BioNanoSci.
DOI 10.1007/s12668-016-0282-1

Lessons from the Whole Exome Sequencing Effort in Populations of Russia and Tajikistan

Eugenia A. Boulygina[1] · **Elena Lukianova**[2] · **Tatyana V. Grigoryeva**[1] · **Maria N. Siniagina**[1] · **Sergey Yu Malanin**[1] · **Elena V. Balanovska**[3] · **Oleg P. Balanovsky**[2,3] · **Vladislav M. Chernov**[1,4]

Abstract In contrast with the traditional methods applied to assessment of population diversity, high-throughput sequencing technologies have a wider application in clinical practice with greater potential to find novel disease-causing variants for multifactorial disorders. Widely used test panels may not meet their goal to diagnose the patient's condition with a full reliability since this method often does not take into account the population frequencies of analyzed genetic markers. Here, we analyzed 57 male individuals of five ethnic groups from Russia and Tajikistan using the whole exome sequencing technique (Ion AmpliSeq Exome), which resulted in detecting more than 299,000 single nucleotide polymorphisms. Samples formed clusters on the PCA plot according to the geographical location of the corresponding populations. Thereby, the methodology of whole-exome sequencing, in general, and the Ion AmpliSeq Exome panel, in particular, could be positively applied for the purposes of population genetics and for detection of the novel clinically relevant variants.

✉ Eugenia A. Boulygina
boulygina@gmail.com

[1] Kazan Federal University, Kremlyovskaya 18, Kazan 420008, Russia

[2] Vavilov Institute of General Genetics, Gubkina 3, Moscow 119333, Russia

[3] Research Centre for Medical Genetics, Moskvorechie 1, Moscow 115478, Russia

[4] Kazan Institute of Biochemistry and Biophysics, Lobachevsky 2/31, Kazan 420111, Russia

Keywords Whole exome sequencing · Diagnostic panel · Ion AmpliSeq Exome · SNP · North Eurasian populations

1 Introduction

Historically, the classical method to discover the pathogenic genome changes is based on the Sanger sequencing of the target gene, which carries a mutation. Constantly updated knowledge bases of multifactorial disorders help to construct a wide spectrum of the diagnostic test panels for the detection of states of many specific disease-associated genetic markers simultaneously. Getting cheaper and more precise, next-generation sequencing technologies enhance the potential of medical diagnostics and bring the modern genome analysis approaches, such as whole genome sequencing (WGS) and whole exome sequencing (WES), into the clinical practice. These techniques allow to reveal the novel pathogenic and benign single nucleotide polymorphisms (SNPs), insertions, and deletions, which characterize the heterogeneous conditions [1] and to respecify the annotation of SNPs' clinical effect for different populations, assigned previously by using microarrays. It has been shown before that the set of disease-associated genes varies according to the patient's ethnicity [2], which brings universality of diagnostic panels into challenge. In this connection, the medical diagnosis should be provided with considering the population frequencies of the pathogenic variants.

Currently, several methods are applied to define population diversity, including Y-chromosomal and autosomal STR profiling [3], mtDNA sequencing [4], and SNP arrays such as Illumina chips [5], human origin array [6], and GenoChip [7]. All these specialized genotyping tools were designed to have the most convenient application in the field of population genetics, but not to detect the clinically relevant SNPs.

Published online: 20 September 2016

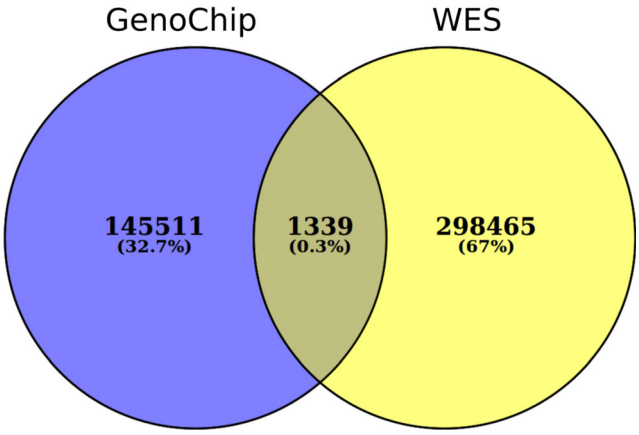

Fig. 1 Intersection between SNPs represented in GenoChip array and those detected for 57 samples using WES

The gene pool of the North Eurasian populations was well studied by many genetic systems, but nevertheless, the medical application of its genome profiling is on the initial stage. Thus, in the present study, we assess a possibility of WES to provide significant convenience for detection of specific population-associated and clinical markers.

2 Material and Methods

Fifty-seven DNA samples (19 male individuals belonging to Kazan Tatars population (from Tatarstan Republic of Russia), 4 to Mishari Tatars (Tatarstan Republic of Russia), 7 to Erzya (Mordovia Republic of Russia), 7 to Moksha (Mordovia Republic of Russia), and 20 to Tajik (Tajikistan)) were obtained from Biobank of indigenous populations of North Eurasia [8]. Blood samples were collected following signed informed consent from the participants who identified their four grandparents as members of the given ethnic group. DNA was obtained with phenol-chloroform extraction. For the target regions, amplification with specific primers pool 50 ng of non-degraded genomic DNA were used. Ion AmpliSeq Exome libraries were prepared and sequenced according to the manufacturer's specifications on the Ion Proton System with mean coverage of 51× and average number of reads on target of 95 %. Trimmed and filtered high quality reads (quality score >14) were aligned to the Human Genome Reference Consortium build 37 using the BWA software package (v.0.7.12). The Genome Analysis Toolkit v.3.3.0 (GATK) was used for further improvement of reads alignment according to GATK Best Practices [9]. For variant calling Unified Genotyper module was used with following GATK hard filtering (QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < −12.5 for heterozygous SNP, ReadPosRankSum < −8.0 for heterozygous SNP, DP < 30 for heterozygous SNP, and DP < 20 for homozygous SNP). Indels were not considered as well as SNPs with more than one variant. For PCA analysis, we applied the additional filters: missing calls less than 0.05, MAF above 0.01, and LD less than 0.2 in the sliding window of 1500 shifting by 150 bp each step. PC1 and PC2 described 2.5 and 2.1 % of total variation, respectively.

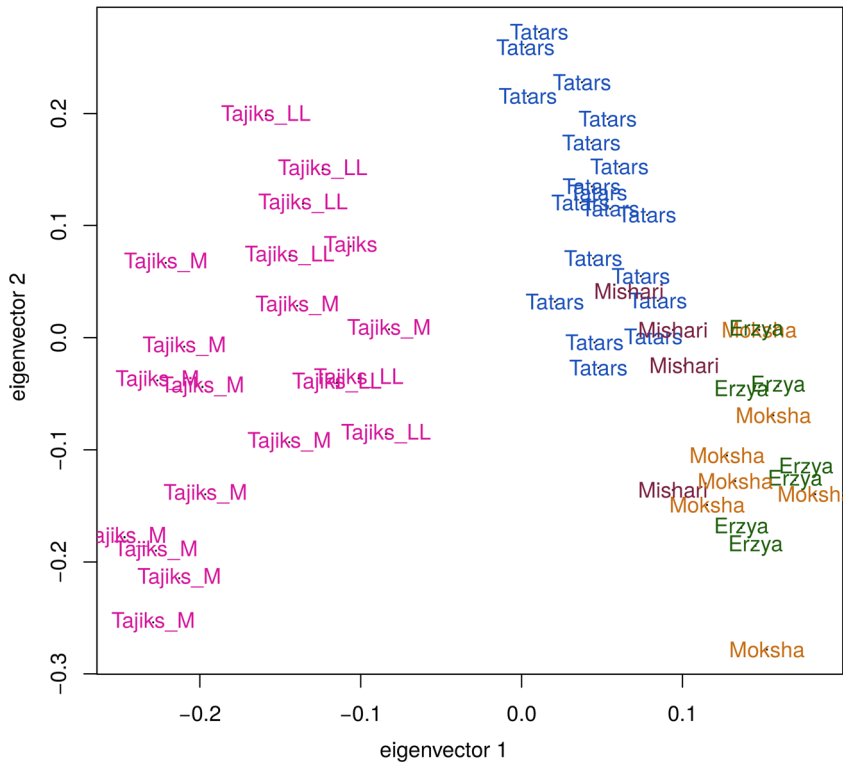

Fig. 2 Principal component plot of the 57 sequenced samples (PC1 and PC2)

Raw data for this project is stored on the Kazan Federal University servers and could be shared according to the Ethics Committee approval upon request.

3 Results and Discussion

SNP calling for all 57 analyzed exomes resulted in more than 299,000 raw variants. Comparing with the GenoChip, one of the most widely used in population genetic studies SNP array revealed very minor overlap of 1 % with WES variants (Fig. 1). This exemplifies how large portion of SNPs present within the coding parts of genome is missed by SNP arrays but could be revealed by whole exome sequencing. Consequently, the exome sequencing technique has a huge potential to reveal the novel ethnic-associated markers.

The principal component analysis of the filtered dataset (20350 SNPs) revealed clear clustering of the individual samples according to their population of origin (Fig. 2). It is notable that both populations from Mordovia cluster together while Mishari Tatars find their place in between these populations and Kazan Tatars population. Samples from Tajikistan are genetically distant, in line with their geographic separation. We observed the difference between mountain Tajiks (designated as Tajiks_M on the plot) and lowland Tajiks (Tajiks_LL), with the latter were more genetically similar to Tatars.

We believe that the AmpliSeq panel, which was designed initially for the medical use, could become a beneficial tool for future research in the field of population genetics, as well.

4 Conclusions

High-throughput WES with the Ion AmpliSeq Exome panel, resulted in dozens of thousands of new SNPs, shows sufficient resolution for the reflection of population diversity, on the one hand, and allows to discover the clinically relevant variants, on the other.

Acknowledgments The research was performed using the equipment of Interdisciplinary Centre for Shared Use of Kazan Federal University. This study has been in part supported by the Russian Foundation for Basic Research (grant 06-04-00890 to OB and VC) and is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

1. Klee, E. W., Hoppman-Chaney, N. L., Ferber, M. J. (2011). Expanding DNA diagnostic panel testing: is more better? *Expert Review of Molecular Diagnostics, 11*(7), 703–709.
2. Girgis, A. H., Wang, M., Fine, A., et al. (2016). Abstract A34: BRCA1 and BRCA2 mutation spectrum across 5, 509 high-risk individuals identifies pathogenic variants associated with ethnicity, age of diagnosis, and type of cancer. *Molecular Cancer Research, 14*(2), A34–A34. doi:10.1158/1557-3125.
3. Balanovsky, O. P., Dibirova, K. D., Dybo, A. V., et al. (2011). Parallel evolution of genes and languages in the Caucasus region. *Molecular Biology and Evolution, 28*(10), 2905–2920.
4. Krzewińska, M., Bjørnstad, G., Skoglund, P. (2015). Mitochondrial DNA variation in the Viking age population of Norway. *Philosophical Transactions of the Royal Society B, 370*(1660), 20130384.
5. Yunusbayev, B., et al. (2015). The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genetics, 11*(4), e1005068.
6. Pickrell, J. K., Patterson, N., Loh, P. R., et al. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Science, 111*(7), 2632–2637.
7. Elhaik, E., et al. (2013). The GenoChip: a new tool for genetic anthropology. *Genome Biology and Evolution, 5*, 1021–1031.
8. Balanovska EV, Zhabagin MK, Agdzhoyan AT et al. Population biobanks: organizational models and prospects of use in gene geography and personalized medicine. *Russ J Genet* (in print).
9. Auwera, G. A., Carneiro, M. O., Hartl, C., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics, 43*, 11.10.1–33. doi:10.1002/0471250953.