细胞器基因组重组检测与重组体图谱绘制教程

——软件ReHRI与ReHRV的使用

# 一、软件安装

## 1.1、下载软件压缩包

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

## 1.2、解压压缩包并进入文件夹内

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

cd ReHRI\_ReHRV\_1.0-main

## 1.3、逐步法安装ReHRI与ReHRV

目前，ReHRI与ReHRV仅能在Linux（v20.0.4）系统下运行，且依赖conda（v24.5.0）软件。以下是安装命令，安装好之后，激活conda环境才能使用ReHRI与ReHRV。如果安装过程中软件的下载速度较慢，可以手工修改ReHRI\_ReHRV\_1.0.yml文件中的channels。

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

conda activate ReHRI\_ReHRV\_1.0

ReHRI\_ReHRV\_1.0为自定义的conda环境名，环境名必须以字母开头，只能包含字母、数字、下划线或英文句点，不能有空格或其他特殊符号。

更新plasmidrender程序：

cp -r bin/plasmidrender $(conda info --base)/envs/ReHRI\_ReHRV\_1.0/lib/  
python3.12/site-packages/（此命令为一行）

chmod +x bin/\*

## 1.5、验证安装结果

测试ReHRI.py的帮助文档

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.

测试ReHRI.py的版本

python bin/ReHRI.py -v

ReHRI version=1.0

测试ReHRV.py的帮助文档

python bin/ReHRV.py -h

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

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

**测试ReHRV.py的版本**

python bin/ReHRV.py -v

ReHRV version=1.0

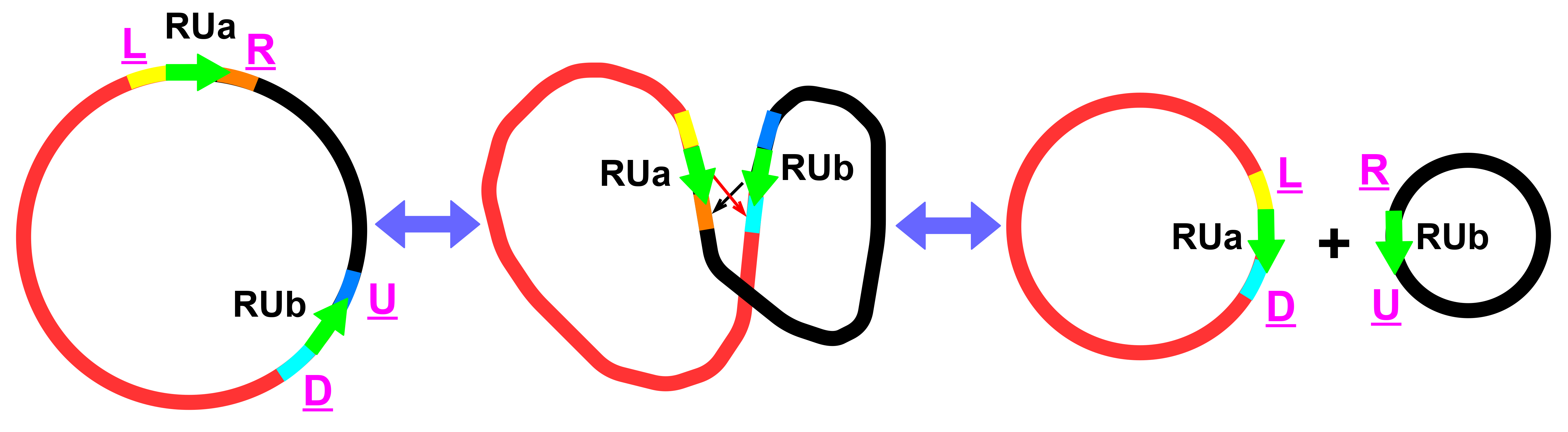
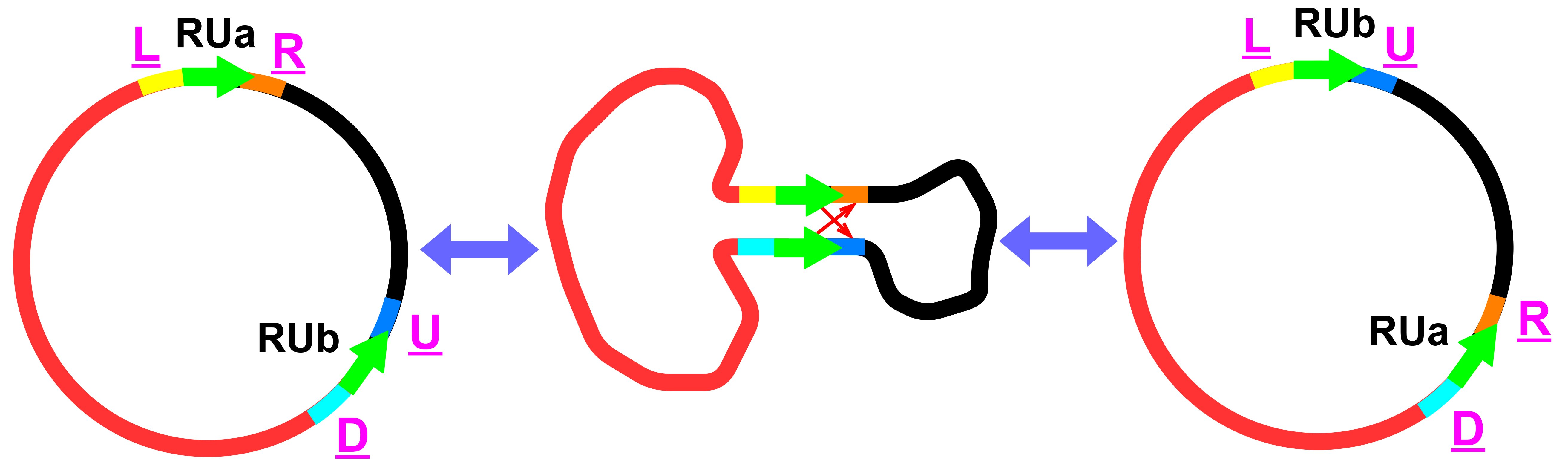
# 二、软件的运行原理

## 2.1、ReHRI的运行原理

ReHRI（**Re**peat-mediated **H**omologous **R**ecombination **I**dentification）设计的运行原理是反向重复序列（IRs）的重组会使中间序列发生倒位（图2-1A），而涉及正向重复序列（DRs）的重组则会产生一对亚基因组分子（图2-1B）。

**A**

**B**

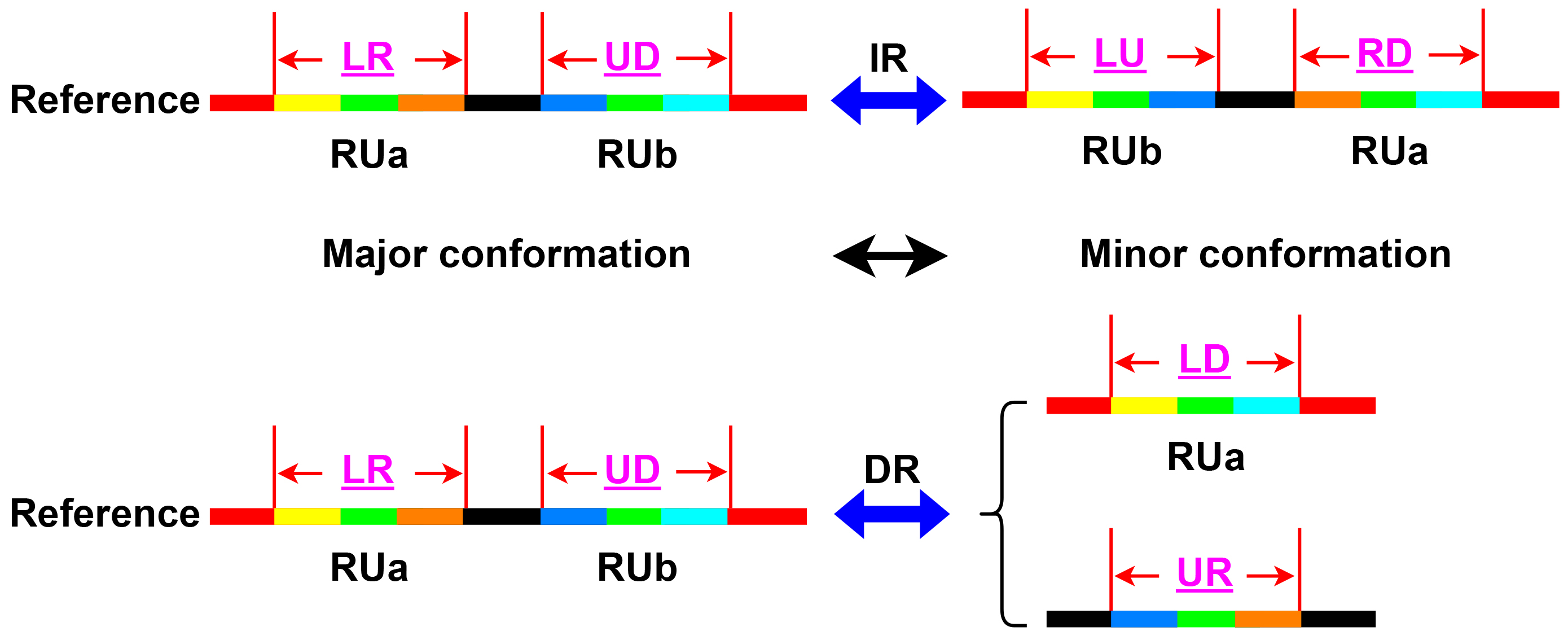


**B**

**图2-1 重复序列介导的环状基因组重组示意图**

A：由反向重复序列（IRs）介导的重组，B：由正向重复序列（DRs）介导的重组。

以重复序列为中心，分别从主要构型（major configuration = mainconfiguration）和次要构型（minor configuration = subconfiguration）中截取一段序列（Trimmed Reference Sequence，TRS），如图2-2所示，主要构型中分别以成对的重复序列RUa与RUb为中心，截取TRS为LR与UD，标记为TRSLR与TRSUD。IR介导的次要构型中，截取TRSLU与TRSRD，分别以重复序列RUb和RUa为中心（图1-1A）。DR介导的次要构型中，截取TRSLD与TRSUR，分别以重复序列RUa和RUb为中心（图2-2B）。



**A**

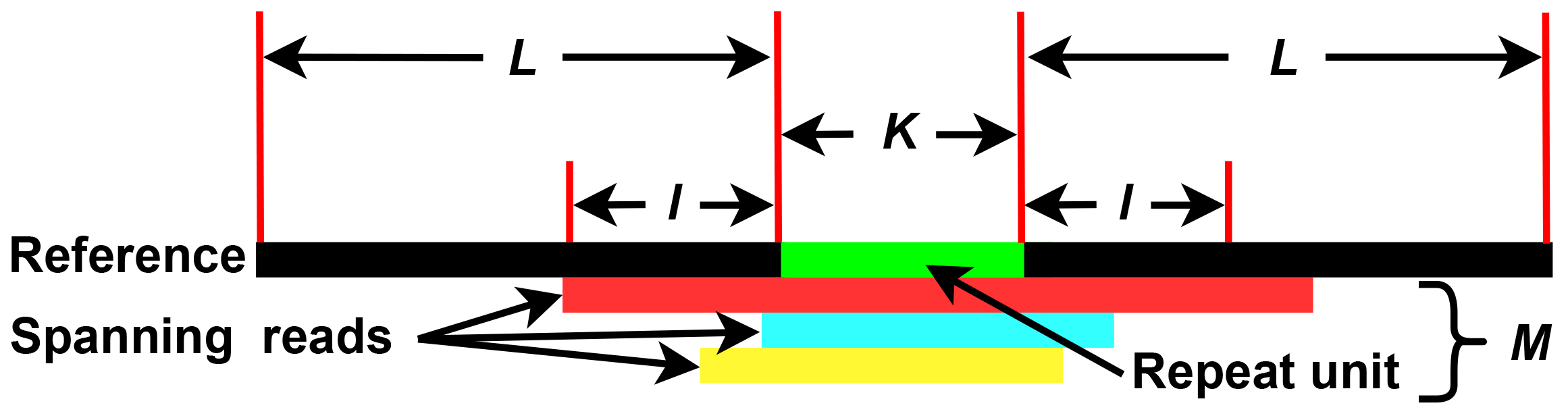
**A**

**B**

**图2-2 TRS截取示意图**

反向重复序列（IRs）介导的重组会使（A）成对重复单元的中间序列发生倒位，而正向重复序列（DRs）介导的重组则会使（B）一条染色体产生一对亚型染色体。LR、UD、LD、UR、LU和RD代表截取的TRS。LR和UD来自主构型。LD和UR来自由DRs介导的亚构型。LU和RD来自由IRs介导的亚构型。主要构型：major configuration = mainconfiguration，次要构型：minor configuration = subconfiguration。

如图2-3所示，在主要构型和次要构型中，以重复序列为中心，左右各截取*L*个碱基，获得TRS后，ReHRI将测序的read映射至TRS，然后从映射至TRS的read中查找可以跨越重复序列的read（*l* > 0），若有read跨越了重复序列（即*M* > 0），则认为该TRS对应的基因组构型存在。



**图2-3 read映射到TRS的示意图**

*L*：表示在一个重复单元两侧截取的序列长度。默认值为1000碱基对（bp）。

*K*：一个重复单元的长度。默认值为5个碱基对（bp）。

*I*：表示一条read跨越一个重复单元左右两侧的长度。当*I* ≥ 1碱基对（bp）时，该read被认为跨越了一个重复单元。

*M*：跨越一个重复单元且*I* ≥ 1碱基对（bp）的read的数量。

## 2.2、重组率的定义与计算

在计算重复序列介导的线粒体基因组重组的概率时，主要构型中，跨越TRSLR和TRSUD中的重复单元RUa与RUb的read数量记录为*M*LR与*M*UD。次要构型中，跨越TRSLD和TRSUR，以及TRSLU和TRSRD中的重复单元的RUa与RUb的read数量记录为*M*LD与*M*UR，以及*M*LU与*M*RD。

IRs介导基因组重组中，重复单元RUa介导重组的概率为：



以上是考虑单链（公式2-1）和双链（公式2-2）两种情况下（+/-代表正负链，下同）重组率的计算公式。

IRs介导基因组重组中，RUb介导重组的概率为：



以上是考虑单链（公式2-3）和双链（公式2-4）两种情况下重组率的计算公式。

DRs介导基因组重组中，RUa介导重组的概率为：



以上是考虑单链（公式2-5）和双链（公式2-6）两种情况下重组率的计算公式。

DRs介导基因组重组中，RUb介导重组的概率为：



以上是考虑单链（公式2-7）和双链（公式2-8）两种情况下重组率的计算公式。

# 三、ReHRI软件的运行

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

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

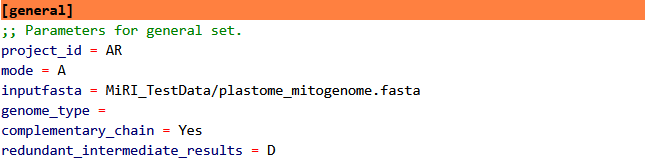
python bin/ReHRI.py -c ReHRI.config.ini -resume

ReHRI运行需要较多的参数，配置文件ReHRI.config.ini用于设置各种参数，但多数参数可以采用默认值，仅少数几个参数需要设置。当程序意外中断后，参数“-redo”允许用户删除之前的中间结果而重新计算，参数“-resume”允许用户基于之前未完成的结果继续计算，直至获得最终结果。

## 3.1、ReHRI的运行模式一

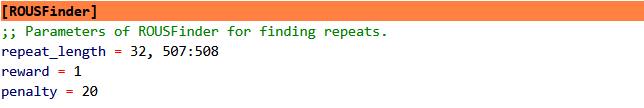
当mode=A时（图3-1），为ReHRI的第一种运行模式。ReHRI将先从基因组内查找正向和反向重复序列，然后再检测可以介导基因组重组的重复序列对。此时，用户必须提供基因组序列文件，查找的重复序列的长度，测序文件以及认定次要构型存在的参数。

基因组序列文件（inputfasta）为fasta格式，同时需要指明基因组的类型（genome\_type）为线性（L）还是环状（C）的。当基因组含多条染色体时，将所有的染色体置于同一个fasta文件内，其他参数可采用默认值，如图3-1所示。



**图3-1** **[general]中的基本参数**

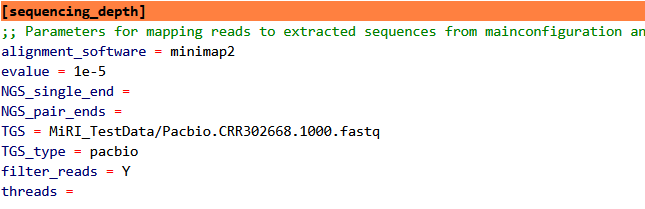
查找重复序列时，其长度（repeat\_length）即为图2-2中的*K*值，可设置为一个区间，如50bp≤长度≤1000bp，设置为50:1000；长度≥50bp，设置为50:。也可以设置为几个长度，如50bp、100bp，可设置为50,100（图3-2）。逗号和冒号必须是英文状态下的，空格不是必须的。其他参数可采用默认值。



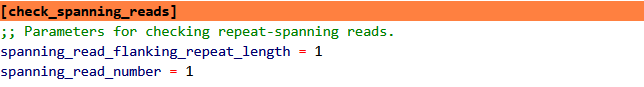
**图3-2 [ROUSFinder]中重复序列长度的参数**

对于测序数据，ReHRI可接受二代（NGS）和三代（TGS）数据，NGS和TGS的数据一次仅能接受其中一种（图3-3）。提供双端数据时，双端数据的文件要以空格隔开。当提供TGS数据时，还要指明数据来自Nanopore测序平台（ont）还是Pacbio测序平台（pacbio）（图3-3）。

检测次要构型存在的参数主要是spanning\_read\_flanking\_repeat\_length和spanning\_read\_number（图3-4）。这是判定次要构型是否存在的重要参数，默认值均为1，后续的重过滤模式（refilter\_mode=Y）中可对其重新设置，以对查询的重复序列介导基因组的重组结果进行多次筛选。其他参数可采用默认值。



**图3-3 [sequencing\_depth]中比对软件和测序数据的参数**



**图3-4 [check\_spanning\_reads]中次要构型存在的认定参数**

spanning\_read\_flanking\_repeat\_length是read跨越重复序列后的长度，即图2-3中的*l*值。spanning\_read\_number是跨越重复单元后的长度≥ *l*的read的数量，即图2-3中的*M*值。此处建议两个参数值均为1，获得的结果可在模式三中重新筛选。

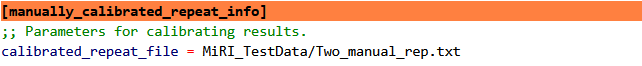
比对软件中，minimap2为默认软件。minimap2、bwa和blast的适用场景如表3-1所示。

**表3-1 minimap2、bwa和blast适用场景比较**

|  |  |  |
| --- | --- | --- |
| **工具** | **最佳应用场景** | **可能漏检原因** |
| minimap2 | 长读长、基因组组装、快速比对 | 短读长、高重复区域、默认参数宽松 |
| BWA | 短读长、变异检测、精确比对 | 长读长处理效率低 |
| BLAST | 同源性搜索、跨物种比较 | 计算耗时，不适合大规模比对 |

## 3.2、ReHRI的运行模式二

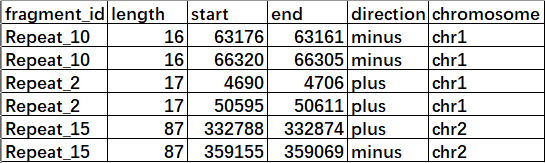
查询重复序列的工作极具挑战，不同的重复序列结果，对介导基因组重组的结果影响较大，而且不同算法查找的重复序列的结果不同。所以，ReHRI设置了可接受用户提供重复序列信息的接口，允许用户自己提供重复序列（图3-5）。此时，需设置mode=C（图3-6）。用户提供的重复序列信息文件为tsv格式（图3-7）。配对的重复序列，采用相同的fragment\_id。当基因组仅有一条染色体的时候，需要去掉chromosome列。ReHRI会对以上fragment\_id相同的重复序列单元所有的两两组合进行检测。如果想特异地检测某些成对重复单元对基因组重组的介导作用，用户可以按照图3-8所示的格式给ReHRI提供配对重复单元信息文件。当基因组仅有一条染色体的时候，需要去掉chromosome和paired\_chromosome两列。



**图3-5 [manually\_calibrated\_repeat\_info]中用户提供重复序列信息**

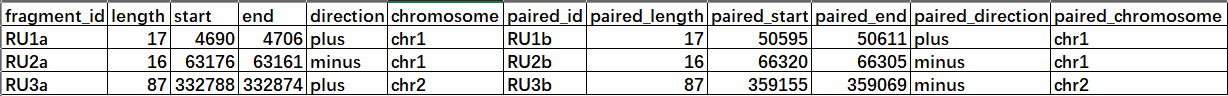


**图3-6 ReHRI 进入模式二的设置参数**



**图3-7** **用户提供的重复序列信息示例（tsv格式）**

chr1，chr2，chr3，……：为染色体的编号，必须采用此格式，其编号顺序表示的是[general]中inputfasta文件中染色体的排列顺序。同一重复序列的不同重复单元的fragment\_id必须相同。

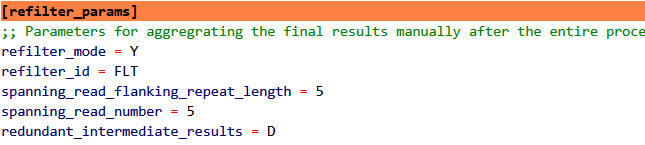


**图3-8 用户提供的配对重复序列信息示例（tsv格式）**

chr1，chr2，chr3，……：为染色体的编号，必须采用此格式，其编号顺序与[general]中inputfasta文件中染色体的排列顺序相同

## 3.3、ReHRI的运行模式三

若对初次筛选结果不满意，用户可设置refilter\_mode = Y（图3-9），对spanning\_read\_flanking\_repeat\_length和spanning\_read\_number重新进行设置，对查询结果重新进行筛选，以获取更满意的结果。



**图3-9 [refilter\_params]中的重过滤条件**

## 3.4、ReHRI示例数据详解

ReHRI所用的各个数据示例数据及其使用场景见表3-2。

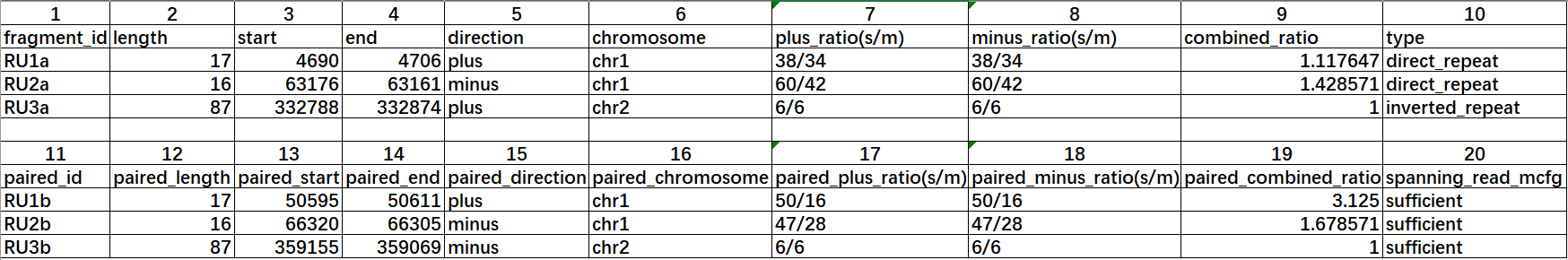
表3-2 ReHRI所用示例数据

|  |  |  |
| --- | --- | --- |
| **文件名** | **文件属性** | **适用场景** |
| CRR302670\_f1.10000.fasta | NGS的forward端测序数据，可以接受fastq或fasta格式 | 测试NGS数据中是否存在支持基因组重组的reads |
| CRR302670\_f1.10000.fastq |
| CRR302670\_r2.10000.fasta | NGS的reverse端测序数据，可以接受fastq或fasta格式 |
| CRR302670\_r2.10000.fasta |
| Pacbio.CRR302668.1000.fasta | TGS的测序数据，可以接受fastq或fasta格式 | 测试TGS数据中是否存在支持基因组重组的reads |
| Pacbio.CRR302668.1000.fastq |
| NC\_000932.1.fasta | 拟南芥的叶绿体基因组 | 测试叶绿体基因组的重组 |
| NC\_037304.1.fasta | 拟南芥的线粒体基因组 | 测试线粒体基因组的重组 |
| plastome\_mitogenome.fasta | 拟南芥的两种细胞器基因组 | 测试多条染色体间的重组 |
| One\_manual\_rep.tsv | 单条染色体内的重复序列 | 测试重复序列所有重复单元对是否能介导基因组重组 |
| One\_manual\_rep\_corr.tsv | 单条染色体内的成对的重复序列 | 测试重复序列指定的重复单元对是否能介导基因组重组 |
| Two\_manual\_rep.tsv | 多条染色体内的重复序列 | 测试多条染色体内/间的重复序列所有重复单元对是否能介导基因组重组 |
| Two\_manual\_rep\_corr.tsv | 多条染色体内/间的成对的重复序列 | 测试多条染色体内/间指定的重复序列对是否能介导基因组重组 |

# 四、ReHRI核心结果解读

ReHRI的运行结果存储在文件夹{project\_id}/final\_repeat-spanning\_results\_{project\_id}中的paired\_repeats\_recomb-supporting\_ratio.tsv文件内。

该结果为一个20列的tsv文件，如图4-1所示。



**图4-1** ReHRI预测的**介导基因组重组的重复序列的信息**

每一列的含义如下：

① fragment\_id：重复单元的编号。

② length：重复单元的长度。

③ start：重复单元在基因组中的起始位置。

④ end：重复单元在基因组中的终止位置。

⑤ direction：在DNA正链或负链上的位置。

⑥ chromosome：基因组中某条染色体的编号。

⑦ plus\_ratio：在DNA正链上，亚构型和主构型中跨越重复序列的read数量之比。

⑧ minus\_ratio：在DNA负链上，亚构型和主构型中跨越重复序列的read数量之比。

⑨ combined\_ratio：DNA两条链上重复序列介导的基因组重组的总体比例。

⑩ type：重复序列的类型（正向重复或反向重复）。

⑪ spanning\_read\_mcfg：主要构型中，跨越重复序列的read数量是否符合用户设置的数量。

注：“配对项（paired\_\*）”指的是介导基因组重组的成对重复单元中的另一个重复单元。

# 五、ReHRI配置文件的详解

用户可通过更加详细的设置.ini配置文件来挖掘ReHRI的性能，表5-1是对.ini配置文件各参数的详细解读。

**表5-1 配置文件中参数详解**

|  |  |  |
| --- | --- | --- |
| **参数类别** | **参数** | **取值与含义** |
| [general] | project\_id（必选参数） | 项目编号，由字母、数字和下划线组成 |
| mode（默认值 A） | 软件运行的模式，取值为\*N/A/R/C\*（大小不敏感）。N：程序不运行；A：程序自动运行；R：仅运行ROUSFinder软件查找重复序列；C：从用户提供的重复序列中查找跨重复序列的read，此时[calibrate\_ROUSFinder\_results]类别中的“calibrated\_repeat\_file”必须提供。 |
| inputfasta（必选参数） | 细胞器基因组序列文件，可包含多条染色体。 |
| genome\_type（默认值C） | 设置基因组为线性（L）或环状（C）。 |
| complementary\_chain（默认值Y） | 查找跨越重复序列的read时，考虑DNA的双链（Y）或不考虑（N）。 |
| redundant\_intermediate\_results（默认值D） | 软件运行时，删除中间结果（D）或者不删除（K）。 |
| [ROUSFinder] | repeat\_length（默认值50:） | 重复序列长度区间。长度≥50bp，设置为50：。长度≤100bp，设置为：100。100bp ≤长度≤ 200bp，设置为100:200。最小值为5bp。 |
| reward（默认值1） | ROUSFinder参数，查找重复序列时，序列比对的奖励值。 |
| penalty（默认值20） | ROUSFinder参数，查找重复序列时，序列比对的惩罚值。 |
| [manually\_calibrated\_repeat\_info] | calibrated\_repeat\_file | 手工校准重复序列结果的输入文件位置，mode=C时为必选参数。 |
| [mainconfiguration] | flanked\_sequence\_length（默认值1000bp） | 在主要构型中，重复序列单元左右两测截取的序列长度，单位为bp。 |
| [subconfiguration] | flanked\_sequence\_length（默认值1000bp） | 在次要构型中，重复序列单元左右两测截取序列的长度，单位为bp。 |
| [sequencing\_depth] | alignment software（默认值minimap2） | 比对软件，可选minimap2，bwa或者blast。 |
| evalue（默认值1e-5） | blast的参数，用于衡量匹配结果的显著性。 |
| NGS\_single\_end | 二代单端测序数据（fastq或fasta格式）。 |
| NGS\_pair\_ends | 二代双端测序数据（fastq或fasta格式）。 |
| TGS | 三代测序数据（fastq或fasta格式）。 |
| TGS\_type | 设置三代测序平台类型（pacbio或ont），TGS参数的补充参数。 |
| filter\_reads（默认值Y） | 是否过滤测序read，过滤后可加快运行速度。 |
| Threads（默认值90%的物理线程） | 线程数，为空时采用默认值。 |
| [check\_spanning\_reads] | spanning\_read\_flanking\_repeat\_length（默认值1bp） | read跨越重复单元的序列长度，为自然数。 |
| spanning\_read\_number（默认值1 bp） | 符合跨越重复单元的read的条数。 |
| [refilter\_params] | refilter\_mode（默认值N） | 是否再次过滤跨越重复单元的read。 |
| refilter\_id | 再次过滤跨越重复单元的项目编号。 |
| spanning\_read\_flanking\_repeat\_length（默认值5 bp） | 再次过滤跨越重复序列的read时，跨越重复序列的长度，为自然数。 |
| spanning\_read\_number（默认值5） | 再次过滤跨越重复序列的read时，符合跨越重复序列的read的条数，为自然数。 |

# 六、ReHRI运行结果的详解

以{project\_id}为名称的文件夹内存储着ReHRI运行后的所有结果。

文件夹final\_repeat-spanning\_results\_{project\_id}内存储着查询到的所有关于重复序列的信息：

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

文件④为核心结果（解读见第五部分），文件③截取自文件④，可用于ReHRV软件绘制重组基因组图谱。文件①存储的是DNA分子的两条链中，有一条链中的重复序列没有可跨越的read。文件②存储的是DNA分子中，正负两条链中的重复序列均没有可跨越的read。文件⑤⑥是来自两条染色体的重复序列，为fasta格式。

文件夹subconfig\_repeat-spanned\_results\_{project\_id}存储的是read映射至次要构型的TRS上的中间结果。mainconfig\_repeat-spanned\_results\_{project\_id}存储的是read映射至主要构型的TRS上的中间结果。每一个文件夹存储一条TRS的映射结果，其文件夹的名字如下所示：

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

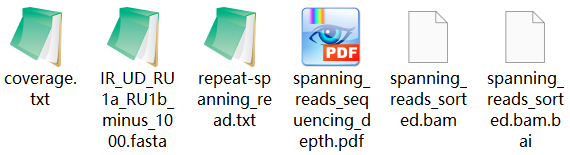
IR\_UD\_RUxxa\_RUxxe\_minus\_1000\_results

文件夹名字各部分的命名规则如表6-1所示：

**表6-1 结果文件夹名中各字符的含义**

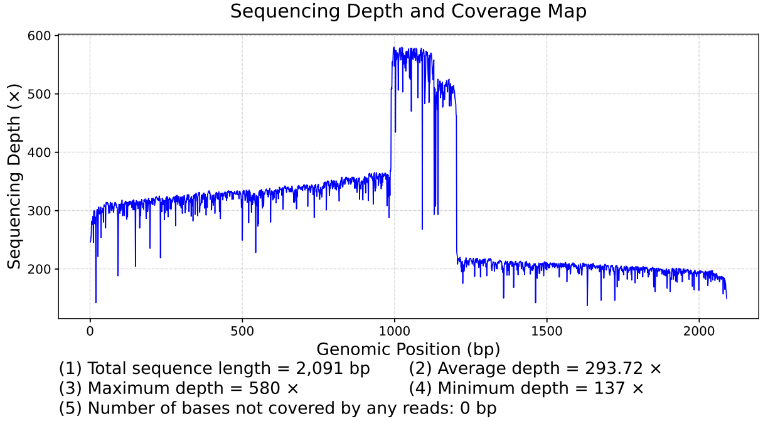
|  |  |
| --- | --- |
| **符号** | **符号的含义** |
| DR | 正向重复序列 |
| IR | 反向重复序列 |
| LU | 次要构型中，以反向重复单元RUb为中心截取的TRS，见图2-3 |
| RD | 次要构型中，以反向重复单元RUa为中心截取的TRS，见图2-3 |
| LD | 次要构型中，以正向重复单元RUa为中心截取的TRS，见图2-3 |
| UR | 次要构型中，以正向重复单元RUb为中心截取的TRS，见图2-3 |
| LR | 主要构型中，以正向重复单元RUa为中心截取的TRS，见图2-3 |
| UD | 主要构型中，以正向重复单元RUb为中心截取的TRS，见图2-3 |
| RU | 重复单元，即Repeat Unit |
| xx | 编号，为自然数 |
| a/b/c … | 同一重复序列的不同重复单元 |
| plus | 表示DNA的正链 |
| minus | 表示DNA的负链 |
| 1000 | 重复序列左右两侧截取的序列长度，即图2-3中的*L* |
| results | 文件夹名的后缀 |

每个文件夹内包含了TRS序列（fasta格式），跨越TRS中重复序列的read映射至TRS的测序深度，read映射至TRS的bam文档（图6-1）。



**图6-1 read映射至每一条TRS后的各种结果**

跨越TRS中重复序列的read映射至TRS的测序深度如图6-2所示，其测序深度的数值保存在coverage.txt中。bam文档可用Tablet等软件可视化read映射至TRS的实际情况（图6-3）。



**图6-2 跨越TRS中重复序列的read映射至TRS的测序深度**

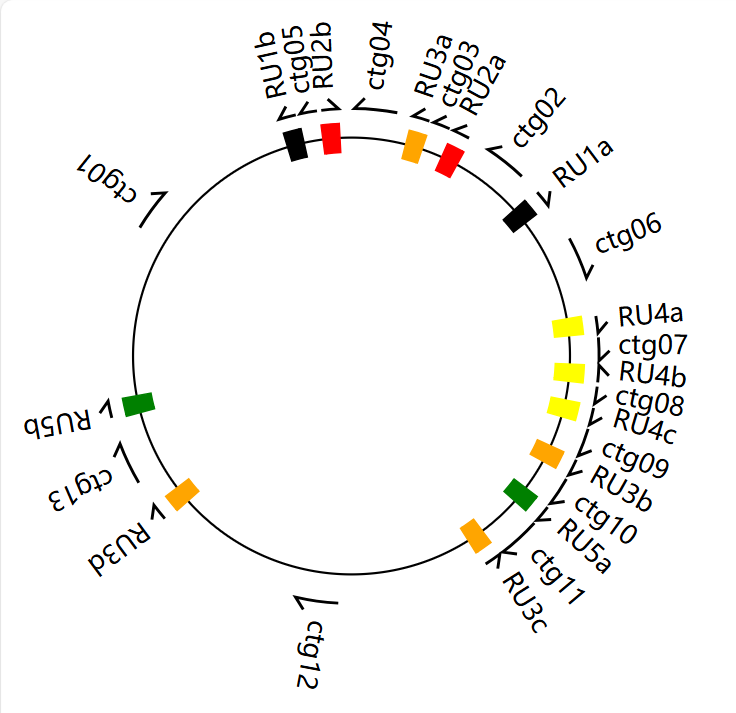


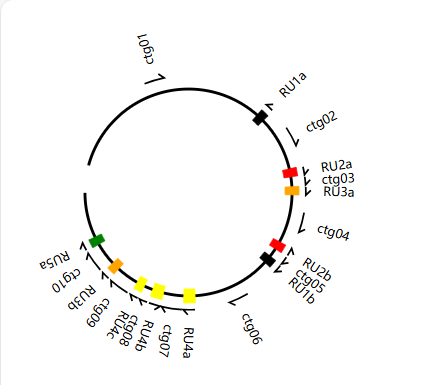
**图6-3 跨越TRS中重复序列的read映射至TRS的可视化结果**

ReHRI对初次筛选的结果再次进行筛选的时候，即ReHRI执行模式三后，其结果存放在以{project\_id}为名的文件夹内的refiltered\_repeat-spanning\_results\_{FLT}文件夹内。其结果的解读参考对文件夹final\_repeat-spanning\_results\_{project\_id}内结果的解读。

# 七、ReHRV绘制重复序列的基因组图谱

软件ReHRV（**Re**peat-mediated **H**omologous **R**ecombination **V**isualization）可以从ReHRI软件的结果出发，绘制环状基因组重组后的示意图，以展示由重复序列介导后线粒体基因组的各种亚型的基因组图谱。图谱以环状表示，箭头表示DNA分子正链重组前后的走向，重复单元（Repeat unit，RU）用带颜色的方块表示，相同颜色的方块表示同一重复序列的不同单元，不同的颜色表示不同的重复序列，ctg表示两相邻重复单元的间区（图7-1A）。线性染色体的图谱为一个带缺口的环状图谱（图7-1B）。图谱半径的大小表示基因组序列的长短。





**B**

**A**

**图7-1 ReHRV绘制的基因组图谱示意图**

## 7.1、ReHRV的运行

运行ReHRV的命令行如下：

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

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

ReHRV的参数以配置文件.ini的形式提供。绝大多数参数提供了默认值，仅有限几个参数需要用户提供。参数“-redo”允许用户在程序意外中断后，重新进行重复序列基因图谱的绘制。此过程会删除前一次运行产生的中间结果。

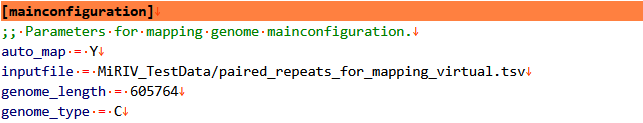
## 7.2、ReHRV的配置文件

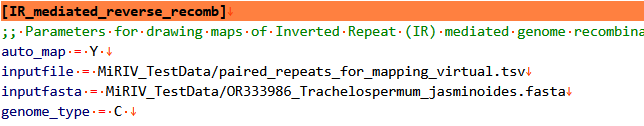
ReHRV可分别对主要构型（[mainconfiguration]模式）（图7-2A）、IRs介导产生的次要构型（[IR\_mediated\_reverse\_recomb]模式）（图7-2B）和DRs介导产生的次要构型（[DR\_mediated\_recomb\_1to2]模式）（图7-2C）绘制其基因组图谱。绘制图谱的时候，需要用户提供重复序列的位置信息、基因组的序列（fasta格式）、基因组的长度以及基因组的类型（即线性还是环状结构）。

每一个模式均设置了auto\_map参数，用于控制是否运行（Y/N/M）相应的模式。auto\_map=N时，ReHRV不运行相应的模式。auto\_map=Y，ReHRV将自动绘制所有的基因组图谱。auto\_map=M，ReHRV将在用户的指导下绘制用户指定的基因组图谱。

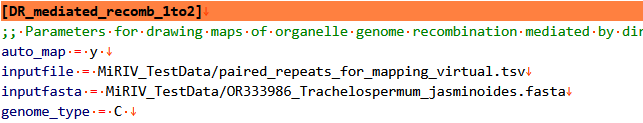
当遇到两条及多条染色体在DR介导下形成一条染色体的情况时，由于ReHRV每次只允许用户提供两条染色体，所以需要用户运行多次ReHRV（[DR\_mediated\_recomb\_2to1]模式），以将多条染色体重组为一条染色体。两条染色体重组为一条染色体的时候，其中一条染色体必须为环状结构（C），参数设置如图7-2D所示（chr1\_type、chr2\_type）。当两条染色体均为线性（L）的时候，ReHRV仅能实现两条染色体间序列的交叉重组（[DR\_mediated\_recomb\_2to2]模式），结果仍然是两条线性染色体，参数设置如图7-2E所示。

**A**

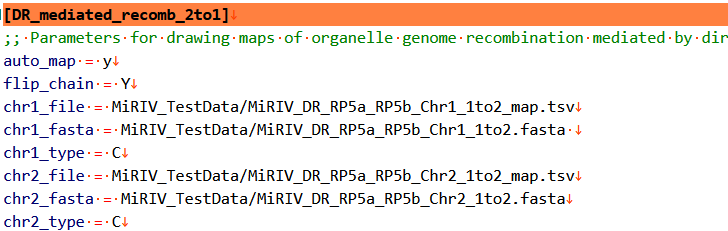




**B**

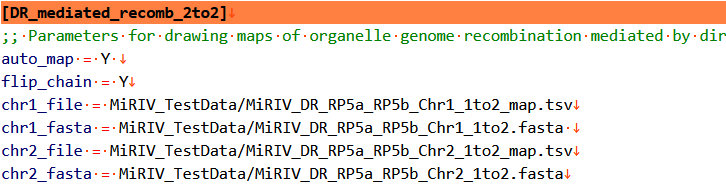


**C**



**D**

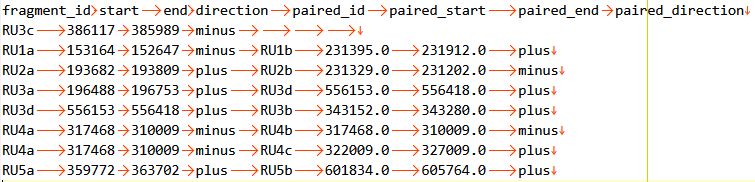
**E**



**图7-2 ReHRV绘制各种基因组图谱时所需的参数**

在[DR\_mediated\_recomb\_2to2]和[DR\_mediated\_recomb\_2to1]的两个模式中，由于两条染色体可以自由旋转，所以所有可以介导基因组重组的重复单元之间都可以以正向重复的形式介导两条染色体形成一条染色体。所以，当令flip\_chain=Y时，即允许所有的可以介导基因组重组的重复单元之间均以正向重复的形式介导两条染色体形成一条染色体。

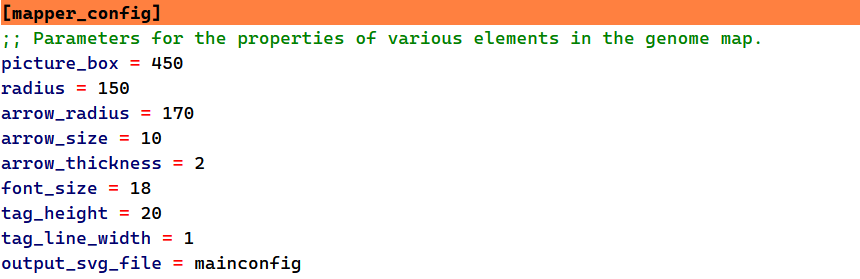
所有绘图模式中，inputfile、chr1\_file和chr2\_file的格式均为tsv格式的8列表，如图7-3所示。每一行表示一对重复序列，配对的重复序列可以缺省。表头和重复序列的名字必须与图7-3所示一致。



**图7-3 ReHRV绘图用8列表示意图（tsv格式）**

## 7.3、基因组图谱中各元素的参数设置

[mapper\_config]选项用于设置绘制的基因组图谱中各元素的属性参数。各参数值如图7-4所示：



**图7-4 [mapper\_config]中基因组图谱各元素的设置参数**

各参数的默认值及其含义如表7-1所示：

**表7-1 [mapper\_config]各选项的默认值及其含义**

|  |  |
| --- | --- |
| **参数** | **参数的取值** |
| picture\_box | 输出图像的大小（正方形一边的长度），\*\*默认值=280\*\* |
| radius | 基因组图谱的半径，决定图片的大小，\*\*默认值=150\*\* |
| arrow\_radius | 箭头所在圆的半径，\*\*默认值=170\*\* |
| arrow\_size | 箭头的大小，\*\*默认值=10\*\* |
| arrow\_thickness | 箭头线的粗细，\*\*默认值=2\*\* |
| font\_size | 字体大小，\*\*默认值=18\*\* |
| tag\_height | 标签（环状扇形）的高度，\*\*默认值=20\*\* |
| tag\_line\_width | 标签（环状扇形）轮廓宽度，\*\*默认值=1\*\* |

## 7.4、重复序列的颜色参数

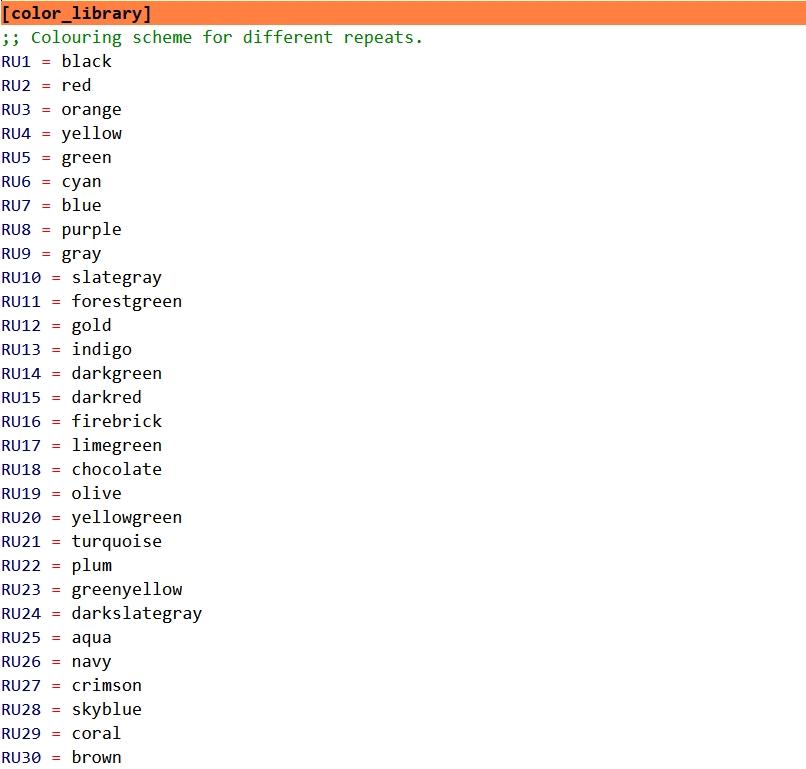
[color\_library]选项用于设置基因组图谱中重复序列的颜色。颜色可用60种内置颜色的英语单词表示，也可以用RGB和16进制数值表示。

内置的60种颜色筛选自python的webcolors库，分别为：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.

RGB的颜色数值设置如：0.0.0（黑），255.0.0（赤），255.165.0（橙），255.255.0（黄），0.255.0（绿），0.255.255（青），0.0.255（蓝），128.0.128（紫）。

16进制表示颜色的数值如：#000000（黑），#FF0000（赤），#FFA500（橙），#FFFF00（黄），#00FF00（绿），#00FFFF（青），#0000FF（蓝），#800080（紫）。

ReHRV最多允许30个重复单元同时上色，具体的设置如图7-5所示：



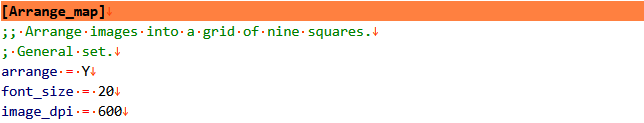
**图7-5 基因组图谱内重复序列上色方案**

## 7.5、多个基因组图谱排版的参数

[Arrange\_map]用于设置多个图谱排版的参数，图谱按照九宫格排版（图7-6A）。

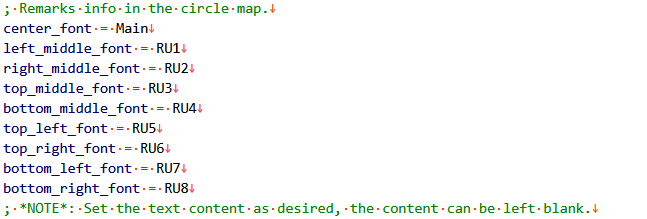
arrange：是否要对多个基因组图谱进行排版。Yes/Y表示排版，No/N表示不排版。font\_size：设置标签字体大小。image\_dpi：排版后图片的分辨率。可采用默认值。

每个基因组图谱可以在其中间位置设置简单标签（图7-6B），标签可由大小写字母、数字和下划线组成。每个基因组图谱可以独立设置在排版后图片中的位置（图7-6C）。

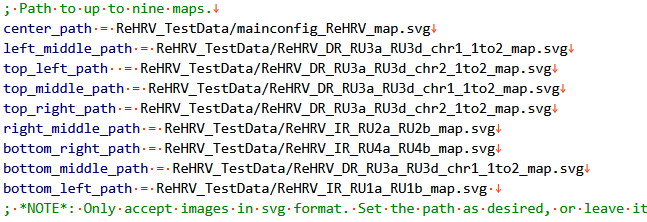


**A**

**B**

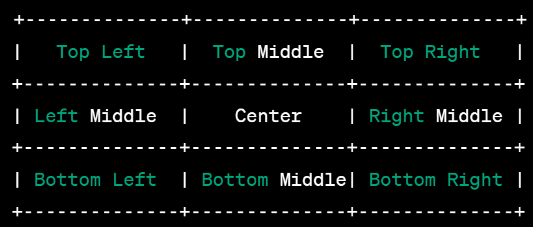


**C**



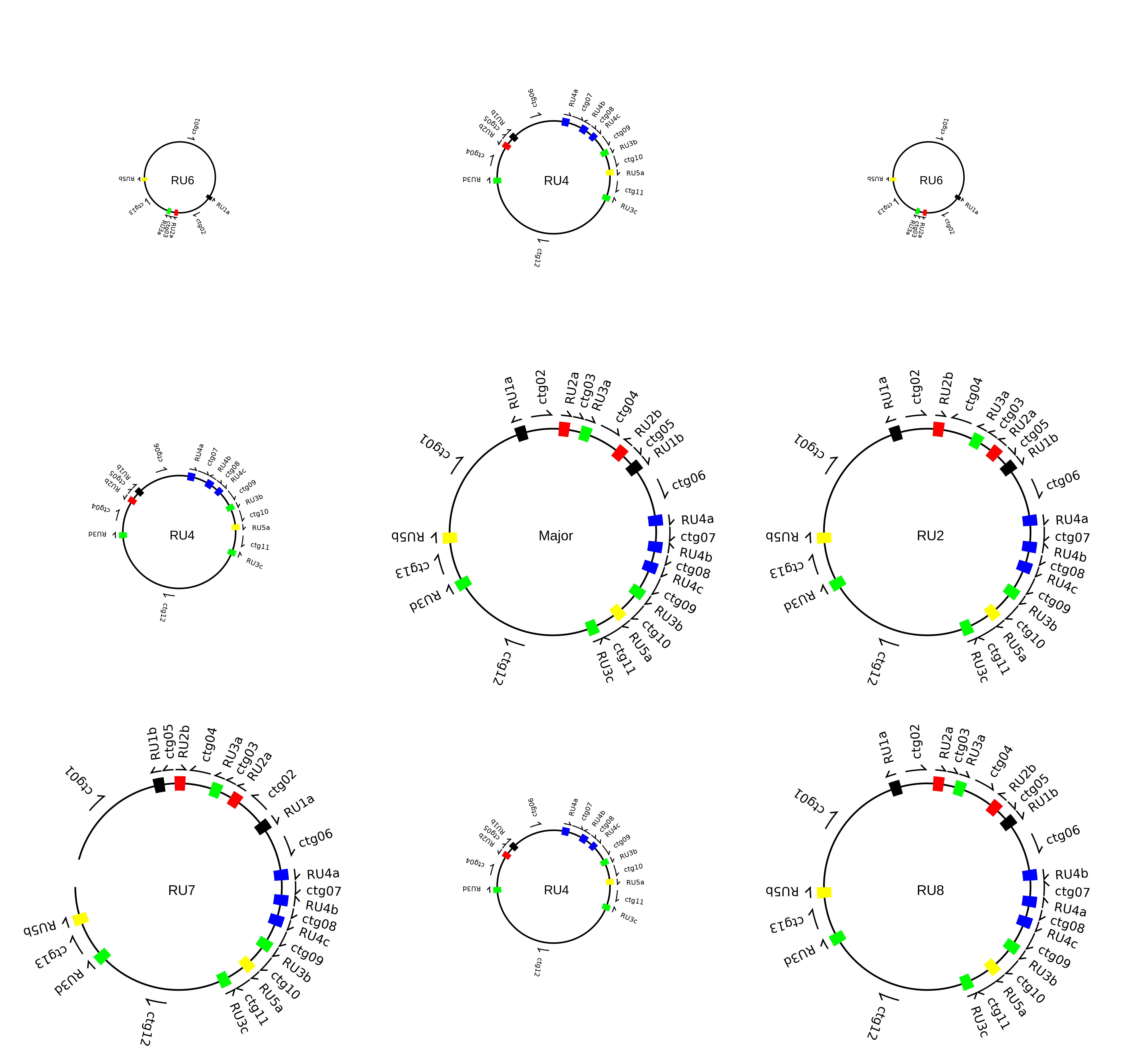
**图7-6 多个基因组图谱排版的参数设置**

基因组图谱在九宫格内的位置如下图7-7所示：



**图7-7 基因组图谱在九宫格内的位置示意图**

图谱在九宫格内排版后的效果如图7-8所示：



**图7-8 九宫格排版后的效果图**

## 7.6、各模式绘图结果解读

[mainconfiguration]、[IR\_mediated\_reverse\_recomb]、[DR\_mediated\_recomb\_1to2]、[DR\_mediated\_recomb\_2to1]、[DR\_mediated\_recomb\_2to2]、[Arrange\_map]几个模式运行后的结果存储在以{project\_id}为名字的文件夹内：

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

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

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

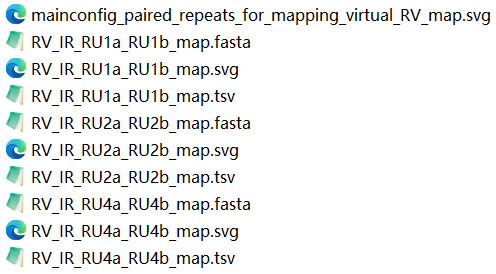
[DR\_mediated\_recomb\_2to1]：DR\_2to1\_{project\_id}

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

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

以[DR\_mediated\_recomb\_1to2]的结果为例，结果有两种：第一种是主要构型中重复序列的基因组图谱，它是根据用户输入的8列表产生，svg格式，可用网页浏览器和Adobe Illustrator CS6等软件打开；第二种是产生的两条染色体（chr1、chr2）的相关信息，每条染色体对应三个文件，分别是fasta格式的序列，svg格式的重复序列的基因组图谱以及基因组图谱对应的8列表（见图7-9）。

其它模式的输出结果类似于[DR\_mediated\_recomb\_1to2]的结果。其中，[DR\_mediated\_recomb\_2to1]模式中需要将三条及以上条序列转换为一条染色体时，该模式的前一次运行结果中的8列表（tsv格式）和fasta格式的文件，可作为该模式下一次运行时的输入文件。从而达到将多条染色体重组为一条染色体的目的。



**图7-9 各模式输出的结果的示意图**

输出的结果包括基因组主构型的图谱（svg格式），次要构型基因组对应的基因组序列（fasta格式），基因组图谱（svg格式）和重复序列配对信息（tsv格式）。