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型基因组分子（图 2-1B）。

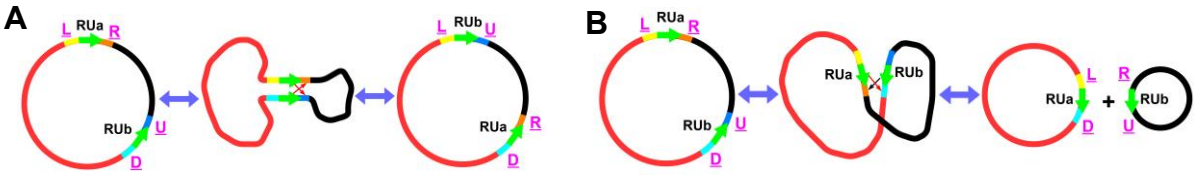

图 2-1 重复序列介导的环状基因组重组示意图

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 TRS_{LR} 与 TRS_{UD}。IR 介导的次要构型中，截取 TRS_{LU} 与 TRS_{RD}，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 TRS_{LD} 与 TRS_{UR}，分别以重复序列 RUa 和 RUb 为中心（图 2-2B）。

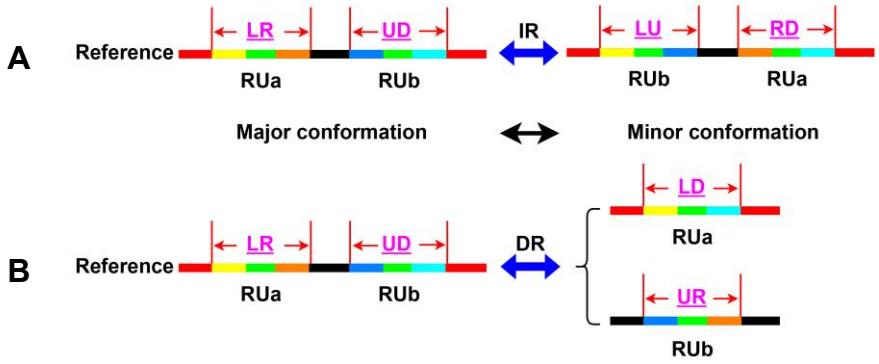

图 2-2 TRS 截取示意图

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（$I > 0$），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。

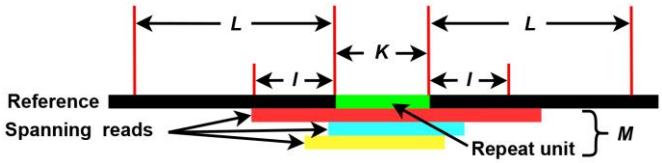

图 2-3 read 映射到 TRS 的示意图

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 $I \geq 1$ 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 $I \geq 1$ 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 TRS_{LR} 和 TRS_{UD} 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 TRS_{LD} 和 TRS_{UR}，以及 TRS_{LU} 和 TRS_{RD} 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：

$$p = \frac{M_{RD}^{-}}{M_{LR}^{+}} \times 100\% \quad (2-1)$$

$$p = \frac{M_{RD}^{+} + M_{RD}^{-}}{M_{LR}^{-} + M_{LR}^{+}} \times 100\% \quad (2-2)$$

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：

$$p = \frac{M_{\mathrm{LU}}^{+}}{M_{\mathrm{UD}}^{-}} \times 100\% \quad (2-3) \qquad p = \frac{M_{\mathrm{LU}}^{+} + M_{\mathrm{LU}}^{-}}{M_{\mathrm{UD}}^{-} + M_{\mathrm{UD}}^{+}} \times 100\% \quad (2-4)$$

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：

$$p = \frac{M_{\mathrm{LD}}^{+}}{M_{\mathrm{LR}}^{+}} \times 100\% \quad (2-5) \qquad p = \frac{M_{\mathrm{LD}}^{+} + M_{\mathrm{LD}}^{-}}{M_{\mathrm{LR}}^{+} + M_{\mathrm{LR}}^{-}} \times 100\% \quad (2-6)$$

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：

$$p = \frac{M_{\mathrm{RU}}^{+}}{M_{\mathrm{UD}}^{+}} \times 100\% \quad (2-7) \qquad p = \frac{M_{\mathrm{RU}}^{+} + M_{\mathrm{RU}}^{-}}{M_{\mathrm{UD}}^{+} + M_{\mathrm{UD}}^{-}} \times 100\% \quad (2-8)$$

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。

```
[general]
;; Parameters for general set.
project_id = RI
mode = A
inputfasta = ReHRI_TestData/plastome_mitogenome.fasta
genome_type =
complementary_chain = Yes
redundant_intermediate_results = D
```

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。

```
[ROUSFinder]
;; Parameters of ROUSFinder for finding repeats.
repeat_length = 32, 507:508
reward = 1
penalty = 20
```

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）。

ReHRI 将 reads 比对到 TRSs 上所用的比对软件有三款，分别是 minimap2，bwa 和 blast。其中，minimap2 是默认选项。三款软件有各自的使用场景，详见表 3-1。

```
[sequencing_depth]
;; Parameters for mapping reads to TRS from mainconfiguration and subconfigurat
alignment_software = minimap2
evalue = 1e-5
NGS_single_end =
NGS_pair_ends =
TGS = ReHRI_TestData/Pacbio.CRR302668.1000.fastq
TGS_type = pacbio
filter_reads = Y
threads =
```

图 3-3 [sequencing_depth]中比对软件和测序数据的参数

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、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

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。

```
[check_spanning_reads]
;; Parameters for checking repeat-spanning reads.
spanning_read_flanking_repeat_length = 1
spanning_read_number = 1
```

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。

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。

```
[manually_calibrated_repeat_info]
;; Parameters for calibrating results.
calibrated_repeat_file = MiRI_TestData/Two_manual_rep.txt
```

图 3-5 [manually_calibrated_repeat_info]中用户提供重复序列信息

```
[general]
;; Parameters for general set.
project_id = AR
mode = C
```

图 3-6 ReHRI 进入模式二的设置参数

fragment_id	length	start	end	direction	chromosome
Repeat_10	16	63176	63161	minus	chr1
Repeat_10	16	66320	66305	minus	chr1
Repeat_2	17	4690	4706	plus	chr1
Repeat_2	17	50595	50611	plus	chr1
Repeat_15	87	332788	332874	plus	chr2
Repeat_15	87	359155	359069	minus	chr2

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。

fragment_id	length	start	end	direction	chromosome	paired_id	paired_length	paired_start	paired_end	paired_direction	paired_chromosome
RU1a	17	4690	4706	plus	chr1	RU1b	17	50595	50611	plus	chr1
RU2a	16	63176	63161	minus	chr1	RU2b	16	66320	66305	minus	chr1
RU3a	87	332788	332874	plus	chr2	RU3b	87	359155	359069	minus	chr2

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。

```
[refilter_params]
;; Parameters for aggregrating the final results manually after the entire proce
refilter_mode = Y
refilter_id = FLT
spanning_read_flanking_repeat_length = 5
spanning_read_number = 5
redundant_intermediate_results = D
```

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。

1	2	3	4	5	6	7	8	9	10
fragment_id	length	start	end	direction	chromosome	plus_ratio(s/m)	minus_ratio(s/m)	combined_ratio	type
RU1a	17	4690	4706	plus	chr1	38/34	38/34	1.117647	direct_repeat
RU2a	16	63176	63161	minus	chr1	60/42	60/42	1.428571	direct_repeat
RU3a	87	332788	332874	plus	chr2	6/6	6/6	1	inverted_repeat
11	12	13	14	15	16	17	18	19	20
paired_id	paired_length	paired_start	paired_end	paired_direction	paired_chromosome	paired_plus_ratio(s/m)	paired_minus_ratio(s/m)	paired_combined_ratio	spanning_read_mcfg
RU1b	17	50595	50611	plus	chr1	50/16	50/16	3.125	sufficient
RU2b	16	66320	66305	minus	chr1	47/28	47/28	1.678571	sufficient
RU3b	87	359155	359069	minus	chr2	6/6	6/6	1	sufficient

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③截取自文件④，可用于 ReHRV 软件绘制重组基因组图谱。文件④为核心结果（解读见第五部分）。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）。

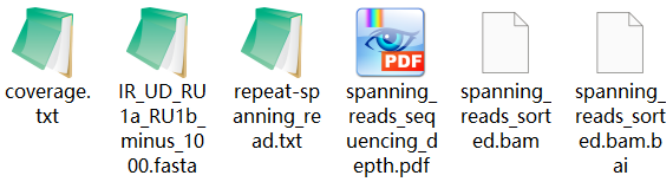

图 6-1 read 映射至每一条 TRS 后的各种结果

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）。

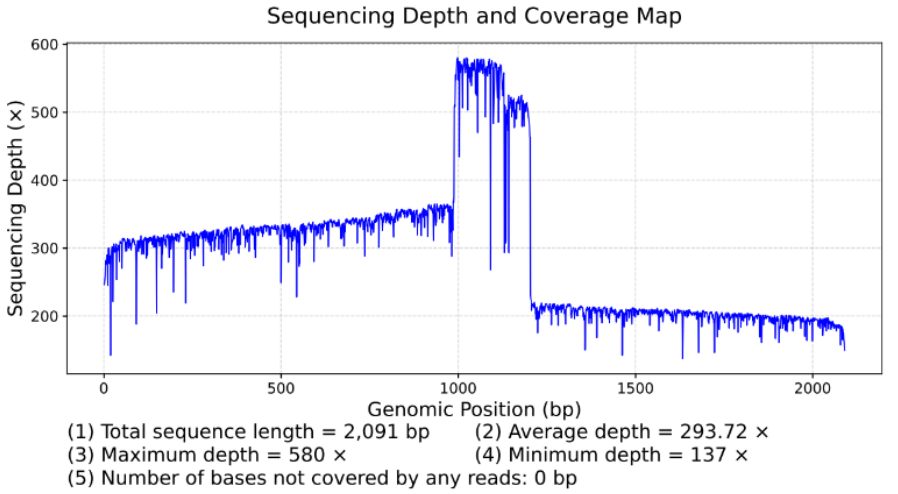

图 6-2 跨越 TRS 中重复序列的 read 映射至 TRS 的测序深度

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。

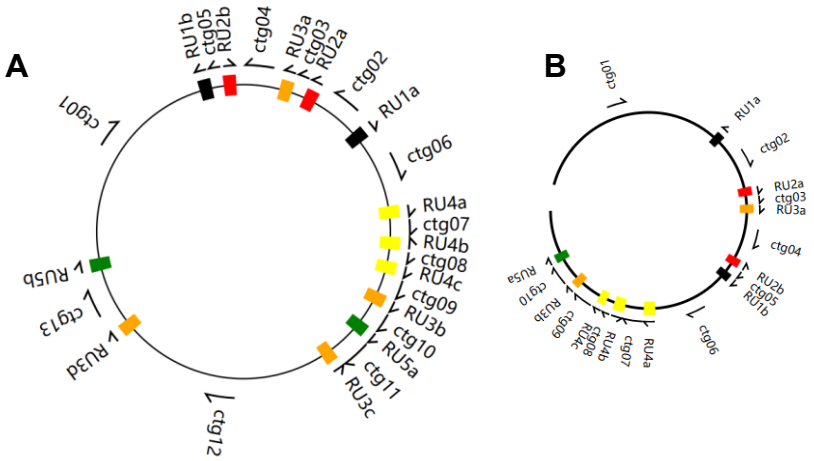

图 7-1 ReHRV 绘制的基因组图谱示意图

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

```
[mainconfiguration]
;;·Parameters·for·mapping·genome·mainconfiguration.
auto_map·=·Y
inputfile·=·ReHRV_TestData/paired_repeats_for_mapping_virtual.tsv
genome_length·=·605764
genome_type·=·C
```

B

```
[IR_mediated_reverse_recomb]
;;·Parameters·for·drawing·maps·of·Inverted·Repeat·(IR)·mediated·genome·recombina
auto_map·=·Y·
inputfile·=·ReHRV_TestData/paired_repeats_for_mapping_virtual.tsv
inputfasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2.fasta·
genome_type·=·C·
```

C

```
[DR_mediated_recomb_1to2]
;;·Parameters·for·drawing·maps·of·organelle·genome·recombination·mediated·by·dir
auto_map·=·y·
inputfile·=·ReHRV_TestData/paired_repeats_for_mapping_virtual.tsv
inputfasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2.fasta·
genome_type·=·C·
```

D

```
[DR_mediated_recomb_2to1]
;;·Parameters·for·drawing·maps·of·organelle·genome·recombination·mediated·by·dir
auto_map·=·y
flip_chain·=·Y
chr1_file·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2_map.tsv
chr1_fasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2.fasta·
chr1_type·=·C
chr2_file·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr2_1to2_map.tsv
chr2_fasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr2_1to2.fasta
chr2_type·=·C
```

E

```
[DR_mediated_recomb_2to2]
;;·Parameters·for·drawing·maps·of·organelle·genome·recombination·mediated·by·dir
auto_map·=·Y·
flip_chain·=·Y
chr1_file·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2_map.tsv
chr1_fasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr1_1to2.fasta·
chr2_file·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr2_1to2_map.tsv
chr2_fasta·=·ReHRV_TestData/ReHRV_DR_RU5a_RU5b_chr2_1to2.fasta
```

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。

```
fragment_id	start	end	direction	paired_id	paired_start	paired_end	paired_direction
RU3c	386117	385989	minus
RU1a	153164	152647	minus	RU1b	231395.0	231912.0	plus
RU2a	193682	193809	plus	RU2b	231329.0	231202.0	minus
RU3a	196488	196753	plus	RU3d	556153.0	556418.0	plus
RU3d	556153	556418	plus	RU3b	343152.0	343280.0	plus
RU4a	317468	310009	minus	RU4b	317468.0	310009.0	minus
RU4a	317468	310009	minus	RU4c	322009.0	327009.0	plus
RU5a	359772	363702	plus	RU5b	601834.0	605764.0	plus
```

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：

```
[mapper_config]
;; Parameters for the properties of various elements in the genome map.
picture_box = 450
radius = 150
arrow_radius = 170
arrow_size = 10
arrow_thickness = 2
font_size = 18
tag_height = 20
tag_line_width = 1
output_svg_file = mainconfig
```

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：

```
[color_library]
;; Colouring scheme for different repeats.
RU1 = black
RU2 = red
RU3 = orange
RU4 = yellow
RU5 = green
RU6 = cyan
RU7 = blue
RU8 = purple
RU9 = gray
RU10 = slategray
RU11 = forestgreen
RU12 = gold
RU13 = indigo
RU14 = darkgreen
RU15 = darkred
RU16 = firebrick
RU17 = limegreen
RU18 = chocolate
RU19 = olive
RU20 = yellowgreen
RU21 = turquoise
RU22 = plum
RU23 = greenyellow
RU24 = darkslategray
RU25 = aqua
RU26 = navy
RU27 = crimson
RU28 = skyblue
RU29 = coral
RU30 = brown
```

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
```
[Arrange_map]
;;·Arrange·images·into·a·grid·of·nine·squares.
;·General·set.
arrange·=·Y
font_size·=·20
image_dpi·=·600
```

B
```
;·Remarks·info·in·the·circle·map.
center_font·=·Main
left_middle_font·=·RU1
right_middle_font·=·RU2
top_middle_font·=·RU3
bottom_middle_font·=·RU4
top_left_font·=·RU5
top_right_font·=·RU6
bottom_left_font·=·RU7
bottom_right_font·=·RU8
;·*NOTE*:·Set·the·text·content·as·desired,·the·content·can·be·left·blank.
```

C
```
;·Path·to·up·to·nine·maps.
center_path·=·ReHRV_TestData/mainconfig_ReHRV_map.svg
left_middle_path·=·ReHRV_TestData/ReHRV_DR_RU3a_RU3d_chr1_1to2_map.svg
top_left_path··=·ReHRV_TestData/ReHRV_DR_RU3a_RU3d_chr2_1to2_map.svg
top_middle_path·=·ReHRV_TestData/ReHRV_DR_RU3a_RU3d_chr1_1to2_map.svg
top_right_path·=·ReHRV_TestData/ReHRV_DR_RU3a_RU3d_chr2_1to2_map.svg
right_middle_path·=·ReHRV_TestData/ReHRV_IR_RU2a_RU2b_map.svg
bottom_right_path·=·ReHRV_TestData/ReHRV_IR_RU4a_RU4b_map.svg
bottom_middle_path·=·ReHRV_TestData/ReHRV_DR_RU3a_RU3d_chr1_1to2_map.svg
bottom_left_path·=·ReHRV_TestData/ReHRV_IR_RU1a_RU1b_map.svg·
;·*NOTE*:·Only·accept·images·in·svg·format.·Set·the·path·as·desired,·or·leave·it
```

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：

```
+--------------+--------------+--------------+
|   Top Left   |  Top Middle  |  Top Right   |
+--------------+--------------+--------------+
| Left Middle  |    Center    | Right Middle |
+--------------+--------------+--------------+
| Bottom Left  | Bottom Middle| Bottom Right |
+--------------+--------------+--------------+
```

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：

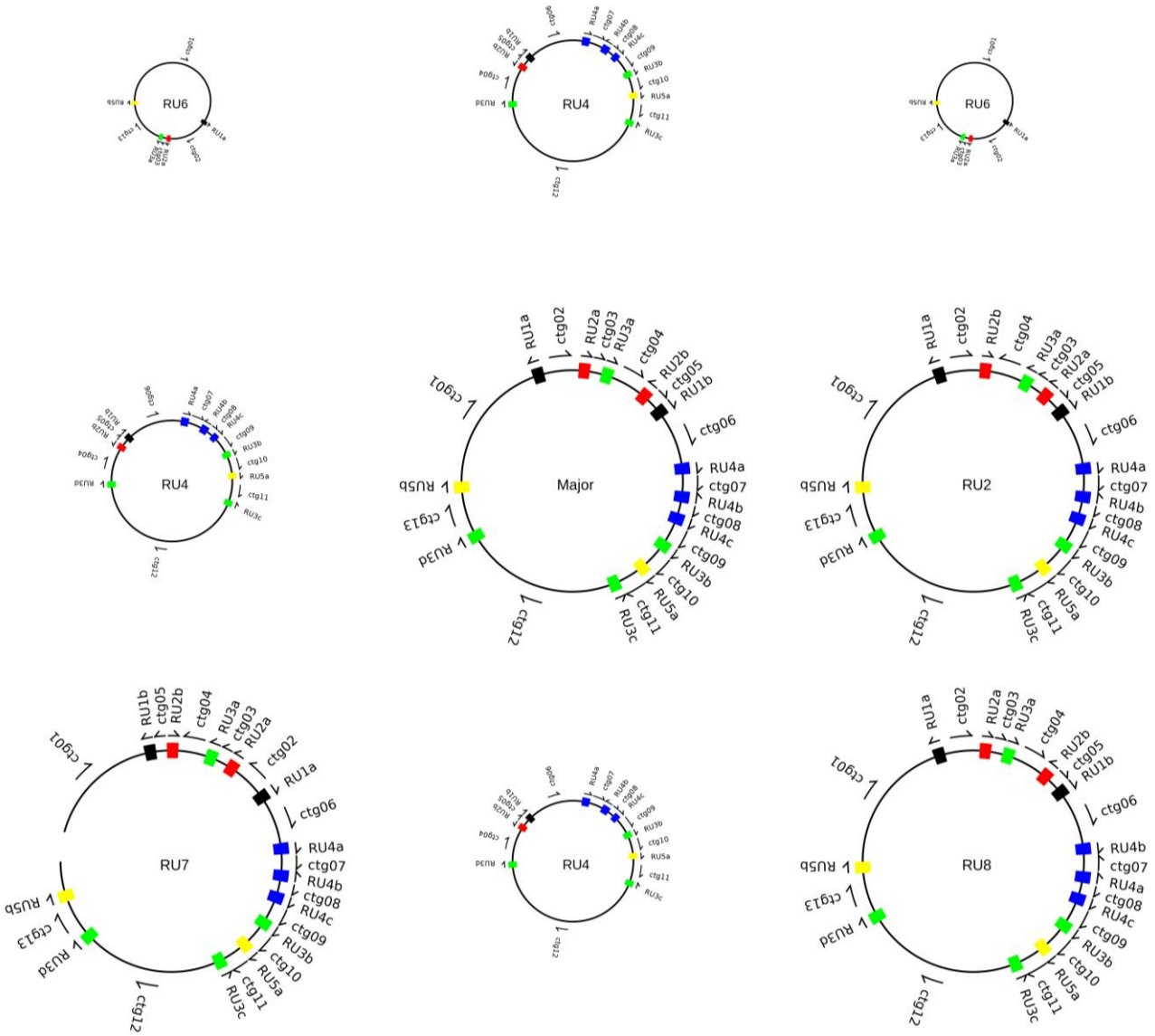

图 7-8 九宫格排版后的效果图

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。

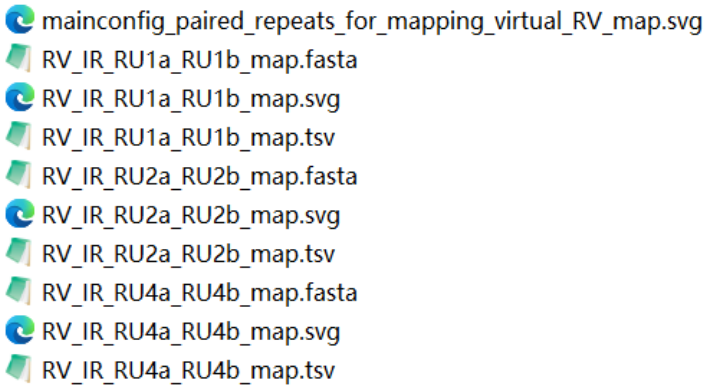

图 7-9 各模式输出的结果的示意图