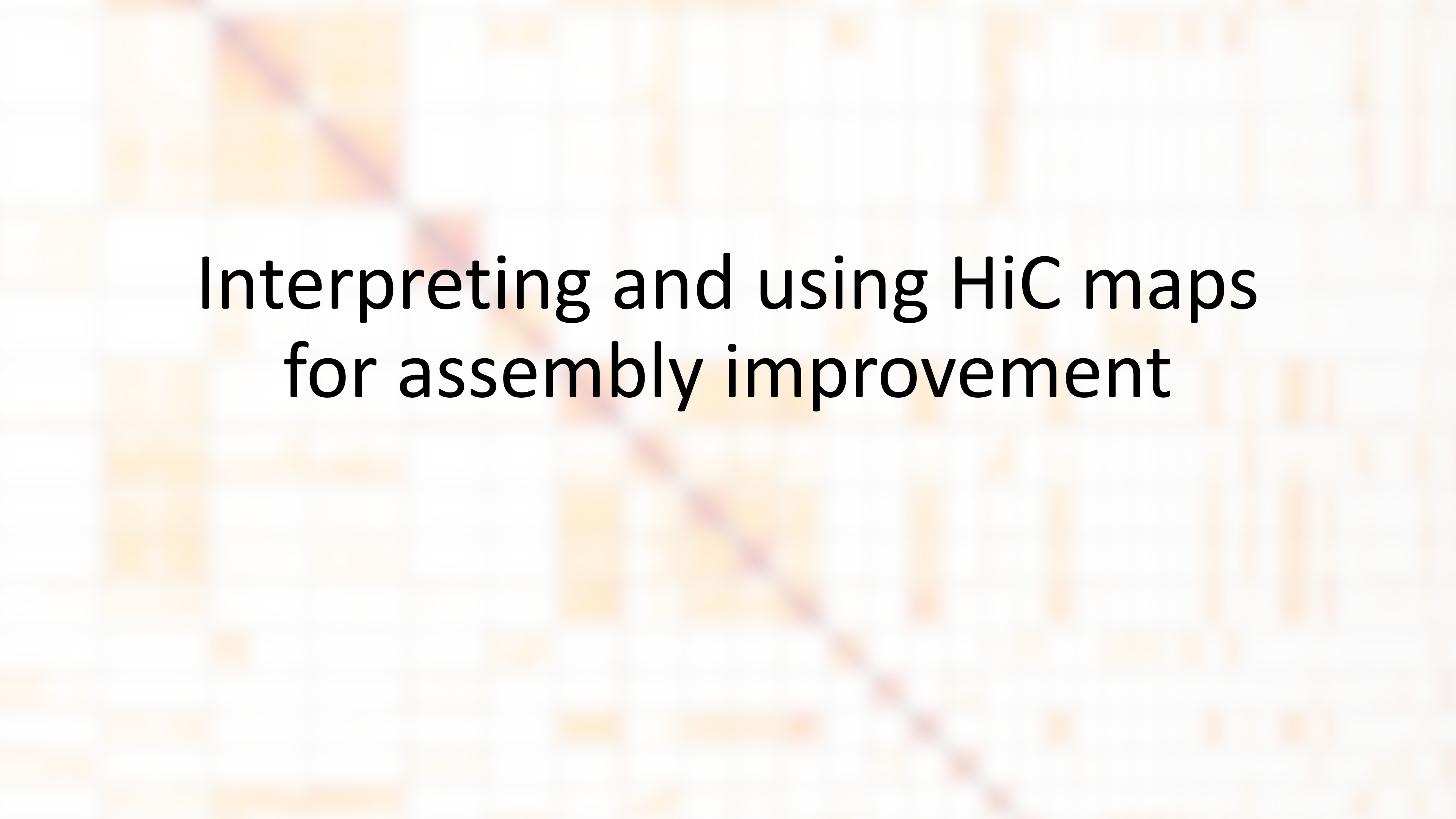


Interpreting and using HiC maps for assembly improvement

Hi-C maps – things to think about

Automated assemblies are imperfect – HiC maps help us to find and fix:

- Mis-joins
- Missed joins
- Retained haplotypic duplication
- Sex chromosomes (heterogametic samples)
- Contamination

Pretext

https://github.com/wtsi-hpag/PretextView

Pretext is an app that allows inspection of HiC reads mapped to an assembly

Pretext can be used to make realtime changes to assemblies:

1. It can be used to break and join scaffolds
2. It can be used to re-orientate scaffolds or parts thereof
3. Additional tracks (bedgraphs) can be injected into pretext maps to view alongside the HiC data, eg: Coverage, Telomeres, Sequence gaps, Repeat density

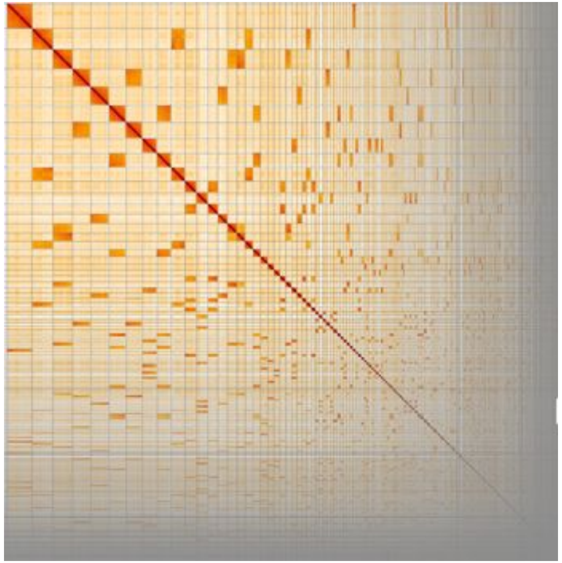
Pretext can also be used to dump an agp (describes the assembly of a larger sequence object from smaller objects) of changes to an assembly

This agp can then be passed to our in-house script rapid_pretext2tpf.py which will then allow the edited map to be turned back into a corrected fasta

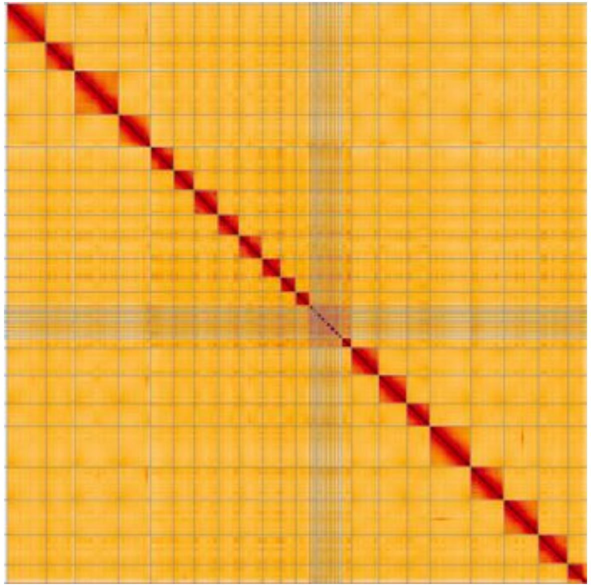
Caveats:

- Pretext is limited by resolution
- Mis-assemblies may be hidden in large genomes.
- Very small scaffolds may not have sufficient resolution to correctly determine placement/orientation.

Interpreting a HiC map



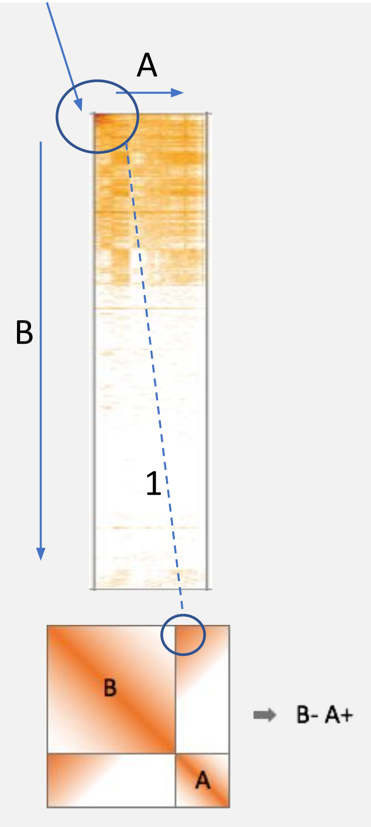
- Squares on centre diagonal show self matches, eg chr1 vs chr1.
- Squares off diagonal show relationship between different chromosomes/scaffolds.
- The darker the off-diagonal square, the stronger the relationship between the scaffolds.
- Horizontal and vertical lines delineate chromosome/scaffold boundaries.
- This top plot shows many off-diagonal relationships as this assembly has not yet been scaffolded.

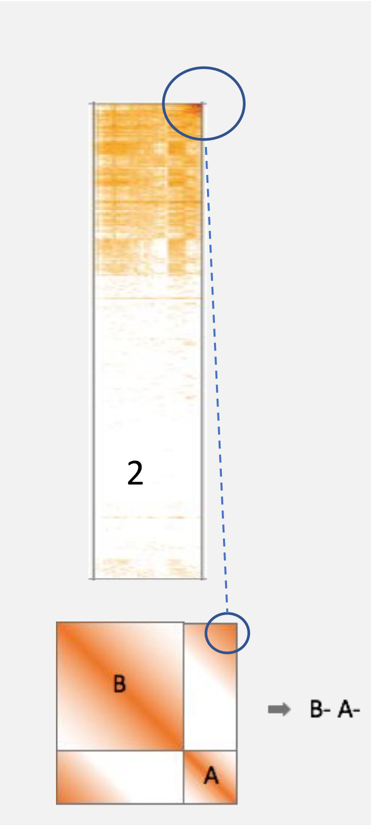


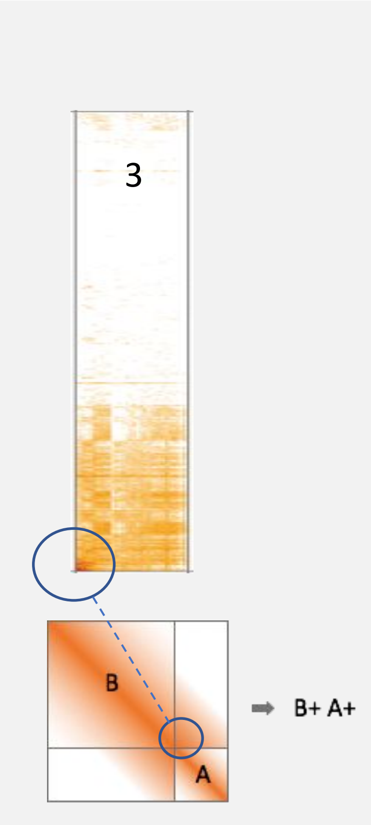
This is what a good genome looks like once all possible joins have been made. In other words, there are no significant off-diagonal associations remaining (apart from small repetitive regions which we are not able to resolve).

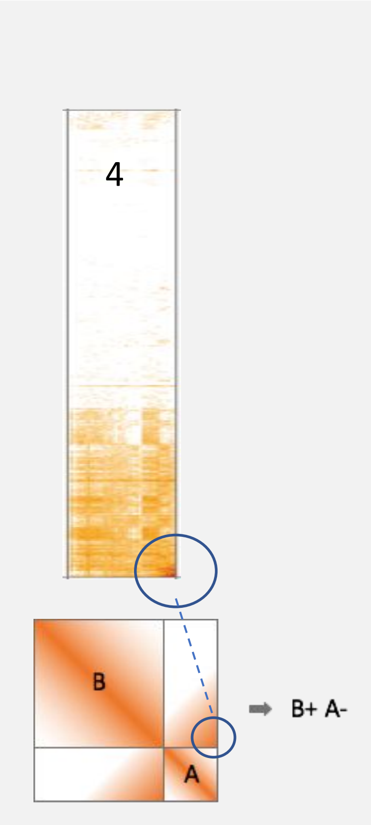
Basic joining

The 4 HiC maps below show scaffold A on the X axis compared with scaffold B on the Y axis. The bright spot shows high affinity indicative of a join as opposed to low affinity which simply shows an association. The order and orientation of the scaffolds can be determined from the location of the bright spot.
Once these joins are actioned correctly, if a new map were to be created, the strong signal would move onto the centre diagonal (as is already the case in scenario 3).

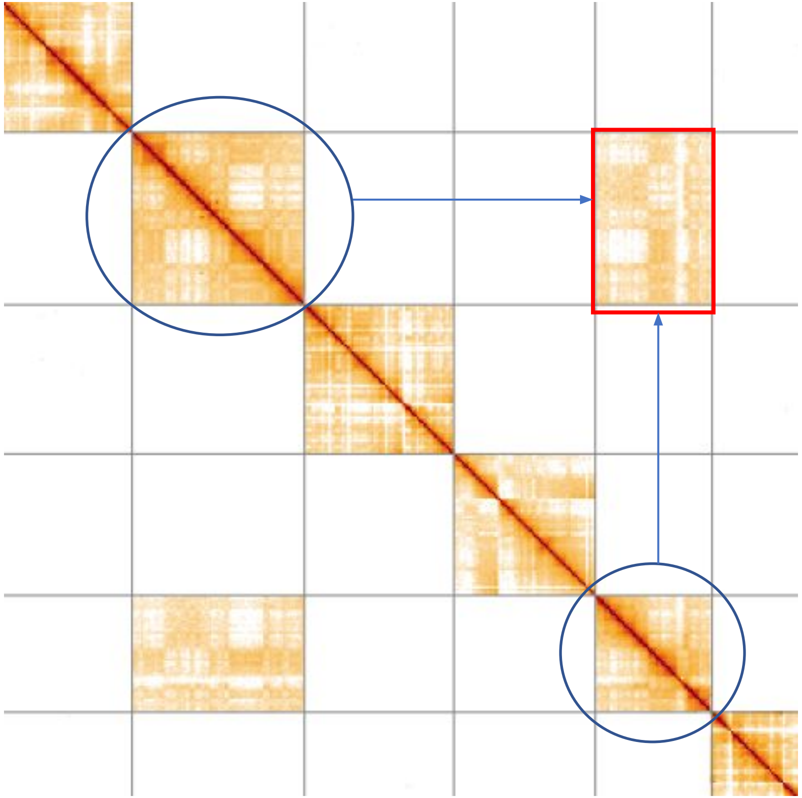








Same chromosome but not immediately adjacent

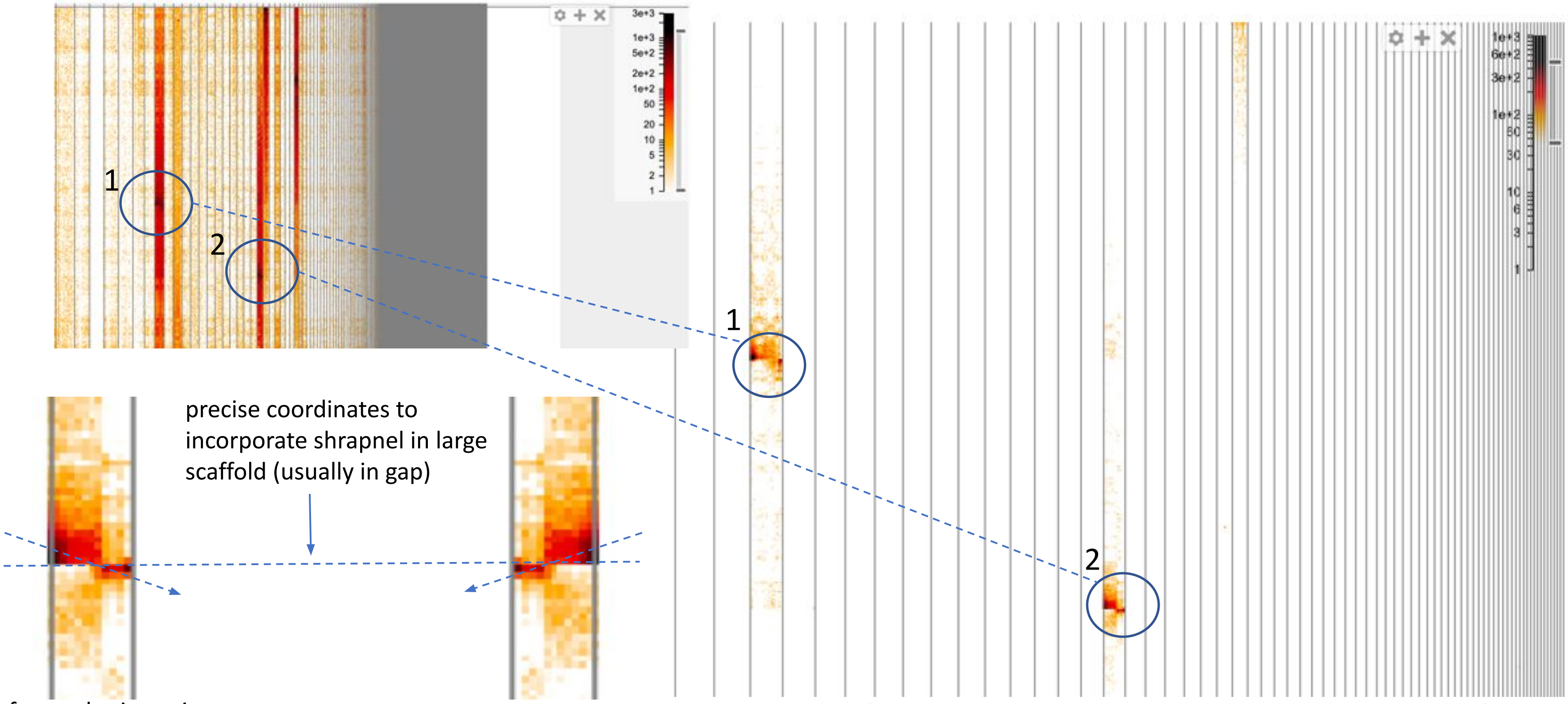


The red outlined square shows a comparison between 2 scaffolds (circled). As there is no strong affinity, but rather a general association (ie no bright spot, just general colouration), we can conclude that the 2 scaffolds

1) belong on the same chromosome
2) are not immediately adjacent to each other.

We don’t have sufficient information here to order or orient them – due to a lack of strong affinity. We might look for other scaffolds belonging to the same chromosome to see if there is stronger affinity between them and the red highlighted scaffold to allow correct scaffolding. Gene order from a close relative may also give clues.

Incorporation of small scaffolds into larger scaffolds



Shrapnel

precise coordinates to incorporate shrapnel in large scaffold (usually in gap)

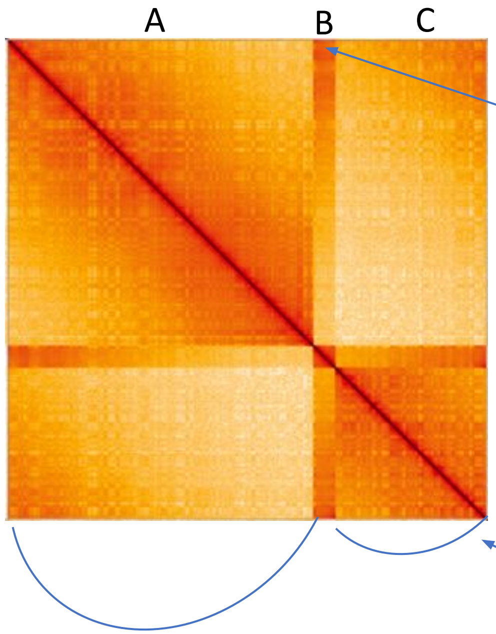
forward orientation

reverse orientation

(Zoom in on shrapnel. Scaffolds delineated by vertical bars)

Rearrangements - scenario 1

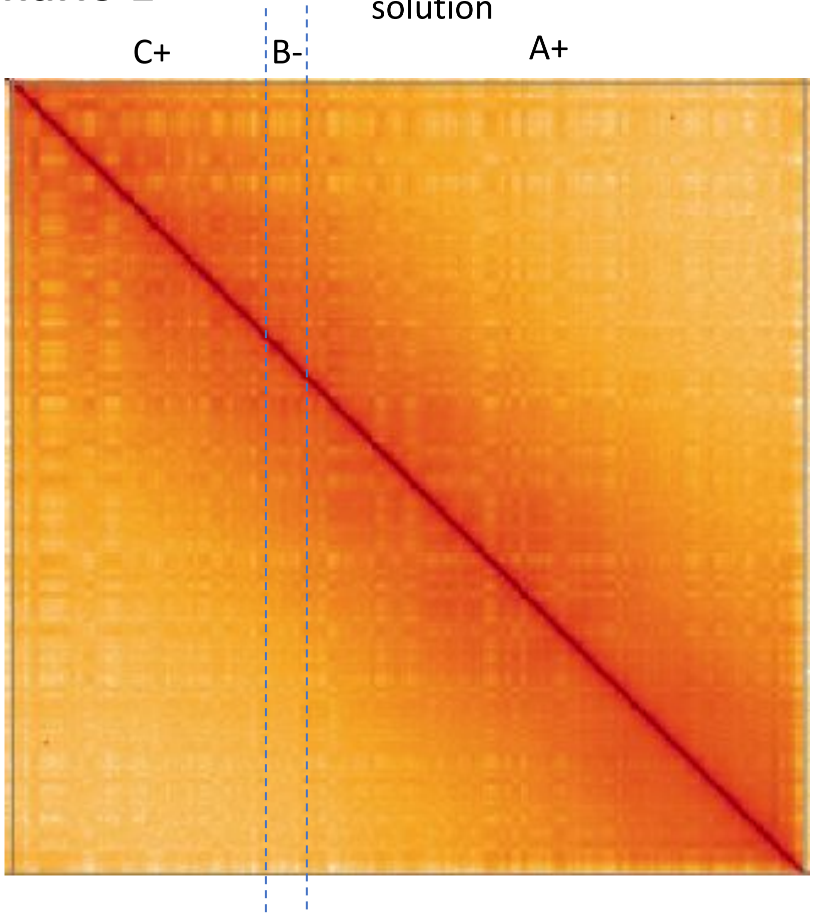
Assembly problem



Strong signal off the centre diagonal is usually indicative of a problem. Zooming in on the centre diagonal at junctions between A/B/C would show breaks in affinity

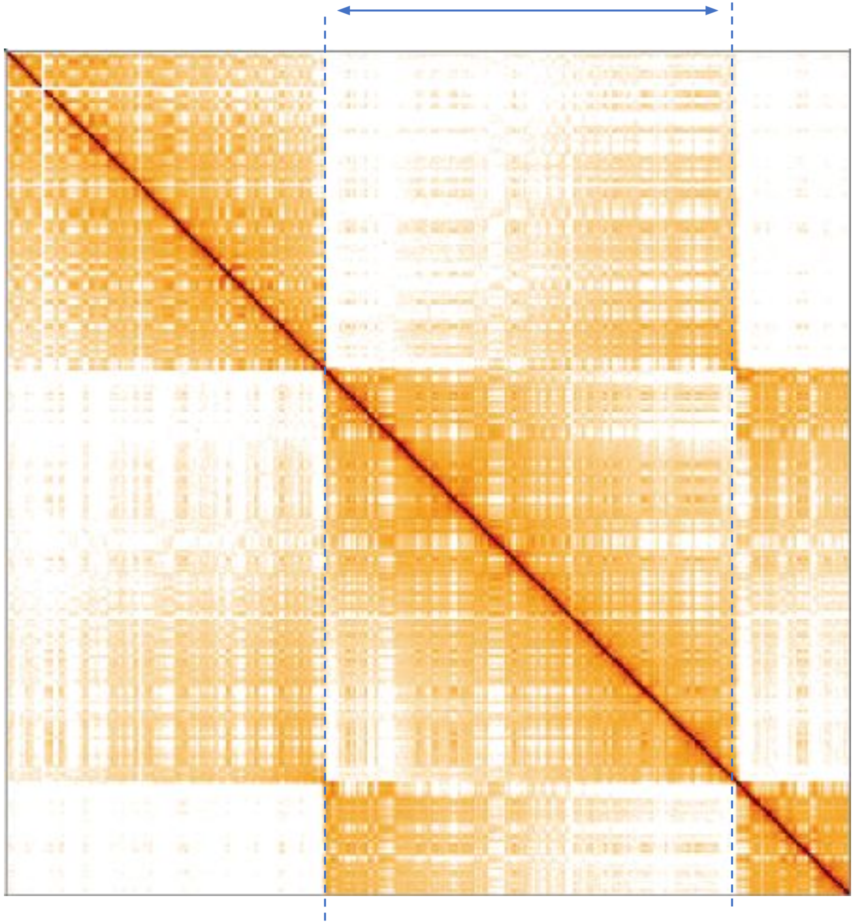
strongest affinity (blue links) between B/r and C/r and B/l and A/l, leading to the solution on the right.

solution



In the solution, the strongest signal is now confined to the centre diagonal.

Rearrangements - scenario 2



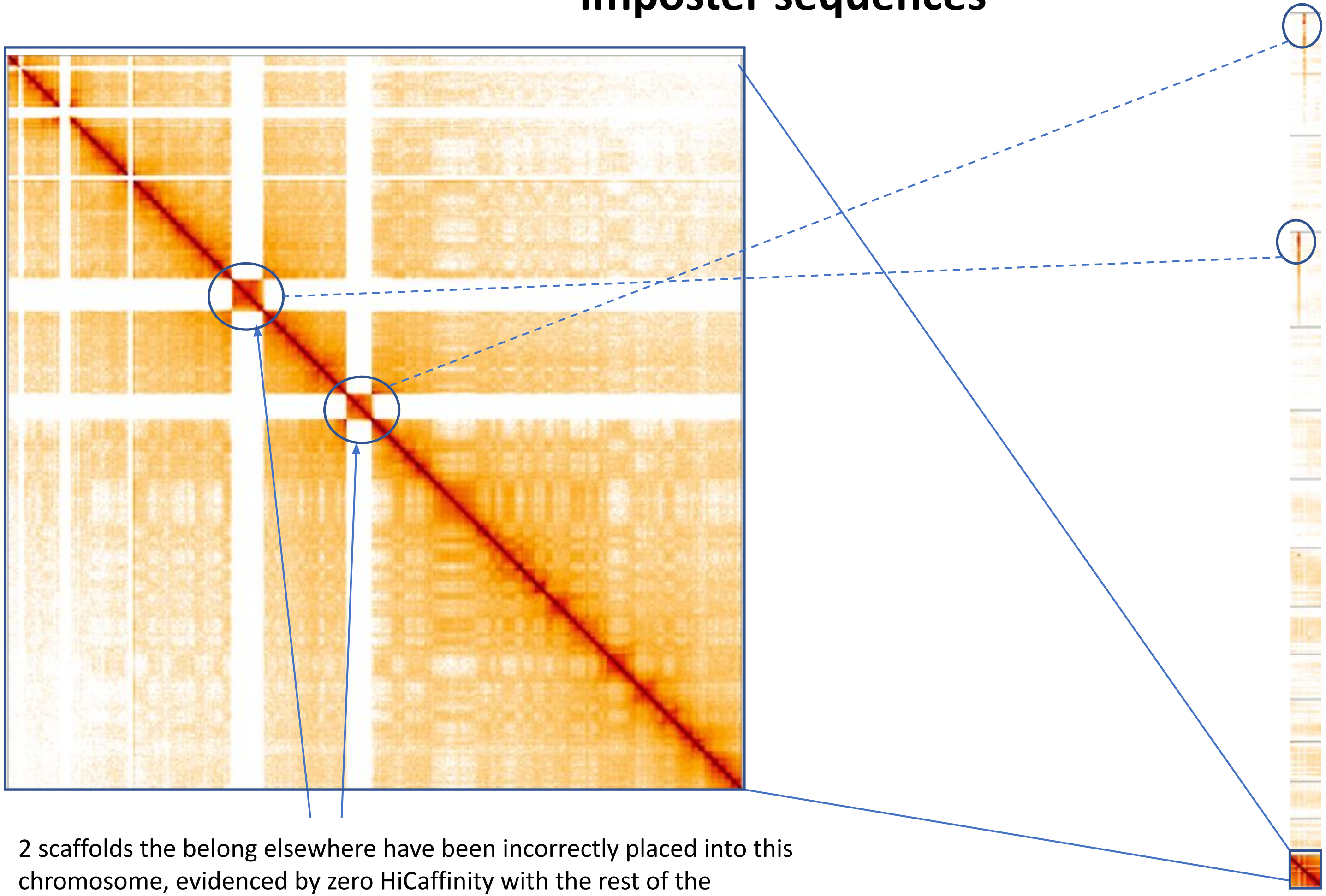
Three distinct sections are visible. The 3 sections are in the correct order, but the 2nd section needs to be inverted.

Sometimes additional evidence can be gleaned from the shrapnel as often these smaller pieces sit in the gaps between the larger pieces.

Often, this additional information is necessary to solve the puzzle and involves a lot of moving around the HiC map.

In this case, all the evidence needed to solve the puzzle can be seen within the self comparison of this one scaffold.

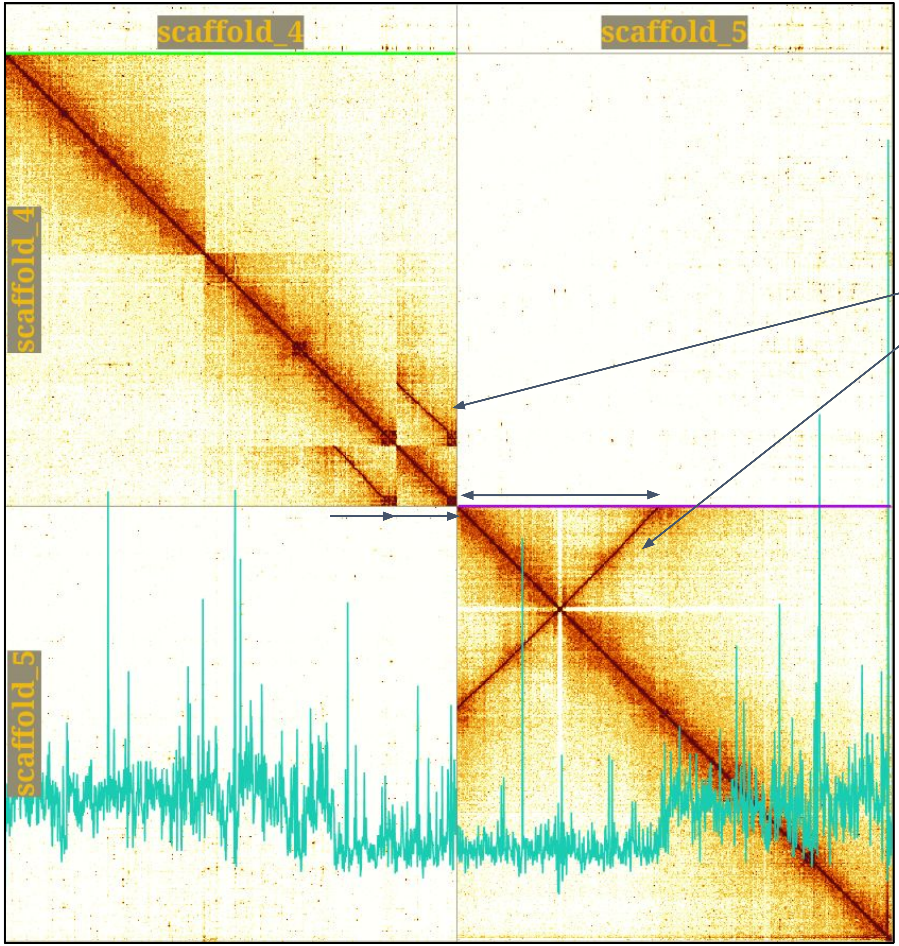
Imposter sequences



Inspection of the off diagonal map shows that these pieces really belong in scaffolds 1 and 3.

2 scaffolds the belong elsewhere have been incorrectly placed into this chromosome, evidenced by zero HiCaffinity with the rest of the chromosome.

Haplotypic Duplication



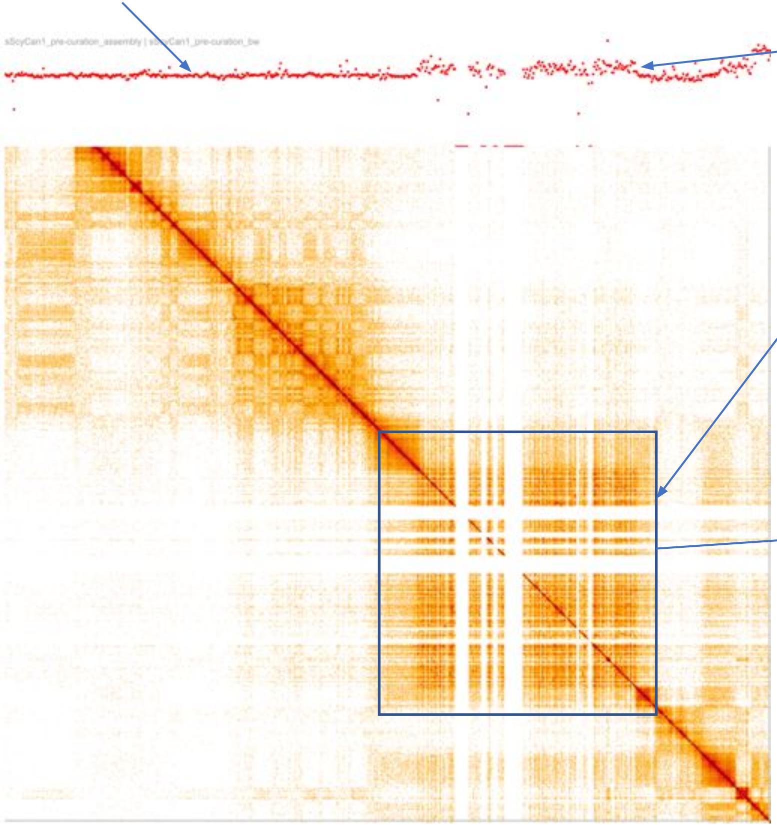
Duplicated sequences - these have half diploid coverage depth and the inflexion point coincides with a gap – the signatures we normally see with haplotypic duplication

One haplotype from each pair should be removed to create a more accurate representation of the chromosome

Coverage track

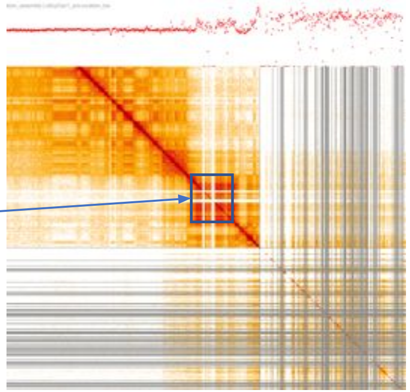
Region of low complexity/tandem repeat

Normal coverage outside the repeat



Elevated coverage due to collapsed assembly in tandem repeat

Distinctive HiC pattern due to this tandem repeat – high affinity between all parts of the repeat (due to many reads and their potential to map readpairs all over the repeat), and reduced affinity with the rest of the chromosome

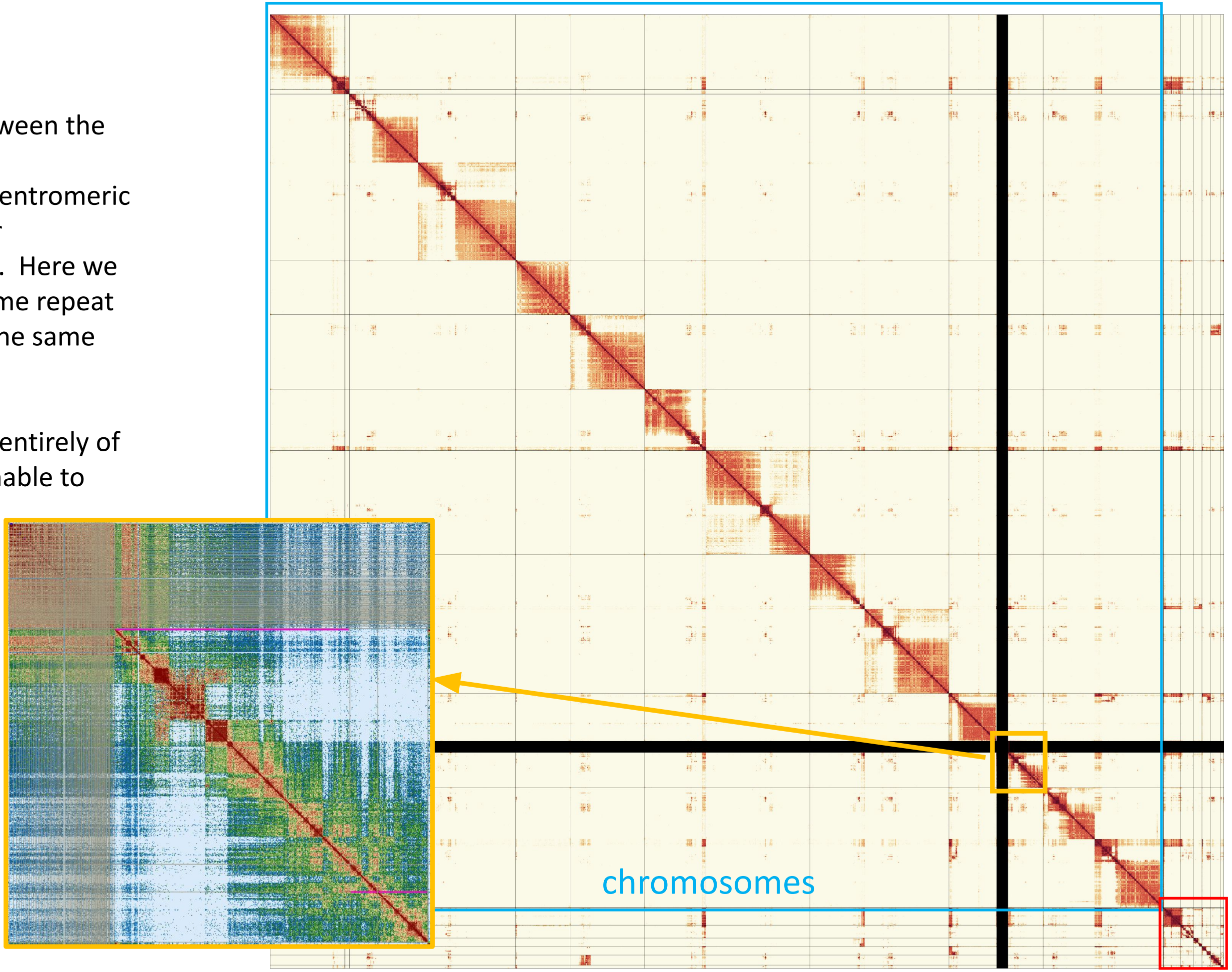


Zooming out a little, many shrapnel contigs have affinity with this repeat and also have high coverage. This looks like a large, highly repetitive region of the genome that is resistant to assembly

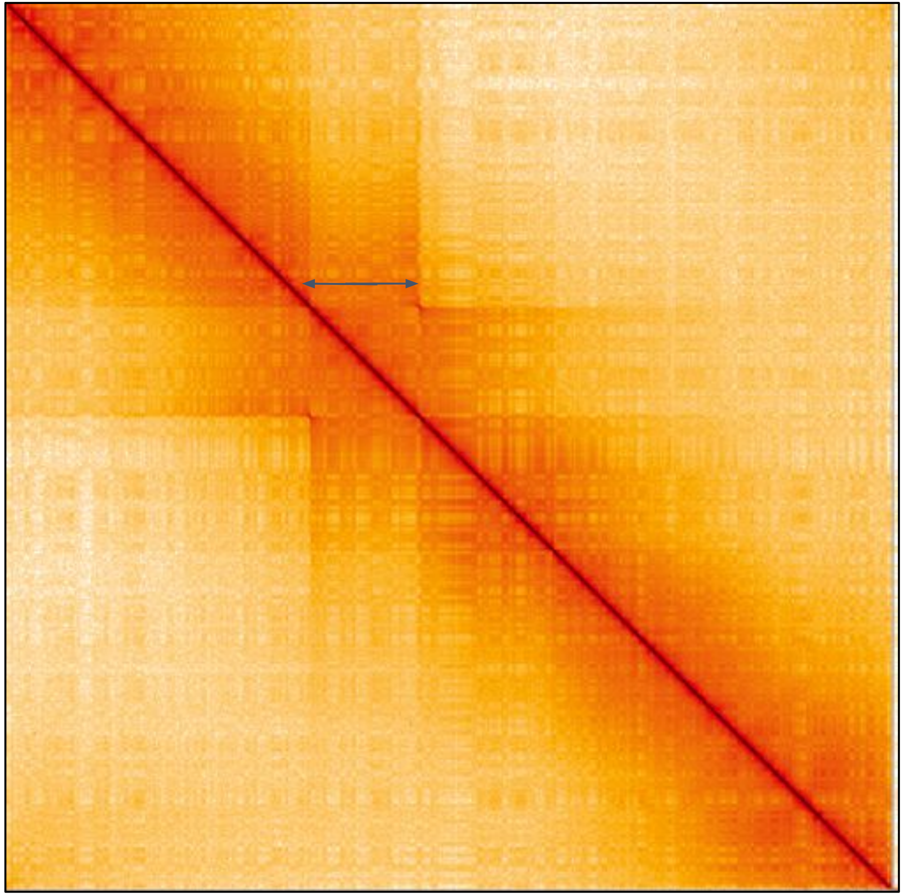
Satellite sequence has weak association with particular chromosome

In this case, there is some association between the chromosome arms of each chromosome. Furthermore, the satellite repeats in the centromeric regions are typically unique to a particular chromosome, enabling them to be placed. Here we highlight 91 scaffolds composed of the same repeat type that we can see from HiC belong to the same chromosome.

There remain several scaffolds composed entirely of satellite sequence which we have been unable to place.



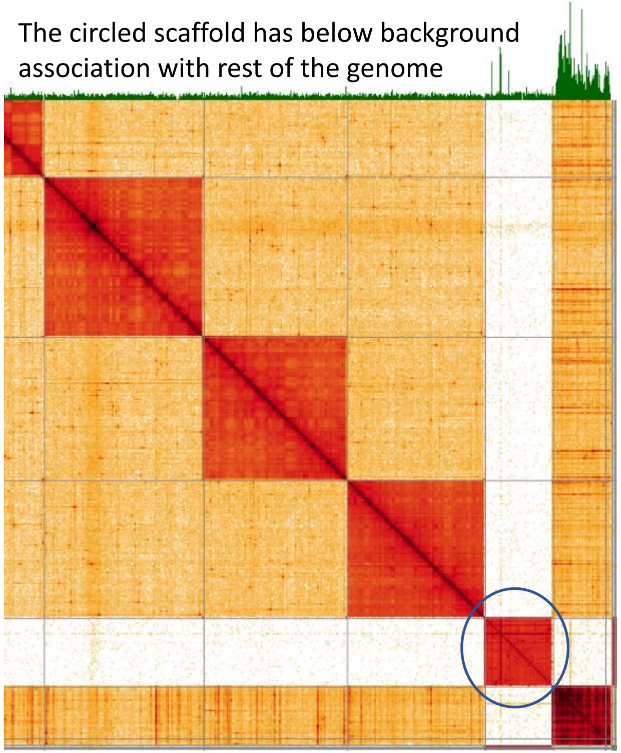
HiC inversion between sister chromatids 1



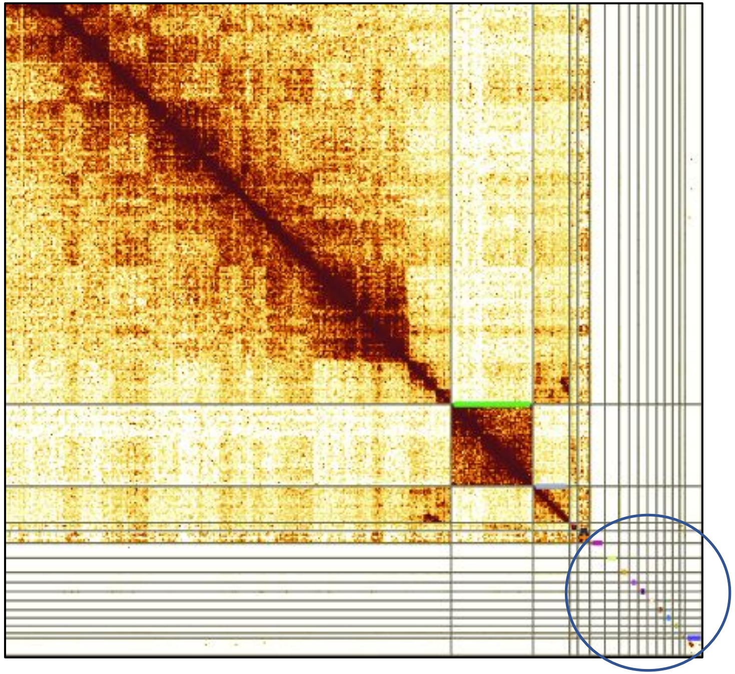
A portion of the sister chromatid is inverted therefore HiC always looks correct (from the centre diagonal) and incorrect (off-diagonal signal) whichever orientation.

Contamination - How to identify it

The circled scaffold has below background association with rest of the genome



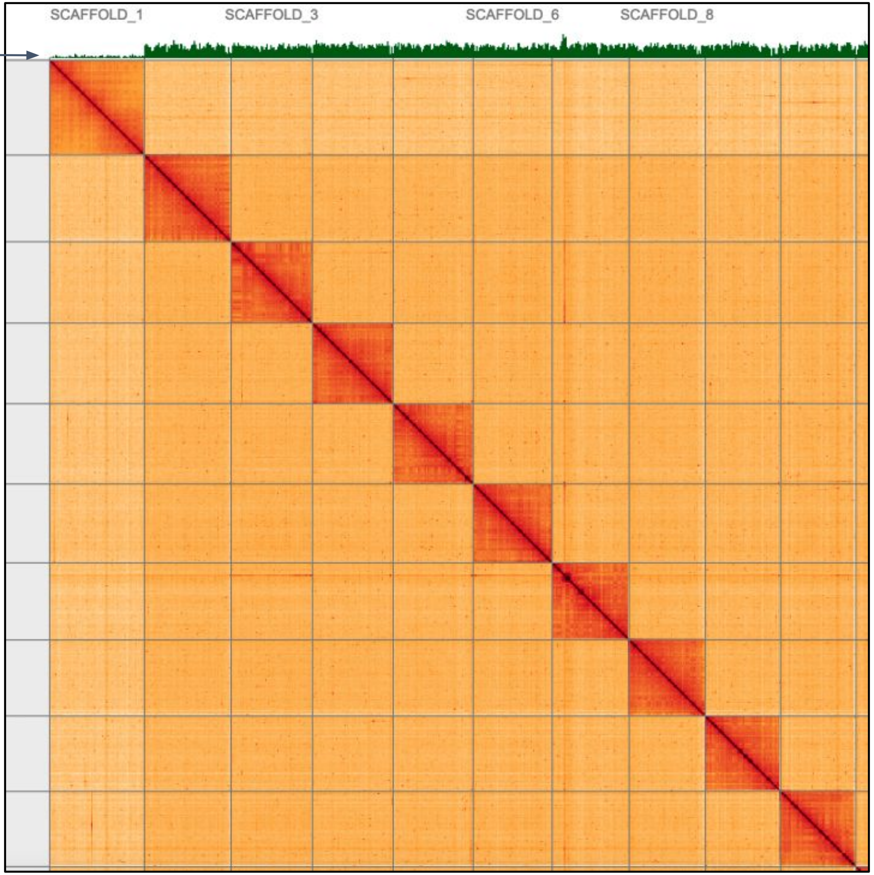
The circled scaffolds have no HiC signal indicating that they are possible contaminants



Despite assemblies being processed through our decontamination pipeline in some scenarios, where the decon results are inconclusive, a contaminant scaffold/s may accidentally be left in the assembly. If contamination is suspected due to the appearance of the scaffolds in the HIC map e.g. as in either scenario above the scaffolds of interest can be interrogated using Blob Toolkit (BTK), blast etc. If they are confirmed as contamination they will be removed during the curation process.

Sex Chromosomes

Half coverage

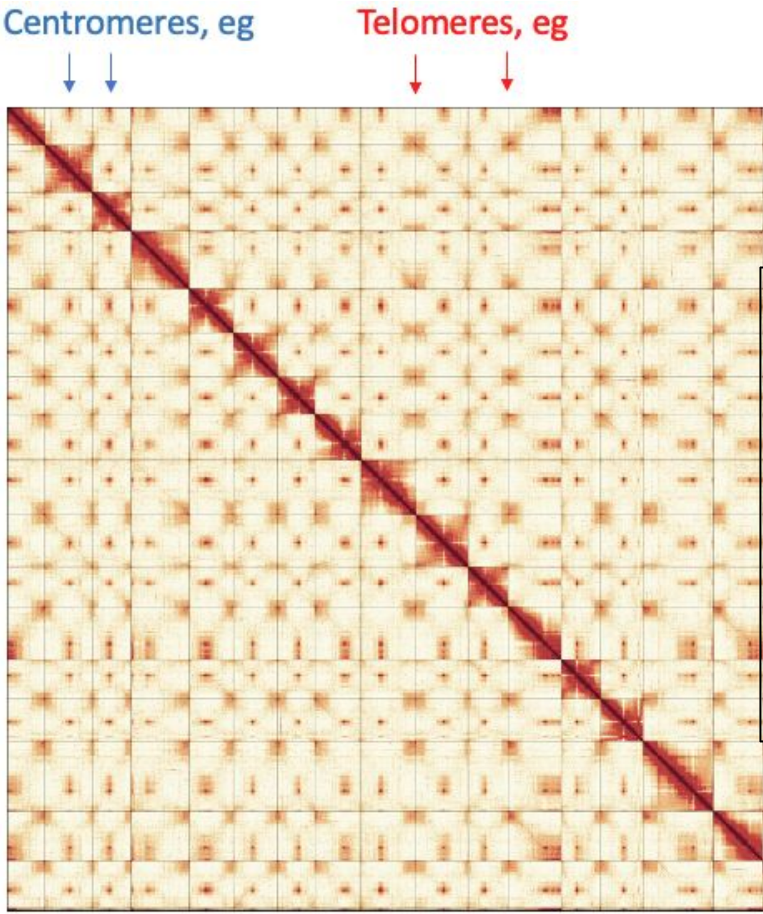


Seq coverage

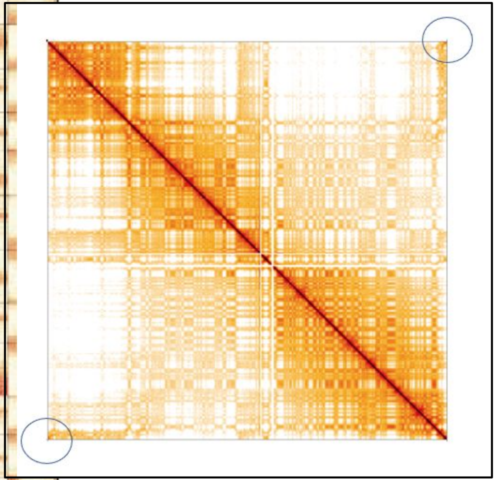
In heterogametic samples sex chromosomes can be identified in HiC maps as they have half depth coverage

For homogametic samples sex chromosomes have diploid coverage so have to be identified by alternate methods e.g. synteny to closely related species where sex chromosome turnover is known to be low

Centromeres and Telomeres



Centromeres and telomeres often light each other up as can be seen in this assembly.



This association provides good evidence for determining if chromosomes are assembled correctly