Krait User Manual

Version 0.8.6

August 22, 2017

Microsatellites Investigation and Primer Design

Lianming Du

Chengdu University

Table of Contents

1 Preface

1.1 Introduction

Krait is a robust and ultrafast tool with a user-friendly graphic interface for genome-wide investigation of microsatellites, which attempts to overcome the limitations of the currently available tools. Krait is written in Python and can be run as a standalone desktop application on Windows, Linux or Mac systems without dependencies. The microsatellite search engine is written in C and compiled as Python modules for import into Krait. Krait has many features: 1) Identification of perfect SSRs, imperfect SSRs (iSSRs), compound SSRs (cSSRs) and VNTRs from extremely large genome. 2) Locating the SSRs in gene coding region. 3) Design primer for microsatellite 4) Statistical analysis and plotting. 5) Supporting gzip compressed fasta as input file. 6) Supporting export FASTA, GFF3 or CSV. 7) Downloading DNA sequence from NCBI database.

1.2 License

Krait is distributed under a license called GNU General Public License, version 2 (GPL2).

1.3 Requirements

Krait is an operating system independent application and works on Windows, Linux and Mac systems. We recommend at least 1GB of RAM and 5GB of available hard disk space. In order to process very large genomes, a faster processor (64 bit) and larger amount of physical memory will be needed.

1.4 Installation

On windows, the preferred way to install Krait is to download the installer directly from https://github.com/lmdu/krait. Then double click the downloaded installer to install the program following the on-screen instructions.

Download the latest version of Krait for your operating system (Windows, Linux or Mac) from the following URL: and then decompress the downloaded file. On Windows, On Linux or Mac, just simply double click krait to run the program not required installation.

1.5 User Interface

The general components of the Krait main user interface are shown in Figure 1.

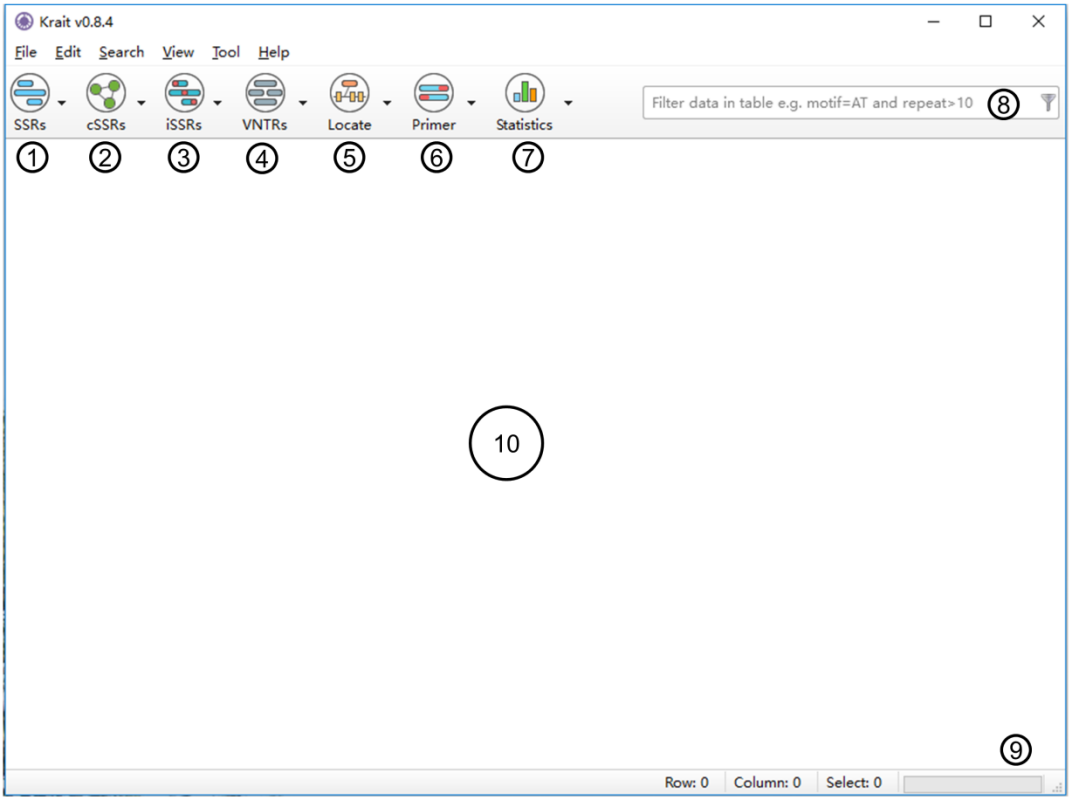

Figure 1: Krait main user interface

1. Search for perfect microsatellites or show results if you have searched.
2. Search for compound microsatellites or show results if you have searched.
3. Search for imperfect microsatellites or show results if you have searched.
4. Search for VNTRs or show results if you have searched.
5. Locate perfect, compound, imperfect microsatellites or VNTRs in coding region of genes.
6. Design primers for microsatellites screened from results.
7. Perform statistical analysis or show results if you have performed.
8. Filter results using user inputted conditions.
9. The task progressing.
10. Display the result of analysis in table.

2 Import sequences

To start a new analysis you just have to import your sequence data from files into Krait. The application

accepts text or gzip compressed files containing one or more DNA sequences in FASTA format (Figure 2). These files may have the extension .fasta, .fna, .fa, .txt, .fasta.gz, .fna.gz, .fa.gz. A sequence in FASTA format begins with a single-line description or header starting with a "**>**" character. The rest of the header line is arbitrary but should be informative. Avoid strange characters in the sequence header, such as '&' or '\' and use 'N' to denote in-determinations in the sequences.

```
>gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
```

Figure 2: An example for the FASTA format

2.1 Import One FASAT File

Step 1: Go to **File Menu -> Import Fasta**.

Step 2: Select a FASTA formatted file containing your sequences.

Step 3: Click **Open** to import FASTA file.

2.2 Import Multiple FASTA Files

Step 1: Put your FASTA files in a folder

Step 2: Go to **File Menu -> Import Fastas in Folder**.

Step 3: Select the folder to load FASTA files in that folder.

2.3 Import Fasta from NCBI

Krait provides a tool for user to download the FASTA formatted DNA sequence from NCBI nucleotide database (Figure 3).

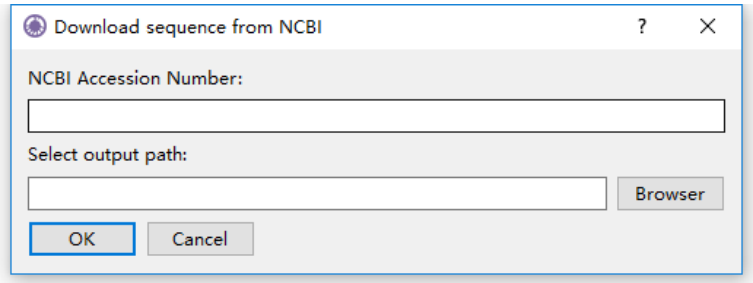

Figure 3: Downloading sequence from NCBI database

Step 1: Go to **Tool Menu -> Download Sequence from NCBI** to open download dialog.

Step 2: Input accession, version or GI number of a sequence in NCBI nucleotide database.

Step 3: Click **Browser** to select an output file path.

Step 4: Click **OK** to start download file and wait for finish.

Step 5: Once download finished, go to **File Menu -> Import Fasta** to select your downloaded file to import.

3 Search for Repeats

Krait allows user to identify perfect microsatellites (SSRs), compound microsatellites (cSSRs) and imperfect SSRs (iSSRs) as well as VNTRs.

3.1 Search for SSRs

3.1.1 Start SSR Search

Step 1: Import fasta sequence file (See 2 Input Files).

Step 2: Go to **SSRs (toobar) -> Specify Minimum Repeats** to specify minimum repeats.

Step 3: Click SSR search bubtton  to start search SSRs.

Tips: Search Menu -> Search for SSRs and **SSRs (toolbar) -> Search for SSRs** will remove the previous searched SSR results and then search for SSRs again.

3.1.2 SSR Search Results

After SSR search finished, a table containing results will be displayed. An example was shown in Figure 4. The result table contains 9 columns.

id	sequence	standard	motif	type	repeat	start	end	length
1	1	A	A	1	18	1833	1850	18
2	1	AG	AG	2	8	6264	6279	16
3	1	AAC	AAC	3	9	6481	6507	27
4	1	AG	AG	2	22	6812	6855	44
5	1	AC	CA	2	21	7007	7048	42

Figure 4: An example for SSR search result

id: unique identifier generated by Krait.

sequence: the name of sequence where SSR was found.

standard: the standardized motif.

motif: repeat unit of SSR.

type: SSR type: mononucleotide, dinucleotide etc corresponding to motif length.

repeat: number of repeats.

start, **end**: start and end position of SSR in original sequence.

length: the length of SSR (bp).

3.1.3 Show SSR Results

If you have searched SSRs, you can click SSR search button  or go to **SSRs (toolbar) -> Show Perfect SSRs** or go to **View Menu -> Show Perfect SSRs** to display SSR results in table.

3.1.4 Remove SSR Results

You can go to **SSRs (toolbar) -> Remove Perfect SSRs** or **View Menu -> Remove Perfect SSRs** to remove searched SSR results.

3.2 Search for cSSRs

3.2.1 Start cSSR Search

Step 1: Search perfect SSRs (See 3.1.1).

Step 2: Go to **cSSRs -> Specify Maximum Distance** to specify a maximum distance (dMAX) allowed between two perfect SSRs.

Step 3: Click cSSR search button  to start search cSSRs.

Tips: Search Menu -> Search for cSSRs and **cSSRs (toolbar) -> Search for cSSRs** will remove the previous searched cSSR results and then search for cSSRs again.

3.2.2 cSSR Seach Results

After cSSR search finished, a table containing results will be displayed. An example was shown in Figure 7. The result table contains 10 columns.

id	sequence	start	end	motif	complexity	length	gap	component	structure
1	1	13846	13882	GT-TG	2	36	1	22-23	(GT)8-(TG)10
2	1	15503	15547	AC-CA	2	44	1	33-34	(AC)11-(CA)11
3	1	15561	15634	AC-CA-AC	3	70	4	35-37	(AC)12-(CA)12-(AC)11
4	1	16924	16969	GT-GT	2	44	2	41-42	(GT)7-(GT)15
5	1	23023	23063	TG-GT	2	40	1	59-60	(TG)12-(GT)8

Figure 5: An example for cSSR search result

id: unique identifier generated by Krait.

sequence: the name of sequence where cSSR was found.

start, **end**: start and end position of cSSR in original sequence.

motif: the motifs of individual SSR in a cSSR.

complexity: the number of individual SSR in a cSSR.

length: the length of cSSR.

gap: total bases of distance between two SSR.

component: the SSR id of individual SSRs comprised cSSR.

structure: the detail of SSRs comprised cSSR.

3.2.3 Show cSSR Results

If you have searched cSSRs, you can click cSSR search button  or go to **cSSRs (toolbar) -> Show Compound SSRs** or **View Menu -> Show Compound SSRs** to dispay cSSR result in table.

3.2.4 Remove cSSR Results

You can go to **cSSRs (toolbar) -> Remove Compound SSRs** or **View Menu -> Remove Compound SSRs** to remove searched cSSR results.

3.3 Search for iSSRs

3.3.1 Start iSSR Search

Step 1: Import fasta sequence file (See 2 Input Files).

Step 2: Go to **iSSRs (toobar) -> Specify Search Parameters** to specify minimum length and repeats of seed, maximum consecutive edits, gap penalty and minimum score required.

Step 3: Click cSSR search bubtton  to start search cSSRs.

Tips: Search Menu -> Search for iSSRs and **iSSRs (toolbar) -> Search for iSSRs** will remove the

previous searched iSSR results and then search for iSSRs again.

3.3.2 iSSR Search Results

After iSSR search finished, a table containing results will be displayed. An example was shown in Figure 6. The result table contains 13 columns.

id	sequence	standard	motif	type	start	end	length	match	subsitution	insertion	deletion	score
1	1	A	A	1	1831	1850	20	19	1	0	0	18
2	1	AAAC	TTTG	4	3225	3250	26	24	1	1	1	19
3	1	AAAC	TGTT	4	4261	4290	30	25	3	0	2	18
4	1	AAT	AAT	3	4892	4913	22	19	3	0	0	16
5	1	AG	AG	2	6251	6279	29	23	6	0	0	17

Figure 6: An example for iSSR search result

id: unique identifier generated by Krait.

sequence: the name of sequence where SSR was found.

standard: the standardized motif.

motif: repeat unit of SSR.

type: SSR type: mononucleotide, dinucleotide etc. corresponding to motif length.

start, **end**: start and end position of SSR in original sequence.

length: the length of SSR (bp).

match: the number of matches.

substitution: the number of substitutions.

insertion: the number of insertions.

deletion: the number of deletions.

score: the score of iSSR.

3.3.3 Show iSSR Results

If you have searched iSSRs, you can click iSSR search button  or go to **iSSRs (toolbar) -> Show Imperfect SSRs** or **View Menu -> Show Imperfect SSRs** to dispay iSSR result in table.

3.3.4 Remove iSSR Results

You can go to **iSSRs (toolbar) -> Remove Imperfect SSRs** or **View Menu -> Remove Imperfect SSRs** to remove searched cSSR results.

3.4 Search for VNTRs

3.4.1 Start VNTR Search

Step 1: Import fasta sequence file (See 2 Input Sequences).

Step 2: Go to VNTRs (toobar) -> Specify Search Parameters to specify minimum and maximum length of motif and minimum repeats.

Step 3: Click VNTR search bubtton  to start search VNTRs.

Tips: Search Menu -> Search for VNTRs and **VNTRs (toolbar) -> Search for VNTRs** will remove the previous searched VNTR results and then search for VNTRs again.

3.4.2 VNTR Search Results

After VNTR search finished, a table containing results will be displayed. An example was shown in Figure 7. The result table contains 8 columns.

id	sequence	motif	type	repeat	start	end	length
1	1	AAAAAAA	7	2	1833	1846	14
2	1	GGGTTAAA	8	2	5106	5121	16
3	1	AGAGAGAG	8	2	6264	6279	16
4	1	AACAACAAC	9	3	6481	6507	27
5	1	AGAGAGAG	8	5	6812	6851	40

Figure 7: An example for VNTR search result

id: unique identifier generated by Krait.

sequence: the name of sequence where SSR was found.

motif: repeat unit of SSR.

type: VNTR type, motif length.

repeat: number of repeats.

start, end: start and end position of SSR in original sequence.

length: the length of SSR (bp).

3.4.3 Show VNTR Results

If you have searched VNTRs, you can click VNTR search button  or go to **VNTRs (toolbar) -> Show VNTRs** or go to **View Menu -> Show VNTRs** to display VNTR results in table.

3.4.4 Remove VNTR Results

You can go to **VNTRs (toolbar) -> Remove VNTRs** or **View Menu -> Remove VNTRs** to remove searched VNTR results.

3.5 Settings for Search

The setting panel also can be opened by **Edit Menu -> Preferences**. The setting panel was shown in Figure 8.

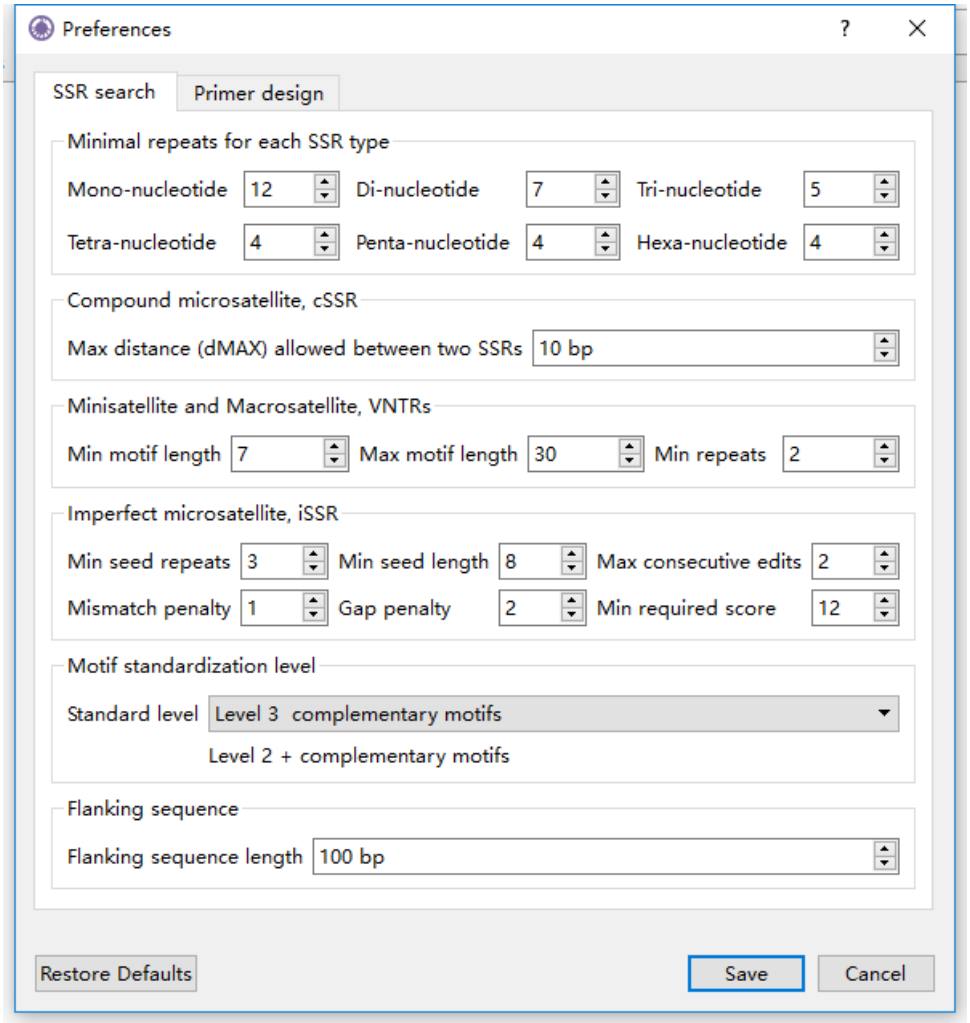

Figure 8: The setting panel for search

The setting panel contains 6 subpanels:

Minimum repeats Panel: specify minimum repeats for each type (di- to hexa-nucleotide) of perfect SSRs.

cSSR Panel: specify maximum distance allowed between two perfect SSRs.

VNTR Panel: set minimum and maximum length of motifs and minimum repeats.

iSSR Panel: set minimum repeats and length of extended seed, maximum consecutive edits allowed in an extension, mismatch penalty, gap penalty and minimum required score.

Motif Standardization Level Panel: set the motif standardization level (0-4), the level only affects the result of perfect and imperfect SSRs search.

Flanking Sequence Panel: set the flanking sequence length, used to design primer, export FASTA and display the details of SSR.

4 Locating Repeats in Genes

4.1 Start Locating Repeats

Step 1: Search SSRs, cSSRs, iSSRs or VNTRs (See 3 Search for Repeats)

Step 2: Go to **Locate (toolbar) -> Import Annotation File** to select a GTF or GFF file corresponding to the imported sequence file.

Step 3: Click the locate button  to start locating.

4.2 Locating Results

If locating task finished, the repeats located in gene exon, intron, CDS or UTRs will be marked as different colors in table. An example was shown in Figure 9.

id	sequence	standard	motif	type	repeat	start	end	length
1	1	A	A	1	18	1833	1850	18
2	1	AG	AG	2	8	6264	6279	16
3	1	AAC	AAC	3	9	6481	6507	27
4	1	AG	AG	2	22	6812	6855	44
5	1	AC	CA	2	21	7007	7048	42
6	1	AC	CA	2	16	7592	7623	32
7	1	ACGC	CACG	4	4	7636	7651	16
8	1	AC	GT	2	29	7975	8032	58
9	1	AC	GT	2	16	9521	9552	32
10	1	AC	AC	2	9	9707	9724	18

Figure 9: An example for locating result

4.3 Select by Color

This function allows selection of repeats on the basis of their color i.e. the region in where the repeat

located. You can go to **Locate (toolbar) -> Show repeats in CDS, Exon, UTR or Intron** to screen as shown in Figure 10.

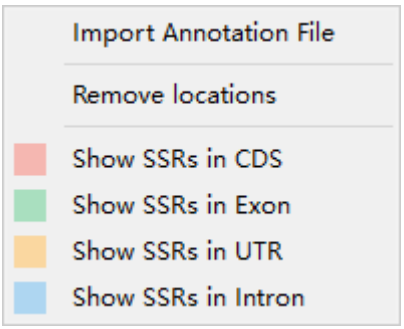

Figure 10: Selection by Colors

4.4 Remove Locating Results

You can go to **Locate (toolbar) -> Remove Locations** to remove locating results.

5 Primer Design

5.1 Start Primer Design

Step 1: Select tandem repeats in table that you want to design primer.

Step 2: Go to **Primer (toolbar) -> Specify Primer3 Setttings** to set parameters for primer3.

Step 3: Click design primer button  to start design primer for selected tandem repeats.

5.2 Primer Design Results

After primer design, a table contains primer details will be display. The result table contains 12 columns. An example for primer design was shown in Figure 11.

id	target	entry	product	forward	tm1	gc1	stability1	reverse	tm2	gc2	stability2
1	issr-1	1	100	CCACAATAAC...	59.05	52.38	4.96	AAGCGCTCAC...	60	47.62	4.09
2	issr-1	2	100	CCACAATAAC...	59.05	52.38	4.96	AAGCGCTCAC...	58.56	50	4.7
3	issr-1	3	101	CCACAATAAC...	59.05	52.38	4.96	AAAGCGCTCA...	59.19	47.62	4.7
4	issr-1	4	101	CCACAATAAC...	59.05	52.38	4.96	AAAGCGCTCA...	58.1	45	4.58
5	issr-1	5	103	CCACAATAAC...	59.05	52.38	4.96	CAAAAGCGCT...	60.03	45.45	4.58

Figure 11: An example for primer design

id: unique identifier generated by Krait.

target: ssr type added id of the tandem repeat.

entry: the primer number for tandem repeats. A tandem repeats can contains multigroup primers.

product: the product size of primer.

forward: the sequence of forward primer.

reverse: the sequence of reverse primer.

tm1,2: the temperature of the forward primer and reverse primer.

gc1,2: the GC content of the forward primer and reverse primer.

stability1,2: the end stability of the forward primer and reverse primer.

5.3 Show or Remove Primer Design Results

If you have designed primers, you can go to **Primer (toolbar) -> Show Designed Primer** to display primer results. The primer results can be removed by **Primer (toolbar) -> Remove Designed Primer**.

5.4 Settings for Primer Design

The settings for primer design shown in Figure 12 are the same with Primer3 tags. Click the setting name will redirect to the corresponding tags in Primer3 manual http://primer3.sourceforge.net/primer3_manual.htm.

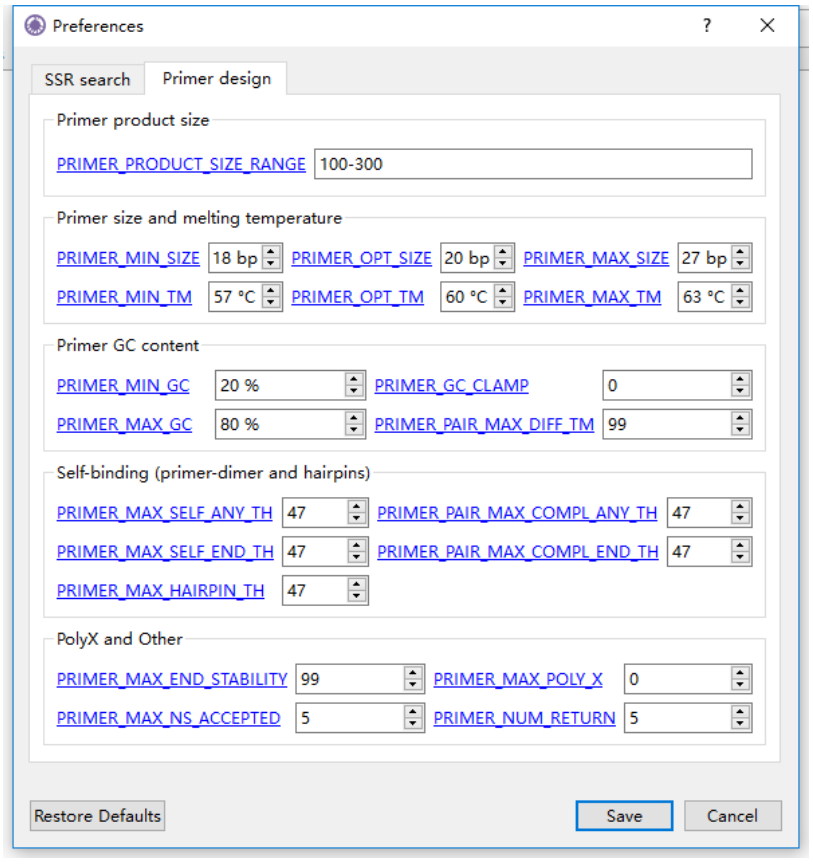

Figure 12: Setting panel for primer design

6 Filter Data in Table

6.1 Start Filter Data

Step 1: Verify that the current display is a table with rows.

Step 2: Input the conditions into input box as shown in Figure 13.

Step 3: Press **Enter** to filter rows in table.

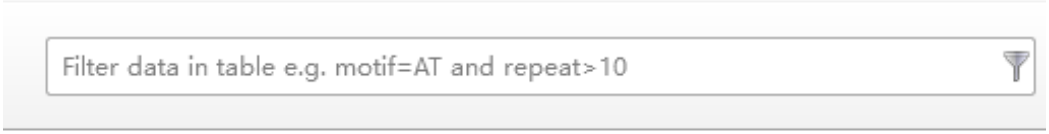

Figure 13: Filter condition input box.

6.2 Construct Filter Conditions

Krait provides comparison operators and logical operators that allow you to construct the condition. The comparison operators include **=, >, <, >=, <=**; the logical operators include **AND**, **OR**. There is a special operator **IN** can be used. The condition is case-insensitive. Thus you can use column name and operators to construct a condition to filter table rows. The numeric column can be allowed to use all comparison operators and the text column only be allowed to use **=** and IN. Examples:

1) Suppose you want to find SSRs whose motif is ATG, you can construct a condition:

   motif=ATG.

2) If you want to find SSRs whose motif is ATG and repeats greater than 10, you can construct a condition:

   motif=ATG and repeat>10

3) If you want to find SSRs whose motifs are ATG, AT, AAAG and repeats between 10 and 15, you can construct a condition:

   motif in (ATG,AT,AAAG) and repeats>=10 and repeats<=15

4) If you want to find SSRs whose length less than 12 or length greater than 20, you can construct a condition:

   length<12 or length>20

7 Statistical Analysis

7.1 Perform statistical Analysis

If you have searched SSRs, cSSRs, iSSRs or VNTRs, you can just click the statistics button  to start statistical analysis.

If you have performed statistical analysis, click the statistics button will display the statistical results. If you want to perform statistical analysis again, you can go to **Statistics (toobar) -> Statistical Analysis**.

7.2 Statistical Analysis Results

Once statistical analysis finished, a statistical report contains several tables and graphs will be generated and displayed as shown in Figure 14.

The summary statistical analysis report

Item	Description	Number
Total number of sequences	Fasta	1061
Total length of sequences	A+T+C+G+N	1371719383 bp
Total valid length of sequences	A+T+C+G	1369631918 bp
Unkown bases (Ns) in sequences	2087465 bp	0.15%
GC content	(G+C)/(A+T+C+G) not include Ns	36.64%

The summary of perfect SSRs

Item	Description	Number
Total number of perfect SSRs	counts	1045912
Total length of perfect SSRs	bp	32616790
The average length of SSRs	total ssr length/total ssr counts	31.19
SSRs per sequence	total SSR counts/sequence counts	986
The percetage of sequence occupied by SSRs	ssr total length/total sequence size	2.38
Relative abundace	total SSRs/total valid length	763.64 loci/Mb
Relative density	total SSR length/total valid length	23814.27 bp/Mb

Figure 14: A part of statistical report

You can go to **File Menu -> Export Statistical Report** to export PDF formatted statistical report.

7.3 Show or Remove Statistical Results

If you have performed statistical analysis, you can go to **Statistics (toolbar) -> Show Statistical Result** to display statistical results. The statistical results can be removed by **Statistics (toolbar) -> Remove Statistical Result**.

8 Export Output Files

Krait allows user to export Table (CSV, GTF, GFF) and FASTA formatted files.

8.1 Export Table

Step 1: Select rows in table that you want to export.

Step 2: Go to File Menu -> Export Selected as Table and select a format (CSV, GTF, GFF or TEXT) and provide an output file name.

Step 3: Click **Save** to export to output file.

8.2 Export FASTA

Step 1: Select rows in table that you want to export.

Step 2: Go to **File Menu -> Export Selected as Fasta** and provide an output file name.

Step 3: Click **Save** to export to output file

9 Feedback

Report issues: https://github.com/lmdu/krait/issues.

Contact email: adullb@qq.com