

Genome-wide Sequence Analyses of the Yakut Ethnic Group as a Tool for the Personalized Medicine in the Region

Vladimir L. Osakovskiy^{1,*}, Tatyana M. Sivtseva¹, Mariya M. Okhotina¹, Tatyana K. Davydova² and Raisa N. Zakharova¹

¹North-Eastern Federal University, Research Center of the Medical Institute, 677027 Yakutsk, Republic of Sakha (Yakutia), Russia

²Yakutsk Scientific Center for Complex Medical Problems, Laboratory of Neurodegenerative Diseases, 677018 Yakutsk, Republic of Sakha (Yakutia), Russia

*Corresponding author. Email: iz_labgene@mail.ru

ABSTRACT

The article presents the results of the first genome-wide research that purposefully analysed the Yakut genome. The structural features of the genomic DNA molecule (the number and characteristics of SNPs, Indels, CNV, and SD) and gene variants characteristic of the Yakut ethnic group are shown. The frequency of some medically relevant gene variants in the Yakut population differed from the Russian, world Asian and European populations. The results will be useful for improving the approaches of personalized medicine in the region and contribute to the introduction of genomic methods into clinical practice.

Keywords: *personalized medicine, genome-wide sequence analyses, Yakut ethnic group, Yakut genome, genomic methods*

1. INTRODUCTION

Improving and increasing the availability of genome-wide sequencing methods contributes to the development of personalized medicine in various regions, taking into account ethnic and population characteristics. In 2016, the project “Genome Russia” was initiated with the goal to creating a database of whole genome sequences of at least 3 thousand people from different ethnic and regional populations across Russia. One of the first the samples of representatives of the Yakut ethnic group were analyzed. Yakuts – the Asian population inhabiting a vast territory with

extreme climatic conditions in the northeast of Russia. The work was performed in the Theodosius Dobzhansky Center for Genomic Bioinformatics, St. Petersburg State University with the support by the Russian Federation megagrant (Genome Russia Grant No. 1/52/1647.2016), Russian Science Foundation (grant No. 17-14-01138), and by the Ministry of Education and Science of Russian Federation (projects No. 17.6344.2017 / 8.9 (2016–2019) и FSRG-2020-0016 (2020–2022)) for employees from the North-Eastern Federal University [1].

The work considers the most important structural changes in genomic DNA: SNP, Indels, CNV and SD,

the most common in which are single nucleotide polymorphisms (SNPs) [2]. The next common type is insertion or deletion of nucleotides in the genomic DNA (Indels). Other major types of structural changes in the DNA molecule are copy-number variants (CNV) and segmental duplication (SD). They lead to a lengthening of the DNA molecule and represent a source of genetic variations that affect gene dosage and penetrance (an indicator of the phenotypic manifestation of a gene) [3]. These structural changes in the genome provide individual genetic differences in humans and are one of the mechanisms of the adaptation process and the evolution of organisms. The proportion of pathological SNP, Indels and SD in the human genomes is small, but in different populations, they can vary depending on the ethnic group and geography of residence. Large and rare chromosome Indels and SD are involved in impaired development of the nervous system, in particular, are associated with intellectual disorders [4, 5].

This article presents the main results of the analysis of the whole genome sequences of representatives of the Yakut ethnic group obtained in the project "Genome Russia". The main medically relevant gene variants have been identified, the frequency of carriage of which in a heterozygous state, in the Yakut population, according to the data obtained, differs from the Russian population and other world populations. The identified variants require further study in terms of clinical manifestations and use in the development of personalized medicine methods in the region.

2. MATERIALS AND METHODS

We sampled family trios (two biological parents and their full aged children) from ethnic Yakuts with confirmation of belonging to the ethnos in the previous three generations. Using the "family trio" scheme facilitates more accurate analyzing of SNP data and identification of the haplotype structure. Samples were

obtained from 4 family trios and 6 unrelated individuals, total 18 samples. The research protocol and informed consent documents were approved by the Ethics Committee of St. Petersburg State University (No. 65/2015).

DNA was isolated from blood samples using a MagCore HF16 automated nucleic acid extractor (RBC Bioscience).

Whole genome sequencing of DNA samples and bioinformatic processing of data were carried out at the Theodosius Dobzhansky Center for Genomic Bioinformatics and the Biobank Resource Center Research Park of St. Petersburg State University. The methods used in detail are published in the journal *Genomics* in 2020 [1].

3. RESULTS AND DISCUSSION

3.1 Structural Features of the Genomic DNA Molecule of the Yakut Ethnos Revealed by Whole Genome Sequencing

Analysis of the sequence of the whole genome made it possible to quantify SNPs, indels, as well as CNV and SD in 18 samples of genomic DNA from representatives of the Yakut ethnic group. In this article, data are presented in comparison with representatives of two northern Russian populations from Pskov (22 samples) and Novgorod (20 samples).

Analysis revealed about 8 million SNPs and 2 million indels per population. 3–4 % of these SNPs were considered as novel as compared to dbSNP. After filtering the raw data, in the Yakut samples, 5,916,168 known and 227,138 new SNPs, 74,290 known and 421,237 new indels were revealed (Fig. 1).

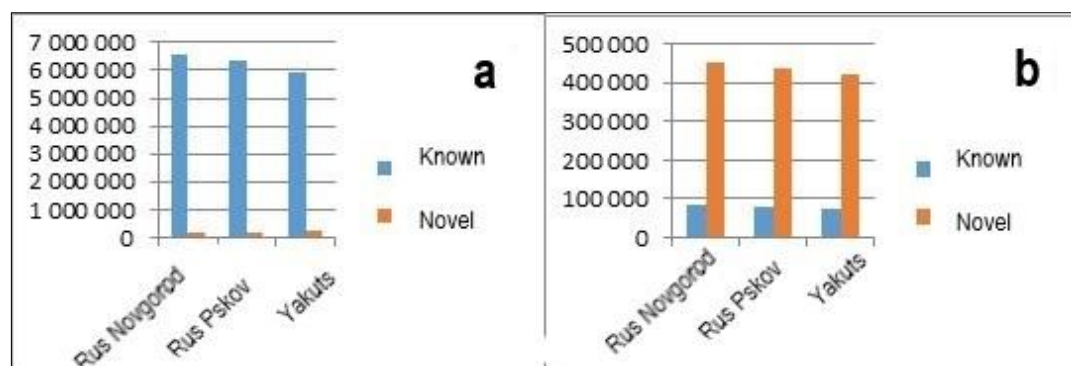


Figure 1. The number of known and novel SNPs (a) and Indels (b) obtained after filtration in three populations

Moreover, the number of overlapping SNPs and indels in Yakut samples is lower compared to the Pskov and Novgorod data. The same trend was observed for

long indels and indicates the unique composition of indels in each genome.

According to the results of CNV and SD studies, segments of unique duplication regions with a length of 1.9 million bases in Novgorod, 1.6 million bases in Pskov and up to 3.0 million bases in Yakut DNA samples were identified. For example, duplications on the 10th chromosome are associated with the gene of a protein rich in leucine, which is involved in the differentiation of melanocyte cells. A mutation of this gene is associated with autosomal ocular albinism. Samples of Yakut representatives have longer SD segments and, accordingly, a larger number of presented genes (157 genes in duplicated regions of the DNA molecule). The enrichment of serine proteinase genes in these regions was revealed. The broad substrate specificity of serine proteases allows them to participate in important physiological processes, such as activation of the cascade of blood coagulation proteins, destruction of fibrin, activation of complement system proteins, and the formation of protein hormones. This may be one of the adaptive benefits of the body.

Assessment of the average length of extended regions of SNP homozygosity showed longer homozygous stretches in the Yakut genomes (median

length=127kbp) than Pskov and Novgorod samples (median length = 119kbp), which indicates the evidence of a recent founder event or population contraction. This event leads to a shift in the genetic structure of the gene pool of the ancient population to a structure poorer in the variety of genotypes, however, providing a better adaptation of its carriers to survival in existing natural conditions.

3.2 Medically relevant gene variants in Yakut genomes

A total 894 medically relevant gene variants, annotated in the Human Gene Mutation Database (HGMD) [6] as disease-causing mutations, were detected within the groups of healthy Yakut and the Russian Pskov, Novgorod representatives. On average, 75 disease-causing mutations were profiled per person. The frequency of some variants in the Yakut representatives was statistically significantly different from the Russian, as well as from the world European and Asian populations (table 1). The identified pathological variants were heterozygous and not clinically manifested in the studied individuals.

Table 1. Frequency of minor alleles of disease -associated variants in the studied samples of the Yakut population

Gene	Variant ID	MAF Yakuts	MAF 1000G EAS	MAF 1000G EUR	Min <i>p</i> -value
<i>OCA2</i>	rs74653330	0.214	0.027	0.01	1.49E-04
<i>SBF1</i>	rs200488568	0.107	0.001	0	6.96E-05
<i>WDR33</i>	rs117753184	0.179	0.026	0	0.001
<i>TBC1D31</i>	rs10101626	0.714	0.183	0.195	2.41E-09
<i>ABCA13</i>	rs141576983	0.464	0.023	0.002	2.67E-13
<i>HLA-DPB1</i>	rs9277535	0.39	0.61	0.27	0.029
<i>STIM1</i>	rs11030122	0.11	0.35	0.33	0.007
<i>IL28B</i>	rs8099917	0,04	0,08	0,17	>0.05

Note. MAF – minor allele frequency, 1000G EAS – frequency in East Asian populations according to 1000 Genomes project, 1000G EUR – frequency in European populations according to the 1000 Genomes project. Min *p*-value – minimum *p*-value for Fisher exact test of allele count difference between Yakut compared with 1000G EAS.

Type II albinism oculocutaneous associated with the *OCA2* gene, rs74653330 (C/T). The *OCA2* gene encodes a membrane protein that is involved in the transport of tyrosine to the melanosome, where tyrosine is used to synthesize and accumulate the melanin pigment [7]. Mutations in this gene are associated with eye color. Some genetic variants of the *OCA2* gene, one of which the T allele (rs74653330), can lead to autosomal recessive type II albinism oculocutaneous. Among the Yakut population, a high frequency of occurrence of this allele in heterozygous form was revealed (table 1).

Neurodegenerative disease Charcot-Marie-Tooth of type 4B3 (CMT4B) associated with the *SBF1* gene, rs200488568 (T/C). The disease is a severe demyelinating peripheral neuropathy, characterized by a slowed rate of nerve conduction, axon loss, and a

characteristic violation of myelin formation. One of the genes associated with this pathology is *SBF1*, rs200488568 (T/C) [8]. The C allele of the *SBF1* gene is the main cause of the new subtype of CMT disease – CMT4B3. The product of the *SBF1* gene is a protein related to myotubularin – a phosphatase enzyme. The product of myotubularins enzymatic activity regulates the intracellular endosomal-lysosomal membrane transport in the axon. Mutations disrupt this transport process and induce demyelinating peripheral neuropathy. In the Yakut population, the frequency of distribution of heterozygous carriers of the recessive C allele is quite high (Table 1).

Coronary artery calcification (CCA) associated with the *WDR33* gene, rs117753184 (A/T). Calcification of coronary arteries is considered as a factor complicating cardiovascular diseases. Genetic variants associated with CCA have been identified: SNP rs1333049 on

chromosome 9p21 and SNP rs9349379 in the PHACTR1 gene on chromosome 6p24. There is evidence of SNP associations with CCA and myocardial infarction at several other loci, including 3q22 (*MRAS* gene), 13q34 (*COL4A1* / *COL4A2* genes) and 1p13 (*SORT1* gene). The molecular nature of the action of these genes on the calcification of the coronary artery has not been disclosed. The *WDR33* gene (A / T), known as the gene modifying messenger RNA before translation, is also associated with CCA [9]. In Yakutia, the frequency of distribution of the recessive allele of the gene is increased (table 1).

Diabetic kidney disease (uromodulin level in urine) associated with the *TBC1D31* gene, rs10101626 (G / T). The product of the *TBC1D31* gene is a protein with an unknown molecular function. Defects in this gene are associated with autosomal dominant renal impairment, medullary cystic kidney disease-2 (MCKD2), and familial juvenile hyperuricemic nephropathy (FJHN).

Astigmatism associated with the *ABCA13* gene, rs141576983 (G / T). The *ABCA13* gene is the largest active transport protein of the cell and intracellular membrane [10]. Normally, the carrier pumps out cholesterol from the cell, which in the blood plasma binds to the Apo1 protein with the formation of high density lipids (HDL). Mutated forms of the protein (T allele of the *ABCA13* gene) reduce the pumping of cholesterol from the cell, contributing to its accumulation in the cells

and an increase in the size of the tissue in which this process proceeds. Deformation changes in the tissues of the visual organ associated with impaired lipid metabolism can cause astigmatism. The frequency of distribution of this recessive allele among the population of Yakutia is also increased (table 1).

3.3 Identification of recessive alleles of genes associated with infectious diseases

The predisposition to infectious diseases is determined by the genetic component and the influence of environmental factors. In the work, the frequency of alleles of marker genes associated with the phenotypes of infectious diseases and increased susceptibility or resistance to specific infectious disease, revealed in genome-wide studies (GWAS) and single-gene variants studies, was analyzed [11]. The analysis was performed on genes associated with 18 infections. Comparative studies revealed similarities and differences in the distribution of these polymorphisms of the European (EUR), East Asian (EAS) and South Asian (SAS) populations (Figure 2.1). For example, the *STIM1* gene variant (rs11030122), associated with T-cell deficiency in classic Kaposi's sarcoma, in the Yakut population has a lower frequency than in European and Asian populations (Table 1).

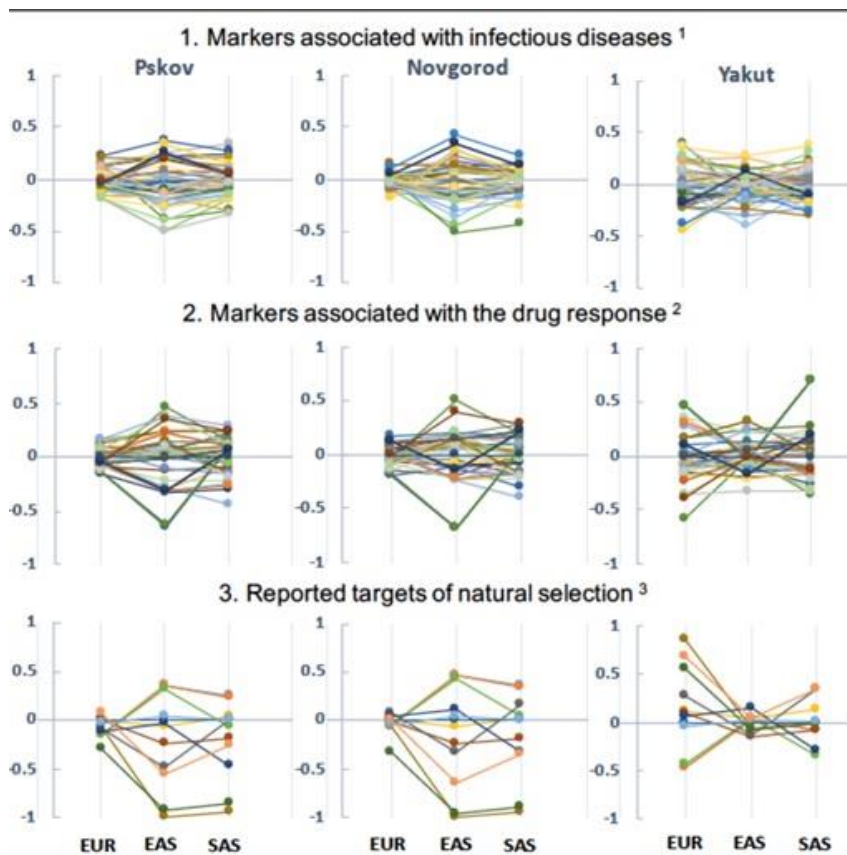


Figure 2. Allele frequency differences of Russian populations versus modern Eurasian populations from 1000G using functional gene variants from infectious disease, pharmacogenomics and selected genes

In the Yakut population, genes, associated with viral hepatitis B and C are of interest, because Yakutia is recognized as an endemic region with a high prevalence of these infections. The variant of the G allele (rs9277535) of the *HLA-DPBI* (A / G) gene, which is involved in the development of chronic hepatitis B, has a lower frequency in European populations, including Russians, unlike the Yakut one, where the frequency is quite high (table 1). Cells with the G allele of the *HLA-DPBI* gene, infected with hepatitis B, weakly present the viral antigen to immune cells and, therefore, carriers of this allele have an increased risk of maintaining the active virus in the liver cells. For hepatitis C: some SNPs localized in the genomic DNA molecule near the *IL28B* gene can have a positive effect of treatment with hepatitis C virus. This SNP effect is associated with an enhanced immune response to the IFN λ – the product of the *IL28B* gene and its secretion by blood monocytes on infected hepatocytes. Earlier studies were conducted in the Yakut population confirming the connection of this polymorphism with the clinical course of chronic hepatitis C [12]. The frequency of the minor allele of SNP rs8099917 (T / G) of the *IL28B* gene, associated with the low effectiveness of the treatment of hepatitis C, in healthy Yakut representatives was lower than in European populations, but did not differ from Asian populations (table 1). The reasons for the increased level of severe outcomes in the Yakut population require further clinical and genetic studies.

Although the frequencies of most genetic markers in Yakuts were often close to the immediate neighbors population of East Asia, the frequencies of alleles in genes significantly associated by GWAS with infectious disease phenotypes are (somewhat surprising) more similar to that of South Asian (SAS) populations, while the single-gene variants associated with increased susceptibility or resistance to specific infectious disease phenotypes do not appear similar to any reference group (EUR, EAS or SAS).

The data obtained in this project can also be used to study the frequencies of pharmacogenomic biomarkers and marker genes associated with loci of quantitative traits and belonging to the category of polygenic adaptation of complex anthropometric traits (Figure 2.2, 2.3) [1].

In general, the Yakut population demonstrates a significant deviation from all database populations (EUR, EAS and SAS) for genes associated with pharmacogenomics and infectious diseases, but a more closely clustering with EAS for genes associated with natural selection (Fig. 2). This is the result of adaptive selection of a cluster of genes that ensure the survival of

the organism to the extreme climatic conditions of Yakutia.

4. CONCLUSION

Here we presented the results of the first genome-wide studies that purposefully analyzed the Yakut genome. The structural features of the genomic DNA and genetic variants characteristic of the Yakut ethnic group are shown. The results will be useful for improving the approaches of personalized medicine in the region and will contribute to the implementation of genomic methods in clinical practice [13].

The main characteristics of the Yakut genome:

1. In the Yakut genomes after filtration of raw data, about 6 million SNPs were identified, of which 3.8% were new and 495.5 thousand insertions and deletions, most of which were not previously described. Yakut DNA molecules are characterized by fewer SNPs and indels overlapping with the European set.
2. The genomic DNA of the Yakut ethnic group, unlike the European ones, is enriched in long indels and SD.
3. Genomes of representatives of the Yakut ethnic group show moderate genetic homogeneity.
4. In the genomes of Yakut healthy carriers, a number of recessive alleles of genes associated with diseases or impaired functions were identified. A high frequency of carriage of such gene variants as *OCA2*, rs74653330, *SBF1*, rs200488568, *WDR33*, rs117753184, *TBC1D31*, rs10101626, *ABCA13*, rs141576983 was revealed and their clinical manifestations should be confirmed.
5. Comparison of the frequencies of the minor allele in genes associated with infectious diseases and pharmacogenomics showed significant differences between the Yakut, European, East Asian and South Asian world populations. In Yakuts the frequencies of alleles in genes significantly associated by GWAS with infectious disease phenotypes are more similar to that of South Asian (SAS) populations, while the single-gene variants associated with increased susceptibility or resistance to specific infectious disease phenotypes do not appear similar to any of the EUR, EAS or SAS groups.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Project FSRG-2020-0016 "Wide-genome studies of the gene

pool of the indigenous population of the Arctic coast of Yakutia" 2020–2022 years).

REFERENCES

- [1] D.V. Zhernakova, V. Brukhin, S. Malov et al., *Genomics* 112 (2020) 442–458.
- [2] B.S. Shastry, *J. of Hum. Genet.* 52(11) (2007) 871–880.
- [3] J.S. Beckmann, X. Estivill, S.E. Antonarakis, *Nat. Rev. Genet.* 8 (2007) 639–646.
- [4] A.K. MacLeod, G. Davies, A. Payton et al., *Plos one* 7(12) (2012) e37385.
- [5] N.M. Williams, I. Zaharieva, A. Martin et al., *Lancet* 376 (2010) 1401–1408.
- [6] P.D. Stenson, E.V. Ball, M. Mort et al., *Hum. Mutat.* 21 (2003) 577–581.
- [7] S.T. Lee, R.D. Nicholls, S. Bunday et al., *N. Engl. J. Med.* 330(8) (1994) 529–534.
- [8] K. Nakhro, J.M. Park, Y.B. Hong et al., *Neurol.* 81(2) (2013) 165–173.
- [9] C.J. O'Donnell, M. Kavousi, A.V. Smith et al., *Circulat.* 124(25) (2011) 2855–2864.
- [10] E.K. Berge, H. Tian, G.A. Graf et al., *Sci.* 290 (2000) 1771–1775.
- [11] S.J. Chapman, A.V.S. Hill, *Nat. Rev. Genet.* 13 (2012) 175–188.
- [12] S.I. Semenov, A.I. Fedorov, V.L. Osakovsky et al., *J. of Microbiol., Epid. and Imm.* 2 (2017) 86–92.
- [13] S.A. Kostyuk, *Med. news* 4 (2016) 11–14.