Introduction

CATG (Collinearity-based Assembly correcTor GUI) is a GUI application base on Qt with PySide6. It is a tool that can adjust assembly with collinearity and generate tour files for assembly.

Installation

Download pre-build binary files

User can download executable file with following links.

1. Windows user

- https://github.com/sc-zhang/CATG/releases/download/v1.2.2/CATG-v1.2.2.exe
- https://zenodo.org/records/13621059/files/CATG-v1.2.2.exe?download=1

2. Mac user (Apple silicon)

- https://github.com/sc-zhang/CATG/releases/download/v1.2.2/CATG-v1.2.2.arm.dmg
- https://zenodo.org/records/13621059/files/CATG-v1.2.2.arm.dmg?download=1

3. Mac user (Intel silicon)

- https://github.com/sc-zhang/CATG/releases/download/v1.2.2/CATG-v1.2.2.Intel.dmg
- https://zenodo.org/records/13621059/files/CATG-v1.2.2.Intel.dmg?download=1

4. Ubuntu user

- https://github.com/sc-zhang/CATG/releases/download/v1.2.2/CATG-v1.2.2.bin
- https://zenodo.org/records/13621059/files/CATG-v1.2.2.bin?download=1

Data preparation

The files required with this tool are two bed files, one anchors file and one agp file, details of them are:

1. The JCVI package is required to generate ref.bed and qry.bed file from gff3 files with command below,

```
python -m jcvi.formats.gff bed --type=mRNA --key=Name ref.gff3.gz -o ref.bed
```

Notice: Both gff files should be chromosome or pesudo-chromosome level, and if the annotation of the new assembly (query one) is missing, a simple way to use this tool is using gmap to mapping reference CDS to the query genome (chromosome level).

2. The anchors file is also generated by JCVI package, with 4 files that qry.bed, qry.cds, ref.bed and ref.cds, then run command below,

```
python -m jcvi.compara.catalog ortholog qry ref
```

More details of JCVI could be found in JCVI.

3. The AGP file is a text file which can be found with many assembly tools like ALLHiC, which recorded the contig positions and orients on chromosomes. More details could be found in AGP file format
4. The qry.bed, ref.bed, qry.agp, qry.ref.anchors files are all we need.

Usage

Get test data

1. Test data could be found in following links.

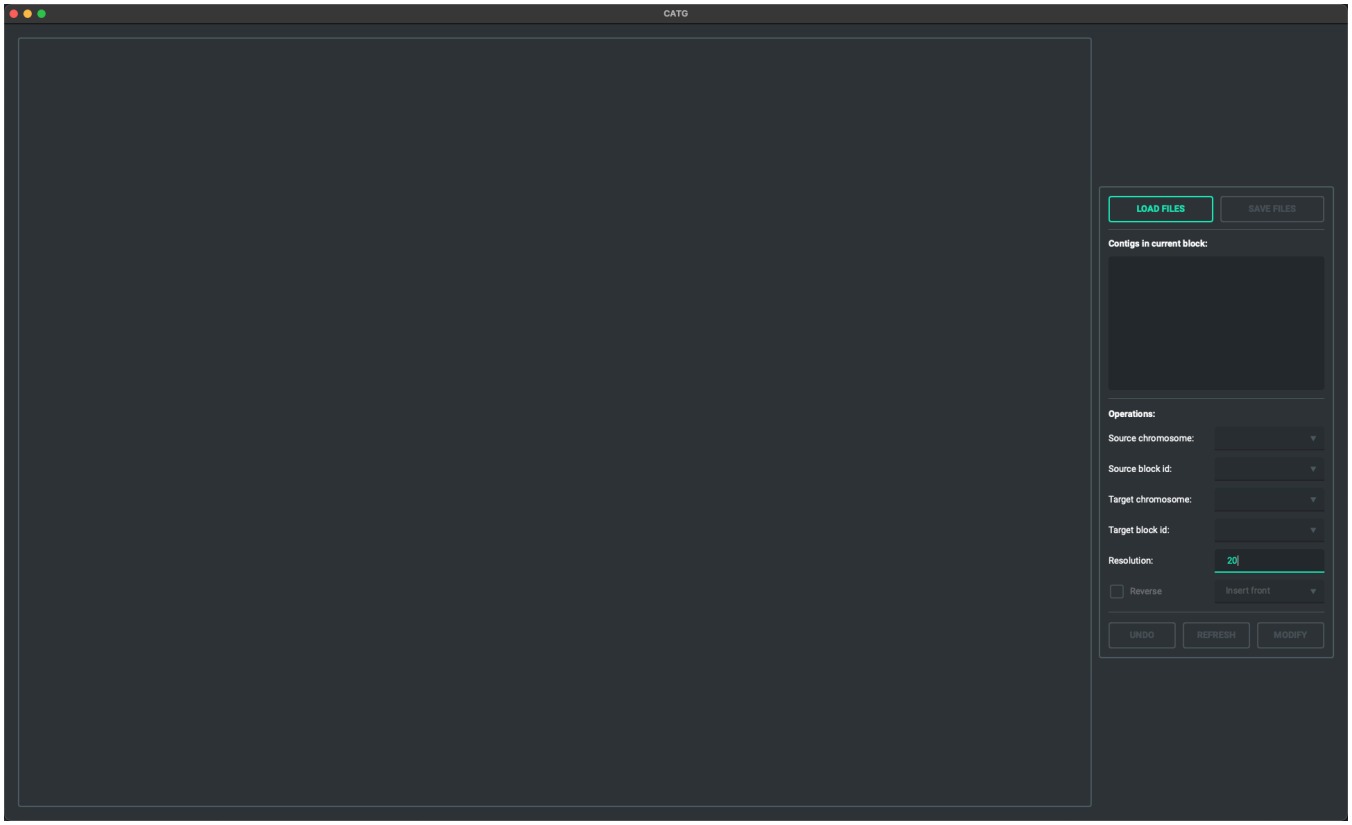
- https://github.com/sc-zhang/CATG/archive/refs/tags/v1.2.2.zip
- https://zenodo.org/records/13621059/files/sc-zhang/CATG-

v1.2.2.zip?download=1

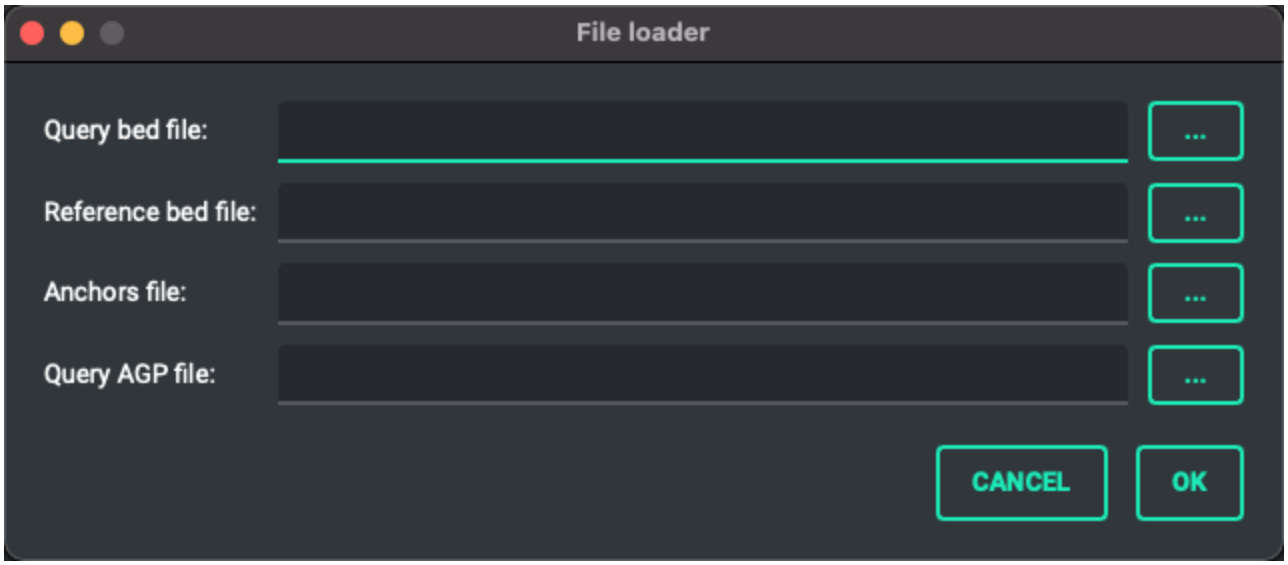
2. Unzip this compressed file, a file named "test.tar.gz" could be found in test folder, then unzip test.tar.gz, four files: qry.agp, qry.bed, qry.ref.anchors, ref.bed could be found which can be used with this tool.

Load files

1. Click `LOAD FILES` button on the main form



2. The "File loader" window is shown below, user can select files by clicking `...` button or just drag file into text box.

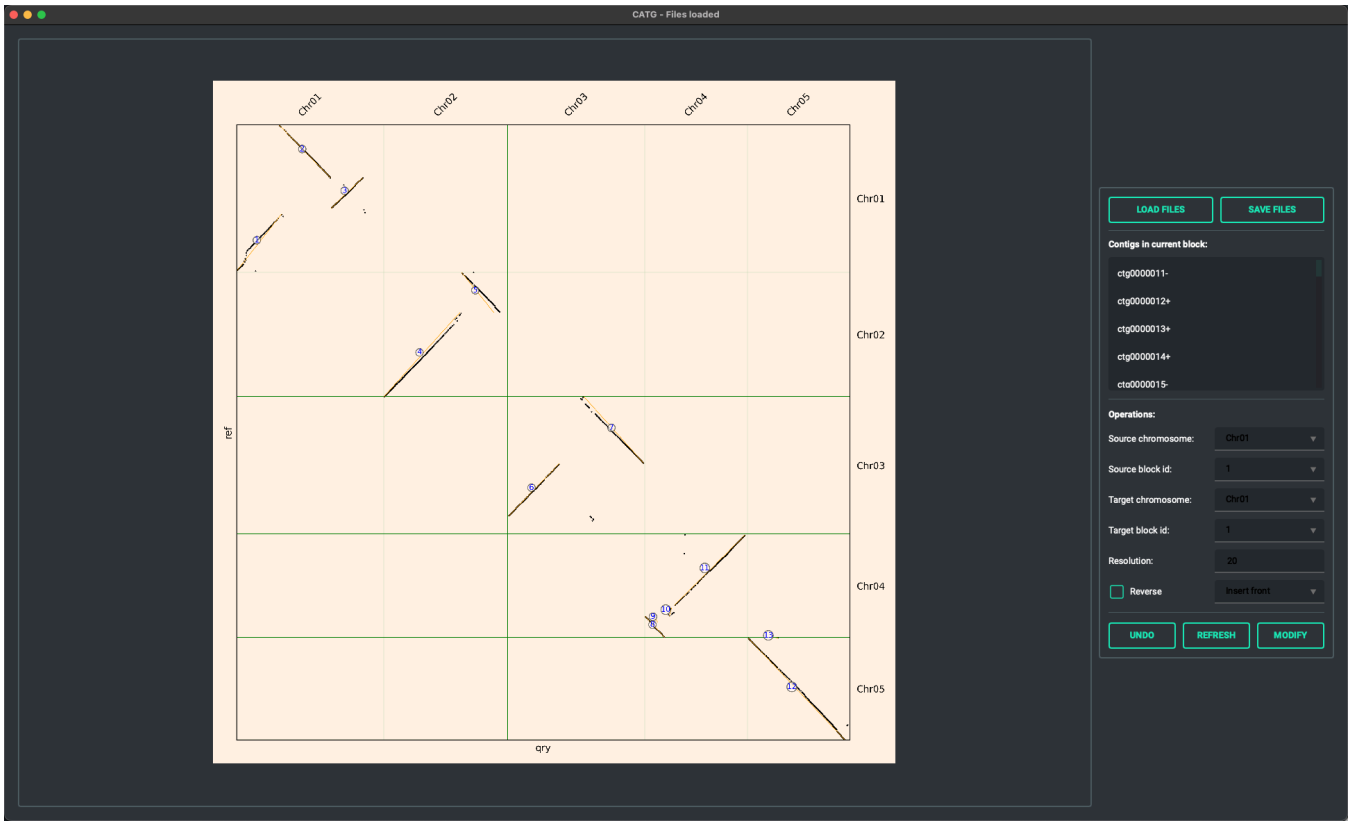


Notice:

The generation of required four files could be found in "Data preparation" section.

- "Query bed file" should be the bed file for new assembly, for test data is "qry.bed"
- "Reference bed file" should be the bed file for reference genome, for test data is "ref.bed"
- "Anchors file" should be the anchors file, for test data is "qry.ref.anchors"
- "Query AGP file" should be the AGP file for new assembly, for test data is "qry.agp"

3. After click `OK` button, the files would be loaded, and an collinearity figure while contigs cluster and marked would be shown.



Opeartions

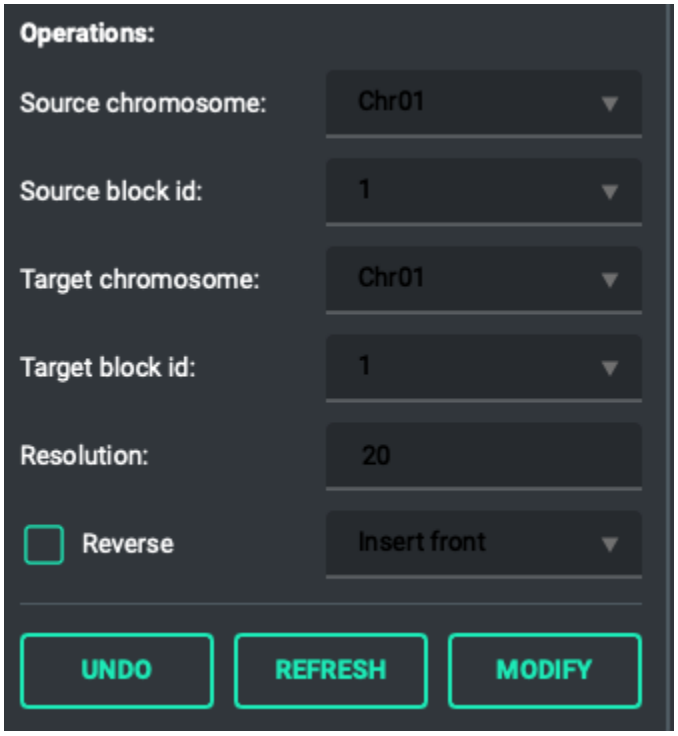
There are 9 opeartions can be selected, which can be used to correct assembled genome manually.

1. **Insert front**
   Move block with source block id from source chromosome to target chromosome and insert it in front of target block.
2. **Insert back**
   Move block with source block id from source chromosome to target chromosome and insert it after target block.
3. **Insert head**
   Move block with source block id from source chromosome to target chromosome and insert it to the head of target chromosome.
4. **Insert tail**
   Move block with source block id from source chromosome to target chromosome and insert it to the tail of target chromosome.
   Operate 1-4 can work with Reverse checkbox, if Reverse checkbox is set

checked, the block from source chromosome will be reverse complement before insert to target positiong.

5. **Source chromosome**
6. **Source block**
   These two operate only affect while Reverse checkbox is set checked, then it will reverse the source chromosome or source block.
7. **Swap chromosome**
8. **Swap block**
   These two operate can swap regions or chromosomes, and Reverse option won't affect.
9. **Delete block**
   Delete block from source chromosome.

After any opeartion, the contigs would be re-cluster and the blocks would remark automatically.

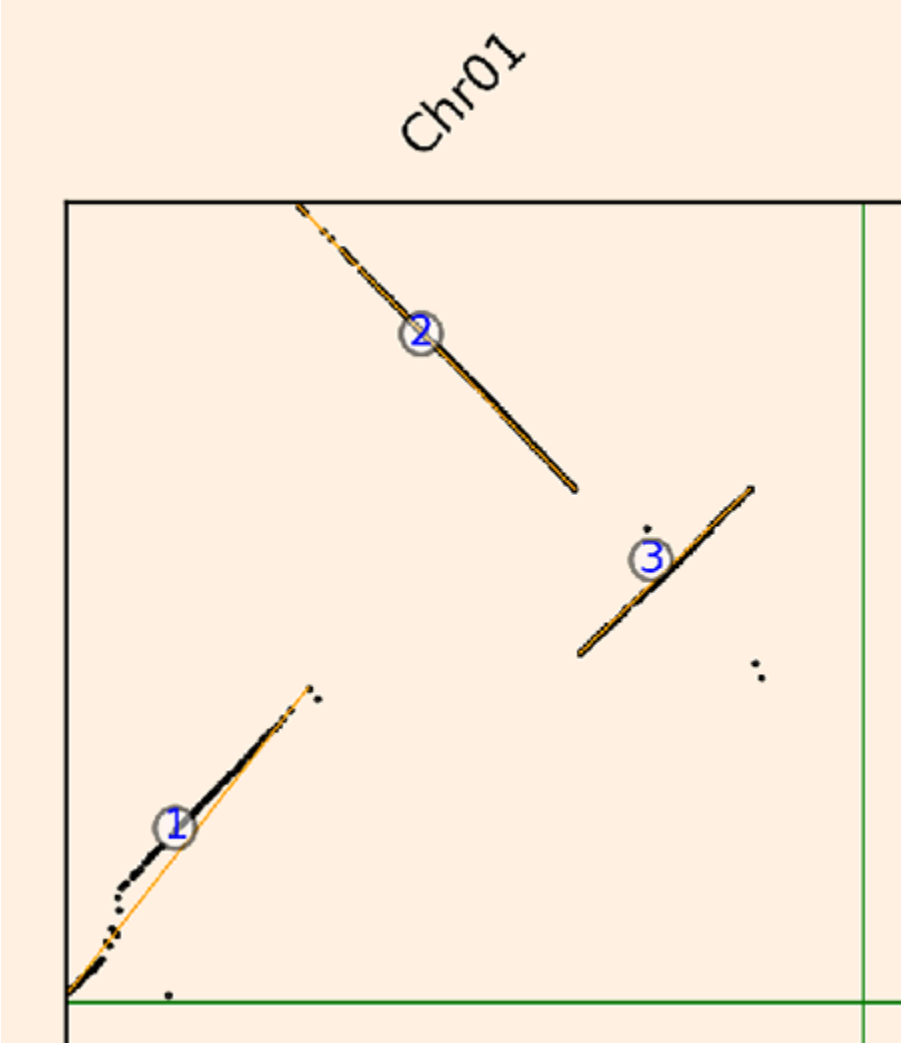


1. Insert front

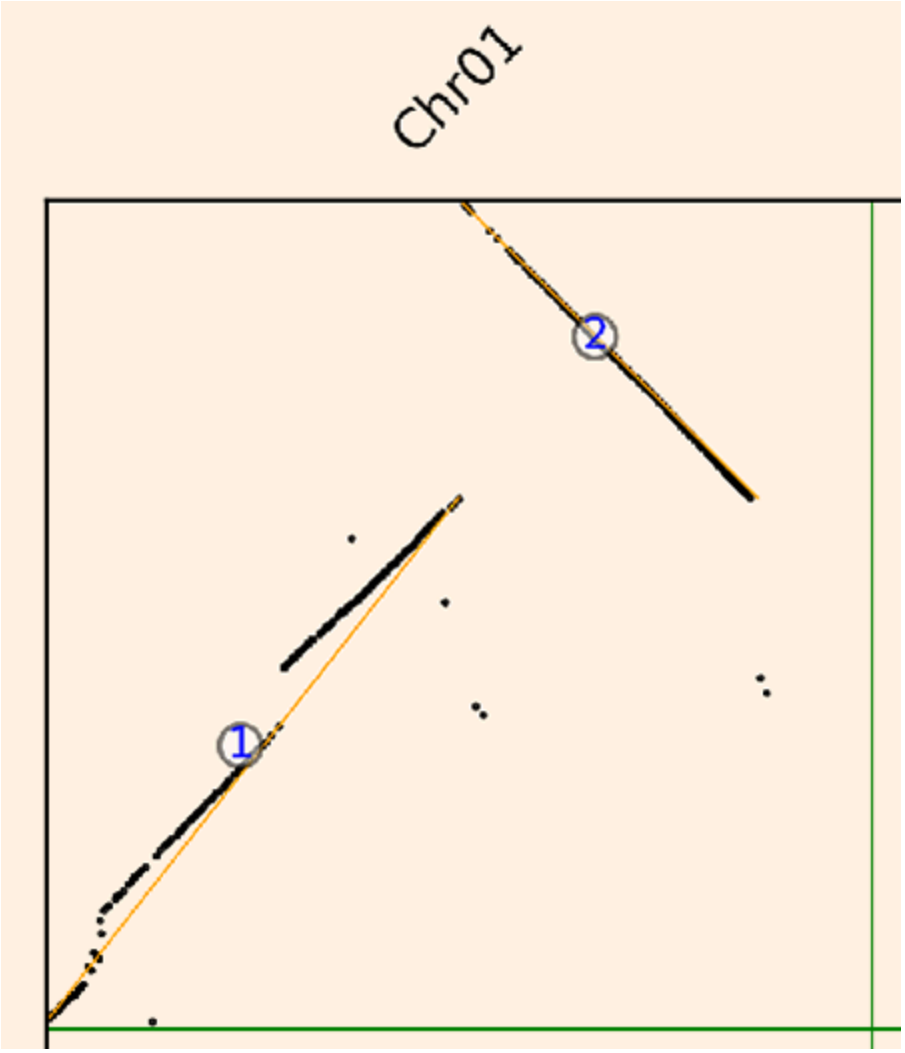
This operation can move one block to the front of another block.
Set "Source chromosome", "Source block id", "Target chromosome" and "Target block id", then select "Insert front" option, click `MODIFY` button.

There is an example that move block 3 in chromosome 1 to the front of block 2 in chromosome 1:

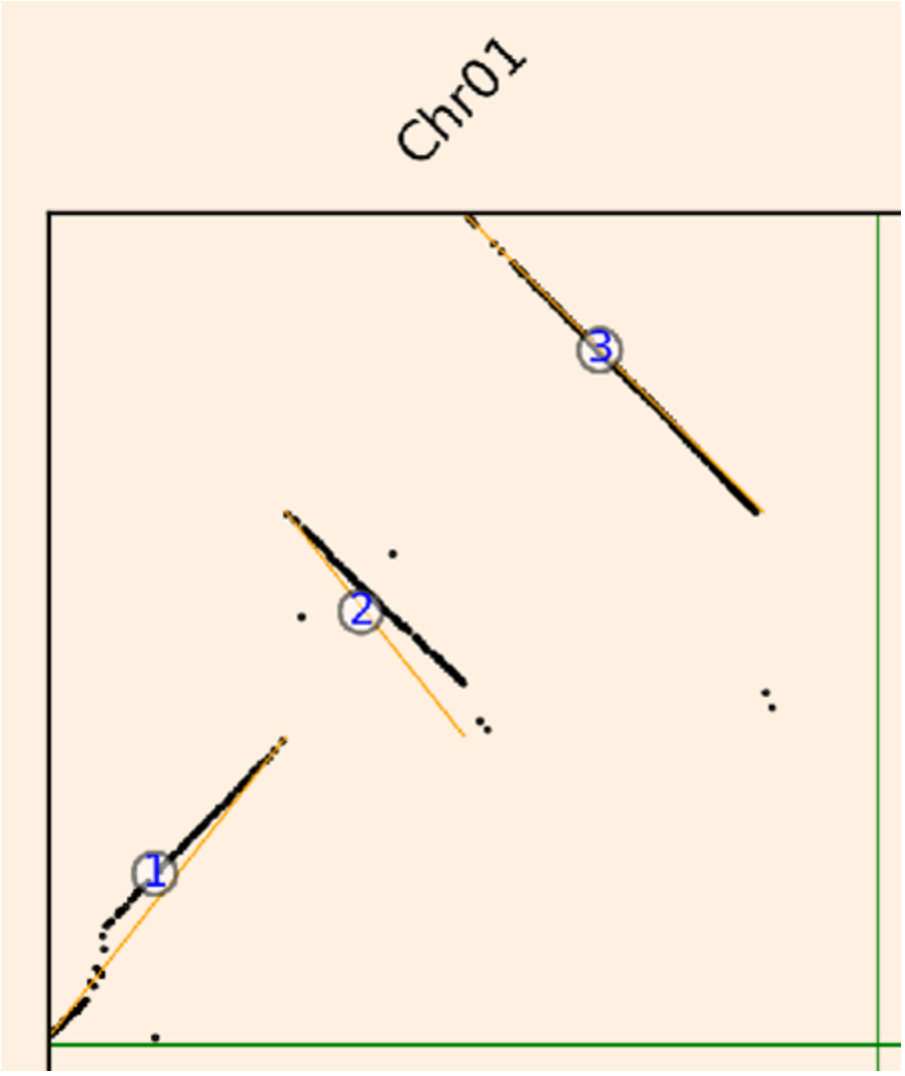
- Before



- After



- `Reverse` checked



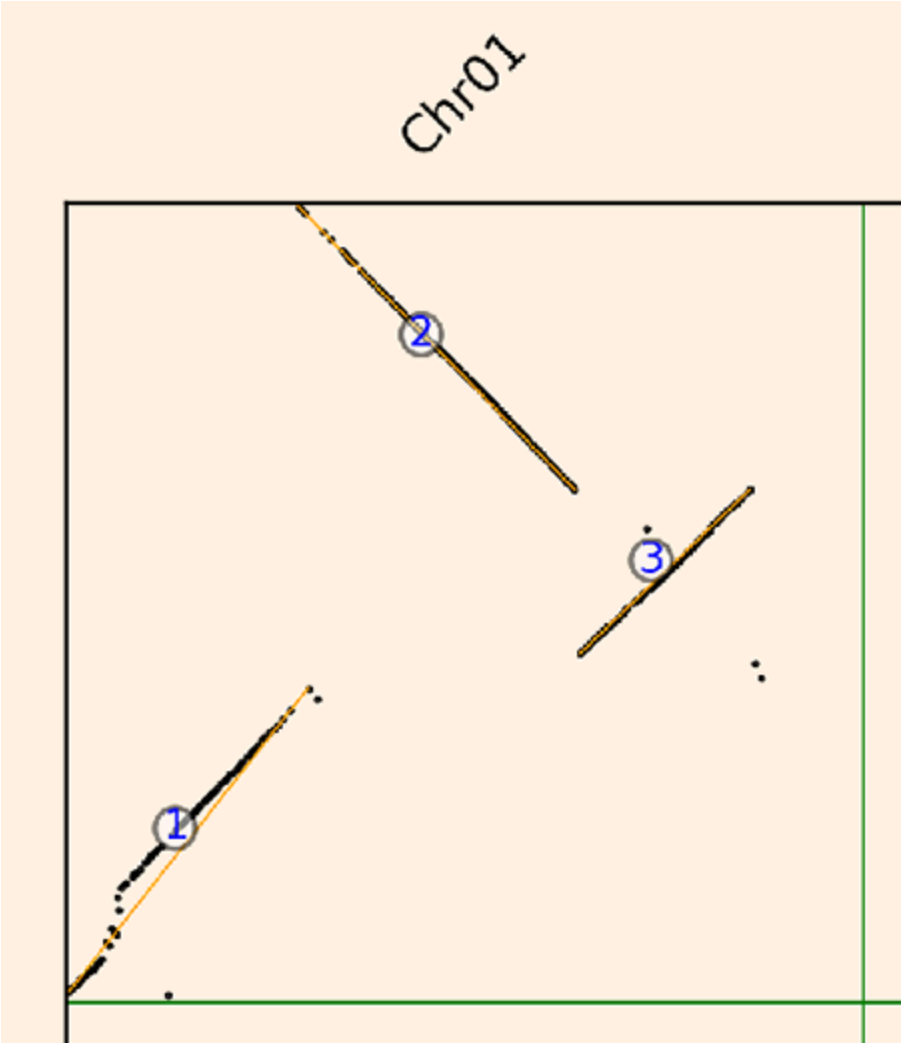
2. Insert back

This operation can move one block to the back of another block.

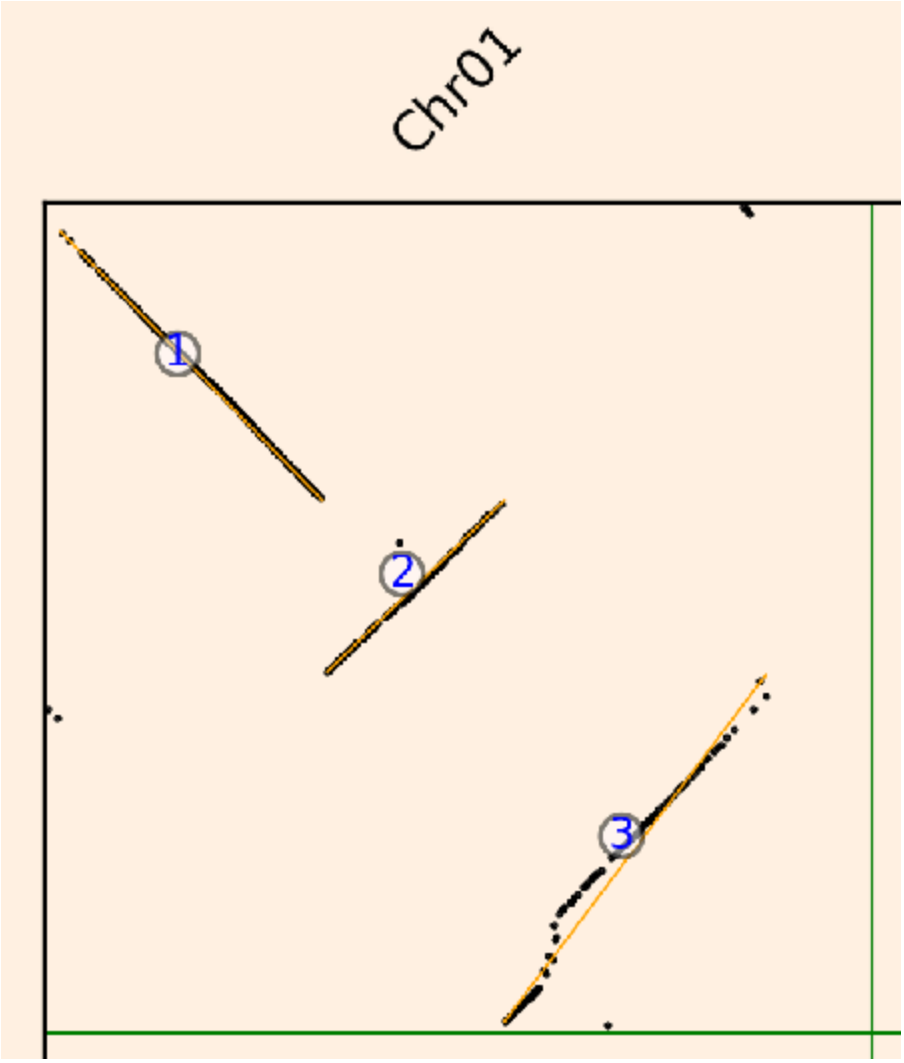
Set "Source chromosome", "Source block id", "Target chromosome" and "Target block id", then select "Insert back" option, click `MODIFY` button.

There is an example that move block 1 in chromosome 1 to the back of block 3 in chromosome 1:

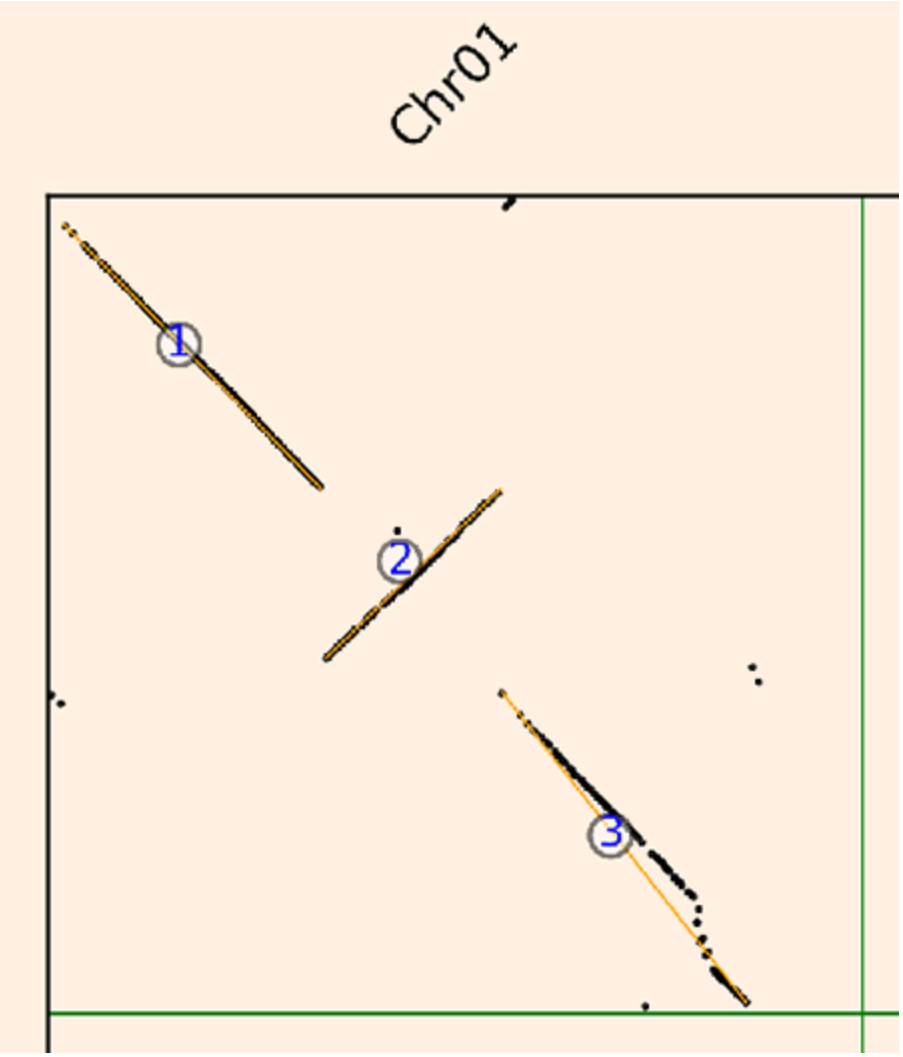
- Before



- After



- `Reverse` checked



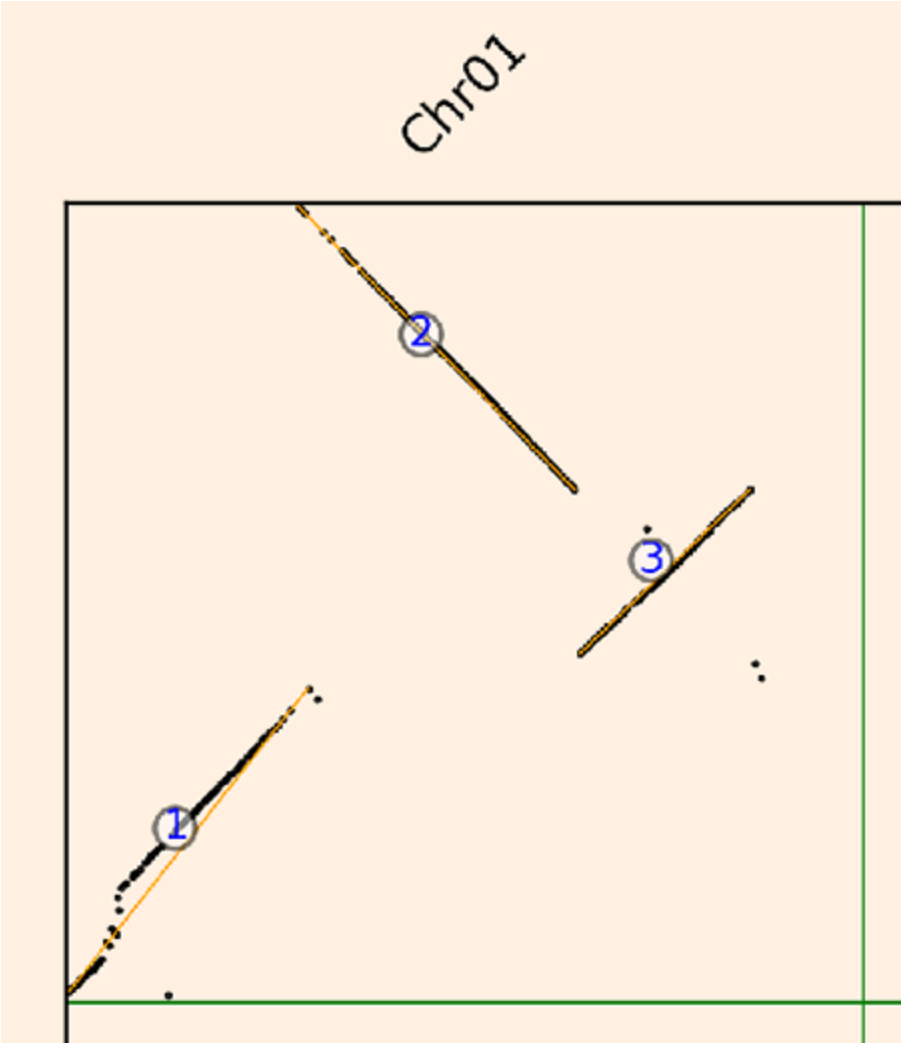
3. Insert head

This operation can move one block to the head of chromosome

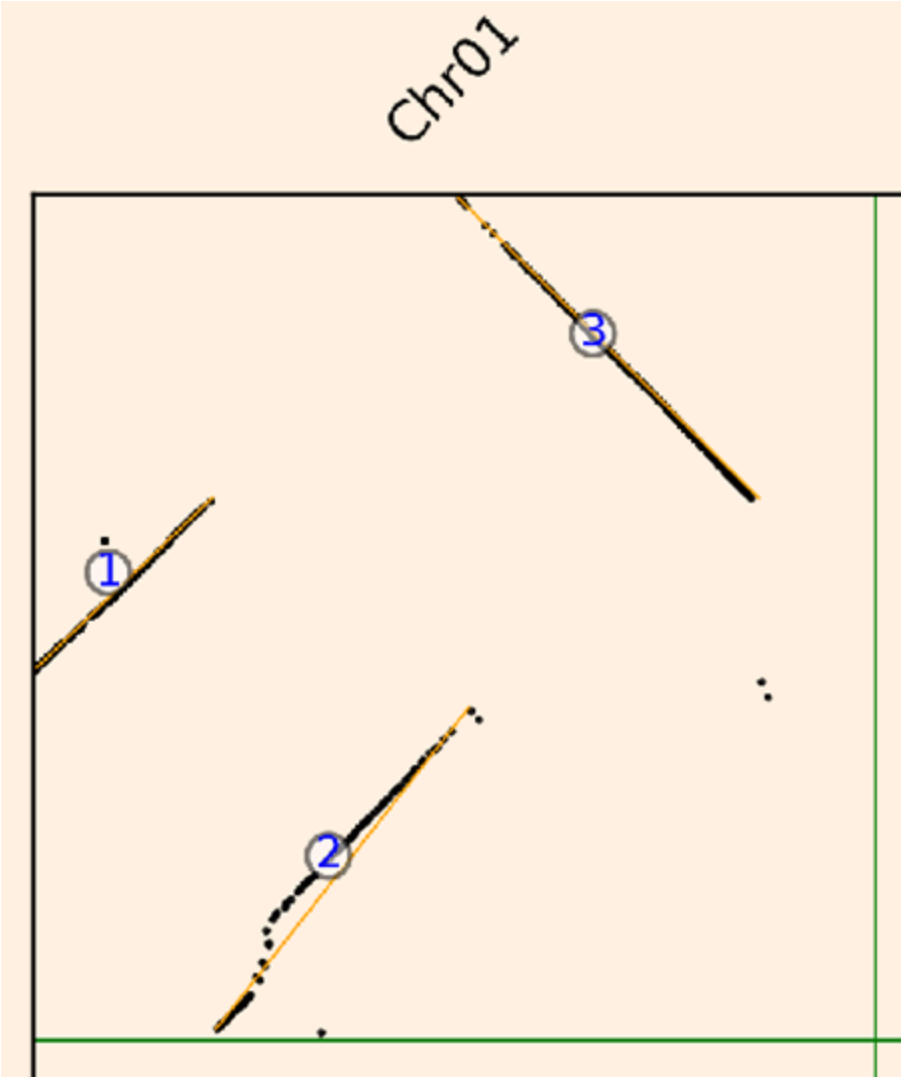
Set "Source chromosome", "Source block id", "Target chromosome" and "Target block id", then select "Insert head" option, click `MODIFY` button.

There is an example that move block 3 in chromosome 1 to head of chromosome 1:

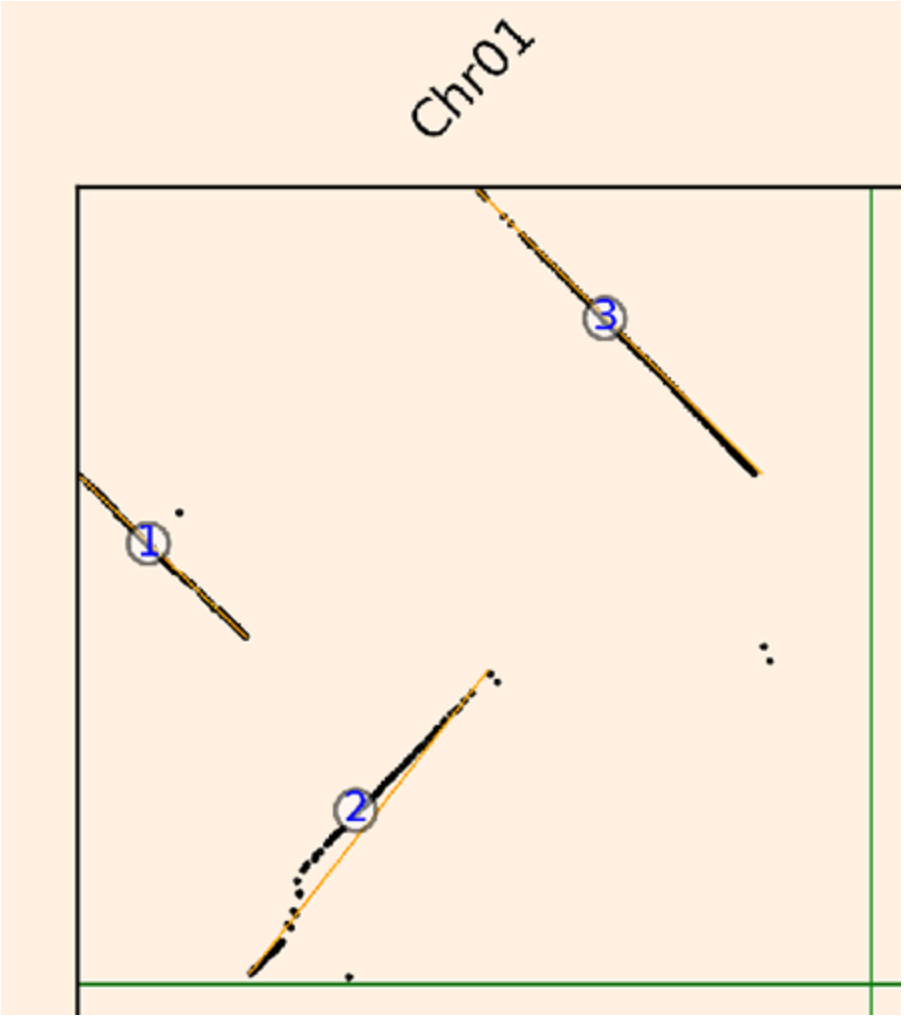
- Before



- After



- `Reverse` checked



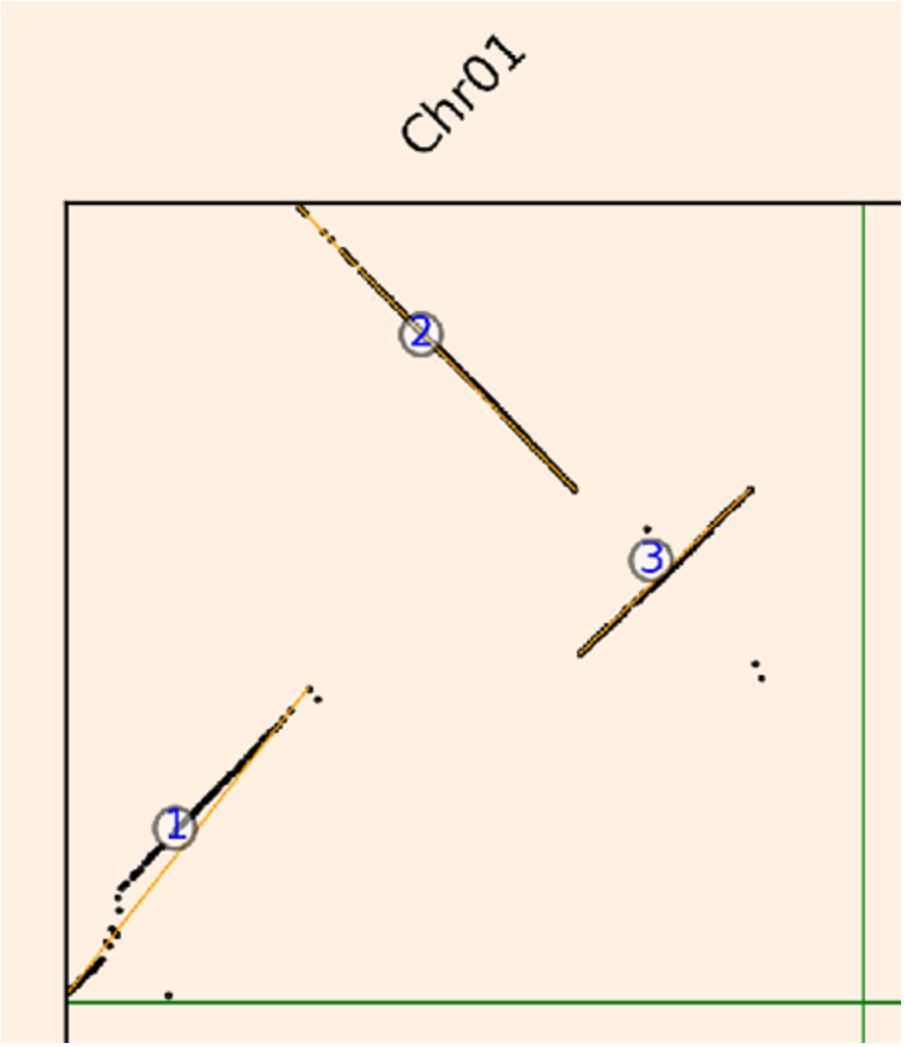
4. Insert tail

This operation can move one block to the tail of chromosome

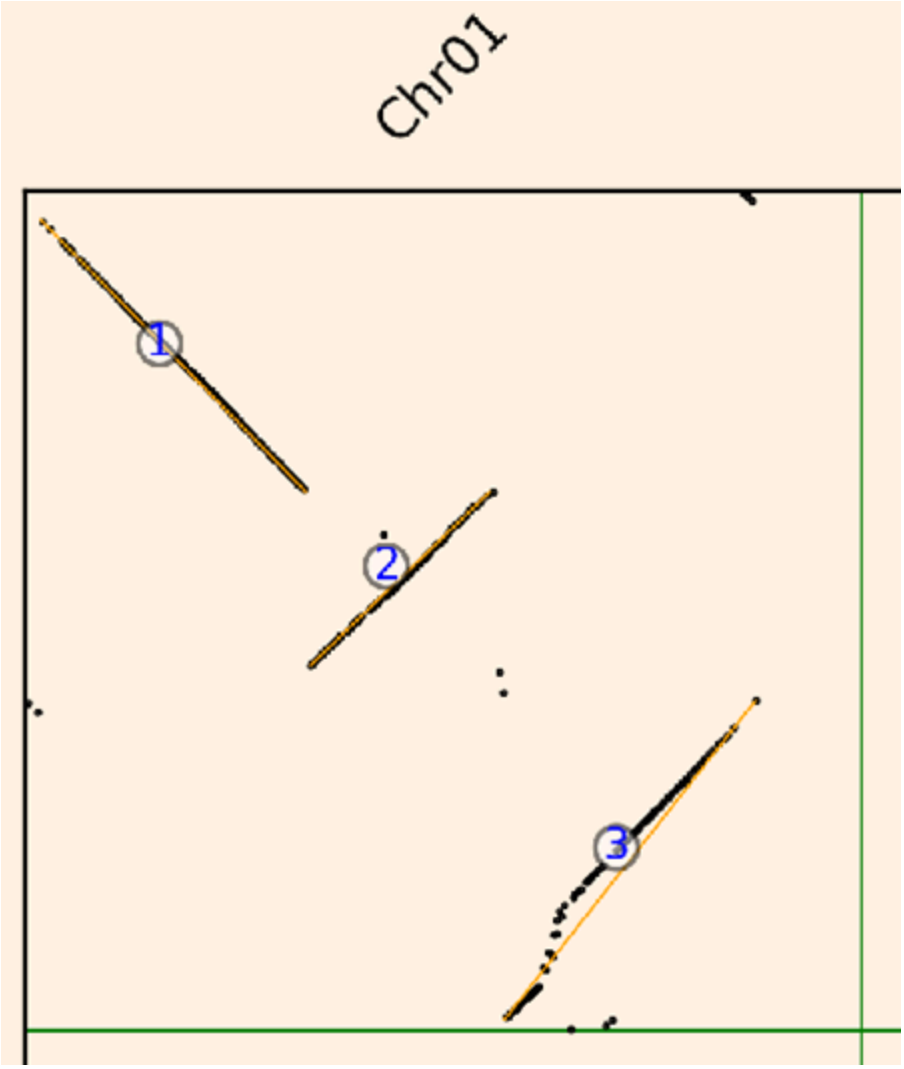
Set "Source chromosome", "Source block id", "Target chromosome" and "Target block id", then select "Insert tail" option, click `MODIFY` button.

There is an example that move block 1 in chromosome 1 to tail of chromosome 1:

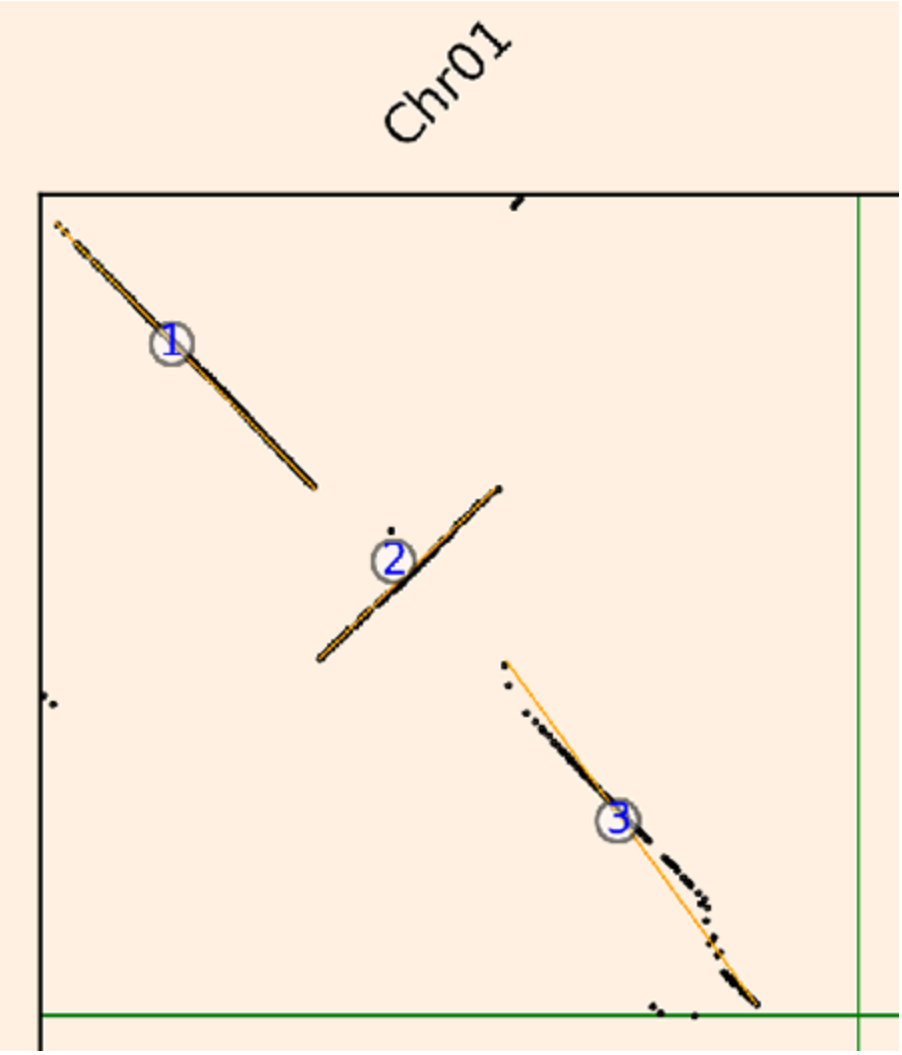
- Before



- After



- `Reverse` checked



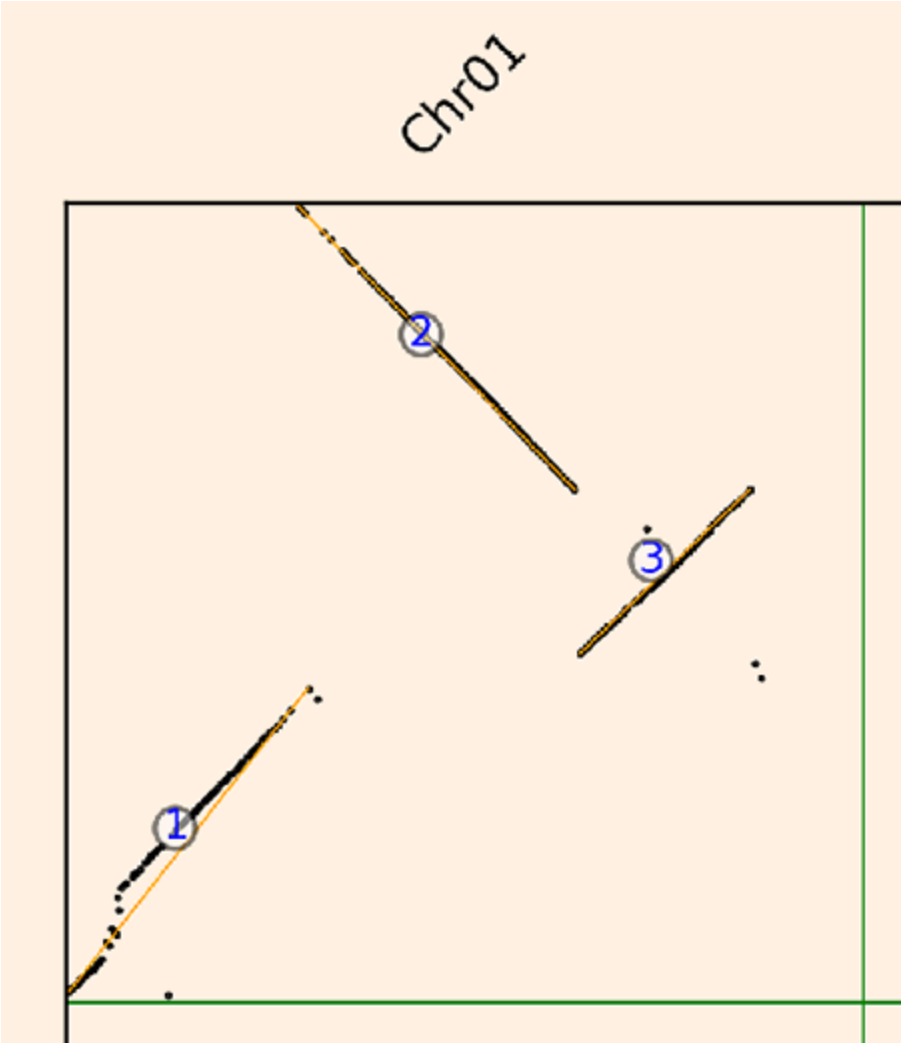
5. Reverse chromosome

This operation reverse one chromosome

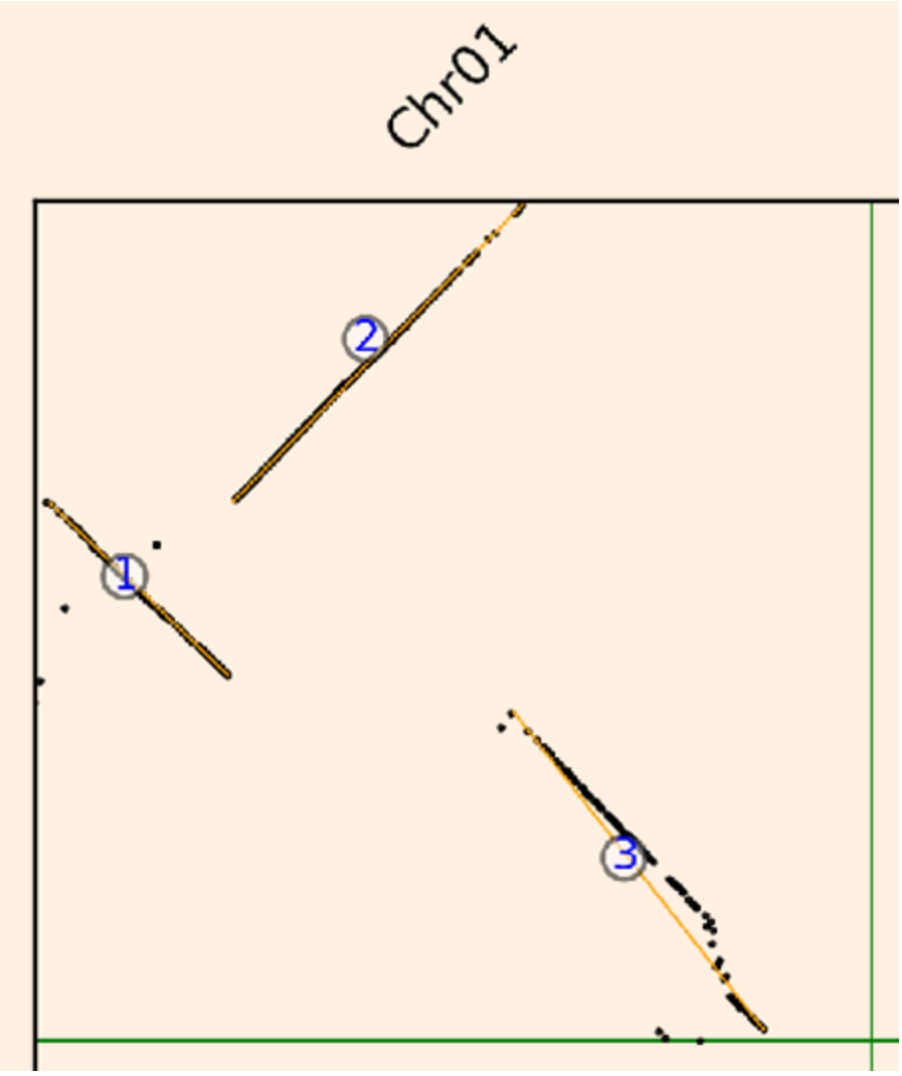
Set "Source chromosome", then select "Source chromosome" option, checked "Reverse" checkbox, click `MODIFY` button.

There is an example that reverse the chromosome 1:

- Before



- After



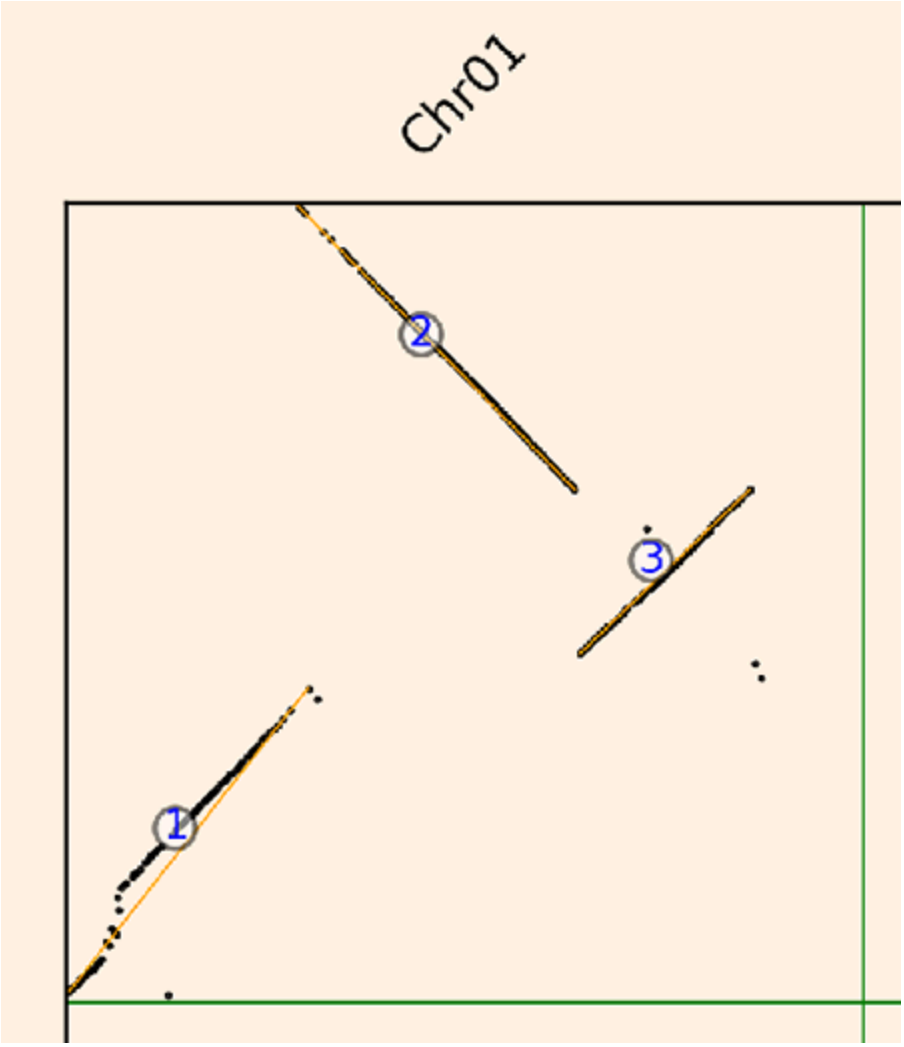
6. Reverse block

This operation reverse one chromosome

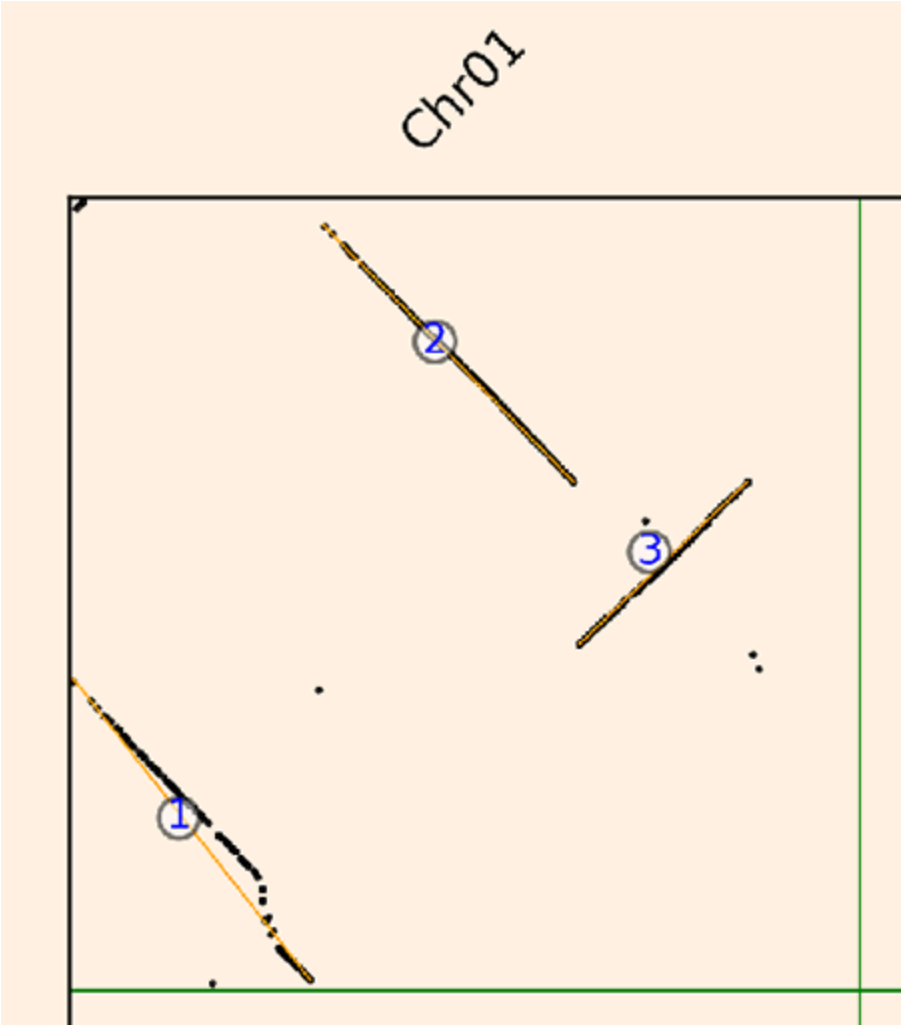
Set "Source chromosome" and "Source block id", then select "Source block" option, checked "Reverse" checkbox, click `MODIFY` button.

There is an example that reverse the block 1 in chromosome 1:

- Before



- After



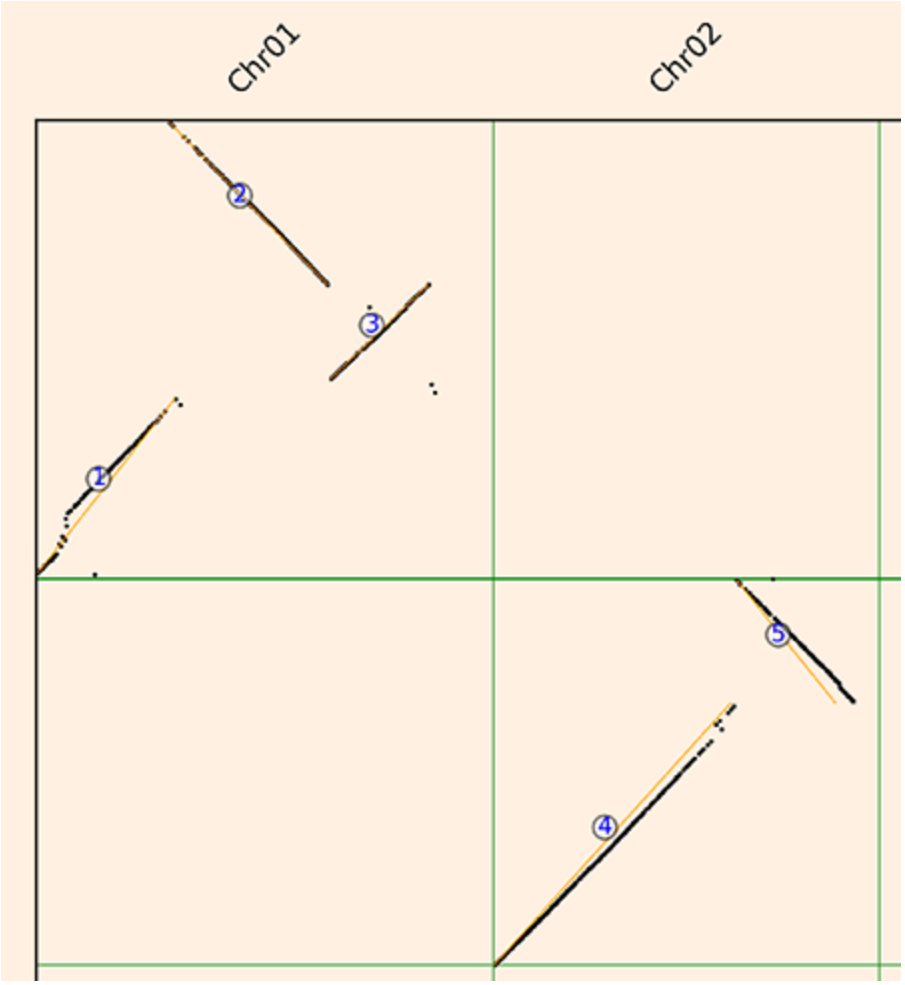
7. Swap chromosomes

This operation swap two chromosomse

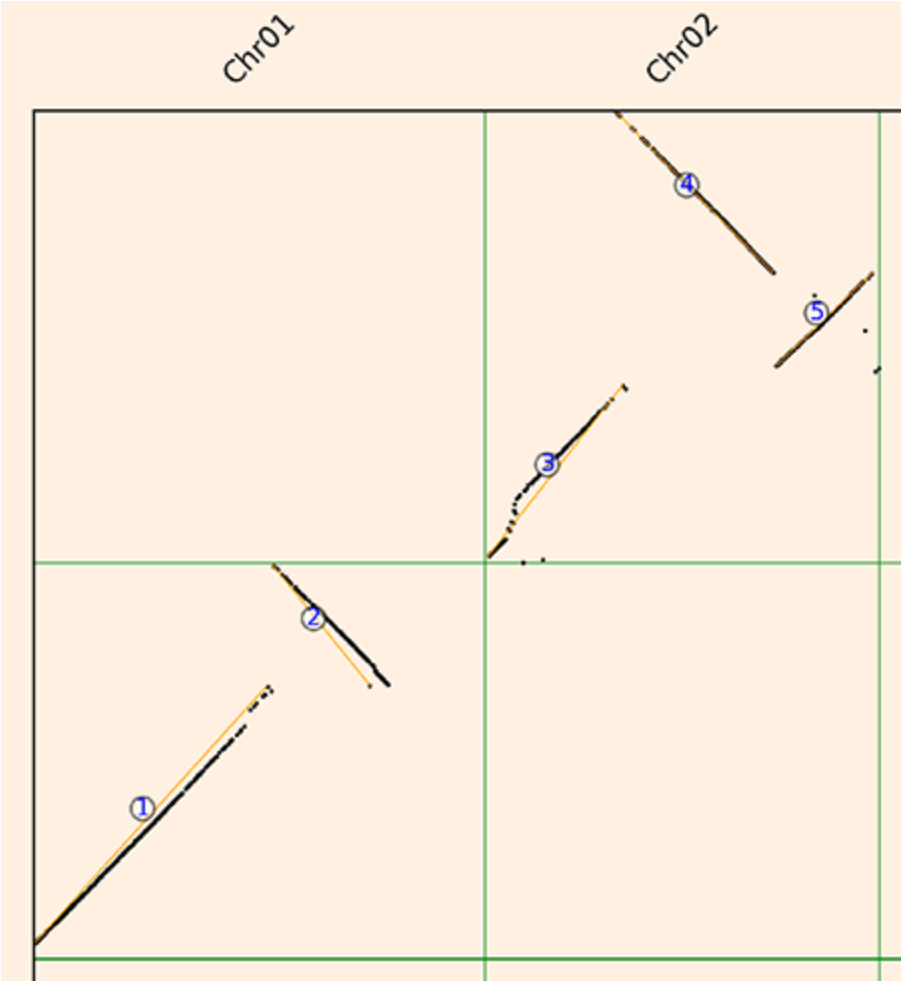
Set "Source chromosome" and "Target chromosome", then select "Swap chromosome" option, click `MODIFY` button.

There is an example that swap chromosome 1 and chromosome 2:

- Before



- After



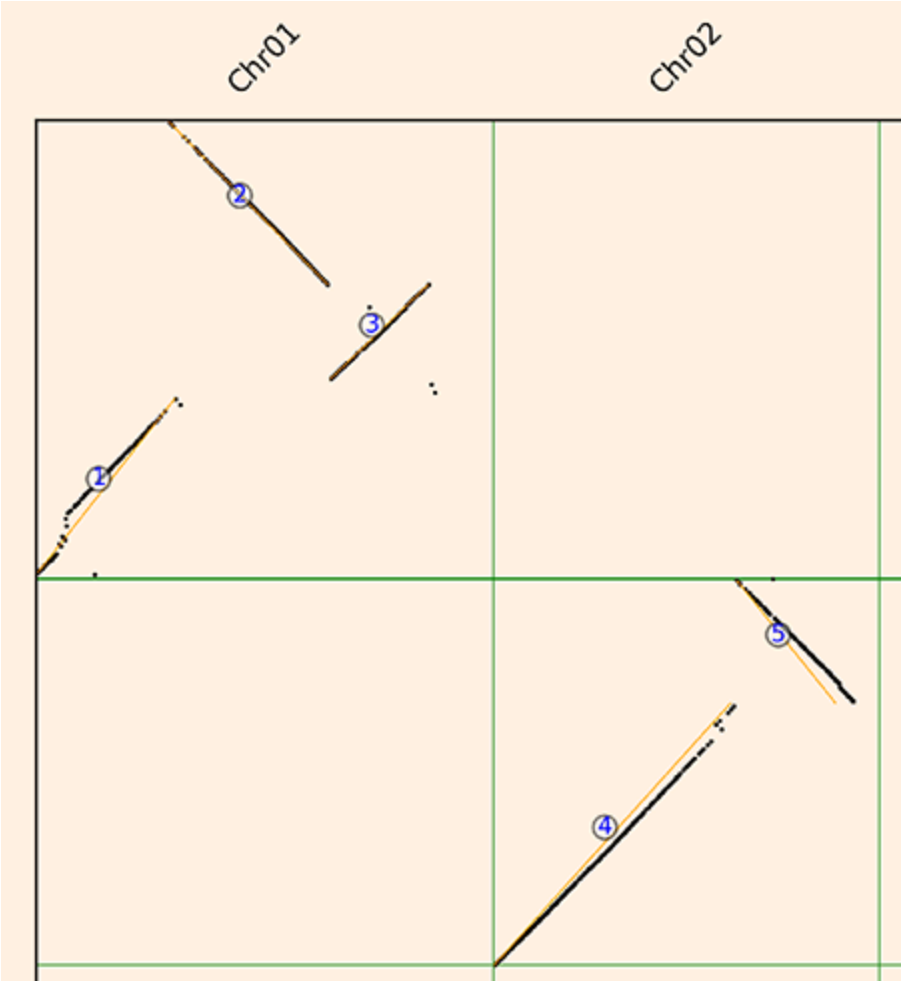
8. Swap blocks

This operation swap two blocks

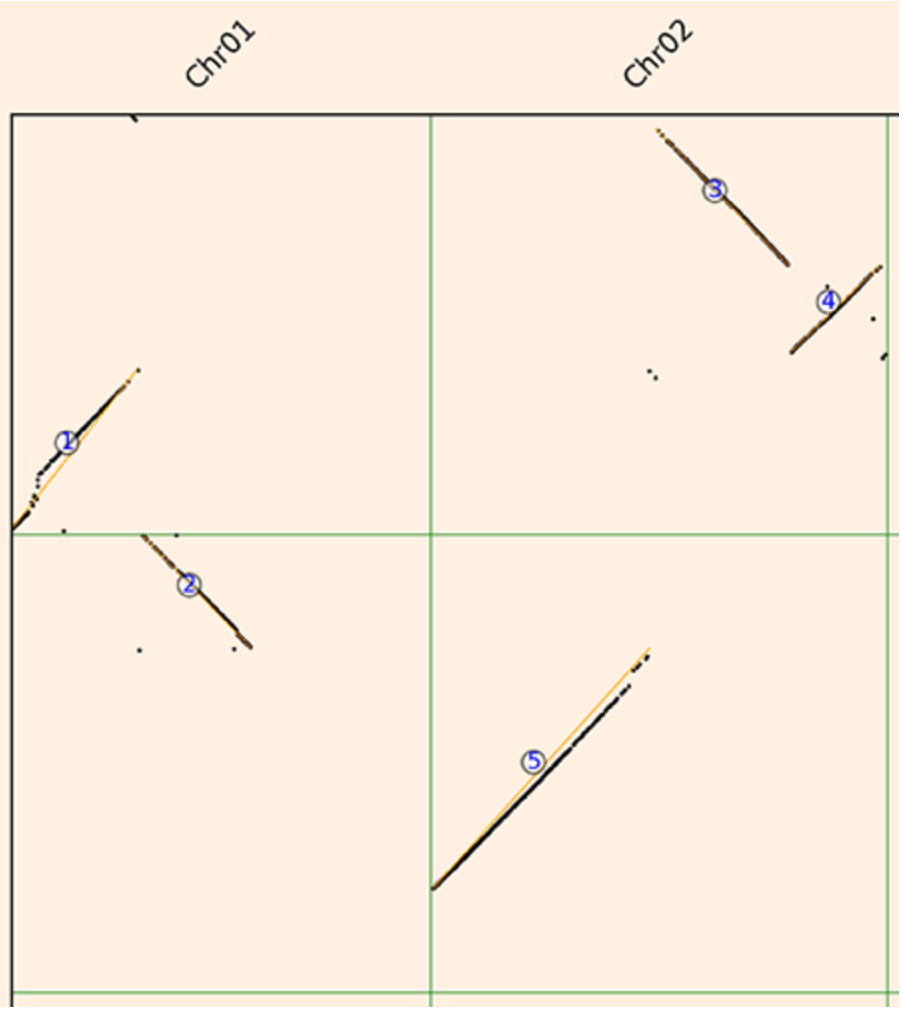
Set "Source chromosome", "Source block id", "Target chromosome" and "Target block id", then select "Swap block" option, click `MODIFY` button.

There is an example that swap the block 1 in chromosome 1 and the block 4 in chromosome 2:

- Before



- After



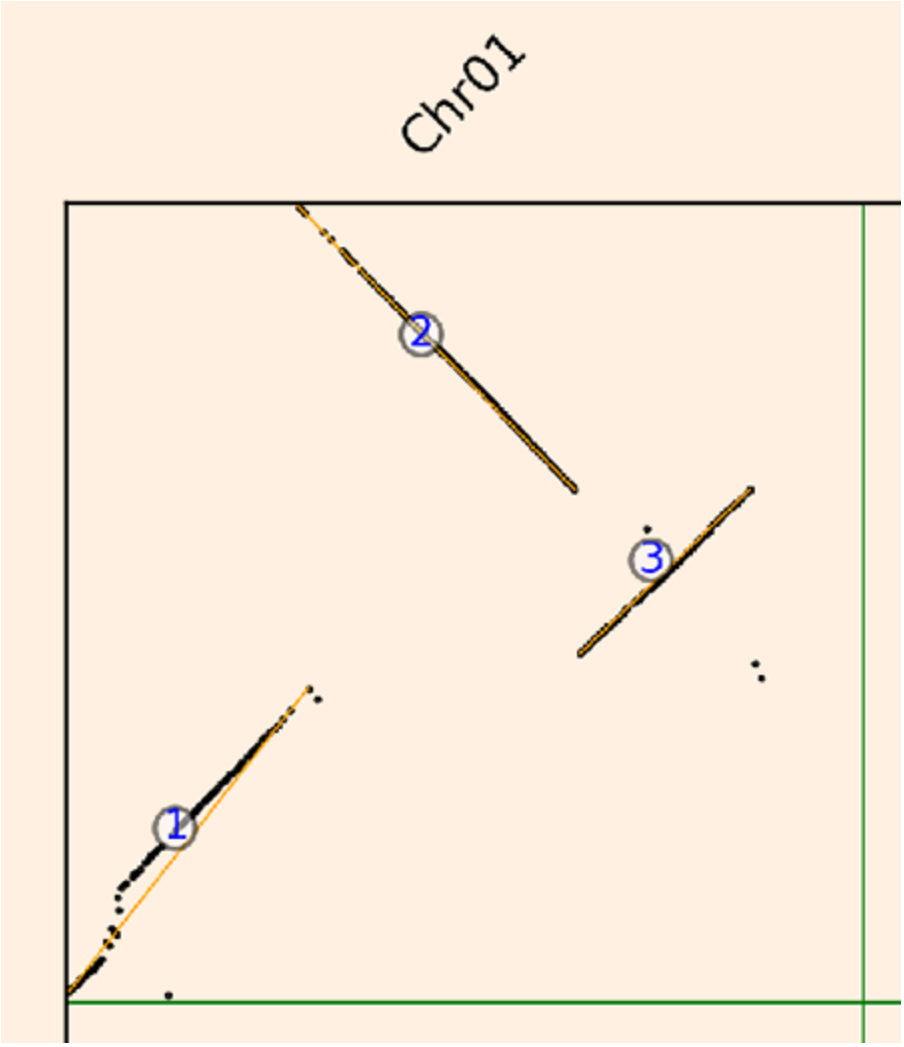
9. Delete block

This operation delete block

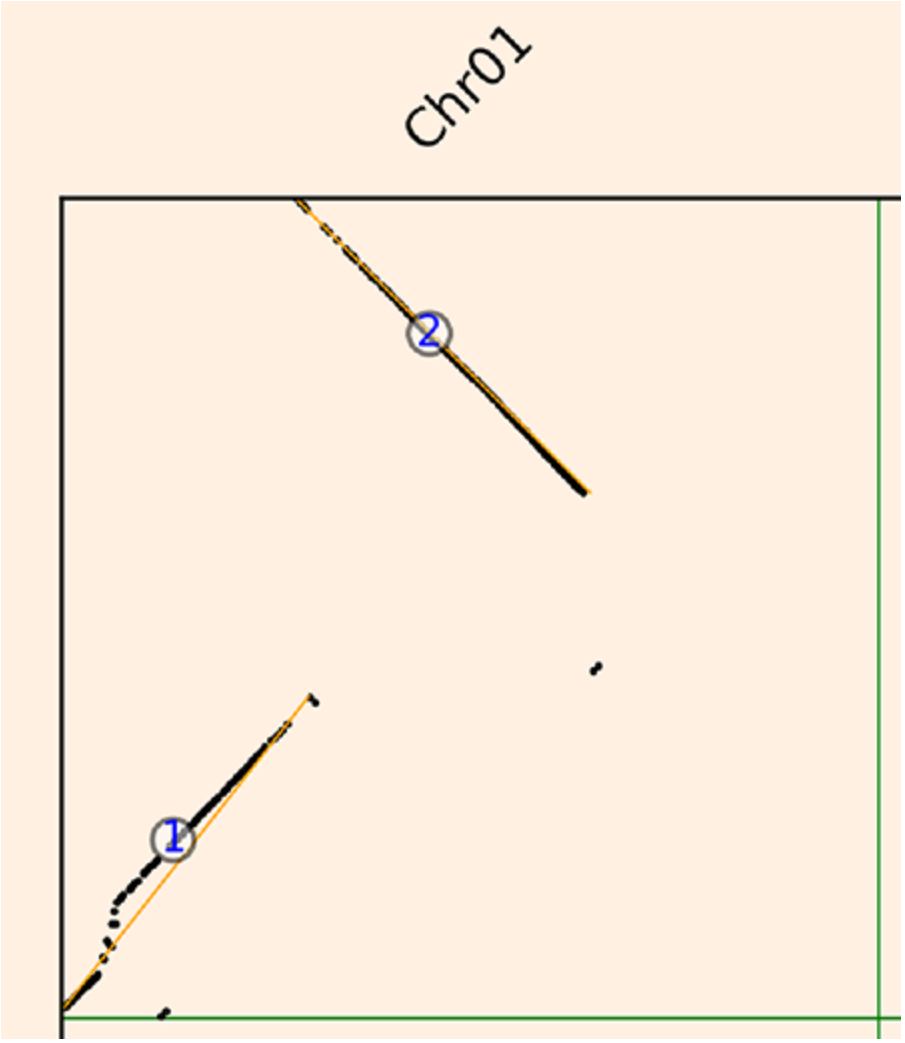
Set "Source chromosome" and "Source block id", then select "Delete block" option, click `MODIFY` button.

There is an example that delete the block 3 in chromosome 1:

- Before



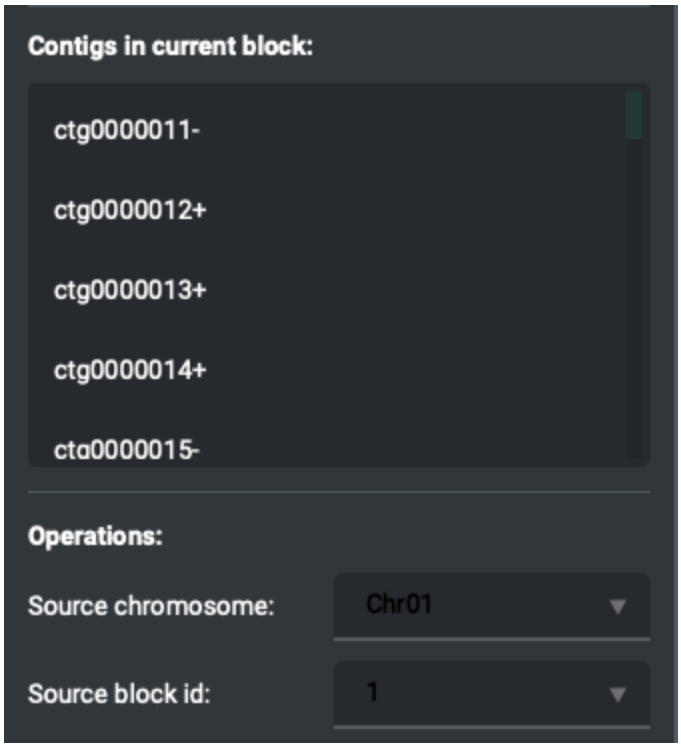
- After



Other operations

1. Check the contigs in one block

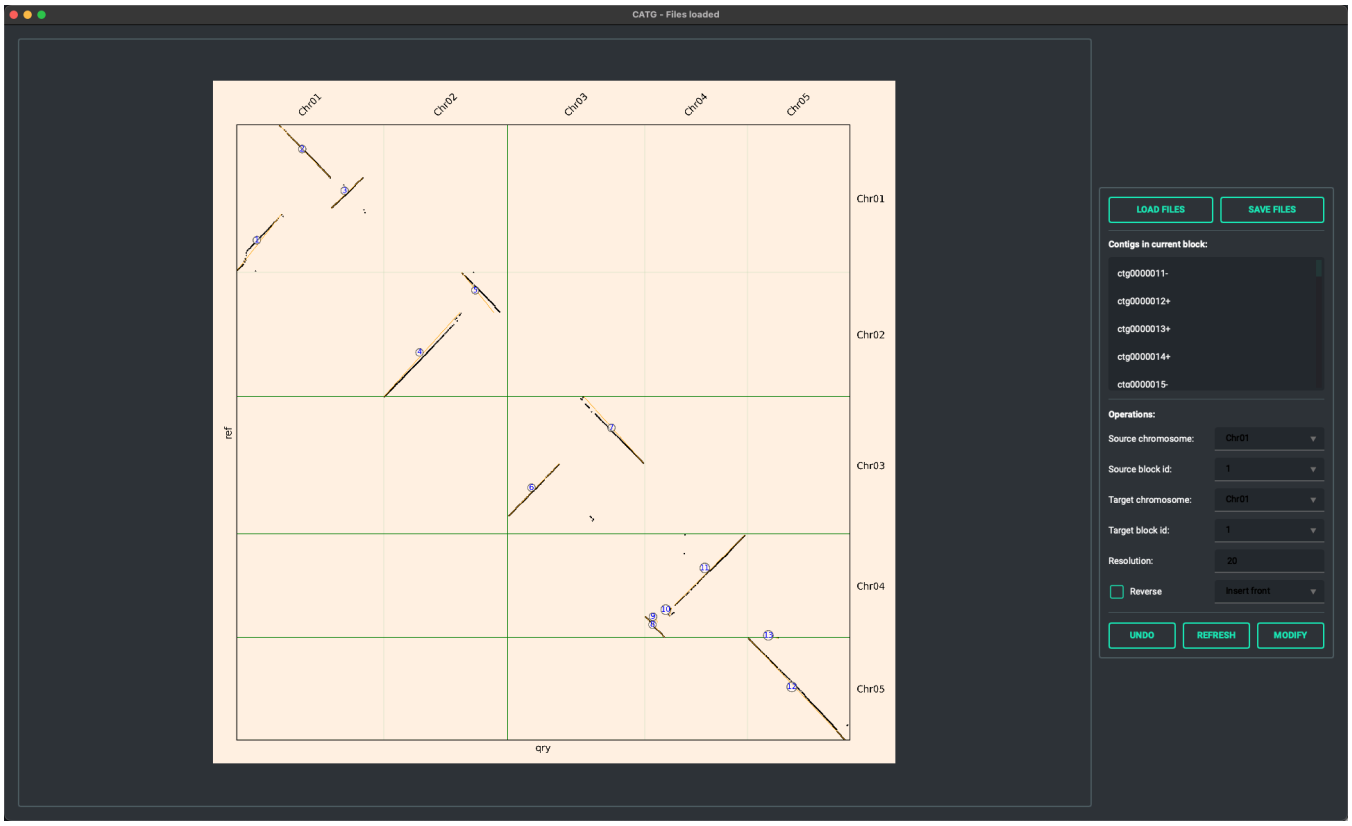
Once "Source chromosome" and "Source block id" was set, the "Contigs in current block" list would be automatically updated and show the order and orientation of contigs in this block.



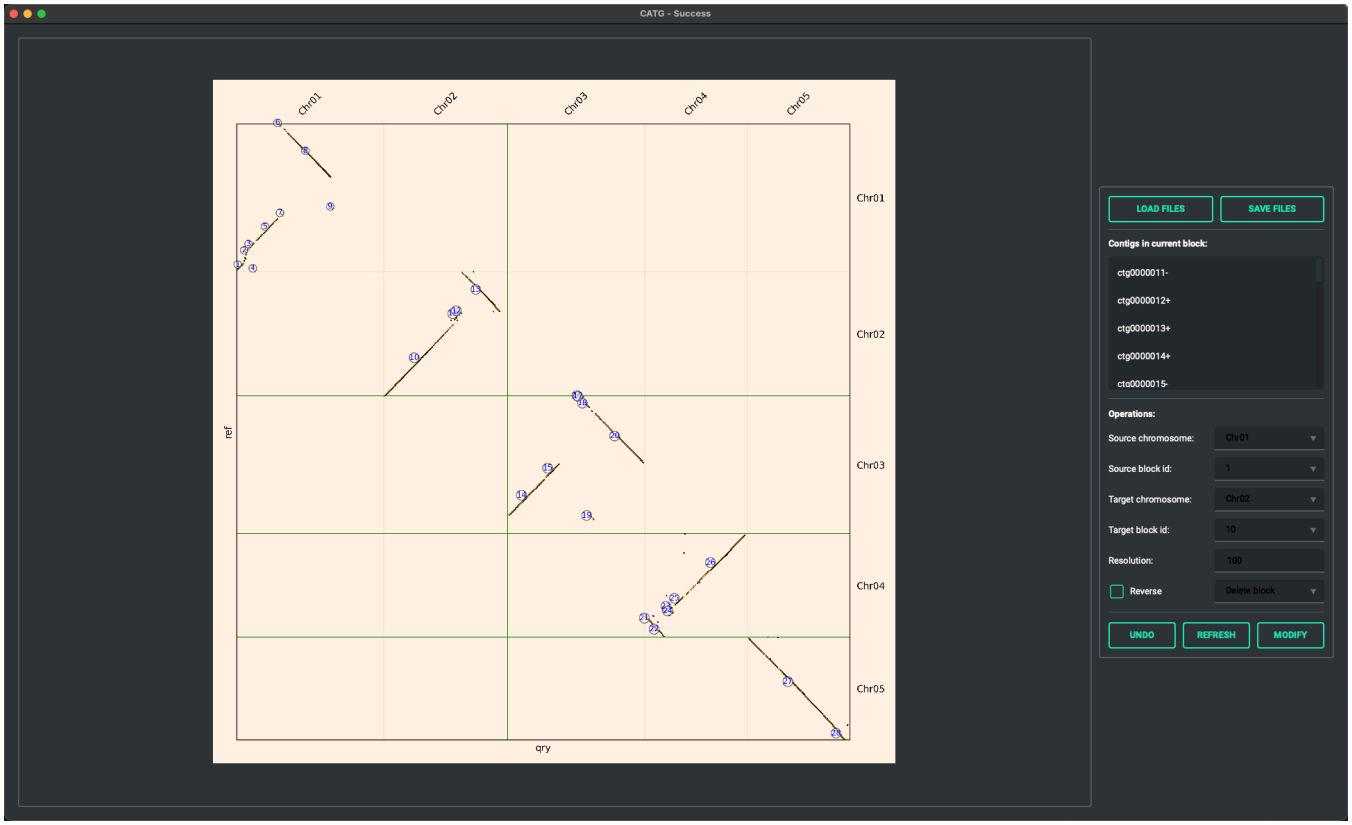
2. Re-cluster with other resolution

Once different resolution was set, and click the `REFRESH` button, the contigs would be re-cluster, and the blocks would be remark.

- Before


CATG - Files loaded
Chr01
Chr02
Chr03
Chr04
Chr05
ref
qry
LOAD FILES
SAVE FILES
Contigs in current block:
ctg0000011-
ctg0000012+
ctg0000013+
ctg0000014+
cta0000015-
Operations:
Source chromosome:
Chr01
Source block id:
1
Target chromosome:
Chr01
Target block id:
1
Resolution:
20
Reverse
Insert front
UNDO
REFRESH
MODIFY

- After

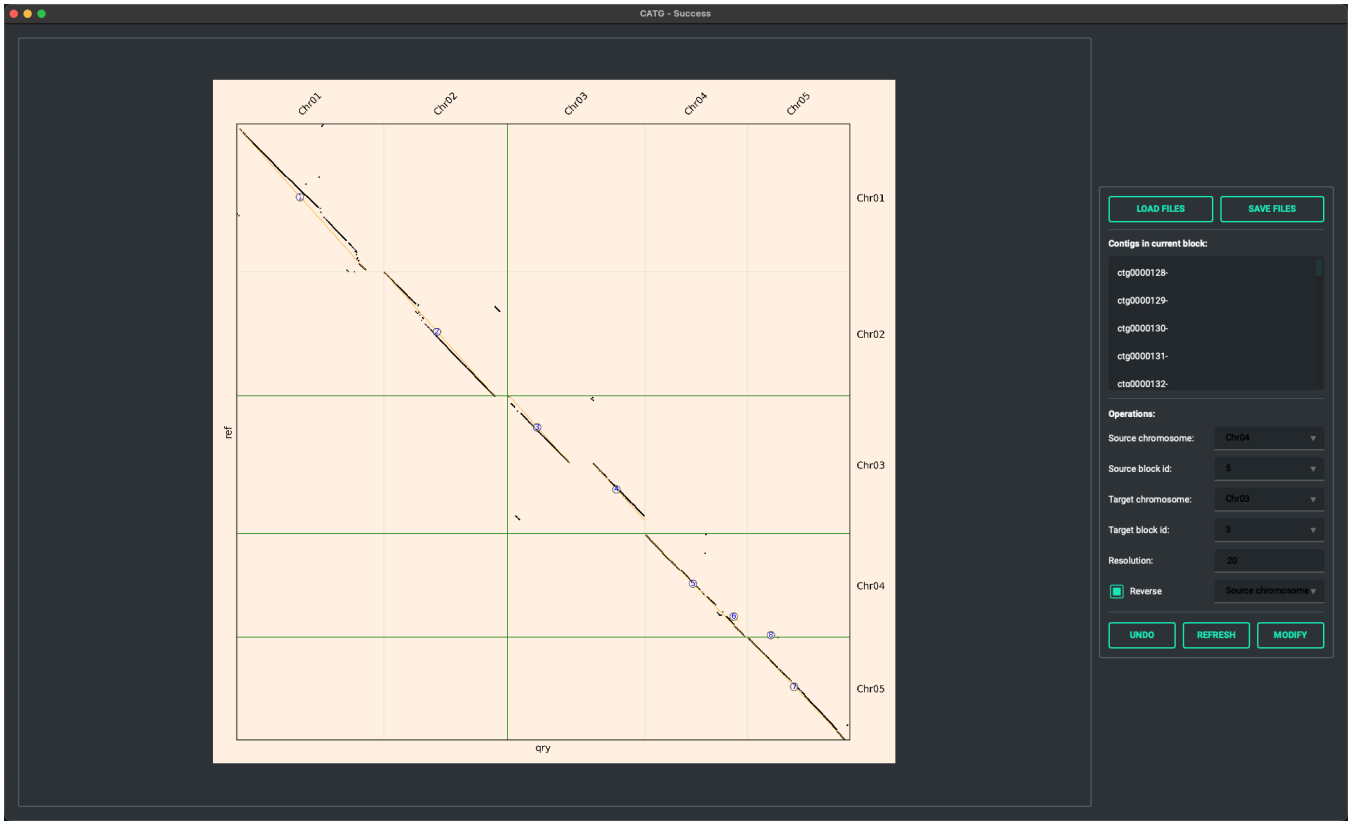

CATG - Success
Chr01
Chr02
Chr03
Chr04
Chr05
ref
qry
LOAD FILES
SAVE FILES
Contigs in current block:
ctg0000011-
ctg0000012+
ctg0000013+
ctg0000014+
cta0000015-
Operations:
Source chromosome:
Chr01
Source block id:
1
Target chromosome:
Chr02
Target block id:
10
Resolution:
100
Reverse
Delete block
UNDO
REFRESH
MODIFY

3. Undo

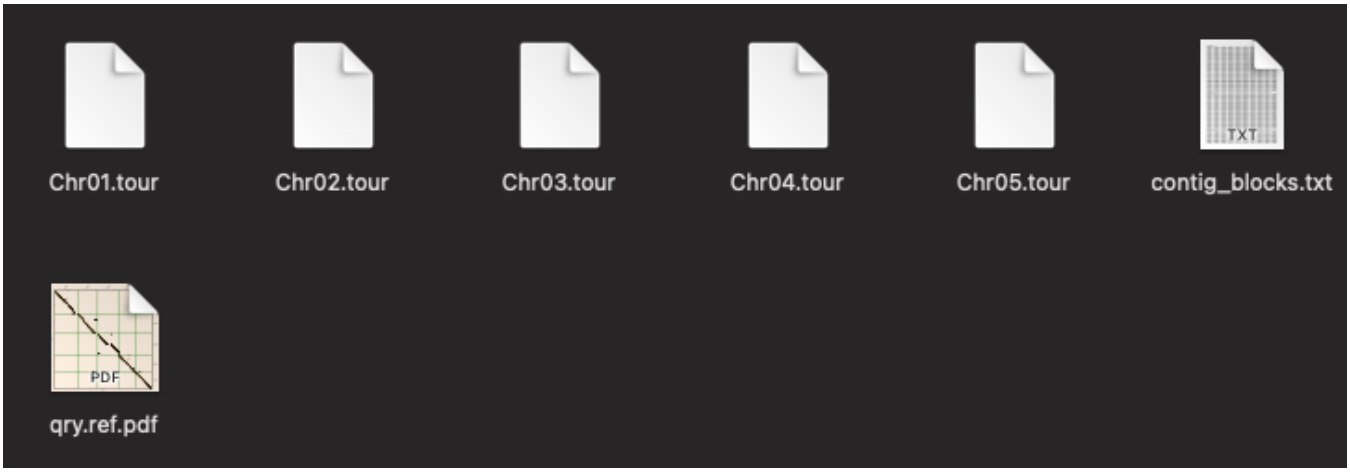
If user did a wrong operation, the `UNDO` button would cancel current operation and turn back to previours result.

Save results

After correction like below



Click the `SAVE FILES` button would let user select target position to save files.



The *.tour files can be used to generate final assembly by ALLHiC_build with the contig-level assembly (the one which was used to scaffold the genome before correction, the id and length of contigs should match which recorded in AGP file).