re-Searcher: GUI-based bioinformatics tool for simplified genomics data mining of VCF files

Daniyar Karabayev [Equal first author, 1], **Askhat Molkenov** [1], **Kaiyrgali Yerulanuly** [1, 2], **Asset Daniyarov** [1], **Aigul Sharip** [1], **Ainur Seisenova** [1], **Zhaxybay Zhumadilov** [1, 3], **Ulykbek Kairov** [Corresp. Equal first author, 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] School of Medicine, Nazarbayev University, Nur-Sultan, Kazakhstan

Corresponding Author: Ulykbek Kairov
Email address: ulykbek.kairov@nu.edu.kz

Background. 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. Different standard data types and file formats have been developed to store and analyze sequence and genomics data. Variant Call Format (VCF) is the most widespread genomics file type and standard format containing genomic information and variants of sequenced samples.

[p]**Results.** 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. re-Searcher solves this problem by pre-processing VCF files by chunks and using simple graphical user interface (GUI). re-Searcher is user-friendly multiplatform GUI software that helps to analyze large VCF files. The software including source code as well as tested VCF files and additional information are publicly available on the GitHub repository https://github.com/LabBandSB/re-Searcher[p]

re-Searcher: GUI-based bioinformatics tool for simplified genomics data mining of 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] School of Medicine, Nazarbayev University, Nur-Sultan, Kazakhstan

Corresponding Author:
Ulykbek Kairov[1]
53 Kabanbay Batyr ave., Nur-Sultan, 010000, Kazakhstan
Email address: ulykbek.kairov@nu.edu.kz

Abstract

Background. 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. Different standard data types and file formats have been developed to store and analyze sequence and genomics data. Variant Call Format (VCF) is the most widespread genomics file type and standard format containing genomic information and variants of sequenced samples.

Results. 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. re-Searcher solves this problem by pre-processing VCF files by chunks and using simple graphical user interface (GUI). re-Searcher is user-friendly multiplatform GUI software that helps to analyze large VCF files. The software including source code as well as tested VCF files and additional information are publicly available on the GitHub repository https://github.com/LabBandSB/re-Searcher

Introduction

Recent achievements in high-throughput sequencing technologies [1, 2] have led to the generation of massive amounts of genomic data [3, 4] available for the research community. Many omics databases have been developed and have collected freely accessible datasets [5, 6] for analysis by the bioinformatics community. Modern bioinformatics tools and methods are in high demand for analyzing and interpreting the big omics data generated by the different types of multi-omics platforms available. Different standard data types and file formats have been

developed to store and analyze sequence and genomics data. Variant Call Format (VCF) [7] is a tab-delimited text file format that is often used in bioinformatics to store genomic variants. A VCF file consists of the header, including meta-information lines and field definition lines (column names), and the body (data section). An arbitrary number of meta-information lines start with '##' and provide a description of the VCF file. The body of the file consists of eight mandatory columns: chromosome (CHROM), starting position of a variant (POS), variant identifiers (ID), the reference allele (REF), a list of alternate alleles (ALT), a PHRED-scaled quality score (QUAL), filter information regarding variant validity (FILTER), and annotation information (INFO). Additional columns describing samples can also be added. Each row of the file describes specific genomic variants (SNVs, INDELs, CNVs, and other structural variants) at the given chromosome and genomic position.

VCF files often store information about numerous samples and can therefore reach huge sizes – gigabytes, or sometimes terabytes. This creates an issue for the readability of VCF files and further analysis for non-programmers, as manual data extraction and analysis using Microsoft Excel or other table processing software may not be possible due to the RAM capacity limitation of standard computers. We introduce the re-Searcher bioinformatics tool, specifically developed for simplified mining and analysis of big-size VCF files with a multi-platform user-friendly graphical user interface (GUI). re-Searcher has been developed for the broader biomedical community and solves the problem of working and analyzing genomics data stored in VCF format.

Materials & Methods

Implementation

re-Searcher application was written in Python 3 [8] with the implementation of Tkinter [9] package to build the GUI, and Pandas [10] library to extract columns. re-Searcher solves the problem of analyzing large VCF files by not loading the whole file directly into RAM, but instead pre-processing it in chunks and utilizing a simple and intuitive GUI (Figure 1).

The main advantage of re-Searcher in comparison with other tools is the presence of a simple and user- friendly GUI (Figure 1) instead of a command line interface, as well as a lack of confused installation procedures typical for existing tools. 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 (Figure 2). re-Searcher browses and opens VCF files with extensions ".txt" or ".vcf", before performing the following filtering and extraction options:

Header extraction

VCF files can be large and, if a user needs to know only certain information in a file header (e.g. a particular meta-line or sample ID), the software can extract only the header from the original VCF file and save it into a new file.

Keyword search

If genomic variants need to be filtered according to the presence of a keyword, the software can find these rows and extract them into a new VCF file. Users may input multiple keywords by accessing the entry field or by uploading a .txt file with keywords.

Sample extraction

If the user needs only particular sample IDs in a VCF file, the software can extract the necessary sample columns into a new file. Similar to a keyword search, users may input multiple sample IDs by accessing the entry field or by uploading a .txt file with the IDs.

Genotype format conversion

re-Searcher can convert numeric genotype (GT) for- mat into letter format. The conversion option is one of the most used operations when working with VCF files, for example, in further comparative analysis of genetic variants or SNPs. The original GT format in VCF files is numeric (0/0, 0/1, 1/1 or 1/2 for multiple nucleotide variants), where 0 is a reference (REF) allele and 1 is an alternative (ALT) allele [7]. Since multiple nucleotide variants are very rare compared to SNPs, we realized the conversion function to only work with SNVs (Figure 3). The final output file is a filtered and processed VCF file and is generated with a complement log file containing the file processing information, name of specified file and work directory with outputs.

Results and Discussion

We have compared the main features of re-Searcher with other existing and open source tools (VIVA [11], VCFtools [7], GEMINI [12], BrowseVCF [13], VCF.Filter [14], VCF-Miner [15]) and prepared a detailed table (Table 1). The compared tools have been developed and implemented in different programming languages (Julia, C++, Perl, Java and Python) and dedicated libraries. VCFtools is one of the most cited and advanced tools for processing VCF files, but it requires additional computational skills for effective usage. VIVA is the only tool that provides the possibility for advanced visualization and plotting figures. In addition to re-Searcher, other tools with a GUI available for users are BrowseVCF, VCF.Filter, and VCF-Miner. From these, only re-Searcher, BrowseVCF and VCF.Filter tools support multiple operational systems (Windows, MacOS, Linux). Searching the whole VCF file by keyword and the corresponding extraction of data based on keywords are features available on re-Searcher and BrowseVCF, whereas genotype conversion is a unique feature of re-Searcher.

re-Searcher is a multi-platform tool and can be run on MacOS, Windows and Linux operating systems. Performance of re-Searcher has been evaluated on these PC platforms (Windows 10 Pro OS and Linux Mint 17 OS-based PC: CPU Intel Core i5-8250U 1.80 GHz, RAM 4Gb and MacOS Catalina based PC: CPU Intel Core i7, 3.2 GHz, RAM 8 Gb) with different VCF file sizes. Different size VCF files (0.081 Gb, 0.814 Gb, 1.320 Gb, 1.980 Gb and 7.950 Gb) were used as input datasets for evaluating re-Searcher performance. The results of the performance benchmarking are shown in Table 2.

We have used big VCF files from the 1000 Genomes Project [16] and then generated the different sized testing VCF files from this dataset. The Linux-based systems had the fastest execution time for different operations in comparison with Windows and MacOS systems.

Availability

re-Searcher executable software including source code, tested VCF files and additional information are publicly available on the GitHub repository https://github.com/LabBandSB/re-Searcher. re-Searcher is free bioinformatics tool and open to all users without login and registration requirements and do not require an installation of additional tools.

Conclusions

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. Several improvements such as visualization options (clustering and plotting functions) with Principal Component Analysis and heatmap methodologies incorporated as the new web application are under future development of re-Searcher.

Acknowledgements

This work is dedicated to the blessed memory of Dr.Vasily Ogryzko.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016 6;17:333–351. Available from: http://www.nature.com/articles/nrg.2016.49.
2. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends in Genetics. 2018 9;34:666–681. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168952518300969.
3. Gao GF, Parker JS, Reynolds SM, Silva TC, Wang LB, Zhou W, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. Cell Systems. 2019 7;9:24–34.e10. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2405471219302017.
4. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93. Available from: https://doi.org/10.1038/s41586-020-1969-6.
5. Molkenov A, Zhelambayeva A, Yermekov A, Mussurova S, Sarkytbayeva A, Kalykhbergenov Y, et al.. Transcriptomic Databases; 2019. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128096338202082.

6. Rigden DJ, Fern´andez XM. The 27th annual Nucleic Acids Research database issue and molecular biology database collection. Nucleic Acids Research. 2020 1;48:D1–D8. Available from: https://academic.oup.com/nar/article/48/D1/D1/5695332.
7. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 8;27:2156–2158. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr330.
8. Rossum GV, Drake FL. Python 3 Tutorial. Python Software Foundation (docs@pythonorg). 2012;p. 1–136.
9. Lundh F. An Introduction to Tkinter. Review Literature And Arts Of The Americas. 1999;p. 166.
10. McKinney W, Team PD. Pandas - Powerful Python Data Analysis Toolkit. Pandas - Powerful Python Data Analysis Toolkit. 2015;p. 1625.
11. Tollefson GA, Schuster J, Gelin F, Agudelo A, Ragavendran A, Restrepo I, et al. VIVA (VIsualization of VAriants): A VCF File Visualization Tool. Scientific Reports. 2019 12;9:12648. Available from: http://www.nature.com/articles/s41598-019-49114-z.
12. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Computational Biology. 2013 7;9:e1003153. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003153.
13. Salatino S, Ramraj V. BrowseVCF: a web-based application and workflow to quickly prioritize disease-causative variants in VCF files. Briefings in Bioinformatics. 2016 7;p. bbw054. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/ bib/bbw054.
14. Mu¨ller H, Jimenez-Heredia R, Krolo A, Hirschmugl T, Dmytrus J, Boztug K, et al. VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. Nucleic Acids Research. 2017 7;45:W567–W572. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkx425.
15. Hart SN, Duffy P, Quest DJ, Hossain A, Meiners MA, Kocher JP. VCF-Miner: GUI-based application for mining variants and annotations stored in VCF files. Briefings in Bioinformatics. 2016 3;17:346–351. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv051.
16. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74. Available from: https://doi.org/10.1038/nature15393.

Figure 1

Main window of re-Searcher interface

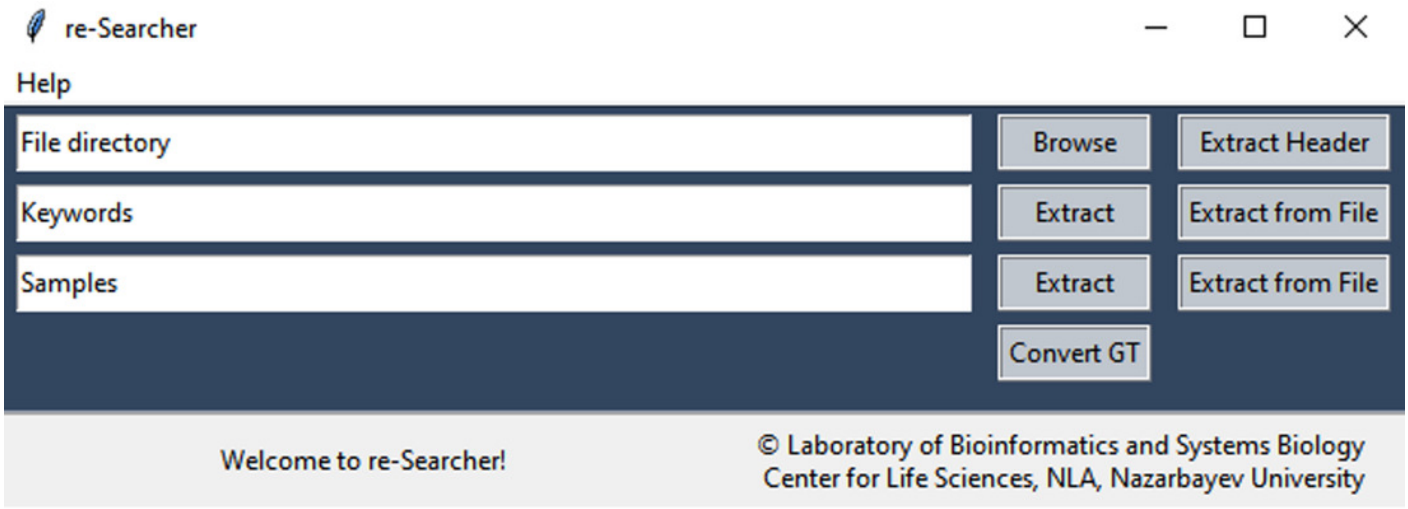


Figure 2

Data processing workfow of re-Searcher

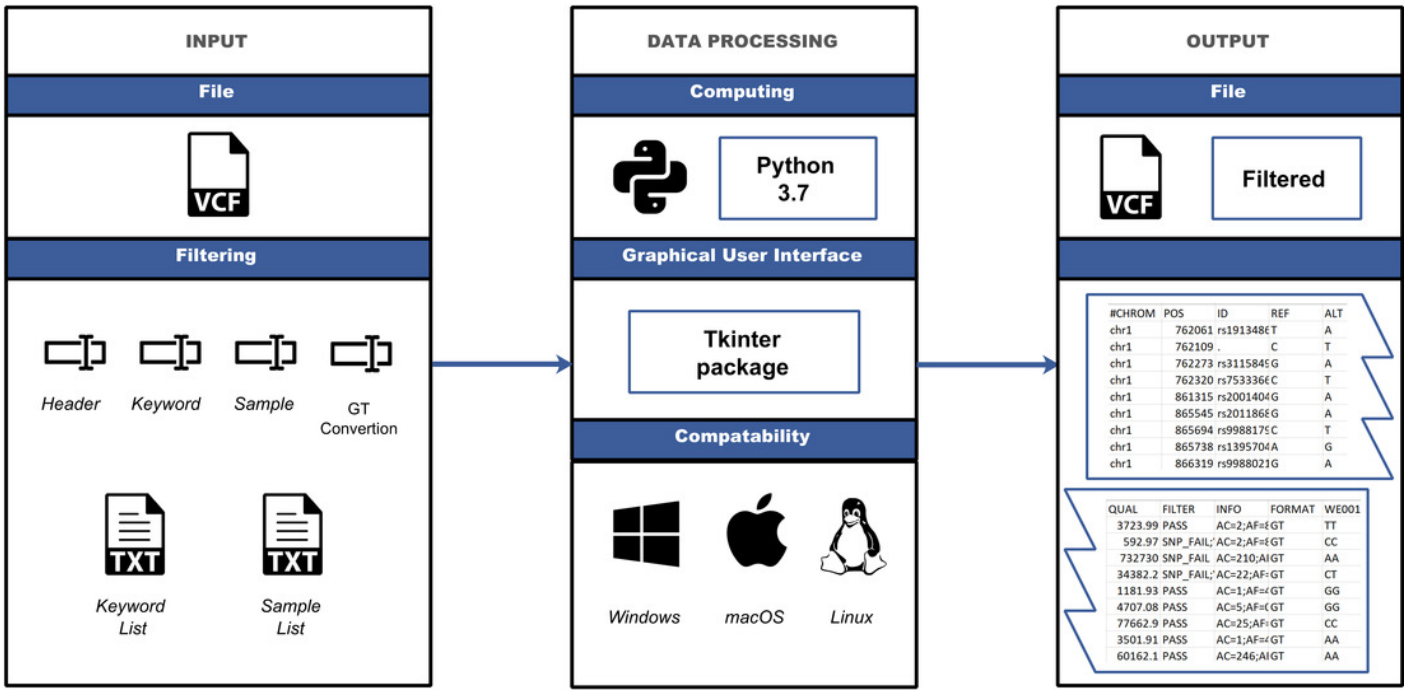
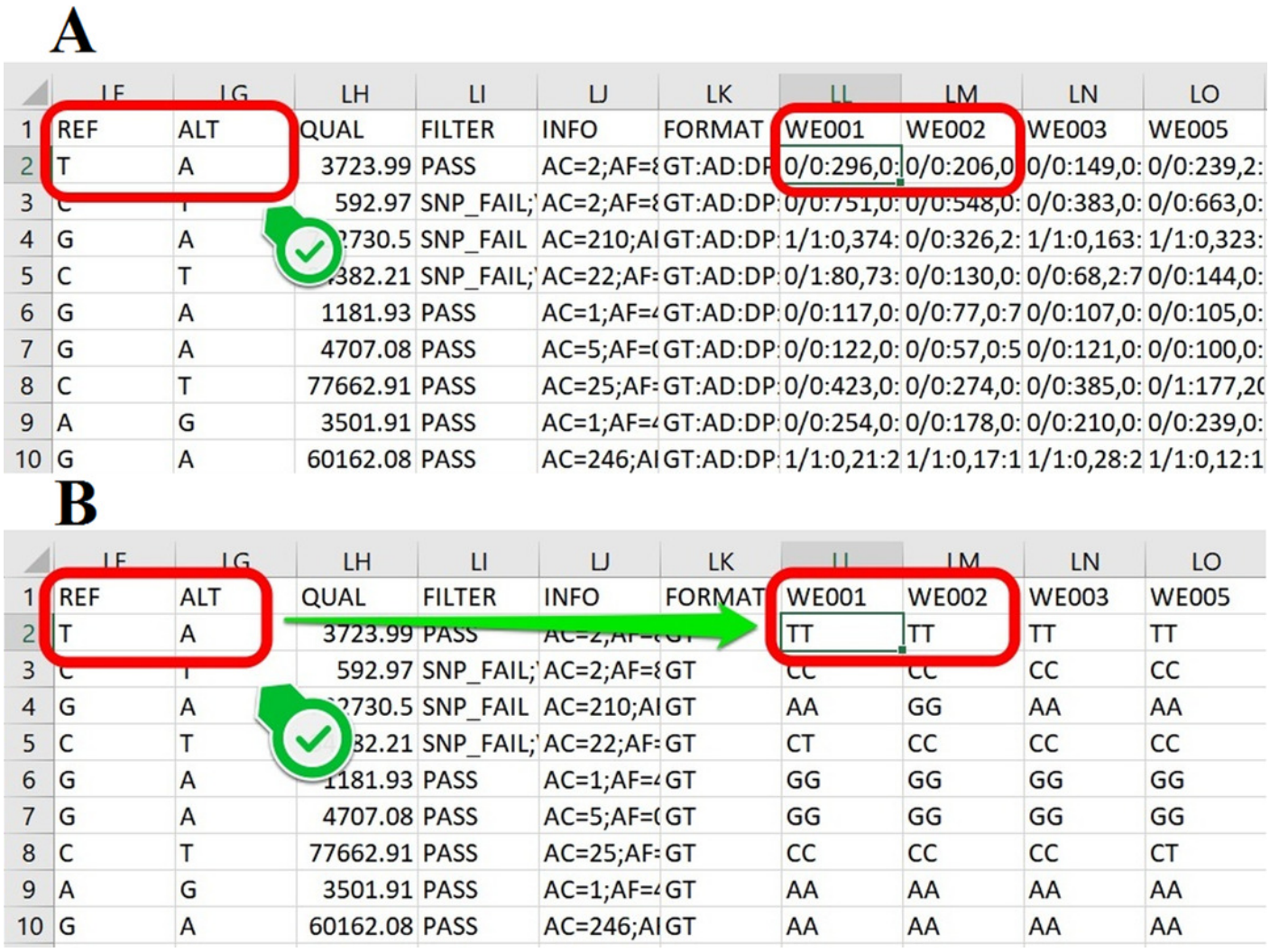


Figure 3

Results of genotype (GT) conversion option by re-Searcher:

(A) before and (B) after VCF file processing. Conversion changes are highlighted by red and green marks and arrows.



A

	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO
1	REF	ALT	QUAL	FILTER	INFO	FORMAT	WE001	WE002	WE003	WE005
2	T	A	3723.99	PASS	AC=2;AF=8	GT:AD:DP	0/0:296,0:	0/0:206,0	0/0:149,0:	0/0:239,2:
3	C	T	592.97	SNP_FAIL;'	AC=2;AF=8	GT:AD:DP	0/0:751,0:	0/0:548,0:	0/0:383,0:	0/0:663,0:
4	G	A	2730.5	SNP_FAIL	AC=210;AI	GT:AD:DP	1/1:0,374:	0/0:326,2:	1/1:0,163:	1/1:0,323:
5	C	T	382.21	SNP_FAIL;'	AC=22;AF=	GT:AD:DP	0/1:80,73:	0/0:130,0:	0/0:68,2:7	0/0:144,0:
6	G	A	1181.93	PASS	AC=1;AF=4	GT:AD:DP	0/0:117,0:	0/0:77,0:7	0/0:107,0:	0/0:105,0:
7	G	A	4707.08	PASS	AC=5;AF=(	GT:AD:DP	0/0:122,0:	0/0:57,0:5	0/0:121,0:	0/0:100,0:
8	C	T	77662.91	PASS	AC=25;AF=	GT:AD:DP	0/0:423,0:	0/0:274,0:	0/0:385,0:	0/1:177,20
9	A	G	3501.91	PASS	AC=1;AF=4	GT:AD:DP	0/0:254,0:	0/0:178,0:	0/0:210,0:	0/0:239,0:
10	G	A	60162.08	PASS	AC=246;AI	GT:AD:DP	1/1:0,21:2	1/1:0,17:1	1/1:0,28:2	1/1:0,12:1

B

	LF	LG	LH	LI	LJ	LK	LL	LM	LN	LO
1	REF	ALT	QUAL	FILTER	INFO	FORMAT	WE001	WE002	WE003	WE005
2	T	A	3723.99	PASS	AC=2;AF=(	GT	TT	TT	TT	TT
3	C	T	592.97	SNP_FAIL;'	AC=2;AF=8	GT	CC	CC	CC	CC
4	G	A	2730.5	SNP_FAIL	AC=210;AI	GT	AA	GG	AA	AA
5	C	T	382.21	SNP_FAIL;'	AC=22;AF=	GT	CT	CC	CC	CC
6	G	A	1181.93	PASS	AC=1;AF=4	GT	GG	GG	GG	GG
7	G	A	4707.08	PASS	AC=5;AF=(	GT	GG	GG	GG	GG
8	C	T	77662.91	PASS	AC=25;AF=	GT	CC	CC	CC	CT
9	A	G	3501.91	PASS	AC=1;AF=4	GT	AA	AA	AA	AA
10	G	A	60162.08	PASS	AC=246;AI	GT	AA	AA	AA	AA

Table 1(on next page)

Comparison of re-Searcher's features with similar tools

1 **Table 1. Comparison of re-Searcher's features with similar tools**
2

Categories	Features	re-Searcher	VIVA	VCFtools	GEMINI	BrowseVCF	VCF.Filter	VCF-Miner
Technical Aspects	Compatibility with operation system	Windows, MacOS, Linux	Windows, MacOS, Linux	Windows, MacOS, Linux	Windows, MacOS, Linux	Windows, MacOS, Linux	Windows, MacOS, Linux	Windows
	Language	Python	Julia	C++, Perl	Python	Python, JavaScript, CSS, HTML5	Java	Java
	Interface	GUI	CLI, Jupyter Notebook	CLI	CLI, Web Browser	GUI, CLI	GUI	GUI
	Works offline	✓	✓	✓	✗	✗	✓	✓
	Portable launcher	✓	✗	✗	✗	✓	✗	✗
Functionality	Search by keyword	✓	✗	✗	✗	✓	✗	✗
	Sample selection	✓	✓	✓	✓	✓	✓	✓
	GT (genotype) format conversion	✓	✗	✗	✗	✗	✗	✗
	Visualization	✗	✓	✗	✗	✗	✗	✗
	Export filtered VCF file	✓	✗	✓	✗	✗	✓	✓

3
4

Table 2(on next page)

re-Searcher multi-platform run time comparison

OS – operational system, GT – genotype, sec – seconds, Gb – gigabyte.

Table 2. re-Searcher multi-platform run time comparison

OS – operational system, GT – genotype, sec – seconds, Gb – gigabyte.

OS	VCF file size (Gb)	Execution Time (sec)			
		Extract header	Keyword extraction	Sample ID extraction	GT conversion
Linux	0.081	2.227	9.995	15.243	38.375
	0.814	2.497	38.106	21.334	117.186
	1.320	2.462	67.260	61.159	206.868
	1.980	2.115	75.331	104.167	330.527
	7.950	6.145	366.137	200.347	1117.168
Windows	0.081	14.865	22.482	4.785	49.642
	0.814	18.898	48.820	21.398	139.641
	1.320	21.054	44.669	59.329	238.192
	1.980	10.919	53.958	90.446	339.275
	7.950	16.308	502.996	309.177	1320.197
MacOS	0.081	9.627	9.297	10.423	16.298
	0.814	5.262	20.916	19.705	116.82
	1.320	6.286	35.433	53.247	186.181
	1.980	3.231	37.923	119.136	286.544
	7.950	5.457	148.612	254.128	1130.225