juicebox.md

This repo is intended as a central location for creating and maintaining documentation about analysis methods, software, demos, and troubleshooting tips and work-arounds, etc.

Curating genome assemblies in Juicebox

Terminology

- **contig**: A contiguous series of `A` / `T` / `C` / `G` (and/or `a` / `t` / `c` / `g`) nucleotides without any intervening subsequence(s) of unknown (*i.e.*, `N`) nucleotides.

- **scaffold**: One or more contigs ordered and oriented to represent a larger (potentially complete) subsequence of the biological nucleotide sequence (*e.g.*, a chromosome).
- **paired-end**: A sequencing format whereby DNA libraries are produced with known sequence adapters ligated to the 5' and 3' ends of a genomic DNA fragment (sometimes called an 'insert'). These libraries are often constructed from a pool of shredded and sized-selected (from 100 bp up to 1 kb) fragments without any circularization procedure, such that sequencing the ends of the completed library--priming from the known 5' and 3' adapters--produces a pair of sequencing reads in forward-strand and reverse-strand (*i.e.*, FR) orientation.
- **mate-pair**: A sequencing format similar to paired-end sequencing, but because the chemistry targets larger (1 kb to 50 kb) fragments, cirucularization and second shearing step are required, producing reads in reverse-strand and forward-strand (*i.e.*, RF) orientation.
- **matrix resolution**: the bin size used to contruct a Hi-C contact map matrix (Rao et al. 2014).
- **map resolution**: the smallest locus size such that 80% of loci have at least 1,000 contacts (Rao et al. 2014).

Introduction

The genome assembly process traditionally relied heavily on scaffolding contigs with multiple paired-end or mate-pair sequencing libraries in order to jump over unassemblable genomic repeats (this is why they are sometimes referred to as 'jumping' libraries) to order-and-orient contigs into sub-chromosomal approximations of the true underlying genomic sequence. This process was highly computation-focused and many different orthogonal datatypes (irregular sequencing coverage, genetic linkage maps, restriction-based physical maps, etc.) were used to identify and correct scaffolding misassemblies. Identifying these misassemblies was often an interative task because the diversity of data types used and their different degrees of sequence resolution; therefore, manual intervention to correct scaffolding misassemblies was an *ad hoc* process focusing on increasing subsets of the assembly. As a result, few tools could work at all genomic scales and allow users to interactively modify their assemblies, until Hi-C and Juicebox Assembly Tools (JBAT) were invented.

In this tutorial, the reader will learn the anatomy of the Juicebox desktop application, how to interact with and customize its graphical user interface (GUI), how to navigate a genome assembly, interpret common Hi-C contact patterns, and correct misassemblies with JBAT. Using Juicebox for feature annotation is not yet described here (in the meantime, see the official Juicebox wiki).

Dependencies

Tool Inventory

Juicebox

> Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst*. 2016 Jul;**3**(1):99-101. doi: 10.1016/j.cels.2015.07.012. PMID: 27467250.
>
> Dudchenko O, Shamim MS, Batra SS, Durand NC. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Biorxiv. 2018. doi: 10.1101/254797.

Online manual:

- See: gitbook.io
- See: Genome Assembly Cookbook, Chapter 5 (pg. 30)

Obtaining the input files for Juicebox (JBAT)

Juicebox Assembly Tools (JBAT) requires two input file types:

1. **.hic**: Files ending in `.hic`. A binary container file format storing matrices of Hi-C contact counts between pairwise genomic bins at pre-computed widths/resolutions (typically at 2.5 Mb, 1 mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, 5 k, 2.5 kb, and 1 kb). JBAT mode in Juicebox is enabled only with a certain prepartion of .hic file: one which combines the input genomic assembly sequences onto a single pseudo chromosome coordinate system named 'assembly'. *Files prepared directly with `juicer_tools.jar pre` are incompatible with JBAT and will not be editable in Juicebox.*

- To create a compatible .hic file with the Juicer + 3D-DNA workflow, go to the Juicer and 3D-DNA scaffolding page.
- To create a compatible .hic file from a BAM file, see this nice step-by-step guide.

2. **.assembly**: Files ending in `.assembly`. A text file representing sequence names and lengths, their orders, orientations, and groupings in a genome assembly.

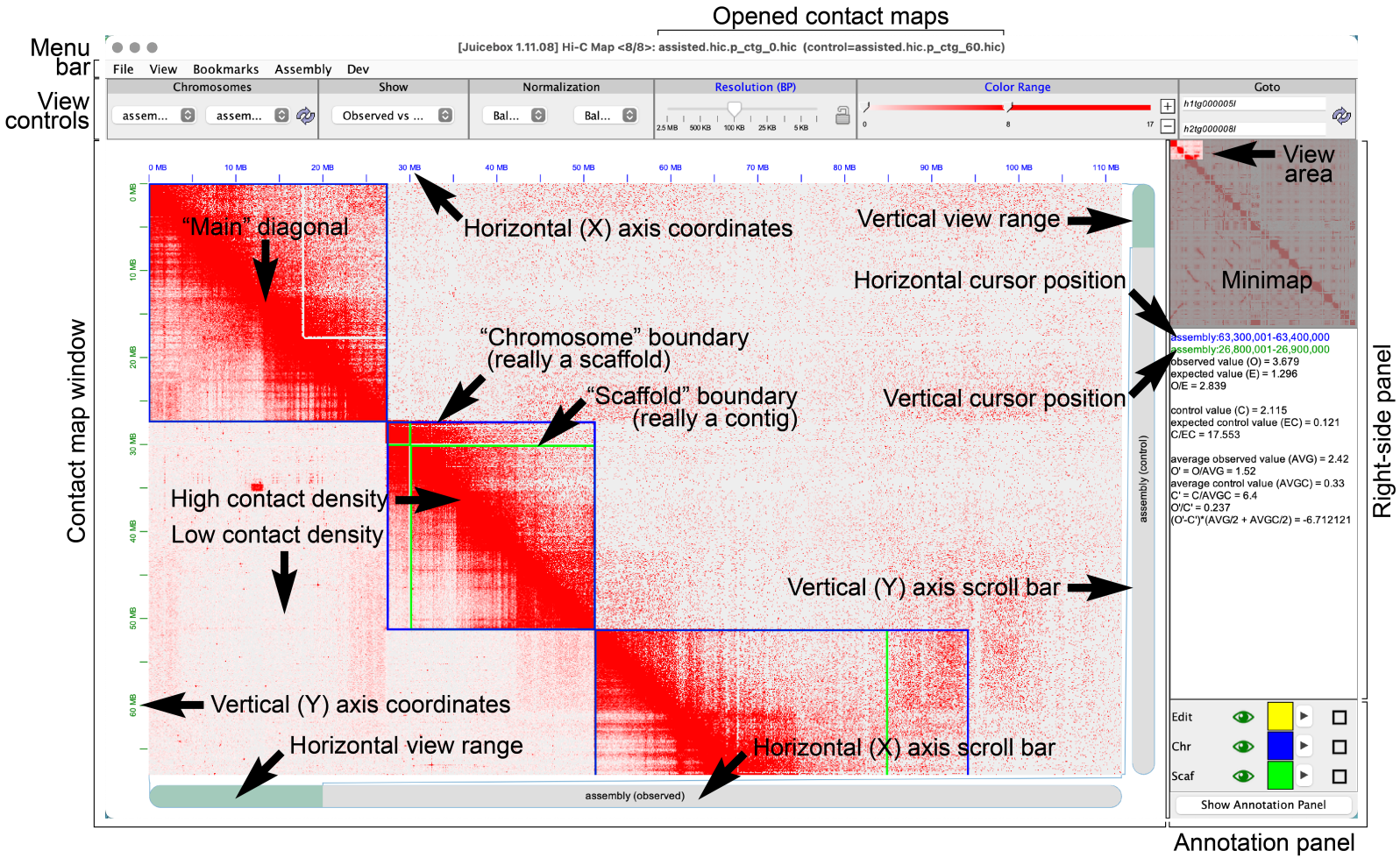
- **map .assembly**: An .assembly-formatted file representing the *physical* state (subsequence connectedness, order, and orientation) of sequences (contig/scaffold/chromosome) in the FASTA file the Hi-C reads were mapped to.
- **modified .assembly**: An .assembly-formatted file representing the *virtual* state of sequences after a series of breaking and/or ordering-and-orienting operations to be applied to the genome assembly FASTA.
- **NOTE**: If the genome sequence the Hi-C reads were aligned to have been pre-scaffolded (*e.g.*, with mate-pairs, fosmids, and/or a third-party Hi-C service), it is recommended to create a map .assembly of that sequence representing contigs. Such an .assembly file can be obtained using the ARTISANAL repo:

```
$ 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-fasta -c yourgenome.fasta yourgenome.contig
```

Introduction to the Juicebox desktop app

TODO: A disclaimer about how contig/scaffold/chromosome sequences are represented as blue/green boxes. Why Scaf/Chr in annotation panel instead of contig/scaff.

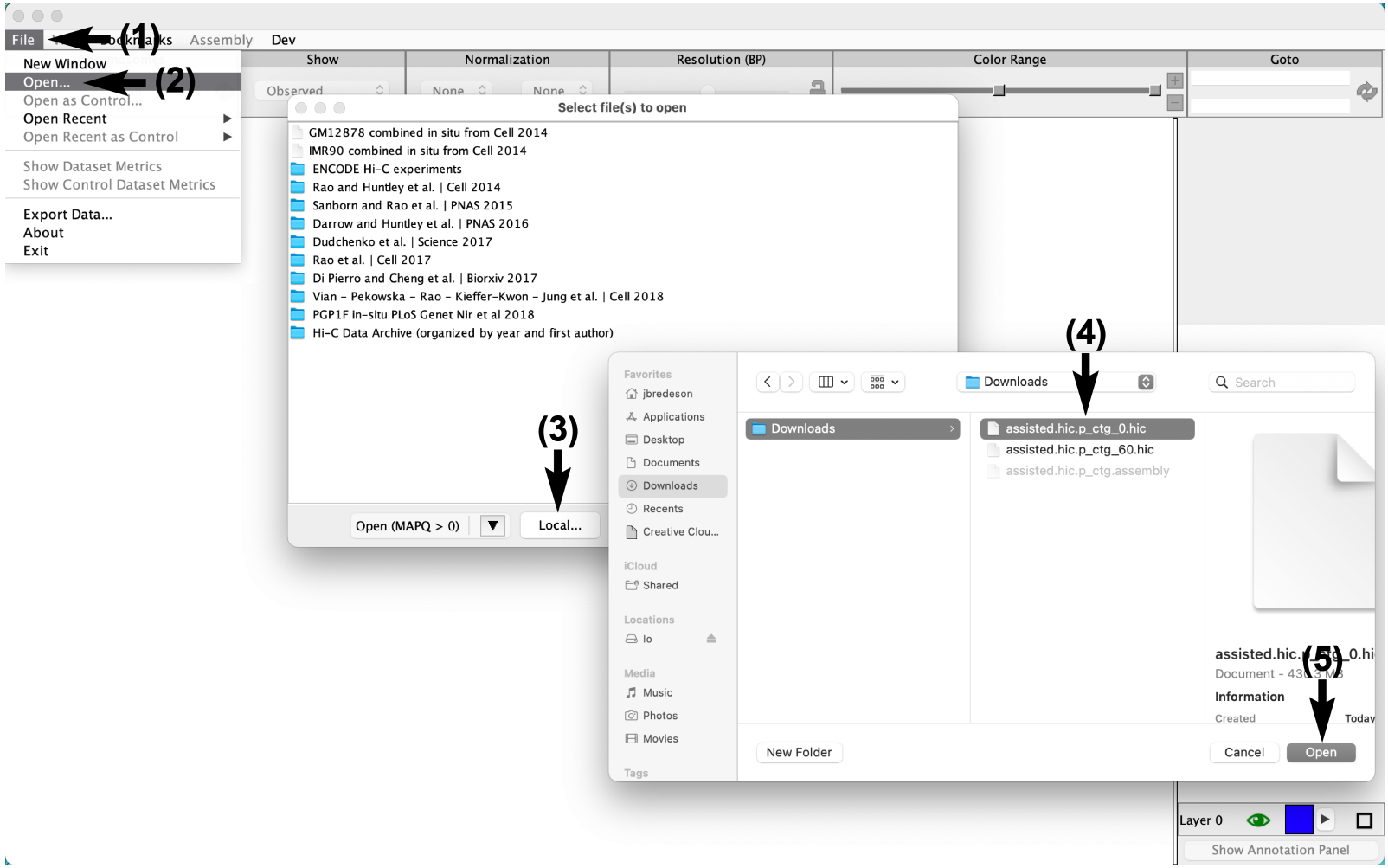
Anatomy of Juicebox



- **Application bar**:
- **Application name and version**:
- **File format**:
- **Opened contact maps**:
  - **Observed**:
  - **Control**:
- **Menu bar**:
- **View controls**:

- **Contact map panel**:
- **Contact density map**:
- **Low and high densities**:
- **Main Diagonal**:
- **Chromosome boundaries**:
- **Scaffold boundaries**:
- **Horizontal (X) axis coordinates**:
- **Vertical (Y) axis coordinates**:
- **Horizontal (X) axis scroll bar**:
- **Horizontal (X) axis view range**:
- **Vertical (Y) axis scroll bar**:
- **Vertical (Y) axis view range**:
- **Right-side panel**:
- **Minimap and view area**:
- **Information subpanel**:
    - **Horizontal and vertical cursor positions**:
    - **Contact density values**:
- **Annotation panel**:

Opening an 'Observed' .hic contact map file



The first .hic contact map file to be loaded is the 'Observed' map. In the context of feature annotation, this map corresponds to the treatment dataset; in the genome assembly context, it's useful to load in a contact map of reads mapped at low mapping quality (MapQ >= 0) as an Observed map and load a high-stringency (MapQ >= 60) map a 'Control' (described below). Whether the user plans on curating a genome assembly or performing feature annotation, loading an Observed map is required.

1. Click on the **File** drop-down tab in the main menu bar.
2. Select the **Open** option in the drop-down menu. A file navigation window will appear.
3. Click on the **Local** button to load files from one's computer (the folders presented in the window organize published feature annotation datasets accessible via remote server and cannot be used to navigate to the

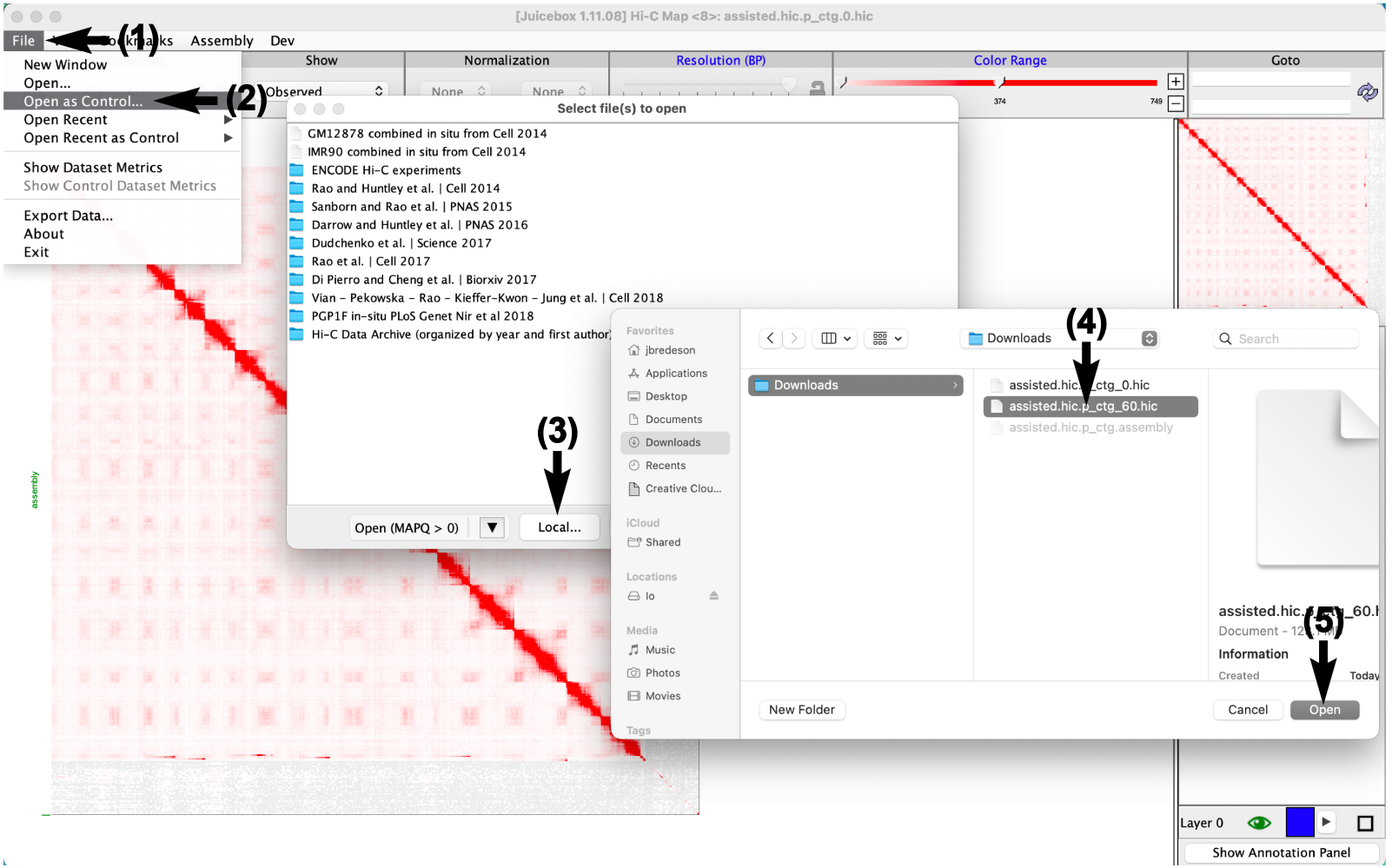
user's desired dataset on their local computer). A new file navigation window will appear allowing the user to navigate their file system to the file to be opened.

4. Click on *filename* to select it for loading. The filename will become highlighted.
5. Click **Open** to load the file.

If the user wishes to re-open an Observed .hic file loaded in a previous Juicebox session, the user can instead:

1. Click on the **File** drop-down tab.
2. Hover the mouse cursor over the **Open Recent** in the drop-down menu to expose a menu of recent files.
3. Click on the *filename* to load.

Loading a 'Control' .hic file



Optionally, a second .hic contact map file can be visualized simultaneously with the Observed contact map. This second map is the 'Control' map and, in the context of feature annotation, corresponds to a control dataset. In the context of genome assembly, it's often useful to load a dataset stringently filtered for read mapping quality (MapQ >= 60) to disambiguate true high density contacts vs. repetitve contacts.

1. In the main menu bar, click on the **File** drop-down tab.
2. Select the **Open as Control** option in the drop-down menu. A pop-up file navigation window will appear.
3. Click on the **Local** button to load files from one's computer (the folders presented in the window contain published feature annotation datasets accessible via remote server and cannot be used to navigate to the user's desired dataset on their local computer). A new file navigation window will appear allowing the user to

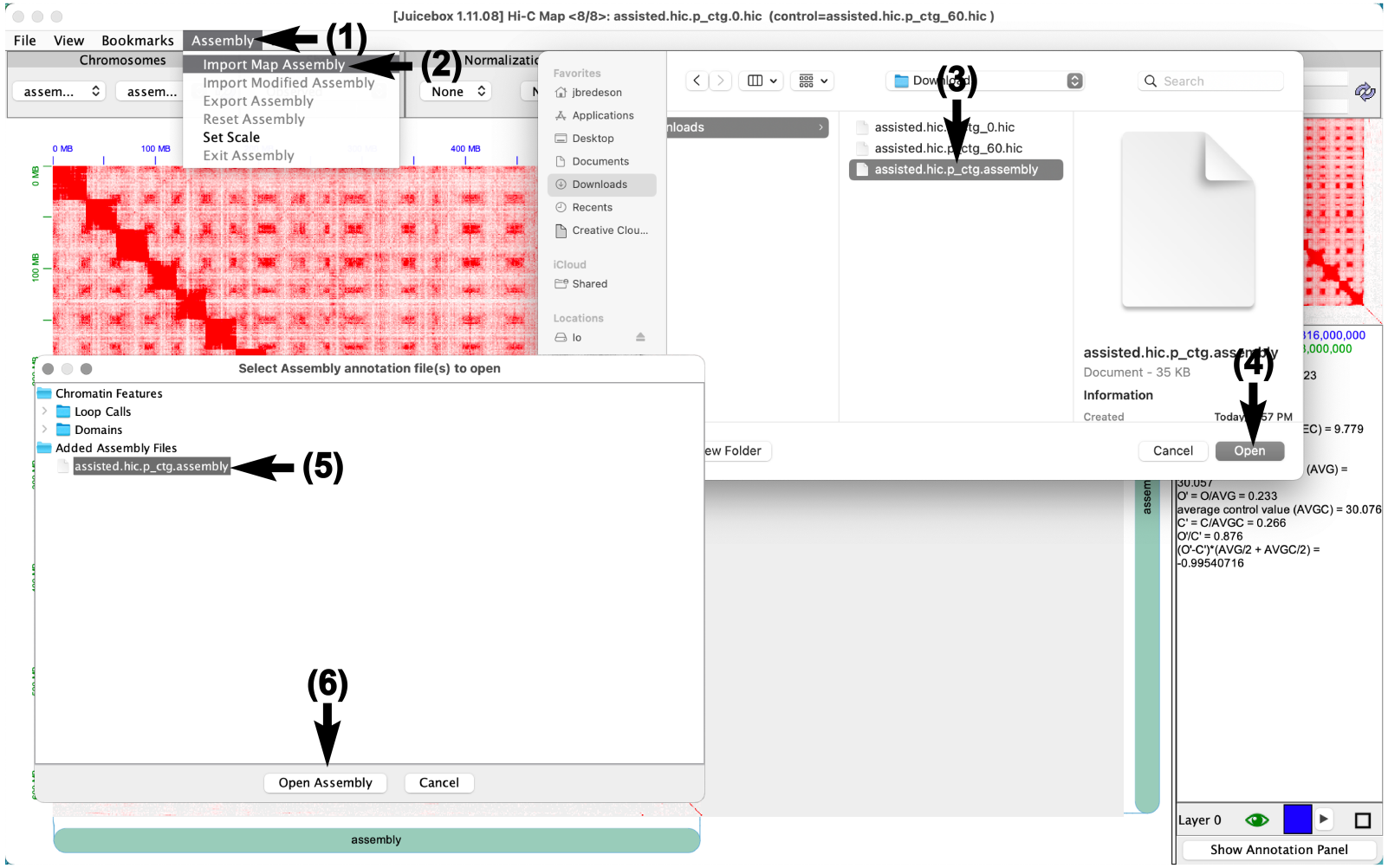
navigate their file system to the file to be opened.

4. Click on *filename* to select it for loading. The filename will become highlighted.
5. Click **Open** to load the file.

If the user wishes to re-open a Control .hic file loaded in a previous Juicebox session, the user can instead:

1. Click on the **File** drop-down tab.
2. Hover the mouse cursor over the **Open Recent as Control** in the drop-down menu to expose a menu of recent files.
3. Click on the *filename* to load.

Opening a 'map' .assembly file

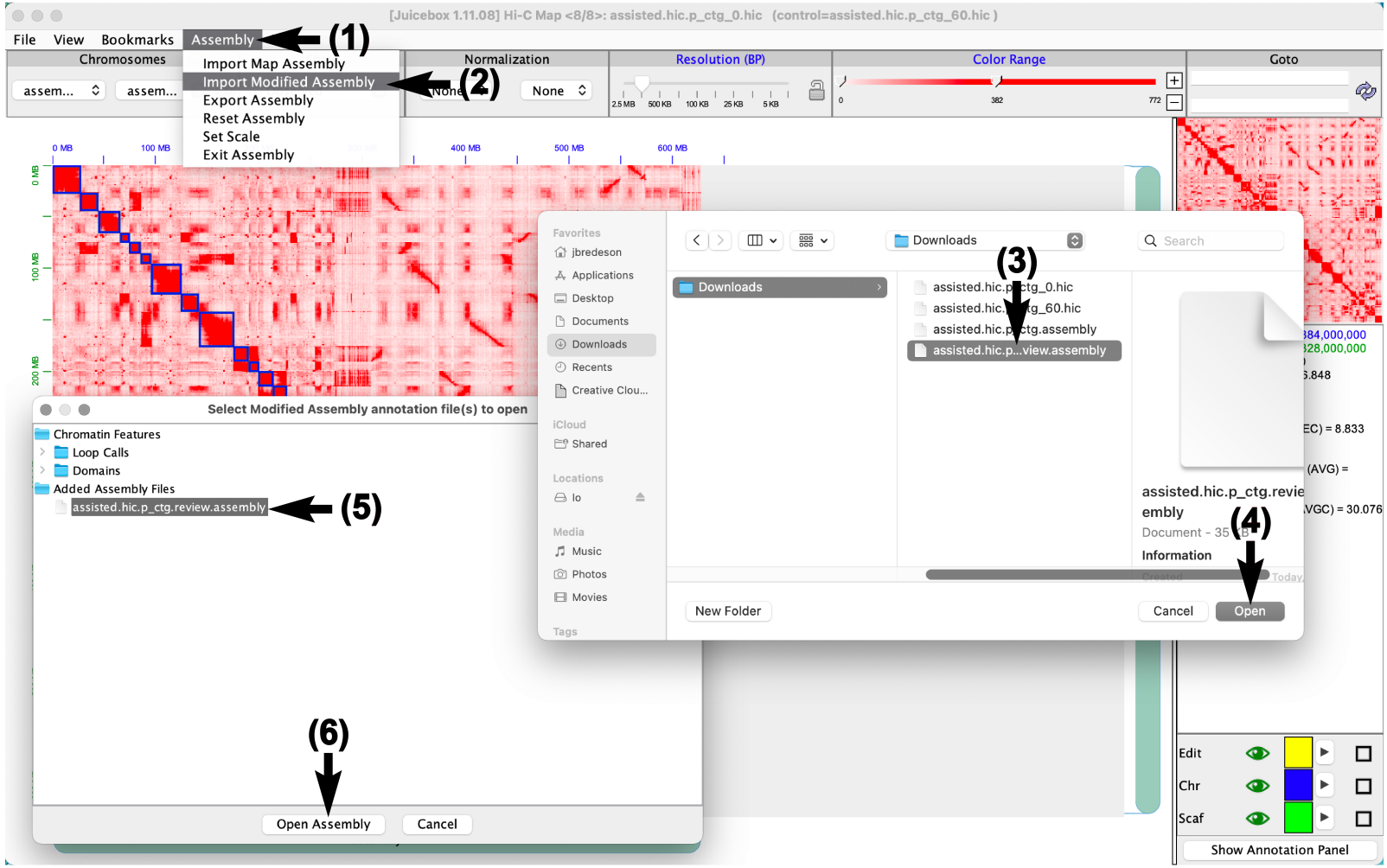


A .hic contact density map is just a matrix of contact density values for a whole genome, no sequence boundary information is explicitly represented. The .assembly file provides this information and *map*s the initial coordinate boundaries of the sequences onto the matrix so that rows and columns corresponding to the sequences can be easily re-ordered. *A map .assembly file must be loaded before any modification to a sequence assembly can be made*. Loading an .assembly file is not required for feature annotation, however, and (in fact) Juicebox will not permit the contact matrix to manipulated.

1. In the menu bar, click on the **Assembly** drop-down tab.
2. Select **Import Map Assembly** in the drop-down menu. A file navigation window will appear.
3. Navigate to, and click, on the *filename* of the .assembly file to be loaded.

4. Click the **Open** button. A second pop-up navigation window will appear.
5. Click the *filename* to highlight the .assembly file to be loaded.
6. Click the **Open Assembly** button.

Loading a 'modified' .assembly file

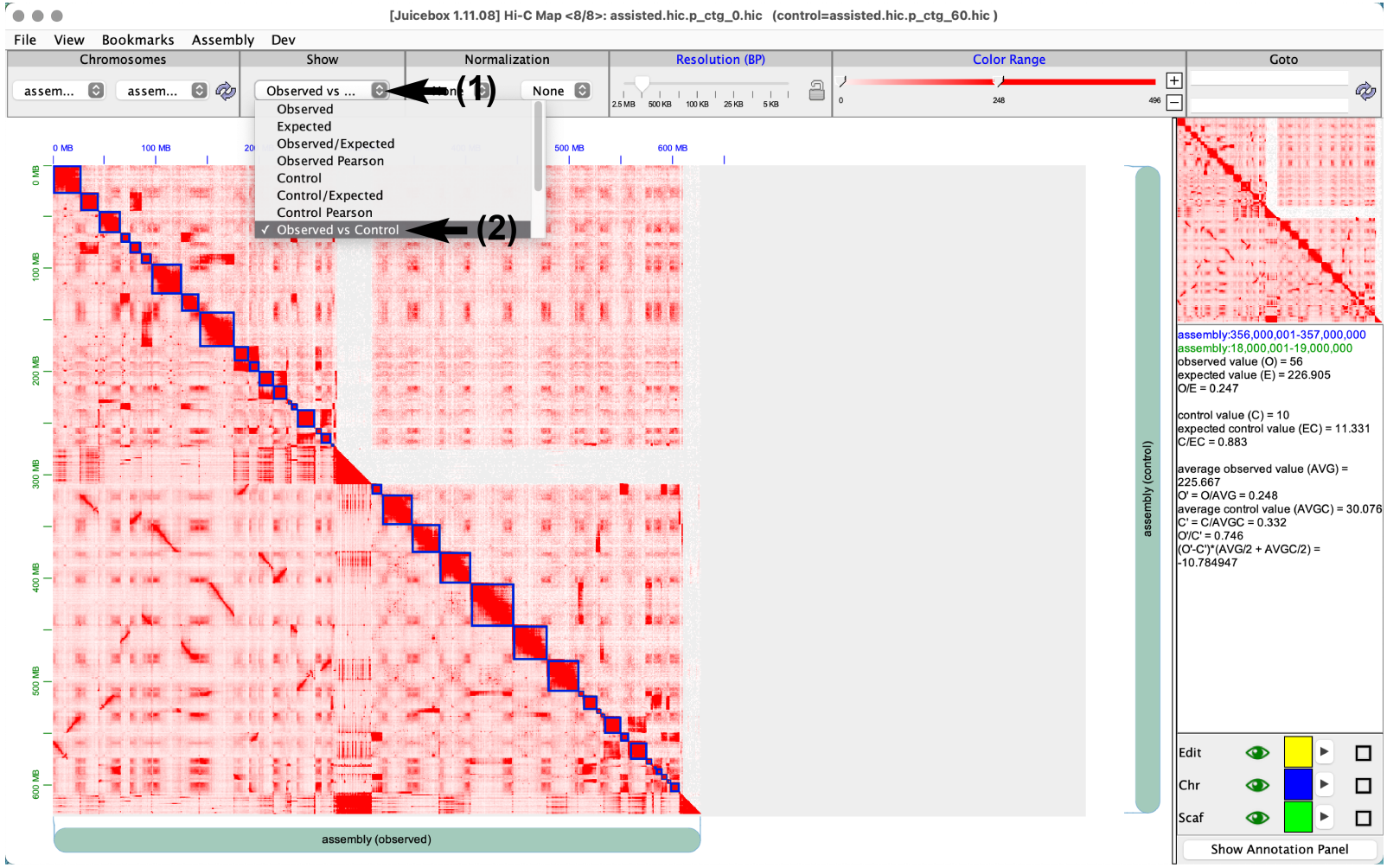


A map .assembly file maps the initial sequence boundary coordinates onto the contact map, a 'modified' .assembly file stores the boundary coordinates after a series of assembly modifications (*e.g.*, moves, invertions, and breaks).

A map .assembly file must be loaded before a modified .assembly file.

1. In the menu bar, click on the **Assembly** drop-down tab.
2. Select **Import Modified Assembly** in the drop-down menu. A file navigation window will appear.
3. Navigate to, and click, on the *filename* of the .assembly file to be loaded.
4. Click the **Open** button. A second pop-up navigation window will appear.
5. Click the *filename* to highlight the .assembly file to be loaded.
6. Click the **Open Assembly** button.

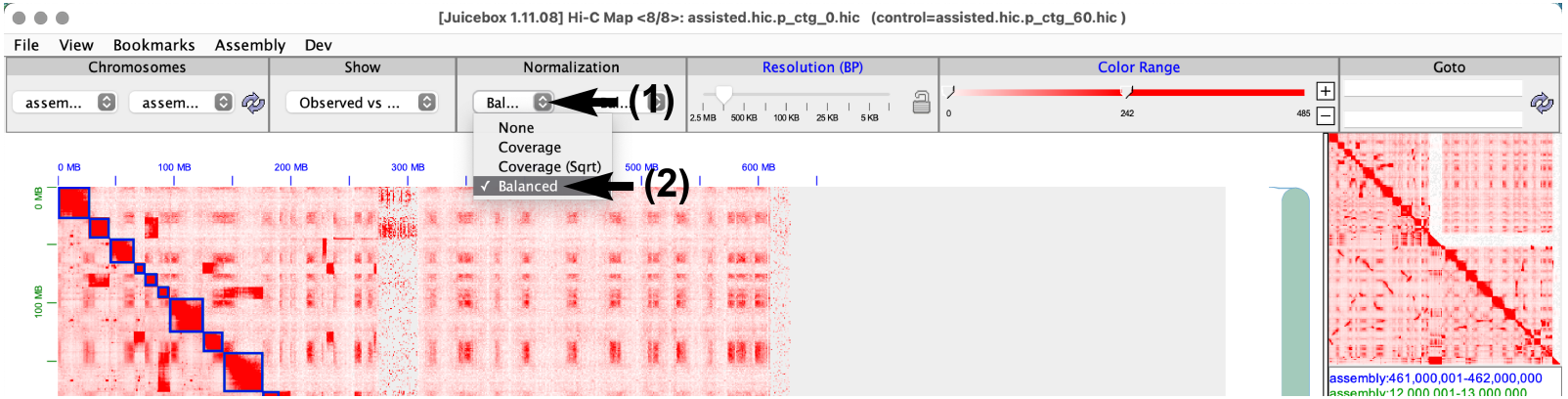
Changing contact map(s) displayed with 'Show'



Once both Observed and Control .hic maps have been loaded:

1. Click on the **Show** selection drop-down menu in the view controls bar. An enumerative drop-down list will appear with all available options.
2. Select **Observed vs Control** to show the Observed and Control maps side-by-side (the user may have to scroll to find this selection).

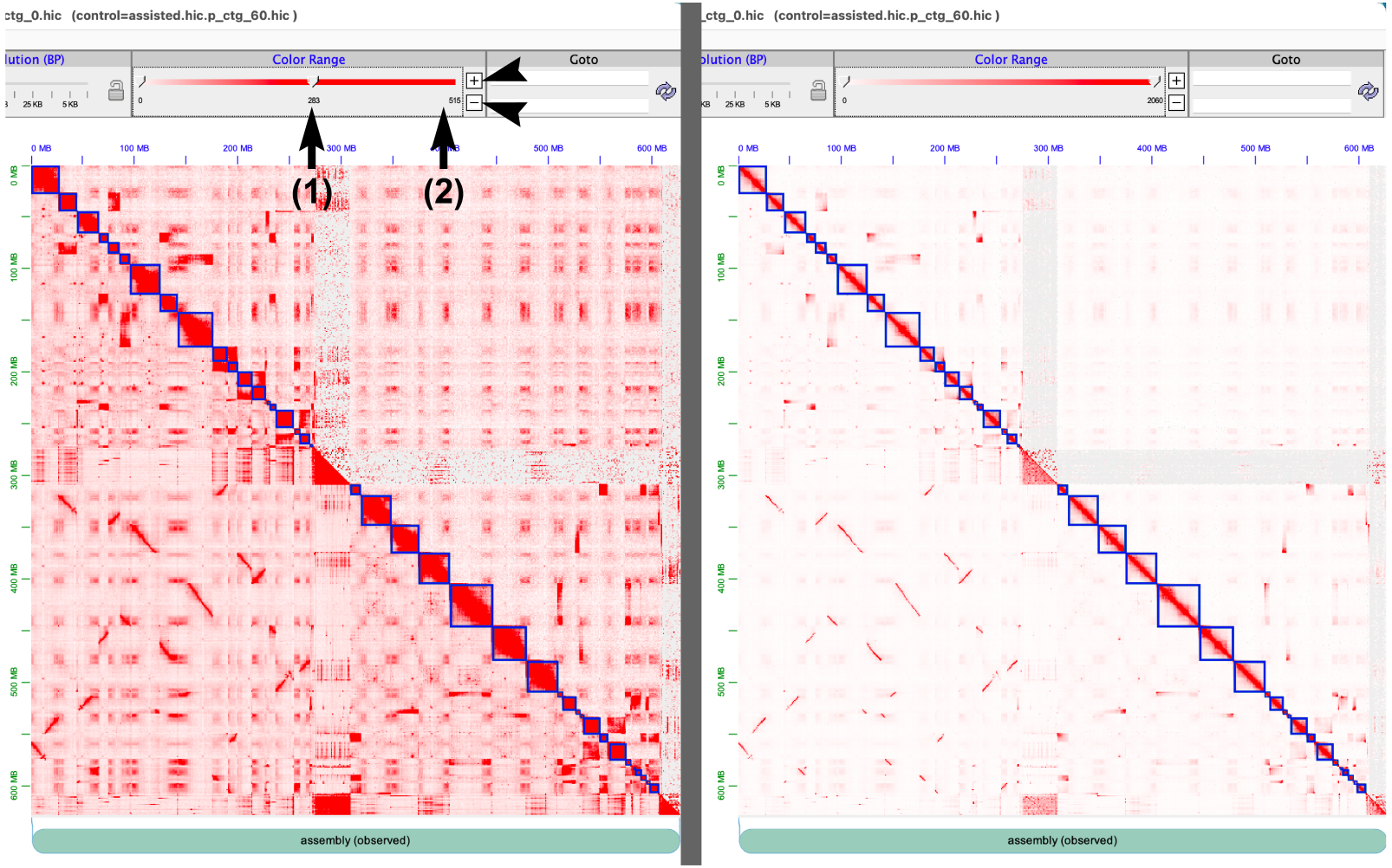
Enabling contact map normalization



Because of Hi-C artefacts, genomic repeats, and biological variability in local background contact densities, true enriched contact puncta can sometimes be obscured. Statistical contact map normalization can often help enhance the signal-to-noise in such datasets.

1. Click on the **Normalize** selection drop-down menu in the view controls bar. Under the Normalize subpanel, there are two drop-down boxes; the left-most sets the normalization for the Observed map, the right-most sets the normalization for the Control map (if loaded). Clicking on the drop-down menu will cause an enumerative drop-down list to appear with all available options.
2. Select **Balanced** to show the Knight-Ruiz balanced contact map.

Adjusting contact map color range/saturation

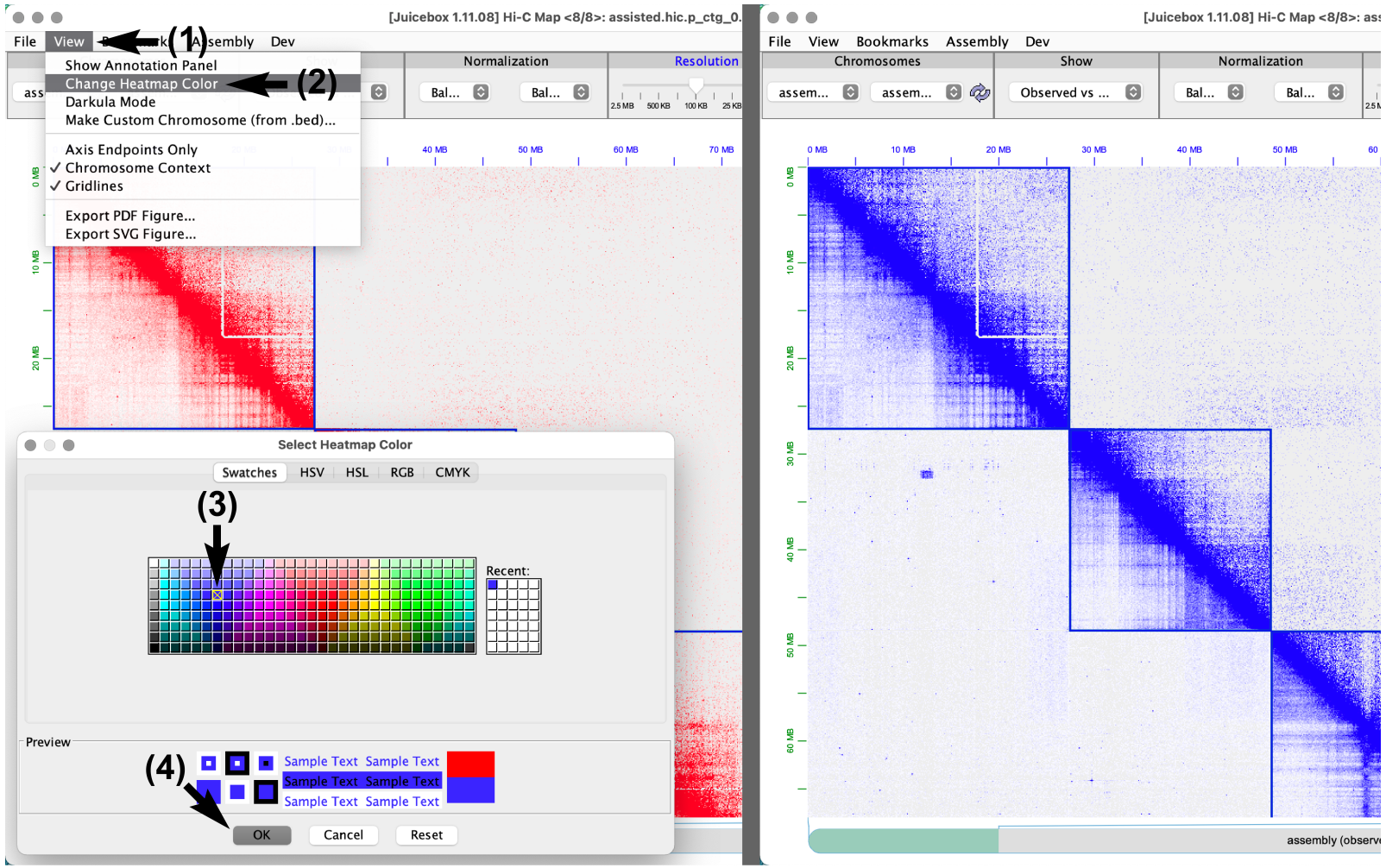


Using the Color Range slider, the user can increase or decrease the contact density saturation. Clicking the [+] or [-] buttons (black arrow heads) to the right of the slider will increase or decrease the slider's dynamic range, respectively.

1. Press-and-hold with the left mouse button on the **Color Range** slider wedge in the view control bar.
2. Drag the wedge to the desired saturation level. Moving the indicator wedge left increases saturation, moving it right decreases saturation.

TIP: Alternatively, the user can click once directly on the Color Range slider saturation level desired.

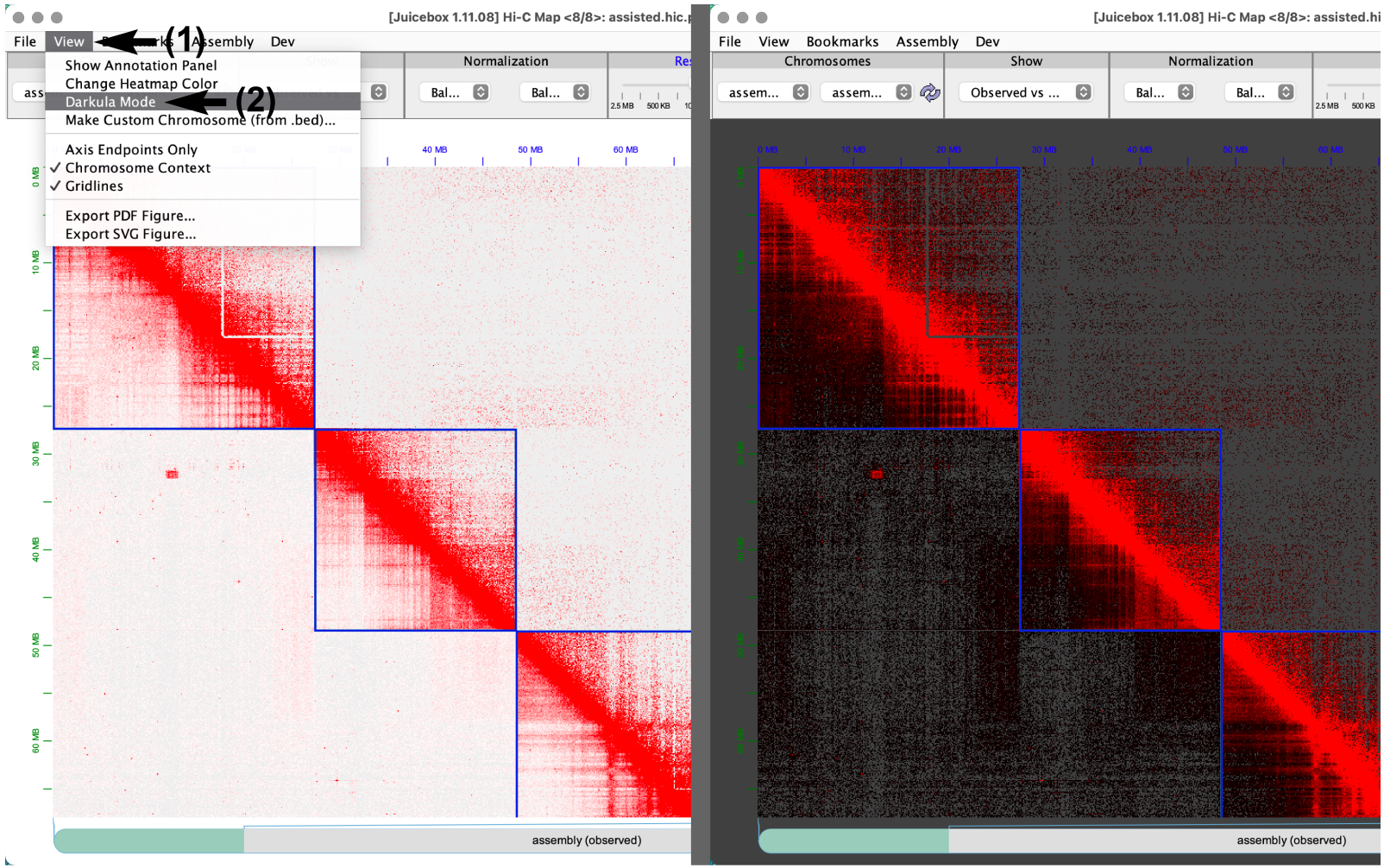
Changing contact map color



Users can change the contact map pixel color to suit their style preferences or increase visilibilty and accessibility (the default red contact map appears yellow to colorblind users and may be difficult to see).

1. Click on the **View** drop-down tab in the menu bar.
2. Select **Change Heatmap Color**. A pop-up window will appear.
3. Select a desired color from the available options.
4. Click the **OK** button.

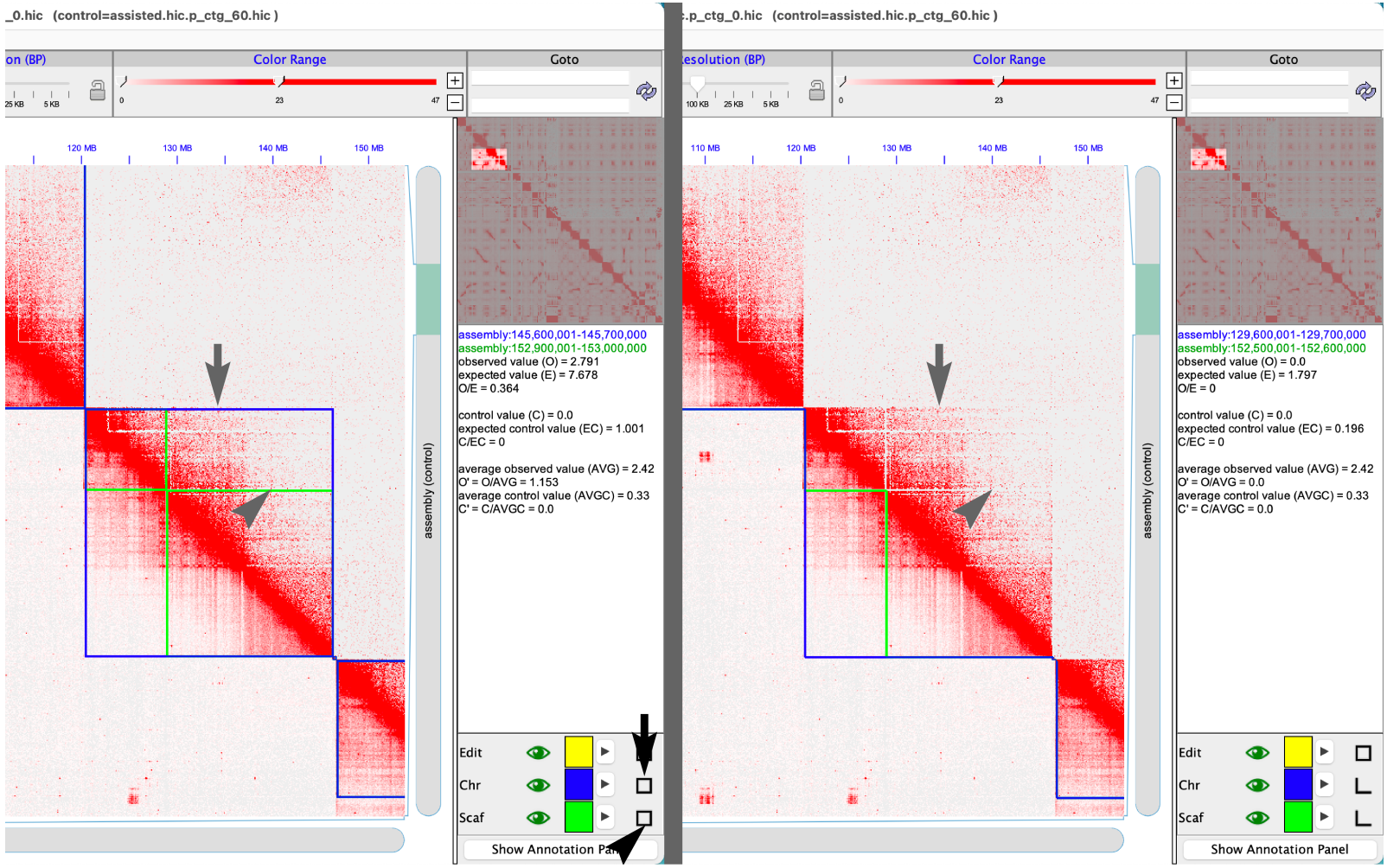
Darkula (night) mode



When working in low-light conditions, the white contact map background may appear too bright for some users. The background can be converted to a dark grey using Darkula mode.

1. Click on the **View** drop-down tab in the menu bar.
2. Select the **Darkula Mode** option.

Toggling scaffold/chromosome boundary visibility

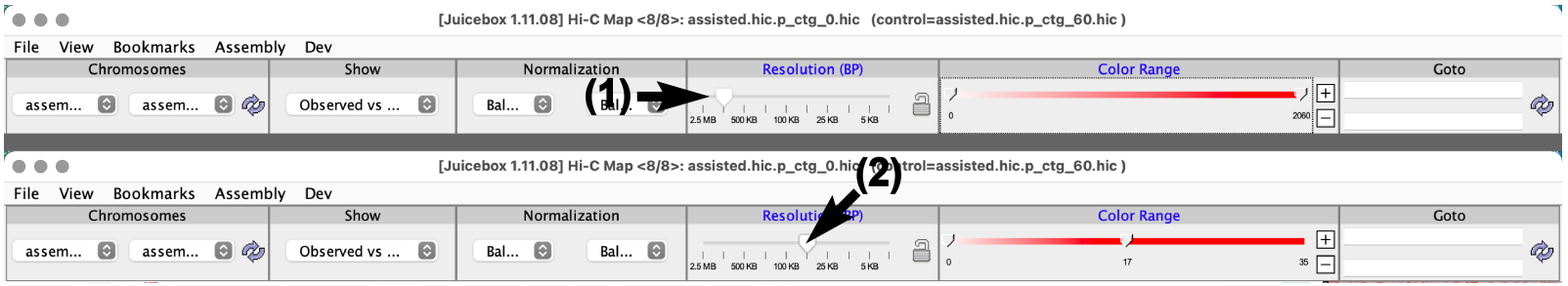


When examining contacts near the corners of adjacent sequences, the chromosome (blue) and scaffold (green) boundaries can sometimes obstruct one's view. These boundaries can be toggled on/off above (1 click) and below (2 clicks) the diagonal.

- Toggle chromosome boundaries (grey arrow) by clicking the black-outlined box to the right of the 'Chr' item (black arrow) in the annotation panel.
- Toggle scaffold boundaries (green arrow head) by clicking the black-outlined box to the right of the 'Scaf' item (black arrow head) in the annotation panel.

TIP: The 'Scaf' or 'Chr' boundaries can be made invisible/visible by toggling the green eye icon to the right of the respetive item.

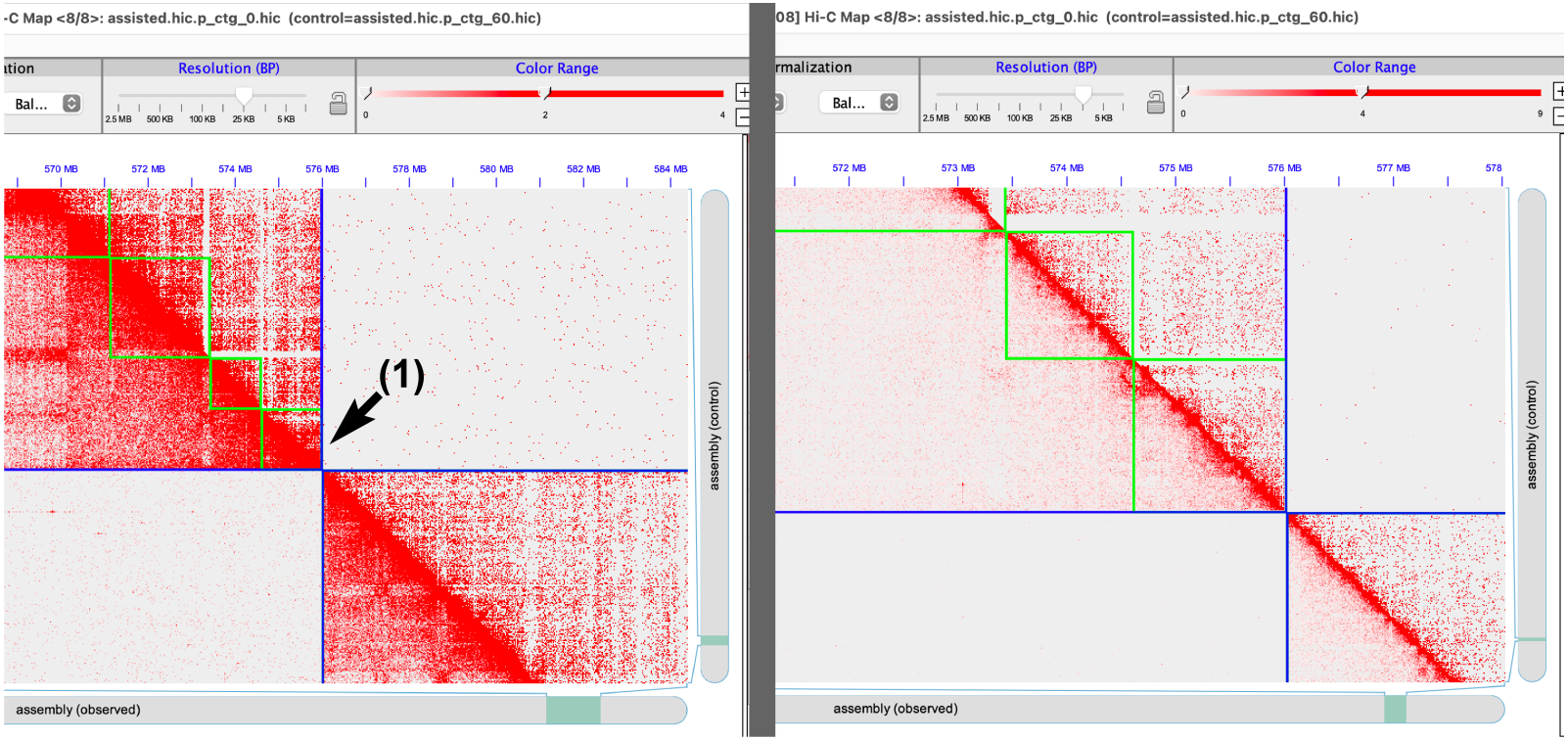
Zooming-in/-out with the 'Resolution' slider



1. Press-and-hold with the left mouse button on the **Resolution (BP)** slider wedge in the view control bar.
2. Drag the wedge to the desired zoom level. Moving the indicator wedge left will zoom-out (to low resolution), moving it right will zoom-in (to high resolution).

TIP: Alternatively, the user can click once directly on the Resolution (BP) slider zoom level desired.

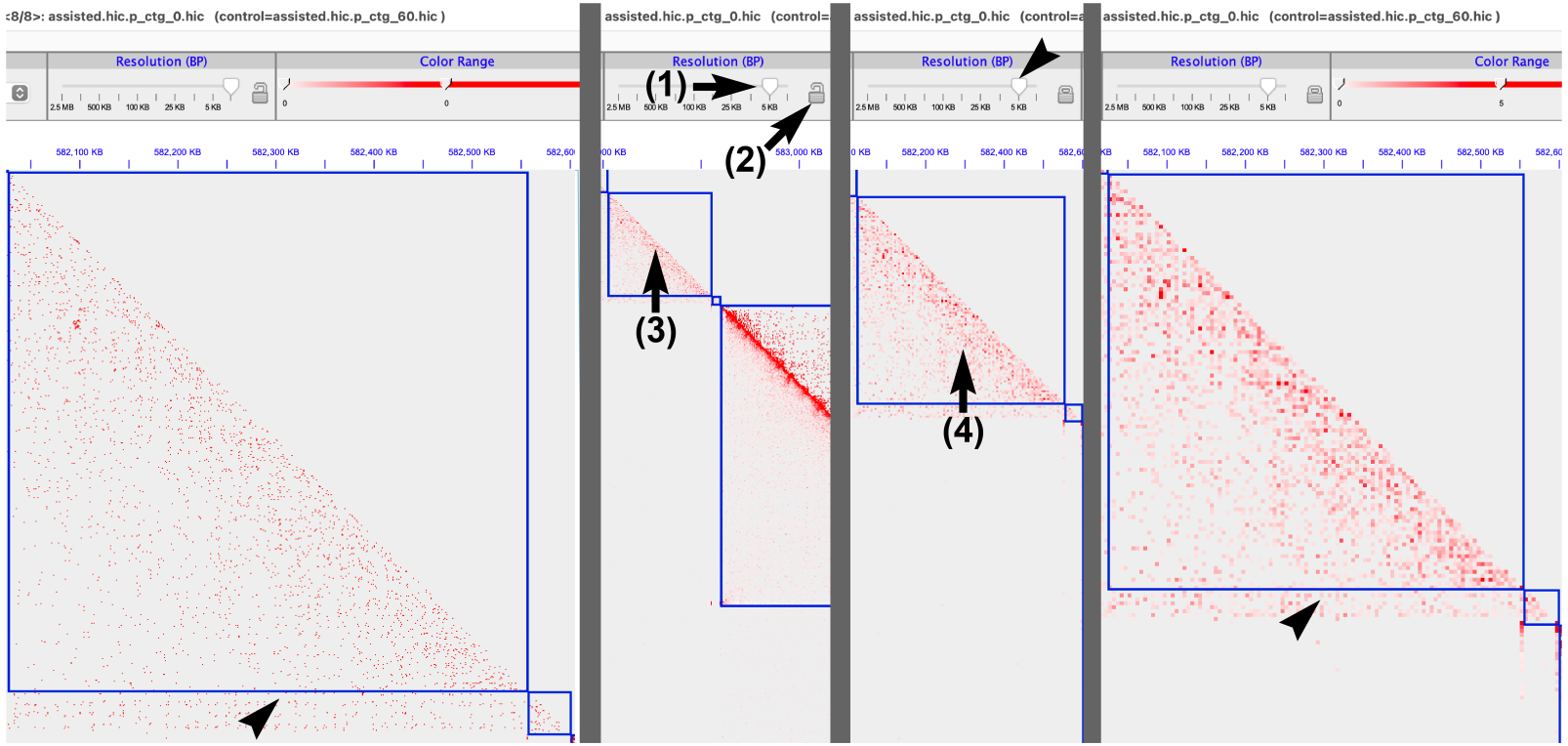
Zooming-in with a mouse double-click



If the user needs to zoom-in to higher resolution than that on the resolution slider (often 1kb resolution), or the user wants to zoom-in on a single point, they can double-clicking in the contact map area. The contact map viewing area is then centered on the double-clicked point.

1. Position the mouse cursor over the desired zoom position
2. Double-click with the left mouse button.

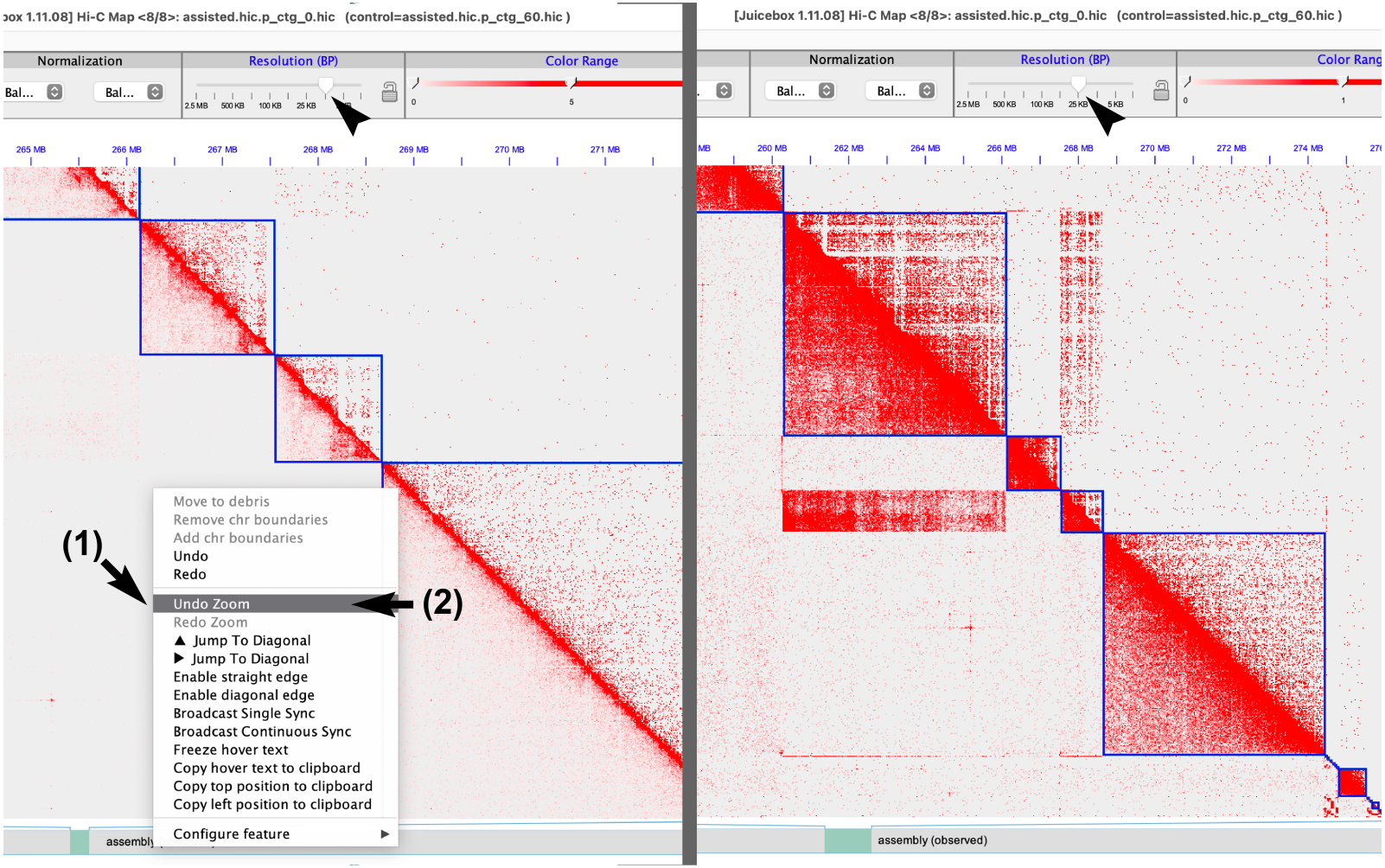
Zoom with the 'Resolution' lock



In areas of sparse contacts (black arrow head, left-most figure panel), particularly for small sequences, zooming-in may cause these sparse contact pixels to appear smaller and harder to see. In such cases, it's useful to maintain a low matrix resolution at a more zoomed-in level; the contact pixels then increase in size and visibility (black arrow head, right-most figure panel). This can be achieved with the Resolution Lock. *Resolution lock only takes affect when zooming in by double-clicking on the contact map window area, it will not apply when using the resolution slider.*

1. Zoom-out one or two resolution level(s) above the desired resolution level.
2. Click on the lock icon in the **Resolution (BP)** subpanel in the view control bar to toggle on the matrix resolution lock.
3. Position the mouse cursor over the desired zoom position, then double-click with the left mouse button.
4. Countinue double-clicking until a satisfactory resolution is achieved. Note that the resolution slider indicator wedge (black arrow head, third figure panel) does not change position when the resolution lock is enabled.

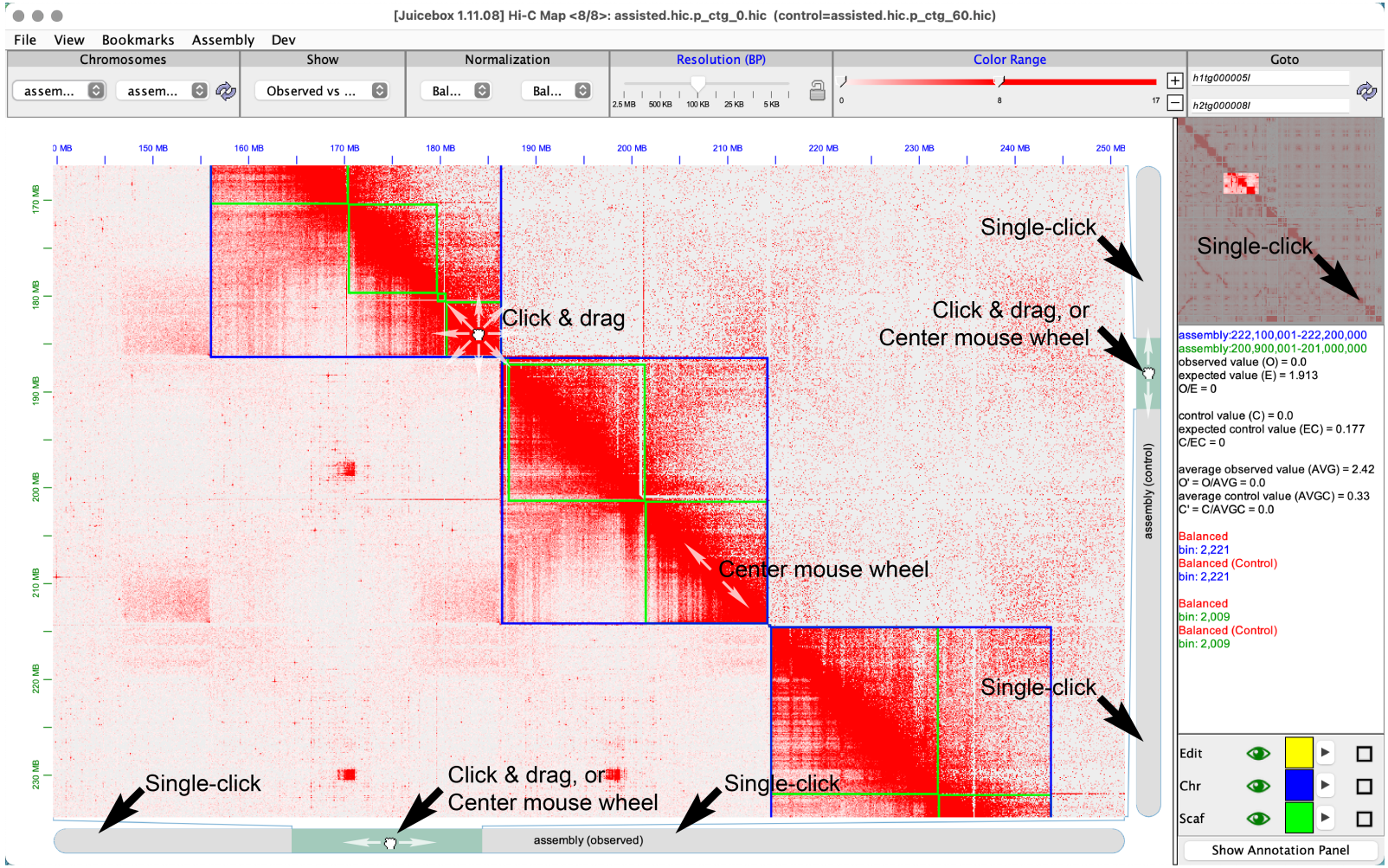
Zooming-out with 'Undo Zoom'



Another method of zooming-out is available to complement zooming-in with a double-click:

1. In the contact map window area, right-click (*control* + click on Mac) to **open the context menu**.
2. Select **Undo Zoom**. Note that the resolution slider changes position to reflect the change in zoom level.

Navigating the contact map



Grab-and-drag

Fine-scale navigation can be performed by pressing-and-holding the mouse cursor in the contact map area (the cursor will become a white fist), then dragging in the contact map area. This allows the user to move in all directions.

Along the diagonal

The user can use the center mouse-wheel (two-finger scroll on a Mac trackpad) to move along the y=x diagonal.

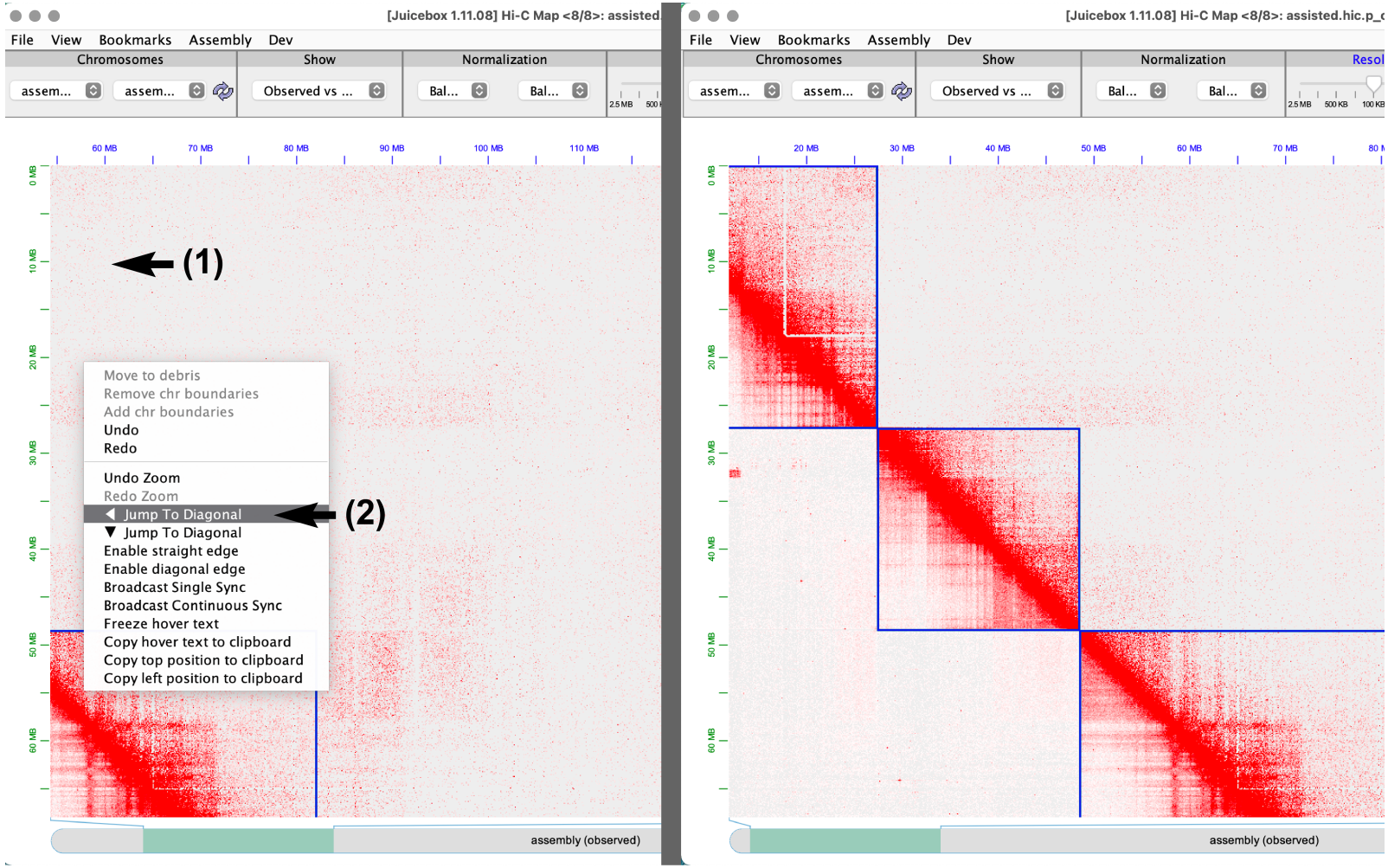
Along horizontal and vertical axes

The user can press-and-hold (the cursor will become a white fist) on the aqua-highlighted segments in bottom- or right-side grey chromosome-glyph scroll bars, then drag to move in the X and Y directions, respectively. The user may also single-click on the grey portions of the scroll bars.

Jump to position in minimap

Course-scale movement can be achieved by single-clicking on the desired position in the minimap map area.

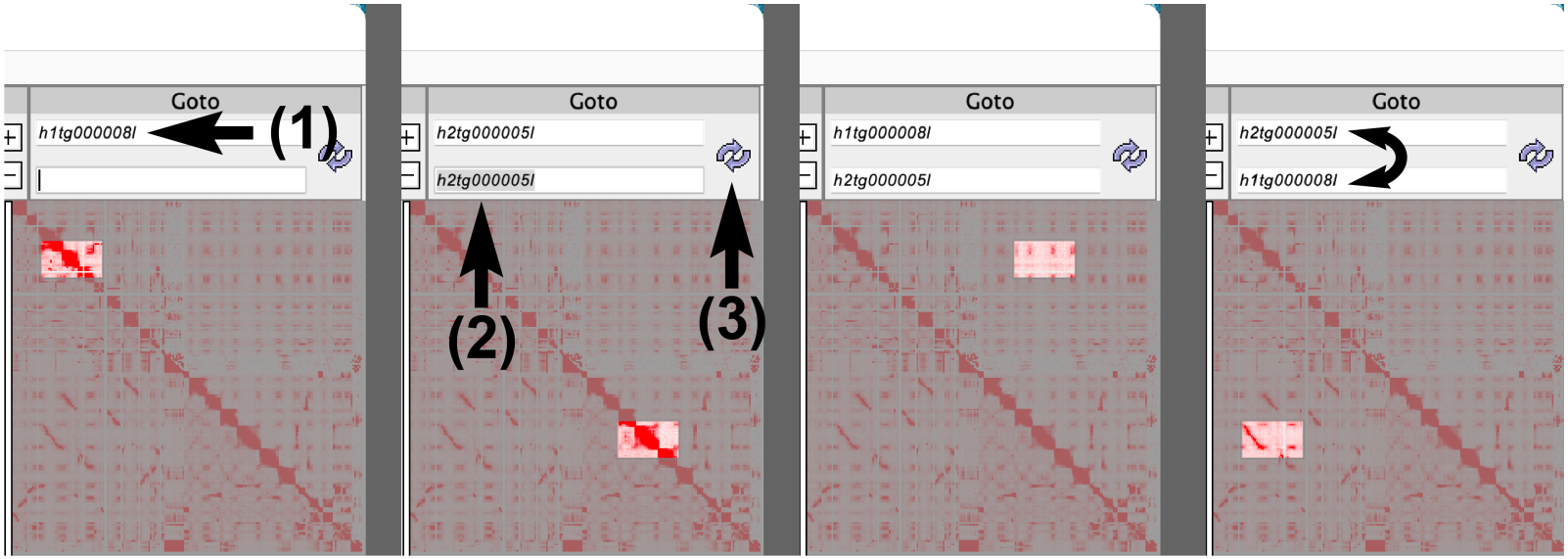
Navigating with 'Jump to Diagonal'



When the user finds themselves far from the main diagonal, they can jump back to safety using the 'Jump to Diagonal' option.

1. Right-click anywhere in the contact map window area to open the **context menu**.
2. Select **Jump to Diagonal**. When the centerpoint of the contact map area is above or to the right of the main diagonal, the context menu will present the user with left-ward and down-ward Jump to Diagonal options; when the centerpoint is below or to the left of the main diagonal, the user will be presented with right-ward and up-ward facing Jump arrow heads.

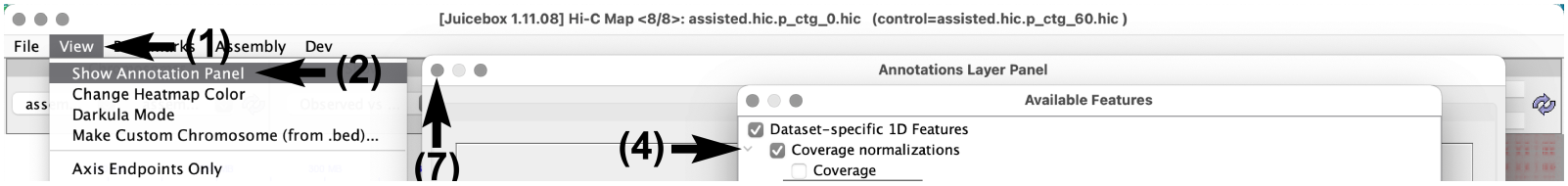
Searching contigs with 'Goto'

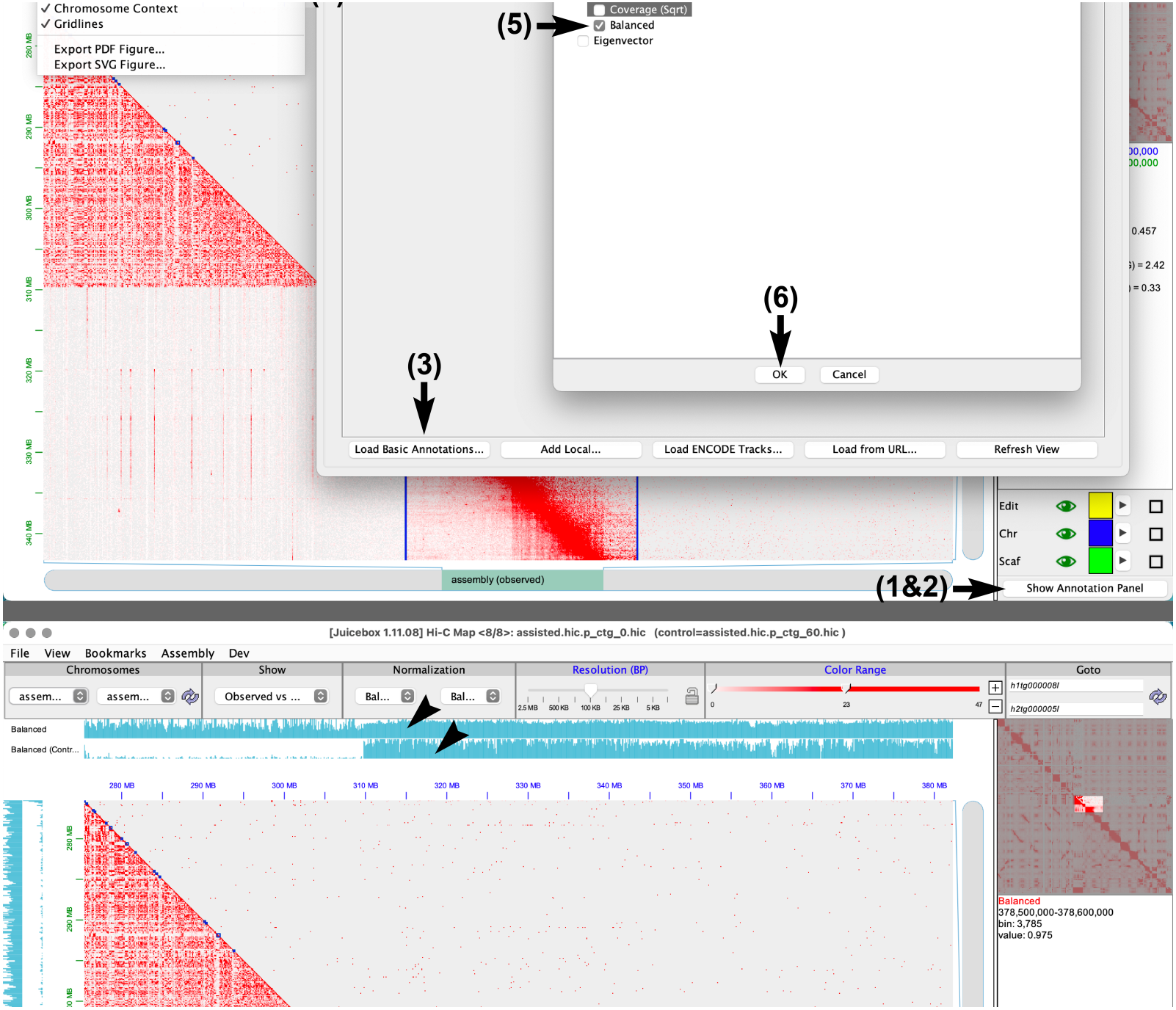


When the user needs to find a particular sequence by name, the 'Goto' subpanel in the view controls bar will be useful.

1. If the user inputs only one sequence name (leaving the other blank) and presses *Return*, the viewing window will jump along the diagonal to the query sequence's position.
2. Inputting the same sequence name into both fields is quivalent to the above.
3. Clicking on the cycling arrows is equivalent to hitting *Return* on the keyboard.
4. When two different sequence names are inputted, the view window jumps to a position off the diagonal. If a sequences are inputted in order of their position along the diagonal, the viewing window jumps horizontally; if inputted in opposite order (second first, first last) then the viewing window jumps vertically.

Displaying coverage tracks





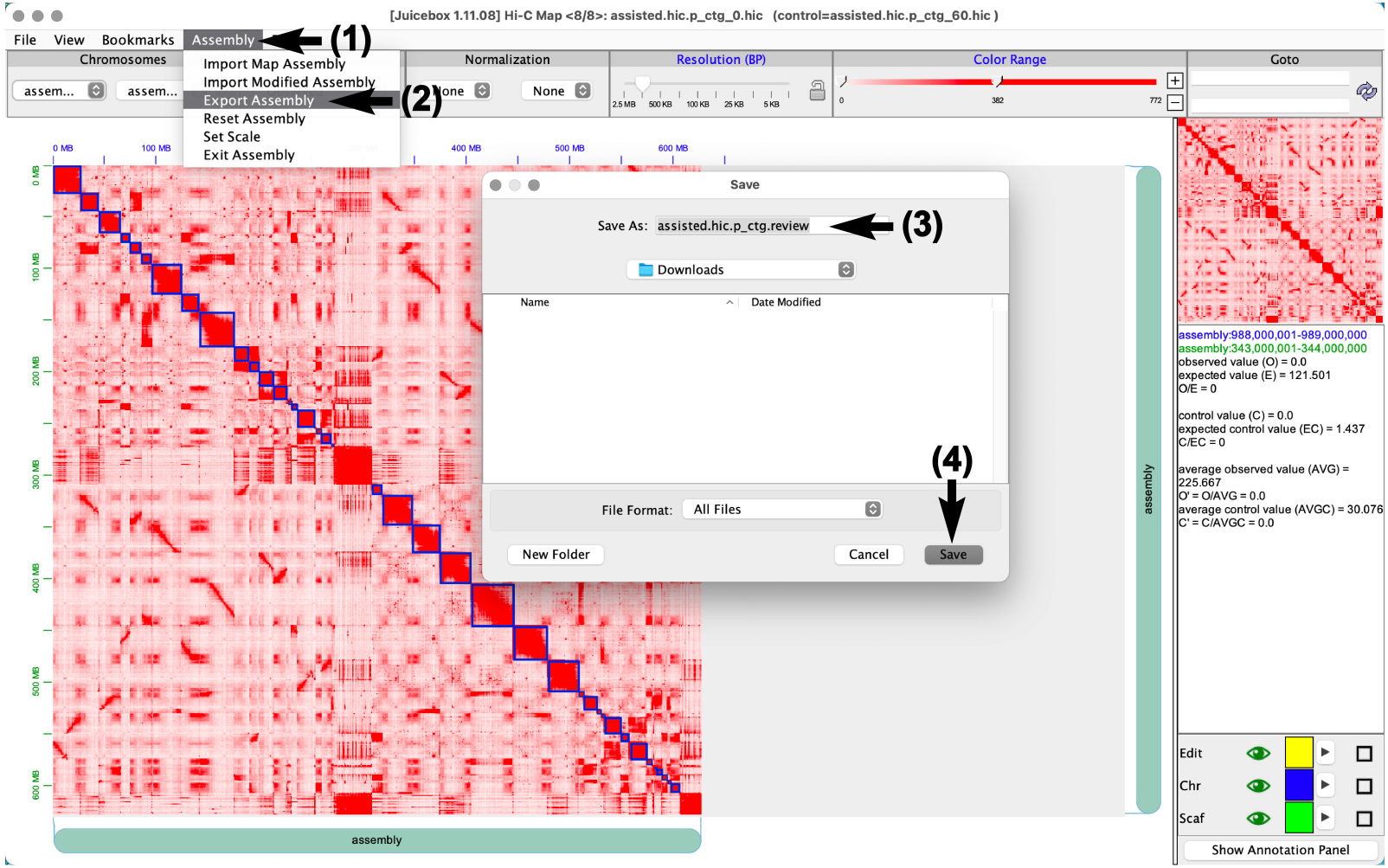
If the user suspects that a sequence has higher- or lower-than-expected depth, it may be useful to examing the Hi-

C read depth of coverage track(s). The coverage track(s) can be enabled with the following steps:

1. Click **View** drop-down tab in the menu bar.
2. Select **Show Annotation Panel**. The annotation panel window will appear, presenting the user with the options to choose 1D or 2D annotations.
3. Click the **Load Basic Annotations...** button under the 1D annotations menu. This will open a second window of available 1D annotations.
4. Click on the grey right-ward facing chevron next to the **Coverage normalizations** node to expand the annotation tree.
5. Click the **Balanced** checkbox to toggle on the coverage track(s).
6. Click **OK**
7. Close the annotation panel window.

TIP: Alternatively, the user can click the **Show Annotations Panel** on the bottom of the right-side panel to open the annnotation panel, then perform steps 3-7 above.

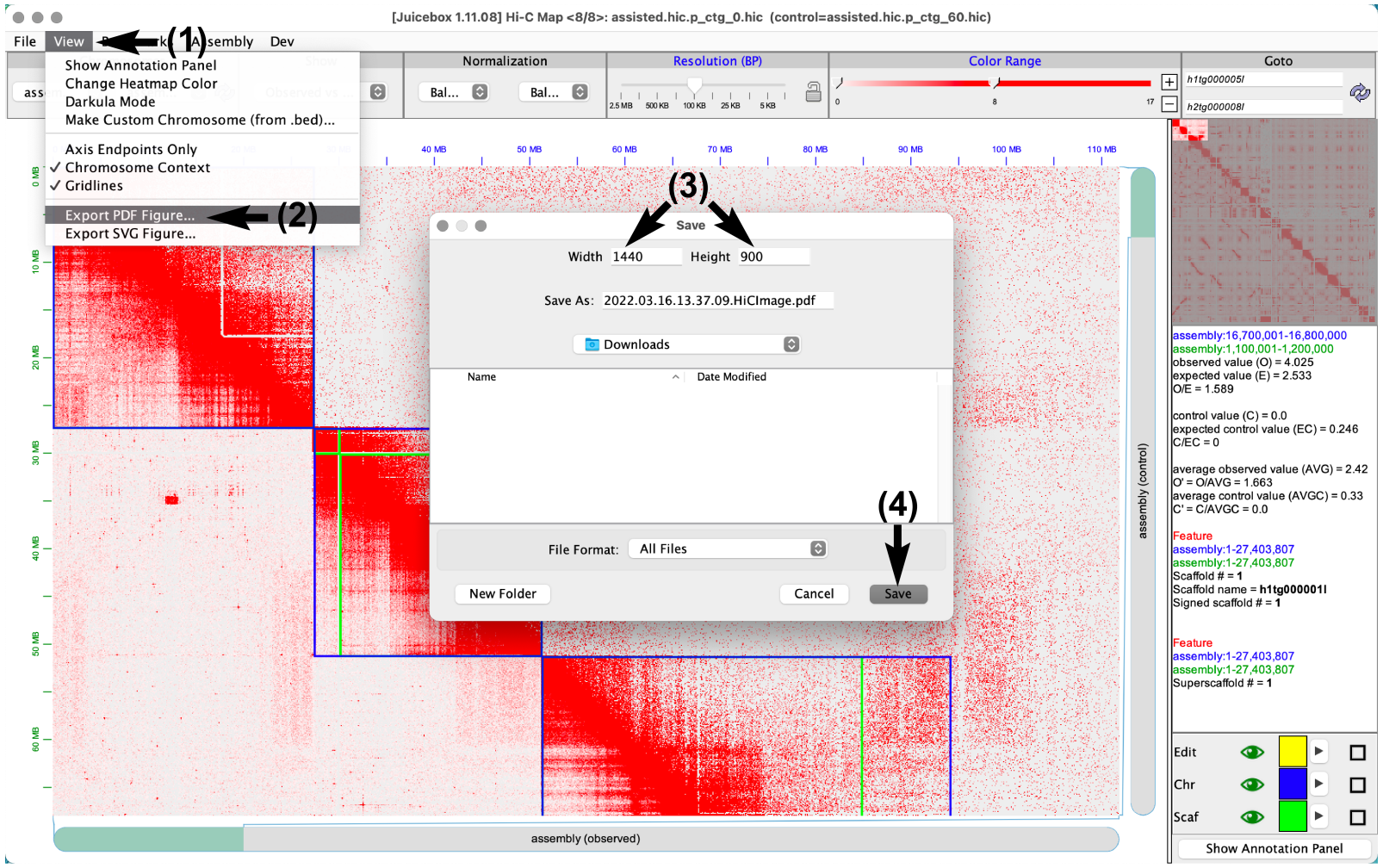
Exporting a modified (reviewed) .assembly file



Save often--after every confident assembly change--becuase Juicebox can crash at unpredictable times, depending on computer RAM, file sizes, genome size, etc.

1. Click on the **Assembly** drop-down tab in the menu bar.
2. Select **Export Assembly**. A 'Save' window will appear.
3. Accept the suggested **Save As** file name (recommended while actively modifying an assembly), or input your own (when finished with curation).
4. Click **Save**

Exporting a PDF/SVG file

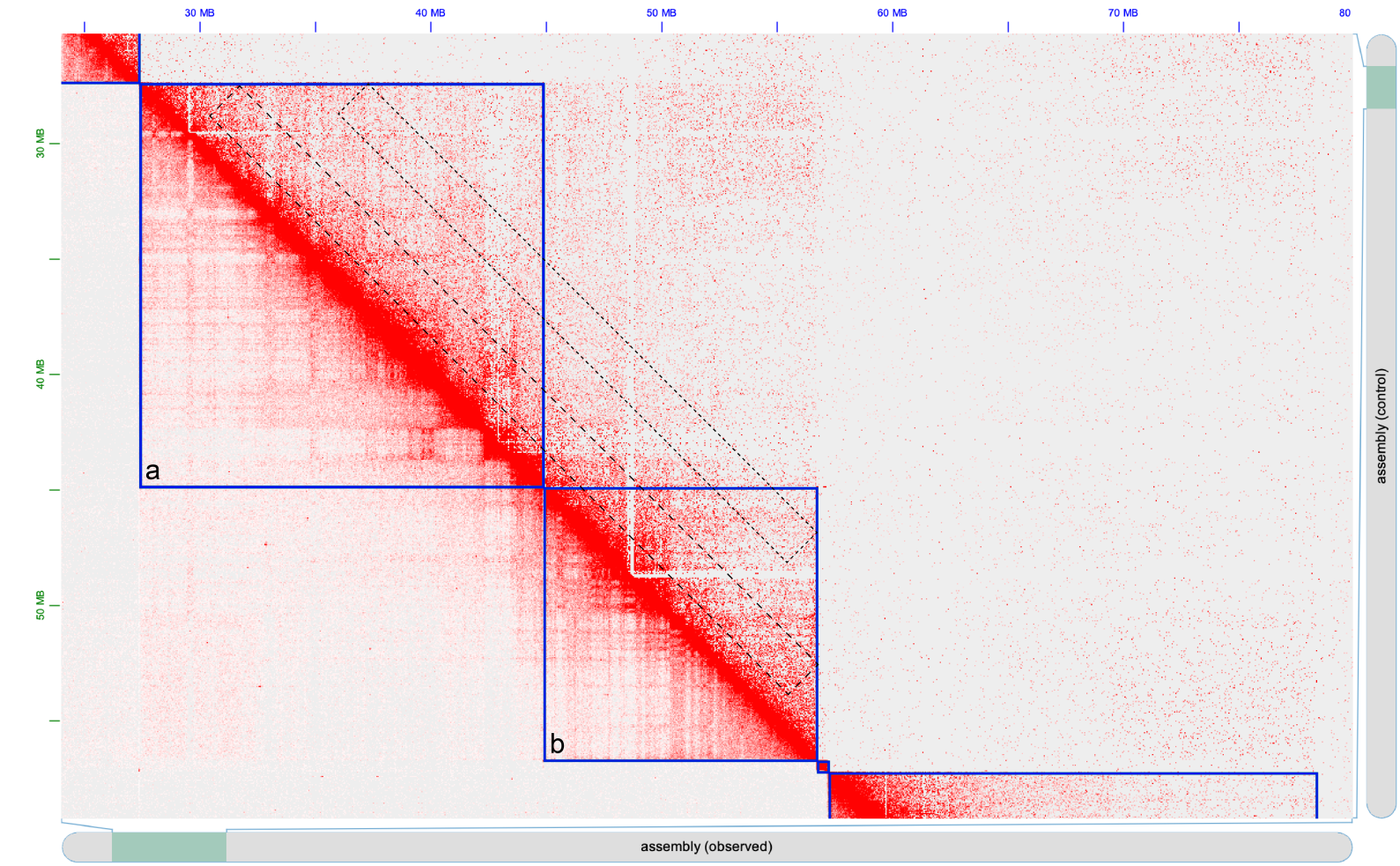


1. Click on the **View** drop-down tab in the menu bar.
2. Select **Export PDF Figure** or **Export SVG Figure**. A 'Save' window will appear.
3. In the **Width** and **Height** input fields, the current screen dimensions (in pixels) is filled-in by default. If the contact map is larger than the screen, increase the dimensions of the figure as desired (this may take some trial-and-error). Optionally, change the output file name in the *Save As* input field.
4. Click **Save**

Common Hi-C contact patterns

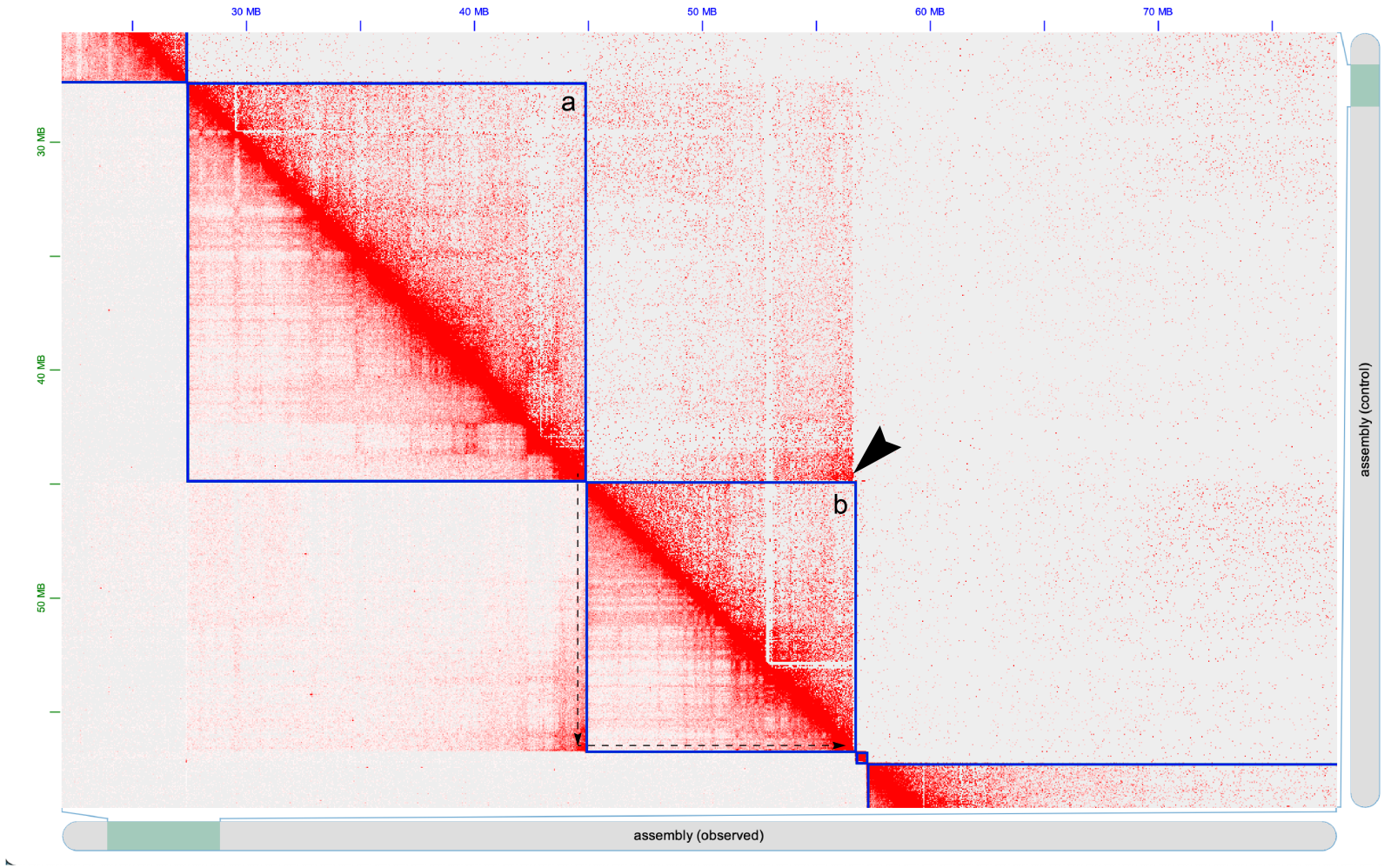
An handful of common patterns associated with misassemblies are described in pages 35-36 of the Genome Assembly Cookbook, but a more exhaustive range of patterns is catalogued here. In each of the following figures, a Hi-C chromatin contact density heatmap (a.k.a., a 'contact map') is visualized with Juicebox. Contact densites are proportional to pixel intensities (red represents dense contacts, white as sparse). A permissive (MapQ0) 'Observed' contact map is plotted below the diagonal and a stringent (MapQ >= 60) contact map plotted above as 'Control'.

Contigs in correct order and orientation



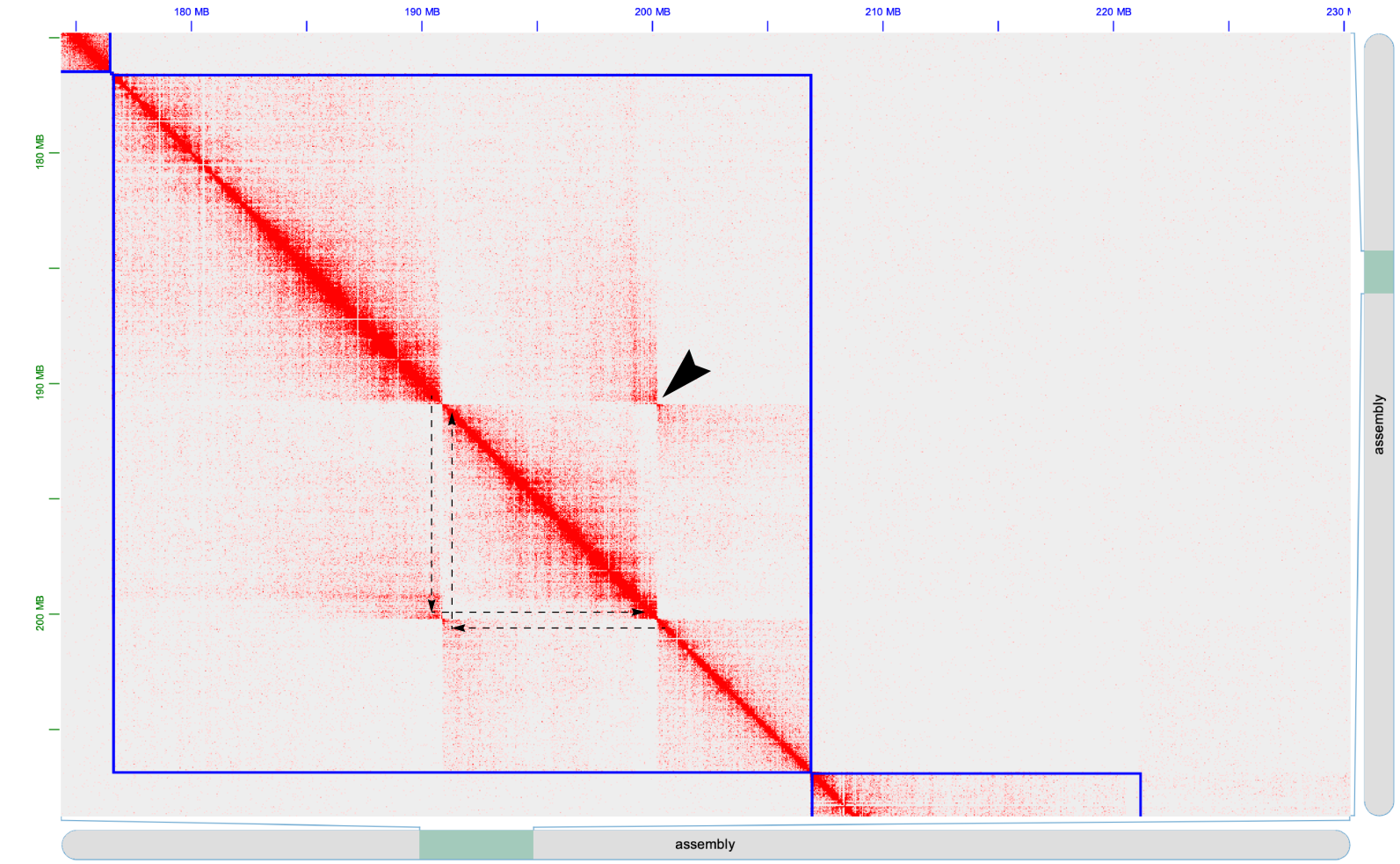
A contact map demonstrating two adjacent contigs (*a* and *b*) in correct relative order and orientation. The structural correctness of this arrangement is supported by the same contact patterns extending between contigs as can be observed within contigs. Compare densities at different points within each of the dashed-outline and dotted-outline boxes. Note the strongest contacts between contigs *a* and *b* is along the diagonal, between the lower-right corner of *a* and upper-left corner of *b*). Contig boundaries are denoted here with blue boxes.

Single-contig inversion



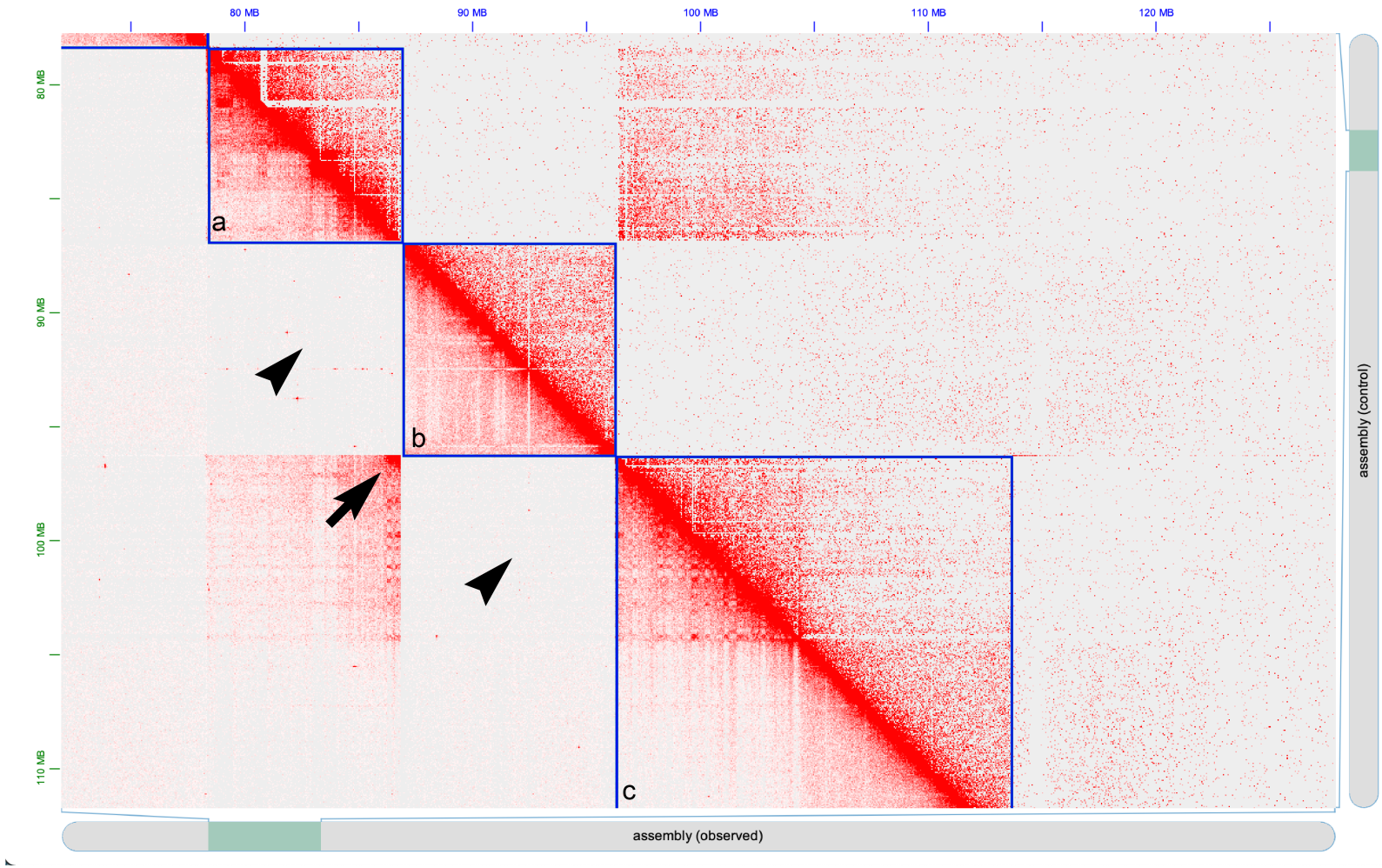
A contact map illustrating an inversion of one contig (*b*) relative to the other (*a*) contig. This is apparent from the high density of shared contacts (black arrow head) between the lower-right corner of the *a* contig and the lower-right of the *b* contig (dashed guide arrows shown). Contig boundaries are denoted here with blue boxes and contact density is proportional to pixel intensity (red represents dense contacts, white represents sparse contacts).

Internal inversion



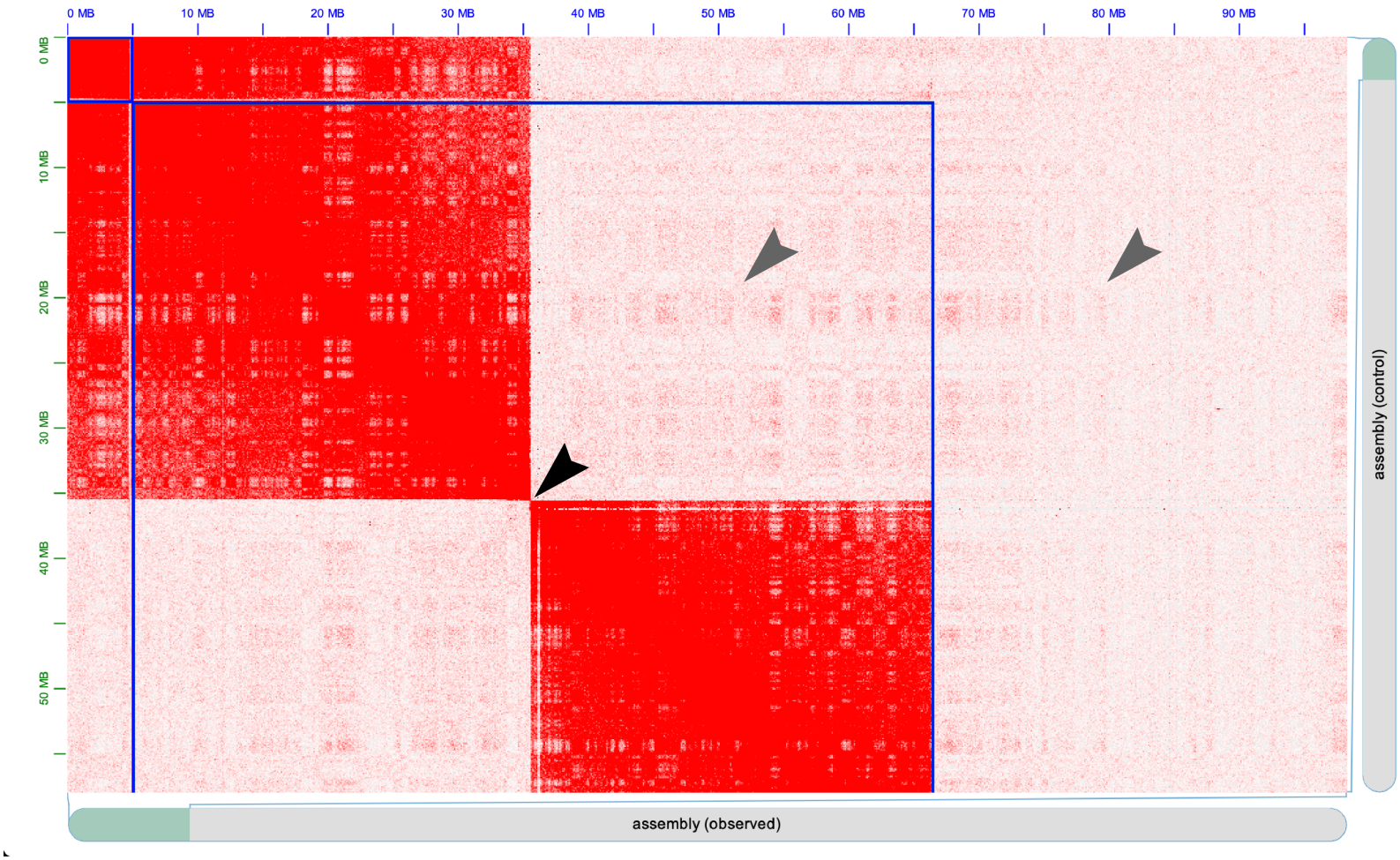
A contact map revealing an inversion within a contig (blue box). This structural difference is apparent from the distinct 'bowtie' contact pattern (black arrow head) resulting from contacts with otherwise correctly ordered and oriented flanking subsequences (dahsed guide arrows shonw). Such inversions may be due to haplotypic differences––when the assembly represents a single-haplotype mosaic of the underlying genome––or misassembly––which is more likely with a phased assembly or inbred genome.

Mis-ordered contigs



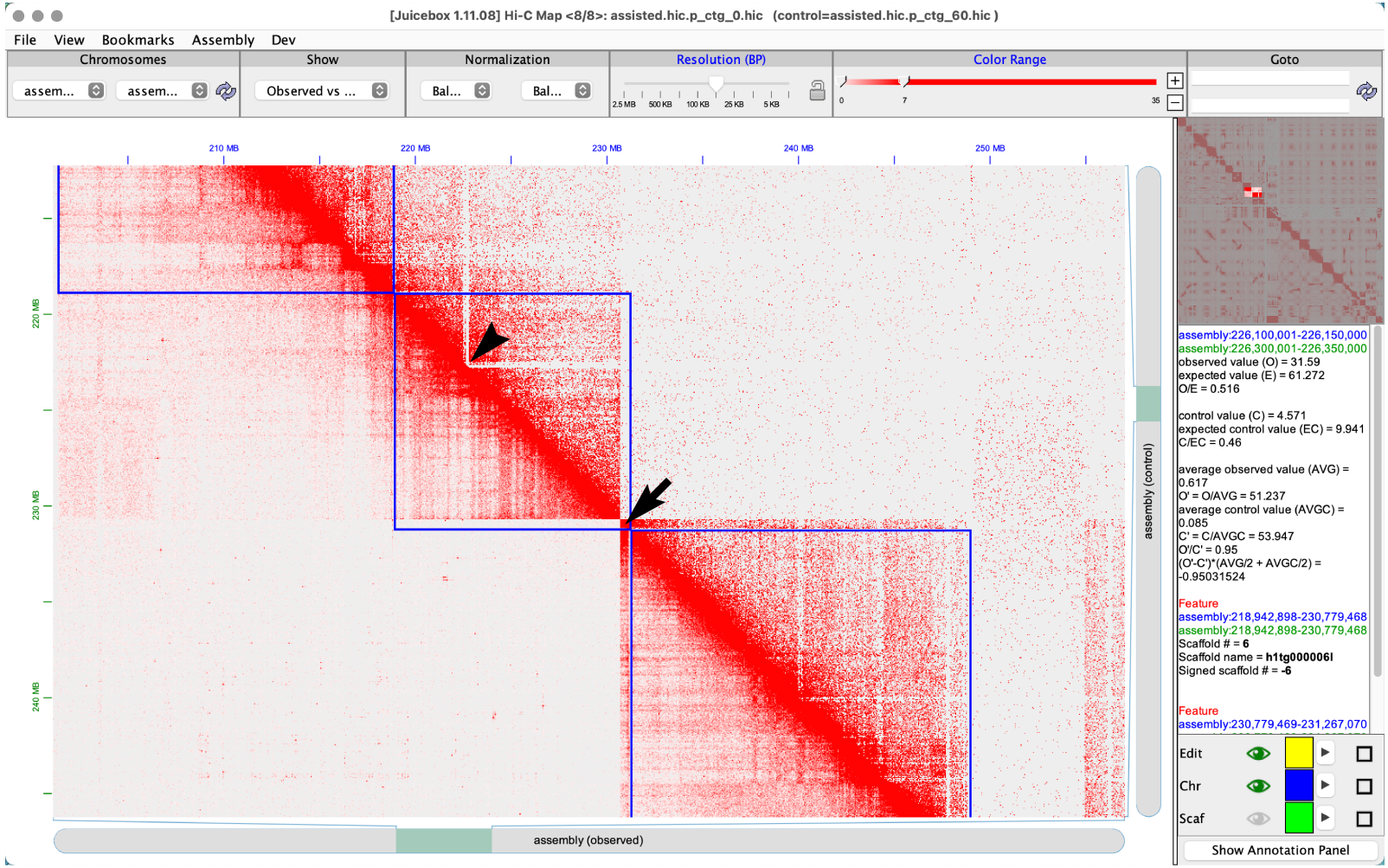
A contact map demonstrating two contigs (*a* and *c*, blue boxes), adjacent to one another in their underlying linear genome sequence, interrupted by the insertion of another contig (*b*). Note the strongest contacts between the *a* and *c* contigs (black arrow) flanks the *b* contig, which has few contacts with the two surrounding contigs (black arrow heads). Permissive (MapQ0) contact map plotted below the diagonal, and stringent (MapQ60) contact map plotted above as control; Contact density is proportional to pixel intensity (red represents dense contacts, white represents sparse contacts).

Contigging misjoin error



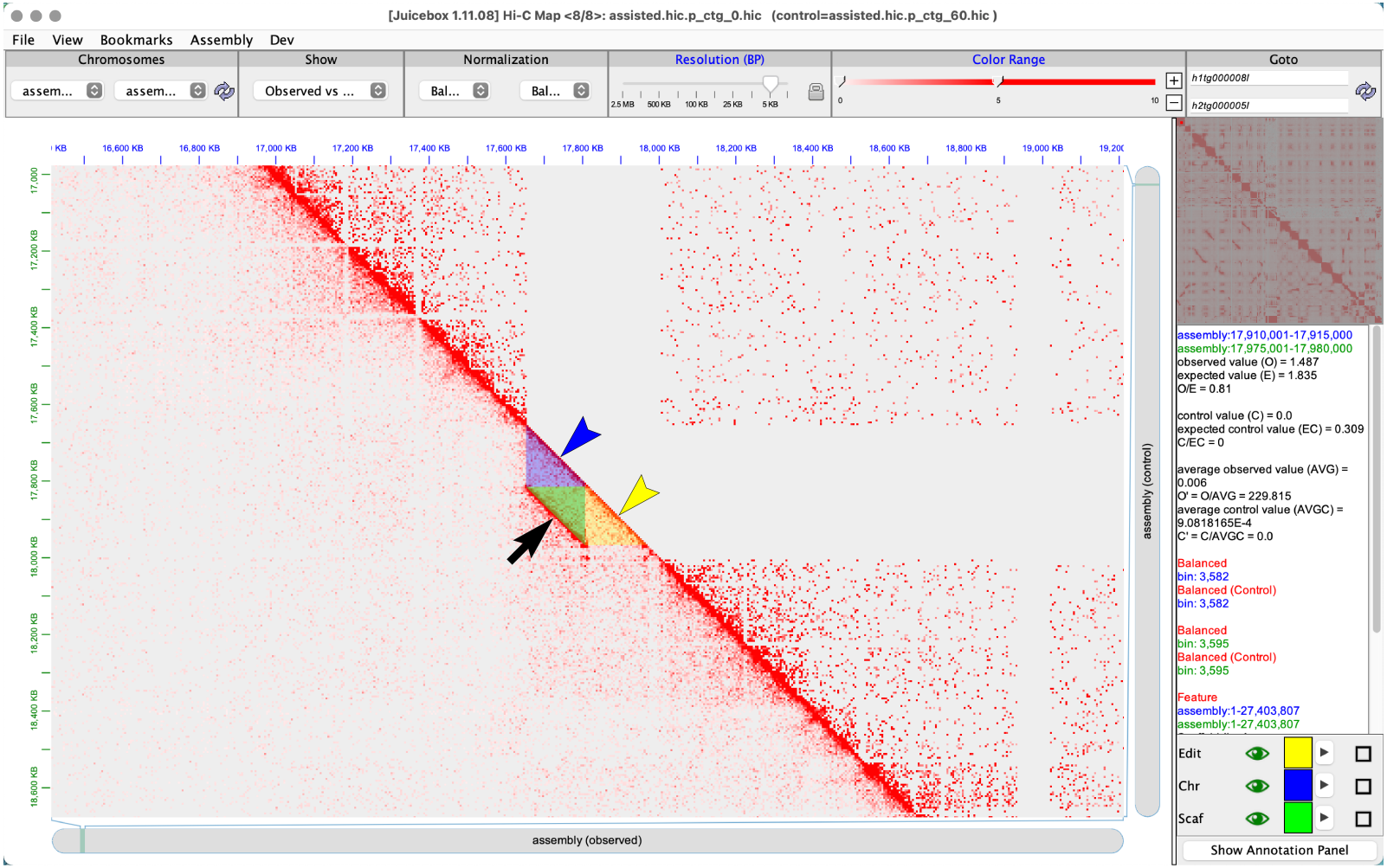
A contact map revealing an assembly misjoin (black arrow head) roughly in the middle of the contig (blue box). The structural error is discernable by the way the upper and lower halves of the contig have strong contacts within themselves (upper-left and lower-right quadrants), but the contact density between contig halves is no more dense than background (compare densities between grey arrow heads). Contact density is proportional to pixel intensity (red represents dense contacts, white represents sparse contacts).

Contigging misjoin error with positive evidence



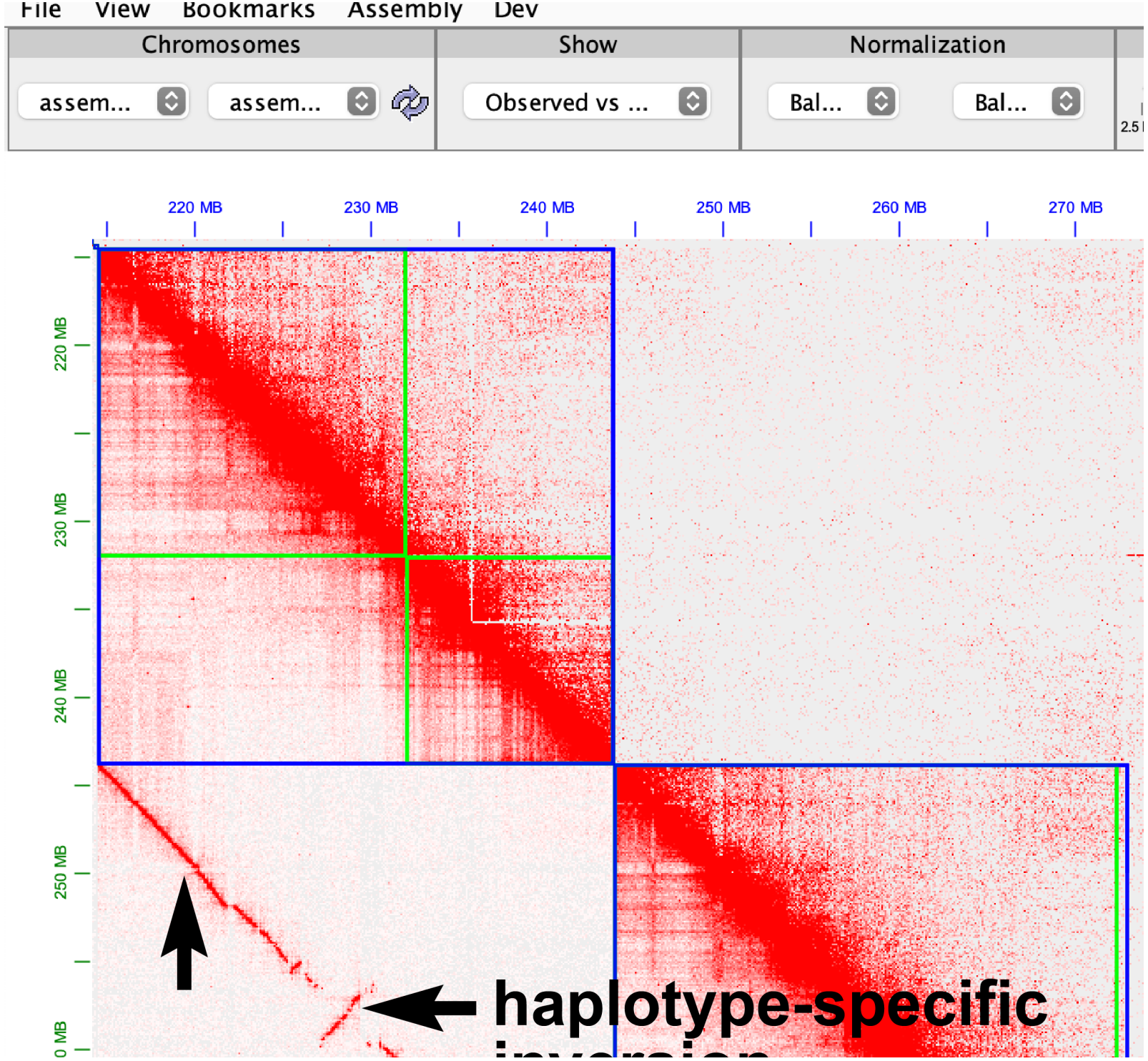
A contact map similar to that described above; in this case, however, positive evidence of a misjoin error is observed between the bottom-right subsequence (black arrow) of a large contig (center blue box) as strong shared contacts with a second large sequence (bottom blue box). Note that the large subsequence of the center contig and bottom contig do not share similarly dense contacts. It is good practice to find positive evidence of a misjoin than break contigs at any horizontal/vertical low-density stripe (black arrow head), which can also be artifacts of repetitive subsequences within an otherwise correctly-assembled sequence.

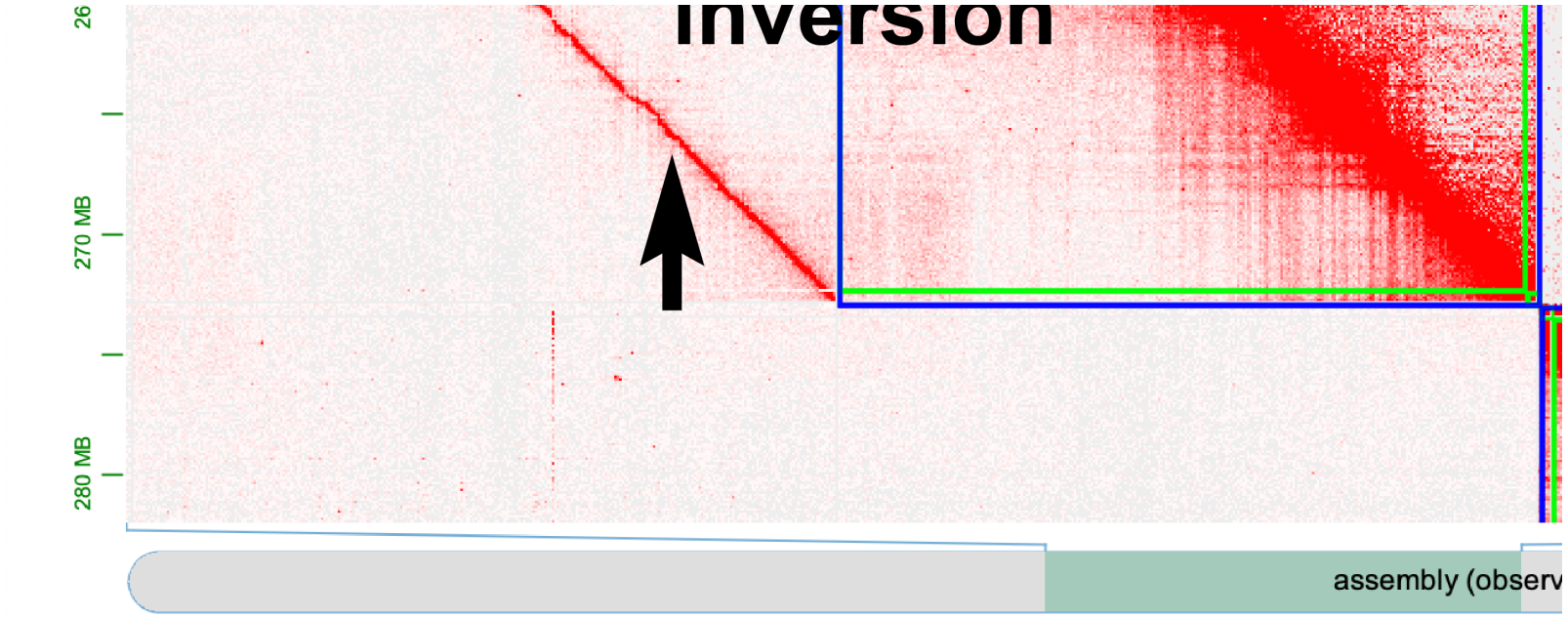
Tandem segmental duplication



A contact map showing the presence of a two-copy segmental tandem duplication. Here, a duplicate segment was inserted into the chromosome in the same orientation as the original segment, which is apparent by the line of contacts (black arrow) parallel to the main diagonal. The two tandem copies are shown with blue and yellow arrow heads.

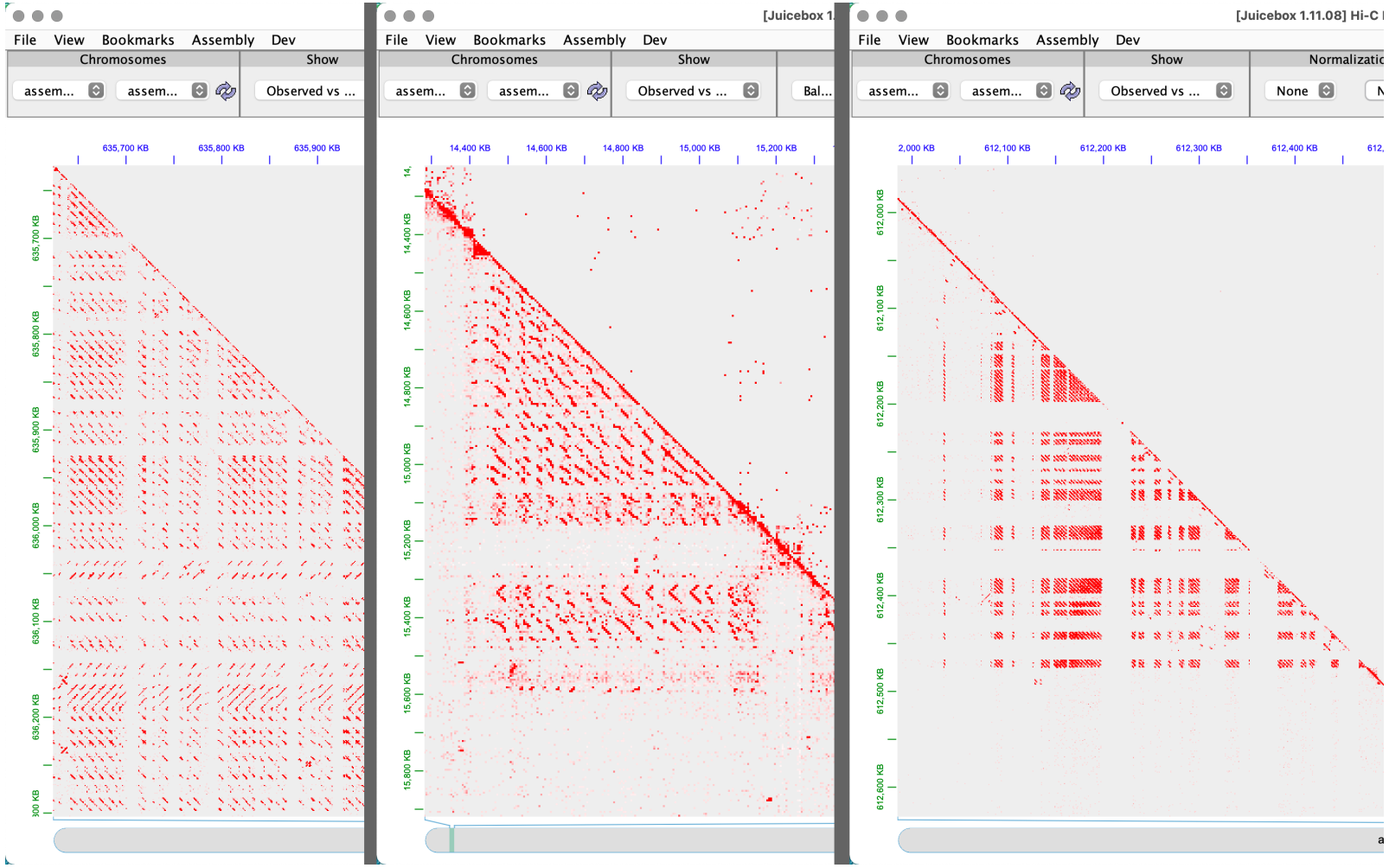
Homologous sequences


File View Bookmarks Assembly Dev
Chromosomes
Show
Normalization
assem...
assem...
Observed vs ...
Bal...
Bal...
220 MB
230 MB
240 MB
250 MB
260 MB
270 MB
haplotype-specific



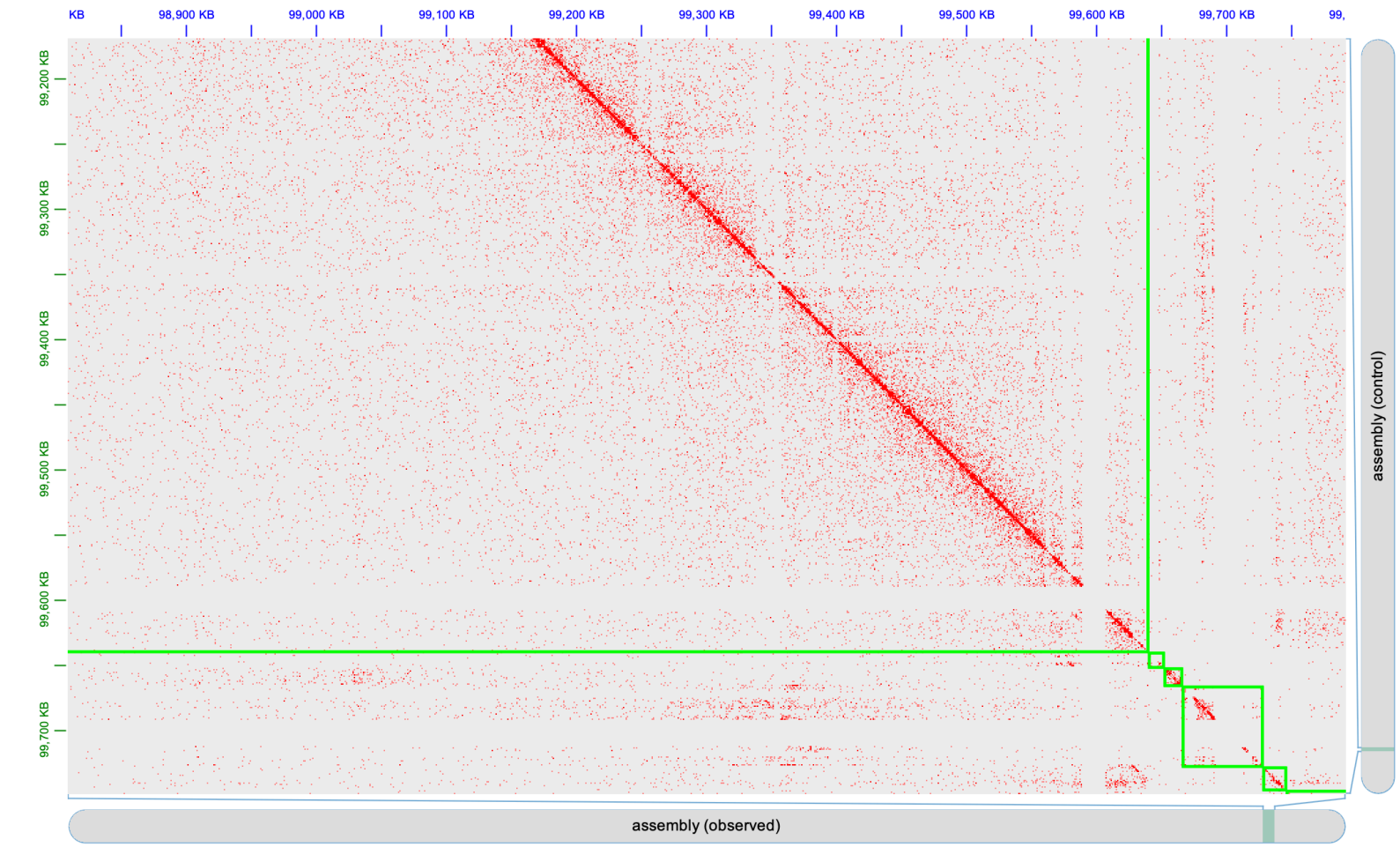
A contact map showing two homologous chromosomes (blue boxes) in collinear orientation. Note the line of 'contacts' parallel to, and below, the main diagonal (upward-facing black arrows; notably, an internal inversion can also be seen in one homolog relative to the other). This line is the result of Hi-C read pairs cross-mapping between haplotypic sequences that share local similarity. The line of contacts is present below the main diagonal (MapQ0) and absent above (Map60) because the mapping qualities of those reads is low, reflecting the uncertainty in their mapped positions.

Repetitive sequences



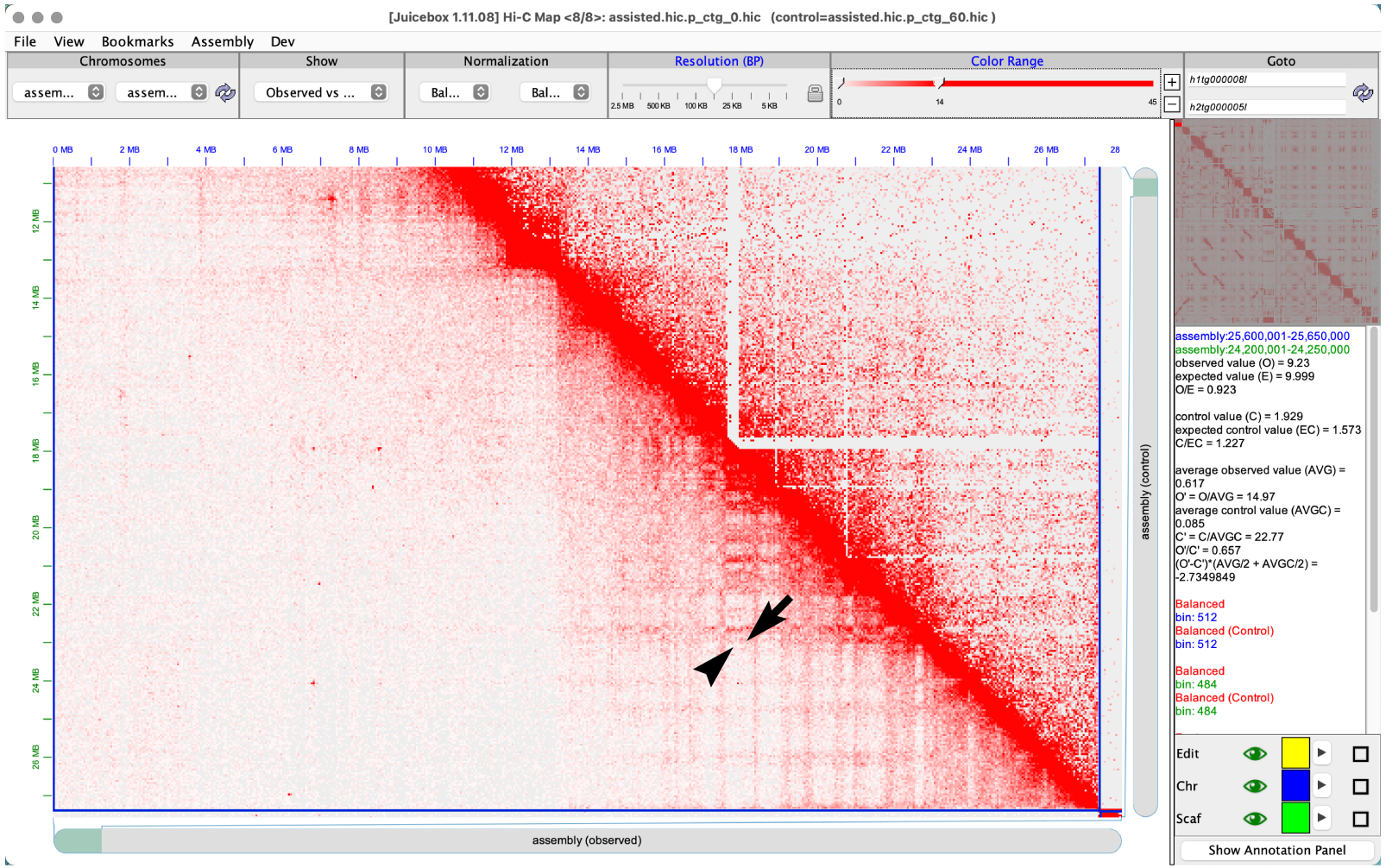
A contact map exposing the repetitive nature of the underlying sequences. Repetitive sequences are characterized by the many diagonal (or anti-diagonal) lines repeating in the horizontal and vertical dimensions. Note the presence of permissively-mapped (MapQ0) 'contacts' (actually, mismapping) below the diagonal that are absent above the diagonal (MapQ60), revealing the poor mappabilty of these sequences.

Redundant sequences



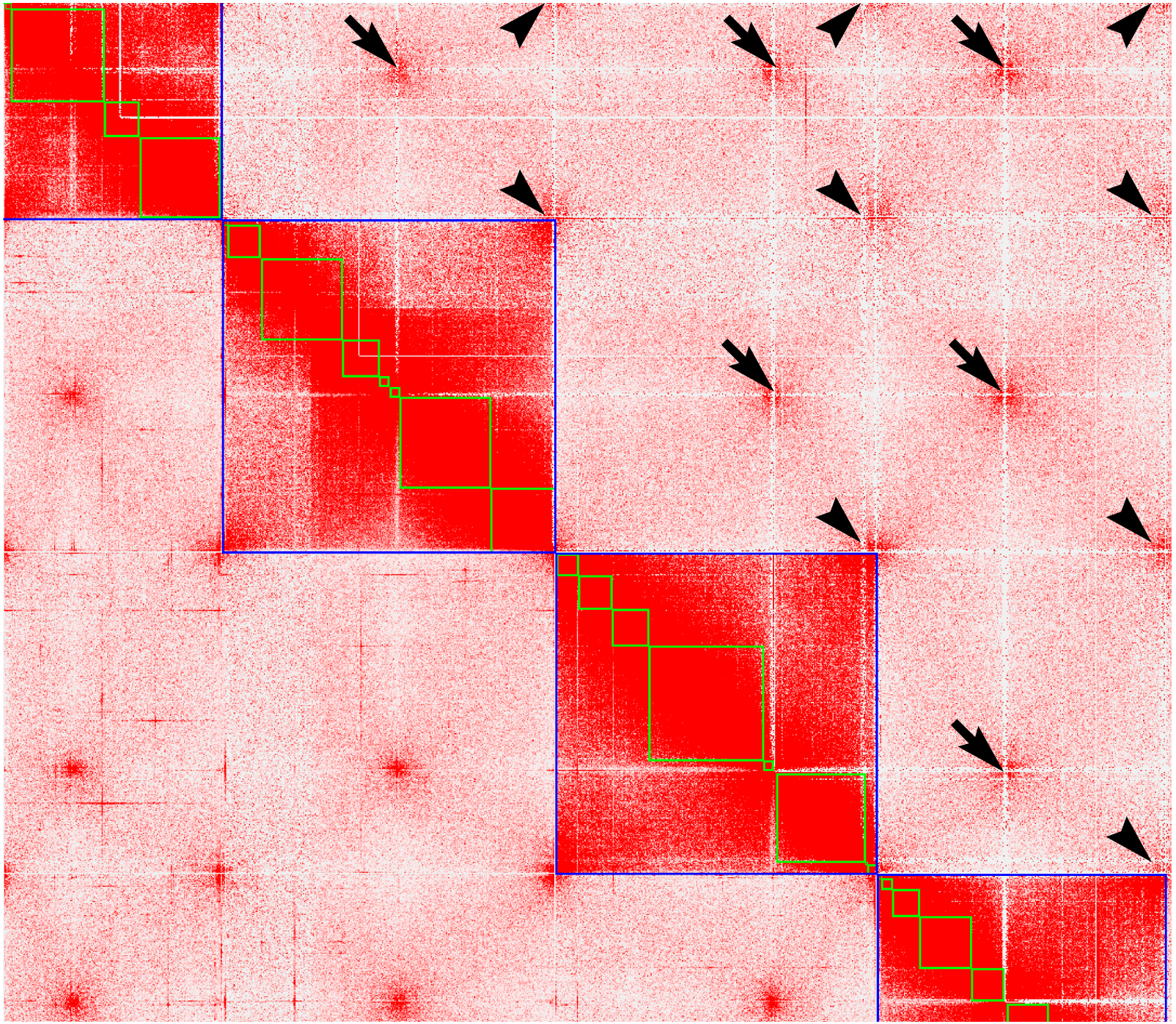
A contact map showing haplotypic sequence redundancy. This view shows a collection of four small, redundant sequences ordered after one larger sequence (green boxes) along the diagonal. See that the densest contacts shared between the largest and smaller contigs localize below the largest contig (below the diagonal; above the diagonal, to the right of the largest contig), within the extent of the contig, indicating that the smaller sequences best map internal to the largest sequence (the smaller contigs are redunant to internal subsequences of the large contig). Permissive (MapQ0) contact map plotted below the diagonal, and stringent (MapQ60) contact map plotted above as control; Contact density is proportional to pixel intensity (red represents dense contacts, white represents sparse contacts).

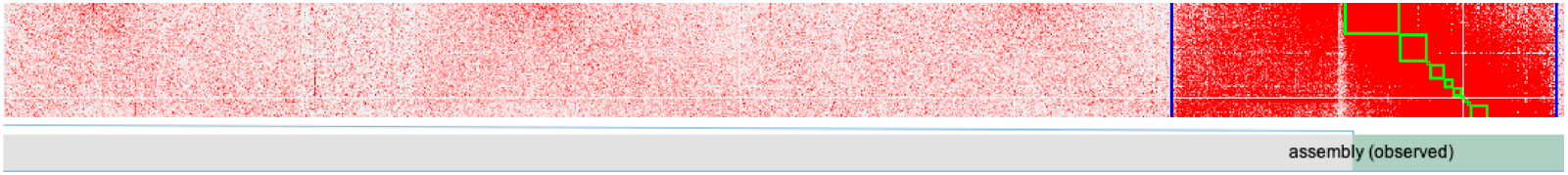
A/B compartmentalization



A contact map with a plaid/checkerboard pattern of alternating high (black arrow) and low (black arrow head) contact densities, illustrating the biological phenomenon of A/B chromatin compartmentalization. Here, chromatin is bundled into small subcompartments that alternate between 'A' active and 'B' repressed transcriptional states (see inset). Because many chromatin bundles (domains) may occupy the same subcompartment, these sequences contact more frequently than more linearly-adjacent loci, creating the plaid/checkerboard pattern.

Centromere-centromere and telomere-telomere clustering





> A contact map of inter-chromosomal contacts between four chromosome-scale scaffolds (blue boxes). In many species, centromeric loci can be approximated via centromere-centromere clustering contact puncta (black arrows). Likewise, telomere-telomere clustering (arrow heads) can be observed.

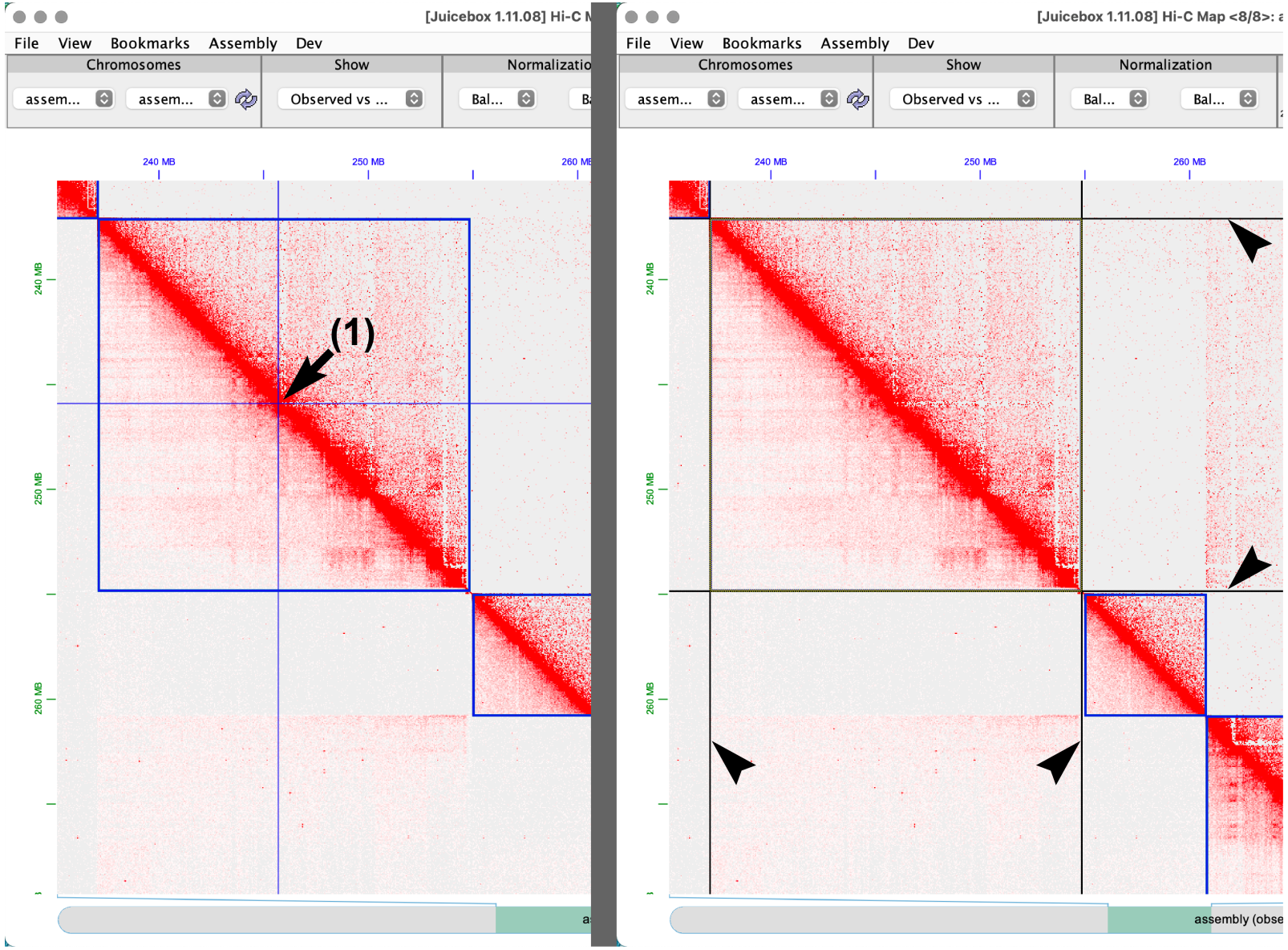
Editing assemblies

Useful shortcuts

- **Undo**: *Ctrl* + *u* (*command* + *u* on a Mac)
- **Redo**: *Ctrl* + *r* (*command* + *r* on a Mac)

(De-)Selecting sequences

Selecting one sequence

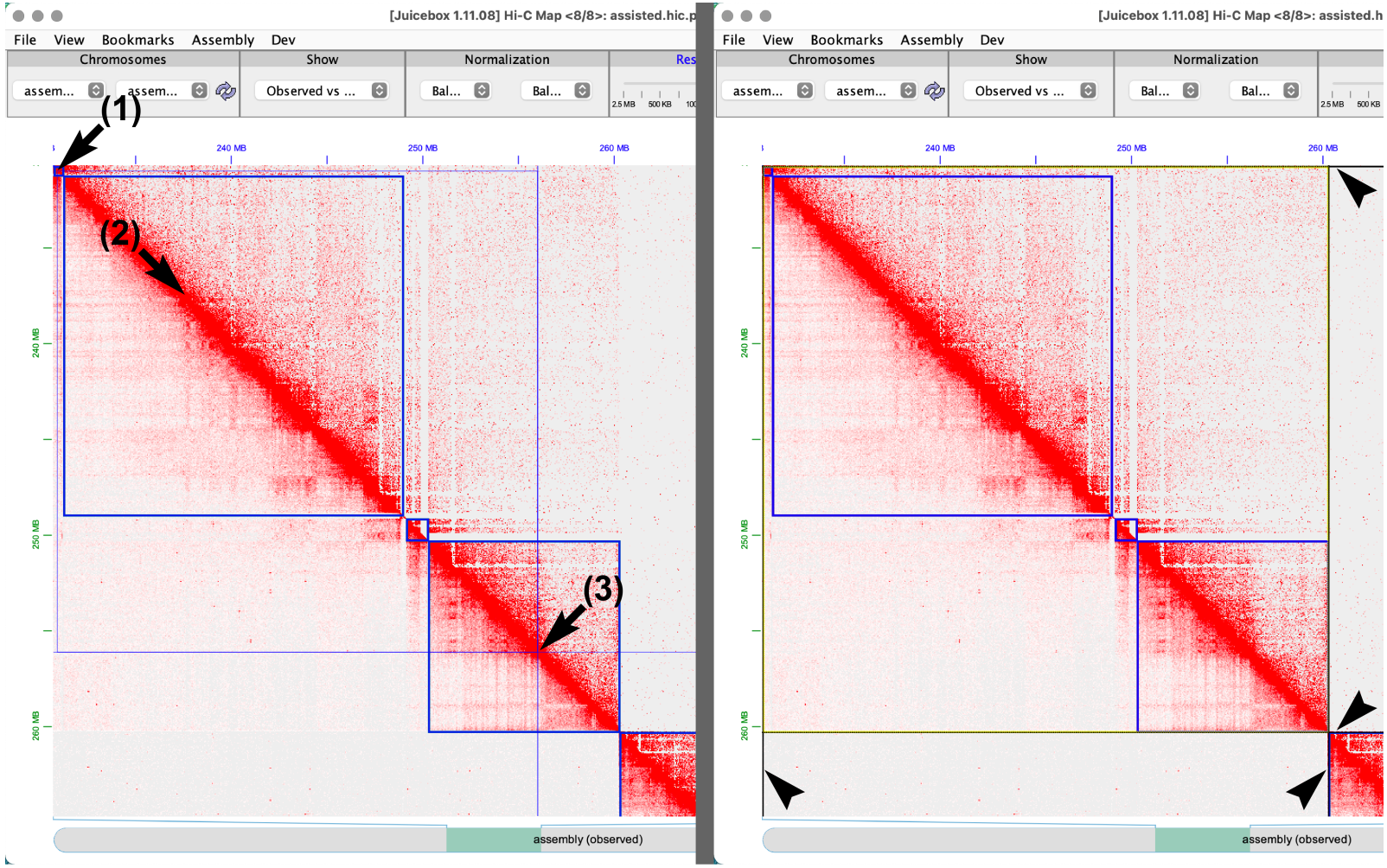


1. Hold the *shift* key (the blue horizontal and vertical straight-edge lines will appear), then single-click with the left mouse button the sequence to be selected. Release the *shift* key). The sequence selected will become bracketed by horizontal and vertical black lines that extend in the horizontal and vertical directions (black arrow heads) from the selected sequence. These act as useful guides when scrolling through the contact map.

De-select one or more sequence(s)

1. Press-and-hold the *shift* key.
2. Click the contact map area anywhere outside of the scaffold (green) or chromosome (blue) boundary boxes.
3. Release the *shift* key.

Selecting multiple sequences

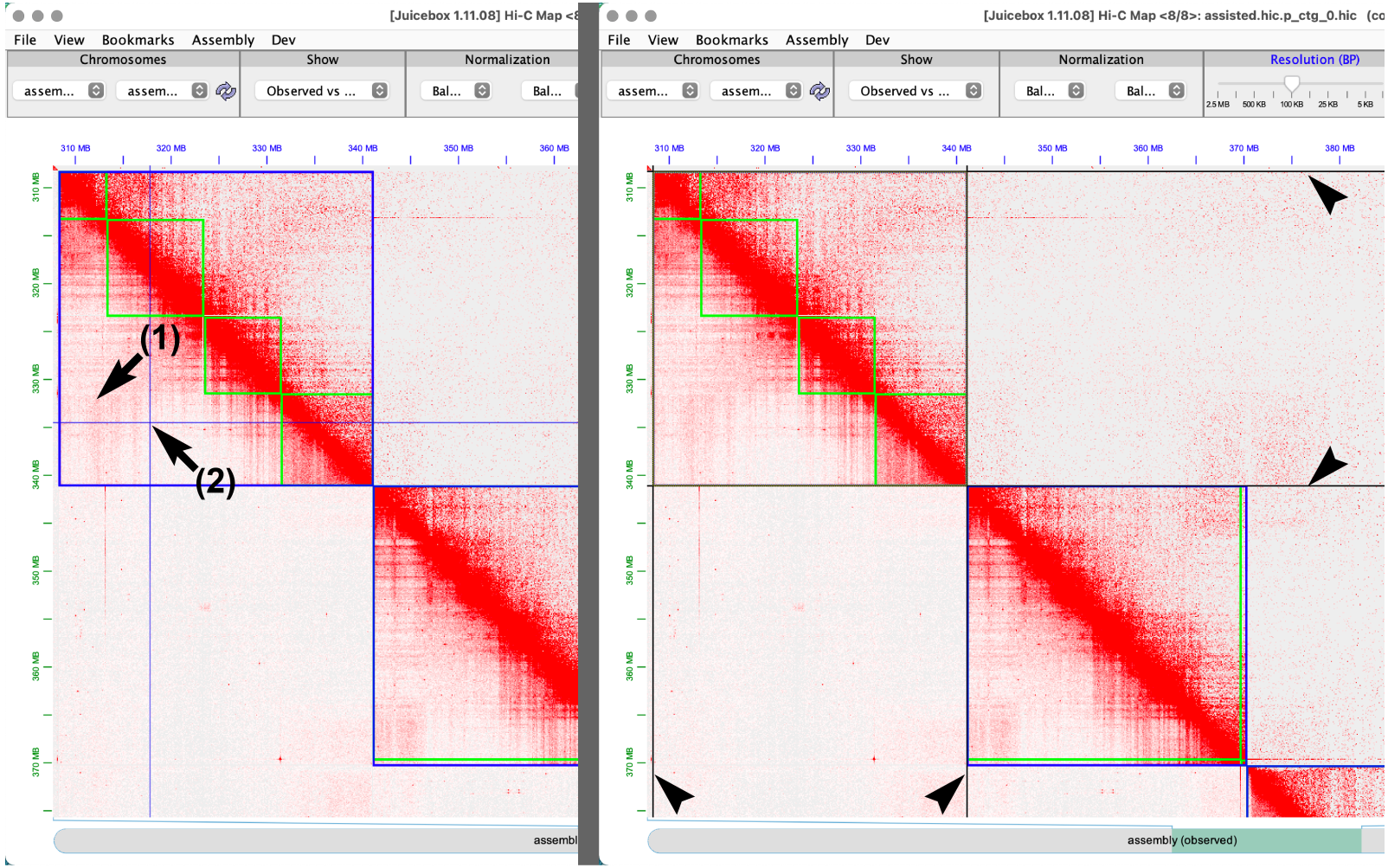


1. Hold the *shift* key (the blue horizontal and vertical straight-edge lines will appear), then press-and-hold the

left mouse button on the first of the contigs to be selected.

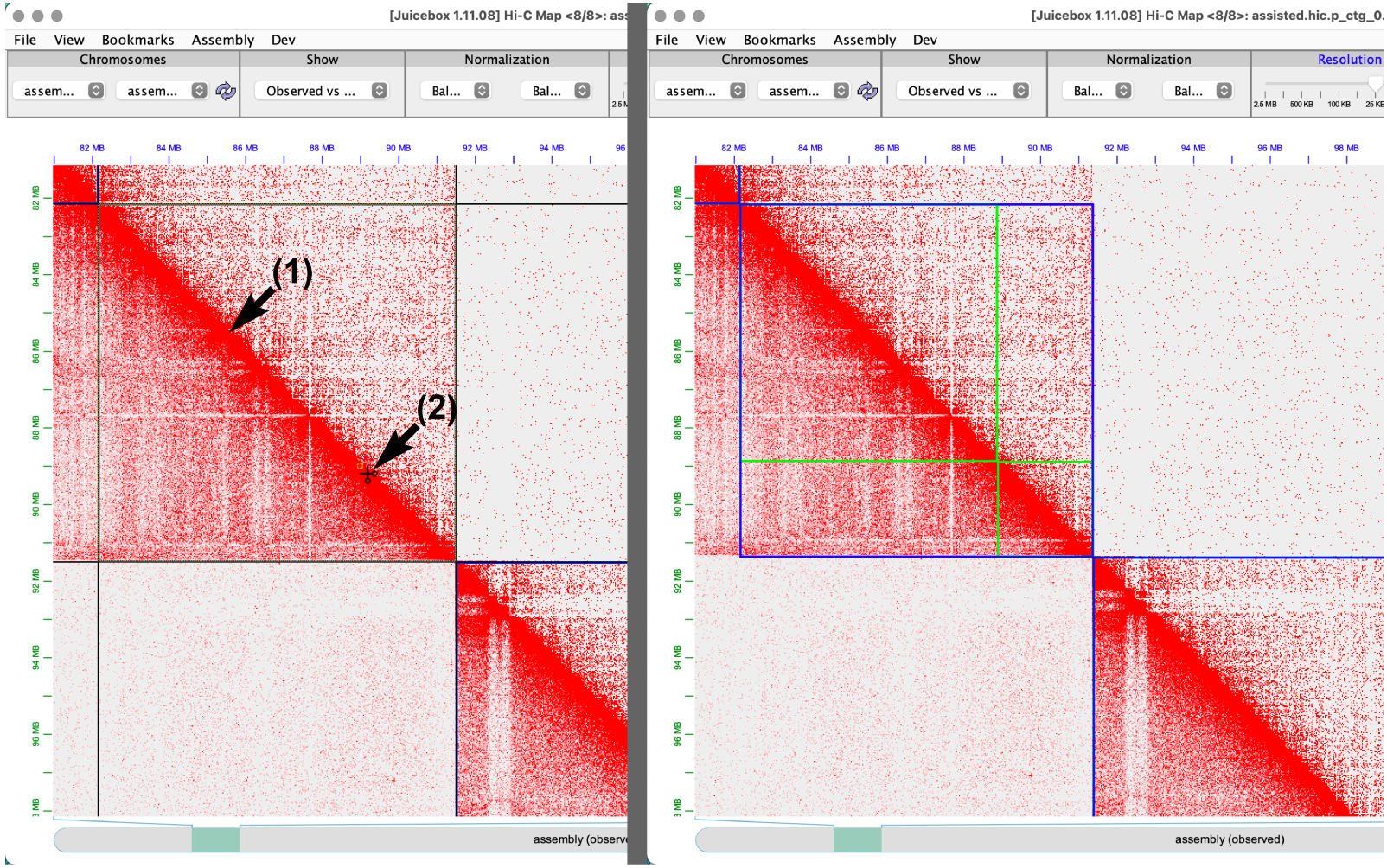
2. Drag the mouse cursor along the diagonal to the inside the last of the contigs to be selected.
3. Release the left mouse button. Similarly to when selecting a single contig, the contigs will become bracketed by horizontal and vertical black lines that extend in the horizontal and vertical directions (black arrow heads) from the selected sequences.
4. Finally, release the *shift* key.

Selecting all sequences in a chromosome



1. Hold the *shift* key (the blue horizontal and vertical straight-edge lines will appear),
2. Click anywhere within a blue box (but outside of any green box). Similarly to when selecting a single contig, the contigs will become bracketed by horizontal and vertical black lines that extend in the horizontal and vertical directions (black arrow heads) from the selected sequences.
3. Release the *shift* key.

Cutting/breaking a contig

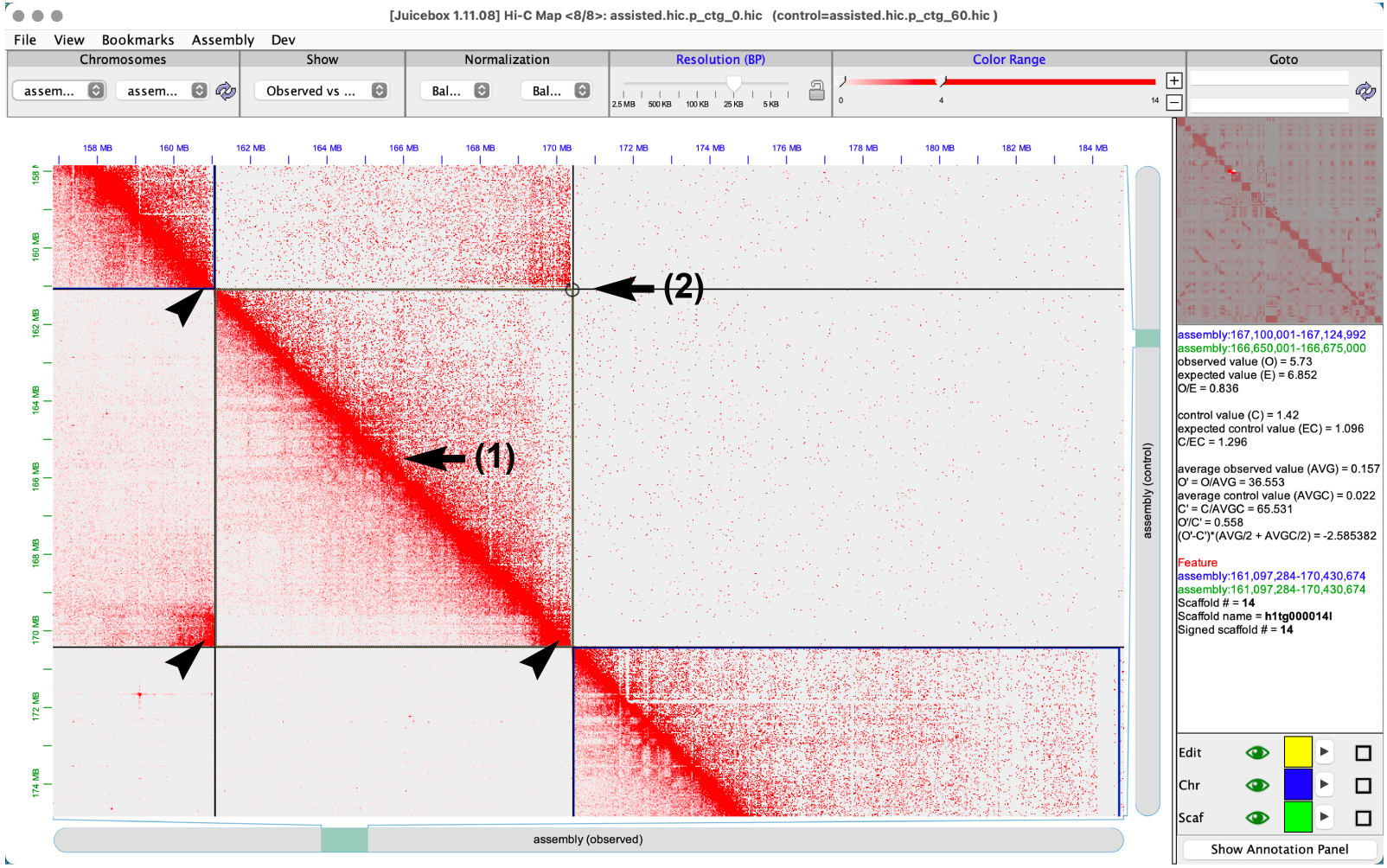


1. Select (as described previously) the misjoined sequence to be cut/broken.
2. Hover the mouse cursor over the main diagonal until the cursor changes to a pair of small black scissors with a yellow box between the blades. This yellow box is the cutting region, and its size can be increased or decreased with the center mouse wheel (or two-finger scroll on a Mac tackpad).
3. Click on the position along the diagonal where a cut/break is desired.

After breaking, what appears to be left are two new sequences. In reality, the original sequence is cut into *three* sequences. These sequences will have `:::fragment_#` appended to their original sequence name; additionally, the sequence from the cut region will have a `:::debris` tag appended. For example, performing a single cutting operation on a sequence named `contig004` will produce the following three new sequences:

1. The subsequence of `contig004` above/left-of the cut named `contig004:::fragment_1`.
2. The subsequence of `contig004` within the cut region named `contig004:::fragment_2:::debris` (which is sent to the 'debris' stack at the end of the file).
3. The subsequence of `contig004` below/right-of the cut named `contig004:::fragment_3`.

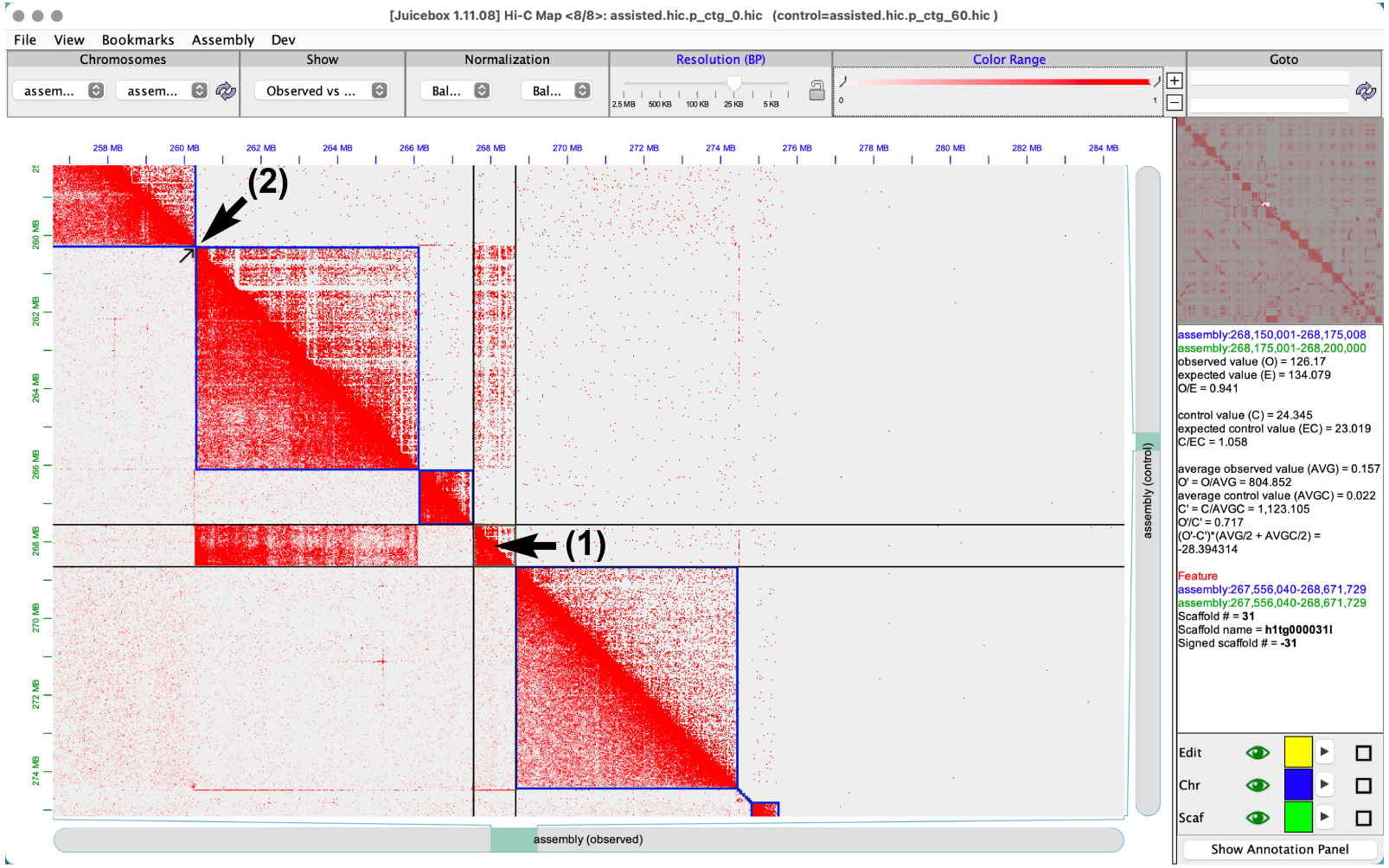
Inverting contig orientation



1. Select (as described previously) the misoriented sequence to be inverted.
2. Hover the mouse cursor over the upper-right or lower-left corners of the selected sequence. The cursor will change into a small black ouroboros-like arrow.
3. Click once on the upper-right or lower-left corner with the left mouse button. The contig will then invert.

Moving misplaced sequences

Moving sequences short distances

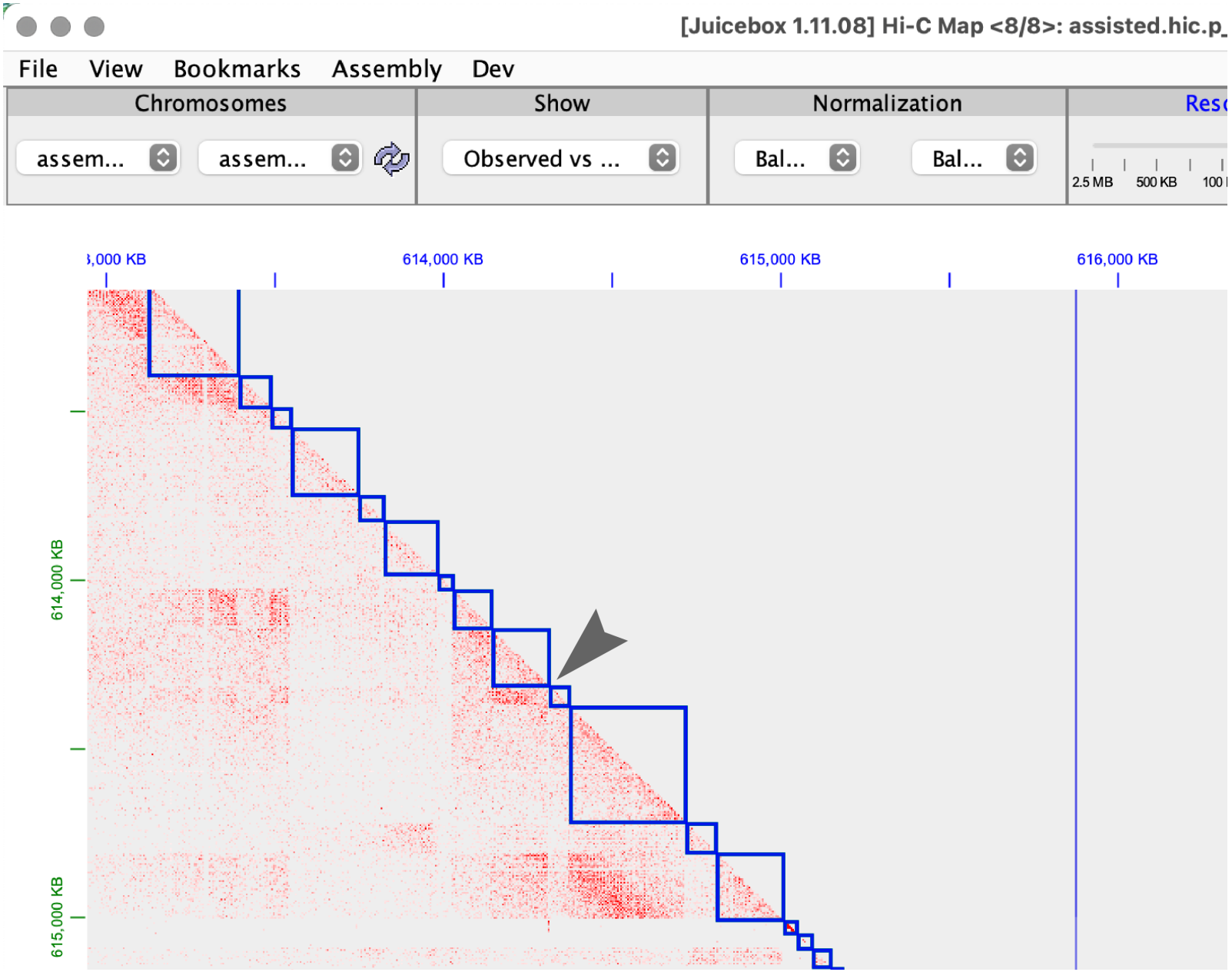


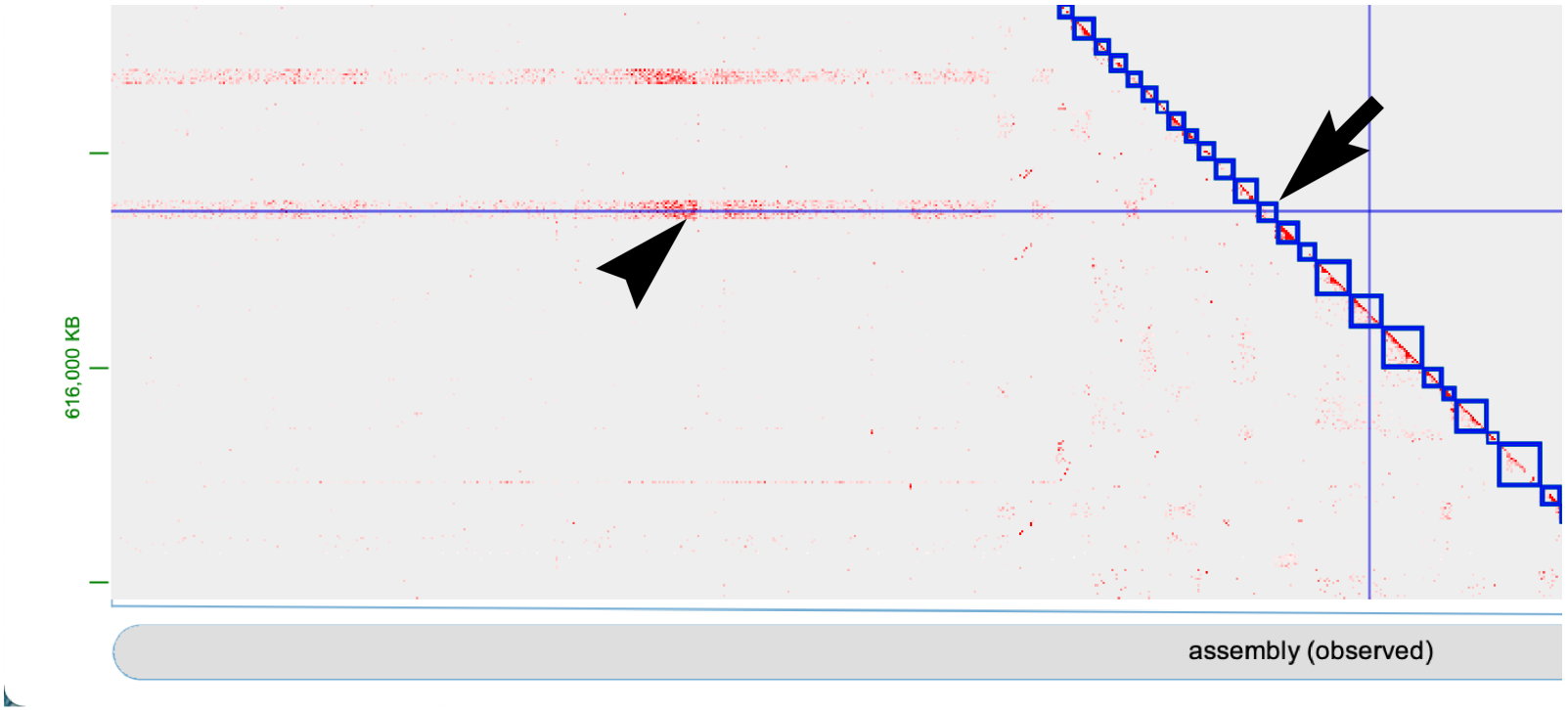
1. Select (as described previously) the misplaced sequence to be moved.
2. Hover the mouse cursor at the boundary between two adjacent sequences, where the selected sequence should be inserted. The cursor will change into an upper-right (if positioned just below the diagonal) or lower-left (if positioned just above the diagonal) -facing insertion arrow.
3. Click once with the left mouse button. The selected sequence will move to the insertion point between the two contigs.

TIP: If there are very small (possibly not visible) sequence(s) at the insertion point between two visible sequences,

the cursor may not change to the insertion arrow. In this case, the user must zoom-in on the desired insertion point (with the misplaced contig still selected) until the insertion arrow appears.

Moving small sequences precisely

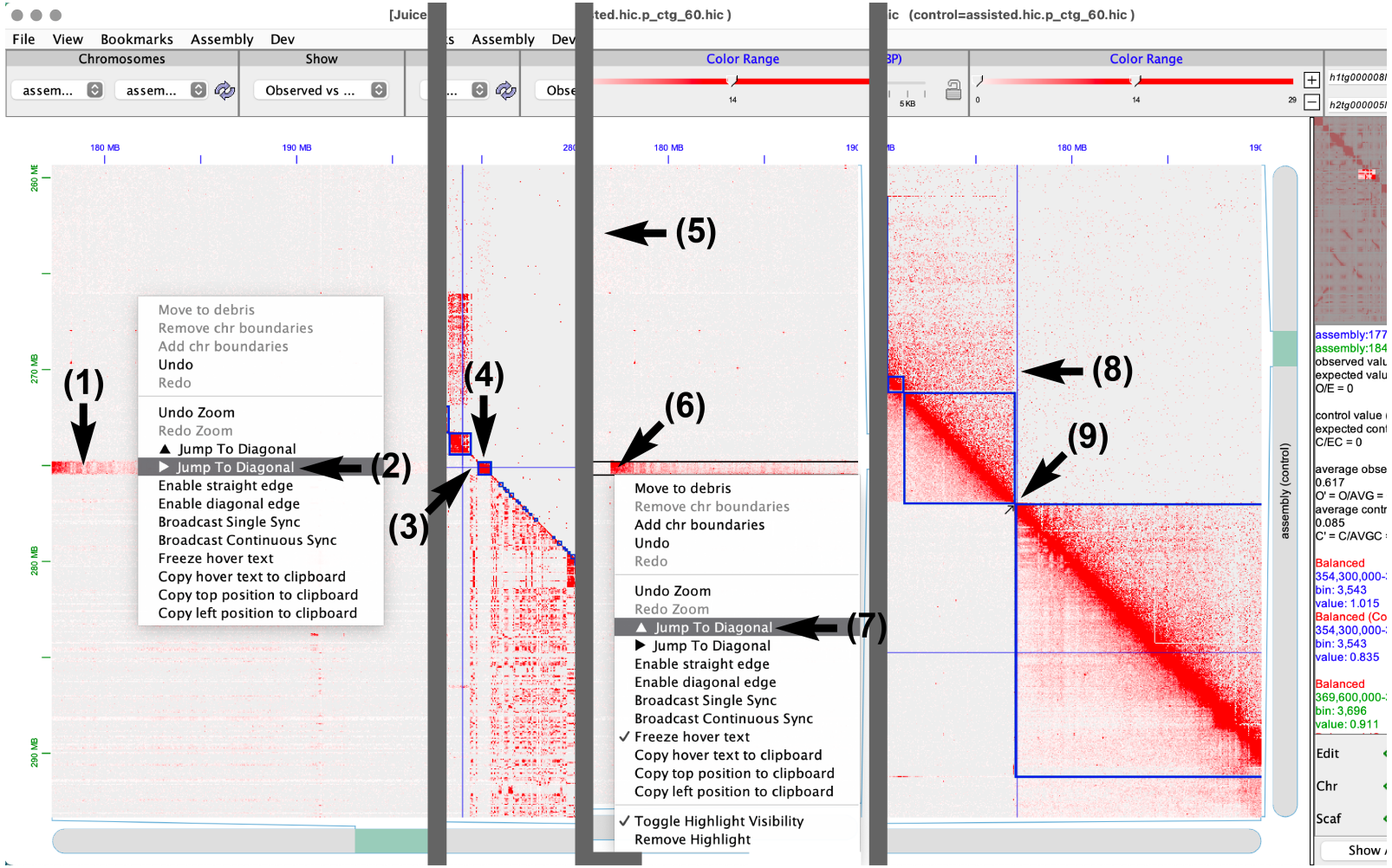




When working in an area with small sequences, contacts between one small misplaced sequence (black arrow) and one or more other contigs will appear as a horizontal or vertical contact stripe (black arrow head). Sometimes identifying which sequence the contacts are from can be difficult if it is embedded in many other small contigs (as shown above). A tip to identify the correct, misplaced sequence is to:

1. Hover the mouse cursor over the contact stripe (black arrow head).
2. Press-and-hold *shift* to activate the blue horizontal and vertical straight-edge.
3. Use the straight-edge to identify the correct sequence on the diagonal (black arrow), then select it (as described previously).
4. Insert (as described previously) the misplaced sequence at the insertion point with the densest contacts in the stripe (grey arrow head).

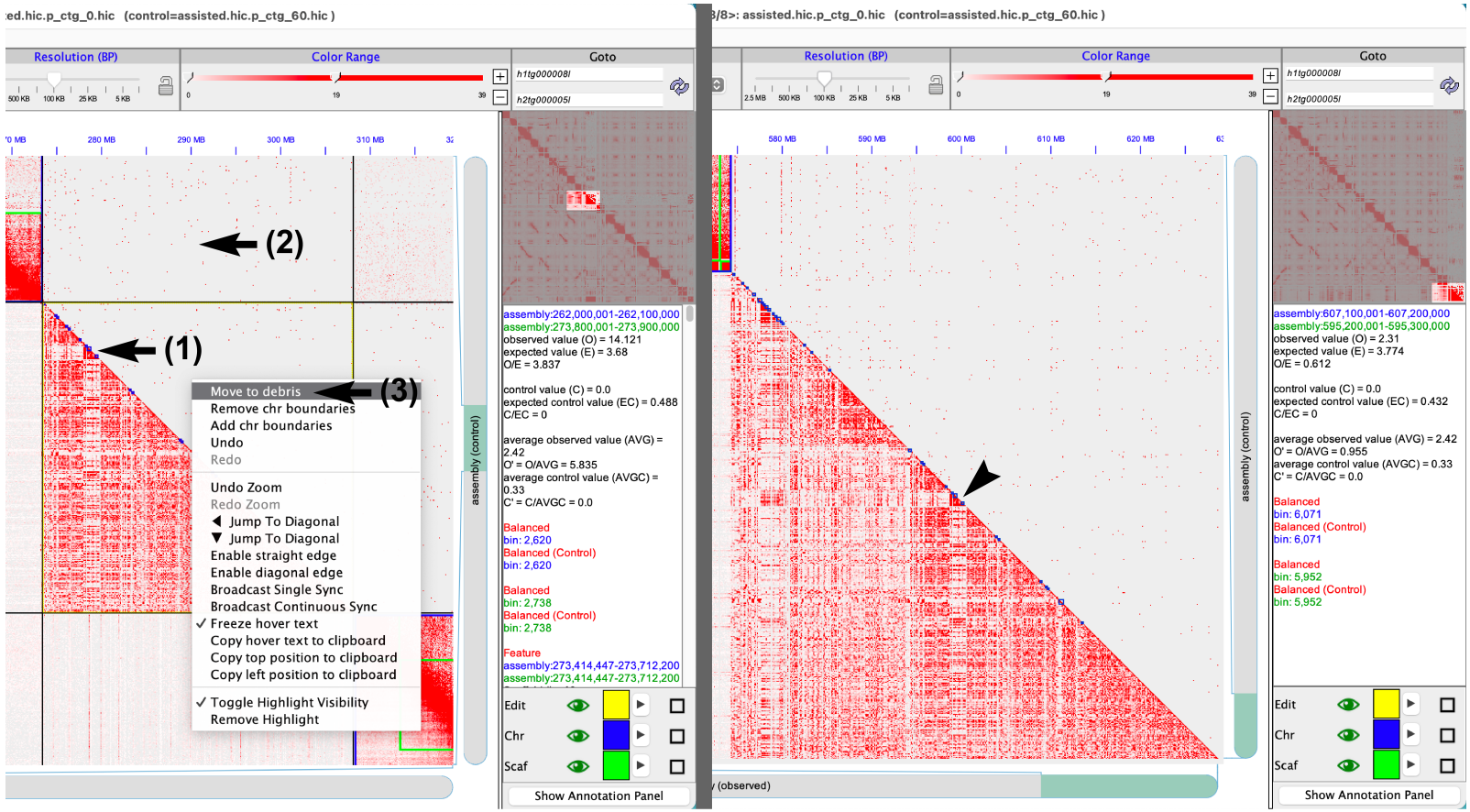
Moving sequences long distances



It's not uncommon for a user to scroll through the contact map and unexpectedly observe a stripe of contacts between two very positionally-distant sequences (*e.g.*, a small contig placed in the wrong chromosome). With such large distances--and especially when scaffolding small contigs at the end of the contact map not yet placed into chromosomes--a great amount of horizontal and vertical scrolling might be required. This becomes tedious very quickly. The following is a series of moves that can be used to efficiently identify the correct misplaced sequence and its appropriate insertion point, while reducing the amount of unnecessary scrolling.

TODO

Moving sequences to 'debris'

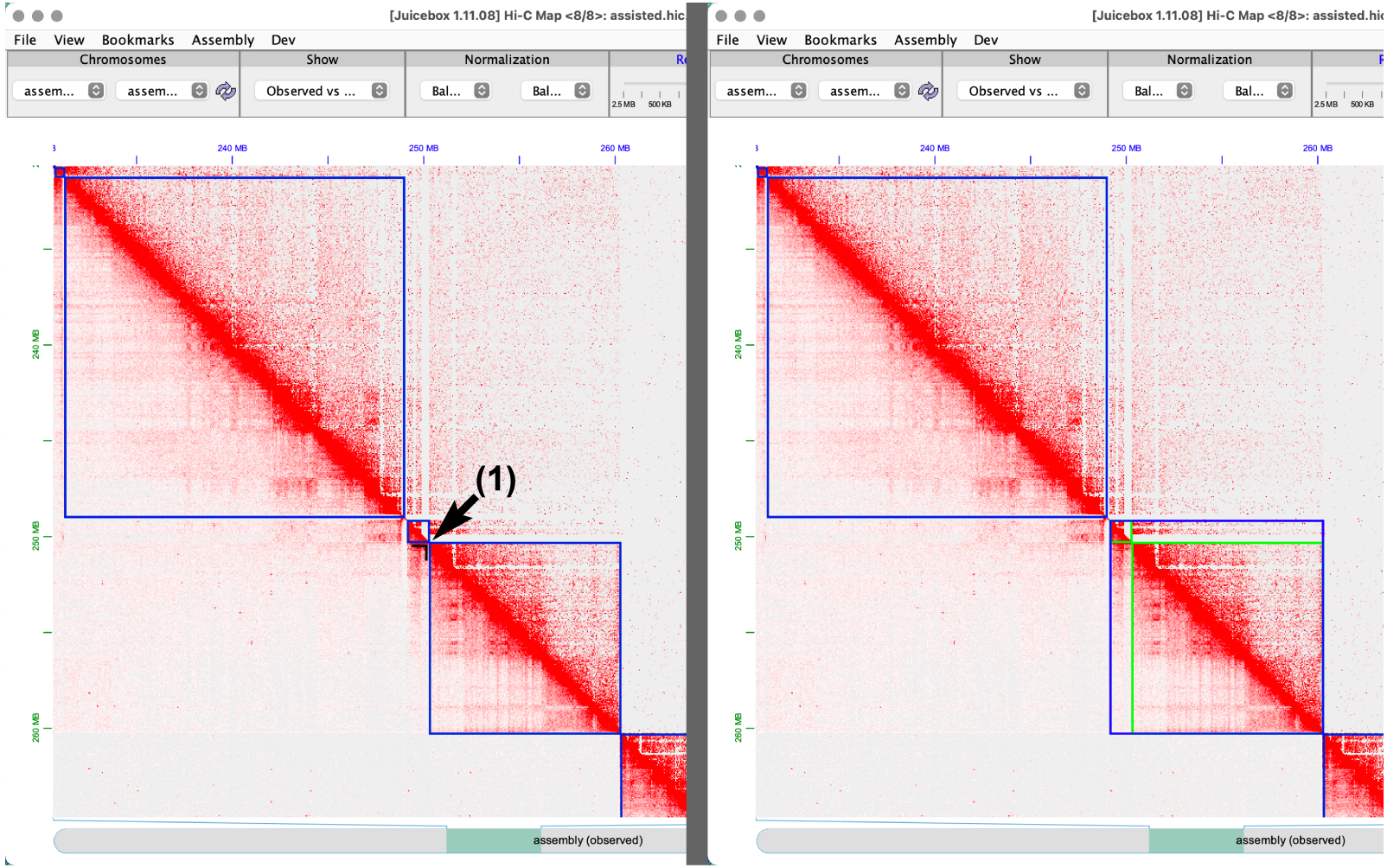


Depending on the scaffolding software used, blocks of small repetitive contigs can be inserted between larger scaffolds, cluttering the contact map. Moving these extraneous sequences to the end of the assembly helps to keep the JBAT workspace tidy:

1. Select (as described previously) the small repetitive sequences to be moved to debris.
2. Click with the right mouse button (*control* + click on Mac) anywhere in the contact map window area to **open the context menu**.
3. Select the **Move to debris** option.

Adding/removing chromosome boundaries

Adding/removing boundaries between two contigs

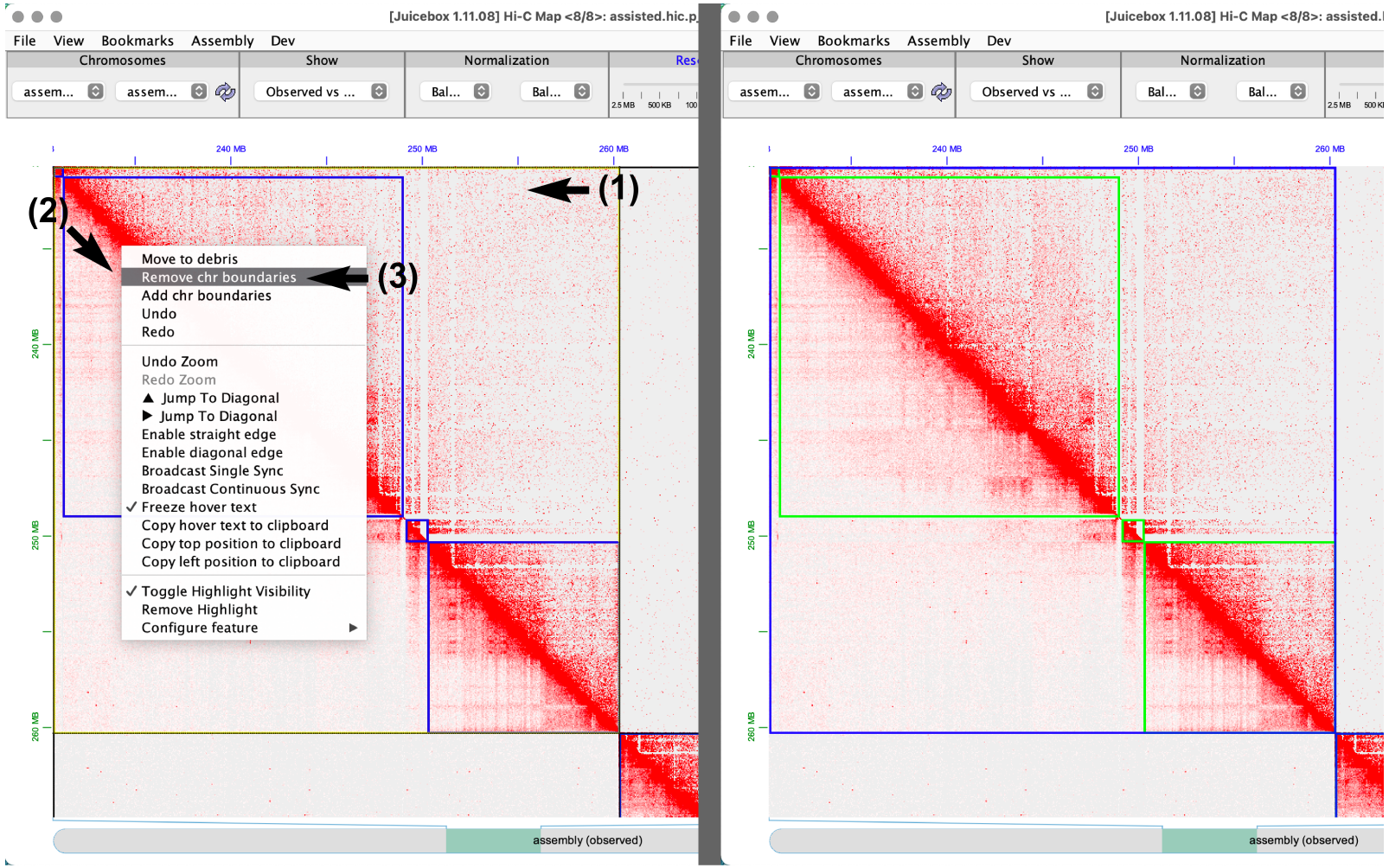


After ordering sequences belonging to the same chromosome, these sequences can be orgazined into chromosome-scale (blue boxes) scaffolds.

1. Hover the mouse cursor at the boundary between two adjacent sequences, where the chromosome boundary should be added/removed. The cursor will change into an upper-right (if positioned just below the diagonal) or lower-left (if positioned just above the diagonal) -facing L-shaped arrow head.

2. Click once with the left mouse button. The chromosome boundary between the two sequences will be added/removed.

Removing boundaries between many contigs



1. Select (as described previously) the two or more sequences belonging to the same chromosome.
2. Click with the right mouse button (*control* + click on Mac) anywhere in the contact map window area to **open the context menu**.
3. Select on of the **Add chr boundaries** or **Remove chr boundaries** options.

Apply modified .assembly file changes

To apply the manually-curated changes made in Juicebox (stored in `yourgenome.review.assembly` file) to `yourgenome.fasta` and generate the corrected .fasta file, there are two methods:

1. Using the 3D-DNA tool:

```
$ nohup 3d-dna/run-asm-pipeline-post-review.sh \
    --stage finalize \
    --mapq 60 \
    --gap-size 100 \
    --review yourgenome.review.assembly \
    yourgenome.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
```

2. Using an ARTISANAL tool, which has the benefit of writing a liftover .chain file:

```
# The following script requires the pysam module be installed
$ assembly-to-fasta yourgenome.review.assembly yourgenome.fasta yourgenome.review
```