Phylogenetic relationships of ATM polymorphic loci haplotypes in Eurasian and African populations

Marina V. Olkova[1,*], Sergey M. Koshel[2], Georgiy Yu. Ponomarev[1], Andrei A. Alimov[1]

[1] Research Centre for Medical Genetics, Moscow, 115478, Russia

[2] Department of Cartography and Geoinformatics, Lomonosov Moscow State University, Moscow, 119991, Russia

Corresponding author email: genetics@inbox.ru

ABSTRACT

The ATM gene encodes a key kinase involved in DNA repair and cell cycle control, and mutations in it lead to the development of ataxia-telangiectasia and increase susceptibility to cancer. In this context, the study of the distribution patterns of ATM haplotypes in metapopulations living in Eurasia and Africa using population genetic and genogeographic methods appears to be a socially important scientific endeavor.

Polymorphic markers of the ATM gene were studied in 3876 individuals from 46 metapopulations, ATM gene haplotypes and genotypes based on ATM haplotypes were determined, and population frequencies of haplotypes and genotypes were calculated. An analysis of phylogenetic relationships of haplotypes, the Tajima test to assess population changes, and mapping of population frequencies of ATM gene haplotypes and genotypes were performed. The study showed the presence of a linkage block of 28 polymorphic markers of the ATM gene in the Eurasian and African populations. Two dominant ATM haplotypes in Eurasia, which are the result of independent evolution of two branches from the ancestral ATM

haplotype, are formed by mutually exclusive alleles and correspond to the "yin-yang" concept. The greatest diversity of ATM haplotypes is observed on the African continent. It is suggested that the process of haplotype formation occurred on the African continent, with subsequent migration events to the territory of Eurasia.

Key words: ATM haplotypes, LD-block, “yin-yang” concept, mutually exclusive alleles, evolutionary separation

INTRODUCTION

The human ATM gene (ataxia-telangiectasia mutated), which encodes a serine-threonine protein kinase, is located on chromosome 11, locus 11q22-23. Activation of the gene’s protein product occurs as a result of double-stranded DNA breaks, leading to the initiation of signaling pathways and the phosphorylation of numerous substrates (several hundred have been identified to date). The final result of cascade reactions is cell cycle arrest and DNA repair. In addition to double-stranded DNA breaks, activation of the ATM protein can be caused by oxidative stress, deficiency of a number of metabolites, hypoxia, and some other factors [1]. ATM is also involved in the activation of the spindle checkpoint during mitosis. Loss of ATM function results in a defective spindle checkpoint and increases the likelihood of aneuploidy [2].

Mutations in the ATM gene are associated with the development of autosomal recessive Louis-Barr syndrome (ataxia-telangiectasia) and with the risk of developing cancer, which is one of the reasons for the research interest in studying the distribution of frequencies of polymorphic variants of this gene in populations [3–6]. One of the most important stages of genetic data research is the linkage disequilibrium (LD) analysis, which allows to identify new clinically significant variants of the ATM gene associated with changes in the function of the gene's

protein product and disease severity. As a result of previous scientific studies, blocks of LD in the ATM gene have been identified and described in a region of ~150 kb, covering the entire length of the gene [7–9]. In particular, Tornston et al. identified seven haplotypes of the ATM gene, two of which were present in 82% of the 93 DNA samples examined from representatives of seven human populations living on all continents. Two of the seven identified haplotypes were found only in African populations with a frequency of 33%. The presence of blocks of LD of varying length was also demonstrated for a number of other autosomal regions of the human genome. In particular, Bonnen's article assessed the LD of non-coding polymorphic markers for 5 genes (TP53, RAG51, BRCA1, BRCA2 and ATM) associated with the cancer development. It was shown that for all genes, taking into account the fact that the study included markers evenly distributed along the gene, there was no tendency for the LD value to decrease with increasing distance. The authors concluded that the LD value can vary widely regardless of the physical distribution of the markers along the length of the DNA sequence of the gene. The observed phenomenon testified to the existence of polymorphic loci in the genome, linked in blocks, with the degree of linkage independent of distance. A similar conclusion was reached by Gabriel et al., who reported the presence of SNP blocks with a small number of common haplotypes, evenly distributed across the human genome. The study was conducted using DNA samples from 275 representatives of the American population of African, European and Asian origin. A link was found between the origins of the populations and specific haplotypes. Haplotypes with mutually exclusive alleles of polymorphic markers, called "yin" and "yang", were also identified [10–12].

Our study aims to perform a population analysis of combined into haplotypes SNP variants of the ATM gene, associated with oncological pathology. In addition, a genogeographic comparison of the North Eurasian gene pool of ATM haplotypes with the gene pools of the peoples of Eurasia and Africa was conducted. Phylogenetic analysis based on all identified

76 ATM haplotypes was performed to better understand the structural changes accumulated in the
77 gene during the evolutionary process.
78
79 ## MATERIALS AND METHODS
80 **Ethical approvals**
81 This study was conducted under the condition that written informed consent was obtained from
82 all subjects. This consent was approved by the Ethics Committee of the Research Center for
83 Medical Genetics.
84
85 **Collection of DNA genotype data and populations**
86 1975 DNA samples from individuals representing 27 populations living in Russia and adjacent
87 countries were provided by the Northern Eurasia Biobank. Population sample sizes ranged
88 from 41 to 129 individuals. Individual allelic variants of 269 polymorphic markers of the ATM
89 gene were extracted from a genome-wide genotyping dataset using the Infinium Omni5Exome
90 BeadChip Kit (Illumina Inc., San Diego, California).
91 Genotype information for 1901 individuals from 19 populations in Eurasia and Africa, with
92 population sample sizes ranging from 85 to 113 individuals, was obtained from the 1000
93 Genomes Project Phase 3 data (Ensembl).
94 The general list of populations, their abbreviated names and the size of population samples are
95 presented in Table 1.
96 The “ancestral” haplotype was determined based on the “ancestral” alleles of the polymorphic
97 markers specified in the Ensembl project, identified based on a three-stage multiple alignment
98 of EPO (Enredo, Pecan, Ortheus).
99

Statistical data processing

Microsoft Excel (Microsoft Corporation, Redmond, WA) and RStudio 2024.12.1+563 - R version 4.4.2 (RStudio, PBC) were used for statistical data processing.

DNA sample genotyping and polymorphic ATM loci quality assessment

Calculation of the percentage of genotyping of markers and DNA samples was performed using the snpStats package in the RStudio software. Polymorphic markers and samples with a genotyping quality of less than 95% were excluded from the study.

Isolation of haplotypes. Calculation of population frequencies of ATM haplotypes and genotypes

The isolation of haplotypes and the calculation of their population frequencies were performed in the Haploview 4.1 program (Broad Institute, Cambridge).

The "confidence interval method" (Gabriel et al.) was used to isolate blocks of polymorphic variants forming a haplotype. A haplotype block was defined if 95% of SNP pairs within it had D' > 0.8.

Calculation of population frequencies of genotypes based on ATM haplotypes was performed with RStudio software using the plyr package.

Phylogenetic analysis

Phylogenetic networks of ATM haplotype representation in populations were constructed in RStudio using the pegas and geneHapR packages according to the “parsimony principle” (parsimony network). The preparatory files were filled with data on population frequencies of haplotypes, multiplied by 1000.

125 **Calculation of Tajima test for populations and individual "yin" and "yang" branches**

126 The Tajima test was calculated based on the population haplotype frequency data obtained in
127 Haploview 4.1 software. The population number of haplotypes of each type was obtained by
128 multiplying the population frequency of the haplotype by 1000. The Tajima test D-value and
129 p-value were calculated using the pegas package (RStudio).

130

131 **Cartographic analysis**

132 Cartographic visualization of the population frequencies of haplotypes and genotypes of the
133 ATM gene was performed using the original GeneGeo cartographic package [13]. Frequency
134 calculations for the entire mapped area were performed using the weighted average
135 interpolation method with a weighting function inversely proportional to the cube of the
136 distance and an influence radius of 4000 km. To avoid the influence of map projection
137 distortions calculations, all distances in the interpolation process were calculated on a sphere.
138 The Winkel triple projection with the central meridian 60^0 E was used to create the maps.

139

140 **RESULTS**

141 **Determination of ATM gene haplotypes and their population frequencies**

142 The genotype data of the combined population of Northern Eurasia allowed the identification
143 of a linkage disequilibrium block (LD-block) of 28 linked polymorphic loci using Haploview
144 4.1 software – rs189037, rs228590, rs623860, rs600931, rs228592, rs672655, rs627418,
145 rs664677, rs609261, rs645485, rs619972, rs599558, rs595747, rs609655, rs227061, rs227062,
146 rs227064, rs227068, rs227070, rs227074, rs227075, rs425538, rs419716, rs227040, rs664143,
147 rs227094, rs227092 and rs4585, located over 145,795 bp from the 5′-untranslated region to the
148 3′-untranslated region of the ATM gene (Figure S1).

149 The value of the square correlation coefficient (r^2) between all pairs of loci in the LD block
150 exceeded 0.8. Of 268,435,456 possible haplotypes for 28 polymorphic loci of the LD block in
151 the combined population of peoples living in the territory of Northern Eurasia, 4 haplotypes
152 were found, and 2 of them were predominant: G-A-C-C-C-A-T-C-T-A-C-C-C-G-A-G-T-A-T-
153 A-T-C-A-T-A-C-G-G and A-G-T-T-A-G-C-T-C-G-T-T-T-A-G-A-C-G-G-G-C-A-C-C-G-A-T-
154 T with a frequency of 39.2% and 54.5%, respectively. The alleles in the two most frequent
155 haplotypes were mutually exclusive, which allows us to classify them as "yin"/"yang"
156 haplotypes. The remaining two detected haplotypes were much less frequent in the population
157 (with a frequency of about 2%) and were the result of recombination between the "yin" and
158 "yang" haplotypes.
159 Similarly, haplotypes were determined, and their population frequencies were calculated for 28
160 polymorphic markers of the ATM haplotype block in each of the 27 populations of Northern
161 Eurasia and in 19 populations of Eurasia and Africa using data from the 1000 Genomes Project
162 Phase 3 (Ensembl). From the total set of haplotypes detected in all of the above Eurasian and
163 African populations, rare (with a population frequency of up to 5%) found in single
164 populations, as well as haplotypes that most likely arose as a result of recombination between
165 the "yin-yang" haplotypes, were excluded. As a result, a list of 11 haplotypes of the ATM gene
166 (Figure 1) was determined, occurring in 46 populations of Eurasia and Africa. The "ancestral"
167 haplotype (H0) of the human ATM gene was determined using the Ensembl genome browser,
168 which indicates the "ancestral" alleles for each of the 28 linked loci of this gene.
169
170 **ATM haplotypes-based phylogenetic analysis**
171 Based on the isolated haplotypes, a phylogenetic network was constructed reflecting the
172 evolutionary variability of the ATM gene haplotype block in the human population and the
173 frequency ratio of ATM haplotypes in modern peoples of Eurasia and Africa (Figure 2).

The presence of two independently evolving phylogenetic branches, which we called I and J branches, was discovered, leading from the "ancestral" haplotype to the most common in Eurasia H6i and H5j ("yin" and "yang") haplotypes (Figure 3). All haplotypes analyzed were found in African populations (Figure 4, S2a).

In addition to the main "yin/yang" haplotypes, the H2i and H4i (branch I), H3j (branch J) haplotypes were also present in the Eurasia area. Of these three haplotypes, only H2i was found in populations of European origin (Figure S2b), in contrast to populations of Asian origin (Figure S2c). All three haplotypes are rare (up to 2%) in Eurasian populations. Based on these data, it can be concluded that the formation of the ancestral haplotype and the subsequent divergent evolution of two isolated branches probably occurred on the territory of the African continent.

Tajima test for populations and individual branches of “yin” and “yang” haplotypes

For all populations of Eurasia and Africa, both divided by origin and geographical location, and combined by these criteria, a positive value of the D parameter of the Tajima test (from 4.7 to 9.3) with a p value $<< 0.05$ was observed, which may be due to a long-term subdivision of the population into several subgroups with limited gene flow between them, balancing selection, a sharp reduction in population size, or other genetic and demographic processes.

For each of these populations, a significant deviation from the neutral model of evolution was observed. At the same time, when considering each of the two evolutionary branches leading from the “ancestral” haplotype to the H6i and H5j haplotypes separately, this test showed a statistically insignificant result for the D parameter and p-value ($p >> 0.05$), which indicating that both divergence paths (I and J) of the “ancestral” ATM haplotype conform to the neutral model of evolution (Table S1).

Geographical analysis of population frequencies of ATM haplotypes and genotypes

The frequencies of ATM haplotypes, as well as their homozygous and heterozygous combinations, are presented in Tables 2-3 and on maps (Fig. 5-6, Supplement: Fig. S3-S4).

"Yin" and "yang" haplotypes

Cartographic analysis of the prevalence of population frequencies of haplotypes H6i and H5j, as well as homozygous carriers of these haplotypes showed similar results.

Haplotype H5j was in southern Eurasia with maxima in the southern of Western Europe (~ 69%), in Transcaucasia (~ 72%) and in the north-west of South Asia (~ 70%). There was also a significant increase in the frequency of this haplotype in the populations of the indigenous populations of southern Siberia (with a maximum among Tuvans and Tofalars ~ 66%) and Kamchatka (Chukchi, Koryaks, Itelmens ~ 59%), with a significant decrease in East Asia to a frequency of ~36%. The H5j haplotype is the second most common in Africa, although its population frequencies are noticeable lower than in Eurasia - from 18% in most African populations to 24% in the Mende people living on the coast of West Africa.

The H6i haplotype has two geographic maxima - on the east coast of Asia (~ 56% in the Dai people in southwest China) and in eastern Europe (~ 49% in the Bashkirs). Between the maxima, there is a marked decrease in population frequencies in central Siberia, extending to the northern coast of Eurasia and the south of the continent - along the northern border of the Alpide belt. The lowest population frequency of the H6i haplotype in Eurasia is in Eastern Transcaucasia (~ 21%). The H6i haplotype is the fourth most common in African populations – from 4% in the Mende of Sierra Leone to 12% in the Yoruba of Nigeria.

In summary, it can be said that in the Eurasian area, with the exception of East Asia, the H5j haplotype predominates between the two mutually exclusive and most common haplotypes. In the populations of East Asia, the H6i haplotype predominates. In Africa, the population

224 frequencies of the "yin" and "yang" haplotypes are significantly lower than in Eurasian
225 populations. On average, the H5j haplotype is 2.8 times less common in African populations
226 than in Eurasian populations, and the H6i haplotype is 4.6 times less common.
227
228 Heterozygous combination of "yin" and "yang" haplotypes
229 The heterozygous genotype H6i/H5j turned out was one of the most common in Eurasian
230 populations (from 16.4% to 52.5%). The exception was 9 out of 42 Eurasian populations (Table
231 5). The prevalence of homozygotes for the H5j haplotype was found in 2 populations of
232 northwestern South Asia (41.7% in Gujaratis of India and 51% in Punjabis of Pakistan), in
233 populations of Transcaucasia (42% in Eastern Transcaucasia and 44.6% in the west), in the
234 Italian population of southern Europe (51.4%), as well as in 3 metapopulations of Northern
235 Eurasia - "Chukchi, Koryaks, Itelmens" (34.5%), "Mari, Chuvash" (35.3%) and "Tuvans,
236 Tofalars" (41.3%). The homozygous genotype for the H6i haplotype turned out was the most
237 common in only one population - the East Asian Han from Beijing, where its frequency was ~
238 34%. In Africa, the frequency of H6i/H5j heterozygotes was an order of magnitude lower than
239 in Eurasian (on average ~ 3.5%) than in Eurasia (~ 37%), although in the Yoruba, the H6i/H5j
240 genotype was the most common population genotype (6.5%), despite the complete absence of
241 homozygotes for the H6i haplotype and a very low population frequency of homozygotes for
242 the H5j haplotype (~ 1%). This may be due to the manifestation of balancing selection or
243 migration processes.
244 The peaks of the H6i/H5j genotype in Eurasia are located on the coasts: in the east of the
245 Indochinese peninsula (~ 52.5%), in the Spanish (~ 46.7%) and British (~ 47.3%) populations
246 of Western Europe, and in the Arkhangelsk region (~ 51.4%) of Northern Europe. There is also
247 a significant increase in heterozygote frequencies in the Eastern European Tatar population
248 (48.1%).

The Eurasian minimum of population frequencies of "yin-yang" heterozygotes is observed among the Chukchi, Koryaks and Itelmens (~ 16.4%), which is unusual, given the population frequencies of homozygous carriers of these haplotypes, which are at the "average European" levels.

Haplotype H3j

Haplotype H3j differs from H5j at 3 positions - rs189037, rs672655, rs619972. With a frequency of 1% to 7%, H3j is found in most Eurasian populations of Asian origin (with the exception of the West Siberian Plain and Vietnam) but has not been found in European populations. Homozygotes for H3j have not been found, apparently because of the low frequencies of the haplotype.

In Africa, this haplotype ranks 5th in population frequency ranging - from 4.5% in the Luhya population of Kenya to 11% in the Yoruba of Nigeria.

Geography of population frequencies of heterozygotes for the haplotype H3j in combination with H6i, H5j, H4i and H2i

The geographical distribution of the H3j haplotype in combination with H6i and H5j generally corresponds to that of the H3j haplotype in populations. The H3j/H6i genotypes are more common along the line running from southern Asia to northern Eurasia, with the highest frequencies among Sri Lankan Tamils (7.8%) and Bashkirs (7.1%). The H3j/H5j genotype is observed with the highest frequency (10.5%) in the Bengal population of South Asia.

The H3j/H4i and H3j/H2i genotypes are found only in Africa, the former with a frequency of 1-2% and the latter with a frequency of 2-6%.

Haplotype H2i

The most common in Africa, H2i (ranging from 19% to 31%), was found at very low levels (~2%) in some populations of East Asia (southern China and Vietnam) and Europe (UK, Western Caucasus, south-eastern central Russia, Mordovia and Tatarstan).

Prevalence of heterozygous carriers of the haplotype H2i in combination with H6i, H5j and H4i

The geographical distribution of heterozygotes for mixed haplotypes of the most common African H2i and the predominant Eurasian H6i and H5j is quite revealing.

The population frequencies of the H2i/H6i genotype are widespread, although at low frequencies, not only in African populations, where its occurrence is 2-5%, but also in some Eurasian populations, where the frequency of this genotype is 1-2%.

Heterozygous carriers of the H2i/H5j haplotypes are more common in Africa than heterozygotes for H2i/H6i, the range of population frequency for H2i/H5j is from 5 - 7%, and in the Mandinka population of Gambia this genotype is the most common.

Carriers of the H2i/H6i genotype have also been found in all populations of East Asia, Western Europe and some European populations of Northern Eurasia with a population frequency of 1-2%.

The H2i/H4i genotype is found in African populations with a frequency of 3-5%, with only the Mandinka having a frequency of about 2%. This genotype was also found in the southern East Asia and in the British and Italian populations of Europe with a frequency of 1-2%.

Haplotype H4i

Haplotype H4i was found to be the third most common in Africa. It differs from the "East Asian" H6i in 3 positions (rs623860, rs664677, rs609655) and its population frequencies in

Africa range from 9% in the Yoruba to 13.5% in the Mandinka of the Gambia and the Mende of Sierra Leone.

In Eurasia, this haplotype was found only in the Bashkir population (frequency ~1%), probably due to the heterogeneity of this population, which was formed by different ethnic groups.

Population features of ATM haplotypes in Africa

Haplotypes H1i, H3i, H5i, H1j, H2j and H4j were found only in African populations. The "ancestral" haplotype H0 is located in the phylogenetic network between haplotypes H1j and H1i.

The H1j haplotype, found in African populations with a frequency of 3% to 9% (on the west coast), differs from the "ancestral" haplotype in only one position (the G allele in the polymorphic marker rs645485, characteristic of H5j). The H1j haplotype was only found homozygous in the Mende of Sierra Leone, where the population frequency of its homozygous carriers was ~3.5%. A less common haplotype, H1i, is most common on the East and West African coasts (7-8%) and is almost absent in the central part of the continent (0 - 1%). H1i is genetically distant by 4 nucleotides from the ancestral H0 in branch I due to the alleles of the polymorphic markers rs227062, rs227074, rs227040 and rs4585. Homozygous carriers of the H1i haplotype have been found at a population frequency of about 1% in two African peoples, the Mandinka of Gambia and the Luhya of Kenya.

The H2j haplotype was found in the Luhya and Mende populations with a frequency of 1%.

The H3i haplotype was the rarest in the Mende population (~ 1%), it was slightly more common among the Yoruba and Ishan of Nigeria (3 - 3.5%), and its population frequency increased to 5 - 6% among the Luhya and Mandinka on the east and west coasts.

H4j is most common among the Mandinka (~ 7%) in West Africa, with its presence gradually decreasing in the east among the Luhya (1.5%).

In contrast, H5i has its maximum frequency on the east coast (~ 10%), while in other African populations its population frequencies are ~ 4 - 5%.

In general, the African continent is characterized by the greatest diversity of ATM haplotypes and the originality of their population distribution.

DISCUSSION

The study of ATM haplotypes in human populations of Eurasia and Africa from the standpoint of the "yin-yang" concept showed the presence in all populations of a conservative LD-block consisting of 28 extra-exon loci of the ATM gene located throughout the DNA sequence of the gene, including 5' and 3' untranslated regions of the ATM gene. Analysis of the phylogenetic relationships between 11 identified ATM haplotypes revealed the presence of two independent branches of ATM gene evolution from the ancestral haplotype, leading to the formation and widespread distribution of haplotypes with mutually exclusive alleles according to the "yin-yang" principle. In the samples analyzed outside the African continent, the dominant haplotypes in all human populations were the H6i haplotype (from 21 to 56%) and the H5j haplotype (from 32 to 72%), while the population frequency of the remaining haplotypes did not exceed 7% in Eurasian populations. In the sample from the African continent, 11 haplotypes were found, reflecting all stages of gene evolution from the "ancestral" haplotype. The most common was the H2i haplotype, its median frequency in African populations was 23%. On the territory of Eurasia, the H2i haplotype was found in only 7 populations, where its occurrence did not exceed 2%, which may be the result of evolutionary and demographic processes. Combining the obtained information on the evolutionary variability of ATM haplotypes with data on their frequencies in modern populations of Africa and Eurasia, it can be assumed that the process of formation of existing ATM haplotypes occurred on the African continent. Phylogenetic analysis of ATM haplotypes demonstrates the existence of two isolated ancestral

human populations in the past, followed by their merging into a single mixed population on the African continent (Figure 3). Similar conclusions were reached in the project [11], which was conducted using data from the human genome of the 1000 Genomes Project, as well as the genomes of Denisovans and Neanderthals - Eastern and Western. Dutta et al. found an abundance of mutually exclusive pairs of common haplotypes for population-frequent (>25%) polymorphic markers in the human genome. Moreover, these "yin" and "yang" haplotypes had close population frequencies in Eurasia. At the same time, this study identified a category of haplotypes that are mostly observed in Africa. These haplotypes are composed of fragments of "yin" and "yang" haplotypes, hence the term "mosaic" used by the authors.

The results of the Tajima test ($D > 4.7$, $p << 0.05$) for the combined population and individual populations of the world in our study confirm the fact of the probable division of the ancestral human population into several independent groups, and statistically insignificant D-values in the calculation of the Tajima test for haplotypes of the "yin" and "yang" branches ((- 0.6) and (- 0.8), at $p >> 0.05$) suggest a neutral nature of evolution for each of the branches.

The combined mixed population of carriers of ATM haplotypes, based on phylogenetic data, continued to exist on the African continent. Only a part of the ATM haplotypes from the "African" set were found on the territory of Eurasia. As a result of genetic drift and, possibly, balancing selection, the "yin" and "yang" haplotypes became dominant in Eurasia, while the others either disappeared completely or survived in small quantities in individual populations.

The peoples of Asia differ from those of European origin, firstly, by the widespread occurrence of the H3j haplotype (up to 7%) and, secondly, by the prevalence of the H6i haplotype in East Asia. It is important to note that in South Asia, the dominant haplotype among peoples of European origin, H5j, prevails, which allows us to consider the modal haplotype H5j in the population as a distinctive Indo-European feature. An analysis of literary data on the connection of the analyzed polymorphic loci with oncological diseases revealed one important

374 contradiction. In particular, in populations of East Asian origin, the H5j haplotype, containing
375 the rs189037, rs609261 and rs4585 loci, is associated with the manifestation of negative effects
376 in oncopathology. In peoples of European origin, this association is not observed, which may
377 be due to the presence of an additional factor, as yet unknown, which acts as an additive in East
378 Asian populations. It is also worth paying attention to the prevalence of heterozygotes for the
379 "yin" and "yang" haplotypes in most Eurasian populations. There are scientific data on the
380 positive effect of heterozygous carrier references between the GA genotype and longevity,
381 which the authors associate with a moderate level of ATM protein expression [14]. A study by
382 Ding et al. showed that heterozygous genotype for rs189037 is associated with a reduced risk
383 of type 2 diabetes in older adults [15].
384

385 **CONCLUSION**

386 Our study of evolutionary changes in the ATM gene haplotypes, represented by high-frequency
387 polymorphic loci throughout the DNA sequence of this gene, indicates a very ancient,
388 significant in time African period of isolation of two branches of the ancestral human
389 population and the subsequent merging of these branches into a single population space on the
390 African continent. Repeated migration events to the territory of Eurasia probably occurred after
391 the unification of ancient isolated subpopulations. Some differences in the composition of ATM
392 haplotypes in the east and west of the Eurasian continent, as well as in individual populations,
393 are probably due to the effect of gene drift.
394

395 **DECLARATION OF INTERESTS**

396 Authors have no conflict of interest to declare.
397
398

ACKNOWLEDGMENTS

Authors express the gratitude to the “Biobank of North Eurasia” for providing DNA samples, and to E.V. Balanovskaya for her comments and suggestions that help improve the article. The research was carried out within the state assignment of MSHE of RF for RCMG.

AUTHOR CONTRIBUTIONS

Marina V. Olkova – article idea, bioinformatics analysis, primary genogeographic analysis, writing the article text

Sergey M. Koshel – genogeographic analysis

Georgiy Yu. Ponomarev – statistical work, text editing

Andrei A. Alimov – research supervision, article editing

DATA AND CODE AVAILABILITY

The 1000 Genomes data can be obtained from the 1000 Genomes Project (Ensembl) Phase 3 data. The published article includes all data on Northern Eurasian populations analyzed during this study. R code are available at https://github.com/molkova/ATM-haplotypes/tree/main.

416 **REFERENCES**

417 1. Ueno S, Sudo T, Hirasawa A. ATM: Functions of ATM Kinase and Its Relevance to
418 Hereditary Tumors. Int J Mol Sci. 2022;23(1):523. https://www.mdpi.com/1422-
419 0067/23/1/523

420 2. Yang C, Tang X, Guo X, Niikura Y, Kitagawa K, Cui K, et al. Aurora-B Mediated ATM
421 Serine 1403 Phosphorylation Is Required for Mitotic ATM Activation and the Spindle
422 Checkpoint. Mol Cell. 2011;44(4):597–608.
423 https://linkinghub.elsevier.com/retrieve/pii/S1097276511008148

424 3. Zhao Z liang, Xia L, Zhao C, Yao J. ATM rs189037 (G > A) polymorphism increased the
425 risk of cancer: an updated meta-analysis. BMC Med Genet. 2019;20(1):28.
426 https://bmcmedgenet.biomedcentral.com/articles/10.1186/s12881-019-0760-8

427 4. Tao Y, Mei Y, Ying R, Chen S, Wei Z. The ATM rs189037 G>A polymorphism is
428 associated with the risk and prognosis of gastric cancer in Chinese individuals: A case–
429 control study. Gene. 2020;741:144578.
430 https://www.sciencedirect.com/science/article/pii/S037811192030247X

431 5. Zhou K, Bellenguez C, Spencer CCA, Bennett AJ, Coleman RL, Tavendale R, et al.
432 Common variants near ATM are associated with glycemic response to metformin in type
433 2 diabetes. Nat Genet. 2011;43(2):117–20. https://doi.org/10.1038/ng.735

434 6. Lee SA, Lee KM, Lee SJ, Yoo KY, Park SK, Noh DY, et al. Antioxidant Vitamins Intake,
435 Ataxia Telangiectasia Mutated (ATM) Genetic Polymorphisms, and Breast Cancer Risk.
436 Nutr Cancer. 2010;62(8):1087–94. https://doi.org/10.1080/01635581.2010.492088

437 7. Bonnen PE, Story MD, Ashorn CL, Buchholz TA, Weil MM, Nelson DL. Haplotypes at
438 ATM Identify Coding-Sequence Variation and Indicate a Region of Extensive Linkage

439 Disequilibrium. Am J Hum Genet. 2000;67(6):1437–51.
440 https://www.sciencedirect.com/science/article/pii/S0002929707632133

441 8. Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL. Haplotype and Linkage
442 Disequilibrium Architecture for Human Cancer-Associated Genes. Genome Res.
443 2002;12(12):1846–53. http://genome.cshlp.org/lookup/doi/10.1101/gr.483802

444 9. Thorstenson YR, Shen P, Tusher VG, Wayne TL, Davis RW, Chu G, et al. Global
445 Analysis of ATM Polymorphism Reveals Significant Functional Constraint. Am J Hum
446 Genet. 2001;69(2):396–412.
447 https://linkinghub.elsevier.com/retrieve/pii/S0002929707610854

448 10. Zhang J, Rowe WL, Clark AG, Buetow KH. Genomewide Distribution of High-
449 Frequency, Completely Mismatching SNP Haplotype Pairs Observed To Be Common
450 across Human Populations. Am J Hum Genet. 2003;73(5):1073–81.
451 https://linkinghub.elsevier.com/retrieve/pii/S0002929707619703

452 11. Dutta R, Mainsah J, Yatskiv Y, Chakrabortty S, Brennan P, Khuder B, et al. Intricacies in
453 arrangement of SNP haplotypes suggest “Great Admixture” that created modern humans.
454 BMC Genomics. 2017;18(1):433.
455 http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3776-5

456 12. Curtis D, Amos W. The human genome harbours widespread exclusive yin yang
457 haplotypes. Eur J Hum Genet. 2024;32(6):691–6.
458 https://www.nature.com/articles/s41431-023-01399-5

459 13. Koshel SM. Geoinformacionnye tehnologii v genogeografii // Sovremennaya
460 geograficheskaya kartografiya. 2012. P. 158–166.

461 14. Chen T, Dong B, Lu Z, Tian B, Zhang J, Zhou J, et al. A functional single nucleotide
462 polymorphism in promoter of ATM is associated with longevity. Mech Ageing Dev.
463 2010;131(10):636–40.
464 https://www.sciencedirect.com/science/article/pii/S0047637410001582

465 15. Ding X, Hao Q, Yang M, Chen T, Chen S, Yue J, et al. Polymorphism rs189037C > T in
466 the promoter region of the ATM gene may associate with reduced risk of T2DM in older
467 adults in China: a case control study. BMC Med Genet. 2017;18(1):84.
468 https://bmcmedgenet.biomedcentral.com/articles/10.1186/s12881-017-0446-z

469

FIGURES

	rs189037	rs228590	rs623860	rs600931	rs228592	rs672655	rs627418	rs664677	rs609261	rs645485	rs619972	rs599558	rs595747	rs609855	rs227061	rs227062	rs227064	rs227068	rs227070	rs227074	rs227075	rs425538	rs419716	rs227040	rs664443	rs227094	rs227092	rs4585	freq
ALLELE	A/G	G/A	T/C	T/C	A/C	G/A	C/T	T/C	C/T	G/A	T/C	T/C	T/C	A/G	G/A	A/G	C/T	G/A	G/T	G/A	C/T	A/C	C/A	C/T	G/A	A/C	T/G	T/G	
H0	G	A	T	C	A	A	C	T	T	A	C	T	T	A	A	A	C	A	G	G	C	A	A	C	G	A	G	T	10
H1i	G	A	T	C	A	A	C	T	T	A	C	T	T	A	A	G	C	A	G	A	C	A	A	T	G	A	G	G	209
H1j	G	A	T	C	A	A	C	T	T	G	C	T	T	A	A	A	C	A	G	G	C	A	A	C	G	A	G	T	334
H2i	G	A	T	C	C	A	C	T	T	A	C	C	T	A	A	G	C	A	G	A	C	A	A	T	G	C	G	G	1288
H2j	G	A	T	C	A	A	C	T	T	G	C	C	C	A	A	A	C	A	G	G	C	A	A	C	A	A	G	G	22
H3i	G	A	T	C	C	A	C	T	T	A	C	C	T	A	A	G	T	A	G	A	C	A	A	T	G	C	G	G	190
H3j	G	G	T	T	A	A	C	T	C	G	C	T	T	A	G	A	C	G	G	G	C	A	C	C	G	A	T	T	897
H4i	G	A	T	C	C	A	T	T	T	A	C	C	C	A	A	G	T	A	T	A	T	C	A	T	A	C	G	G	600
H4j	G	G	T	T	A	A	C	T	C	G	T	T	T	A	G	A	C	G	G	G	C	A	C	C	G	A	T	T	212
H5i	G	A	C	C	C	A	T	C	T	A	C	C	C	A	A	G	T	A	T	A	T	C	A	T	A	C	G	G	284
H5j	A	G	T	T	A	G	C	T	C	G	T	T	T	A	G	A	C	G	G	G	C	A	C	C	G	A	T	T	22911
H6i	G	A	C	C	C	A	T	C	T	A	C	C	C	G	A	G	T	A	T	A	T	C	A	T	A	C	G	G	16667

Figure 1 – ATM haplotypes

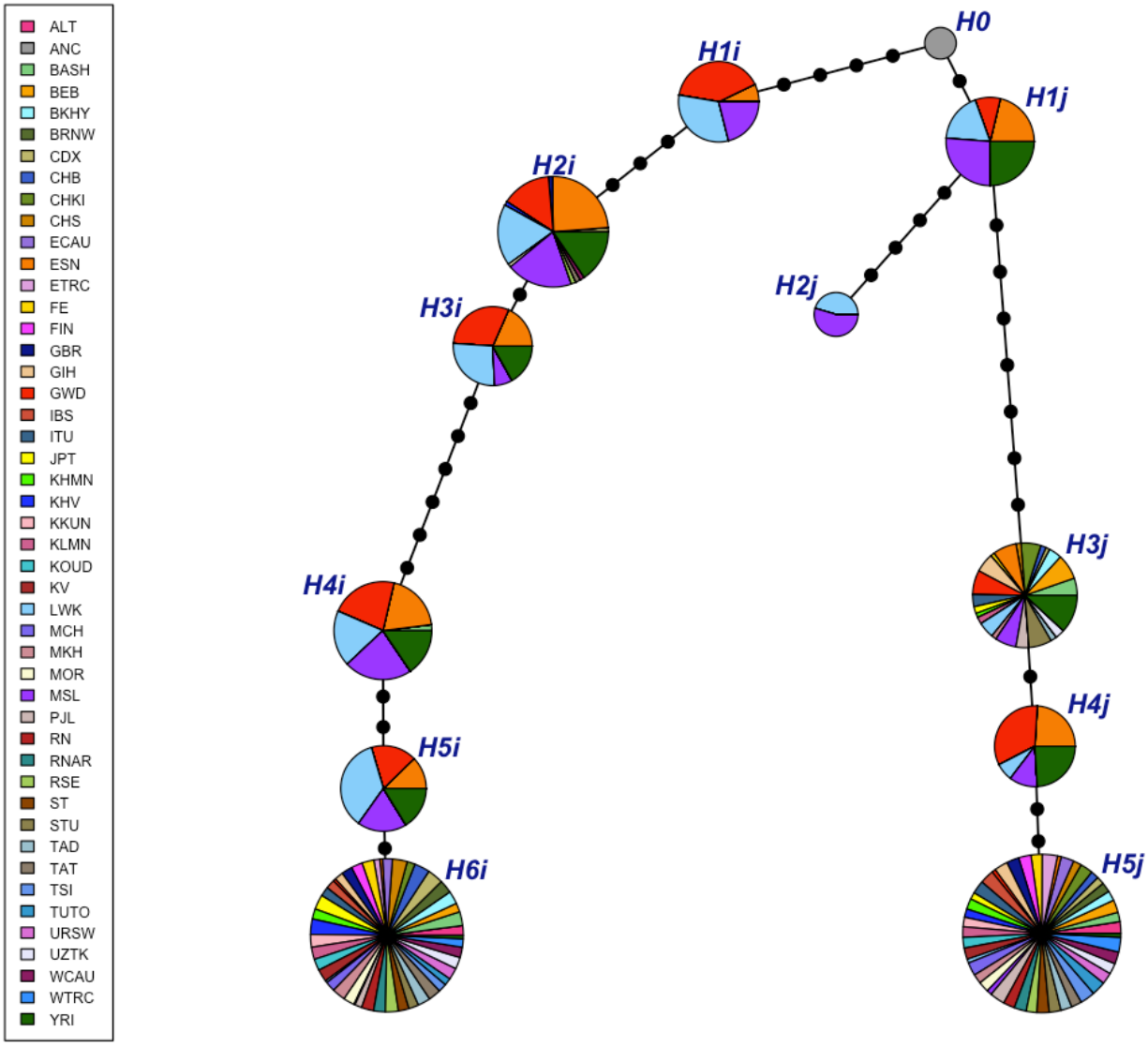

Figure 2 - Phylogenetic network of ATM haplotypes in Eurasian and African populations

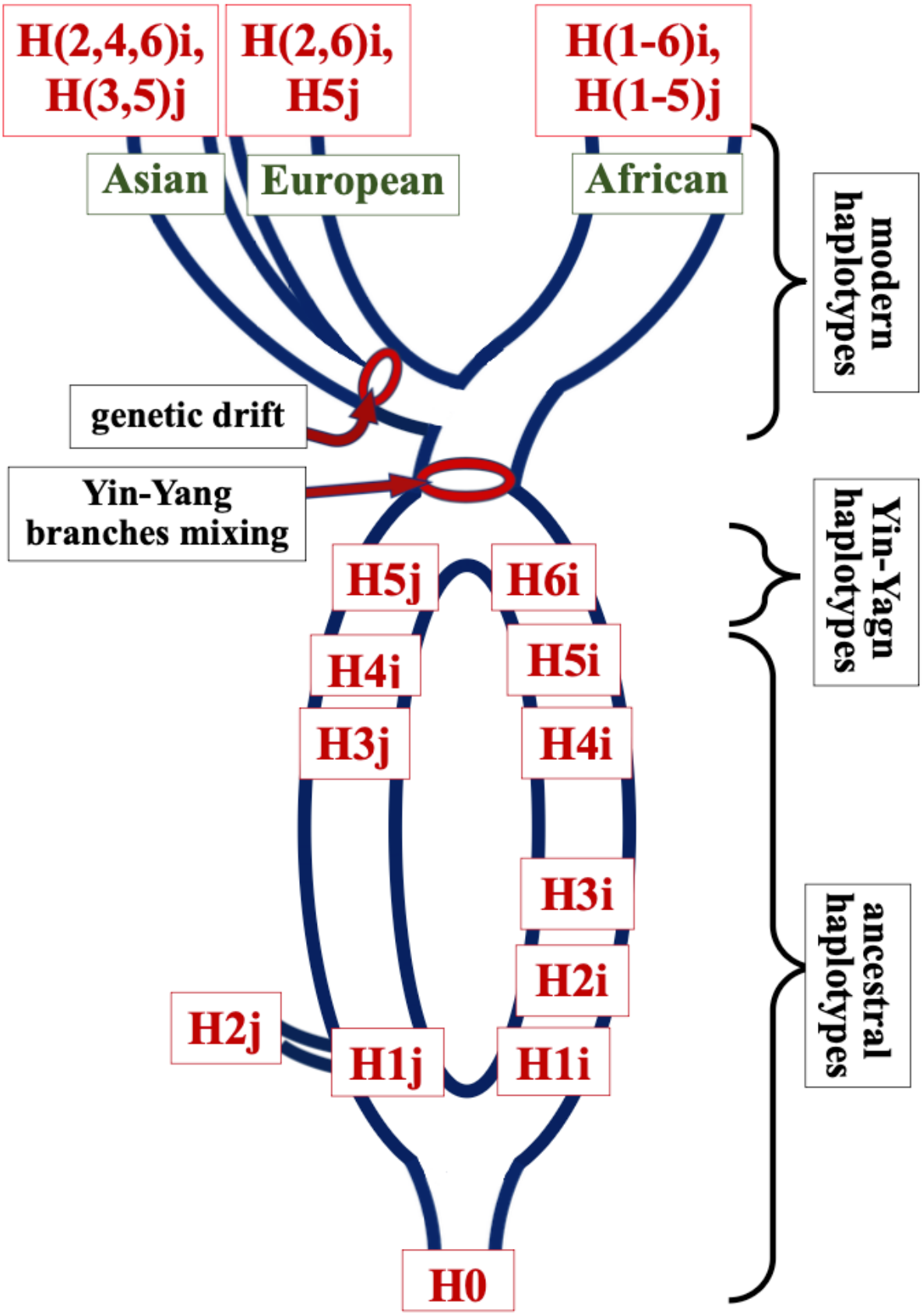

Figure 3 - Scheme of phylogenetic relationships of ATM haplotypes

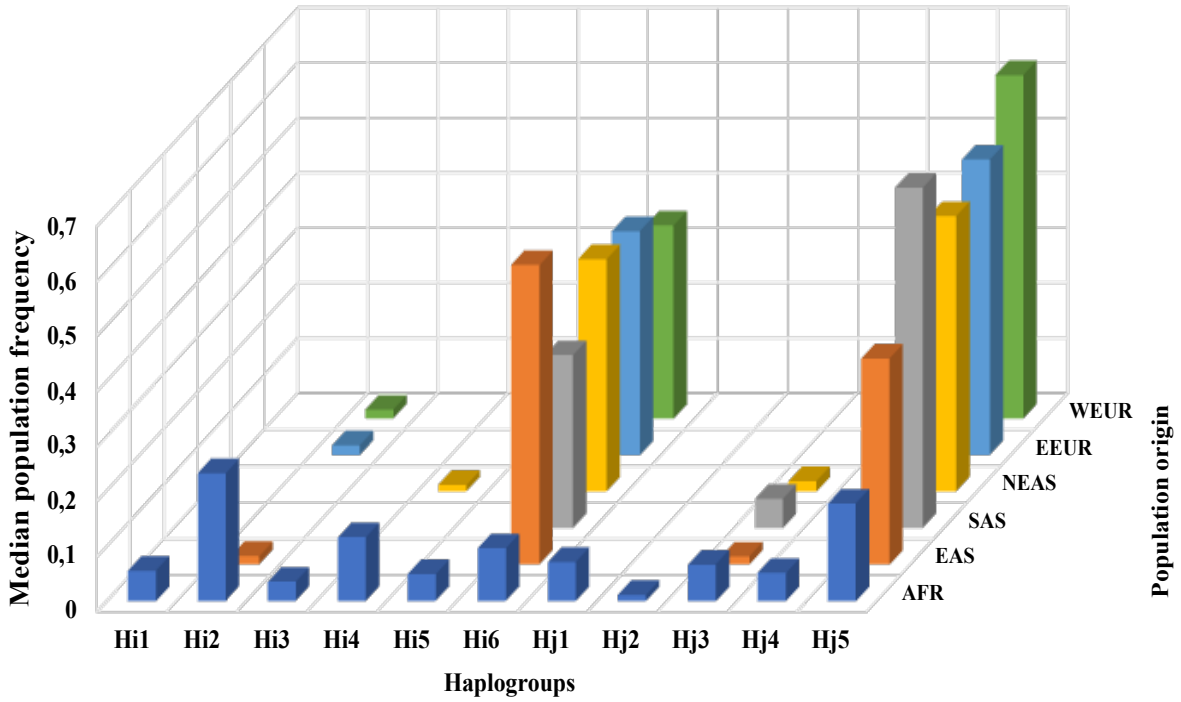

Figure 4 – Medians of population frequencies of ATM haplotypes in populations by origin:

AFR – African,

EAS – East Asian,

EEUR – Eastern European,

NEAS – Asian of Northern Eurasia,

SAS – South Asian,

WEUR – Western European

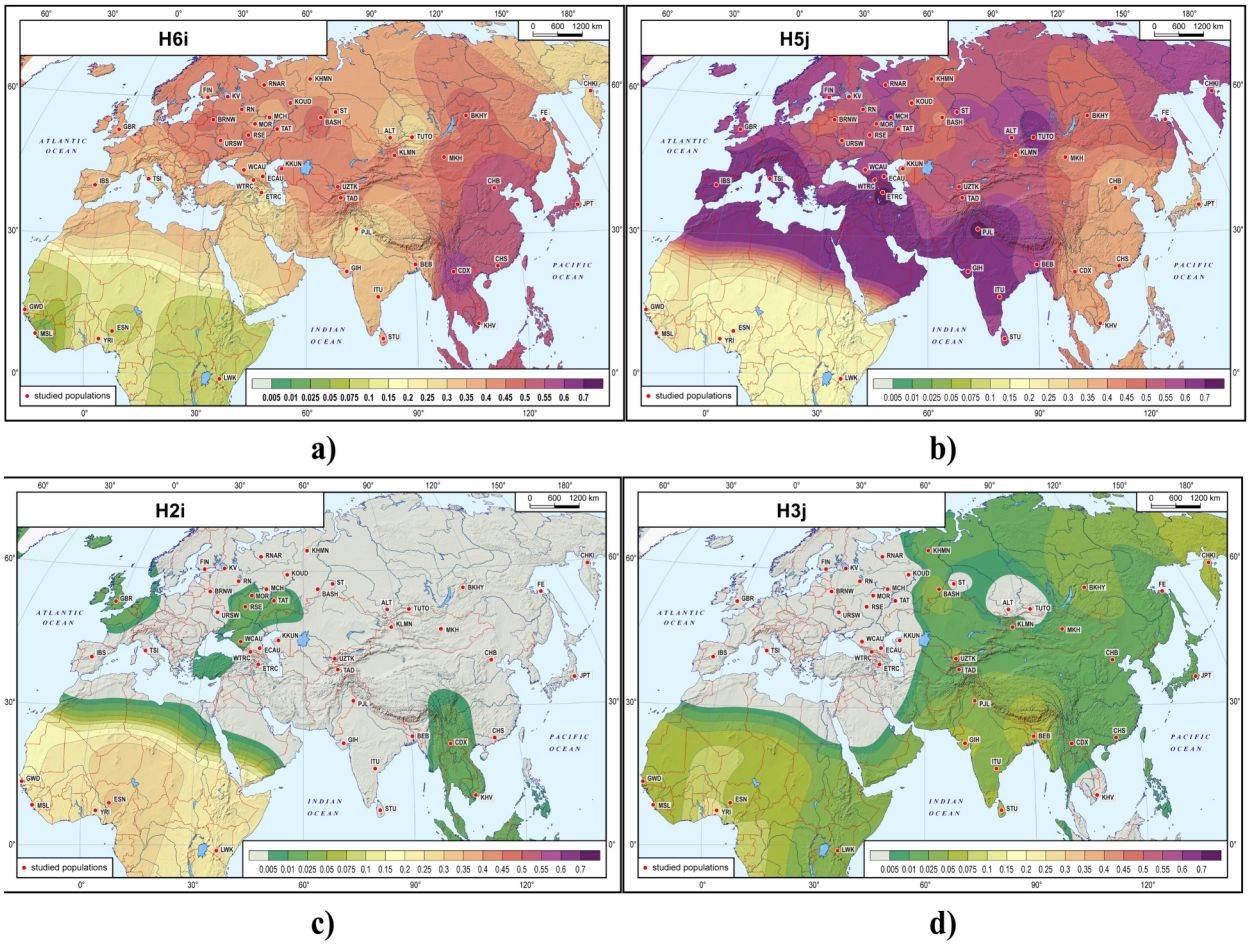

Figure 5 - Geographic distribution of the most common ATM haplotypes

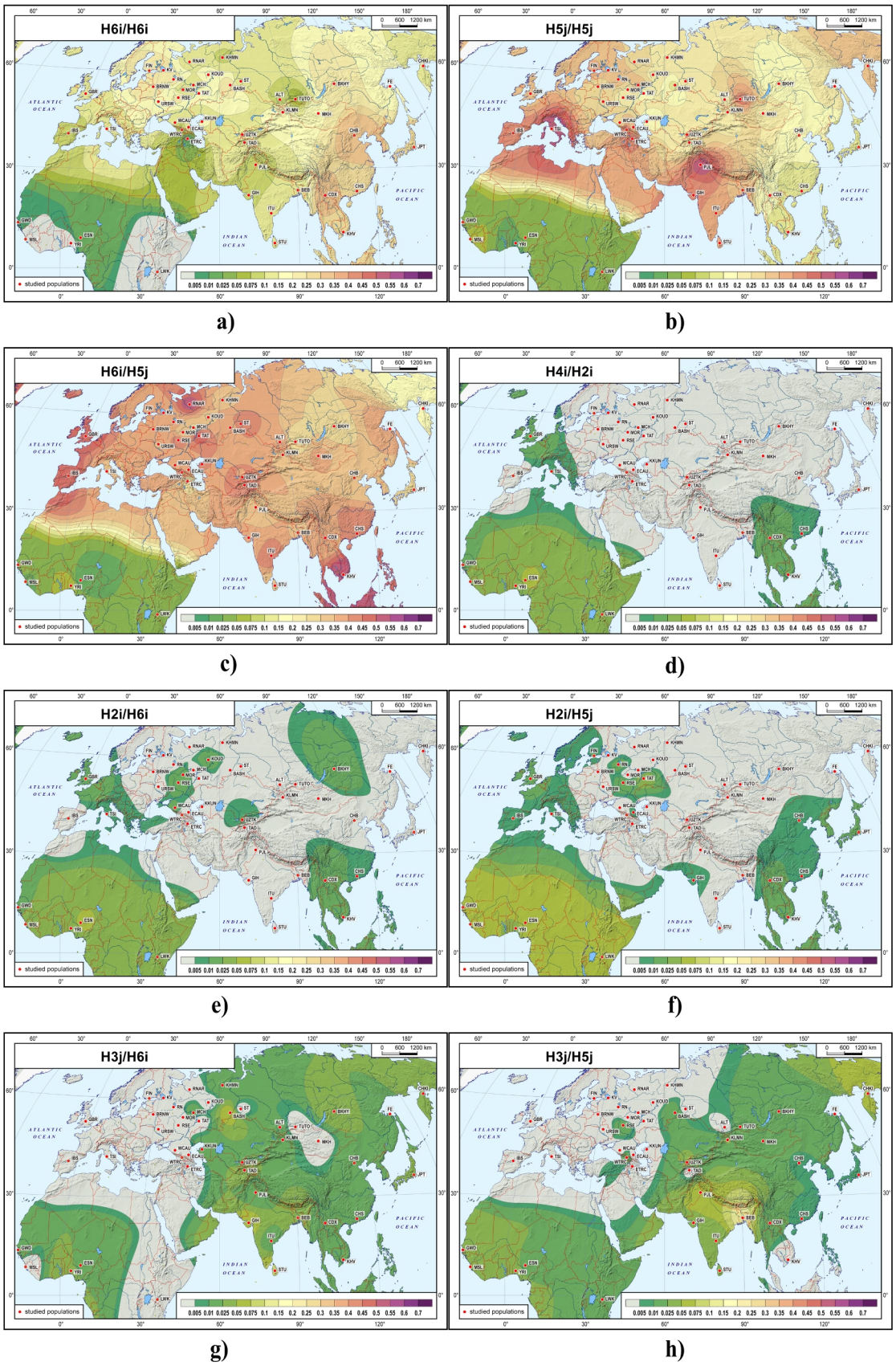

Figure 6 - Geographic distribution of the most widespread Eurasian ATM genotypes

491 **TABLES**

Abbreviation	Population	Population size
ALT	Peoples of Altai	83
ANC	Ancestral (Ensembl)	-
BASH	Bashkirs	44
BEB	Bengali	86
BKHY	Buryats, Khamnigans, Yakuts	59
BRNW	Belarusians and Russians of the North-West of the European Russia	61
CDX	Chinese Dai in Xishuangbanna, China	93
CHB	Han Chinese in Bejing, China	103
CHKI	Chukchi, Koryaks, Itelmen	75
CHS	Southern Han Chinese, China	105
ECAU	Peoples of the North-East Caucasus	123
ESN	Esan in Nigeria	99
ETRC	Peoples of Eastern Transcaucasia	87
FE	Peoples of the Far East	104
FIN	Finnish in Finland	99
GBR	British in England and Scotland	91
GIH	Gujarati Indian	103
GWD	Gambian in Western Division, The Gambia	113
IBS	Iberian populations in Spain	107
ITU	Indian Telugu	102
JPT	Japanese in Tokyo, Japan	104
KHMN	Khanty, Mansi, Nenets	63
KHV	Kinh in Ho Chi Minh City, Vietnam	99
KKUN	Kazakhs, Karakalpaks, Uighurs, Nogays	53
KLMN	Mongols (other groups) and Kalmyks	79
KOUD	Komi and Udmurts	84
KV	Karelians and Vepsians	59
LWK	Luhya in Webuye, Kenya	99
MCH	Mari and Chuvash	53
MKH	Mongols (Khalkha)	49
MOR	Mordovians	41
MSL	Mende in Sierra Leone	85
PJL	Punjabi in Lahore, Pakistan	96
RN	Russians of the North of the European Russia	84
RNAR	Russians from the north of the Arkhangelsk region	75
RSE	Russians of the South-East of Central Russia	62
ST	Siberian Tatars	68
STU	Sri Lankan Tamil	102
TAD	Tajiks, Pamiris, Yagnobis	72
TAT	Tatars	60
TSI	Toscani in Italy	107

492

493 Table 1 - Studied populations

Population	H1i	H1j	H2i	H2j	H3i	H3j	H4i	H4j	H5i	H5j	H6i
ALT										0,556	0,344
BASH						0,048	0,012			0,44	0,488
BEB						0,07				0,587	0,314
BKHY						0,036				0,464	0,455
BRNW										0,458	0,467
CDX			0,016			0,011				0,36	0,559
CHB						0,015				0,374	0,544
CHKI						0,055				0,591	0,28
CHS						0,014				0,39	0,524
ECAU										0,604	0,346
ESN	0,015	0,071	0,308		0,035	0,066	0,116	0,051	0,035	0,177	0,096
ETRC										0,715	0,209
FE						0,011				0,517	0,422
FIN										0,581	0,399
GBR			0,016							0,588	0,379
GIH						0,053				0,631	0,306
GWD	0,084	0,031	0,186		0,058	0,067	0,133	0,071	0,049	0,177	0,097
IBS										0,659	0,322
ITU						0,034				0,618	0,333
JPT						0,019				0,322	0,514
KHMN						0,012				0,492	0,359
KHV			0,015							0,399	0,545
KKUN										0,442	0,45
KLMN						0,019				0,494	0,43
KOUD										0,499	0,402
KV										0,533	0,44
LWK	0,066	0,061	0,232	0,01	0,051	0,045	0,111	0,015	0,101	0,182	0,076
MCH										0,616	0,353
MKH						0,014				0,443	0,471
MOR			0,012							0,499	0,39
MSL	0,044	0,088	0,25	0,012	0,014	0,059	0,135	0,024	0,053	0,241	0,041
PJL						0,036				0,703	0,25
RN										0,537	0,431
RNAR										0,568	0,412
RSE			0,018							0,5	0,456
ST										0,571	0,375
STU						0,064				0,554	0,353
TAD						0,017				0,517	0,441
TAT			0,019							0,497	0,442
TSI										0,692	0,299
TUTO										0,659	0,254
URSW										0,536	0,411
UZTK						0,026				0,5	0,408
WCAU			0,017							0,565	0,365
WTRC										0,687	0,295
YRI		0,083	0,199		0,032	0,106	0,093	0,051	0,046	0,176	0,12

Table 2 - Population frequencies of ATM haplotypes

See Table 1 for metapopulation abbreviations

Population	Population frequency of homozygous genotype									Population frequency of heterozygous genotype					
	H6i/H6i	H5j/H5j	H4i/H4i	H3i/H3i	H2iH2i	H1j/H1j	H1i/H1i	H5i/H5i	H4j/H4j	H6i/H5j	H2i/H6i	H2i/H5j	H3j/H6i	H3j/H5j	H4i/H2i
ALT	0,088	0,238								0,400			0,013		
BASH	0,238	0,143								0,357			0,071	0,024	
BEB	0,116	0,337								0,349			0,035	0,105	
BKHY	0,232	0,161								0,286	0,018		0,036	0,018	
BRNW	0,217	0,250								0,383					
CDX	0,333	0,151								0,355	0,022	0,011	0,011	0,011	0,022
CHB	0,340	0,175								0,330		0,010	0,019	0,010	
CHKI	0,145	0,345								0,164			0,036	0,055	
CHS	0,267	0,114								0,448	0,010	0,010	0,019	0,010	0,010
ECAU	0,123	0,298								0,360		0,009			
ESN	0,020	0,040	0,020	0,010	0,101					0,010	0,051	0,071	0,010	0,020	0,051
ETRC	0,012	0,420								0,272					
FE	0,178	0,211								0,378			0,011	0,011	
FIN	0,202	0,364								0,394		0,010			
GBR	0,121	0,341								0,473	0,022	0,011			0,022
GIH	0,097	0,417								0,359		0,010	0,058	0,049	
GWD	0,009	0,035	0,044	0,009	0,018		0,009		0,009	0,044	0,018	0,053	0,018	0,027	0,018
IBS	0,084	0,411								0,467		0,009			
ITU	0,108	0,382								0,422			0,020	0,039	
JPT	0,279	0,115								0,279		0,019	0,029	0,010	
KHMN	0,097	0,210								0,339			0,016		
KHV	0,263	0,101								0,525	0,010	0,020	0,010		0,010
KKUN	0,154	0,154								0,404			0,019		
KLMN	0,203	0,215								0,342			0,013	0,025	
KOUD	0,159	0,244								0,341	0,017				
KV	0,224	0,328								0,362					
LWK		0,030	0,020		0,061		0,010	0,010		0,030	0,040	0,051		0,020	0,040
MCH	0,137	0,353								0,294			0,020		
MKH	0,200	0,244								0,356				0,022	
MOR	0,098	0,244								0,317	0,024				
MSL		0,059	0,024		0,082	0,035				0,024	0,047	0,071		0,047	0,047
PJL	0,083	0,510								0,313			0,010	0,063	
RN	0,213	0,338								0,363		0,013			
RNAR	0,135	0,297								0,514					
RSE	0,211	0,228								0,421	0,018	0,018		0,018	
ST	0,107	0,250								0,429					
STU	0,147	0,363								0,314			0,078	0,039	
TAD	0,136	0,220								0,424			0,034		
TAT	0,154	0,173								0,481		0,038			
TSI	0,121	0,514								0,346	0,009	0,009			0,009
TUTO	0,063	0,413								0,254				0,016	
URSW	0,195	0,285								0,350					
UZTK	0,143	0,235								0,398	0,010		0,020	0,031	
WCAU	0,101	0,292								0,360	0,011				
WTRC	0,071	0,446								0,411				0,018	
YRI		0,009	0,009		0,019					0,065	0,046	0,046	0,046	0,037	0,046

Table 3 - Population genotype frequencies based on ATM haplotypes

See Table 1 for metapopulation abbreviations

SUPPLEMENT

Figures

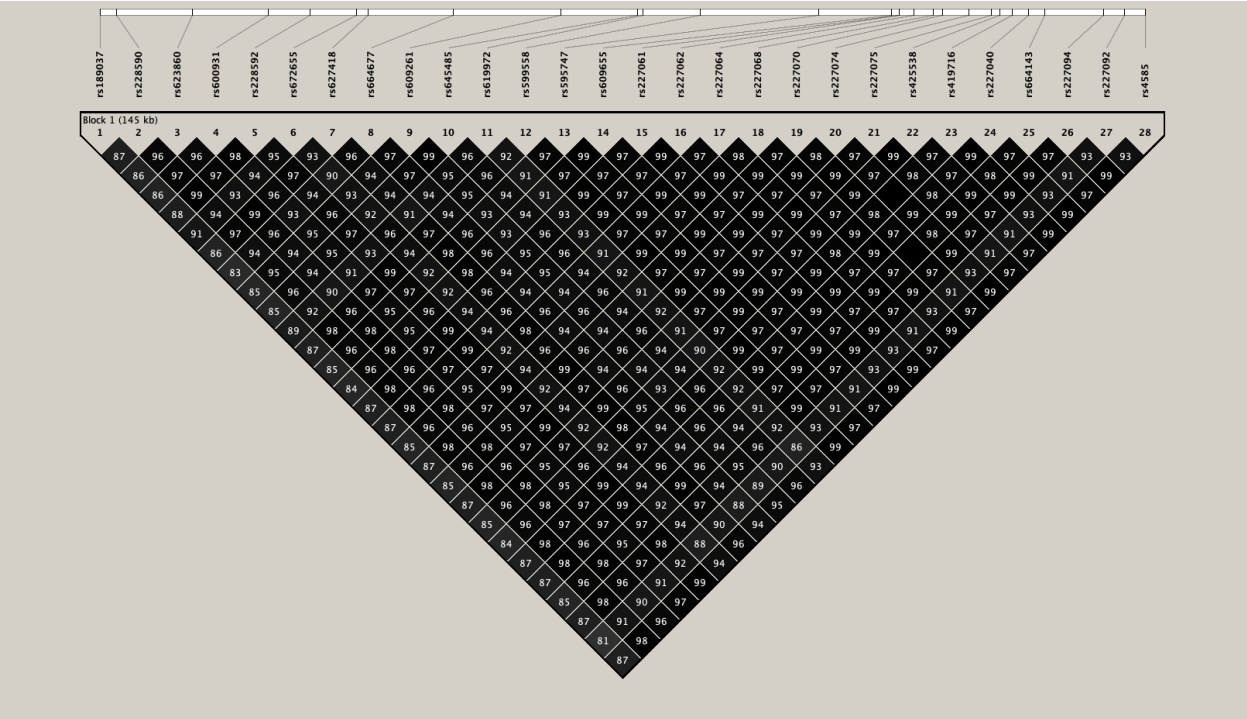

Figure S1 - LD-block of ATM SNPs. The numbers in the cells are the $\hat{r}^2$ value multiplied by 100

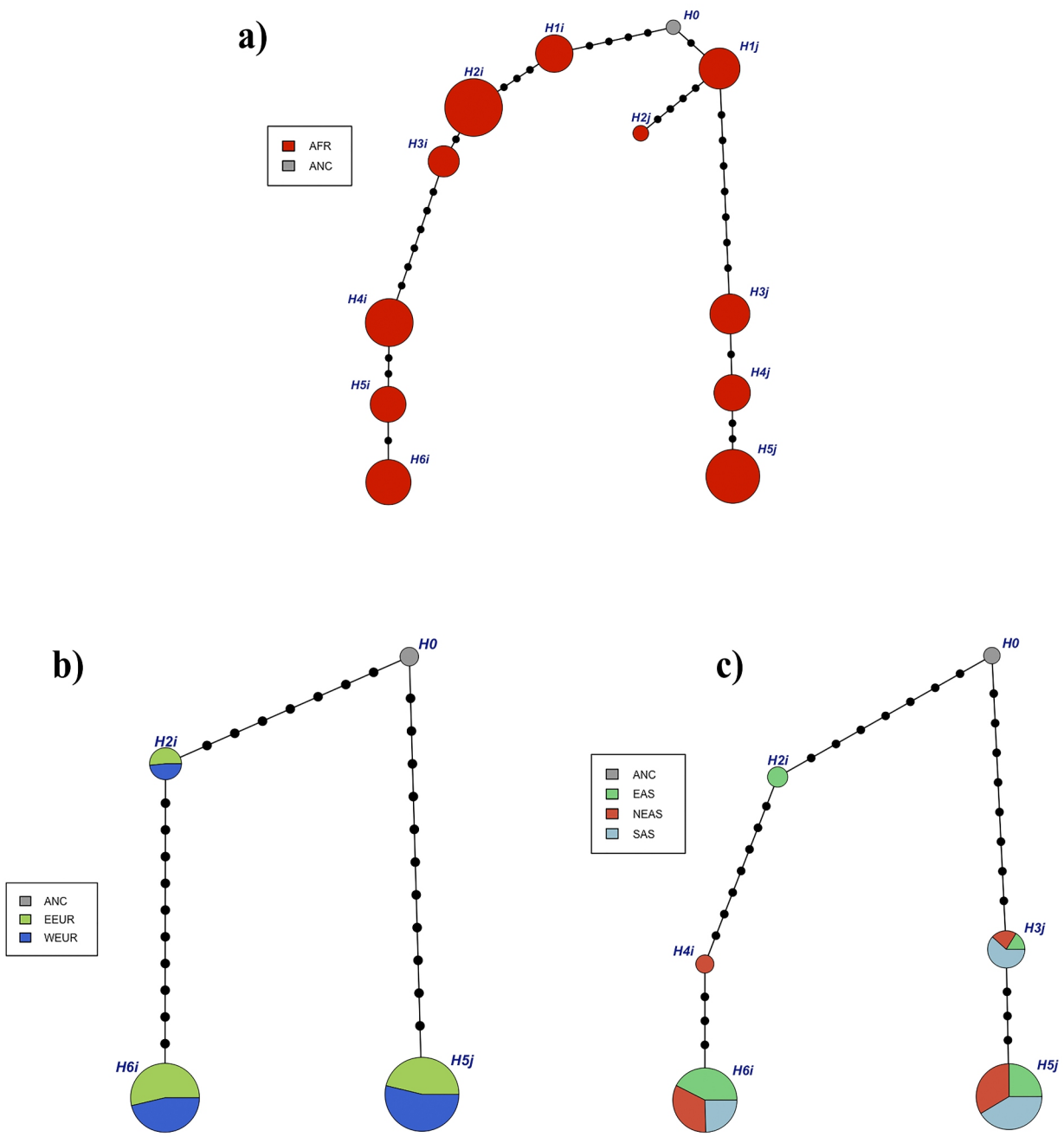

Figure S2 - Phylogenetic networks based on medians of population frequencies of ATM haplotypes for populations of African (a), European (b) and Asian (c) origin

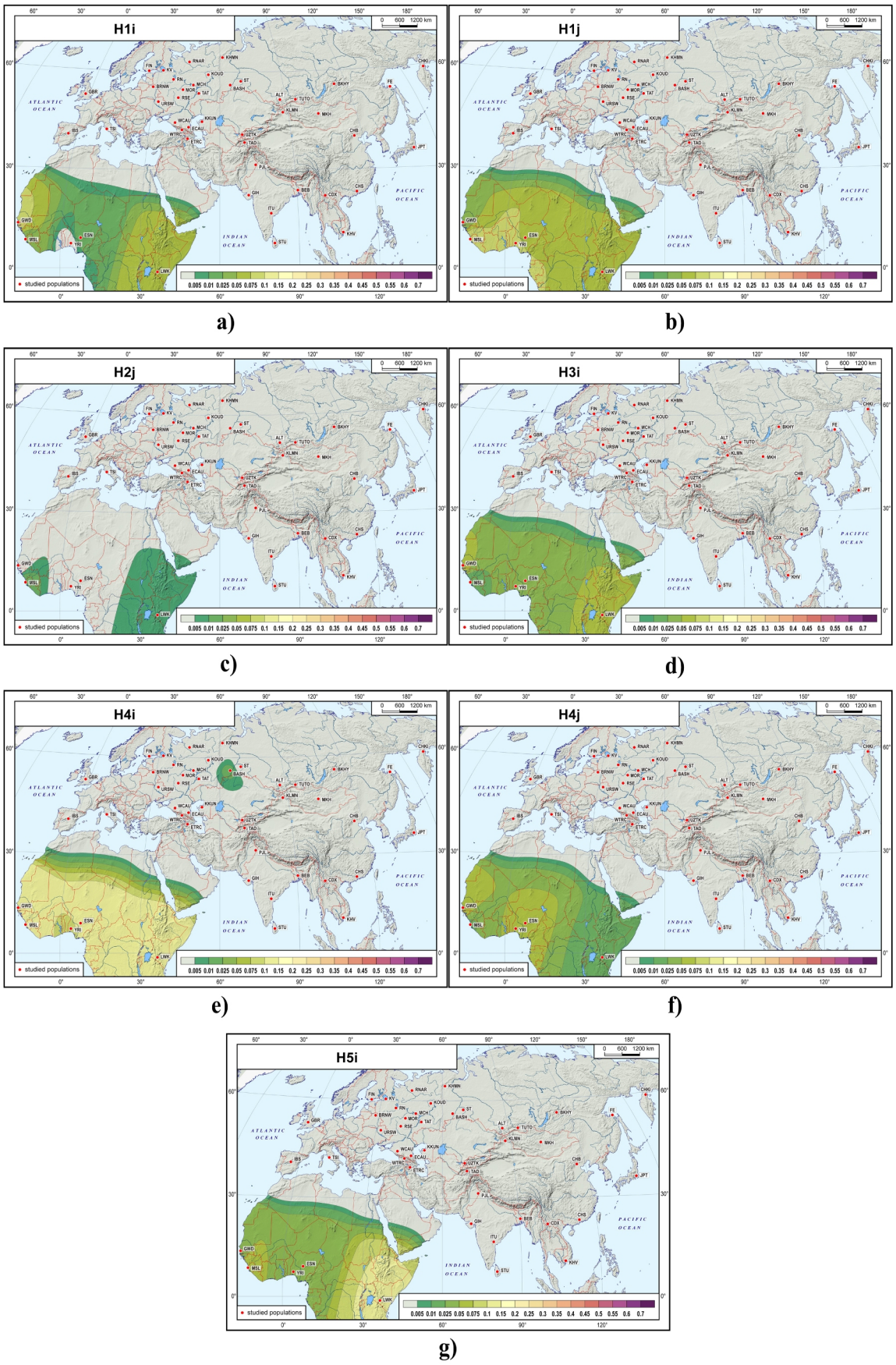

Figure S3 - Geographic distribution of African-specific ATM haplotypes

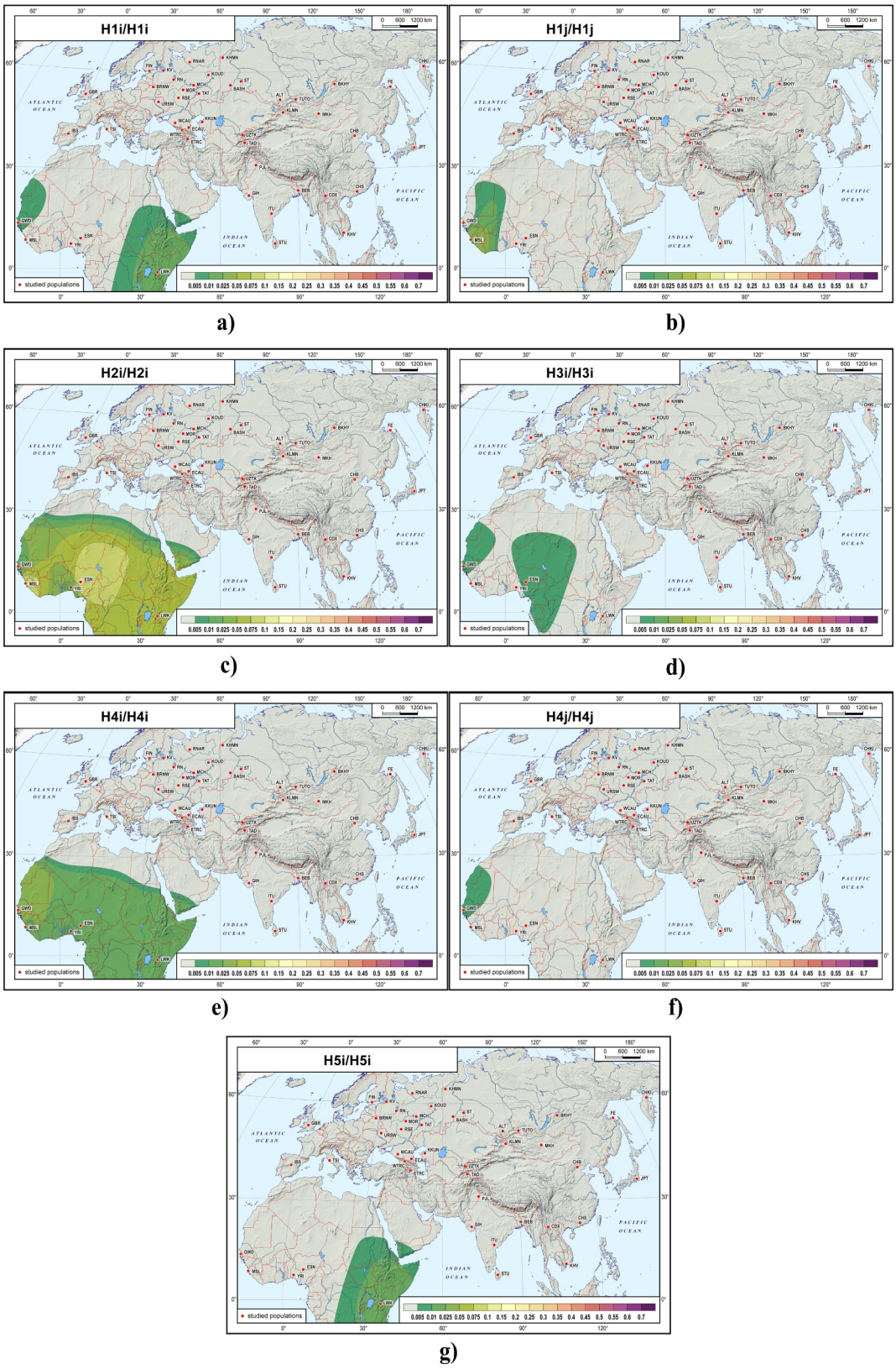

Figure S4 - Geographic distribution of ATM genotypes defined by African-specific haplotypes

514 **Tables**

Population	D	p-value	Origin
Combined population of Eurasia and Africa	6,5	<< 0.05	
Africa	6,1	<< 0.05	AFR
Northern Eurasia	8,9	<< 0.05	EEUR and NEAS
EUR&EAS&SAS	8,4	<< 0.05	EUR, EAS and SAS
Set of ATM haplotypes in the I branch	-0,6	> 0.05	ATM branch I
Set of ATM haplotypes in the J branch	-0,9	> 0.05	ATM branch J
EUR	7,0	<< 0.05	EUR
EAS	7,8	<< 0.05	EAS
SAS	6,7	<< 0.05	SAS
FIN	6,5	<< 0.05	EUR
GBR	6,4	<< 0.05	
IBS	5,7	<< 0.05	
TSI	5,3	<< 0.05	
BEB	5,6	<< 0.05	SAS
GIH	5,5	<< 0.05	
ITU	5,8	<< 0.05	
PJL	4,6	<< 0.05	
STU	6,1	<< 0.05	
CDX	6,5	<< 0.05	EAS
CHB	6,5	<< 0.05	
CHS	6,7	<< 0.05	
JPT	6,5	<< 0.05	
KHV	6,6	<< 0.05	
MSL	4,9	<< 0.05	AFR
ESN	4,8	<< 0.05	
GWD	5,2	<< 0.05	
LWK	4,8	<< 0.05	
YRI	5,4	<< 0.05	
ALT	6,3	<< 0.05	NEAS
BASH	6,8	<< 0.05	
BKHY	6,7	<< 0.05	
CHKI	6,0	<< 0.05	
FE	6,7	<< 0.05	
KHMN	6,7	<< 0.05	
KKUN	6,7	<< 0.05	
KLMN	6,7	<< 0.05	
MKH	6,7	<< 0.05	
ST	6,4	<< 0.05	
TAD	6,7	<< 0.05	
TUTO	5,5	<< 0.05	
UZTK	6,6	<< 0.05	
BRNW	6,8	<< 0.05	EEUR
ECAU	6,1	<< 0.05	
ETRC	4,6	<< 0.05	
KOUD	6,7	<< 0.05	
KV	6,6	<< 0.05	
MCH	6,0	<< 0.05	
MOR	6,6	<< 0.05	
RN	6,6	<< 0.05	
RNAR	6,5	<< 0.05	
RSE	6,7	<< 0.05	
TAT	6,7	<< 0.05	
URSW	6,5	<< 0.05	
WCAU	6,3	<< 0.05	
WTRC	5,3	<< 0.05	

515

516 Table S1 - Tajima's test for ATM haplotypes

517 Abbreviations for metapopulations can be found in Table S1, and for populations by origin:

518 AFR – African, EAS – East Asian, EEUR – Eastern European, NEAS – Asian of Northern

519 Eurasia, SAS – South Asian, WEUR – Western European