Tutorial for Organelle Genome Recombination Detection and Recombinant Map Construction

----Using ReHRI and ReHRV Software

# I. Software Installation

## 1. Download the Software Package

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

## 2. Unzip and Enter the Software Directory

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

cd ReHRI\_ReHRV\_1.0-main

## 3. Install and Activate the Conda Environment

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

conda activate ReHRI\_ReHRV\_1.0

## 4. Update Components

sh install.sh

rm install.sh

## 5. Verify Installation

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.

python bin/ReHRV.py -h

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

**ReHRV:** A tool to map the confgiure 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.

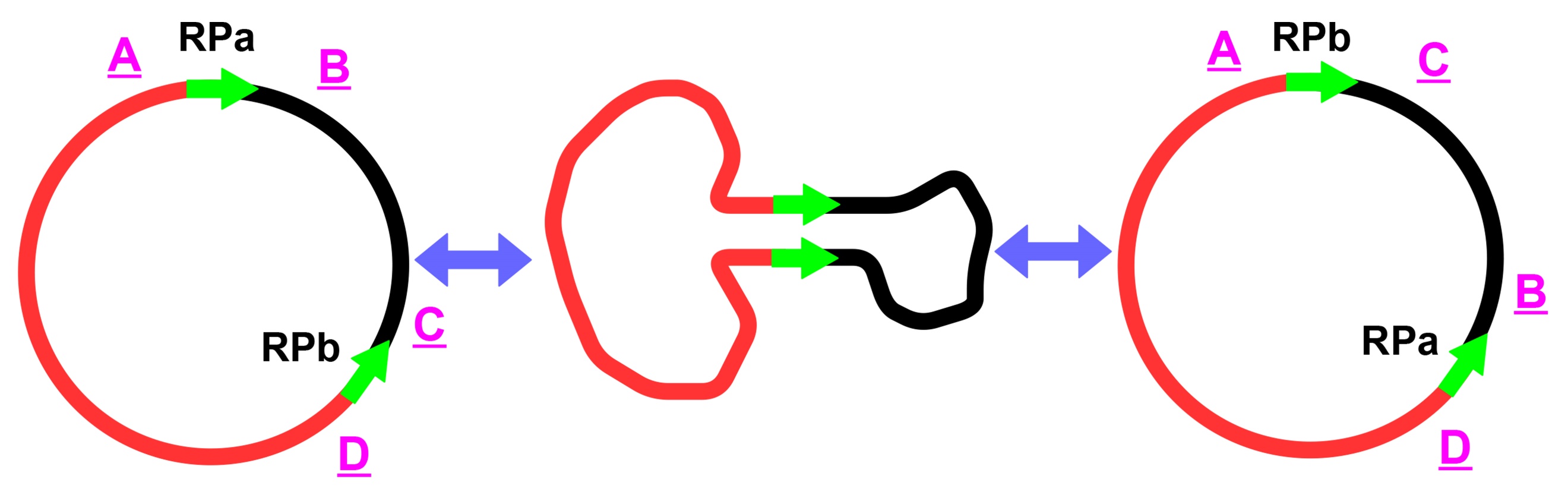
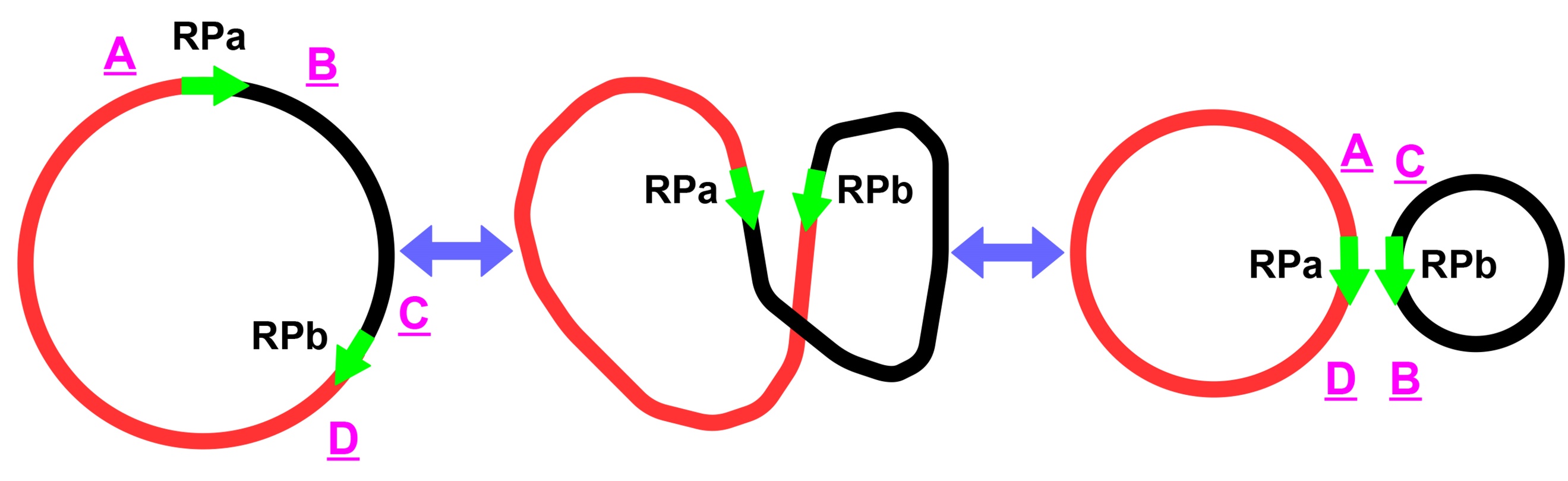
-v, --version Show the version number and exit.

# II. Operating Principles

## 1. Operating principle of ReHRI

The operational principle of ReHRI (**Mi**togenome **R**ecombination **I**dentification) is based on the following molecular mechanisms: (1) Recombination mediated by inverted repeats (IRs) induces inversion of intervening sequences (Figure 2-1A), while (2) recombination involving direct repeats (DRs) produces paired subgenomic molecules (Figure 2-1B)

**A**



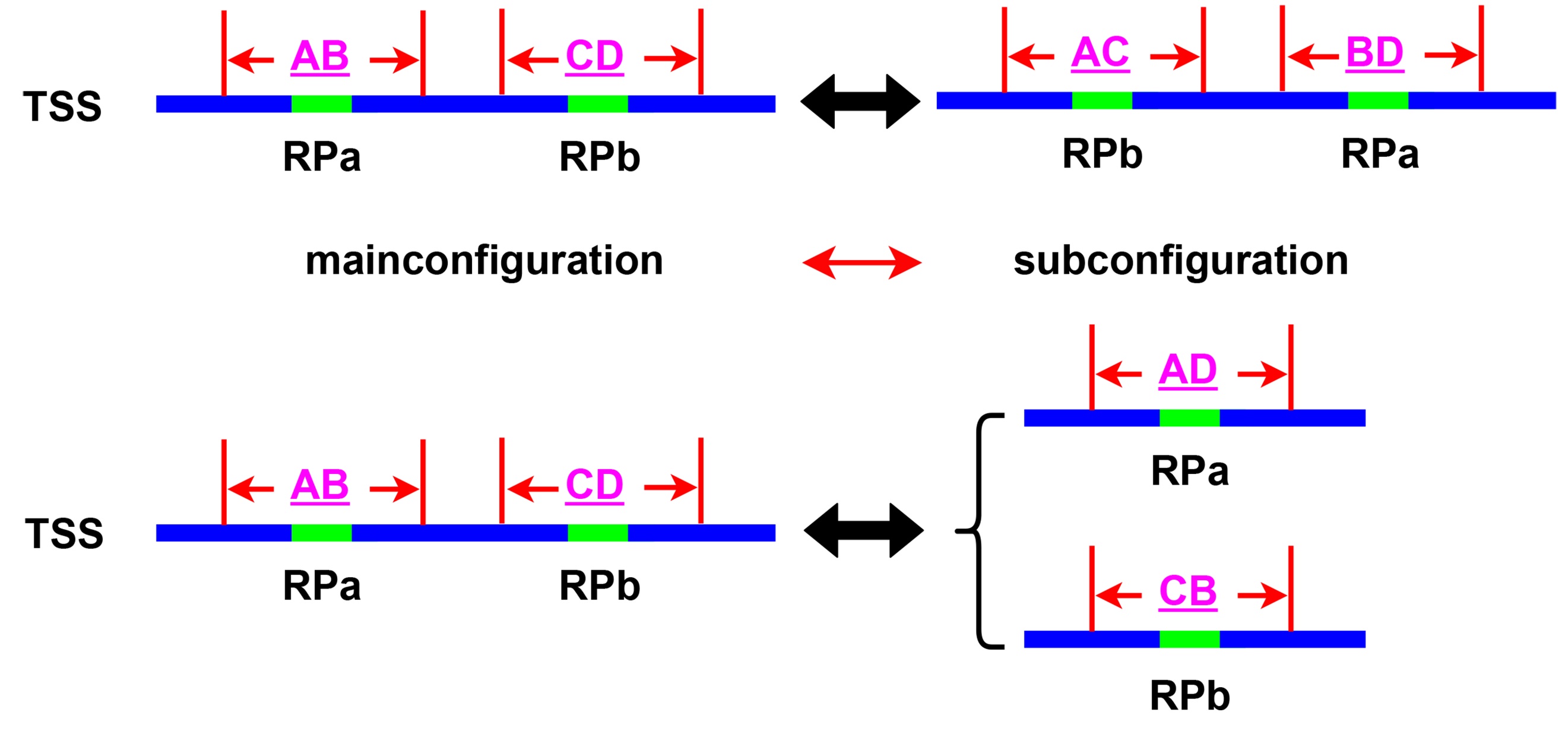
**B**

**A**

**Figure 2-1 Schematic diagram of repeat-mediated circle genome recombination.**

A: Recombination mediated by IRs, B: Recombination mediated by DRs.

Centered on a repeat unit, a trimmed short sequence (TSS) is intercepted from the mainconfiguration and the subconfiguration respectively. As shown in Figure 2-2, in the mainconfiguration, taking the paired repeat units RPa and RPb as the centers respectively, the intercepted TSSs are AB and CD, which are marked as TSSAB and TSSCD. In the subconfiguration mediated by IRs, TSSAC and TSSBD are intercepted, centered on the repeat units RPb and RPa respectively (Figure 1-1A). In the subconfiguration mediated by DRs, TSSAD and TSSCB are intercepted, centered on the repeat units RPa and RPb respectively (Figure 2-2B).



**A**

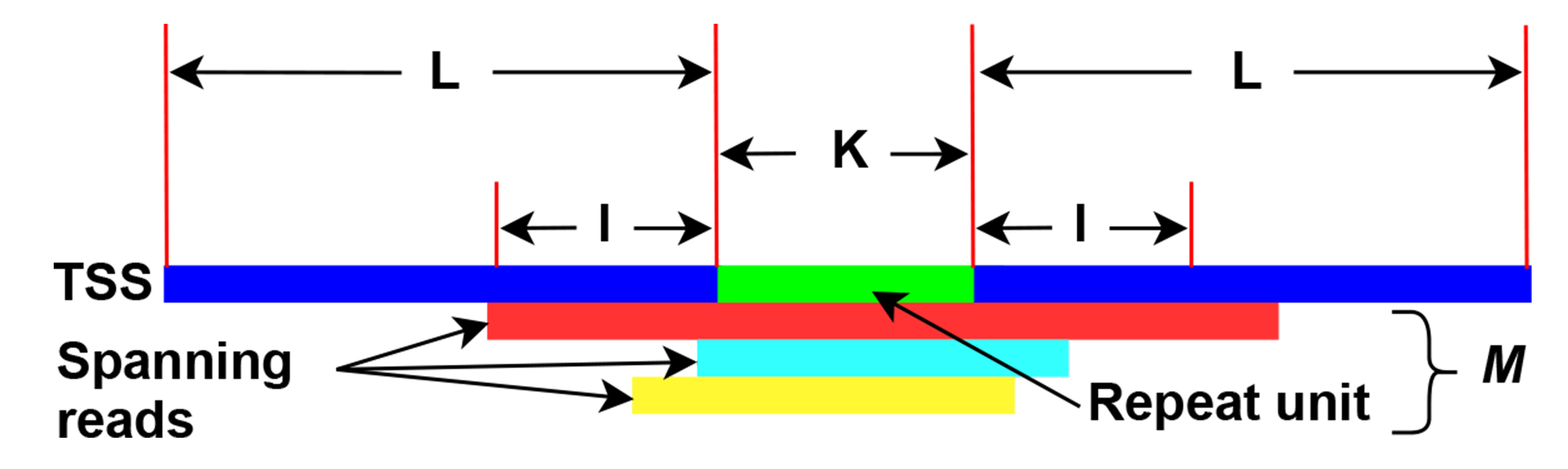
**A**

**B**

**Figure 2-2 Schematic diagram of TSS interception.**

Recombination involving IRs inverts the intervening sequences (A), while recombination involving DRs generates a pair of subgenomic molecules (B). AB, CD, AD, CB, AC, and BD represent TSS. AB and CD are from the mainconfiguration. AD and CB are from the subconfiguration mediated by DRs. AC and BD are from the subconfiguration mediated by IRs.

In both the mainconfiguration and the subconfiguration, with the repeat unit as the center, *L* bases are intercepted on each of the left and right sides. After obtaining the TSS, ReHRI maps the sequencing reads to the TSS (Figure 2-3). Then, it searches for reads that can span the repetitive sequence (*l* > 0) among the reads mapped to the TSS. If there are reads that span the repetitive sequence (i.e., *M* > 0), it is considered that the genome configuration corresponding to this TSS exists.



**Figure 2-3 Schematic diagram of reads mapped to TSS (**trimmed a short sequence**).**

*L*: represents the length of the sequences intercepted on both sides of a repeat unit. Default value is 1000 bp.

*K*: The length of a repaet unit. Default value is 50 bp.

*I*: represents the length by which a read spans the left and right sides of a repeat unit. A read is considered to have spanned a repeat unit when *l* ≥ 1 bp.

*M*: number of the read which spans a repeat unit with *l* ≥ 1 bp.

## 2. Definition and Calculation of Recombination Rate

When calculating the probability of repeat-mediated mitochondrial genome recombination, in the mainconfiguration, the number of reads spanning the repetitive sequences RPa and RPb in TSSAB and TSSCD are recorded as *M*AB and *M*CD, respectively. In the subconfiguration, the number of reads spanning the repeat units RPa and RPb in TSSAD and TSSCB, as well as in TSSAC and TSSBD, are recorded as *M*AD and *M*CB, and *M*AC and *M*BD, respectively.

In DR-mediated genome recombination, the probability of recombination mediated by RPa is:



The above are the calculation formulas for the recombination rates in the cases of considering single-strand (2-1) and double-strand (2-2).

In DR-mediated genome recombination, the probability of recombination mediated by RPb is:



The above are the calculation formulas for the recombination rates in the cases of considering single-strand (2-3) and double-strand (2-4).

In IR-mediated genome recombination, the probability of recombination mediated by RPa is:



The above are the calculation formulas for the recombination rates in the cases of considering single-strand (2-5) and double-strand (2-6).

In IR-mediated genome recombination, the probability of recombination mediated by RPb is:



The above are the calculation formulas for the recombination rates in the cases of considering single-strand (2-7) and double-strand (2-8).

# III. Operation of the ReHRI Software

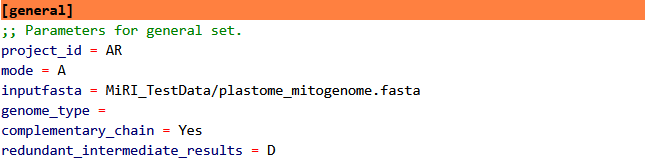
python bin/ReHRI.py -c ReHRI.config.ini

ReHRI requires numerous parameters for its operation. The configuration file ReHRI.config.ini is utilized to set various parameters. However, most of these parameters can adopt default values, and only a few need to be configured.

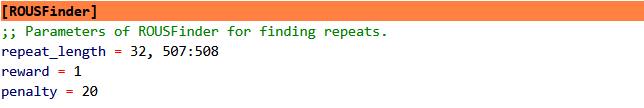
## 1. Operation Mode One of ReHRI

When mode = A (Figure 3-1), it is the first operation mode of ReHRI. ReHRI will first search for direct and inverted repeats within the genome, and then detect repeat pairs that can mediate genome recombination. At this time, users must provide the genome sequence file, the length of the repetitive sequences to be searched, the sequencing file, and the parameters for determining the existence of the subconfiguration.

The genome sequence file (inputfasta) is in .fasta format. Meanwhile, it is necessary to specify whether the genome type (genome\_type) is linear (L) or circular (C) (Figure 3-1). When the genome contains multiple chromosomes, all chromosomes should be placed in the same fasta file.

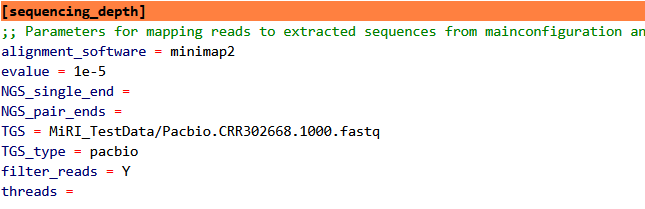


**Figure 3-1 Setting of Parameters in [general] Section**

The length of the repeats to be searched (repeat\_length), which is the value of *K* in Figure 2-2, can be set as a range. For example, if 50 bp ≤ length ≤ 1000 bp, it can be set as 50:1000; if length ≥ 50 bp, it can be set as 50: ; several lengths can also be set. For instance, if the lengths are 50bp and 100bp, it can be set as 50, 100 (Figure 3-2).

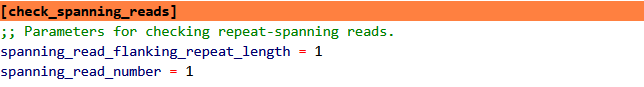
**Figure 3-2 Setting of Repeat Length in [ROUSFinder] section**

For sequencing data, ReHRI can accept next-generation sequencing (NGS) and third-generation sequencing (TGS) data. Only one of NGS and TGS data can be accepted at a time (Figure 3-3). When providing paired-end data, the files of paired-end data should be separated by spaces. When providing TGS data, it is also necessary to specify whether the data comes from the Nanopore sequencing platform (ont) or the Pacbio sequencing platform (pacbio) (Figure 3-3).



**Figure 3-3 Setting of the repeat length in [ROUSFinder] section**

The parameters for detecting the existence of minor configurations are mainly spanning\_read\_flanking\_repeat\_length and spanning\_read\_number (Figure 3-4). These are important parameters for determining whether a minor configuration exists. Their default values are both 1. In the subsequent re-filtering mode (refilter\_mode = Y), they can be reset to conduct multiple screenings on the results of genome recombination mediated by the queried repetitive sequences.



**Figure 3-4 Parameters for determining the existence of minor configurations in [check\_spanning\_reads] section**

The applicable scenarios of minimap2, bwa, and blast in alignment software are shown in Table 3-1.

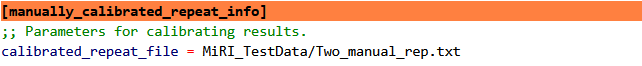
**Table 3-1 Comparison of Applicable Scenarios of minimap2, bwa, and blast**

|  |  |  |
| --- | --- | --- |
| **Tool** | **Best Application Scenarios** | **Possible Reasons for Missed Detection** |
| minimap2 | Long-read sequencing, genome assembly, rapid alignment | Short-read sequencing, highly repetitive regions, relaxed default parameters |
| BWA | Short-read sequencing, variant detection, precise alignment | Low efficiency in processing long-read sequencing |
| BLAST | Homology search, cross-species comparison | Time-consuming calculation, not suitable for large-scale alignment |

The "spanning\_read\_flanking\_repeat\_length" is the length of the read after it spans the repeat, that is, the value of "*l*" in Figure 2-3. The "spanning\_read\_number" is the number of reads whose length after spanning the repeat is greater than or equal to "*l*", that is, the value of "*M*" in Figure 2-3.

## 2. Operation Mode Two of ReHRI

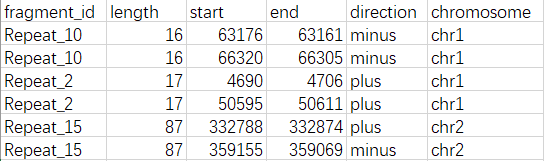
The work of querying repeats is extremely challenging. Results of different repeats have a great impact on the results of mediating genome recombination, and the results of the repeats found by different algorithms also vary. Therefore, ReHRI has set up an interface that can accept the repeat sequence information provided by users (Figure 3-5), allowing users to provide the repeats themselves. At this time, it is necessary to set mode = C (Figure 3-6). The repeat sequence information file provided by users is in tsv format (Figure 3-7). For the same repeat, the same fragment\_id should be used. When there is only one chromosome in the genome, the chromosome column needs to be removed. If users specifically want to detect the mediating role of certain paired repeat units in genome recombination, they can provide ReHRI with a repeat information file in the format shown in Figure 3-8. When the genome has only one chromosome, the chromosome and paired\_chromosome columns should be removed.



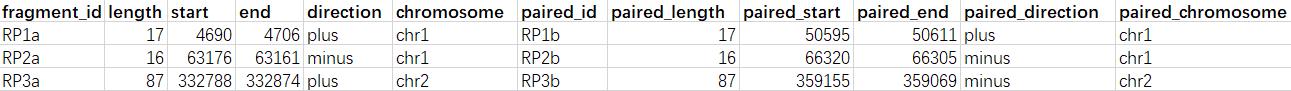
**Figure 3-5 Repeats provided by users in [manually\_calibrated\_repeat\_info] section**



**Figure 3-6 Parameter settings for ReHRI to enter Operation Mode Two**

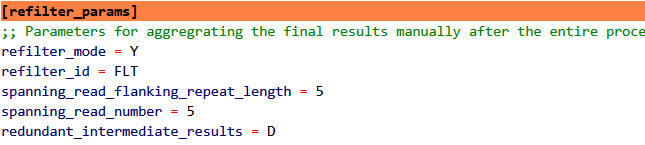


**Figure 3-7 Example of repeat information provided by users (TSV format)**



**Figure 3-8 Example of paired repeats provided by users (TSV format)**

## 3. Operation mode three of ReHRI

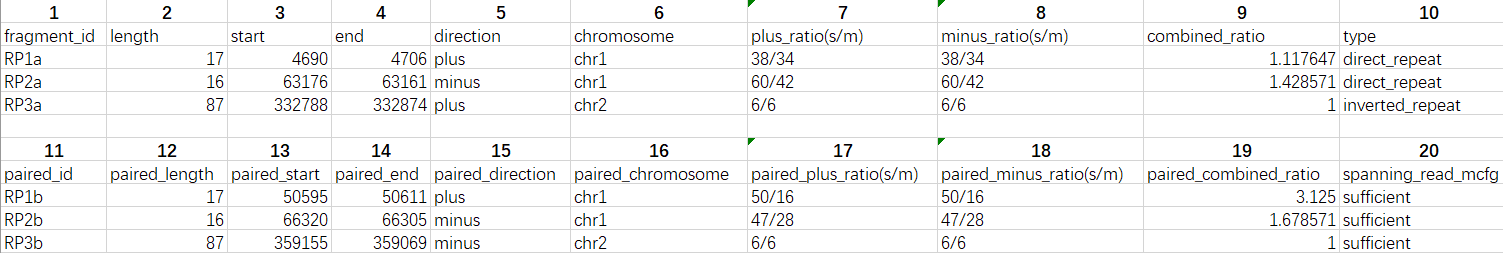
If users are not satisfied with the initially set screening conditions, they can set refilter\_mode = Y (Figure 3-9), reset the spanning\_read\_flanking\_repeat\_length and spanning\_read\_number, and re-clean the query results to obtain more satisfactory results.

**Figure 3-9 Refiltering conditions in [refilter\_params] section**

# IV. Interpretation of the Core Results of ReHRI

The operation results of ReHRI are stored in the file "paired\_repeats\_recomb-supporting\_ratio.tsv" within the folder "{project\_id}/final\_repeat-spanning\_results\_{project\_id}".

This result is a TSV file with 20 columns, as shown in Figure 4-1.



**Figure 4-1 Information of the Repeated Sequences Mediating Genomic Recombination**

The meanings of each column are as follows:

① fragment\_id: The ID of repeat unit.

② length: The length of repeat unit.

③ start: The start positin of repeat unit in the genome.

④ end: The end positin of repeat unit in the genome.

⑤ direction: The postion in plus or minus strand of DNA.

⑥ chromosome: The ID of a chromosome in genome.

⑦ plus\_ratio(s/m): The number of reads that spans repeats in subconfiguration(s) and mainconfiguration(m) in the plus strand of DNA.

⑧ minus\_ratio(s/m): The number of reads that spans repeats in subconfiguration(s) and mainconfiguration(m) in the minus strand of DNA.

⑨ combined\_ratio: The overall ratio of repeat-mediated genome recombination on the two strands of DNA.

⑩ type: The type of repeat (direct or inverse repeat).

⑪ spanning\_read\_mcfg: In the mainconfiguration, whether the number of reads spanning the repeated sequences meets the quantity set by the user.

Note: The “paired” represents the other repeat unit in a pair of repeat units that mediate genome recombination.

# V. Detailed Explanation of the 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 each parameter in the .ini configuration file.

**Table 5-1 Detailed Explanation of Parameters in the Configuration File**

|  |  |  |
| --- | --- | --- |
| **Parameter Category** | **Parameter** | **Value and Meaning** |
| [general] | project\_id (Required) | Project ID, consisting of letters, numbers, and underscores. |
| mode (default value: A) | The operation mode of the software, with values *N/A/R/C* (case-insensitive). N: The program does not run; A: The program runs automatically; R: Only run the ROUSFinder software to find repeated sequences; C: Find reads spanning repeated sequences from the repeated sequences provided by the user. In this case, the "calibrated\_repeat\_file" in the [calibrate\_ROUSFinder\_results] category must be provided. |
| inputfasta (Required) | Organelle genome sequence file, which can contain multiple chromosomes. |
| genome\_type (default value: C) | Set the genome as linear (L) or circular (C). |
| complementary\_chain (default value: Y) | When finding reads spanning repeated sequences, consider the double-strands of DNA (Y) or not (N). |
| redundant\_intermediate\_results (default value: D) | During software operation, delete intermediate results (D) or not (K). |
| [ROUSFinder] | repeat\_length (default value: 50: ) | The length range of repeated sequences. For lengths ≥ 50bp, set it to 50:. For lengths ≤ 100bp, set it to :100. For 100bp ≤ length ≤ 200bp, set it to 100:200. The minimum value is 5bp. |
| reward (default value: 1) | A parameter of ROUSFinder, the reward value for sequence alignment when finding repeated sequences. |
| penalty (default value: 20) | A parameter of ROUSFinder, the penalty value for sequence alignment when finding repeated sequences. |
| [manually\_calibrated\_repeat\_info] | calibrated\_repeat\_file | The location of the input file for manually calibrated repeated sequence results, which is a required parameter when mode = C. |
| [mainconfiguration] | flanked\_sequence\_length (default value: 1000bp) | In the mainconfiguration, the length of the sequences intercepted on both the left and right sides of the repeated sequence unit, with the unit of bp. |
| [subconfiguration] | flanked\_sequence\_length (default value: 1000bp) | In the subconfiguration, the length of the sequences intercepted on both the left and right sides of the repeated sequence unit, with the unit of bp. |
| [sequencing\_depth] | alignment software (default value: minimap2) | Provide the alignment software, which can be minimap2, bwa, or blast. |
| evalue (default value: 1e-5) | A parameter of blast, used to measure the significance of matching results. |
| NGS\_single\_end | The location of the NGS single-end data file, which needs to be provided separately. |
| NGS\_pair\_ends | The location of the NGS paired-end data file, which needs to be provided separately. |
| TGS | The location of the TGS data file, which needs to be provided separately. |
| TGS\_type | Set the type of the TGS sequencing platform, a supplementary parameter for the TGS parameter, with values of pacbio or ont. |
| filter\_reads (default value: Y) | Whether to filter sequencing reads. Filtering can speed up the operation. |
| threads | The number of threads for program operation. When left blank, the default value is used. |
| [check\_spanning\_reads] | spanning\_read\_flanking\_repeat\_length (default value: 1 bp) | The length of the read spanning the repeated sequence, which is a natural number. |
| spanning\_read\_number (default value: 1 bp) | The number of reads that meet the requirement of spanning the repeated sequence. |
| [refilter\_params] | refilter\_mode (default value: N) | Whether to refilter the reads spanning the repeated sequence. |
| refilter\_id | The ID of the project for refiltering the reads spanning the repeated sequence. |
| spanning\_read\_flanking\_repeat\_length (default value: 5 bp) | When refiltering the reads spanning the repeated sequence, the length of the read spanning the repeated sequence, which is a natural number. |
| spanning\_read\_number (default value: 5 bp) | When refiltering the reads spanning the repeated sequence, the number of reads that meet the requirement of spanning the repeated sequence. |

# VI. Detailed Explanation of ReHRI Operation Results

All the results after the operation of ReHRI are stored in the folder named {project\_id}.

All the information about the repeated sequences that have been queried is stored in the folder "final\_repeat-spanning\_results\_{project\_id}":

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. File ③ is excerpted from File ④ and can be used for the ReHRV software to draw the recombination genome map. In File ①, it stores the situation that among the two strands of the DNA molecule, there is no read spanning the repeated sequence on one of the strands. In File ②, it stores the situation that for the DNA molecule, there are no reads spanning the repeated sequences on both the positive and negative strands. Files ⑤ and ⑥ are the queried repeated sequences from two chromosomes, in the fasta format.

The folder "subconfig\_repeat-spanned\_results\_{project\_id}" stores the intermediate results of reads mapped to the TSS of the subconfigurations. The folder "mainconfig\_repeat-spanned\_results\_{project\_id}" stores the intermediate results of reads mapped to the TSSs of the mianconfigurations. Each folder stores the mapping results of one TSS, and the names of these folders are as follows:

DR\_AD\_RPxxa\_RPxxb\_plus\_1000\_results

DR\_AD\_RPxxa\_RPxxb\_minus\_1000\_results

DR\_CB\_RPxxa\_RPxxe\_plus\_1000\_results

DR\_CB\_RPxxa\_RPxxe\_minus\_1000\_results

IR\_AC\_RPxxb\_RPxxc\_plus\_1000\_results

IR\_AC\_RPxxb\_RPxxc\_minus\_1000\_results

IR\_BD\_RPxxa\_RPxxb\_plus\_1000\_results

IR\_BD\_RPxxa\_RPxx1b\_minus\_1000\_results

DR\_AB\_RPxxa\_RPxxb\_plus\_1000\_results

DR\_AB\_RPxxa\_RPxxb\_minus\_1000\_results

DR\_CD\_RPxxa\_RPxxe\_plus\_1000\_results

DR\_CD\_RPxxa\_RPxxe\_minus\_1000\_results

IR\_AB\_RPxxa\_RPxxb\_plus\_1000\_results

IR\_AB\_RPxxa\_RPxxb\_minus\_1000\_results

IR\_CD\_RPxxa\_RPxxe\_plus\_1000\_results

IR\_CD\_RPxxa\_RPxxe\_minus\_1000\_results

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

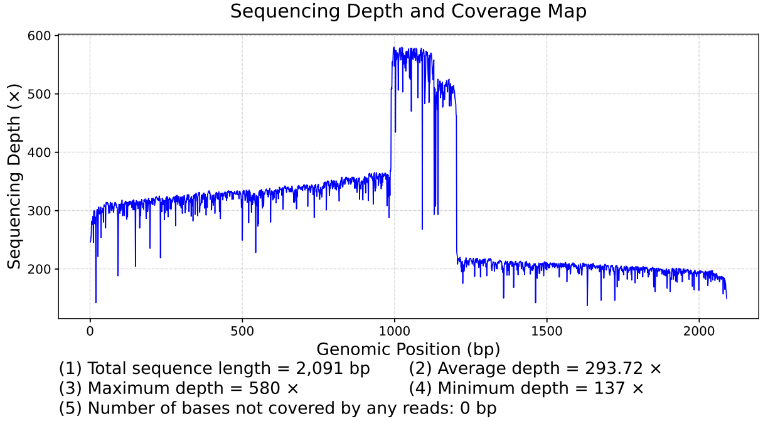
**Table 6-1 Meanings of Each Character in the Name of the Result Folder**

|  |  |
| --- | --- |
| **Symbol** | **Meaning of the Symbol** |
| DR | Direct Repeat |
| IR | Inverted Repeat |
| AC | In the subconfiguration, the TSS intercepted centered on the inverted repeat RPb, as shown in Figure 2-3 |
| BD | In the subconfiguration, the TSS intercepted centered on the inverted repeat RPa, as shown in Figure 2-3 |
| AD | In the subconfiguration, the TSS intercepted centered on the direct repeat RPa, as shown in Figure 2-3 |
| CB | In the subconfiguration, the TSS intercepted centered on the direct repeat RPb, as shown in Figure 2-3 |
| AB | In the mainconfiguration, the TSS intercepted centered on the direct repeat RPa, as shown in Figure 2-3 |
| CD | In the mainconfiguration, the TSS intercepted centered on the direct repeat RPb, as shown in Figure 2-3 |
| RP | Represents the repeated sequence |
| xx | Represents the serial number of the repeated sequence, which is a natural number |
| a/b/c … | Different repeat units of the same repeated sequence |
| plus | Represents the positive strand of DNA |
| minus | Represents the negative strand of DNA |
| 1000 | The length of the sequences intercepted on both sides of the repeated sequence, that is, L in Figure 2-3 |
| results | The suffix of the folder name |

Each folder contains the TSS sequence (in fasta format), the sequencing depth of the reads spanning the repeated sequences in the TSS mapped to the TSS, and the bam document of the reads mapped to the TSS (Figure 6-1).



**Figure 6-1 Various Results after Reads Mapped to Each TSS**



**Figure 6-2 Sequencing Depth of Reads Spanning the Repeat in the TSS**

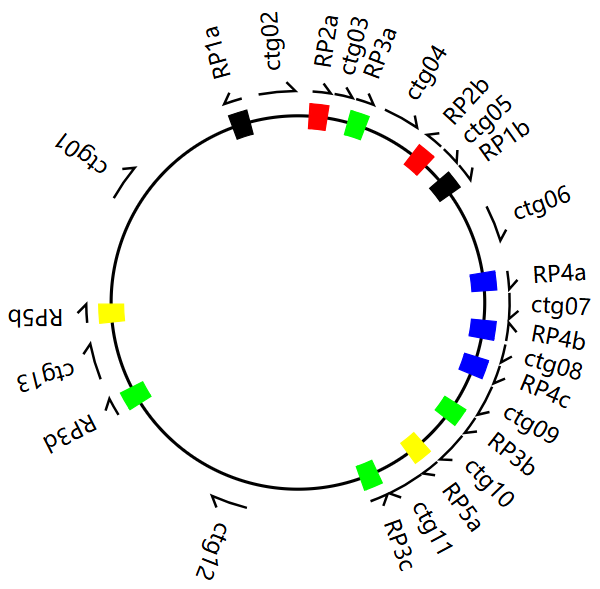
The sequencing depth of the reads spanning the repeated sequences in the TSS mapped to the TSS is shown in Figure 6-2, and the values of the sequencing depth are saved in the "coverage.txt". The bam document can be used to visualize the actual situation of reads mapped to the TSS with software such as Tablet, as shown in Figure 6-3.

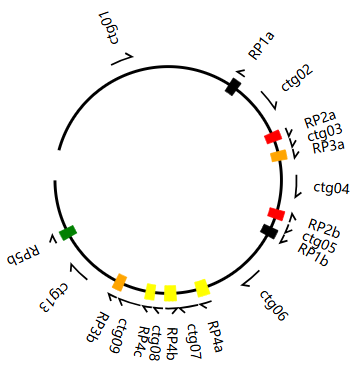


**Figure 6-3 Visualization Results of Reads Spanning the Repeated Sequences in the TSS**

# VII. Draws maps by ReHRV

The software ReHRV (**Mi**togenome **R**ecomb**i**nation **V**isulization) can, based on the results of the ReHRI software, draw a schematic diagram of the circular genome after recombination to display the genome maps of various subconfigurations of the mitochondrial genome mediated by repeated sequences. The map is represented in a circular shape. The arrows indicate the directions of the positive strand of the DNA molecule before and after recombination. The repeated sequences are represented by colored squares. Squares of the same color represent different units of the same repeated sequence, and different colors represent different repeated sequences (Figure 7-1 A). The map of a linear chromosome is a circular map with a gap (Figure 7-1 B). The size of the radius of the map indicates the length of the genome sequence.





**B**

**A**

**Figure 7-1 Schematic Diagram of the Genome Map Drawn by ReHRV**

## 1. Operation of ReHRV

The command line for running ReHRV is as follows:

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

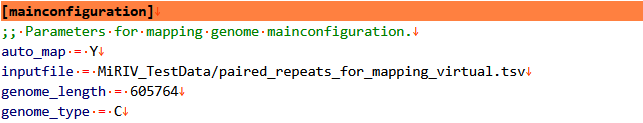
The parameters of ReHRV are provided in the form of a configuration file with the extension .ini. The vast majority of the parameters have default values, and only a limited number of parameters need to be provided by the user.

## 2. Configuration file of ReHRV

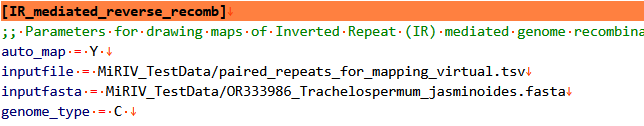
ReHRV can respectively draw the genome maps for the mainconfiguration (Figure 7-2 A), the subconfiguration mediated by IR (Figure 7-2 B), and the subconfigurations mediated by DR (Figure 7-2 C-E). When drawing the maps, users are required to provide the location information of the repeated sequences, the genome sequence (in fasta format), the length of the genome, and the type of the genome (that is, whether it is a linear or circular structure).

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

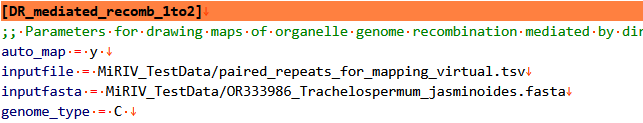
**A**

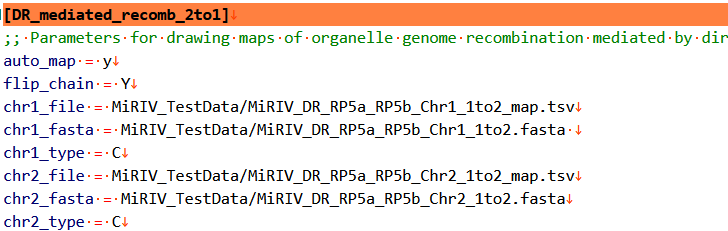


**B**



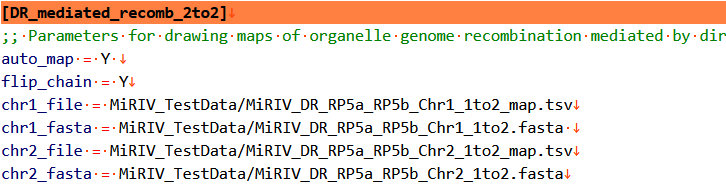
**C**





**D**

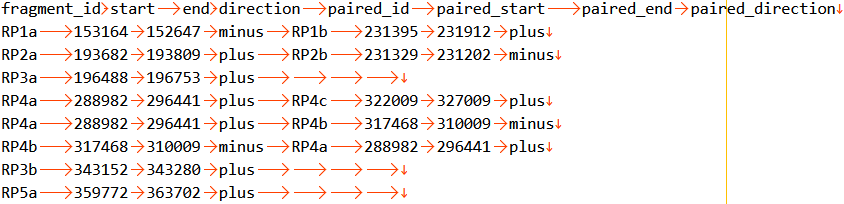
**E**



**Figure 7-2 Parameters Required When ReHRV Draws Various Maps**

When two or more chromosomes form a single chromosome mediated by DR, since ReHRV only allows the provision of two chromosomes at a time, users need to run ReHRV multiple times to recombine multiple chromosomes into one. When recombining two chromosomes into one, one of the chromosomes must be in a circular structure (C), and the parameter settings are shown in Figure 7-2D (chr1\_type, chr2\_type). When both chromosomes are linear (L), ReHRV can only achieve cross-recombination of sequences between the two chromosomes, and the result is still two linear chromosomes. The parameter settings are shown in Figure 7-2E.

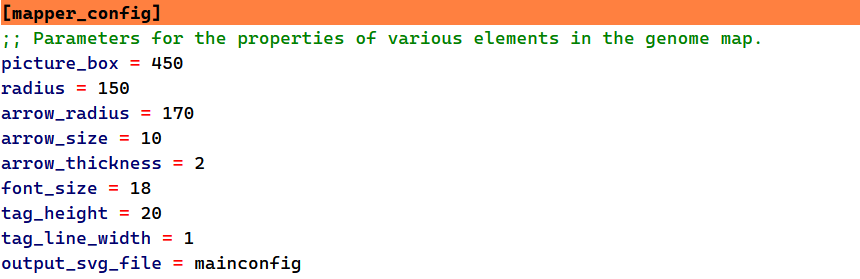
In the two modes of [DR\_mediated\_recomb\_2to2] and [DR\_mediated\_recomb\_2to1], the two chromosomes can rotate freely. Therefore, all repeat units can mediate the formation of a single chromosome from the genome through direct repeats. So, set flip\_chain = Y to allow all repeat units to mediate the formation of a single chromosome from the genome through direct repeats.

In all drawing modes, the formats of inputfile, chr1\_file, and chr2\_file are all 8-column tables in TSV format, as shown in Figure 7-3. Each line represents a pair of repeated sequences, and paired repeated sequences can be omitted. The table headers and the names of the repeated sequences must be the same as those shown in Figure 7-3.

**Figure 7-3 Schematic Diagram of the 8-Column Representation for Drawing (TSV Format)**

## 3. Parameter Settings for Each Element in the Map

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



**Figure 7-4 Settings of Each Option in [mapper\_config] section**

Default values of each parameter and their meanings are shown in Table 7-1:

**Table 7-1 Default Values and Meanings of Each Option in [mapper\_config] section**

|  |  |
| --- | --- |
| **Parameters** | **Parameter Values** |
| picture\_box | The size of the output image (the length of one side of the square), default value = 280 |
| radius | The radius of the genome map, which determines the size of the picture, default = 150 |
| arrow\_radius | The radius of the circle where the arrow is located, default = 170 |
| arrow\_size | The size of the arrow, default = 10 |
| arrow\_thickness | The thickness of the arrow line, Default = 2 |
| font\_size | The size of the font, Default = 18 |
| tag\_height | The height of the label (circular sector), Default = 20 |
| tag\_line\_width | The thickness of the outline of the label (circular sector), default = 1 |

## 4. Color parameters of each repeat

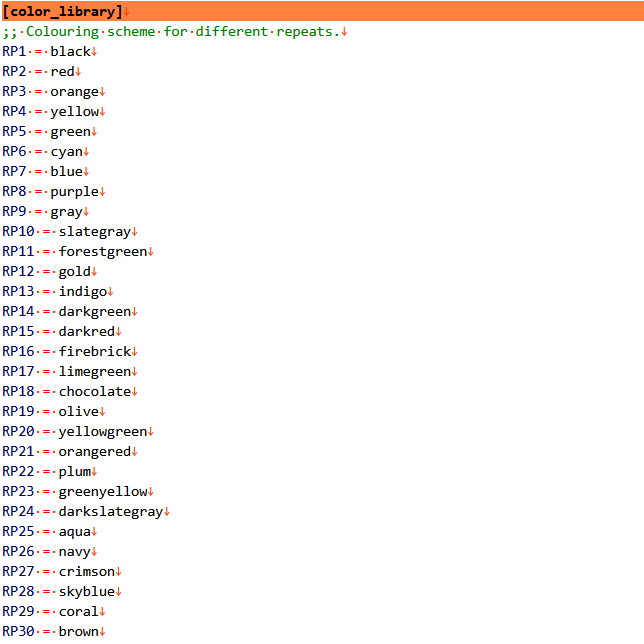
The [color\_library] section is used to set the colors of the repeat sequence units in the genome map. The colors can be represented by the English words of the 60 built-in colors, or by RGB values and hexadecimal codes.

The 60 built-in colors are screened from Python's webcolors library, and they are as follows: 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 settings in RGB are 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 colors represented in hexadecimal are as follows: #000000 (black), #FF0000 (red), #FFA500 (orange), #FFFF00 (yellow), #00FF00 (green), #00FFFF (cyan), #0000FF (blue), #800080 (purple).

ReHRV allows a maximum of 30 repeats to be colored simultaneously. The specific settings are shown in Figure 7-5:



**Figure 7-5 Coloring Scheme of Repeats in Genome Map**

## 5. Parameters for the layout of multiple maps

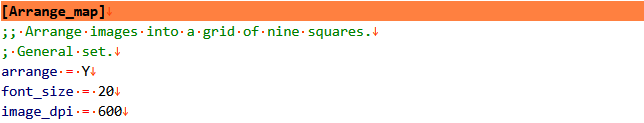
[Arrange\_map] is used to set the parameters for the layout of multiple maps. The maps are arranged in a nine-square grid layout (Figure 7-6 A):

arrange: Indicates whether to arrange multiple maps. "Yes/Y" means to perform the layout, and "No/N" means not to perform the layout.

font\_size: The size of the font for simple labels.

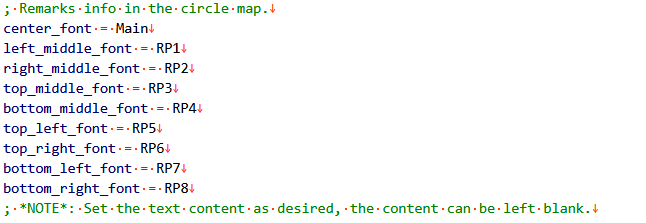
image\_dpi: The resolution of the image after the maps are arranged.

Each map can have a simple label set at its center (Figure 7-6 B). Moreover, the position of each map in the arranged image can be set independently (Figure 7-6 C). The set labels can consist of uppercase and lowercase letters, numbers, and underscores.

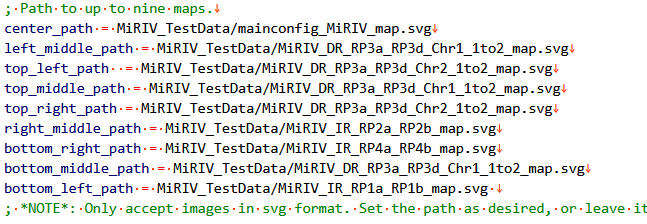


**A**

**B**

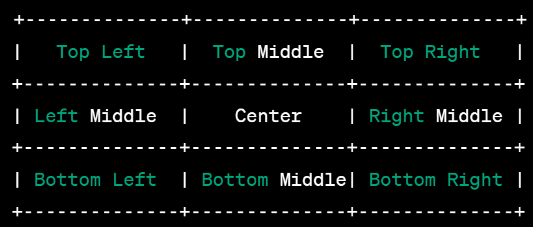


**C**



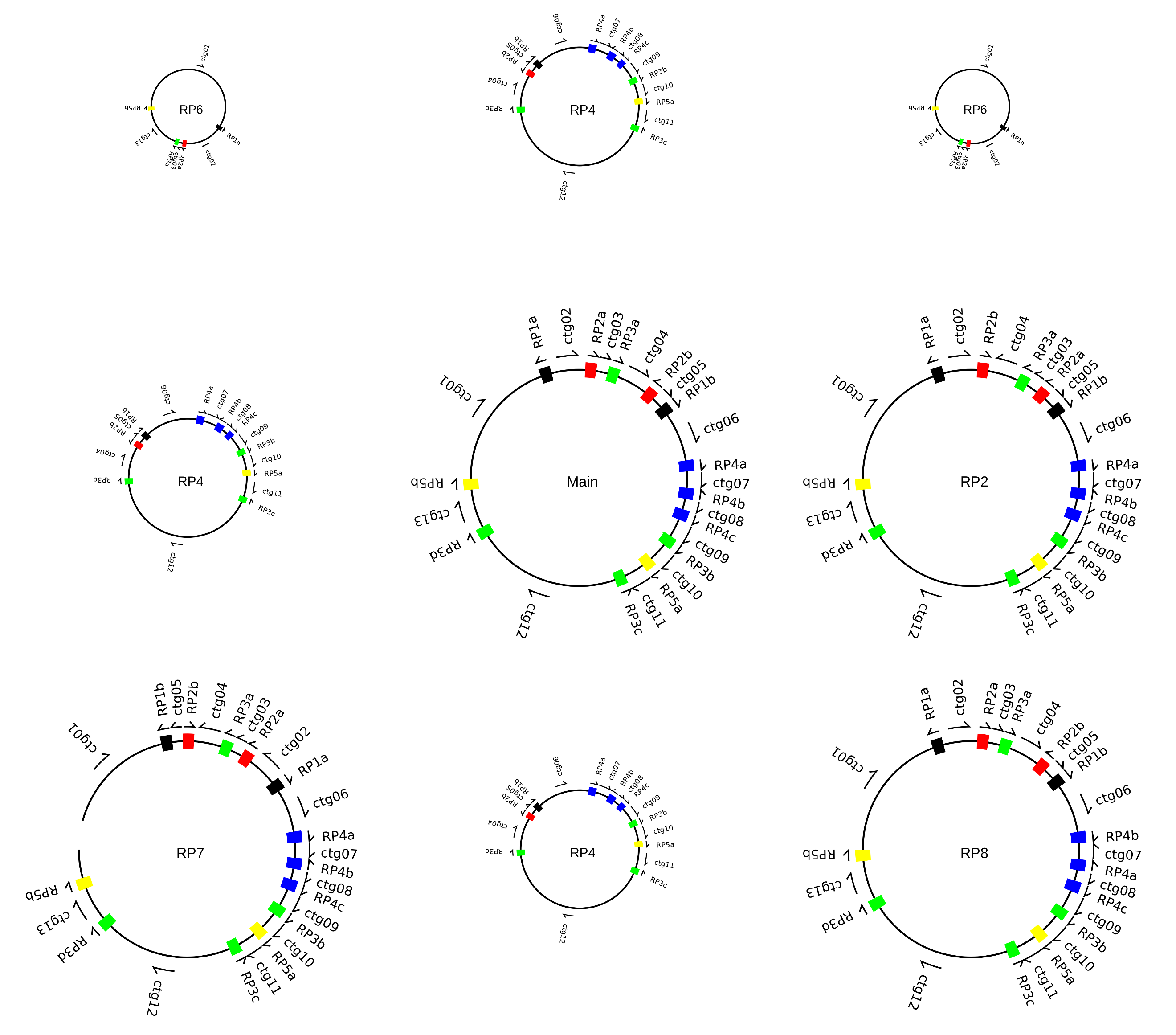
**Figure 7-6 Parameter Settings for the Layout of Multiple Maps**

The positions of the maps within the nine-square grid are shown in Figure 7-7 as follows:



**Figure 7-7 Schematic Diagram of the Positions of the Maps within the Nine-Square Grid**

The effect of the maps after being arranged within the nine-square grid is shown in Figure 7-8:



**Figure 7-8 Rendering Diagram after the Nine-square Grid Layout**

## 6. Interpretation of the Drawing Results in Each Mode

The results after running the modes of [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}

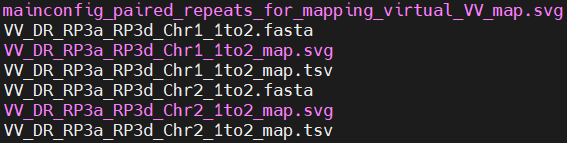
Taking the results of [DR\_mediated\_recomb\_1to2] as an example, there are three types of results:

Firstly, it is the map of the repetitive sequences of the mainconfiguration. It is generated according to the 8-column list (8CT.tsv) input by the user, in SVG format, and can be opened by software such as web browsers and Adobe Illustrator CS6.

Secondly, it is the first chromosome (chr1) generated from the input sequence. There are a total of three files, namely the sequence in FASTA format, the map of the repetitive sequences in SVG format, and the corresponding 8-column list of the map.

Thirdly, it is the result file of the second chromosome (chr2) generated from the input sequence.

The output results of other modes are similar to those of [DR\_mediated\_recomb\_1to2]. Among them, in the [DR\_mediated\_recomb\_2to1] mode, when three or more sequences need to be converted into one chromosome, the 8-column list (in TSV format) and the file in FASTA format in the previous step's results can be used as the input files for ReHRV in the next run.



**Figure 7-9 Schematic diagrams of the results output in each mode**

The outputfile include a map of the genomic mainconfiguration, as well as the genomic sequences, genomic maps about repeats, and pair information of repeats corresponding to the genomic subconfiguration.