Russian Journal of Genetics, Vol. 37, No. 7, 2001, pp. 807–822. Translated from Genetika, Vol. 37, No. 7, 2001, pp. 974–990.

HUMAN GENETICS

The Russian Gene Pool: Gene Geography of Surnames

O. P. Balanovsky[1], A. P. Buzhilova[2], and E. V. Balanovskaya[3]

[1] *Research Center, for Medical Genetics Russian Academy of Medical Sciences, Moscow, 115478 Russia*
[2] *Institute of Archeology, Russian Academy of Sciences, Moscow, 117036 Russia*
[3] *Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, 123182 Russia; fax: 324-07-02; e-mail: balanovska@mtu-net.ru*

Received December 7, 2000

Abstract—Surnames are traditionally used in population genetics as "quasi-genetic" markers (i.e., analogs of genes) when studying the structure of the gene pool and the factors of its microevolution. In this study, spatial variation of Russian surnames was analyzed with the use of computer-based gene geography. Gene geography of surnames was demonstrated to be promising for population studies on the total Russian gene pool. Frequencies of surnames were studied in 64 sel'sovets (rural communities; a total of 33 thousand persons) of 52 raions (districts) of 22 oblasts (regions) of the European part of Russia. For each of 75 widespread surnames, an electronic map of its frequency was constructed. Summary maps of principal components were drawn based on all maps of individual surnames. The first 5 of 75 principal components accounted for half of the total variance, which indicates high resolving power of surnames. The map of the first principal component exhibits a trend directed from the northwestern to the eastern regions of the area studied. The trend of the second component was directed from the southwestern to the northern regions of the area studied, i.e., it was close to latitudinal. This trend almost coincided with the latitudinal trend of principal components for three sets of data (genetic, anthropological, and dermatoglyphical). Therefore, the latitudinal trend may be considered the main direction of variation of the Russian gene pool. The similarity between the main scenarios for the genetic and quasi-genetic markers demonstrates the effectiveness of the use of surnames for analysis of the Russian gene pool. In view of the dispute between R. Sokal and L.L. Cavalli-Sforza about the effects of false correlations, the maps of principal components of Russian surnames were constructed by two methods: through analysis of maps and through direct analysis of original data on the frequencies of surnames. An almost complete coincidence of these maps (correlation coefficient $\rho = 0.96$) indicates that, taking into account the reliability of the data, the resultant maps of principal components have no errors of false correlations.

INTRODUCTION

The method of the use of surnames as analogs of genetic markers was developed in detail by Crow and Mange [1]. It has been further developed and widely used by Russian researchers. Yu.G. Rychkov, A.A. Revazov, E.K. Ginter, their coworkers, and many other researchers used surnames as quasi-genetic markers (the term coined by A.A. Revazov). Direct comparison of the results obtained with the use of two types of markers, genetic and quasi-genetic, demonstrated that the use of surnames for studying the gene-pool structure can yield reliable results. The estimates of interpopulation diversity with respect to genes ($F_{ST} \times 10^2$) and surnames ($f_r \times 10^2$) were close to each other. For example, in Adygeans, $F_{ST} = 0.69$ and $f_r = 0.60$; in Lithuanians, $F_{ST} = 1.24$ and $f_r = 1.41$; in Primorye populations, $F_{ST} = 5.73$ and $f_r = 5.85$; in Armenians, $F_{ST} = 6.16$ and $f_r = 6.00$; in Altaians, $F_{ST} = 6.59$ and $f_r = 7.51$; in Nivkhs, $F_{ST} = 6.73$ and $f_r = 5.88$; and in Evenks, $F_{ST} = 8.29$ and $f_r = 8.95$ [2, 3]. Surnames (or clan affiliation as their analog) were successfully used when studying the genetic structure of Evenk, Lithuanian, Armenian, Altaian, Udegeian, Nivkh, Russian, Adygean, Mari, Central Asian, and other populations (see [3] for review and [4–12]). Three main methods were used to study the structure of the gene pool based on the data on surnames: the estimations of inbreeding and genetic distances and analysis of principal components.

We combined these three methods with two new powerful methods, namely, computer-based mapping and analysis of spatial-temporal dynamics in successive generations, to obtain a comprehensive gene-geographic technique of analysis of quasi-genetic markers. This approach was tested on the Adygean gene pool, which was an ideal model for studying the population structure. Indeed, the Adygeans constitute a subdivided population with a distinct hierarchical structure and pronounced endogamy, their gene pool has been comprehensively studied (there were the data on all surnames in all villages where the Adygeans lived compactly), six-generation pedigrees had been obtained, etc. The results demonstrated that cartographic analysis of surnames was effective for studying the gene-pool structure [2, 13–15]. This allowed us to tackle the gene-geographic analysis of Russian surnames, which are a more complex object of study.

The frequencies of Russian surnames were used when studying the spatial variation of inbreeding and the burden of hereditary pathology in some raions of the Arkhangel'sk, Kostroma, Kirov, and Kursk oblasts

1022-7954/01/3707-0807$25.00 © 2001 MAIK "Nauka / Interperiodica"

(see, e.g., [6–9, 16]). However, the analysis was restricted to individual raions or oblasts, because the "hypervariability" of Russian surnames indicated that their analysis is only applicable to studying the gene pools of relatively small population groups [11, 12]. Only recently, 75 prevalent surnames were studied in the total Russian population [17]. This analysis demonstrated that the geographic variation of prevalent Russian surnames reflected not only the events that had occurred at the district level, but also the ethnic history of the entire Russian population: the zones of accumulation of certain surnames have been found to be associated with the main anthropological types of the Russians [17]. The successful qualitative analysis of Russian surnames (the presence or absence of the given surname in the given region) allows us to do the next step and quantitatively analyze the variation of the frequencies of the same 75 surnames in the geographic area of the Russian ethnic group using electronic maps. The gene-geographic approach that proved to be effective when studying the Adygean gene pool will help us to determine the actual possibilities and restrictions of using Russian surnames for studies on the entire Russian gene pool.

In the framework of anthroponimics (the branch of onomastics dealing with human names), some patterns of the development of surnames have been found, including their origin in certain social groups, an earlier appearance of surnames in economically developed regions, etc. Development of surnames takes a long time. Russian surnames formed during several hundred years. The surnames of princes originated from the names of their principalities as early as in the 14th century. Later, the surnames of both princes and boyars were formed from their patronymics. The majority of surnames of the Russian nobility appeared in the 16th and 17th centuries. Most clergy received surnames in the late 18th century, the earliest originating from the names of churches. The first surnames of "noble merchants" are dated to the 16th century. However, the majority of Russians had, since ancient times, the so-called "street" surnames (in addition to family names and Christian names). The street surnames were not officially regulated, because they were not compulsorily registered; however, they were transmitted in generations. Many people had complex, compound surnames consisting of derivatives from the family name (e.g., Tolstoy [stout]), patronymic (e.g., Ivanov [son of Ivan]), and street surname (e.g., Korobeinikov [of the family of peddlers]). In 1861, a law was adopted on assigning surnames to the entire population of Russia. In the case of peasants, street surnames, surnames derived from patronymics, or landlords' surnames usually became their official surnames. However, peasants' surnames were being officially registered until the early 20th century, and the process was only accomplished when all Russian citizens were given passports.

The spread and geographic distribution of Russian surnames have been extensively studied. Usually, the zone of accumulation or origin of a surname was interpreted in terms of dialectology. In some cases, the areas where surnames originating from dialect words were found coincided with the geographic areas of the corresponding dialects. Groups of surnames with common finals were also analyzed. For example, surnames ending in "-ochkin" are more frequent in southwestern regions, surnames ending in "-ykh" or "-ikh" are typical of northern and southern regions of Russia but are absent in central regions, etc. Note that the geographic areas of some surnames and their groups coincide with the territories of the former Russian principalities. Apparently, having once emerged, surnames spread within communities of administrative regions [18–20].

However, there is another method of studying surnames. According to this method, mass samples of surnames are simultaneously studied over large areas by means of mapping, instead of studying arbitrarily chosen surnames or their groups. This study was a variant of such a "mass" analysis of modern Russian surnames with the use of new computer-based gene-geographic methods, namely, continuous mapping generalized cartographic analysis.

In summary, two questions should be clarified before studying the Russian gene pool based on the data on surnames: (1) whether it is possible, in principle, to the study the gene-pool structure (inbreeding, spatial patterns, and differentiation) based on the data on surnames and (2) how suitable Russian surnames are for studying the Russian gene pool. All the aforementioned data demonstrate that we may answer positively to the first question (applicability of surnames in general). Some studies have also demonstrated that Russian surnames can be effectively used for analysis of small populations. We believe that the results of this study demonstrate that Russian surnames can also be used for analysis of the Russian gene pool as a whole. To solve this task, we first had to demonstrate that the spatial distribution of Russian surnames is regular, rather than random or mosaic. Putting it in genetic terms, we may say that this distribution is determined by their spread due to drift and migration rather than by "mutations." Second, we had to prove the existence of common patterns of distribution for all surnames, because such patterns might only be determined by the history of the population and should be absent if the surname frequencies were random. Third, it was advisable to compare the information on the gene-pool structure obtained based on surname distribution with the result of similar analysis using genetic markers. If the results were to be similar, this would prove the adequacy of the use of surname to study the structure of the gene pool.

Thus, the methodological purpose of our study was to estimate the suitability of Russian surnames for studying the structure of the gene pool. However, the main goal was to obtain (based on the data on surnames) new information on the structure of the Russian

The most widespread Russian surnames selected for the study (in the alphabetic order)

Abramov	Efimov	Kotov	Noskov	Sokolov
Afanas'ev	Egorov	Kovalev	Novikov	Solov'ev
Aleksandrov	Ershov	Kozlov	Osipov	Stepanov
Alekseev	Fedorov	Krotov	Pavlov	Tarasov
Andreev	Filippov	Krylov	Pestov	Tikhonov
Anokhin	Golubev	Kudryashov	Petrov	Timofeev
Antonov	Gorbachev	Kurochkin	Polyakov	Trifonov
Balashov	Grigor'ev	Kuz'min	Popov	Tsvetkov
Belov	Gulyaev	Kuznetsov	Prokof'ev	Vasil'ev
Borisov	Gusev	Lebedev	Romanov	Veselov
Bykov	Il'in	Makarov	Savel'ev	Vinogradov
Chernov	Ivanov	Mikhailov	Semenov	Volkov
Chistyakov	Kalinin	Morozov	Shcherbakov	Vorob'ev
Denisov	Kiselev	Nikitin	Sidorov	Voronin
Druzhinin	Kosterov	Nikolaev	Smirnov	Yakovlev

gene pool, the degree of its differentiation, and the main trends of its variation.

MATERIALS AND METHODS

The data used in this study have been collected by A.P. Buzhilova for many years and were partly supplemented by other authors of this study. We mapped the data on the frequencies of surnames in 64 sel'sovets (rural communities; a total of 33 thousand persons). We studied indigenous Russian populations of 52 raions (districts) belonging to 22 oblasts (regions) of the European part of Russia.

The geography of the populations studied followed the main route of the Russian Anthropological Expedition [21], which studied the Russian population in the "ethnic zone of formation of the Russian population in the 11th through 14th centuries" [21, p. 24] still remains an exemplar of population studies. At this stage, we did not study the Russian populations that were separated from the main area and, according to V.V. Bunak, "became fixedly inhabited by Russians in the epoch when the Russian ethnic type had already been formed. Although all of these territories … are undoubtedly interesting anthropologically, they are not of decisive importance…" [21, p. 24]. Analysis of genetic markers [22, 23] has demonstrated that the gene pools of these populations reflected the characteristics of recent migrations; it will be more correct to study these regions after the variation patterns of the main core of the Russian ethnos have been determined. Thus, the study evenly encompasses the main area of formation and "indigenous" spread of the Russians. The corresponding populations are indigenous and have not been significantly affected by migrations during the past century. In our maps, the populations studied are indicated by asterisks.

For the study, we chose the surnames that were found in more than two populations (in contrast to the studies [11, 12], where the surname frequency served as a selection criterion). Hereinafter, these 75 surnames are referred to as widespread Russian surnames (table); we do not indicate their frequencies because the volume of data is too large. Classification of Russian surnames and their detailed analysis in terms of population migrations were performed by A.P. Buzhilova and described in [17]. We used computer-based gene-geographic methods to draw 75 maps of the frequencies of the widespread surnames in the studied geographic area of the Russian ethnos. The methods of gene-geographic mapping are described in [24–27]. The electronic map is a homogenous grid covering the area studied. The frequency of each given surname was calculated for each node of the grid by means of interpolation using orthogonal polynomials of zero power. When calculating the surname frequency at a given node of the map, we used the data on all population studied within a distance of 2000 km from this node; however, the weight coefficient of the surname frequency in each population was inversely proportional to the sixth power of the distance between the population and the node. A reliability map, which ranked the grid nodes according to the statistical significance of the frequency values at these nodes, helped us to estimate the reliability of the interpolation. The reliability map was constructed at a significance level of $\alpha = 0.5$; the nodes with the probability of correct prediction of $P > 0.95$ were used for cartographic analysis. For all maps of surnames, we used a unified homogenous scale of surname frequencies with the following borders of intervals: < 0.001, 0.011,

0.021, > 0.021. Since all frequencies are mapped using the same scale, the maps are easily comparable with one another. The minimum (MIN), maximum (MAX), and mean (MEAN) frequencies of the surnames in the area studied are indicated in the statistical inset of each map. We used all of the 75 maps of individual surnames to calculate and map principal components, i.e., new generalized parameters reflecting the most general characteristics of the spatial variation of the entire set of surnames studied. For this purpose, we calculated a correlation matrix, eigenvalues, and eigenvectors based on the original 75 maps (d.f. = 74). We only used the frequency values from reliable regions of the maps ($P > 0.95$). Afterwards, we calculated principal components for all node of the homogenous grid to obtain maps of principal components. The correlation matrix of the maps was used for the calculations (i.e., the mapped values were normalized for variances and deviations from the mean values); the variance the principal components was normalized, and the solution was optimized [26]. Important characteristics of principal components are an absence of correlation between them and the fact that first principal components describe the major part of variation of the entire set of maps.

RESULTS AND DISCUSSION

The main goal of the study was to analyze the architectonics of the Russian gene pool based on the data on all of the widespread Russian surnames. However, first we had to estimate the informativeness of the map of an individual surname. It was necessary to find out whether the map of a given surname revealed any patterns of the spatial variation of its frequency, or maps of individual surnames were so mosaic that only analysis of general parameters (principal components, genetic distances, etc.) of all maps could reveal the elements that formed the structure of the Russian gene pool. These questions were caused by a "hypervariability" of surnames. In terms of genetics, surnames may be regarded as one locus with a large number of alleles characterized by high mutation rates. The mean frequencies of most alleles are even lower than 1% polymorphism level. For example, the frequency of even the most common surname (Ivanov) was only $p = 0.0095$ in a sample of 200 000 of St. Petersburg citizens in 1910; the least frequent of the widespread surnames was Naumov ($p = 0.0006$) [20]. In our study, rural communities with the mean population size of only 500 persons served as elementary populations; therefore, we expected sharp fluctuations of surname frequencies. Even neighboring communities could considerably differ with respect to the frequencies of surnames, and each oblast was represented, on average, by three rural communities in our study. Therefore, a mosaic distribution of the frequency of each surname was the most probable. This must not decrease the effectiveness of the generalized analysis. For example, genes of hereditary diseases usually have low frequencies; however, their cartographic analysis has proved to be effective [28]. If, notwithstanding the sampling errors, the areas of surnames were not to be mosaic, this would directly indicate that Russian surnames are suitable for analysis of the gene pool. Indeed, the absence of mosaic pattern would confirm that the variation of surnames largely depended on the same microevolutionary factors as the variation of genes, i.e., the ratio between drift and migrations.

Cartographic Analysis of Individual Surnames

To estimate the informativeness of individual surnames, consider the maps of distribution of six surnames chosen at random (Figs. 1–6). According to the available data on the frequencies of surnames [20], two of these surnames (Ivanov and Petrov) rank first and third, respectively, with respect to their frequency in the Russian population, Grigor'ev is on the 14th place, Nikitin and Kuznetsov share the 23d place, and Kovalev does not fall within 100 most frequent Russian surnames [20]. Thus, the surnames chosen considerably differ from one another in their frequencies.

Figure 1 shows the geographic distribution of the surname Nikitin. The distribution of this surname was restricted by the southern parts of the Pskov and Novgorod oblasts. Thus, the surname Nikitin had a distinct geographic location. This was also true for some other surnames.

For example, the distribution of the surname Kovalev was strikingly compact, being restricted to the Bryansk and Smolensk oblasts (Fig. 2). We may hypothesize that the surname Kovalev has emerged within this zone. Note that the compact distribution of this surname was not found until its frequency was mapped [17].

Figure 3 shows the distribution of the surname Kuznetsov. In this case, we did not find a distinct geographic pattern. Zones of accumulation of this surname alternated with zones of its complete absence. Comparison of the latter two maps is interesting in terms of the origin of Russian surnames. The surnames Kovalev and Kuznetsov are generally assumed to be alternative (both of them originate from the words that mean "smith" in different dialects). However, the mapping demonstrated that the areas of these surnames partly overlapped: the southwestern maximum of the frequency of the surname "Kuznetsov" was located within the main zone of the distribution of the surname Kovalev.

The surname Ivanov (Fig. 4) is of special interest because this is the most frequent Russian surname. The name Ioann (a dated form of Ivan) is found 79 times in the Russian Orthodox Church calendar and used to be the most frequent man's name in Russia (with an average frequency of 15%): between the 17th and mid-20th century, it was more than twice as frequent as the next most frequent name, Vasilii (calculated from the data reported in [29]). It was natural to expect that the most frequent and polyphyletic surname (i.e., a surname that

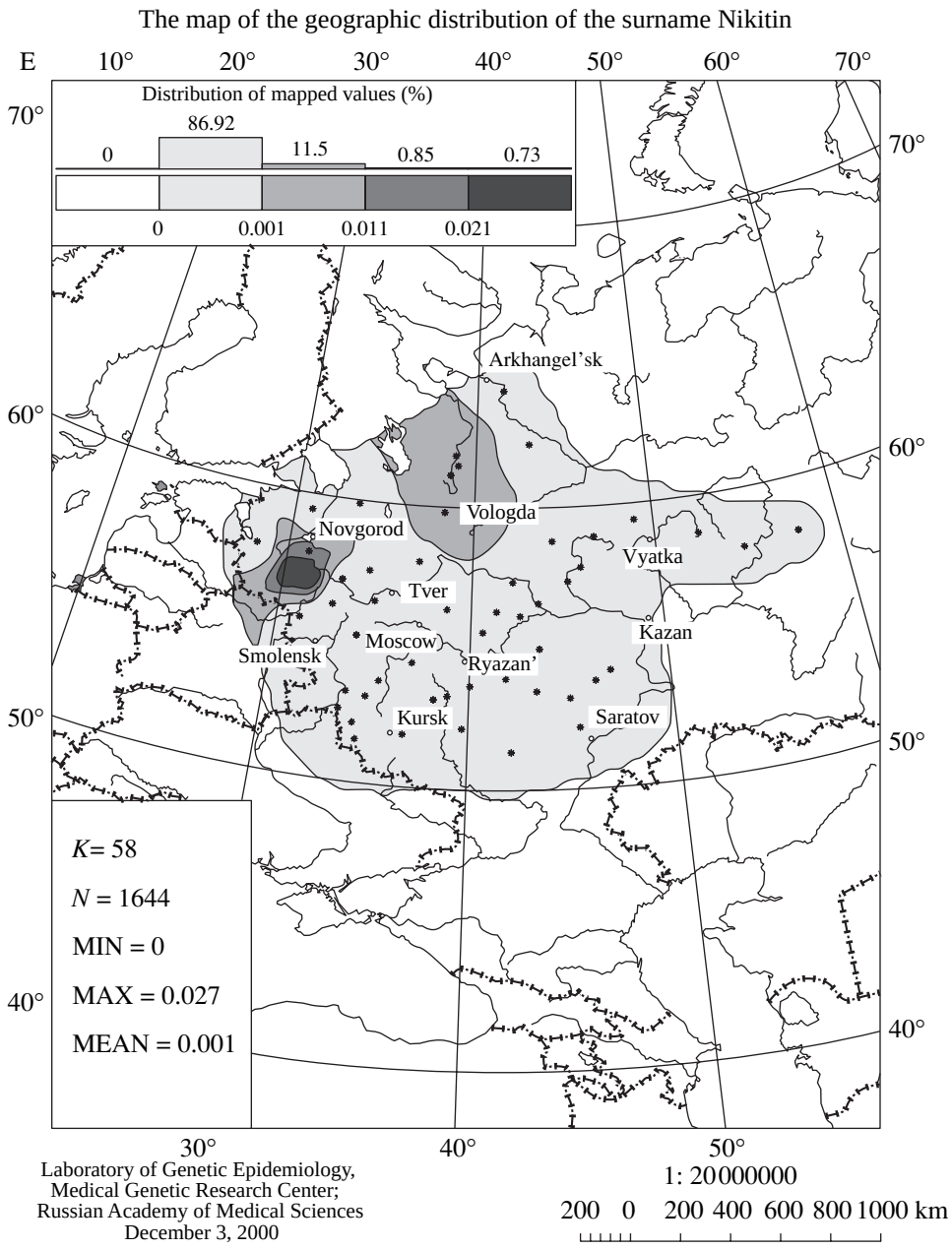

Fig. 1. Geographic distribution of the surname Nikitin. The interval inset: the first interval (<0), low-reliability zones; the second interval (0–0.001), zones where the surname is almost absent; the third interval (0.001–0.011), zones with the surname frequencies lower than 1%; the fourth interval (0.011–0.021), zones with the surname frequencies from 1 to 2%; the fifth interval (>0.021), zones with the surname frequencies higher than 2%; the proportion of the total area corresponding to the given interval is indicated above each interval in the histogram. The statistical inset: *K*, the number of the populations mapped; *N*, the number of grid nodes in the high-reliability zones; MIN, MAX, and MEAN, the minimum, maximum, and mean gene frequencies, respectively.

independently originated from the most frequent Christian name at many sites throughout the area) would not exhibit any geographic pattern. It is all the more surprising that the surname Ivanov is almost absent in some regions on the map. The zones of high frequencies are located in the northwest and (with somewhat lower frequencies) in the northeast. In the southern regions, the only "islet" of this surname is located near Kursk.

Figure 5 shows two zones of the highest frequencies of the surname Grigor'ev. The major zone encompasses the northwestern regions of Russia, and the minor zone is located in the Nizhni Novgorod oblast. The major zone is almost continuous, with a low frequency being found only in a part of the Tver oblast.

Figure 6 shows the vast zone of distribution of the surname Petrov in the northwestern part of Russia.

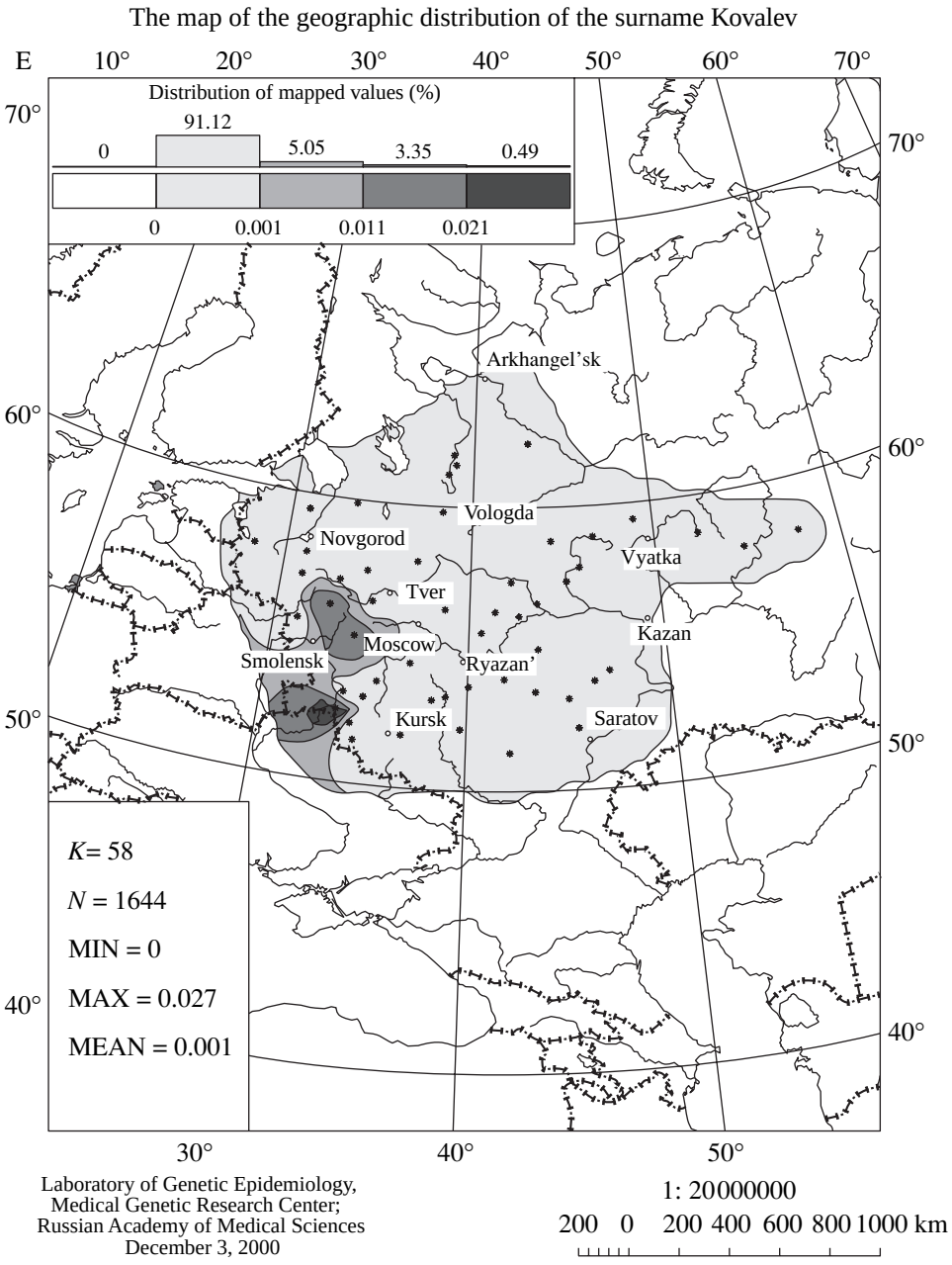

Fig. 2. Geographic distribution of the surname Kovalev (see Fig. 1 and the text for designations)

However, the surname Petrov is almost absent in the Tver oblast and neighboring regions, i.e., in the center of this zone. Thus, the area of this surname is ring-shaped and has maximums near Moscow and near Vologda.

The main conclusion that we may draw from the analysis of the maps of randomly chosen surnames is that each Russian surname has its typical geographic location. Some surnames exhibited extremely distinct geographic distributions, whereas none of the surnames studied was ubiquitous, i.e., none of them exhibited a mosaic or homogenous distribution throughout the area.

The results of our analysis, including the "accumulation zones" found in this study, are certainly preliminary and require further study and refinement. To correctly assess the areas of distribution of each Russian surname, it is necessary to use more data sources and to increase the total sample size by an order of magnitude. Taking into account that the population samples used in this study were insufficient, it seemed improbable to find any accumulation zones encompassing several neighboring populations. The fact that we have nevertheless found some geographic patterns of distribution of individual surnames indicates that surnames are

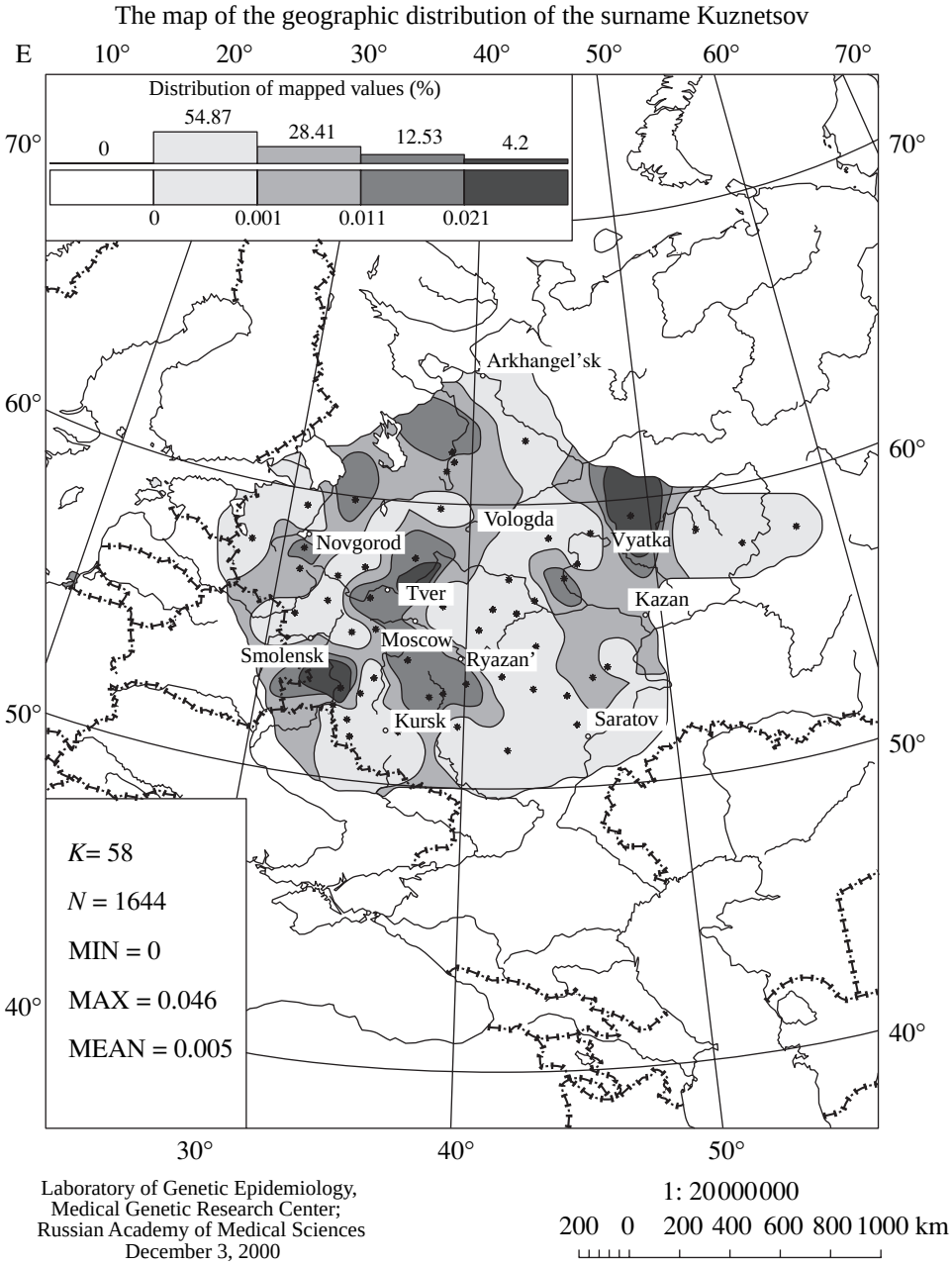

Fig. 3. Geographic distribution of the surname Kuznetsov (see Fig. 1 and the text for designations).

more appropriate for studying the Russian gene pool, including its history, than it was expected. It is of fundamental importance that the mapping of surname frequencies allowed us to objectively estimate their spatial variation. In the given case, a map is indispensable and often the only tool at the researcher's disposal.

Thus, the cartographic analysis of individual surnames revealed geographic patterns of their variation, which demonstrated that quasi-genetic markers were sufficiently informative when studying the spatial structure of the Russian gene pool. This allowed us to turn to the main goal of the study and to reveal the gene-pool structure through the generalized cartographic analysis of the entire set of surnames studied.

Analysis of Principal Components

For this purpose, we analyzed the principal components of the variation of the widespread surnames. The gene-geographic maps of principal components are maps of new generalized parameters that describe the major portion of the diversity of all surnames and correspond to the main scenarios of their variation. Since the original set of data consisted of 75 maps of sur-

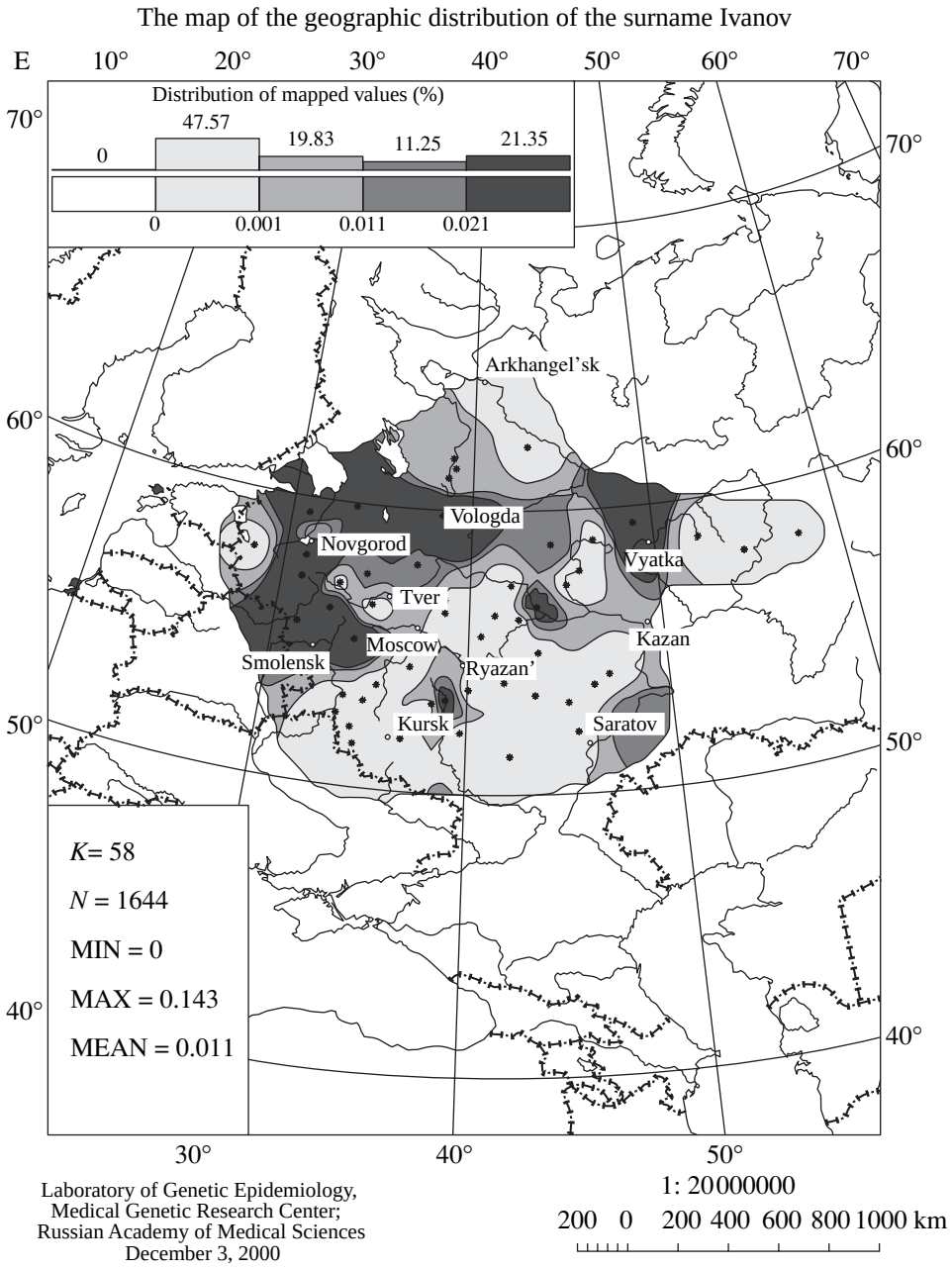

Fig. 4. Geographic distribution of the surname Ivanov (see Fig. 1 and the text for designations).

names, we constructed 75 maps of principal components. The effectiveness of the method may be estimated if we compare all components with respect to the proportion of variation described by each of them (S_i). If the next component describes a considerably lower proportion of the total variance than the preceding component, and the accumulated variance rapidly increases, then the method is effective. If the proportions of the variance of the consecutive components only slightly differ from one another (in our case, by about 1/74 or 1.3%), this means that the original characters are distributed chaotically, and we cannot distinguish the main vectors of the gene-pool variation. If this were the case, the new 75 generalized parameters would not differ fundamentally from the original data, and the method of principal components would not be informative.

Figure 7 shows the variance that is described by the first ten components. The percentages of the total variance accounted for by each component (S_i) are shown in the histogram and indicated on the left ordinate axis. The accumulated variance ($\sum S_i$, in percents) is described by the curve, and its values are shown on the

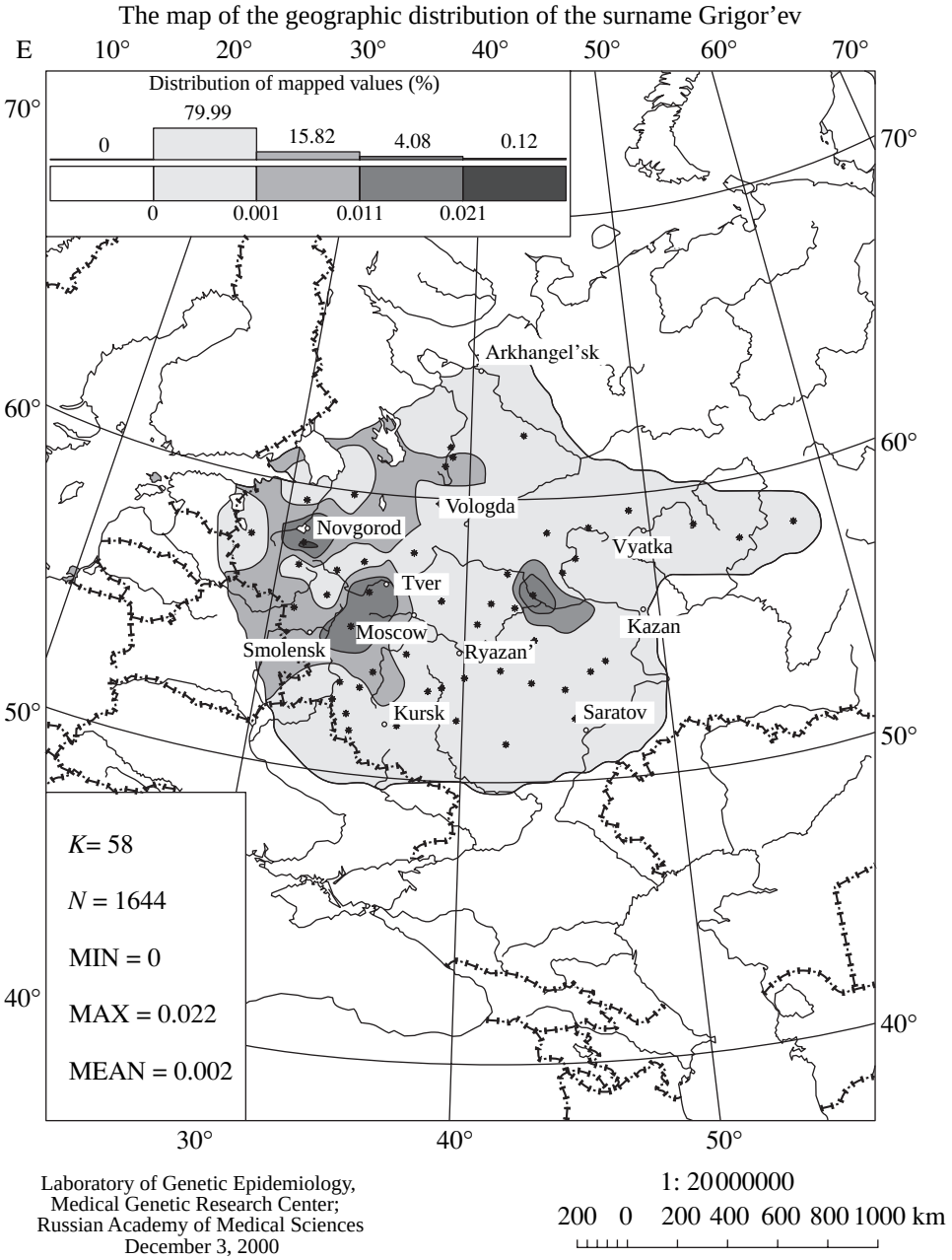

Fig. 5. Geographic distribution of the surname Grigor'ev (see Fig. 1 and the text for designations).

right ordinate axis. As is seen from the graph, the accumulated variance quickly reaches the plateau: the first two principal components account for as much as 30% of the total variation (instead of 3% expected in the case of ineffectiveness of the method); the first five components, almost half of the total variation (instead of 7% for an ineffective method); and the first ten components, two thirds of the total variation of 75 characters; so that the remaining 65 principal components add little information. The results obtained indicate that the method of principal components is highly effective for analysis of Russian surnames.

The map of the first principal component (Fig. 8) shows the main trend of the gene-pool geographic variation, which is close to longitudinal; it is directed from the northwestern to the eastern regions of the area studied. This trend exhibits the strongest correlations with the surnames Vasil'ev, Mikhailov, Fedorov, Aleksandrov, Andreev, and Il'in. Note that all of these surnames are "calendar," i.e., they originate from Christian names contained in the church calendar.

The map of the second principal component (the second most important "scenario" of the gene-pool geographic variation) is characterized by latitudinal

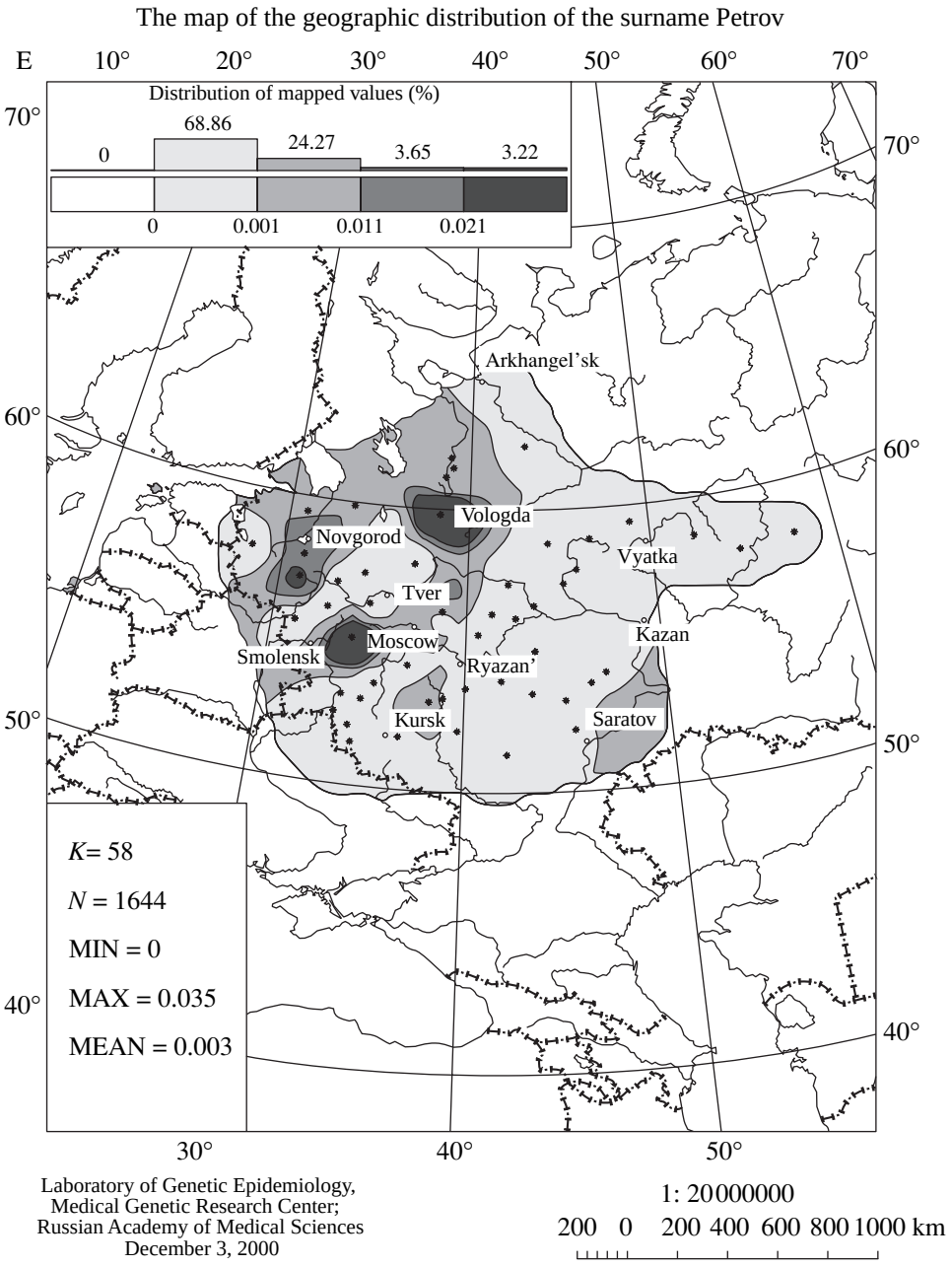

Fig. 6. Geographic distribution of the surname Petrov (see Fig. 1 and the text for designations).

variation directed from the southwestern to the northern regions of the area studied. This trend exhibits the strongest correlations with the surnames Denisov, Novikov, Kovalev, Kozlov, Anokhin, and Chistyakov (note that only one of them is a "calendar" name).

Before we turn to interpretation of the maps of principal components, note that a polysystemic approach (comparison of the maps of different systems of characters) is fundamentally important when analyzing the gene-pool structure [30–33]. If different systems of characters reveal a unified structure of the gene pool (although some characters may be studied better than others, sample sizes may be different, the characters themselves may differ in informativeness, etc.), then we may postulate that we have found the actual architectonics of the gene pool, whose features have forced their way through all the imperfectness of our studies [30–33].

In the given case, we had the rare possibility to carry out such a polysystemic study, because the maps of principal components for three important systems of characters (immunobiochemical genetic markers, dermatoglyphical traits, and anthropological traits) had already been drawn for the Russian gene pool [30, 32, 34].

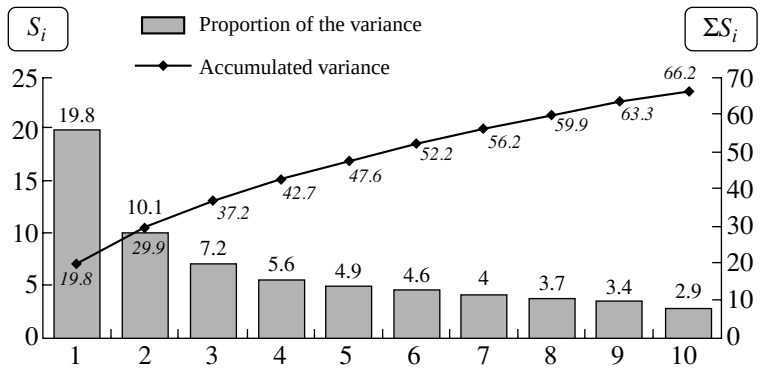

Fig. 7. Distribution of the total variance of Russian surnames among the first ten principal components. The histogram and the left ordinate axis show the percentages of the variance (S_i) described by each of the ten principal components. The curve and the right ordinate axis show the accumulated variance ($\sum S_i$; in percents of the total variance).

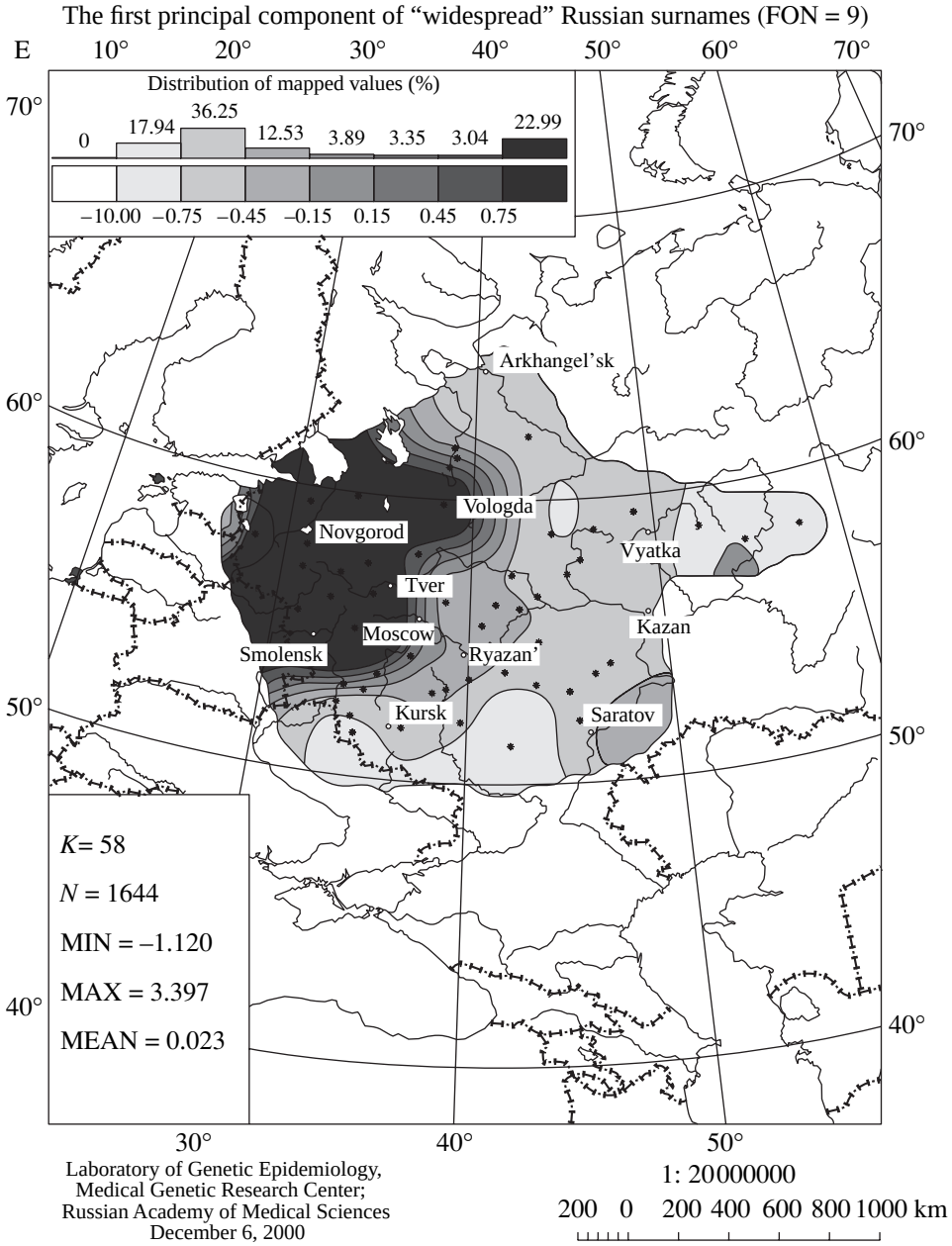

Fig. 8. The map of the first principal component of the variation of Russian surnames.

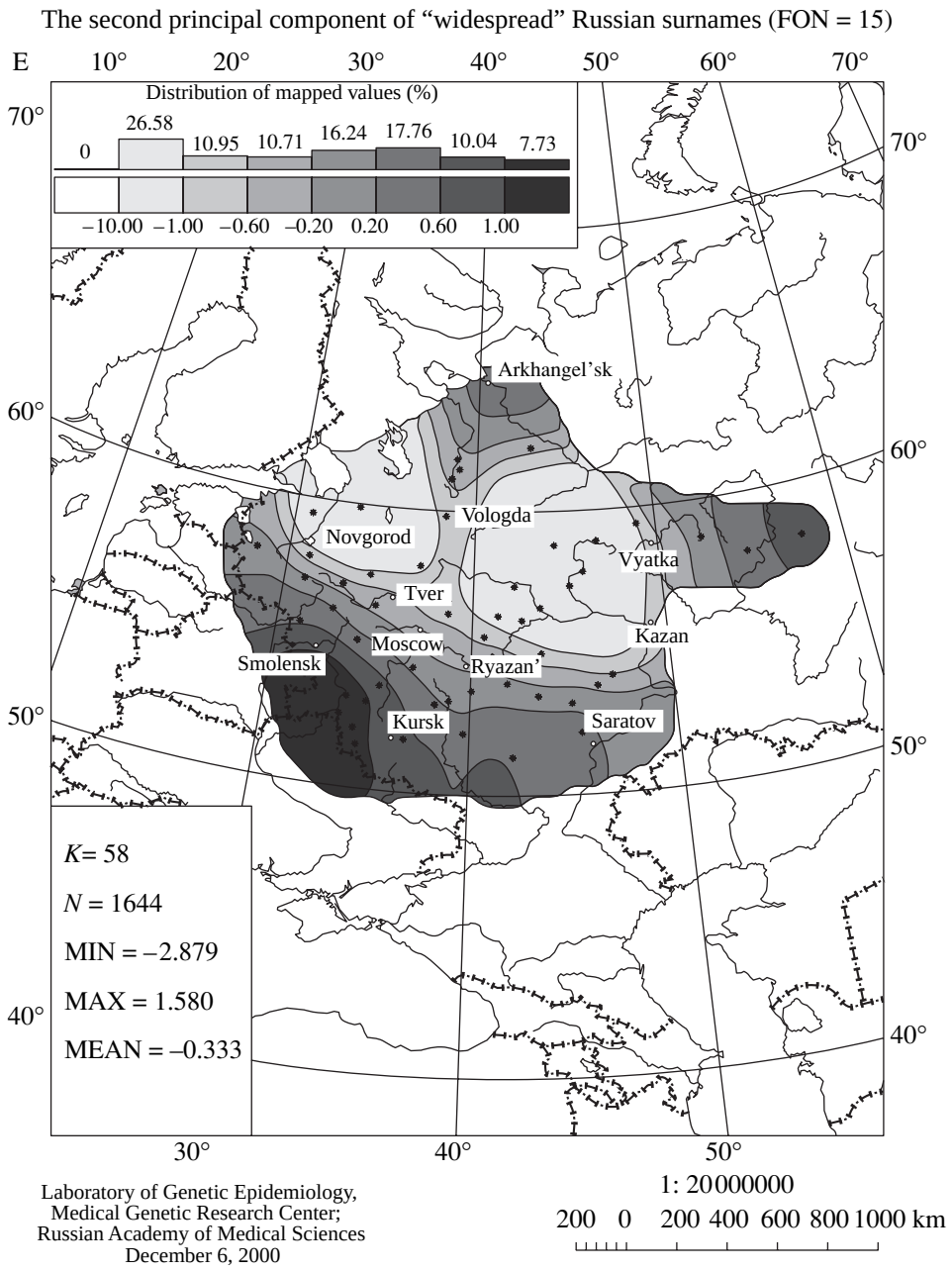

Fig. 9. The map of the second principal component of the variation of Russian surnames

Comparison of the generalized maps demonstrated that the first principal components of all of the three systems are strongly correlated and exhibit the same (latitudinal) variation trend. The map of the second principal component of the variation of surname frequencies also fits this pattern. This map (Fig. 9) is enough to conceive the main trend of the variation of the Russian gene pool reflected by all of the four so different systems of characters. The maps of these four systems only differ from one another with respect to two elements: (1) the relative position of the southern core (which is located in some maps farther to the east or west than in others; in some maps, there are both southwestern and southeastern cores) and (2) either more or less pronounced the northern or "Pomor" core (it is especially marked in the map of surnames).

The similarity of maps of four systems of characters (genetic, dermatoglyphical, anthropological, and anthroponymical) indicates that the latitudinal variation with a special position of the northernmost population group is the true main structure of the Russian gene pool. This latitudinal variation of the Russian gene pool agrees with another fundamentally important systems of characters, namely, the linguistic system. This is a system of

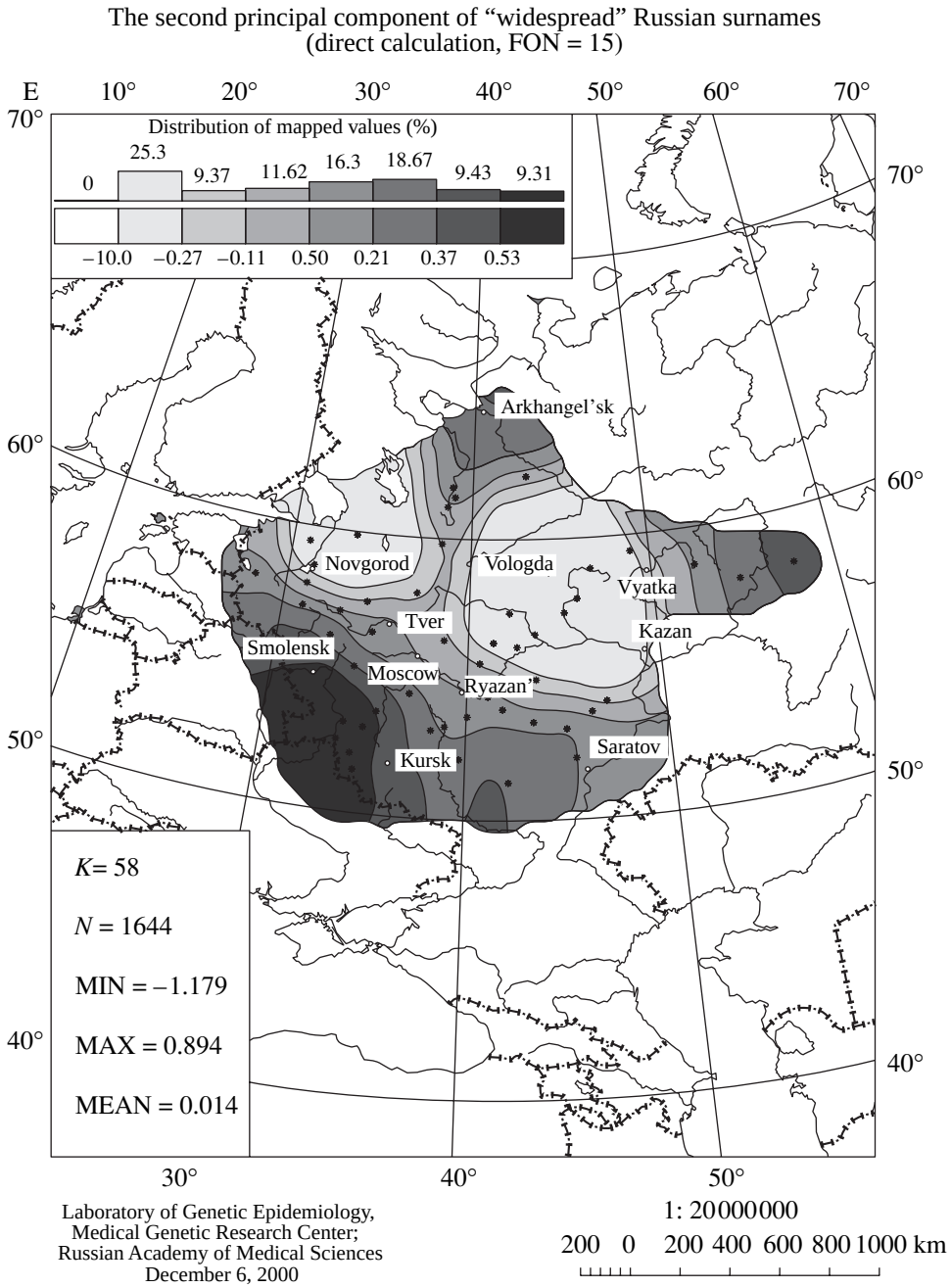

Fig. 10. The map of the second principal component of the variation of Russian surnames constructed by direct calculation, i.e., from the original data on the surname frequencies in rural populations.

various Russian dialects, the main groups of which are southern, central, and northern dialects (see, e.g., [35]).

In addition, we should emphasize the specificity of the Russian gene pool. In contrast to the Russian gene pool, the total Eastern European gene pool exhibits a longitudinal (west–east) variation trend of the systems of immunobiochemical markers, DNA markers, dermatoglyphical traits, and anthropological traits [32, 33, 36, 37].

Thus, analysis of various sets of data has confirmed that the latitudinal variation that is displayed by the second principal component of the surname frequencies is a true variation pattern of the Russian gene pool. This is additional evidence for the effectiveness of surnames as a model character for studying the gene-pool spatial structure.

Regarding the first principal component of surname variation, note that its trend is different from the trends of the first and second principal components for other systems of characters [30]. The map (Fig. 8) demonstrates that, judging by quasi-genetic markers, the northwestern region is almost isolated, which is consid-

erably less pronounced in the maps of other characters [31]. Possibly, the northwestern core has been determined by the criterion for widespread surnames used in our study (the presence of the surname in at least two populations), because this region was characterized by the smallest spectrum of surnames. Therefore, a greater proportion of "northwestern" surnames might have been included in the list of widespread surnames. However, it is possible that the distribution of surnames reflects not only common trends, but also other patterns, which are not reflected by anthropological and genetic characters (as is the case with dermatoglyphical characteristics in Eastern Europe) [33]. Additional studies on larger sets of surnames and populations are required to determine whether the distribution of the first principal component of the surname distribution actually reflects the true gene-geographic trend.

The Problem of False Correlation

In conclusion, consider an important methodical question concerning the mapping of principal components. Sokal and Cavalli-Sforza raised this question in their discussion on the effect of false correlations [38–40] caused by the interpolation procedure during mapping.

The number of population studied is always lower than the number of nodes on the map; therefore, data interpolation is necessary for constructing the cartographic model. Sokal notes that this procedure may lead to false correlations between maps. Since principal components are calculated from the matrix of correlations *between maps*, then the principal components will also be distorted due to the false-correlation error. According to Sokal, if the principal components were calculated directly from the original data (without maps) and the map of principal components were constructed based on these values, we would avoid the false-correlation error. However, Sokal agrees that this method of calculation is only possible in the rare cases when the entire set of characters is comprehensively studied in all populations.

Although we agree with Sokal in general, we think that the solution proposed by him (a rejection of the maps of principal components) is an extreme measure. This problem should be theoretically and experimentally studied in more detail. This experiment cannot be performed with genetic markers, because different sets of markers have been studied in different populations (the matrix is singular) and, hence, the matrix of principal components cannot be calculated directly. In contrast, all quasi-genetic markers were studied in all populations, which allowed us to test the effect of false correlations experimentally.

For this purpose, we constructed two variants of the map of principal components, calculated directly and from the surname maps, respectively. Figure 9 shows the result of calculation from the maps, i.e., construction of the maps of individual characters and calculation of principal components based on the resultant maps, the correctness of which was questioned by R. Sokal. Figure 10 shows the map of the same component obtained by direct calculation from the original data. This map serves as the correctness standard.

Comparison of these maps demonstrated their close similarity: the correlation coefficient was 0.963. In other words, the estimation of principal components based on maps of characters and through direct calculation yielded the same results. Thus, the experiment demonstrated that calculation of principal components from interpolated data did not lead to the false-correlation error. Note that we used a special technique for assessment of the reliability of mapping to select the regions with highly reliable prognosis ($P > 0.95$), and only data on these regions were used when calculating the principal components. The use of only "reliable" regions guaranteed the absence of false correlations.

Undoubtedly, intense research is required to estimate the limits of applicability of the principal-component method in the general case. However, the almost complete coincidence of the maps shown in Figs. 9 and 10 indicates that the resultant maps are free from false-correlation errors (provided that the reliability of information is taken into account). Therefore, the maps of principal component presented in this study are entirely correct and reflect the actual spatial variation of the Russian gene pool.

ACKNOWLEDGMENTS

This study was supported by the Russian Humanitarian Scientific Foundation, the Russian Foundation for Basic Research and the State Program "Frontiers in Genetics."

REFERENCES

1. Crow, J.F. and Mange, A.P., Measurement of Inbreeding from Frequency of Marriages between Person of the Same Surname, *Eug. Quart.*, 1965, vol. 12, pp. 199–203.
2. Balanovskaya, E.V., Pocheshkhova, E.A., Balanovsky, O.P., *et al.*, Gene Geographical Analysis of a Subdivided Population: II. Geography of Accidental Inbreeding (Inferred from Frequencies of Family Names in Adyghes, *Genetika* (Moscow), 2000, vol. 36, no. 8, pp. 1126–1139.
3. Rychkov, Yu.G., Space and Time in Gene Geography, *Vestn. Akad. Med. Nauk SSSR*, 1984, no. 7, pp. 11–15.
4. Luzina, F.A., Hereditary Polymorphism and Genetic Processes in the Indigenous Population of the Altai Highlands, *Abstract of Cand. Sci. (Biol.) Dissertation*, Moscow: Mos. State Univ., 1987.
5. Kazachenko, B.N., Revazov, A.A., Tarlycheva, L.V., and Lavrovskii, V.A., The Use of Family Names to Study the Factors in Dynamics of the Population Structure, *Genetika* (Moscow), 1980, vol. 16, no. 11, pp. 2049–2057.
6. Revazov, A.A., Paradeeva, G.M., and Rusakova, G.I., Possibility of Using Russian Family Names as a Quasi-

genetic Marker, *Genetika* (Moscow), 1986, vol. 22, no. 4, pp. 699–703.

7. El'chinova, G.I., Paradeeva, G.M., and Revazov, A.A., Medical Genetic Study of the Population of the Kostroma Oblast: 9. Interpretation of the Matrix of Genetic Distances, *Genetika* (Moscow), 1988, vol. 24, no. 11, pp. 2043–2049.
8. Ginter, E.K., Mamedova, R.A., El'chinova, G.I., *et al.*, A Load of Autosomal Recessive Disorders and Its Association with Consanguineous Marriages in the Population of the Kirov Oblast, *Genetika* (Moscow), 1993, vol. 29, no. 6, pp. 1042–1046.
9. Ginter, E.K., Mamedova, R.A., El'chinova, G.I., and Brusintseva, O.V., The Population Genetic Structure and Specific Features of the Spatial Distribution of Autosomal Recessive Disorders in the Kirov Oblast, *Genetika* (Moscow), 1994, vol. 30, no. 1, pp. 107–111.
10. El'chinova, G.I., Kravchuk, O.I., Startseva, E.A., *et al.*, Meadow Maris: Genes, Family Names, and Migrations, *Genetika* (Moscow), 1996, vol. 32, no. 10, pp. 1421–1422.
11. El'chinova, G.I., Kadoshnikova, M.Yu., Mamedova, R.A., *et al.*, On the Frequency-Based Criterion for Selecting Family Names to Study the Population Genetic Structure, *Genetika* (Moscow), 1991, vol. 27, no. 2, pp. 358–360.
12. El'chinova, G.I., Kadoshnikova, M.Yu., and Mamedova, R.A., Identification of Specific Features of the Population Genetic Structure by Means of Description of the Genetic Landscape, *Genetika* (Moscow), 1991, vol. 27, no. 11, pp. 1994–2001.
13. Balanovsky, O.P., Pocheshkhova, E.A., Nurbaev, S.D., and Balanovskaya, E.V., Spatial Variation of the Accidental Inbreeding Coefficient (A Cartographic Analysis of Quasigenetic Markers), *Zhizn' populyatsii v geterogennoi srede* (Life of a Population in Heterogeneous Environment), Ioshkar-Ola: Periodika Marii El, 1998, part 2, pp. 64–66.
14. Pocheshkhova, E.A., Balanovskaya, E.V., Balanovsky, O.P., and Nurbaev, S.D., The Adyghe Gene Pool: The Past in the Present, *Rasa: mif ili real'nost'? Trudy 1 Mezhdunarodnoi konferentsii Rossiiskogo otdeleniya Evropeiskoi antropologicheskoi assotsiatsii* (Race: A Myth of a Reality?: Proc. 1st Int. Conf. of the Russian Division of the Eur. Anthropol. Assoc.), Moscow: Staryi Sad, 1998, pp. 71–72.
15. Pocheshkhova, E.A., Genetic Demographic Analysis of the Subdivided Adyghe Population, *Cand. Sci. (Biol.) Dissertation*, Moscow, 1998.
16. Churnosov, M.I., The Genetic Demographic Structure and Distribution of Multifactorial Traits in Populations of the Kursk Oblast, *Abstract of Doctoral (Med.) Dissertation*, Moscow, 1997.
17. Buzhilova, A.P., Geography of Russian Family Names, *Vostochnye slavyane. Antropologiya i etnicheskaya istoriya* (Eastern Slavs: Anthropology and Ethnical History), Moscow: Nauchnyi Mir, 1999, pp. 135–152.
18. Nikonov, V.A., *Imya i obshchestvo* (Name and Society), Moscow, 1974.
19. Nikonov, V.A., *Geografiya familii* (Geography of Family Names), Moscow, 1988.
20. Superanskaya, A.V. and Suslova, A.V., *Sovremennye russkie familii* (Modern Russian Family Names), Moscow: Nauka, 1984.
21. *Proiskhozhdenie i etnicheskaya istoriya russkogo naroda: po antropologicheskim dannym* (The Origin and Ethnical History of the Russian People: Anthropological Data), Bunak, V.V., Ed., Moscow: Nauka, 1965.
22. Balanovskaya, E.V., Balanovsky, O.P., Spitsyn, V.A., *et al.*, The Russian Gene Pool: Gene Geography of Serum Gene Markers (*HP, GC, PI, TF*), *Genetika* (Moscow), 2001, vol. 37 (in press).
23. Balanovskaya, E.V., Balanovsky, O.P., Spitsyn, V.A., *et al.*, The Russian Gene Pool: Gene Geography of Erythrocytic Gene Markers (*ACP1, PGM1, ESD, GLO1, 6-PGD*), *Genetika* (Moscow), 2001, vol. 37 (in press).
24. Serbenyuk, S.N., Koshel', S.M., and Musin, O.R., The MAG Programs for Constructing Digital Models of Geological Fields, *Geodez. Kartograf.*, 1991, no. 4, pp. 44–46.
25. Balanovskaya, E.V., Nurbaev, S.D., and Rychkov, Yu.G., Computer Technology of Gene Geographical Studies of the Gene Pool: I. Statistical Information of Maps, *Genetika* (Moscow), 1994, vol. 30, no. 7, pp. 951–965.
26. Balanovskaya, E.V. and Nurbaev, S.D., Computer Technology of Gene Geographical Studies of the Gene Pool: IV. Populations in the Space of the Major Components, *Genetika* (Moscow), 1997, vol. 33, no. 12, pp. 1693–1710.
27. Nurbaev, S.D. and Balanovskaya, E.V., Computer Technology of Gene Geographical Studies of the Gene Pool: V. Estimation of Map Reliability, *Genetika* (Moscow), 1998, vol. 34, no. 6, pp. 825–838.
28. Zinchenko, R.A., El'chinova, G.I., Balanovskaya, E.V., *et al.*, The Effect of the Population Genetic Structure on the Load of Monogenic Hereditary Disorders in Russian Populations, *Vestn. Rus. Akad. Med. Nauk*, 2000, no. 5, pp. 5–11.
29. Suslova, A.V. and Superanskaya, A.V., *O russkikh imenakh* (On Russian Names), Leningrad: Lenizdat, 1991.
30. Balanovskaya, E.V., Balanovsky, O.P., Nurbaev, S.D., *et al.*, The Russian Gene Pool: Data of Various Disciplines, *II S"ezd VOGiS: Tezisy dokladov* (Proc. II Meeting of VOGiS), St. Petersburg, 2000, vol. 2, pp. 311–312.
31. Balanovsky, O.P., Buzhilova, A.P., and Balanovskaya, E.V., Gene Geographical Analysis of Russian Family Names, *II (IV) rossiiskii s"ezd meditsinskikh genetikov. Tezisy dokladov* (Proc. II (IV) Russ. Meeting of Medical Geneticists), Kursk, 2000, pp. 153–154.
32. Balanovskaya, E.V. and Balanovsky, O.P., Gene Geography of East European Ethnoses and Peculiar Features of the Russian Gene Pool, in *Zhenshchiny v fundamental'noi nauke* (Women in Basic Research), St. Petersburg, 2000, pp. 81–82.
33. Balanovsky, O.P., Deryabin, V.E., Dolinova, N.A., *et al.*, The Gene Pool of East European Ethnoses from the Genetic and Anthropological Data, *II S"ezd VOGiS: Tezisy dokladov* (Proc. II Meeting of VOGiS), St. Petersburg, 2000, vol. 2, pp. 312–313.
34. Deryabin, V.E., Modern East Slavonic Ethnoses, in *Vostochnye slavyane. Antropologiya i etnicheskaya istoriya*

(Eastern Slavs: Anthropology and Ethnical History), Moscow: Nauchnyi Mir, 1999, pp. 30–59.

35. Kasatkin, L.L., Russian Dialects, in *Voprosy antropologii, dialektologii i etnografii russkogo naroda* (Problems of Anthropology, Dialectology, and Ethnography of the Russian Ethnos), Moscow, 1998, pp. 37–100.
36. Rychkov, Yu.G., Balanovskaya, E.V., Nurbaev, S.D., and Shneider, Yu.V., Historical Gene Geography of Eastern Europe, in *Vostochnye slavyane. Antropologiya i etnicheskaya istoriya* (Eastern Slavs: Anthropology and Ethnical History), Moscow: Nauchnyi Mir, 1999, pp. 109–135.
37. Limborska, S., Slominsky, P., Balanovskaya, E., *et al.*, Study of DNA Diversity in East European Populations, *Human Genome Meeting, 1999*, Brisbane, 1999, p. 55.
38. Sokal, R.R., Oden, N.L., and Thomson, B.A., A Problem with Synthetic Maps, *Hum. Biol.*, 1999, vol. 71, no. 1, pp. 1–13.
39. Rendine, S., Piazza, A., Menozzi, P., and Cavalli-Sforza, L.L., A Problem with Synthetic Maps: Reply to Sokal *et al.*, *Hum. Biol.*, 1999, vol. 71, no. 1, pp. 14–25.
40. Sokal, R.R., Oden, N.L., and Thomson, B.A., Problems with Synthetic Maps Remain: Reply to Rendine *et al.*, *Human Biol.*, 1999, vol. 71, no. 3, pp. 447–453.