

Original Research

Genetic Differentiation of Populations of Three Megalopolises by DNA Markers of the Y-Chromosome in Connection with the Problem of Developing Genetic Databases

Irina G. Udina ^{1,*}, Marina A. Gubina ², Alesya S. Gracheva ¹

1. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Gubkin str. 3, 119991, Russia; E-Mails: irina_udina@mail.ru; palesa@yandex.ru
2. Federal Research Center Institute of Cytology and Genetics of the Siberian Department of the Russian Academy of Sciences, Novosibirsk, Prospekt Lavrentyeva 10, 630090, Russia; E-Mail: marina@bionet.nsc.ru

* **Correspondence:** Irina G. Udina; E-Mail: irina_udina@mail.ru**Academic Editor:** Ying S. Zou

OBM Genetics

2026, volume 10, issue 1

doi:10.21926/obm.genet.2601324

Received: June 04, 2025**Accepted:** January 06, 2026**Published:** January 12, 2026

Abstract

The purpose of the study was to consider the distribution of Y-chromosome DNA markers in samples from the populations of the three largest megalopolises (Moscow, Saint Petersburg, and Novosibirsk) in the Russian Federation, in the context of developing genetic databases. The study aimed to compare the frequency profiles of 18 Y-chromosome STRs (Short tandem repeats) and the level of genetic differentiation. Based on F_{ST} estimates for the distribution of these 18 Y-chromosome STRs, the senior generations of Moscow and Novosibirsk were found to be similar. Statistically significant differences were detected between the Novosibirsk sample and the samples from Saint-Petersburg and from the young generation of Moscow ($F_{ST} = 0.0087$, $p < 0.05$ and $F_{ST} = 0.0084$, $p < 0.01$, respectively). For probability prediction of Y-chromosome haplotypes, analysis of 18 STR-haplotypes using Internet-predictor - Whit Athey's Haplogroup Predictor was performed. The distribution of Y-chromosome haplogroups in the studied samples from the three megalopolises largely corresponds to the Russian gene pool, with the most frequent haplogroups being R1a, R1b, E1b1b1, N, I1, I2, J1, and J2 — predominantly R1a. This aligns with the predominance of ethnic Russians within the



© 2026 by the author. This is an open access article distributed under the conditions of the [Creative Commons by Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is correctly cited.

populations of these large urban centers. Significant differences for two samples representing senior generations in two megalopolises were detected by the “Southern origin” haplogroups (G2a, G2c, J1, J2, L, O2, O3, Q, R2, and T): 3.4% in Novosibirsk and 11.2% in Moscow ($G = 6.1081$, $df = 1$, $p < 0.05$). These frequencies are notably higher in the sample from the young generation of Moscow (21%) and in the sample from Saint-Petersburg (16%). The observed distribution of Y-chromosome DNA markers is consistent with the observed migration parameters, reinforcing the conclusion regarding the necessity of developing reference databases of Y-chromosome DNA markers specific to each megalopolis, based on both molecular studies and genetic-demographic questionnaires. These databases must be subject to timely updates to account for the changes in the gene pool, particularly those driven by migration. Such an approach is particularly relevant for monitoring the gene pool dynamics of populations in megalopolises and ensuring the ongoing applicability of genetic databases.

Keywords

Novosibirsk; Moscow; population; megalopolis; migration; gene pool; Y-chromosome haplogroups; STR; reference database

1. Introduction

Affiliation to a definite haplogroup is determined by the presence of the main substitutions (SNP – Single nucleotide polymorphism of Y-chromosome), which are cataloged in haplogroup nomenclature [1, 2]. These definitive markers for Y-chromosome haplogroups are localized in the Non-Recombinant Region of the Y-chromosome (NRY) and are inherited as a single block. A standard method for detecting Y-chromosome haplogroups is based on the definition of SNP markers with relatively low mutation rates and the accumulation of mutations in the absence of recombination across generations. During the transmission of the Y chromosome across generations, not only does a specific SNP profile accumulate, but linked STR-markers — markers with higher mutation rates — also accumulate. This provides a theoretical basis for predicting affiliation with a specific haplogroup from the STR haplotype of the human Y chromosome. This recent advancement relies on sophisticated computational models and large reference databases to predict an individual's haplogroup affiliation from their STR profile across various populations [3-9].

For comparison of probable detection results or prediction of Y-chromosome haplogroups, various predictors were used; overall, good agreement between prediction of haplogroup affiliation by STR-haplotype and the direct SNP-detection method was observed [10, 11]. The reliability of Y-chromosome haplogroup prediction is typically assessed through two main criteria: the estimated probability of correct assignment and the concordance of haplogroups predicted across different predictors [3, 4, 11]. To minimize the risk of incorrect haplogroup assignment, it is recommended to utilize multiple predictors or a sufficiently large number of STR loci [3, 4, 11].

Individuals belonging to the same Y-chromosome haplogroup are typically related, either as full siblings or as paternally descended descendants of a common male ancestor. These defining SNPs are crucial for tracing ancient migration routes and understanding demographic processes, thereby contributing significantly to our knowledge of modern human population formation. Simultaneous

analysis of variability in STR and SNP markers of Y chromosome is applied for studying of subclade differentiation within haplogroup and dating of events related to history of human settlement due to rapid mutation rate of STR [12-14].

Due to the uniqueness and high informativity of Y-chromosome DNA markers, they are successfully used in forensic genetics. Databases, including these markers, are widely developed worldwide [15, 16]. The resolution of modern DNA technologies is very high, allowing DNA identification and establishing kinship. At present, more detailed analysis of the Y chromosome based on sequencing data opens additional possibilities for population studies [17]. The probability of a random coincidence of a single genotype by DNA-markers in the test sample with the genotype of another resident in the population might be estimated based on the frequency data of corresponding DNA-markers in a concrete population. Reference database for application in DNA-identification might be equivalent to a representative population database, in which informative DNA-markers for performing forensic examination are included [18, 19]. Autosomal STR panels [20], and also, molecular markers of mtDNA and Y-chromosome are applied [19, 21-23].

The study of Y-chromosome haplogroups allows us to analyze the genetic structure of populations of mixed origin, which are historically formed as a result of the mixing of different-origin population flows [24, 25]. For megalopolises, as populations of mixed type, specific characteristics of genetic and demographic processes, including a large population, multi-ethnicity, a high percentage of interethnic marriages, intensive migration, territorial subdivision, and population reproduction through external migration, are typical [26]. Migration is established as the main factor of megalopolis gene pool dynamics [27]. According to the results of the All-Russian census of 2010, the migration rate for the population of Saint Petersburg was 0.438, for Moscow – 0.418 [28, 29].

Study of molecular markers and haplogroups of the Y chromosome is very informative for assessing migration to megalopolises and detecting the impact of various gene flows only of male-line gene pool of megalopolis populations line [30, 31]. Developing reference forensic databases for megalopolises is in high demand both in the social sphere for kinship establishment and in forensics, as threats from artificial disasters, terrorist attacks, and crime are growing. The specificity of developing reference databases must take into account the structure of megalopolises and characteristic genetic-demographic processes that determine the dynamics of their gene pool.

Earlier, it was shown that the regions of migration attraction of the three megalopolises under consideration – Moscow, Novosibirsk, and Saint-Petersburg – with a prevailing Russian population differ noticeably, which evidences the necessity of developing megalopolis-specific reference databases [28, 29, 32]. It is essential to account for the age distribution of megalopolis residents when studying the distribution of DNA markers to develop reference databases, because varying migration flows influence different generations.

The population of Novosibirsk, the third most populous megalopolis of the Russian Federation, has previously been characterized in relation to genetic-demographic processes and their parameters. For the young generation of the Novosibirsk population, gender differences in intensity of migration and its ethno-regional composition were demonstrated [32].

The aim of investigation was to study distribution of 18 STR-haplotypes and predicted haplogroups of Y-chromosome in the population of Novosibirsk (representing senior generation), and to perform comparative analysis with earlier studied samples from senior and young generations of Muscovites [30] and inhabitants of Saint - Petersburg [31], for comparison of

distribution of these markers of Y-chromosome in the population of megalopolises in Siberia and the European part of Russia.

2. Materials and Methods

2.1 Characteristics of the Samples Studied

In 2001-2004 years in the frames of the budget fund of the Novosibirsk Research Institute of therapy and prevention – branch of the Institute of Cytology and Genetics of Siberian department of the Russian Academy of Sciences, population study of Novosibirsk inhabitants was performed based on medical institutions, which included genetic-demographic questionnaire survey and collection of blood samples (additional collection of samples was conducted in 2014-2016). For the study, men residing in various regions of Novosibirsk were selected ($N = 116$). Criterium of inclusion – healthy inhabitants of Novosibirsk, of male gender, aged over 18 years old, who are non-relatives. Average age of men in the sample was 57.1 ± 0.6 years and average year of birth was 1944.9 ± 0.6 (years of birth: 1934-1958).

Comparative data from earlier studies of Moscow samples were included [29, 30]. United sample from senior generations, which included two samples from the male population of Muscovites: the first ($N = 73$) with an average year of birth 1950.7 ± 2 (1914-1984), and the second ($N = 96$) with average year of birth 1981 ± 0.35 (1956-1989), were studied. Additionally, a sample from the young generation, consisting of newborns born in 2017-2018 ($N = 400$), was investigated. Two samples from the senior generations of Moscow included Muscovites born in the Soviet Union before its collapse, which caused intensive migration into the megalopolis. Samples were combined because earlier, the absence of significant differences in the haplogroup profiles of the Y chromosome was demonstrated [30]. Established age parameters from the senior generations provide opportunities for studying the gene pool in the male populations of two megalopolises before the modern intensive migration from regions of the North Caucasus and Transcaucasia, and subsequent migration from countries of Central Asia generated by the collapse of the Soviet Union (1991). The sample from the young generation of the Moscow population represents residents born a generation after 1991 – the year of the collapse of the Soviet Union and the emergence of intensive migration from the North Caucasus and Transcaucasia, and from Central Asia into the population of the megalopolis [29, 30]. For comparison, the earlier studied sample from Saint-Petersburg, which included residents born before and after 1991 ($N = 150$), was included [28, 31].

2.2 Genetic-Demographic Questionnaire Survey

For residents of megalopolises, genetic-demographic data were collected by questionnaire survey and included data about date and place of birth, ethnic affiliation, and also, analogous data for the ancestors in two previous generations.

2.3 Genotyping by 18 STR of the Y-Chromosome

Variability of 18 STRs of the Y chromosome was studied. For residents of Novosibirsk, total DNA was isolated from peripheral blood lymphocytes using the standard method with phenol-chloroform extraction and proteinase K [33]. Genotyping of the samples of DNA by 18 STR of Y-chromosome (DYS389I, DYS389II, DYS390, DYS19, DYS385A, DYS385B, DYS456, DYS437, DYS438,

DYS447, DYS448, DYS449, DYS391, DYS392, DYS393, DYS439, DYS635 and DYS576) performed on the basis of OOO «Gordiz» (Moscow) by multiplex PCR with fluorescent primers and subsequent laser-induced fluorescent detection by means of automatic genetic analyzer («Nanofor 05» produced by company «Synthol»). For typing 18 STR of human Y-chromosome, a multiplex set for amplification «18 STR of human Y-chromosome» produced by OOO «Gordiz» (Moscow) was used [34].

2.4 Prediction of Haplogroups of the Y-Chromosome

For the identified 18 STR-genotypes, prediction of the probability of affiliation to haplogroups of the Y chromosome was performed by Internet predictor - Whit Athey's Haplogroup Predictor [35]. In complicated cases, haplogroup prediction was performed using an additional predictor, NEVGEN.ORG [36], especially for predicting subhaplogroups of haplogroup N.

2.5 Statistical Analysis

Statistical analysis was performed in the software «WinPepi2» [37], «Arlequin 3.5.2.2» [38], «GenAlEx 6.5» [39], and «MS Excel».

2.6 Ethics Statement

All procedures performed in the study with the participation of people comply with the ethical standards of the institutional and/or national committee, as outlined in the Helsinki Declaration in 1964 and its subsequent changes, or comparable ethical standards. The study is approved by the Ethical Committee (protocol № 16 from 26.11.2019) of Novosibirsk Research Institute of Therapy and Preventive Medicine (branch of the Institute of Cytology and Genetics, Siberian Department of RAS).

From each participant included in the study, informed agreement was received. Demographic data obtained from the questionnaire survey and as a result of the study are stored anonymously.

3. Results

3.1 Genetic Differentiation of the Studied Samples by 18 STR Haplotypes

Genetic differentiation among the samples, as measured by the variability of 18 STRs, was estimated using F_{ST} . Two samples of the senior generation in Moscow and Novosibirsk were considered, and a sample of the young generation in Moscow and a sample from Saint Petersburg [30, 31] were used for comparison (Table 1).

Table 1 Estimates of genetic differentiation F_{ST} * for pairwise comparisons of the studied samples from the population of the three megalopolises.

Pairwise comparisons of samples	F_{ST}	$P \pm SE$
Moscow (senior generations) – Moscow (young generation)**	0.0035	0.1081 ± 0.0297
Moscow (senior generations) – Novosibirsk	0.0033	0.4955 ± 0.0412
Moscow (senior generations) – Saint-Petersburg***	0.0050	0.1802 ± 0.0332
Moscow (young generation) – Novosibirsk	0.0084	0.0090 ± 0.0091 ****
Moscow (young generation) – Saint-Petersburg	0.0018	0.5676 ± 0.0526
Novosibirsk – Saint-Petersburg	0.0087	0.0270 ± 0.0139 ****
All samples	0.0019	0.8280 ± 0.0128

Note. * – F_{ST} calculated using the Arlequin 3.0 program (Excoffier and Lischer, 2010); ** – previously published data were used [30]. *** – previously published data were used [31]. **** – statistically significant differences taking into account the error by thresholds of validity: $p < 0.05$ and $p < 0.01$.

3.2 Prediction of Y-Chromosome Haplogroups by 18 STR Haplotypes

Statistically significant differences were found in frequency profiles between the sample from Novosibirsk and the young generation of Moscow. No statistically significant differences were found between the samples from Novosibirsk and from the sample from the senior generations of Moscow by the variability of 18 STR of the Y-chromosome, as well as between the sample from the senior generations and the young generation of Moscow. Analysis of R_{ST} values between the samples from the senior generations from Moscow and from Novosibirsk revealed zero differences. Between the sample from Novosibirsk and the young generation of Moscow, $R_{ST} = 0.005$; between the samples from the senior and young generations of Moscow, $R_{ST} = 0.007$.

For probability detection of haplogroups of the Y chromosome, the obtained STR-haplotypes in the studied samples were analyzed with Internet-predictor. Over 92% of the samples of inhabitants of Novosibirsk are predicted with probability 100%, 6% with probability 99%, 2% with probability over 86%. Therefore, the prediction was performed with high reliability. In Novosibirsk, with a prevailing Russian population, distribution of haplogroups of the Y chromosome, on the whole, corresponds to the Russian gene pool with the most frequent haplogroups R1a, R1b, E1b1b1, N, I1, I2, (excluding haplogroups J1 and J2) — predominantly R1a (Table 2) [40].

Table 2 Frequency of haplogroups of the Y chromosome in the population of three megalopolises.

Haplogroup	Novosibirsk		Moscow (senior generations)**		Mocow (young generation)**		Saint-Petersburg***	
	<i>N</i>	<i>P</i>	<i>N</i>	<i>P</i>	<i>N</i>	<i>P</i>	<i>N</i>	<i>P</i>
C3*	1	0.009	0		7	0.018	2	0.013
E1b1b	5	0.043	3	0.018	17	0.043	15	0.100
G2a*	0		3	0.018	7	0.018	3	0.020
G2c*	0		0		1	0.003	1	0.007
H*	0		0		1	0.003	0	
I1	9	0.078	10	0.059	21	0.053	14	0.093
I2a	0		0		1	0.003	2	0.013
I2a(xI2a1)	12	0.103	22	0.130	35	0.088	9	0.060
I2a1	0		0		2	0.005	0	
I2b(xI2b1)	0		0		1	0.003	0	
I2b1	2	0.017	1	0.006	3	0.008	0	
J1*	2	0.017	2	0.012	16	0.040	3	0.020
J2a*	0		5	0.030	15	0.038	8	0.053
J2a1b*	1	0.009	2	0.012	3	0.008	1	0.007
J2ab1	0		0		1	0.003	0	
J2a1bh*	0		1	0.006	0		0	
J2a4(xbh)*	0		0		3	0.008	0	
J2b*	0		0		7	0.018	3	0.020
J2a1(xJ2a1-bh)*	0		1	0.006	0		0	
L*	0		0		2	0.005	1	0.007
N	15	0.129	25	0.148	29	0.073	17	0.113
R1a	63	0.543	79	0.466	181	0.452	62	0.414
R1b	6	0.052	10	0.059	26	0.065	7	0.047
R2*	0		0		4	0.010	0	
T*	0		3	0.018	5	0.013	0	
O2*	0		1	0.006	0		0	
O3*	0		0		5	0.013	0	
Q*	0		1	0.006	7	0.018	2	0.013
Total	116	1.000	169	1.000	400	1.000	150	1.000

Note. * – “Southern origin” haplogroups; ** – quoted from [30]; *** – quoted from [31].

The studied sample of the male population of Novosibirsk is characterized by the practical absence in the gene pool of modern gene flows from the Caucasus and Central Asia, which are marked in our studies of megalopolises by uncharacteristic haplogroups of the Y chromosome, designated as “Southern origin” haplogroups (C3, G2a, G2c, H, J1, J2, L, O2, O3, Q, R2, and T). These haplogroups were previously identified in Moscow and Saint-Petersburg with a frequency of 21 and 16%, respectively [30, 31]. In Novosibirsk, the frequency of these haplogroups (represented by haplogroups C3, J1, and J2) is only 3.4%. In Novosibirsk, haplogroup C3 (M217) was detected in one

resident with a probability of 99.9%. (according to the NEVGEN.ORG predictor). This haplogroup was not detected in the sample from the senior generation of Moscow. However, in contrast to the sample from Novosibirsk, for the sample from the senior generations of Moscow, “Southern origin” haplogroups were identified with a frequency of 11.2% (haplogroups represented: G2a, J1, J2, T, O2, and Q).

The obtained data on the distribution of Y-chromosome haplogroups agree well with the questionnaire data on the places of origin and ethnic composition of Novosibirsk residents.

3.3 Differentiation of the Studied Samples by “Southern Origin” Haplogroups

Statistically significant differences in the proportion of “Southern origin” haplogroups between sample from Novosibirsk and all the samples studied was observed: from the senior generations of Moscow ($G = 6.1081$, $df = 1$, $p < 0.05$), the young generation of Moscow ($G = 25.1640$, $df = 1$, $p < 0.001$), sample from Saint-Petersburg ($G = 12.0926$, $df = 1$, $p < 0.01$), as well as, between samples from the senior generations and the young generation of Moscow ($G = 8.1565$, $df = 1$, $p < 0.05$). According to questionnaire data, the migration coefficient for the sample from the senior generation in Moscow was 0.485, and the migration distance was 774.43 ± 61.52 km; for the sample from the young generation, it was 0.649 and 1781.59 ± 105.99 km, respectively [29]. A higher proportion of «Southern origin» haplogroups in the sample from the young generation of Moscow compared to the sample from senior generations is in agreement with the ratio of migration coefficients detected for the samples. Analogous data for the sample from Saint-Petersburg, 0.356 and 1741 ± 130.81 km [28], and from Novosibirsk were 0.300 and 893 ± 181 km. In Figure 1, the geographic positions of the three megalopolises (Moscow, Saint-Petersburg, and Novosibirsk) are shown, along with the migration coefficients and the proportions of the “Southern origin” haplogroups of the Y-chromosome for the studied samples.

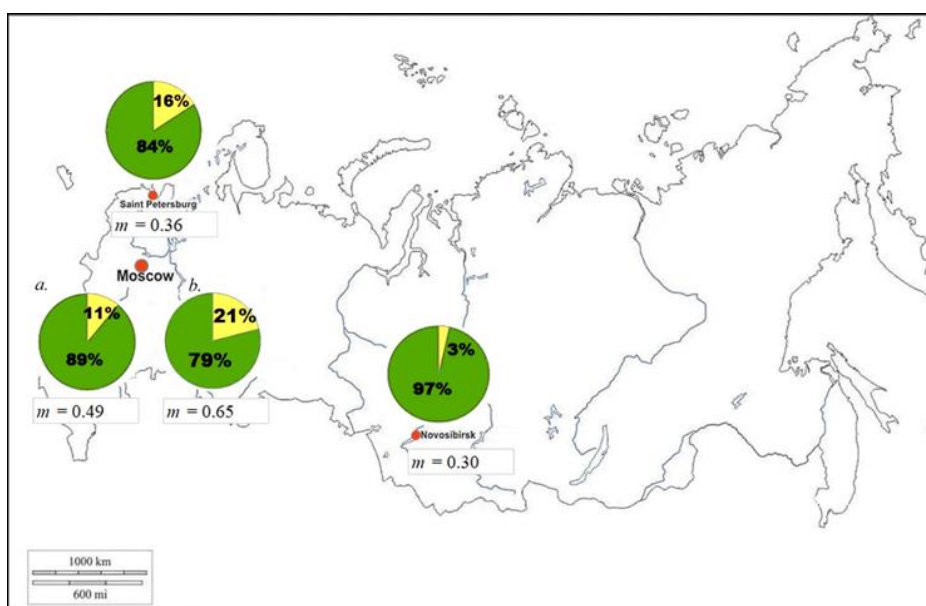


Figure 1 Geographic position of the three megalopolises (Moscow, Saint-Petersburg, and Novosibirsk), migration coefficients (m), proportion of the “Southern origin” haplogroups of the Y chromosome determined for the samples from the population of megalopolises is indicated in yellow. *a.* – senior generations, *b.* – young generation.

In Figure 2, the more detailed distribution of frequency features for Y-chromosome large haplogroups is shown. The statistically significant increase in the considered component for the sample from the young generation (21%) compared to the sample from the senior generations (11.2%) in Moscow is due to the presence of migrants from the North Caucasus, Transcaucasia, and Central Asia, confirmed by the pedigrees of the studied Moscow residents [29, 30]. The noted peculiarities of the senior generations of Moscow (megapolis on the European territory of the Russian Federation) compared to the sample of Novosibirsk (Siberian megapolis) demonstrate the presence of higher proportion “Southern origin” haplogroups, which is in good agreement with the questionnaire data showing the presence of ancestors of Jewish or Armenian nationality, natives of Dagestan, or of unknown nationality in the carriers of these haplogroups.

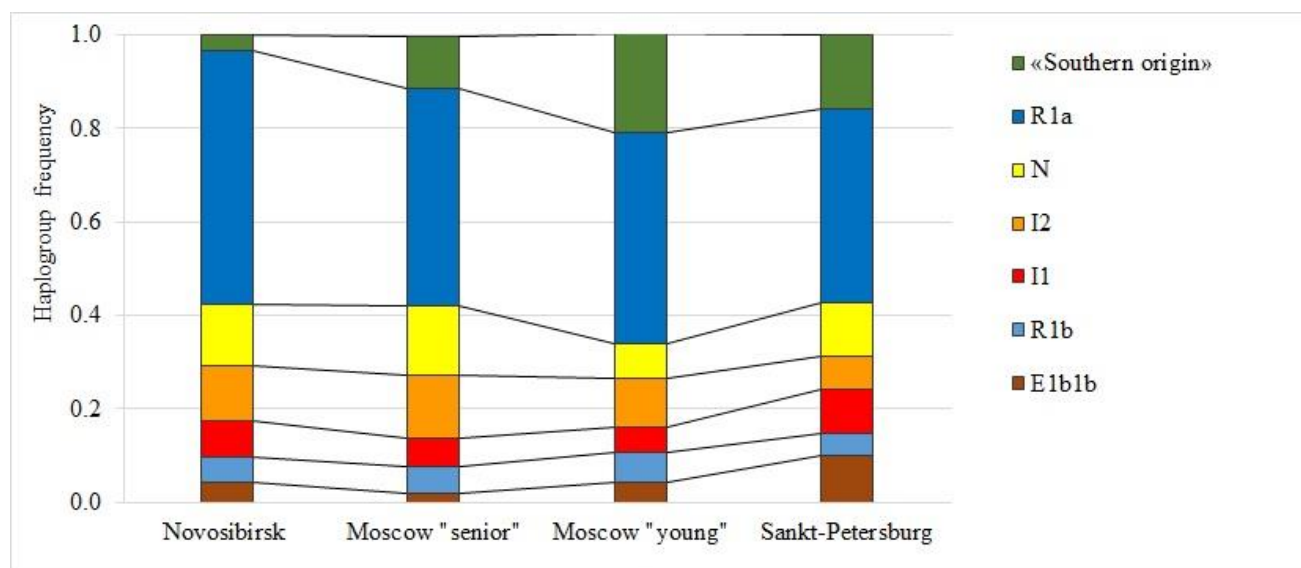


Figure 2 Frequencies of large haplogroups of the Y chromosome in the studied samples from the population of three megalopolises: Novosibirsk, Moscow (senior and young generations), and Saint-Petersburg.

3.4 Geographic Peculiarities of Y-Chromosome Haplogroups Distribution

In Moscow and Novosibirsk, the ratios of haplogroups I1 and I2 are 0.48 and 0.64, respectively, and in Saint Petersburg - 1.3. Haplogroup E1b1b has the highest frequency in Saint-Petersburg (10.0%) and a much lower frequency in the total sample of Moscow (3.5%) [30, 31] and Novosibirsk (4.3%).

In Novosibirsk, the N1a1 (M46) subhaplogroup of haplogroup N (12.9%) was the most frequent, with a frequency of 10.3%, compared to N1a2 (CTS6380) (2.6%). In the sample from the senior generations of Moscow, only the N1a1 subhaplogroup was present [40].

4. Discussion

The results obtained allow us to state that in the senior generations, the gene pools of the male population of megalopolises, both in the European part of the Russian Federation (Moscow) and in Siberia (Novosibirsk), show similarity in the spectra of Y-chromosome STR-haplotypes, but demonstrate differences by the presence of “Southern origin” haplogroups. High informativity and

polymorphism of microsatellites, simultaneous study of variability of 18 STR markers contributed to obtaining a statistically significant estimate of F_{ST} – the level of genetic differentiation for the sample from Novosibirsk, the young Moscow sample, and the sample from Saint-Petersburg. The absence of differences in STR haplotypes between the senior generations of Moscow and the Novosibirsk sample was established for both F_{ST} and R_{ST} estimates.

The difference in the proportions of «Southern origin» haplogroups between the compared samples is consistent with estimates of the migration coefficient. Compared to the sample from the young generation, the sample from the senior generations of Moscow revealed a lower representation of “Southern origin” haplogroups, which penetrated megalopolises with a prevailing Russian population by the flows of migrants from other ethnic groups (Armenians, Jews, Dagestanis, and of unknown ethnicity) before 1991 [30].

For Saint-Petersburg, we compared the frequency profiles of Y-chromosome haplogroups in samples of men with Russian grandfathers and fathers, and only with Russian fathers, based on questionnaire data: an increase in the presence of “Southern origin” haplogroups from 4.2% to 10% was observed, respectively. In the total sample, “Southern origin” Y-chromosome haplogroups are present at a frequency of 16%, demonstrating their accumulation due to migration flows [31].

Earlier, the haplogroups of the Y-chromosome of “Southern origin” under consideration were identified as introduced with migrant flows from the North Caucasus, Transcaucasia, and Central Asia into the gene pools of the male populations of the Moscow and Saint-Petersburg megalopolises in the European territory of the Russian Federation [30, 31]. Our study of the frequency profile of Y-chromosome haplogroups in the sample from Novosibirsk confirmed that the accumulation of “Southern origin” Y-chromosome haplogroups occurs due to gene flows brought to the megapolis by modern waves of migration from the Southern regions of the country to the megalopolises under consideration, where these haplogroups originated and occur with high frequency. With rare exceptions, the haplogroups under consideration are not characteristic of the Russian population [40], which constitutes the overwhelming majority of the residents of the megalopolises under consideration.

In the Novosibirsk population, migrants from the North Caucasus, Transcaucasia, and Central Asia are not represented in the studied sample, as indicated by questionnaire data, consistent with the low representation of the haplogroups under consideration. For the Novosibirsk sample, the lowest migration coefficient estimate was obtained (0.300), consistent with the lowest frequency of “Southern origin” haplogroups among the samples studied.

The age features of the Novosibirsk sample suggest the practical absence of modern migration waves from the North Caucasus, Transcaucasia, and Central Asia, a conclusion also supported by the questionnaire data and consistent with the absence of haplogroups penetrating the gene pool of megalopolises. On the contrary, the presence of “Southern origin” haplogroups in the sample from the senior generations of Moscow is due to earlier migrations to Moscow (before 1991) as the capital, which is more attractive to migrants from various regions of the country and neighboring regions. The sample from the young generation of Moscow shows the highest migration coefficient (0.649), consistent with the highest frequency of “Southern origin” haplogroups (21%).

In the studied samples from the megalopolis population, the frequency profiles of Y-chromosome haplogroups show differences associated with geographical location, including regional peculiarities in distribution and differences in migration parameters.

Subhaplogroups of haplogroup N N1a1 and N1a2 are spread in the population of the megalopolises studied, consistent with their ethno-geographic peculiarities. Haplogroup N1a1, in addition to being found among indigenous peoples of Siberia and Finno-Ugrians, is also found in Russian populations across various regions, with frequencies ranging from 13 to 27%. Haplogroup N1a2 is found at high frequency among Samoyeds, Finno-Ugrians of the Urals and Western Siberia, and other indigenous peoples of Siberia, and is virtually absent from Russian populations, except for the northern ones [40]. The revealed peculiarities of the distribution of N1a1 and N1a2 subhaplogroups in megalopolises were distinguished by the geographical location of the considered megalopolises and the geographical distribution of haplogroups.

Haplogroup C3 is distributed at high frequency among the peoples of Central Asia and among some indigenous peoples of Siberia [41].

The results of our study confirm the necessity of developing specific reference genetic databases for three megalopolises by Y-chromosome DNA markers based on the established features of haplogroup distribution and haplotype frequency profiles by 18 STRs of the Y-chromosome [42], determined by the geographical location of megalopolises and geographical peculiarities of Y-chromosome haplogroup distribution. Essential factors in the dynamics of gene pools across generations are migration parameters.

This conclusion aligns with the previously demonstrated significant ethno-territorial subdivisions in megalopolises [43-45], which underscores the probable need for distinct genetic databases (including reference databases) for each region within these megalopolises.

The study of genetic-demographic processes within each megalopolis allows for predicting the dynamics of the frequencies of specific Y-chromosome haplogroups in the megalopolis's gene pool over generations, particularly driven by migration [21].

4.1 Limitations of the Samples

The limitation of the sample for Novosibirsk in the study must be considered. However, the residents from all regions of Novosibirsk are included in the sample. The sample consists of residents of a particular megalopolis of a specific age. The high informativity and polymorphism of microsatellites, along with the simultaneous analysis of 18 STR markers, contributed to obtaining reliable results. The sample limitation may have led to the underrepresentation of rare haplogroups in the megalopolis's gene pool, with frequencies below 1%, which might not significantly affect our main conclusion based on the summarized frequency of "Southern origin" haplogroups. Additionally, molecular data are consistent with questionnaire data analysis.

For further studies to obtain more reliable and systematic results, it will be relevant to consider larger samples that are representative of megalopolises' gene pools, and to investigate alternative generations from each megalopolis with comparable age characteristics.

5. Conclusions

Detection of the frequencies of Y-chromosome markers in concrete generations of the megalopolis population is important for developing population reference databases. When creating population reference databases of DNA data on the Y-chromosome in megalopolises, the identified parameters of migration and changes in the gene pool under the influence of migration processes should be taken into account. For developing population reference databases for DNA-identification

in megalopolises, it is necessary to obtain genetic-demographic questionnaire data, simultaneously with molecular genetic analysis. For obtaining representative sample for the population of megalopolis, special approaches might be used accumulating data on territorial and ethnic subdivision [46] and formation of the groups of alternative age.

Acknowledgments

Authors are grateful to the citizens of Novosibirsk, who took part in the questionnaire survey and who support biological samples for the study.

Author Contributions

Irina G. Udina: Conceptualization, investigation, methodology, writing – original draft, formal analysis, writing – review and editing. Alesya S. Gracheva: Investigation, software, methodology, writing – review and editing. Marina A. Gubina: Investigation, methodology, writing – editing. All authors have read and approved the published version of the manuscript.

Funding

The study is performed in the frames of the theme of State Task of IOGen RAS.

Competing Interests

The authors declare no conflict of interest.

References

1. Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* 2002; 12: 339-348.
2. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 2008; 18: 830-838.
3. Doğan S, Ašić A, Doğan G, Besic L, Marjanovic D. Y-chromosome haplogroups in the Bosnian-Herzegovinian population based on 23 Y-STR loci. *Hum Biol.* 2016; 88: 201-209.
4. Dogan S, Babic N, Gurkan C, Goksu A, Marjanovic D, Hadziavdic V. Y-chromosomal haplogroup distribution in the Tuzla Canton of Bosnia and Herzegovina: A concordance study using four different in silico assignment algorithms based on Y-STR data. *Homo.* 2016; 67: 471-483.
5. Doğan S, Doğan G, Ašić A, Bešić L, Klimenta B, Hukić M, et al. Prediction of the Y-Chromosome Haplogroups within a recently settled Turkish Population in Sarajevo, Bosnia & Herzegovina. *Coll Antropol.* 2016; 40: 1-7.
6. Lazim H, Almohammed EK, Hadi S, Smith J. Population genetic diversity in an Iraqi population and gene flow across the Arabian Peninsula. *Sci Rep.* 2020; 10: 15289.
7. Babić Jordamović N, Kojović T, Dogan S, Bešić L, Salihefendić L, et al. Haplogroup prediction using Y-chromosomal short tandem repeats in the general population of Bosnia and Herzegovina. *Front Genet.* 2021; 12: 671467.

8. Primorac D, Škaro V, Projić P, Missoni S, Zanki IH, Merkaš S, et al. Croatian genetic heritage: An updated Y-chromosome story. *Croat Med J*. 2022; 63: 273-286.
9. Efetov KA, Kharlamov SG, Efremov IA. Polymorphism of microsatellite loci of Y-chromosome in Crimean Karaites and Crimean Tatars (in Russian). *Patogenez*. 2023; 21: 62-74.
10. Athey TW. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. *J Genet Geneal*. 2006; 2: 34-39.
11. Emmerova B, Ehler E, Comas D, Votrubova J, Vanek D. Comparison of Y-chromosomal haplogroup predictors. *Forensic Sci Int Genet Suppl Ser*. 2017; 6: e145-e147.
12. Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF. High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol*. 2004; 21: 164-175.
13. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, et al. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: Inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet*. 2004; 74: 1023-1034.
14. Grugni V, Battaglia V, Hooshier Kashani B, Parolo S, Al-Zahery N, Achilli A, et al. Ancient migratory events in the Middle East: New clues from the Y-chromosome variation of modern Iranians. *PloS One*. 2012; 7: e41252.
15. Willuweit S, Roewer L. The new Y chromosome haplotype reference database. *Forensic Sci Int Genet*. 2015; 15: 43-48.
16. de Knijff P. On the forensic use of Y-chromosome polymorphisms. *Genes*. 2022; 13: 898.
17. Hallast P, Ebert P, Loftus M, Yilmaz F, Audano PA, Logsdon GA, et al. Assembly of 43 human Y chromosomes reveals extensive complexity and variation. *Nature*. 2023; 621: 355-364.
18. Kilchevsky AV, Yankovsky NK. Developing the innovative gene geographical and genomic technologies for identification and revealing the personal features by studying the gene pools of the regional populations. *Russ J Genet*. 2021; 57: 1361-1369.
19. Butler JM. Recent advances in forensic biology and forensic DNA typing: INTERPOL review 2019–2022. *Forensic Sci Int Synergy*. 2023; 6: 100311.
20. Tsybovskii IS, Veremeichik VM, Kotova SA, Kritskaya SV, Evmenenko SA, Udina IG. Developing forensic reference database by 18 autosomal STR for DNA identification in Republic of Belarus. *Russ J Genet*. 2017; 53: 275-284.
21. Hammer MF, Chamberlain VF, Kearney VF, Stover D, Zhang G, Karafet T, et al. Population structure of Y chromosome SNP haplogroups in the United States and forensic implications for constructing Y chromosome STR databases. *Forensic sci int*. 2006; 164: 45-55.
22. Parson W, Dür A. EMPOP—A forensic mtDNA database. *Forensic Sci Int Genet*. 2007; 1: 88-92.
23. Udina IG, Gracheva AS, Kurbatova OL. Scientific and applied significance of the analysis of Y-chromosome variation for criminalistics, ethnodemographic studies and historical reconstructions. *Soc Area*. 2023; 9. doi: 10.15838/sa.2023.1.37.7.
24. Caputo M, Sala A, Corach D. Demand for larger Y-STR reference databases in ethnic melting-pot countries: Argentina as a test case. *Int J Leg Med*. 2019; 133: 1309-1320.
25. Moutsouri I, Keravnou A, Manoli P, Bertoncini S, Michailidou K, Christofi V, et al. Comparative Y-chromosome analysis among Cypriots in the context of historical events and migrations. *PLoS One*. 2021; 16: e0255140.

26. Kurbatova OL, Pobedonostseva EY, Veremeichyk VM, Prudnikova AS, Atramentova LA, Tsybovsky IS, et al. Genetic demography of populations of three megalopolises in relation to the problem of creating genetic databases. *Russ J Genet.* 2013; 49: 448-456.
27. Kurbatova OL, Yankovsky NK. Migration as the main factor of the Russia's urban population dynamics. *Russ J Genet.* 2016; 52: 726-745.
28. Kurbatova OL, Udina IG, Gracheva AS, Pobedonostseva EY, Borinskaya SA. Genetic demography of the population of St. Petersburg: Migration processes. *Russ J Genet.* 2019; 55: 1119-1129.
29. Kurbatova OL, Gracheva AS, Pobedonostseva EY, Udina IG. Genetic demography of the population of Moscow: Migration processes. *Russ J Genet.* 2021; 57: 1443-1453.
30. Udina IG, Gracheva AS, Kurbatova OL. Frequencies of Y-chromosome haplogroups and migration processes in three generations of Moscow residents. *Russ J Genet.* 2022; 58: 1365-1372.
31. Udina IG, Gracheva AS, Borinskaya SA, Kurbatova OL. Distribution peculiarities of Y-chromosome haplogroups in the population of St. Petersburg in connection with the problem of creation of reference databases. *Russ J Genet.* 2023; 59: 1216-1221.
32. Kurbatova OL, Udina IG, Gracheva AS. Genetic and demographic population parameters of Novosibirsk (in Russian). *Genetika.* 2018; 54: S74-S84.
33. Maniatis T, Fritsch EF, Sambrook J. Molecular cloning. A laboratory manual. Cold Spring Harbor Laboratory; 1982.
34. COrDIS YSTR. Set of reagents for multiplex analysis of 18 STR-markers of human Y-chromosome (in Russian) [Internet]. COrDIS YSTR; 2020 [cited date 2024 November 25]. Available from: <https://gordiz.ru/wp-content/uploads/2020/06/instrukciya-cordis-ystr-231113.pdf>.
35. Haplogroup Predictor. Homepage [Internet]. Haplogroup Predictor; [cited date September 6]. Available from: <http://www.hprg.com/hapest5/>.
36. Y-DNA Haplogroup Predictor - NEVGEN.ORG. Homepage [Internet]. Y-DNA Haplogroup Predictor - NEVGEN.ORG.; [cited date 2024 September 5]. Available from: <https://nevgen.org/>.
37. Abramson JH. WINPEPI updated: Computer programs for epidemiologists, and their teaching potential. *Epidemiol Perspect Innov.* 2011; 8: 1.
38. Excoffier L, Lischer HE. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour.* 2010; 10: 564-567.
39. Peakall R, Smouse PE. GenAEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—An update. *Bioinformatics.* 2012; 28: 2537-2539.
40. Balanovskaya EV, Balanovskii OP. Russian gene pool on the Russian Plain. Moscow, Russian: Luch; 2007.
41. Balaganskaya OA, Damba LD, Zhabagin MK, Agdzhoyan AT, Yusupov YM, Sabitov ZM, et al. Mongolian trace in gene pool of populations along the steppe of Eurasia. *Modern Probl Sci Educ.* 2016; 4: 211.
42. Udina IG, Gracheva AS, Kurbatova OL. Developing forensic databases for megalopolises considering migration (Y-chromosome haplogroups). *Proceedings of the 13th ISABS Conference on Applied Genetics and Mayo Clinic Lectures in Translational Medicine; 2024 June 17-22; Split, Croatia; Presentation number FG10; ABS-146-ISABS-2024.*
43. Gracheva AS, Pobedonostseva EY, Udina IG, Kurbatova OL. Territorial subdivision of the megalopolis population by the ethnic trait in relation to the problem of creating genetic databases: Moscow. *Russ J Genet.* 2020; 56: 1520-1529.

44. Gracheva AS, Pobedonostseva EY, Udina IG, Kurbatova OL. Territorial subdivision of the megalopolis population by the ethnic trait in relation to the problem of creating genetic databases. St. Petersburg. Russ J Genet. 2019; 55: 1536-1544.
45. Gracheva AS, Pobedonostseva EY, Udina IG, Kurbatova OL. Territorial subdivision of the megalopolis population by the ethnic trait in relation to the problem of creating genetic databases. Novosibirsk (in Russian). Genetika. 2018; 54: S85-S90.
46. Kish L. Sample surveys versus experiments, controlled observations, censuses, registers, and local studies. Aust J Stat. 1985; 27: 111-122.