the Near East: in the majority of cases during the Early Upper Palaeolithic (EUP) in conjunction with the initial AMH dispersal and a smaller part during the Neolithic epoch (in the course of Neolithisation, Richards et al., 2000). The genetic landscape has been reshaped in the Mesolithic/Late Palaeolithic times, during the repopulation of Europe from the southern European refugia. The only European haplogroup that presumably had emerged in Europe (haplogroup V) became spread across the entire continent during Mesolithic/Late Palaeolithic recolonisation (Torroni et al., 2001). The western European origin of this haplogroup (the Franco-Cantabrian refugium) is presumed, based on its high frequency in this area as well as the occurrence of its phylogenetic predecessor (pre-V lineages). However, a recent accumulation of genetic data on previously poorly studied Eastern Europe, enabled the present author (Balanovsky, 2008) to suppose the occurrence of an additional East European centre of origin of this haplogroup. This finding is based on even higher frequency and yet again, on the presence of pre-V lineages in East European steppe area. Impossibility to distinguish between the western and eastern European homelands emphasised the important feature of the European mitochondrial landscape – its extreme homogeneity.

Indeed, when additional data from different European populations became available the genetic similarity in haplogroup frequencies (and identity in haplogroup spectra) has been immediately recognised (Simoni et al., 2000). As a result, mtDNA studies has appeared dealing with Europe as a whole, comparing it with the Near East or other areas, while attempts to trace genetic processes within Europe encountered problems (Helgason et al., 2000). This development was rather discouraging for archaeologists and linguists who were typically interested in a genetic support of existing hypothesis in their respective disciplines, although on a much smaller scale. Fortunately, the paper entitled "In search of geographic patterns in European mitochondrial DNA" (Richards et al., 2002) made the point that with the emergence of a larger dataset (with more than 3,000 individual mtDNAs) a spatial structuring became more apparent (e.g. the south-north difference was acknowledged among macro-regions of Europe: Mediterranean area, Central Europe, Scandinavia, and, surprisingly, the Basque Country).

Presently one can affirm that size of the dataset is the key factor. Having at our disposal

the database six times larger than previously possessed, comprising 20,000 European mtDNAs (Balanovska, Zaporozhchenko, Pshenichnov, Balanovsky; MURKA Mitochondrial Database and Integrated Software, unpublished) we were able to recognise a much clearer patterning (Fig. 4). European populations altogether occupying all parts of this plot provide a geometrical illustration of the genetic variation in Europe.



**Figure 4.** Genetic relationships of European populations from mitochondrial DNA perspective.  
**A.** The multidimensional scaling plot (geometric distances between points display the genetic distances between corresponding populations). Ellipses mark populations belonging to the same linguistic group.  
**B.** The idealized approximation of the plot (A) by flower-like structure. Black core – proto-Indo-European population; dotted ellipses – hypothetical extinct linguistic groups.

The most remarkable feature is that the pattern distinctly resembles linguistic groups: populations cluster together according to their linguistic group of the Indo-European family. Three largest clusters are formed by Romanic, Slavic and Germanic speakers, while Baltic and Celtic speakers form smaller clusters and Albanians form a "cluster" of its own outside any other cluster. There are a few exceptions: Romanians, Aromuns and Sicilians lie outside the Romanic cluster while Estonians join the Slavic cluster. In both cases the geographical distance (remoteness for Romanians and Aromuns, proximity for Estonians) had probably a stronger impact on genetics than the linguistic affiliation.

Among the non-Indo-European populations of Europe the Basques found their place outside any other cluster but close to the Romance one (not surprisingly, considering the geographic proximity again). Finno-Ugric and Turkic speakers are not shown on this plot because of their extreme genetic variation, but on another plot they lie apart of Indo-Europeans.

This linguistic structuring of European mitochondrial DNA follows a remarkable "flower shape" pattern: all clusters looking like petals around the "core" of the flower. The possible explanation of this pattern is that genetic and linguistic differentiations were parallel processes or, in better words, – two aspects of the same process, related to the multiplication and differentiation of proto-Indo-European population in Europe. It is well known, that in many particular cases distribution of genes is opposed to distribution of language (especially in cases of the language replacement by the elite dominance model). However, in very general view, almost all Europe is populated by speakers of one linguistic family and almost all Europe is genetically homogenous. This allows speculations (like our flower-like interpretation of the genetic plot) which consider genetic and linguistic evolution as generally parallel processes, disregarding partial exceptions. Such speculations inevitably oversimplify both processes but could serve as a starting point for more detailed studies.

Therefore, one can accept as a working hypothesis the differentiation of proto-Indo-European language into linguistic groups being accompanied by genetic differentiation resulting in a clear clustering pattern (Fig. 4). This allows one to introduce time frames into the formation of the European mitochondrial landscape. It would coincide with origin and differentiation of European branches of IE family, i.e. covers the last 5-6 millennia (Starostin et al., their linguistic database is available at <http://starling.rinet.ru/main.html>).

This does not necessarily imply the Neolithisation (for example, major changes during the Bronze Age is one of alternative explanations), but lends credence to the hypotheses advocating a relatively recent origin (or at least late major reshaping) of the mitochondrial pool in Europe.

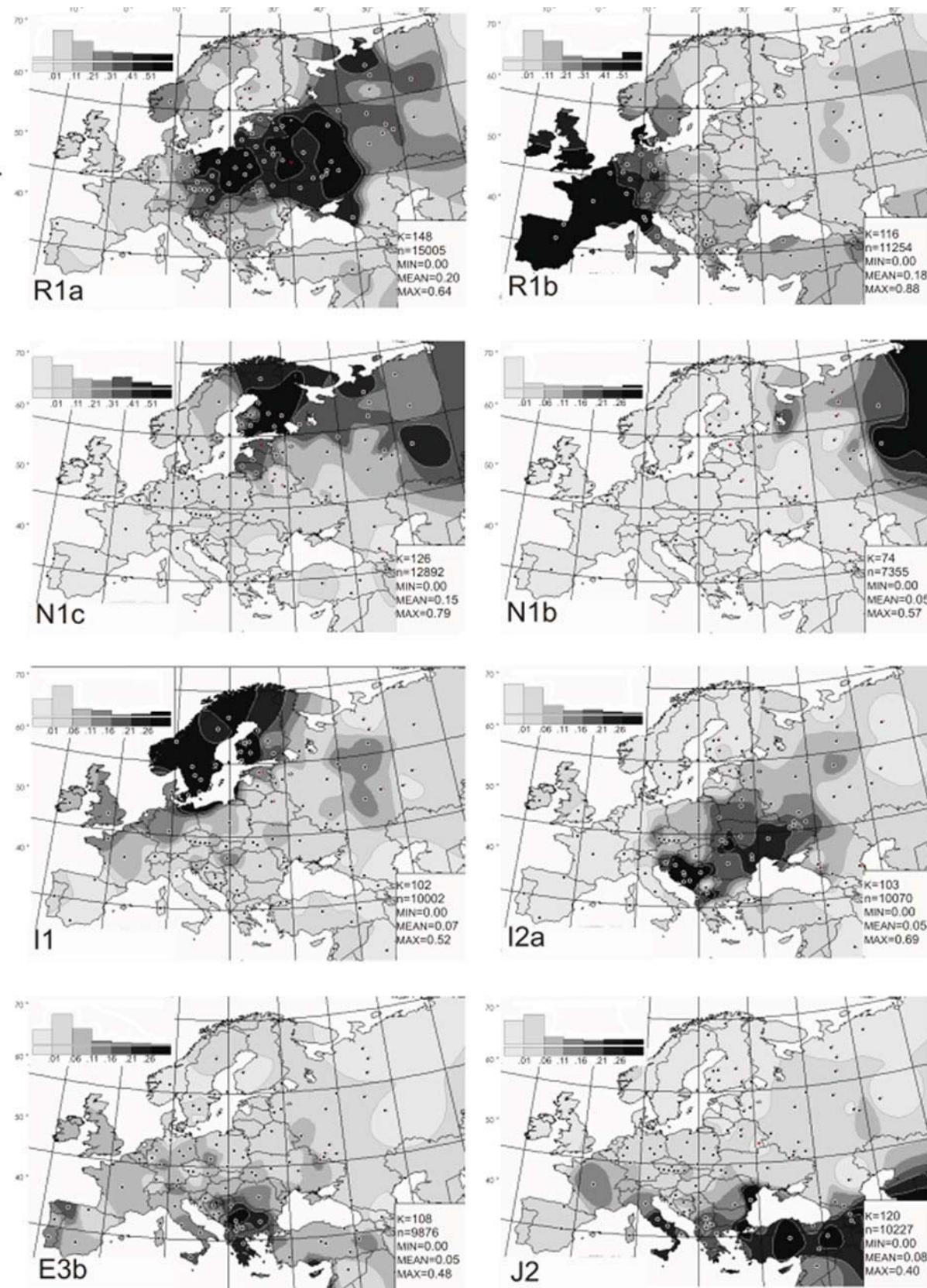
One may note that the significance of the linguistic factor is quite obvious on the graph (Fig. 4 A). However, the idea of a single proto-population totally depends on the flower-like structure of this graph (Fig. 4 B) and should be therefore considered as one of the plausible hypotheses.

### Y chromosomal landscape of the Europe

While the "homogeneity" is the principal feature of mitochondrial pool, the Y chromosomal pool is characterized by a high heterogeneity. As with mtDNA, there are seven Y chromosomal haplogroups dominating in Europe. But while frequencies of mitochondrial haplogroups are quite similar across Europe, Y chromosomal haplogroups follow a clear geographical pattern (Fig. 5). Neither classical markers, nor mitochondrial haplogroups demonstrated such obvious and elegant trends. Therefore, Y chromosome became an effective instrument in population genetics.

One should remember that European gene pool cannot be homogeneous and heterogeneous at the same time. The question is to what degree different markers are able to reveal the existing degree of variations. Having dozens of autosomal markers, classical population geneticists achieved reasonable resolution in assessing the variation between populations (Cavalli-Sforza et al., 1994). Mitochondrial DNA failed to reveal a difference between populations and successfully operates only at a higher hierarchical level: separating regions (Richards et al., 2002) and linguistic groups (present study, fig. 4). Y chromosome operates much better and separates even subpopulations within the same ethnic group (Balanovsky et al., 2008). Recent studies based on half of million autosomal markers became able to separate even individuals within subpopulations (Novembre et al., 2008).

This high differentiation power of Y chromosome (i.e. clear geographical clines of its haplogroups) was revealed already in the early large-scale studies (Semino et al., 2000; Rosser et al., 2000). These clines have been recently summarized in a panel of frequency distribution maps (Balanovsky et al., 2008). Two main haplogroups, accounting altogether almost for a



**Figure 5.** Geographic distribution of European Y chromosomal haplogroups (modified from Balanovsky et al., 2008).

K – number of studied populations; n – number of studied individuals; MIN, MEAN, and MAX – minimal, mean and average frequency on the map.

half of the total European Y chromosomal pool are distributed along the west-to-east axis. Haplogroup R1b accounts for roughly 50% of the Y chromosomal pool in Western Europe and decreases eastward, while R1a reaches the same high frequency in the east (Fig. 5) and decreases westward.

Analysis of another type of Y chromosomal markers (microsatellite variation) also proved the western and eastern domains to be main features of the Y chromosomal pool (Roewer et al., 2005). As it was stressed above, the "interpretation by association" should be made with caution. That is why attributing these domains to Late Palaeolithic re-colonisation from two principal refugia (the south-western and south-eastern ones) can be considered as a possible but not yet proven hypothesis. (One of other possibly hypotheses is attributing these genetic domains to descendants of Late Neolithic Bell Beaker and Corded Ware cultures).

These two principal European haplogroups R1b and R1a are shared between Europe and other regions (Central Asia, Near East, India and North Africa). But two other haplogroups, I1 and I2a (according to previously used nomenclature the same haplogroups were labelled as I1a and I1b, respectively) are restricted to Europe, where they had likely originated.

While R1b and R1a occupy the west and the east, I1 and I2a predominate in the Europe's north and south, respectively. I1 which is frequent in Scandinavia and southern Baltic area has attracted less attention due to obviously late colonization of this region. In contrast, the distribution of haplogroup I2a (Balkan haplogroup) has been widely debated. As southeast European autochthonous haplogroup it could not be attributed to Neolithic immigrants (or any other immigrants) into Europe. We will discuss it in more details below.

Three remaining haplogroups (E, J, and N1c) are not evenly spread across the entire Europe but are restricted to distinct areas. For this and other reasons they are believed to mark later migration waves into Europe which did not cover the entire continent.

The haplogroup N1c (N3 or TAT, according to previous nomenclatures) is restricted to north-east Europe (mainly Finnic speakers) and Siberia. During the last decade it remained unclear whether it marks an eastward migration from Europe or the opposite westward migration trend. In 2007 Rootsi and colleagues have shown that this haplogroups could be deeply rooted in East Asian phylogeny and therefore the occurrence of this haplogroup in Europe may be

attributed to the Asian influence. Authors supposed step-by-step migration from North China to Eastern Europe, which started in early Holocene and underwent a secondary expansion on its long way. Derenko and colleagues (2007) studied microsatellite variation associated with this haplogroup in more detail and tried to estimate its age. They identified two variants, one of which migrated into Europe 6-10 ky ago, while the second (less frequent) variant was shown to come by the way of a smaller and more recent migration, 2-4 ky ago. Although these time estimations should be taken with great caution, the both studies (Rootsi et al., 2007; Derenko et al., 2007) agree that north-east Europe had a significant (or even predominant) genetic legacy in South Siberian/Central Asian populations.

This creates a problem for "two systems" approach, because from mitochondrial perspective Siberian/East Asian haplogroups appeared in low frequencies and only in the eastern edge of Europe and did not account for a significant portion of the gene pool anywhere else in Europe. (When low frequency of typical East Asian haplogroup F was found on Croatian isles and in even lower frequencies on Croatian mainland, this was considered as a paradox and a special paper (Tolk et al., 2001) tried to explain it by possible medieval gene flow caused by trade routes of Venice). That is why a significant Asian presence in Europe, concluded from haplogroup N1c remains one of the main inconsistencies between Y chromosomal and mitochondrial genetic systems. From our point of view, this problem could be resolved if one takes into consideration the fact that genetic boundary between Europe and Asia lies much eastern than Ural Mountains. The western Central Asia (the Altai Mountains in particular) could be therefore considered as a genetically intermediary in present time and primary "European" zone in the past. This view explains why Y chromosomal haplogroup N (despite its origin in East Asia 20-30 ky ago) in pre Neolithic or Neolithic times could be the characteristic haplogroup for Caucasoid populations in Eurasian steppe west from the Altai and also for Mongoloid populations east from the Altai. From the western part of this area the haplogroup N1c could spread northward and north-westward by a number of migrations suggested for this area. This view also explains why these migrations did not bring East Eurasian mitochondrial haplogroups into Europe: the source area having mainly Western Eurasian haplogroups even in the contemporary gene pool. It was more the case in earlier times before Turkic speakers brought East Eurasian haplogroups by

their expansion into this area two millennia ago onward.

Two last Y chromosomal haplogroups to be discussed are E and J. They predominate in North Africa and Near East, and in Europe they are found mainly in Mediterranean area. Not all sub-branches of these two haplogroups reached Europe, but mainly one branch of haplogroup E (namely, E-V13) and two branches of haplogroup J (J-M241 and J-410).

While J-410 follows the separate pattern, the other two haplogroups (and also haplogroup I2a, mentioned above) are concentrated in the Balkans and have not been found in neighbouring regions with any significant frequencies. The ages of these haplogroups estimated from their STR diversity are: E-V13 from 4 to 7,5 ky; J-M241 from 3,5 to 6 ky; I2a from 5,5 and 10 ky (Battaglia et al., 2008). This roughly coincides with time of Neolithic transition in this part of Europe. The model suggested by Battaglia and colleagues states that Neolithic cultural package was adopted by local Mesolithic populations of the Balkans which started grow in numbers, expanding farming across the entire Balkan peninsula and later transmitting this package to other Mesolithic populations of Europe. This model explains why these three haplogroups are restricted to the Balkans, why they exhibit decreasing frequency towards other part of Europe and why their age is similar to that of the Neolithic transition.

However, the internal logic of this model is opposite to those applied by Richards and colleagues in relation to mitochondrial DNA. Indeed, Richards and colleagues proved that mitochondrial haplogroups whose diversity was accumulated in Europe in situ are of Palaeolithic age; and from this fact they concluded that present-day Europeans are descendants of Palaeolithic population of Europe (Richards et al., 2000). Eight years later, Battaglia and colleagues proved that Y chromosomal haplogroups whose diversity was also accumulated in Europe in situ are of Neolithic age; but from this contrasting fact they concluded also that present day Europeans are descendants of Palaeolithic population of Europe (Battaglia et al., 2008). Both studies are reasonably substantiated and their conclusions look correct. However this example illustrates that genetic studies need more robust and universal logic, at least when dealing with such complex process like the Neolithisation. In this particular case the possible logical compromise lays in the fact that concept of Battaglia and co-authors actually implies both, cultural diffusion and demic diffusion models. Although the au-

thors did not formulate this explicitly, their concept implies, that cultural diffusion took place between regions (Anatolia and Balkans, Balkans and Central Mediterranean) while the demic diffusion occurred within regions.

### Ancient DNA

Analysis of ancient DNA (aDNA) provides direct data on the former European gene pool, which are free from assumptions and speculations which often accompany deductions of past genetic processes, based on the contemporary genetic pattern. This advantage of aDNA may trigger a revolution in population genetics and if this did not happen so far, this was due to limited quality and quantity of available aDNA evidence.

Problems with the quality (authenticity) of aDNA data are dramatic because of the possible contamination by modern DNA. For this reason some aDNA results may be false, and many early aDNA papers were criticized exactly from this point of view. This problem could be partially resolved in few high-standard laboratories only, which have special equipment to minimize risk of contamination. Cross-checking, i.e. independent analysis of the same ancient sample in different aDNA labs is the second condition. The third one is implementing the "modern DNA free" style of excavation into the practice of the archaeological fieldwork. Having these three conditions met, one can reach reasonable degree of authenticity of the aDNA results.

The quantity problem consists in the scarcity and limited sample sizes of the aDNA data. Again, this problem could be solved only partially, by increasing the number of aDNA studies and average sample size per study. Fortunately, both factors tended to increase in the last decade, still trace amounts and high fragmentation of ancient DNA samples hinder its high-throughput analysis.

Because of these limitations aDNA at least presently cannot be the main source of genetic knowledge about the Neolithisation. But it is already one of the important sources on this problem. Indeed, analyses of Neandertal mitochondrial DNA (Krings et al., 1997; Ovchinnikov et al., 2000), though being criticized for probable mistakes in sequencing, put an end to a lengthy discussion of the possible assimilation of the Neandertal populations by anatomically modern humans. Specificity of Neandertal mitochondrial type (Currat, Excoffier, 2004) and absence of this type in present day Europeans (Behar et al., 2007) allow to root European gene

pool in AMH colonization of the Europe, disregarding the previous epochs.

Direct genetic data on first Neolithic groups in Europe are of course most promising source for choosing between the demic and cultural diffusion models of Neolithisation. Such data are now available for Neolithic population of the Iberian peninsula (Sampietro et al., 2007) and Neolithic population of the Central Europe (sites of Linear Band Ceramic with age of 7.0 – 7.5 ky; Haak et al., 2005). Iberian Neolithic population was shown to be genetically similar to the present day Iberian population. In contrast, LBK population in Central Europe was shown to be genetically distinct from the present day population of that or any other region of Europe). The most remarkable feature of the Neolithic population was mitochondrial haplogroup N1a found in 6 of 24 individuals. This haplogroup is virtually absent in present-day Europe. The mathematical simulation has shown that if this Neolithic population was source of present-day Europeans they could not have lost this haplogroups by stochastic genetic drift.

It was therefore concluded (Haak et al., 2005) that Neolithic LBK population did not become parental for present-day European gene pool, but became dissolved in pre-existing European populations. This conclusion is therefore in agreement with cultural diffusion model in assuming that since Neolithic farmers arrived in Europe, the farming was adopted by aboriginal populations and first farmers did not leave any considerable genetic legacy in their new homeland.

The study by Haak and colleagues did answer the question: "what happened with first farmers after their arrival in Europe". To address the another question, "where these first farmers came from", the consequent study was performed (Haak, pers. comm.). Based on extended dataset (44 individual mtDNAs from different sites of early LBK culture) it was found that this population is genetically similar to present day populations of Northern Mesopotamia, southern Caucasus and eastern Anatolia. Although the genetic composition of this area could be disturbed after the Neolithic period by subsequent migrations, it is reasonable to suppose that inner areas of the Near East were homeland for migrating groups who finally brought these mitochondrial lineages into the LBK population of the Central Europe.

Of course, this data give rise to many new questions, and currently available aDNA data are not sufficient to address them. The moderate optimism is based on increasing number

and quality of aDNA data which might allow better chronological and geographical resolution of genetic processes in the near future.

## Conclusions

The increasingly accumulating genetic data on extant and extinct (aDNA) European populations are most frequently discussed in terms of two opposite concepts: demic diffusion and cultural diffusion models of Neolithisation. In hands of Cavalli-Sforza and his colleagues the genetic mirror reflected Neolithic expansion across Europe (demic diffusion); but in hands of present-day writers this mirror reflects mainly the Palaeolithic legacy of Europeans and cultural diffusion model is needed to explain spread of farming.

Understanding the genetic history of Europe implies clarifying relative significance and patterns of each of the following processes:

1. the initial dispersal of AMH in Europe (Upper Palaeolithic);
2. the restructuring of the genetic landscape during the Mesolithic repopulation of the Europe from two-four refugia;
3. the importance of the Neolithic expansion viewed as the spread of early farming communities or spread of Neolithic cultural package;
4. the role of post-Neolithic human movements within Europe;
5. the "oriental" influence in different epochs – from Palaeolithic to Medieval times.

To address these questions population genetics operated with autosomal (classical) markers in the past and autosomal (DNA) markers may become the new standard in the future, while the present day studies are based on mitochondrial DNA and Y chromosomal variation.

Analysis of mtDNA demonstrated that most of European haplogroups came from the Near East during the Upper Palaeolithic times and Neolithic migration of Near Eastern farmers did not contribute much into the European gene pool. The south-east – northwest cline within Europe, as established by many genetic markers, is not considered anymore as the trace of Neolithic expansion, because Palaeolithic colonists used likely the same geographical route.

Y chromosomal data reveal distinct domains of prehistoric movements within Europe. Particularly, two different haplogroups predominate in Western versus Eastern Europe, and one may speculate about two secondary homelands, associating them with Mesolithic refugia or centres of later expansions.

Southeast Europe (the Balkans) which deserves special attention as gates into Europe, is populated by three different Y chromosomal haplogroups exhibiting similar patterns: being autochthonous for Europe these haplogroups started to expand in time frames comparable with the Neolithisation; it was supposed that this expansion might took the form of Balkan's Mesolithic population adopting farming from their Anatolian neighbours.

Analysis of ancient DNA indicated that first Central European farmers (LBK) were of Near Eastern origin but did not left recognisable descendants. The early farmers in Iberia (and possible in other areas of late Neolithisation) were of aboriginal European genetic type.

Genetic mirror shows a controversial picture: even in this summary "indigenous" Balkan populations adopted farming without immigrant farmers, but "immigrant" gene pool was found in first farmers even north of Balkans (in Central Europe). Nevertheless most lines of reasoning show that Neolithisation did not change drastically the European gene pool and consequently did not involve large-scale population movements. Since, if one would like to obtain further information about these (relatively minor) movements from genetic data it is necessary to be equipped with a large genetic databases and a good dose of scepticism not to rush to conclusions.

## References

- Ammerman A.J. and L.L Cavalli-Sforza (1984) *Neolithic Transition and the Genetics of Populations in Europe*, Princeton. N.J.: Princeton University Press.
- Balanovskaya E.V. and S.D Nurbaev (1997a) "A computer-aided methods for gene-geographic study of a population gene pool in space of principal components", *Genetika (Moscow)* Vol 33: 1456-1469.
- Balanovskaya E.V. and S.D Nurbaev (1997b) "The selective structure of gene pool: possibilities of studying", *Genetika (Moscow)* Vol. 33: 1572-1588.
- Balanovsky O, S. Rootsi, A. Pshenichnov, T. Kivisild, M. Churnosov, I. Evseeva, E. Pocheshkhova, M. Boldyreva, N. Yankovsky, E. Balanovska and R. Villems R (2008) "Two sources of the Russian patrilineal heritage in their Eurasian context" *American Journal of Human Genetics* Vol. 82(1):236-50.
- Balanovsky O. (2008) "Gene pool of the high latitudes in Eurasia or where Saami came from?", in *Way to North: the environments and the earliest inhabitants of Arctic and Subarctic*. Moscow, Institut of Geography. P. 277-282 (in Russian).
- Barbujani G, G. Bertorelle and L.Chikhi (1998) "Evidence for Paleolithic and Neolithic gene flow in Europe", *American Journal of Human Genetics* Vol 62(2):488-92.
- Barbujani G. and G. Bertorelle (2001) "Genetics and the population history of Europe", *Proceedings of the National Academy of Sciences USA* Vol 98(1):22-5.
- Battaglia V, S. Fornarino, N. Al-Zahery, A. Olivieri, M.Pala, M.N. Myres, R.J. King, S.Rootsi, D. Marjanovic, D.Primorac, R.HadziSelimovic, S.Vidovic, K.Drobic, N.Durmishi, A.Torroni, A.S. Santachiara-Benerecetti, P.A. Underhill and O. Semino (2008) "Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe", *European Journal of Human Genetics Advance online publication* 24 December 2008; doi: 10.1038/ejhg.2008.249
- Behar D.M, S. Rosset, J. Blue-Smith, O. Balanovsky, S.Tzur, D. Comas, R.J. Mitchell, L. Quintana-Murci, C.Tyler-Smith, R.S. Wells and Genographic Consortium. (2007) "The Genographic Project public participation mitochondrial DNA database" *PLoS Genetics* Vol 3(6):e104.
- Cavalli-Sforza L.L., P. Menozzi and A. Piazza A. (1994) *History and Geography of Human Genes*. Princeton: Princeton University Press. 1059 p.
- Chikhi L, R.A. Nichols, G. Barbujani and M.A. Beaumont (2002) "Y genetic data support the Neolithic demic diffusion model", *Proceedings of the National Academy of Sciences USA* Vol 99(17):11008-13.
- Childe V.G. 1928. *The Dawn of European Civilization*. New York, Knopf
- Curat M, and L. Excoffier (2004) "Modern humans did not admix with Neanderthals during their range expansion into Europe" *PLoS Biology* Vol 2(12):e421.
- Derenko M, B. Malyarchuk, G. Denisova, M. Wozniak, T. Grzybowski, I.Dambueva, and I. Zakharov (2007) "Y-chromosome haplogroup N dispersals from south Siberia to Europe", *Journal of Human Genetics*. Vol 52(9):763-70.
- Haak W, P. Forster, B. Bramanti, S. Matsumura, G. Brandt, M. Tänzer, R. Villems, C. Renfrew, D. Gronenborn, Alt, K.W. and J. Burger (2005) "Ancient DNA from the first European farmers in 7500-year-old Neolithic sites", *Science* Vol 310(5750):1016-8.
- Helgason A, S. Sigurethardóttir, J. Nicholson, B. Sykes, E.W. Hill, D.G. Bradley, V. Bosnes, J.R. Gulcher, R. Ward, and K. Stefánsson (2000) "Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland", *American Journal of Human Genetics* Vol 67(3):697-717.
- Krings M, A. Stone, R.W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo (1997) "Neandertal DNA sequences and the origin of modern humans", *Vol Cell* 90(1):19-30.
- Lewontin R.C., and J. Krakauer J. (1975) "Testing the heterogeneity of F values", *Genetics*. Vol 80: 397-398.
- Menozzi P,A. Piazza A and L.L. Cavalli-Sforza L.L. (1978) "Synthetic maps of human gene frequencies in Europe", *Science* Vol 201: 786-792.
- Novembre J, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, M Stephens, and C.D. Bustamante (2008) "Genes mirror geography within Europe", *Nature* Vol 456(7218):98-101.
- Ovchinnikov IV, Götherström A, Romanova GP, Kharitonov VM, Lidén K, Goodwin W. (2000) "Molecular analysis of Neanderthal DNA from the northern Caucasus", *Nature* 404(6777):490-3.
- Richards M, H. Côte-Real, P. Forster, V. Macaulay, H. Wilkinson-Herbots, A. Demaine, S. Papiha, R. Hedges, H.J. Bandelt, Sykes B. (1996) "Paleolithic and neolithic lineages in the European mitochondrial gene pool", *American Journal of Human Genetics*. 59(1):185-203.
- Richards M, Macaulay V, Hickey E, Vega E, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, G. Stefanescu, J. Hatina, M. Belledi, A. Di Rienzo, A. Novelletto, A. Oppenheim, S. Norby, N. Al-Zaheri, S. Santachiara-Benerecetti, R. Scozari, A. Torroni, and H.J. Bandelt (2000) "Tracing European founder lineages

- in the Near Eastern mtDNA pool", *American Journal of Human Genetics* Vol 67(5):1251-76.
- Richards M, Macaulay V, A. Torroni, and H.J. Bandelt. (2002) "In search of geographical patterns in European mitochondrial DNA", *American Journal of Human Genetics* Vol 71(5):1168-74.
- Roewer L, P.J. Croucher, S. Willuweit, T.T. Lu, M. Kayser, R. Lessig, P. de Knijff, M.A. Jobling, C. Tyler-Smith and M. Krawczak M. (2005) "Signature of recent historical events in the European Y-chromosomal STR haplotype distribution", *Hum Genet.* 116(4):279-91.
- Rootsi S, L.A. Zhivotovsky, M. Baldovic, M. Kayser, I.A. Kutuev, R.Khusainova, M.A. Bermisheva, Gubina, M. S.A. Fedorova, A.M. Ilumäe, E.K. Khusnutdinova, M.I. Voevoda, L.P. Osipova, M. Stoneking, A.A. Lin, V. Ferak, J. Parik, T. Kivisild, P.A. Underhill, and R. Villems. (2007) "A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe", *European Journal of Human Genetics* Vol 15(2):204-11.
- Rosser Z.H., T. M.E. Zerjal, M Hurler, Adojaan, D. Alavantic, A. Amorim, W. Amos, M. Armenteros, E. Arroyo, G. Barbujani, L. Beckman, J. Beckman, E Bertranpetit, D.G. Bosch, G. Bradley, H.B. Brede, Cooper, P. Côte-Real, R. de Knijff, Decorte, Y.E. Dubrova, O. Evgrafov, A. Gilissen, S. Glisic, M. E.W. Gölge, Hill, A Jeziorowska, L. Kalaydjieva, M. Kayser, T. Kivisild, S.A. K. Kravchenko, A. Krumina, V. Kucinskas, J. Lavinha, L.A. Livshits, P. Malaspina, S. Maria, McElreavey, T.A. Meitinger, A.V. Mikelsaar, R.J. Mitchell, K. Nafa, J. Nicholson, S. Nørby, A. Pandya, J. Parik, P.C. Patsalis, L. Pereira, B. Peterlin, G. Pielberg, M.J. Prata, C. Previderé, L. Roewer, S. Rootsi, D.C. Rubinsztein, J. Saillard, F.R. Santos, G. Stefanescu, B.C. Sykes, A. Tolun, R. Villems, C. Tyler-Smith, and M.A. Jobling. (2000) "Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language", *American Journal of Human Genetics* Vol 67(6):1526-43.
- Rychkov Yu.G., and E.V. Balanovskaya. (1992) "Genetic geography of USSR population", *Genetika (Moscow)* Vol . 28. (1) : 52-75.
- Sampietro M.L., O. Lao, D. Caramelli, M. Lari, R. Pou, M. Marti, J. Bertranpetit, and C. Lalueza-Fox (2007) "Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe", *Vol Proc Biol Sci.* Vol 274(1622):2161-2167.
- Semino O, G. Passarino, P.J. Oefner, A.A. Lin, S. Arbuzova, L.E. Beckman, G. De Benedictis, P. Francalacci, A. Kouvatsi, S. Limborska, M. Marcikiae, A. Mika, B. Mika, D. Primorac, A.S. Santachiara-Benerecetti, L.L. Cavalli-Sforza, and P.A. Underhill (2000) "The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective", *Science* Vol 290(5494):1155-1159.
- Simoni L, F. Calafell, D. Pettener, J. Bertranpetit, and G. Barbujani (2000) "Geographic patterns of mtDNA diversity in Europe", *American Journal of Human Genetics* Vol 66(1):262-278.
- Sokal R.R., and N.L Oden. (1978) "Spatial autocorrelation in biology: I. Methodology", *Biological Journal of Linnean Society* Vol. 10. P. 199-228.
- Sokal R.R., N.L Oden, and B.A. Thomson (1999) "A problem with synthetic maps", *Human Biology* Vol 71 (1): 1-13.
- Tolk H.V., L. Barac, M. Pericic, I.M. Klaric, B. Janicijevic, H. Campbell, I. Rudan, T. Kivisild, R. Villems, and P. Rudan (2001) "The evidence of mtDNA haplogroup F in a European population and its ethnohistoric implications", *European Journal of Human Genetics* Vol 9(9):717-23.
- Torroni A, H.J. Bandelt, V. Macaulay, M. Richards, M. Cruciani, C. Rengo, V. Martinez-Cabrera, R. Villems, T. Kivisild, E. Metspalu, J. Parik, H.V. Tolk, K. Tambets, P. Forster, B. Karger, P. Francalacci, P. Rudan, B. Janicijevic, O. Rickards, M.L. Savontaus, K. Huoponen, V. Laitinen, S. Koivumäki, B. Sykes, E. Hickey, A. Novelletto, P. Moral, D. Sellitto, A. Coppa, N. Al-Zaheri, A.S. Santachiara-Benerecetti, O. Semino, and R. Scozzari (2001) "A signal, from human mtDNA, of postglacial recolonization in Europe", *American Journal of Human Genetics* Vol 69(4):844-52.
- Yamazaki T. and T. Maryama (1973) "Evidence that enzyme polymorphisms are selectively neutral", *Nature* Vol 245: 140-141.

# Geographic distribution and molecular evolution of ancestral Y-chromosome haplotypes in the Low Countries

Gerhard Mertens and  
Hugo Goossens

## Abstract

By means of a sample of 225 men with origins in the Low Countries we describe the regional Y-chromosomal differences in this area of West Europe. Haplogroups, 10-marker haplotypes and geographic location were retrieved from genealogical websites. Data were analyzed by freely available population genetics software. This showed generally insignificant genetic distances between the populations in the different regions of the Low Countries, corresponding to the very limited geographic barriers to migration. A small but significant genetic difference could be demonstrated between populations on different sides of the Germanic-Latin language border which runs through the Low Countries. Comparison of the molecular structure of haplotypes revealed quasi absence of migration for thousands of years for certain paternal lineages in the region of Brabant.

## Introduction

The Low Countries is the historical name for countries on low-lying land around the delta of the Rhine, Scheldt and Meuse rivers that has been called geopolitically BENELUX after the Second World War (Fig. 1). From the genealogical perspective, we however prefer the former nomination, which will be used throughout the paper.

The genetic and genealogical connections between the Low Countries and Russia may seem limited at first. Still, with Napoleons Grande Armée, about 25000 soldiers of the Low Countries, then part of the French Empire, marched to Moscow in 1812 (Zamoyski, 2004). Only 1000 returned home while probably most of them perished, but a number may have survived as prisoners and left their Y-chromosome in Russian progeny.

In our paper, we will give a description of Y chromosomal haplotypes of men with documented ancestry in the Low Countries and study

the relation between these haplotypes at the level of DNA structure, a process called molecular evolution. While many genetic genealogy population studies focus on (Y-chromosomal) haplogroups, we will concentrate on haplotypes of short tandem repeats (STR).

We will extend on the recent paper in the Journal of Genetic Genealogy (Deboeck, 2008) on the Flemish, who indeed are a major population of the Low Countries but certainly not its only constituent.

## Material and methods

**Dataset.** Y-chromosomal data posted on genealogical websites were retrieved. These were Ysearch (most haplotypes), Ybase and the BeNeLux and Flanders-Flemish Projects from Family Tree DNA. Only Y-chromosomal haplotypes from men with a most distant male ancestor with a recorded place of birth in the Low Countries were retained. The year of birth of these ancestors varied between 1200 and 1922, with 75% of cases before 1739.

As expected, not all Y-chromosomes deposited on different websites had been typed for



**Figure 1.** Situation of BENELUX countries (red) in West Europe

the same markers. Between 9 and 74 STRs had been tested. To retain a statistically significant number of individuals from different geographic regions in the Low Countries, we included Y-chromosomes for which the following 10 Y-STRs were available: DYS19, DYS385 a, DYS385 b, DYS389 I, DYS389 II, DYS390, DYS391, DYS392, DYS393 and DYS439. If we would have required more markers, then the number of individuals that could be included would seriously drop, precluding the possibility to make relevant comparisons within the Low Countries.

If the corresponding Y-haplogroup was mentioned on the website, it was recorded. Alternatively, it was deduced from the Y-STRs using Whit Athey's Haplogroup Predictor.

**Geographical regions.** In order to be included in the study, the place of birth of the old-

est known ancestor had to be located in the territory of present day Belgium, the Netherlands, Luxemburg or the "Nord" Department of France. Indeed, the latter is the area which is also called French Flanders, an originally Dutch speaking region that was part of the medieval County of Flanders until it was annexed to France in 1678. Though today few speakers of the regional Flemish dialect remain, toponyms and patronyms still unmistakably prove the Flemish origin of the region. An examples of the former is the famous town of Dunkirk (Duinkerke), while the surname of France's World War II great patriot and later president general De Gaulle was originally Vandewalle - a pure Flemish name - which was changed to the (very) French sounding De Gaulle.

The places of birth found on the websites were ordered into one of the presently existing

provinces. Because an insufficient number of persons could be found for several provinces, provinces were combined to larger "regions" with the objective to permit meaningful comparisons between the Y-haplotypes in these regions. Provinces were united to "regions" according to geographical, historical and cultural logic, as follows:

Friesland, Groningen, Drente	"North Provinces"
Gelderland, Overijssel	"East Provinces"
Noord-Holland, Zuid-Holland, Utrecht	"Holland"
Noord-Brabant, Antwerpen, Brabant	"Brabant"
Nederlands Limburg, Belgisch Limburg	"Limburg"
Oost-Vlaanderen, West-Vlaanderen, Zeeland, Nord	"Flanders"
Liege, Luxembourg, Namur, Hainaut	"Wallonia"

Dutch is the language of all regions of the Low Countries, except for the region of Wallonia, where French is spoken.

Figure 2 illustrates the location of the actual provinces and the combined "regions" we will use in the rest of the study. As previously mentioned, we have enlarged the area of "FLANDERS" (yellow) with French Flanders which belongs to the territory of present day France.

**Statistical analysis.** Allelic frequencies of the 10 Y-STRs were determined by simple counting. As a measure for the genetic distance between the populations in the different regions of the Low Countries, we used the parameter FST. This was calculated using Arlequin, a software for population genetics data analysis by Laurent Excoffier of the University of Geneva. It can be freely downloaded, a manual included. Data input is done by a .txt file which can be simply saved from the Excel table with the haplotypes of Y-STRs of each individual. The program's "calculation settings" were set to "genetic structure", "population comparisons" and "compute pairwise FST".

To analyze the relation between the different regional populations, a Neighbor Joining method phylogenetic tree was constructed using Neighbor. This is one of the programs of Phylip, a free package of programs for inferring phylogenies from Joe Felsenstein from the University of Washington. The input for the calculations by Neighbor was the matrix of FST genetic



Figure 2. Present day provinces (above), combined to 7 larger "regions" (below)

distances obtained from the Arlequin analysis. Subsequently, an unrooted phylogenetic tree was plotted with the output file of the Neighbor analysis by means of Drawtree, another program of the Phylip package.

The molecular evolution of the different haplotypes was analyzed with Network, which is a software - also available for free - to generate evolutionary trees from genetic, linguistic and other data; this one from the Fluxus Technology company. All haplotypes of Y-STRs have to be entered manually, then the network is calculated and finally the tree is drawn.

## Results and discussion

**Haplogroups.** Table 1 gives the 10 Y-STR haplotypes of 225 men with ancestors in the Low Countries, as well as their corresponding haplogroup. For each haplotype, the province and region of its bearer are mentioned.

The haplogroup frequencies for the Low Countries as a whole were 57% haplogroup R1b, 16% haplogroup I, 5% haplogroup R1a, 6% haplogroup E, 6% haplogroup J, 8% haplogroup G and less than 1% (1 individual) for the haplogroups L, A, K and N. It may seem bizarre to observe a haplogroup A - the father of all human haplogroups, and originating in central Africa - in a man of supposed Low Country origin, but this is not so surprising since this haplogroup has also been found in men from Yorkshire (King, 2007), illustrating the complexity of human migration patterns.

The haplogroup frequencies in this Low Countries population sample of 225 correspond well with the data presented by Deboeck (2008) in his study on 228 Flemish (57% R1b, 15% I, 4% R1a, 5% E, 6% J, 4% G). The frequency distribution of the Low Countries can be situated in between Germany and Great Britain, as geo-

graphically the Low Countries are located between their east and west neighbors. It is also noteworthy that according to linguistics, Dutch language has an intermediate position between German and English. All the former nicely illustrates the "genes, people and languages" concept, formulated by the father of human population genetics, Luigi Cavalli-Sforza (1997) of Princeton University.

**Allelic frequencies.** Histograms of allelic frequencies for the 10 typed Y-STRs in the 7 regions of the Low Countries are given in Figure 3. For the purpose of comparison, we included a population of previously described (Mertens, 2007) African residents of Belgium. These people have their roots in Congo, Nigeria and Ghana.

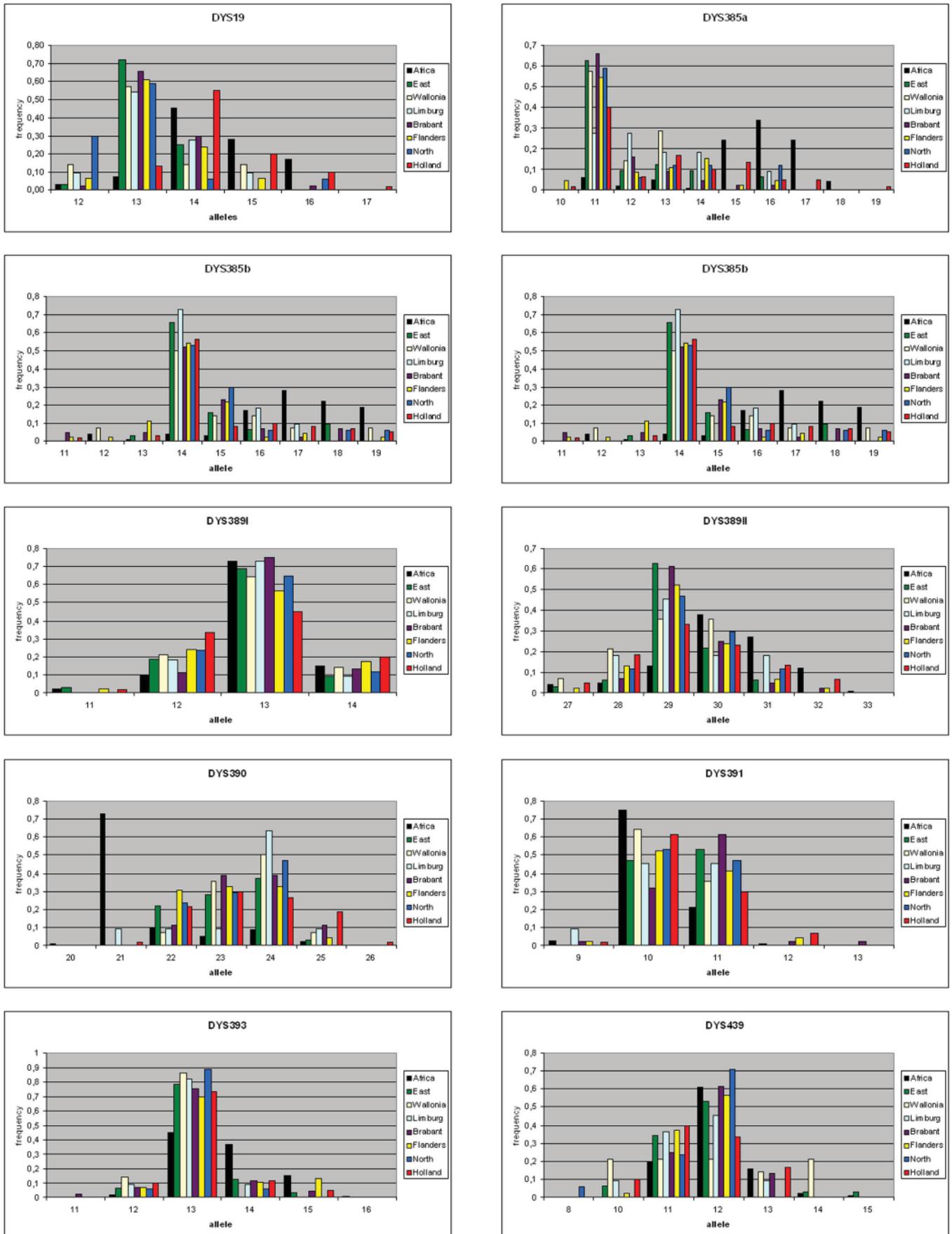
As can be expected from the common origin of all members of the Homo sapiens species, most alleles are observed in most populations. Some alleles are however significantly more frequent in one than in another population - or better "metapopulation" (a group of populations of the same species which interact at some level) - and can be considered ancestry informative markers. So we see that allele 13 of DYS19, allele 11 of DYS385a and allele 14 of DYS385b are rather specific for the European metapopulation, while allele 21 of DYS390 and allele 11 of DYS392 are typical of the African metapopulation.

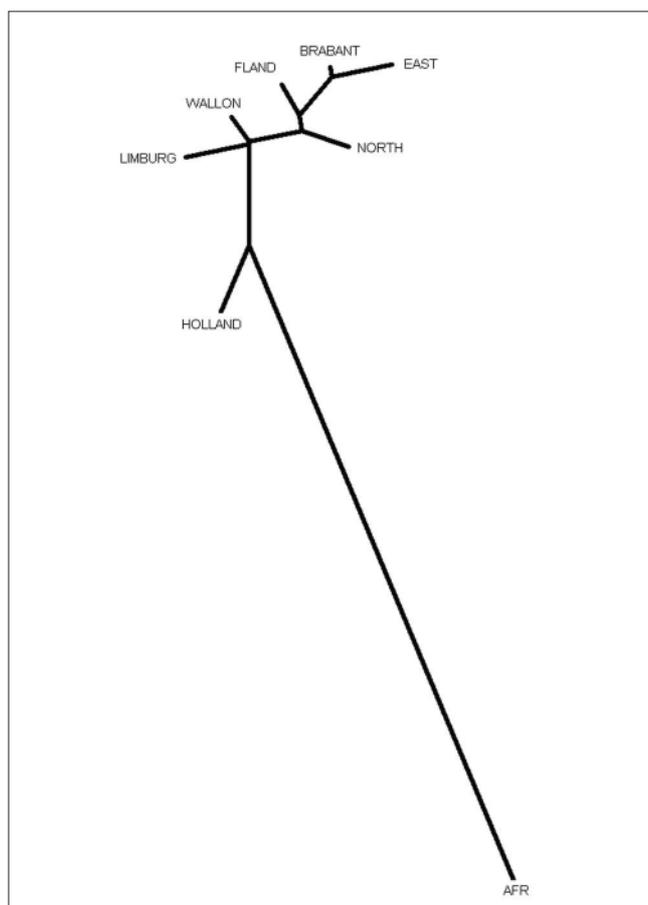
**Genetic distances.** Table 2 is the matrix of pairwise  $F_{ST}$  distances. Each number in the matrix is a relative measure for the genetic difference between a pair of populations. This so-called  $F_{ST}$  genetic distance is calculated from the allelic frequencies of the 10 tested Y-chromosomal STRs. Similar distributions of allelic frequencies of a pair of populations result in a

	1	2	3	4	5	6	7	8
1. Africans	0							
2. East Provinces	0.23855	0						
3. Wallonia	0.20205	0.00621	0					
4. Limburg	0.18039	-0.01024	-0.02334	0				
5. Brabant	0.24588	-0.01311	0.02447	0.01374	0			
6. Holland	0.15904	0.02688	-0.00499	-0.00183	0.05338	0		
7. Flanders	0.20191	-0.01127	0.00497	-0.00491	0.00572	0.00982	0	
8. North Provinces	0.20300	-0.01249	-0.00817	-0.01333	0.00237	0.01223	-0.01478	0

**Table 2.** Matrix of  $F_{ST}$  genetic distances

**Figure 3.** Allelic frequencies of 10 Y-chromosomal STRs in the Low Countries and in Africans





**Figure 4.** Neighbor Joining method phylogenetic tree

smaller genetic distance between these populations. This is based on the population genetic concept of migration or gene flow. It implies that increased migration between populations, including reproduction within the receiving population, causes differences in allelic distribution between populations to decrease. In lay terms, this refers to admixing or assimilation of populations. Generally, genetic distances correlate with physical distances between populations. Indeed, the farther populations live apart, the more impractical it becomes to exchange individuals for procreation.

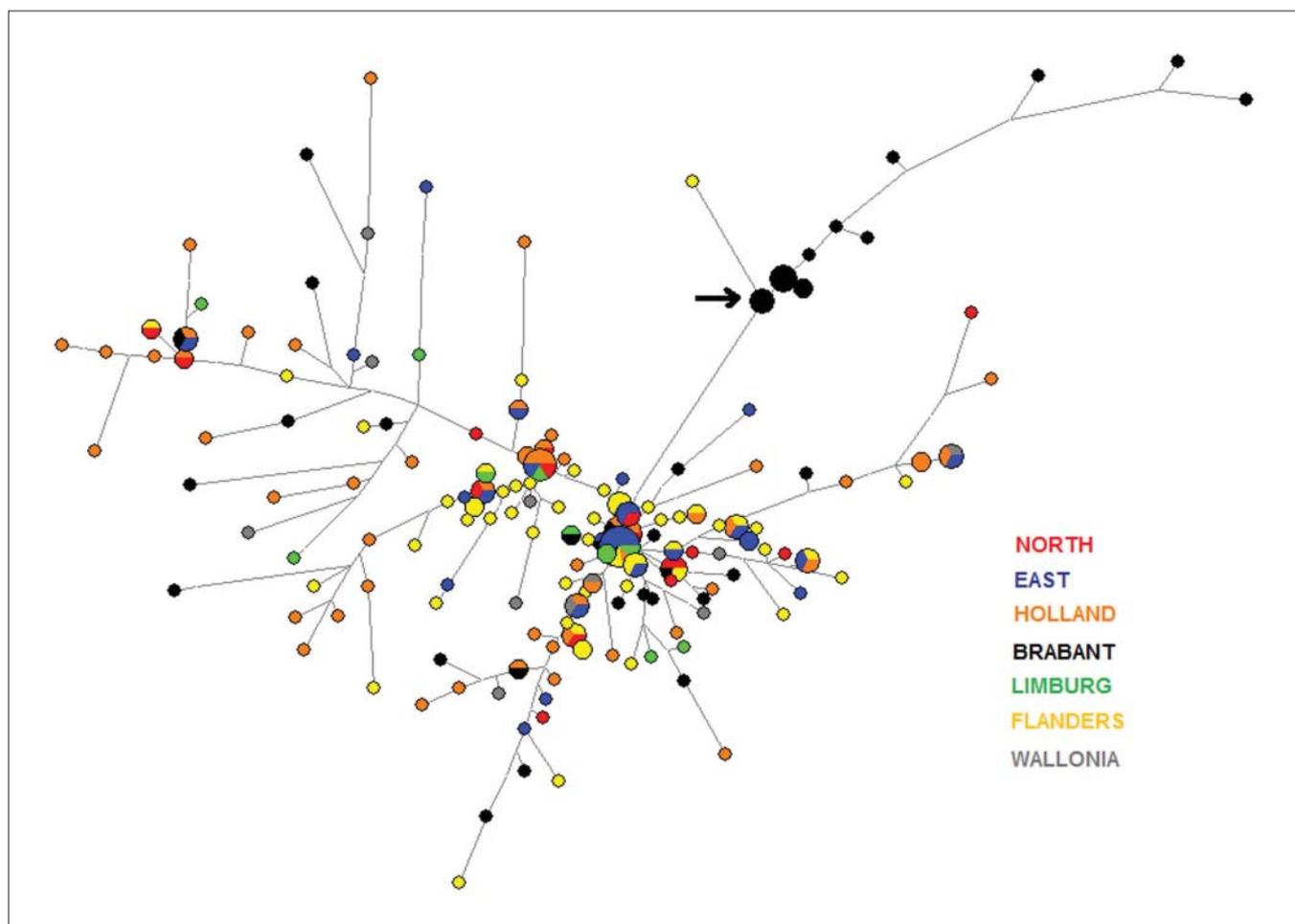
The matrix should be read as follows: the genetic distance between Africans and Brabant is 0.24588; between East Provinces and Holland it is 0.02688, etc.

The largest genetic distances are observed versus the population of Africans. This is not surprising in view of the large geographic distance between Africa and Holland, which is clearly larger than from Flanders to Holland. Compared with this African – Low Countries genetic distance, the distance between the different Low Countries populations are small, even insignificantly small in a number of cases. Indeed, since

a negative value for a genetic distance – as is obtained for 11 out of 28 of the pairwise comparisons – is a physical impossibility (such as a negative value for the weight of an object), the genetic distance between several Low Countries populations is virtually non-existent. On the other hand, it can be noted that the genetic distance between populations on different sides of the Germanic-Romanic language border – Brabant versus Wallonia and Flanders versus Wallonia – has a positive value. This corresponds to the fact that not only geographic distance may limit admixture between populations, but also cultural elements including language, can form a relative barrier between people for mating and reproduction.

**Phylogenetic tree (Phylip).** If the matrix of genetic distances (Table 2) is further analyzed using the Neighbor Joining methodology, the phylogenetic tree of Figure 4 is obtained. Branch lengths are directly related to the genetic distance between populations, which in turn are a measure for dissimilarity between allelic frequency distributions. It graphically shows that, compared with the branch length to Africans, the Low Countries populations are closely related.

**Tree of molecular evolution.** Figure 5 is a tree produced with Network software, showing the relation between the different Y-chromosomal haplotypes observed in the sample of 225 men with ancestors in the Low Countries. As opposed to the distance matrix (Table 2) and the Neighbor Joining tree (Figure 4), it is not based on frequencies of alleles of individual Y-STRs. The only element in the tree referring to frequencies, is the area of each circle, corresponding to the number of times a certain 10-marker-haplotype was observed. The smallest circles are haplotypes which occurred once in the examined sample of 225 men. The Network tree is based on the molecular mechanism for formation of new alleles and thus haplotypes. This is the so-called "replication slippage model" where, with the frequency of a molecular clock, every 500 to 1000 generations, the number of repeats of a Y-chromosomal STR changes by a slip of the DNA transcription system. This implies that a 14 repeat unit of the DYS19 marker can evolve or change by mutation (loss or gain of 1 repeat unit) to a 13 or 15 repeat unit. In the tree, branch lengths are proportional to the structural differences (number of repeat unit differences for all 10 Y-STRs of the 10-marker-haplotypes) between haplotypes. Median vec-



**Figure 5.** Network method tree of molecular evolution

tors - haplotypes theoretically expected in the tree to explain the transition from one haplotype to the other - were not drawn because otherwise the tree would become too cluttered.

By using colors corresponding to Low Country regions, an element for geographic analysis was introduced in the tree.

Though a root for the tree or an ancestral haplotype for the Low Countries, cannot be demonstrated, there are some inferences to be made. Generally colors appear evenly dispersed over the tree, which correlates with extensive migration between the regions. There are indeed few geographic barriers in the Low Countries limiting or opposing migration. A circle at the periphery of a branch implies that the haplotype is of more recent origin than a circle closer to the center of the tree. Though certainly not very clear-cut, the orange circles are generally situated somewhat more in the periphery of the tree, while the yellow circles are located more centrally. This means that the region of Holland has more "recent", "derived" haplotypes while Flanders is populated by more "ancient", "origi-

nal" haplotypes. This might be explained by a somewhat more recent human settlement in Holland than in Flanders. This is consistent with a repopulation after the last Ice Age of this part of north west Europe starting from the Iberian peninsula, i.e. the south, Flanders lying south of Holland. It is also consistent with the fact that Holland is literally the lowest part of the Low Countries, with large parts below sea level and with frequent flooding. Consequently, it has taken Holland longer than Flanders to reach the same density of population. The upper right branch of the tree, with 10 black circles connected without interposition of circles of other colors, is also of interest. The haplotype with the black arrow is the ancestral haplotype of the 10 haplotypes that have evolved from this haplotype. It is remarkable that out of these 10 descending haplotypes 9 still have a place of residence in the region of Brabant (one - the yellow circle - having migrated to the neighboring region of Flanders). It should equally be realized that each circle may represent one haplotype, but that each haplotype corresponds to a

complete clan of family members with the same paternal lineage. Furthermore, the time scale of formation of the 10 subsequent haplotypes from the first (arrow) is of truly evolutionary magnitude. Indeed, if a mutation rate of 1 per 500 generations for Y-STRs is assumed, it will have taken at least 10 000 years to arrive to all 10 haplotypes of the upper right branch. This conclusion illustrates the lack of migration for vast periods of time for some population groups within the Low Countries.

**Conclusion**

Thanks to the success of genetic genealogy, presently a wealth of genetic data can be found on genealogical websites. This, together with freely available software, permits genuine population genetic research within the reach of the enthusiastic genealogist. We express our thanks to each individual for sharing his genetic data. However, it should also be stressed that the utmost care is required when entering data on websites in order to prevent clerical errors and subsequent faulty conclusions.

**References**

Cavalli-Sforza LL (1997) Genes, peoples, and languages. PNAS, 94:7719-7725  
 Deboeck G (2008) Genetic diversity in Flemish Y-DNA. J

Genet Geneal, 4:55-57  
 King TE, Parkin EJ, Swinfield G, Cruciani F, Scozzari R, Rosa A, Lim S, Xue Y, Tyler-Smith C, Jobling MA (2007). Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. Eur J Human Genetics 15:288-283.  
 Mertens G, Leijnen G, Jehaes E, Rand S, Jacobs W, Van Marck E (2008) A likelihood based approach using Y-chromosomal STRs for metapopulation determination. In Population Genetic Research Progress, Nova Science Publ., Hauppauge  
 Zamoyski A (2004) Moscow 1812: Napoleon's Fatal March. Harper Collins, London

**Web resources**

http://www.ysearch.org/ - Ysearch, a genetic genealogy website  
 http://www.ybase.org/ - Ybase, a genetic genealogy website  
 http://www.familytreedna.com/public/benelux/default.aspx - BeNeLux DNA Project of Family Tree DNA, a genetic genealogy website/company  
 http://www.familytreedna.com/public/Flanders/default.aspx - Flanders-Flemish DNA Project of Family Tree DNA, a genetic genealogy website/company  
 http://www.hprg.com/hapest5/ - Y Haplogroup Prediction from Y-STR values by Whit Athey  
 http://lgb.unige.ch/arlequin/ - Arlequin, a software for population genetics data analysis  
 http://evolution.genetics.washington.edu/phylip.html - Homepage of Phylip, a collection of population genetics programs  
 http://www.fluxus-engineering.com/sharenet.htm - Network, a software generating evolutionary trees and networks from genetic, linguistic, and other data

**Table 1.** Haplotypes, haplogroups and geography of 225 men with origin in the Low Countries

Id	PROVINCE	REGION	D	D	D	D	D	D	D	D	D	D	HG
			Y	Y	Y	Y	Y	Y	Y	Y	Y		
1	Antwerp	BRABANT	17	15	18	12	29	23	10	11	14	11	A
2	Antwerp	BRABANT	13	16	18	13	30	24	10	11	13	13	E3b
3	Antwerp	BRABANT	15	12	15	14	30	23	11	11	13	12	I1b1
4	Antwerp	BRABANT	15	12	15	13	31	23	10	12	15	11	I1b2a
5	Antwerp	BRABANT	15	11	17	13	29	22	10	11	13	12	J1
6	Antwerp	BRABANT	14	13	18	12	29	22	10	14	11	13	L
7	Antwerp	BRABANT	15	11	13	13	29	25	10	11	13	11	R1a
8	Antwerp	BRABANT	14	11	14	13	29	24	12	13	13	12	R1b
9	Antwerp	BRABANT	14	11	14	13	29	23	11	13	13	12	R1b
10	Antwerp	BRABANT	14	11	15	13	30	24	11	13	13	11	R1b
11	Antwerp	BRABANT	15	11	14	14	30	23	11	13	13	12	R1b
12	Antwerp	BRABANT	15	11	15	14	30	24	11	13	14	13	R1b
13	Antwerp	BRABANT	14	11	14	13	30	24	11	13	13	13	R1b
14	Antwerp	BRABANT	14	12	15	13	30	24	11	13	13	12	R1b
15	Antwerp	BRABANT	14	11	13	13	30	24	11	13	13	11	R1b
16	Antwerp	BRABANT	14	11	15	13	28	24	11	13	13	12	R1b
17	Antwerp	BRABANT	14	11	11	13	29	25	10	13	13	12	R1b

18	Antwerp	BRABANT	14	11	15	13	29	23	11	13	13	12	R1b
19	Antwerp	BRABANT	14	11	15	13	29	23	13	13	13	12	R1b
20	Antwerp	BRABANT	14	11	14	13	29	23	11	13	13	12	R1b
21	Antwerp	BRABANT	14	11	16	13	29	23	11	13	14	12	R1b
22	Antwerp	BRABANT	14	11	14	13	29	25	11	13	13	12	R1b
23	Antwerp	BRABANT	14	12	14	13	29	25	11	13	13	11	R1b
24	Flem. Brabant	BRABANT	15	13	15	14	33	22	10	11	13	13	I1a
25	Flem. Brabant	BRABANT	15	14	14	12	28	23	10	11	13	11	I1a
26	Flem. Brabant	BRABANT	14	11	14	12	28	24	11	13	13	11	R1b
27	Flem. Brabant	BRABANT	14	12	14	13	29	23	11	13	13	12	R1b
28	Flem. Brabant	BRABANT	14	11	14	13	29	25	11	13	12	11	R1b
29	Flem. Brabant	BRABANT	14	11	15	14	30	24	10	14	13	13	R1b
30	Flem. Brabant	BRABANT	14	12	14	13	29	24	11	13	13	11	R1b
31	Flem. Brabant	BRABANT	14	11	14	13	30	24	10	13	13	12	R1b
32	Flem. Brabant	BRABANT	14	11	14	13	29	23	10	13	13	12	R1b
33	Wall. Brabant	BRABANT	14	11	14	13	29	23	11	13	13	12	R1b
34	North Brabant	BRABANT	15	12	14	12	29	22	10	11	13	12	G2
35	North Brabant	BRABANT	15	13	16	13	29	23	9	11	12	11	J2
36	North Brabant	BRABANT	14	13	16	14	31	22	10	11	12	12	J2
37	North Brabant	BRABANT	15	11	14	13	29	23	11	12	14	12	R1b
38	North Brabant	BRABANT	14	11	14	13	29	24	11	13	13	12	R1b
39	North Brabant	BRABANT	14	14	14	13	29	24	11	13	13	12	R1b
40	North Brabant	BRABANT	14	11	14	13	30	24	11	13	13	12	R1b
41	North Brabant	BRABANT	15	11	14	13	29	23	11	12	14	12	R1b
42	North Brabant	BRABANT	14	11	11	13	29	24	11	13	13	12	R1b
43	Drenthe	NORTH	14	13	14	12	28	22	10	11	13	12	I1a
44	Drenthe	NORTH	14	11	15	13	30	23	11	12	13	12	R1b
45	Drenthe	NORTH	13	11	14	13	29	23	11	13	13	12	R1b
46	Frisia	NORTH	15	14	14	12	29	22	10	11	14	11	G2
47	Frisia	NORTH	13	13	14	12	28	22	10	12	13	11	I1a
48	Frisia	NORTH	14	11	14	14	31	24	11	13	13	12	R1a
49	Frisia	NORTH	13	12	14	13	30	24	10	11	13	12	R1b
50	Frisia	NORTH	13	16	19	13	30	24	11	11	13	12	R1b
51	Frisia	NORTH	14	11	15	12	29	24	11	13	12	12	R1b
52	Frisia	NORTH	14	11	14	13	29	24	10	13	13	12	R1b
53	Groningen	NORTH	13	16	18	13	30	24	10	11	13	12	E3b
54	Groningen	NORTH	14	14	16	13	29	22	10	11	13	11	I1a
55	Groningen	NORTH	17	11	14	14	31	24	10	11	13	8	R1a
56	Groningen	NORTH	14	11	14	13	29	23	11	13	13	11	R1a
57	Groningen	NORTH	14	11	15	13	29	23	11	13	13	12	R1b
58	Groningen	NORTH	14	11	15	13	30	24	10	13	13	12	R1b
59	Groningen	NORTH	14	11	15	13	29	23	11	13	13	12	R1b
60	Gelderland	EAST	15	12	14	12	31	23	11	12	14	12	G2
61	Gelderland	EAST	15	13	14	12	29	22	10	11	14	11	G2
62	Gelderland	EAST	14	14	15	11	27	22	10	11	13	11	I1a
63	Gelderland	EAST	14	13	14	12	28	22	10	11	13	11	I1a
64	Gelderland	EAST	14	16	18	13	29	25	10	13	12	10	J1
65	Gelderland	EAST	14	13	18	13	30	23	10	11	12	11	J1
66	Gelderland	EAST	15	11	14	13	30	26	11	11	13	10	R1a
67	Gelderland	EAST	15	14	14	12	29	22	10	11	14	11	R1b
68	Gelderland	EAST	14	11	14	13	29	24	11	13	13	11	R1b
69	Gelderland	EAST	14	11	14	13	29	23	11	13	13	11	R1b
70	Gelderland	EAST	14	11	14	13	30	23	11	14	13	12	R1b
71	Gelderland	EAST	14	11	15	13	29	24	10	13	13	12	R1b

72	Gelderland	EAST	14	11	14	13	28	22	11	13	13	12	R1b
73	Gelderland	EAST	14	11	14	13	29	24	11	13	13	12	R1b
74	Gelderland	EAST	14	11	14	13	29	24	11	13	13	12	R1b
75	Gelderland	EAST	14	11	14	13	29	24	11	13	13	12	R1b
76	Gelderland	EAST	14	11	14	13	29	24	11	13	13	12	R1b
77	Gelderland	EAST	14	11	14	13	29	23	11	13	13	11	R1b
78	Gelderland	EAST	14	11	14	13	30	24	11	13	13	12	R1b
79	Gelderland	EAST	14	11	15	14	30	23	11	13	13	12	R1b
80	Gelderland	EAST	15	11	16	13	29	24	10	13	13	11	R1b
81	Gelderland	EAST	15	11	16	13	29	24	10	13	13	11	R1b
82	Overijssel	EAST	13	16	18	12	29	23	10	11	13	13	E3b
83	Overijssel	EAST	14	13	13	13	30	22	10	11	15	12	G2
84	Overijssel	EAST	15	14	15	12	29	22	10	11	14	11	G2
85	Overijssel	EAST	15	11	14	14	31	24	10	14	13	12	R1b
86	Overijssel	EAST	14	11	14	13	29	23	11	13	13	12	R1b
87	Overijssel	EAST	14	11	14	13	29	25	11	13	13	12	R1b
88	Overijssel	EAST	14	11	14	13	29	24	11	13	13	12	R1b
89	Overijssel	EAST	14	11	14	14	30	23	10	13	13	12	R1b
90	Overijssel	EAST	14	12	14	13	29	25	11	13	13	12	R1b
91	Overijssel	EAST	14	12	15	13	29	24	10	13	13	15	R1b
92	French Fl.	FLANDERS	14	11	14	12	28	22	10	11	14	11	I1a
93	French Fl.	FLANDERS	14	13	13	12	29	22	10	11	14	11	I1a
94	French Fl.	FLANDERS	15	16	17	13	29	23	10	12	15	11	I1b2a
95	French Fl.	FLANDERS	14	11	15	14	32	25	10	13	13	11	R1b
96	French Fl.	FLANDERS	14	11	14	13	29	24	11	13	13	11	R1b
97	French Fl.	FLANDERS	14	11	14	13	28	23	12	13	13	11	R1b
98	East Flanders	FLANDERS	13	16	19	13	30	24	11	11	13	12	E3b
99	East Flanders	FLANDERS	13	13	17	13	30	24	10	11	13	12	E3b
100	East Flanders	FLANDERS	15	14	14	12	29	22	10	11	13	12	G2
101	East Flanders	FLANDERS	15	14	14	12	29	22	10	11	14	12	G2
102	East Flanders	FLANDERS	15	14	14	12	29	22	10	11	14	12	G2
103	East Flanders	FLANDERS	15	13	13	12	28	22	10	11	13	12	I1a
104	East Flanders	FLANDERS	16	12	13	12	28	22	10	11	13	11	I1a
105	East Flanders	FLANDERS	16	14	14	12	29	23	10	12	15	10	I1b2a
106	East Flanders	FLANDERS	14	10	14	14	30	24	10	13	12	12	R1b
107	East Flanders	FLANDERS	14	11	14	13	29	23	11	13	14	12	R1b
108	East Flanders	FLANDERS	14	11	15	13	29	23	11	13	13	12	R1b
109	East Flanders	FLANDERS	14	11	14	14	31	23	10	13	13	12	R1b
110	East Flanders	FLANDERS	12	11	15	14	31	23	11	13	13	12	R1b
111	East Flanders	FLANDERS	14	11	14	13	30	24	11	13	13	12	R1b
112	East Flanders	FLANDERS	14	11	13	13	30	24	11	13	13	12	R1b
113	East Flanders	FLANDERS	14	11	15	13	29	24	10	13	13	12	R1b
114	East Flanders	FLANDERS	14	11	14	13	29	24	11	13	12	11	R1b
115	East Flanders	FLANDERS	14	11	12	13	29	23	11	13	13	12	R1b
116	East Flanders	FLANDERS	14	12	14	14	28	23	11	13	13	11	R1b
117	East Flanders	FLANDERS	14	11	11	13	29	23	11	13	13	12	R1b
118	West Flanders	FLANDERS	15	14	15	12	29	22	10	11	15	11	G2
119	West Flanders	FLANDERS	15	12	14	12	29	22	10	11	13	11	G2
120	West Flanders	FLANDERS	15	14	15	12	30	22	11	11	15	12	G2
121	West Flanders	FLANDERS	14	12	15	14	30	22	11	11	13	11	I1a
122	West Flanders	FLANDERS	15	11	14	13	29	23	11	13	13	12	R1b
123	West Flanders	FLANDERS	14	11	14	13	29	24	11	13	15	12	R1b
124	West Flanders	FLANDERS	15	11	14	13	29	23	11	13	13	12	R1b
125	West Flanders	FLANDERS	14	11	14	13	28	22	10	13	13	11	R1b

126	West Flanders	FLANDERS	14	11	14	13	29	23	11	13	13	12	R1b
127	West Flanders	FLANDERS	14	11	15	13	29	24	11	13	13	13	R1b
128	West Flanders	FLANDERS	14	11	14	13	29	24	11	13	13	11	R1b
129	West Flanders	FLANDERS	14	11	14	14	30	23	10	13	13	12	R1b
130	West Flanders	FLANDERS	14	11	14	13	29	24	11	13	13	12	R1b
131	West Flanders	FLANDERS	14	11	14	13	29	24	10	13	13	12	R1b
132	West Flanders	FLANDERS	14	11	14	15	29	24	10	13		12	R1b
133	Zeeland	FLANDERS	13	13	13	13	30	22	10	11	15	12	G2
134	Zeeland	FLANDERS	14	14	15	11	27	22	10	11	13	12	I1a
135	Zeeland	FLANDERS	15	15	15	14	31	24	10	13	15	11	I1b2a
136	Zeeland	FLANDERS	15	13	16	13	29	23	9	11	12	12	J2
137	Zeeland	FLANDERS	16	11	14	13	30	25	10	11	13	11	R1a
138	Zeeland	FLANDERS	14	10	14	13	29	23	12	13	13	13	R1b
139	Zeeland	FLANDERS	14	11	14	13	29	24	10	13	13	11	R1b
140	Zeeland	FLANDERS	14	11	14	13	30	24	11	13	13	12	R1b
141	North Holland	HOLLAND	13	17	18	13	30	24	10	11	13	12	E3b
142	North Holland	HOLLAND	13	16	19	13	31	24	10	11	13	13	E3b
143	North Holland	HOLLAND	13	15	17	13	31	25	10	11	13	12	E3b
144	North Holland	HOLLAND	15	19	19	13	31	24	10	11	12	11	E3b
145	North Holland	HOLLAND	15	14	14	12	29	22	10	11	14	11	G2
146	North Holland	HOLLAND	15	13	14	12	28	22	10	11	14	11	G2
147	North Holland	HOLLAND	14	13	14	12	28	23	10	11	13	12	I1a
148	North Holland	HOLLAND	14	14	14	12	28	22	10	11	13	11	I1a
149	North Holland	HOLLAND	14	13	14	12	28	22	10	11	13	12	I1a
150	North Holland	HOLLAND	16	13	16	12	29	25	10	11	13	11	J2
151	North Holland	HOLLAND	14	13	15	14	31	21	10	11	12	13	J2
152	North Holland	HOLLAND	14	15	16	12	27	23	10	13	13	12	K
153	North Holland	HOLLAND	14	11	13	13	29	23	11	14	15	10	N
154	North Holland	HOLLAND	17	10	14	14	31	25	10	12	13	10	R1a
155	North Holland	HOLLAND	15	11	14	13	30	24	11	15	13	13	R1b
156	North Holland	HOLLAND	14	17	14	13	29	23	12	13	13	12	R1b
157	North Holland	HOLLAND	15	11	13	12	29	24	11	13	13	12	R1b
158	North Holland	HOLLAND	14	11	16	13	29	23	11	13	13	11	R1b
159	Utrecht	HOLLAND	13	16	17	14	32	24	9	11	13	13	E3b
160	Utrecht	HOLLAND	14	14	15	11	27	22	10	11	13	11	I1a
161	Utrecht	HOLLAND	16	12	16	12	29	23	10	11	12	11	J2
162	Utrecht	HOLLAND	14	12	15	13	29	25	10	13	13	13	R1b
163	Utrecht	HOLLAND	14	11	14	13	29	24	11	13	13	12	R1b
164	South Holland	HOLLAND	13	15	18	13	30	24	10	11	13	12	E3b
165	South Holland	HOLLAND	13	17	18	14	31	24	10	11	13	12	E3b
166	South Holland	HOLLAND	13	16	18	14	31	23	10	11	13	12	E3b
167	South Holland	HOLLAND	14	11	14	13	29	24	11	13	13	13	G2
168	South Holland	HOLLAND	14	13	14	12	28	22	10	11	13	11	I1a
169	South Holland	HOLLAND	15	13	14	12	28	22	10	11	13	11	I1a
170	South Holland	HOLLAND	14	13	14	12	28	22	10	11	13	11	I1a
171	South Holland	HOLLAND	14	12	14	12	28	23	10	11	13	11	I1a
172	South Holland	HOLLAND	14	14	14	12	28	22	10	11	13	11	I1a
173	South Holland	HOLLAND	14	13	14	12	28	22	10	11	12	11	I1a
174	South Holland	HOLLAND	16	13	17	12	28	25	11	11	13	11	I1b1
175	South Holland	HOLLAND	15	15	15	14	30	22	10	12	15	11	I1b2a
176	South Holland	HOLLAND	15	15	17	14	32	23	10	12	14	11	I1b2a
177	South Holland	HOLLAND	15	15	14	14	32	23	10	12	14	11	I1b2a
178	South Holland	HOLLAND	15	15	16	13	31	22	10	12	15	11	I1b2a
179	South Holland	HOLLAND	13	15	19	12	29	22	12	11	12	11	J2

180	South Holland	HOLLAND	14	14	17	13	29	23	11	11	12	12	J2
181	South Holland	HOLLAND	15	11	14	12	30	25	11	11	13	10	R1a
182	South Holland	HOLLAND	15	11	14	13	29	25	10	11	13	10	R1a
183	South Holland	HOLLAND	16	11	14	13	30	25	11	11	13	10	R1a
184	South Holland	HOLLAND	14	11	14	13	30	25	10	13	13	13	R1b
185	South Holland	HOLLAND	14	11	14	13	29	23	11	13	13	12	R1b
186	South Holland	HOLLAND	14	11	14	13	29	23	11	13	13	12	R1b
187	South Holland	HOLLAND	14	11	14	13	29	23	10	13	13	12	R1b
188	South Holland	HOLLAND	14	11	11	13	29	24	10	13	13	12	R1b
189	South Holland	HOLLAND	14	11	14	13	29	23	11	13	13	12	R1b
190	South Holland	HOLLAND	14	11	14	12	27	25	12	13	13	13	R1b
191	South Holland	HOLLAND	16	14	15	14	32	24	10	12	14	11	R1b
192	South Holland	HOLLAND	14	11	14	14	30	24	11	13	14	12	R1b
193	South Holland	HOLLAND	14	11	16	13	30	26	12	13	13	12	R1b
194	South Holland	HOLLAND	14	11	14	14	30	23	10	13	13	11	R1b
195	South Holland	HOLLAND	14	11	14	12	30	23	11	13	13	13	R1b
196	South Holland	HOLLAND	16	11	14	13	30	25	11	11	13	10	R1b
197	South Holland	HOLLAND	14	11	14	13	30	23	10	13	13	13	R1b
198	South Holland	HOLLAND	14	11	14	13	29	24	11	13	13	11	R1b
199	South Holland	HOLLAND	14	12	14	14	30	24	11	13	14	11	R1b
200	South Holland	HOLLAND	14	11	14	13	29	24	10	13	13	12	R1b
201	Fl. Limburg	LIMBURG	16	12	16	15	31	24	10	11	13	11	I1a
202	Fl. Limburg	LIMBURG	14	13	16	13	29	24	10	11	12	10	J2
203	Fl. Limburg	LIMBURG	15	11	14	13	29	24	11	13	13	12	R1b
204	Neth. Limburg	LIMBURG	13	16	17	13	30	24	11	11	13	13	E3b
205	Neth. Limburg	LIMBURG	15	14	14	12	29	22	10	11	15	11	G2
206	Neth. Limburg	LIMBURG	14	13	14	12	28	21	10	11	13	11	I1a
207	Neth. Limburg	LIMBURG	14	12	14	13	28	25	11	13	13	11	R1b
208	Neth. Limburg	LIMBURG	14	12	14	13	31	23	10	13	13	12	R1b
209	Neth. Limburg	LIMBURG	15	11	14	13	29	24	11	13	13	12	R1b
210	Neth. Limburg	LIMBURG	14	14	14	13	30	24	9	13	13	12	R1b
211	Neth. Limburg	LIMBURG	14	11	14	13	29	24	11	13	13	12	R1b
212	Hainaut	WALLONIA	16	12	12	13	28	23	10	11	13	10	I1b1
213	Hainaut	WALLONIA	16	12	16	12	27	24	10	14	13	11	R1b
214	Hainaut	WALLONIA	14	13	16	13	29	24	11	13	13	12	R1b
215	Liege	WALLONIA	14	13	17	13	30	23	10	11	12	10	J2
216	Liege	WALLONIA	14	11	15	14	30	24	10	13	13	12	R1b
217	Liege	WALLONIA	13	11	14	13	29	24	11	13	13	14	R1b
218	Liege	WALLONIA	14	11	14	13	29	23	10	13	13	12	R1b
219	Luxemburg	WALLONIA	13	13	19	12	30	22	10	11	12	13	J1
220	Luxemburg	WALLONIA	15	11	14	13	30	25	11	11	13	10	R1a
221	Luxemburg	WALLONIA	15	13	14	12	28	23	10	11	13	11	R1b
222	Luxemburg	WALLONIA	14	11	14	13	28	23	11	13	13	11	R1b
223	Luxemburg	WALLONIA	14	11	15	13	29	24	10	14	13	14	R1b
224	Namur	WALLONIA	14	11	14	13	29	24	10	13	13	13	R1b
225	Namur	WALLONIA	14	11	14	14	30	24	11	13	13	14	R1b

# Whit Athey, developer of Y-haplogroup predictor

Denis Grigoriev

## Abstract

This interview is with Dr. Whit Athey, Editor of the Journal of Genetic Genealogy (JoGG - <http://www.jogg.info>), the developer of the Y-haplogroup predictor, and the administrator of several surname projects. You will learn when he began to be engaged in genetic genealogy, what his plans are for future developments of the Y-haplogroup predictor, and what he expects to see in the future for genetic genealogy. Also, Dr. Athey will tell what complexities arise in managing surname projects and what interesting, and sometimes unexpected, results can be discovered.

"I have always been interested in molecular biology, and my graduate work, though primarily in physics, was partly in molecular biology. When the article by Cann, Stoneking, and Wilson came out about 20 years ago, I was really struck by the potential for a better understanding of human origins. However, at that time I was heavily involved in other things, so I was just an interested bystander for many years.

I bought Bryan Sykes's book, *The Seven Daughters of Eve*, when it was published in 2001, and this rekindled my interest. I almost ordered the mtDNA sequencing that his company was offering, but in those days it was rather expensive, so again I did not get personally involved. However, I developed a presentation that I called «The Human Family», and presented it several times in 2001 and 2002 to small local groups.

In 2003 I finally ordered both Y-STR tests and mtDNA sequencing for myself, and I started a surname project for my own surname. I started five other projects during 2004 and 2005, mostly to characterize the Y profiles of other surnames of interest to me, but also to charac-

terize some of the mtDNA lineages of my ancestors, through testing of some of my cousins."

So has begun the story of Whit Athey with whom I communicated in August-September, 2008, more known to Russian DNA-genealogists as the developer of the Y-haplogroup predictor program. Following is the interview.

**D.G. – Many know you as developer of the Y-haplogroup predictor. How and when you have started to be engaged in a predictor? You did it independently or with someone else?**

**W.A. –** I started thinking about the problem of haplogroup prediction almost as soon as I saw my own results in late 2003. The company where I tested could not predict my haplogroup, so I wanted to find a way to do it myself. I tried several approaches before developing the one that I put on-line in about September 2004. The method was described in the Journal of Genetic Genealogy in early 2005 (<http://www.jogg.info> – unfortunately only available in English at present). This method calculated a "haplogroup fitness score" that measured how well a Y-STR haplotype "fit" into any given haplogroup, using a 0 to 100 scale. A fitness score of 100 means



Whit Athey

that the haplotype has exactly the modal values of the haplogroup and is therefore a perfect "fit". Typical scores for the actual haplogroup range from about 40 to 60.

The next major version of the program incorporated a Bayesian probability calculation and is also published in the *Journal of Genetic Genealogy* (2006). The program now returns both a fitness score and a Bayesian probability that the haplotype is in a particular haplogroup.

Until late 2006 I had done all of the programming in an Excel spreadsheet and then converted it to an executable program for the web site. With the addition of more and more haplogroups and markers, this approach was starting to require too much time to execute and take too long to download. Doug McDonald, who is a very capable programmer, offered to take my program and convert it into the C+ language, so that it could download and execute in a manner that would be much more efficient than my old version. This new version went on-line in early 2007 and has worked well since then.

I now have 23 haplogroups in the program. The only limitation to adding more haplogroups is that the program requires the allele frequency distributions for each marker in each haplogroup. A new haplogroup cannot be added until I can collect a few dozen haplotypes, including all

of the markers that people normally test, from people who are members of that haplogroup. The latest additions to the program were Haplogroups C3 and G1, which were added in June 2008. I am trying to collect a sufficient number of haplotypes from some sub-haplogroups of Haplogroup O that they can be added to the program. I would also like to add Haplogroup N2, which should be of interest to many people from Central Russia. At present the program provides a prediction for Haplogroup "N", but in fact, it is really Haplogroup N3, which is common in northwest Russia. I would like to have both N2 and N3 separately in the program.

You asked me about the possibility of adding Haplogroup O3 to the program. In order to add O3 or O3a3 to the program, I would need a few dozen 67-marker haplotypes that were confirmed or strongly predicted to be O3 or O3a3. I could even work with 37-marker haplotypes. Using the characteristics of these haplotypes on the basic markers, I could probably identify others in the SMGF database in order to provide data on about 10 more markers. The same for N2. I do have a collection of O3 minimal haplotypes from research studies, so that is a beginning.

In regard to the validation of the predictor program, I have done a little, but I would like to do more. The challenge is to select a test dataset that is not biased. The initial validation was done on 100 R1b haplotypes and 50 I1-M253 haplotypes. However, there needs to be a validation involving many haplogroups. These should not include haplotypes that have been used in my program, so this right away makes the selection difficult. Still, it is probably possible to put together several more datasets for validation purposes. I have just not had the time to follow up on this.

#### **D.G. - Tell about your other activities, please.**

**W.A.** - As you are aware, my other major activity in the genetic genealogy field is that I serve as Editor of the *Journal of Genetic Genealogy* (or JoGG). We publish a free on-line journal with two issues, Fall and Spring, each year. Almost all of our authors are "amateurs" rather than professional geneticists, though we welcome submissions from anyone. JoGG provides a way for the amateurs to publish their work in this field. In JoGG, the only thing that matters is the quality of the work and its presentation — we do not care what your field of formal training may be (or even if you have formal training in any field).

Enter any combination of one or more markers, or use the string entry form below

Area Selection

393	390	19	391	385a	385b	426	388
<input type="text" value="0"/>							
439	389 1	392	389 2	458	459a	459b	455
<input type="text" value="0"/>							
454	447	437	448	449	464a	464b	464c
<input type="text" value="0"/>							
464d	460	H4	YCAIIa	YCAIIb	456	607	576
<input type="text" value="0"/>							
570	CDYa	CDYb	442	438	531	578	395a
<input type="text" value="0"/>							
395b	590	537	641	472	406	511	425
<input type="text" value="0"/>							
413a	413b	557	594	436	490	534	450
<input type="text" value="0"/>							
444	481	520	446	617	568	487	572
<input type="text" value="0"/>							
640	492	565	461	462	A10	635C4	1B07
<input type="text" value="0"/>							
441	445	452	463	485	495	505	508
<input type="text" value="0"/>							
522	532	533	540	556	643		
<input type="text" value="0"/>							

Enter a string with allele values separated by *single* spaces, commas, tabs, returns ("Enter"), ampersands (&) or semicolons in the area below, select the proper value for the existence of extra alleles and the separator character, and click "Submit".  
 Markers must be in the order of the above form.

Separator:

Haplo-group	Fitness score	Probability (%)
C3	-	-
E1b1a	-	-
E1b1b	-	-
G1	-	-
G2a	-	-
G2c	-	-
H	-	-
I1	-	-
I2a (xI2a1)	-	-
I2a1	-	-
I2b (xI2b1)	-	-
I2b1	-	-
J1	-	-
J2a1b	-	-
J2a1h	-	-
J2a1 (x bh)	-	-
J2b	-	-
L	-	-
N	-	-
Q	-	-
R1a	-	-
R1b	-	-
T	-	-

Haplogroup Predictor (<http://www.hprg.com/hapest5/hapest5b/hapest5.htm>)

**D.G. - As the editor of JOGG, you probably communicate with different experts in the field of genetic genealogy. Have you communicated with any from Russia?**

**W.A.** – Unfortunately, I have not had much contact with Russian amateurs or professionals. As editor of JoGG, I have corresponded with one of your compatriots, who seems to be quite expert in mtDNA. He has served as a reviewer for JoGG. He may possibly be a professional geneticist, but I really know very little about his background. As I said in regard to JoGG and its authors, the only important characteristics are what an author or reviewer knows about the subject. We don't care how she or he became knowledgeable. If any of your readers have an interesting project, I would invite them to write an article for JoGG. It would probably be a good idea to check with me to make sure the subject would be appropriate for JoGG, before devoting a lot of time to this. I have had a very brief correspondence with Malyarchuk and he was very helpful in obtaining an article for me from

a Russian journal. I could not read the article, but I was primarily interested in one of the data tables, which was available no where else.

I think that population geneticists who work in Russia have an incredibly interesting and diverse group of populations to work with, all within or very near your borders. I would think that Russia is one of the best places in the world to work as a population geneticist.

**D.G. - What developments do you expect in the next few years in genetic genealogy?**

**W.A.** – These things are difficult to predict - I am often surprised at developments that actually occur, and surprised with others fail to materialize as quickly as I hope.

One area where I do expect some nice developments is in the price for testing services. I believe that the cost will continue to fall for basic tests, or else the price may remain almost the same while more results are returned. I hope that will mean that many more people in the

world can participate.

Family Tree DNA has announced that they are working on a project that would allow the sequencing of small sections of the Y chromosome for individuals. This project is called the "Walk Through the Y" project and it could be a very significant development. We are hoping that many new Y SNPs will be discovered.

I believe that we may see more use of "gene chips" to produce hundreds or thousands of Y-SNP results at once. This is already available, but at the present time it is expensive and not very focused on SNPs on the Y chromosome.

I also believe very strongly that "amateurs" like us will continue to make real contributions to this field. At first we were only "consumers" of information that the "professionals" generated. More and more important discoveries are coming from amateurs and I believe that this phenomenon will grow even more in the future.

**D.G. - I know that you actively communicated with most of the companies offering DNA-tests, but you yourself nevertheless have primarily used Family Tree DNA. Why?**

**W.A.** - I have most of my testing done at FTDNA and all of my projects have an official "home" there. However, I have had testing done at nearly all of the labs and I believe that our community is best served by having a variety of options available to us. The competition helps to drive developments — each company knows that it must innovate if it is to continue to be successful. I have located my project web sites at independent sites, rather than accepting one of the project web page templates that FTDNA offers. It would be much easier to accept the template because all of the FTDNA results are transferred automatically. However, I use independent sites so that I can include results from any lab.

I have what I believe is a good relationship with Bennett Greenspan and Thomas Krahn of FTDNA. I can send an e-mail to either of them and expect to receive a thoughtful answer. I try not to send e-mail to them except when really needed because I am sure that their mailboxes are very full every day. There are several staff members at FTDNA with whom I most often communicate with in regard to the more mundane problems that arise.

I have found that most of the companies I have worked with have provided good support by being available to answer questions and investigate problems. Most of them welcome constructive feedback. I am happy to be able to

communicate with each of them.

**D.G. - Tell, please, what you consider as the most important development of your family projects? What difficulties have you met?**

**W.A.** - The most important development for my Athey project, most of whose members are in Haplogroup G2a3, is that we finally found a cluster of Y-STR results from people of another surname that match those of my Athey participants. Since the two families lived in different places, Ireland and England, from at least the year 1500, and my Athey ancestor immigrated from Ireland to North America in 1661, there was no opportunity for a connection between the two families until we go back prior to the year 1500. The other family is named Whitfield and is the same family that included the Rev. George Whitfield, one of the founders of Methodism (who had no male children himself). This is quite ironic since my own given name [first name] is Whitfield.

The main challenge I face in all of my projects is recruitment of participants. There are still many lines from my immigrant ancestors that are not represented in my projects, and locating appropriate people to test, and then persuading them to be tested, is difficult, even when I offer to help pay the cost. I have two people that I have located who would be very valuable to one of my projects, and I can't persuade them to be tested.

**D.G. - Tell more in detail about your project. The number of members of your projects constantly grows, I think. How many members are now in your projects? Are there any Russians or any with Russian roots? Whether you can recollect any interesting cases from DNA-genealogy? Whether there were some unusual DNA-results?**

**W.A.** - My Athey surname project has about 30 participants, 21 of whom have profiles in Haplogroup G2a3-U8 that all match with each other (I am in this cluster). My largest project is for the surname Owen (a Welch name), which has about 150 participants. My other projects are small with only a few participants. These are for the surnames Folmar (from German, Vollmar), Rodgers, and Perdue (from the French, Perdieu).

My own mtDNA profile puts me in Haplogroup U5a1a. This haplogroup is found in central and eastern Europe, including Russia. In fact, the closest match to my full-sequence profile is in a Russian research subject from a study

a few years ago by Malyarchuk and Derenko. If my mtDNA line really comes from Russia, and I have no other evidence that it does, it must have left many centuries ago. I am not aware that I have any recent ancestors from Russia. My Y haplogroup (G2a3) may have come from the Caucasus region many centuries ago.

Here is a link to my Athey project web site: <http://www.hprg.com/athey/> You will see that our main cluster has some unusual values, even for Haplogroup G2a, which is already fairly unusual (except in the Caucasus).

I am also giving you a photo of me with Garland Boyette from last year's FTDNA con-



Garland Boyette and Whit Athey

ference in Houston. Garland is very unusual in having a F\* haplogroup, and he runs a very successful Boyette project. I have followed this project closely, and sponsored one of my cousins as a participant in the project. One of my great-great-great-grandmothers was a Boyett, and my cousin had this same odd F\* haplotype as about 25 other Boyetts in the project. I had corresponded with Garland many times before finally meeting him in person in Houston. I was very surprised to discover that he is a black man, whereas all of the other Boyett participants are white. He and I are about seventh cousins. Garland's Boyett ancestor was a slave, fathered by his Boyett owner, and Garland actually started the project to prove this theory. Incidentally, Garland travels often to your southern border areas, and is presently working on a project in Kazakhstan. Garland also is the administrator of the F\* project and I have worked with him on trying to figure out which of the major haplogroups within F that his line is closest to.

**D.G. - What is more interesting for you yDNA or mtDNA and why?**

**W.A.** - For my genealogical research, the Y is much more useful and interesting because

the Y follows the surname, and it adds a new dimension to researching particular surnames. I have had many more new discoveries from the Y studies than mtDNA.

The mtDNA lineages are more difficult to trace using traditional genealogical tools because, at least in our culture, the names of the women change at each generation. However, while mtDNA matches that one finds in a database are not likely to be useful genealogically, I find that mtDNA can be quite useful for hypothesis testing. That is, if there is a relationship you are trying to prove or disprove a few generations back, it may be useful to trace matrilineal lines from the person in question, down to living people, and test the mtDNA of those persons. I did that recently for a great-great-great-grandmother of mine, who was a matrilineal ancestor of my father. Another researcher had claimed this same woman as an ancestor, but I doubted the claim. If his claim was correct then a certain female cousin of his, should have had the same mtDNA as my father. I paid for a test of the other researcher's cousin, and sure enough, she did not match my father. This is what I mean by "hypothesis testing". In other cases, the hypothesis you want to test might require a positive match, or in another situation it might require that two people not match. In these very specific circumstances that you set up to test a particular theory, a mtDNA match can be highly significant, whereas mtDNA matches in the general population are usually meaningless for genealogical purposes. I think that mtDNA is actually underutilized and underappreciated for this kind of application.

For anthropological applications, I am equally interested in the Y and mtDNA research. Both have important roles to play. I am very interested to read articles on both types of DNA.

**D.G. - I think it will be interesting for our readers to know about your biography. Where are you live now and where are your roots from?**

**W.A.** - I presently live about 35 km north of Washington and about 45 km southwest of Baltimore. The area was fairly rural when I moved here 20 years ago, but now the expanding city has caught up with us.

I was born in a rural farming area in the south, in the state of Alabama. The nearest village to my house had a population of about 100 people. I attended a public university in Alabama where I studied physics. People sometimes ask me why, with my farm background, I was attracted to physics. I tell them that when you

are working in the hay fields in summer and the temperature is 38 degrees C, or you are trying to push a reluctant cow up a loading ramp to a truck, physics can seem quite attractive!

I was also attracted to molecular biology when I was a teenager, and if there had been a curriculum in molecular biology at the time, I would probably have studied that instead of physics. I kept this interest through college, and when I was in graduate school in physics, I decided to try to do my research project in a joint physics-biochemistry program, and finally, I could combine many of my interests.

Talking with Dr. Athey about different companies, I agree that presence of many companies allows us to capture any aspects of DNA-testing (including a price question), and also features of different regions. Having one's own web-site of the project, or "representation" in any community, it is possible to accumulate the information for the project, without breaking rules of the companies. The data can be collected worldwide.

© Denis Grigoriev  
August 2008 - June 2009

# Comment on "Geographic distribution and molecular evolution of ancestral Y chromosome haplotypes in the Low Countries" by Gerhard Mertens and Hugo Goossens

Guido J Deboeck<sup>1</sup>

Dear Editor-in-Chief

In "*Geographic distribution and molecular evolution of ancestral Y chromosome haplotypes in the Low Countries*" by Gerhard Mertens and Hugo Goossens, (published in number 1 of volume 1 of the RJGG), the authors present the results of an analysis of a Y chromosome dataset for individuals with genetic roots in the Benelux.

Many historians, familiar with the history of the Low Countries, as well as many countrymen of the authors, will be surprised (and possibly be dismayed) about the geographic distribution that was created and adopted in this paper. Since this geographic distribution was the basis for all comparisons of Y chromosome haplotypes, we (a number of countrymen and history buffs) have serious concerns with regard the analysis of the data on which the conclusions are based. The paper is too short with regard to the history of this part of the world, which is necessary to explain some of the results.

The authors wrote that "*in order to be included in the study, the place of birth of the oldest known ancestor had to be located in the territory of present day Belgium, the Netherlands, Luxemburg or the North Department of France*". However, to make meaningful comparisons between Y chromosome haplotypes, the authors constructed artificial "regions" based on their

own "geographical, historical and cultural logic". Unfortunately this logic is not provided in the paper. We respectfully submit that these artificial groupings can hardly be justified on the basis of established geographical, historical and cultural history of the Low Countries.

This paper uses data from ancestors born between 1200 and 1922 (with 75% of cases before 1739) from locations in present day Belgium, Netherlands, Luxembourg, and the French Department du Nord, grouped in several "regions". Some groupings e.g. Brabant, Limburg, Holland, Flanders are obvious. However, the distribution of provinces over the North Provinces and the East Provinces is not obvious. To the extent that geography is the main determinant of common characteristics the presented groupings are appropriate. The difficulty arises, however, because over time several of these locations belonged to principalities, or moved between principalities of which the groupings have no relationship to the regions to which they are assigned for this study.

For example, part of *Overijssel* was in the 14th century part of the 'bishopric of Utrecht', after having been part of Gelderland. Gelderland is grouped as part of the East Provinces while Utrecht is grouped as part of the region Holland. Some ancestors of *Overijssel* may, however, have more common characteristics with those of Holland than with the East Provinces. Also, during its early history *Overijssel* included much of modern-day Drenthe. Therefore, some older ancestors born in *Overijssel* may have more common elements with ancestors from the North Provinces than with those of the East Provinces. So, the geographic delineations adopted in this paper may not be appropriate for some an-

---

Received: July 4 2009; accepted: July 4 2009; published July 8 2009

<sup>1</sup>The associate editor for JOGG, Arlington, VA in collaboration with several others who reviewed this paper for an earlier submission to JOGG.

E-mail reactions to [gdeboeck@mac.com](mailto:gdeboeck@mac.com)

cestor from older periods. Another example: *French Flanders* contained not only sites in the Department du Nord, but also some of the Department du Pas de Calais. The annexation by France of French Flanders took place over some time period. Regarding *Brabant*: To the extent that the Germanic-Latin language border has a significant genetic difference, then the sites of Waals Brabant should be removed from the region Brabant and moved to Wallonia. A question could also be raised regarding the inclusion of Zeeland in the Flanders region. Would it not be preferable to include only Zeeuws Vlaanderen or the territories South of the Scheldt in the Flanders region and the remainder in the Holland region with which it seems to have many historical and cultural links?

Nothing is said about *Flevoland*, which is part of the East Provinces. If Flevoland was won only after 1922, then there would be no ancestors in the sample! Flevoland actually exists only since 1986. After a flood in 1916, it was decided that the Zuiderzee, an inland sea within the Netherlands, would be enclosed and reclaimed. In 1932, this work was completed, which closed off the sea completely. The Zuiderzee was subsequently called the lake at the end of the river IJssel. The first part of the new lake that was reclaimed was the Northeast polder. This new land included the former islands of Urk and Schokland and was included in the province of Overijssel. After this, other parts were reclaimed: the Southeastern part in 1957 and the Southwestern part in 1968. There was an important change in these post-war projects from the earlier Noordoostpolder reclamation: a narrow body of water was preserved along the old coast to prevent coastal towns from losing their access to the sea, so that Flevopolder became an artificial island joined to the mainland by bridges. The municipalities on the three parts voted to become a separate province, which happened in 1986.

The traditional view of a *clear-cut division between Frisians in the north, Franks in the south and Saxons in the east, common in the 19th- and early 20th-century historiography, has proven problematic*. Archeological evidence suggests dramatically different models for different regions, with demographic continuity for some parts of the country and depopulation and possible replacement in other parts notably the coastal areas of Frisia and Holland. Much of the western Netherlands was barely inhabited between the end of the Roman period and around 1100. Around 1000, farmers from Flanders and Utrecht began purchasing the swampy land,

draining it and cultivating it. This process happened quickly and the uninhabited territory was settled in only a few generations. They built independent farms that were not part of villages, something unique in Europe at the time. Before this happened the language and culture of most of the people who lived in the area that is now Holland were Frisian. The area was known as "West Friesland" (Westfriesland). As settlement progressed, the area quickly became Dutch. This area became known as "Holland" in the 12th century. (The part of North Holland situated north of the 'IJ' is still colloquially known as West Friesland)."

The historic factors calling for a distinction between the *East Provinces and the North Provinces* are somewhat ambiguous. Flemish migration, before and after the split of the Seventeen Provinces, and that of French Huguenots was mainly in the direction of Holland and Zeeland. Immigration from Germany and Scandinavia was for a large part to large cities in Holland. However, there is much less information on immigration from Germany and Scandinavia than on the flows from the South. There is evidence of substantial influx of German immigrants in the eastern provinces, in particular Gelderland, but not excluding the North Provinces. The East Provinces have more cities and much higher populations than the North Provinces. In an interesting article "Founder mutations among the Dutch" in *European Journal of Human Genetics* (2004) 12, 591 – 600, Zeegers and al. attribute regional differences to a geographic division by the major rivers, the Maas and the Rhine, as they crested barriers to migration. The location of the rivers does not seem to favor a separation between the East Provinces and the North Provinces. This article is also interesting for what its title indicates: founder mutations.

The authors wrote that "*Dutch is the language of all regions of the Low Countries, except for the region of Wallonia, where French is spoken.*" However, French is not only the language of the region Wallonia, but also part of Brabant and is the majority language in Brussels. Regarding Luxembourg, part of the region Wallonia, it should be mentioned that it is composed of both the Belgian province and the Grand Duchy of Luxembourg. The language most commonly used by the natives of Luxembourg is Luxembourgish, which is also an official language besides French. As French Flanders is included in Flanders, the latter is not unilingual Dutch.

The analysis in this paper is based on a number of freely available genetic software

packages and the results are interpreted in the context of the history of this region. However, there are a number of flaws both in the use of the software and the interpretation of the results. Specifically,

The term "ancestry informative markers" is mainly used for loci showing polymorphism in one population but almost not in other populations. This term is therefore not appropriate for explaining the increased frequency of certain alleles in populations. This high frequency could have resulted from genetic isolation and a low population size. Furthermore, the out-of-Africa migrations that predate the arrival of Homo sapiens in Europe has also resulted in a decrease of genetic diversity (genetic bottleneck) with a result that the non-African populations show difference in genetic diversity with respect to the African populations.

Usually when  $F_{st}$  values are determined it is common practice to determine the significance of the differences by doing permutation tests. It is therefore difficult to see if there are any significant differences (except with the African population) among the groups from the Benelux from the figures in table 2. Furthermore, negative values indicate no difference at all between the data. It is unclear from the paper if this has been taken into account for calculating the Neighbour Joining (NJ) tree.

The NJ tree shows some results that contrasts with publications concerning the Dutch population by De Knijff et al. The largest differences were seen between the Northern provinces and the Southern while in this paper it is an East-West (East-Holland) difference. It is difficult from the data in the paper to identify if this difference is due to a problem in the analysis performed with the genetic software programs or is due to the sample set.

The network produced in figure 5 is probably not correct, which is also substantiated by the remark of the authors that "if median vectors are included the tree becomes very cluttered". This is mainly due to insufficient knowledge of the analysis of Y chromosome haplotypes with the network software by the authors. Not all Y-STRs can be used for this analysis and one must incorporate also differences in mutation rate in order to get a network that reflects the evolutionary history of the haplotypes. In addition, one should make use of the haplogroup designation to explore in more detail the evolutionary history and to come to meaningful conclusions with regard to the history of the populations analysed. This can be done by constructing phylogenetic networks for each haplogroup or to

include the haplogroup designation in the haplotype. The recent paper by King and Jobling (Molecular Biology and Evolution 26: 1093-1102, 2009) could be helpful for this analysis. Therefore, any conclusions (historical and ancestral haplotypes) obtained from the network in figure 5 is premature and should be regarded as not scientifically substantiated.

*In a nutshell, the historical context of what the Low Countries went thru is missing and hence makes it more difficult for readers not familiar with this part of the world to grasp the complexities the paper is trying to unravel. The paper also omits any comparisons with Great Britain, France, Germany which are the direct neighbors of the former Low Countries. There are a number of flaws in the use of the freely available software, not to mention wrong interpretations of results.*

Sincerely,

Guido J Deboeck et al.  
Arlington, VA

## Y-haplogroups of carriers of the Aryan language

A.A. Aliev,  
A.S. Smirnov

### What will be discussed

Ancient history of the Aryan language, the ancestor of Nuristani, Iranian and Indo-Aryan languages (Fig. 1) is still the object of scrutiny.

A study of its history has always faced the problem of localization of its ancestral homeland, the area of origin of Old Aryan.

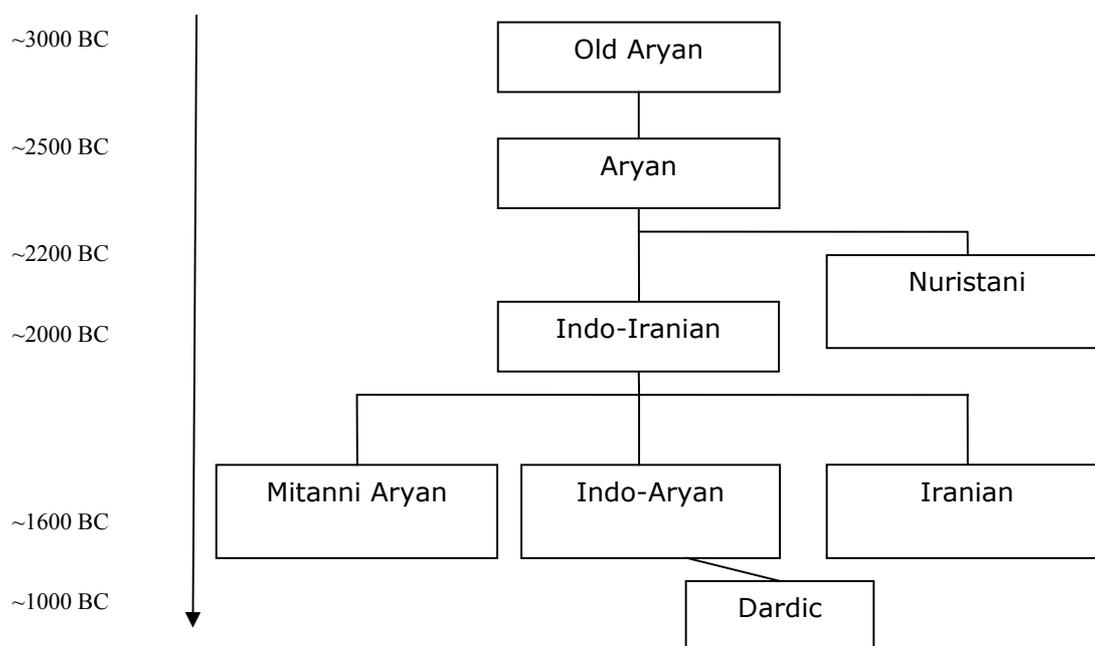


Fig. 1. Aryan language group

Currently considered two basic hypotheses:

1) Steppe («Kurgan») hypothesis. According to this hypothesis, the area of the initial spreading of the Aryan language was the Russian Plain and zone of so called Andronovo culture in the end of III millennium BC beginning of I millennium BC from Southern Urals to Central Asia [1, 2].

2) «Bactrian-Margianian» hypothesis. According to this hypothesis the area of the initial spreading of the Aryan language was the zone of Bactrian-Margianian culture in the end of III millennium BC and beginning of II millennium BC in south of Central Asia and Afghanistan [2, 3].

Recently, addressing this issue has attracted the scientific basis of DNA genealogy, based on

opinion that haplogroups of original Aryan language speakers could at least partially be preserved in the modern native speakers of the Aryan group languages.

Among the works in Russian on the topic should be allocated to articles of Dr. A.A. Klyosov «Where did the Slavs and Indo-Europeans come from and where is their ancestral home? The answer is provided by DNA genealogy» [4] and «Another proof of the transition of the Aryans (haplogroup R1a1) in India and Iran from Russian Plain» [5]. Klyosov binds the spreading of Aryan languages in Iran and India with the migration of carriers Y-haplogroup R1a1 (M17) from the Russian Plain.

The main arguments for his theory are based on the high (over 60%) prevalence of haplogroup R1a1 among Ukrainians, the people of the Pamir and the Brahmins. According to Klyosov's calculations, the age of the common ancestor of Brahmin R1a1 is  $4050 \pm 500$  years and the age of the common ancestor of Slavs is  $4750 \pm 500$  years. The older age of Slavic R1a1 may indicate the direction of R1a1 migration from the Russian Plain across the Urals and Central Asia to northwestern India, which took place not later than the II millennium BC.

It should be noted that according to Klyosov's calculations, the age of the common ancestor of the South Asian R1a1 was significantly higher than 4 thousand years and is above 12 thousand years [6]. According to Zhivotovskiy's calculations this age is higher [7, 8, 9]. This excludes the allegation that R1a1 appears in India along with the «Aryan invasion» during the migration from the Russian Plain. In other words, haplogroup R1a1 was among the Indian population long before the invasion of Indo-Aryans.

If we consider a set of haplogroups of modern north Indian Brahmins, as the most likely candidates for the direct descendants of the ancient Aryans, it consist of 68% R1a1, 21% J2, 16% H1, 3.6% G2a [7-9]. As we see, among this set contains a typical North Indian haplogroups (R1a1, H1), and «Middle East» haplogroups (J2 and G2a), which argues in favor of the hypothesis of mixed origin of people of this caste.

In addition to the Brahmins, pagan Kalashs, the endogamic Dard people in the mountainous Pakistan can be alleged genetic descendants of carriers of Old Aryan language. The set of Kalash haplogroups consists of L3a (22.7%), H1\* (20.5%), R1a (18.2%), G (18.2%), J2 (9.1%) [10].

Based on this data, the Kalashs and the Brahmins have approximately the same set of «local» and «Middle Eastern» haplogroups represented in different proportions.

There is a lack of consensus regarding the ancestral home of the Aryan languages. There are also drawbacks of the «steppe» hypothesis and «Bactrian-Margianian» hypothesis.

#### Disadvantages of «steppe» hypothesis

The hypothesis of «steppe» homeland of the Aryan language has a number of linguistic and archaeological inconsistencies.

According to this hypothesis, the Aryan language split within the Russian Plain and the Indo-Aryans and the Iranians, not mingling with each other individually, but along the same path, through the Urals migrated to the Central Asian oases. Then the Indo-Aryans migrated across the Hindu Kush in the Punjab, and the Iranians settled in the Iranian plateau. For Mitanni Aryans «proposed» path of the invasion was from the Russian plain through the Caucasus to Mesopotamia.

Hypothesis does not take into account the Old Nuristanis, who were the ancestors of the modern Nuristani tribes living in the modern boundaries of Afghanistan and Pakistan. If the Indo-Aryans and Iranians lived in the south of the Russian Plain, it means that Old Nuristanis should have been separated from them earlier. According to diffusion of ancient migrations we could meet the Nuristanis anywhere. But nevertheless the region they live is the valleys of the same Hindu Kush, the contiguous territory of residence Indo-Aryans and Iranians. The probability of distribution in one region of the three related groups who independently migrated

thousands of kilometers from the outside is completely negligible.

In addition, look at the geography of the Avesta and the Rig Veda, the only sources of our knowledge of the Aryans. Avesta and the Rig Veda describe the same region, covering the rivers, starting in the mountain systems of Pamirs, Hindu Kush and the Himalayas.

Vedic (Indo-Aryan) and Avestan (Iranian) languages are very close. It can not be the result of their separate existence and their separate migrations over the centuries and thousands of kilometers from their original homeland. This state of Indo-Iranian borderlands could not be the result of an independent migration of Indo-Aryans and Iranians, who separated thousands of kilometers away. It seems to be a plausible assumption that the presence of three different Aryan groups in one region is not accidental and not a result of their separated invasions from the outside.

Localization of Aryan homeland not on the Russian plain, but in the Central Asian area of Andronovo culture is also faced with other kinds of linguistic and archaeological confusion. Andronovo graves do correlate with the Aryan funeral rituals. Aryans used the cremation, not burying corpses in the land. According to the Avesta the desecration of land by dead matter is the ultimate sin. In the reconstructed Old Aryan language is viewed significant impact of the Semitic language system, which is possible only in conditions of close contact. According to the hypothesis Szemerényi, the transformation of Indo-European vocalism \*e \*o \*a and in Old Aryan occurred under the influence of Semitic languages with a triangular *a~i~u* system [11].

The ethnonym «arya» origins from the Indo-European \*ario-s «friend, equal, noble» has anomalous structure for the Proto-Indo-European and has Afro-Asiatic origin (for example, in Ugaritic «Ary» means «relative, friend»). In addition, south of Central Asia, where the presence of the Aryans is undeniable, there is no presence of the Andronovo. Along with Semitic influence in the Old Aryan language we can identify the substrate [12] which has the anomalous non- Indo-European structure of the

roots. This substrate is not clearly attributable to any presently known language families. Analysis of the semantics of substrate words allows dividing them into four categories:

- 1) words associated with the cult of the Soma/haoma, and such gods like Indra, Sarva;
- 2) names of animals – «camel», «donkey»;
- 3) irrigation and land reclamation terminology – canals, wells, sleeves;
- 4) all architectural and construction terms related to stationary houses with walls of brick and gravel.

Such cultural and linguistic contacts imply interaction of Old Aryan with Semitic languages on one hand, and interaction of Old Aryan with the unknown language of the civilized people familiar with farming land reclamation and construction of buildings of brick on the other hand, linguistically and archaeologically excluding «pastoral» Andronovo culture from the list of Aryan cultures due to its distance from Mesopotamia, the main area of distribution of Semitic languages in antiquity.

#### Disadvantages of «Bactrian-Margiana» hypothesis

«Bactrian-Margiana» hypothesis localizes the Aryan homeland in Margiana civilization (BMAC). This civilization had its own distinctive features such as brick construction, land reclamation, cultivation of donkeys and camels. It corresponds to the substrate terminology detected in the Aryan language. In addition, the distribution area of Margiana civilization is consistent with the toponyms of the Avesta and the Rig Veda and with the possible ways of further migration of the Aryans in the Pamir and Hindu Kush.

This hypothesis however does not take into account the aforementioned influence of the Semitic language system.

But where was the Aryan language born? How could R1a1 get from South Asia to the Russian Plain? Why among the Brahmins and the

Kalash, in addition to «local» haplogroups, present «Middle East» haplogroups J2, G2a? Where and how could Aryans have contact with the Semitic languages, as well as with the «substrate» language?

All these issues require the development of a unified system of events that would take into account all these facts.

#### The search for ancestral homeland

How can you find the ancestral homeland of the Aryan language? To do this, define the region, which corresponds with conditions of formation of Aryan language.

The presence of Semitic influence in Old Aryan allows to locate an ancestral home in the area, where could be contacts between Old Aryan and Semitic in III-II millennium BC. According to the hypothesis of TV Gamkrelidze and VV Ivanov [9], not later than the VI-V millennium BC in the contact area of Asia Minor and Northern Mesopotamia allocated Proto-Indo-European language, which is associated with the archaeological culture of Tell-Halaf in northern Syria (V millennium BC).

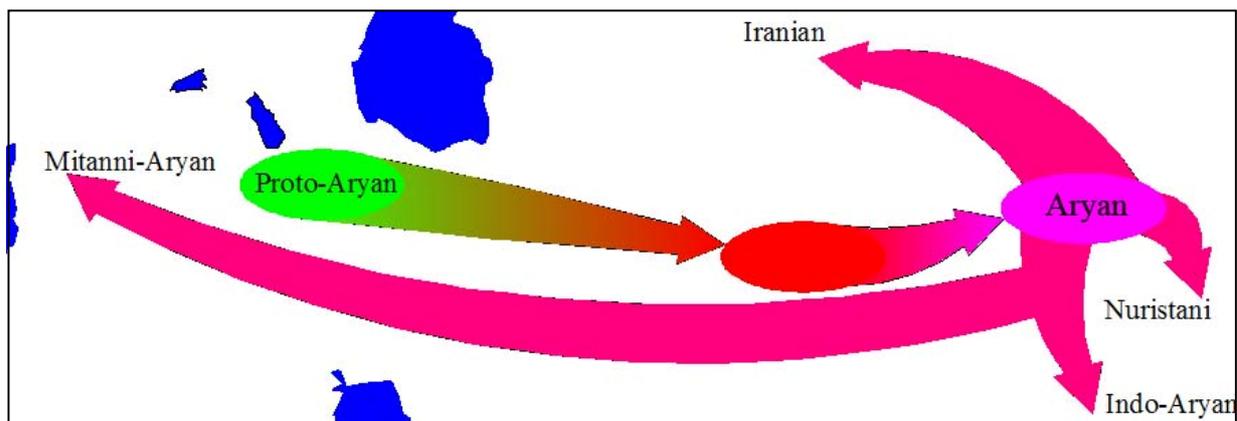
Proceeding from this hypothesis and taking into account all the facts presented, the initial area of spreading of the Old Aryan language most likely would be the northern part of the

Iranian plateau, where the appearance of Old Aryan tribes refers to the first half of the III millennium BC. The authors compare their appearance with the north-Iranian culture denoted as «Hissar II B» in VI-III millennium BC. [13, 14].

Hence, through Afghanistan Old Aryans could go further to the east to the Hindu Kush.

In the process of migration from Northwestern Iran through the Middle East Old Aryan language superimposed on the local Margiana substrate and as result was the Aryan language. In the culture of the Aryans was borrowed many new elements. After some time, the Aryan language migrated toward the Pamir and Hindu Kush, where occurred its disintegration into Nuristani, Mitanni Aryan and Indo-Iranian dialects. Judging by the disparate localization of late Aryan dialects, Aryans were equipped with chariots and horses and could make migrations in the east (India) and west (Mitanni) directions.

From Indo-Iranians (or Indo-Aryans) archaeologically mapped Gandhara culture or the culture of the Swat valley, which existed in the period 1600-500 BC in the territory of modern Pakistan. Pottery of this culture reveals its obvious similarity with the pottery Margiana civilization [15].



**Fig. 2.** Alleged scheme of migrations of Aryan Languages and haplogroups

To find a set of Y-haplogroups speakers of Old Aryan language, let's try to link its alleged ancestral home in north-western Iran with the spreading of Y-haplogroup in this area in the III-II millennium BC. According to preliminary data, they can be attributed to haplogroup J2a, J2b, G2a, R1b1b2 and R1a1. The age of these haplogroups in the Middle East is more than 10 thousand years [16].

#### Haplogroup J2

Haplogroup J2 (J2a, J2b) is currently the predominant (over 30%) in western Iran, is also represented in Afghanistan, among the Brahmins of the North-western India and Pakistan and Kalashs, [9, 17, 18].

#### Haplogroup G2a

In the Middle East with a frequency of 10-20% is found among the Kurds, Persians, Pashuns, Kalashs, Punjabis. In a small percentage it fixed among the Brahmins [10].

#### Haplogroup T

Among the peoples of the Middle East is currently a fairly rare haplogroup in amounts up to 8% noted among the southern Iranians (2.5%), the Pashtuns and Indo-Aryan Bhils in the North-West India (3.8%) [19].

#### Haplogroup R1b1b2

Submitted in Turkey (16.3%) [20], Iraq (11.3%) [21] and other countries in West Asia. In Central Asia, was found in Turkmenistan – 36.7% [12], Uzbeks – 9.8% [12], Tatars – 8.7% [22], Uighurs – up to 19.4% [23], as well as in the Bashkir [24]. In Pakistan – 6.8% [25] in India is insignificant – 0.55% [26].

Summarizing the above, it may be noted that haplogroup R1a1, J2 and G2a present among almost all modern speakers of the Aryan group languages.

To determine the possible presence of haplogroups J2a, J2b, G2a among the speakers of Old Aryan language the most important criterion is the age of the most common ancestor of Indian populations. It should be at least 4 thousand years. According to the A.A. Klyosov [18], age of J2a and J2b in India is more than 6 thousand years, which correlate with the alleged scheme. Klyosov notes the similarity of the Iranian and Indian J2 and indicates their migration from the Middle East through Iran to India. It is significant that this fact was rejected by Klyosov in that article [18], which is dictated, apparently, by his preconceived concept of haplogroups R1a1, as the only one haplogroup of the Aryan tribes.

Unfortunately, accurate data on the age of haplogroup G2a in India are not given, therefore, based on known data we can conclude that the initial speakers of Old Aryan language might have haplogroups J2 and, possibly, G2a.

#### Haplogroup R1a1 and Aryans

The emergence of haplogroup R1a1 among the speakers of Aryan language deserves special consideration. Migrated from the north-Iranian homeland to the east, Old Aryan speakers could assimilate with the local populations, which could lead to including new haplogroups in the Aryan gene pool.

Modern distribution of haplogroups in the Middle East shows that the frequency of haplogroup R1a1, starting with a small percentage in Western Iran (5%) gradually increases to almost 60% in Pakistan and Northern India [27], being present in different ethnic groups.

In this respect, it is not an unreasonable assumption that in the territory of Afghanistan or Pakistan in the II millennium BC speakers of the Aryan dialects interacted with the local population of R1a1.

Subsequently haplogroup R1a1 could be in the Y-DNA of Indo-Aryan tribes who invaded the north-western India not later the II millennium BC. Aryan migration from the North-western Iran through Afghanistan to India says infiltra-

tion such haplogroups as J2 and G2a and relatively young (later – Brahmin) branch of R1a1, brought by the dominant Aryan tribes (Fig. 3).

Given the extent of the Aryan languages of the Middle East, we can conclude that Old Aryan tribes, who invaded Afghanistan from the north-

western Iran, were at a high level of social organization that allowed them to transmit their language to the indigenous population of the Middle East and North India by large-scale assimilation.

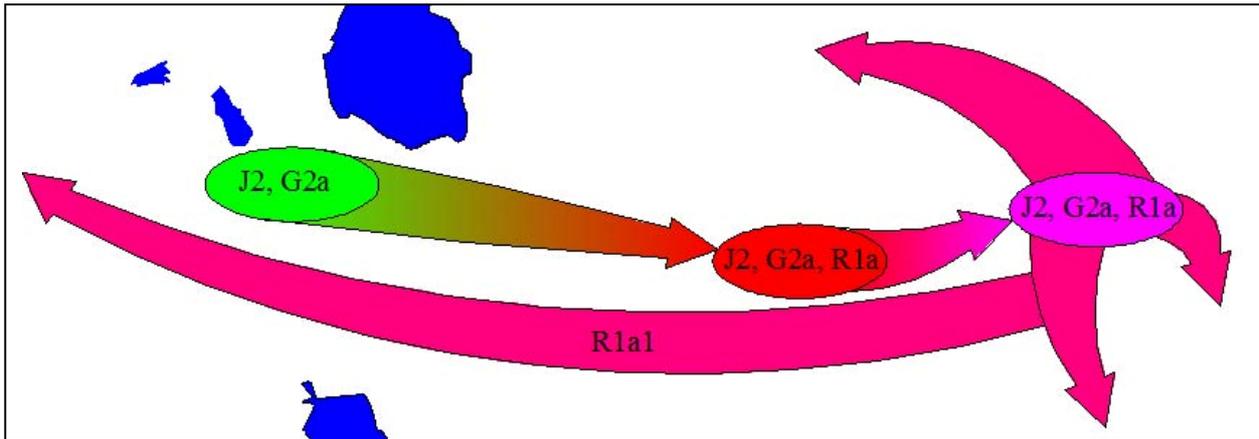


Fig. 3. Alleged scheme of migrations Y-haplogroups of Old Aryan language speakers

### **R1a1 in Eastern Europe as a consequence of migration of speakers of «ancient European dialects» through Central Asia to Europe**

As noted, the Eastern European (East Slavic) R1a1 is more ancient than that of Brahmins. How can we explain this fact? Let put forward the following assumption.

In the aforementioned hypothesis of Gamkrelidze and Ivanov allocation of «ancient European» dialect (the ancestors of Germanic, Italo-Celtic and Balto-Slavic languages) from Indo-European language occurred one of the first and went on with their subsequent migration to the east, through Central Asia and the Volga region to Europe. In this way the migration of the western group of Indo-European languages can be explained by their ancient lexical influence with Altaic, Finno-Ugric and the Yenisei languages [13]. Assuming the initial presence of the ancient Indo-European dialects in the Middle East, it is logical to allow the presence of «Middle Eastern» haplogroups among the speakers of «ancient European» dialects. The most suitable haplogroup is R1b1b2.

When moving «ancient European» tribes through the Middle East and Central Asia in the IV-III millennium BC they assimilated and included in their community the carriers of haplogroup R1a1. They gradually migrated to the north and further west and reached the modern Ukraine. This is indirectly confirmed by the fact that R1b1b2 is present in the gene pool of some Turkic peoples of Central Asia and the Finno-Ugric peoples of Russia [28, 29], located in the ways of «ancient European» tribes in Europe.

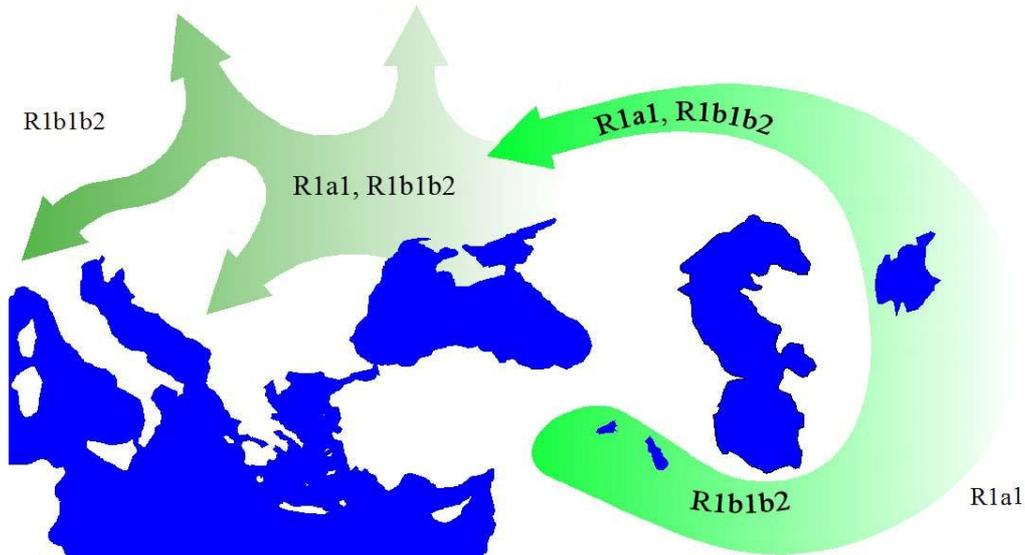
Summarizing the above, it can be assumed that there were two waves of migration of carriers R1a1 from the Middle East. The first wave in IV-III millennium BC migrated to the north with the speakers of «ancient European» dialects. Second wave in III-II millennium BC migrated with Aryans to the Pamir and Hindu Kush.

R1b1b2 haplogroup is prevalent among the peoples of Central and Western Europe, but the age of its subclades does not exceed 4500 years, which is comparable with the age of Slavic R1a1 [4, 30]. This may serve as indirect confirmation of the fact that these two haplogroups

at the same time about 5 thousand years ago migrated in Europe from Asia.

Migration of «ancient European» dialects from Central Asia to Europe accompanied by long intermediate settling in an area of the Northern Black Sea coast, not later than III – II

Yamna culture. By their gene pool at this stage they were carriers of haplogroup R1a1 and R1b1b2. Later R1a1 became dominant among the Slavic tribes, and R1b1b2 – among the speakers of Indo-European languages of Central and Western Europe (Fig. 4).



millennium BC. Archaeologically speakers of «ancient European» dialects can be compared to

**Fig. 4.** Alleged scheme of migrations Y-haplogroups of speakers of «ancient European» dialects

### Total reasoning

Given all these factors, the authors propose the following unified system of events:

1) The combination of linguistic and archaeological data homeland of Old Aryan language could be located on the territory of North-Western Iran in the region of culture Hissar B in III millennium BC where Old Aryans migrated to the east, south of Central Asia, in the area of Margiana civilization and beyond to the region of the Pamir and Hindu Kush.

2) Most likely, Old Aryans have several haplogroups and their gene pool can consist of subclades of J2 (and, possibly, G2a). These haplogroups are also represented among the Brahmins and the age of these populations is over 12 thousand years. In the gene pool of the orig-

inal speakers of «ancient European» dialects present haplogroup R1b1b2.

3) During the migration of speakers of «ancient European» dialects through the Middle East and Central Asia, and further through the Volga and the northern Black Sea region to Europe, in their gene pool was involved the R1a1 haplogroup. Later the R1a1 haplogroup become dominant among the eastern Slavs, R1b1b2 – among the peoples of Central and Western Europe.

4) In the period of stay of the ancient Aryans in the territory of Margiana in II millennium BC in their gene pool could be included haplogroup R1a1, later this became dominant among the Brahmins.

## References

1. Денисов И. В. Некоторые проблемы археологии бронзового века Волго-Уралья и Ведийско-Авестийские сказания // В центре Евразии: Сборник научных трудов / Отв. Ред. В. А. Иванов. — Стерлитамак: Стерлитамак. Гос. пед. ин-т, 2001. С. 4-21.
2. Кузьмина Е.Е. Откуда пришли индоарии. Москва, 1994.
3. Michael Witzel, The home of the Aryans, Harvard University.
4. Сарианиди В.И. Древние земледельцы Афганистана, М., Наука, 1977 г.
5. Сарианиди В.И. В поисках страны Маргуш. М., 1993.
6. Клёсов А.А. Откуда появились славяне и «индоевропейцы» и где их прародина? Ответ дает ДНК-генеалогия. Вестник Российской Академии ДНК-генеалогии. 1, 400-477 (2008а).
7. А. А. Клёсов. Ещё одно доказательство перехода ариев (гаплогруппа R1a1) в Индию и Иран с Русской равнины. Вестник Российской Академии ДНК-генеалогии. Том 2, № 7, декабрь, 2009 г.
8. Клёсов А.А. Древнейшие восточно-азиатские ветви гаплогруппы R1a. Вестник Российской Академии ДНК-генеалогии. т. 2, №5, 2009 г.
9. Underhill et al. (2009), «Separating the post-Glacial co-ancestry of European and Asian Y chromosomes within haplogroup R1a», *European Journal of Human Genetics*, doi:10.1038/ejhg.2009.194.
10. Sengupta et al. (2005), «Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists», *Am. J. Hum. Genet.* 78 (2): 202-21, doi:10.1086/499411, PMID 16400607.
11. Sharma et al. (2009), «The Indian origin of paternal haplogroup R1a1(\*) substantiates the autochthonous origin of Brahmins and the caste system», *J. Hum. Genet.* 54 (1): 47-55, doi:10.1038/jhg.2008.2, PMID: 19158816.
12. Firasat S, Khaliq S, Mohyuddin A, et al (January 2007). «Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan». *Eur. J. Hum. Genet.* 15 (1): 121-6. doi:10.1038/sj.ejhg.5201726. PMID 17047675. PMC 2588664.
13. O. Szemerényi. Structuralism and substratum – Indo-Europeans and Semites in the Ancient Near East (LINGUA. International Review of General Linguistics, v. 13: 1-29) 1977. ero же Sprachtypologie, funktionelle Belastung und die Entwicklung indo-germanischer Lautsysteme («Textes et Memoires», v. V. Acta Ir»-nica, Leiden. E. J. Brill : 339—393), 1977.
14. A. Lubotsky. The Indo-Iranian substratum in Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations. Papers presented at an international symposium held at the Tvärminne Research Station of the University of Helsinki 8-10 January 1999. (Mémoires de la Société Finno-ougrienne 242.) Chr. Carpelan, A. Parpola, P. Koskikallio (eds.). Helsinki 2001, 301-317.
15. Thomas L. H. 1970. New evidence for dating the Indo-European dispersal in Europe (Indo-European and Indo-Europeans. Papers presented at the Third Indo-European Conference at the University of Pennsylvania, ed. by G. Cardona, H. M. Hoenigswald, and A. Senn, Philadelphia. University of Pennsylvania Press: 199-215).
16. Т.В. Гамкрелидзе, Вяч. Вс. Иванов. Индоевропейский язык и индоевропейцы. Реконструкция и историко-типологический анализ праязыка и протокультуры. Тбилиси, 1984.
17. Parpola A.: Margiana and the Aryan Problem. 1993. International Association for the Study of the Cultures of Central Asia Information Bulletin 19:41-62.
18. Bryant E. The Quest for the Origins of Vedic Culture. — Oxford University Press, 2001. — ISBN 0-19-513777-9
19. International Society of Genetic Genealogy (ISOGG) - Y-DNA Haplogroup R and its Subclades – 2009;
20. Ornella Semino et al., «Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area», *American Journal of Human Genetics* 74:1023–1034, 2004.
21. Semino O, Passarino G, Oefner PJ, et al. (November 2000). «The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective». *Science* 290 (5494): 1155–9. doi:10.1126/science.290.5494.1155. PMID 11073453.
22. J. R. Luis et al.: The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations (Errata), *American Journal of Human Genetics*.
23. Sengupta et al. – Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists., *Am J Hum Genet.* 2006 February.
24. А.А.Клёсов. Гаплогруппа J2 в Индии и России. Возраст предков. Вестник Российской Академии ДНК-генеалогии, Том 2, № 5, 2009 г.
25. Sanghamitra Sahoo, Anamika Singh, G. Himabindu, Jheelam Banerjee, T. Sitalaximi, Sonali Gaikwad, R. Trivedi, Phillip Endicott, Toomas Kivisild, Mait Metspalu, Richard Villems, and V. K. Kashyap, «A prehistory of Indian Y chromosomes: Evaluating demic diffusion scenarios,» *Proceedings of the National Academy of Sciences of the United States of America.* Published online on January 13, 2006, 10.1073/pnas.0507714103.[1] (cf. Supporting Figure 3 in online data supplement).
26. 76/523, Y-Chromosome Excavating Y-chromosome haplotype strata in Anatolia, Cinnioglu et al. 2004.
27. R. Spencer Wells et al., «The Eurasian Heartland: A continental perspective on Y-chromosome diversity,» *Proceedings of the National Academy of Sciences of the United States of America* (en:August 28, en:2001).
28. Tambets et al. (2004).
29. Yali Xue, Tatiana Zerjal, Weidong Bao, Suling Zhu, Qunfang Shu, Jiujiu Xu, Ruofu Du, Songbin, Pu Li, Matthew E. Hurles, Huanming Yang, Chris Tyler-Smith, «Male demography in East Asia: a north-south contrast in human population expansion times,» *Genetics* 2006.
30. A. S. Lobov et al. — Y chromosome analysis in subpopulations of Bashkirs from Russia, 2005.
31. Qamar et al. (2002), Cruciani et al. (2004), Semino et al. (2004), Underhill et al. (2000).
32. 13/176 in Pakistan and 4/728 in India, 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, Sengupta et al. 2008.

27. Underhill et al. (2009), «Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a», *European Journal of Human Genetics*, doi:doi:10.1038/ejhg.2009.194.
- Qamar et al. (2002), «Y-Chromosomal DNA Variation in Pakistan», *The American Journal of Human Genetics*, doi:10.1086/339929.
- Wells et al. (2001), «The Eurasian Heartland: A continental perspective on Y-chromosome diversity», *Proc. Natl. Acad. Sci. U. S. A.* 98 (18): 10244-9, doi:10.1073/pnas.171305098.
28. V. N. Kharkov, V. A. Stepanov, O. F. Medvedeva, M. G. Spiridonova, M. I. Voevoda, V. N. Tadinova and V. P. Puzyrev, «Gene pool differences between Northern and Southern Altaians inferred from the data on Y-chromosomal haplogroups», *Russian Journal of Genetics*, Volume 43, Number 5 / May, 2007.
29. Rosser et al. (2000) Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language. *Am J Hum Genet*; 67:1526-1543.
30. В. Веренич. К проблеме ирландских гаплотипов группы R1b1b2 (атлантического модального гаплотипа): опыт ДНК-генеалогического исследования. *The Russian Journal of Genetic Genealogy (Русская версия)*, Vol 1, No 2 (2009).

## Origin of “Jewish” clusters of E1b1b1 (M35) haplogroup

A.A. Aliev

### Preface

Among the haplogroups represented among modern Jews, the frequency of more than 10% can be divided into three of them. They are J1 (M267), J2 (M172) and E1b1b1 (M35) [1]. According to recent studies, J1 and J2 claim the role of «haplogroups of Abraham», the legendary ancestor of the Jews and Arabs [2]. Despite the fact that the genealogical aspect of Jewish history is studied in sufficient detail [1, 3], the question about appearing in the Jewish community of the various «Jewish clusters» of subclades of the E1b1b1 haplogroup so far have been neglected. This leads us to the question of how and when were they formed?

### «Jewish» clusters and their ancestors

The cluster is a set of haplotypes, which goes back to a separated ancestor, a sort of independent branch within the tree of subclade.

According to Haplozone E3b Project [4], it is known that there are four subclades of E1b1b1\* haplogroup (M35), within which, among others, there are several «Jewish» clusters: E1b1b1\* (unclassified), E1b1b1a3\* (V22), E1b1b1c1\* (M34) and E1b1b1c1a\* (M84).

**Unclassified subclade E1b1b1\*** can be considered as a subclade of E1b1b1\* (M35) haplogroup with an unidentified SNP-mutation. Therefore determination of its age is still difficult to ascertain. This cluster has been found in Iraq, two persons (out of 218 tested) [5].

**Subclade E1b1b1a3\* (V22)** originated about 5100 years ago in Egypt. Later, its representatives settled in different countries, including Palestine, where V22 is found among the Palestinian Arabs and Samaritans [6, 7].

**Subclade E1b1b1c1\* (M34)** presumably originated in the late period of the Upper Paleolithic (about 10 thousand years ago). The highest frequency and diversity of its haplotypes observed are among the population of Lebanon, Syria and the adjoining region of Turkey [8, 9, 10].

Judging by various open Y-DNA projects [4, 11], **subclade E1b1b1c1a\* (M84)** is mainly represented in the same region as its ancestor subclade M34.

Given that all listed subclades have been from the Middle East and the Eastern Mediterranean, you can prevent their presence in the region even in the era of the formation of Jewish nation.

According to the classification of the E3b Project [4] «Jewish» clusters are designated as E1b1b1\*-C, E1b1b1\*-D, E1b1b1a3\*-E, E1b1b1c1\*-D1, E1b1b1c1a\*-A, E1b1b1c1a\*-B and E1b1b1c1a\*-C\*. Judging by their surnames, representatives of these clusters are Ashkenazim [12].

Single members with non-Jewish surnames, obviously, are the baptized Jews. This fact indicates that the listed clusters presumably originated during the times of mass migrations of the ancestors of modern Ashkenazim from the Middle East to Europe, deep into the Germanic lands. To identify the circumstances of the emergence of these clusters in Europe is necessary to calculate their ages.

For study we used 37- and 67-marker haplotypes from the Haplozone E-M35 databases. The times of most recent ancestors (TMRCA) were calculated by A.A. Klyosov's algorithm [13] (see Table 1.) and assumes that one generation is 25 years. The calculation for the cluster E1b1b1c1a-A did not take place due to the small number of haplotypes (N=2).

**Table 1**

Cluster	TMRCA	Modal haplotype (12 markers)
E1b1b1*-C (N=12, 67 m.)	(1075±175 years) X century AD	13-24-14-10-16-17-11-12-13-14-11-32
E1b1b1*-D (N=10, 37 m.)	(1825±300 years) II-III centuries AD	14-24-13-10-15-18-11-12-11-12-11-30
E1b1b1a3*-E (N=14, 37 m.)	(1125±200 years) IX-X centuries AD	14-24-14-10-17-18-11-12-12-12-11-29
E1b1b1c1*-D1 (N=32, 67 m.)	(1000±130 years) XI century AD	14-25-13-9-17-18-11-12-12-13-11-30
E1b1b1c1a-B (N=24, 67 m.)	(1125±150 years) IX-X centuries AD	13-24-13-10-17-18-11-12-12-13-11-30
E1b1b1c1a-C (N=9, 37 m.)	(1800±300 years) III century AD	13-25-13-10-16-16-11-12-12-13-11-31

As we can see, two of six clusters (E1b1b1\*-D and E1b1b1c1a-C) appeared in the II-III centuries AD. The last (E1b1b1\*-C, E1b1b1a3\*-E, E1b1b1c1\*-D1 and E1b1b1c1a-A) appeared in the IX-XI centuries AD. What happened in the Jewish history in these periods? Who could be the ancestors of these clusters?

### Bar Kochba and the period of Geonim

Based on historical evidence, the emergence of clusters E1b1b1\*-D and E1b1b1c1a-C in **II-III centuries AD** can be associated with one of the two waves of migration in Central Europe. One wave came from Gaul, from the area of the river Rhine, where the Latin-speaking Jews as citizens of the Roman Empire have lived since the beginning of AD. Another wave is linked to the uprising in Judea, led by Bar Kochba (132-135 years AD). After the suppression of this uprising the Jewish population was stolen into slavery in Rome. Jerusalem was plowed, and in its place the new city of Aelia Capitolina was built. Conversion to Judaism first widely prac-

ticed throughout the empire, was now banned. Scaling missionary activities of Judaism came to an end.

**IX-XI centuries** in Jewish history occur in the so-called sunset of the period of Geonim (583-1040 years AD) [14]. Gaons was the Jewish religious leaders in the VI - XI centuries. They had the highest authority in the interpretation of the Talmud, and they were heads of yeshivas (the highest religious schools to study the Talmud). Their centers were the cities of Sura and Pumbedita in the territory of modern Iraq. In IX-XI centuries, the situation in the Baghdad Caliphate had deteriorated therefore the bulk of the Jewish population had to move far to the west, to Europe in search of a better life. Around 1040 yeshiva of Sura was closed completely, therefore this year was widely regarded as the date of the end of the period of Geonim. It then followed that the Center for the Study of Torah was moved from the land of Israel and the banks of the Tigris and Euphrates rivers in Europe.

## Summing

All listed gives grounds to assert the following:

1) Various subclades of E1b1b1 (M35) haplogroup reveal an ancient presence in the Middle East in the times of the formation of the Jewish nation.

2) «Jewish» clusters of E1b1b1 (M35) haplogroup (E1b1b1\*-C, E1b1b1\*-D, E1b1b1a3\*-E, E1b1b1c1\*-D1, E1b1b1c1a\*-A, E1b1b1c1a\*-B and E1b1b1c1a\*-C) are Ashkenazim. They originated at different times: E1b1b1\*-D and E1b1b1c1a-C are originated in the II-III centuries AD; E1b1b1\*-C, E1b1b1a3\*-E, E1b1b1c1\*-

D1, E1b1b1c1a\*-B were originated in the IX-XI centuries AD.

3) TMRCA of these clusters can bind their appearance in Europe with the following historical facts:

a) Relocation of the Jews from Gaul (at the beginning of BC) or the massive capture of Jews into slavery after the suppression of the Bar Kochba (clusters E1b1b1\*-D and E1b1b1c1a\*-C);

b) Start of the resettlement of Jews in Europe due to decline of the caliphate of Baghdad and the end of period of Geonim (clusters E1b1b1\*-C, E1b1b1a3\*-E, E1b1b1c1\*-D1, E1b1b1c1a\*-B).

## References

1. Coffman-Levy. A mosaic of people: the Jewish story and a reassessment of the DNA evidence. *Journal of Genetic Genealogy* 1: 12-33, 2005.
2. А.А. Клёсов. Загадки «модального гаплотипа кознов». *Вестник Российской Академии ДНК-генеалогии*, № 3, 2008 г.
3. Клёсов А. А. Происхождение евреев с точки зрения ДНК-генеалогии.
4. Haplozone E-M35.
5. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS (2004) Origin, Diffusion, and Differentiation of YChromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area.
6. С.В. Лутак, А.А. Клёсов. Гаплогруппа E1b1b1a (M78) – современные потомки древних египтян. *Вестник Российской Академии ДНК-генеалогии* 2 (4): 639-669, апрель 2009 г.
7. B. Bonn'e-Tamir et al. Maternal and Paternal Lineages of the Samaritan Isolate: Mutation Rates and Time to Most Recent Common Male Ancestor, 2003. P. Shen et al. Reconstruction of Patrilineages and Matrilineages of Samaritans and Other Israeli Populations From Y-Chromosome and Mitochondrial DNA Sequence Variation, 2004.
8. C. Cinnioğlu et al. (2003), «Excavating Y-chromosome haplotype strata in Anatolia». *Hum Genet* (2004) 114 : 127-148. DOI 10.1007/s00439-003-1031-4.
9. Flores et al. (2005), «Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan», *J Hum Genet* 50: 435-441, doi:10.1007/s10038-005-0274-4.
10. Mirvat El-Sibai, Daniel E. Platt, Marc Haber, Yali Xue, Sonia C. Youhanna, R. Spencer Wells, Hassan Izaabel, May F. Sanyoura, Haidar Harmanani, Maziar Ashrafian A. Bonab, Jaafar Behbehani, Fuad Hashwa, Chris Tyler-Smith, Pierre A. Zalloua. Geographical Structure of the Y-chromosomal Genetic Landscape of the Levant: a coastal-inland contrast. *Annals of human genetics*, 2009.
11. The E-M35 Phylogeny Project.
12. Происхождение еврейских фамилий; «Ashkenazic Family Names Origin and Development».
13. Клёсов А.А. Руководство к расчёту времен до общего предка гаплотипов Y-хромосомы и таблица возвратных мутаций. *Вестник Российской Академии ДНК-генеалогии* Том 1, № 5, 2008 ноябрь.
14. Сесиль Рот. История евреев с древнейших времён по Шестидневную войну. Пер. с англ., 1978 г.

## Modern carriers of haplogroup E1b1b1c1 (M34) are the descendants of the ancient Levantines

A.A. Aliev,  
Bob Del Turco

### Who, where, when

The homeland of haplogroup E1b1b1c1 (M34) is placed in a relatively small region of the Middle East, covering south-east Asia Minor and the Levant areas (Syria and Palestine) [1]. This opinion is based on the fact that it is here presented as the haplogroup E1b1b1c1 \* (M34), and its known subclades: E1b1b1c1a \* (M84), E1b1b1c1a1 \* (M136) and E1b1b1c1b \* (M290) [2, 3, 4, 5, 6]. It may be the result of the long-term presence of this haplogroup. The haplogroup was found in the Eastern Mediterranean countries, in the European Mediterranean countries, the British Isles [7, 8, 9, 10, 11] as well as on the Arabian peninsula, but with relatively low diversity [12, 13, 14].

The following paper will help to better understand the history of the haplogroup and how it occurred.

### Time of occurrence

First, let's determine and clarify the time of the first appearance of this haplogroup. That is, to calculate the age of the most recent common ancestor of all modern carriers of this haplogroup. It requires a large sample of «long» haplotypes (preferably 67-marker). It should be noted that the calculated age will determine only the approximate date, after which the haplogroup could not appear, the lower temporal boundary, which may not always coincide with its true age. For this calculation we will use the manual [15], assuming that one generation is 25 years.

If you have to deal with rare haplogroups, one will inevitably have to be content with modest samples. But with too few (less than 10) numbers of haplotypes used for analysis, calculating the age of the haplogroup has no meaning: in this case the age would be underestimated.

Analysis of the history of the settlement of this haplogroup is easier if we can identify one or more clusters in this haplogroup. Cluster is the independent branch, forming one separate haplotype. Calculating the age of the cluster, and knowing the geography of its spread can more clearly elucidate the history of the settlement of the carriers of this haplogroup. How exactly? It will become clear below.

According to the classification of the Haplozone E3b Project [16], the known haplogroup clusters of E1b1b1c1 are identified as E1b1b1c1\*-A, E1b1b1c1\*-B, E1b1b1c1\*-C, E1b1b1c1\*-D1 («Jewish cluster») and E1b1b1c1\*-D2.

Each of these clusters has its own peculiarities.

E1b1b1c1\*-A is the «European» cluster, discovered among the Germans and the Spaniards;

E1b1b1c1\*-B is the «Arabian» cluster found among the Arabs from Persian Gulf countries;

E1b1b1c1\*-C is the «British» cluster found among the British and Irish.

E1b1b1c1\*-D1 is the «Jewish» cluster found among Ashkenazi. History of this cluster (about 1000 years ago) was considered in another paper [17] and will not be considered here.

E1b1b1c1\*-D2 is the «mixed» cluster, found both among Europeans and people of the Levant and Turkey.

Let us calculate the ages of these clusters. It is necessary to mention that due to the small number of haplotype clusters of E1b1b1c1\*-A, E1b1b1c1-B and E1b1b1c1-C, the probability is very high that we will observe very «rejuvenation» ages. For example, the cluster E1b1b1c1\*-A (N=4, 37 markers), such age is  $3525 \pm 650$  years. Ancestor of E1b1b1c1-B (N=2, 67 markers) lived  $350 \pm 175$  and E1b1b1c1-C (N=3, 25 markers) –  $750 \pm 400$  years ago. Although the ages are approximate, it shows that these clusters occurred in different epochs.

More plausible results can be expected when calculating the age of the cluster E1b1b1c1\*-D2, whose sample consisted of 32 67-marker haplotypes. Its age was  $3850 \pm 450$  years. The mixture of nations in this cluster indicates that the founding father was born 3400-4300 years ago in the Levant. Part of his descendants later migrated to Europe. This confirms the close age proximity of the cluster E1b1b1c1\*-A, equal to  $3525 \pm 650$  years. Apparently, the emergence of these two clusters are linked to the same period in the history of the Middle East.

To determine the age of the common ancestor of all E1b1b1c1, authors compiled samples with the involvement of 9-markers of Lebanese, Syrian, Palestinian and Turkish haplotypes from the papers [2, 3], the modal haplotypes of all noted clusters and haplotypes are not related to any of the known clusters and designated as E1b1b1c1\*-Miscellaneous (N=51, 9 markers). Age of the most recent common ancestor of all modern carriers of E1b1b1c1\* is  $7000 \pm 850$  years.

### Ethnolinguistic portrait of the ancient Levant

The analysis suggests the presence of haplogroup E1b1b1c1 in the peoples of western Asia Minor with V millennium BC. e. These people are remarkable as the creators of the first civilizations of the world, and laid the foundations of social and cultural development of mankind. The presence of numerous archaeological and written records will trace the historical path of these peoples.

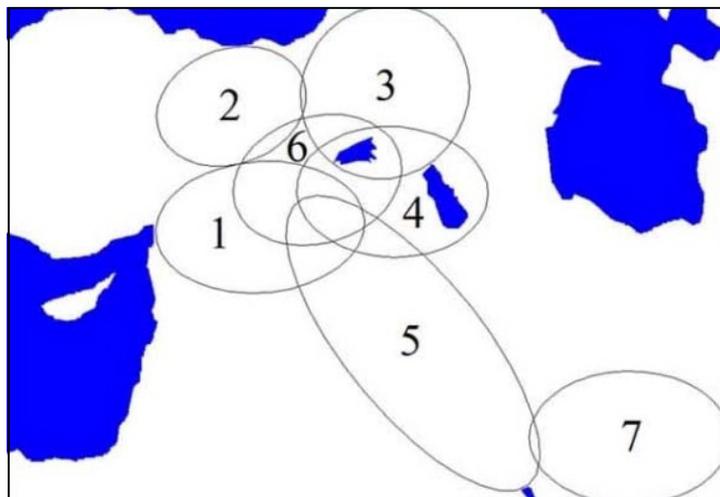
According to the hypothesis of TV Gamkrelidze and VV Ivanov, not later than the V-IV millennium BC in the Middle East began the gradual disintegration of Indo-European, Proto-Semitic and Proto-Kartvelian languages and their interaction with other languages of the region [18, 19] (Fig. 1).

Over time, these proto-languages began to break up into smaller subgroups. Among all these languages, Semitic, Hurrian and Indo-European languages have relevance to the area of Levant and Southeast Anatolia, so one can assume that the ancient native of haplogroup E1b1b1c1 could be from this area [20]:

The people who lived in the III-II millennium BC spoke the **North-west Semitic languages**, Proto-Canaanite and Aramaic subgroups – Amorite, Ugaritic, Old Canaanite, Phoenician, Hebrew, Moab, and the Aramaic dialects. Sometimes Pre-Jewish ancient peoples of Palestine – the Amorites, the Phoenicians, Moabites etc. – collectively called the Canaanites. The greatest contribution to world civilization of the Canaanites is the invention of alphabetic writing.

*Amorites* are known as the founders of the first royal dynasty of Babylon, the most famous representative of which was Hammurabi, the creator of the Code of Laws.

The *Phoenicians* created a powerful civilization with advanced craft and maritime trade. The Phoenician alphabet became one of the first recorded in the history of systems of syllabic phonetic letters.



**Fig. 1.** Alleged scheme of spreading of languages in the Middle East in the IV-III millennium BC.  
1 – Semitic («Paleo-Canaanian»), 2 – Proto-Chattian, 3 – Proto-Kartvelian 4 – Hurro-Urartian,  
5 – Sumerian, 6 – Proto-Indo-European; 7 – Elamite

The Jews are known as the founders of Judaism – one of the first monotheistic religions of the world, from which later emerged Christianity and Islam. Canaanian pantheon influenced Jewish demonology, in which the names of Canaanite gods turned into epithets of servants of evil – Beelzebub, Vaalberit, Astaroth, etc.

*Aramaeans* never formed a united nation and did not have a single state. Nevertheless, their language playing the role of lingua franca in large parts of the Middle East. It was the official language of the Persian Achaemenid Empire, the spoken language of Palestine at the time of Jesus Christ.

The **Indo-European family** in this area in the III-II millennium BC was presented by both the Hittite and Mitanni Aryan language.

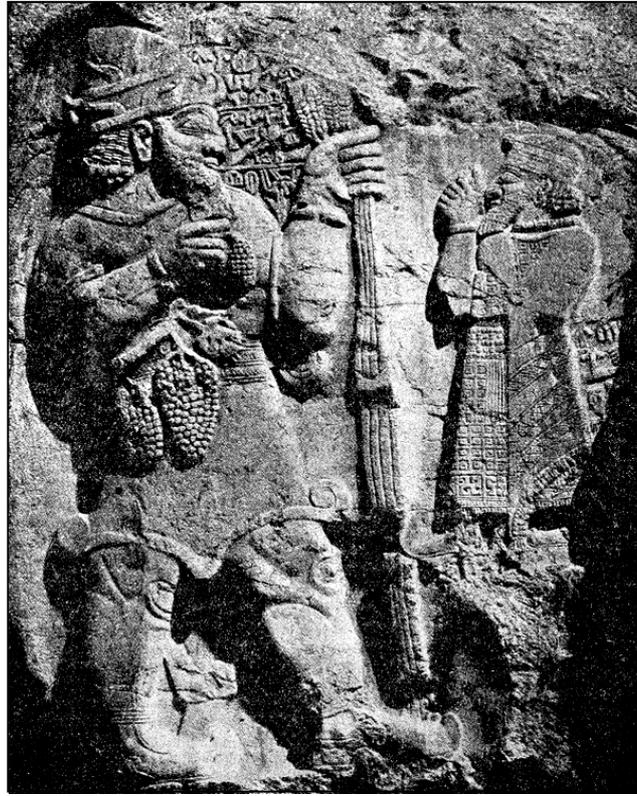
*Hittites* were the first Indo-European people who created a state – the Hittite kingdom. From Chattites they took over the processing technology of iron, which was the guarded secret of Chattites.

*Mitanni Aryans* were part of the population of the ancient kingdom of Mitanni, who spoke with a separate Aryan language. They were known due to handicrafts of training horses.

Assuming an ancestry of the Proto-Indo-European language in North Syria, the migration of people from the Middle East to Europe around 3400-4300 years ago and the formation of clusters E1b1b1c1-D2 and E1b1b1c1-A can be attributed with migration speakers of *Proto-Hellenic* and «*Old Balkanian*» dialects from Asia Minor to the Balkans.

The **Hurrian** population, along with the Semitic, lived in the II millennium BC in parts of northern Syria and southeastern Anatolia. Hurrians created the kingdom Urkish and Nawar, Mitanni and Kizzuwatna, as well as a number of city-states from Palestine to Mesopotamia. In the Bible, among the inhabitants of Pre-Jewish Palestine are noted Horites the small groups of Semitized Hurrits retains its tribal designation until the first centuries of the I millennium BC.

By their anthropological type ancient and modern populations of Levant and Asia Minor refers to the type of the Armenoid (or Assyroid) race, known by ancient monuments in Asia Minor. This type is characterized by a pronounced brachycephaly, enhanced development of hair on the face and body, a unique form of the nose (look at Hittite reliefs, Fig. 2) [21].



**Fig. 2.** The external appearance of the ancient Levantines (Hittite king, praying god of fertility)

Although with time almost all the ancient languages of the Levant later were replaced by Arabic, Turkish and Kurdish languages, physical displacement of the population did not happen.

It is evidenced by the continuous presence of haplogroup E1b1b1c1 for thousands of years and remained unchanged anthropological type of the population.



**Fig. 3.** Modern Levantine

So today, Palestinians, Jordanians, Lebanese, Syrians, and some Turks, are among those of haplogroup E1b1b1c1 and direct descendants of ancient peoples – the creators of the civilizations of the Eastern Mediterranean.

### Dry residue

1) Haplogroup E1b1b1c1\* (M34) was born in an area of modern south-east Turkey, Syria, Lebanon and Palestine about 7000 years ago.

2) In ancient times, carriers of this haplogroup were peoples of civilizations of the Eastern Mediterranean and Asia Minor, who spoke the Northwest Semitic languages, Anatolian and Mitannian Aryan languages, as well as Hurrites language.

3) Approximately 3400-4300 years ago some carriers of E1b1b1c1 migrated from the Middle East to Europe, poured into the European nations, forming clusters E1b1b1c1-D2 and E1b1b1c1-A.

### References

1. A. Lancaster. Y Haplogroups, Archaeological Cultures and Language Families: a Review of the Possibility of Multidisciplinary Comparisons Using Case of M-35. *Journal of Genetic Genealogy*, 5(1):35-65, 2009.
2. Flores et al. (2005), «Isolates in a corridor of migrations: a high-resolution analysis of Y-chromosome variation in Jordan», *J Hum Genet* 50: 435-441, doi:10.1007/s10038-005-0274-4
3. C. Cinnioğlu et al. (2003), «Excavating Y-chromosome haplotype strata in Anatolia». *Hum Genet* (2004) 114 : 127-148. DOI 10.1007/s00439-003-1031-4
4. Mirvat El-Sibai, Daniel E. Platt, Marc Haber, Yali Xue, Sonia C. Youhanna, R. Spencer Wells, Hassan Izaabel, May F. Sanyoura, Haidar Harmanani, Maziar Ashrafian A. Bonab, Jaafar Behbehani, Fuad Hashwa, Chris Tyler-Smith, Pierre A. Zalloua. Geographical Structure of the Y-chromosomal Genetic Landscape of the Levant: a coastal-inland contrast. *Annals of human genetics*, 2009.
5. Coffman-Levy (2005), «A mosaic of people: the Jewish story and a reassessment of the DNA evidence», *Journal of Genetic Genealogy* 1: 12-33
6. Cruciani et al. (May 2004), «Phylogeographic Analysis of Haplogroup E3b (E-M215) Y Chromosomes Reveals Multiple Migratory Events Within and Out Of Africa», *American Journal of Human Genetics*.
7. Flores et al. (2004), «Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography», *European Journal of Human Genetics* 12: 855-863, doi:10.1038/sj.ejhg.5201225
8. Gonçalves et al. (2005), «Y-chromosome Lineages from Portugal, Madeira and Açores Record Elements of Sephardim and Berber Ancestry», *Annals of Human Genetics* 69: 443-454, doi:10.1046/j.1529-8817.2005.00161.x
9. Semino et al. (2004), «Origin, Diffusion, and Differentiation of Y-Chromosome Haplogroups E and J: Inferences on the Neolithization of Europe and Later Migratory Events in the Mediterranean Area», *American Journal of Human Genetics* 74: 1023-1034, doi:10.1086/386295
10. Bosch et al. (2006), «Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns», *Annals of Human Genetics* 70: 459-487, doi:10.1111/j.1469-1809.2005.00251.x
11. Haplozone E3b Project
12. Haplozone E3b, Arabian-E-Y Dna Project, Arab DNA Project
13. Cadenas et al. (2007), «Y-chromosome diversity characterizes the Gulf of Oman», *European Journal of Human Genetics* 16: 1-13, doi:10.1038/sj.ejhg.5201934
14. Arredi et al. (2004), «A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa», *American Journal of Human Genetics* 75: 338-345, doi:10.1086/423147
15. Клёсов А. А. Руководство к расчёту времен до общего предка гаплотипов Y-хромосомы и таблица возвратных мутаций. *Вестник Российской Академии ДНК-генеалогии* Том 1, № 5, 2008, ноябрь.
16. Haplozone E-M35
17. Aliev AA. Origin of «Jewish» clusters of E1b1b1 (M35) haplogroup. *RJGG Vol.2 No.1 2010*.
18. Т. В. Гамкрелидзе, Вяч. Вс. Иванов. Индоевропейский язык и индоевропейцы. Реконструкция и историко-типологический анализ праязыка и протокультуры. Тбилиси, 1984.
19. Иллич-Свитыч. В. М. Опыт сравнения ностратических языков. М., Наука, 1971.
20. Дьяконов И.М. Языки древней Передней Азии. М.: Издательство «Наука», Главная редакция восточной литературы, 1967.
21. В. П. Алексеев. География человеческих рас. М., Мысль, 1974.