ISSN 1607-6729, Doklady Biochemistry and Biophysics, 2017, Vol. 474, pp. 213–216. © Pleiades Publishing, Ltd., 2017.
Original Russian Text © A.S. Zlobin, S.Sh. Sharapov, V.P. Guryev, M.R. Bevova, Y.A. Tsepilov, T.M. Sivtseva, U.A. Boyarskih, E.A. Sokolova, Y.S. Aulchenko, M.L. Filipenko, V.L. Osakovsky, 2017, published in Doklady Akademii Nauk, 2017, Vol. 474, No. 4, pp. 505–509.

Population Specific Analysis of Yakut Exomes

A. S. Zlobin[a, b] *, S. Sh. Sharapov[a, b], V. P. Guryev[c], M. R. Bevova[c], Y. A. Tsepilov[a,b], T. M. Sivtseva[d], U. A. Boyarskih[a,e], E. A. Sokolova[a, e], Y. S. Aulchenko[a, b], M. L. Filipenko[a, e], and V. L. Osakovsky[d]

Presented by Academician V.K. Shumnyi October 18, 2016

Received November 23, 2016

Abstract—We studied the genetic diversity of the Yakut population using exome sequencing. We performed comparative analysis of the Yakut population and the populations that are included in the "1000 Genomes" project and we identified the alleles specific to the Yakut population. We showed, that the Yakuts population is a separate cluster between Europeans and East Asians.

DOI: 10.1134/S1607672917030188

Yakuts are a genetically isolated population, the majority of representatives of which reside on the territory of the Republic of Sakha (Yakutia). Yakuts are the largest indigenous people of Siberia on the territory of the Russian Federation according to the results of the national census of 2010. Previously, the nucleotide sequences of the mitochondrial DNA and Y chromosome have been studied to determine the genetic diversity in the Yakut population, and over 500 000 autosomal single nucleotide polymorphism (SNP) variants were genotyped using microarrays [1, 2]. The results of analysis of autosomal SNPs confirmed the genetic similarity of Yakuts and peoples of Southern Siberia. However, the data obtained using the microarray technique provide information only on the previously known polymorphic variants, whereas the next generation sequencing provide information on the previously unknown variants, including those that are unique to this population. Earlier, the sequence of whole genome of representatives of the Yakut population was used only to identification of copy number variations [3].

[a] *Novosibirsk State University, Novosibirsk, 630090 Russia*
[b] *Institute of Cytology and Genetics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*
[c] *European Research Institute for the Biology of Ageing, University Medical Center Groningen, Groningen, The Netherlands*
[d] *Research Institute of Health, Ammosov North-Eastern Federal University, Yakutsk, Russia*
[e] *Institute of Chemical Biology and Fundamental Medicine, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*
**e-mail: defrag12@gmail.com*

In the framework of this study, we analyzed the exome sequences of 12 representatives of the Yakut ethnic group. We identified novel genetic variants and their heterozygosity and analyzed the types of substitutions. We also compared the genetic structure of the Yakut population and the populations that are included in the "1000 Genomes" project.

Blood samples for sequencing were collected by the personnel of the Genetic Research Laboratory of the Research Institute of Health, North-Eastern Federal University. Blood samples were obtained from the representatives of the Yakut ethnic group living in the Vilyuiskii District of the Republic of Sakha (Yakutia). A total of 12 subjects (7 men and 5 women) were selected for further analysis. Seven individuals in the sample suffered from the Vilyui encephalomyelitis. The average age of the sample was 47 years. Sample preparation and sequencing were performed in the Laboratory of Pharmacogenomics of the Institute of Chemical Biology and Fundamental Medicine, Russian Academy of Sciences, and in the Beijing Genomics Institute (China) using the Illumina HiSeq2000 system (Illumina Inc., United States) with the generation of the primary nucleotide sequences in the FASTQ format. Samples 1–7 and 8–12 were sequenced in different time by using two enrichment sets (Agilent SureSelect Human All Exon V6 and V5, respectively; Agilent Technology, United States). In the first step, reads were aligned to the human genome reference sequence (assembly GRCh37) using the Bowtie2 software [4]. The aligned sequences were sorted and converted using the SAMtools format [5]. SNP variants (insertions/deletions) in the data obtained were distinguished using GATK [6].

Data were visualized by the principal component analysis using the KING software [7]. The homozygosity segments (also known as runs of homozygosity,

Table 1. Overall substitution statistics for 12 samples

Polymorphism	Total number			Not represented in the dbSNP database			Located in exome regions		
Sample	homozygous	heterozygous	total number	homozygous	heterozygous	total number	homozygous	heterozygous	total number
1	78099	165242	243341	3693	10606	14299	10358	5771	16129
2	75352	131445	206797	2908	10797	13705	10213	5897	16110
3	78944	251961	330905	6428	10459	16887	9869	6073	15942
4	76072	152241	228313	3478	10806	14284	10331	5920	16251
5	73150	126508	199568	2735	10729	13464	10215	6003	16218
6	73083	126893	199976	2796	10232	13028	10246	5867	16113
7	78814	134831	213645	2603	11499	14102	10220	5922	16142
8	42293	179316	221609	5017	4501	9518	8905	5820	14725
9	40505	118841	159346	2139	3890	6029	8658	5846	14504
10	41093	120563	161656	2597	4069	6666	8963	5769	14732
11	44581	97270	141851	1779	4141	5920	9597	5351	14948
12	39309	113562	152871	2262	3939	6201	8591	5948	14539
Total			746396			56949			35409
Mean	61775	143223	204990	3203	7972		9681	5849	

The total number of substitutions as well as the number of SNPs in homozygous and heterozygous states are shown. Counts were performed for all detected substitutions, for the first found substitutions (not represented in the dbSNP database), and for the substitutions located in the exome sequences.

ROH) were determined using the PLINK software with the default settings (no more than 5 missed genotypes and not more than one heterozygote per segment).

As a result, we found 746396 substitutions (SNPs and short insertions/deletions) located on the autosomes, which have passed the quality control (the probability of error was less than 1 per 10000 substitutions; the number of reads at a given position in at least one sample was greater than or equal to 15).

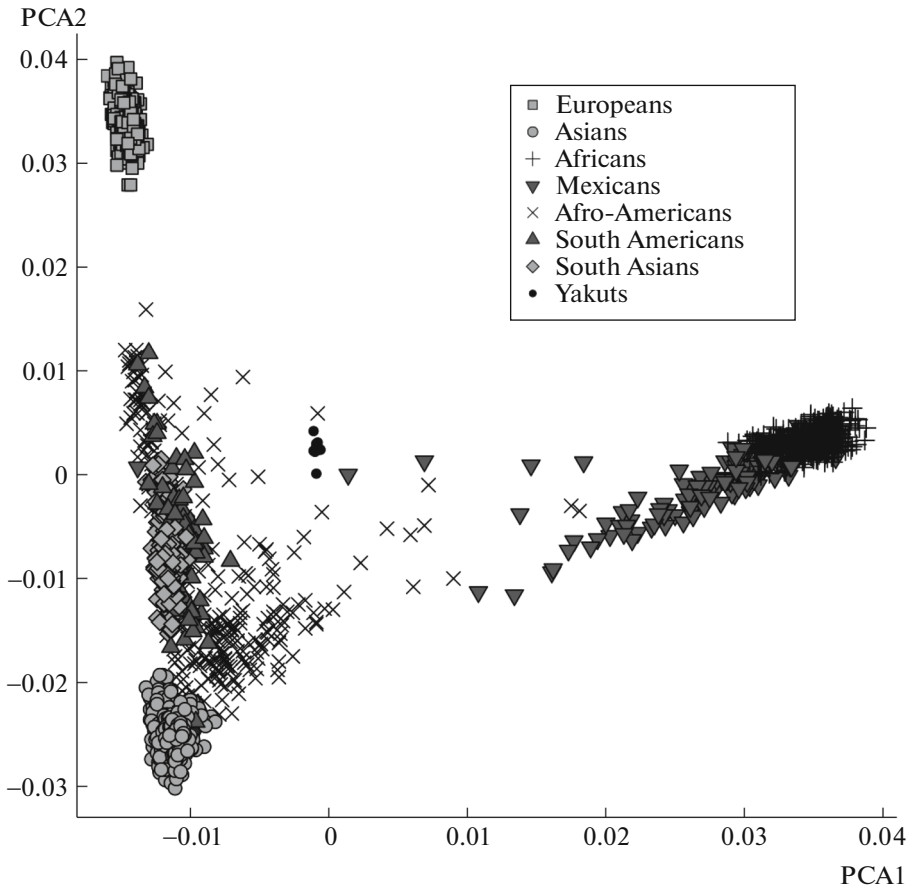
Table 2. Classification and number of LoF variants

Sample	LoF variants			Stop codons	ORF shift
	homozygous	heterozygous	total number		
1	199	430	629	94	268
2	196	422	618	114	252
3	188	436	624	104	254
4	216	439	655	112	276
5	212	442	654	113	280
6	217	412	629	107	280
7	201	443	644	110	278
8	109	216	325	65	140
9	103	206	309	58	122
10	120	201	321	60	130
11	89	218	307	53	116
12	105	187	292	57	124
Total			1364	299	617

At the next step, we determined the number of homozygous and heterozygous polymorphic variants for each sample (Table 1). Of that number, 56949 variants (7.6%) were absent in the dbSNP database. Of these, the average number of homozygous and heterozygous alleles per sample was 3233 and 8097, respectively. Among the new variants, we identified 4.1% synonymous substitutions (2335), 2.4% (1367) mutations at splicing sites, 26.3% (14978) nonsense substitutions, and 9.6% (5467) missense substitutions. Table 2 shows the number of alleles resulting in the loss of gene function (the so-called Loss of function variants, LoF).

Next, we analyzed the number of ROH and, as a result, found 153 ROH 1008 to 8029 bp long. Eight of 12 samples had long homozygosity segments (5000 kb). Such segments are usually observed in genetically isolated populations [8].

For a comparative analysis, we used only the autosomal SNPs. The transition-to-transversion ratio was found to be 2.05 (437613 transitions and 213720 transversions), which is consistent with to the published data [9]. For further analysis we used the SNPs containing at least 15 reads at a given position (at least in 8 out of 12 samples). Of the 98879 SNPs, only those that are in the exome regions (exome sequencing includes



Results of comparative analysis of Yakut and other populations from the "1000 Genomes" project plotted on the basis of two first principal components.

the reads of not only exome but also flanking regions) were taken for analysis. As a result, 5988 SNPs were used in comparative populational genetic analysis (figure). The figure shows the location of the Yakut population relative to other human populations.

In our study, we obtained 746396 polymorphism variants, which were quality controlled and annotated, in 12 representatives of the Yakut ethnic group. Among all identified variants 56949 SNPs were absent in the dbSNP database. The average number of homozygotes for the alternative allele in one person was 61 775. The analysis of ROH has revealed two longest homozygosity regions, which might be due to endogamicity and provided additional evidence for the existence of "bottleneck" in the history of Yakuts population [10].

Comparative analysis of 12 exomes showed that the Yakut population is a sufficiently isolated cluster that is grouped with the clusters representing South America and located between the Asian and European populations. This result agrees with the previous data obtained as a result of analysis of 500000 autosomal variants in Yakuts.

ACKNOWLEDGMENTS

Part of the study was performed under the State task of the Ministry of Education and Science of the Russian Federation no. 2014/257 (project no. 3095).

REFERENCES

1. Fedorova, S.A., *Geneticheskie portrety narodov Respubliki Sakha, Yakutiya: analiz linii mitokhondrial'noi DNK i Y-khromosomy* (Genetic Portraits of People of the Republic of Sakha, Yakutia: Analysis of Lines of Mitochondrial DNA and Y-Chromosome), YaNTs SO RAN, 2008.
2. Fedorova, S.A., Reidla, M., Metspalu, E., Metspalu, M., Rootsi, S., Tambets, K., Trofimova, N., Zhadanov, S.I., Hooshiar Kashani, B., Olivieri, A., Voevoda, M.I., Osipova, L.P., Platonov, F.A., Tom-sky, M.I., Khusnutdinova, E.K., Torroni, A., and Villems, R., Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia, *BMC Evol. Biol.,* 2013, vol. 13, no. 1, p. 127.
3. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., Jorde, L.B., Posukh, O.L.,

Sahakyan, H., Watkins, W.S., Yepiskoposyan, L., Abdullah, M.S., Bravi, C.M., Capelli, C., Hervig, T., Wee, J.T.S., Tyler-Smith, C., van Driem, G., Romero, I.G., Jha, A.R., Karachanak-Yankova, S., Toncheva, D., Comas, D., Henn, B., Kivisild, T., Ruiz-Linares, A., Sajantila, A., Metspalu, E., Parik, J., Villems, R., Starikovskaya, E.B., Ayodo, G., Beall, C.M., Di Rienzo, A., Hammer, M., Khusainova, R., Khusnutdinova, E., Klitz, W., Winkler, C., Labuda, D., Metspalu, M., Tishkoff, S.A., Dryomov, S., Sukernik, R., Patterson, N., Reich, D., and Eichler. E.E., Global diversity, population stratification, and selection of human copy number variation, *Science*, 2015, vol. 349, no. 6253.

4. Langmead, B. and Salzberg, S.L., Fast gapped-read alignment with Bowtie 2, *Nat. Methods*, 2012, vol. 9, no. 4, pp. 357–359.
5. Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y.M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., Retterer, J., Rodriguez, A., Rogers, S., Salazar, A., Tsai, M., and Myers, R.M., Quality assessment of the human genome sequence, *Nature*, 2004, vol. 429, no. 6990, pp. 365–368.
6. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J., A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet.,* Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved, 2011, vol. 43, no. 5, pp. 491–498.
7. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen. W.M., Robust relationship inference in genome-wide association studies, *Bioinformatics*, 2010, vol. 26, no. 22, pp. 2867–2873.
8. McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A.K., Farrington, S.M., Ru-dan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S.H., Dunlop, M.G., Wright, A.F., Campbell, H., and Wilson, J.F., Runs of homozygosity in European populations, *Am. J. Hum. Genet. Elsevier*, 2008, vol. 83, no. 3, pp. 359–372.
9. Stoltzfus, A. and Norris, R.W., On the causes of evolutionary transition: transversion bias, *Mol. Biol. Evol., Oxford University Press*, 2016, vol. 33, no. 3, pp. 595–602.
10. The dating of the first phase of ethnogenesis of the Yakuts and the beginning of the colonization of the territory of Yakutia [Electronic resource]. http://new.chronologia.org/volume10/turin_dat_jakuty.php#52 (accessed: 10.05.2015).

Translated by M. Batrukova