A USER-FRIENDLY TOOL FOR SIMPLIFIED GENOMICS DATA MINING FROM LARGE VCF FILES

Daniyar Karabayev[1], Askhat Molkenov[1], Kaiyrgali Yerulanuly[1,2], Asset Daniyarov[1], Aigul Sharip[1], Ainur Seisenova[1], Zhaxybay Zhumadilov[1,3] and Ulykbek Kairov[1]

[1]Laboratory of Bioinformatics and Systems Biology, Center for Life Sciences, National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan
[2]L.N. Gumilyov Eurasian National University, Nur-Sultan, Kazakhstan
[3]Corporate fund "University Medical Center", Nazarbayev University, Nur-Sultan, Kazakhstan

Introduction: High-throughput sequencing platforms generate a massive amount of high-dimensional genomic datasets that are available for analysis. Modern and user-friendly bioinformatics tools for analysis and interpretation of genomics data becomes essential during the analysis of sequencing data. Variant Call Format (VCF) is a standard format containing genomic information and variants of sequenced samples. Existing tools for processing VCF files don't usually have an intuitive graphical interface, but instead have just a command-line interface that may be challenging to use for the broader biomedical community interested in genomics data analysis. We present re-Searcher, a new bioinformatics application with a user-friendly GUI developed to simplify genomic data mining from VCF files.

Methods: re-Searcher application was written in a Python 3. Pandas library solves the problem of analyzing large VCF files by not loading the whole file directly into RAM, but instead pre-processing it in chunks. Simple and intuitive GUI was built using Tkinter library.

Results: The generalized workflow of the re-Searcher consists of several steps: selecting an input file, setting up necessary filtering parameters, data processing, and exporting a filtered output VCF file. re-Searcher browses and opens VCF files with extensions .txt or .vcf, before performing the following filtering and extraction options: header extraction, keyword search, sample extraction, and genotype format conversion.

Conclusion: Exploring and analyzing VCF files generated after the bioinformatics processing of sequencing data is one of the important steps performed by researchers during analysis and meta-analysis of genotype/phenotype associations. We have developed and introduced an easy-to-use bioinformatics tool, re-Searcher, with several unique features for mining big VCF files and realized with a simple graphical user interface that makes it easily available for clinicians and researchers without any computational skills. The software publicly available on the GitHub repository (https://github.com/LabBandSB/re-Searcher)