Genomics 112 (2020) 442–458

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Genome-wide sequence analyses of ethnic populations across Russia

Daria V. Zhernakova[a,b,*], Vladimir Brukhin[a], Sergey Malov[a,c], Taras K. Oleksyk[a,d,r], Klaus Peter Koepfli[a,e], Anna Zhuk[a,f], Pavel Dobrynin[a,e], Sergei Kliver[a], Nikolay Cherkasov[a], Gaik Tamazian[a], Mikhail Rotkevich[a], Ksenia Krasheninnikova[a], Igor Evsyukov[a], Sviatoslav Sidorov[a], Anna Gorbunova[a,g], Ekaterina Chernyaeva[a], Andrey Shevchenko[a], Sofia Kolchanova[a,d], Alexei Komissarov[a], Serguei Simonov[a], Alexey Antonik[a], Anton Logachev[a], Dmitrii E. Polev[h], Olga A. Pavlova[h], Andrey S. Glotov[u], Vladimir Ulantsev[i], Ekaterina Noskova[i,j], Tatyana K. Davydova[s], Tatyana M. Sivtseva[k], Svetlana Limborska[l], Oleg Balanovsky[m,n,o], Vladimir Osakovsky[k], Alexey Novozhilov[p], Valery Puzyrev[q], Stephen J. O'Brien[a,t,*]

[a] Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation
[b] Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[c] Department of Mathematics, St. Petersburg Electrotechnical University, St. Petersburg, Russian Federation
[d] Biology Department, University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico
[e] National Zoological Park, Smithsonian Conservation Biology Institute, Washington, DC, USA
[f] Vavilov Institute of General Genetics, Russian Academy of Sciences, St. Petersburg Branch, St. Petersburg, Russian Federation
[g] I.I. Mechnikov North-Western State Medical University, St. Petersburg, Russian Federation
[h] Centre Biobank, Research Park, St. Petersburg State University, St. Petersburg, Russian Federation
[i] Computer Technologies Laboratory, ITMO University, St. Petersburg, Russian Federation
[j] JetBrains Research, St. Petersburg, Russian Federation
[k] Institute of Health, North-Eastern Federal University, Yakutsk, Russian Federation.
[l] Department of Molecular Bases of Human Genetics, Institute of Molecular Genetics, Russian Academy of Sciences, Moscow, Russian Federation
[m] Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russian Federation
[n] Research Centre for Medical Genetics, Moscow, Russian Federation
[o] Biobank of North Eurasia, Moscow, Russian Federation
[p] Department of Ethnography and Anthropology, St. Petersburg State University, St. Petersburg, Russian Federation
[q] Research Institute of Medical Genetics, Tomsk National Research Medical Center, Russian Academy of Science, Tomsk, Russian Federation
[r] Department of Biological Sciences, Oakland University, Rochester, MI 48309, USA
[s] Federal State Budgetary Scietific Institution, "Yakut science center of complex medical problems", Yakutsk, Russian Federation
[t] Guy Harvey Oceanographic Center, Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, 8000 North Ocean Drive, Ft Lauderdale, Florida 33004, USA
[u] Laboratory of biobanking and genomic medicine of Institute of translation biomedicine, St. Petersburg State University, St. Petersburg, Russian Federation

ABSTRACT

The Russian Federation is the largest and one of the most ethnically diverse countries in the world, however no centralized reference database of genetic variation exists to date. Such data are crucial for medical genetics and essential for studying population history. The Genome Russia Project aims at filling this gap by performing whole genome sequencing and analysis of peoples of the Russian Federation.

Here we report the characterization of genome-wide variation of 264 healthy adults, including 60 newly sequenced samples. People of Russia carry known and novel genetic variants of adaptive, clinical and functional consequence that in many cases show allele frequency divergence from neighboring populations. Population genetics analyses revealed six phylogeographic partitions among indigenous ethnicities corresponding to their geographic locales. This study presents a characterization of population-specific genomic variation in Russia with results important for medical genetics and for understanding the dynamic population history of the world's largest country.

* Corresponding authors at: Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation.
E-mail addresses: dashazhernakova@gmail.com (D.V. Zhernakova), lgdchief@gmail.com (S.J. O'Brien).

https://doi.org/10.1016/j.ygeno.2019.03.007
Received 25 December 2018; Accepted 15 March 2019
Available online 19 March 2019
0888-7543/ © 2019 Elsevier Inc. All rights reserved.

1. Introduction

The Russian Federation (Russia) has one of the most ethnically diverse indigenous human populations within a single country. According to the 2010 census, 195 ethnic groups are represented on Russian territory. The migrations of the last millennia have created a complex patchwork of human diversity that represents today's Russia. The (pre) historic milestones that founded modern Russian populations include settlement of Northern areas of Eurasia by anatomically modern humans, the eastward expansion of the Indo-European speakers, the westward expansion of the Uralic and Altai language families and centuries of admixture between them [1–6]. The migration routes for peopling Northern and Central Eurasia and the Americas inevitably passed through this territory, followed by the waves of great human migrations together with the exchange of knowledge and technology (and likely the genes) along the Silk Road [7,8]. Studies of population ancestry and structure in Russia would further provide genomic links to the lost Neanderthal and Denisovan cultures discovered in Russia's fossil beds [9,10].

There remain ongoing discussions about the origins of the ethnic Russian population. The ancestors of ethnic Russians were among the Slavic tribes that separated from the early Indo-European Group, which included ancestors of modern Slavic, Germanic and Baltic speakers, who appeared in the northeastern part of Europe ca. 1500 years ago. Slavs were found in the central part of Eastern Europe, where they came in direct contact with (and likely assimilation of) the populations speaking Uralic (Volga-Finnish and Baltic-Finnish), and also Baltic languages [11–13]. In the following centuries, Slavs interacted with the Iranian-Persian, Turkic and Scandinavian peoples, all of which in succession may have contributed to the current pattern of genome diversity across the different parts of Russia. At the end of the Middle Ages and in the early modern period, there occurred a division of the East Slavic unity into Russians, Ukrainians and Belarusians. It was the Russians who drove the colonization movement to the East, although other Slavic, Turkic and Finnish peoples took part in this movement, as the eastward migrations brought them to the Ural Mountains and further into Siberia, the Far East, and Alaska. During that interval, the Russians encountered the Finns, Ugrians, and Samoyeds speakers in the Urals, but also the Turkic, Mongolian and Tungus speakers of Siberia. Finally, in the great expanse between the Altai Mountains on the border with Mongolia, and the Bering Strait, they encountered paleo-Asiatic groups that may be genetically closest to the ancestors of the Native Americans [14]. Today's complex patchwork of human diversity in Russia has continued to be augmented by modern migrations from the Caucasus, and from Central Asia, as modern economic migrations take shape [15].

There have been several studies of genetic history of Russia using microarrays, microsatellites, Y-chromosome and mitochondrial genome sequences [16–35] and more recently using whole genome sequencing [36–38]. Most studies have focused on profiling specific ethnicities, but a centralized reference dataset of genomic variation of most Russian populations is currently lacking. Furthermore, a number of medically-relevant candidate genes with variants specific to groups within Russia have been reported [39–42]. To further expand on these reports, we initiated the Genome Russia Project [43,44], with the goal of sequencing the whole genomes of approximately 3500 individuals, including family trios, to assess the genetic diversity across the Russian Federation and to reveal functional genomic variation of medical significance.

In the current study, we annotated whole genome sequences of individuals currently living on the territory of Russia and identifying themselves as ethnic Russian or as members of a named ethnic minority (Fig. 1). We analyzed genetic variation in three modern populations of Russia (ethnic Russians from Pskov and Novgorod regions and ethnic Yakut from the Sakha Republic), and compared them to the recently released genome sequences collected from 52 indigenous Russian populations [36,37]. The incidence of function-altering mutations was explored by identifying known variants and novel variants and their allele frequencies relative to variation in adjacent European, East Asian and South Asian populations. Genomic variation was further used to estimate genetic distance and relationships, historic gene flow and barriers to gene flow, the extent of population admixture, historic population contractions, and linkage disequilibrium patterns. Lastly, we present demographic models estimating historic founder events within Russia, and a preliminary HapMap of ethnic Russians from the European part of Russia and Yakuts from eastern Siberia.

2. Results

Our study presents analyses of the whole genome sequences (WGS) at 30× coverage of 60 newly sequenced individuals from three populations: Pskov region (western Russia), Novgorod region (western Russia), and Yakutia (eastern Siberia), and comparing these to 204 individuals from 52 populations including both Russians and other ethnic groups (Table 1, Supplemental Table S1; Fig. 1). Samples of Pskov, Novgorod, and Yakut populations were collected as family trios (two biological parents and their adult child) upon obtaining informed consent and IRB approval, and with a stated three (or more) generation homogenous ancestry from the same ethnic group and the same region [45]. The genomes of all study participants were explored for known disease-associated mutations, as well as for medically important 'loss-of-function' coding variants including SNPs, short indels (< 20 bp), longer indels (20–100 bp), copy-number variants (CNVs) and segmental duplications (SDs).

Variant calling and genotyping of SNPs and short indels in Pskov, Novgorod, and Yakut genomes revealed 8 million SNPs and 2 million indels per population (Supplemental Table S2; Supplemental material). Between 3 and 4% of these SNPs were classified as novel as compared to dbSNP (Supplemental Fig. S1a-b). Overall, over 10.5 million SNPs and 2.8 million short indels were found in all 60 samples from three Genome Russia populations combined (Supplemental Fig. S1c). As might be predicted, the number of overlapping SNPs and indels was higher when comparing Pskov and Novgorod than when comparing Yakut with the western Russia populations, in line with the geographic separation of these populations (Supplemental Fig. S1c). The same trend was observed for long indels (see Supplemental material).

In addition, we resolved CNVs and aggregated them into segmental duplication (SD) profiles for the 60 Genome Russia samples. This resulted in regions of SDs spanning around ~214 Mbs in each dataset (Supplemental Table S3). The highest number of SDs (~3 Mb) was observed for Yakuts. We further compared SD profiles between populations using V statistics (Vst, see Methods and Supplemental material), and observed relatively strong differences between Yakut and the two western populations of Pskov and Novgorod, consistent with expectations based on geography (Supplemental Fig. S2).

The collection of identified SNPs was used to inspect quantitative distinctions among 264 individuals from across Eurasia (Fig. 1) using Principal Component Analysis (PCA) (Fig. 2). The first and the second eigenvectors of the PCA plot are associated with longitude and latitude, respectively, of the sample locations and accurately separate Eurasian populations according to geographic origin. East European samples cluster near Pskov and Novgorod samples, which fall between northern Russians, Finno-Ugric peoples (Karelian, Finns, Veps etc.), and other Northeastern European peoples (Swedes, Central Russians, Estonian, Latvians, Lithuanians, and Ukrainians) (Fig. 2b). Yakut individuals map into the Siberian sample cluster as expected (Fig. 2a). To obtain an extended view of population relationships, we performed a maximum likelihood-based estimation of ancestry and population structure using ADMIXTURE [46] (Fig. 2c). The Novgorod and Pskov populations show similar profiles with their Northeastern European ancestors, while the Yakut ethnic group showed mixed ancestry similar to the Buryat and Mongolian groups.

We further assessed ancestral divergence between the populations in western Russia (Pskov and Novgorod), and Siberia (Yakuts) by choosing 'Ancestry Informative Markers' (AIMs) from the major

Fig. 1. Map of the Russian Federation with locales of indigenous ethnic groups. Sample collections locales are indicated as colored circles with 2 letter code (see Supplemental Table 1). White dotted lines are arbitrary boundaries separating major population partitions suggested by the phylogenetic analyses (see text). Populations code colors correspond to geographic areas: A) Pink – North-Eastern Siberia; B) Green – Eastern Siberia; C) Brown – Western Siberia; D) Orange – Volga-Ural; E) Red – Western Russia; F) Blue – Caucasus. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Number of samples used in the study.

Sample group	Region	Number populations	Number samples	Number unrelated samples	Number families
GR Pskov	Western Russia	1	22	14	7
GR Novgorod	Western Russia	1	20	15	5
GR Yakuts	East Siberia	1	18	14	4
Mallick et al.	Many	18	31	32	0
Pagani et al.	Many	45	173	174	0
Total		55	264	249	16

GR means Genome Russia.

Eurasian population groupings represented in 1000 Genome Project [47]. As expected European-specific AIMs were concentrated in the western Russia (Pskov and Novgorod) populations compared to the Yakut samples; while the converse was observed for East Asia-specific AIMs (Supplemental Fig. S3).

Possible admixture sources of the Genome Russia populations were addressed more formally by calculating F3 statistics, which is an allele frequency-based measure, allowing to test if a target population can be modeled as a mixture of two source populations [48]. Results showed that Yakut individuals are best modeled as an admixture of Evens or Evenks with various European populations (Supplemental Table S4). Pskov and Novgorod showed admixture of European with Siberian or Finno-Ugric populations, with Lithuanian and Latvian populations being the dominant European sources for Pskov samples (Supplemental Table S4).

2.1. Medically relevant gene variants

A total 894 medically relevant gene variants, annotated in the Human Gene Mutation Database (HGMD) [49] as disease-causing mutations, were detected within the Pskov, Novgorod and Yakut populations, resulting in 1776 known disease-associated variants occurring within all samples (Supplemental Table S5). We profiled the distribution of disease-associated mutations from HGMD (disease-causing mutations – HGMD-DM) in the 60 Genome Russia individuals, which showed an average of 75 HGMD-DM variants per individual (Supplemental Fig. S4). In addition, 31 unique variants in 29 genes were classified as pathogenic after manual curation of HGMD-DM variants (Supplemental Table S6). Forty three (43) of the 60 individuals carried at least one pathogenic variant: 41 as heterozygous, one compound heterozygote, and one homozygous case (both in the *ABCA4* gene, associated with age-related macular degeneration; [50]. Notably, three of eighteen Yakut participants were heterozygous for a pathogenic variant in the *SBF1* gene (MAF = 0.17 compared to MAF < 0.003 in gnomAD database [51]; Table 2, Supplemental Table S6), related to Charcot-Marie-Tooth disease 4b3 type [52].

All variants conformed to Mendelian expectations in the trios. We validated each of the four disease-causing mutations showing the most

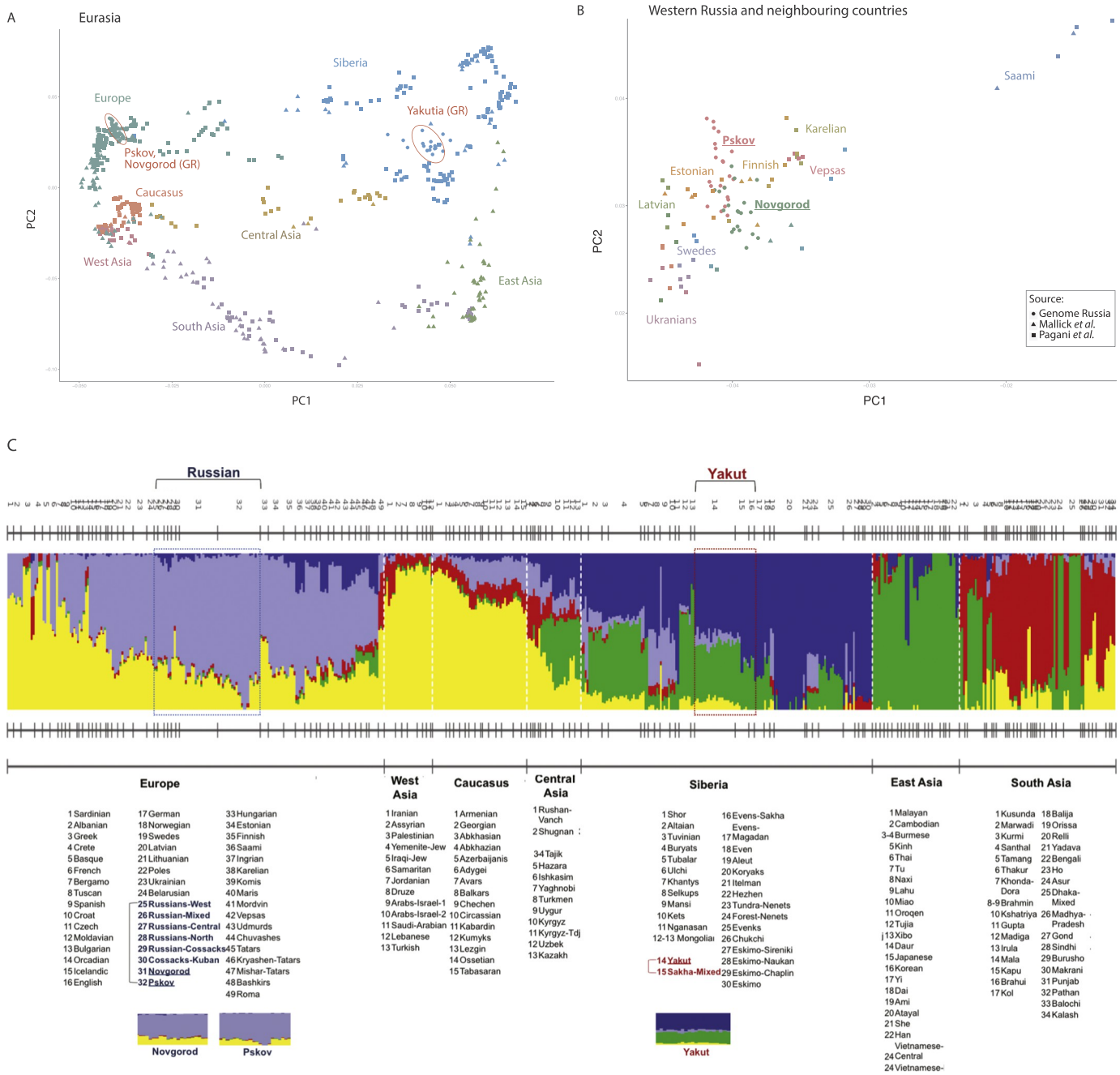

Fig. 2. Sample relatedness based on genotype data. (a,b) Principal Component plot of 574 modern Russian genomes. Colors reflect geographical regions of collection; shapes reflect the sample source. Red ovals show the location of Genome Russia samples. (a) Eurasia; (b) Western Russia and neighboring countries. (c) Population structure across samples in 178 populations from five major geographic regions (k = 5). Samples are pooled across three different studies that covered the territory of Russian Federation (Mallick et al. 2016 [36], Pagani et al. 2016 [37], this study). The optimal k-value was selected by value of cross validation error. Russian samples from all studies (highlighted in bold dark blue) show a slight gradient from Eastern European (Ukrainian, Belorussian, Polish) to North European (Estonian Karelian, Finnish) structures, reflecting population history of northward expansion. Yakut samples from different studies (highlighted in bold red) also show a slight gradient from Mongolian to Siberian people (Evens), as expected from their original admixture and northward expansions. The samples originated from this study are highlighted, and plotted in separated boxes below. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

significant allele frequency difference together with other variants listed in Table 2 by Sanger sequencing, which confirmed genotypes and MAF for each. The Genome Russia subjects carrying functional mutations were all healthy, which would be explained by heterozygotes for recessive alleles, age-dependent penetrance (e.g. for *ABCA4)*, and/or other gene–environmental interactions.

2.2. Loss-of-function SNPs

Coding gene variants were annotated for their effect on proteins using the Ensembl Variant Effect Predictor (VEP) tool [53]. We investigated their occurrence in large SNP databases (1000 Genomes – 1000G [47], Exome Aggregation Consortium – ExAC [51], Genome Aggregation Database – gnomAD [51]), and reported associations with diseases and complex traits, focusing on loss-of-function (LoF) SNPs and

Table 2
Disease-associated and function-altering variants in Genome Russia samples.

Category of discovery	Phenotype	Location	Variant id	ref/ alt	Gene	MAF Pskov	MAF Novgorod	MAF Yakuts	MAF 1000G EUR	MAF 1000G EAS	Details	Min *p*-value
Medically relevant gene variants	Albinism oculocutaneous II	15q13.1	rs74653330	C/T	*OCA2*	0.04	NA	0.214	0.01	0.027	ST5, ST6	1.49E-04
	Charcot-Marie-Tooth disease 4b3	22q13.33	rs200488568	T/C	*SBF1*	0	0	0.107	0	0.001	ST5, ST6	6.96E-05
	Age-related macular degeneration	1p22.1	rs28938473	G/A	*ABCA4*	0	0.07	0	0.006	0	ST5, ST6	0.0204
	tyrosinemia type I	15q25.1	rs11555096	C/T	*FAH*	0.14	NA	0	0.019	0	ST5, ST6	0.00268
Lof SNPs	Coronary artery calcification	2q14.3	rs117753184	A/T	*WDR33*	0	0	0.179	0	0.026	ST7	0.00104
	Diabetic kidney disease; Urinary uromodulin levels	8q24.13	rs10101626	G/T	*TBC1D31*	0.18	0.1	0.714	0.195	0.183	ST7	2.41E-09
	Astigmatism; cerebrospinal fluid clustering measurement; coronary artery bypass, vein graft stenosis	7p12.3	rs141576983	G/T	*ABCA13*	0	0	0.464	0.002	0.023	ST7	2.67E-13
Long indels	Complement C2 deficiency	6p21.33	rs572361305		*C2*	0	0.1	0	0.007	0	ST10	0.002296
Population-specific phenotypes	Lactose intolerance	2q21.3	rs4988235	A/G	*MCM6*	0.36	0.47	0.04	0.508	0	Fig. 3	0.027
	Warfarin dosage sensitivity	16p11.2	rs9923231	C/T	*VKORC*	0.25	0.2	0.86	0.388	0.885	Fig. 3	0.0121
	Skin pigmentation	5p13.2	rs16891982	C/G	*SLC45A2*	1	1	0.07	0.938	0.006	Fig. 3	0.0178
	Retinitis pigmentosa	1p36.11	rs3816539	G/A	*DHDDS*	0.11	0.07	0.96	0.235	0.709	Fig. 3	0.00599
	Short stature syndrome	2p24.3	rs369698072	C/T	*NBAS*	0	0	0.071	NA (ExAC: 0)	NA (ExAC: 1.3e-4)	NA	8.03E-06
Infectious diseases	Hepatitis B infection	6p21.32	rs9277535	A/G	*HLA-DPB1*	0.11	0.17	0.39	0.27	0.61	ST11	0.0292
	Kaposi's sarcoma	11p15.4	rs11030122	C/G	*STIM1*	0.54	0.47	0.11	0.33	0.35	ST11	0.00761
Pharmacogenomics	Tamoxifen outcomes in breast cancer	10q22.3	rs11593840	A/G	*LRMDA*	0.57	0.37	0.43	0.41	0.18	ST12	0.00212
	Irinotecan in Colorectal Cancer	2q37.1	rs6742078	G/T	*UGT1A1*	0.29	0.27	0.46	0.3	0.13	ST12	2.50E-05
		2q37.1	rs887829	C/T		0.29	0.27	0.46	0.298	0.13	ST12	2.50E-05
	Trastuzumab Lapatinib in Breast Cancer treatment	10q26.13	rs3135718	C/T	*FGFR2*	0.46	0.37	0.07	0.43	0.4	ST12	2.47E-04

Variants described in multiple sections of the paper are listed in the table (column one corresponds to the section), showing variant and overlapping gene ids, phenotype associated with the variant or the gene. Allele frequency (AF) for Genome Russia is given for the alternative allele. Details column gives the table/Fig. with more information on these variants. The last column gives the minimum *p*-value for Fisher exact test of allele count difference between either Novgorod and Pskov compared with 1000G EUR or Yakut compared with 1000G EAS. The population AFs showing the minimum *p*-value are underlined.

indels. Of 82,574 coding SNPs identified in the combined cohort, 2145 SNPs were identified as high-confidence loss-of-function variants (stop codon, frameshifts, splice alterations; see Methods). For the subsequent analyses, we selected only the 758 LoF SNPs that had an allele count of two or more and did not fail Mendelian inheritance expectations in any of the Genome Russia trios. One hundred and one (101) of these LoF SNPs were not reported in 1000G, ExAC or gnomAD (Supplemental Table S7).

We detected 34 LoF SNPs showing elevated allele frequencies in Genome Russia populations compared to that in the European (EUR), East Asian (EAS), or South Asian (SAS) populations of 1000G (18 SNPs with MAF > 10 fold, and 17 SNPs with MAF = 5–10× greater than in human genetic population databases). Implicated genes, minor allele frequency (MAF) and allele counts for each observed LoF SNP, their allele frequencies in public databases, specific disease phenotype associated with genes in GWAS catalog, including five genes that are scored as LoF-intolerant [51] are listed in Supplemental Table S7. For example, a LoF variant rs117753184 of *WDR33*, an RNA editing gene previously associated with coronary artery calcification, carries a stop codon allele in Yakut at a MAF of 18%, but occurs at 3% frequency in East Asia and is even less frequent in European and South Asian populations (Table 2, Supplemental Table S7). This gene is considered as "LoF intolerant" according to ExAC [51] and may have clinical consequences that are not yet confirmed. Other LoF SNPs in Supplemental Table S7 are also potential candidates for both population differentiation and clinical influence.

2.3. Insertions and deletions

As many as 757 short insertion-deletion mutations (< 20 bp) were annotated as LoF among the Genome Russia populations. Indel calling is known to be error prone, therefore we performed additional filtering by applying alignment-free k-mer-based genotyping (see Methods). In addition, novel insertions and deletions (indels) that failed Mendelian inheritance compliance were filtered out, leaving a total of 308 indels for which at least two alleles were present among the 43 unrelated Genome Russia individuals (Supplemental Table S8). We identified longer indels (20–100 bp) in the Pskov, Novgorod and Yakut populations and annotated the indels with Ensembl VEP [53] (Supplemental Tables S9,10). Each population had 1600–1900 long indels, of which < 1% overlapped exons and about 80% were previously recorded in dbSNP (Supplemental Table S9). Exon overlapping long indels were detected in 26 genes, with six genes having long indels located within exons in two or more populations (*AGBL5, CHIT1, DNAH9, ENOSF1, PLCH2, and ZNF683*). The majority of samples in the three populations were heterozygotes for the long indels overlapping with exons (Table 2, Supplemental Table S10).

2.4. Population-specific biomedical phenotypes

Certain diseases and heritable traits have different occurrence in different populations due to genetic drift, adaptation or migration [54–57]. Variant frequencies with population-specific patterns can lead to differences in traits or disease prevalence in different populations that can influence tailored clinical treatment specific for particular populations. To date, complete Russian genomes have not been interrogated for the presence and incidence of medically significant variants. Here, we offer a first step in making the personalized approach in genomic medicine for this part of the world. To illustrate how differences in population history can affect frequencies of important physiological traits, we examined four familiar loci in depth: *MCM6*, *VCORC1*, *SLC45A2*, and *DHDDS* (Fig. 3).

LCT, a gene that regulates adults' tolerance to lactose and milk products, is a well-known example of selection-based differentiation [58,59]. However, the first mutation associated with the lactose tolerance phenotype in Europeans −13.910: C > T (rs4988235) is not located in the *LCT* gene, but rather 14 kb upstream, within intron 13 of the mini-chromosome maintenance complex component 6 (*MCM6*) gene. The lactose tolerance *A* allele has been strongly selected in European populations within the last 10,000 years since the dawn of agriculture and modern civilization in the Fertile Crescent of the Middle East [60]. This *LCT- MCM6* variant has been suggested to be one of the strongest signals of natural selection in the human genome [58,59]. As expected, the non-European chromosomes in EAS and Yakut were all nearly fixed for the ancestral *G* (lactose intolerant) allele, while CEU and FIN had a higher frequency for the *A* (lactose tolerant) allele. Pskov and Novgorod populations show intermediate A allele frequencies (41%), lower than CEU (74%) or FIN (59%) possibly indicating the effect of admixture (Fig. 3a, Table 2).

Another example of a population-specific allele frequency difference important for medical drug dosage prediction, is response to warfarin (also called coumadin). Warfarin is a popular anti-coagulant that has severe side-effects, such as bleeding, if used in an inappropriate dosage. Response to warfarin depends on several factors, including genetic variants in the *CYP2C9* and *VKORC1* genes that are commonly used to predict the correct dose [61,62]. Carriers of the *VKORC1*-T allele, which is predominant in Asia, require a substantially lower dose of warfarin than Europeans, where the *VKORC1*-C allele predominates [61,62]. As expected, EAS and Yakut populations showed a higher frequency of the *VKORC1*-T allele (86% and 88% respectively), compared to the CEU (43%) and FIN (31%). The two western Genome Russian populations (Pskov and Novgorod) showed a T allele frequency (24%) similar to the Finnish population (Fig. 3b, Table 2). This likely means that warfarin dosage for Pskov and Novgorod individuals needs to be similar to that for Finns, while a lower dosage in Yakuts is expected to be effective, similar to populations of East Asia.

Dramatic population stratification is also apparent for *SLC45A2,* a gene related to lighter skin pigmentation. Within European populations *SLC45A2* has been shown to be under strong selection, as evidenced by multiple genome-wide scans of selection (reviewed in [63]. Analyses of ancient Eurasian genomes found that the allele associated with light skin pigmentation has likely reached fixation in modern Europeans from very low frequency during the Neolithic period, due to strong selection pressure over the past ~4000 years [57]. Not surprisingly, the light skin allele (G) is nearly fixed in both populations from western Russia (100%), while in the Yakut population the low frequency (7%) reflects some level of an ancestral genetic component shared with Europeans, since this variant is completely absent from the East Asian populations (EAS) (Fig. 3c, Table 2).

Single-nucleotide mutation in the gene that encodes cis-prenyltransferase (*DHDDS*) has been identified as the cause for non-syndromic recessive retinitis pigmentosa (RP) [64,65]. The 757G > A: recessive missense variant in the *DHDDS* gene (rs3816539) associated with retinitis pigmentosa pathology is reduced in frequency in western Russian populations (9% vs 20% in CEU and 28% in FIN) compared to other European populations, and increased in the Yakut population (96% vs 71% in EAS) compared to the other Asian populations (Fig. 3d). It is not clear if this reversal derives from genetic drift or natural selection.

The Yakut population is known to have higher allele frequencies for some variants, which sometimes leads to hereditary pathologies [39–42]. For example, rs369698072 in the *NBAS* gene is associated with short stature syndrome in Yakuts [40]. While this variant is extremely rare in European and Asian populations, it has a MAF of 7% in our Yakut samples, which is significantly higher than in other populations ($p = 8.03 \times 10^{-6}$, see Table 2).

2.5. Russian gene variants that are associated with infectious diseases, pharmacogenomics, and natural selection across the globe

Table 2 also summarizes Russian gene variants that convey notable infectious disease, natural selection and pharmacogenomic phenotypes

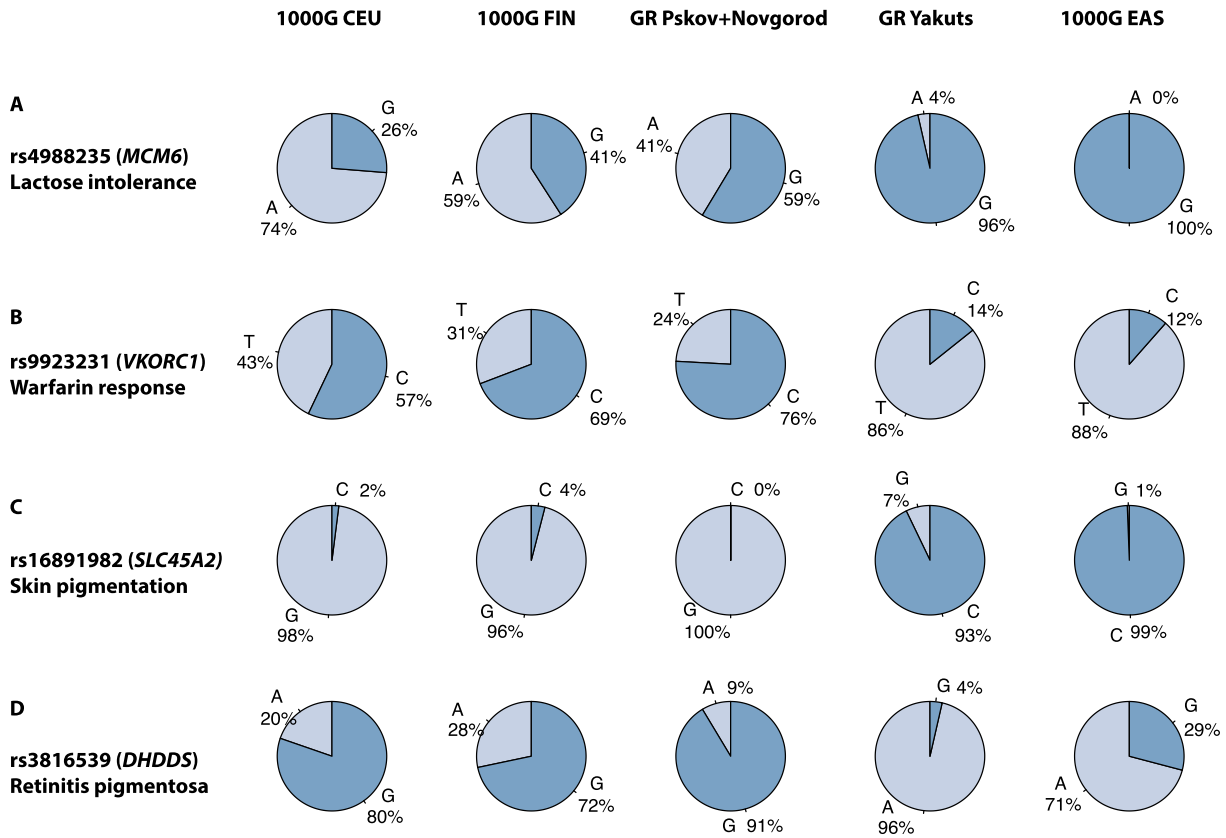

Fig. 3. Differences in Genome Russia allele frequencies of SNPs in notable genes with important phenotypes differentiate among Eurasian ethic groups.
Allele frequencies for populations of Pskov and Novgorod (combined) and Yakut are shown together with allele frequencies of 1000G populations: Europeans (CEU), Finnish (FIN), East Asians (EAS) and South Asians (SAS) for four SNPs: (a) rs4988235, located in *MCM6* gene. This SNP is associated with adult type lactose intolerance. G allele tags the lactose intolerant haplotype [58,59]; (b) rs9923231, located in *VKORC1* gene. This SNP is associated with Warfarin response. T allele carriers need reduced dose of warfarin; (c) rs16891982 located in *SLC45A2* gene. G allele related to lighter skin pigmentation; (d) rs3816539 located in DHDDS gene. A allele is associated with retinitis pigmentosa.

distinctive in world populations [54,55,66,67] (see also Supplemental Tables 11–13). Many of these show divergent MAF of Russian populations compared to parental EUR and EAS database frequencies, which we confirmed by Sanger sequencing (Supplemental material). These gene frequency distinctions may reflect historic genetic drift, occasional founder events, or possible assortative mating effects [56] that would, upon replication, be relevant to the health impact in Russian communities.

We compared the variation in MAF for variants previously associated with human infectious disease, natural selection and pharmacogenomic phenotypes [54,55,67] from Supplemental Tables S11–13 in Russian versus 1000G populations to search for patterns of overall allele frequency change during the founding of Russian populations (Supplemental Fig. S5). While allele frequencies in western Russians (Pskov and Novgorod) resembled the European reference, we observed a different pattern for allele frequencies of Yakut versus the EAS allele frequencies. For example, we observed a rather tight cluster (indicating near invariance) among all alleles in Novgorod and Pskov versus their EUR neighbors for all three gene categories, while variance of the same alleles from EAS and SAS is considerably larger (Supplemental Fig. S5, left and center plots). The Yakut population shows larger substantial deviation from all database populations (EUR, EAS and SAS) for infectious disease and pharmacogenomics associated genes, but tighter clustering of the selected alleles with EAS. For the Novgorod and Pskov populations, the pattern may likely be interpreted as indicating that all the studied alleles were adapted and set before the recent founding of these populations in Russia from EUR predecessors with little drift or perturbation effects or MAF changes since. This explanation also seems to hold for the tight clustering of Yakut and EAS for the 'selection' alleles. However, if affirmed, the absence of clustering for Yakut-EAS for the alleles mediating infectious disease and pharmacogenomics phenotypes would suggest these important gene variants were altered by selection, drift or other demographic factors in more recent times after the original founder events.

2.6. Phylogeography of Russian peoples

To further explore the relationships of individuals within and between different regions of Russia (Fig. 1), we constructed neighbor-joining trees based on pairwise nucleotide differences of ~3.8 M homologous SNPs (after filtering, see Methods) from 231 unrelated individuals representing 55 ethnicities. The resulting topology (Fig. 4a) showed a stepwise arrangement of individuals into six phylogeographic clusters ordered from eastern Asia to western Europe, corresponding to the six regions separated by white dashed lines in Fig. 1. Individuals from each of the six geographic locations were clustered together as monophyletic clades, indicating recent isolation and restricted gene flow between them since.

The family trio design of our project allowed us to accurately phase SNP data and identify the haplotype structure of our samples, which have been suggested to perform as well as or better than unlinked SNPs in reconstructing historical relationships of populations [68]. We created a haplotype-based phylogenetic tree with fineSTRUCTURE [68] using the same Russian genomes plus 308 additional neighboring Eurasian genomes [36,37](Fig. 4b). The analysis largely re-affirmed the geographic clusters obtained in the neighbor-joining tree (Fig. 4b).

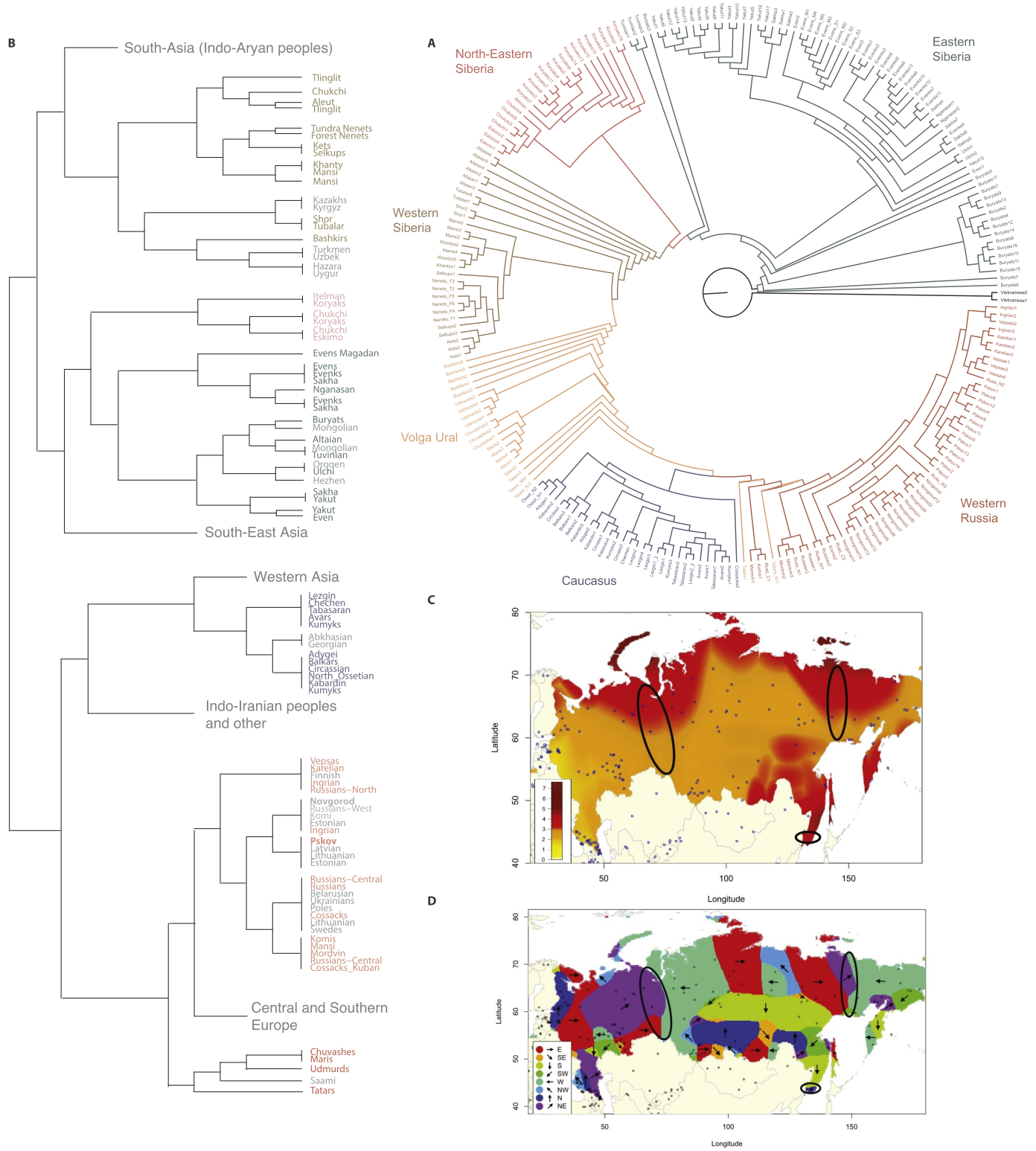

Fig. 4. Phylogenetic analyses of samples from the territory of Russia.
(a) Neighbor-joining tree showing relationships among 231 Russian-ancestry individuals based on pairwise nucleotide divergence for 3,779,316 homologous SNPs. The tree was rooted using two individuals from Vietnam. Colors are used to differentiate among individuals originating from six major geographic regions across the Russian Federation (see Fig. 1): Eastern Siberia, North-Eastern Siberia, Western Siberia, Volga-Ural region, the Caucasus, and Western Siberia. The separation between the three eastern regions (Eastern and North-Eastern and Western Siberia) and the western regions (Volga-Ural, Caucasus and Western Russia) is centered along the Ural Mountains. (b) Haplotype-based tree of samples from the territory of Russia and neighboring countries. (c,d) The heatmaps of gene flow barriers show for each point at the geographical map the interpolated differences in allele frequencies (AF) between the estimated AF at the point with AFs in the vicinity of this point. (c) The maximum difference in AFs over all directions is plotted. (d) The direction of the maximal difference in allele frequencies is coded by colors and arrows.

Individuals from Pskov clustered together with Estonians, Latvians, and Lithuanians, highlighting their close contact. Novgorod samples also showed similarities to Finno-Ugric groups such as Estonians and Ingrians. Yakuts clustered together with Evens, and Evenks (Fig. 4b), but the haplotype sharing also shows a close relation of Yakuts to Mongols, Buryats, Altaians, and Tuvinians, consistent with their postulated Turkic founders being from the Lake Baikal region [69]. The two phylogenetic approaches, when supplemented with Fig. 2, add confidence to a definitive population structure that indicates appreciable population isolation in recent times.

The genetic distinctions seen in the phylogenetic analyses suggested there appears to have been strong geographic isolation restricting gene flow between certain groups. To assess this possibility, we imputed allele frequencies (AF) at each point of a geographical grid from the available AF and estimate the differences in AF between the predicted population at each point on the grid and other predicted populations in the geographic vicinity of that point in eight directions, using a method derivative to the one used in Pagani et al. [37]. Three putative gradients indicative of restricted gene flow "barriers "were identified within Russia (Fig. 4c) taking into account the direction of the most rapid allele frequency change and the maximal directional allele frequency changes (Fig. 4d). The most intense gene flow restriction occurs on the West of Siberia (corresponding to the Urals and Ob' river), that separates the area populated predominantly by ethnic Russians (or Russian-speaking descendants of other eastern European groups) from the native people of Siberia and the Arctic. A second detected gene flow restriction lies in North Eastern Siberia (along the Lena River and Verkhoyansk Mountain Range). A third gene flow restriction barrier was identified at the Russian border at the South Far East (Fig. 4c). In this last case the direction of maximal local divergence in allele frequencies is changing from North to South if we follow across the restriction gradient region from South to North (Fig. 4d). For the other two restricted regions, the direction of maximal local divergence in allele frequencies is changing from East (North East) to West if we follow across the barrier from West to East (but not vice versa) and the maximal whole genome allele frequency changes are large. It is notable that the three areas of restricted gene flow correlate with geographic and climatic features, which may provide physical barriers for human migrations.

2.7. Demographic history

The demographic history of a population's founder events or population bottlenecks can influence the genetic diversity, the length of haplotype blocks generated by linkage disequilibrium, and the genome-wide patterning of endemic variation. We noted moderately high SNP variation in the Novgorod and Pskov samples compared to Yakut (Supplemental Fig. S1), raising the prospect of a population bottleneck or an historic founder event in the Yakut population's past. Another possible reason for the difference in variant numbers is reference bias, as the human reference genome reflects more European genetic variation. The distinctiveness of the study populations prompted a closer look at the patterns of SNP variation across the genomes. First, we computed the average length of extended regions of SNP homozygosity and noted that the Yakut population displayed relatively longer homozygous stretches (median length = 127 kbp) than the western Russian samples (median length = 119 kbp; Supplemental Fig. S6). When SNP density was plotted across the entire genome of Yakut and compared to Novgorod and Pskov, there were multiple chromosomal regions of the Yakut genomes with diminished SNP variation that would corroborate the evidence of a recent founder event or population contraction (Supplemental Figs. S6, 7).

To assess demographic history within a population, coalescence rates were calculated and scaled by mutation rate and generation time (Fig. 5). Patterns of whole-genome sequence variation were used to model population history using the diffusion approximation to the allele frequency spectrum built without inferring ancestral alleles (folded AFS) for three populations: Yakut, Novgorod and Pskov (n = 14 unrelated for each). The GADMA approximation software tool [70] was used to compare the expected allele frequency and the observed allele frequency spectrum (AFS) over the parameter value space by computing a composite-likelihood score for the best plausible evolutionary scenarios (Fig. 5). The scenarios were simulated with the AFS data and the results were used to calculate the likelihoods of best fit for each model. Pskov and Novgorod show nearly identical histories. The best model and reconstruction combined the Western Russia population and indicated patterns implying a common "out of Africa" coalescence date at 70,000 years BP followed by a split and asymmetric migration from western Russia (Pskov –Novgorod) toward Yakutia 6900 years ago, followed by slower population growth and very limited migration events between the relatively isolated populations. A more recent split between Pskov and Novgorod occurred around 1200 years ago and was followed by population growth in both populations. All three populations have subsequently increased effective population size, most probably following postulated founder events and expansion in Russian regions around that period.

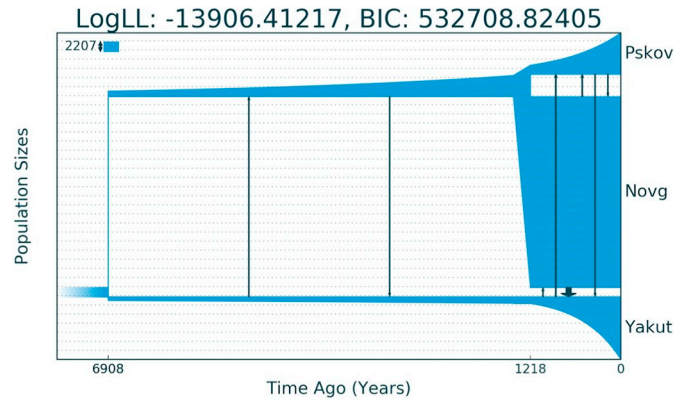

Fig. 5. Demographic history of Yakut and Pskov-Novgorod populations.
The GADMA approximation of populations' demographic history for three Genome Russia populations based upon comparison of expected allele frequency and the allele frequency spectrum. The best composite likelihood scenario suggests a founder event 6900 years BP and split between western Russian Populations and Yakut ancestors approximately 1200 years BP, coincident with the establishment of human settlements in Russian regions.

2.8. Haplotype map

An important goal of the Genome Russia Project is to construct a haplotype map (HapMap) of ethnic Russians and several smaller ethnic minorities within the Russian Federation for further use in gene association as well as population studies. We analyzed western Russians (Novgorod and Pskov) and Yakuts separately and created two haplotype maps. We also assessed the variation in SNP density within the Western Russians as compared with Yakut (Supplemental Fig. S7) and haplotype length (Supplemental Fig. S8). A Haploview LD structure of a homologous regions on chromosome 17 is presented in Supplemental Fig. S9. Although these haplotypes are based upon a limited number of trios they present an illustration of the comparative differences between the large ethnic Russian populations (Novgorod and Pskov) and the more isolated ethnic group of Yakuts. With more extensive sampling, we expect that the precision, accuracy and utility of the LD patterns will increase substantially. Variation and LD structure of Genome Russia samples can be visualized using a genome track at http://garfield.dobzhanskycenter.org/genomerussia/.

3. Discussion

We present here an analysis of genomic patterns and inference of 264 people representing 55 ethnic groups living across the Russian Federation today in an initial step toward developing a comprehensive database describing genetic diversity in Russia (Fig. 1). Using whole genome sequences, we quantified the variation, and catalogued that for medical and population-based studies. Our first goal was to inspect variation of importance in medical diagnostics by screening known disease-associated variants and "loss-of-function" mutations within all genes. In a sampling of healthy Russian study volunteers, we present known variants mined from the HGMD database [49] as well as predicted loss-of-function SNPs, short and long insertion/deletion variants, segmental duplications and copy number variation regions (Table 2; Supplemental Tables S5–10). We demonstrate the occurrence and frequencies in trait- and disease-associated variants as an illustration of the medical genomic information relevant to diagnostics and prognostics of the Russian populations (Fig. 3).

In 1950, Hermann Muller introduced the terms "Our Load of Mutations", also termed "Genetic Load", to assess and describe the accumulation of damaging or fitness-lowering variants that influence population survival, individual health and biomedical cost [71]. From the genome data offered here (Table 2; Supplemental Tables S5–13) we can begin to impute the genetic load qualitatively and quantitatively as a preview of the more comprehensive analyses anticipated in the Genome Russia Project, in parallel with the human genome sequencing initiatives begun in many world communities.

We and others have applied the tools of molecular evolutionary genetics to address the many anthropological conundrums that involve Russian peoples and their recent ancestry [16–35]. We estimated relative ancestry of populations directly and compared the relatedness and phylogenetic distinctions among different ethnicities using multiple phylogenetic algorithms (Figs. 2, 4). The results demonstrate a separation of ethnic groups along geographic regions, which is further indicated by imputation of gene flow barriers across the Russian landscape. Coalescence calculations conform to archaeological estimates as affirming a recent isolation and separation of certain populations (Fig. 5). The Yakut genomes display moderate genetic homogeneity, most of which may be explained by founder events and genetic drift also mentioned in previous publications [23,27].

Our data lend support to historical records that suggest that the ethnic Russian people had early contact with groups speaking Uralic and Baltic languages and their subsequent expansion from the Central Eastern Europe to North, South and Eastern frontiers, followed by encounters with Uralic and also Baltic speaking populations [11,12]. This historical contact inevitably contributed to admixture and to the patterns seen in the current local genome diversity in western Russia. While ethnic Russian populations cluster with the West Europeans in the PCA plot (Fig. 2a, b), the groupings are not tight; rather they are spread along an axis indicating divergence and admixture (Fig. 2b). The neighbor-joining and fineSTRUCTURE trees show that Novgorod and Pskov define distinct clusters that group with their immediate neighbors: Novgorod with the Uralic (Komi, Ingrian, Estonian) and Pskov with the Baltic people (Estonian, Latvian and Lithuanian) (Fig. 4a and b). At the same time, the Uralic populations very likely received the genetic contributions from the Russians and other Slavs, with whom they share branches of the phylogenetic trees (Fig. 2b; Fig. 4b). In addition, other peoples that came in contact in the area carry the evidence of historic admixture (e.g. Scandinavian and Finns: Fig. 2b).

The occurrence of Uralic admixture in Novgorod corroborates the historic evidence. In the middle of the 9th century, Novgorod was an important trade post on the route from the Baltic Sea to Constantinople in the Byzantine Empire. At the time, various Finnish, Baltic, and Slavic tribes populated the area [13]. The presence of Uralic admixture in Novgorod is justified by the historic contacts and gene flow that occurred for at least a millennium. Pskov is the westernmost region in modern Russia, and the presence of Estonians, Latvians and Lithuanians in the same branch on the phylogenetic tree probably indicates the same gene flow in the area, as the three modern Baltic countries had historic contact.

When we examined the Yakut ethnic population in Siberia, our analysis supports the prior historic evidence on the origin of the Yakut people who live in the Sakha Republic (Yakutia) in eastern Siberia (North East Asia) and that practiced animal husbandry and semi-nomadic lifestyle. The ancestors of Yakuts were Turkic people with some Mongolian admixture that migrated from the Yenisey river to the Lake Baikal region, and expanded to the north, as far as Kolyma river [72]. Our data clearly shows this to be the case, indicating among other things a historic admixture between the Yakut and the native North Siberian people such as Evens and Evenks (Fig. 4b). On the other hand, some individuals are closer to the Altaic people and the Mongolians, supporting the earlier theories of Yakut origins [72].

Lastly, we present preliminary haplotype maps of the three groups: ethnic Russians represented by trios collected in Novgorod and Pskov regions, as well as for the Yakut population. The population-specific HapMap will assist in the identification of the causal or operative variants resolved by genome-wide association studies.

4. Materials and methods

4.1. Sample description

We sampled family trios (two biological parents and their full aged children) of ethnic Russians from Pskov and Novgorod and Yakuts from Yakutia in Siberia (Table 1). The two ethnic Russian populations originated from the western part of the Russian Federation, namely the Pechora district of Pskov region and Starorussky district of Novgorod region. Yakut population is a representative of East Siberia and was collected in various locations in Yakutia (Sakha) Rebublic.

The research protocol and informed consent documents were approved by the Institutional Review Board (IRB) of the Saint-Petersburg State University (#65/2015).

DNA was extracted from blood samples using MagCore HF16 Automated Nucleic Acid Extractor (RBC Bioscience).

4.2. Data processing

4.2.1. Sequencing

One μg of each DNA sample was used as starting material for whole genome library preparation. DNA was sheared using an M220 Focused-ultrasonicator™ and microTUBE-50 tubes (Covaris, Inc.). The targeted library insert size was 350 bp. Genomic DNA libraries were constructed using TruSeq DNA PCR-Free Library Preparation Kits (Illumina, Inc., USA). All laboratory procedures were conducted in accordance with the protocol "TruSeq DNA PCR-Free Library Prep Reference Guide" (Illumina Part # 15036187 Rev. D, 2015). The final libraries were quantified using the KAPA library quantification kit for Illumina sequencing platforms (KAPA Biosystems, Inc., USA) and sequenced on the Illumina HiSeq 4000 platform (PE 2 × 150 bp; Illumina, Inc., USA) at the Resource Center Biobank of the Research Park of Saint-Petersburg State University, Russia, in accordance with the protocol "Illumina HiSeq 4000 System Guide" (Illumina Part # 15066496, Rev. 02 RUS, 2016).

4.2.2. Data analysis infrastructure

Data analysis was performed at the Theodosius Dobzhansky Center for Genome Bioinformatics of Saint-Petersburg State University. For our project, we developed a closed protected network to securely perform data analyses. The protected network does not have any connections to the Internet and to other segments of the computer network. It is divided into two subnetworks located in two separate buildings, one of which contains the main storage system and the second one contains a

backup storage system. Access to the network is granted from eight dedicated desktops for researchers of the Dobzhansky Center. Data analyses were performed on a server cluster containing 192 CPU cores and 1.5 Tb of memory. For data storage, we use storage systems with a total capacity of 150 Tb.

4.2.3. Raw read quality control and filtration

The initial quality control of the raw sequence reads was assessed using FastQC [73]. The distribution of 23-mer coverage was calculated and visualized by KrATER [https://pypi.python.org/pypi/KrATER/0.1], based on the Jellyfish [74] k-mer counter. Adapter occurrence was estimated using Cookiecutter [75]. As adapter occurrence was low and had little impact on the genome alignment, we skipped the adapter removal stage. Finally, only reads with a mean quality score equal to or higher than 20 (Q20) were retained.

Overall, six parameters were measured to assess the quality of the sequencing data:

1. Mode of coverage
2. Estimated mean coverage (calculated only for the non-repetitive region of genome using 23-mer distribution);
3. Variance coefficient of coverage (estimation of uniformity of coverage);
4. Fraction of read pairs with both reads retained after filtration (estimation of sequencing quality);
5. Fraction 23-mers with errors (estimation of sequencing error rate);
6. Fraction of read pairs without adapters or "N"s (estimation of library preparation and sequencing quality).

Several samples contained low quality tiles (according to FastQC) in some sequencing lanes. For these samples additional filtration was applied. All reads from low quality tiles were removed before the filtration steps described above.

4.2.4. Read alignment

We mapped raw reads that passed our quality control measures to the GRCh38 human reference genome using Bowtie2 2.2.8 [76] with the "–very-sensitive" and "-X 800" option and obtained one BAM file per sample. We obtained alignment statistics from BAM files using a combination of samtools-1.3 [77], BEDTools2–2.25.0 [78] and custom scripts written in Python 2.7.

4.2.5. Variant calling and genotyping

We sorted and indexed the individual BAM files using Sambamba 0.6.1 [79]. We used the SAMtools 1.3 mpileup utility with options *-q 37 -Q 30 -t AD,INFO/AD,ADF,INFO/ADF,ADR,INFO/ADR,DP,SP* and the BCFtools 1.3 call utility [77] with options *-v -m -f GQ,GP* for joint genotyping of samples in each population. To get a set of high-confidence variants, we selected only the variants that passed all of the following filters: (1) QUAL > 40, (2) FORMAT/GQ > 20, (3) FORMAT/DP > 10 and (4) FORMAT/SP < 20 by using BCFtools view utility.

We also filtered out variants by the universal mask (using an approach similar to [22]) which contained low-mappability and low-complexity genomic regions and covered 24% of the human genome. The regions of low mappability were identified in the following way: for each position in the genome, all 151-mers covering it were mapped back to the reference human genome using the Bowtie2 aligner with the same options as used for the read alignment and the ratio of the uniquely mapped 151-mers was calculated. If the ratio was < 0.5, then the position was considered to belong to a low-mappability region. The low-complexity genomic regions were obtained by merging three sets of regions: homopolymers of 7 bp or longer, low-complexity regions identified using DustMasker [80], and RepeatMasker-annotated low-complexity and microsatellite regions, and adding 10 bp to their flanks [81]. Statistics on the universal mask and its components are given in Supplemental Table S2.

To check the quality of genotype data and correctness of gender and family relation data of the DNA samples, we assessed the percentage of missing genotypes per sample, looked for potential outliers using PCA on genotype data and compared the identity-by-descent and gender predicted from genotype and stated in the phenotype data using PLINK 1.9 [82]. We also assessed the percentage of known and novel SNPs, indels and singleton variants in the three datasets. Genotype statistics were collected using BCFtools 1.3 [77] and PLINK1.9 [82]. For some of the additional analyses, we needed a dataset in which genotypes were merged. Merging was performed in PLINK 1.9. Indels were aligned using LeftAlignAndTrimVariants in the GATK toolkit [83] prior to merging.

To reduce the number of false positives in the list of LoF indels we validated the them by an alternative alignment-free genotyping method. For indel verification, we computed all 23-mers based on raw reads using Jellyfish software [74] and constructed de Bruijn graphs based on these kmers. For each indel we searched for unique flanking regions in the reference genome and filtered out all indels located in repeated regions or located in regions with several closely located SNVs. Next, each indel was confirmed by the presence of a bubble structure between two unique paths in the constructed graph, while for missing indels we expected only one unique path between two flanks.

4.2.6. Copy number variation and segmental duplication discovery

Copy number variants (CNVs) and segmental duplications were identified for the 60 individuals from Pskov, Novgorod and Yakuts. The human genome reference assembly GRCh38 was hard-masked from the repetitive regions using RepeatMasker [81] and Tandem Repeats Finder [84] software. Potential repeats also were identified with a k-mer approach by alignment of 36-mers using mrFast [85] and masking out the overrepresented fragments from the assembly. The copy number values (CNVs) were evaluated using mrCaNaVaR [85] in non-overlapping windows of 1Kbp of unmasked sequence. From each read of length 100 bp we selected two non-overlapping k-mers. The flanking regions of length 9 bp of potentially lower quality were excluded from the analysis.

Population genetic analysis on CNVs was performed using Vst statistic, which estimates the proportion of variance attributable to variation between populations [86]. Analysis was based on average CNV values in windows of 100Kbp and involved 15 unrelated samples from Novgorod, 16 from Pskov and 14 from Yakutia. Segmental duplications (SD) were defined as regions that span at least 10Kbp in genomic coordinates of increased average copy number value in comparison to the mean copy number value in control (non-repetitive) regions of the corresponding individual with correction for dispersion.

4.2.7. Long indel calling

We called genomic variants in the Pskov, Novgorod and Yakut populations using Platypus [87] with default options except for *–assemble* = 1, which enables local read assembly functionality. We filtered the obtained variants in the following series of steps: (1) indels called by Platypus (with "PASS" tag in "FILTER" field); (2) indels successfully normalized; (3) long indels (20 to 100 bp); (4) indels with quality scores (QUAL) > 40; (5) indels with minimal genotype quality (GQ) > 20; (6) indels outside of low complexity and low mappability regions. For steps (1), (2), (4) and (5) we used BCFtools utilities [77]. In step (2), we normalized indels using the BCFtools norm utility with the following options: *–check-ref x -m-*. In step (3), we selected long indels (20 to 100 bp) using a custom script. An indel was considered to have length from 20 to 100 bp if the difference between the lengths of the reference allele and the alternative allele was greater or equal to 20 bp and less or equal to 100 bp. In step (6), we filtered out indels located in low-complexity and low-mappability genomic regions using the universal mask described above and the BEDTools [78] intersect utility with the options *-v -header*.

To determine the conformance of long indels to Mendelian laws of

inheritance, we used BCFtools [77] with plugin "mendelian."

We considered a long indel from our datasets to be present in a database, if its coordinates as well as reference and alternative alleles exactly matched those of some long indel in the database. As 1000 Genomes Phase 3 [47], ExAC [51] and gnomAD [51] databases use GRCh37 genomic coordinates, we employed UCSC Genome Browser utility liftOver to convert genomic coordinates of long indels in our datasets from GRCh38 to GRCh37. We intersected sets of long indels using the BCFTools [77] isec utility with options *–n = 2 –w1*. For long indel annotation and filtration, we used the Ensembl Variant Effect Predictor (VEP) version 84 [53]. In Supplemental Table S9, we considered a long indel to be novel/existing, if it missed/had an *rs* identifier in its VEP annotation.

4.2.8. Creating a combined dataset of SNPs

For SNP-based analyses, including both mining of putatively medically-relevant variants and population genetics, we combined our 60 Genome Russia individuals (20 each from Novgorod, Pskov and Yakutsk) with the two published whole genome sequencing datasets of Pagani et al. [37] and Mallick et al. [36]. In all subsequent analyses (except for fineSTRUCTURE, PCA and ADMIXTURE, see below) we used only the samples collected on the territory of the Russian Federation: 31 samples from Mallick et al. [36] and 173 samples from Pagani et al. [37]. To merge the genotype data, we performed the following steps: (1) all genotype data from the two published datasets were converted to PLINK format and merged using the PLINK v.1.9 merge utility [88] (when possible we swapped the alleles and removed the SNPs when alleles were discordant); (2) we lifted the merged genotype data SNP coordinates to GRCh38 using the UCSC liftOver tool; (3) the resulting genotype data was merged with the Genome Russia genotype data of 60 samples using PLINK in the same way as described in (1). (4) Samples from the territory of Russia were extracted for further analyses; and (5) we applied the universal mask to remove low mappability and low complexity regions (as described above).

By running a preliminary PCA, we identified a batch effect associated with the genotypes distinguishing the individuals from the Pagani et al. [31] study versus the Mallick et al. [32] study. To correct for this batch effect, we ran a chi-square test and removed SNPs with *p*-values higher than 0.05 after Bonferroni correction. An additional PCA showed that this resolved the batch effect.

4.2.9. Genotype phasing

We performed phasing of genotype sets using SHAPEIT v2 [89] without reference panels. We phased the following datasets: (1) Pskov and Novgorod individuals together and (2) Yakut individuals (both (1) and (2) were used for population-specific haplotype map creation); (3) all Genome Russia individuals combined with published Eurasian samples from [36,37] (for fineSTRUCTURE tree construction). Each dataset was phased in the following way: (1) genotypes were filtered using PLINK v1.9 [88] to remove samples with call rates < 95%, families with a Mendel error rate > 5%, and to remove SNPs with MAF < 5% and a Mendel error rate > 5%; (2) SNP positions were mapped to hg19 using the UCSC liftOver tool to make the coordinates consistent with SHAPEIT v2 recombination maps; (3) only autosomes were included; (4) SHAPEIT v2 was run using default options using genetic maps based on data from the 1000 Genomes Project (1000G). Chromosome X was phased for haplotype map construction using the SHAPEIT –chrX parameter. One family from Pskov included two children, and only one of them was kept for phasing.

4.2.10. Variant annotation

The set of variants obtained after filtration were annotated using the Ensembl VEP [53] and the Loss-Of-Function Transcript Effect Estimator plugin to obtain potential LoF variants. We annotated the variants with MAFs from the 1000G phase3 [47], ExAC [51] and gnomAD [51] databases and with associated diseases obtained from the GWAS catalog [90], ClinVar [91] and HGMD [49].

We compared MAFs of variants identified in our data with those from 1000G phase 3 populations [47] using a chi-square sum test and testing codominant, dominant, recessive and allelic models. To gain more statistical power, we combined the individuals from Pskov and Novgorod. MAF estimation was performed after removing children from trios.

To identify disease-causing mutations in our samples, we used the Human Gene Mutation Database (HGMD) Professional version 2016.2 [49]. We considered an HGMD variant as potentially pathogenic if it was annotated as a "*Disease-causing Mutation*" (DM). Pathogenic status of variants was accepted after manual curation according to American College of Medical Genetics and Genomics (ACMG) recommendations [92]. Literature search was performed in PubMed using the dbSNP reference SNP ID number, gene name or disorder name as a search term. Pathogenic HGMD-DM variants were screened for variants recommended by the ACMG to be returned to patients in genome and exome sequencing studies [93].

4.3. Population data analysis

We performed population genetic analysis on the data obtained from the Genome Russia Project ($n = 60$) (Pagani et al. [37] ($n = 173$), and Mallick et al. [36] ($n = 31$), which together provide a widespread geographical sampling of Eurasia peoples. We merged whole genome data as described above. We further reduced the number of SNPs by removing those with a call rate < 95% and MAF < 5%. We performed LD pruning using PLINK v1.9 [88] indep-pairwise 1000 50 0.2 to select independent SNPs for non-phased data analysis. Finally, we reduced the number of individuals representing the Genome Russia Project by excluding progeny (Table 1).

4.3.1. Principal component analysis (PCA)

We performed explanatory PCA based on all Eurasian samples using the SNPRelate [94] R package on the pruned set of SNPs.

4.3.2. Admixture analysis

We used the unsupervised ADMIXTURE [46] algorithm to estimate genetic structure in the Genome Russia multilocus SNP dataset relative to the data from Pagani et al. [36,37]. A total of 557 individuals were included in our final dataset. Analyses were done for K values ranging from 2 to 10, each with 200 bootstrap replications. The best fitting K was selected according to the value of the cross validation error (Supplemental Table S14).

4.3.3. F3 statistics

We calculated the F3 statistics using Pskov, Novgorod and Yakut populations as targets and all possible pairs of Eurasian populations as sources using qp3pop from AdmixTools [95] with default settings. This analysis was performed on the genotype file filtered using the following filters: call rate > 0.95, MAF > 0.05 and HWE $p > 10^{-4}$; LD pruning was performed using the following parameters in plink: indep-pairwise 1000 50 0.5. Only the F3 results with *Z*-score < −3 were reported.

4.3.4. Identification of ancestry informative markers

Ancestry informative markers (AIMs) were identified based on 1000G phase3 [47] EUR, EAS and SAS data by identifying SNPs with allele count difference higher than 0.5 in each possible population pair. These SNPs were considered as AIMs for the population with higher allele counts.

4.3.5. Genetic distance and the identity by state (IBS) analysis

We used Nei's ${D_A}^{1/2}$ distance [96] to evaluate genetic differences in each pair of individuals and obtain a hierarchical cluster analysis with

complete linkage. The D_A genetic squared distance is actually a mean value of the squared Hellinger distance between allele frequencies over the whole genome. Moreover, we performed identity by state (IBS) analysis. The results obtained from the IBS analysis were very similar to the results obtained with the D_A squared distance.

The allele frequencies of each population were obtained as the half of the mean value of alleles available coded as 0, 1 and 2 for common homozygote, heterozygote and minor homozygote, respectively. Undefined variants were excluded from the analysis. An efficient methodology to evaluate allele frequencies at any point on the Earth is the kernel smoothing method. We determined that the Nadaraja–Watson estimator used in [22] to evaluate the expected allele frequencies f_a(x) at any point x of the grid from the observed allele frequencies f_{ka} of k-th population located at some point x_k does not work well if population locations x_k are not uniformly distributed on the Earth. We evaluated the expected values f_a(x) from the observed values of allele frequencies f_{ka} using a generalized linear smoother [97]:

$$f_a(x) = \sum_{k=1}^{m} w_k(x) f_{ka} \quad (1)$$

where $w_k(x) = \frac{1}{\sigma d_k}\varphi\left(\frac{x-x_k}{\sigma}\right) / \sum_{k=1}^{m} \frac{1}{\sigma d_k}\varphi\left(\frac{x-x_k}{\sigma}\right)$; $d_k = \frac{1}{d}\sum_{k=1}^{m} \frac{1}{\sigma_0}\varphi\left(\frac{x-x_k}{\sigma_0}\right)$ is the kernel density estimator with the Gaussian kernel φ. In contrast to the Nadaraja–Watson estimator, this approach is much more robust to splitting populations (up to single individuals). We selected heuristically the parameter of kernel width for density $\sigma_0 = 500$ km and the main parameter of kernel width $\sigma = 1000$ km. We mapped the genetic IBS distances based on allele frequencies between the Pskov, Novgorod, and Yakutia populations and the evaluated allele frequencies at any point of the geographic grid.

4.3.6. Gene flow barriers analysis

We improved and extended the framework for studying genetic differences between widely distributed populations of any size, originally developed in Pagani et al. [37], to investigate gene flow barriers on a grid. For any node x_{ij} in the geographical grid we draw a small circle $S_r(x_{ij})$ of radius r, set $d = 8$ equally spaced points of the Earth in the small circle in the 8 directions S, SE, E, NE, N, NW, W, SW from the node and calculated directional and mean increments for any node of the grid. The nodes were selected equally spaced in geographical coordinates with approximately 25 km between a node at the equator and its four neighbors. The distance (in kilometers) between the nodes depends on the latitude and becomes smaller in high latitudes. The distance between each node of the grid and the eight surrounding points-on-the-circle is fixed to $r = 500$ km.

The allele frequencies are obtained sequentially for each node and 8 points around it on the circle (9 points) by using the formula (1) above, which requires to evaluate allele frequencies at any point by its geographic coordinates. Finally, taking mean values of the increments for all loci, we obtained the directions of the smallest and largest divergences and the mean divergence in the area by using the following formula:

$$\Delta^2 f(x) = \frac{1}{L}\sum_a \frac{1}{\pi R r^2 d} \sum_{y \in S(x)} (f_a(x) - f_a(y))^2,$$

where $S(x)$ is the set of d points on the small sphere around x; R is the radius of Earth and L is the number of loci.

We call "barrier" a line by crossing which the genetic difference is maximal. In order to get more detailed results, we looked for directions of maximal changes in allele frequencies at each point on the grid. A true barrier should be accompanied by a rapid change of the evaluated allele frequencies with the appropriate change direction of the maximal difference in allele frequencies in its neighborhood. First, we examined the gradient direction change to the inverse (or closely inverse) one, which formed the boundary, in such a way that the gradients were directed outward with respect to the boundary. The local difference in allele frequencies Δf (x) (or, more precisely, the maximal directional difference in allele frequencies (gradient) Δf(x,ymax)) displays the magnitude of the barrier and the higher value of the ratio f(x,ymax)/Δf (x,ymin) in the neighborhood of the bound (the ratio should be equal to 1 on the bound) displays a sharpness of the barrier. Larger values of the ratio near the border should mean more gene flow along the border.

4.3.7. Neighbor-joining tree

We estimated the phylogenetic relationships among 231 individuals of Russian ancestry, including 60 individuals from the Genome Russia Project (Novgorod, Pskov and Yakut), using the neighbor-joining algorithm [98]. The majority of individuals included were derived from the studies by Mallick et al. [36] and Pagani et al. [37]. Individuals showing a high proportion of admixture between two or more populations (based on the ADMIXTURE results reported in the Mallick et al. and Pagani et al. studies) were excluded from the analysis. A data matrix of 3,779,316 homologous SNPs was assembled after filtering the full SNP data set according to call rate (> 95%), minor allele frequency (> 5%) and Hardy-Weinburg equilibrium (p < 1e-4). Neighbor-joining trees based on pairwise nucleotide divergence were constructed using PAUP* v4.0a159 [99]. Ties were broken randomly and no topological constraints were defined during the tree search. Tree files generated from PAUP* v4.0a159 (.tre) were saved and then visualized in FigTree v1.4.3 [100]. Two Vietnamese individuals were used to root the distance tree.

4.3.8. Haplotype-based tree

We used fineSTRUCTURE v2 [68] to create a haplotype-based tree. This was done on the set of 60 individuals from Pskov, Novgorod, and Yakutia combined with all samples from Eurasia from [36,37]. We phased the data as described above and removed children, which resulted in 573 samples used in the analysis. We ran fineSTRUCTURE with default settings using the 1000G phase 3 recombination maps. Chromopainter was run within fineSTRUCURE command automatically with default settings. To visualize the resulting tree, we used R scripts provided by the authors of fineSTRUCTURE.

4.4. Haplotype estimation

We inferred haplotypes from multilocus SNP genotypes by using the SHAPEIT2 tool [89] as described in the previous sections. This was done separately for the Pskov + Novgorod and Yakutia individuals. Haplotype structure analysis was performed in the Haploview software [101]. Haplotype blocks were estimated using the Solid Spine LD algorithm [101] between each pair of SNPs within 100 kb distance.

4.5. Identification of runs of homozygosity

Runs of homozygosity (ROHs) for the three populations (Novgorod, Pskov, and Yakutia) were identified using the PLINK2 software [88]. Biallelic SNPs were considered for the analysis. PLINK2 was launched with the following options: –geno 0.05 –homozyg-density 1000 –homozyg-window-het 1 –homozyg-kb 10 –homozyg-window-snp 20. These options correspond to filtering out the variants with > 5% of missing call rates and requiring runs of homozygosity to contain at least one SNP per 1 Mb on average and be at least 10 kbp long. The sliding windows consisting of 20 SNPs and containing at most one heterozygous SNP were used to scan every individual for runs of homozygosity.

5. Data access

The datasets supporting the results of this article are publicly available at http://genomerussia.spbu.ru/dataaccess.html.

Acknowledgements

The scientists at the Dobzhansky Center were supported, in part, by the Russian Science Foundation grant (project no. 17-14-01138) and by St. Petersburg State University (Genome Russia Grant no. 1.52.1647.2016). WGS was performed at Research Resource Centre "Centre Biobank", and data analysis was done at Computing Center, Research park, St. Petersburg State University. OB was supported by the Russian Science Foundation (RSF) grant 17-14-01345, Russian Foundation for Basic Research (RFBR) grant 16-04-00890 and by the State assignments of Russian Ministry of Science for the VIGG (0112-2019-0001) and for the RCMC. EN and VU were financially supported by the Government of Russian Federation (Grant 08-08). VO and TMS were supported in part by the Ministry of Education and Science of the Russian Federation (Project No. 17.6344.2017/8.9)

Disclosure declaration

The authors declare no competing interests.

Author contributions

Conceptualization: SJOB, VB, TKO, DVZ, SL. Sample collection: VB, NC, AG, IE, AS, SK, MR, AL, AN, TKD, TMS, VO, SL. DNA sample preparation and sequencing: IE, AL, DEP, AG. Data Analyses: DVZ, SM, TKO, KPK, AZ, PD, SK, NC, GT, MR, KK, IE, SSid, AG, EC, AK, SSim, AA, VU, EN, SJOB. Writing and editing: DVZ, VB, SM, TKO, KPK, PD, SK, GT, MR, KK, IE, SSid, AG, OB, AN, SJOB. Project administration: SJOB, VB, VP.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2019.03.007.

References

[1] M.E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P.B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. Della Casa, P. Dąbrowski, P.R. Duffy, A.V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolář, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtchyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, Ł. Pospieszny, T.D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V.I. Soenov, V. Szeverényi, G. Tóth, S.V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S. Brunak, R. Nielsen, K. Kristiansen, E. Willerslev, Population genomics of Bronze age Eurasia, Nature 522 (2015) 167–172, https://doi.org/10.1038/nature14507.

[2] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R.G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S.L. Pichler, R. Risch, M.A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K.W. Alt, D. Reich, Massive migration from the steppe was a source for Indo-European languages in Europe, Nature 522 (2015) 207–211, https://doi.org/10.1038/nature14317.

[3] C. Gamba, E.R. Jones, M.D. Teasdale, R.L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, A. Whittle, J. Dani, P. Raczky, T.F.G. Higham, M. Hofreiter, D.G. Bradley, R. Pinhasi, Genome flux and stasis in a five millennium transect of European prehistory, Nat. Commun. 5 (2014) 5257, https://doi.org/10.1038/ncomms6257.

[4] B. Yunusbayev, M. Metspalu, E. Metspalu, A. Valeev, S. Litvinov, R. Valiev, V. Akhmetova, E. Balanovska, O. Balanovsky, S. Turdikulova, D. Dalimova, P. Nymadawa, A. Bahmanimehr, H. Sahakyan, K. Tambets, S. Fedorova, N. Barashkov, I. Khidiyatova, E. Mihailov, R. Khusainova, L. Damba, M. Derenko, B. Malyarchuk, L. Osipova, M. Voevoda, L. Yepiskoposyan, T. Kivisild, E. Khusnutdinova, R. Villems, The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia, PLoS Genet. 11 (2015) e1005068, , https://doi.org/10.1371/journal.pgen.1005068.

[5] P. Skoglund, H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M.T.P. Gilbert, A. Götherström, M. Jakobsson, Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe, Science 336 (2012) 466–469, https://doi.org/10.1126/science.1216304.

[6] M. Raghavan, P. Skoglund, K.E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T.W. Stafford Jr., L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P.F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L.P. Osipova, S.A. Fedorova, M.I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, E. Willerslev, Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans, Nature 505 (2014) 87–91, https://doi.org/10.1038/nature12736.

[7] M. Mezzavilla, D. Vozzi, N. Pirastu, G. Girotto, P. d'Adamo, P. Gasparini, V. Colonna, Genetic landscape of populations along the silk road: admixture and migration patterns, BMC Genet. 15 (2014) 131, https://doi.org/10.1186/s12863-014-0131-6.

[8] D. Xu, S. Wen, The Silk Road: Language and Population Admixture and Replacement, in: Lang, Genes Northwest. China Adjac. Reg., Springer Singapore, Singapore, 2017, pp. 55–78, https://doi.org/10.1007/978-981-10-4169-3_4.

[9] K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P.H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J.C. Mullikin, S.H. Vohr, R.E. Green, I. Hellmann, P.L.F. Johnson, H. Blanche, H. Cann, J.O. Kitzman, J. Shendure, E.E. Eichler, E.S. Lein, T.E. Bakken, L.V. Golovanova, V.B. Doronichev, M.V. Shunkov, A.P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains, Nature 505 (2014) 43–49, https://doi.org/10.1038/nature12886.

[10] D. Reich, R.E. Green, M. Kircher, J. Krause, N. Patterson, E.Y. Durand, B. Viola, A.W. Briggs, U. Stenzel, P.L.F. Johnson, T. Maricic, J.M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E.E. Eichler, M. Stoneking, M. Richards, S. Talamo, M.V. Shunkov, A.P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia, Nature 468 (2010) 1053–1060, https://doi.org/10.1038/nature09710.

[11] Paul M. Barford, The Early Slavs : Culture and Society in Early Medieval Eastern Europe, Cornell University Press, 2001.

[12] J.P. Mallory, In Search of the Indo-Europeans : Language, Archaeology and Myth, Thames and Hudson, 1991.

[13] D.A. Machinskiy, Migration of the Slavs in the 1st millennium AD e. (from written sources with the use of archeological data), in: V.D. Korolyuk, L.V. Zaborovsky (Eds.), Form. Early Feudal Slav. Natl. Nauka, Moscow, 1981, pp. 39–51.

[14] M.C. Dulik, S.I. Zhadanov, L.P. Osipova, A. Askapuli, L. Gau, O. Gokcumen, S. Rubinstein, T.G. Schurr, Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians, Am. J. Hum. Genet. 90 (2012) 229–246, https://doi.org/10.1016/j.ajhg.2011.12.014.

[15] I.I. Korel', L.V. Korel', Modern contrasts in Russia's interregional migration, Reg. Res. Russ. 5 (2015) 147–153, https://doi.org/10.1134/S2079970515020057.

[16] P. Flegontov, P. Changmai, A. Zidkova, M.D. Logacheva, N.E. Altınışık, O. Flegontova, M.S. Gelfand, E.S. Gerasimov, E.E. Khrameeva, O.P. Konovalova, T. Neretina, Y.V. Nikolsky, G. Starostin, V.V. Stepanova, I.V. Travinsky, M. Tříska, P. Tříska, T.V. Tatarinova, Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient north Eurasian ancestry, Sci. Rep. 6 (2016) 20768, https://doi.org/10.1038/srep20768.

[17] V. Orekhov, A. Poltoraus, L.A. Zhivotovsky, V. Spitsyn, P. Ivanov, N. Yankovsky, Mitochondrial DNA sequence diversity in Russians, FEBS Lett. 445 (1999) 197–201, https://doi.org/10.1016/S0014-5793(99)00115-5.

[18] I. Morozova, A. Evsyukov, A. Kon'kov, A. Grosheva, O. Zhukova, S. Rychkov, Russian ethnic history inferred from mitochondrial DNA diversity, Am. J. Phys. Anthropol. 147 (2012) 341–351, https://doi.org/10.1002/ajpa.21649.

[19] M.V. Golubenko, V.P. Puzyrev, V.B. Salyukov, A.N. Kucher, N.O. Sanchat, Distribution of deletion-insertion polymorphism of mitochondrial DNA intragenic region V among indigenous population of the Tuva republic, Russ. J. Genet. 36 (2000) 293–297.

[20] E.L. Loogväli, U. Roostalu, B.A. Malyarchuk, M.V. Derenko, T. Kivisild, E. Metspalu, K. Tambets, M. Reidla, H.V. Tolk, J. Parik, E. Pennarun, S. Laos, A. Lunkina, M. Golubenko, L. Barać, M. Peričić, O.P. Balanovsky, V. Gusar, E.K. Khusnutdinova, V. Stepanov, V. Puzyrev, P. Rudan, E.V. Balanovska, E. Grechanina, C. Richard, J.P. Moisan, A. Chaventré, N.P. Anagnou, K.I. Pappa, E.N. Michalodimitrakis, M. Claustres, M. Gölge, I. Mikerezi, E. Usanga, R. Villems, Disuniting uniformity: A pied cladistic canvas of mtDNA haplogroup H in Eurasia, Mol. Biol. Evol. 21 (2004) 2012–2021, https://doi.org/10.1093/molbev/msh209.

[21] B. Malyarchuk, A. Litvinov, M. Derenko, K. Skonieczna, T. Grzybowski, A. Grosheva, Y. Shneider, S. Rychkov, O. Zhukova, Mitogenomic diversity in Russians and poles, Forensic Sci. Int. Genet. 30 (2017) 51–56, https://doi.org/10.1016/j.fsigen.2017.06.003.

[22] H. Sahakyan, B.H. Kashani, R. Tamang, A. Kushniarevich, A. Francis, M.D. Costa, A.K. Pathak, Z. Khachatryan, I. Sharma, M. Van Oven, J. Parik, H. Hovhannisyan, E. Metspalu, E. Pennarun, M. Karmin, E. Tamm, K. Tambets, A. Bahmanimehr, T. Reisberg, M. Reidla, A. Achilli, A. Olivieri, F. Gandini, U.A. Perego, N. Al-Zahery, M. Houshmand, M.H. Sanati, P. Soares, E. Rai, J. Šarac, T. Šarić, V. Sharma, L. Pereira, V. Fernandes, V. Černý, S. Farjadian, D.P. Singh, H. Azakli, D. Üstek, N.E. Trofimova, I. Kutuev, S. Litvinov, M. Bermisheva, E.K. Khusnutdinova, N. Rai, M. Singh, V.K. Singh, A.G. Reddy, H.V. Tolk, S. Cvjetan, L.B. Lauc, P. Rudan, E.N. Michalodimitrakis, N.P. Anagnou, K.I. Pappa, M.V. Golubenko, V. Orekhov, S.A. Borinskaya, K. Kaldma, M.A. Schauer, M. Simionescu, V. Gusar, E. Grechanina, P. Govindaraj, M. Voevoda, L. Damba, S. Sharma, L. Singh, O. Semino, D.M. Behar, L. Yepiskoposyan, M.B. Richards, M. Metspalu, T. Kivisild, K. Thangaraj, P. Endicott, G. Chaubey, A. Torroni,

R. Villems, Origin and spread of human mitochondrial DNA haplogroup U7, Sci. Rep. 7 (2017) 1–9, https://doi.org/10.1038/srep46044.
[23] V.P. Puzyrev, V.A. Stepanov, M.V. Golubenko, K.V. Puzyrev, N.R. Maximova, V.N. Kharkov, M.G. Spiridonova, A.N. Nogovitsina, MtDNA and Y-chromosome lineages in the Yakut population, Russ. J. Genet. 39 (2003) 816–822, https://doi.org/10.1023/A:1024761305958.
[24] V. Pankratov, S. Litvinov, A. Kassian, D. Shulhin, L. Tchebotarev, B. Yunusbayev, M. Möls, H. Sahakyan, L. Yepiskoposyan, East Eurasian ancestry in the middle of Europe: genetic footprints of steppe nomads in the genomes of Belarusian Lipka Tatars, Nat. Publ. Gr. (2016) 1–11, https://doi.org/10.1038/srep30197.
[25] P. Triska, N. Chekanov, V. Stepanov, E.K. Khusnutdinova, G.P.A. Kumar, V. Akhmetova, K. Babalyan, E. Boulygina, V. Kharkov, M. Gubina, I. Khidiyatova, I. Khitrinskaya, E.E. Khrameeva, R. Khusainova, N. Konovalova, S. Litvinov, A. Marusin, A.M. Mazur, V. Puzyrev, D. Ivanoshchuk, M. Spiridonova, A. Teslyuk, S. Tsygankova, M. Triska, N. Trofimova, E. Vajda, O. Balanovsky, A. Baranova, K. Skryabin, T.V. Tatarinova, E. Prokhortchouk, Between Lake Baikal and the Baltic Sea: genomic history of the gateway to Europe, BMC Genet. 18 (2017) 110, https://doi.org/10.1186/s12863-017-0578-3.
[26] K. Tambets, B. Yunusbayev, G. Hudjashov, A.-M. Ilumäe, S. Rootsi, T. Honkola, O. Vesakoski, Q. Atkinson, P. Skoglund, A. Kushniarevich, S. Litvinov, M. Reidla, E. Metspalu, L. Saag, T. Rantanen, M. Karmin, J. Parik, S.I. Zhadanov, M. Gubina, L.D. Damba, M. Bermisheva, T. Reisberg, K. Dibirova, I. Evseeva, M. Nelis, J. Klovins, A. Metspalu, T. Esko, O. Balanovsky, E. Balanovska, E.K. Khusnutdinova, L.P. Osipova, M. Voevoda, R. Villems, T. Kivisild, M. Metspalu, Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations, Genome Biol. 19 (2018) 139, https://doi.org/10.1186/s13059-018-1522-1.
[27] S.A. Fedorova, M. Reidla, E. Metspalu, M. Metspalu, S. Rootsi, K. Tambets, N. Trofimova, S.I. Zhadanov, B. Kashani, A. Olivieri, M.I. Voevoda, L.P. Osipova, F.A. Platonov, M.I. Tomsky, E.K. Khusnutdinova, A. Torroni, R. Villems, Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia, BMC Evol. Biol. 13 (2013) 127, https://doi.org/10.1186/1471-2148-13-127.
[28] V.N. Pimenoff, D. Comas, J.U. Palo, G. Vershubsky, A. Kozlov, A. Sajantila, Northwest Siberian Khanty and Mansi in the junction of west and east Eurasian gene pools as revealed by uniparental markers, Eur. J. Hum. Genet. 16 (2008) 1254–1264, https://doi.org/10.1038/ejhg.2008.101.
[29] C. Der Sarkissian, O. Balanovsky, G. Brandt, V. Khartanovich, A. Buzhilova, S. Koshel, V. Zaporozhchenko, D. Gronenborn, V. Moiseyev, E. Kolpakov, V. Shumkin, K.W. Alt, E. Balanovska, A. Cooper, W. Haak, G. Consortium, Ancient DNA reveals prehistoric gene-flow from siberia in the complex human population history of North East Europe, PLoS Genet. 9 (2013) e1003296, , https://doi.org/10.1371/journal.pgen.1003296.
[30] A. Kushniarevich, O. Utevska, M. Chuhryaeva, A. Agdzhoyan, K. Dibirova, I. Uktveryte, M. Möls, L. Mulahasanovic, A. Pshenichnov, S. Frolova, A. Shanko, E. Metspalu, M. Reidla, K. Tambets, E. Tamm, S. Koshel, V. Zaporozhchenko, L. Atramentova, V. Kučinskas, O. Davydenko, O. Goncharova, I. Evseeva, M. Churnosov, E. Pocheshchova, B. Yunusbayev, E. Khusnutdinova, D. Marjanović, P. Rudan, S. Rootsi, N. Yankovsky, P. Endicott, A. Kassian, A. Dybo, C. Tyler-Smith, E. Balanovska, M. Metspalu, T. Kivisild, R. Villems, O. Balanovsky, O. Balanovsky, et al., PLoS One 10 (2015) e0135820, , https://doi.org/10.1371/journal.pone.0135820.
[31] A.V. Khrunin, D.V. Khokhrin, I.N. Filippova, T. Esko, M. Nelis, N.A. Bebyakova, N.L. Bolotova, J. Klovins, L. Nikitina-Zake, K. Rehnström, S. Ripatti, S. Schreiber, A. Franke, M. Macek, V. Krulišová, J. Lubinski, A. Metspalu, S.A. Limborska, A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe, PLoS One 8 (2013) e58552, , https://doi.org/10.1371/journal.pone.0058552.
[32] L. Roewer, S. Willuweit, C. Krüger, M. Nagy, S. Rychkov, I. Morozowa, O. Naumova, Y. Schneider, O. Zhukova, M. Stoneking, I. Nasidze, Analysis of Y chromosome STR haplotypes in the European part of Russia reveals high diversities but non-significant genetic distances between populations, Int. J. Legal Med. 122 (2008) 219–223, https://doi.org/10.1007/s00414-007-0222-2.
[33] A. Fechner, D. Quinque, S. Rychkov, I. Morozowa, O. Naumova, Y. Schneider, S. Willuweit, O. Zhukova, L. Roewer, M. Stoneking, I. Nasidze, Boundaries and clines in the west Eurasian Y-chromosome landscape: insights from the European part of Russia, Am. J. Phys. Anthropol. 137 (2008) 41–47, https://doi.org/10.1002/ajpa.20838.
[34] O. Balanovsky, S. Rootsi, A. Pshenichnov, T. Kivisild, M. Churnosov, I. Evseeva, E. Pocheshkhova, M. Boldyreva, N. Yankovsky, E. Balanovska, R. Villems, Two sources of the Russian patrilineal heritage in their Eurasian context, Am. J. Hum. Genet. 82 (2008) 236–250, https://doi.org/10.1016/j.ajhg.2007.09.019.
[35] B. Malyarchuk, T. Grzybowski, M. Derenko, M. Perkova, T. Vanecek, J. Lazur, P. Gomolcak, I. Tsybovsky, Mitochondrial DNA phylogeny in eastern and Western Slavs, Mol. Biol. Evol. 25 (2008) 1651–1658, https://doi.org/10.1093/molbev/msn114.
[36] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J.P. Spence, Y.S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I.G. Romero, A.R. Jha, D.M. Behar, C.M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O.L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M.S. Abdullah, A. Ruiz-Linares, C.M. Beall, A. Di Rienzo, C. Jeong, E.B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B.M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J.T.S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M.F. Hammer, T. Kivisild, W. Klitz, C.A. Winkler, D. Labuda, M. Bamshad, L.B. Jorde, S.A. Tishkoff, W.S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The Simons Genome Diversity Project: 300 genomes from 142 diverse populations, Nature 538 (2016) 201–206, https://doi.org/10.1038/nature18964.
[37] L. Pagani, D.J. Lawson, E. Jagoda, A. Mörseburg, A. Eriksson, M. Mitt, F. Clemente, G. Hudjashov, M. DeGiorgio, L. Saag, J.D. Wall, A. Cardona, R. Mägi, M.A.W. Sayres, S. Kaewert, C. Inchley, C.L. Scheib, M. Järve, M. Karmin, G.S. Jacobs, T. Antao, F.M. Iliescu, A. Kushniarevich, Q. Ayub, C. Tyler-Smith, Y. Xue, B. Yunusbayev, K. Tambets, C.B. Mallick, L. Saag, E. Pocheshkhova, G. Andriadze, C. Muller, M.C. Westaway, D.M. Lambert, G. Zoraqi, S. Turdikulova, D. Dalimova, Z. Sabitov, G.N.N. Sultana, J. Lachance, S. Tishkoff, K. Momynaliev, J. Isakova, L.D. Damba, M. Gubina, P. Nymadawa, I. Evseeva, L. Atramentova, O. Utevska, F.-X. Ricaut, N. Brucato, H. Sudoyo, T. Letellier, M.P. Cox, N.A. Barashkov, V. Škaro, L. Mulahasanovic, D. Primorac, H. Sahakyan, M. Mormina, C.A. Eichstaedt, D.V. Lichman, S. Abdullah, G. Chaubey, J.T.S. Wee, E. Mihailov, A. Karunas, S. Litvinov, R. Khusainova, N. Ekomasova, V. Akhmetova, I. Khidiyatova, D. Marjanović, L. Yepiskoposyan, D.M. Behar, E. Balanovska, A. Metspalu, M. Derenko, B. Malyarchuk, M. Voevoda, S.A. Fedorova, L.P. Osipova, M.M. Lahr, P. Gerbault, M. Leavesley, A.B. Migliano, M. Petraglia, O. Balanovsky, E.K. Khusnutdinova, E. Metspalu, M.G. Thomas, A. Manica, R. Nielsen, R. Villems, E. Willerslev, T. Kivisild, M. Metspalu, Genomic analyses inform on migration events during the peopling of Eurasia, Nature 538 (2016) 238–242, https://doi.org/10.1038/nature19792.
[38] E.H.M. Wong, A. Khrunin, L. Nichols, D. Pushkarev, D. Khokhrin, D. Verbenko, O. Evgrafov, J. Knowles, J. Novembre, S. Limborska, A. Valouev, Reconstructing genetic history of Siberian and Northeastern European populations, Genome Res. 27 (2017) 1–14, https://doi.org/10.1101/gr.202945.115.
[39] D. Kumar, D. Kumar, Genomics and Health in the Developing World, Oxford University Press, 2012, https://www.oupjapan.co.jp/en/node/1808 , Accessed date: 1 February 2018.
[40] N. Maksimova, K. Hara, I. Nikolaeva, T. Chun-Feng, T. Usui, M. Takagi, Y. Nishihira, A. Miyashita, H. Fujiwara, T. Oyama, A. Nogovicina, A. Sukhomyasova, S. Potapova, R. Kuwano, H. Takahashi, M. Nishizawa, O. Onodera, Neuroblastoma amplified sequence gene is associated with a novel short stature syndrome characterised by optic nerve atrophy and Pelger-Huët anomaly, J. Med. Genet. 47 (2010) 538–548, https://doi.org/10.1136/jmg.2009.074815.
[41] H. Kondo, N. Maksimova, T. Otomo, H. Kato, A. Imai, Y. Asano, K. Kobayashi, S. Nojima, A. Nakaya, Y. Hamada, K. Irahara, E. Gurinova, A. Sukhomyasova, A. Nogovicina, M. Savvina, T. Yoshimori, K. Ozono, N. Sakai, Mutation in *VPS33A* affects metabolism of glycosaminoglycans: a new type of mucopolysaccharidosis with severe systemic symptoms, Hum. Mol. Genet. (2016), https://doi.org/10.1093/hmg/ddw377 ddw377.
[42] V.P. Puzyrev, N.R. Maksimova, Hereditary diseases among Yakuts, Genetika 44 (2008) 1317–1324 http://www.ncbi.nlm.nih.gov/pubmed/19062529 , Accessed date: 5 March 2018.
[43] T.K. Oleksyk, V. Brukhin, S.J. O'Brien, The genome Russia project: closing the largest remaining omission on the world genome map, Gigascience 4 (2015) 53, https://doi.org/10.1186/s13742-015-0095-0.
[44] T.K. Oleksyk, V. Brukhin, S.J. O'Brien, Putting Russia on the genome map, Science 350 (2015) 747, https://doi.org/10.1126/science.350.6262.747-a.
[45] S. Leslie, B. Winney, G. Hellenthal, D. Davison, A. Boumertit, T. Day, K. Hutnik, E.C. Royrvik, B. Cunliffe, D.J. Lawson, D. Falush, C. Freeman, M. Pirinen, S. Myers, M. Robinson, P. Donnelly, W. Bodmer, P. Donnelly, W. Bodmer, The fine-scale genetic structure of the British population, Nature 519 (2015) 309–314, https://doi.org/10.1038/nature14230.
[46] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19 (2009) 1655–1664, https://doi.org/10.1101/gr.094052.109.
[47] A. Auton, G.R. Abecasis, D.M. Altshuler, R.M. Durbin, D.R. Bentley, A. Chakravarti, A.G. Clark, P. Donnelly, E.E. Eichler, P. Flicek, S.B. Gabriel, R.A. Gibbs, E.D. Green, M.E. Hurles, B.M. Knoppers, J.O. Korbel, E.S. Lander, C. Lee, H. Lehrach, E.R. Mardis, G.T. Marth, G.A. McVean, D.A. Nickerson, J.P. Schmidt, S.T. Sherry, J. Wang, R.K. Wilson, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J.G. Reid, Y. Zhu, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, N. Gupta, N. Gharani, L.H. Toji, N.P. Gerry, A.M. Resch, J. Barker, L. Clarke, L. Gil, S.E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W.M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R.E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, R. Sudbrak, M.W. Albrecht, V.S. Amstislavskiy, T.A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, L. Fulton, R. Fulton, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, S. Balasubramaniam, J. Burton, P. Danecek, T.M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, C.J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, C.L. Campbell, Y. Kong, A. Marcketta, F. Yu, L. Antunes, M. Bainbridge, A. Sabo, Z. Huang, L.J.M. Coin, L. Fang, Q. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, E.P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A.N. Ward, J. Wu, M. Zhang, M.J. Daly, M.A. DePristo, R.E. Handsaker, E. Banks,

G. Bhatia, G. del Angel, G. Genovese, H. Li, S. Kashin, S.A. McCarroll, J.C. Nemesh, R.E. Poplin, S.C. Yoon, J. Lihm, V. Makarov, S. Gottipati, A. Keinan, J.L. Rodriguez-Flores, T. Rausch, M.H. Fritz, A.M. Stütz, K. Beal, A. Datta, J. Herrero, G.R.S. Ritchie, D. Zerbino, P.C. Sabeti, I. Shlyakhter, S.F. Schaffner, J. Vitti, D.N. Cooper, E.V. Ball, P.D. Stenson, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Kahn, L. Murray, J. Peden, R. Shaw, E.E. Kenny, M.A. Batzer, M.K. Konkel, J.A. Walker, D.G. MacArthur, M. Lek, R. Herwig, L. Ding, D.C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. Frederick Willems, J.T. Simpson, M.D. Shriver, J.A. Rosenfeld, C.D. Bustamante, S.B. Montgomery, F.M. De La Vega, J.K. Byrnes, A.W. Carroll, M.K. DeGorter, P. Lacroute, B.K. Maples, A.R. Martin, A. Moreno-Estrada, S.S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F.C.L. Hyland, D.W. Craig, A. Christoforides, N. Homer, T. Izatt, A.A. Kurdoglu, S.A. Sinari, K. Squire, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E.G. Burchard, R.D. Hernandez, C.R. Gignoux, D. Haussler, S.J. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares, E.T. Dermitzakis, S.E. Devine, H. Min Kang, J.M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, J.L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D.K. Xifara, T.K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, B.L. Browning, S.R. Browning, F. Hormozdiari, P.H. Sudmant, E. Khurana, C. Tyler-Smith, C.A. Albers, Q. Ayub, Y. Chen, V. Colonna, L. Jostins, K. Walter, Y. Xue, M.B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A.O. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, C. Hartl, K. Shakir, J. Degenhardt, S. Meiers, B. Raeder, F. Paolo Casale, O. Stegle, E.-W. Lameijer, I. Hall, V. Bafna, J. Michaelson, E.J. Gardner, R.E. Mills, G. Dayama, K. Chen, X. Fan, Z. Chong, T. Chen, M.J. Chaisson, J. Huddleston, M. Malig, B.J. Nelson, N.F. Parrish, B. Blackburne, S.J. Lindsay, Z. Ning, Y. Zhang, H. Lam, C. Sisu, D. Challis, U.S. Evani, J. Lu, U. Nagaswamy, J. Yu, W. Li, L. Habegger, H. Yu, F. Cunningham, I. Dunham, K. Lage, J. Berg Jespersen, H. Horn, D. Kim, R. Desalle, A. Narechania, M.A. Wilson Sayres, F.L. Mendez, G. David Poznik, P.A. Underhill, L. Coin, D. Mittelman, R. Banerjee, M. Cerezo, T.W. Fitzgerald, S. Louzada, A. Massaia, G.R. Ritchie, F. Yang, D. Kalra, W. Hale, X. Dan, K.C. Barnes, C. Beiswanger, H. Cai, H. Cao, B. Henn, D. Jones, J.S. Kaye, A. Kent, A. Kerasidou, R. Mathias, P.N. Ossorio, M. Parker, C.N. Rotimi, C.D. Royal, K. Sandoval, Y. Su, Z. Tian, S. Tishkoff, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J.C. Martinez-Cruzado, R.A. Mathias, A. Hennis, H. Watson, C. McKenzie, F. Qadri, R. LaRocque, X. Deng, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. Sisay Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. Tịnh Hiê'n, S.J. Dunstan, N. Thuy Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, J. Schieffelin, D.S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L.D. Brooks, A.L. Felsenfeld, J.E. McEwen, Y. Vaydylevich, A. Duncanson, M. Dunn, J.A. Schloss, A global reference for human genetic variation, Nature 526 (2015) 68–74, https://doi.org/10.1038/nature15393.

[48] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, L. Singh, Reconstructing Indian population history, Nature 461 (2009) 489–494, https://doi.org/10.1038/nature08365.

[49] P.D. Stenson, E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S.T. Thomas, S. Abeysinghe, M. Krawczak, D.N. Cooper, Human gene mutation database (HGMD ®): 2003 update, Hum. Mutat. 21 (2003) 577–581, https://doi.org/10.1002/humu.10212.

[50] R. Allikmets, N.F. Shroyer, N. Singh, J.M. Seddon, R.A. Lewis, P.S. Bernstein, A. Peiffer, N.A. Zabriskie, Y. Li, A. Hutchinson, M. Dean, J.R. Lupski, M. Leppert, Mutation of the Stargardt disease gene (ABCR) in age-related macular degeneration, Science 277 (1997) 1805–1807 http://www.ncbi.nlm.nih.gov/pubmed/9295268 , Accessed date: 10 January 2018.

[51] M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, T. Tukiainen, D.P. Birnbaum, J.A. Kosmicki, L.E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D.N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M.I. Kurki, A.L. Moonshine, P. Natarajan, L. Orozco, G.M. Peloso, R. Poplin, M.A. Rivas, V. Ruano-Rubio, S.A. Rose, D.M. Ruderfer, K. Shakir, P.D. Stenson, C. Stevens, B.P. Thomas, G. Tiao, M.T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D.M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J.C. Florez, S.B. Gabriel, G. Getz, S.J. Glatt, C.M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M.I. McCarthy, D. McGovern, R. McPherson, B.M. Neale, A. Palotie, S.M. Purcell, D. Saleheen, J.M. Scharf, P. Sklar, P.F. Sullivan, J. Tuomilehto, M.T. Tsuang, H.C. Watkins, J.G. Wilson, M.J. Daly, D.G. MacArthur, E.A. Consortium, Analysis of protein-coding genetic variation in 60,706 humans, Nature 536 (2016) 285–291, https://doi.org/10.1038/nature19057.

[52] A.M. Alazami, F. Alzahrani, S. Bohlega, F.S. Alkuraya, SET binding factor 1 (SBF1) mutation causes Charcot-Marie-tooth disease type 4B3, Neurology 82 (2014) 1665–1666, https://doi.org/10.1212/WNL.0000000000000331.

[53] W. Mclaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The ensembl variant effect predictor, Genome Biol. (2016) 1–14, https://doi.org/10.1186/s13059-016-0974-4.

[54] U.A. Meyer, U.M. Zanger, M. Schwab, Omics and drug response, Annu. Rev. Pharmacol. Toxicol. 53 (2013) 475–502, https://doi.org/10.1146/annurev-pharmtox-010510-100502.

[55] S. Fan, M.E.B. Hansen, Y. Lo, S.A. Tishkoff, Going global by adapting local: a review of recent human adaptation, Science 354 (2016) 54–59, https://doi.org/10.1126/science.aaf5098.

[56] Matthew R. Robinson, Aaron Kleinman, Mariaelisa Graff, Anna A.E. Vinkhuyzen, David Couper, Michael B. Miller, Wouter J. Peyrot, Abdel Abdellaoui, Brendan P. Zietsch, Ilja M. Nolte, Jana V. van Vliet-Ostaptchouk, Harold Snieder, The LifeLines Cohort Study, Genetic Investigation of Anthropometric Traits (GIANT) consortium, Sarah E. Medland, Nicholas G. Martin, Patrik K.E. Magnusson, William G. Iacono, Matt McGue, Kari E. North, Jian Yang, Peter M. Visscher, Genetic evidence of assortative mating in humans, Nat. Hum. Behav. 1 (2017), https://doi.org/10.1038/s41562-016-0016 0016.

[57] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S.A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, K. Sirak, C. Gamba, E.R. Jones, B. Llamas, S. Dryomov, J. Pickrell, J.L. Arsuaga, J.M.B. de Castro, E. Carbonell, F. Gerritsen, A. Khokhlov, P. Kuznetsov, M. Lozano, H. Meller, O. Mochalov, V. Moiseyev, M.A.R. Guerra, J. Roodenberg, J.M. Vergès, J. Krause, A. Cooper, K.W. Alt, D. Brown, D. Anthony, C. Lalueza-Fox, W. Haak, R. Pinhasi, D. Reich, Genome-wide patterns of selection in 230 ancient Eurasians, Nature 528 (2015) 499–503, https://doi.org/10.1038/nature16152.

[58] J.T. Troelsen, Adult-type hypolactasia and regulation of lactase expression, Biochim. Biophys. Acta Gen. Subj. 1723 (2005) 19–32, https://doi.org/10.1016/j.bbagen.2005.02.003.

[59] N.S. Enattah, T. Sahi, E. Savilahti, J.D. Terwilliger, L. Peltonen, I. Järvelä, Identification of a variant associated with adult-type hypolactasia, Nat. Genet. 30 (2002) 233–237, https://doi.org/10.1038/ng826.

[60] P.S. Bellwood, First Farmers : The Origins of Agricultural Societies, Blackwell Pub, 2005, https://www.wiley.com/en-us/First+Farmers%3A+The+Origins+of+Agricultural+Societies-p-9780631205661 , Accessed date: 12 July 2018.

[61] N.A. Limdi, M. Wadelius, L. Cavallari, N. Eriksson, D.C. Crawford, M.-T.M. Lee, C.-H. Chen, A. Motsinger-Reif, H. Sagreiya, N. Liu, A.H.B. Wu, B.F. Gage, A. Jorgensen, M. Pirmohamed, J.-G. Shin, G. Suarez-Kurtz, S.E. Kimmel, J.A. Johnson, T.E. Klein, M.J. Wagner, International Warfarin Pharmacogenetics Consortium, Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups, Blood 115 (2010) 3827–3834, https://doi.org/10.1182/blood-2009-12-255992.

[62] H. Takahashi, G.R. Wilkinson, E.A. Nutescu, T. Morita, M.D. Ritchie, M.G. Scordo, V. Pengo, M. Barban, R. Padrini, I. Ieiri, K. Otsubo, T. Kashima, S. Kimura, S. Kijima, H. Echizen, Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans, Pharmacogenet. Genomics 16 (2006) 101–110, https://doi.org/10.1097/01.fpc.0000184955.08453.a8.

[63] R.A. Sturm, D.L. Duffy, Human pigmentation genes under environmental selection, Genome Biol. 13 (2012) 248, https://doi.org/10.1186/gb-2012-13-9-248.

[64] B.L. Lam, S.L. Züchner, J. Dallman, R. Wen, E.C. Alfonso, J.M. Vance, M.A. Peričak-Vance, Mutation K42E in dehydrodolichol diphosphate synthase (DHDDS) causes recessive retinitis pigmentosa, Adv. Exp. Med. Biol. (2014) 165–170, https://doi.org/10.1007/978-1-4614-3209-8_21.

[65] S. Züchner, J. Dallman, R. Wen, G. Beecham, A. Naj, A. Farooq, M.A. Kohli, P.L. Whitehead, W. Hulme, I. Konidari, Y.J.K. Edwards, G. Cai, I. Peter, D. Seo, J.D. Buxbaum, J.L. Haines, S. Blanton, J. Young, E. Alfonso, J.M. Vance, B.L. Lam, M.A. Peričak-Vance, Whole-exome sequencing links a variant in DHDDS to retinitis pigmentosa, Am. J. Hum. Genet. 88 (2011) 201–206, https://doi.org/10.1016/j.ajhg.2011.01.001.

[66] Y. Li, L. Si, Y. Zhai, Y. Hu, Z. Hu, J.-X. Bei, B. Xie, Q. Ren, P. Cao, F. Yang, Q. Song, Z. Bao, H. Zhang, Y. Han, Z. Wang, X. Chen, X. Xia, H. Yan, R. Wang, Y. Zhang, C. Gao, J. Meng, X. Tu, X. Liang, Y. Cui, Y. Liu, X. Wu, Z. Li, H. Wang, Z. Li, B. Hu, M. He, Z. Gao, X. Xu, H. Ji, C. Yu, Y. Sun, B. Xing, X. Yang, H. Zhang, A. Tan, C. Wu, W. Jia, S. Li, Y.-X. Zeng, H. Shen, F. He, Z. Mo, H. Zhang, G. Zhou, Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese, Nat. Commun. 7 (2016) 11664, https://doi.org/10.1038/ncomms11664.

[67] S.J. Chapman, A.V.S. Hill, Human genetic susceptibility to infectious disease, Nat. Rev. Genet. 13 (2012) 175–188, https://doi.org/10.1038/nrg3114.

[68] D.J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data, PLoS Genet. 8 (2012) e1002453,, , https://doi.org/10.1371/journal.pgen.1002453.

[69] A.I. Gogolev, Basic stages of the formation of the Yakut people, Anthropol. Archeol. Eurasia. 31 (1992) 63–69, https://doi.org/10.2753/AAE1061-1959310263.

[70] E. Noskova, V. Ulyantsev, K.P. Koepfli, S.J. O'Brien, P. Dobrynin, GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data, BioRxiv (2019) 407734, https://doi.org/10.1101/407734.

[71] H.J. Muller, Our load of mutations, Am. J. Hum. Genet. 2 (1950) 111–176 http://www.ncbi.nlm.nih.gov/pubmed/14771033 , Accessed date: 10 January 2018.

[72] V.N. Khar'kov, V.A. Stepanov, O.F. Medvedev, M.G. Spiridonova, N.R. Maksimova, A.N. Nogovitsyna, V.P. Puzyrev, The origin of Yakuts: analysis of Y-chromosome haplotypes, Mol. Biol. (Mosk) 42 (2008) 226–237 http://www.ncbi.nlm.nih.gov/pubmed/18610830 , Accessed date: 1 February 2018.

[73] S. Andrews, FastQC, (2010).

[74] G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics 27 (2011) 764–770, https://doi.org/10.1093/bioinformatics/btr011.

[75] E. Starostina, G. Tamazian, P. Dobrynin, S. O'Brien, A. Komissarov, Cookiecutter: A Tool for Kmer-Based Read Filtering and Extraction, BioRxiv, (2015).

[76] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with bowtie 2, Nat.

Methods 9 (2012) 357–359, https://doi.org/10.1038/nmeth.1923.
[77] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[78] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842, https://doi.org/10.1093/bioinformatics/btq033.
[79] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats, Bioinformatics 31 (2015) 2032–2034, https://doi.org/10.1093/bioinformatics/btv098.
[80] A. Morgulis, E.M. Gertz, A.A. Schäffer, R. Agarwala, A fast and symmetric DUST implementation to mask low-complexity DNA sequences, J. Comput. Biol. 13 (2006) 1028–1040, https://doi.org/10.1089/cmb.2006.13.1028.
[81] A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0, n.d. http://www.repeatmasker.org.
[82] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575, https://doi.org/10.1086/519795.
[83] G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K.V. Garimella, D. Altshuler, S. Gabriel, M.A. DePristo, From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, Curr. Protoc. Bioinforma. 43 (2013), https://doi.org/10.1002/0471250953.bi1110s43 11.10. 1–33.
[84] G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Res. 27 (1999) 573–580 http://www.ncbi.nlm.nih.gov/pubmed/9862982 , Accessed date: 5 March 2018.
[85] C. Alkan, J.M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J.O. Kitzman, C. Baker, M. Malig, O. Mutlu, S.C. Sahinalp, R.A. Gibbs, E.E. Eichler, Personalized copy number and segmental duplication maps using next-generation sequencing, Nat. Genet. 41 (2009) 1061–1067, https://doi.org/10.1038/ng.437.
[86] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, M.E. Hurles, Global variation in copy number in the human genome, Nature 444 (2006) 444–454, https://doi.org/10.1038/nature05329.
[87] A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S.R.F. Twigg, W. WGS500 Consortium, A.O.M. Wilkie, G. McVean, G. Lunter, Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications, Nat. Genet. 46 (2014) 912–918, https://doi.org/10.1038/ng.3036.
[88] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience 4 (2015) 7, https://doi.org/10.1186/s13742-015-0047-8.
[89] J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J.E. Huffman, I. Rudan, R. McQuillan, R.M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A.F. Wright, V. Vitart, P. Navarro, J.-F. Zagury, J.F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M.S. Sandhu, J. Marchini, A general approach for haplotype phasing across the full Spectrum of relatedness, PLoS Genet. 10 (2014) e1004234, , https://doi.org/10.1371/journal.pgen.1004234.
[90] J. Macarthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. Mcmahon, A. Milano, J. Morales, Z.M. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog), 45 (2017) 896–901, https://doi.org/10.1093/nar/gkw1133.
[91] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-salomon, W. Rubinstein, D.R. Maglott, ClinVar : public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (2016) 862–868, https://doi.org/10.1093/nar/gkv1222.
[92] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W.W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H.L. Rehm, ACMG Laboratory Quality Assurance Committee, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, Genet. Med. 17 (2015) 405–424, https://doi.org/10.1038/gim.2015.30.
[93] S.S. Kalia, K. Adelman, S.J. Bale, W.K. Chung, C. Eng, J.P. Evans, G.E. Herman, S.B. Hufnagel, T.E. Klein, B.R. Korf, K.D. McKelvey, K.E. Ormond, C.S. Richards, C.N. Vlangos, M. Watson, C.L. Martin, D.T. Miller, Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 Update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics, Genet. Med. 19 (2017) 249–255, https://doi.org/10.1038/gim.2016.190.
[94] X. Zheng, D. Levine, J. Shen, S.M. Gogarten, C. Laurie, B.S. Weir, A high-performance computing toolset for relatedness and principal component analysis of SNP data, Bioinformatics 28 (2012) 3326–3328, https://doi.org/10.1093/bioinformatics/bts606.
[95] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history, Genetics 192 (2012) 1065–1093, https://doi.org/10.1534/genetics.112.145037.
[96] M. Nei, F. Tajima, Y. Tateno, Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data, J. Mol. Evol. 19 (1983) 153–170 http://www.ncbi.nlm.nih.gov/pubmed/6571220 , Accessed date: 10 January 2018.
[97] A.A. Georgiev, Consistent nonparametric multiple regression: the fixed design case, J. Multivar. Anal. 25 (1988), https://ac.els-cdn.com/0047259X88901558/1-s2.0-0047259X88901558-main.pdf?_tid=ecc80f32-f615-11e7-807a-00000aacb360&acdnat=1515596148_a0cba068a606f3fb100ae704efcf4027 , Accessed date: 10 January 2018.
[98] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425, https://doi.org/10.1093/oxfordjournals.molbev.a040454.
[99] D.L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), (2003).
[100] Rambaut, FigTree v. 1.4.0, http://Tree.Bio.Ed.Ac.Uk/Software/Figtree/, (2012).
[101] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263–265, https://doi.org/10.1093/bioinformatics/bth457.