

Diversity of genomic variants and population genetics of ethnic and regional groups across Russia

D.V. Zhernakova^{1,2}, V. Brukhin¹, S. Malov^{1,3}, T.K. Oleksy^{1,4}, K.P. Koepfli^{1,5}, A. Zhuk¹, P. Dobrynin¹, S. Kliver¹, N. Cherkasov¹, G. Tamazian¹, M. Rotkevich¹, K. Krashenninnikova¹, I. Evsyukov¹, S. Sidorov¹, A. Gorbunova^{1,6}, E. Chernyaeva¹, A. Shevchenko¹, S. Kolchanova^{1,4}, A. Komissarov¹, S. Simonov¹, A. Antonik¹, A. Logachev¹, D.E. Polev⁷, A.S. Glotov⁷, V. Ulantsev⁸, E. Noskova^{8,9}, T.K. Davydova¹⁰, T.M. Sivtseva¹⁰, S. Limborska¹¹, O. Balanovsky^{12,13,14}, V. Osakovsky¹⁰, A. Novozhilov¹⁵, V. Puzyrev¹⁶, N. Kropachev¹⁷, S.J. O'Brien^{1*}

¹ *Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia*

² *University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands*

³ *Department of Mathematics, St. Petersburg Electrotechnical University, St. Petersburg, Russia*

⁴ *Biology Department, University of Puerto Rico at Mayaguez, Mayaguez, Puerto Rico*

⁵ *Smithsonian Conservation Biology Institute, National Zoological Park, Washington, USA*

⁶ *I.I. Mechnikov North-Western State Medical University, St. Petersburg, Russia*

⁷ *Research Resource Center for Molecular and Cell Technologies, Research Park, St. Petersburg State University, St. Petersburg, Russia*

⁸ *Computer Technology Department, St. Petersburg Electrotechnical University, St. Petersburg, Russia*

⁹ *Jet Brains Research, St. Petersburg, Russia*

¹⁰ *Institute of Health, North-Eastern Federal University, Yakutsk, Russia*

¹¹ *Department of Molecular Bases of Human Genetics, Institute of Molecular Genetics, RAS, Moscow, Russia*

¹² *Vavilov Institute of General Genetics, RAS, Moscow, Russia*

¹³ *Research Centre for Medical Genetics, Moscow, Russia*

¹⁴ *Biobank of North Eurasia, Moscow, Russia*

¹⁵ *Department of Ethnography and Anthropology, St. Petersburg State University, St. Petersburg, Russia*

¹⁶ *Institute of Medical Genetics, Tomsk Research Center RAS, Tomsk, Russia*

¹⁷ *Office of the Rector, St. Petersburg State University, St. Petersburg, Russia*

* *e-mail: lgdchief@gmail.com*

Key words: genome-wide variation, Russian Federation, disease-causing mutations, gene flow, geological barriers

The Russian Federation spans 11 time zones and is the home of ~146,000,000 people: 80 % are the ethnic Russians and the remainder identify themselves as one of ~200 indigenous ethnic minorities. Despite the large population size and high ethnic diversity, no centralized reference database of functional and endemic genetic variation has been established to date. Such data are crucial for medical genetic purposes and would be essential for studying population history. The Genome Russia Project aims at filling this gap by performing high coverage whole genome sequencing and analysis of peoples of the Russian Federation. Here we report general methodology and inferences of genome-wide variation (SNPs, indels, and copy number variation) from 264 healthy adults, including 60 newly sequenced samples consisting of family trios from three geographic regions: Pskov, Novgorod and Yakutia, now contributed to the 1000 Genomes database. People of Russia are shown to carry known and novel genetic variants of adaptive, clinical and functional consequence. We identified 31 SNPs associated with disease-causing mutations from Human Gene Mutation Database; 758 loss-of-function SNPs and more than 20 SNPs associates with diseases, drug response and other phenotypes showing appreciable occurrence or allele frequency divergence from the neighboring Eurasian populations. Principal component and phylogenetic analyses of overall variation revealed strong geographic partitions among indigenous ethnicities corresponding to the geographic locales where they have lived. Allele frequency spectra identified strong constraints to gene flow that correspond to the geological barriers (e. g. the Ural Mountains and Verkhoyansk mountain range). The first haplotype-based genetic maps have been generated for western ethnic Russian and Yakut populations that can be used for identification of functional gene variants associated with diseases, and as a tool for the population genetic analyses. This study presents a genomic characterization of population-specific variation in Russia with results important for medical genetics as well as for population natural history studies.