

Curating assemblies in Juicebox with Juicebox Assembly Tools (JBAT)

Jessen V. Bredeson (LBNL/DOE-JGI), 2022-03-17

File types

- **.fasta** : Text file of nucleotide sequences (may contain scaffolding gaps).
- **.assembly** : Text file representing the order and orientations of contigs as vectors of signed contig indices. Does not contain any nucleotide sequence. Input file to Juicebox.
 - **'Map' .assembly** : An .assembly-formatted file representing the *physical* state (subsequence connectedness, order, and orientation) of sequences (contig or scaffold) in the .fasta file that the Hi-C reads were mapped to.
 - **'Modified' .assembly** : An .assembly-formatted file representing the *virtual* state of sequences after a series of breaking and/or ordering-and-orienting operations.
- **merged_nodups.txt** : SAM-like file of Hi-C read pair mapping information output by Juicer. Intermediate file, precursor of .hic file. Not visualizable.
- **.hic** : Binary container file of contact matrices at pre-computed resolutions. Input file to Juicebox.

Generating a .hic file

- Using Juicer + 3D-DNA (see Genome Assembly Cookbook):
https://aidenlab.org/assembly/manual_180322.pdf
- From .bam files:
<https://github.com/Yujiaxin419/ALLHiC/wiki/Manually-refine-ALLHiC-scaffold-assembly-through-juicebox#hic-files>

To create a 'map' .assembly file representing individual contigs in a pre-scaffolded genome assembly, download artisanal (<https://bitbucket.org/bredeson/artisanal>), then do:

```
git clone https://bredeson@bitbucket.org/bredeson/artisanal.git
```

```
# Build versioned artisanal scripts:
```

```
pushd artisanal  
make install PREFIX=$PWD  
source ./activate  
popd
```

```
# The following script requires the pysam module be installed  
assembly-to-fastq -c genome.fasta genome
```

Introduction to the Juicebox desktop app

<https://github.com/aidenlab/Juicebox/wiki/Download>

Anatomy of the Juicebox desktop app

Opened contact maps

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

Menu bar: File View Bookmarks Assembly Dev

View controls: Chromosomes Show Normalization Resolution (BP) Color Range Goto

Chromosomes: assem... assem...

Show: Observed vs ...

Normalization: Bal... Bal...

Resolution (BP): 2.5 MB 500 KB 100 KB 25 KB 5 KB

Color Range: 0 8 17

Goto: h1tg000005l h2tg000008l

View area

Minimap

assembly:63,300,001-63,400,000
assembly:26,800,001-26,900,000
observed value (O) = 3.679
expected value (E) = 1.296
O/E = 2.839

control value (C) = 2.115
expected control value (EC) = 0.121
C/EC = 17.553

average observed value (AVG) = 2.42
O' = O/AVG = 1.52
average control value (AVGC) = 0.33
C' = C/AVGC = 6.4
O'/C' = 0.237
(O'-C')*(AVG/2 + AVGC/2) = -6.712121

assembly (control)

assembly (observed)

Right-side panel

Edit

Chr

Scaf

Show Annotation Panel

Annotation panel

Menu bar

View controls

Contact map window

Horizontal (X) axis coordinates

Vertical view range

Horizontal cursor position

Vertical cursor position

Vertical (Y) axis scroll bar

Horizontal (X) axis scroll bar

Vertical (Y) axis coordinates

Horizontal view range

"Main" diagonal

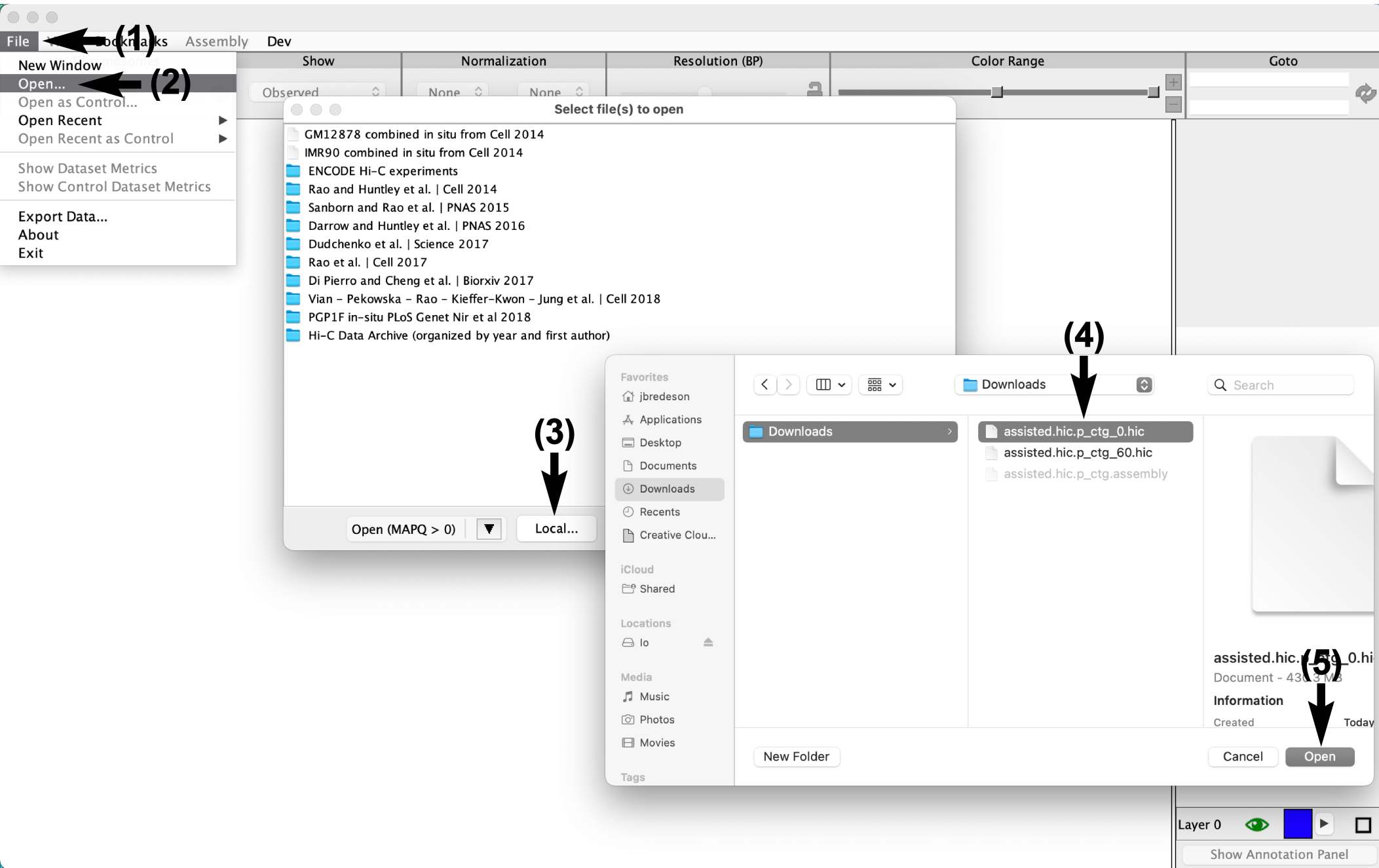
"Chromosome" boundary (really a scaffold)

"Scaffold" boundary (really a contig)

High contact density

Low contact density

Opening an 'Observed' .hic contact map file



Opening a 'Control' .hic contact map file

The screenshot shows the Juicebox 1.11.08 interface. The main window displays a heatmap with a red diagonal line. The menu bar includes 'File', 'Bookmarks', 'Assembly', and 'Dev'. The 'File' menu is open, showing options like 'New Window', 'Open...', 'Open as Control...', 'Open Recent', 'Open Recent as Control', 'Show Dataset Metrics', 'Show Control Dataset Metrics', 'Export Data...', 'About', and 'Exit'. A 'Select file(s) to open' dialog box is open, showing a list of files and folders. The 'Downloads' folder is selected, and the file 'assisted.hic.p_ctg_60.hic' is highlighted. The 'Open' button is visible at the bottom right of the dialog box. The background heatmap shows a red diagonal line, indicating a control dataset. The 'Resolution (BP)' and 'Color Range' sliders are visible at the top. The 'Goto' field is empty. The 'Layer 0' control is visible at the bottom right.

(1) File

(2) Open as Control...

(3) Local...

(4) assisted.hic.p_ctg_60.hic

(5) Open

Opening a 'map' .assembly file

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg.0.hic (control=assisted.hic.p_ctg_60.hic)

File View Bookmarks Assembly **(1)**

Chromosomes

assem... assem...

Import Map Assembly **(2)**

Import Modified Assembly

Export Assembly

Reset Assembly

Set Scale

Exit Assembly

Normalizati

None

Downloads **(3)**

assisted.hic.p_ctg_0.hic

assisted.hic.p_ctg_60.hic

assisted.hic.p_ctg.assembly

assisted.hic.p_ctg.assembly

Document - 35 KB

Information

Created Today 5:57 PM

Cancel **(4)** Open

Select Assembly annotation file(s) to open

Chromatin Features

Loop Calls

Domains

Added Assembly Files

assisted.hic.p_ctg.assembly **(5)**

(6)

Open Assembly Cancel

assembly

Layer 0

Show Annotation Panel

30.057

O' = O/AVG = 0.233

average control value (AVGC) = 30.076

C' = C/AVGC = 0.266

O/C' = 0.876

(O'-C')*(AVG/2 + AVGC/2) =

-0.99540716

Opening a 'modified' .assembly file

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View Bookmarks **Assembly** ← (1)

- Import Map Assembly
- Import Modified Assembly ← (2)
- Export Assembly
- Reset Assembly
- Set Scale
- Exit Assembly

Chromosomes: assem... assem...

Normalization: None None

Resolution (BP): 2.5 MB 500 KB 100 KB 25 KB 5 KB

Color Range: 0 382 772

Goto: [] []

0 MB 100 MB 200 MB 300 MB 400 MB 500 MB 600 MB

0 MB 100 MB 200 MB

384,000,000 328,000,000 3.848 (EC) = 8.833 (AVG) = (VGC) = 30.076

assisted.hic.p_ctg.review.embly Document - 35 KB Information Created Today

Cancel Open

Chromatin Features

- Loop Calls
- Domains

Added Assembly Files

- assisted.hic.p_ctg.review.assembly ← (5)

Open Assembly Cancel

(6)

Changing contact map(s) displayed with 'Show'

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View Bookmarks Assembly Dev

Chromosomes Show Normalization Resolution (BP) Color Range Goto

assem... assem... Observed vs ... (1) None

- Observed
- Expected
- Observed/Expected
- Observed Pearson
- Control
- Control/Expected
- Control Pearson
- ✓ Observed vs Control (2)

0 MB 100 MB 200 MB 300 MB 400 MB 500 MB 600 MB

assembly (observed)

assembly (control)

assembly:356,000,001-357,000,000
assembly:18,000,001-19,000,000
observed value (O) = 56
expected value (E) = 226.905
O/E = 0.247

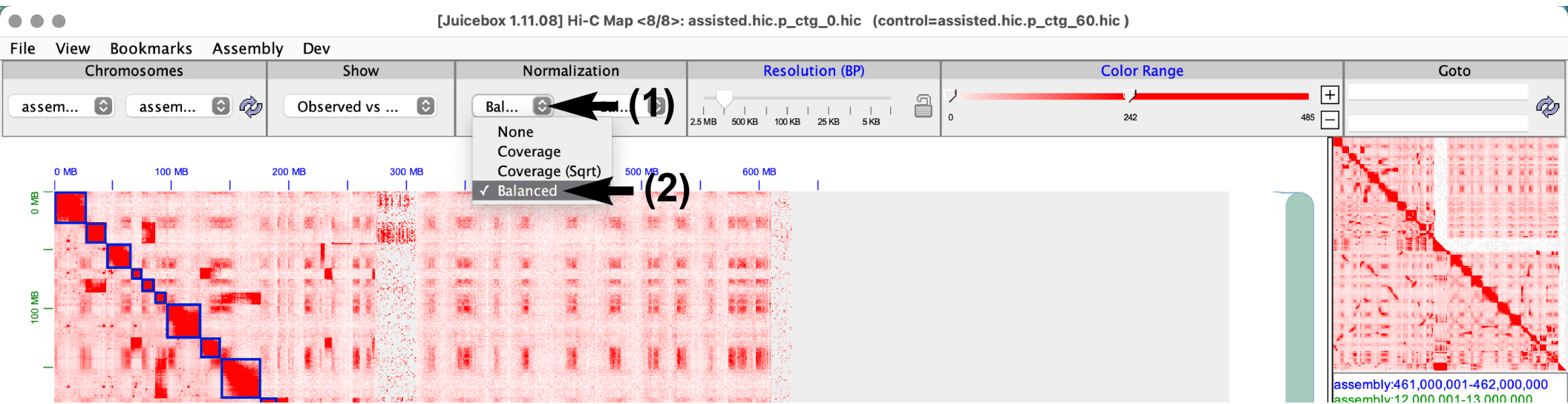
control value (C) = 10
expected control value (EC) = 11.331
C/EC = 0.883

average observed value (AVG) = 225.667
O' = O/AVG = 0.248
average control value (AVGC) = 30.076
C' = C/AVGC = 0.332
O'/C' = 0.746
(O'-C')*(AVG/2 + AVGC/2) = -10.784947

Edit Chr Scaf

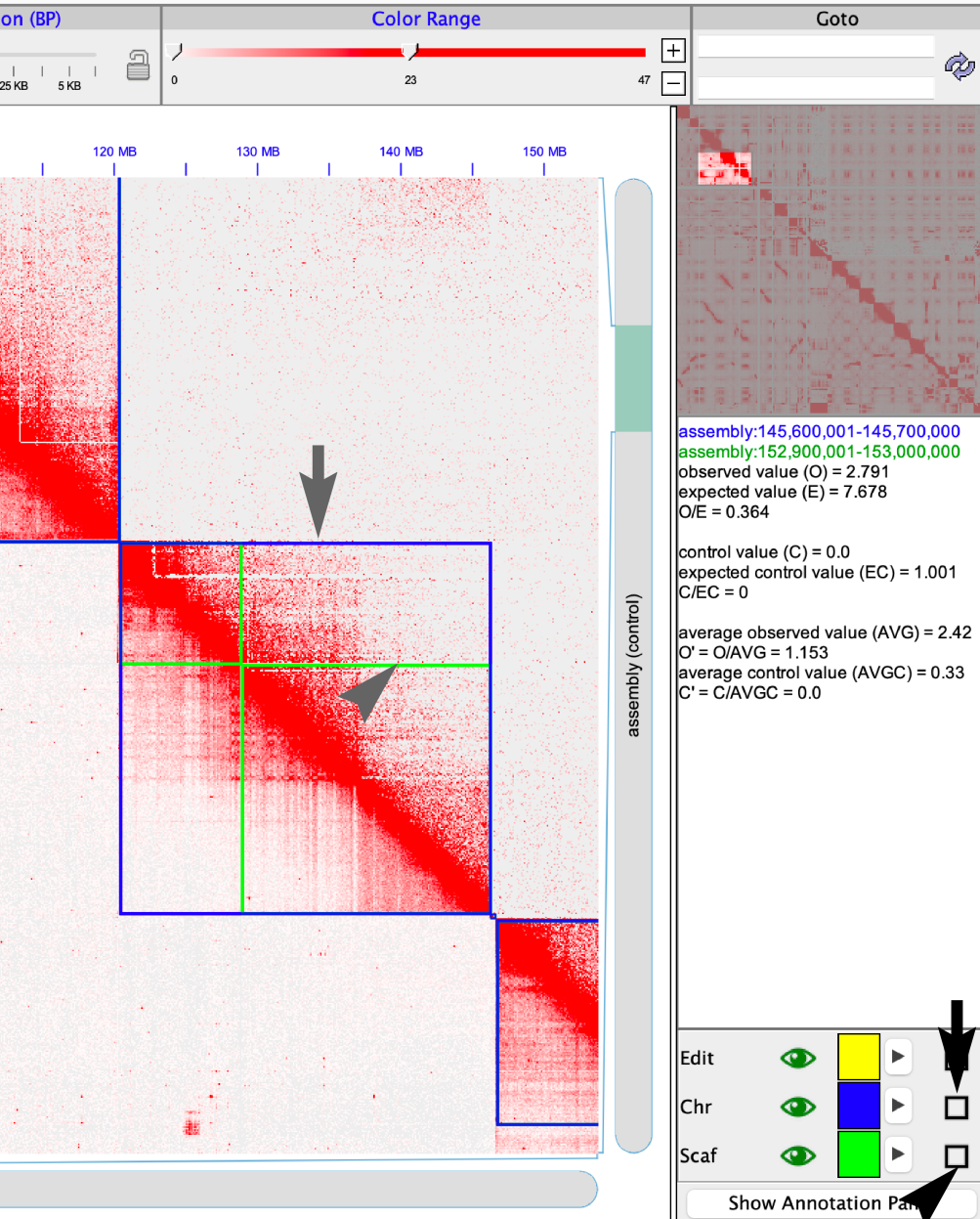
Show Annotation Panel

Enabling contact map normalization

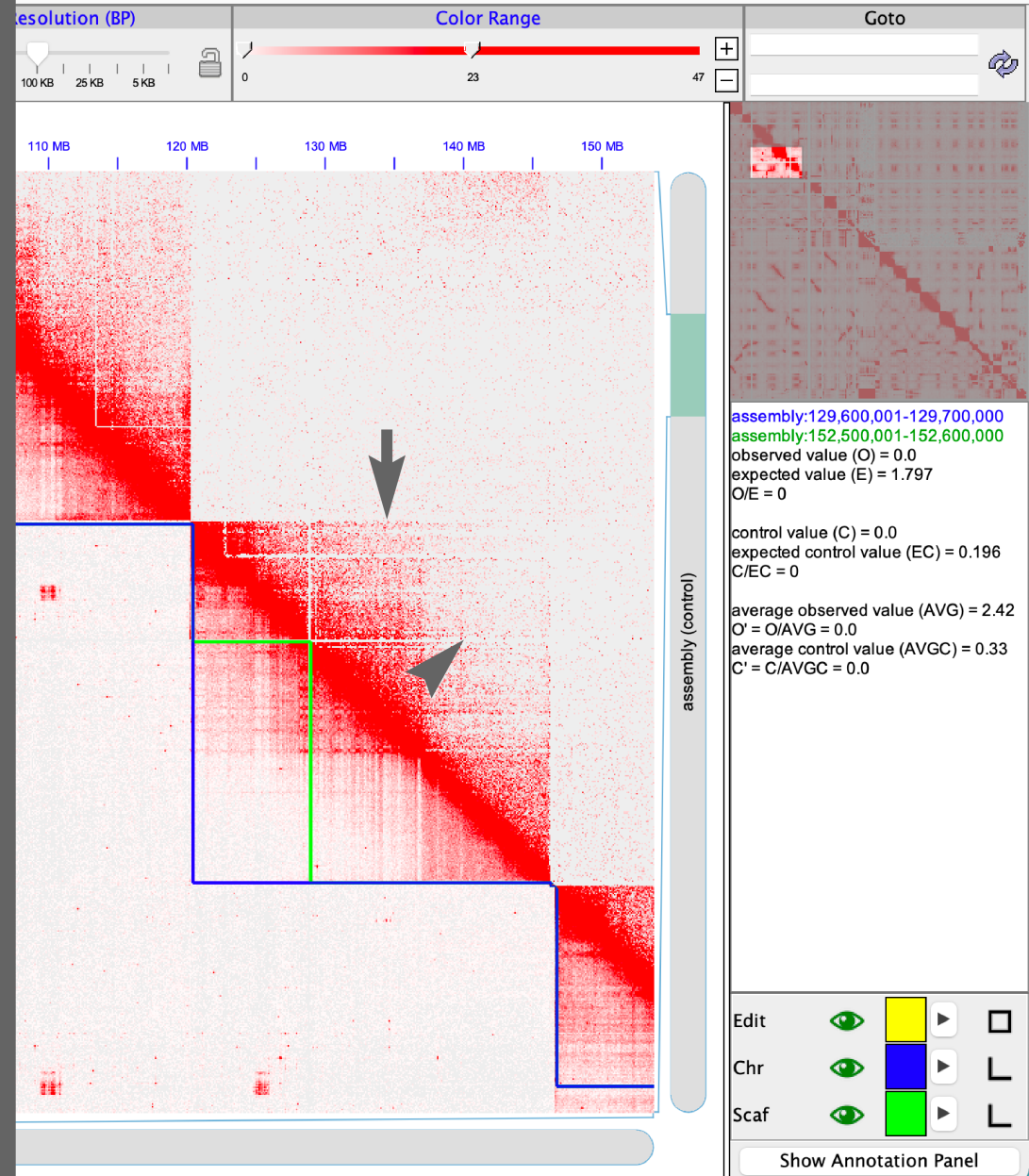


Toggling contig/scaffold boundary visibility

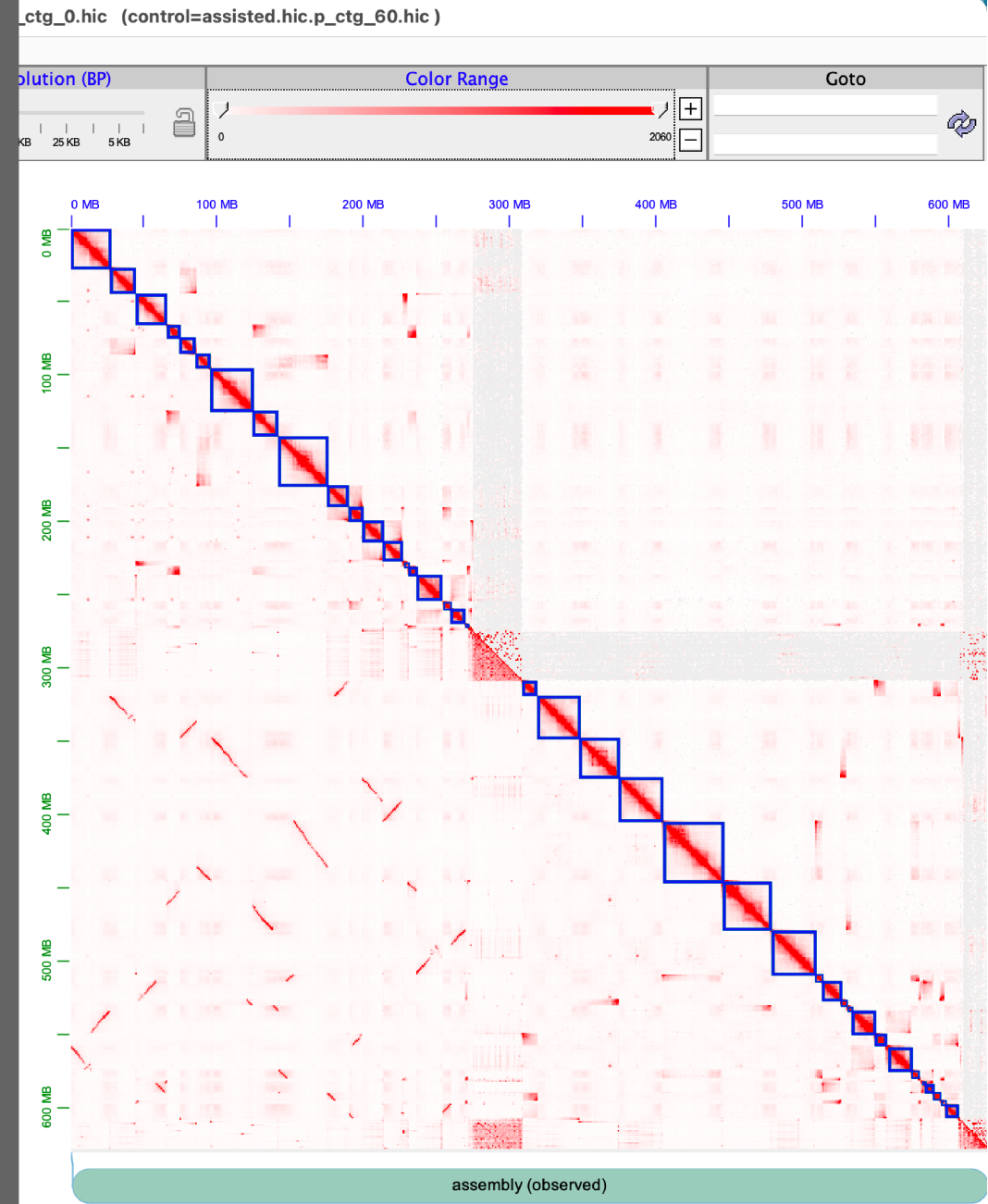
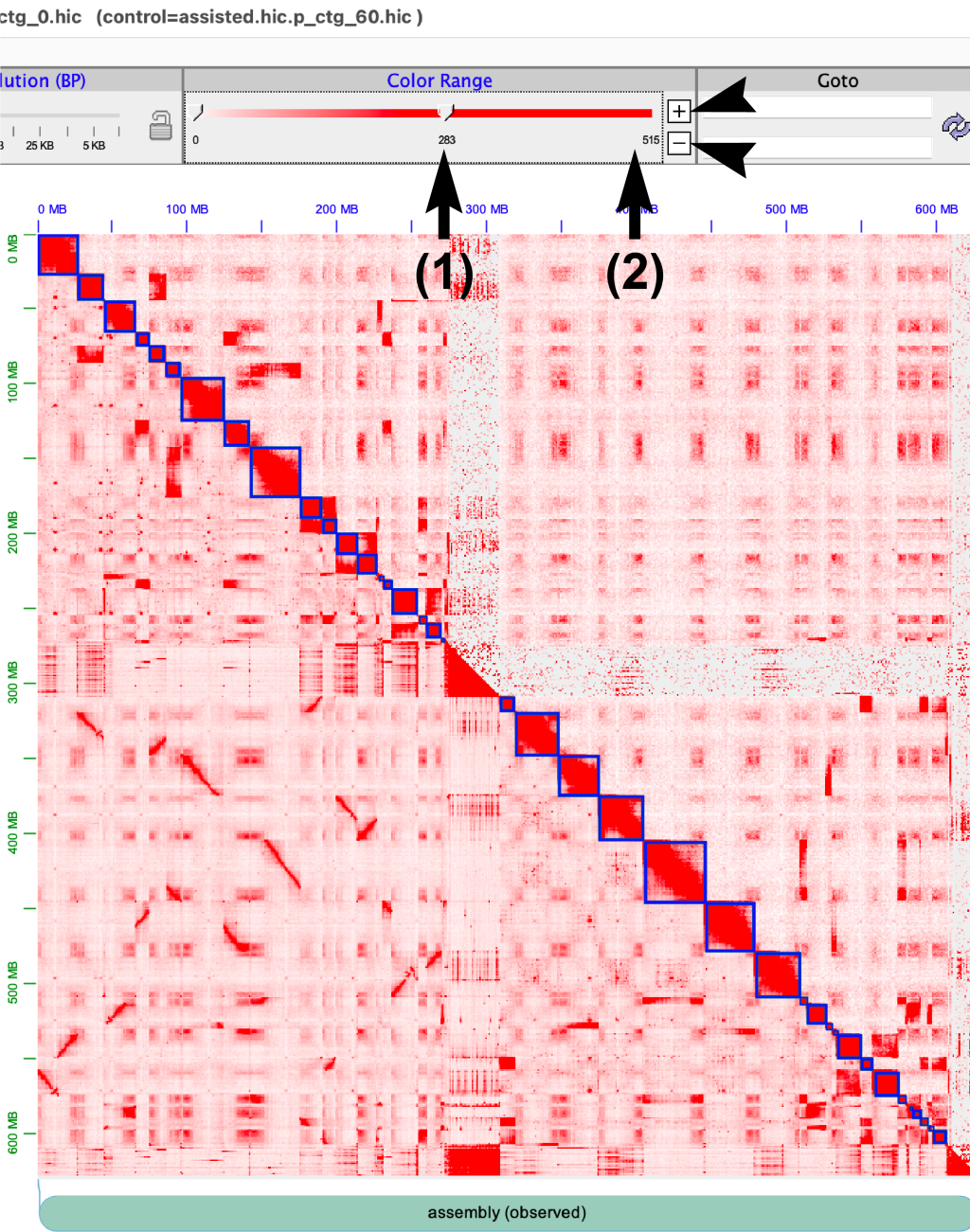
_0.hic (control=assisted.hic.p_ctg_60.hic)



_p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Adjusting contact map color range/saturation



Changing contact map heatmap color

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.

File View **(1)** Assembly Dev

- Show Annotation Panel
- Change Heatmap Color (2)**
- Darkula Mode
- Make Custom Chromosome (from .bed)...

Axis Endpoints Only
✓ Chromosome Context
✓ Gridlines

Export PDF Figure...
Export SVG Figure...

Select Heatmap Color

Swatches HSV HSL RGB CMYK

(3)

Recent:

Preview

(4)

Sample Text Sample Text
Sample Text Sample Text
Sample Text Sample Text

OK Cancel Reset

This screenshot shows the Juicebox Hi-C Map interface. The 'View' menu is open, and the 'Change Heatmap Color' option is highlighted with a black arrow and the number (2). The heatmap is currently displayed in red. Below the main interface, the 'Select Heatmap Color' dialog is open, showing a color palette with a yellow cursor pointing to a blue square, labeled with (3). Below the palette, there are preview boxes with text and a color bar, with an arrow pointing to the first blue box labeled (4). The dialog has 'OK', 'Cancel', and 'Reset' buttons at the bottom.

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.

File View Bookmarks Assembly Dev

Chromosomes Show Normalization

assem... assem... Observed vs ... Bal... Bal...

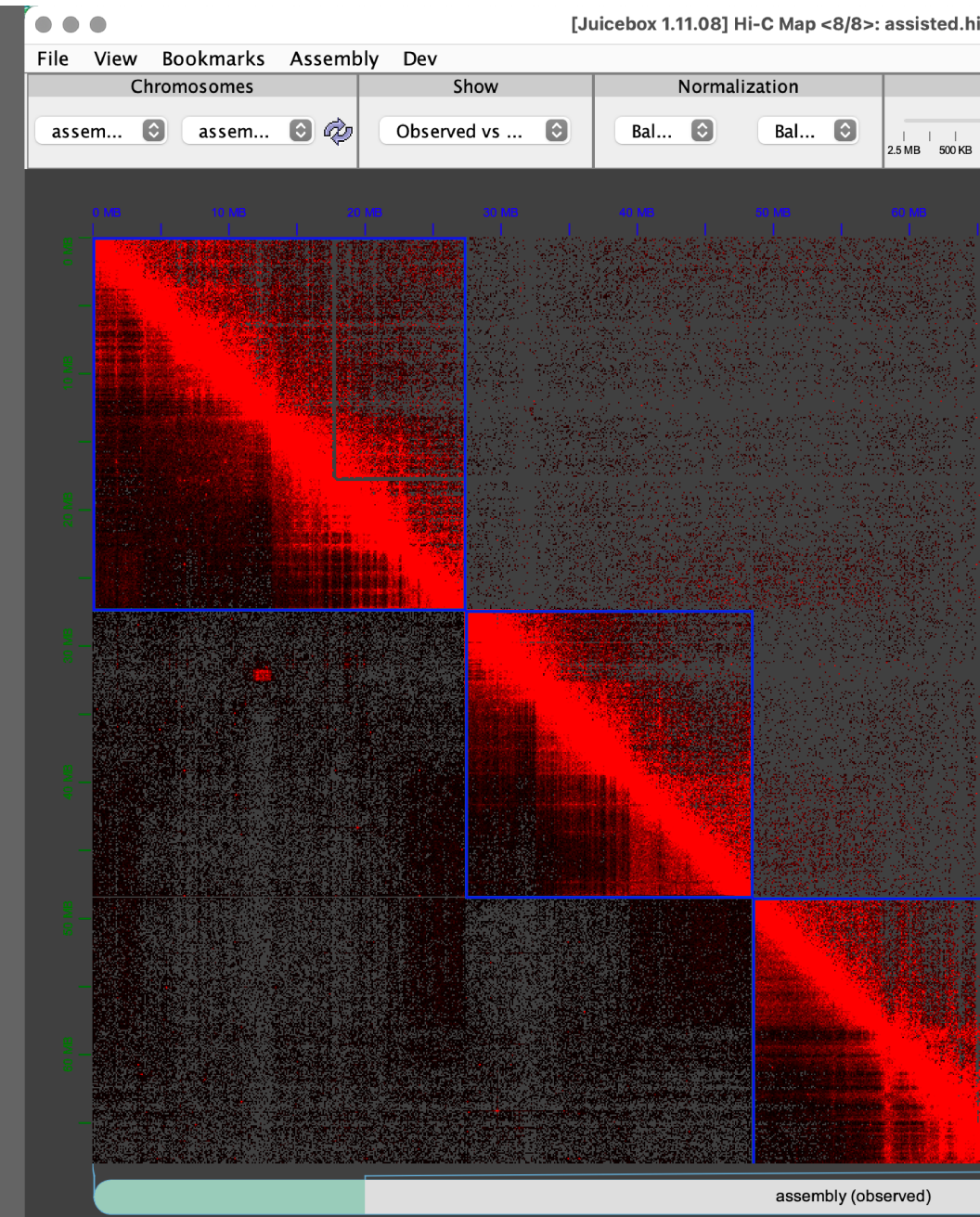
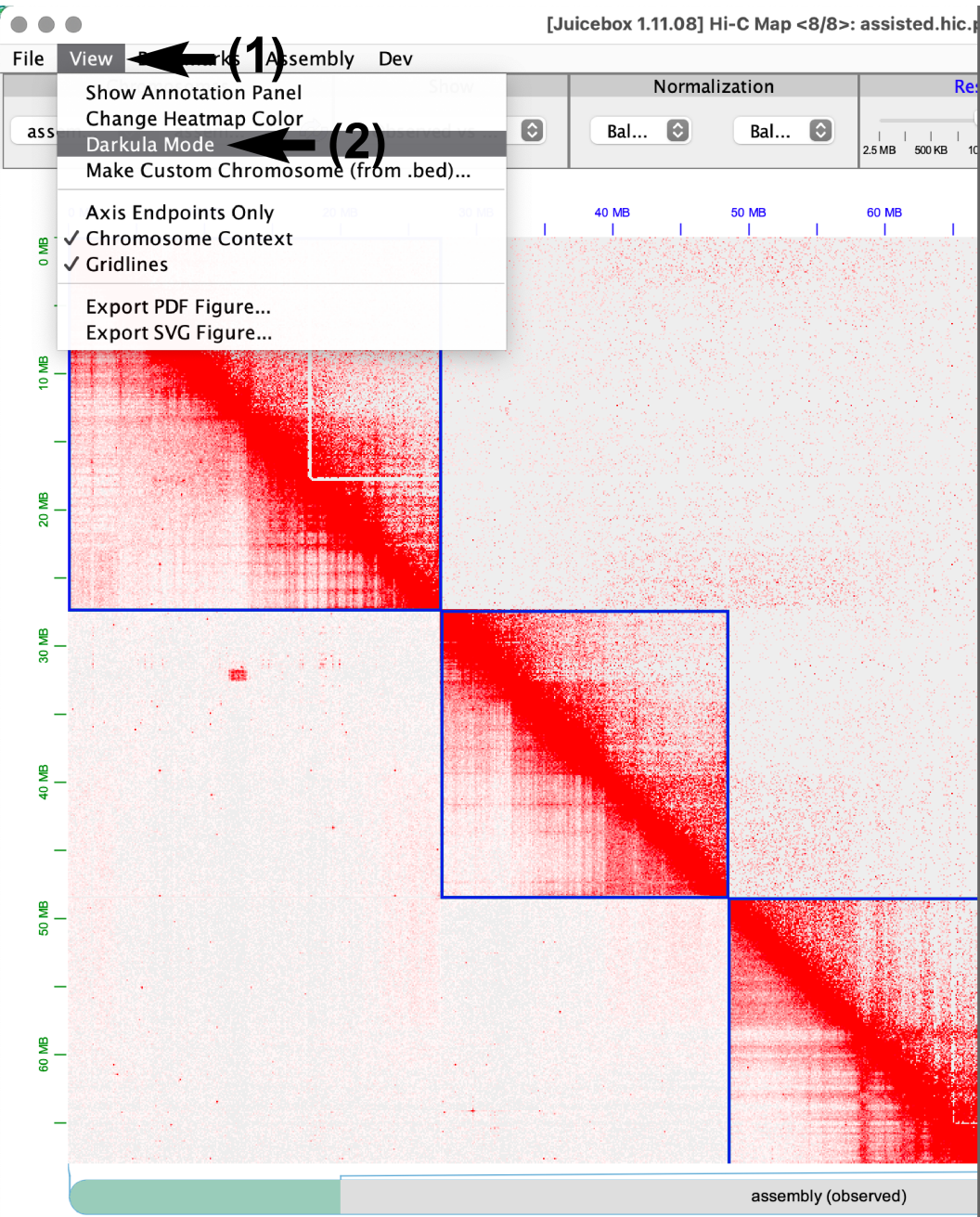
0 MB 10 MB 20 MB 30 MB 40 MB 50 MB 60 MB

0 MB 10 MB 20 MB 30 MB 40 MB 50 MB 60 MB

assembly (observed)

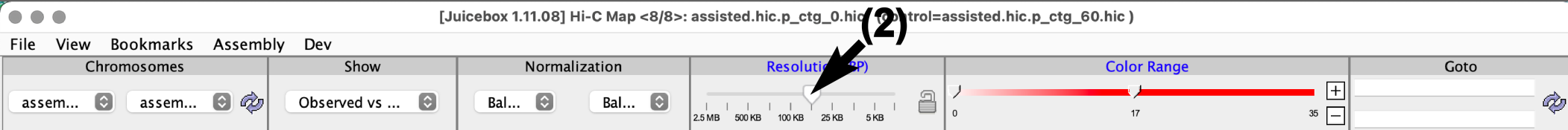
This screenshot shows the Juicebox Hi-C Map interface after the heatmap color has been changed to blue. The 'View' menu is closed, and the 'Show' dropdown is set to 'Observed vs ...'. The heatmap is now displayed in blue. The interface includes a 'Chromosomes' section with 'assem...' and 'assem...' dropdowns, and a 'Normalization' section with 'Bal...' and 'Bal...' dropdowns. The heatmap axes are labeled from 0 MB to 60 MB. The bottom right corner of the window shows 'assembly (observed)'.

Darkula (night) mode



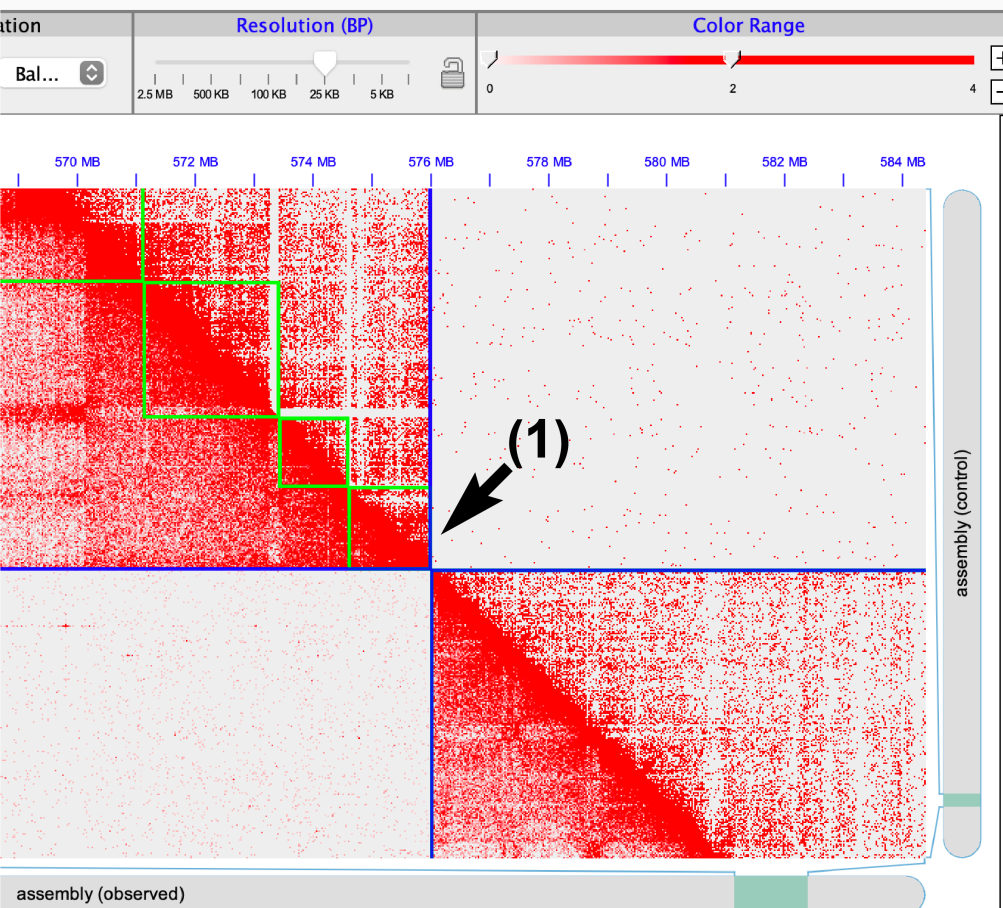
Zooming-in/-out with the 'Resolution' slider

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

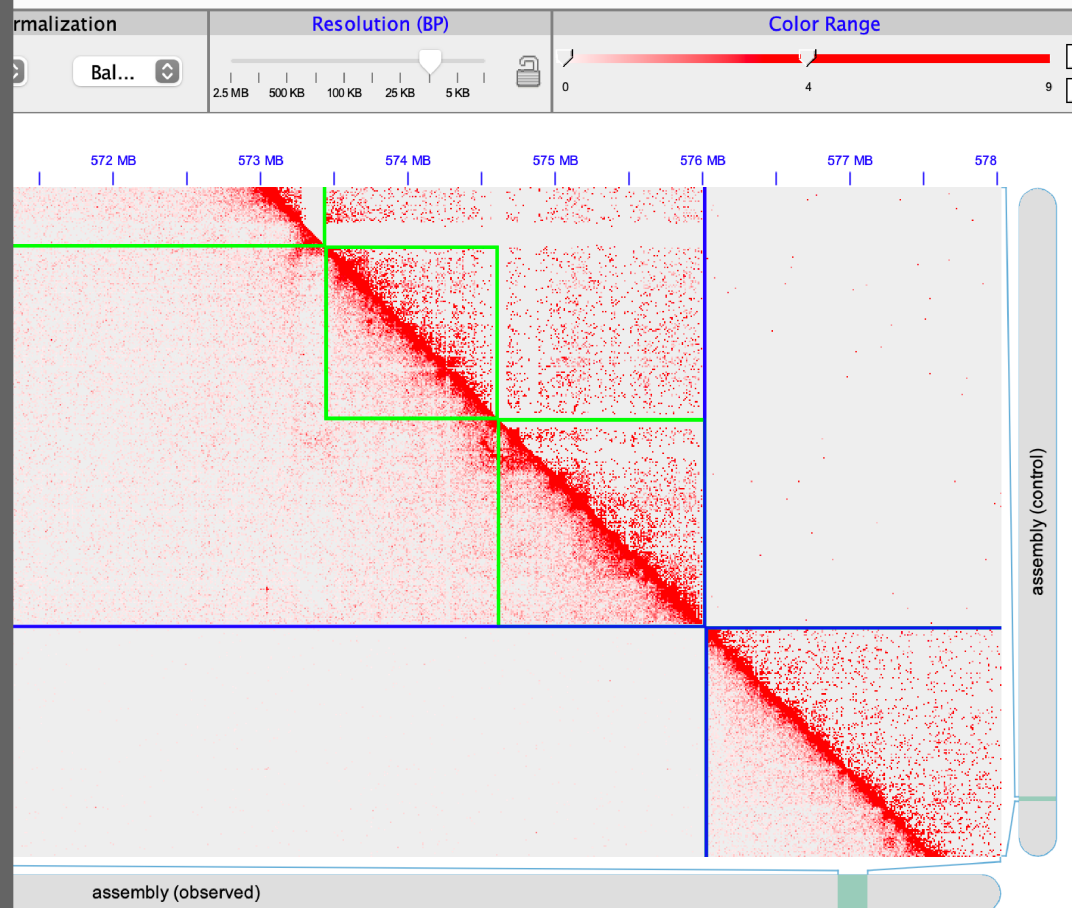


Zooming-in with a mouse double-click

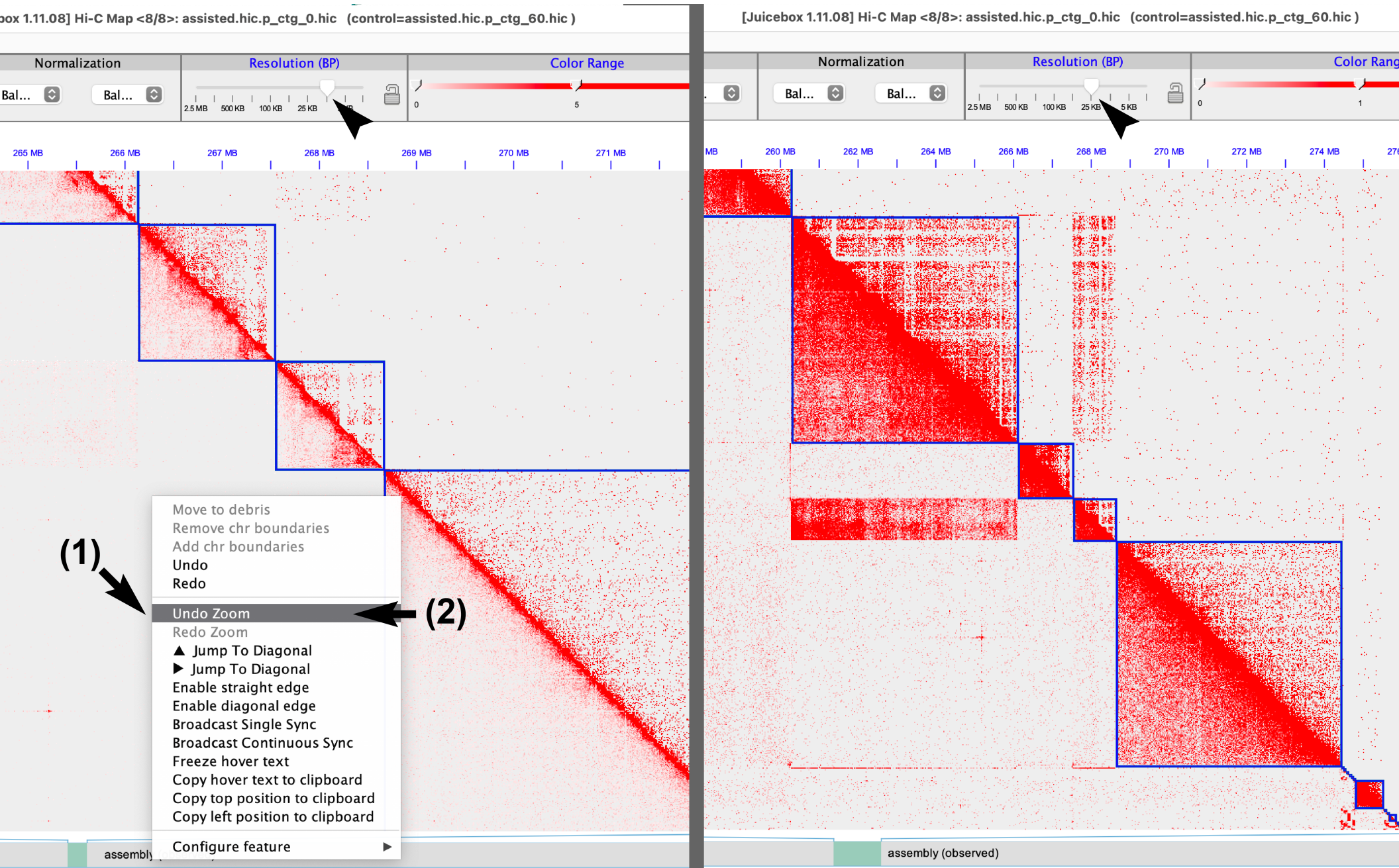
Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



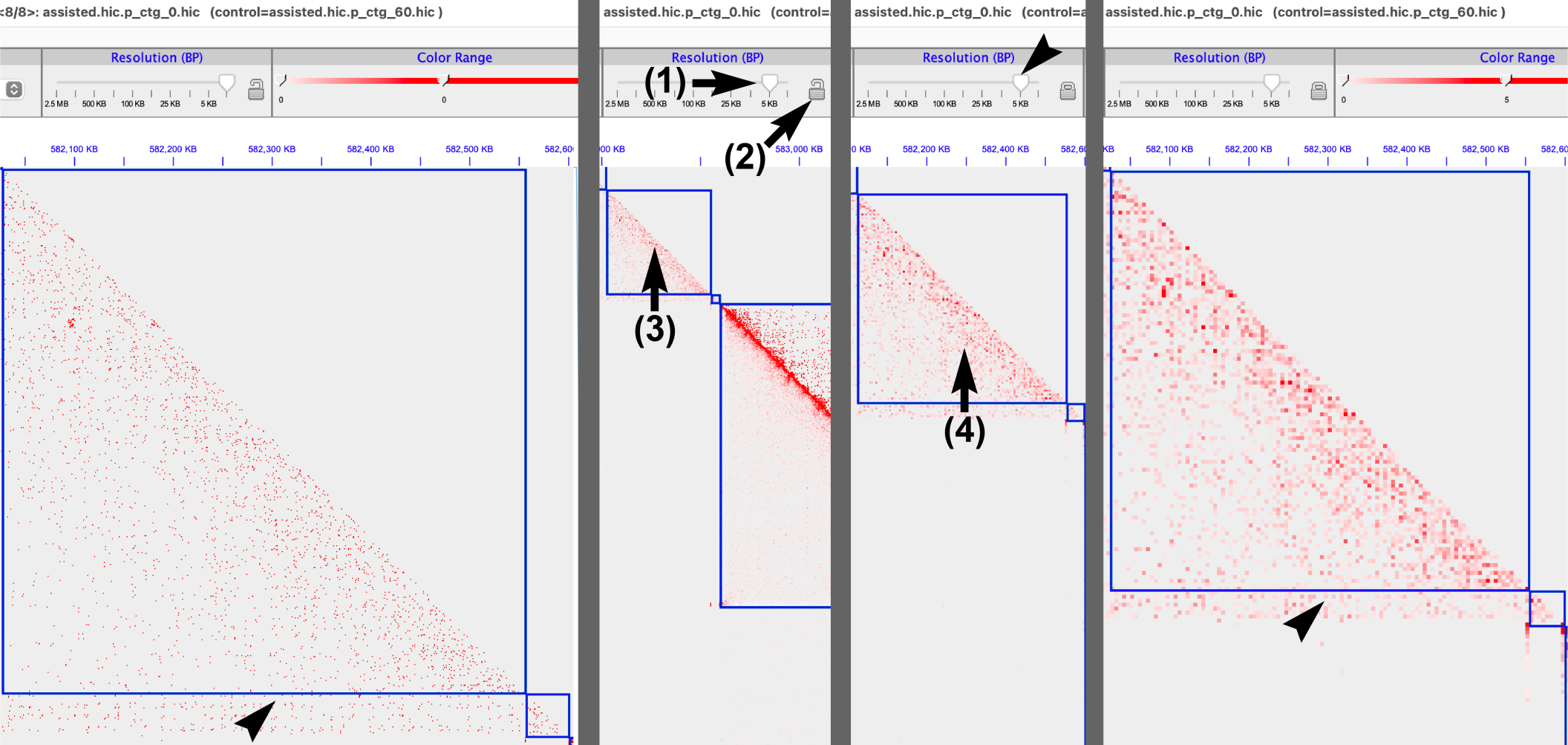
[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Undo zoom / Zooming out

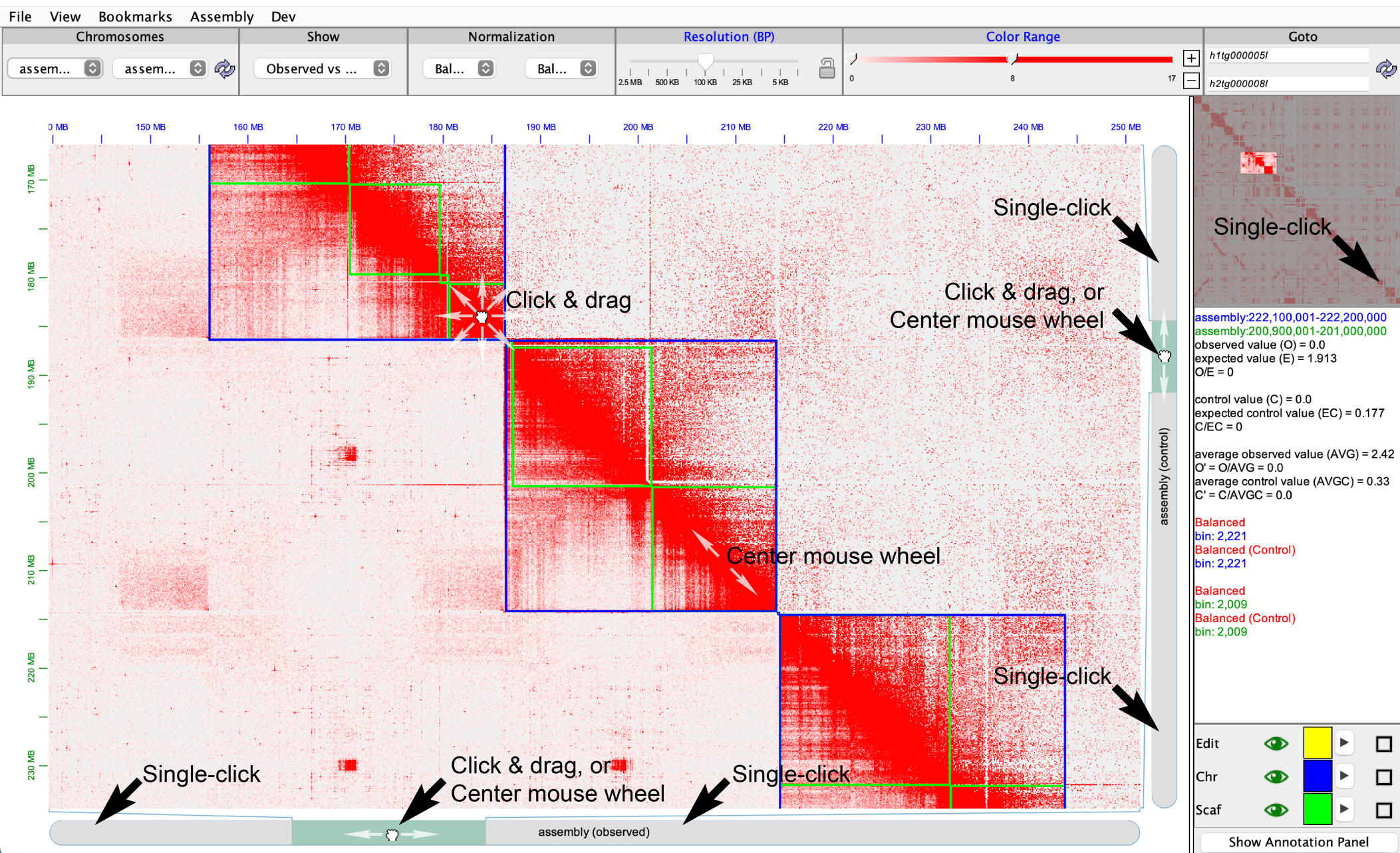


Zoom with the 'Resolution' lock

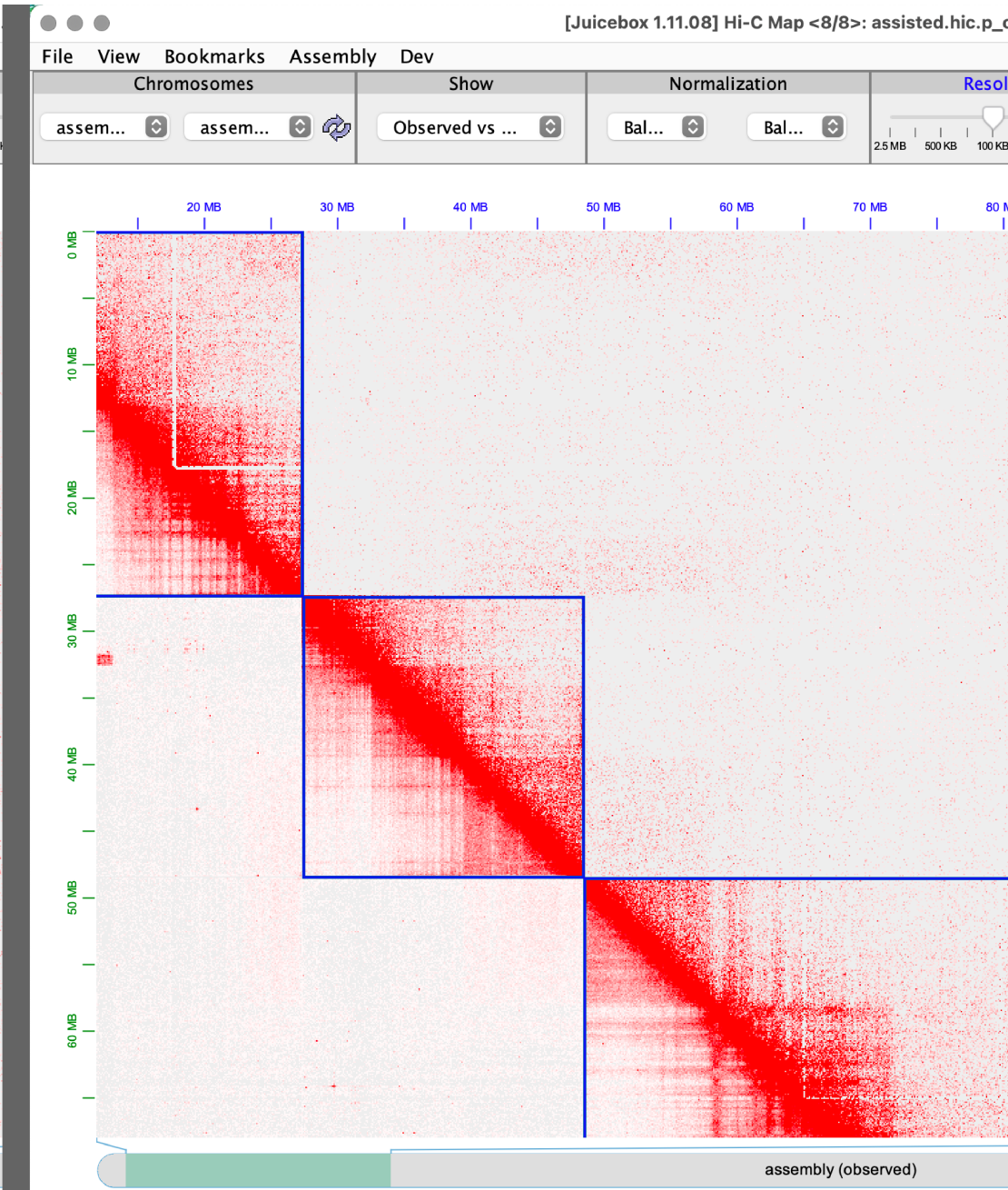
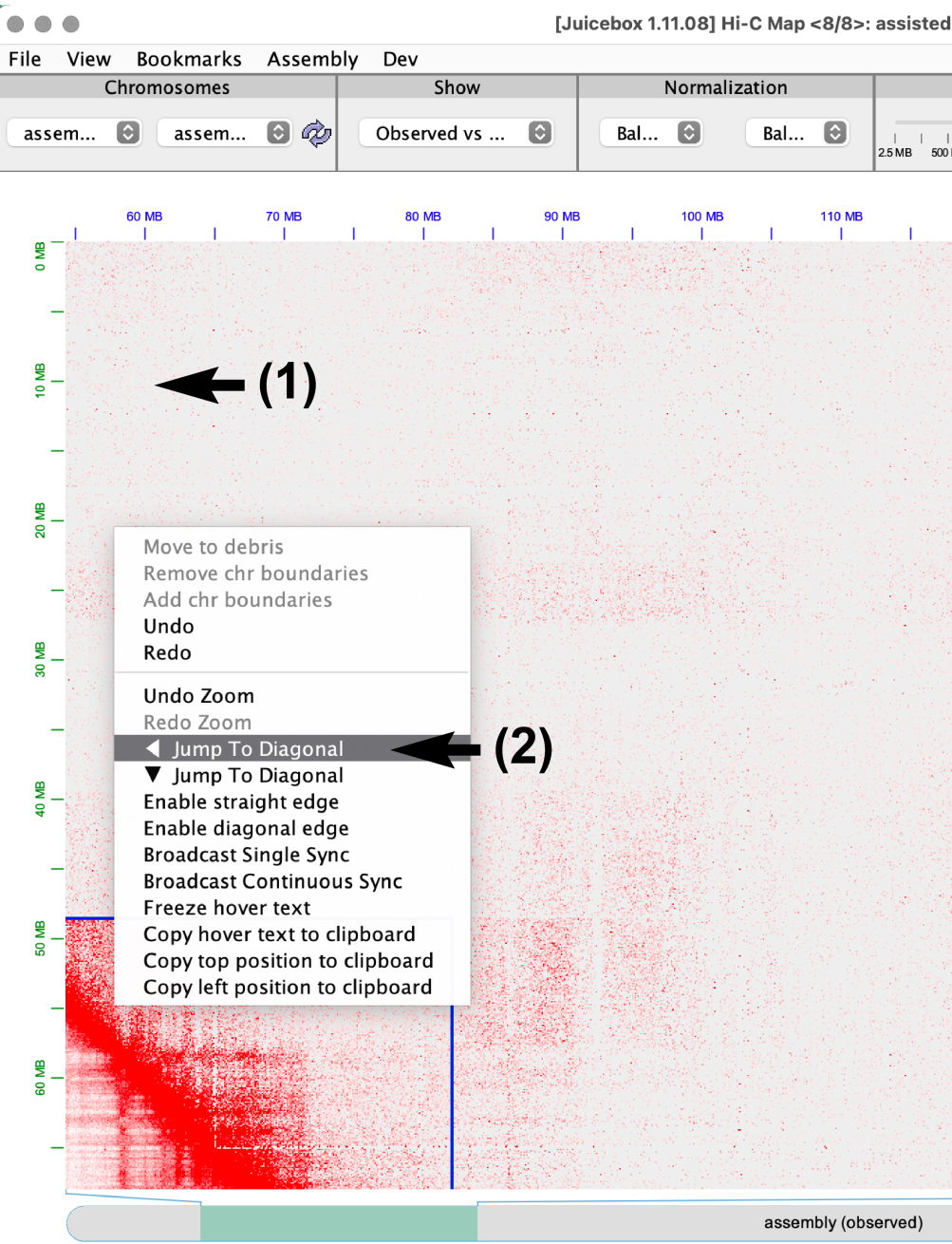


Navigating the contact map

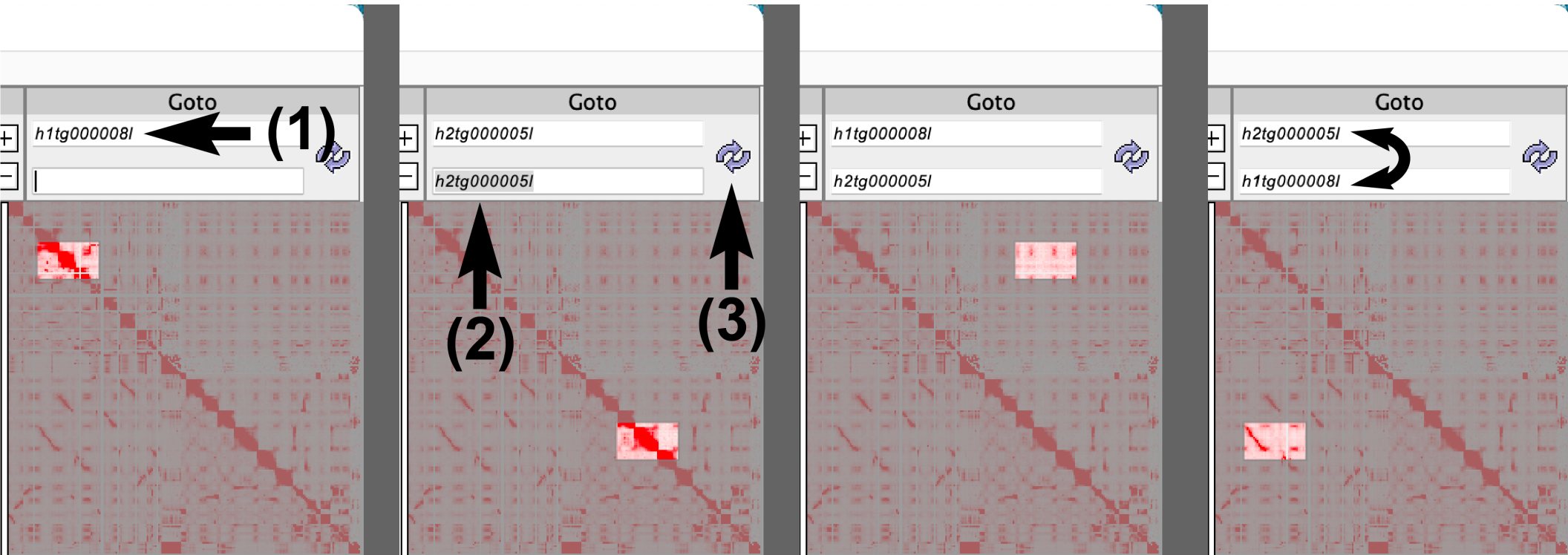
[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Navigating with 'Jump to Diagonal'



Searching contigs with 'Goto'



Displaying coverage tracks

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View **(1)** Assembly Dev

Change Heatmap Color **(2)**
Darkula Mode
Make Custom Chromosome (from .bed)...

Axis Endpoints Only
Chromosome Context
Grid Lines
Export PDF Figure...
Export SVG Figure...

Annotations Layer Panel

Available Features

- Dataset-specific 1D Features
- Coverage normalizations
 - Coverage
 - Coverage (Sqrt)
- Balanced **(5)**
- Eigenvector

(4) → **(7)** →

(6) → OK Cancel

Load Basic Annotations... Add Local... Load ENCODE Tracks... Load from URL... Refresh View

assembly (observed)

(1&2) → Show Annotation Panel

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View Bookmarks Assembly Dev

Chromosomes Show Normalization Resolution (BP) Color Range Goto

assem... assem... Observed vs ... Bal... Bal...

Balanced
Balanced (Contr...

280 MB 290 MB 300 MB 310 MB 320 MB 330 MB 340 MB 350 MB 360 MB 370 MB 380 MB

280 MB
290 MB
300 MB

h1tg0000081
h2tg0000051

Balanced
378,500,000-378,600,000
bin: 3,785
value: 0.975

Exporting a modified (reviewed) .assembly file

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View Bookmarks **Assembly** ← (1)

- Import Map Assembly
- Import Modified Assembly
- Export Assembly** ← (2)
- Reset Assembly
- Set Scale
- Exit Assembly

Normalization: None None

Resolution (BP): 2.5 MB 500 KB 100 KB 25 KB 5 KB

Color Range: 0 382 772

Goto: [] []

Save

Save As: assisted.hic.p_ctg.review ← (3)

Downloads

Name Date Modified

File Format: All Files

New Folder Cancel **Save** ← (4)

assembly

assembly:988,000,001-989,000,000
assembly:343,000,001-344,000,000
observed value (O) = 0.0
expected value (E) = 121.501
O/E = 0

control value (C) = 0.0
expected control value (EC) = 1.437
C/EC = 0

average observed value (AVG) = 225.667
O' = O/AVG = 0.0
average control value (AVGC) = 30.076
C' = C/AVGC = 0.0

Edit Chr Scaf

Show Annotation Panel

Exporting a PDF/SVG file

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

File View **(1)** Assembly Dev

- Show Annotation Panel
- Change Heatmap Color
- Darkula Mode
- Make Custom Chromosome (from .bed)...

Axis Endpoints Only
✓ Chromosome Context
✓ Gridlines

(2) Export PDF Figure...
Export SVG Figure...

Normalization Bal... Bal... Resolution (BP) Color Range Goto

Width 1440 Height 900
Save As: 2022.03.16.13.37.09.HiCImage.pdf
Downloads

Name Date Modified

File Format: All Files

New Folder Cancel **(4)** Save

assembly (observed)

assembly (control)

assembly:16,700,001-16,800,000
assembly:1,100,001-1,200,000
observed value (O) = 4.025
expected value (E) = 2.533
O/E = 1.589

control value (C) = 0.0
expected control value (EC) = 0.246
C/EC = 0

average observed value (AVG) = 2.42
O' = O/AVG = 1.663
average control value (AVGC) = 0.33
C' = C/AVGC = 0.0

Feature
assembly:1-27,403,807
assembly:1-27,403,807
Scaffold # = 1
Scaffold name = h1tg0000011
Signed scaffold # = 1

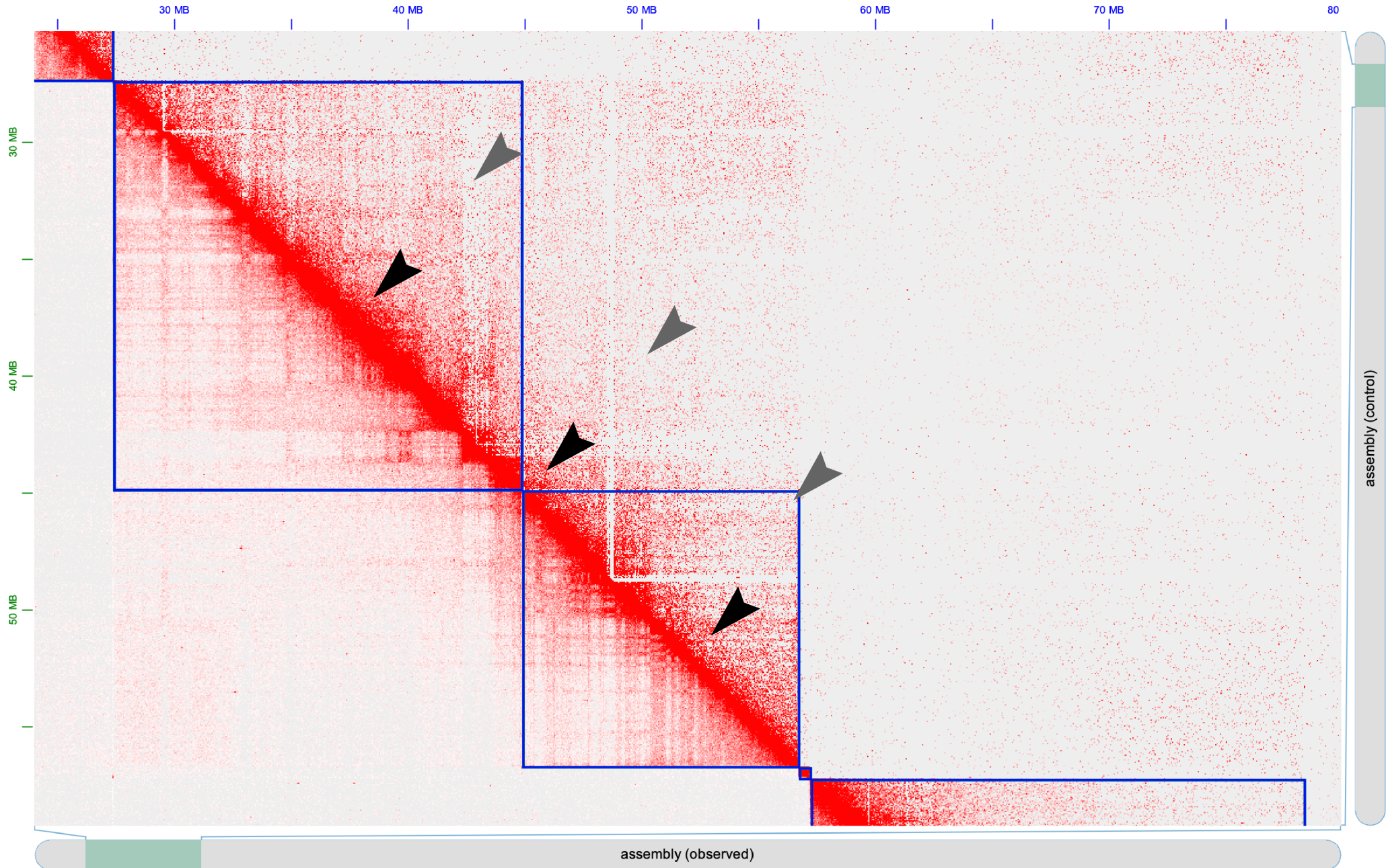
Feature
assembly:1-27,403,807
assembly:1-27,403,807
Superscaffold # = 1

Edit Chr Scaf

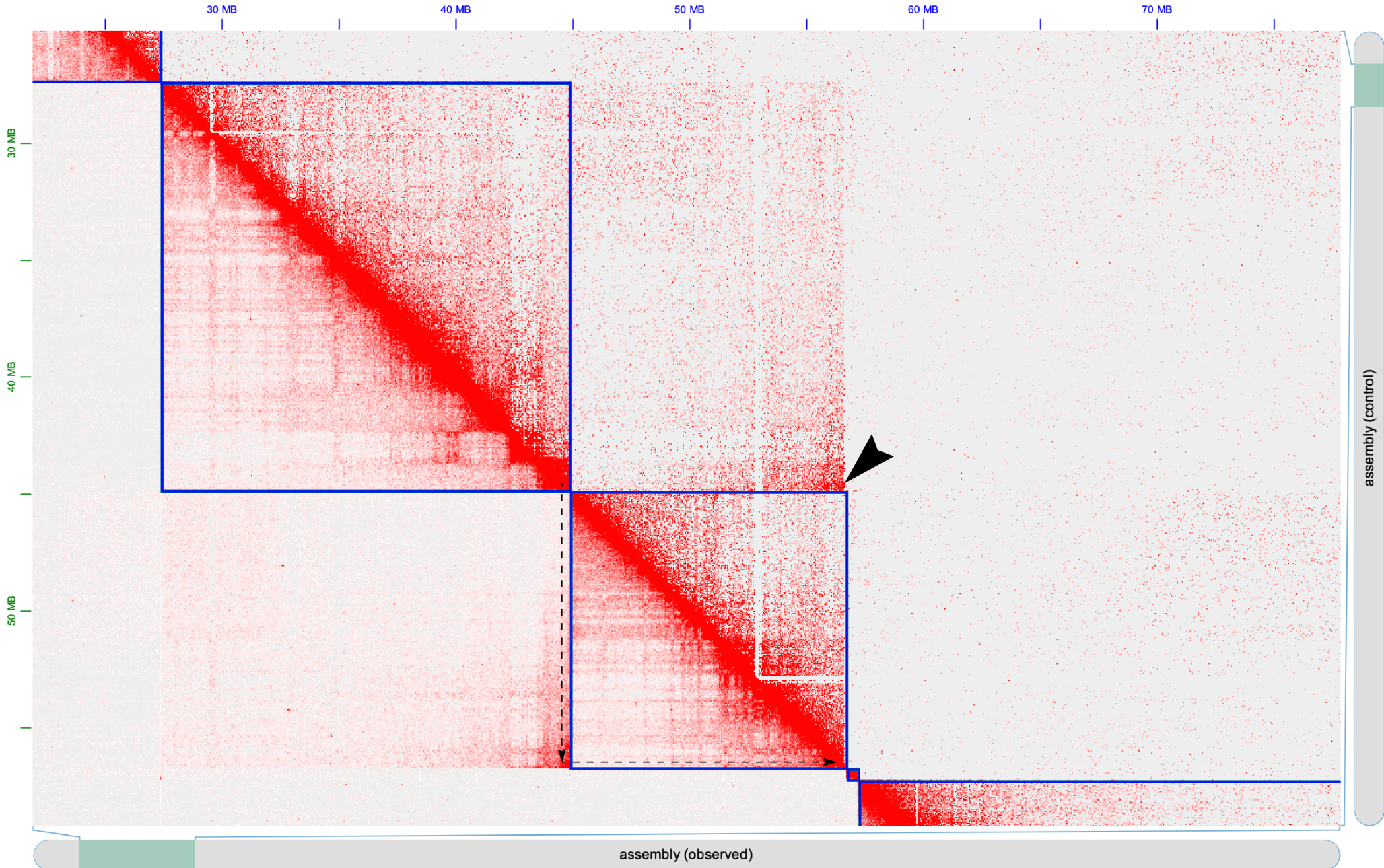
Show Annotation Panel

Common Hi-C contact patterns

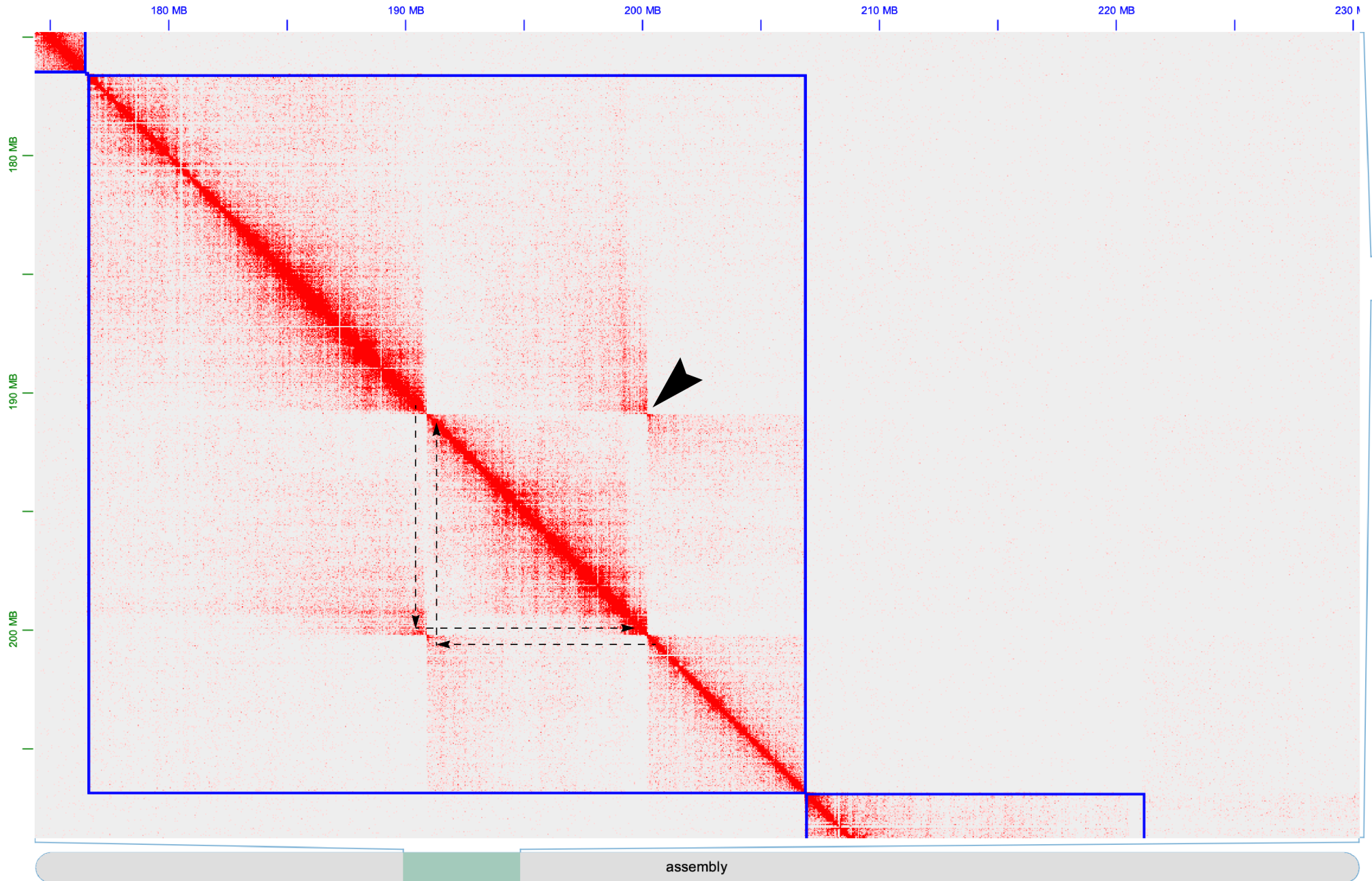
Contigs in correct order and orientation



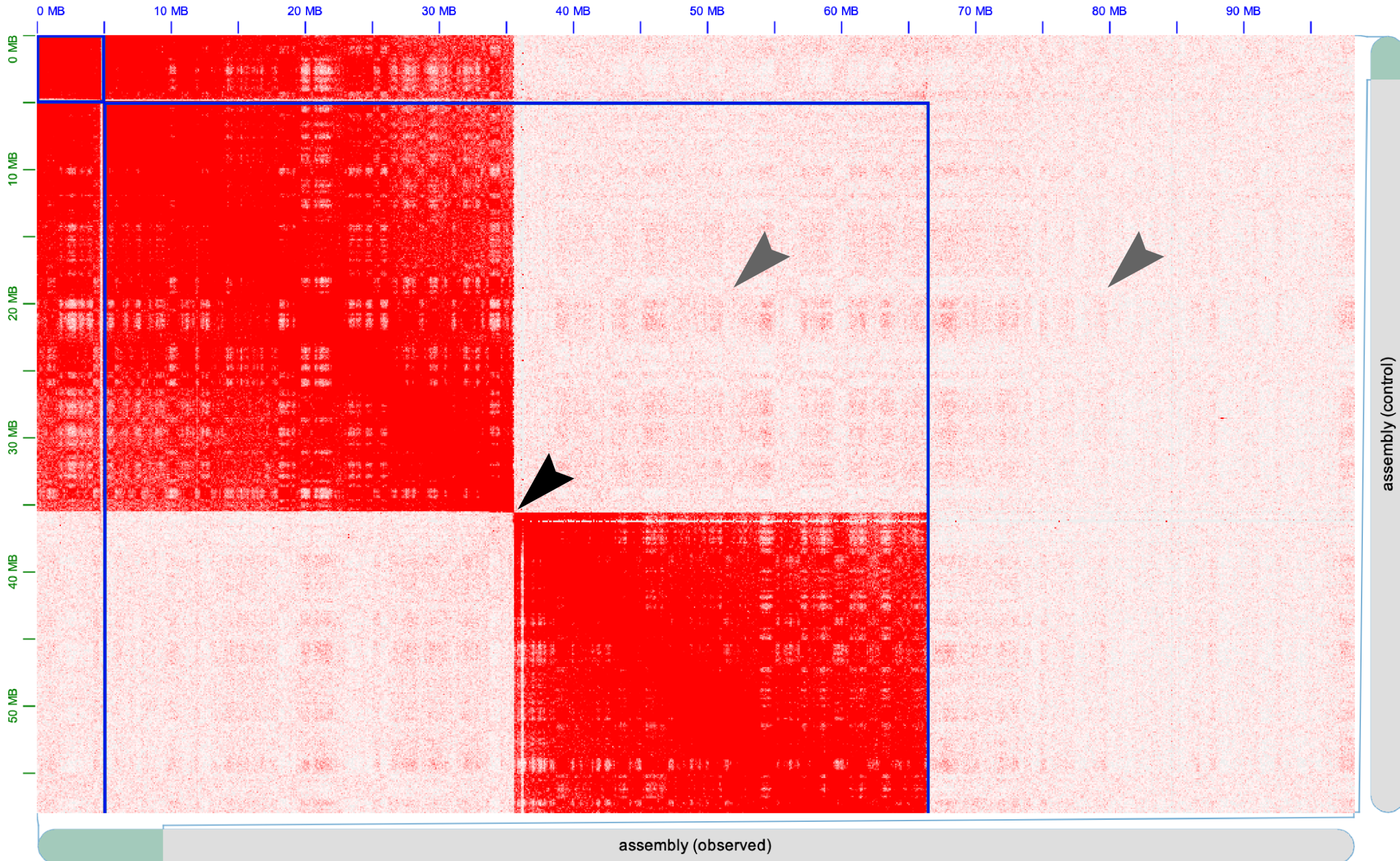
Contig inversion



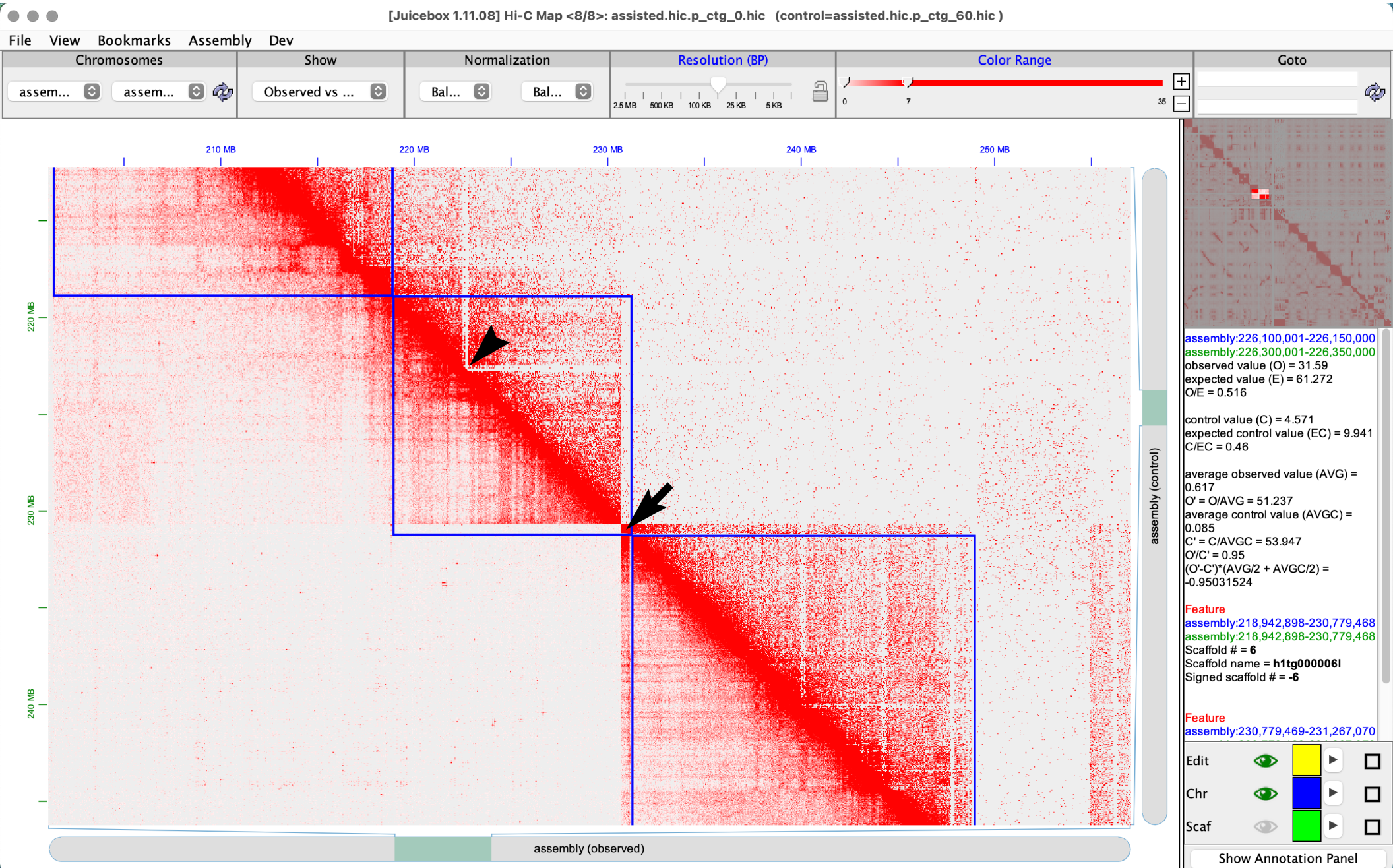
Internal inversion misassembly



Contigging misjoin error

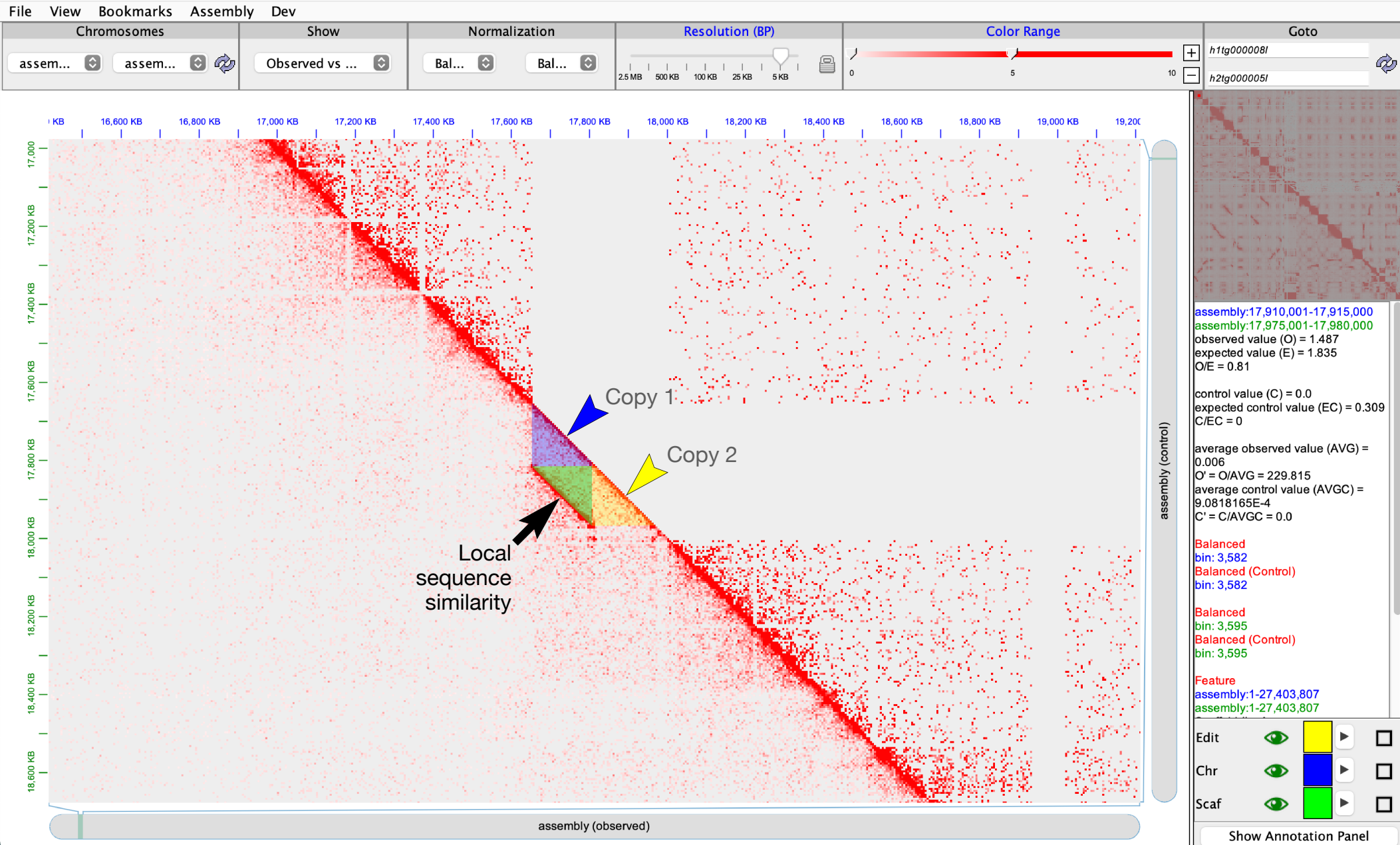


Positive evidence of a contigging misjoin error

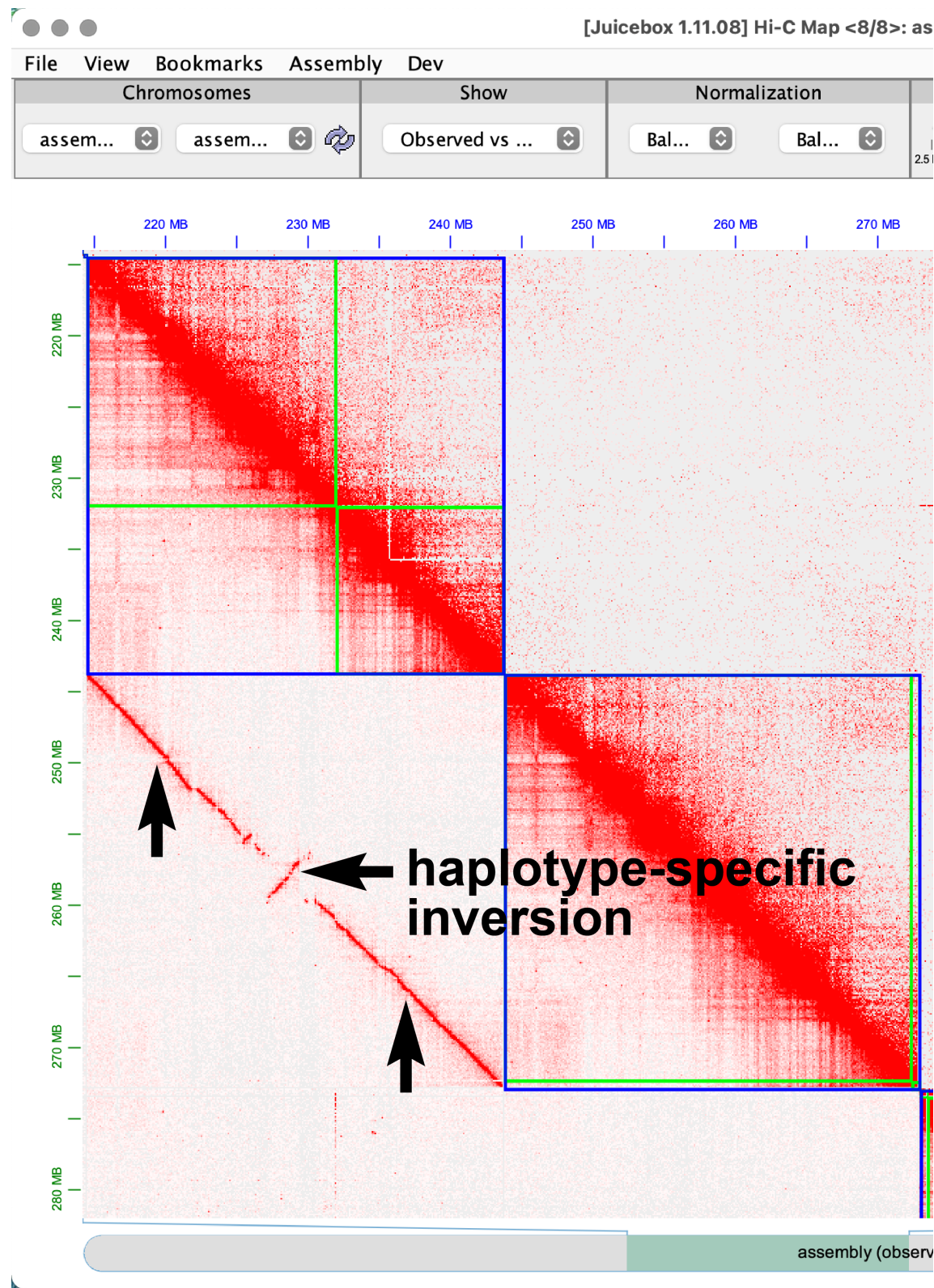


Tandem segmental duplication

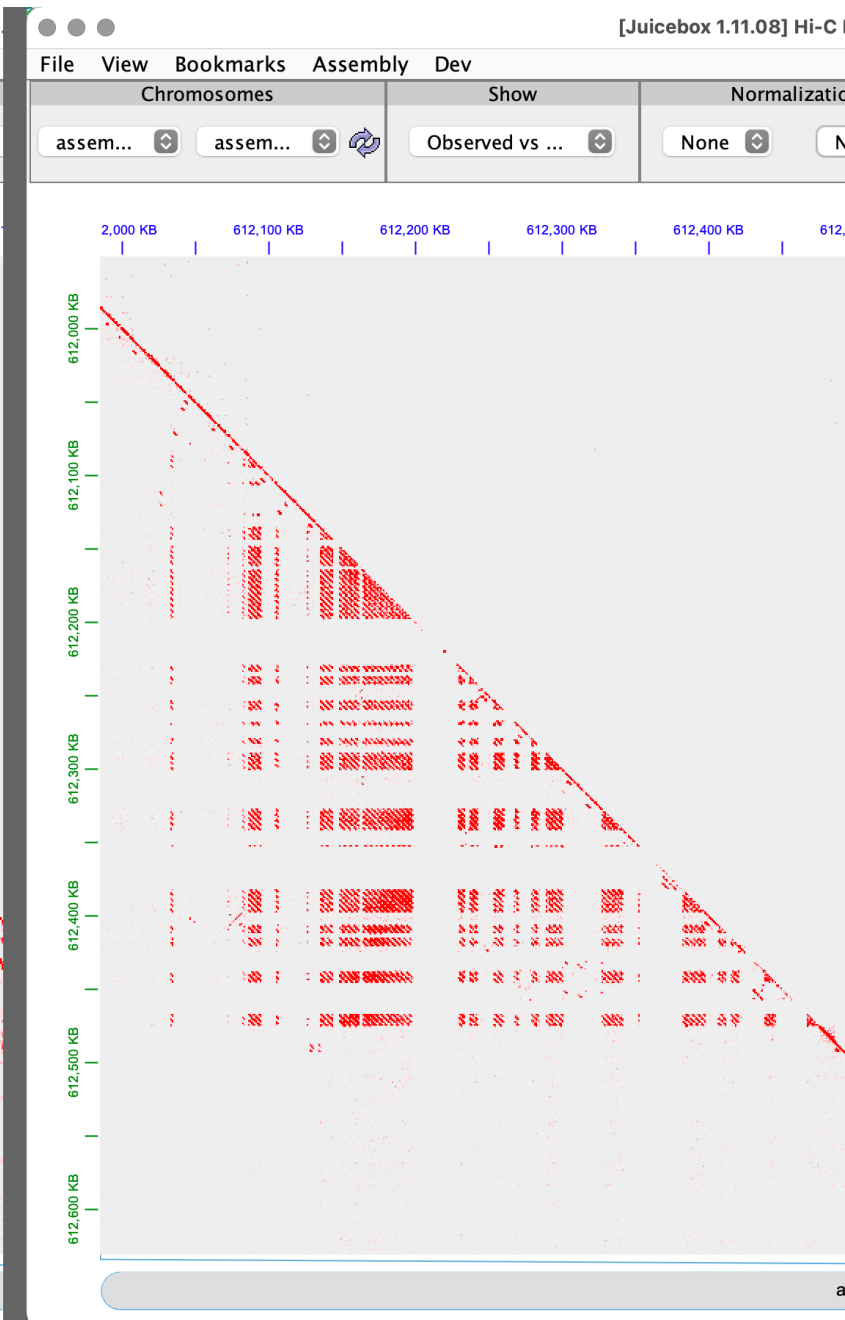
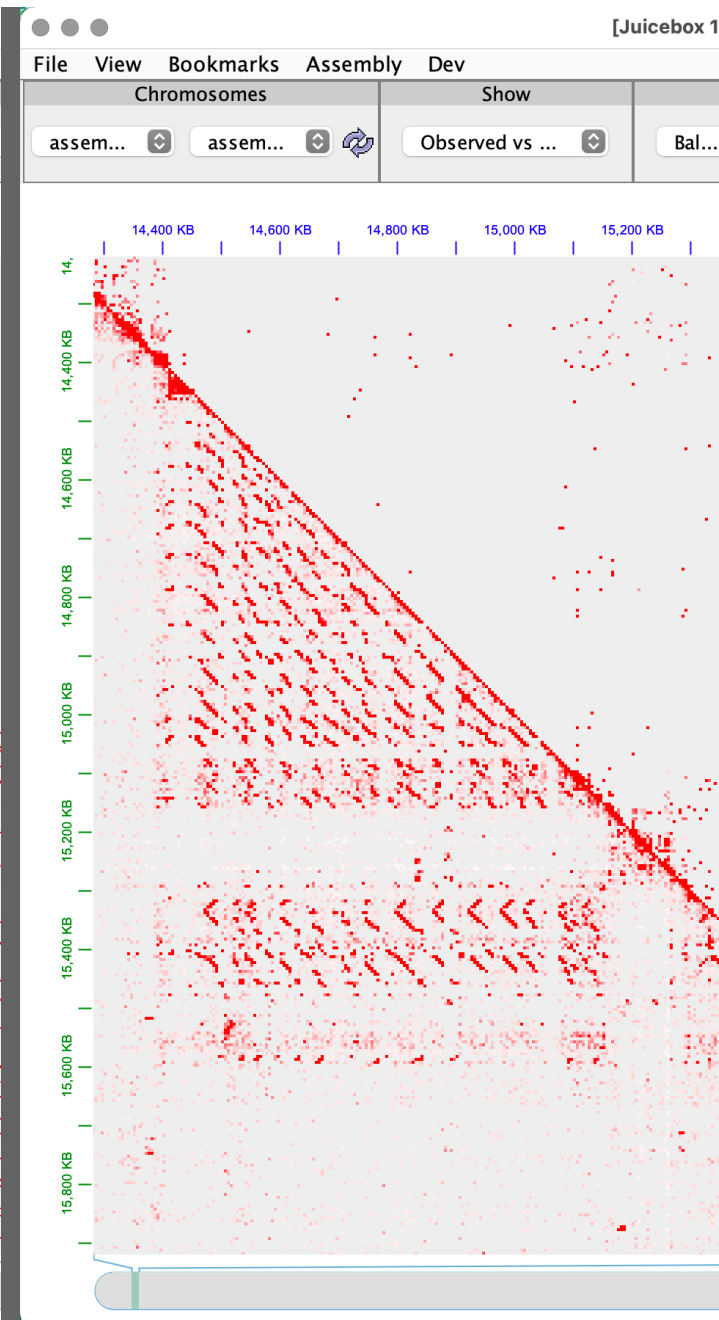
[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Homologous sequences

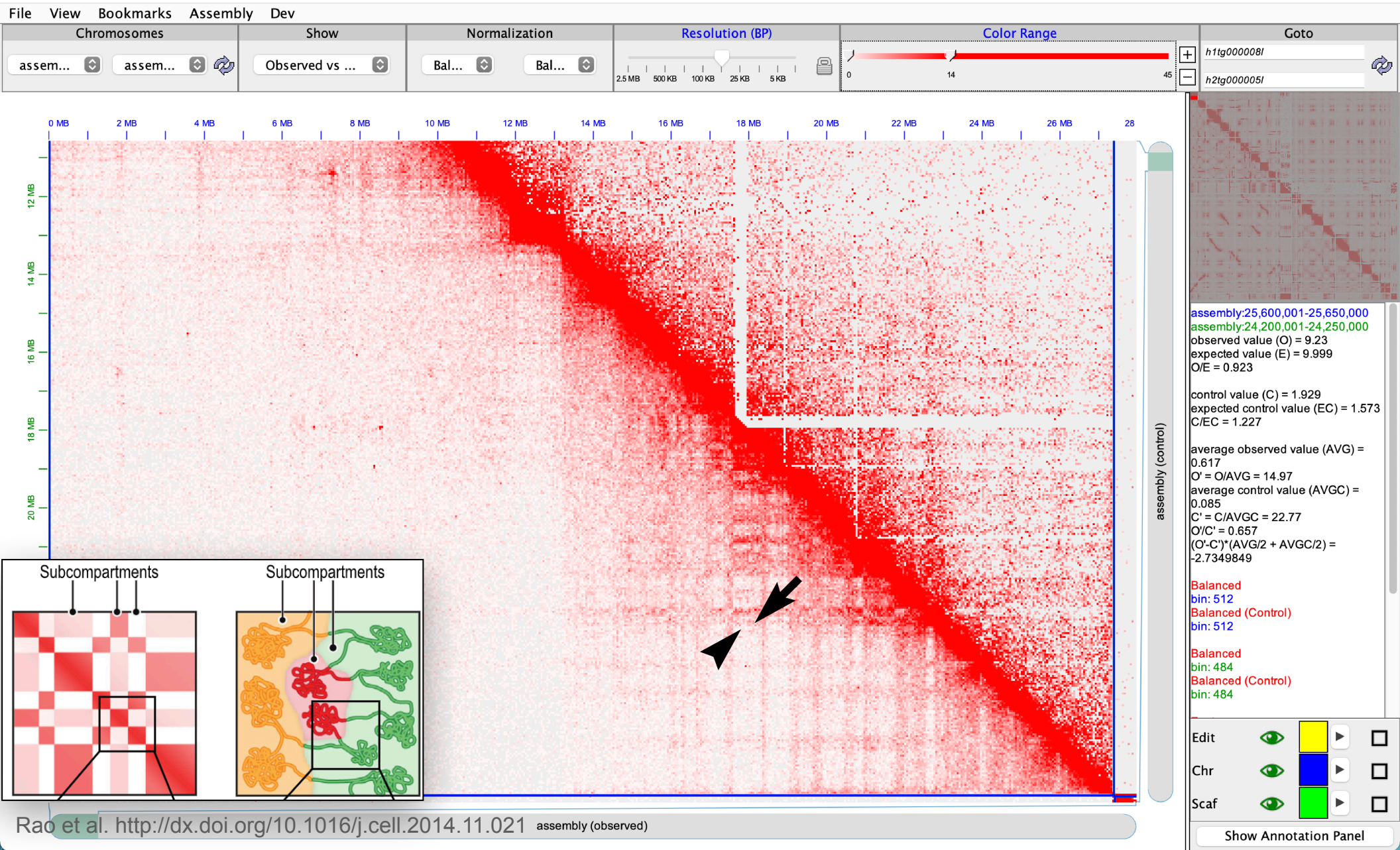


Repetitive sequences

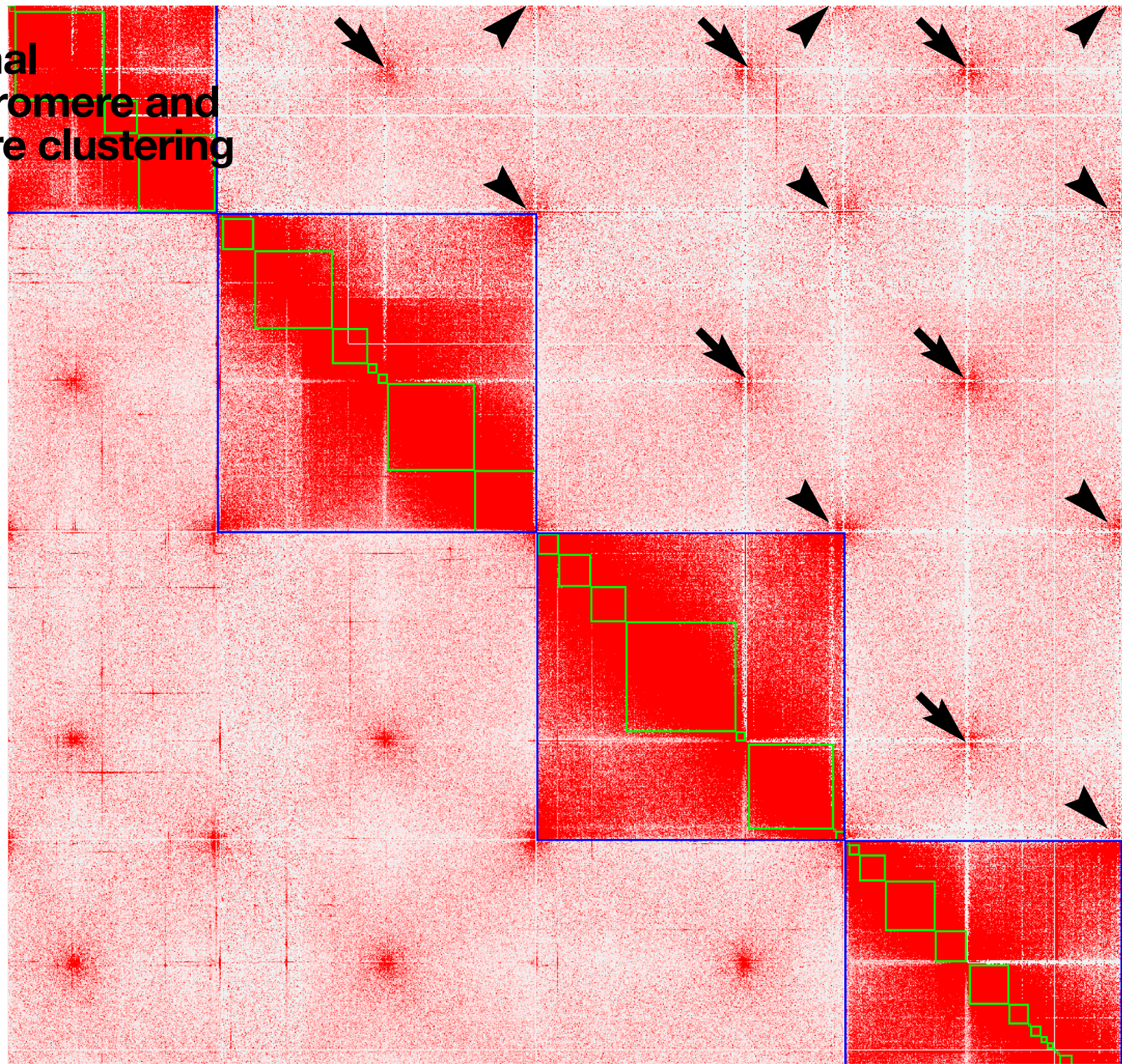


A/B compartmentalization

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

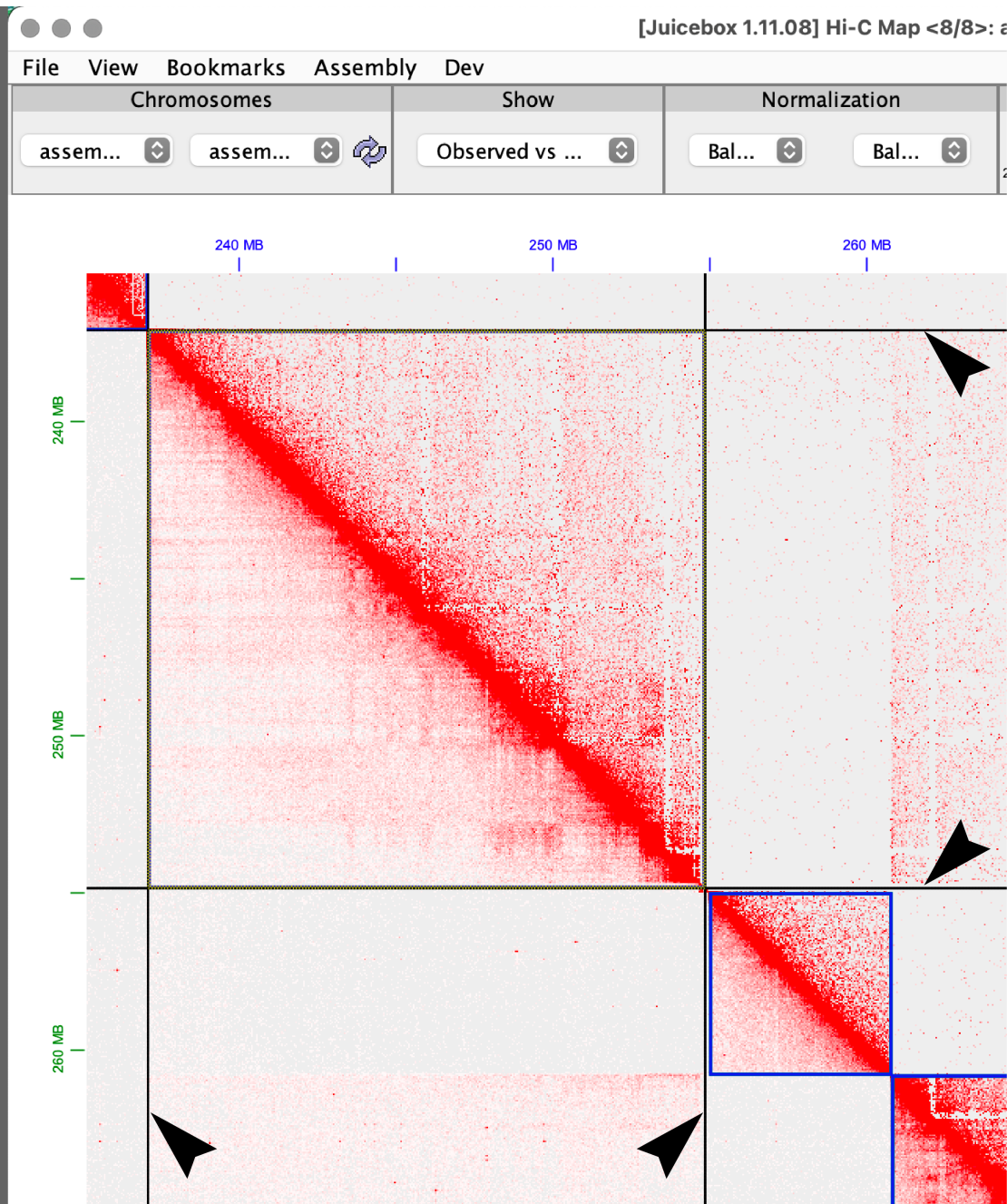
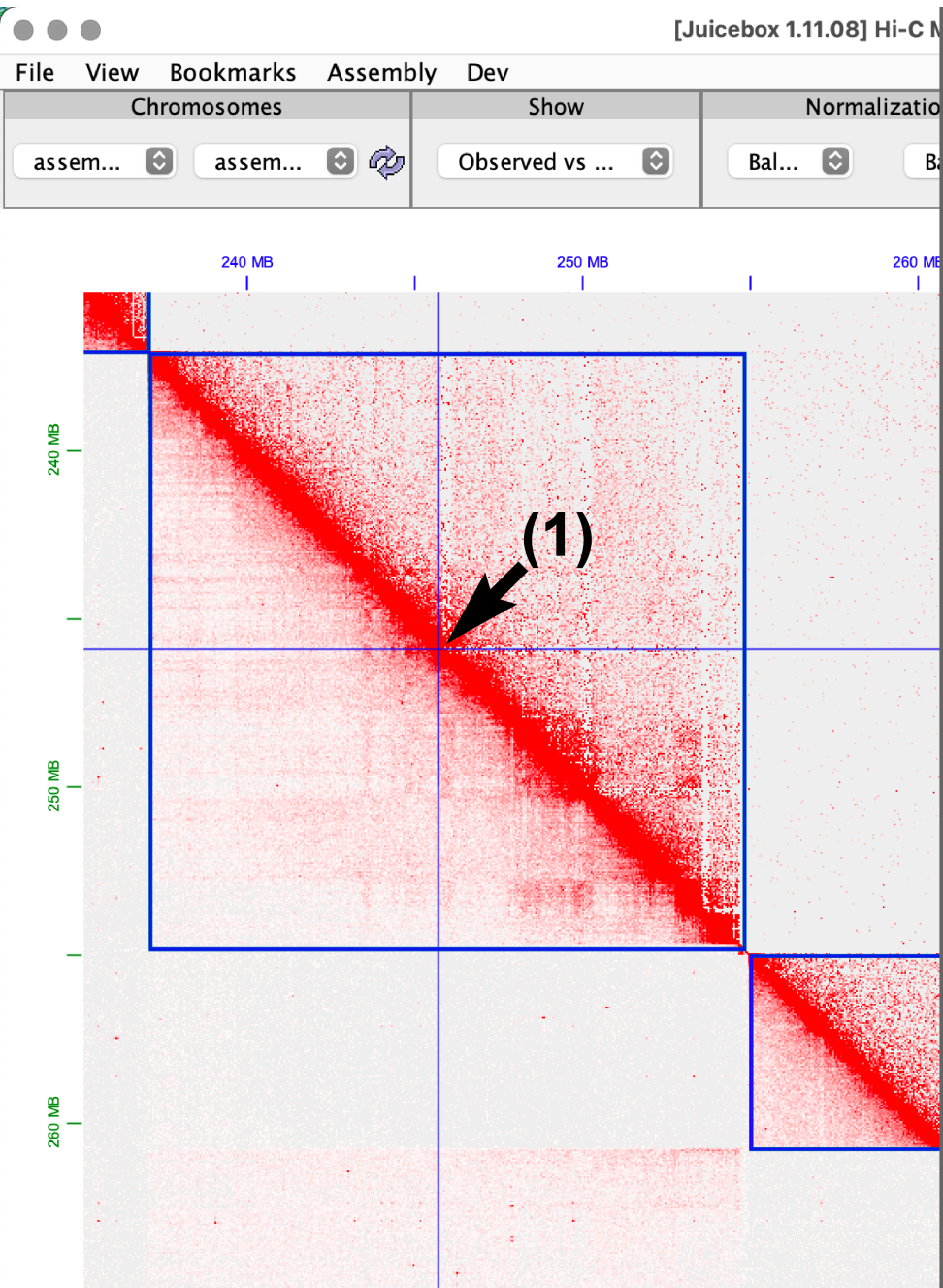


Inter-chromosomal centromere-centromere and telomere-telomere clustering

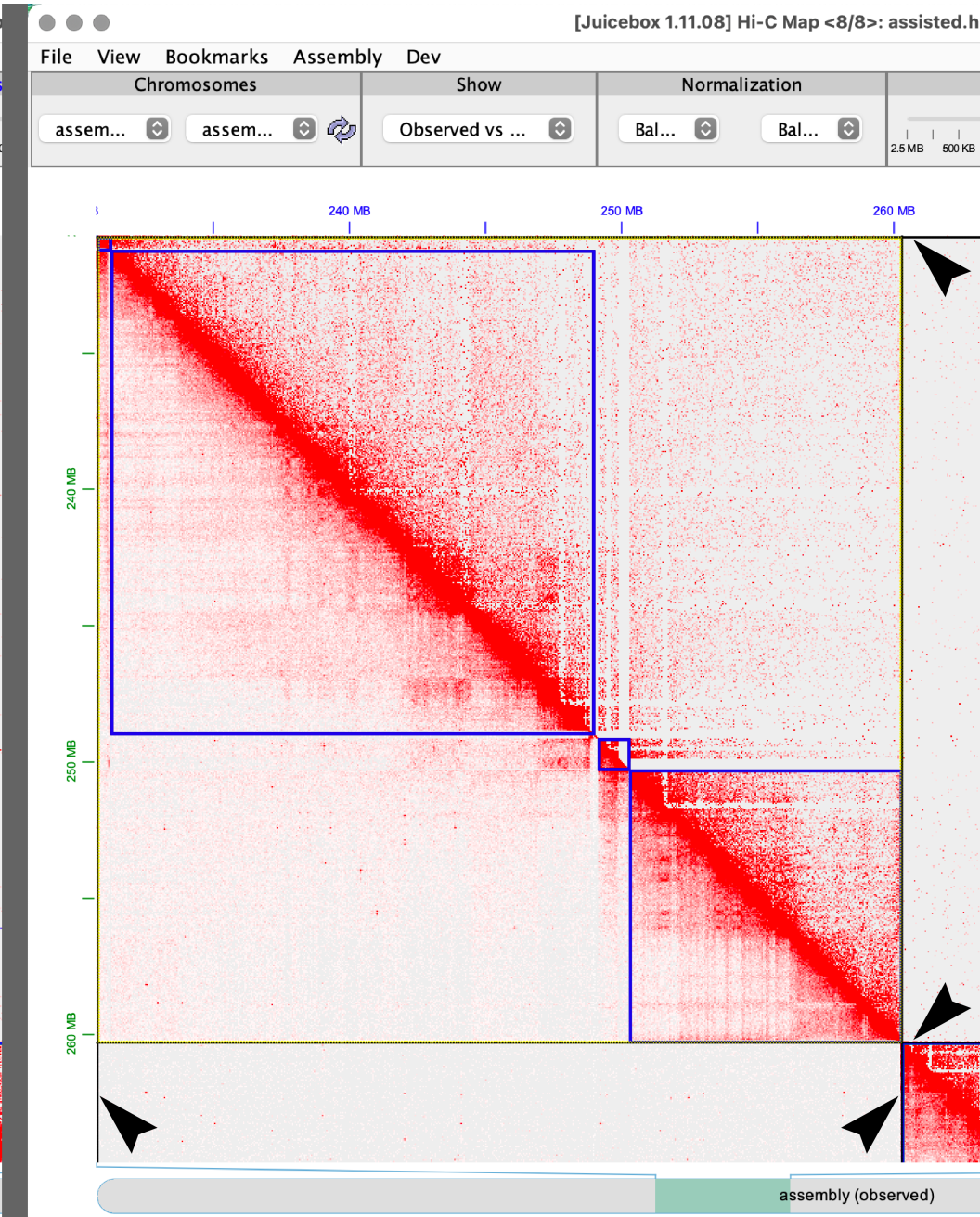
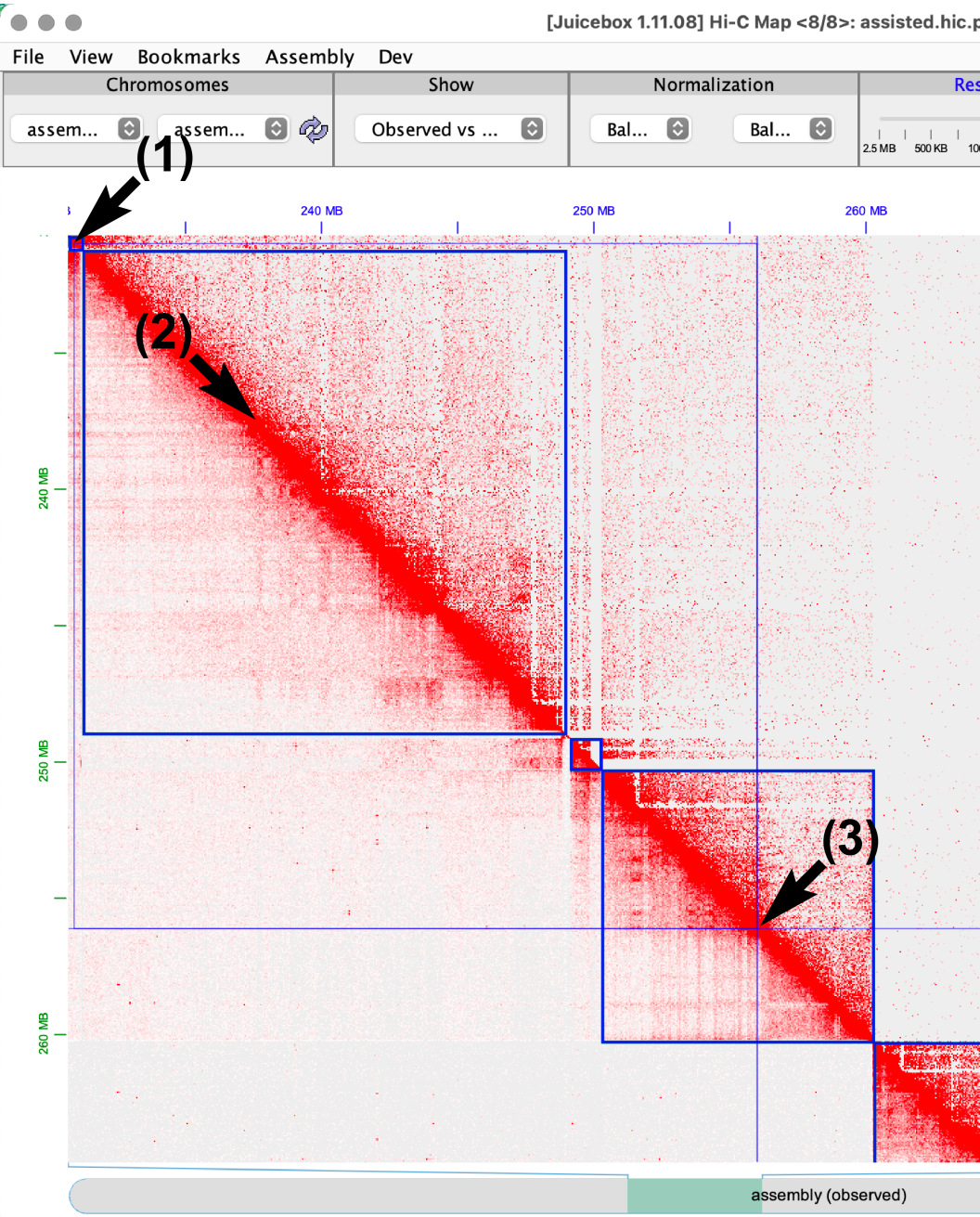


Editing assemblies

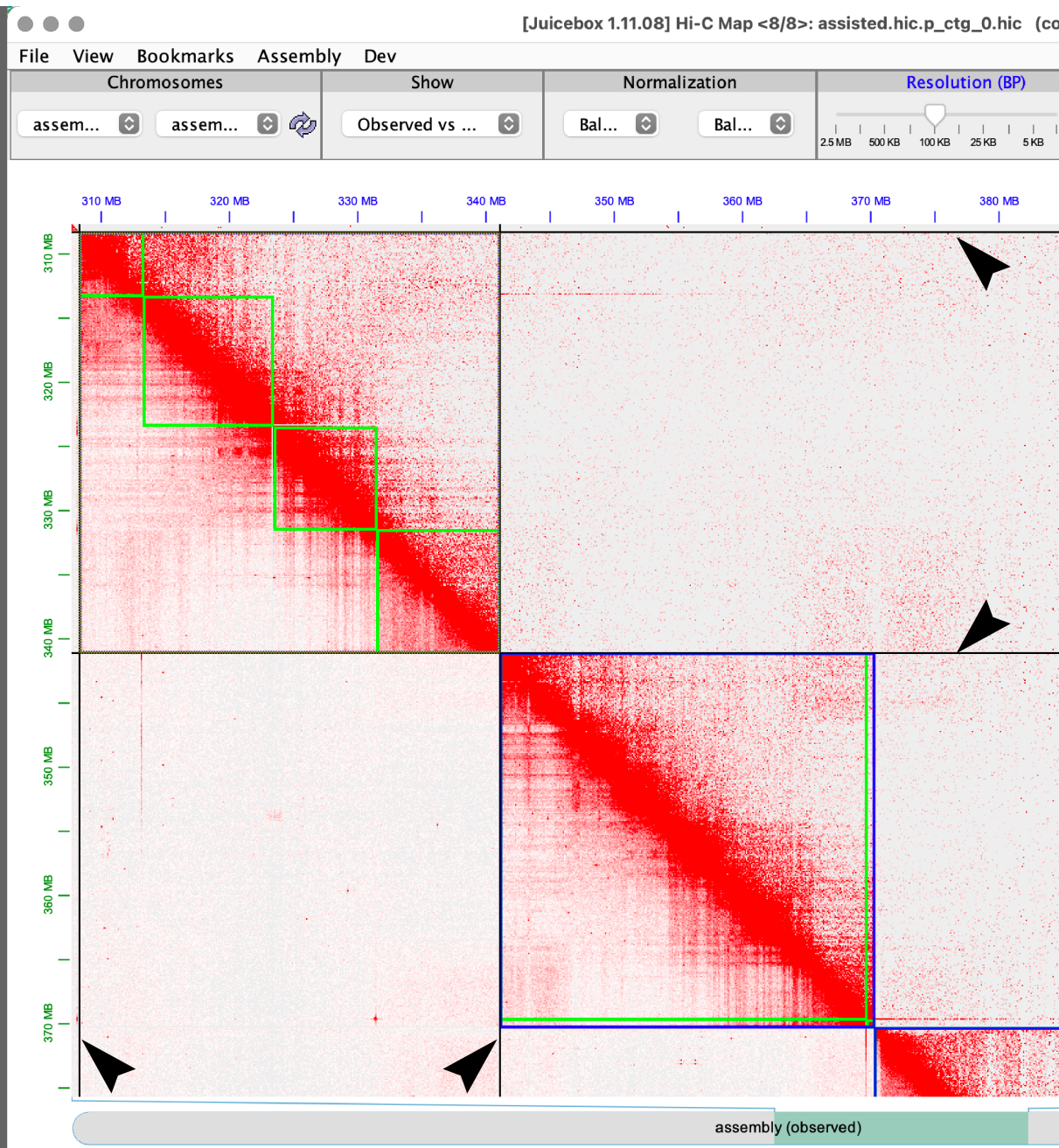
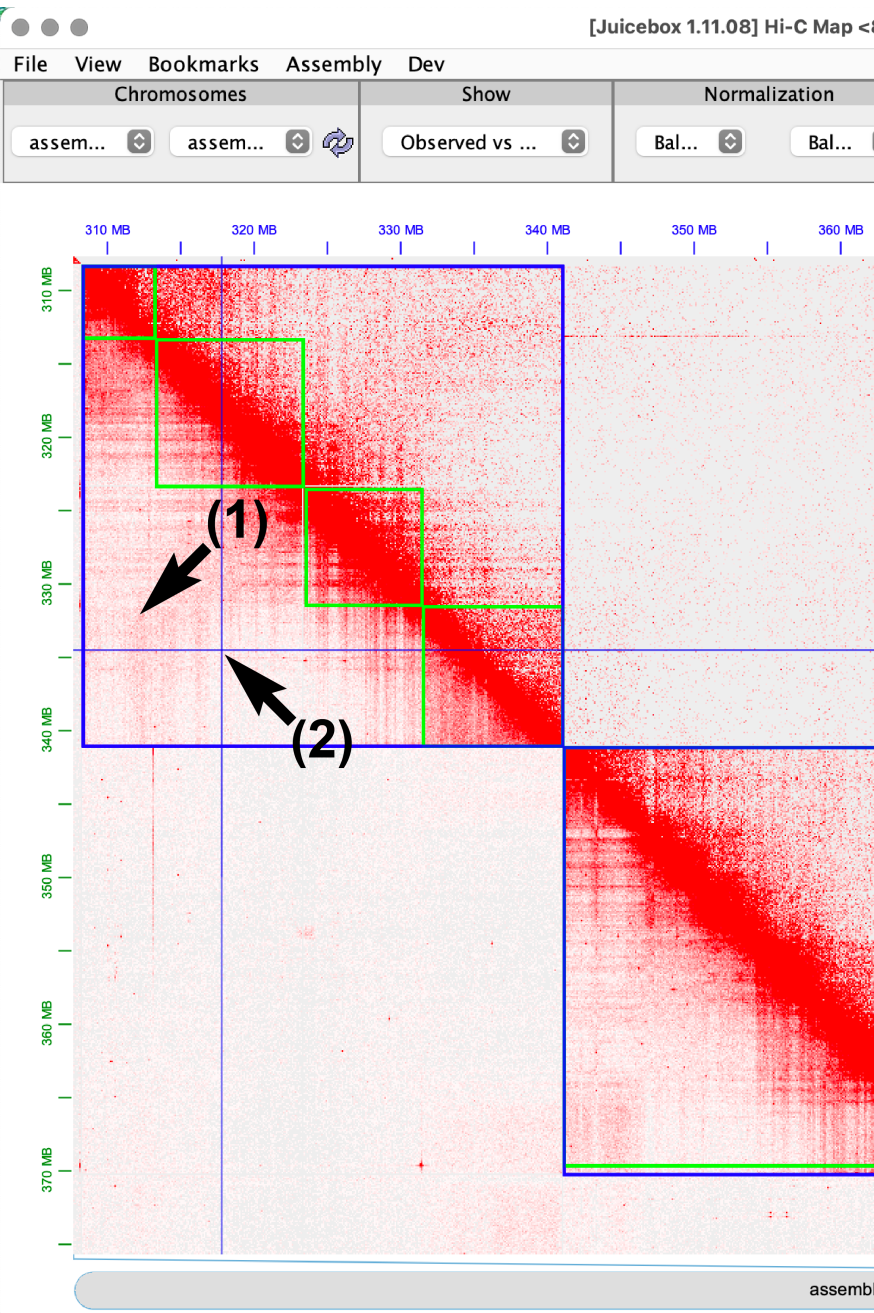
Selecting a single contig



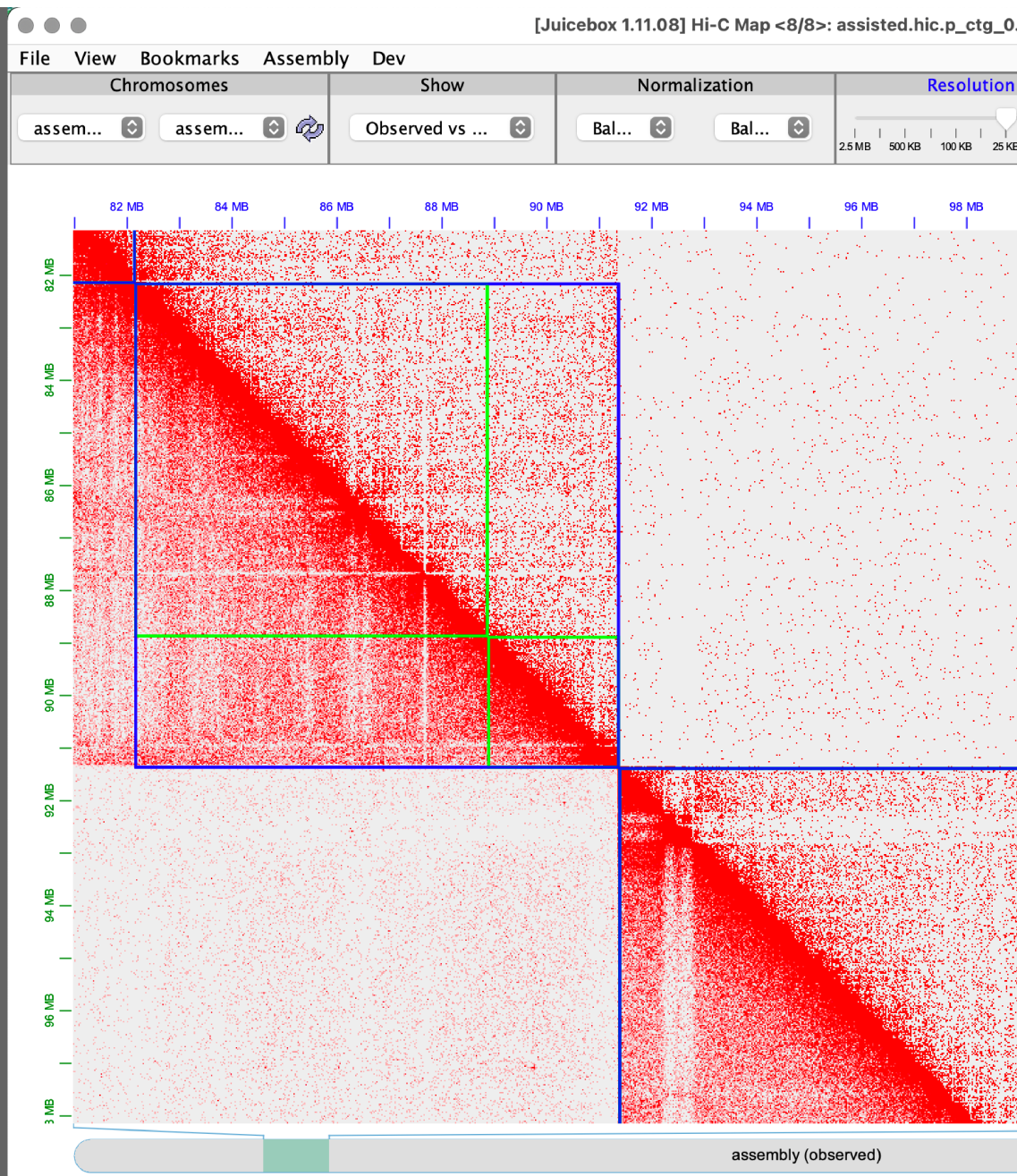
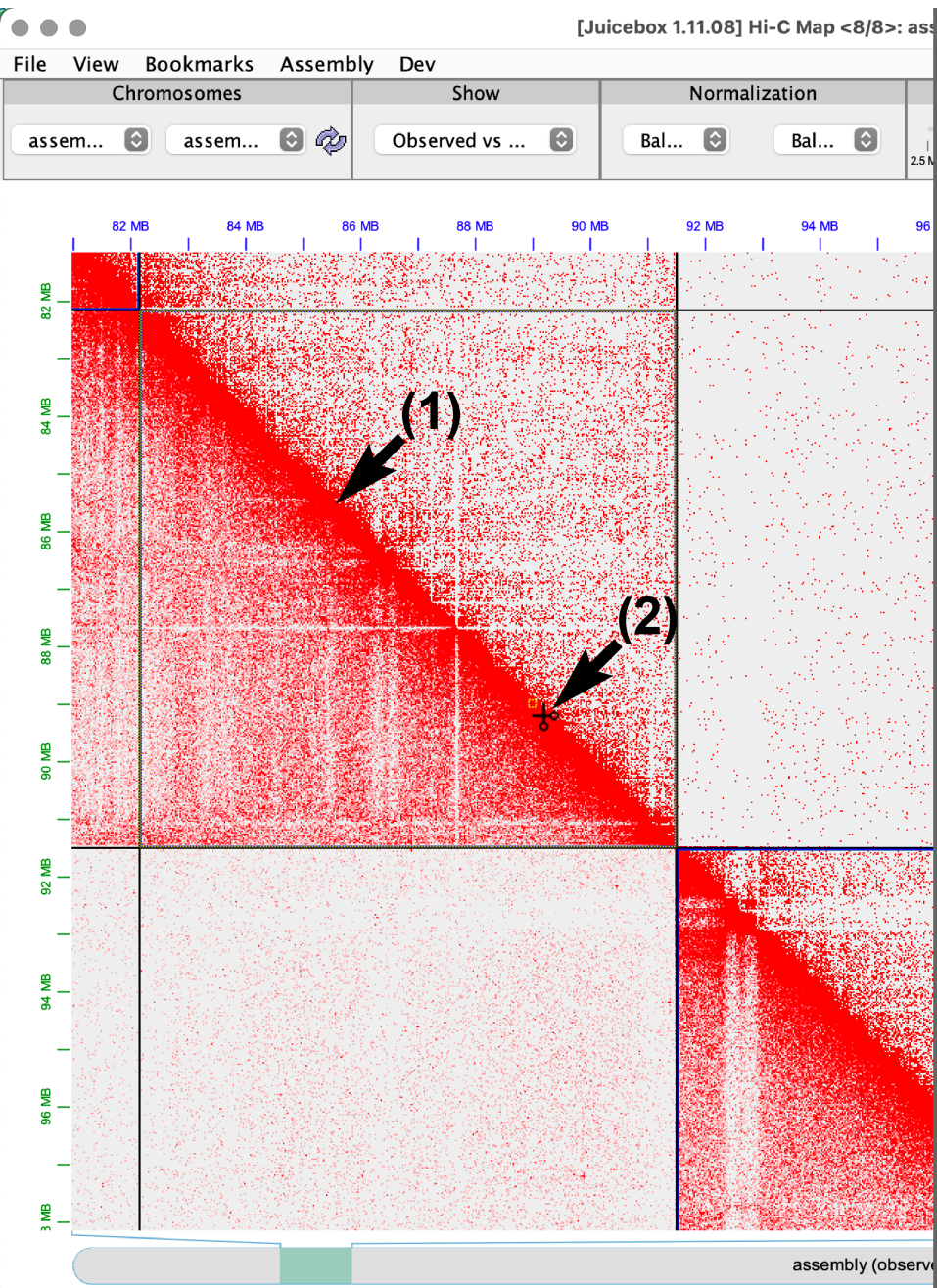
Selecting multiple contigs



Selecting all contigs in a scaffold

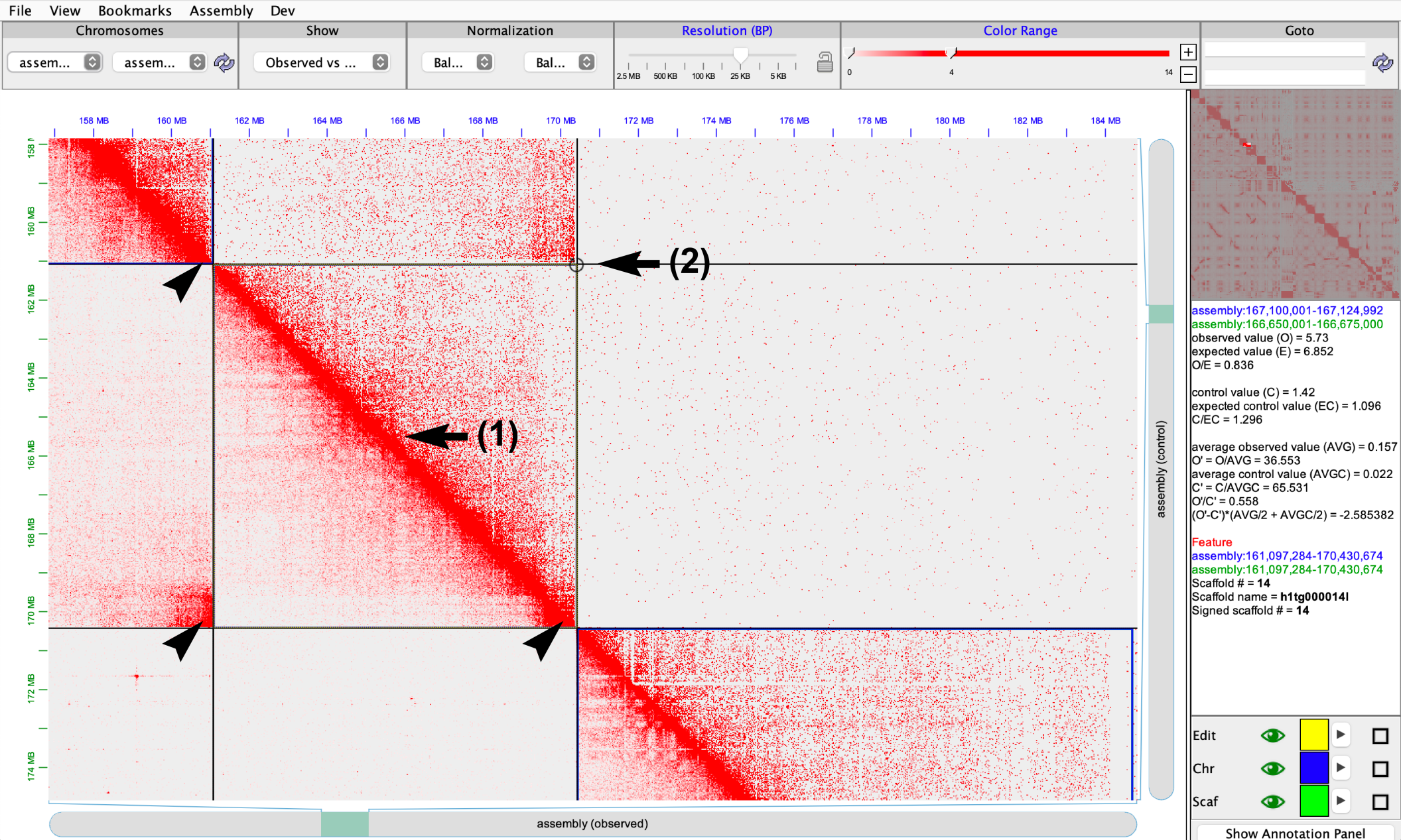


Cutting/breaking a contig



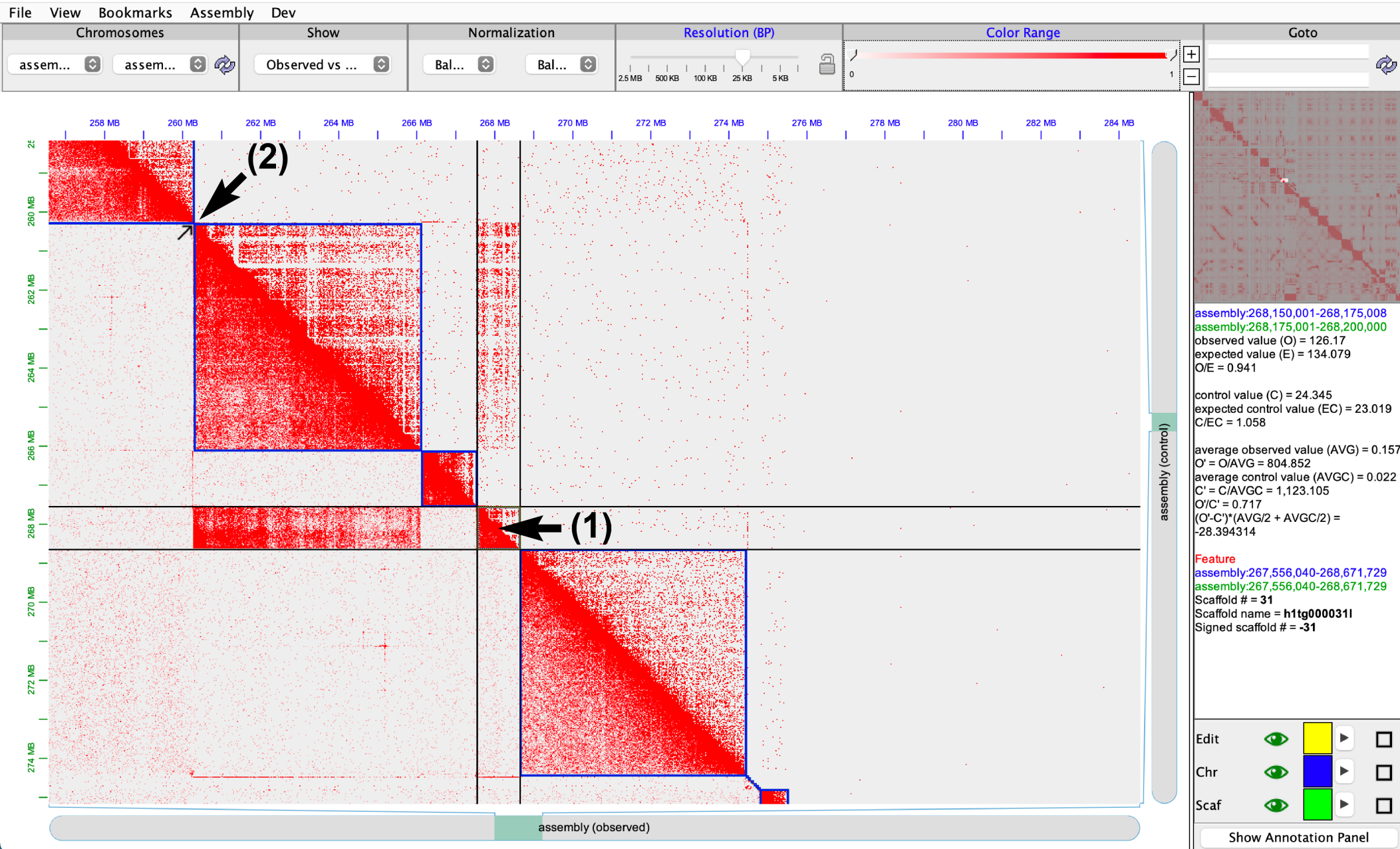
Inverting contig orientation

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Moving misplaced contigs

[Juicebox 1.11.08] Hi-C Map <8/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)



Moving misplaced contigs

Using the straight edge tool(s) to precisely identify small contigs from contact signal “stripes”

- Move to debris
- Remove chr boundaries
- Add chr boundaries
- Undo
- Redo

- Undo Zoom
- Redo Zoom

- ◀ Jump To Diagonal
- ▼ Jump To Diagonal

- Enable straight edge**

- Enable diagonal edge

- Broadcast Single Sync

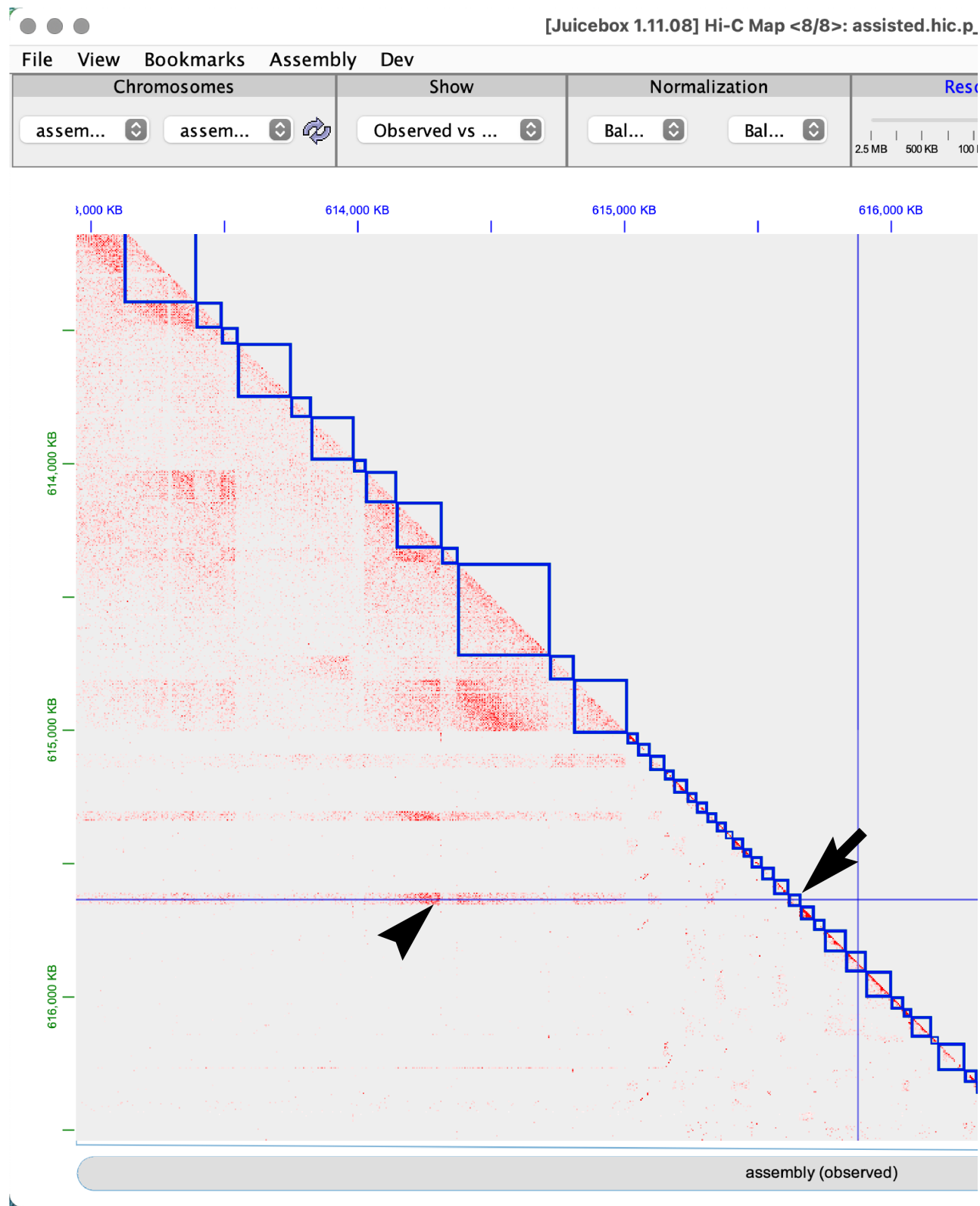
- Broadcast Continuous Sync

- Freeze hover text

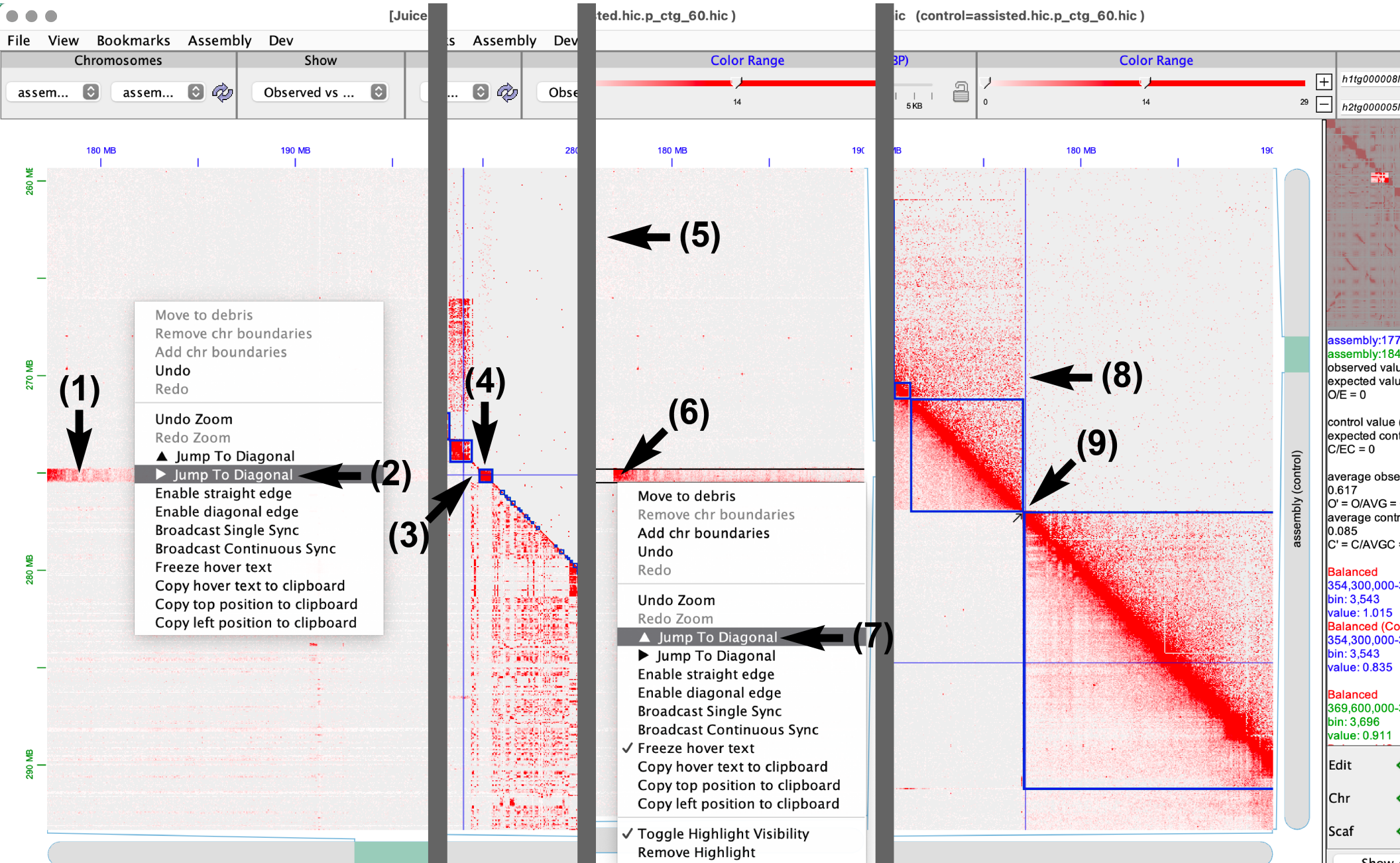
- Copy hover text to clipboard

- Copy top position to clipboard

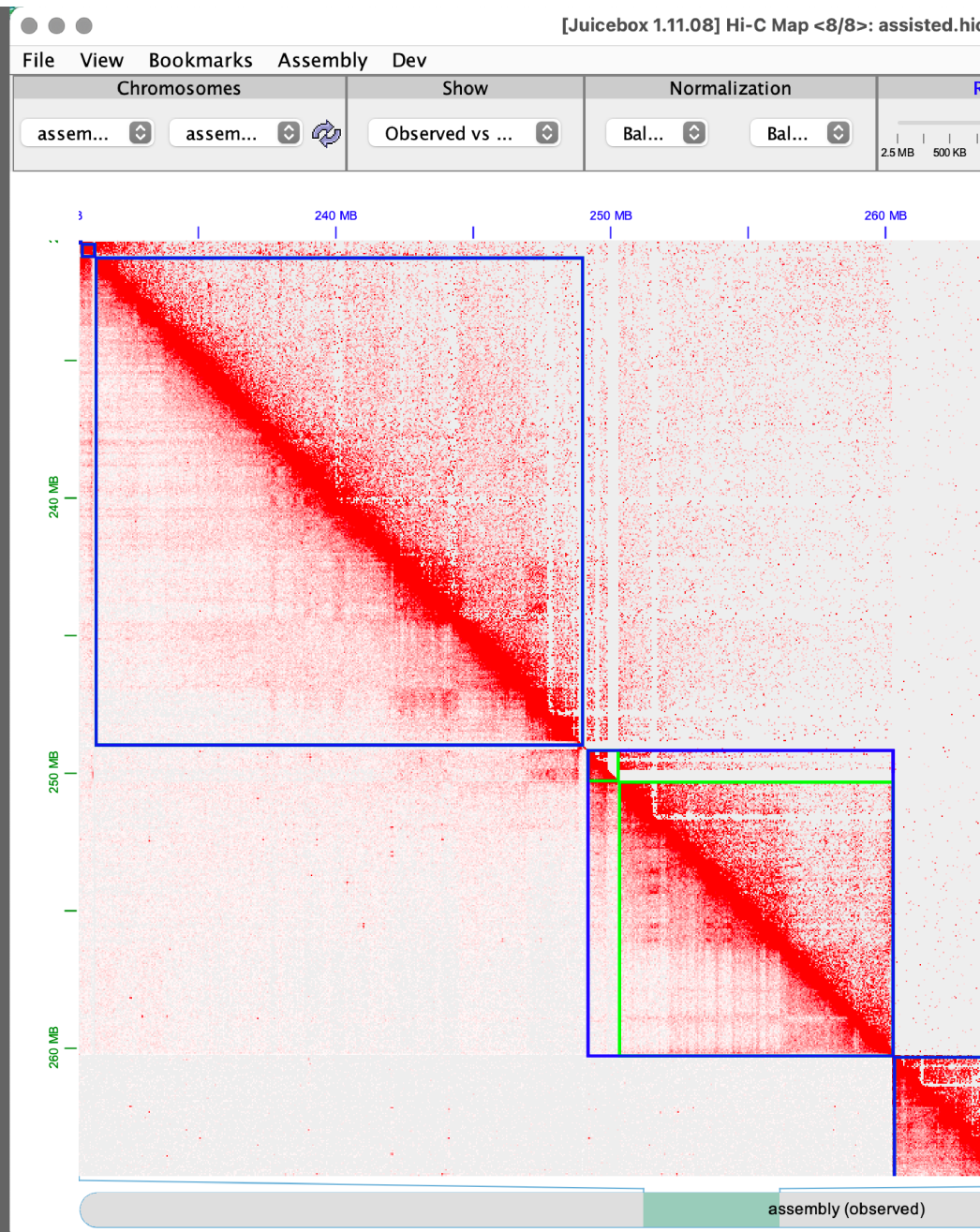
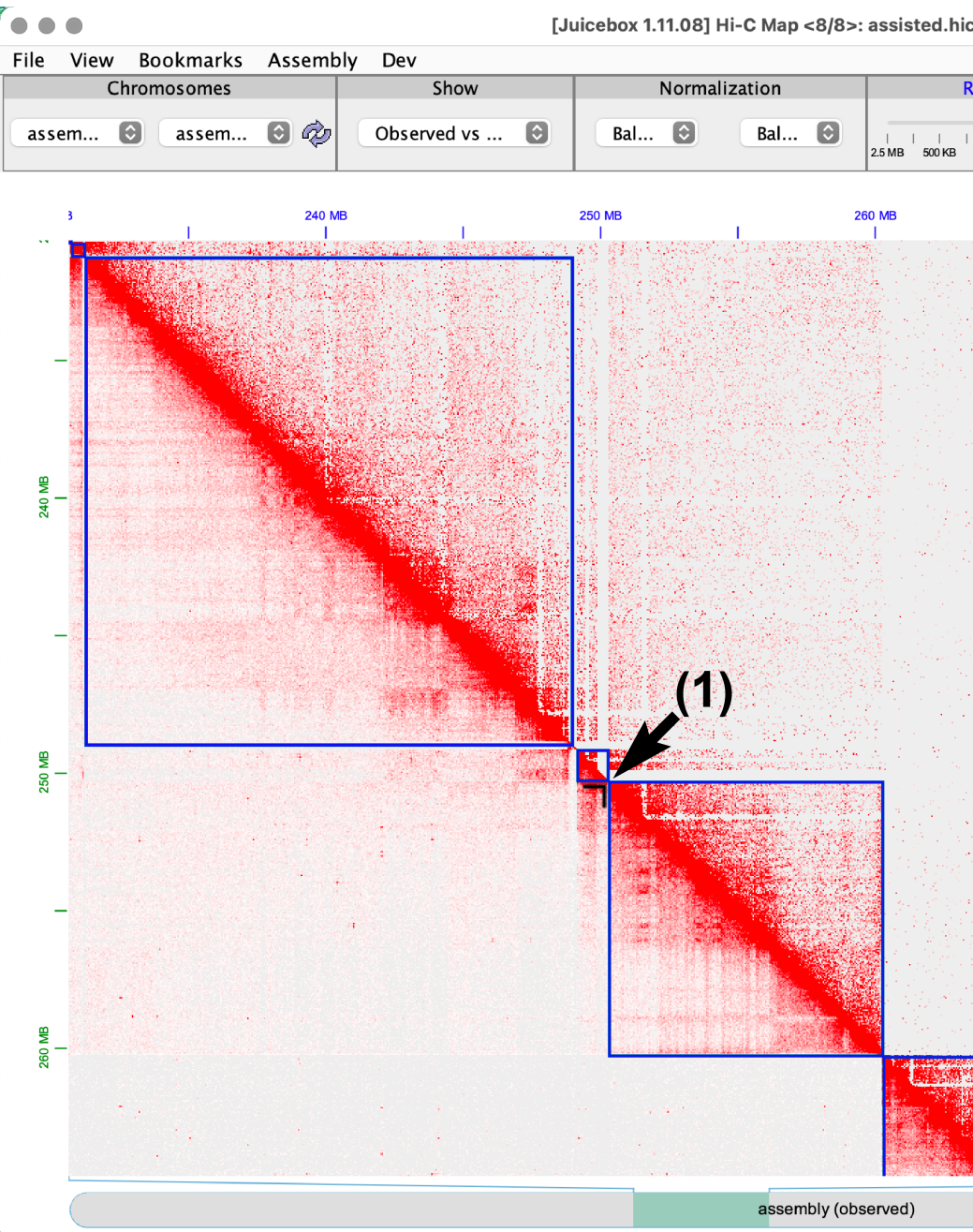
- Copy left position to clipboard



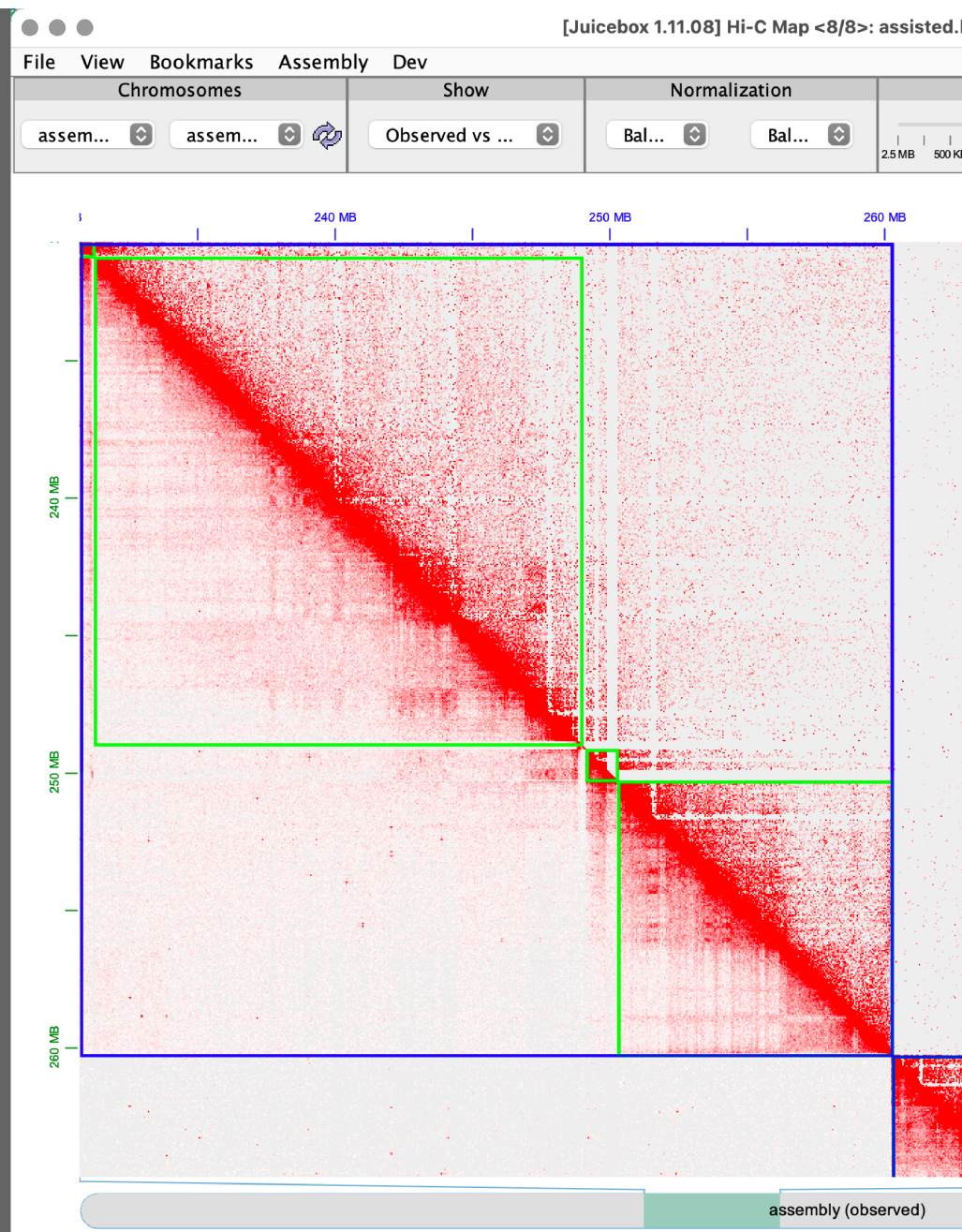
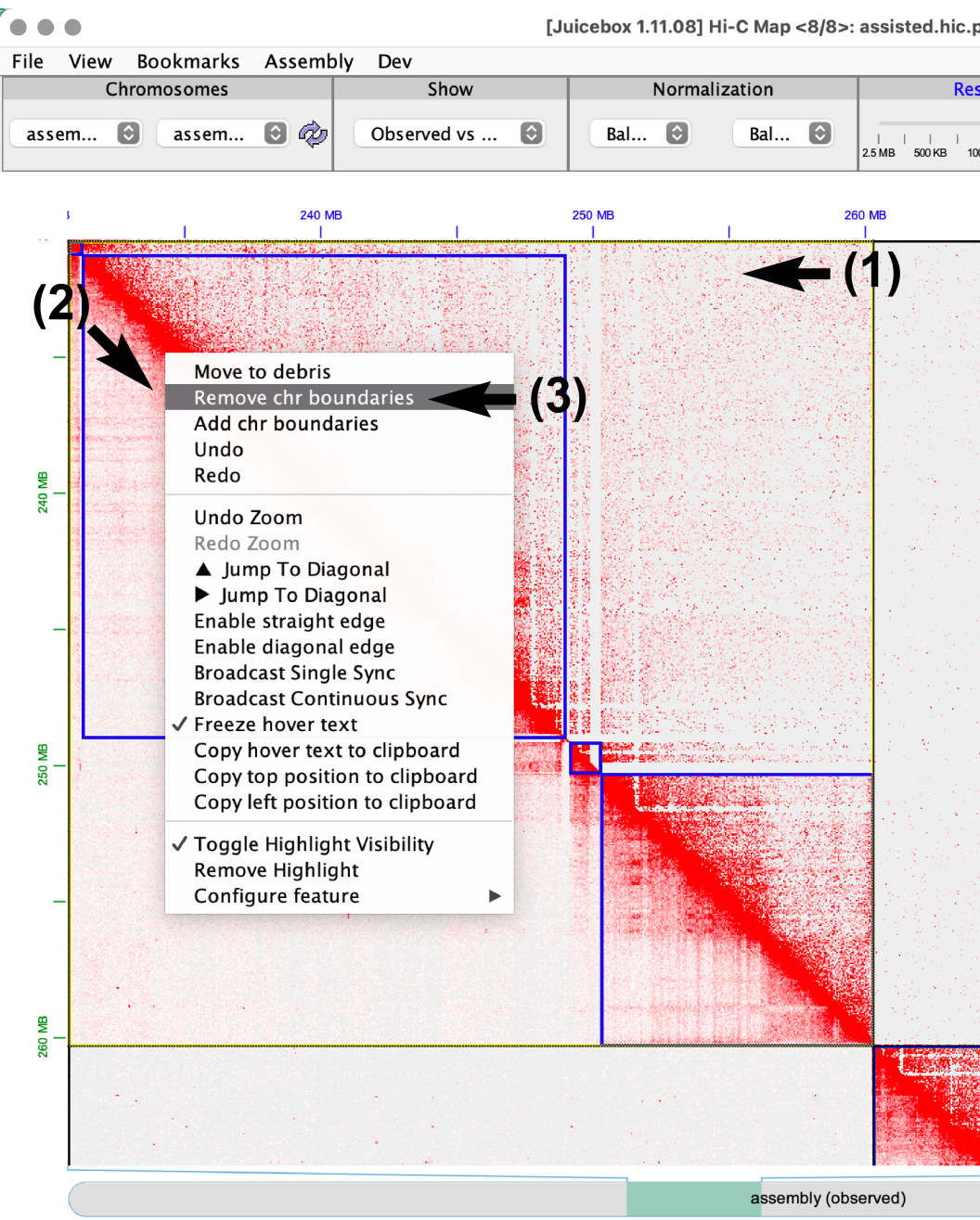
Moving misplaced contigs long distances



Removing a chromosome boundary between two contigs



Removing chromosome boundaries between many contigs



Moving contigs to “debris”

ed.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

Resolution (BP) 500 KB 100 KB 25 KB 5 KB

Color Range 0 19 39

Goto h1tg000008l h2tg000005l

← (2)

← (1)

← (3)

- Move to debris
- Remove chr boundaries
- Add chr boundaries
- Undo
- Redo
- Undo Zoom
- Redo Zoom
- ◀ Jump To Diagonal
- ▼ Jump To Diagonal
- Enable straight edge
- Enable diagonal edge
- Broadcast Single Sync
- Broadcast Continuous Sync
- ✓ Freeze hover text
- Copy hover text to clipboard
- Copy top position to clipboard
- Copy left position to clipboard
- ✓ Toggle Highlight Visibility
- Remove Highlight

assembly (control)

assembly:262,000,001-262,100,000
 assembly:273,800,001-273,900,000
 observed value (O) = 14.121
 expected value (E) = 3.68
 O/E = 3.837

control value (C) = 0.0
 expected control value (EC) = 0.488
 C/EC = 0

average observed value (AVG) = 2.42
 O' = O/AVG = 5.835
 average control value (AVGC) = 0.33
 C' = C/AVGC = 0.0

Balanced
 bin: 2,620
 Balanced (Control)
 bin: 2,620

Balanced
 bin: 2,738
 Balanced (Control)
 bin: 2,738

Feature
 assembly:273,414,447-273,712,200
 assembly:273,414,447-273,712,200

Edit

Chr

Scaf

Show Annotation Panel

3/8>: assisted.hic.p_ctg_0.hic (control=assisted.hic.p_ctg_60.hic)

Resolution (BP) 2.5 MB 500 KB 100 KB 25 KB 5 KB

Color Range 0 19 39

Goto h1tg000008l h2tg000005l

←

assembly (control)

assembly:607,100,001-607,200,000
 assembly:595,200,001-595,300,000
 observed value (O) = 2.31
 expected value (E) = 3.774
 O/E = 0.612

control value (C) = 0.0
 expected control value (EC) = 0.432
 C/EC = 0

average observed value (AVG) = 2.42
 O' = O/AVG = 0.955
 average control value (AVGC) = 0.33
 C' = C/AVGC = 0.0

Balanced
 bin: 6,071
 Balanced (Control)
 bin: 6,071

Balanced
 bin: 5,952
 Balanced (Control)
 bin: 5,952

Edit

Chr

Scaf

Show Annotation Panel

(observed)

Apply modified .assembly changes to a .fasta

To apply the manually-curated changes made in Juicebox (stored in `genome.review.assembly` file) to `genome.fasta` and generate the corrected `.fasta`, do:

```
nohup 3d-dna/run-asm-pipeline-post-review.sh \  
  --stage finalize \  
  --mapq 60 \  
  --gap-size 100 \  
  --review genome.review.assembly \  
  genome.fasta \  
  merged_nodups.txt \  
&>run-asm-pipeline-post-review.log &
```

To see all options (and their descriptions) offered by the post-review script, do:

```
3d-dna/run-asm-pipeline-post-review.sh --help
```

Genome post-review finalization can also be performed with artisanal:

```
assembly-to-fasta genome.review.assembly genome.fasta genome.review
```


Final note: Choosing cell/tissue type(s) for Hi-C

