GenEditScan-GUI User Guide

Version 1.0.2

December 26, 2024

National Agriculture and Food Research Organization (NARO)

GenEditScan-GUI is a tool designed to detect foreign DNA sequences in genome-edited agricultural products through the analysis of NGS data and statistical testing. It features visualization capabilities for sequence fragment count numbers and statistical test results. The tool achieves faster calculations through the use of bit operations and parallel processing. Developed using JavaFX, one of Java's standard development environments, GenEditScan-GUI is compatible with both Windows and Mac operating systems.

Contents

1. Introduction

GenEditScan-GUI is a tool designed to detect foreign DNA sequences in genome-edited agricultural products through the analysis of NGS data and statistical testing. It features visualization capabilities for sequence fragment count numbers and statistical test results. The tool achieves faster calculations through the use of bit operations and parallel processing. Developed using JavaFX, one of Java's standard development environments, GenEditScan-GUI is compatible with both Windows and Mac operating systems.

2. Getting started

[Windows]

Double-click on "GenEditScan_win.vbs" or a shortcut to this vbs file (note: a copy of the vbs file will not function).

[Mac]

Launch "GenEditScan_mac.sh" from the terminal or double-click on "GenEditScan_mac.command".

3. Function

An overview of the features of GenEditScan-GUI is shown in Table 1. When both Mutant and Wild type data are specified in "Count mer," the software performs G-tests after counting the mers and displays the test results along with the counts in a graph. If only Mutant or Wild type data is provided, the G-tests are not performed, and only the counts are displayed in a graph.

Table 1　Summary of GenEditScan-GUI features

Tab menu	Overview of the features	
Count mer	Count mers (sequence fragments) and perform G-tests. Specify the vector file (fasta), Mutant file (fastq), Wild-type file (fastq), K-mer length (fragment length), and the number of calculation threads.	
Draw graph	Graph display	Display graphs of mer counts and G-statistics. It is possible to switch the display to p-values, FDR, or Bonferroni correction. The axis range, title, and graph colors can be customized, and the graphs can be saved as PDF or PNG files.
	Flanking sequences display	Select sequences exceeding the specified FDR-corrected p-values threshold using mouse operations, and count the frequency of surrounding sequence patterns for both mutant and wild-type.

4. Input and output files

Table 2 shows the input and output files for each tab menu of GenEditScan-GUI, and Table 3 provides details about the file contents. For "Count mer," users need to prepare the input files. By providing all three types of input files listed in the table to "Count mer," the software will perform G-tests following the mer counting process. The filenames of the generated files will use the prefix specified in the GUI.

Table 2　Input/output files of GenEditScan-GUI

Tab menu	Input file	Output file
Count mer	・ Vector sequence file (fasta) ・ Mutant NGS data (fastq) ・ Wild-type NGS data (fastq)	・ out_prefix.statistics.txt ・ out_prefix.outside.txt ・ out_prefix.mutant.merFreq.txt ・ out_prefix.wildtype.merFreq.txt
Draw graph	・ out_prefix.statistics.txt ・ out_prefix.outside.txt	・ PDF ・ PNG

Table 3　Overview of the input files

File type/name	Content
Vector sequence file	Vector sequence data in FASTA format. Required
Mutant NGS data	Mutant NGS data in FASTQ format. Required
Wild-type data	Wild-type NGS data in FASTQ format. Required
out_prefix.statistics.txt	A tab-delimited text file containing G-statistics and mer counts for each vector position.
out_prefix.outside.txt	A tab-delimited file containing the results of k-mer and surrounding sequences analysis.
out_prefix.mutant.merFreq.txt	Mer sequences and frequencies of mutant data
out_prefix.wildtype.merFreq.txt	Mer sequences and frequencies of wild-type data
PDF	Created using the "Save graph" function in the Draw graph tab.
PNG	Created using the "Save graph" function in the Draw graph tab.

5. Statistics file

The contents of the G-test result file for k-mer sequence detection, out_prefix.statistics.txt, are shown in Table 4, and an example output is presented in Table 5. The data is output in a tab-delimited format. The first line contains the number of bases in the analyzed k-mer as a comment, and the second line outputs header information in a comment format.

Table 4 Contents of out_prefix.statistics.txt (Header Information)

Items	Content
Pos	Position on the vector sequence
Seq	k-mer sequence
Mutant	k-mer count in the mutant data
WildType	k-mer count in the wild-type data
Gval	G-statistics (Williams Collections)
Pval	P-value
FDR	FDR-adjusted p-value (Benjamin-Hochberg)
Bonferroni	p-value adjusted using the Bonferroni method

Table 5 Example of out_prefix.statistics.txt

#K-mer	20						
#Pos	Seq	Mutant	WildType	Gval	Pval	FDR	Bonferroni
1	T	1	0	0.9769876	0.32294366	0.372447	1.0
2	A	1	0	0.9769876	0.32294366	0.3724235	1.0
3	A	1	0	0.9769876	0.32294366	0.37240002	1.0
4	A	1	0	0.9769876	0.32294366	0.3723765	1.0
5	C	1	0	0.9769876	0.32294366	0.37235302	1.0
1146	G	12	0	16.893818	3.9530132E-5	3.5464566E-4	0.7224132
1147	C	12	0	16.893818	3.9530132E-5	3.5447165E-4	0.7224132
1148	T	12	0	16.893818	3.9530132E-5	3.5429778E-4	0.7224132
1149	C	13	0	18.358307	1.830192E-5	2.1803624E-4	0.3344676
1150	A	13	0	18.358307	1.830192E-5	2.178942E-4	0.3344676
:	:	:	:	:	:	:	:

6. Flanking sequence file

The output example of the analysis result file for sequences surrounding detected k-mers, out_prefix.outside.txt, is shown in Table 6, and the contents of each column are described in Table 7. The data is output in a tab-delimited format.

The first line outputs, in a comment format, the number of bases in the analyzed k-mer, the FDR threshold for analyzing surrounding sequences, and the number of bases added upstream and downstream (per side). From the second line onward, "k-mer sequence rows (gray rows)" and "flanking base rows (white rows)" are output. Note that symbols such as A, B, C, ... are not included in the output.

Table 6 Example of out_prefix.outside.txt

A	B	C	D	E	F	G	H	I
#K-mer	20	FDR	0.01	Bases	5			
2324	2	ACATATGCCCG TCGACCCCA	82	0	111.752	4.04896e-26	1.01017e-25	7.32376e-22
TTGAT	TCACA	72	0	TTGATACATATGCCCG TCGACCCCA TCACA	98.0146	4.15309e-23	1.75028e-22	1.36662e-18
TTCAT	TCACA	1	0	TTCATACATATGCCCGT CGACCCCA TCACA	0.911434	0.339734	0.431766	1
2325	3	CATATGCCCGT CGACCCCAT	82	0	111.752	4.04896e-26	1.01003e-25	7.32376e-22
TGATA	CACAA	75	0	TGATACATATGCCCGT CGACCCCAT CACAA	102.126	5.21069e-24	2.28131e-23	1.71463e-19
TGATA	CACAT	1	0	TGATACATATGCCCGT CGACCCCAT CACAT	0.911434	0.339734	0.431733	1
TCATA	CACAA	1	0	TGATACATATGCCCGT CGACCCCAT CACAA	0.911434	0.339734	0.43175	1
2326	3	ATATGCCCGTC GACCCCATC	81	0	110.381	8.0853e-26	2.00009e-25	1.46247e-21
GATAC	ACAAG	75	0	GATACATATGCCCGTC GACCCCATC ACAAG	102.126	5.21069e-24	2.281e-23	1.71463e-19
GATAC	ACATG	1	0	GATACATATGCCCGTC GACCCCATC ACATG	0.911434	0.339734	0.4317	1
CATAC	ACAAG	1	0	GATACATATGCCCGTC GACCCCATC ACAAG	0.911434	0.339734	0.431716	1
:	:	:	:	:	:	:	:	:

Table 7　Contents of out_prefix.outside.txt

Symbol	k-mer sequence rows (gray rows)	flanking base rows (white rows)
A	Position on the vector sequence	Upstream flanking bases
B	Number of the flanking bases patterns	Downstream flanking bases
C	k-mer sequence	Count of flanking bases patterns in the mutant data
D	Count of k-mer sequence in the mutant data	Count of flanking bases patterns in the wild-type data
E	Count of k-mer sequence in the wild-type data	A sequence with upstream and downstream bases added to the k-mer sequence
F	G-statistics (Williams Collections)	
G	p-value	
H	FDR-corrected p-value (Benjamin-Hochberg)	
I	p-value adjusted using Bonferroni method	

7.　k-mer frequency file

For all patterns of the specified k-mer length on the vector sequence, the count in the NGS data of both mutant and wild-type is output. An example of a k-mer frequency file output is shown in Table 8. The data is output in a tab-delimited format.

Table 8　Example of k-mer frequency file

K-mer sequence	Frequency
AAAAAAAAAGGAGAACACAT	140
AAAAAAAAGCATGAAAAGAT	96
AAAAAAAAGGAGAACACATG	138
AAAAAAAAGGATGATCATGC	109
AAAAAAAATATGTGGTAATT	122
AAAAAAAATCATGAAATCGA	138
AAAAAAACATGTCATAACAA	0
AAAAAAACCACCGCTACCAG	81
AAAAAAACTAAAATAGAGTT	124
AAAAAAACTAAGGAAACATT	108

8. Count mer

The "Count mer" screen is shown in Figure 1, and its features and usage instruction are listed in Table 9.

⑦ The k-mer length is limited to a range of 8-1024 bases.

⑫ The "Maximum number of threads" is set to the lesser of the PC's logical processor count and 8, but this can be modified. The actual number of processors used is capped by the total number of fastq files specified in ② and ④.

⑬ Either the "Stop" or the "Execute" button is enabled, but not both at the same time.

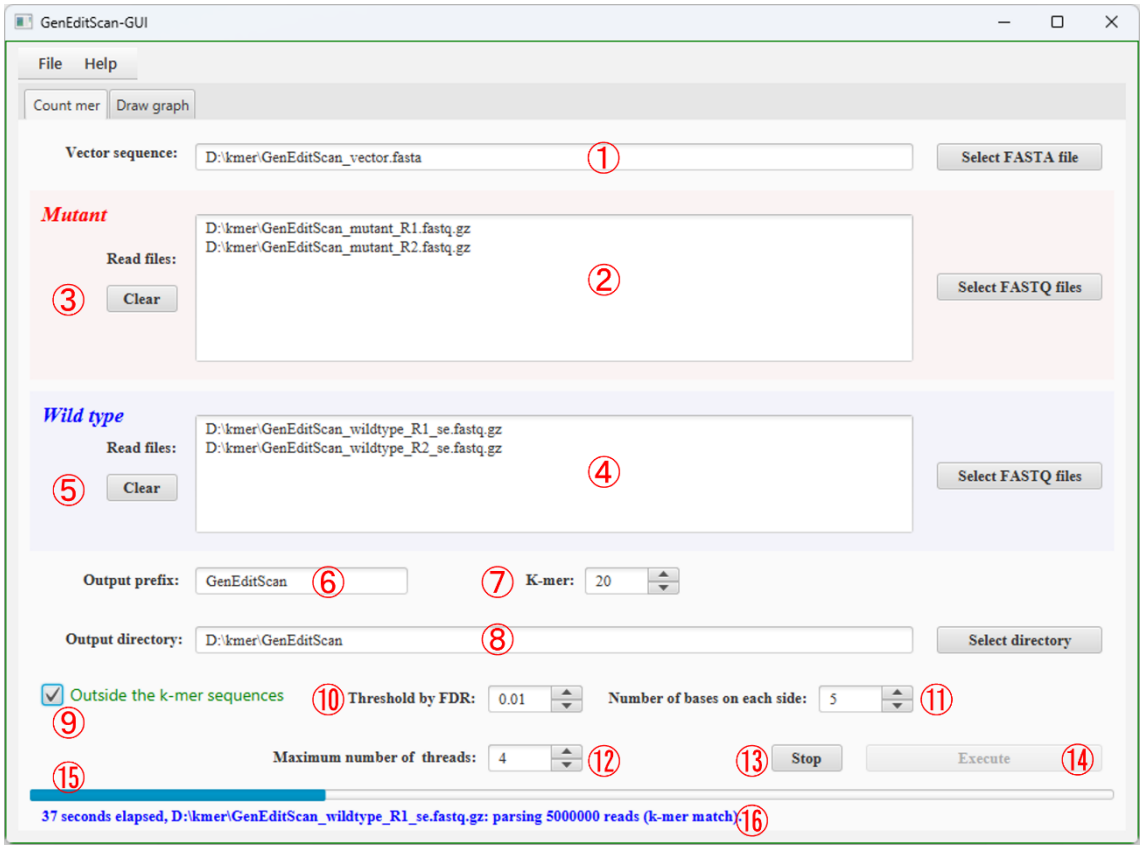


Figure 1 "Count mer" screen

Table 9 Features and usage instruction on "Count mer" screen

No.	Display items	Feature/usage
①	Vector sequence	Select using the "Select FASTA file" button on the right.
②	Mutant Read files	Select using the "Select FASTQ files" button on the right.
③	Clear	Clear the contents of ②
④	Wild type Read files	Select using the "Select FASTQ files" button on the right.
⑤	Clear	Clear the contents of ④

⑥	Output prefix	Prefix for output files
⑦	K-mer	Length of the k-mer sequence to be analyzed (8-1024)
⑧	Output directory	Select using the "Select directory" button on the right.
⑨	Outside the k-mer sequences	Check the box to analyze the sequences flanking the k-mer.
⑩	Threshold by FDR	Threshold of the FDR-corrected p-value for analyzing the flanking bases
⑪	Number of bases on each side	Number of the flanking nucleotide bases to be analyzed.
⑫	Maximum number of threads	The actual number of processors used is capped by the total number of fastq files specified in ② and ④.
⑬	Stop	Stop the analysis
⑭	Execute	Execute the analysis
⑮	Progress bar	Progress of the calculation
⑯	Message	Message of the progress

9. Draw graph

9.1. Graph display

The "Draw graph" screen is shown in Figure 2, and its features and usage instruction are listed in Table 10. This screen is automatically drawn upon completion of the analysis executed in the "Count mer" screen. It is also possible to draw graphs by selecting a pre-calculated result file (⑲). Using “Save graph” function (⑯), the graph can be saved as a vector-based PDF file or a PNG file.

Figure 2　Example of the “Draw graph” screen

Table 10　Features and usage instruction on “Draw graph” screen

No.	Display items	Feature/usage
①	Y-axis (upper)	Label of the Y-axis.
②	auto	Auto-scale the Y-axis of the above figure.
③	specify	Specify the range of the Y-axis in the above figure.
④	Mutant	Line color of the mutant counts in the above figure.
⑤	Wild type	Line color of the wild-type counts in the above figure.
⑥	Y-axis (lower)	Metrics to be displayed in the figure below.

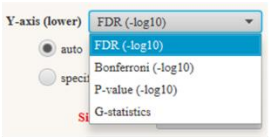
⑦	auto	Auto-scale the Y-axis of the below figure.
⑧	specify	Specify the range of the Y-axis in the below figure.
⑨	Significant	Line color for values below the threshold.
⑩	Not significant	Line color for values not below the threshold.
⑪	Threshold	Value and color of the threshold line.
⑫	X-axis	Label of the X-axis.
⑬	auto	Auto-scale the X-axis.
⑭	specify	Specify the range of the X-axis.
⑮	Rotate yticks	Rotation of Y-axis tick marks.
⑯	Save graph	Save the graph in PNG or PDF.
⑰	Clear	Clear the figures
⑱	Redraw	Redraw the figures.
⑲	Statistics file	Select the file using the "Select statistics.txt" button on the right.
⑳	Outside file:	Display information about the k-mer sequence and its surrounding sequences in a popup window.
㉑	Outside file	Select the file using the "Select outside.txt" button on the right.

When "⑥Y-axis (lower)" is set to G-statistics, examples of the threshold ⑪ applied to G-statistics are shown in Table 11. The values are determined based on the chi-square distribution table with 1 degree of freedom.

Table 11 Chi-square distribution table with 1 degree of freedom (G-statistics)

Probability	0.050	0.025	0.010	0.005
G-statistics	3.84146	5.02389	6.63490	7.87944

The selection menu for "⑥Y-axis (lower)" allows to select the content and title of the Y-axis to be displayed (Figure 3).



Select the statistics to plot from the dropdown menu.

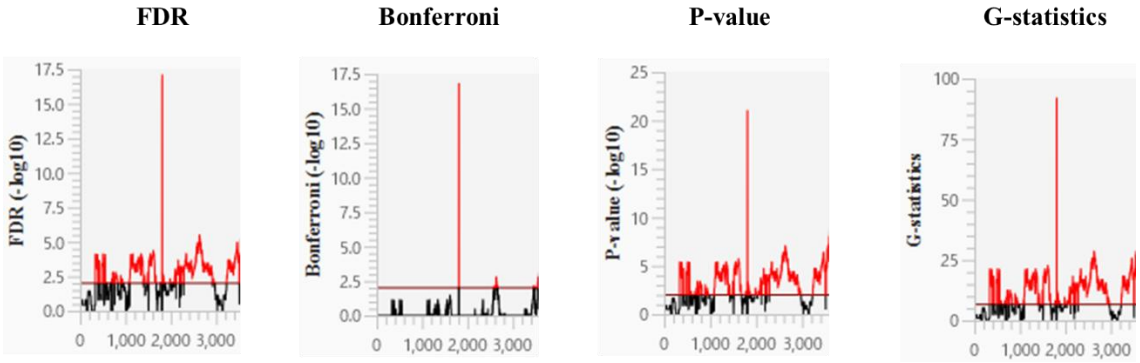


Figure 3　Content and title of the Y-axis.

To change the range of the horizontal axis of the graph, in addition to specifying it through "X-axis" > "specify" in Figure 2, the range can also be adjusted by selecting the desired area using drag and drop (Figure 4).

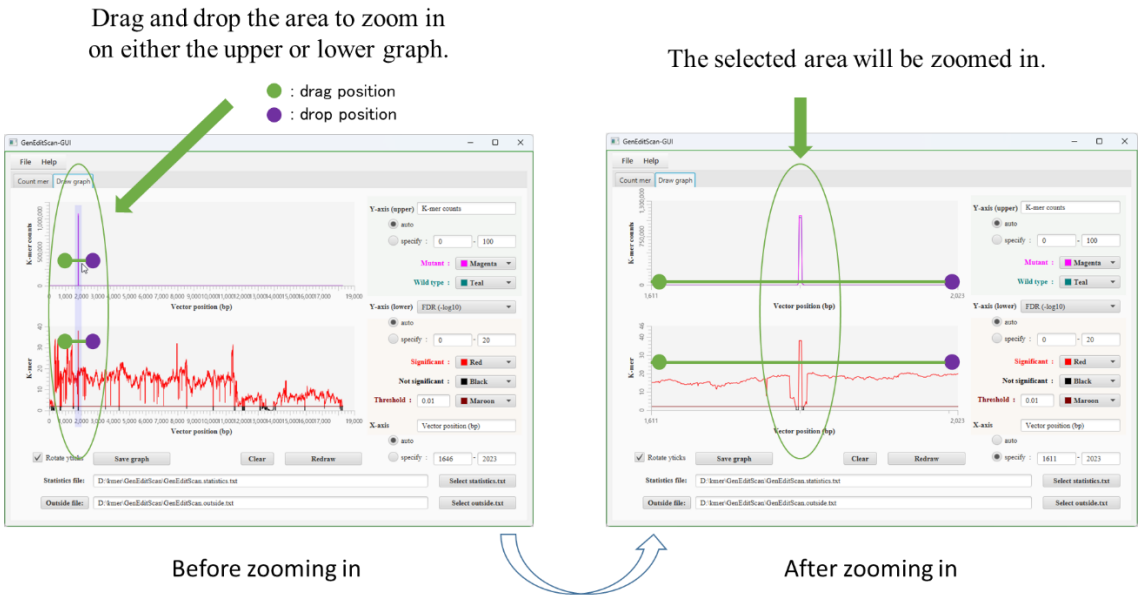


Figure 4　Select the range of the X-axis using drag and drop

To specify the maximum value of the vertical axis on the graph, you can use the "specify" option in "Y-axis (upper)" (upper screen) or "Y-axis (lower)" (lower screen) in Figure 2. Alternatively, the maximum value can be set by double-clicking on the desired position (Figure 5).

Figure 5　Specify the maximum value of the vertical axis by double-clicking

The "Select statistics.txt" button allows selecting the statistics.txt file, which contains the results of the G-test.

The "Select outside.txt" button allows selecting the outside.txt file, which contains the analysis results for the sequences surrounding the detected k-mer sequences.

The "Save graph" button allows choosing the format for saving the graph, either as a PDF or PNG.

9.2. Flanking sequence display

In the lower screen of the "Draw graph" tab, clicking on a region where the p-values are lower than the threshold displays the detected k-mer sequences and their surrounding sequence analysis results (Figure 6).

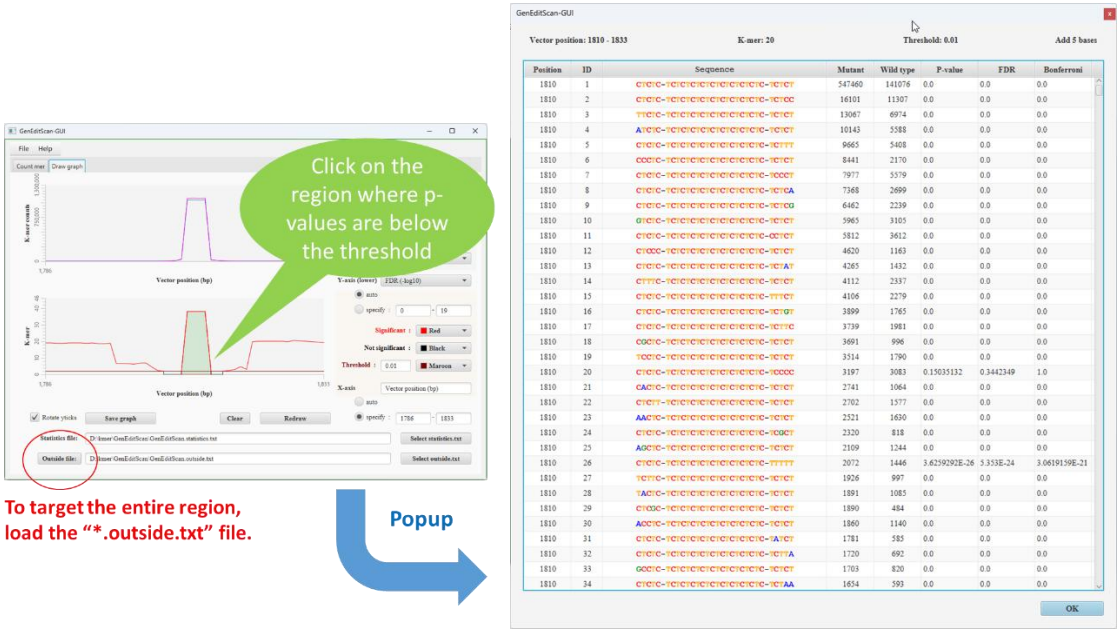


Figure 6　Display the k-mer sequences and their surrounding sequence analysis results

Clicking the ⑳ "Outside file" button in Figure 2 displays a popup window with information about the k-mer sequences and their surrounding sequences searched across the entire vector sequence.

The display content of the detected k-mer sequences and their surrounding sequence analysis results is shown in Figure 7. Each item can be sorted in ascending or descending order by clicking on it. Additionally, the sequences in the "Sequence" column can be copied by right-clicking.

All items are sortable.

Items	Content
Position	Position on the vector
ID	Serial number of the surrounding sequence pattern
Sequence	K-mer + surrounding sequences
Mutant	Frequency of the surrounding sequences in the mutant data
Wild type	Frequency of the surrounding sequences in the wild-type data
P-value	p-value
FDR	FDR-adjusted p-value (Benjamin-Hochberg)
Bonferroni	Bonferroni corrected p-value

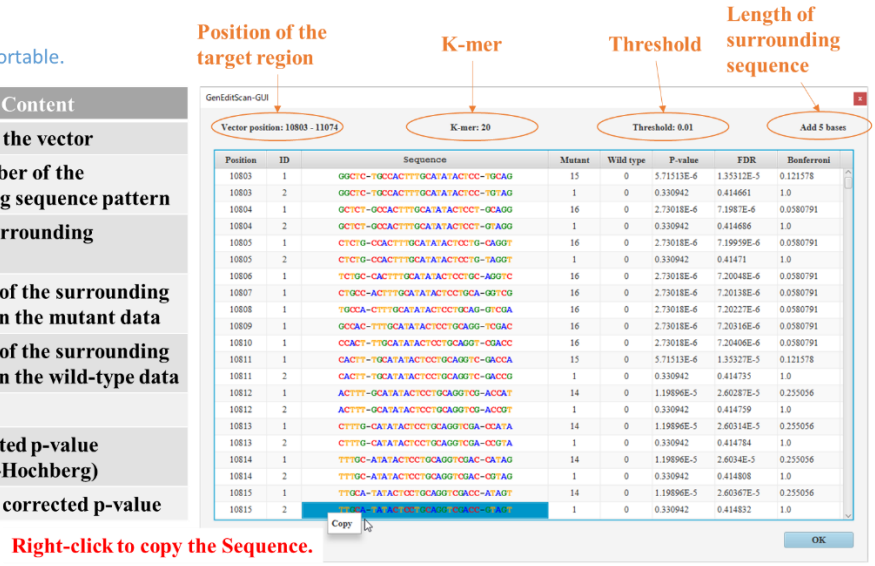


Figure 7 Example of the display content of the detected k-mer sequences and their surrounding sequence analysis results.

10. Menu

10.1. File

The "File" dropdown menu is shown in Figure 8, and its features are listed in Table 12. The screen state at the time of saving can be reproduced using the configuration file.

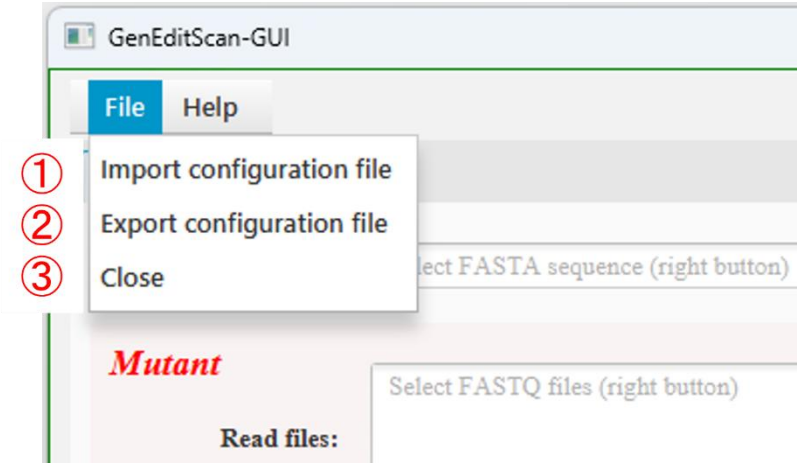


Figure8　Dropdown menu of "File" tab.

Table 12　Features of "File" dropdown menu.

No.	Display items	Features
①	Import configuration file	Import the configuration file
②	Export configuration file	Output the configuration file
③	Close	Exit the program

10.2. Help

The "Help" dropdown menu is shown in Figure 9, and its features are listed in Table 13. The User Guide is displayed using the application associated with PDF files on the PC being used.

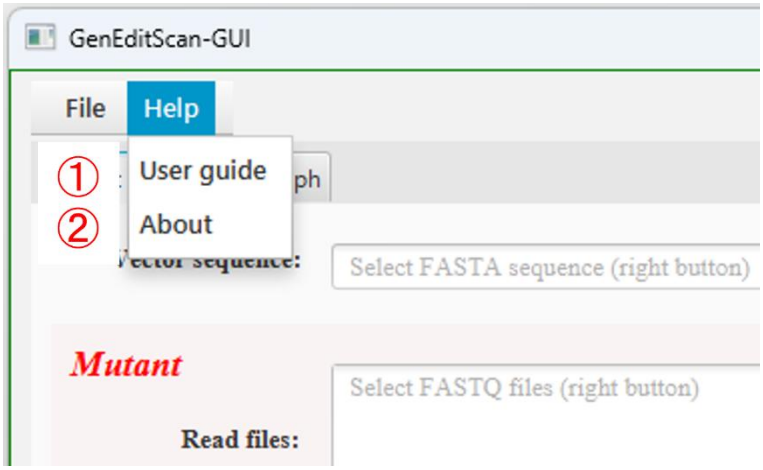


Figure9　Dropdown menu of "Help" tab.

Table 13　Features of "Help" dropdown menu.

No.	Display items	Features
①	User guide	Display the user guide (PDF)
②	About	Display the program's version and license information.

11. License

Copyright 2019-2024 National Agriculture and Food Research Organization (NARO)

12. Open source licenses

GenEditScan-GUI uses the tools listed in Table 14 as external libraries for saving vector-based PDF image files and calculating p-values.

Table 14　Open source licenses

Open source	Licenses
Apache Batik	Copyright 1999-2022 The Apache Software Foundation This product includes software developed at The Apache Software Foundation (http://www.apache.org/). This software contains code from the World Wide Web Consortium (W3C) for the Document Object Model API (DOM API) and SVG Document Type Definition (DTD). This software contains code from the International Organisation for Standardization for the definition of character entities used in the software's documentation. This product includes images from the Tango Desktop Project (http://tango.freedesktop.org/). This product includes images from the Pasodoble Icon Theme (http://www.jesusda.com/projects/pasodoble).
Apache FOP	Copyright 1999-2024 The Apache Software Foundation This product includes software developed at The Apache Software Foundation (http://www.apache.org/).
JFXConverter	Copyright (c) 2016, 2020 Herve Girod All rights reserved.

	Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met: 1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. 2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE. The views and conclusions contained in the software and documentation are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the FreeBSD Project. Alternatively if you have any questions about this project, you can visit the project website at the project page on https://sourceforge.net/projects/jfxconverter/

Colt	Copyright (c) 1999 CERN - European Organization for Nuclear Research. Permission to use, copy, modify, distribute and sell this software and its documentation for any purpose is hereby granted without fee, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation. CERN makes no representations about the suitability of this software for any purpose. It is provided "as is" without expressed or implied warranty.

13. Release notes

The development history of GenEditScan-GUI is shown in Table 15.

Table 15　Version history and release notes

Version	Date	Release notes
1.0.2	December 26, 2024	Refactoring of the source code.
1.0.0	February 29, 2024	・Reassigned version numbers due to the program name change. ・Added output for "sequences with upstream and downstream bases added to the k-mer sequence" in the Outside file. These sequences are intended for copying and performing database searches. ・Display elapsed time and calculation logs below the progress bar. ・Modified to assign the same FDR-adjusted p-value to identical p-values.
2.1.0-beta	February 25, 2022	・Added functionality for multiple comparison correction ・Added graph display options for p-values, and adjusted p-values.
1.2.0-beta	February 22, 2019	First release.