Pharmacogenomics and Personalized Medicine

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Variation of Genomic Sites Associated with Severe Covid-19 Across Populations: Global and National Patterns

Oleg Balanovsky[1–3]
Valeria Petrushenko[1,4]
Karin Mirzaev[2,5]
Sherzod Abdullaev[5]
Igor Gorin[1,4]
Denis Chernevskiy[2]
Anastasiya Agdzhoyan[1,2]
Elena Balanovska[2,3]
Alexander Kryukov[5]
Ilyas Temirbulatov[2,5]
Dmitriy Sychev[5]

[1]Laboratory of Genome Geography, Vavilov Institute of General Genetics, Moscow, Russia; [2]Laboratory of Human Population Genetics, Research Centre for Medical Genetics, Moscow, Russia; [3]Biobank of North Eurasia, Moscow, Russia; [4]Department of Bioinformatics Moscow Institute of Physics and Technology, Moscow, Russia; [5]Department of Clinical Pharmacology and Therapeutics, Russian Medical Academy of Continuous Professional Education, Moscow, Russia

Background: Information about the distribution of clinically significant genetic markers in different populations may be helpful in elaborating personalized approaches to the clinical management of COVID-19 in the absence of consensus guidelines.

Aim: Analyze frequencies and distribution patterns of two markers associated with severe COVID-19 (*rs11385942* and *rs657152*) and look for potential correlations between these markers and deaths from COVID-19 among populations in Russia and across the world.

Methods: We genotyped 1883 samples from 91 ethnic groups pooled into 28 populations representing Russia and its neighbor states. We also compiled a dataset on 32 populations from other regions using genotypes extracted or imputed from the available databases. Geographic maps showing the frequency distribution of the analyzed markers were constructed using the obtained data.

Results: The cartographic analysis revealed that *rs11385942* distribution follows the West Eurasian pattern: the marker is frequent among the populations of Europe, West Asia and South Asia but rare or absent in all other parts of the globe. Notably, the transition from high to low *rs11385942* frequencies across Eurasia is not abrupt but follows the clinal variation pattern instead. The distribution of *rs657152* is more homogeneous. The analysis of correlations between the frequencies of the studied markers and the epidemiological characteristics of COVID-19 in a population revealed that higher frequencies of both risk alleles correlated positively with mortality from this disease. For *rs657152*, the correlation was especially strong ($r = 0.59$, $p = 0.02$). These reasonable correlations were observed for the "Russian" dataset only: no such correlations were established for the "world" dataset. This could be attributed to the differences in methodology used to collect COVID-19 statistics in different countries.

Conclusion: Our findings suggest that genetic differences between populations make a small yet tangible contribution to the heterogeneity of the pandemic worldwide.

Keywords: severe COVID-19, genetic markers, AB0, rs11385942, rs657152, gene geography

Correspondence: Elena Balanovska
Laboratory of Human Population Genetics, Research Centre for Medical Genetics, Moscow, Russia
Tel +7 499 612-81-79
Email balanovska@mail.ru

Dmitriy Sychev
Department of Clinical Pharmacology and Therapeutics, Russian Medical Academy of Continuous Professional Education, Moscow, Russia
Tel +7 495 680-05-99
Email dmitry.alex.sychev@gmail.com

Introduction

The novel coronavirus known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first identified in China in December 2019. The virus was rapidly spreading around the world, and three months later the World Health Organization declared a pandemic.[1] The disease caused by SARS-CoV-2 is referred to as coronavirus disease 2019 (COVID-19). The course of the disease varies from asymptomatic infection to death from respiratory failure. Epidemiological data show that the morbidity and mortality of COVID-19 vary greatly not only between individuals within a population

Received: 26 May 2021
Accepted: 4 October 2021
Published: 4 November 2021

but also between countries. At the time of writing, the daily rate of new COVID-19 cases in Northern Europe (Denmark, Estonia, Finland, Norway) was relatively low, whereas Southern Europe (Italy, Spain, France) was suffering from higher morbidity and higher mortality.[2] Such variation may be due to sociodemographic, environmental (see[3–5] for correlations with vitamin D levels) and genetic factors. A genome-wide association study involving 1610 patients and 2205 control participants from Italy and Spain has identified two key genomic sites associated with severe respiratory failure due to COVID-19.[6] One of them is *rs11385942* (G>GA) at locus 3p21.31, which comprises a cluster of six genes potentially predisposing to severe COVID-19. Notably, the association was demonstrated for the entire 50 kb-long haplotype segment. Such high LD is uncommon for this segment of chromosome 3 and can be explained by the evolutionarily recent introgression of this haplotype into the human gene pool from Neanderthals.[7] The other site (*rs657152*), which demonstrated a less pronounced association with severe COVID-19, is located at the ABO blood group locus 9q34.2. According to Zeberg and Pääbo,[7] the frequencies of these SNP varied among the populations included in the 1000 Genomes project and the highest frequency occurs in South Asia (Bangladesh). Zeberg and Pääbo also cited the report where individuals of Bangladeshi origin in the UK have an about two times higher risk of dying from COVID-19 than the general population.

So, we decided to investigate how these two SNPs associated with severe COVID-19 are distributed in different populations across the world focusing on the least studied populations of North Eurasia.

The datasets we used consisted of 1883 samples representing populations of North Eurasia and the published data on 3088 samples from other world regions. Our goal was to unravel the patterns of geographic variation of these polymorphisms and investigate possible correlations between their frequencies and the number of COVID-19 deaths and recoveries in the studied populations. Information about the distribution of clinically significant genetic markers in different populations may help to elaborate personalized approaches to the treatment of COVID-19 in the absence of consensus guidelines.

The primary goal of this study was to determine the frequencies of *rs11385942* and *rs657152* in different populations of Russia, which is a highly ethnically heterogenous country, and its closest neighbors. The second goal was to look for any possible associations between the population frequencies of *rs11385942* and *rs657152* and the number of COVID-19 cases, recoveries and deaths.

Materials and Methods

Dataset 1: Populations of Russia and Its Neighbor States

Samples representing the populations of Russia and its closest neighbors in North Eurasia (post-Soviet states and Mongolia) were provided by the Biobank of North Eurasia.[8] The study was approved by the Ethics Committee of the Research Centre for Medical Genetics; informed consent was obtained from every donor. The analyzed samples represented a vast territory encompassing much of Russia, its neighbor states and most ethnic groups inhabiting this area (Figure 1). Similar to the sampling strategy of the 1000 Genomes Project, our study focused on indigenous populations. For example, East Siberian populations were represented by Yakuts and Evenks but not ethnic Russians who had moved to Siberia in the past 1–3 centuries.

We genotyped a total of 1883 samples representing 91 populations from Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Lithuania, Mongolia, Russia, Tajikistan, Ukraine, and Uzbekistan. We focused on two markers associated with severe COVID-19: *rs11385942* and *rs657152*.[6] The samples were genotyped for *rs657152* and 250 SNPs in the vicinity of *rs11385942* (chr3: 45700000–45900000, GRCh37) using an Infinium Omni5Exome-4 v1.3 BeadChip Kit (Illumina, USA), which covers over 4.5M SNPs.

To estimate the frequencies of COVID-associated markers, we needed a sample size of 100 chromosomes per population, but most populations included in the study were represented by only a few samples. Therefore, samples of geographically and genetically close populations were pooled. The genetic similarity between those populations was determined by means of genome-wide genotyping for 4.5M SNPs included in the Illumina array. The pooling procedure consisted of the following steps: i) input filtering and pruning with PLINK (geno 0.05, maf 0.01, mind 0.1,[9] indep-pairwise 1500 150 0.2), ii) running PCA with smartpca,[10,11] and iii) identifying groups of populations that were internally genetically homogenous but showed pronounced intergroup variation. Some populations were excluded from the analysis because of their small sample size and genetic features that made pooling impossible. The pooling procedure is described in greater

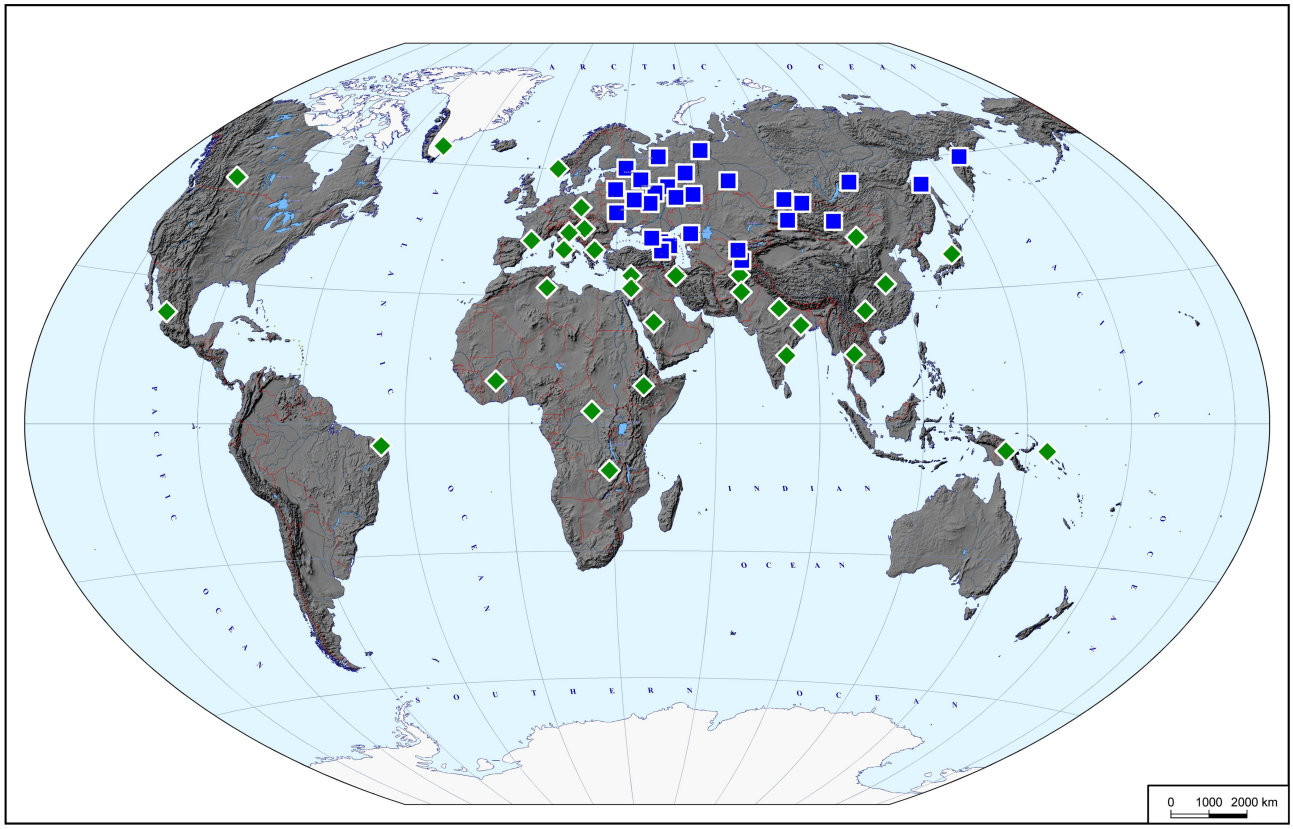

Figure 1 The studied populations. Blue squares show locations of population samples genotyped specifically for this study (the Russian dataset). Green diamonds show locations of population samples described in other sources and reanalyzed in the course of this study (the world dataset).

detail.[12] The final dataset consisted of 1785 samples from 28 pooled populations; the average sample size was 128 chromosomes (64 individuals). These populations, below referred to as the "Russian dataset", are listed in Table S1. Their geography is shown on the map (Figure 1).

Dataset 2: Populations from Other World Regions

In addition to the dataset representing Russian populations, we compiled a dataset of populations from other world regions. We used data from 16 publications in which indigenous populations were studied using genome-wide genotyping arrays by Illumina (Illumina700k, Illumina730k, Illumina660k, Illumina650k, Illumina610k, Illumina550k, and Illumina1M).[13–28] The merged dataset consisted of 3336 samples genotyped for *rs657152* and 66 SNPs in the vicinity of *rs11385942*.

With this dataset, some samples had to be pooled, too, to achieve larger sample sizes. Samples representing the same country were pooled, in most cases. In some cases, samples representing bordering states, like Spain and France, were also pooled. Two large countries, China and India, were represented by 3 pooled populations each. Samples from Russia and its bordering states were excluded from this dataset to avoid double counting. The final dataset included 32 populations with an average sample size of 104 chromosomes. These populations, referred to as the "world dataset", are listed in Table S1. Their geography is shown on the map (Figure 1).

Imputation

One of the two analyzed COVID-19-associated markers, *rs657152*, was genotyped directly in both datasets. The frequencies of this marker were calculated in PLINK and are shown in Table S1. The second marker, *rs11385942*, is not included in any of the currently available genome-wide arrays. During the genome-wide association study that inspired our work, 712,189 SNPs had been genotyped and 170 million SNPs had been imputed.[6] The association with severe COVID-19 was the strongest for one of the imputed SNPs, *rs113859420*. The study reported a ~ 50 kb-long haplotype segment where SNPs were in high LD; *rs113859420* tags this segment as effectively as any other

SNP in the vicinity.[7] The SNP array used to genotype our "Russian" dataset was far denser (4.5M SNPs) than that used in and comprised 250 SNPs in the vicinity of *rs113859420*.[6] Ungenotyped markers were imputed with Beagle; the number of iterations was set to 200.[29] The 1000 Genomes Project dataset was used as a reference.

Since Illumina arrays used to genotype samples included in our "world" dataset were less dense (66 SNPs in the vicinity of *rs113859420*), the quality of imputation had to be assessed first. For that, we compared genotypes of the same samples generated by the 250 SNP and 66 SNP arrays; 1871 (99.4%) of 1883 genotypes coincided between the two sets, confirming that imputation of *rs113859420* genotypes from lower-density Illumina arrays was quite accurate. Those genotypes were used to count *rs113859420* allele frequencies in the populations from the "world" dataset (Table S1).

Cartographic Analysis

Maps showing geographic distribution of the two analyzed COVID-associated markers were created in GeneGeo.[30,31] The maps of North Eurasia (Russia and its closest neighbors) were created using Shepard's method. The weight function was set to 3; the radius of influence was set to 1500 km. For other world regions, frequency distribution maps were constructed using the same method, with the radius of influence extended to 2500 km to extrapolate frequency variation to populations not included in the dataset and the weight function set to 2 for better smoothing so as to see trends.[32]

Statistical Analysis

For the correlation analysis, we needed information about the number of COVID-19 cases per one million population, the number of recoveries per one million population, the number of deaths from COVID-19 per one million population, and the number of deaths per all confirmed COVID-19 cases (this information was last updated on September 18, 2020). COVID-19 statistics were available per region and did not discriminate between indigenous and non-indigenous populations. So, we identified 16 regions in Russia and its neighbor states where indigenous populations constituted a majority (85% on average, according to the latest census) and ran the correlation analysis on those 16 groups (Table S3). To analyze the distribution of COVID-19-associated markers at the global level, we used those 16 groups representing Russia and its neighbor states and all the populations included in the "world" dataset, except Native Americans and Greenlanders (Table S4). Differences in allele frequencies between the groups were tested using Fisher's exact test and GraphPad InStat (GraphPad Software, San Diego, CA, USA); the Bonferroni-adjusted significance threshold was set to 0.05. Correlations between the frequencies of the studied SNPs and epidemiological parameters per 1 million population were tested using Pearson's correlation coefficient calculated in StatSoft Statistica (Dell Statistica, Tulsa, OK, USA); the Bonferroni-adjusted significance threshold was also set to 0.05.

Results

Variation of rs11385942 Frequencies at the Global Level

Rs11385942 is strongly associated with severe COVID-19.[6] We analyzed the geographic distribution of its variants in the world's populations using two datasets: 1883 genotyped samples representing Russia and its neighbor states and the published data on 3088 samples from other regions. The "world" dataset contained information about *rs11385942_GA* frequencies in 60 populations (Table S1). Figure 2 shows the geographic distribution of this marker across the globe. Its highest frequency (20–30%) is observed in South Asia, followed by West Asia and Europe (5–15%). This SNP is found in the populations of North Africa but is rare or undetected in East Asia, North Asia (Siberia), native American populations, and sub-Saharan Africa. According to the yet scarce data on the distribution of *rs11385942_GA* in Southeast Asia and Papua New Guinea, the frequency of this variant in these two regions is elevated. Similar to many other genetic markers, including mitochondrial DNA and Y-chromosomal markers, the geographic distribution of this marker follows the well-known "West Eurasian" pattern.[33,34] South Asian populations have more pronounced genetic affinity with West Eurasians than with East Eurasians. Interestingly, the geographic distribution of *rs11385942* is characterized by its higher frequencies in South Asian populations and lower frequencies in West Asian/European populations of West Eurasia.

To compensate for the inevitably schematic character of this description, the analysis of *rs11385942* variation at the global level was complemented by the detailed analysis of its geographic distribution in Russia.

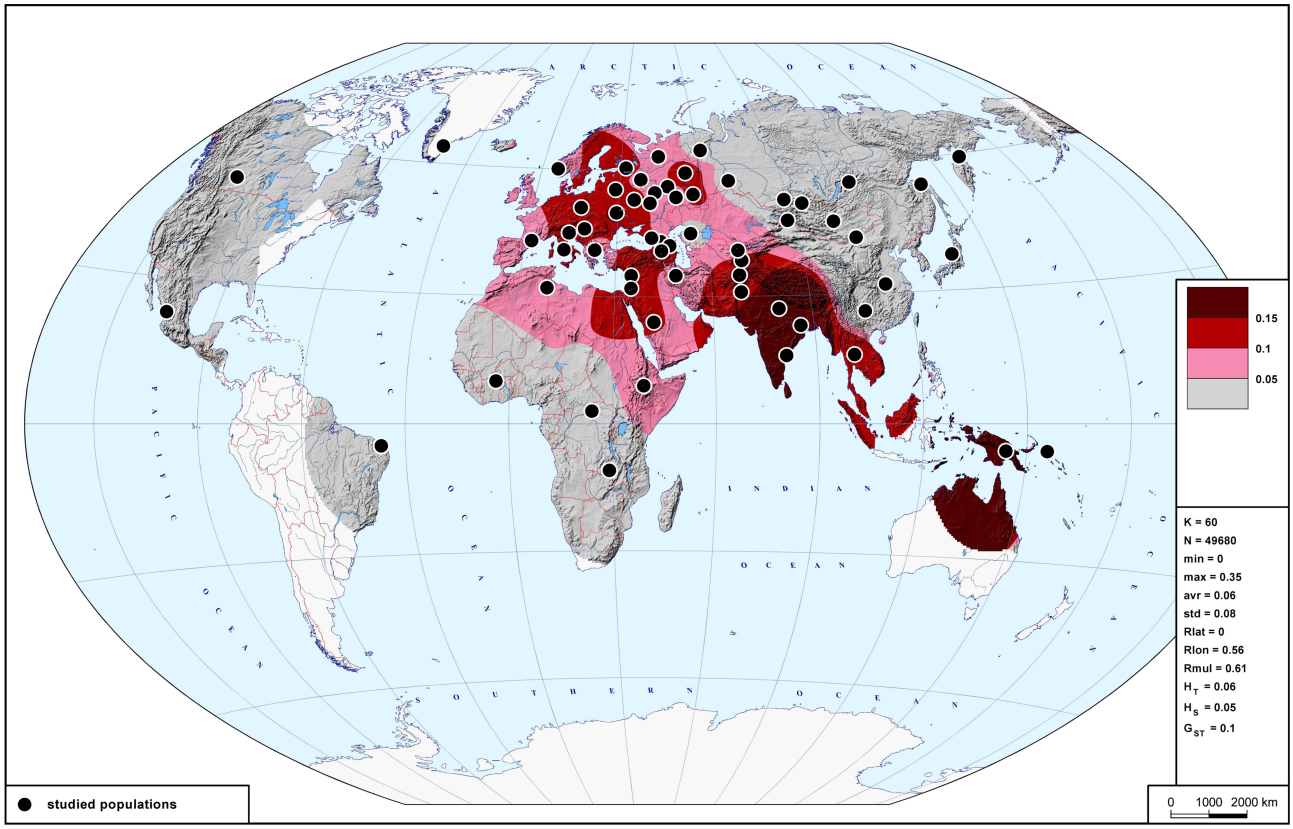

Figure 2 Global variation of rs11385942_GA frequencies. Four colors mark areas of 4 frequency ranges of this risk allele. The black points represent the populations analyzed. Abbreviations in the statistical legend indicate the following.

Abbreviations: K, number of the populations studied; N, the number of the map grid nodes (calculated according to the data on the studied populations); min, the minimal frequency on the map; max, the maximum frequency on the map; avr, the average frequency on the map; std, the standard deviation; Rlat, the coefficient of partial correlation between latitude and frequency on the map; Rlon, the coefficient of partial correlation between longitude and frequency on the map; Rmul, the coefficient of multiple correlation of the frequency with both latitude and longitude on the map.

Variation of rs11385942 Frequencies Across Russia

Our extensive dataset of samples collected in Russia and its neighbor states allowed us to analyze the geographic distribution of *rs11385942* in these countries in greater detail. The frequencies of this SNP are listed in Table S1 and visualized on the map (Figure 3). The map shows that there is no abrupt transition from high *rs11385942* frequencies in Europe to its zero presence in North Asia. The distribution of this SNP follows a clinal variation pattern, ie, its frequency gradually decreases across 7000 kilometers, from 13% to 16% in Ukraine, Belarus and westernmost Russian regions to zero in the Kamchatka and Chukotka peninsulas at the Pacific coast. The average *rs11385942* frequency in the populations inhabiting European Russia is 11% vs 3% in Siberia (Table S1). In Central Asia regions, *rs11385942* frequencies are generally low (1–4%), Tajikistan being a notable exception (14%, which is significantly different from *rs11385942* frequencies in its neighbor states). This is explained by the fact that the population of Tajikistan is geographically and genetically close to South Asian populations, where the frequency of *rs11385942* reaches its peak.

The frequency of *rs11385942* in the southwestern regions of Russia is very similar to that in the bordering Belarus and Ukraine, the difference being statistically insignificant (Tables S1 and S2). In the Caucasus, the lowest frequency of the risk *GA* allele (6%) is observed in the central regions (Chechnya, Ingushetia, North Ossetia), increasing to the east and to the west (≈ 10%), and is particularly high in the South Caucasus (14%).

Variation of rs657152 Frequencies at the Global Level

Figure 4 shows a frequency distribution map for *rs657152*. In comparison with *rs11385942*, the distribution of this marker is more homogeneous. It is quite frequent in almost

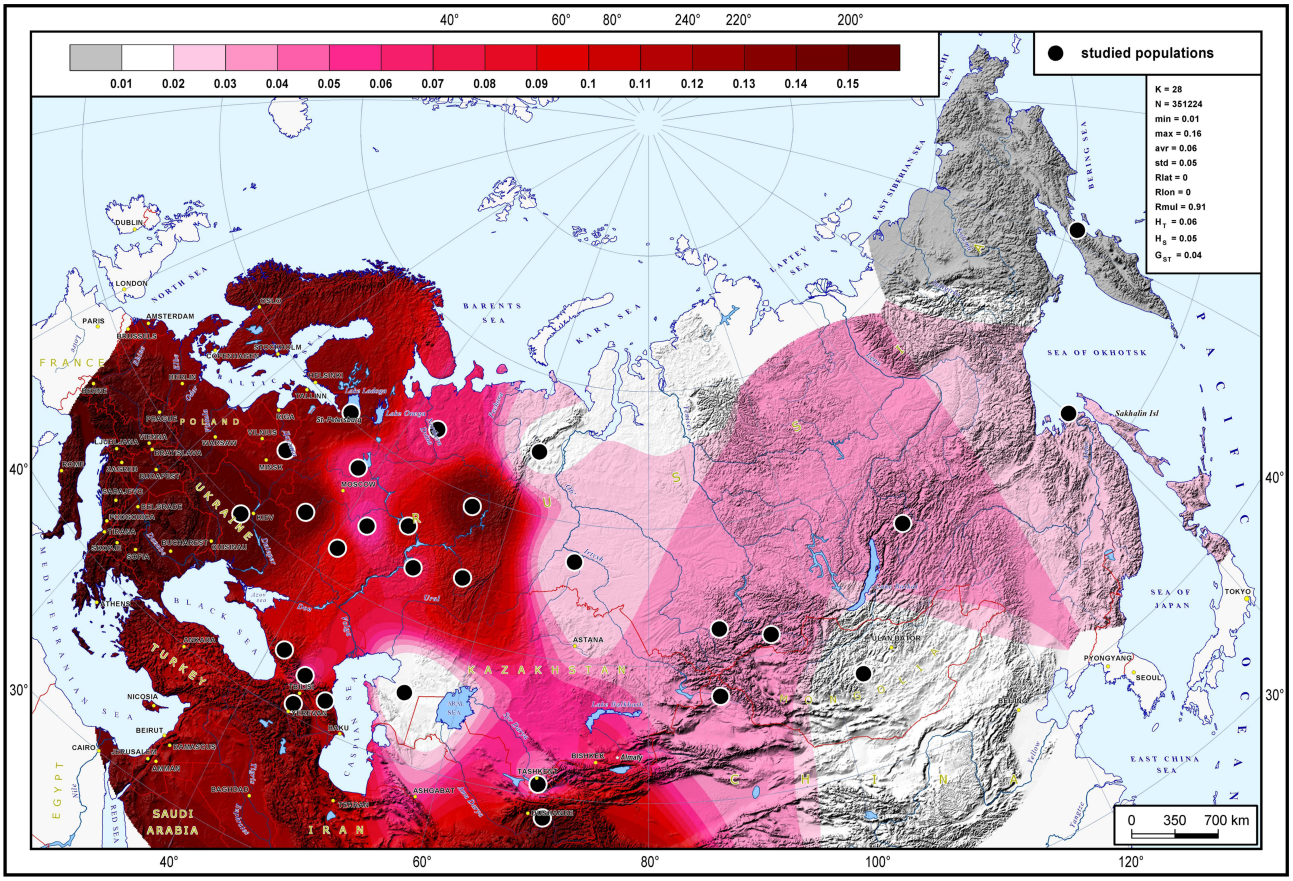

Figure 3 Variation of rs11385942_GA frequencies across Russia and its neighbor states. The frequency spectrum here is more detailed than the one shown on the world map (Figure 2). The black points represent the populations analyzed. Abbreviations in the statistical legend indicate the following.

Abbreviations: K, number of the populations studied; N, the number of the map grid nodes (calculated according to the data on the studied populations); min, the minimal frequency on the map; max, the maximum frequency on the map; avr, the average frequency on the map; std, the standard deviation; Rlat, the coefficient of partial correlation between latitude and frequency on the map; Rlon, the coefficient of partial correlation between longitude and frequency on the map; Rmul, the coefficient of multiple correlation of the frequency with both latitude and longitude on the map; H_T, the total heterozygosity; H_S, the within population heterozygosity; G_{ST}, the coefficient of genetic differentiation among populations (Masatoshi Nei's G_{ST} coefficient).

all populations in the Old World. Its highest frequencies (above 50%) are observed in sub-Saharan Africa. In most Eurasian populations, *rs657152* occurs at 40% to 50% frequencies. At the periphery of Eurasia (Europe's Atlantic fringe, Far East, Southeast Asia), its frequency tends to drop below 40%; this marker is almost absent in Native American and Australasian populations.

Variation of rs657152 Frequencies Across Russia

The frequency of rs657152 in the populations of European Russia varies from 38% to 42%, is comparable with that in Belarus (33%, $p > 0.05$) and is lower than in Ukraine (51%, $p < 0.05$) (Tables S1 and S2). In the Caucasus, the frequency of rs657152 is the highest in the east (Dagestan, 52%) and the lowest in the west (27%). The differences in rs657152 frequencies between the populations of these regions are the most significant, in comparison with other regions included in the analysis (Table S2). In Asia, the frequency of rs657152 reaches 38–39% in Tuvans and Mongols and is 51% in the populations of Uzbekistan and Kyrgyzstan (in comparison with Kazakhstan and Tajikistan, the differences here are insignificant; Tables S1 and S2).

Correlations Between Frequencies of COVID-Associated Markers and Epidemiological Parameters

The frequencies of both markers in different populations of Russia were analyzed in the context of COVID-19 recovery and mortality rates in those populations. Information about the number of COVID-19 cases, recoveries, deaths, and population sizes is provided in Table S3. Weak negative correlations (Table 1) were established

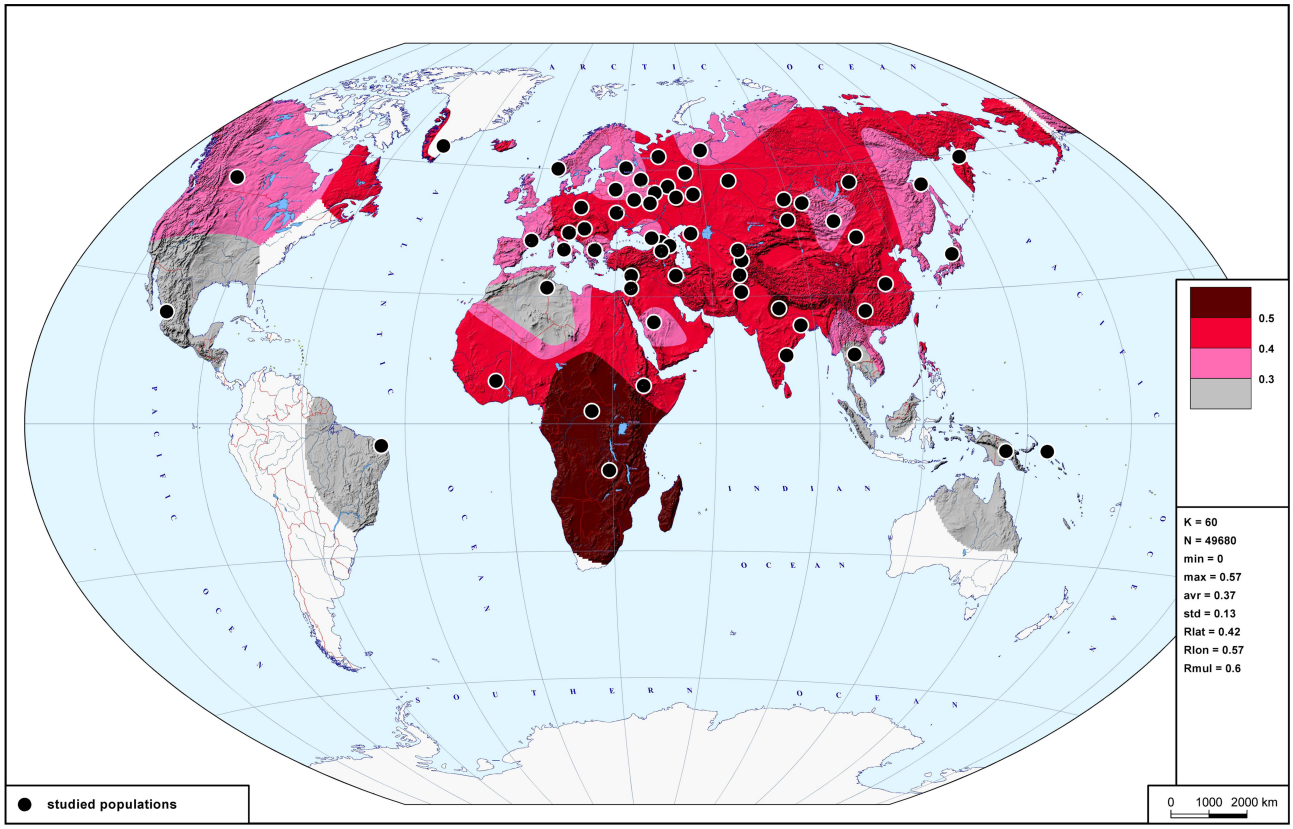

Figure 4 Global variation of rs657152_A frequencies. Four colors mark areas with four intervals of frequencies of this risk allele, according to the scale. The black points represent the populations analyzed. Abbreviations in the statistical legend indicate the following.

Abbreviations: K, number of the populations studied; N, the number of the map grid nodes (calculated according to the data on the studied populations); min, the minimal frequency on the map; max, the maximum frequency on the map; avr, the average frequency on the map; std, the standard deviation; Rlat, the coefficient of partial correlation between latitude and frequency on the map; Rlon, the coefficient of partial correlation between longitude and frequency on the map; Rmul, the coefficient of multiple correlation of the frequency with both latitude and longitude on the map.

between the frequencies of both risk alleles and the absolute number of COVID-19 cases, as well as the number of COVID-19 cases normalized to the population size. By contrast, the correlations between both risk alleles and the mortality rate, calculated as the number of deaths per total COVID-19 cases, were positive. The correlations between both risk alleles and the recovery rate calculated as the number of recoveries normalized to the total number of COVID-19 cases were negative. The correlation between COVID-19 mortality and the frequency of *rs657152* (r = 0.63) was significant (p = 0.01; Table 1), unlike the correlation between COVID-19 mortality and the frequency of *rs11385942* (r = 0.24), which was statistically insignificant (p = 0.38). The correlation between the frequency of *rs657152* and mortality from COVID-19 remained significant after the Bonferroni correction. After the 2 datasets were merged, the established correlations became insignificant (Table 1).

Discussion

The genome-wide association study by Ellinghaus et al has reported a strong association signal at locus 3p21.31

Table 1 Correlations Between COVID-19 Recoveries, Deaths and the Distribution Frequencies of the Studied Genetic Markers

	"Russian" Dataset		"World" Dataset	
Epidemiological parameter	rs11385942_GA	rs657152_A	rs11385942_GA	rs657152_A
Number of COVID-19 cases per 1 million population	−0.18 (p = 0.50)	−0.44 (p = 0.09)	0.11 (p = 0.49)	−0.13 (p = 0.44)
Number of recoveries per 1 million population	−0.18 (p = 0.50)	−0.46 (p = 0.08)	0.13 (p = 0.43)	−0.03 (p = 0.84)
Number of deaths per 1 million population	−0.17 (p = 0.52)	−0.04 (p = 0.89)	0.10 (p = 0.54)	−0.12 (p = 0.45)
Mortality rate (number of deaths per all confirmed COVID-19 cases)	0.24 (p = 0.38)	0.63 (p = 0.01)	−0.03 (p = 0.88)	−0.14 (p = 0.38)

comprising six genes (*SLC6A20, LZTFL1, CCR9, FYCO1, CXCR6, XCR1*): the risk allele *rs11385942-GA* was associated with a genetic predisposition to acute respiratory failure due to COVID-19. The frequency of the risk allele was higher among patients who required mechanical ventilation in comparison with those who were on supplemental oxygen therapy. Among the genes most strongly associated with COVID-19 were *SLC6A20, LZTFL1* and *CXCR6*. The *SLC6A20* gene encodes sodium-amino acid transporter 1, which closely interacts with ACE2, an entry receptor for SARS-CoV-2.[35,36] *LZTFL1* regulates ciliary function and intraflagellar transport in the cell.[37] *CXCR6* controls the distribution of tissue-resident memory T cells in different parts of the lungs, ensuring a stable immune response to pathogen invasion of the respiratory tract.[38] The *GA* risk allele is implicated in the decreased expression of *CXCR6* and the elevated expression of *SLC6A20* and *LZTFL1* in human lung cells, determining the severity of COVID-19.[6]

Ellinghaus et al noticed that the frequency of the *rs11385942* risk allele varies between populations but did not analyze the pattern of this variation. We compiled a dataset of *rs11385942* frequencies in human populations from different corners of the world and found that at the global level its variation follows the West Eurasian pattern. Like many other polymorphisms, *rs11385942* is common in West Eurasia (Europeans, West Asians, and South Asians) but rare or absent in other parts of the globe, including the populations of East Asia, North Asia, native Americans, and sub-Saharan Africans (North African populations are related to West Eurasians and carry *rs11385942* at moderate frequencies). This marker, along with the linked SNPs in the 50kb-long haplotype segment, is the product of Neanderthal admixture,[7] which explains its absence in sub-Saharan Africa never inhibited by Neanderthals. The absence of this SNP in East Eurasia may be attributed to the genetic drift after the divergence of West and East Eurasians. The detailed analysis of Russian populations demonstrates that differences in *rs11385942* frequencies across Eurasian subcontinents are not abrupt. For example, the elevated frequency of this marker in European populations gradually decreases eastward, hitting zero on the Pacific coast.

Another genomic site strongly associated with severe COVID-19 is *rs657152*; it is located at the ABO blood group locus. The association between the ABO blood group system and COVID-19 severity has been reported in a number of studies. For example, the risk of COVID-19-associated respiratory failure was the highest in patients with blood type A; by contrast, patients with blood type O were at the lowest risk for COVID-19-associated respiratory failure.[39,40] However, the cited studies focused primarily on the risk of infection but not on COVID-19 severity.[41] The protective effect of blood type O was attributed to the presence of neutralizing antibodies against N-linked glycans.[6,42] It is also known that there is a link between the ABO blood group locus and the expression of the von Willebrand factor (locus 12p13.31), which, together with factor VIII, promotes clotting in damaged blood vessels. Patients with non-O blood types have higher levels of VWF in their pulmonary endothelial cells than patients with blood type O,[43,44] which may explain the role of the ABO blood group system in COVID-19.

That said, *rs657152* does not directly encode any blood type but can be used to distinguish rare variants of the ABO gene.[45] The fact that the cited GWAS has established a strong association between this SNP and COVID-19 severity whereas other studies have reported associations with "classic" variants of the ABO gene indicates the complexity behind these associations, which might be resolved by future research involving sequencing of the entire ABO locus. Meanwhile, let us look at the frequency of ABO group A in different populations (Figure 5). The most impressive thing about the map in Figure 5 is the vast amount of data used for its construction: the map features 2757 populations, which is almost thirty times higher than the number of populations included in the datasets analyzed in this study. This map was constructed from literature data on blood types accumulated in the 20th century. It shows the true value of old datasets: frequency distribution maps based on modern datasets (like the 1000 Genomes Project or the dataset analyzed in this study) are only roughly accurate, whereas data published in the second half of the 20th century produce a very detailed and reliable picture (Figure 5). The map shows that the highest *rs657152* frequencies (30% and above) occur in Western Europe. Slightly lower *rs657152* frequencies are observed in the Volga-Ural region of Eastern Europe. The populations of North Asia (Siberia), East Asia and sub-Saharan Africa are characterized by overall low *rs657152* frequencies (10–20%). Most South Asian populations carry ABO_A at moderate frequencies (15–20%), but in the populations of West Central Asia the frequency of this allele is as high as in Western Europe.

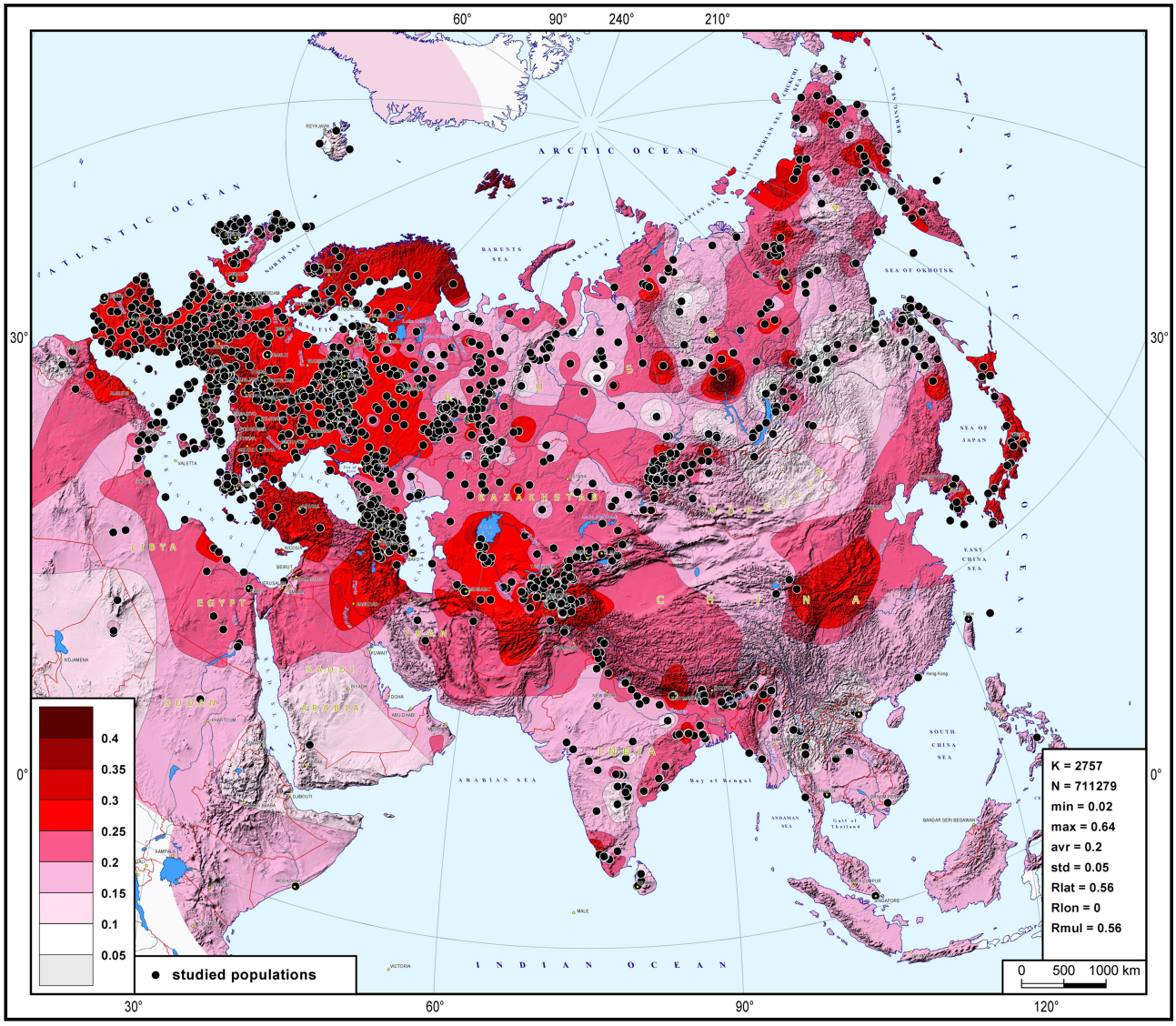

Figure 5 Distribution frequencies of blood group A (the ABO system) in the world. The map was modified from previous study.[46] The black points represent the populations analyzed. Abbreviations in the statistical legend indicate the following.

Abbreviations: K, number of the populations studied; N, the number of the map grid nodes (calculated according to the data on the studied populations); min, the minimal frequency on the map; max, the maximum frequency on the map; avr, the average frequency on the map; std, the standard deviation; Rlat, the coefficient of partial correlation between latitude and frequency on the map; Rlon, the coefficient of partial correlation between longitude and frequency on the map; Rmul, the coefficient of multiple correlation of the frequency with both latitude and longitude on the map.

Conclusion

To sum up, the frequency of the risk allele *rs11385942* is high in South Asia, slightly lower in Europe and West Asia, and low in other world regions. The frequency of the risk allele *rs657152* is the highest in Africa and moderate in Eurasia. Blood type A is most commonly observed in European populations and is moderately frequent in Eurasia. In other words, the frequencies of either of the analyzed risk alleles are the highest or higher among Europeans and South Asians than among the populations of other regions.

While investigating whether the elevated frequency of the studied risk alleles in a population may aggravate the epidemiological situation, we discovered a reasonable pattern of correlations, though we cannot assert that the link between the two was causative. First, we found no correlations between the frequencies of the studied risk alleles and the total number of COVID-19 cases in a population, which did not come as a surprise since these alleles are more associated with the severe course of the disease than with susceptibility to the infection. Second, the correlations with COVID-19 outcomes were as expected: for both markers, the higher frequency of

risk alleles was positively correlated with mortality rates. This was a trend with *rs11385942*; with *rs657152* the correlation was strong (r = 0.6) and significant ($p = 0.01$, Table 1). These reasonable correlations were observed for the "Russian" dataset only: no such correlations were established for the "world" dataset. This could be attributed to the difference in methodology used to report non-severe cases in different countries: underreported mild and moderate COVID-19 cases affect the mortality rate dramatically, making it impossible to analyze correlations with genetic factors. By contrast, within a country, the methodology of epidemiology surveillance is more uniform, and correlations with the frequency distribution of risk alleles become evident. Our samples sizes were relatively small, so statistical noise may have reduced the strength of the established correlations. We believe that even stronger correlations between COVID-19 recovery/mortality rates and the gene pool of the studied populations may transpire in future research and that genetic variation between populations makes a small yet tangible contribution into the heterogeneity of the pandemic in different parts of the globe.

Data Sharing Statement

All data generated or analyzed during the study are included in the published article.

Ethics Declaration

The study was approved by the Ethics Committee of the Research Centre for Medical Genetics (Moscow, Russia). The study complied with the principles of research ethics and the Declaration of Helsinki (1964). Written informed consent was obtained from every donor.

Acknowledgments

This work was supported by the Russian Science Foundation (Project 21-14-00363: dataset compilation, imputation, cartographic analysis), the Russian Ministry of Science and Higher Education (State Assignments for the Research Centre for Medical Genetics and Vavilov Institute of General Genetics: correlation analysis involving confirmed COVID-19 cases) and by the Russian Ministry of Health (State Assignment for the Russian Medical Academy of Continuous Professional Education: correlation analysis using epidemiological parameters per population size).

The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

Balanovsky O, Mirzaev K, Gorin I, Chernevskiy D, Agdzhoyan A, Balanovska E, Temirbulatov I, Sychev D report grants from Russian Science Foundation, during the conduct of the study. The authors declare no other conflicts of interest.

References

1. WHO. [homepage on the Internet]. WHO director-general's opening remarks at the media briefing on COVID-19; 2020. Available fromhttps://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. Accessed October 26, 2021.
2. European centre for disease prevention and control. [homepage on the Internet]. Available from: https://www.ecdc.europa.eu/en/covid-19-pandemic. Accessed October 13, 2021.
3. Ghasemian R, Shamshirian A, Heydari K, et al. Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaften für Hämatologie und Onkologie. *Onkologie*. 2010;33(6):67–108. doi:10.1159/000321409
4. Ilie PC, Stefanescu S, Smith L. The role of vitamin D in the prevention of coronavirus disease 2019 infection and mortality. *Aging Clin Exp Res*. 2020;32(7):1195–1198. doi:10.1007/s40520-020-01570-8
5. Rhodes JM, Subramanian S, Laird E, Griffin G, Kenny RA. Perspective: vitamin D deficiency and COVID-19 severity – plausibly linked by latitude, ethnicity, impacts on cytokines, ACE2 and thrombosis. *J Intern Med*. 2021;289(1):97–115. doi:10.1111/joim.13149
6. Ellinghaus D, Degenhardt F, Bujanda L, et al. Genomewide association study of severe covid-19 with respiratory failure. *N Engl J Med*. 2020;383(16):1522–1534. doi:10.1056/nejmoa2020283
7. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature*. 2020;587(7835):610–612. doi:10.1038/s41586-020-2818-3
8. Balanovska EV, Zhabagin MK, Agdzhoyan AT, et al. Population biobanks: organizational models and prospects of application in gene geography and personalized medicine. *Russ J Genet*. 2016;52(12):1227–1243. doi:10.1134/S1022795416120024
9. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1). doi:10.1186/s13742-015-0047-8
10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904–909. doi:10.1038/ng1847
11. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):e190. doi:10.1371/journal.pgen.0020190

12. Gorin I, Petrushenko V, Zapisetskaya Y, Koshel S, Balanovsky O. Application of the population biobank for analysis of the distribution of the clinically significant DNA markers in the Russian populations: bioinformatic aspects. Bulletin of RSMU (In Print).
13. Behar DM, Metspalu M, Baran Y, et al. No evidence from genome-wide data of a khazar origin for the Ashkenazi jews. *Hum Biol*. 2013;85(6):859–900. doi:10.3378/027.085.0604
14. Chaubey G, Metspalu M, Choi Y, et al. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol Biol Evol*. 2011;28(2):1013–1024. doi:10.1093/molbev/msq288
15. Di Cristofaro J, Pennarun E, Mazières S, et al. Afghan hindu kush: where Eurasian sub-continent gene flows converge. *PLoS One*. 2013;8(10):e76748. doi:10.1371/journal.pone.0076748
16. Fedorova SA, Reidla M, Metspalu E, et al. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol*. 2013;13 (1):127. doi:10.1186/1471-2148-13-127
17. Flegontov P, Changmai P, Zidkova A, et al. Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. *Sci Rep*. 2016;6(1):20768. doi:10.1038/srep20768
18. Haber M, Mezzavilla M, Xue Y, et al. Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations. *Eur J Hum Genet*. 2016;24(6):931–936. doi:10.1038/ejhg.2015.206
19. Kovacevic L, Tambets K, Ilumäe A-M, et al. Standing at the gateway to Europe - the genetic structure of western Balkan populations based on autosomal and haploid markers. *PLoS One*. 2014;9(8):e105090. doi:10.1371/journal.pone.0105090
20. Kushniarevich A, Utevska O, Chuhryaeva M, et al. Genetic heritage of the balto-slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLoS One*. 2015;10(9): e0135820. doi:10.1371/journal.pone.0135820
21. Li JZ, Absher DM, Tang H, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (80-)*. 2008;319(5866):1100–1104. doi:10.1126/science.1153717
22. Metspalu M, Romero IG, Yunusbayev B, et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J Hum Genet*. 2011;89 (6):731–744. doi:10.1016/j.ajhg.2011.11.010
23. Raghavan M, Skoglund P, Graf KE, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*. 2014;505 (7481):87–91. doi:10.1038/nature12736
24. Raghavan M, DeGiorgio M, Albrechtsen A, et al. The genetic prehistory of the New World Arctic. *Science (80-)*. 2014;345 (6200):1255832. doi:10.1126/science.1255832
25. Raghavan M, Steinrücken M, Harris K, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science (80-)*. 2015;349(6250):aab3884–aab3884. doi:10.1126/science.aab3884
26. Rasmussen M, Li Y, Lindgreen S, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. 2010;463 (7282):757–762. doi:10.1038/nature08835
27. Yunusbayev B, Metspalu M, Jarve M, et al. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol Biol Evol*. 2012;29(1):359–365. doi:10.1093/molbev/msr221
28. Yunusbayev B, Metspalu M, Metspalu E, et al. The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLOS Genet*. 2015;11(4):e1005068. doi:10.1371/journal.pgen.1005068
29. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet*. 2018;103 (3):338–348. doi:10.1016/j.ajhg.2018.07.015
30. Balanovsky O, Dibirova K, Dybo A, et al. Parallel evolution of genes and languages in the Caucasus region. *Mol Biol Evol*. 2011;28 (10):2905–2920. doi:10.1093/molbev/msr126
31. Koshel SM. Geoinformation technologies in gene geography. In: *Modern Geographical Cartography, M., «data+»*. 2012:158–166.
32. Balanovska EV, Nurbaev SD. Computer technology of gene geographic studies of the gene pool. III. The isolation of the trend surfaces. *Genetika*. 1995;31(4):536–559.
33. Metspalu M, Kivisild T, Metspalu E, et al. Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet*. 2004;5(1):26. doi:10.1186/1471-2156-5-26
34. Hallast P, Agdzhoyan A, Balanovsky O, Xue Y, Tyler-Smith C. Early replacement of West Eurasian male Y chromosomes from the east. *bioRxiv*. 2019;867317. doi:10.1101/867317
35. Kuba K, Imai Y, Ohto-Nakanishi T, Penninger JM. Trilogy of ACE2: a peptidase in the renin–angiotensin system, a SARS receptor, and a partner for amino acid transporters. *Pharmacol Ther*. 2010;128 (1):119–128. doi:10.1016/j.pharmthera.2010.06.003
36. Vuille-dit-Bille RN, Camargo SM, Emmenegger L, et al. Human intestine luminal ACE2 and amino acid transporter expression increased by ACE-inhibitors. *Amino Acids*. 2015;47(4):693–705. doi:10.1007/s00726-014-1889-6
37. Seo S, Zhang Q, Bugge K, et al. A novel protein LZTFL1 regulates ciliary trafficking of the BBSome and smoothened. *PLoS Genet*. 2011;7(11):e1002358. doi:10.1371/journal.pgen.1002358
38. Wein AN, McMaster SR, Takamura S, et al. CXCR6 regulates localization of tissue-resident memory CD8 T cells to the airways. *J Exp Med*. 2019;216(12):2748–2762. doi:10.1084/jem.20181308
39. Zhao J, Yang Y, Huang H, et al. Relationship between the ABO blood group and the coronavirus disease 2019 (COVID-19) susceptibility. *Clin Infect Dis*. 2021;73(2):328–331. doi:10.1093/cid/ciaa1150
40. Zietz M, Zucker J, Tatonetti NP. Testing the association between blood type and COVID-19 infection, intubation, and death. *medRxiv Prepr Serv Heal Sci*. 2020. doi:10.1101/2020.04.08.20058073
41. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. *JAMA*. 2020;323(13):1239. doi:10.1001/jama.2020.2648
42. Breiman A, Ruve n-Clouet N, Pendu JLE. Harnessing the natural anti-glycan immune response to limit the transmission of enveloped viruses such as SARS-CoV-2. *PLoS Pathog*. 2020;16(5):e1008556. doi:10.1371/journal.ppat.1008556
43. Franchini M, Crestani S, Frattini F, Sissa C, Bonfanti C. ABO blood group and von Willebrand factor: biological implications. *Clin Chem Lab Med*. 2014;52(9). doi:10.1515/cclm-2014-0564
44. Murray GP, Post SR, Post GR. ABO blood group is a determinant of von Willebrand factor protein levels in human pulmonary endothelial cells. *J Clin Pathol*. 2020;73(6):347–349. doi:10.1136/jclinpath-2019-206182
45. SNPedia. [homepage on the Internet]. ABO blood group; 2021. Available from: https://www.snpedia.com/index.php/SNPedia. Accessed October 13, 2021.
46. Balanovska EV, Balanovsky OP. *Russian Gene Pool*. Moscow: Luch Print; 2007:416.