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

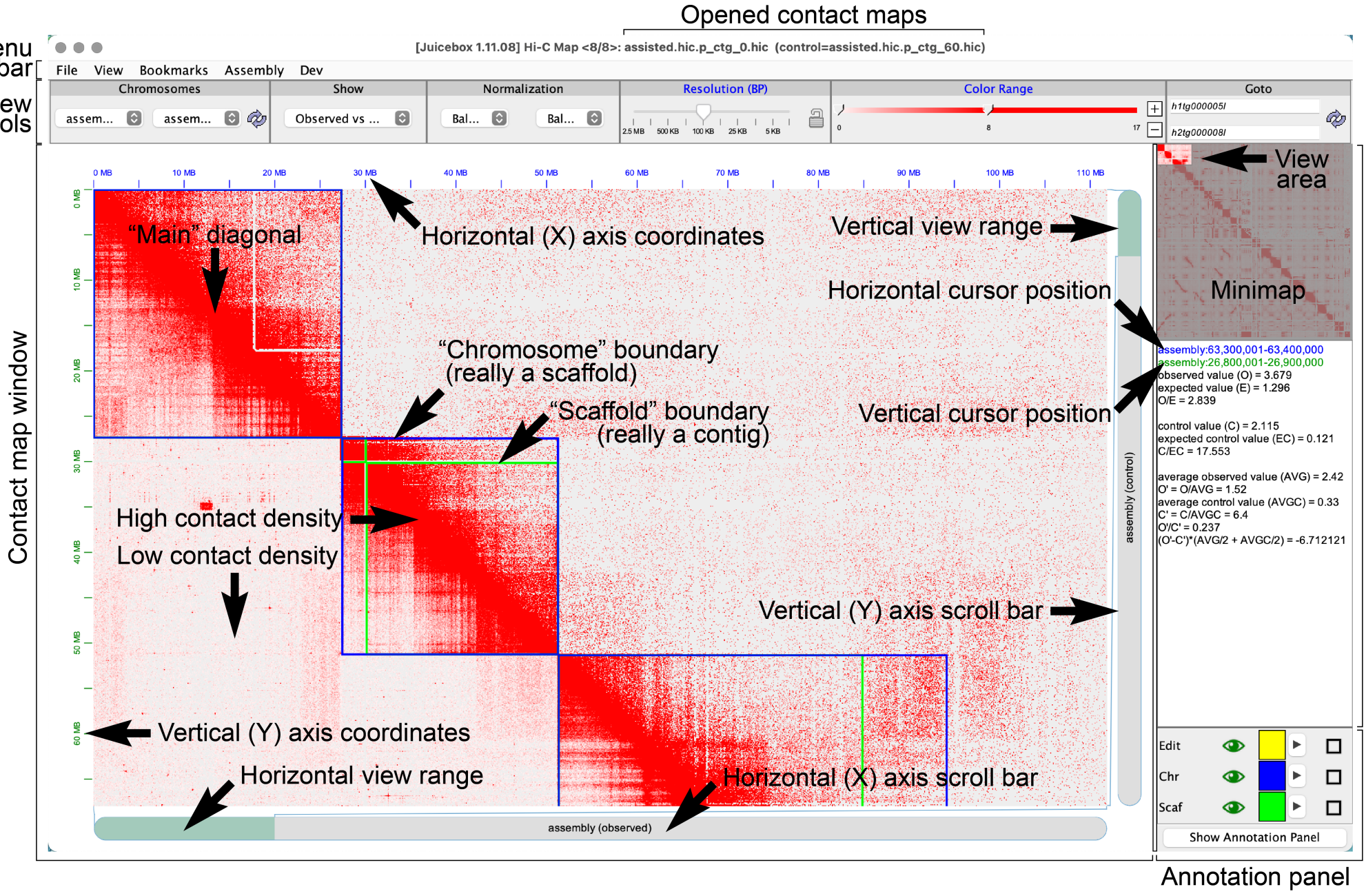
To create a 'map' .assembly file representing individual contigs in a pre-scaffolded genome assembly, download artisanal (https://bitbucket.org/bredeson/artisanal), then do:

```
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-fasta -c genome.fasta genome
```

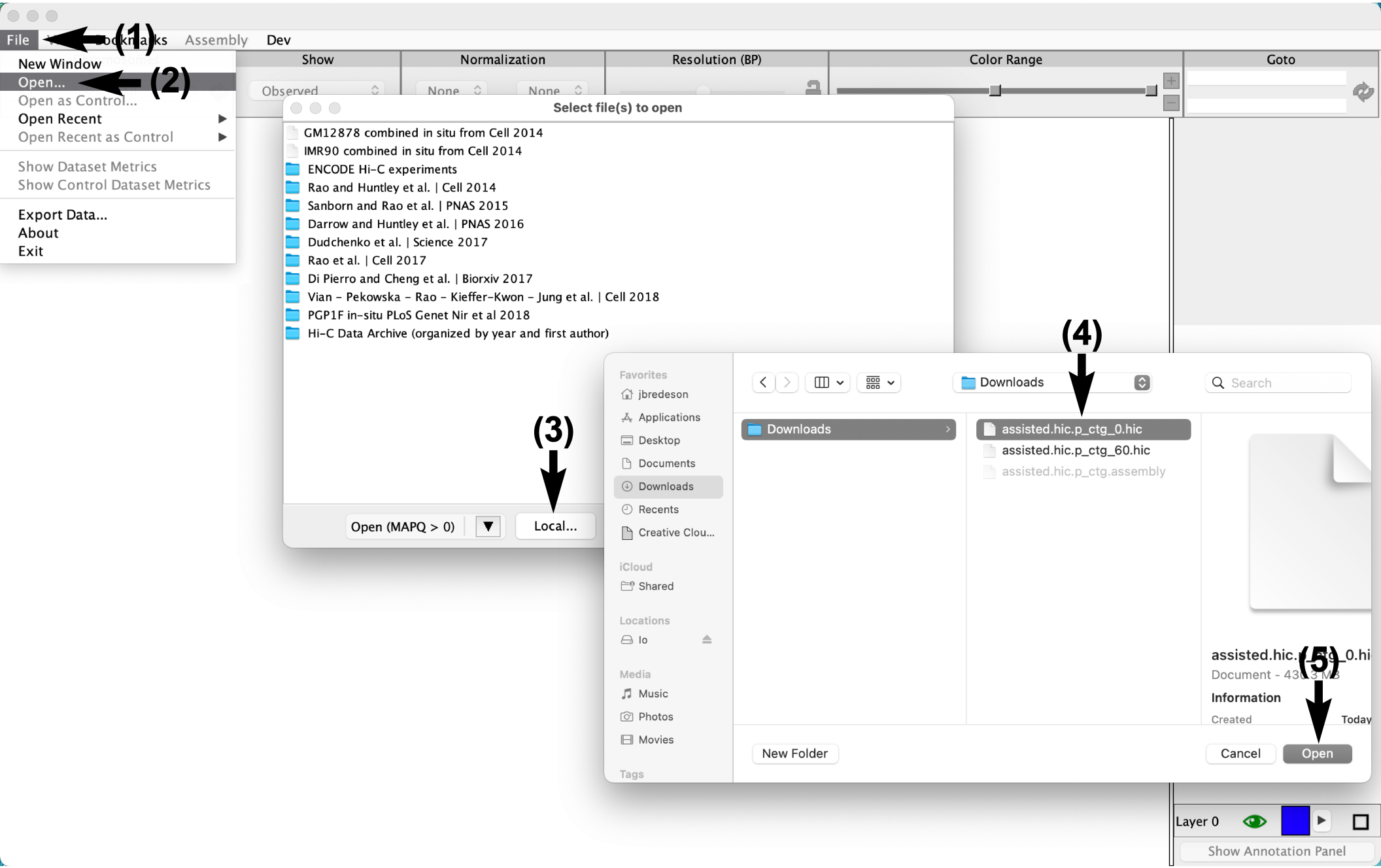
Introduction to the Juicebox desktop app

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

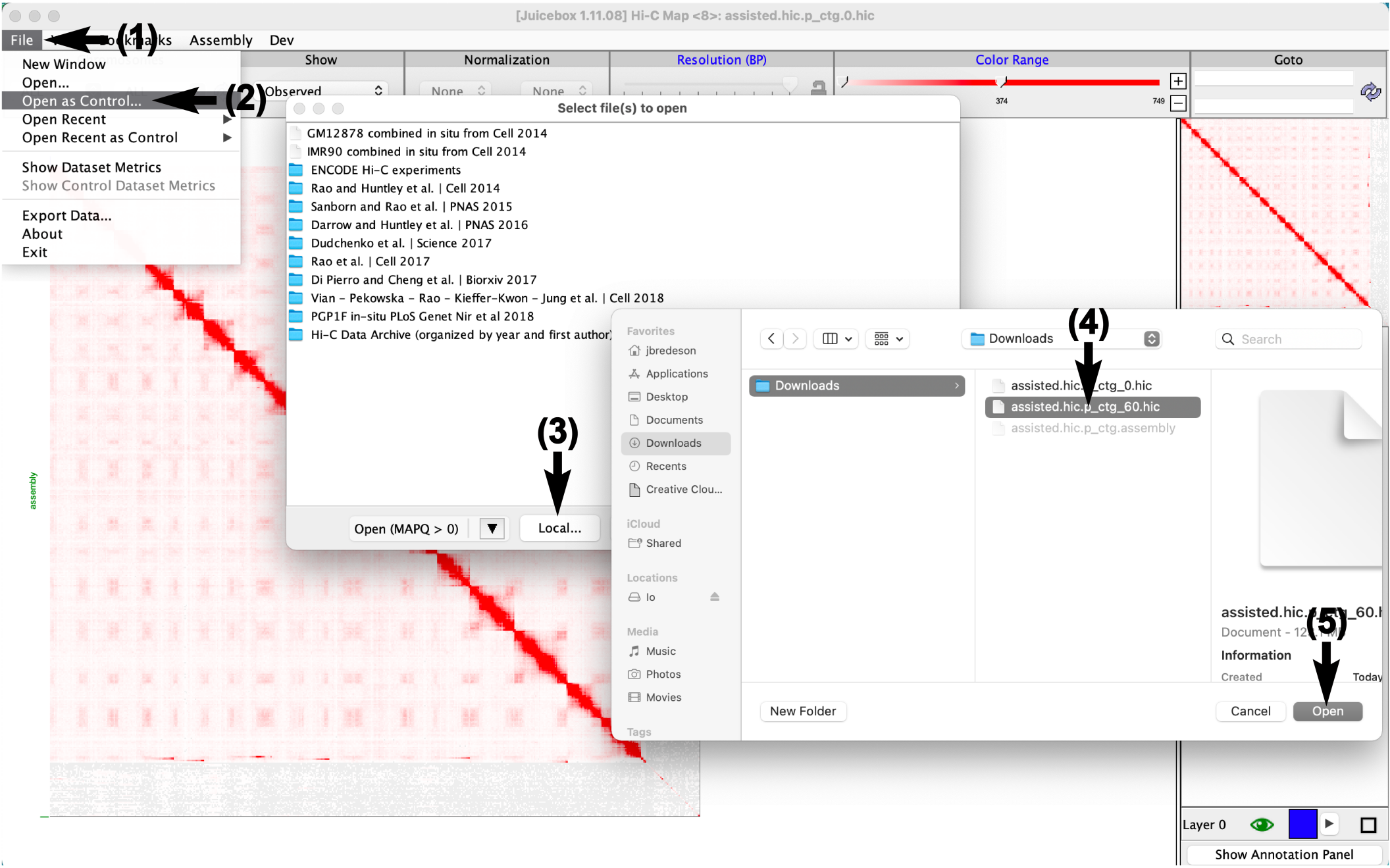
Anatomy of the Juicebox desktop app



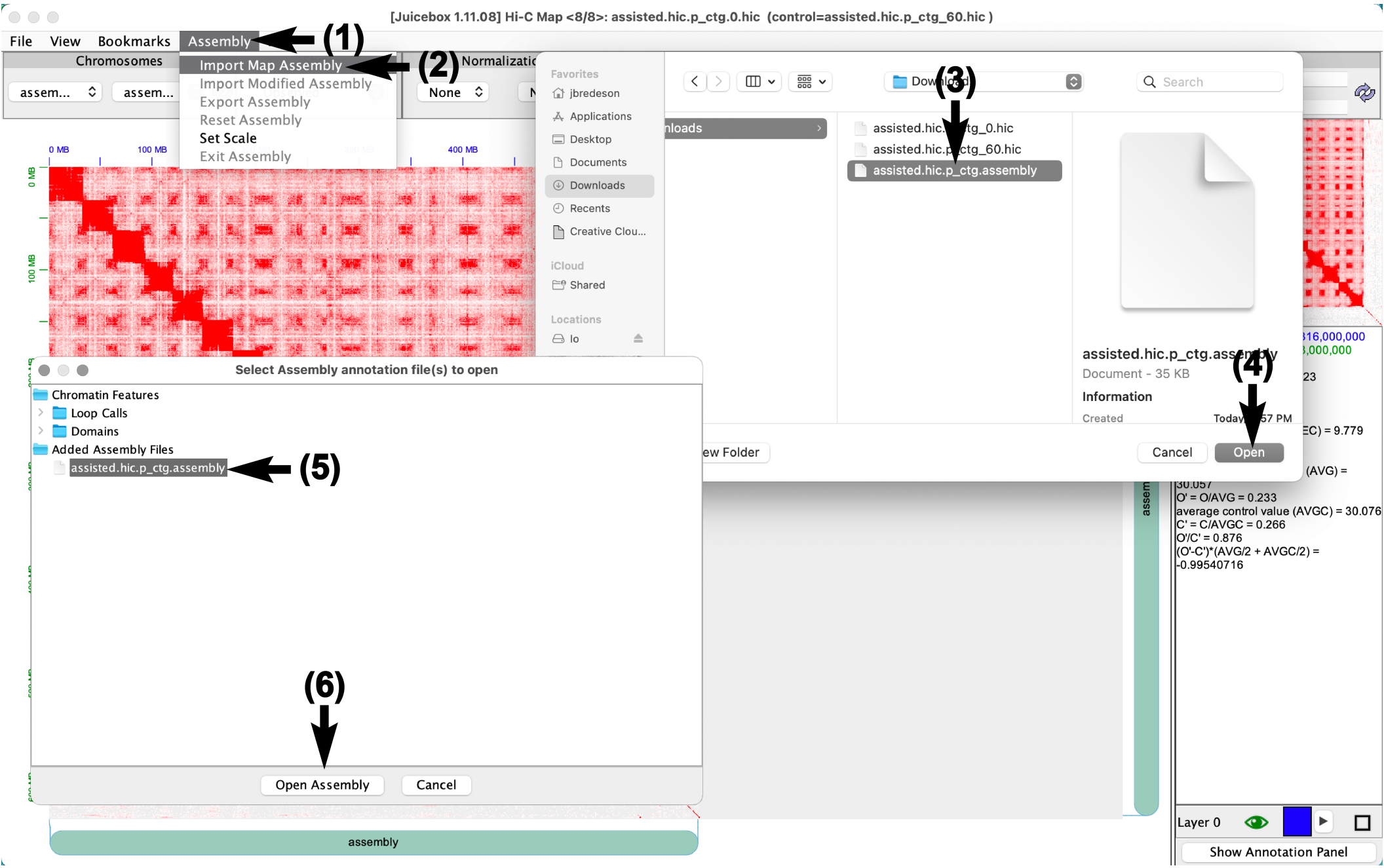
Opening an ‘Observed’ .hic contact map file



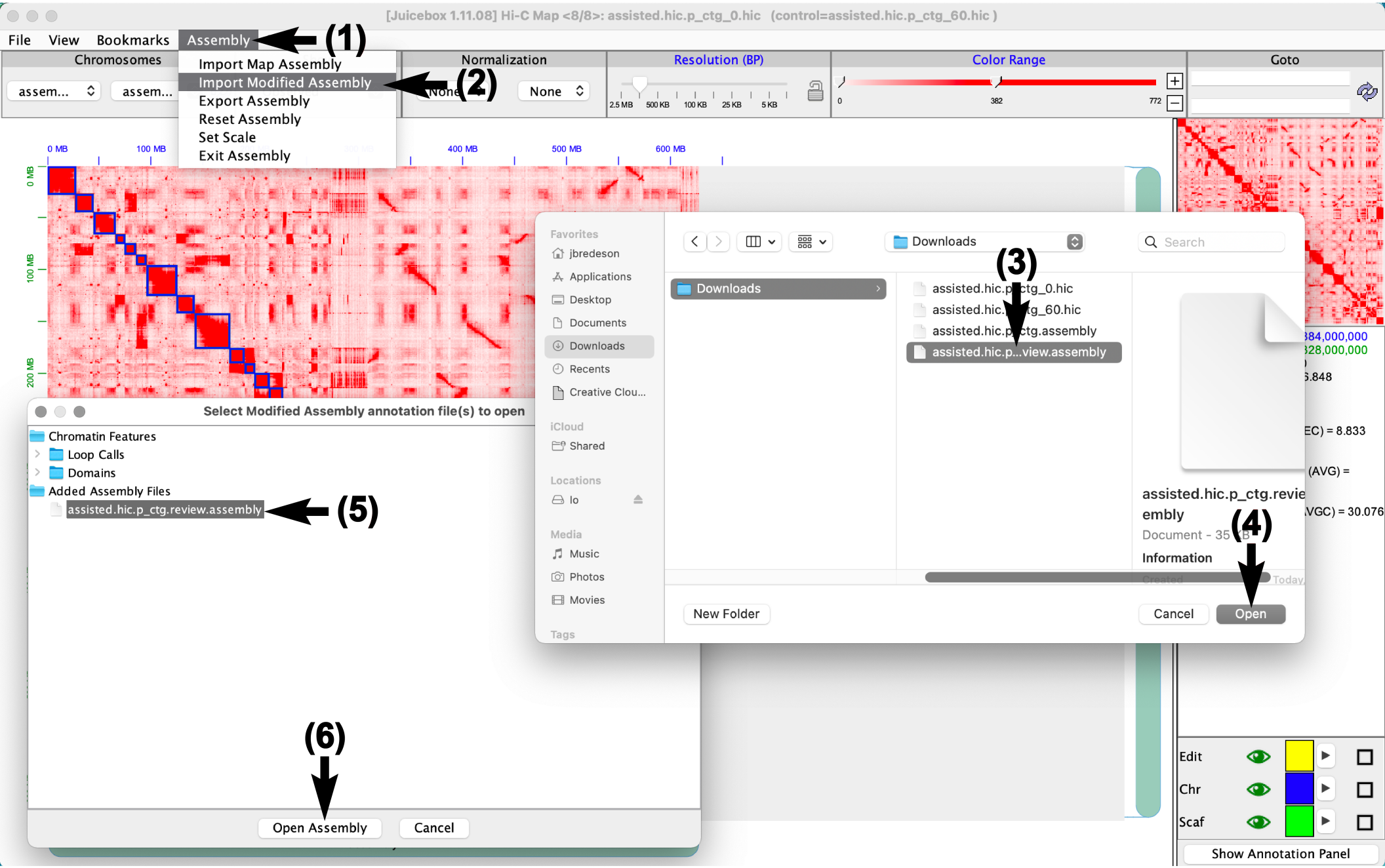
Opening a ‘Control’ .hic contact map file



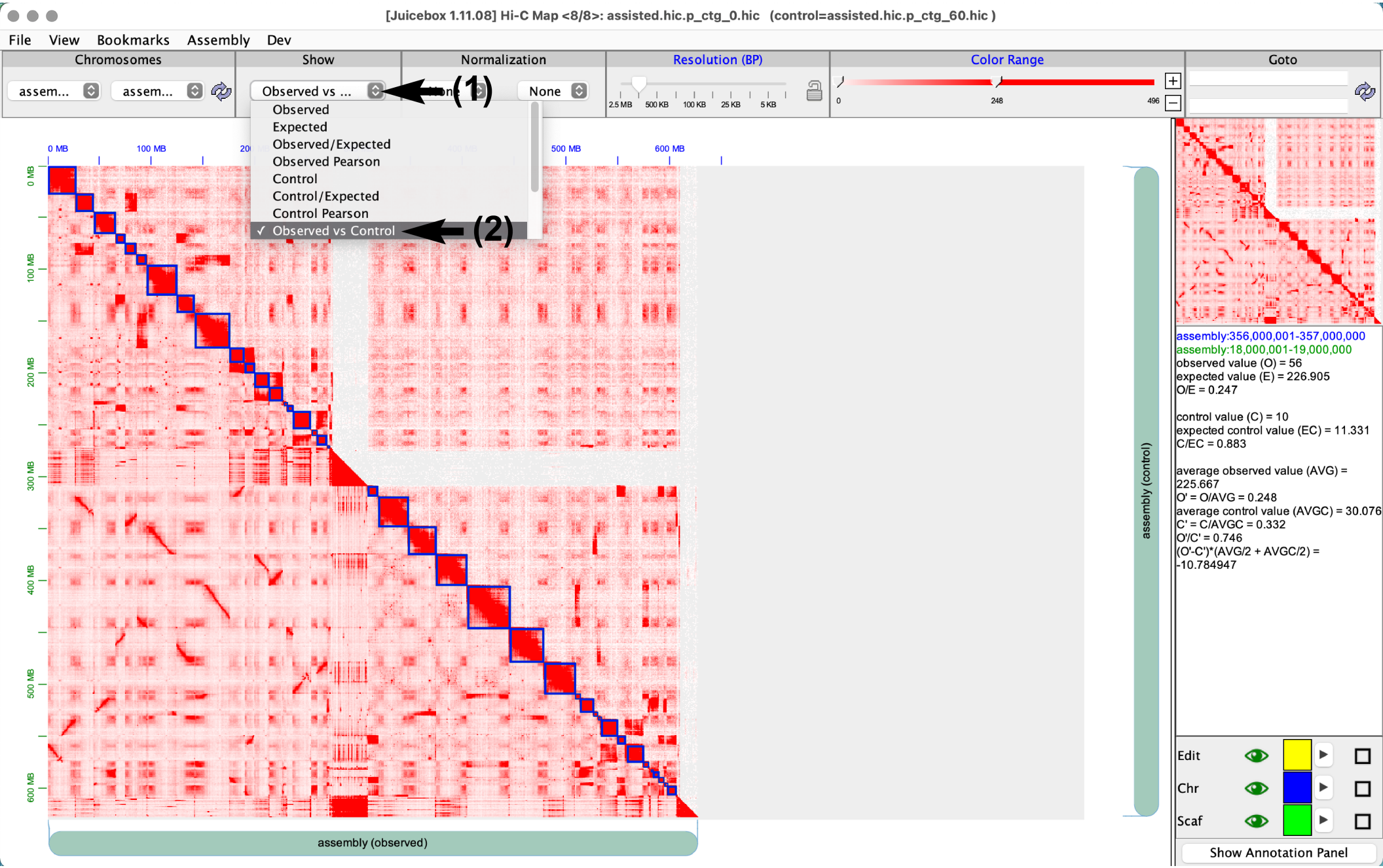
Opening a ‘map’ .assembly file



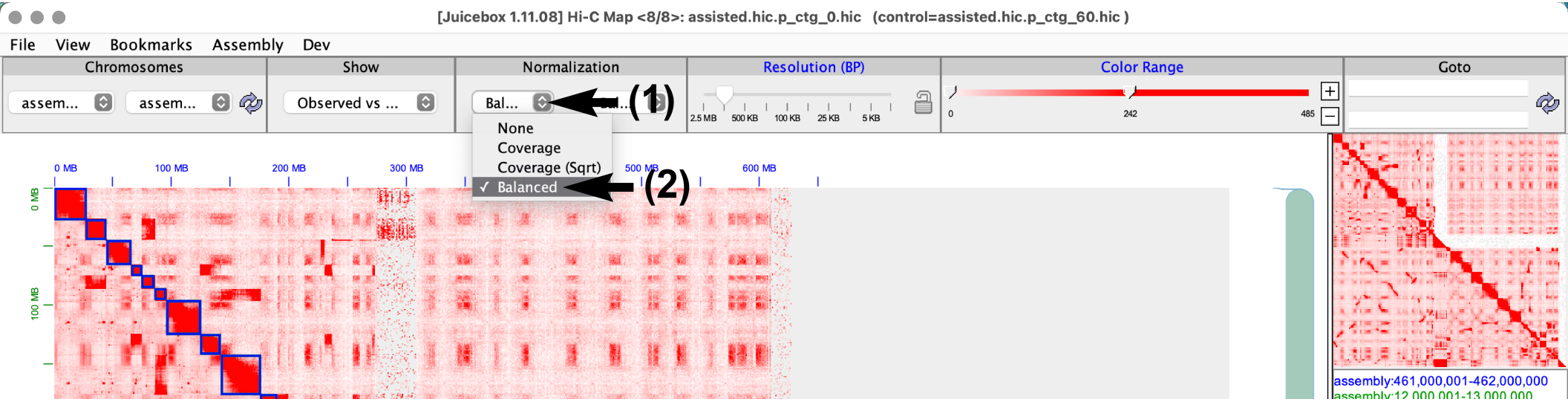
Opening a ‘modified’ .assembly file



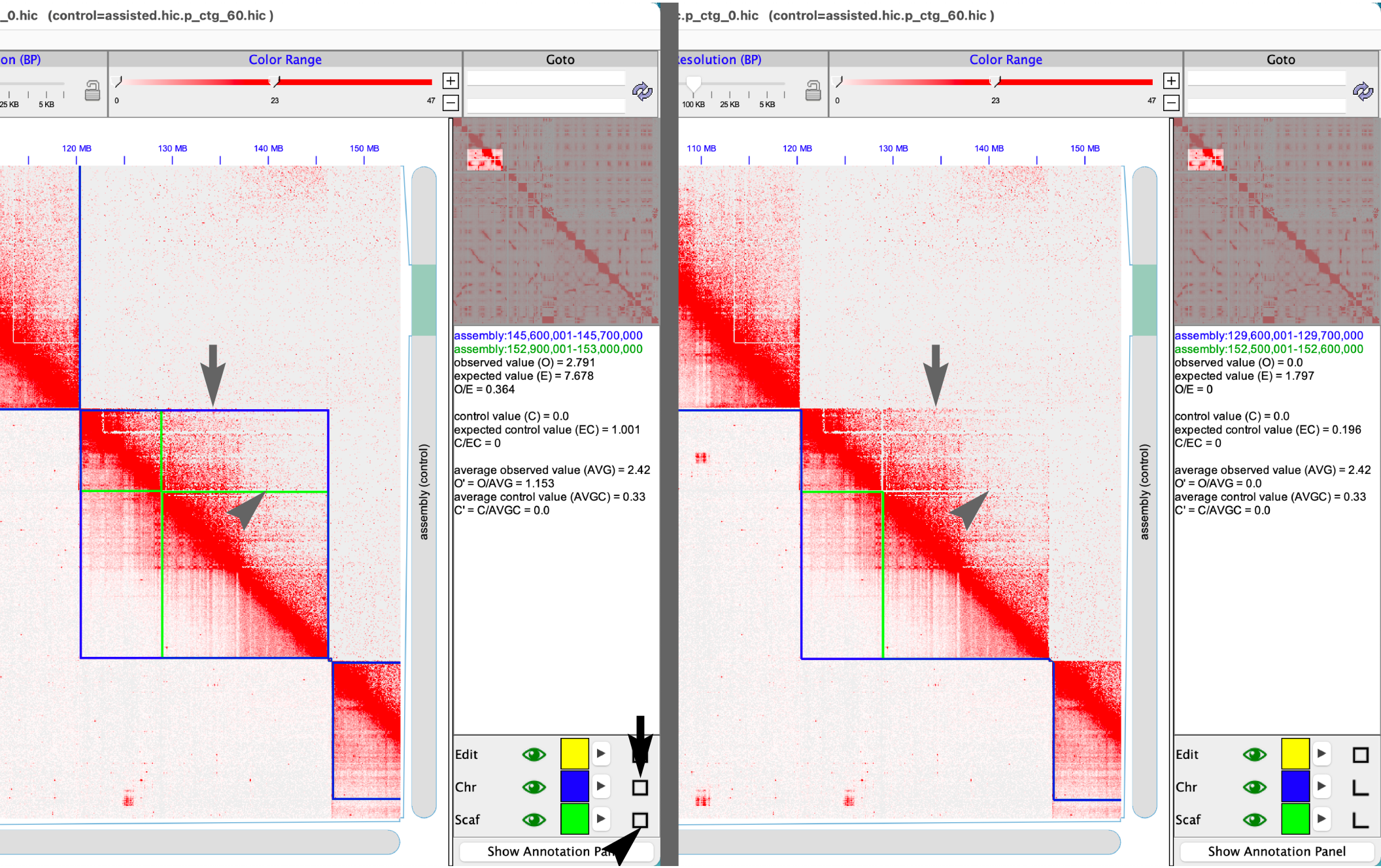
Changing contact map(s) displayed with ‘Show’



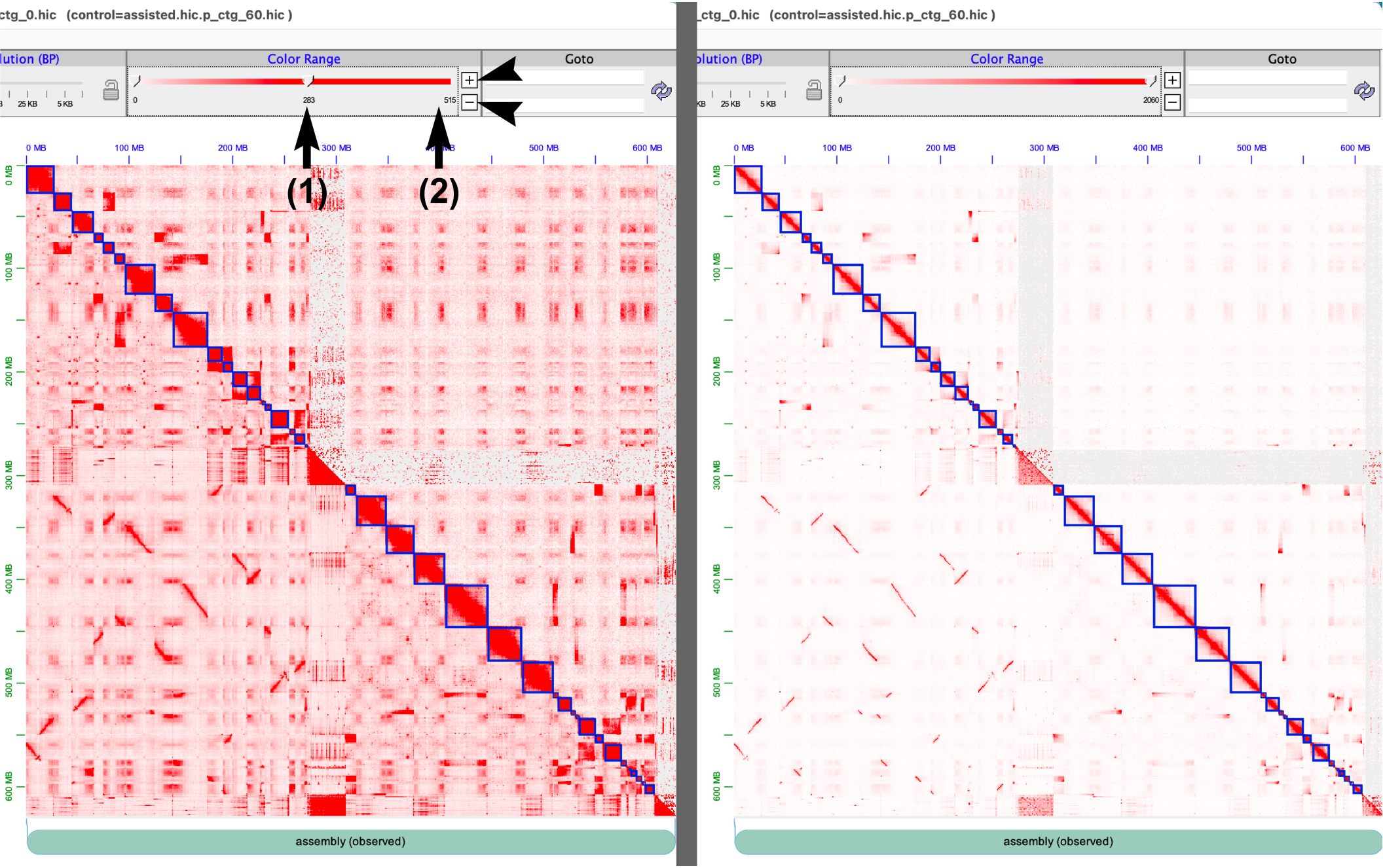
Enabling contact map normalization



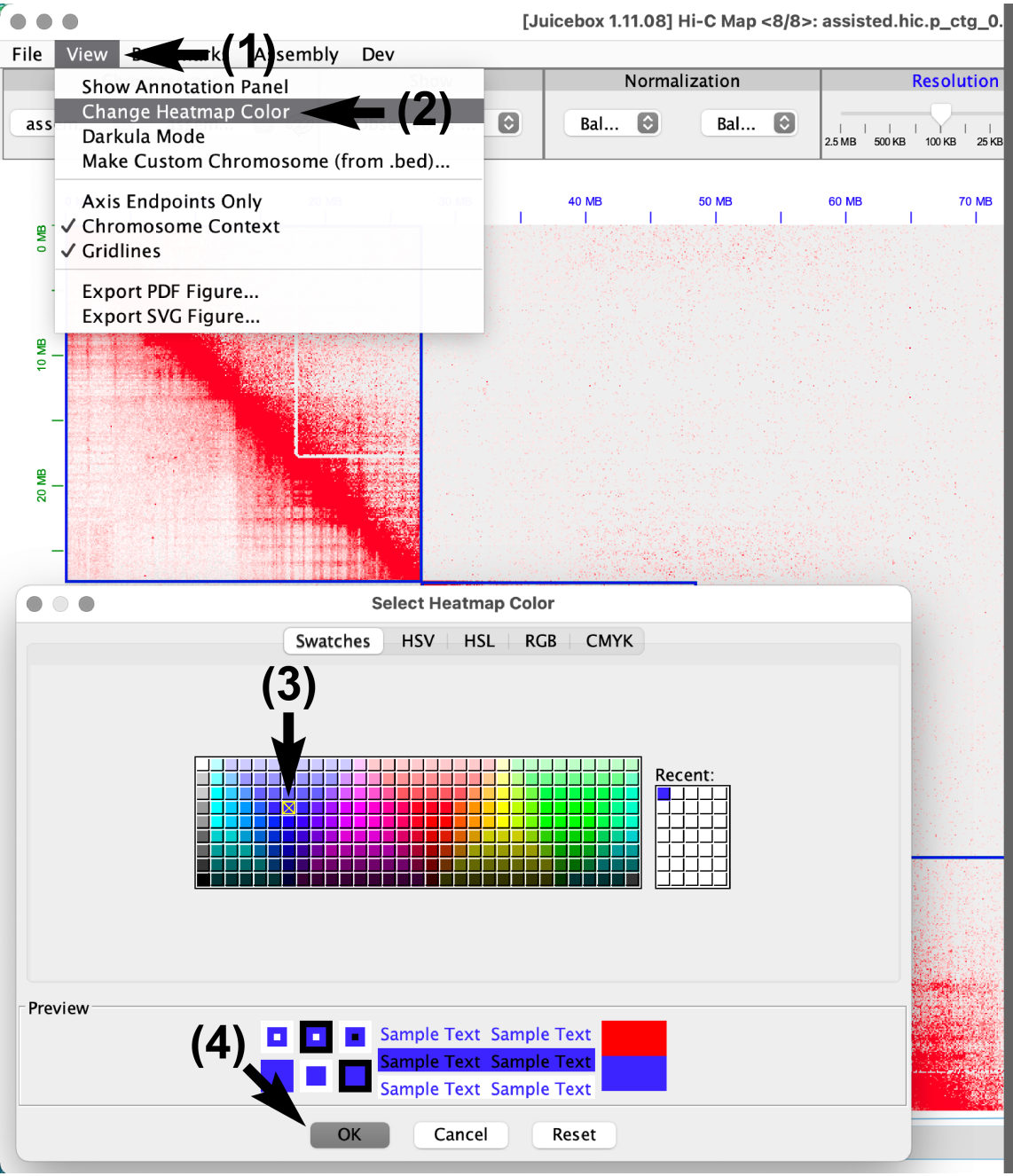
Toggling contig/scaffold boundary visibility

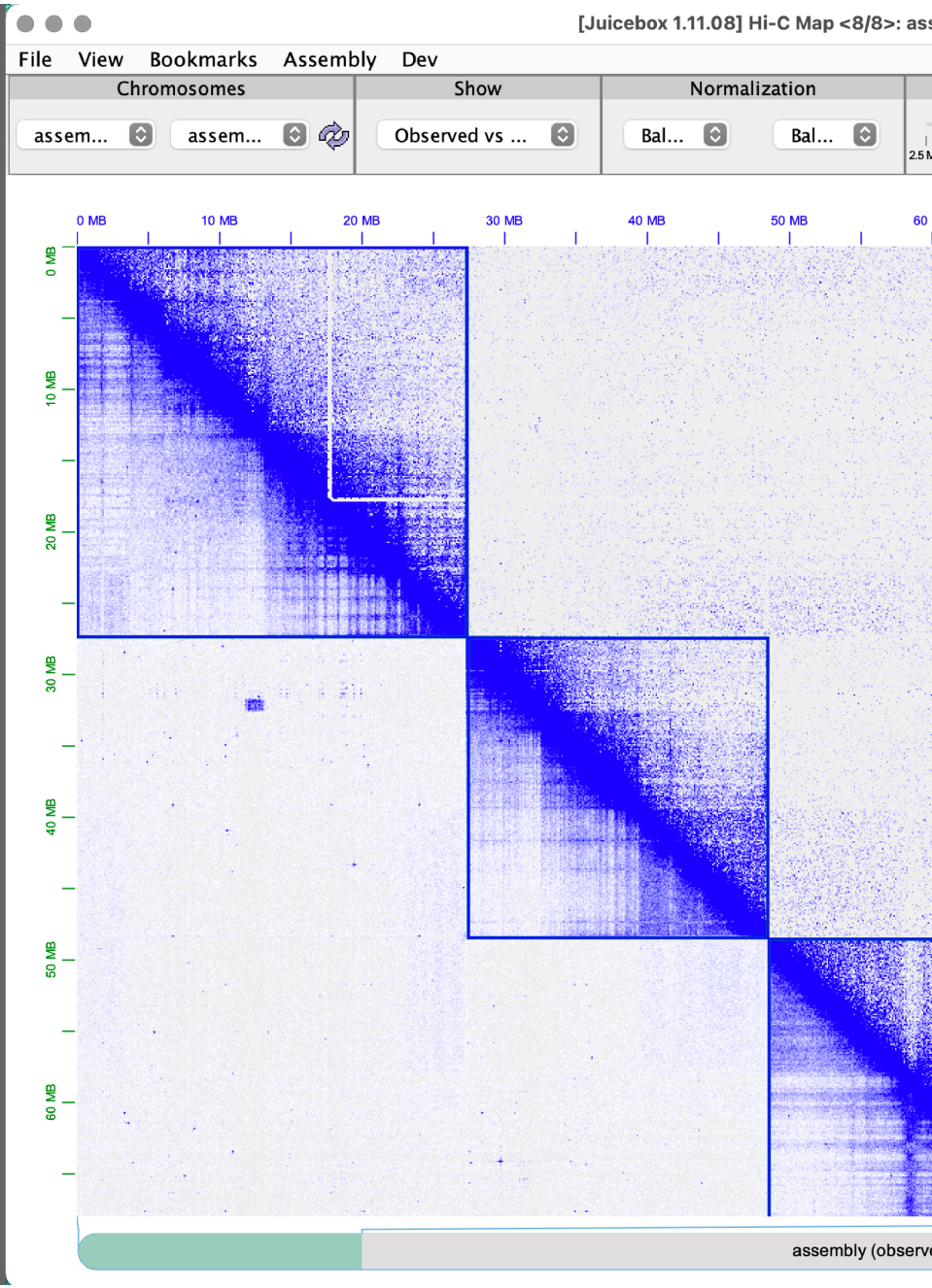


Adjusting contact map color range/saturation

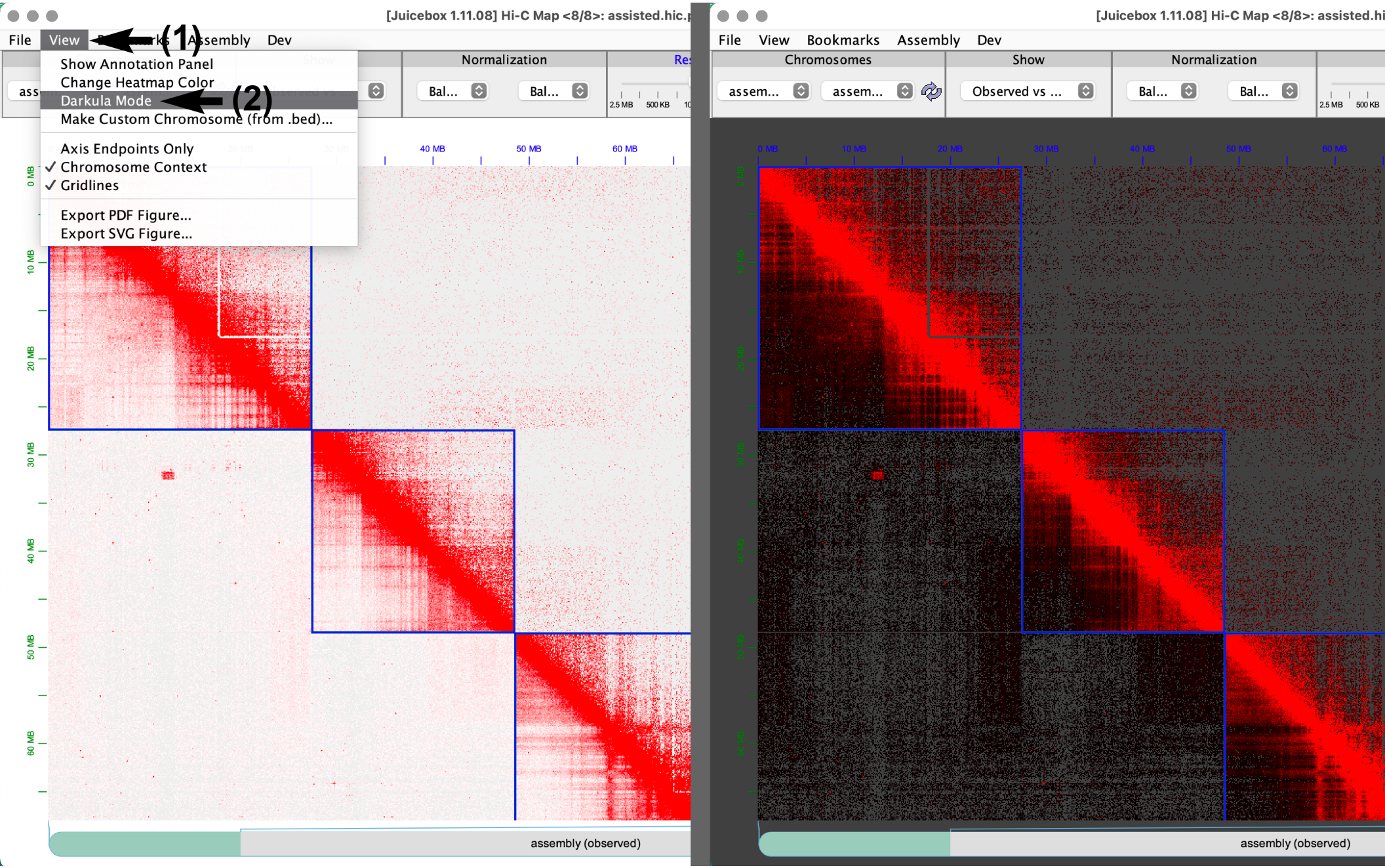


Changing contact map heatmap color

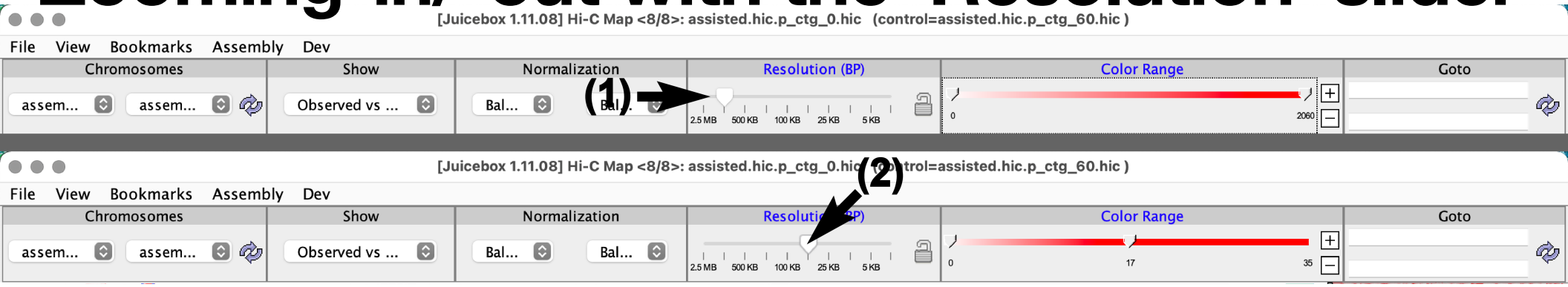




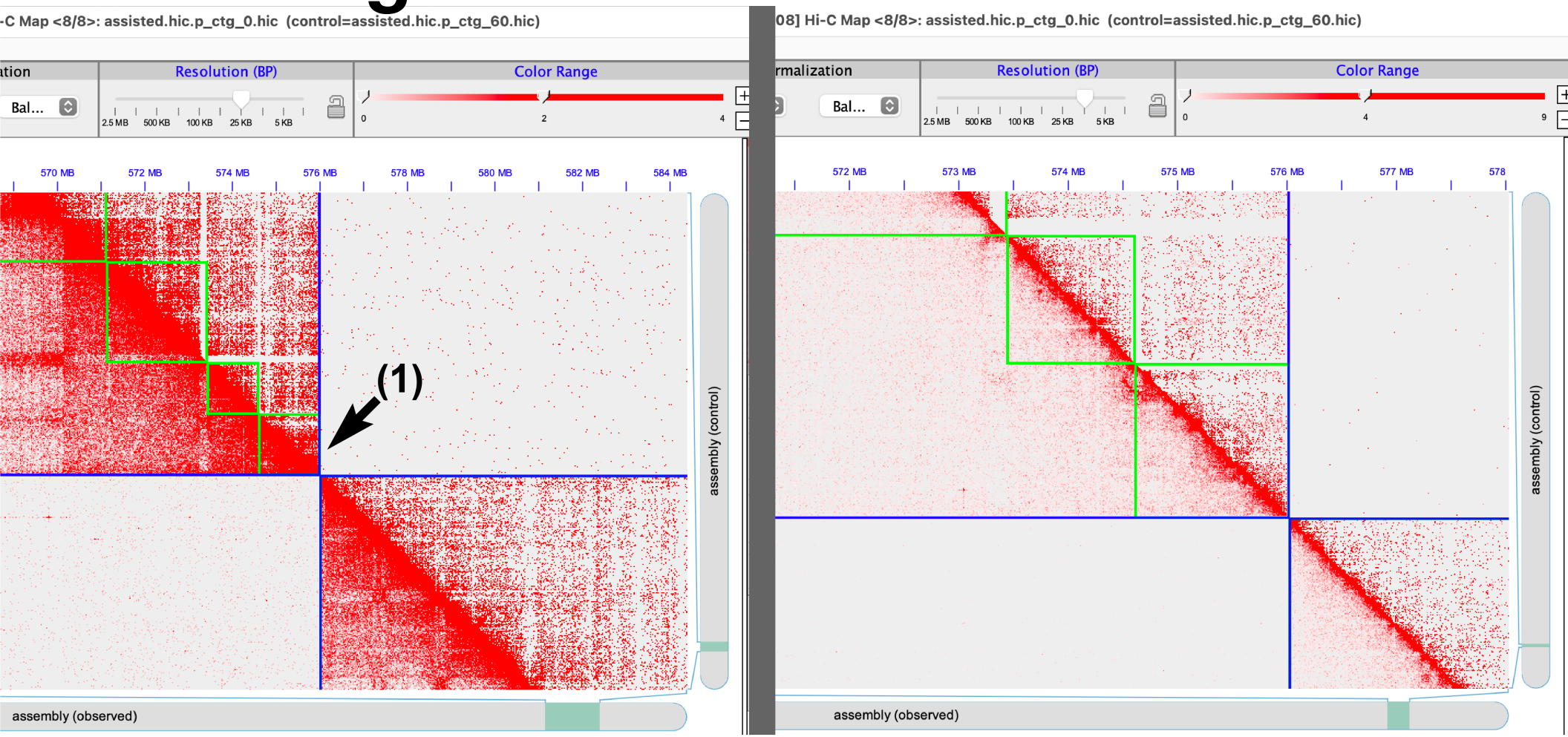
Darkula (night) mode



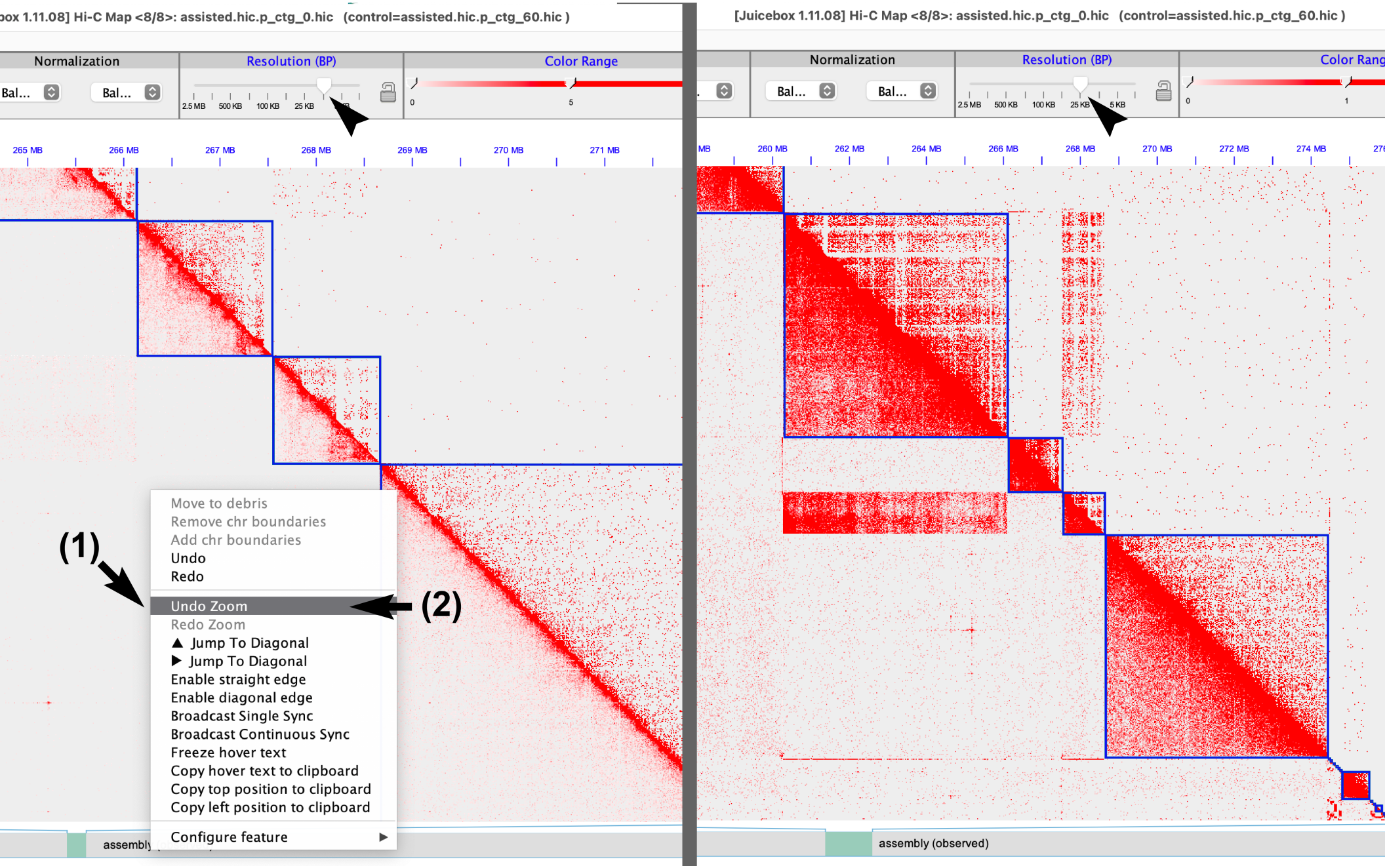
Zooming-in/-out with the ‘Resolution’ slider



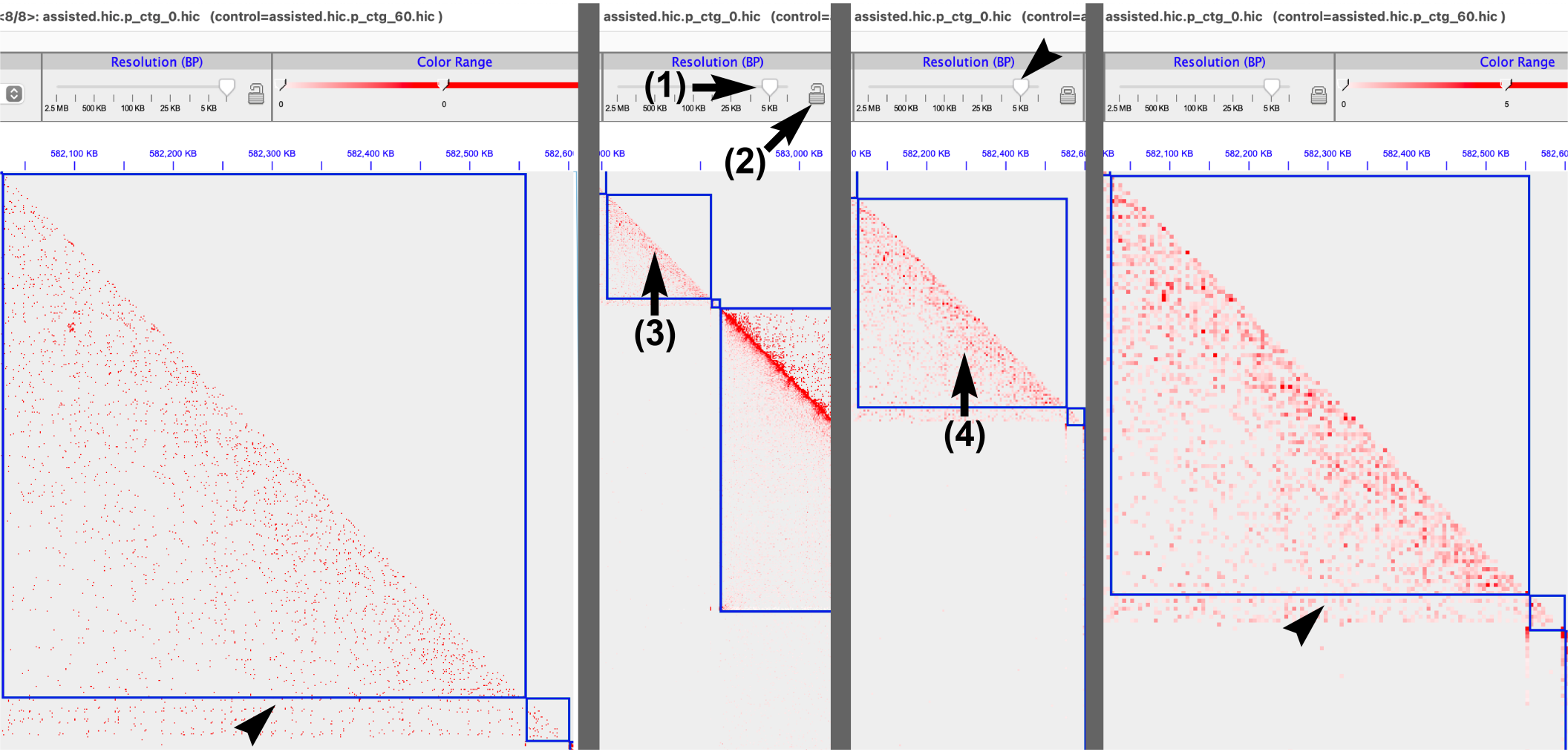
Zooming-in with a mouse double-click



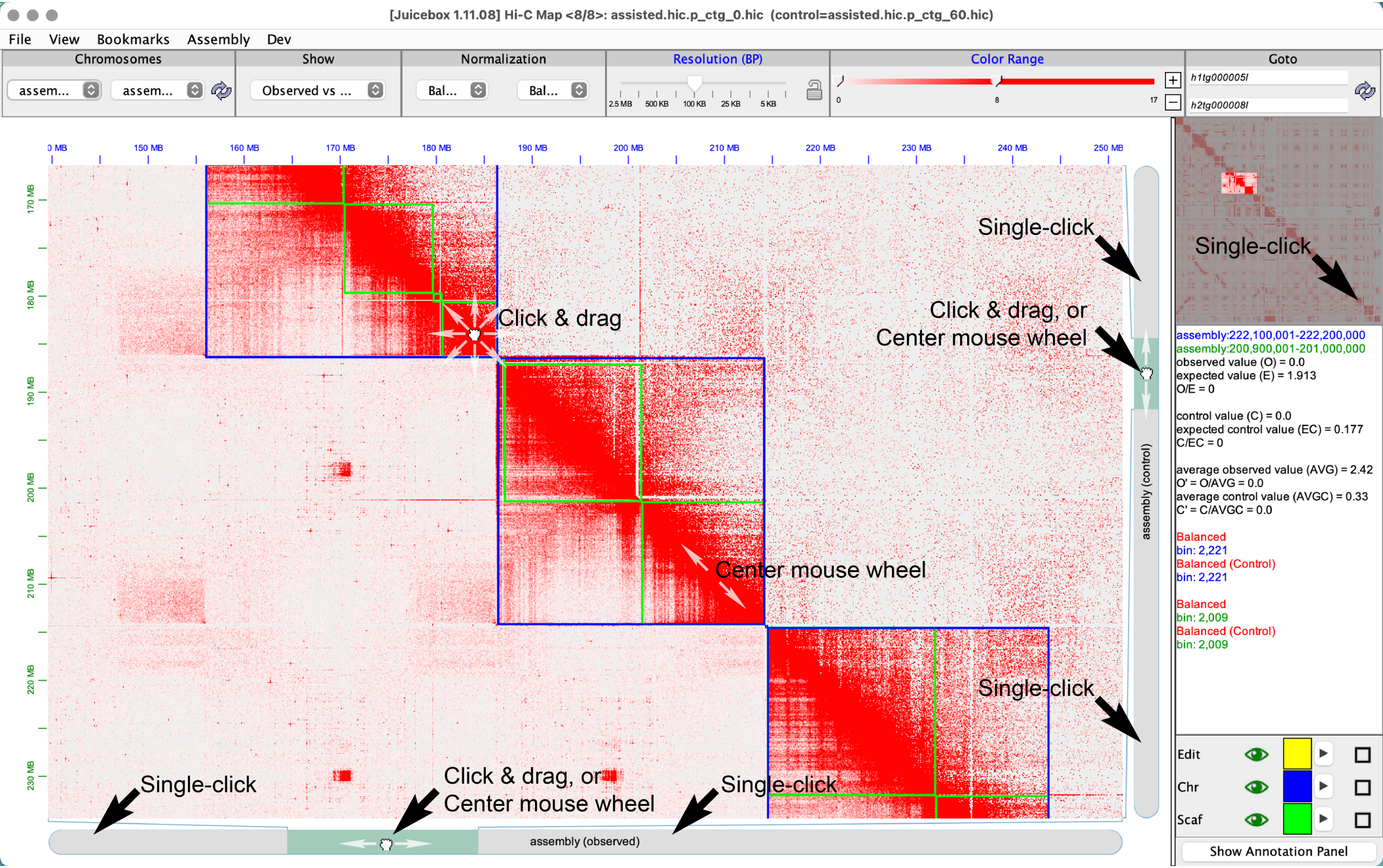
Undo zoom / Zooming out



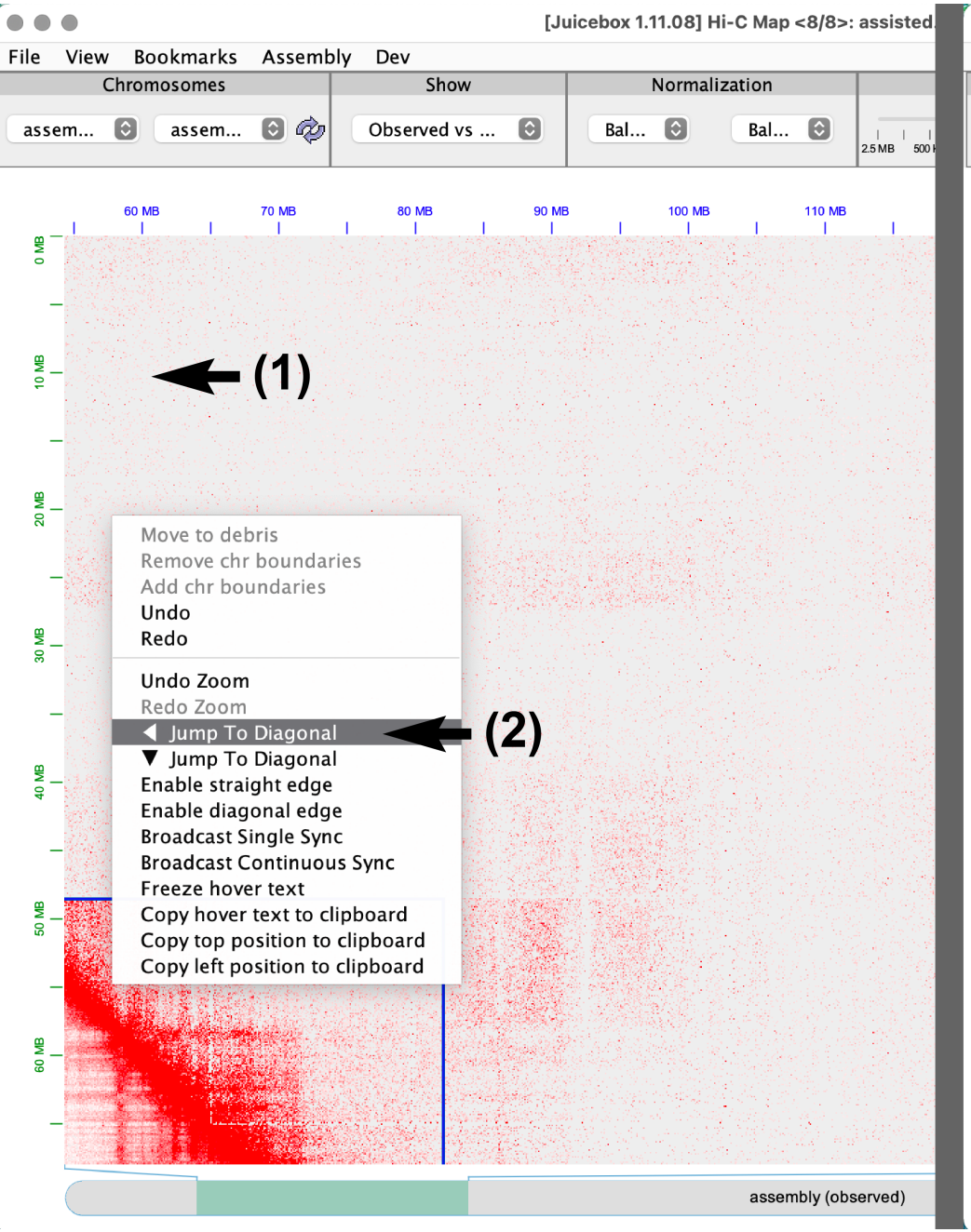
Zoom with the ‘Resolution’ lock

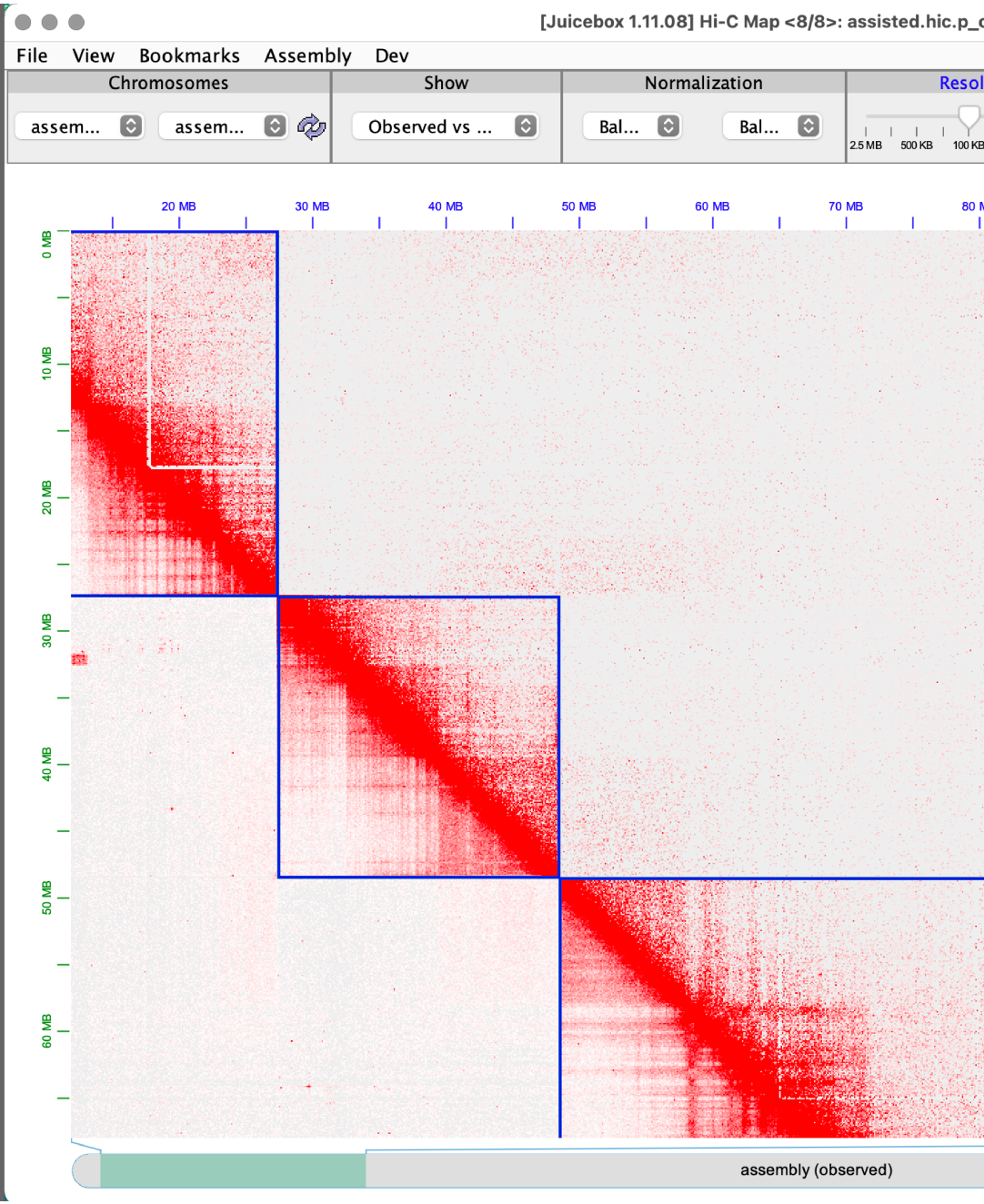


Navigating the contact map

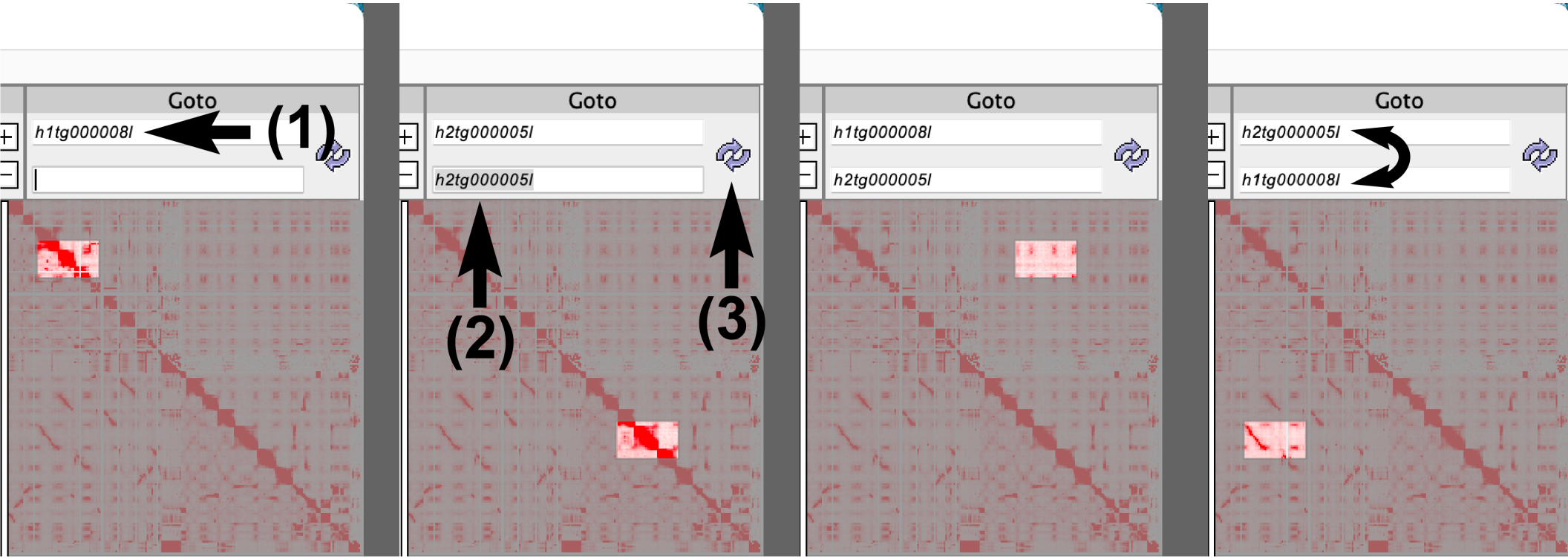


Navigating with ‘Jump to Diagonal’

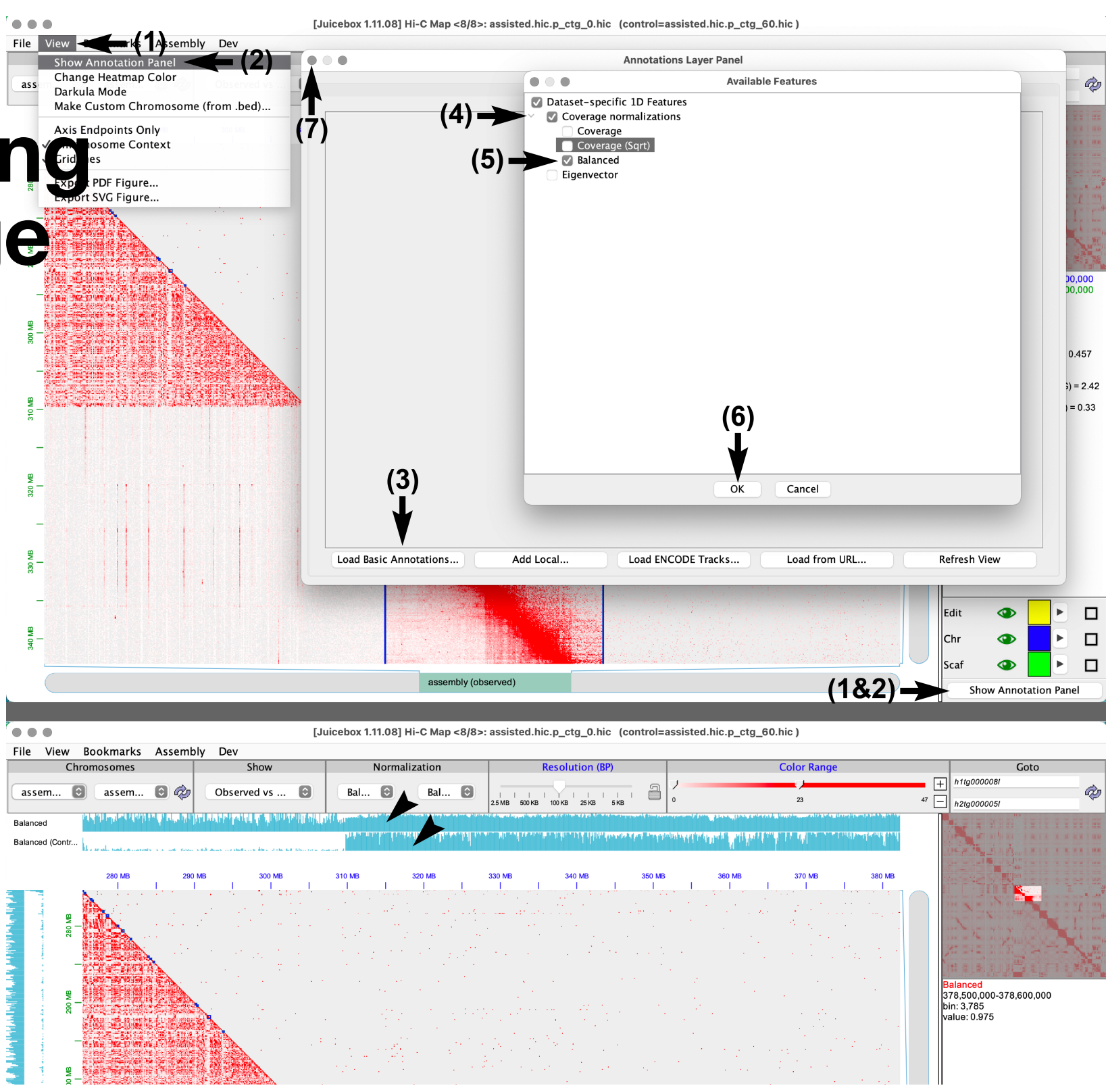




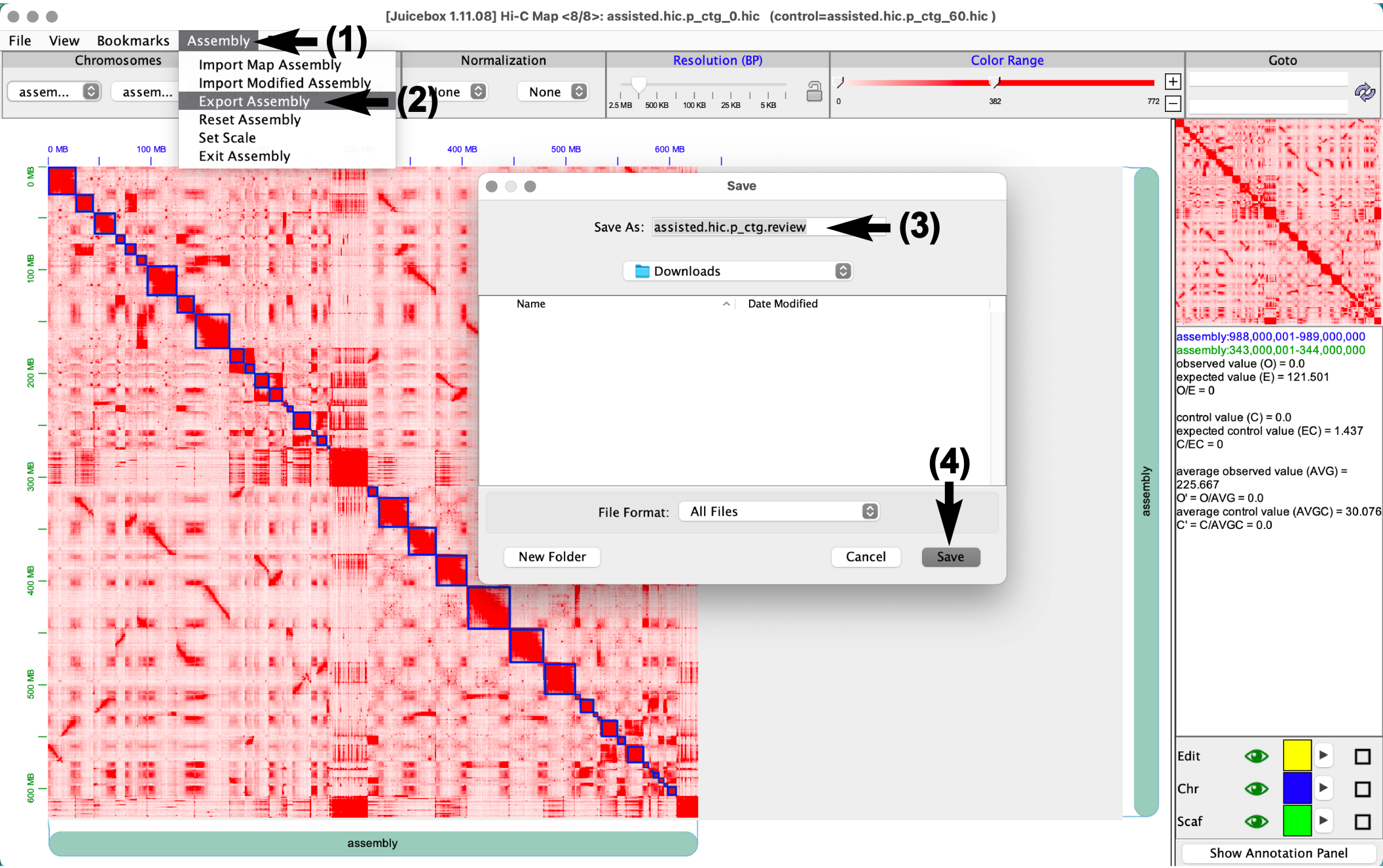
Searching contigs with ‘Goto’



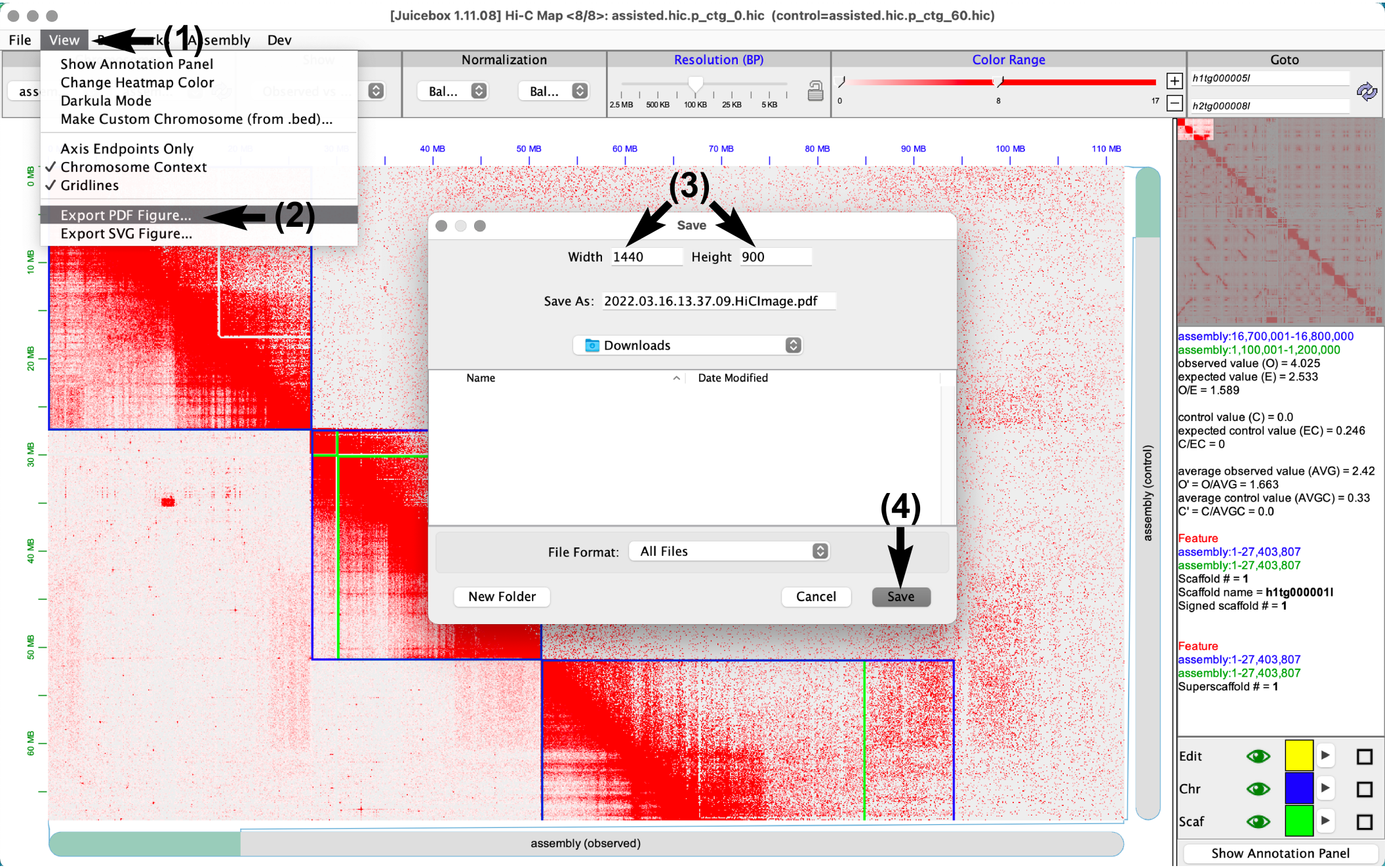
Displaying coverage tracks



Exporting a modified (reviewed) .assembly file

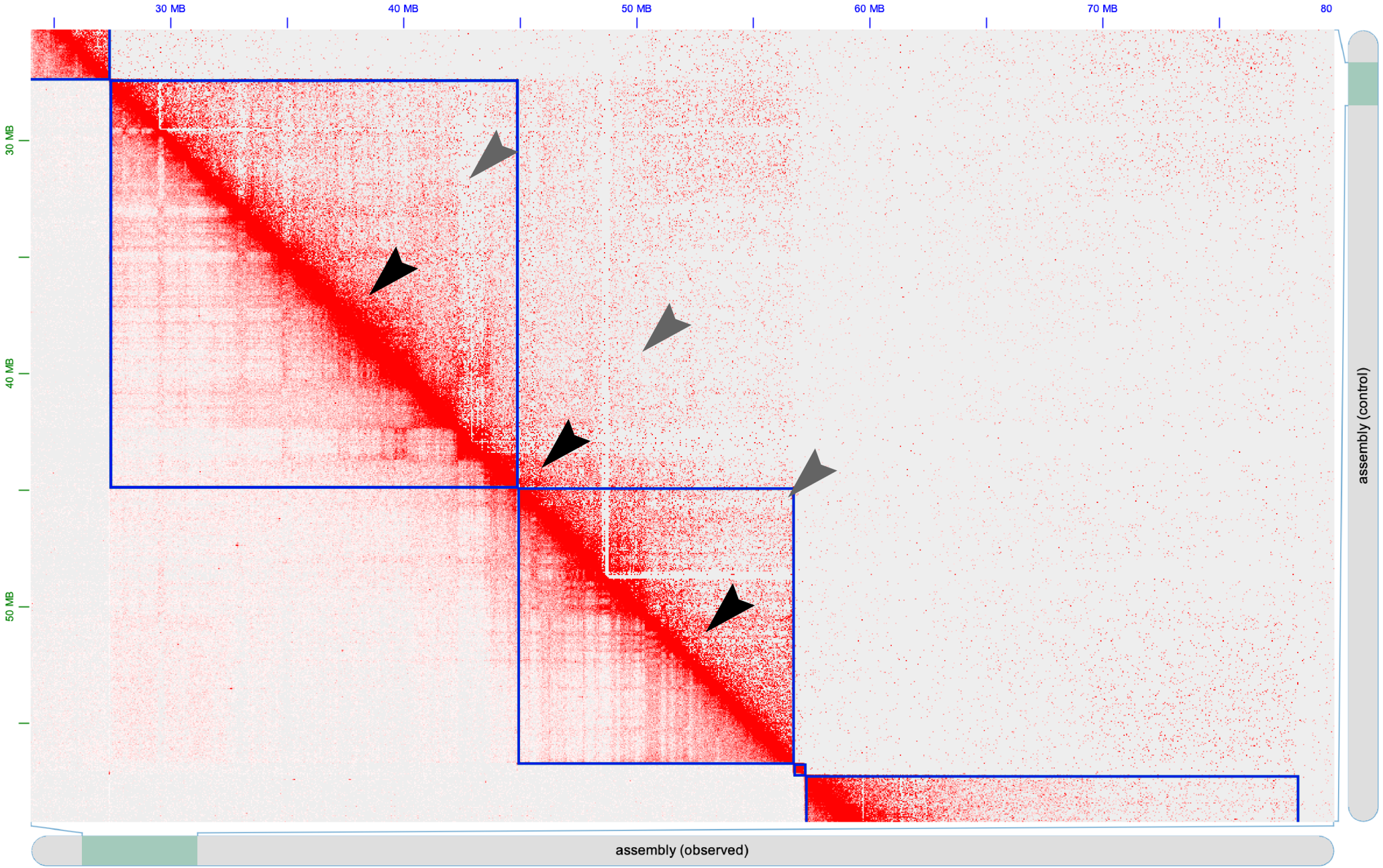


Exporting a PDF/SVG file

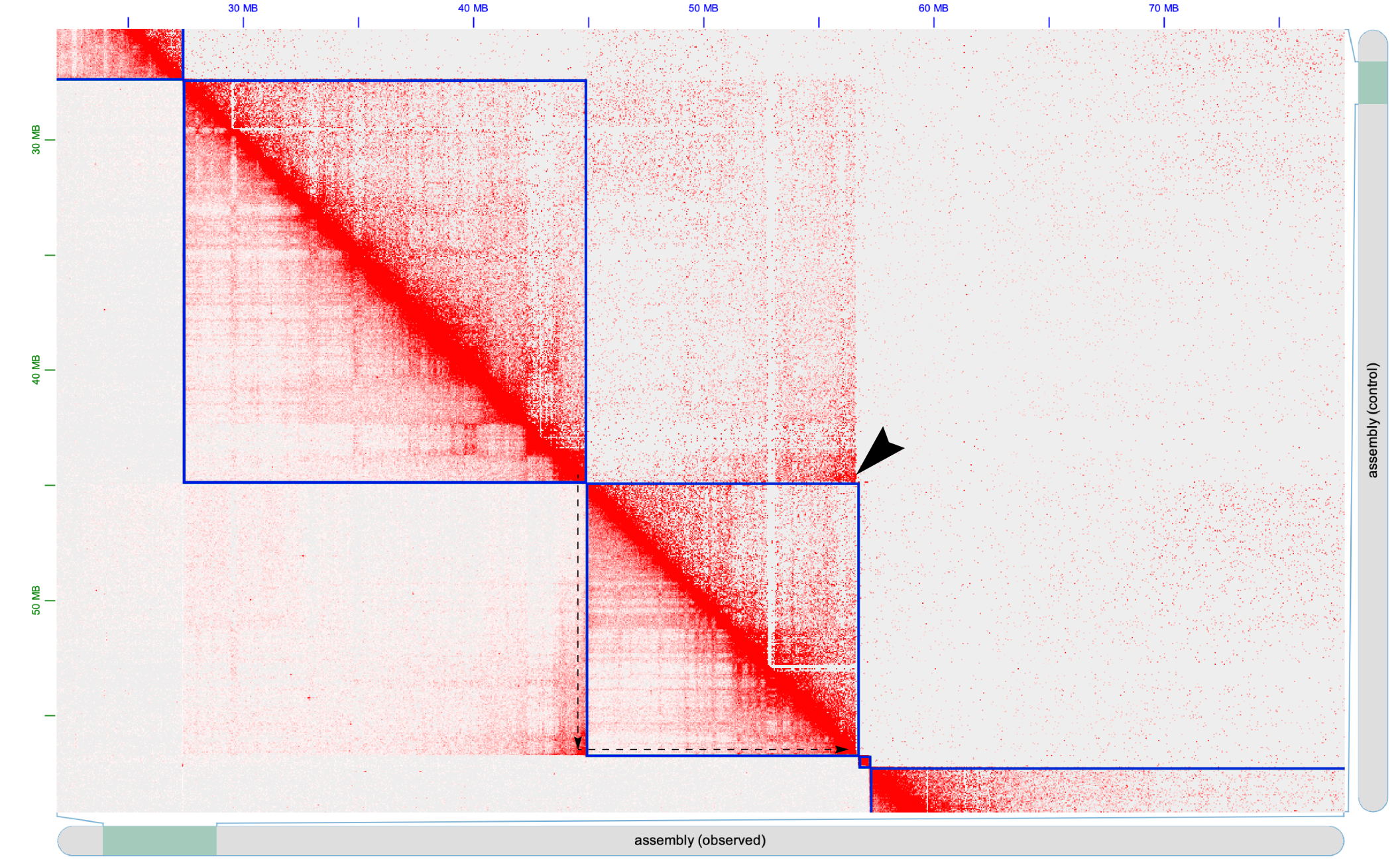


Common Hi-C contact patterns

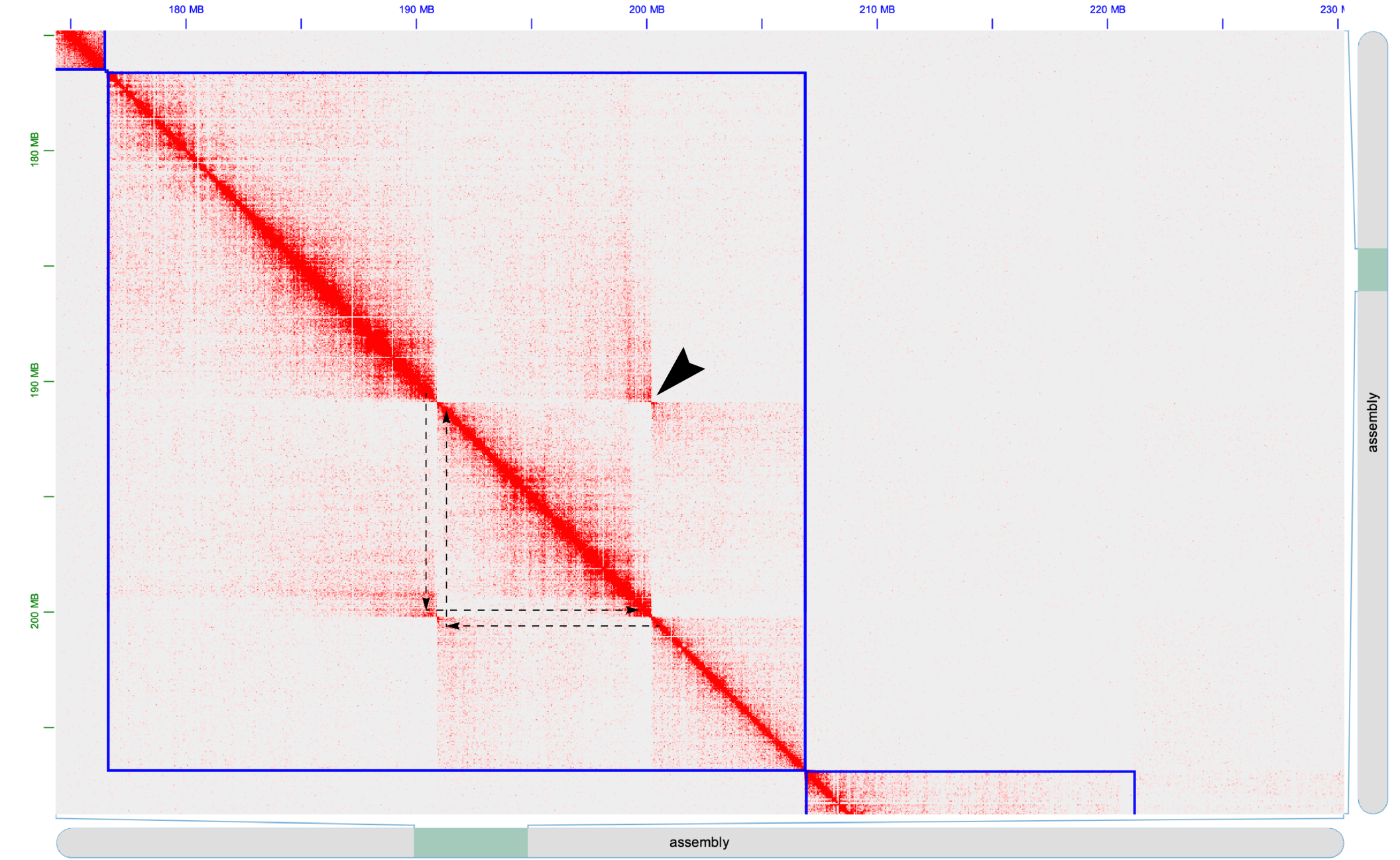
Contigs in correct order and orientation



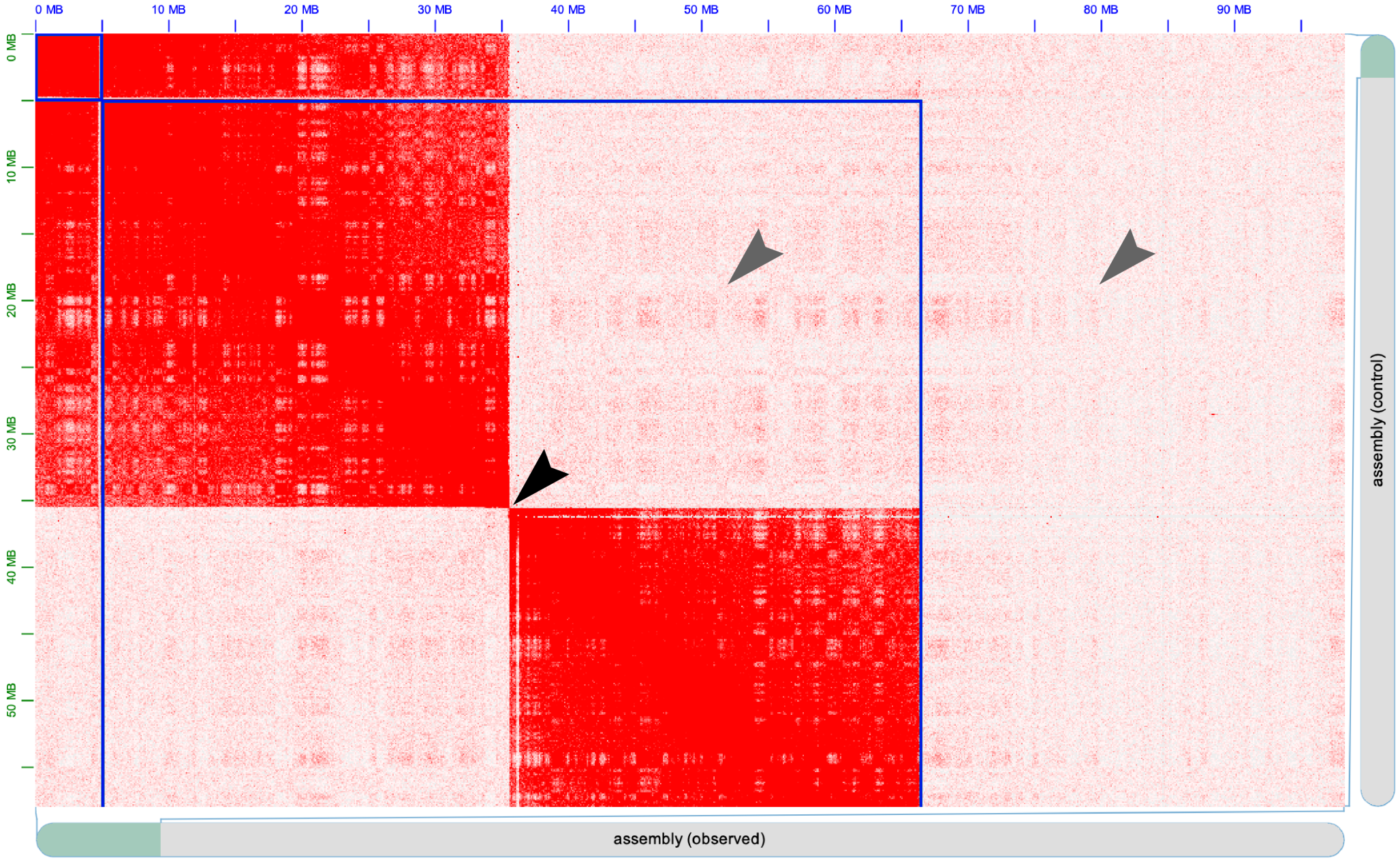
Contig inversion



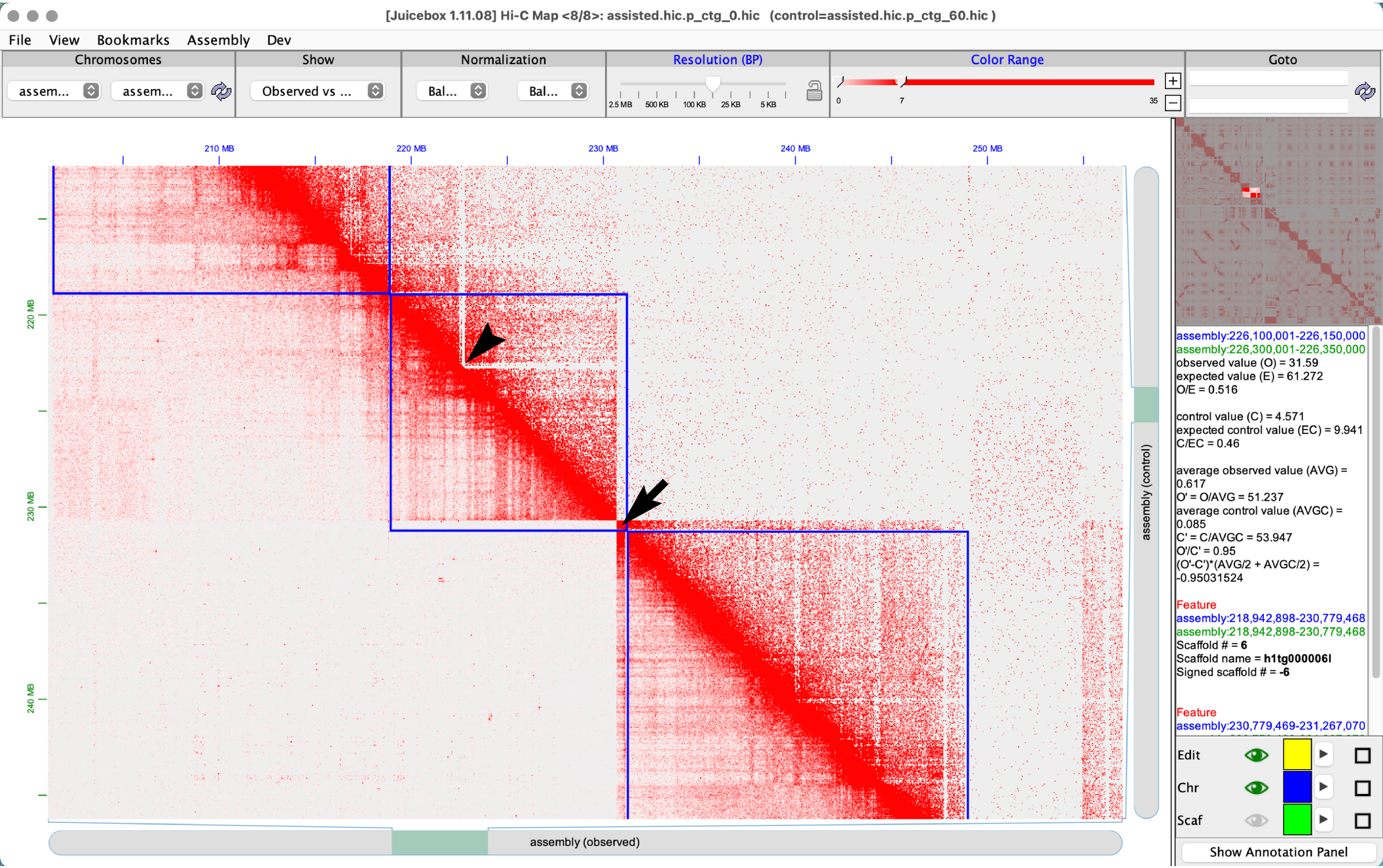
Internal inversion misassembly



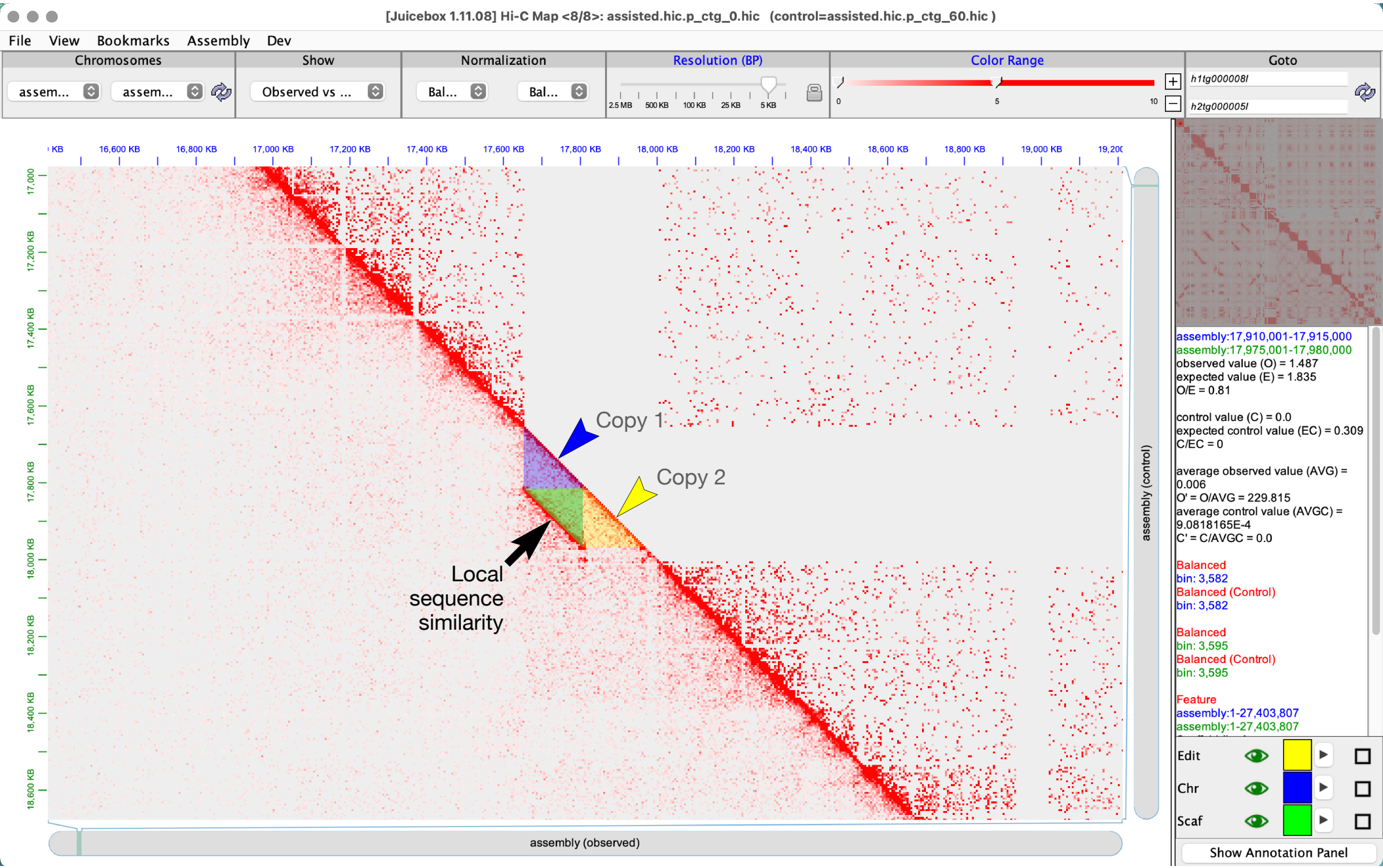
Contigging misjoin error



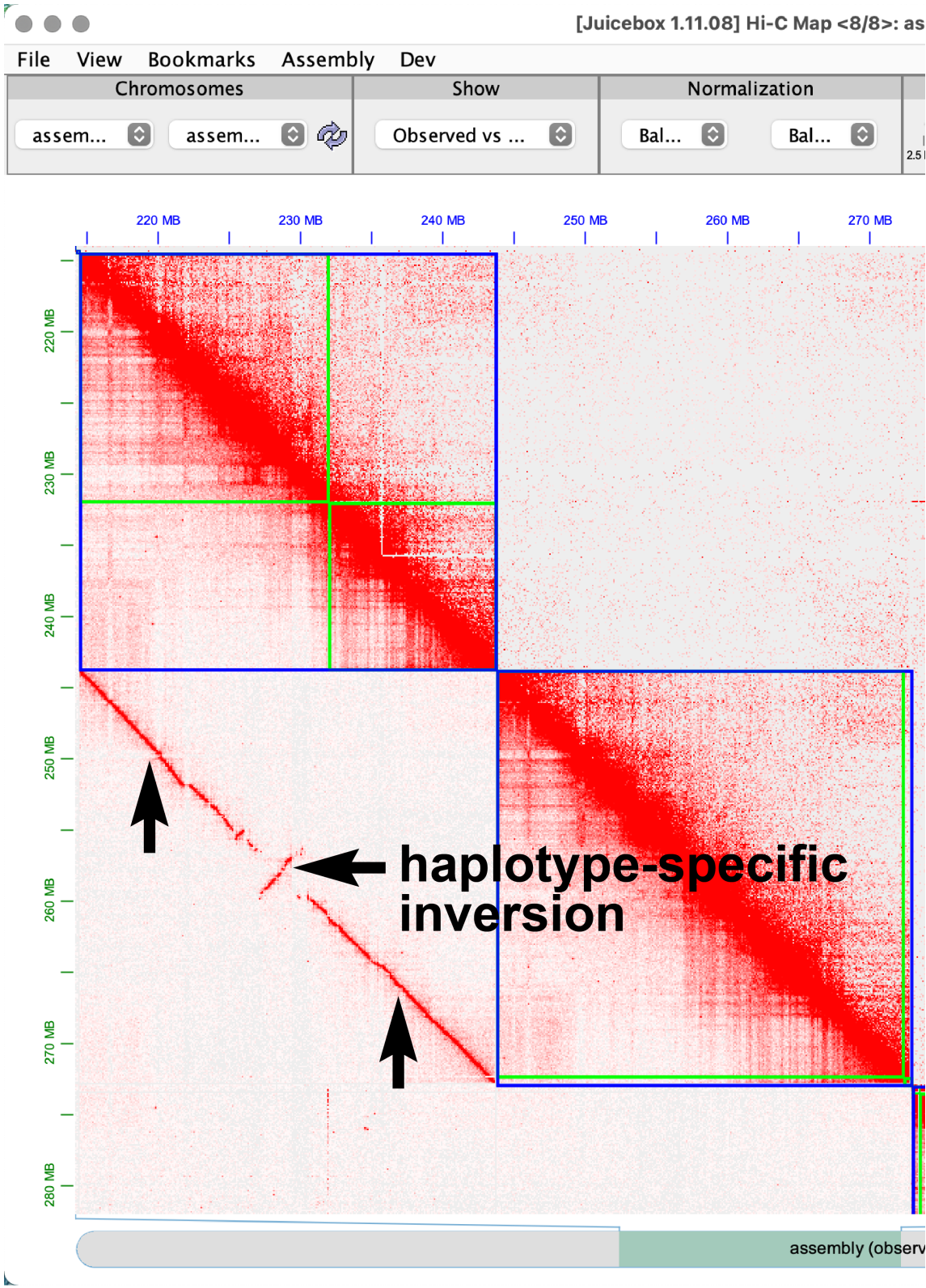
Positive evidence of a contigging misjoin error



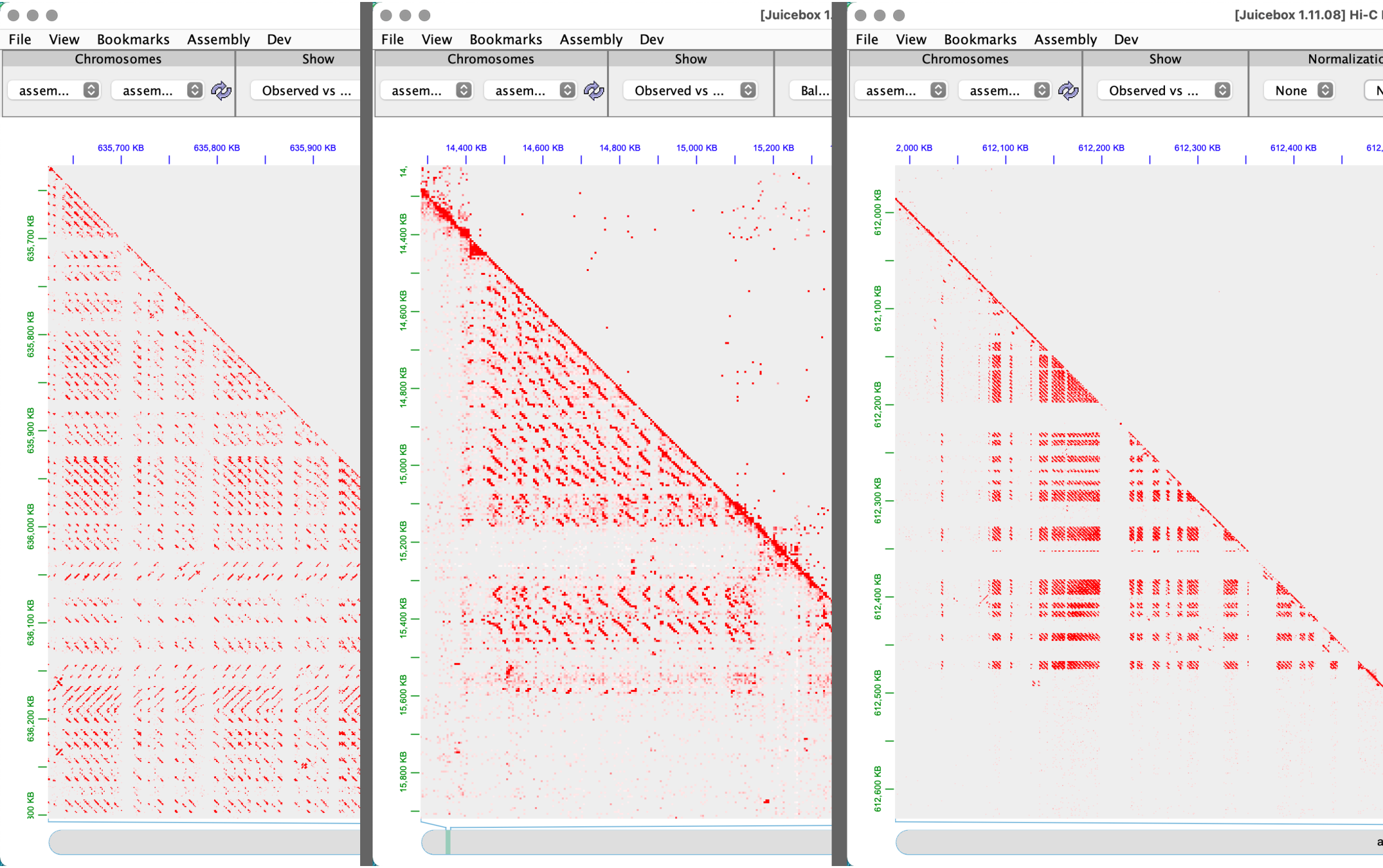
Tandem segmental duplication



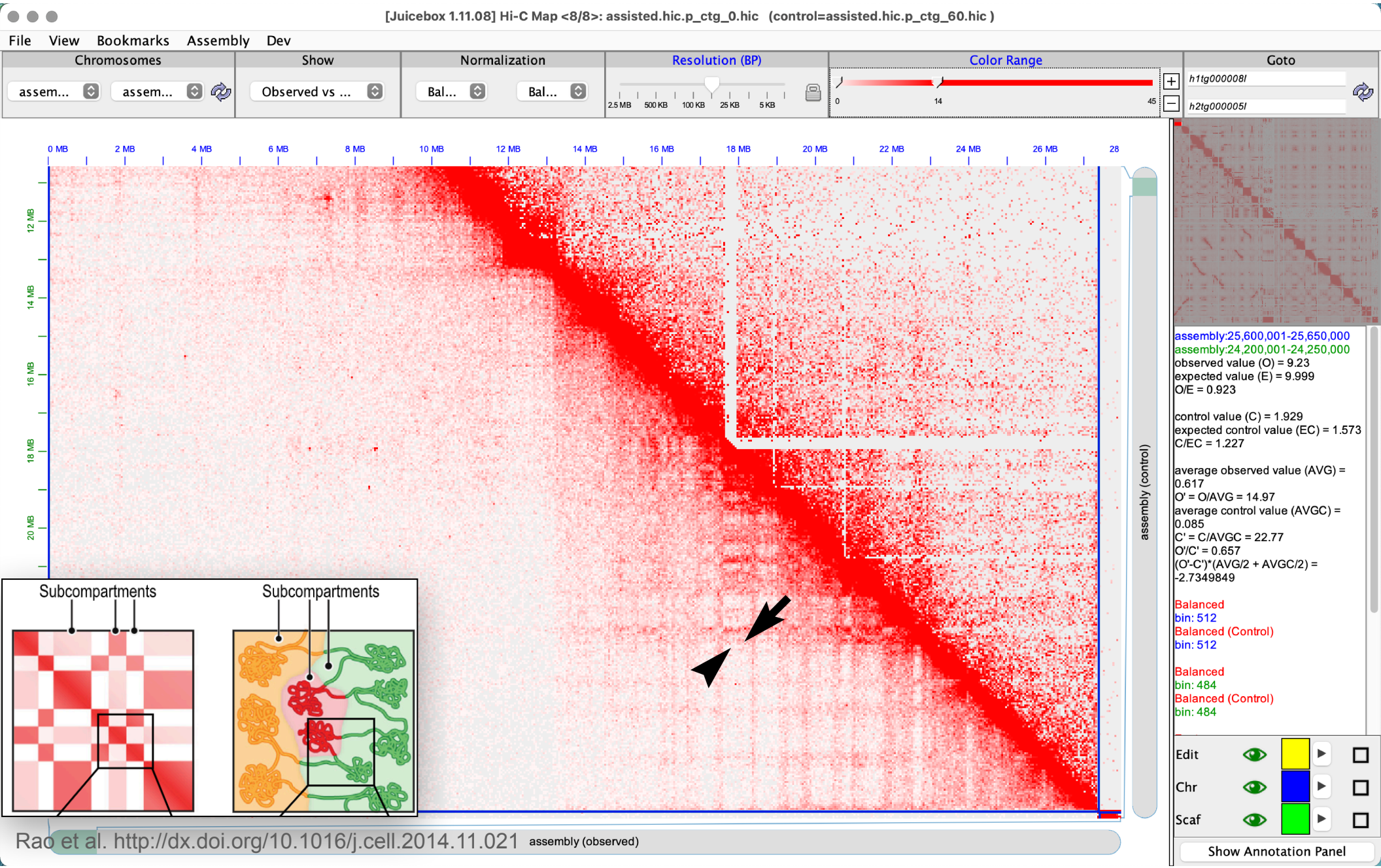
Homologous sequences



Repetitive sequences

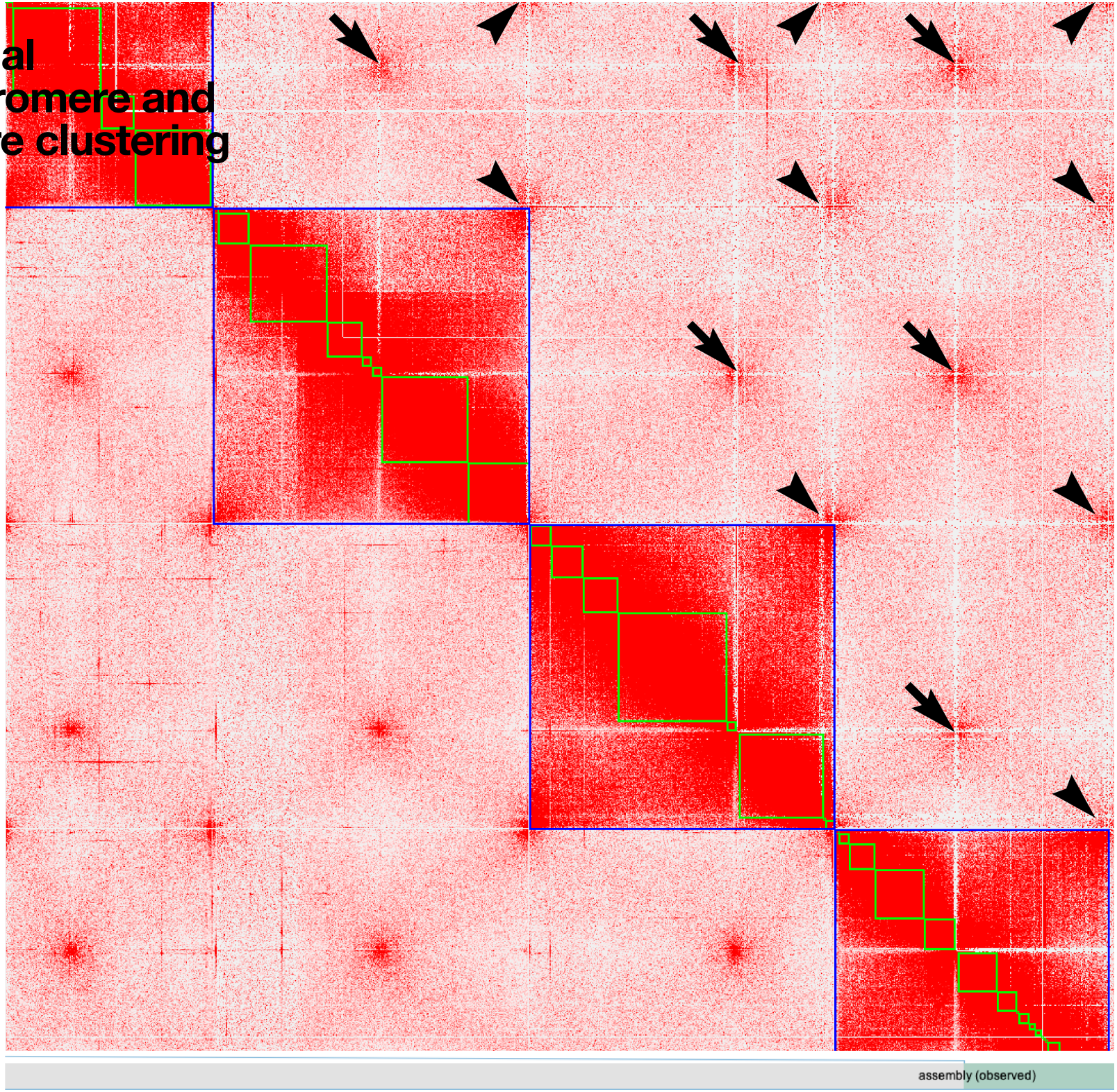


A/B compartmentalization



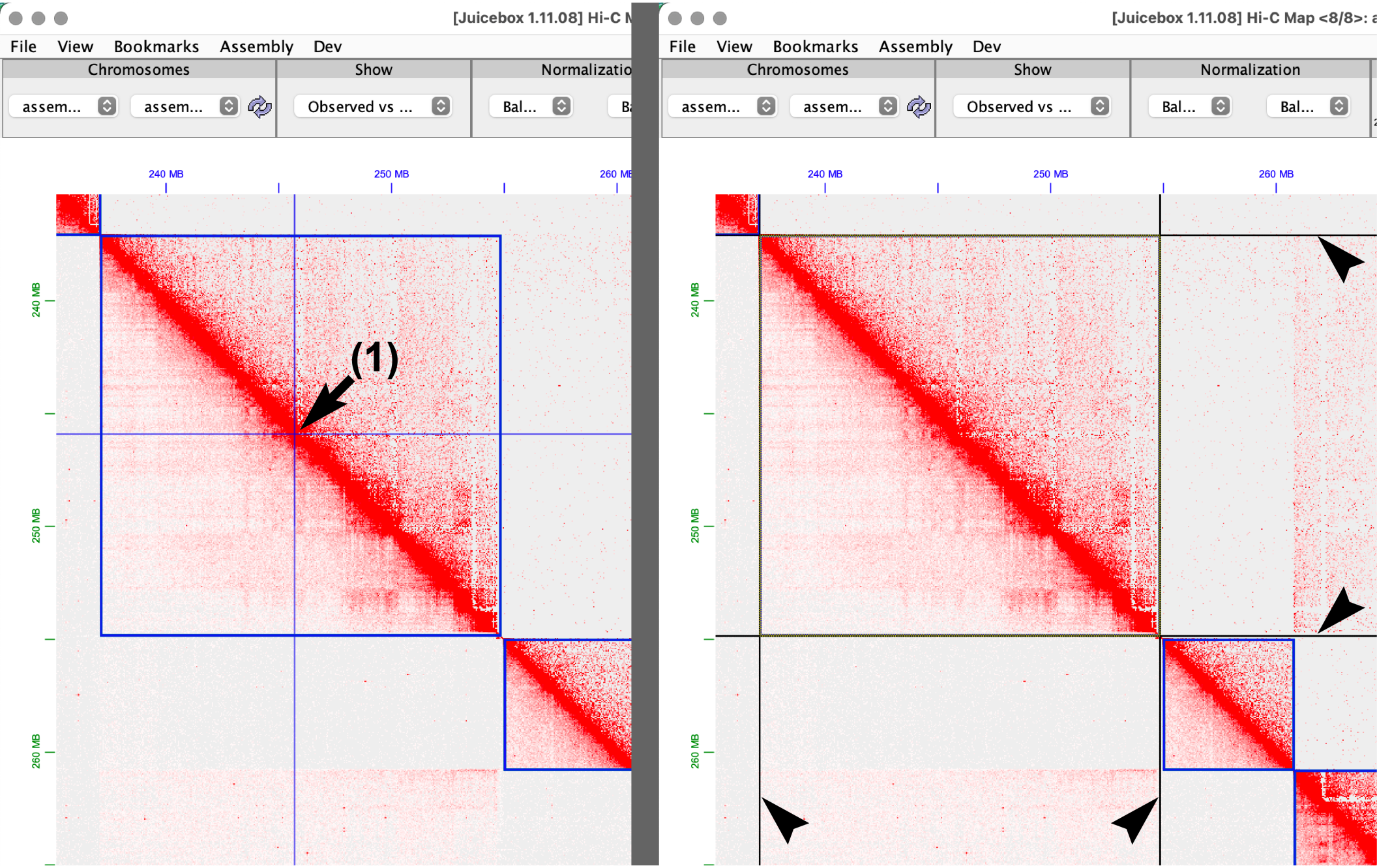
Rao et al. http://dx.doi.org/10.1016/j.cell.2014.11.021

Inter-chromosomal centromere-centromere and telomere-telomere clustering

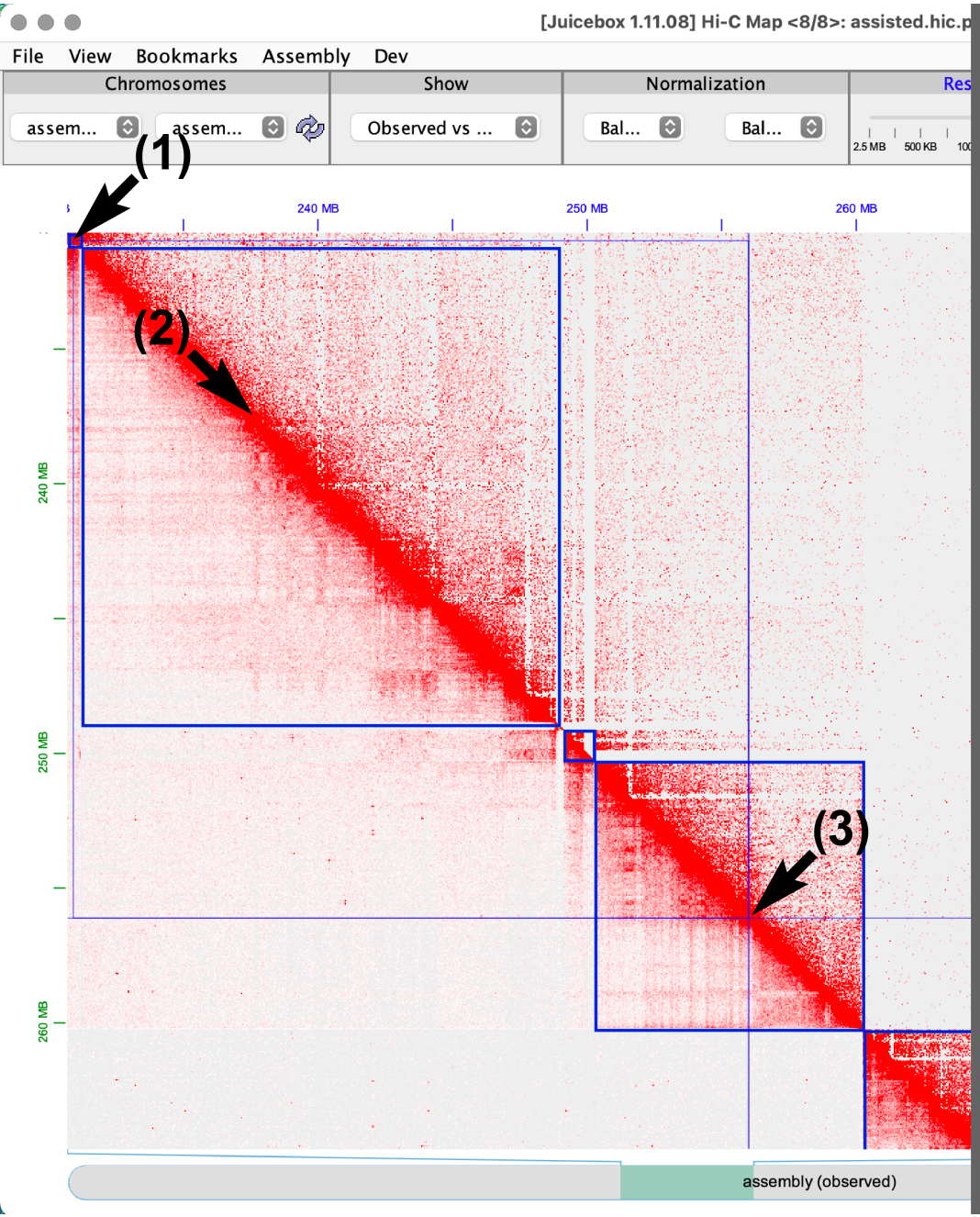


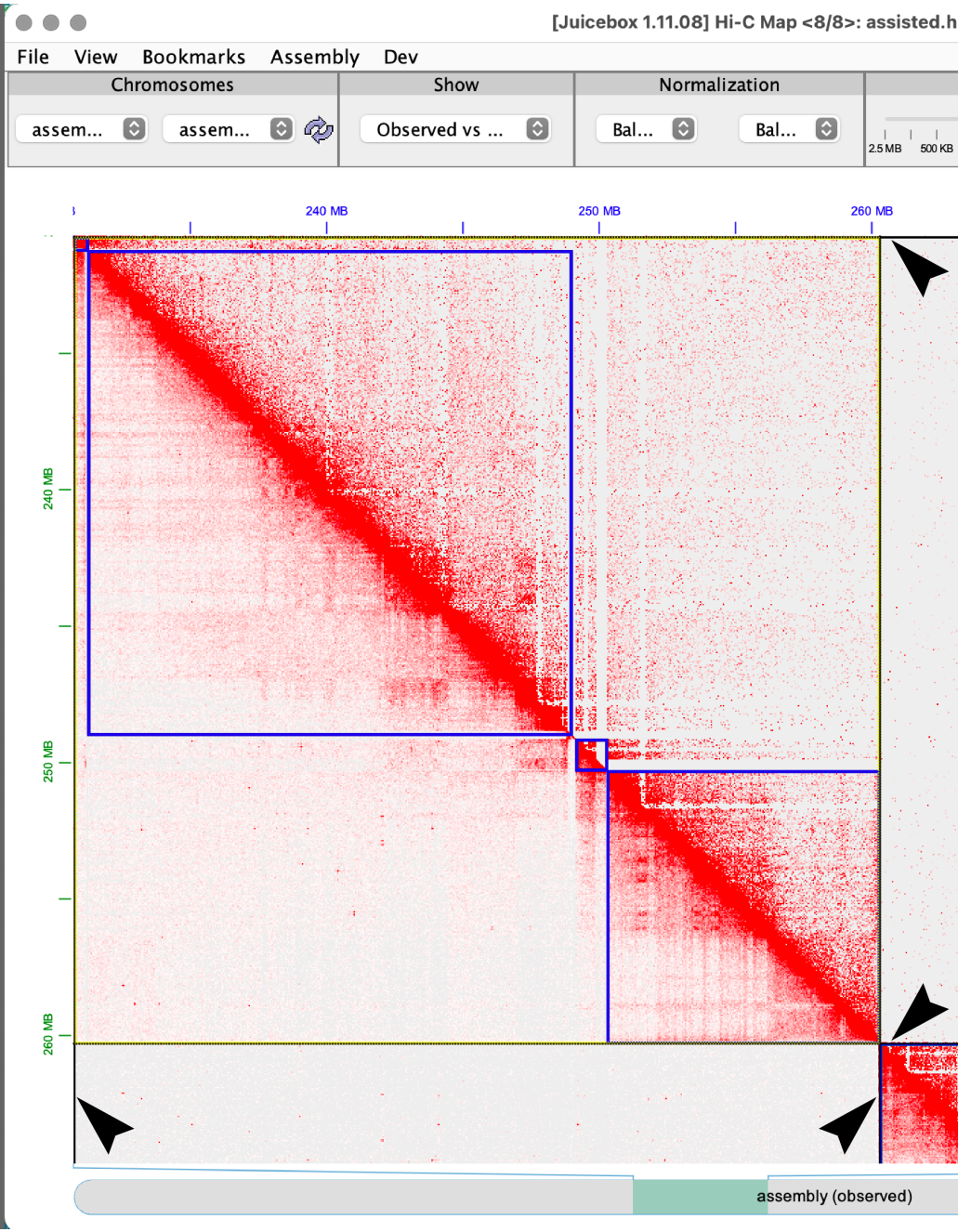
Editing assemblies

Selecting a single contig

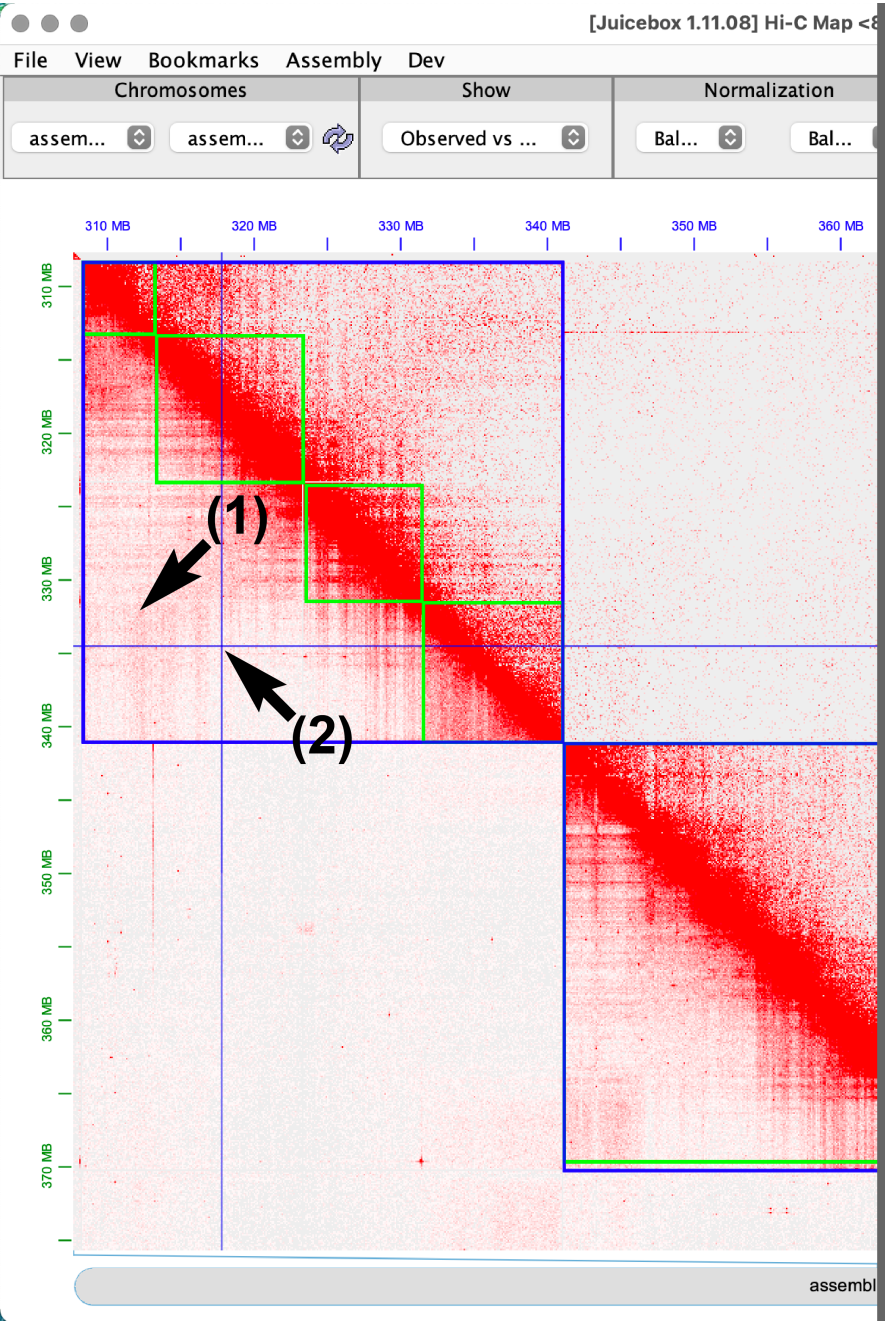


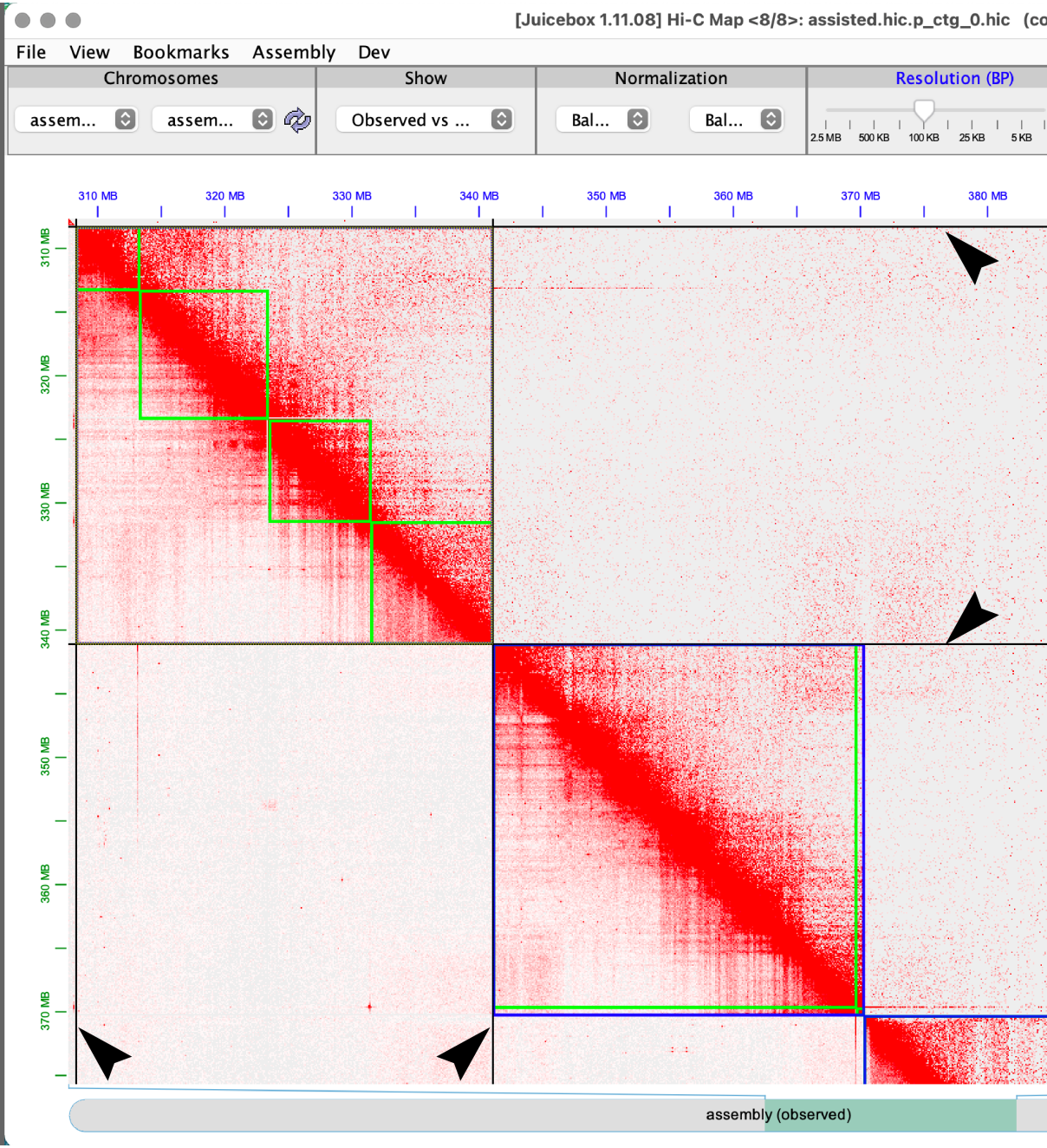
Selecting multiple contigs



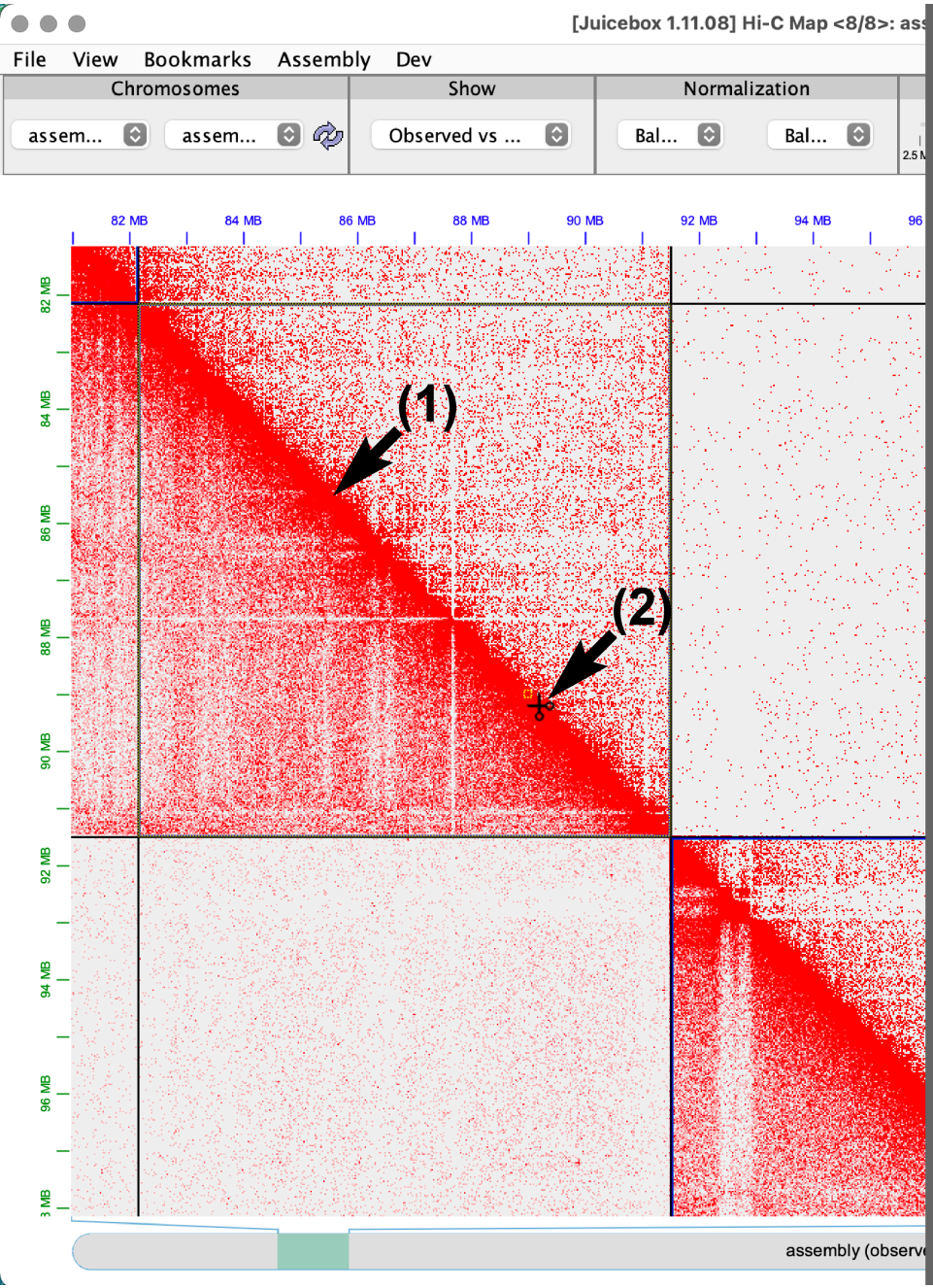


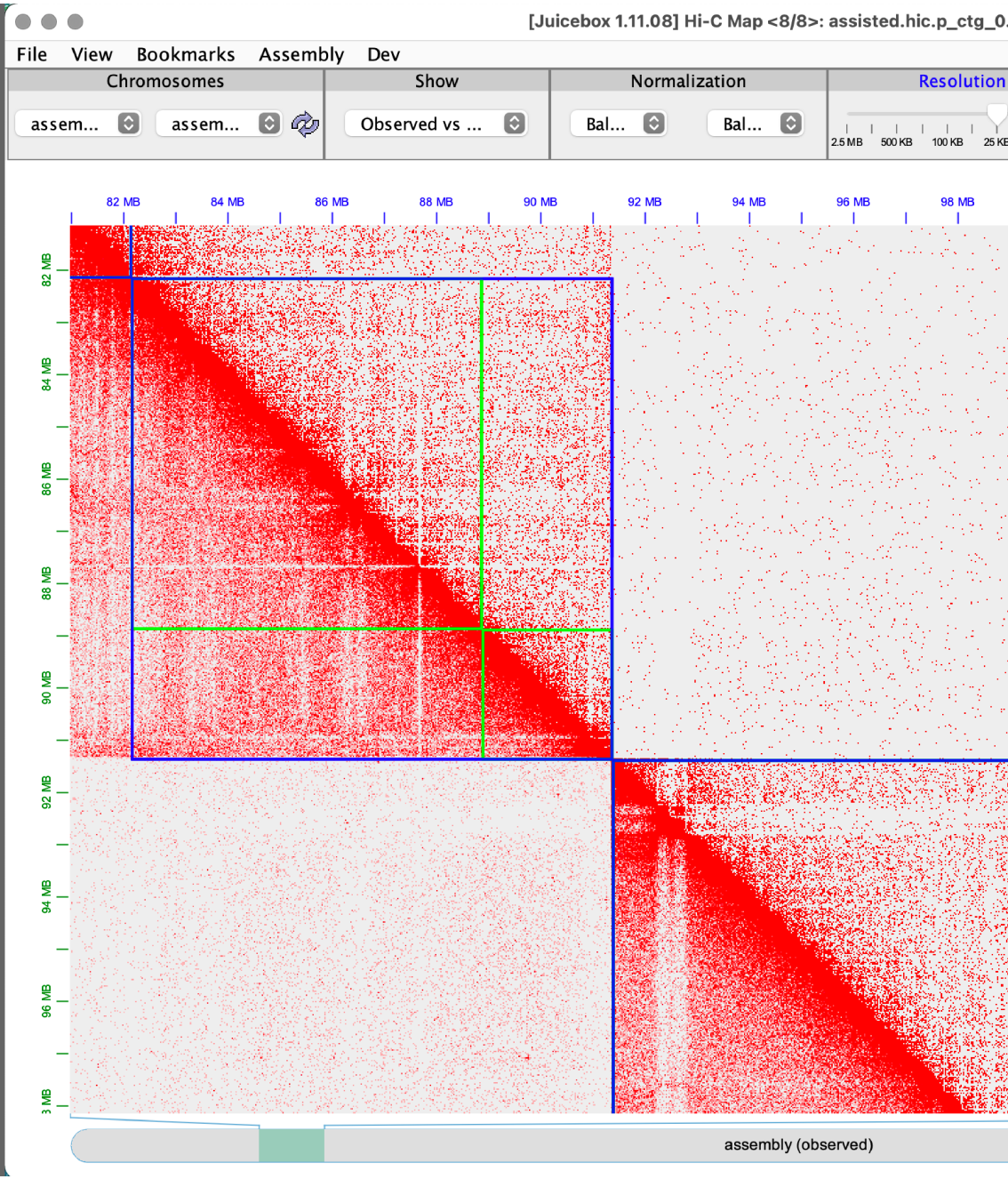
Selecting all contigs in a scaffold



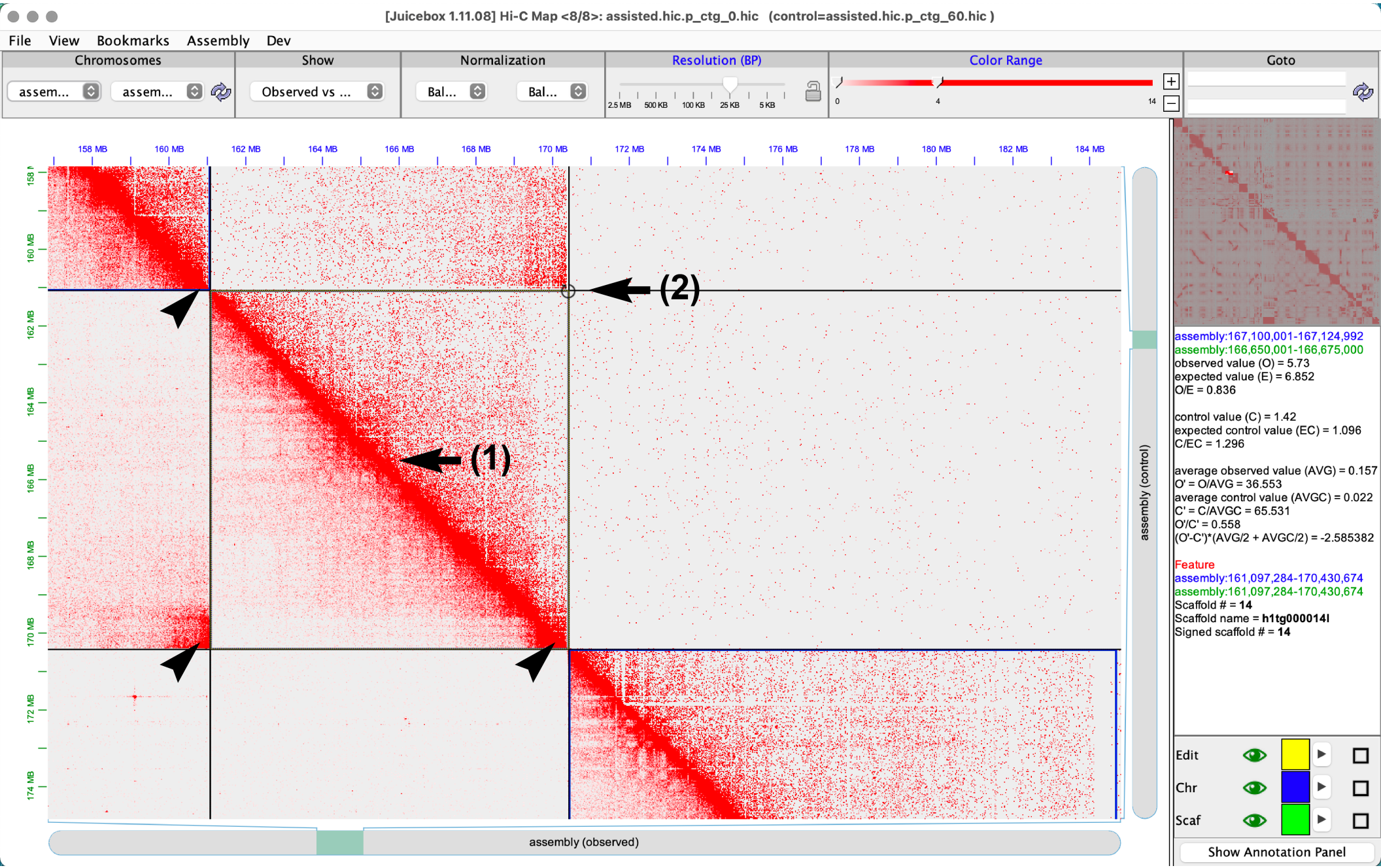


Cutting/breaking a contig

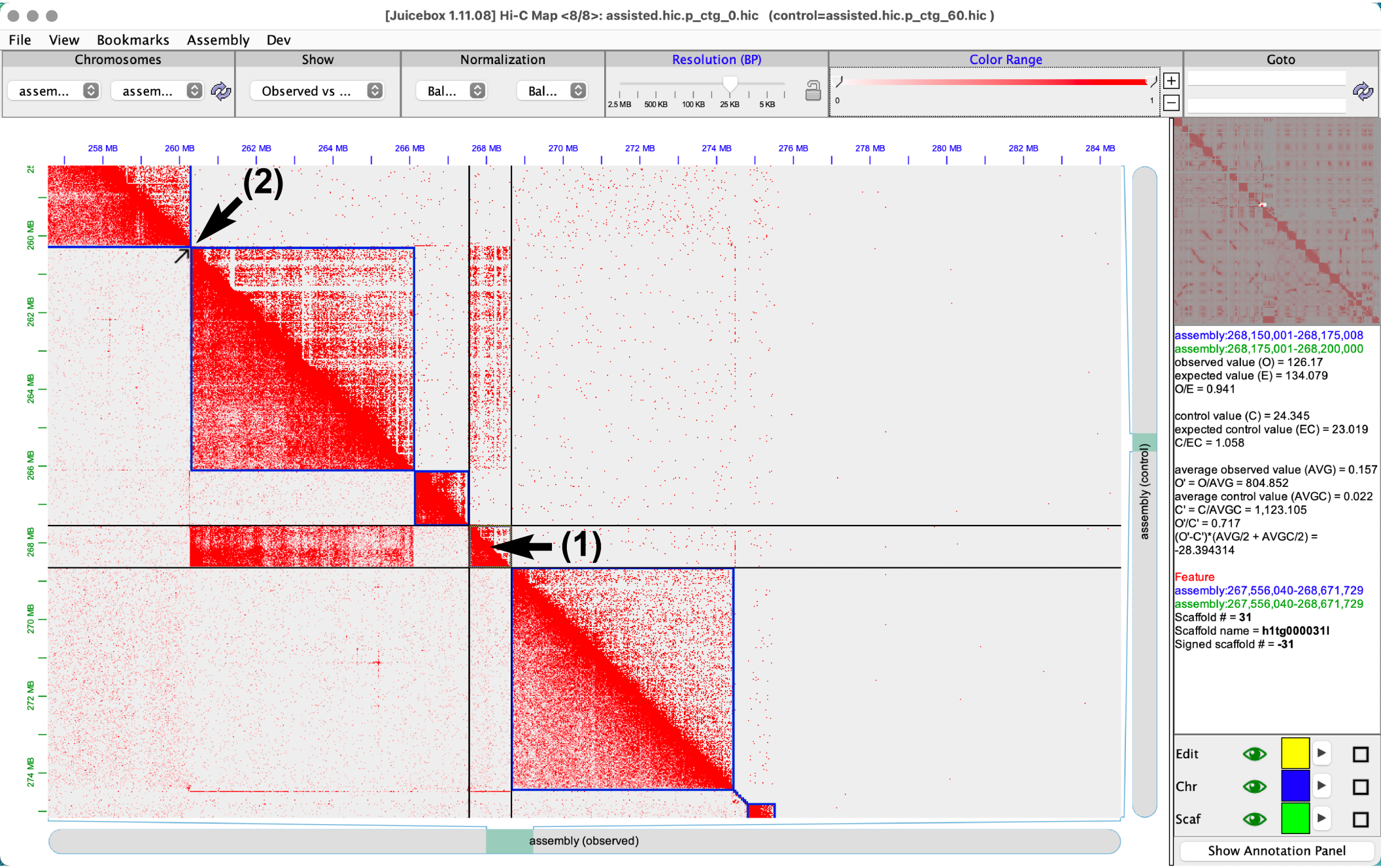




Inverting contig orientation

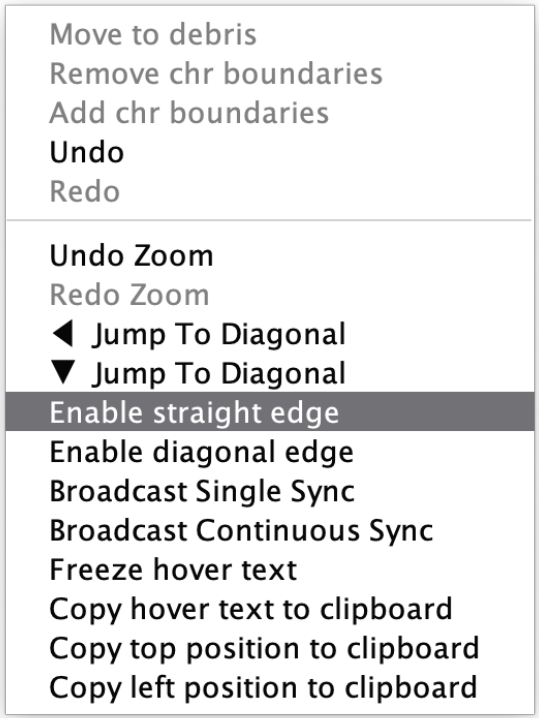


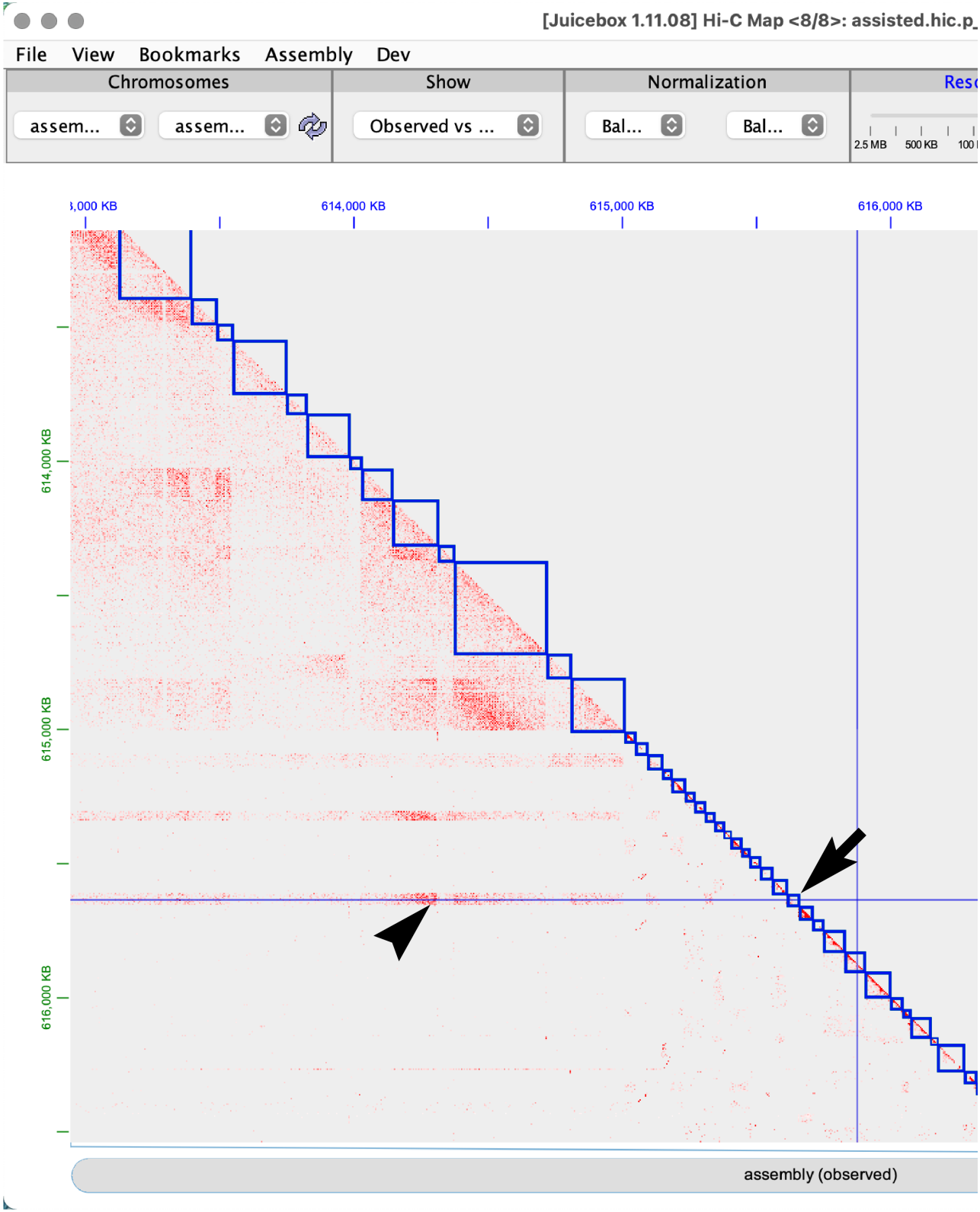
Moving misplaced contigs



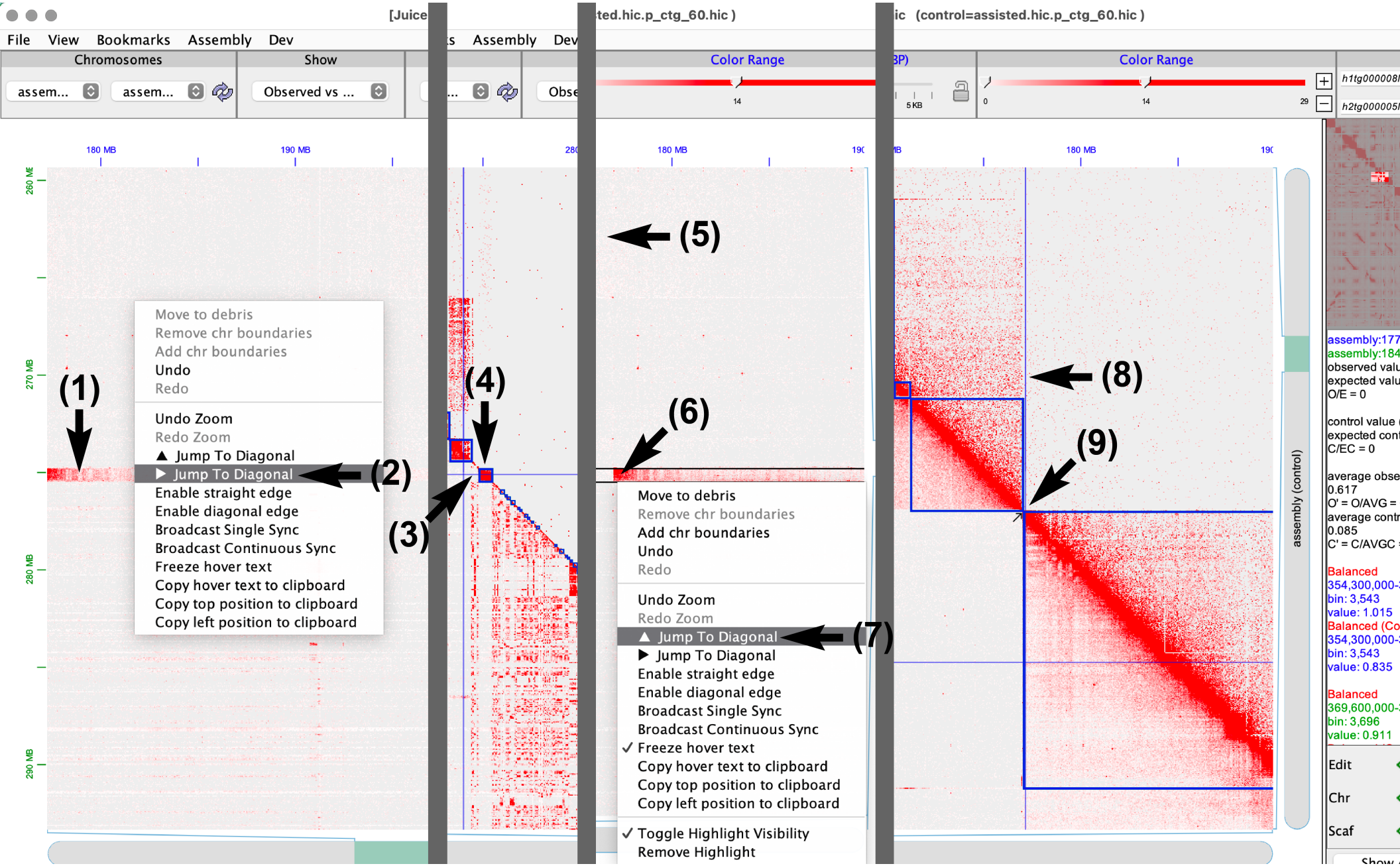
Moving misplaced contigs

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

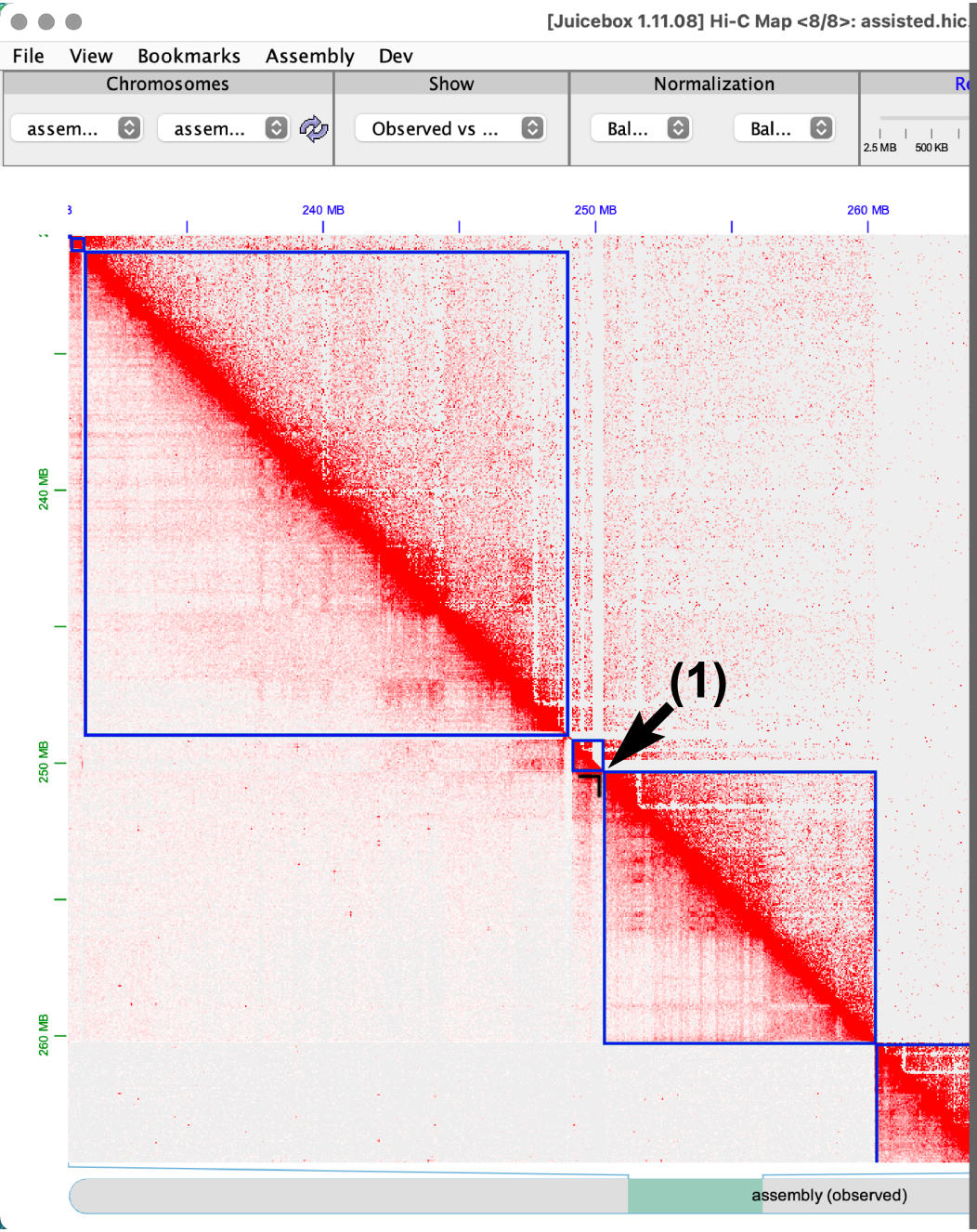


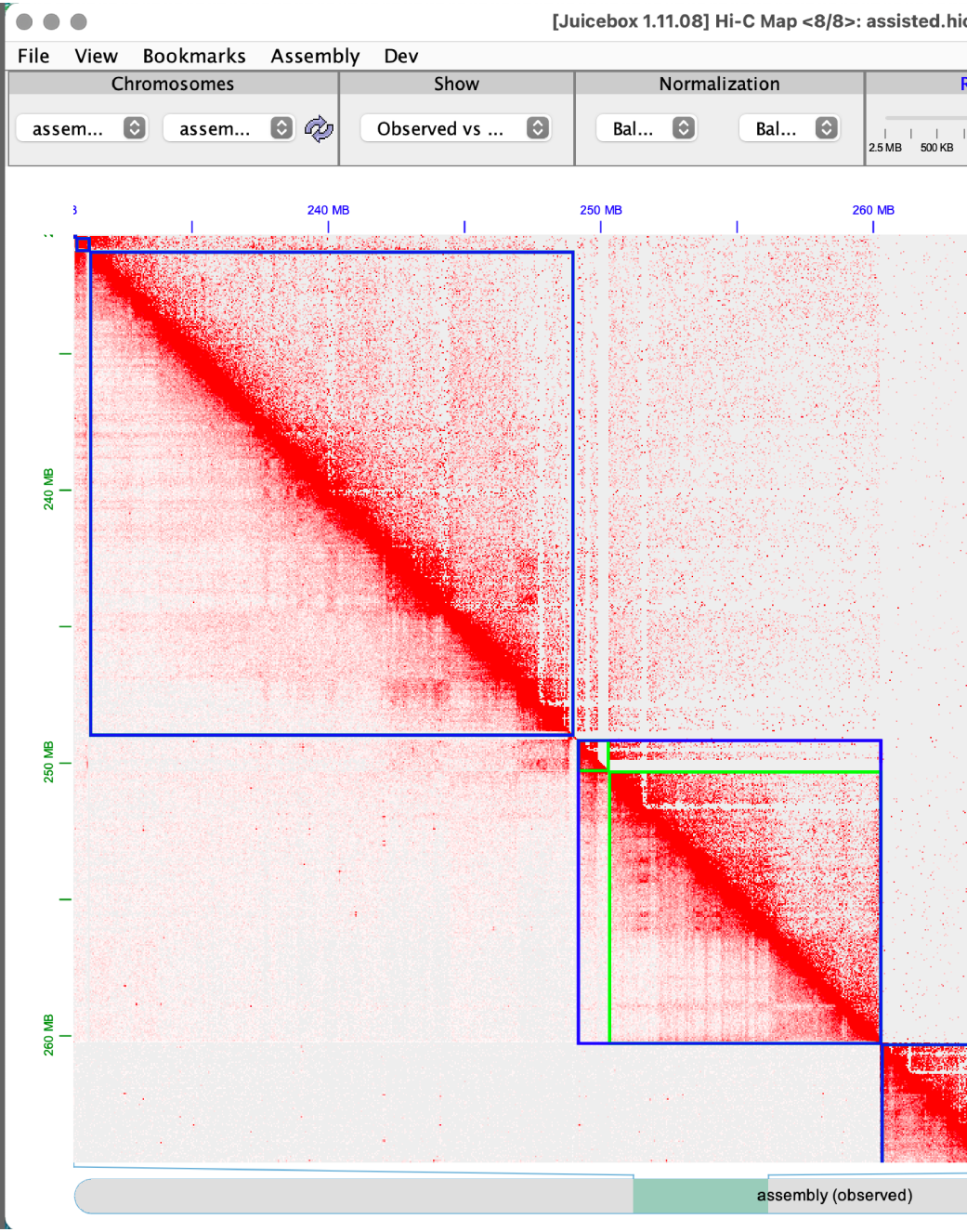


Moving misplaced contigs long distances

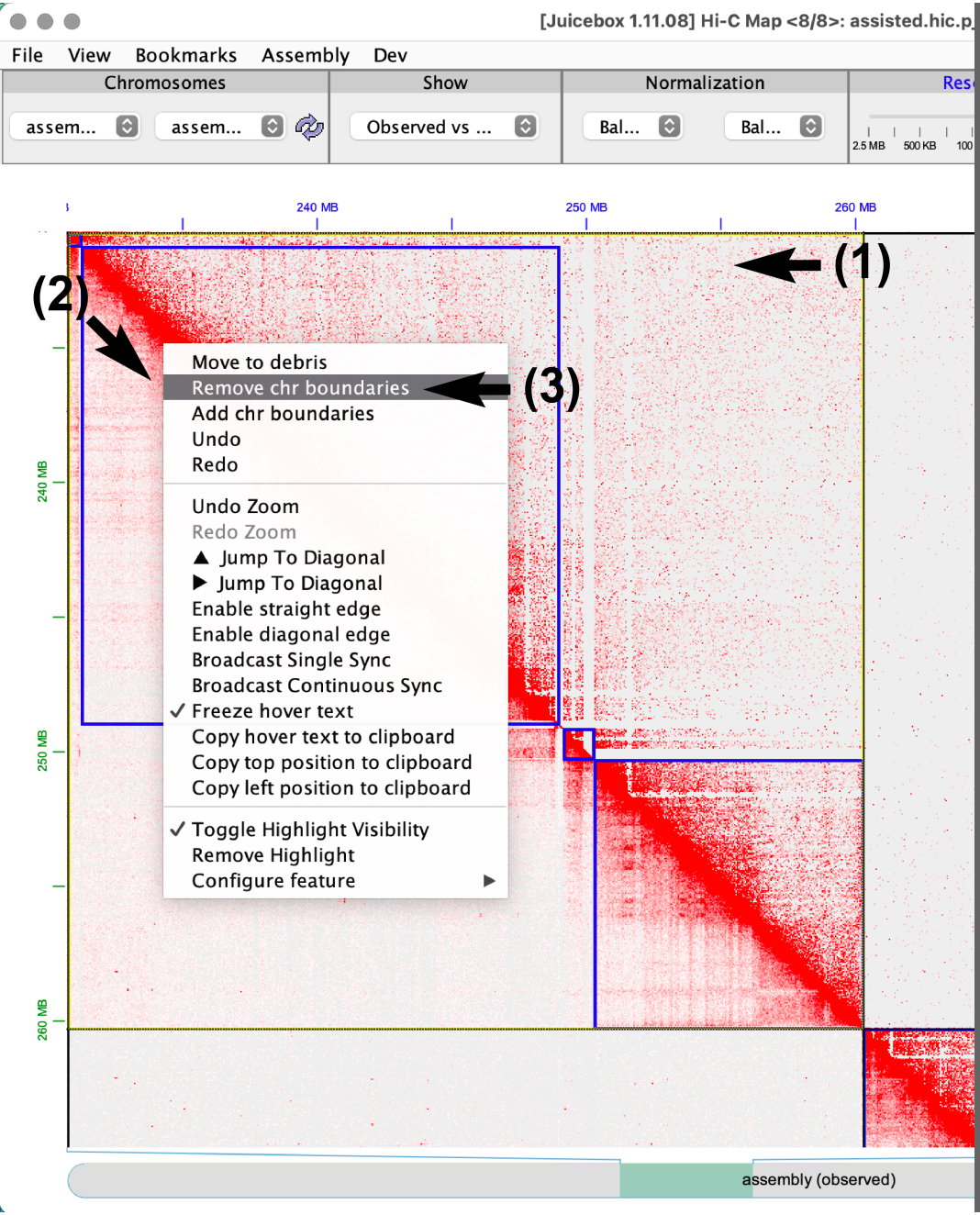


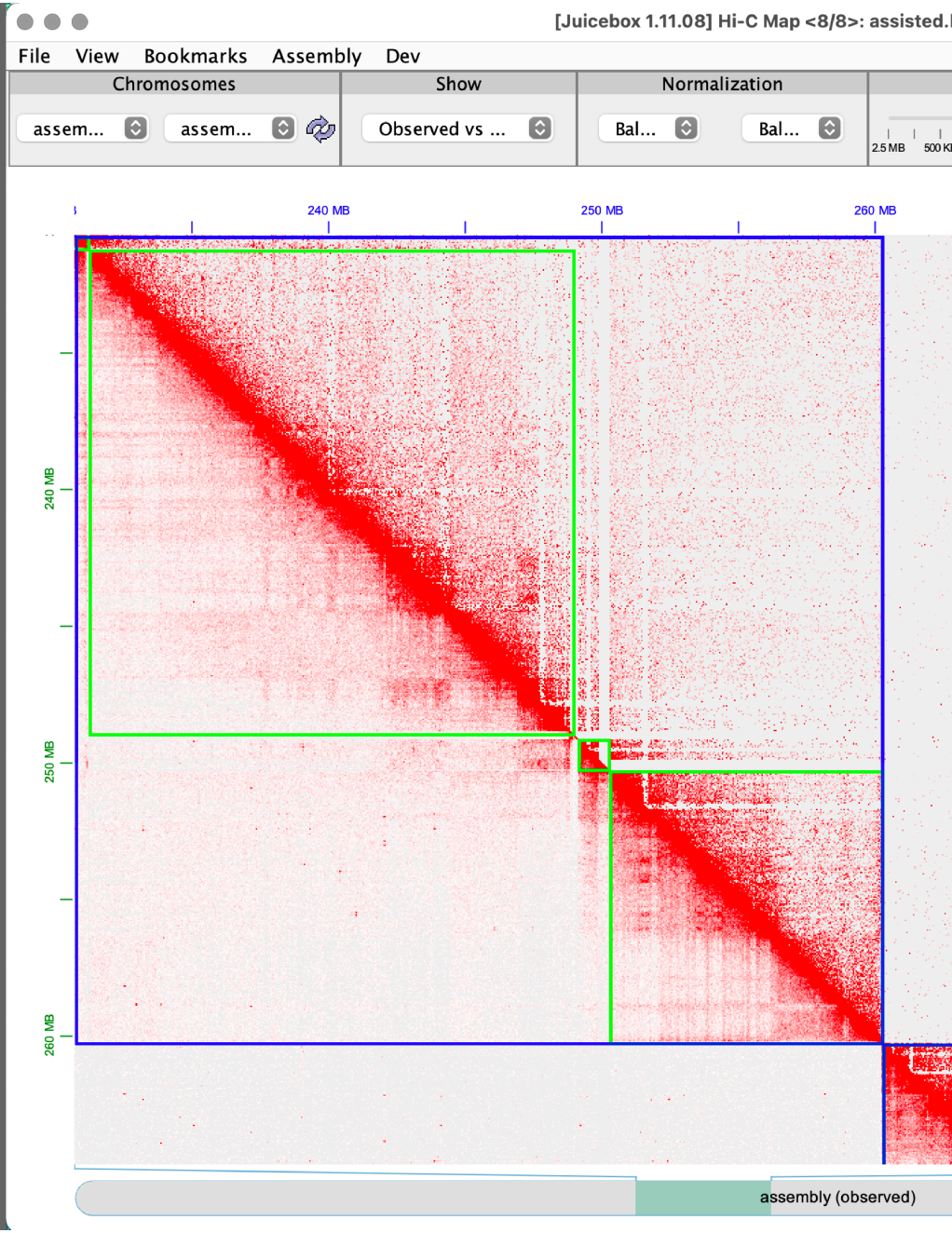
Removing a chromosome boundary between two contigs



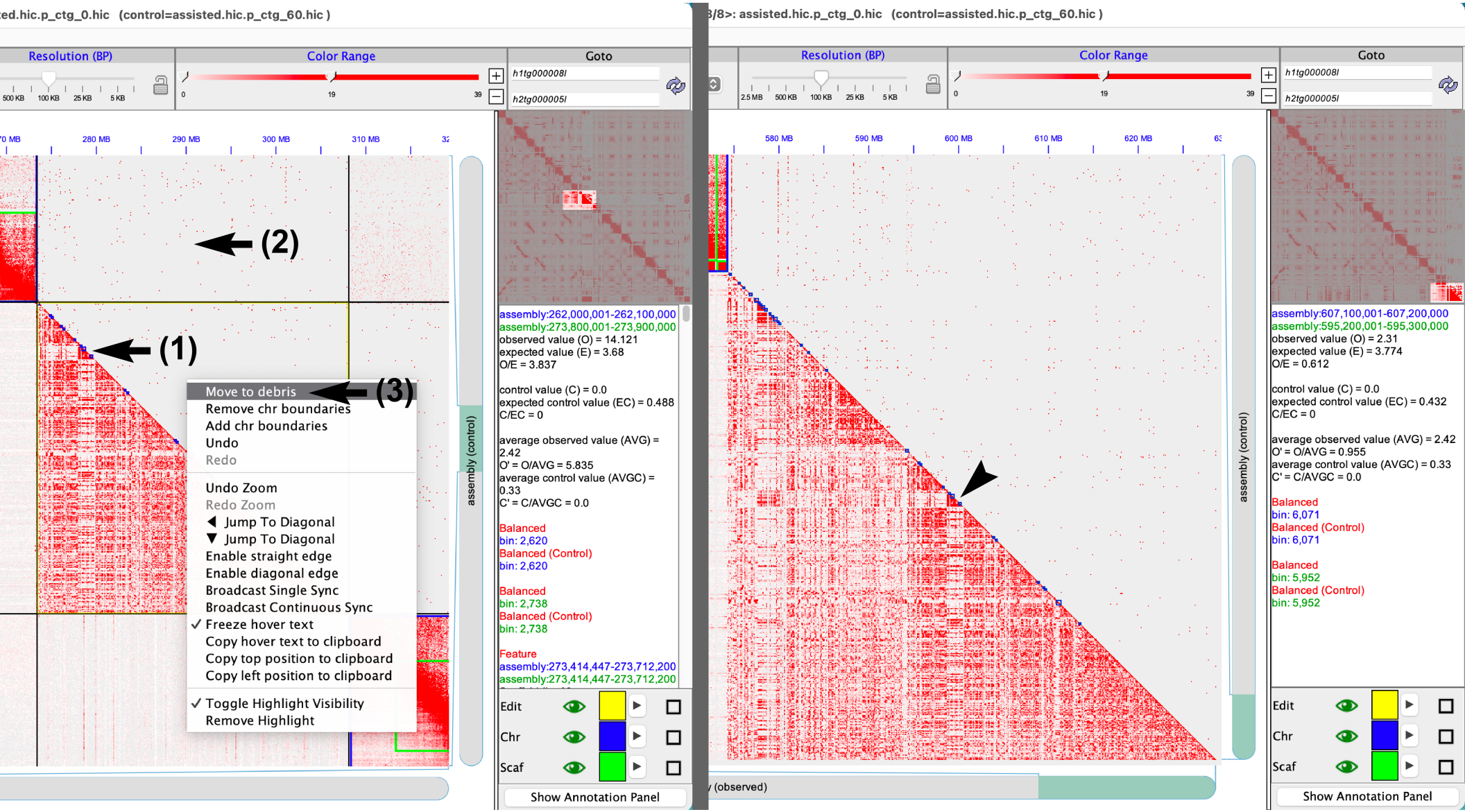


Removing chromosome boundaries between many contigs





Moving contigs to “debris”



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 &
```

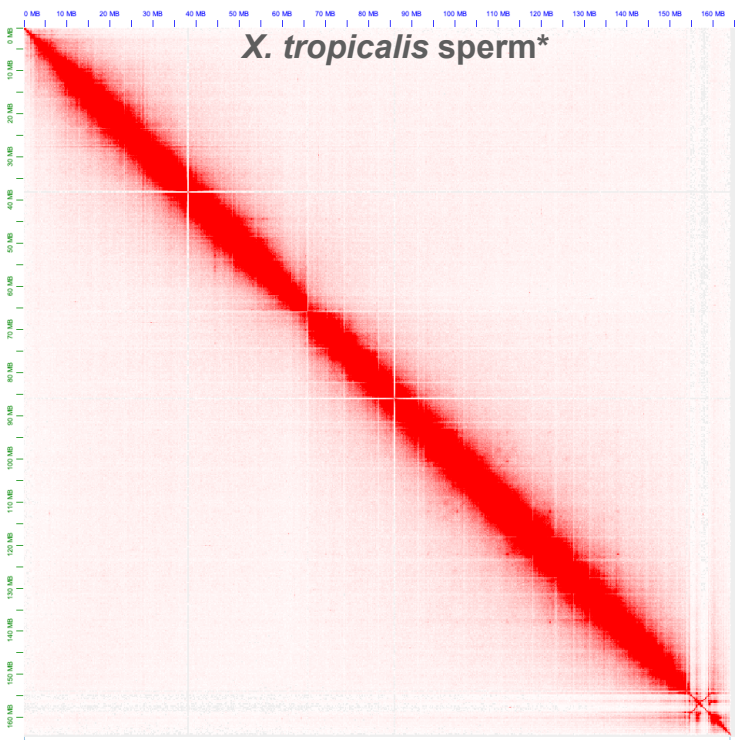
To see all options (and their descriptions) offered by the post-review script, do:

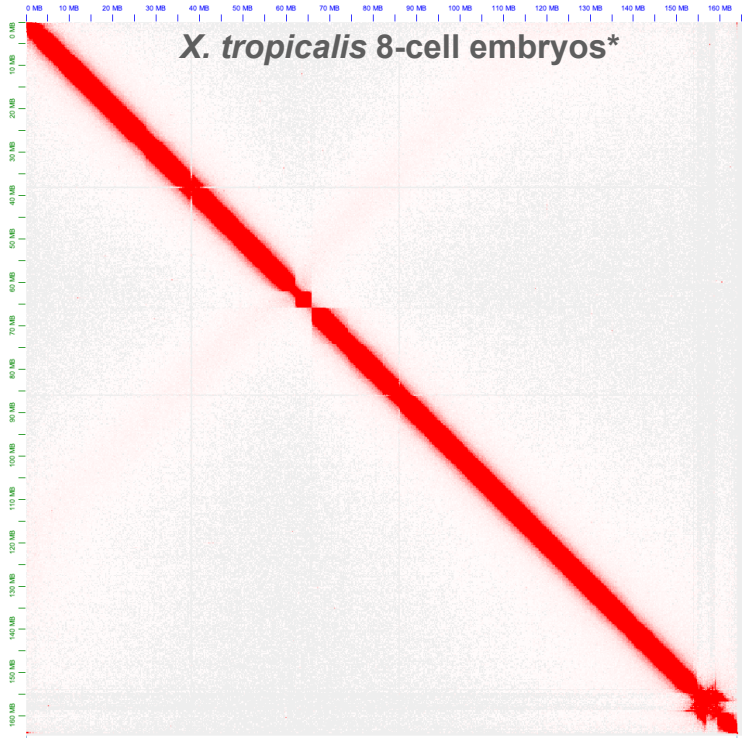
```
3d-dna/run-asm-pipeline-post-review.sh --help
```

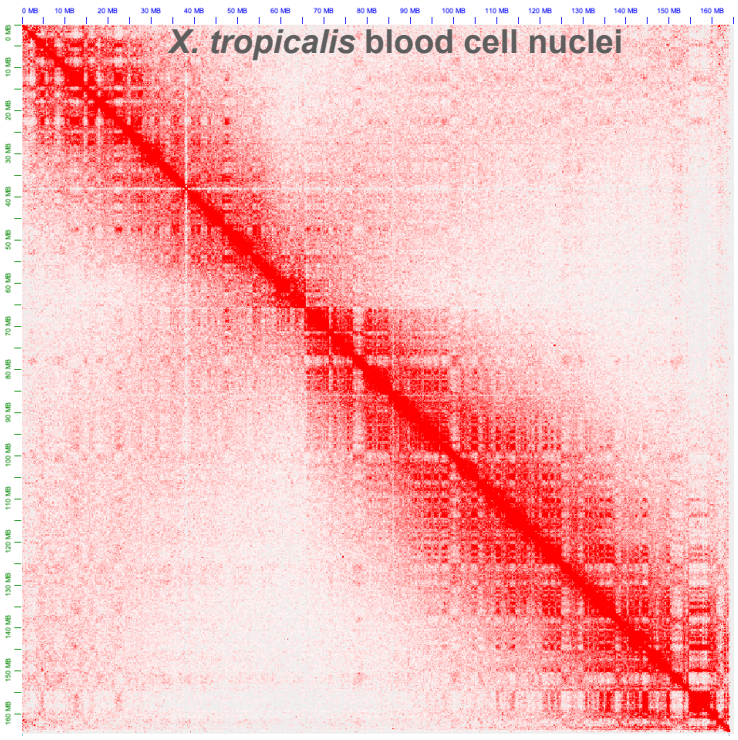
Genome post-review finalization can also be performed with artisanal:

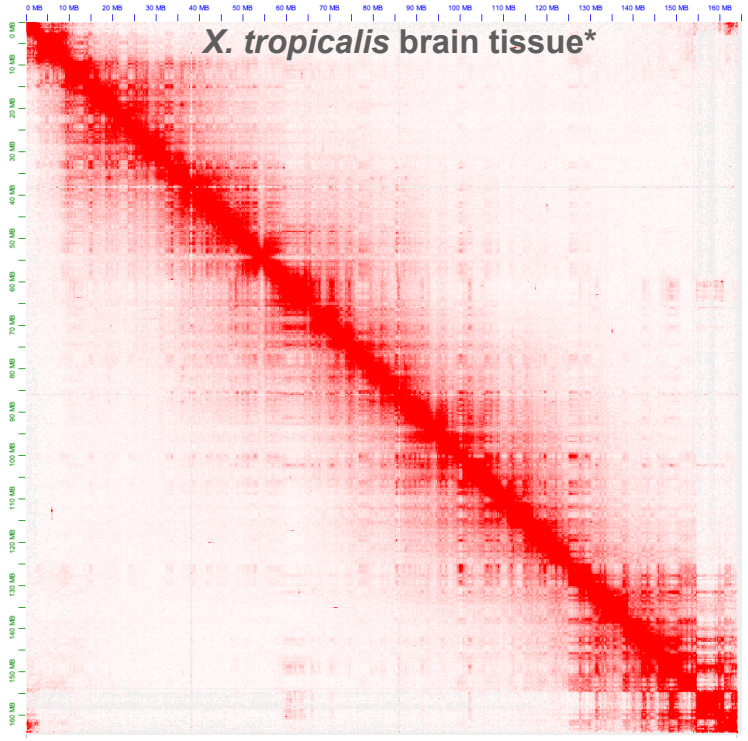
```
assembly-to-fasta genome.review.assembly genome.fasta genome.review
```

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









*Niu et al. https://doi.org/10.1038/s41588-021-00878-z