Name of the project: IRN AP19678041 «Development of software for identification of tandem repeats using whole genome sequencing»

Urgency:

Over the past decades, the idea of the role of repeating sequences in the genome has changed dramatically, and from the category of "junk DNA", repeating elements have a great influence on the functioning and evolution of the genomes of their hosts, contributing to genetic diversity and the emergence of new regulatory elements. Further development of sequencing technology, and in particular third generation sequencing, significantly contributes to the study of tandem repeats, which has led to the emergence of new data for detailed study. It has been established that short tandem repeats account for about 7% of the human genome. Wide representation in the genomes of eukaryotes and prokaryotes, and their high rate of variability, as one of the key factors in genome evolution and genetic diversification, repeats will be systematically evaluated for their role. Since many of these elements are known to be activated in diseases, there is potential for personalized medicine and disease diagnosis regarding genetic changes and expected consequences in the analysis of tandem repeats and identification of biomarker associations and regulation of biological processes in organism. In this regard, the development of advanced and easy-to-use bioinformatics tools for identifying various forms of tandem repeats is an **urgent task**.

The purpose of the proposed project is to develop an open access bioinformatics application for the identification and analysis of tandem repeat variability, including in the original data for third-generation whole genome sequencing.

Expected and achieved results:

Within the framework of the project, bioinformatics algorithms to identify related sequences with different levels of divergence, as well as programming languages and tools to develop the structure and interface of the application will be applied. Confirmation of the reliability of the results obtained in the software being developed will be carried out by standard laboratory molecular genetic methods, including genome-wide sequencing of prokaryotic and eukaryotic genomes, and methods for differentiating tandem repeats using capillary electrophoresis. **The main result** of the project being implemented will be open access software and a user interface for identifying tandem repeats during genome-wide sequencing. The software will allow identifying a variety of target loci with tandem repeats, including in the initial data of genome-wide sequencing and conducting statistical analysis of the identified variants.

During the implementation of the project, at least 2 (two) articles and (or) reviews will be published in peer-reviewed scientific journals included in the 1 (first) and (or) 2 (second) quartile by the impact factor in the Web of Science database and (or) having a CiteScore percentile in the Scopus database of at least

65 (sixty-five); or at least 1 (one) article or review in a peer-reviewed scientific publication included in the 1 (first) quartile in the Web of Science database or a CiteScore percentile in the Scopus database of at least 95 (ninety-five). All bioinformatics codes, scripts will be placed in permanent, open repositories, as well as placed on Github with free access.

Members of the research group:

project supervisor – Ismailova Aisulu, PhD, Associate Professor
ORCID: 0000-0002-8958-1846
Scopus/WoS (Hirsch Index = 3): Scopus Author ID: 56145830200

research group:

1) **Kalendar Ruslan**, Ph.D., Chief Scientific Officer, genetic biologist, Professor (Biology), Associate Professor of Genetics (University of Helsinki)
ORCID: 0000-0003-3986-2460
Scopus/WoS (Hirsch Index = 34): ResearcherID: D-9751-2012
Scopus ID: 6602789279

2) **Beldeubayeva Zhanar**, Leading Researcher, PhD
ORCID: 0000-0003-4056-6220
Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56951278600

3) **Satybaldiyeva (Satekbaeva) Aizhan**, Leading Researcher, PhD
ORCID: 0000-0001-5740-7934
Scopus/WoS (Hirsch Index =2): Scopus Author ID: 56145597900

4) **Shevtsov Vladislav**, Senior researcher, Master of Technical Sciences, doctoral student of the program «Big Data Analytics» of the Department of Information Systems, "S. Seifullin Kazakh Agrotechnical University"
ORCID: 0000-0001-6202-2123
Scopus/WoS (Hirsch Index =3): Scopus Author ID: 57216896596

5) **Golenko Yekaterina**, Senior researcher, Master of Technical Sciences,
ORCID: 0000-0002-4643-4571

6) **Vacancy**, Leading Researcher, IT architect, programmer

7) **Vacancy**, research associate, doctoral student

Information for potential users:

The scope of the developed software: bioinformatics, medical and agricultural genetics, genetics of microorganisms. The results of this project are of great importance, including for the fundamental sciences. The software will effectively identify tandem repeats and establish associations between the diversity of tandem repeats with human genetic diseases and with the genetic diversity of microorganisms and their pathogenicity. The implementation of the project will strengthen the direction of bioinformatics in the country's leading university and create a platform for specialization and career guidance for students.