Road-like pattern of HLA-ABDR-based genetic distances between populations

JACEK NOWAK[1], *RENATA MIKA-WITKOWSKA*[1], *MAŁGORZATA POLAK*[2], *MAŁGORZATA ZAJKO*[1], *MARTA ROGATKO-KOROŚ*[1], *ELŻBIETA GRACZYK-POL*[1], *TOMASZ STĘPKOWSKI*[1], *ANDRZEJ LANGE*[2]

[1]Laboratory of Immunogenetics, Institute of Haematology and Transfusion Medicine, Warsaw, Poland; [2]Lower Silesian Center for Cellular Transplantations, Wrocław, Poland

Abstract

Genetic distances between populations may indicate the optimal direction of hematopoietic stem cell donor search. In some extent straight-line spatial geographic distances are similar to genetic distances. We tested the hypothesis that natural barriers, such as mountains, water reservoirs, desert and permafrost areas can change simple correlation between geography and genetics. We typed 200 healthy individuals of Polish population in HLA-A, B and DRB1 at the allelic level and compared allele group frequencies with 38 world populations. Standard Nei's pairwise genetic distances were estimated and correlated with land road distances in kilometres and straight-line spatial geographic distances. The genetic distances between all 39 populations were highly correlated with road distances (r=0.499, P=0.0012, df. =37). The correlation was much higher when estimation was limited to non-admixed and non-transferred populations (ie. when African American 1998, African American 2007, Argentina Buenos Aires, Belgian Total and Turkish minority in Germany, have been removed) (r=0.831, P=1,2E-09, df.=32). Simultaneously the straight-line spatial geographic distances correlated with genetic distances even higher (r=0.699, P=7.3E-07, df.=32 and r=0.848, P=2.4E-10, df.=32, respectively, for all 39 and 34 non-admixed/non-transferred populations). The preliminary analysis of a limited number of populations (N=20) has shown that the straight-line geographic spatial distances correlated with genetic distances much less than road distances (r=0.442, P=0.051, df.=18 vs. r=0.550, P=0.012, df.=18, respectively for straight-line and road distances). Simultaneously, geographic-like pattern of gene frequency low dimensional correspondence analysis was displayed in three dimensional projection. These results confirm considerable dependence of genetic and geographic distances. The results support the hypothesis that natural gene spread-out must have been more similar to waves and water circle-like propagation than to journeys along single paths. Natural barriers such as water reservoirs and mountains have had limited influence.

Key words: *genetic distance, geographic distance, human leukocyte antigens, population genetics.*

(Centr Eur J Immunol 2008; 33 (3): 114-119)

Introduction

The genetic distance between two populations is a mathematical expression of accumulated differences of allele frequencies [1]. The genetic similarity (low value of the genetic distance) may be due to geographic proximity of populations that facilitates gene spread-out and gene flow or due to common roots of parental populations [2]. The significant correlation between the genetic and geographic spatial distances has been confirmed in multiple studies [2-5].

It is a theory, well documented by the coalescence of mtDNA haplotypes, that modern humans emerged in South-East Africa and populated the continent in several waves [6]. The successful migration of modern humans out of Africa took place about 60-50,000 years ago [6]. The early human groups populating the Middle-East have probably

Correspondence: Jacek Nowak, Institute of Haematology and Transfusion Medicine, 14 Indira Gandhi Street, 02-776 Warsaw, Poland. Phone/fax number: +48 22 646 85 79, Email: szpik@ihit.waw.pl

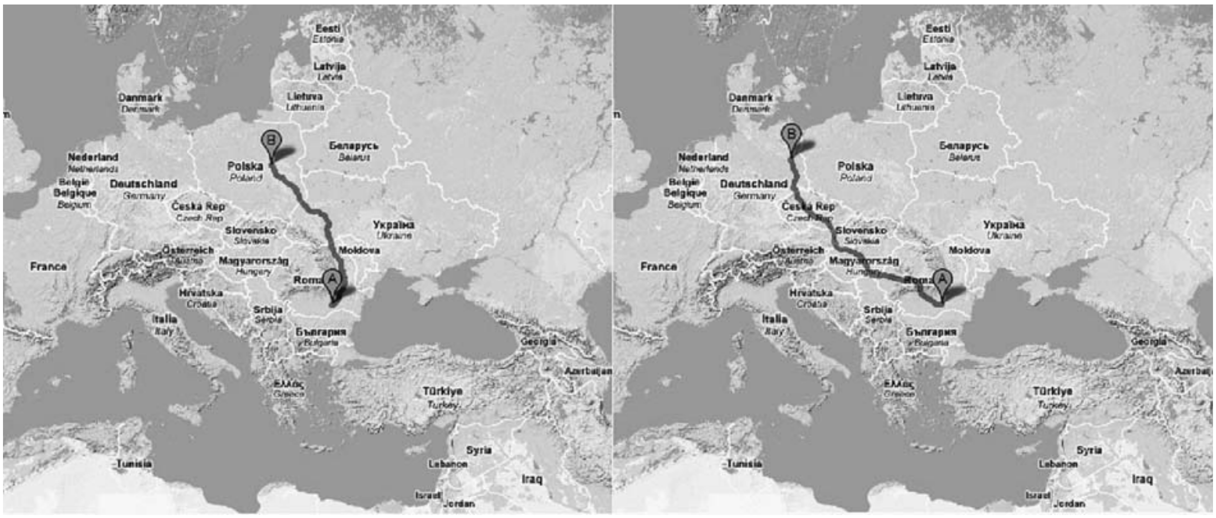

Fig. 1. Current road routes are influenced by natural geographic barriers
The left panel shows that the contemporary best road from the Balkans (A) to Warsaw (B) leads between the Carpathians and the Black Sea. The right panel shows that the contemporary best road from the Balkan area (A) to Berlin (B) leads between the Carpathian Mountains and the Alps. According to Google Map, v. 4.2 (August 2007)

been very small and for the next 30,000 years humans gained tenuous footholds in the continents of the Old World until the agriculture has been developed, starting Neolithic Era [7]. The migrations and thus the gene spread-out had to be dependent on hunting/gathering and agriculture opportunities, and possibly on local natural barriers like mountains, bigger water reservoirs, deserts and permafrost areas. We hypothesised that the current road routes (Figure 1) can simulate peoples' movements and gene spread in the past.

In the current study we tested the hypothesis that the correlation between the Nei's standard genetic distance and geographic distance matrices can be modified by the natural barriers met by populations during expansion. Our model have been current roads (Google Earth, July 2007) and microsatelite [8], Y-chromosome [9] and mtDNA based hypotheses [6] about human African and later on Asiatic expansions to new lands.

Material and Methods

Populations and typing

The analysis comprised N=200 healthy individuals living in Poland and having European Polish ancestors for at least two generations. Only those individuals who signed fully informed consent were recruited. The design of the study was approved by the institutional Ethics Committee. Three HLA loci (A, B and DRB1) were DNA typed up to allele (four-digit) level of resolution as described earlier [3]. Additionally, 38 reference populations were analysed in parallel with the current Polish population. HLA A, B and DRB1 allele group frequency data for the Argentinean population from Buenos Aires, Armenian, Belgian-European, Belgian-Total, Brazilian from Terena tribe, Chinese from Wuhan and from Taiwan, Czech from Prague-1 and Olomouc, Danish, French from Lyon, German from Dreieich, Freiburg, Mannheim and German-Total, Italian-North and Italian from Torino, Japanese from Kyoto, Nagano and Yokohama, Northern Ireland, Romanian, Spanish from Madrid, Turkish from Ankara and Turkish minority from Germany, African American 98 and Welsh populations are from Gjertson and Terasaki [10]. The data for African Americans 07 [11], Bulgarians [12], Georgian Svaneti-Svans [13], Mongolian Khalkha [14] and Spanish from Catalonia [15] populations were published originally and for Croatian, Czech from Prague-2, Finland and Spanish from Andalusia populations were published in the website www.allelefrequencies.net [16]. Sub-Saharan African Dogon population from Mali and Shona from Zimbabwe are taken from 13th and 14th Histocompatibility Workshops web database [17]. For the preliminary correlation analysis 20 populations, the examples representative for all continents and for the final analysis all 39 populations have been involved to show how the number of populations influence the results.

Statistical analyses

Allele frequencies (AFs) were estimated with maximum likelihood (ML) method combined with expectation maximization (EM) algorithm [18] using Arlequin v. 3.1 [19] with 1000 iterations (maximum) and 1×10^{-11} epsilon value. We used AFs of Polish and 38 reference populations to build pairwise Nei's genetic distance matrix [1]. The genetic distances and correspondence analyses were based on frequencies of A, B and DRB1 alleles. The frequency data were integrated to 69 low resolution allelic groups (21 groups in HLA-A, 35 in B and 13 in DRB1) because of the lack of high resolution data for most populations. Frequency correspondence analysis of HLA-A, B and DRB1 groups was performed using Statistica package v. 7.1 (StatSoft, Tulsa, OK). Typical two-way crosstabulation table has been used for 39 rows (populations) and 69 columns (frequencies of ABDR allele groups). The frequencies were first standardized, so that the relative frequencies across all cells sum to 1.0. The analysis represents the entries in the table of relative frequencies in terms of euclidean genetic distances between populations presented in a low-dimensional space. Three-dimensional presentation retained 65.1% of variance (inertia) between populations, including 28.6% for dimension 1, 23.2% for dimension 2 and 13.3% for dimension 3.

Geographic road distances between populations have been estimated in kilometres using Google Earth v. 4.2 (August 2007), choosing capitals of countries or regions as geographic points and their road connections as far as possible. African road distances have been measured along with Sub-Saharan train lines, Africa to Asia connection involved the Suez region [6], Middle East to Eurasia roads passed through Anatolia and Balkans [20-22] and hypothetical isthmus of Beringia has been used as Asia to Northern-America road connection [6]. The geographic straight-line spatial distances have been estimated using geographic latitude (φ) and longitude (λ) of two points, calculating the central angle (D) according to formula: $D=\arccos[(\sin\varphi_1 \times \sin\varphi_2)+(\cos\varphi_1 \times \cos\varphi_2 \times \cos\Delta\lambda)]$, and estimating a great-circle distance (orthodrome) with the coefficient 111.12 for kilometres [23]. Standard correlation coefficients for genetic and geographic road, and straight-line spatial distances have been estimated.

Results

The three-dimensional presentation of correspondence analysis represents euclidean genetic distances [24] between 39 populations (Figure 2), expressing 65.1% of the total variance. Several concentrated clusters of populations can be seen as corresponding to the continents of the populations' origin. The vast majority (51.8%) of the population's variance is displayed for dimensions 1 and 2. All populations except one show similar values of dimension 3, represented at the Figure 2A by nearly equally long vertical projection lines. The dimension 3 for the exceptional population, Terena tribe from Brazil, displayed extremely low value (short vertical projection line). Simultaneously, values of the dimensions 1 and 2 place this population in Eastern-Asiatic quadrant (Figure 2A). European and Caucasus population cluster zoomed in is shown at the Figure 2B. Two populations, the Argentinean population from Buenos Aires and Belgian Total lean toward the lower values of dimensions 1 and 2, that is more typical for analysed African populations. Simultaneously,

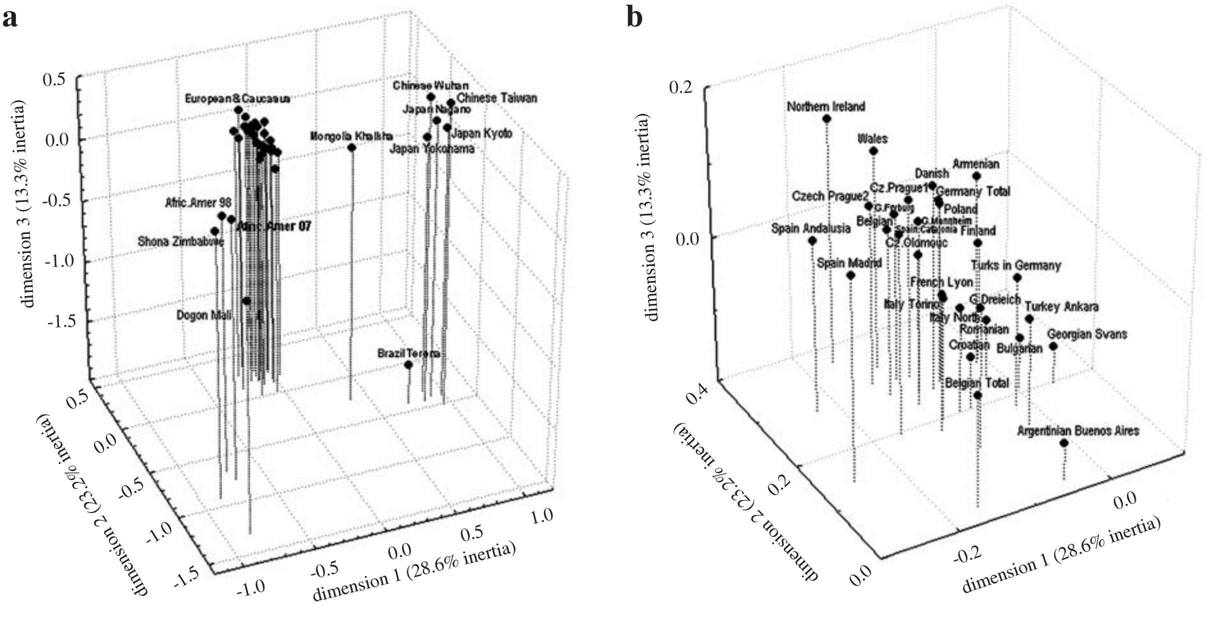

Fig. 2. Correspondence analysis (three dimensional presentation) using HLA-A, B and DRB1 allele frequencies. Panel A, analysis of all 39 populations; Panel B, analysis zoomed in to the European and Caucasus populations. The frequency data were integrated to 69 low resolution allelic groups (21 groups in HLA-A, 35 in B and 13 in DRB1)

the dimension 3 value for Argentinean population show slightly lowered value, toward the South-American Terena. The locations of Belgian Total and Buenos Aires populations in the three dimensional space suggest their substantial European background and African, or African and South American admixture, respectively. The three-dimensional relative locations of the remaining European and Caucasus populations remind to some extent their cartographic locations.

The correlation coefficients between geographic and genetic distances have been influenced by the number of populations involved in the analysis (Table 1). The preliminary analysis of limited number of populations (N=20) has shown that the correlation of the genetic distances with straight-line geographic spatial distances has been much less significant than with road distances (r=0.442, P=0.051, df.=18 vs. r=0.550, P=0.012, df.=18, for straight-line and road distances, respectively). Dubious significance of the correlation when N=15 to 20 populations were analysed, and very sharp P values for N=34 to 39 populations indicate better reliability of the multiple data. The genetic distances between all 39 populations were highly correlated with road distances (r=0.499, P=1.2 × 10^{-3}, df.=37). The correlation was much higher when estimation was limited to non-admixed and non-transferred populations (ie. when African American 1998, African American 2007, Argentina Buenos Aires, Belgian Total and Turkish minority in Germany, have been removed) (r=0.831, P=1,2 × 10^{-9}, df.=32). Simultaneously the straight-line spatial geographic distances correlated with genetic distances even higher (r=0.699, P=7.3 × 10^{-7}, df.=32 and r=0.848, P=2.4 × 10^{-10}, df.=32, for all 39 and 34 non-admixed/non-transferred populations, respectively, Table 1).

Discussion

The similarity of the relative locations of the tested populations at the genetic distance correspondence analysis diagram to the geographic locations was striking. Populations were grouped not only at the continent-like but also at the sub-continent-like level. Current European, Mediterranean and Caucasus populations are genetically placed closely together, indicating common ancestry in not so distant eras. The location of those populations at the genetic correspondence diagram could have been a result of gene spread similar to the migration movements. For the two populations of today's Africa (Dogon from Mali and Shona from Zimbabwe) slightly bigger spacing has been shown indicating higher genetic diversity than for Europeans, but the geographic distance of these two populations is also longer. For Central-Asiatic Khalkha population from Mongolia similar genetic departure from both, European and Eastern-Asiatic populations has been revealed, confirming the hypothesis of step by step primary expansion of humans in Asia. For South-American Terena tribe the similar values of dimensions 1 and 2 suggest common ancestry with Central and Eastern-Asiatic populations, but the simultaneously observed value of the dimension 3, substantially different from Asiatic populations, indicates significant genetic specificity of the Terenas. The interesting finding from the genetic correspondence analysis diagram was the shift of admixed populations toward the admixing populations and consequently the disparity with their current relative geographic locations.

Our preliminary study (Abstract No. 167, this Volume) of straight-line distance correlation has been based on 20 populations and the current study has been supplemented up to 39 available populations. This difference in the volume of genetic data involved makes qualitative change because the square matrices are correlated pairwise and accordingly the number of degrees of freedom is rising. Consequently, in our studies the data with 39 populations are much more reliable than with 20 populations. For 20 population analysis the straight-line geographic distances correlated insignificantly and road distances correlated more strongly (P=0.012) with genetic distances. As a consequence of much more data available, for 39 populations the straight-line geographic distances and road distances correlated very highly significantly, but the straight-lines correlated with genetic distances of three orders of

Table 1. Standard correlation between genetic and geographic (road or straight-line) distances in preliminary (20 populations) and multiple-population (39 populations) data analyses

Groups	Total correlation					Non-admixed/non-transferred				
	N	r	P	df.	t	N	r	P	df.	t
preliminary SLD vs. GD	20	0.442	0.051	18	2.09	15	0.492	0.062	13	2.04
preliminary RD vs. GD	20	0.550	0.012	18	2.80	15	0.631	0.012	13	2.93
all populations SLD vs. GD	39	0.699	7,3E-07	37	5.95	34	0.848	2,4E-10	32	9.06
all populations RD vs. GD	39	0.499	1.2E-03	37	3.50	34	0.831	1,2E-09	32	8.45

GD – genetic distance, RD – road distance, SLD - straight-line distance, N – number of populations, r – coefficient of correlation, P – statistical significance, df. – degrees of freedom, t – statistics in t-test.

magnitude stronger than road distances (P=7,3 × 10^{-7} and P=1.2 × 10^{-3}, respectively). This confirms earlier observations on high dependence of genetic and geographic distances, but it fails to show superiority of road distances beyond straight-line distances. As could be expected the correlations with genetic distances were much improved after removal of those populations shown by correspondence analysis to be admixed (ie. Argentina Buenos Aires, Belgian Total, Turkish minority in Germany) and African-American populations (P=2.4 × 10^{-10} and P=1.2 × 10^{-9}, for the straight-line and road distances, respectively). These values show that road-like population distances do not correlate stronger with genetic distances than straight-line geographic distances.

In the previous studies, the "air distances" between populations were calculated by researchers because of their correlation with road distances [2, 25], the last considered to correlate primarily with genetics. Our data did not confirm this intuitive presumption.

Any analysis of geography/genetics association relies on the validity of the underlying model. Like every model our model is a simplification of a set of evolutionary phenomena that would otherwise be difficult or impossible to address quantitatively. The satellite positioning of geographic objects, the technique currently available, gives the opportunity for valid estimations of both the road distances between populations and their geographic coordinates further used for the central angle and great-circle distance estimations. Although the satellite measurements were precise the main drawback of the model can be uncertainty of presumed migration and expansion ways. For example, it sounds reasonably that the Germanic and Nordic populations on the one hand and Slavic populations on the other hand could expand during Neolithic era from common Proto-Indo-European population located in the Balkans [21], passing Carpathian Mountains from the left and right hand side, respectively (compare with current best ways at Figure 1). However, the secondary re-expansions, documented partly by archeological sources [26], may have passed different directions and routes than primary expansion. Moreover, the ways of continental exits are still hypothetical.

The validity of the genetics component of the model, the HLA allele groups can be subject to some criticism. The gene flow processes affect not only MHC genes but the entire genome. Analyses of single or physically linked alleles are unlikely to contain all the information needed to infer and quantify population processes [2]. Indeed, the greater the number of systems considered, the more robust the inferences about gene flow processes [27]. The plausibility of HLA typing may rely on the fact that A, B and DRB1 *loci* are the most polymorphic *loci* in entire human genome [28]. Moreover, the HLA alleles contributed to the selection of individuals and bigger groups at the immune response level. The HLA frequency selections by pathogens and genetic drift that act strongly in small populations, appear to make this system very informative upon population history. The extreme HLA polymorphism can be potentially powered by high resolution allele and especially haplotype analyses. The informativeness would be facilitated by the fact that the recombination events, that lead to the selection of new haplotypes [29] are of several orders of magnitude more frequent than mutations [30]. However, at the current state of art this kind of analysis was limited by unavailability of multilocus population data at high resolution.

The most apparent cause why straight-line distances correlate with genetics slightly better than proposed road distances is the observation that multiple roads can be chosen to connect any two geographic places. Although some roads are more convenient, also more challenging ways could be chosen by migrating peoples for different reasons. Multiple population re-expansions, those documented by archeological sources [26] and unknown, may have taken different directions and routes. This uncertainty causes that the best simulation so far satisfying genetic similarity of populations is their straight-line geographic distance.

These results confirm considerable dependence of genetic and geographic distances. The results support the hypothesis that natural gene spread must have been more similar to waves and water circle-like propagation than to journeys along single paths. The natural barriers such as water reservoirs and mountains have had limited influence.

Acknowledgment

This study was supported in total by funding from Ministry of Science and Higher Education, grant No. 2 P05E 126 26. We wish to thank Piotr Nowak M.A. for English editing and improvements.

References

1. Nei M (1972): Genetic distance between populations. Am Nat 106: 283-292.
2. Dupanloup I, Bertorelle G, Chikhi L, Barbujani G (2004): Estimating the impact of prehistoric admixture on the genome of Europeans. Mol Biol Evol 21: 1361-1372.
3. Nowak J, Mika-Witkowska R, Polak M et al. (2008): Allele and extended haplotype polymorphism of HLA-A, C, B, DRB1 and DQB1 loci in Polish population and genetic affinities to other populations. Tissue Antigens 71: 193-205.
4. Simoni L, Calafell F, Pettener D et al. (2000): Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66: 262-278.
5. Nowak J, Gronkowska A, Brojer E (2002): Analysis of HLA frequencies in donor population of Unrelated Bone Marrow Donor and Cord Blood Registry of Institute of Haematology and Blood Transfusion. Centr Eur J Immunol 27: 97-115.
6. Forster P (2004): Ice ages and the mitochondrial DNA chronology of human dispersals: a review. Philos Trans R Soc Lond Biol Sci 359: 255-264.
7. Pennington RL (1996): Causes of early population growth. Am J Phys Anthropol 99: 259-274.
8. Belle EM, Landry PA, Barbujani G (2006): Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. Proc Biol Sci 273: 1595-1602.

9. Chikhi L, Nichols RA, Barbujani G, Beaumont MA (2002): Y genetic data support the Neolithic demic diffusion model. Proc Nat Acad Sci U S A 99: 11008-11013.
10. Gjertson DW, Terasaki P. HLA 1998. American Society for Histocompatibility and Immunogenetics. Lenexa, Kansas, USA, 1998.
11. Tu B, Mack SJ, Lazaro A et al. (2007): HLA-A, -B, -C, DRB1 allele and haplotype frequencies in an African American population. Tissue Antigens 69: 73-85.
12. Ivanova Ivanova M, Rozemuller E, Tyufekchiev N et al. (2002): HLA polymorphism in Bulgarians defined by high-resolution typing methods in comparison with other populations. Tissue Antigens 60: 496-504.
13. Sanchez-Velasco P, Leyva-Cobian F (2001): The HLA class I and class II allele frequencies studied at the DNA level in the Svanetian population (Upper Caucasus) and their relationships to Western European populations. Tissue Antigens 58: 223-233.
14. Machulla HK, Batnasan D, Steinborn F et al. (2003): Genetic affinities among Mongol ethnic groups and their relationship to Turks. Tissue Antigens 61: 292-299.
15. Comas D, Mateu E, Calafell F et al. (1998): HLA class I and class II DNA typing and the origin of Basques. Tissue Antigens 51: 30-40.
16. Middleton D, Menchaca L, Rood H, Komerofsky R (2003): www.allelefrequencies.net New Allele Frequency Database. Tissue Antigens 61: 403-407.
17. National Center for Biotechnology Infirmation, dbMHC database, Anthropology, Edition Oct 2007, http://www.ncbi.nlm.nih.gov/projects/mhc/ihwg.cgi?cmd=PRJOV&ID=9
18. Excoffier L, Slatkin M (1995): Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921-927.
19. Excoffier L, Laval G, Schneider S (2005): Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50.
20. Gamkrelidze TV, Ivanov VV (1985): The migrations of tribes speaking Indo-European dialects from their original homeland in the Near East to their historical habitations in Eurasia. Journal of Indo-European Studies 13: 49-91.
21. Diakonov IM (1985): On the original homeland of the speakers of Indo-European. Journal of Indo-European Studies 13: 92-176.
22. Renfrew C (1998): The Tarim basin, Tocharian, and Indo-European origins: a view from the west. In: Mair VH (ed.). The Bronze Age and Early Iron Age Peoples of Eastern-Central Asia. Journal of Indo-European Studies Monograph 26: 158-175.
23. Steinhaus H (1999): Mathematical snapshots. 3rd ed. New York: Dover, pp. 183 and 217.
24. Engelking R: General Topology. Polish Scientific Publishing. Warszawa, 1977.
25. Crumpacker DW, Zei G, Moroni A, Cavalli-Sforza LL (1976): Air distance versus road distance as a geographical measure for studies on human population structure. Geogr Anal 8: 215-223.
26. Kaczanowski P, Kozłowski JK. Oldest history of Polish lands (till 7th century). Fogra, Kraków, 1998.
27. Bertorelle G, Excoffier L (1998): Inferring admixture proportions from molecular data. Mol Biol Evol 15: 1298-1311.
28. IMGT/HLA Database, Edition: July 2007, Available at: http://www.anthonynolan.org.uk/HIG/lists/
29. Nowak J, Mika-Witkowska R, Zajko M et al. (2007): Two different methods of inference of extended HLA haplotypes and their reliability in Polish population. East-West Immunogenetic Conference, Prague 1-3 March 2007, Abstract Book 2007: 42.
30. Nowak J, Kalinka-Warzocha E, Juszczynski P et al. (2007): Association of human leukocyte antigen ancestral haplotype 8.1 with adverse outcome of non-Hodgkin's lymphoma. Genes Chromosomes Cancer 46: 500-507.