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 =3): 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
Scopus/WoS (Hirsch Index =1): Scopus Author ID: 57962978000
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.

The results obtained for the project for 2023

1) A library of classes has been developed that will allow identifying sequences containing target loci with tandem repeats of known nature, as well as predicting tandem repeats for regions with a hidden signature and unknown nature. The identified tandems are classified according to their signature, the nature of the repeat and the heterogeneity of the tandem blocks, for further analysis and identification of allelic variants in the compared genotypes. An algorithm and software code were developed to identify any type of repeats in genomic sequences. In addition, repeat analysis is carried out on whole genome sequences from the NCBI gene bank. Repeat sequences are functionally ubiquitous structural units that are found throughout genomes. However, the diversity of repeats, each with a unique signature and structure, makes them difficult to classify. To overcome this problem, we developed a tool to detect any type of repeats in genomic sequences. A Java tool for identifying repeats and genomic analysis results in various taxonomic species, including the genomes of eukaryotes, fungi, microorganisms, and giant viruses, is available at https://zenodo.org/records/8424601, and the source code is freely available on GitHub at https: //github.com/rkalendar/Repeater.

2) The performance of the developed tandem repeat sequence identification code was tested using several algorithms, including linear nearest neighbor models, Knuth-Morris-Pratt algorithm, Boyer-Moore algorithm, Rabin-Karp algorithm, and suffix trees algorithm. To evaluate the effectiveness of testing, the following parameters were chosen: the speed of each algorithm, the number of tandem repetitions found. In the process of testing the effectiveness of the code for identifying tandem repeat sequences, the following steps were performed:

- Made-up biological sequence data were prepared containing various tandem repeat patterns that needed to be identified.
- A series of tests were carried out where each algorithm was applied to the prepared data. When compared, the algorithm using suffix trees was determined to be the most effective algorithm option for identifying tandem repeats.
- The best algorithm was determined based on the criteria of efficiency in finding tandem repeat patterns and speed.
- An algorithm has been developed for searching for tandem repeats, including methods using suffix trees.

As a result, suffix trees, generalized suffix trees, a multi-line variant of suffix trees, can be used to solve computational biology problems in optimal space and time.

During the current period of the project, 3 scientific articles were published:

1) **Kalendar R**, Karlov GI 2023. Editorial: Mobile Elements and Plant Genome Evolution, Comparative Analyses and Computational Tools, Volume II. ***Frontiers in Plant Science***, 14: 1308536. DOI: 10.3389/fpls.2023.1308536.

https://www.frontiersin.org/articles/10.3389/fpls.2023.1308536/full

WoS IF$_{2022}$=6.627 Q1;

Scopus 88th percentile

https://www.scopus.com/sourceid/21100313905

2) Belyayev A, **Kalendar R**, Josefiová J, Paštová L, Habibi F, Mahelka V, Mandák B, Krak K 2023. Telomere sequence variability in genotypes from natural plant populations: unusual block-organized double-monomer terminal telomeric arrays. BMC Genomics 24, 572 (2023).

https://doi.org/10.1186/s12864-023-09657-y

WoS IF_{2022}=4.4 Q1;

Scopus 76th percentile

https://www.scopus.com/sourceid/21727

3) **Shevtsov V., Ismailova A., Beldeubayeva Zh., Satybaldiyeva A.,** Nurpeisova A. MLVA as a method of genotyping and algorithms for its implementation using genome-wide data. News of the National academy of sciences of the Republic of Kazakhstan. Physico-mathematical series. Volume 4. № 348 (2023). P. 300-312 https://doi.org/10.32014/2023.2518-1726.235