Tutorial on Organelle Genome Recombination Detection and Recombinant Map Construction

——A Guide to Using ReHRI and ReHRV Software

# 1、Install

## 1.1、Download

The URL is: https://github.com/wlqg1983/ReHRI\_ReHRV\_1.0

## 1.2、Unzip the compressed file and enter the folder

unzip ReHRI\_ReHRV\_1.0-main.zip

cd ReHRI\_ReHRV\_1.0-main

## 1.3、Install ReHRI and ReHRV

Currently, ReHRI and ReHRV can only run on Linux (v20.0.4) systems and require conda (24.5.0) as a dependency. Below are the installation commands. After installation, the user must activate the conda environment to use ReHRI and ReHRV. If the software download speed is slow during installation, the user can manually modify the channels in the ReHRI\_ReHRV\_1.0.yml file.

conda env create -f ReHRI\_ReHRV\_1.0.yml

conda activate ReHRI\_ReHRV\_1.0

"ReHRI\_ReHRV\_1.0" is a user-defined conda environment name. The environment name must start with a letter and can only contain letters, numbers, underscores, or periods (no spaces or other special characters).

Update plasmidrender：

cp -r bin/plasmidrender $(conda info --base)/envs/ReHRI\_ReHRV\_1.0/lib/  
python3.12/site-packages/ (This command should be entered as a single line.)

chmod +x bin/\*

## 1.5、Verify the installation results

**Testing ReHRI.py help documentation**

python bin/ReHRI.py -h

**usage:** ReHRI.py [-h] -c CONFIG [-redo] [-resume] [-v]

**ReHRI:** A tool to check spanning reads for supporting subconfig of your organelle genome.

**Options:**

-h, --help show this help message and exit

-c, CONFIG Path to external configuration file.

-redo Delete all previous results and start calculation anew.

-resume Resume from a previous project.

-v, --version Show the version number and exit.

**Testing ReHRI.py version**

python bin/ReHRI.py -v

ReHRI version=1.0

**Testing ReHRV.py help documentation**

python bin/ReHRV.py -h

**usage:** ReHRV.py [-h] -c CONFIG [-redo] [-v]

**ReHRV:** A tool to map the conFig. of your organelle genome.

**Options:**

-h, --help show this help message and exit

-c, CONFIG Path to external configuration file.

-redo Delete all previous results and start calculation anew.

-v, --version Show the version number and exit.

**Testing ReHRV.py version**

python bin/ReHRV.py -v

ReHRV version=1.0

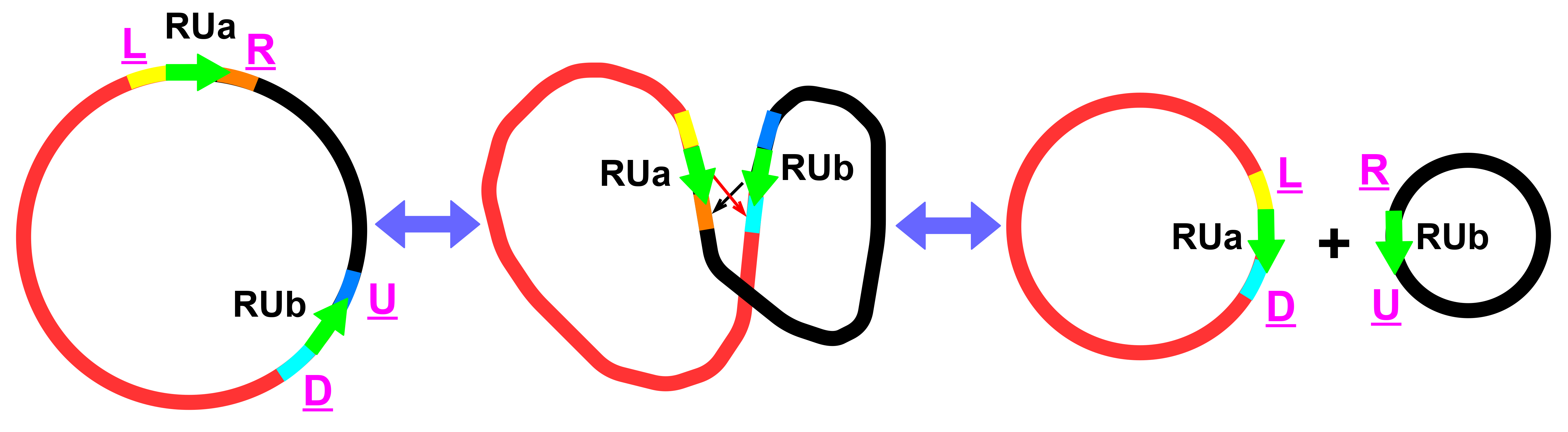
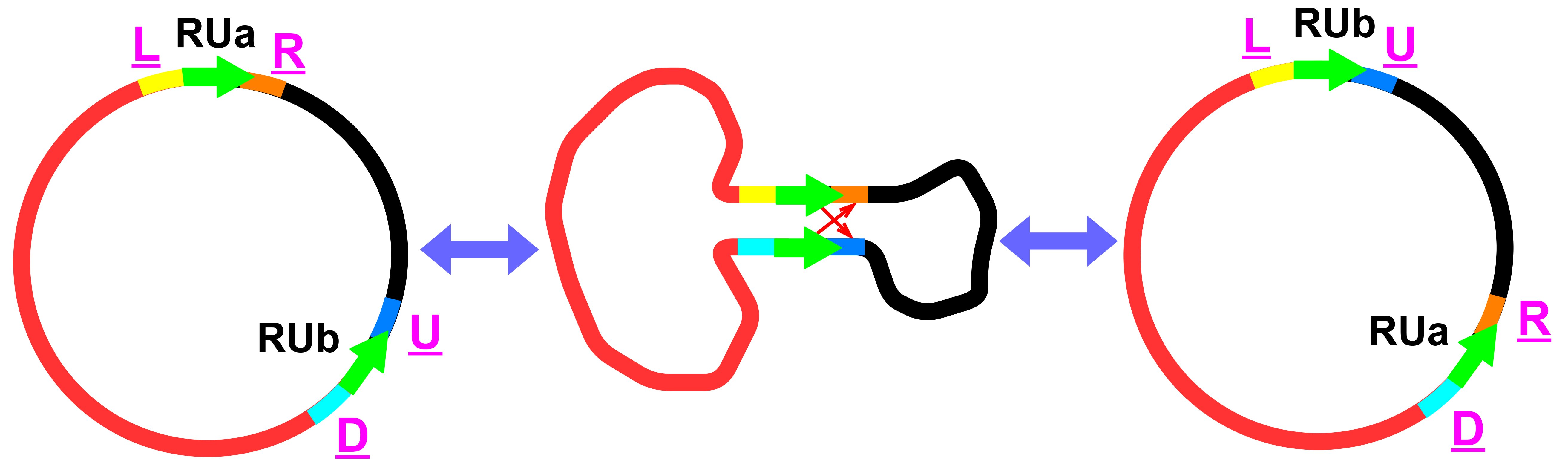
# 2、Software operating principle

## 2.1、ReHRI operating principle

ReHRI (Repeat-mediated Homologous Recombination Identification) operates on the principle that recombination between inverted repeats (IRs) inverts the intervening sequence (Fig. 2-1A), while recombination involving direct repeats (DRs) produces a pair of subgenomic molecules (Fig. 2-1B).

**A**

**B**

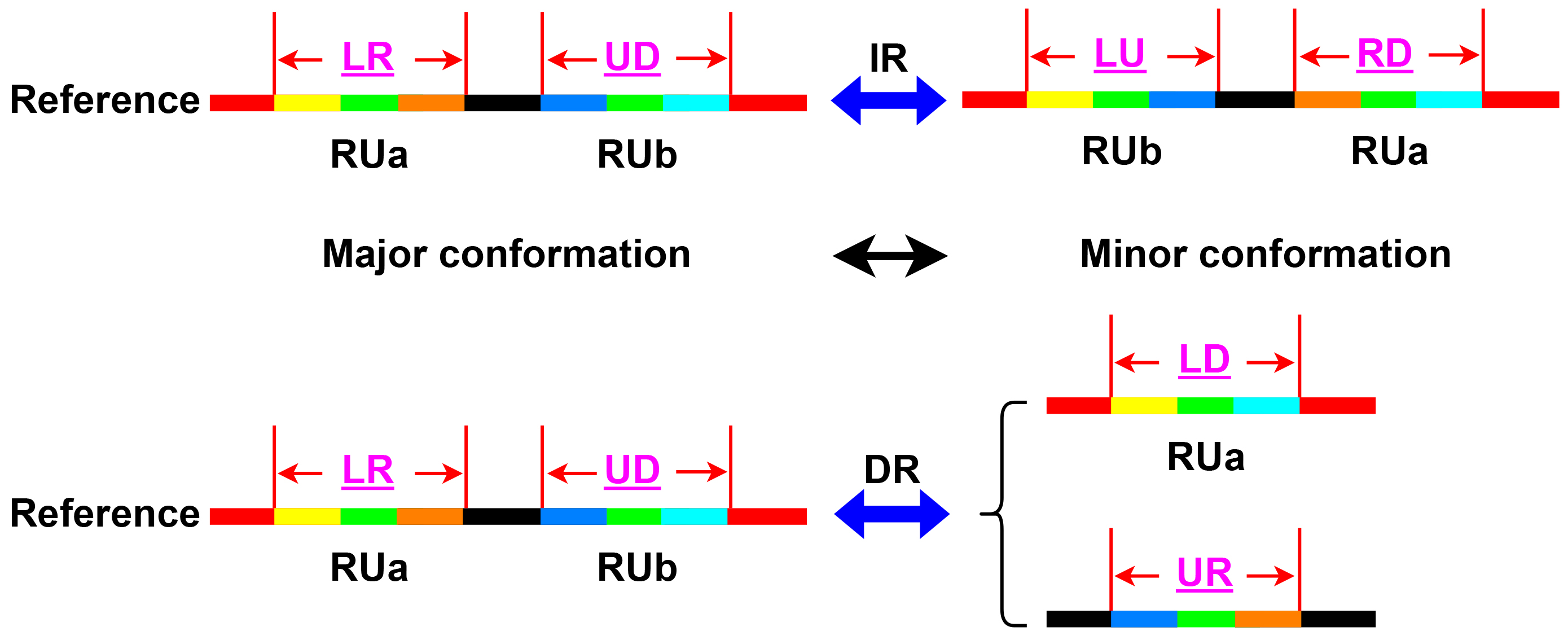


**B**

**Fig. 2-1 Schematic diagram of repeat-mediated circular genome recombination**

A: Recombination mediated by inverted repeats (IRs); B: Recombination mediated by direct repeats (DRs).

Taking the repetitive sequence as the center, a Trimed Reference Sequence (TRS) is extracted from the major configuration (main configuration) and minor configuration (sub configuration), as shown in Fig. 2-2. In the major configuration, paired repetitive sequences RUa and RUb are taken as the center, and TRS is extracted as LR and UD, labeled as TRSLR and TRSUD. In the secondary configuration mediated by IR, TRSLU and TRSRD were truncated with repeat sequences RUb and RUa as centers, respectively (Fig. 1-1A). In the DR mediated secondary configuration, TRSLD and TRSUR were truncated with repeat sequences RUa and RUb as the centers, respectively (Fig. 2-2B).



**A**

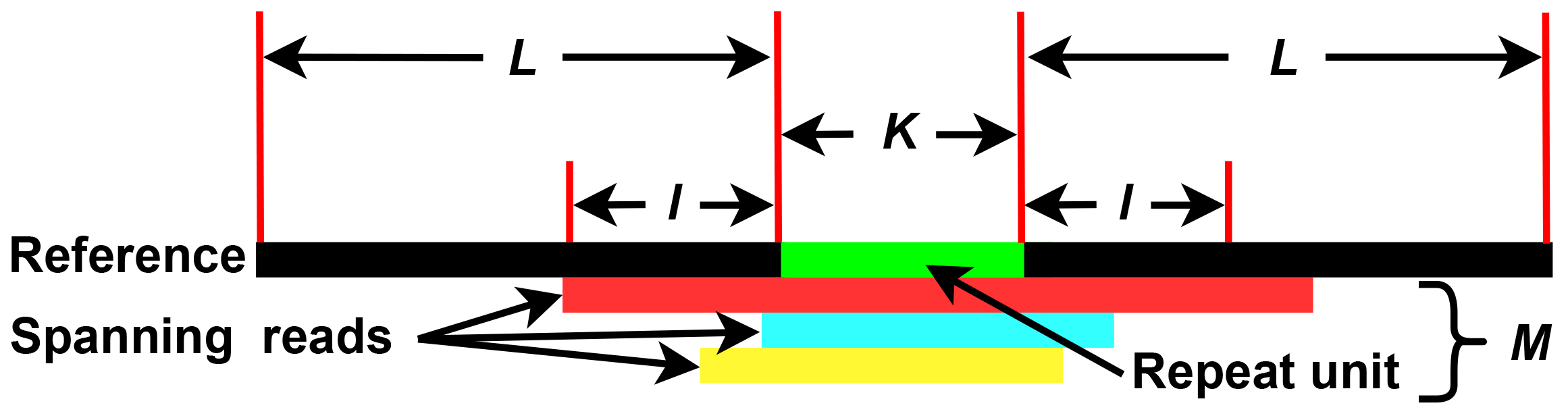
**A**

**B**

**Fig. 2-2 Schematic diagram of TRS interception**

Inverted repeat sequences (IRs) - mediated recombination causes inversion of the intermediate sequence of (A) paired repeat units, while directed repeat sequences (DRs) - mediated recombination causes (B) a chromosome to produce a pair of subtyping chromosomes. LR, UD, LD, UR, LU, and RD represent the truncated TRS. LR and UD come from the main configuration. LD and UR are derived from DRs mediated sub configurations. LU and RD are derived from IRs mediated sub configurations. Main configuration: major configuration = main configuration, minor configuration = sub configuration.

As shown in Fig. 2-3, in both the major and minor configurations, *L* base pairs are extracted on either side centered around the repeat sequence to obtain the TRS. ReHRI then maps the sequenced reads to the TRS and searches for reads that can span the repeat sequence (*l* > 0) from those mapped to the TRS. If any read spans the repeat sequence (i.e., *M* > 0), it is considered that the genomic configuration corresponding to that TRS exists.



## Fig. 2-3 Schematic diagram of mapping read to TRS

## *L*: Indicates the length of a sequence intercepted on both sides of a repeat unit. The default value is 1000 base pairs (bp).

## *K*: The length of a repeat unit. The default value is 5 bp.

## *I*: Represents the length of a read across the left and right sides of a repeat unit. When *I* ≥ 1 base pair (bp), the read is considered to span a repeat unit.

## *M*: The number of reads that span a repeat unit and *I* ≥ 1 bp.

## 2.2、Definition and calculation of recombination rates

When calculating the probability of mitochondrial genome recombination mediated by repeat sequences, the number of reads across repeat units RUa and RUb in TRSLR and TRSUD is recorded as *M*LR and *M*UD in the main configuration. In the secondary configuration, the number of reads for RUa and RUb across TRSLD and TRSUR, as well as repeat units in TRSLU and TRSRD, is recorded as *M*LD and *M*UR, and *M*LU and *M*RD.

The probability of repeat unit RUa mediated recombination in IRs mediated genome recombination is:



The above is the calculation formula for the recombination rate considering two cases: single chain (formula 2-1) and double chain (formula 2-2) (+/- represents positive and negative chains, the same below).

The probability of RUb mediated recombination in IRs mediated genome recombination is:



The above is the calculation formula for the recombination rate considering two cases: single chain (formula 2-3) and double chain (formula 2-4).

The probability of RUa mediated recombination in DRs mediated genome recombination is:



The above is the calculation formula for the recombination rate considering two cases: single chain (formula 2-5) and double chain (formula 2-6).

The probability of RUb mediated recombination in DRs mediated genome recombination is:



The above is the calculation formula for the recombination rate considering two cases: single chain (formula 2-7) and double chain (formula 2-8).

# 3、Running ReHRI software

python bin/ReHRI.py -c ReHRI.config.ini

python bin/ReHRI.py -c ReHRI.config.ini -redo

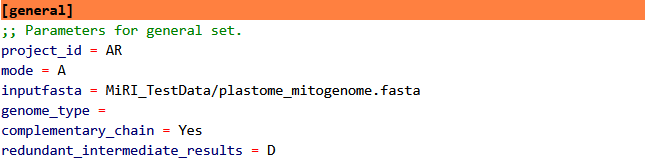
python bin/ReHRI.py -c ReHRI.config.ini -resume

ReHRI requires a number of parameters to run, and the configuration file ReHRI.config.ini is used to set various parameters, but most parameters can be set to default values, and only a few parameters need to be set. When the program is unexpectedly interrupted, the parameter "-redo" allows the user to delete previous intermediate results and recalculate, while the parameter "-resume" allows the user to continue calculating based on previously incomplete results until the final result is obtained.

## 3.1、Operating mode 1 of ReHRI

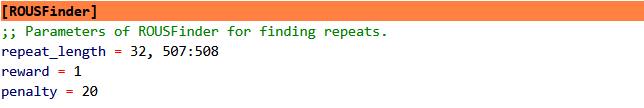
When mode=A (Fig. 3-1), it is the first operating mode of ReHRI. ReHRI will first search for forward and reverse repeat sequences within the genome, and then detect repeat sequence pairs that can mediate genome recombination. At this point, the user must provide the genome sequence file, the length of the duplicated sequence being searched for, the sequencing file, and the parameters for identifying the existence of secondary configurations.

The genome sequence file (inputfasta) should be in fasta format, and it is necessary to indicate whether the genome type (genomeotype) is linear (L) or circular (C). When the genome contains multiple chromosomes, place all chromosomes in the same fasta file, and use default values for other parameters, as shown in Fig. 3-1.



**图3-1** **Parameters in [general]**

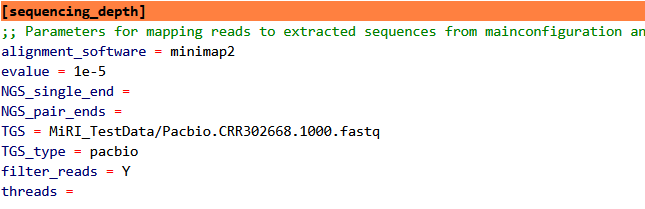
When searching for a duplicate sequence, its length (repeat length) is the K value in Fig. 2-2, which can be set as an interval, such as 50bp ≤ length ≤ 1000bp, set to 50:1000; Length ≥ 50bp, set to 50:. It can also be set to several lengths, such as 50bp, 100bp, and can be set to 50,100 (Fig. 3-2). Commas and colons must be in English, spaces are not mandatory. Other parameters can use default values.



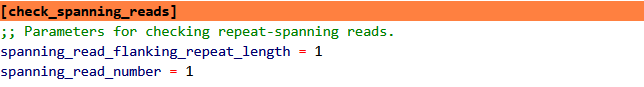
**图3-2 Length of repeat units in [ROUSFinder]**

For sequencing data, ReHRI can accept second-generation (NGS) and third-generation (TGS) data, and only one of NGS and TGS data can be accepted at a time (Fig. 3-3). When providing dual ended data, the files of the dual ended data should be separated by spaces. When providing TGS data, it is also necessary to indicate whether the data comes from the Nanopore sequencing platform (ont) or the Pacbio sequencing platform (pacbio) (Fig. 3-3).

The parameters used to detect the presence of secondary configurations are mainly spanning\_dead\_flanking\_depeat\_length and spanning\_dead\_numbers (Fig. 3-4). This is an important parameter for determining the existence of secondary configurations, with a default value of 1. It can be reset in the subsequent re filtering mode (refiltermode=Y) to perform multiple screenings on the results of genome recombination mediated by the queried repeated sequences. Other parameters can use default values.



**Fig. 3-3 Alignment softwares and sequencing data in [sequencing\_depth]**



**Fig. 3-4 Identification parameters of minor configurations in [check\_stpanning\_deads]**

Spanning\_dead\_flanking\_depeat\_length is the length of the read after crossing the repeated sequence, which is the value of *l* in Fig. 2-3. Spanning-read\_number is the number of reads with a length ≥ *l* after crossing duplicate units, which is the M value in Fig. 2-3. It is suggested that both parameter values be set to 1, and the obtained results can be re filtered in Mode 3.

In the comparison software, minimap2 is the default software. The applicable scenarios of minimap2, bwa, and blast are shown in Table 3-1.

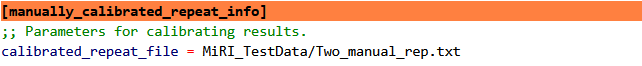
**Fig. 3-1 Comparison of Applicable Scenarios for Minimap2, BWA, and Blast**

|  |  |  |
| --- | --- | --- |
| **Tool** | **Best Application Scenario** | **Possible Missed Detection Reasons** |
| minimap2 | Long reads, genome assembly, fast alignment | Short reads, high-repeat regions, default parameters |
| BWA | Short reads, variant detection | Low efficiency with long reads |
| BLAST | Homology search, cross-species comparison | Time-consuming, not suitable for large-scale alignment |

## 3.2、Operating mode 2 of ReHRI

The task of querying duplicate sequences is extremely challenging, as different duplicate sequence results have a significant impact on the results of mediating genome recombination, and the results of finding duplicate sequences by different algorithms are also different. So, ReHRI has set up an interface that accepts users to provide duplicate sequence information, allowing users to provide duplicate sequences themselves (Fig. 3-5). At this point, it is necessary to set mode=C (Fig. 3-6). The repeat information file provided by the user is in tsv format (Fig. 3-7).

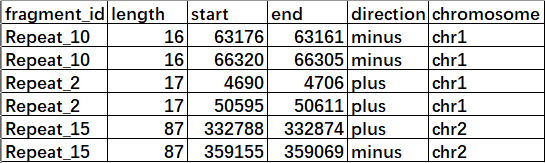
Paired duplicate sequences using the same fragment\_id. When there is only one chromosome in the genome, the chromosome column needs to be removed. ReHRI will detect all pairwise combinations of repeated sequence units with the same fragment\_id. If users want to specifically detect the mediating effect of certain paired repeat units on genome recombination, they can provide ReHRI with a paired repeat unit information file in the format shown in Fig. 3-8. When there is only one chromosome in the genome, the chromosome and paired chromosome columns need to be removed.



**Fig. 3-5 User-provided repeat information in [manually\_calibrated\_peat\_info]**

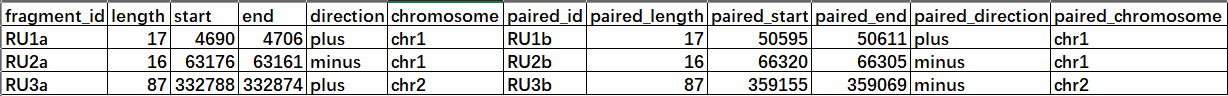


**Fig. 3-6 Parameters for ReHRI entering mode two**



**Fig. 3-7 Example of repeat information provided by users (tsv format)**

chr1, chr2, chr3, ...: For chromosome numbering, this format must be used, and the numbering order represents the arrangement order of chromosomes in the inputfasta file in [general]. The fragment\_id of different repeat units of the same repeat sequence must be the same.

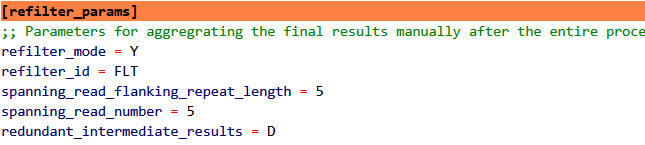


**Fig. 3-8 Example of paired repeat information provided by users (tsv format)**

chr1, chr2, chr3, ...: For chromosome numbering, this format must be used, and the numbering order must be the same as the arrangement order of chromosomes in the inputfasta file in [general].

## 3.3、Operating mode 3 of ReHRI

If the user are not satisfied with the initial filtering results, the user can set refiter\_made=Y (Fig. 3-9), reset the [spanning\_dead\_flanking\_depeat\_length] and [spanning\_dead\_numbers], and re-filter the query results to obtain more satisfactory results.



**Fig. 3-9 Refilter parameters in [refilter\_params]**

## 3.4、Example data explanation of ReHRI

The various data examples and their usage scenarios used by ReHRI are shown in Table 3-2.

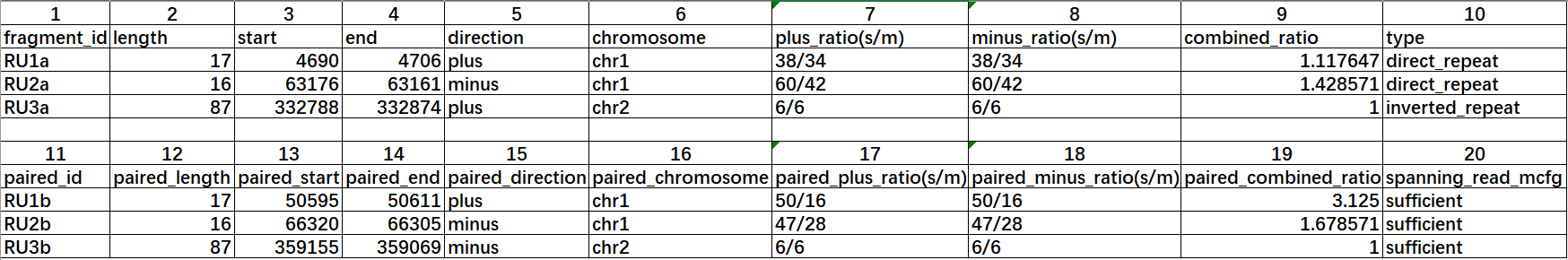
Table 3-2 Example data used for ReHRI

|  |  |  |
| --- | --- | --- |
| **File name** | **File feature** | **Applicable scenarios** |
| CRR302670\_f1.10000.fasta | Forward Reads: Accepts both FASTQ and FASTA formats for detecting recombination-supporting reads. | Test whether there are NGS reads that support genome recombination. |
| CRR302670\_f1.10000.fastq |
| CRR302670\_r2.10000.fasta | Reverse Reads: Accepts both FASTQ and FASTA formats for detecting recombination-supporting reads. |
| CRR302670\_r2.10000.fasta |
| Pacbio.CRR302668.1000.fasta | TGS data can accept fastq or fasta formats. | Test whether there are reads in TGS data that support genome recombination. |
| Pacbio.CRR302668.1000.fastq |
| NC\_000932.1.fasta | *Arabidopsis thaliana* chloroplast genome | Testing recombination in the chloroplast genome |
| NC\_037304.1.fasta | *Arabidopsis thaliana* mitochondrial genome | Testing recombination in the mitochondrial genome |
| plastome\_mitogenome.fasta | *Arabidopsis thaliana* two organelle genomes | Testing recombination between multiple chromosomes |
| One\_manual\_rep.tsv | Repeat sequences within a single chromosome | Testing whether all repeat unit pairs can mediate genome recombination |
| One\_manual\_rep\_corr.tsv | Paired repeat sequences within a single chromosome | Testing whether specified repeat unit pairs can mediate genome recombination |
| Two\_manual\_rep.tsv | Repeat sequences within/across multiple chromosomes | Testing whether all repeat unit pairs can mediate recombination within/across multiple chromosomes |
| Two\_manual\_rep\_corr.tsv | Paired repeat sequences within/across multiple chromosomes | Testing whether specified repeat sequence pairs can mediate recombination within/across multiple chromosomes |

# 4、Interpretation of ReHRI core results

The running results of ReHRI are stored in the paired\peats-recomb-supporting\_iratio.tsv file located in the folder {project\_id}/final-repeat-spaning-results\_{project\_id}.

The result is a 20 column tsv file, as shown in Fig. 4-1.



**Fig. 4-1 Repeat sequence-mediated recombination predicted by ReHRI**

The meaning of each column is as follows:  
① **fragment\_id**: The ID of the repeat unit.  
② **length**: The length of the repeat unit.  
③ **start**: The start position of the repeat unit in the genome.  
④ **end**: The end position of the repeat unit in the genome.  
⑤ **direction**: The strand location (positive or negative).  
⑥ **chromosome**: The chromosome number in the genome.  
⑦ **plus\_ratio**: The ratio of reads spanning the repeat sequence in the minor vs. major configuration on the positive strand.  
⑧ **minus\_ratio**: The ratio of reads spanning the repeat sequence in the minor vs. major configuration on the negative strand.  
⑨ **combined\_ratio**: The overall ratio of repeat-mediated genomic recombination across both DNA strands.  
⑩ **type**: The type of repeat sequence (direct or inverted).  
⑪ **spanning\_read\_mcfg**: Whether the number of reads spanning the repeat sequence in the major configuration meets the user-defined threshold.

**Note:** "Paired entries (paired\_\*)" refer to the other repeat unit in the pair mediating genomic recombination.

# 5、Detailed explanation of ReHRI configuration file

Users can explore the performance of ReHRI by setting the. ini configuration file in more detail. Table 5-1 provides a detailed interpretation of the parameters in the. ini configuration file.

**Table 5-1 Detailed explanation of parameters in the configuration file**

|  |  |  |
| --- | --- | --- |
| **Parameter Category** | **Parameter** | **Values and Descriptions** |
| [general] | project\_id (required) | Project ID, composed of letters, numbers, and underscores. |
| mode (default: A) | Software operation mode, values: N/A/R/C (case-insensitive). N: Program does not run; A: Program runs automatically; R: Only runs ROUSFinder to identify repeat sequences; C: Identifies reads spanning repeat sequences from user-provided repeats. In this case, "calibrated\_repeat\_file" under the [calibrate\_ROUSFinder\_results] category must be provided. |
| inputfasta (required) | Organelle genome sequence file, which may contain multiple chromosomes. |
| genome\_type (default: C) | Sets the genome as linear (L) or circular (C). |
| complementary\_chain (default: Y) | When identifying reads spanning repeat sequences, considers both DNA strands (Y) or not (N). |
| redundant\_intermediate\_results (default: D) | Deletes intermediate results (D) or keeps them (K) during software operation. |
| [ROUSFinder] | repeat\_length (default: 50:) | Repeat sequence length range. Length ≥50bp, set as 50:. Length ≤100bp, set as :100. 100bp ≤ length ≤ 200bp, set as 100:200. Minimum value is 5bp. |
| reward (default: 1) | ROUSFinder parameter: reward value for sequence alignment when identifying repeat sequences. |
| penalty (default: 20) | ROUSFinder parameter: penalty value for sequence alignment when identifying repeat sequences. |
| [manually\_calibrated\_repeat\_info] | calibrated\_repeat\_file | Input file location for manually calibrated repeat sequence results. Required when mode=C. |
| [mainconfiguration] | flanked\_sequence\_length (default: 1000bp) | In the main configuration, the length of sequences extracted from both sides of the repeat unit, in bp. |
| [subconfiguration] | flanked\_sequence\_length (default: 1000bp) | In the sub-configuration, the length of sequences extracted from both sides of the repeat unit, in bp. |
| [sequencing\_depth] | alignment software (default: minimap2) | Alignment software options: minimap2, bwa, or blast. |
| evalue (default: 1e-5) | Blast parameter: evaluates the significance of matching results. |
| NGS\_single\_end | Second-generation single-end sequencing data (fastq or fasta format). |
| NGS\_pair\_ends | Second-generation paired-end sequencing data (fastq or fasta format). |
| TGS | Third-generation sequencing data (fastq or fasta format). |
| TGS\_type | Sets the third-generation sequencing platform type (pacbio or ont), supplementary parameter for TGS. |
| filter\_reads (default: Y) | Whether to filter sequencing reads. Filtering can speed up operation. |
| Threads (Default value: 90% physical threads) | Number of threads. Uses default value if left empty. |
| [check\_spanning\_reads] | spanning\_read\_flanking\_repeat\_length (default: 1bp) | Length of the read spanning the repeat unit, must be a natural number. |
| spanning\_read\_number (default: 1 bp) | Number of reads that meet the criteria for spanning the repeat unit. |
| [refilter\_params] | refilter\_mode (default: N) | Whether to re-filter reads spanning the repeat unit. |
| refilter\_id | Project ID for re-filtering reads spanning the repeat unit. |
| spanning\_read\_flanking\_repeat\_length (default: 5 bp) | When re-filtering reads spanning repeat sequences, the length of the read spanning the repeat sequence, must be a natural number. |
| spanning\_read\_number (default: 5) | When re-filtering reads spanning repeat sequences, the number of reads that meet the criteria for spanning the repeat sequence, must be a natural number. |

# 6、Detailed explanation of ReHRI operation results

The folder named {project\_id} stores all the results of ReHRI after running.

The folder final\_depeat-spanning\_desults\_ {project\_id} stores all the information about duplicate sequences found in the query:

① one\_chain\_without\_sufficient\_spanning\_reads.tsv

② one\_repeat\_unit\_without\_spanning\_reads.tsv

③ paired\_repeats\_for\_mapping.tsv

**④ paired\_repeats\_recomb-supporting\_ratio.tsv**

⑤ repeat\_sequences\_{project\_id}\_chr1.fasta

⑥ repeat\_sequences\_{project\_id}\_chr2.fasta

File ④ is the core result (see interpretation in Part 5), and file ③ is excerpted from file ④ and can be used for drawing recombinant genome maps using ReHRV software. File 1 stores a duplicate sequence in one of the two strands of DNA molecule that has no reads that can be crossed. File 2 stores repeated sequences in both positive and negative strands of DNA molecules that have no reads that can be crossed. Files ⑤⑥ are duplicate sequences from two chromosomes in fasta format.

The folder subclonfig.respeand\_referends\_ {project\_id} stores the intermediate results of the read mapping to the TRS of the secondary configuration. The mainconfig.ref peat-spaned\_desults\_ {project\_id} stores the intermediate results of the read mapping to the TRS of the main configuration. Each folder stores a mapping result of TRS, and its folder name is as follows:

DR\_LD\_RUxxa\_RUxxb\_plus\_1000\_results

DR\_LD\_RUxxa\_RUxxb\_minus\_1000\_results

DR\_UR\_RUxxa\_RUxxe\_plus\_1000\_results

DR\_UR\_RUxxa\_RUxxe\_minus\_1000\_results

IR\_LU\_RUxxb\_RUxxc\_plus\_1000\_results

IR\_LU\_RUxxb\_RUxxc\_minus\_1000\_results

IR\_RD\_RUxxa\_RUxxb\_plus\_1000\_results

IR\_RD\_RUxxa\_RUxx1b\_minus\_1000\_results

DR\_LR\_RUxxa\_RUxxb\_plus\_1000\_results

DR\_LR\_RUxxa\_RUxxb\_minus\_1000\_results

DR\_UD\_RUxxa\_RUxxe\_plus\_1000\_results

DR\_UD\_RUxxa\_RUxxe\_minus\_1000\_results

IR\_LR\_RUxxa\_RUxxb\_plus\_1000\_results

IR\_LR\_RUxxa\_RUxxb\_minus\_1000\_results

IR\_UD\_RUxxa\_RUxxe\_plus\_1000\_results

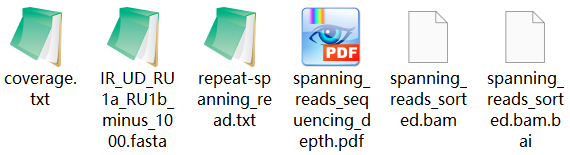
IR\_UD\_RUxxa\_RUxxe\_minus\_1000\_results

The naming rules for each part of the folder name are shown in Table 6-1:

**Table 6-1 Meaning of characters in result folder names**

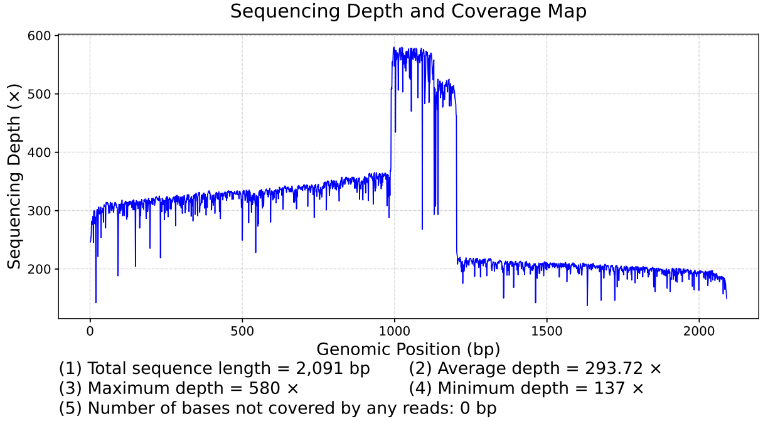
|  |  |
| --- | --- |
| **Symbol** | **Meaning of the Symbol** |
| DR | Direct Repeat sequence |
| IR | Inverted Repeat sequence |
| LU | In the sub-configuration, the TRS extracted with the inverted repeat unit RUb as the center, see Fig. 2-3 |
| RD | In the sub-configuration, the TRS extracted with the inverted repeat unit RUa as the center, see Fig. 2-3 |
| LD | In the sub-configuration, the TRS extracted with the direct repeat unit RUa as the center, see Fig. 2-3 |
| UR | In the sub-configuration, the TRS extracted with the direct repeat unit RUb as the center, see Fig. 2-3 |
| LR | In the main configuration, the TRS extracted with the direct repeat unit RUa as the center, see Fig. 2-3 |
| UD | In the main configuration, the TRS extracted with the direct repeat unit RUb as the center, see Fig. 2-3 |
| RU | Repeat Unit |
| xx | Number, a natural number |
| a/b/c... | Different repeat units of the same repeat sequence |
| plus | Represents the plus strand of DNA |
| minus | Represents the minus strand of DNA |
| 1000 | Length of the sequence extracted from both sides of the repeat sequence, i.e., L in Fig. 2-3 |
| results | Suffix for folder names |

Each folder contains TRS sequences (in fasta format), and reads across repeated sequences in TRS are mapped to the sequencing depth of TRS, and reads are mapped to the BAM document of TRS (Fig. 6-1).



**Fig. 6-1 Various results after mapping read to each TRS**

The read mapping across repeated sequences in TRS to the sequencing depth of TRS is shown in Fig. 6-2, and the sequencing depth values are saved in coverage.txt. The BAM document can be visualized using software such as Tablet to map the actual situation of the read to TRS (Fig. 6-3).



**Fig. 6-2 Sequencing depth of reads spanning repeats in TRS mapped to TRS**

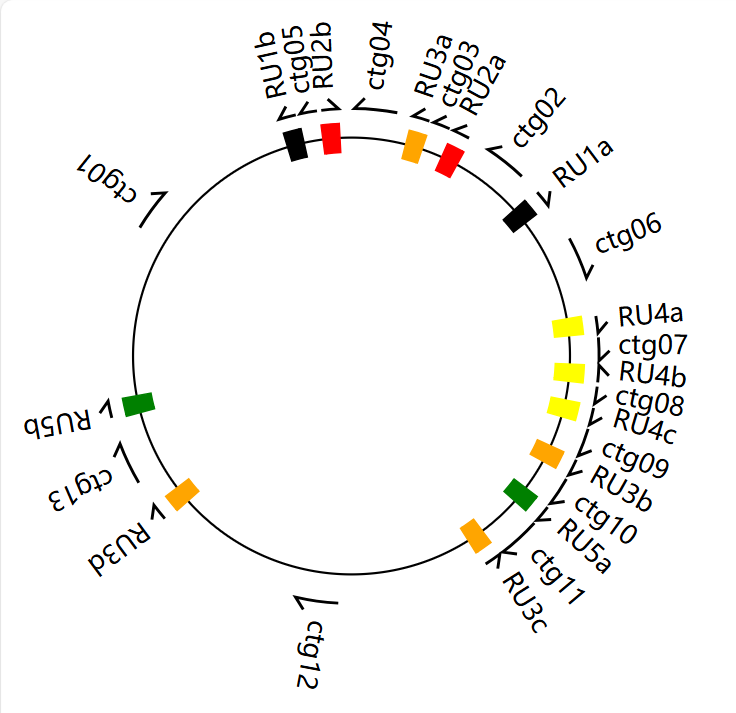


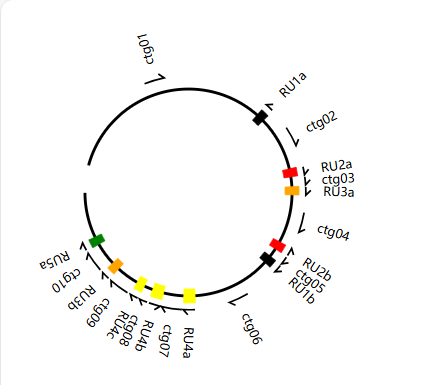
**Fig. 6-3 Visualization results of reads spanning repeats in TRS mapped to TRS**

When ReHRI filters the results of the initial screening again, that is, after ReHRI executes mode three, the results are stored in the folder named {project\_id} in the refilted-peat-spanning-results\_{FLT} folder. The interpretation of the results is based on the interpretation of the results in the folder final\_depeat-spanning-results\_{project\_id}.

# 7、ReHRV generates genomic maps of repeat sequences.

The software ReHRV (Repeat mediated Homologous Recombination Visualization) can draw a schematic diagram of circular genome recombination based on the results of ReHRI software, to display the genomic maps of various subtypes of mitochondrial genomes mediated by repetitive sequences. The graph is represented in a circular pattern, with arrows indicating the direction of the DNA molecule's positive strand before and after recombination. Repeat units (RU) are represented by colored blocks, with the same colored blocks representing different units of the same repeat sequence, and different colors representing different repeat sequences. The “ctg” represents the region between two adjacent repeat units (Fig. 7-1A). The map of linear chromosomes is a circular map with notches (Fig. 7-1B). The size of the graph radius represents the length of the genome sequence.





**B**

**A**

## Fig. 7-1 Schematic diagram of the genome map drawn by ReHRV

## 7.1、Operation of ReHRV

The command line to run ReHRV is as follows:

python bin/ReHRV.py -c ReHRV.config.ini

python bin/ReHRV.py -c ReHRV.config.ini -redo

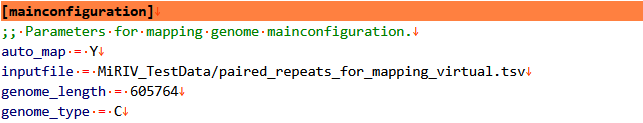
The parameters of ReHRV are provided in the form of a configuration file .ini. The vast majority of parameters provide default values, with only a few parameters requiring user input. The parameter '-redo' allows users to redraw the repeated sequence gene map after an unexpected program interruption. This process will delete intermediate results generated from the previous run.

## 7.2、Configuration file of ReHRV

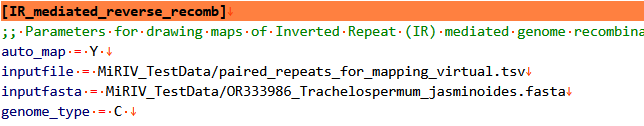
ReHRV can draw its genome maps for the main configuration ([mainconfiguration] mode) (Fig. 7-2A), the secondary configuration mediated by IRs ([IR-mediated\_deverse\_recover] mode) (Fig. 7-2B), and the secondary configuration mediated by DRs ([DR-mediated\_decomb\_1to2] mode) (Fig. 7-2C). When drawing a map, users need to provide information on the location of repeated sequences, the sequence of the genome (in fasta format), the length of the genome, and the type of genome (i.e. linear or circular structure).

Each mode is set with an auto\_map parameter to control whether to run the corresponding mode (Y/N/M). When auto\_map=N, ReHRV does not run the corresponding mode. Auto\_map=Y, ReHRV will automatically draw all genome maps. Auto\_map=M, ReHRV will draw the user specified genome map under the user's guidance.

When encountering the situation where two or more chromosomes form one chromosome under DR mediation, ReHRV only allows users to provide two chromosomes at a time, so users need to run ReHRV multiple times ([DR-mediated\_decomb\_2to1] mode) to recombine multiple chromosomes into one chromosome. When two chromosomes recombine into one chromosome, one of them must have a circular structure (C), and the parameter settings are shown in Fig. 7-2D (chr1\_type, chr2-type). When both chromosomes are linear (L), ReHRV can only achieve cross recombination of sequences between the two chromosomes ([DR-mediated\_decomb\_2to2] mode), resulting in two linear chromosomes. The parameter settings are shown in Fig. 7-2E.

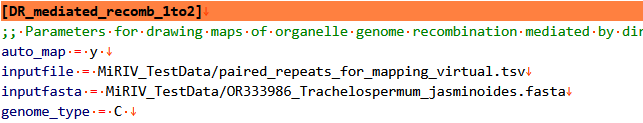


**A**

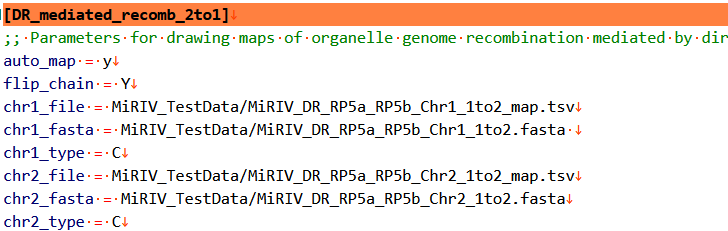


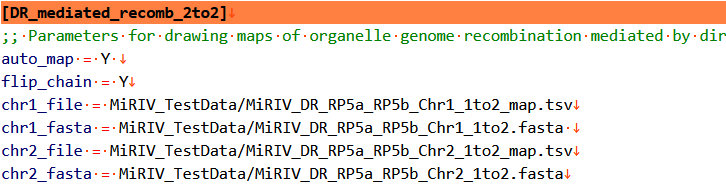
**B**

**C**



**D**



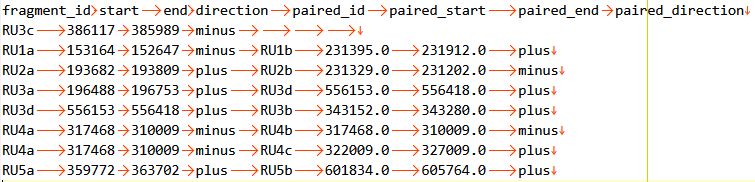


**E**

**Fig. 7-2 Parameters required for ReHRV to draw various genome maps**

In the two modes of [DR-mediated-recomb\_2to2] and [DR-mediated-recomb\_2to1], due to the free rotation of the two chromosomes, all repeating units that can mediate genome recombination can mediate the formation of one chromosome by positively repeating the two chromosomes. So, when flipuchain=Y, it allows all repetitive units that can mediate genome recombination to form one chromosome by mediating two chromosomes in a positive repeat form.

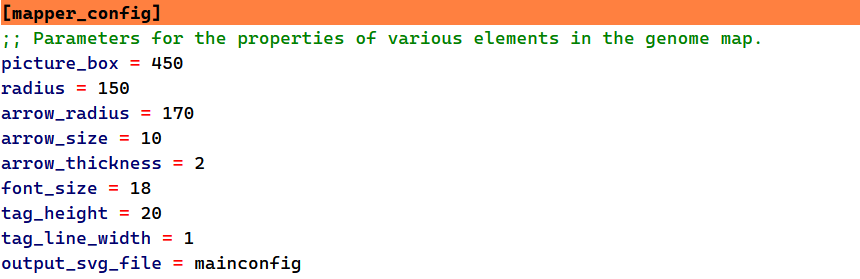
In all drawing modes, the formats of inputfile, chr1\_file, and chr\_2file are the 8 lists in tsv format, as shown in Fig. 7-3. Each line represents a pair of repeated sequences, and paired repeated sequences can be omitted. The names of the header and repeated sequences must be consistent with those shown in Fig. 7-3.



**Fig. 7-3 Schematic diagram of 8 column table for ReHRV (tsv format)**

## 7.3、Parameter for each element in the genome map

The [mapper\_config] option is used to set the attribute parameters of each element in the drawn genome map. The parameter values are shown in Fig. 7-4:



**Fig. 7-4 Parameters of each element in the genome map in [mapper\_config]**

The default values and meanings of each parameter are shown in Table 7-1:

**Table 7-1 Default values and descriptions of [mapper\_config] options**

|  |  |
| --- | --- |
| **Parameters** | **Possible Values** |
| picture\_box | Size of the output image (length of one side of the square), **default=280** |
| radius | Radius of the genomic map, determines the image size, **default=150** |
| arrow\_radius | Radius of the circle where the arrow is located, **default=170** |
| arrow\_size | Size of the arrow, **default=10** |
| arrow\_thickness | Thickness of the arrow line, **default=2** |
| font\_size | Font size, **default=18** |
| tag\_height | Height of the tag (annular sector), **default=20** |
| tag\_line\_width | Outline width of the tag (annular sector), **default=1** |

## 7.4、Color parameters of repeat sequences

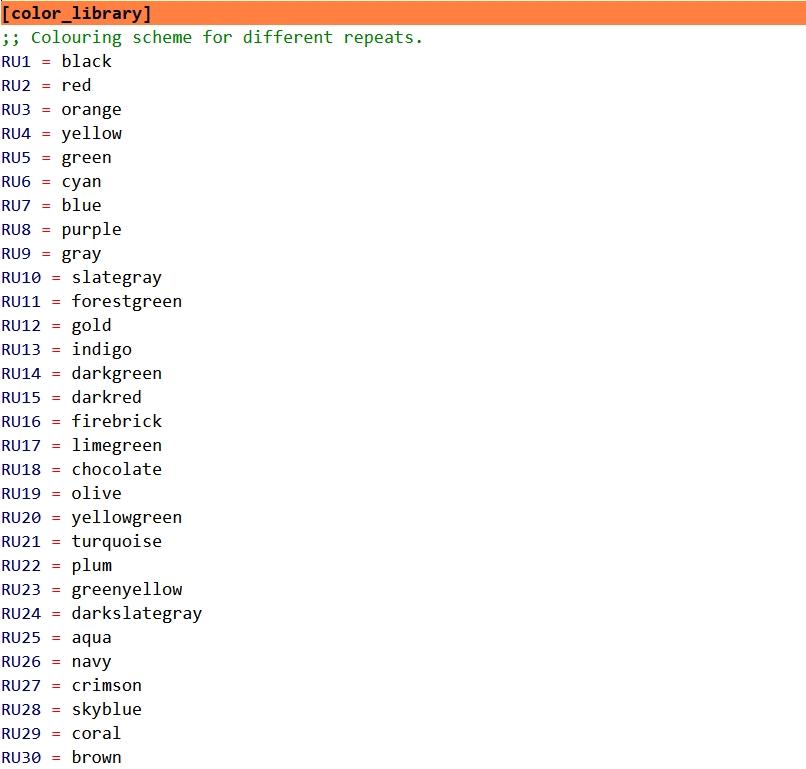
The [color\_ibrary] option is used to set the color of repetitive sequences in the genome map. Colors can be represented using 60 English words with built-in colors, as well as RGB and hexadecimal numerical values.

The built-in 60 color filters are from Python's webcolors library, namely: black，red，orange，yellow，green，cyan，blue，purple，brown，gray，darkslategray，dimgray，navy，indigo，darkgreen，darkred，firebrick，crimson，chocolate，olive，yellowgreen，lawngreen，limegreen，greenyellow，lightseagreen，seagreen，darkseagreen，lightgreen，forestgreen，darkcyan，mediumturquoise，turquoise，aquamarine，mediumaquamarine，aqua，deepskyblue，skyblue，steelblue，cadetblue，royalblue，mediumblue，darkviolet，plum，deeppink，hotpink，pink，palevioletred，mediumvioletred，coral，orangered，darkorange，goldenrod，gold，khaki，darkkhaki，wheat，lightgrey，lightslategray，slategray，darkgray.

The color values for RGB are set as follows: 0.0.0 (black), 255.0.0 (red), 255.165.0 (orange), 255.255.0 (yellow), 0.255.0 (green), 0.255.255 (cyan), 0.0.255 (blue), 128.0.128 (purple).

The hexadecimal values representing colors are: # 000000 (black), # FF0000 (red), # FFA500 (orange), # FFFF00 (yellow), # 00FF00 (green), # 00FFFF (blue), # 800080 (purple).

ReHRV allows up to 30 repeated units to be colored simultaneously, as shown in Fig. 7-5:



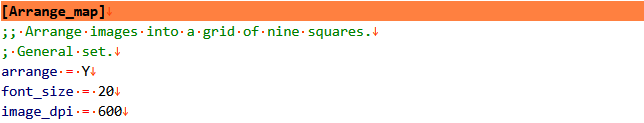
**Fig. 7-5 Coloring scheme for repeat sequences within the genome map**

## 7.5、Parameters for arranging multiple genomic maps

[Arrange\_map] is used to set parameters for multiple graph layouts, with graphs arranged in a nine grid layout (Fig. 7-6A).

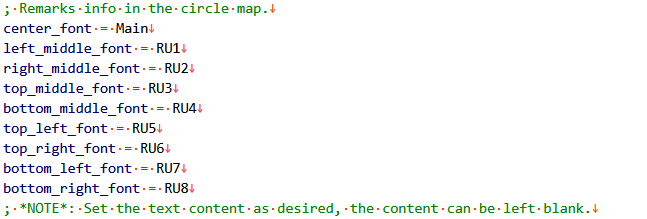
Arrange: Should multiple genome maps be formatted. Yes/Y indicates typesetting, No/N indicates no typesetting. Font\_size: Set the label font size. Image\_depi: The resolution of the typeset image. Default values can be used.

Each genome map can have a simple label set in its middle position (Fig. 7-6B), which can be composed of uppercase and lowercase letters, numbers, and underscores. Each genome map can be independently positioned in the formatted image (Fig. 7-6C).

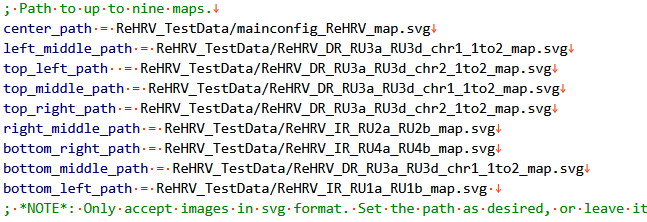


**A**

**B**

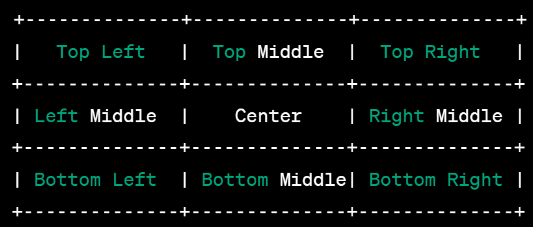


**C**



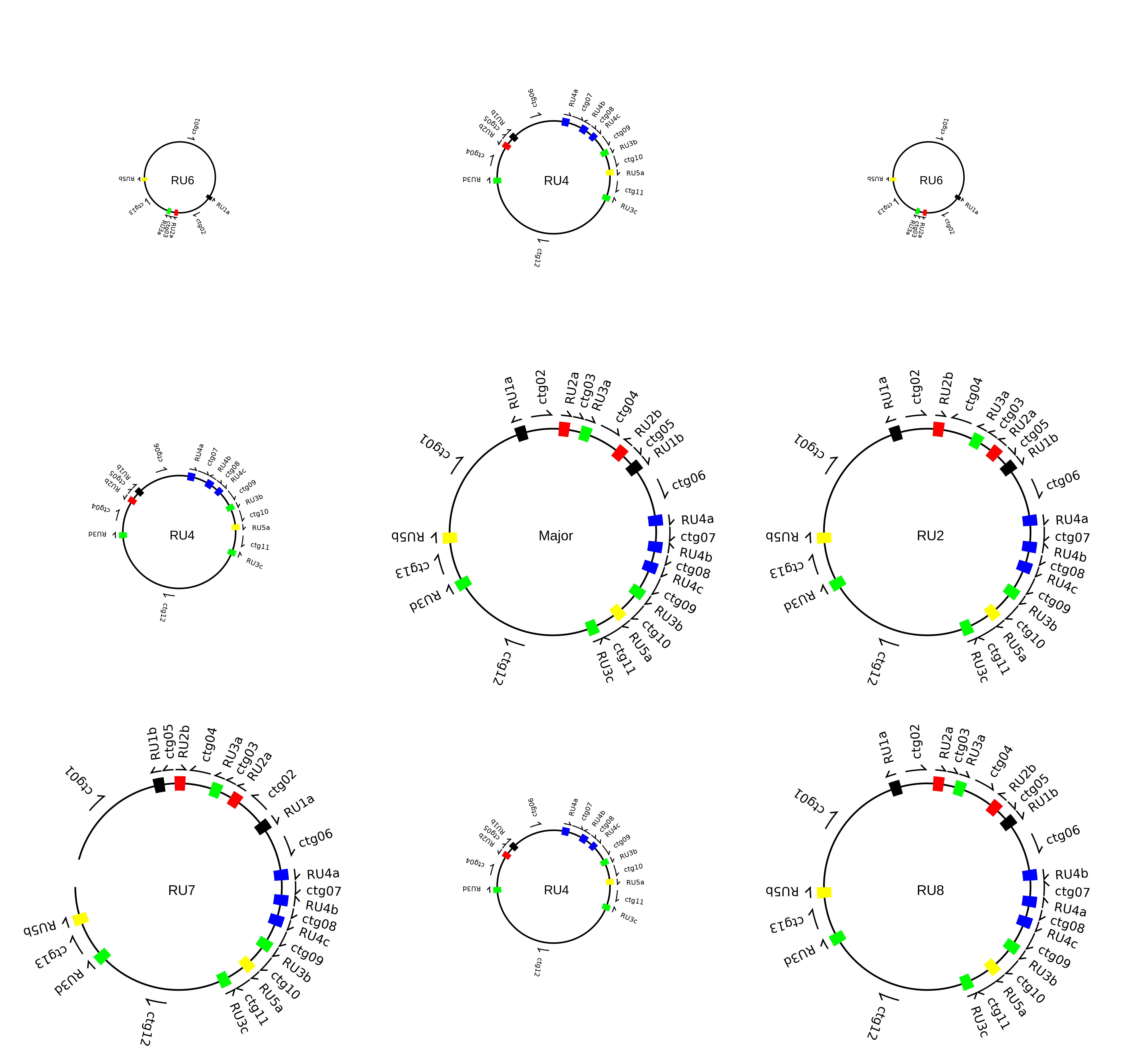
**Fig. 7-6 Parameters for arranging multiple genomic maps**

The positions of genomic maps within the 3×3 grid are shown in Fig. 7-7 below:



**Fig. 7-7 Schematic diagram of the genome map's position within the nine-square grid.**

The effect of the layout of the graph in the nine grid is shown in Fig. 7-8:



**Fig. 7-8 The effect diagram after layout in the nine-square grid.**

## 7.6、Interpretation of drawing results in each mode

The results after running the modes [mainconfiguration], [IR\_mediated\_reverse\_recomb], [DR\_mediated\_recomb\_1to2], [DR\_mediated\_recomb\_2to1], [DR\_mediated\_recomb\_2to2], and [Arrange\_map] are stored in a folder named {project\_id}:

[mainconfiguration]：mainconfig\_{project\_id}

[IR\_mediated\_reverse\_recomb]：Inv\_Rev\_{project\_id}

[DR\_mediated\_recomb\_1to2]：DR\_1to2\_{project\_id}

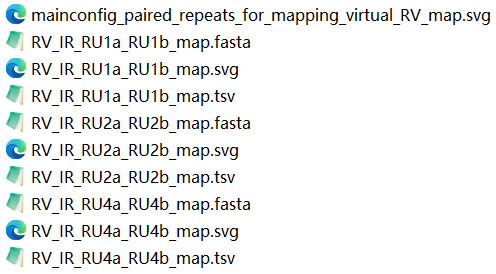
[DR\_mediated\_recomb\_2to1]：DR\_2to1\_{project\_id}

[DR\_mediated\_recomb\_2to2]：DR\_2to2\_{project\_id}

[Arrange\_map]：map\_nine\_squares\_{project\_id}

For example, there are two types of results based on the results of [DR-mediated\_decomb\_1to2]: the first is a genome map of repeated sequences in the main configuration, which is generated based on the user's input of 8 lists in SVG format and can be opened using web browsers and software such as Adobe Illustrator CS6; The second type is the relevant information of the two chromosomes (chr1, chr2) generated, with each chromosome corresponding to three files, namely the fasta format sequence, the svg format repetitive sequence genome map, and the eight lists corresponding to the genome map (see Fig. 7-9).

The output results of other modes are similar to those of [DR-mediated\_decomb\_1to2]. Among them, when three or more sequences need to be converted into one chromosome in the [DR-mediated\_decomb\_2to1] mode, the 8 list (tsv format) and fasta format files from the previous run of this mode can be used as input files for the next run of this mode. Thus achieving the goal of recombining multiple chromosomes into one chromosome.



**Figure 7-9 Schematic diagram of the output results for each mode.**

The output results include the genome main configuration map (SVG format), the genome sequence corresponding to the secondary configuration genome (Fasta format), the genome map (SVG format), and duplicate sequence pairing information (TSV format).