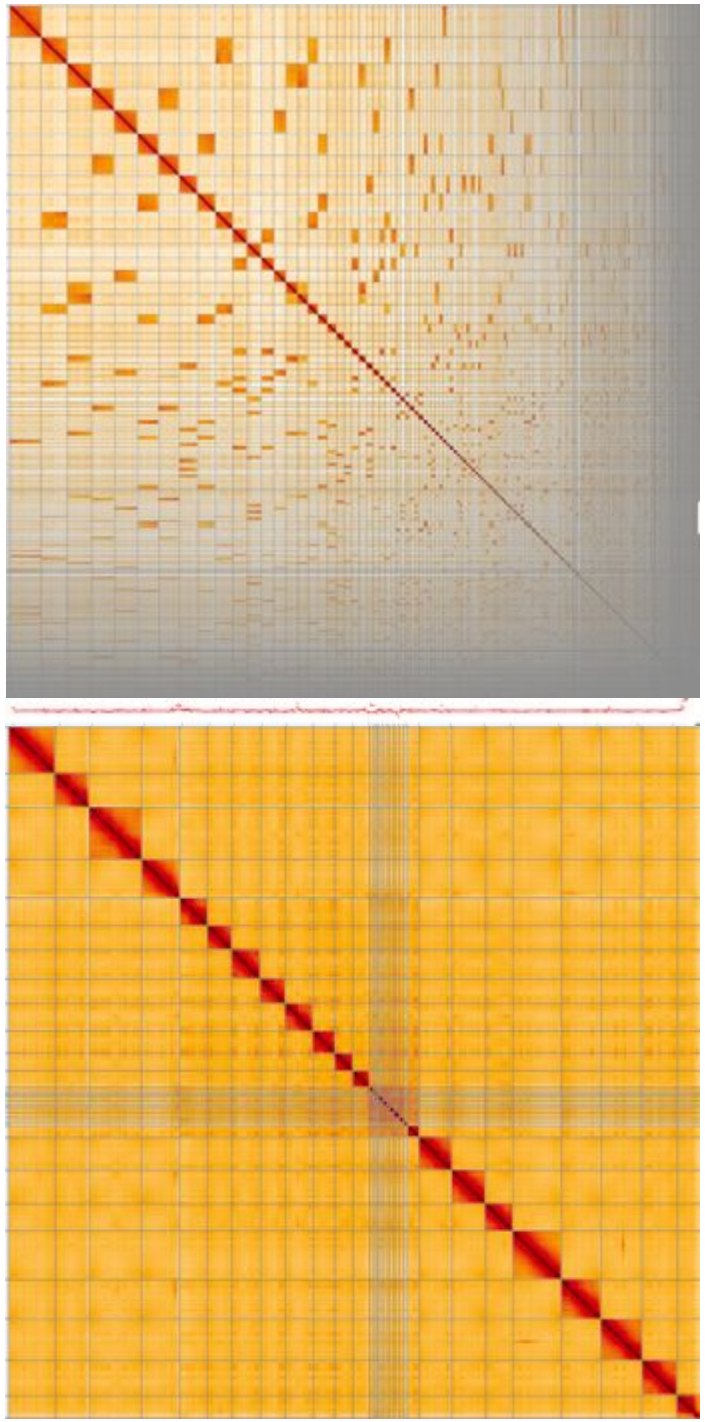
Interpreting and using HiC maps for assembly improvement

Intro

Includes:

- General map interpretation for QC and scaffold joining
- Incorporation of shrapnel
- Background signal/checquerboarding
- Interpreting more challenging cases - considering other evidence
- Choosing colour schemes
- Pretext vs HiGlass



Interpreting a HiC map

Squares on centre diagonal show self matches, eg chr1 vs chr1.

Squares off diagonal show relationship between different chromosomes/scaffolds (eg chr1 vs scaffold52). The darker the off-diagonal square, the stronger the relationship between the scaffolds.

Horizontal and vertical lines delineate chromosome/scaffold boundaries.

We can also decorate the HiC image with a bigwig coverage file (red histogram, on top of bottom plot).

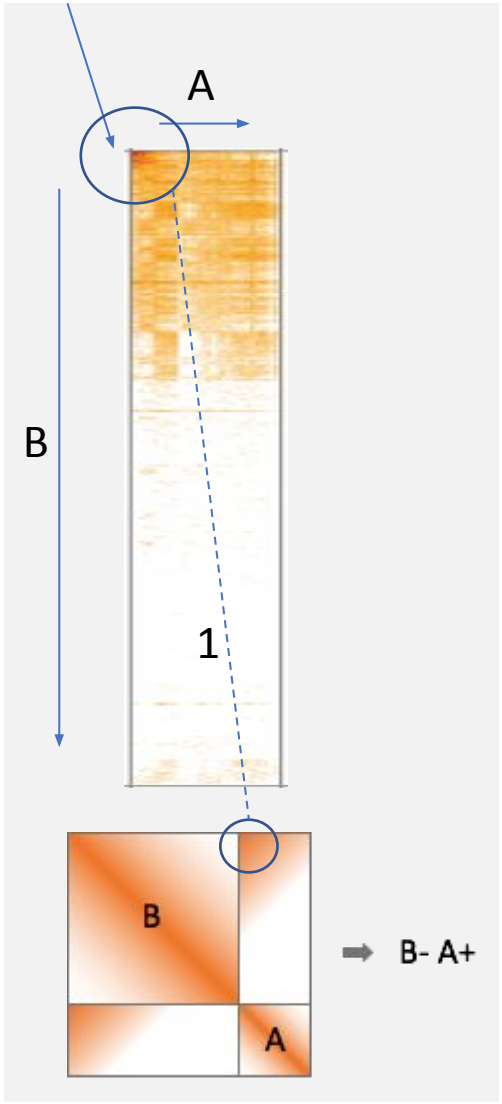
This top plot shows many off-diagonal relationships as this assembly has not yet been scaffolded.

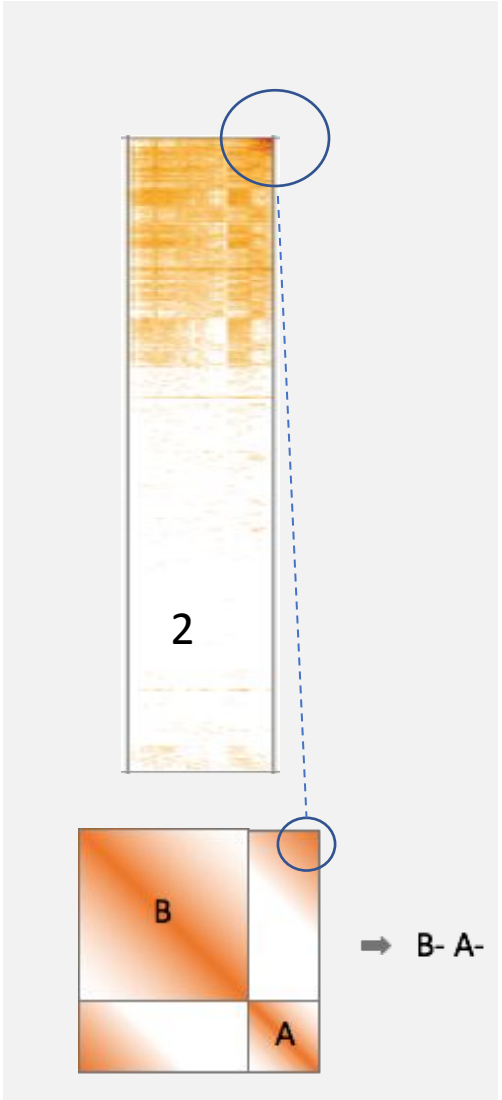
This is what a good genome looks like once all possible joins have been made. In other words, there are no significant off-diagonal associations remaining (apart from small repetitive regions which we are not able to resolve).

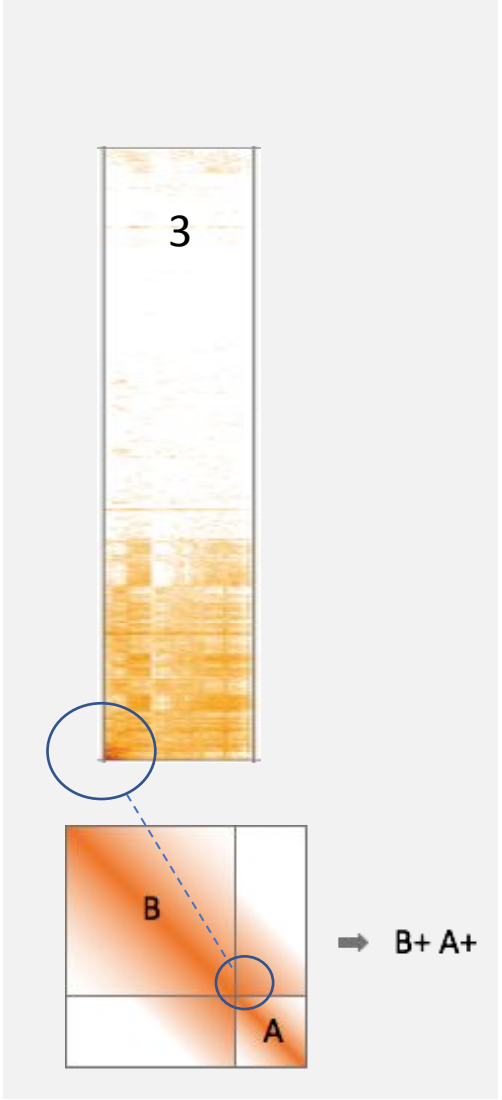
When curating assemblies, we don't rely only on HiC information, but also pay a lot of attention to other evidence, eg optical maps, and synteny.

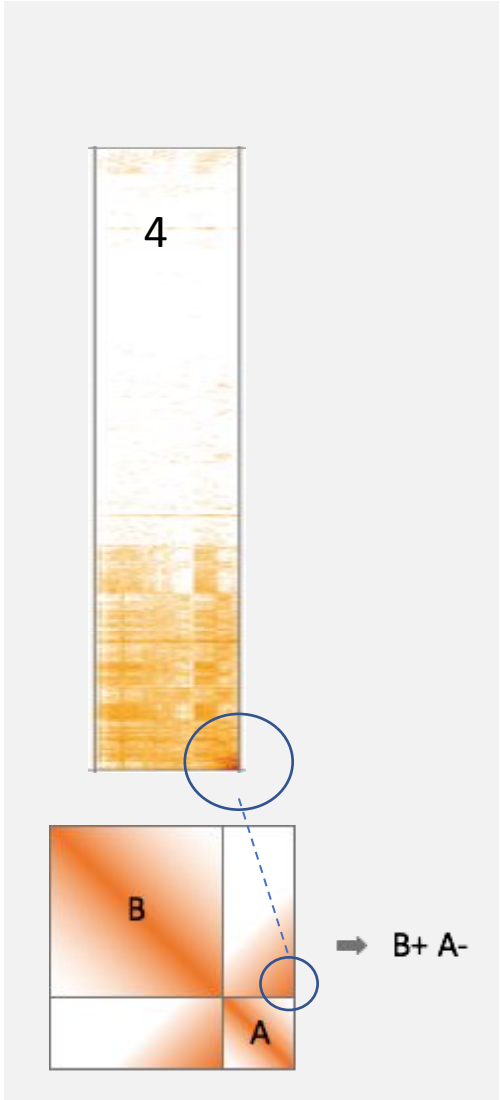
Basic joining

4 HiC maps below show scaffoldA on the X axis compared with scaffoldB 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).

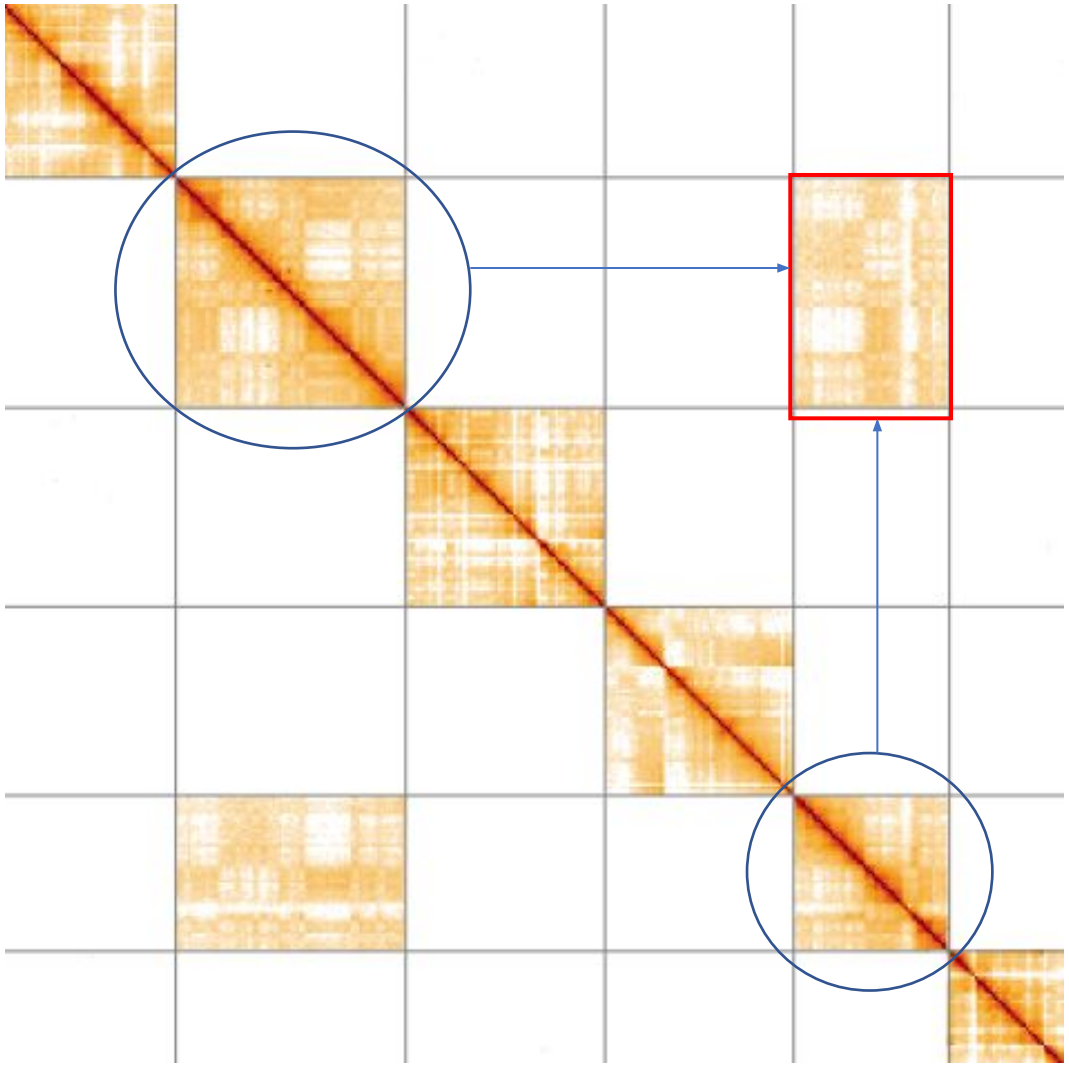








Same chromosome but not immediately adjacent



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

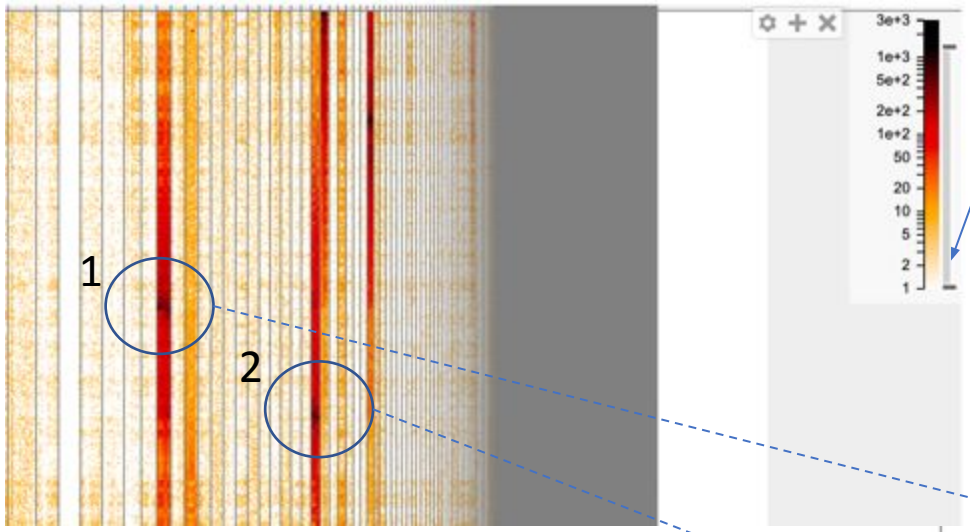
1) belong on the same chromosome
2) are not immediately adjacent to each other.

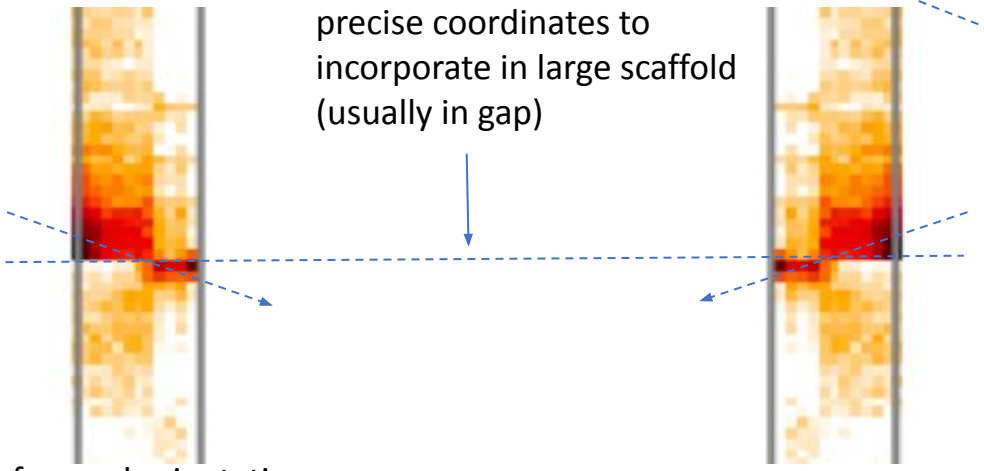
We don't have sufficient information here to order or orient them – due to a lack of strong affinity. We might look for other scaffolds belonging to the same chromosome to see if there is stronger affinity between them and the red highlighted scaffold to allow correct scaffolding. Gene order from a close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

Incorporation of small scaffolds into larger scaffolds

Tweaking colour-bar is crucial to deduce 1) orientation 2) precise coordinates for incorporation

Shrapnel

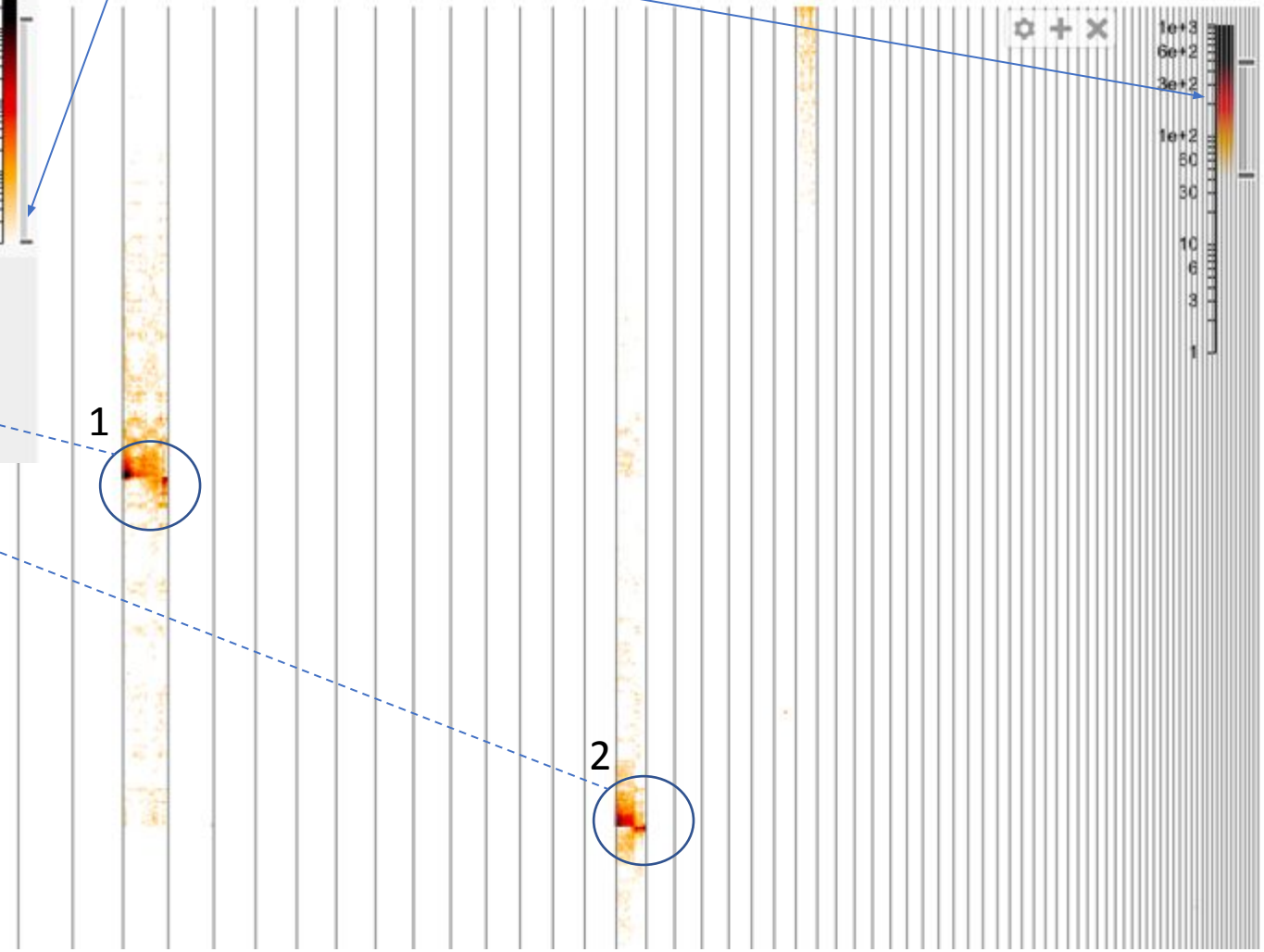




precise coordinates to incorporate in large scaffold (usually in gap)

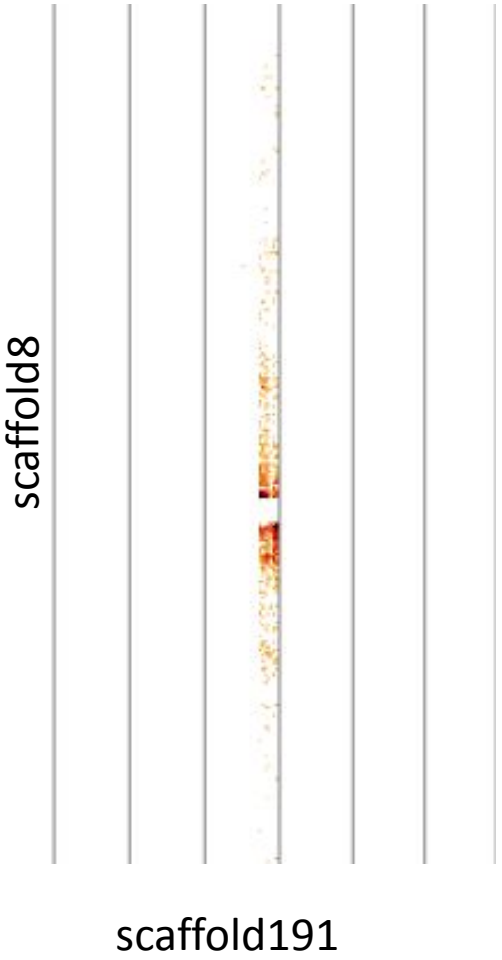
forward orientation

reverse orientation

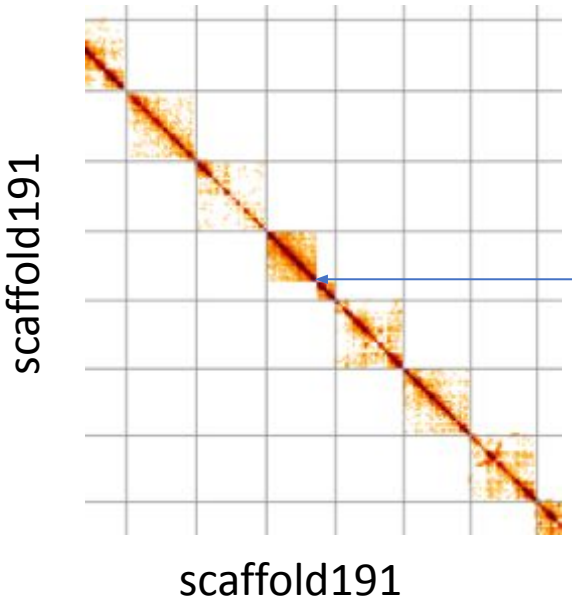


(Zoom in on shrapnel. Scaffolds delineated by vertical bars)

Shrapnel contig needs incorporating, but changes are needed first



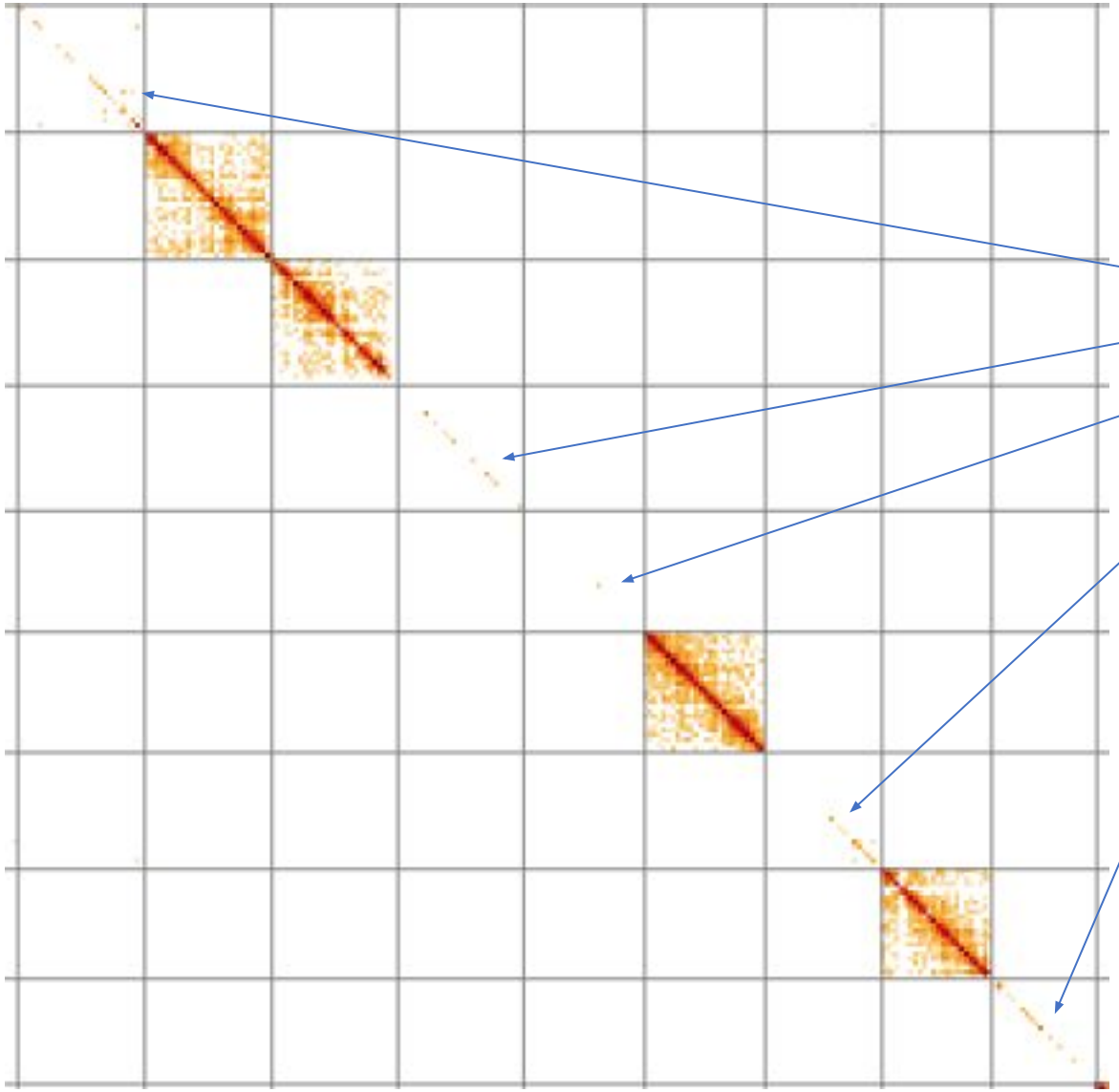
It's easy to spot that scaffold191 needs incorporating into scaffold8, but it's also important to notice that the match doesn't span the whole of scaffold191. This suggests that a break in scaffold191 is needed before the right-hand portion of scaffold191 is incorporated into scaffold8.



Looking at scaffold191 against itself, we can see that there is no affinity between the first 2/3s and the last 1/3 – further confirmation that scaffold191 needs breaking before incorporating it into scaffold8.

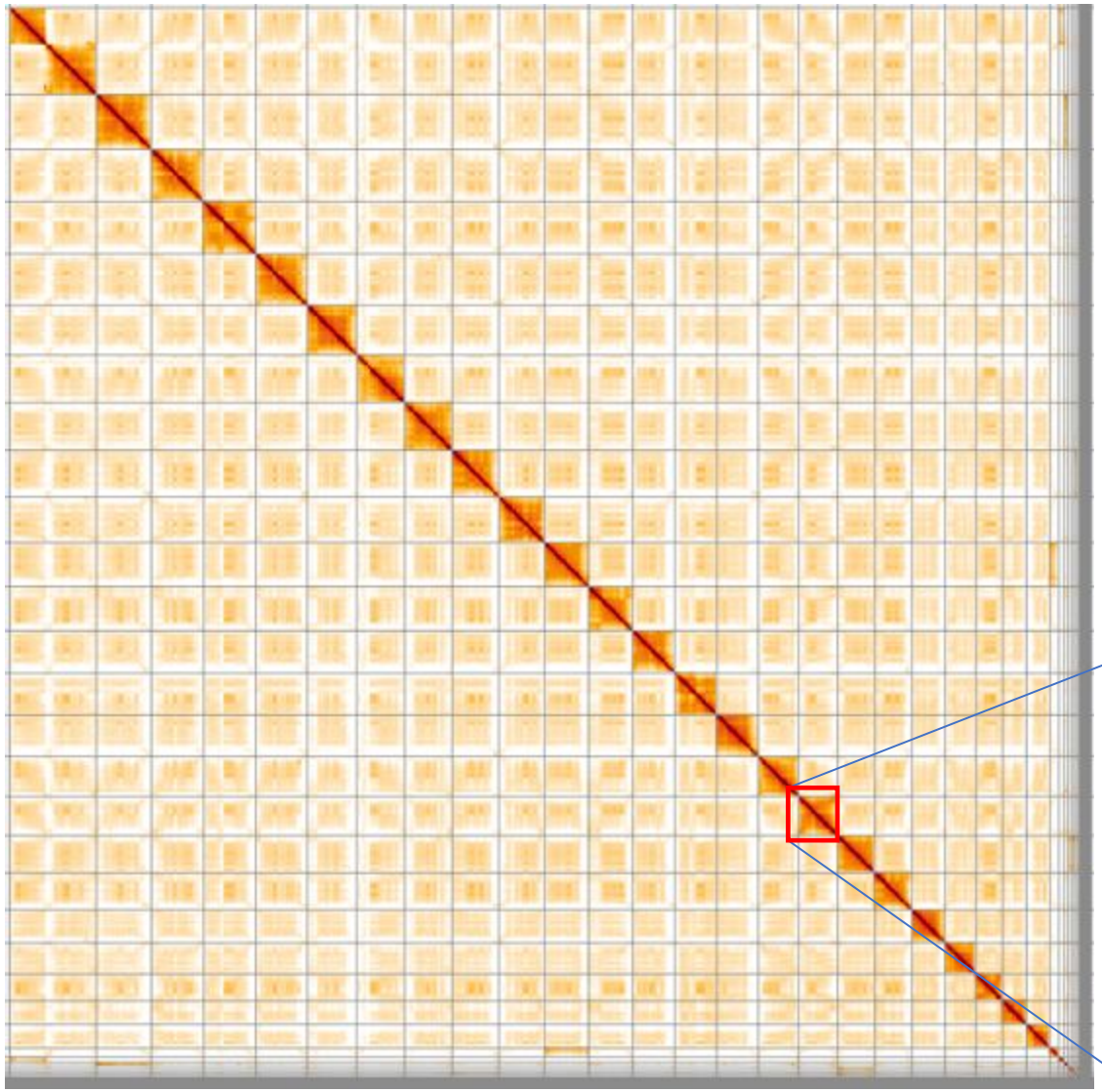
fNotCel1_2 scaffold8 vs scaffold191

Shrapnel



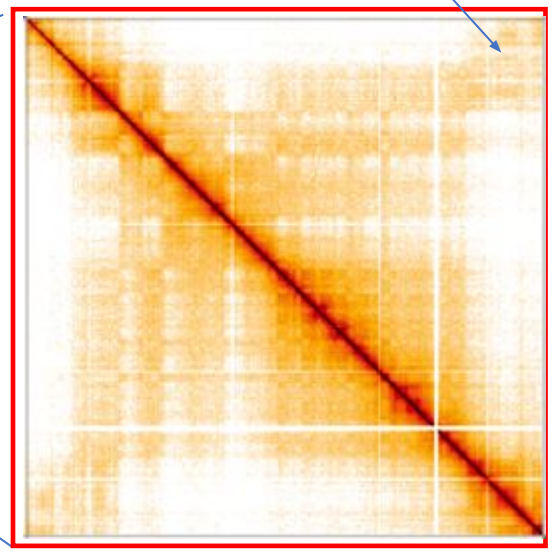
Potential haplotypes or repetitive sequence with weak read mapping. Read mapping is weak in these cases due to multi-mapping reads having been filtered – an indication of repetitive sequence. Looking at other evidence such as coverage will help to determine what is happening in each case. If these scaffolds are deemed to be haplotypes (typically have a high percentage match to another larger scaffold and 50% read coverage), they can be removed from the primary assembly.

Background signal



Background signal tends to be different for each assembly and curators need to factor this in when interpreting HiC maps and looking for genuine signal. In this case, there tends to be a weak association between the subtelomeric regions and they seem to behave differently to the middle region of each chromosome.

Background affinity between subtelomeres is higher than one might expect and could be misleading.



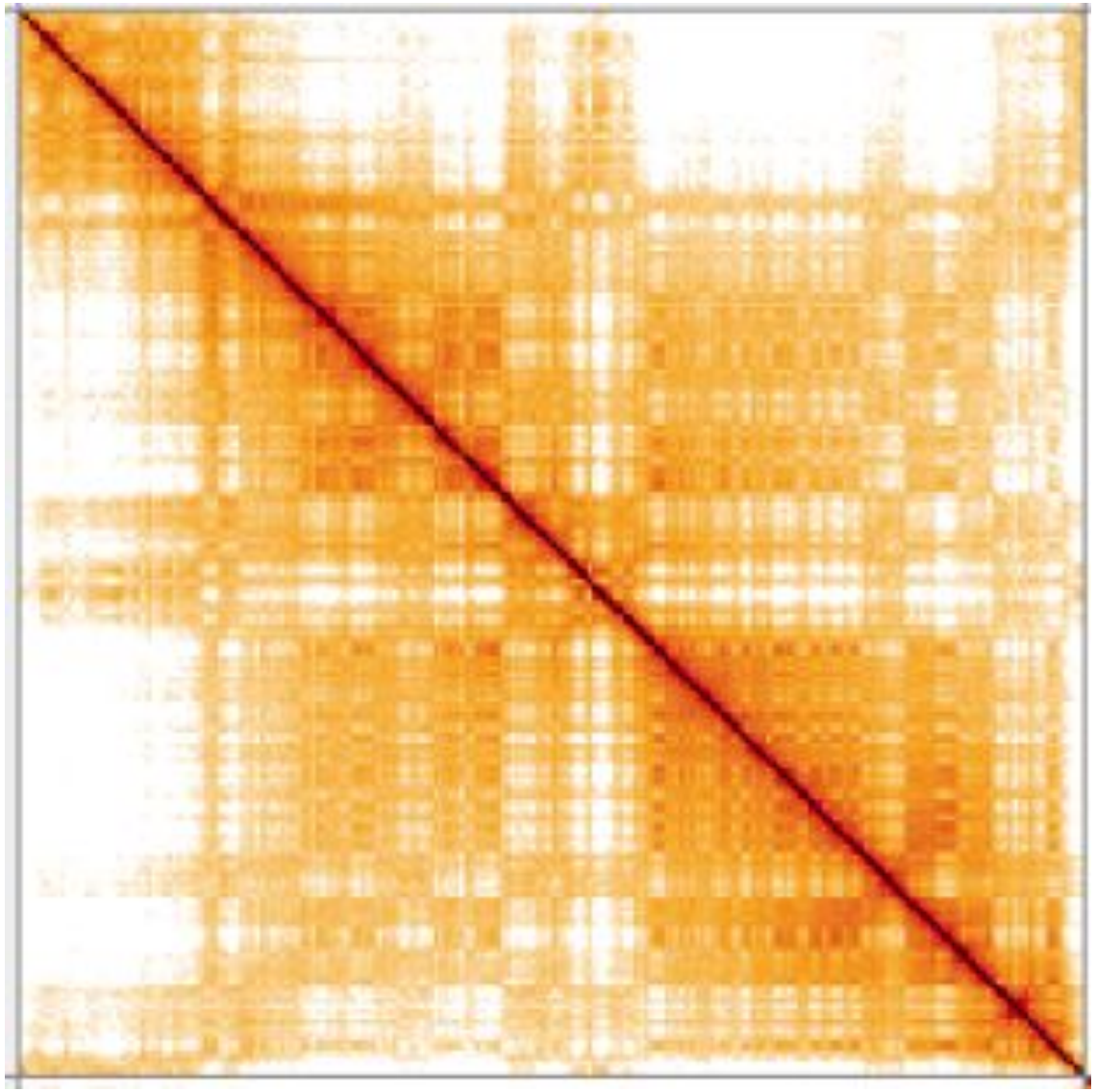
fCycLum1 superscaffold17

For example, here is a smaller scaffold that wants to join to the start of the chromosome, but in this case could be misinterpreted as wanting to join to the r/h end.

Real join

Misleading signal

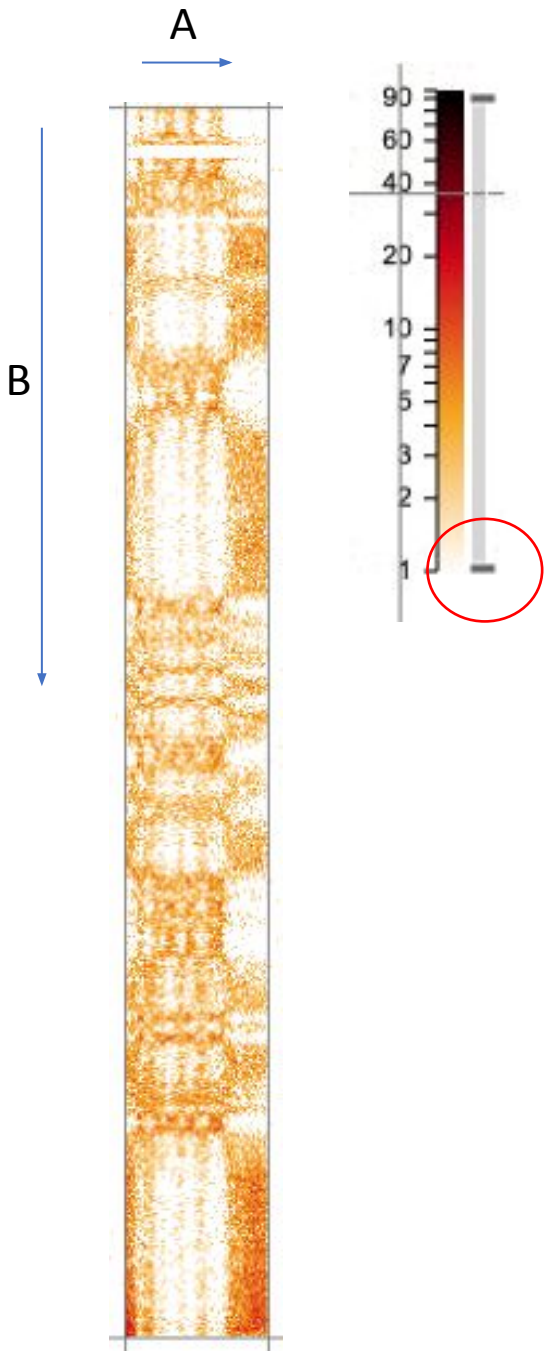
Background signal



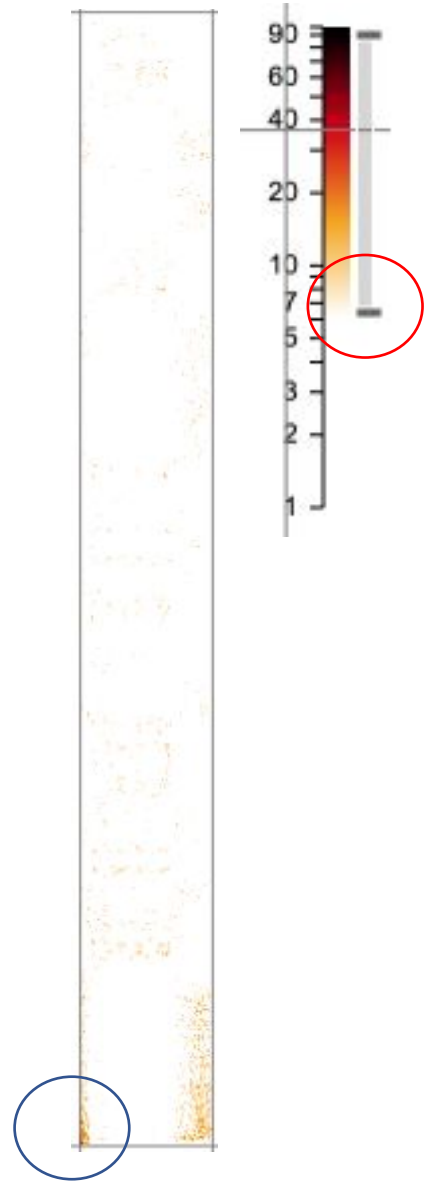
mSciVul1 – scaffold18

Unfortunately HiC contact maps do not always produce a smooth gradient of linkage across the chromosome.
Background signal/cross-hatching/cheqerboarding can be misleading and can make it difficult to determine if the assembly is correct. A close inspection of the diagonal line (zoomed in) combined with inspection of other data (eg BioNano) is important in order not to miss assembly problems.

Chequerboarding

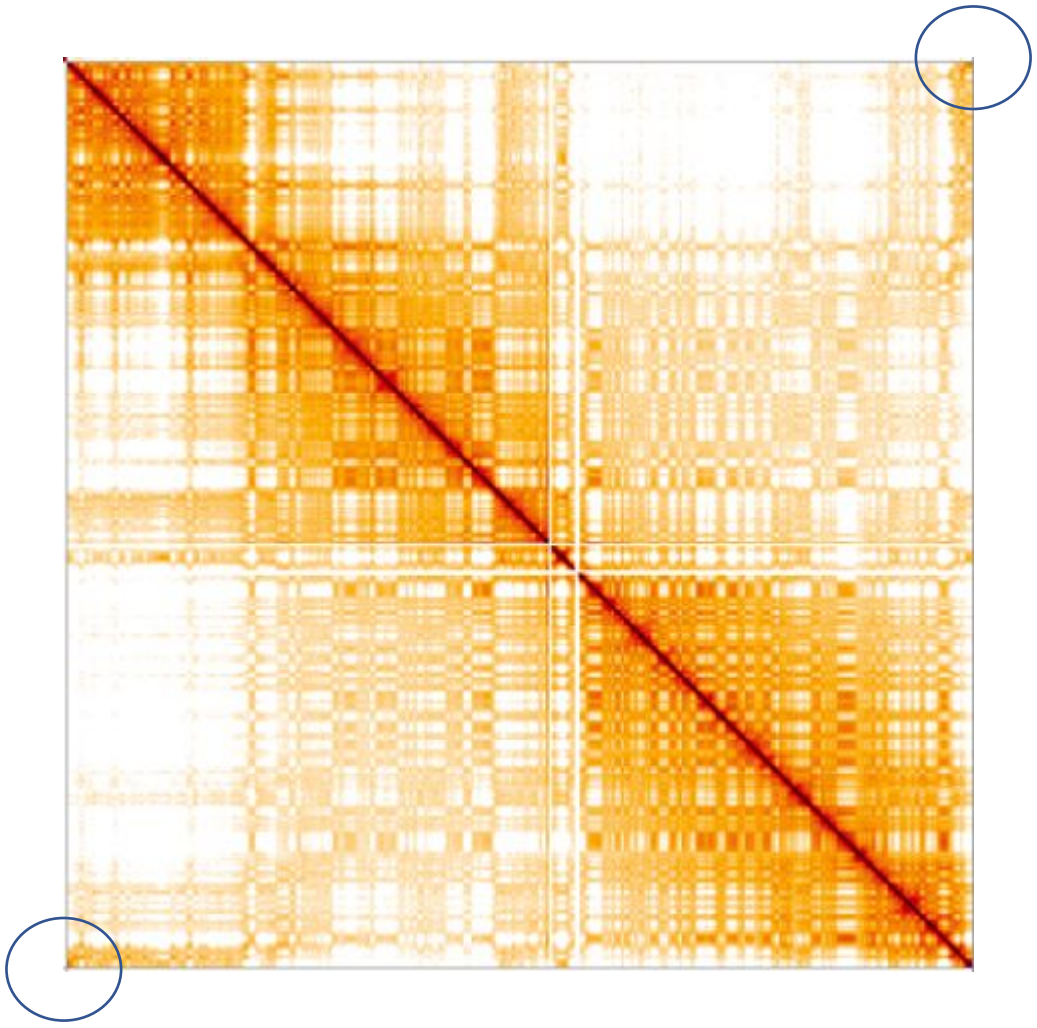


A superficial glance at the relationship between scaffolds A and B might suggest that A/r joins to B/r (A+ B-). However, signal is being lost from the left bottom of the image due to background 'chequerboarding' – ie patches of the contact map where signal seems to be lost. Adjusting the colour bar makes it quite obvious that strongest affinity actually occurs at the bottom left corner, so the solution (supported by Bionano data) is really B+ A-



bPteGut1 – join 77 to 19

Linking between chromosome ends

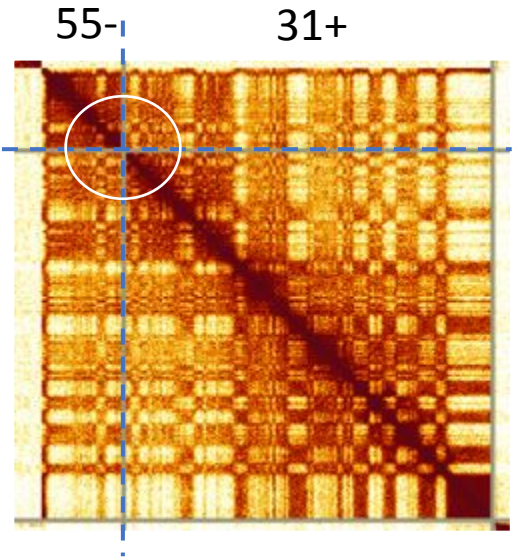


mZalCal1 – scaffold4

We often see affinity (ie off-diagonal signal at a level higher than we'd expect) between chromosome ends on the same chromosome. All evidence suggests that when we see this the chromosome is assembled correctly.

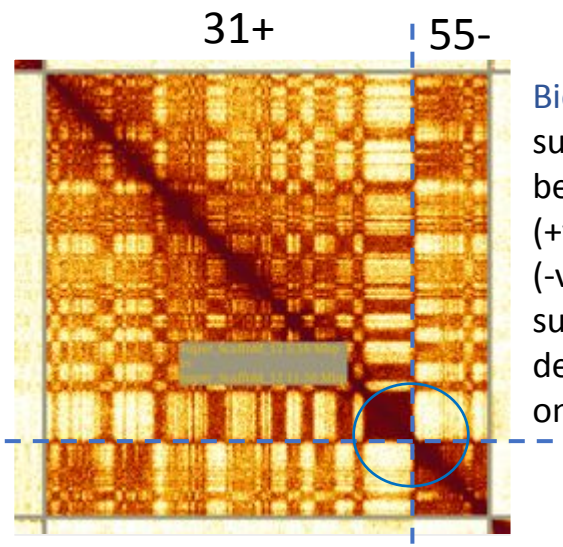
HiC (Pretext) can appear to conflict with Bionano

Apparent best fit with HiC

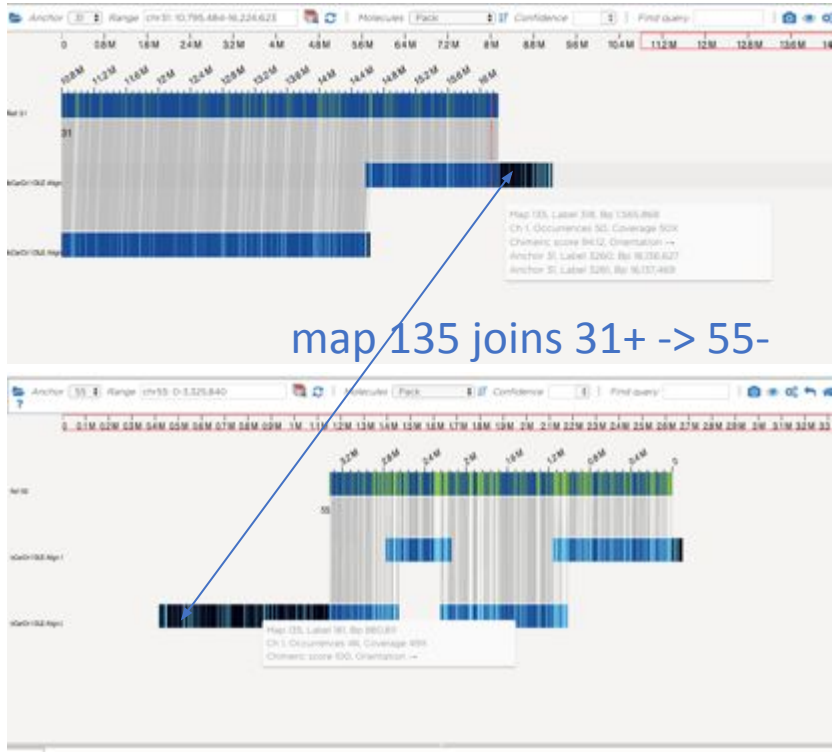


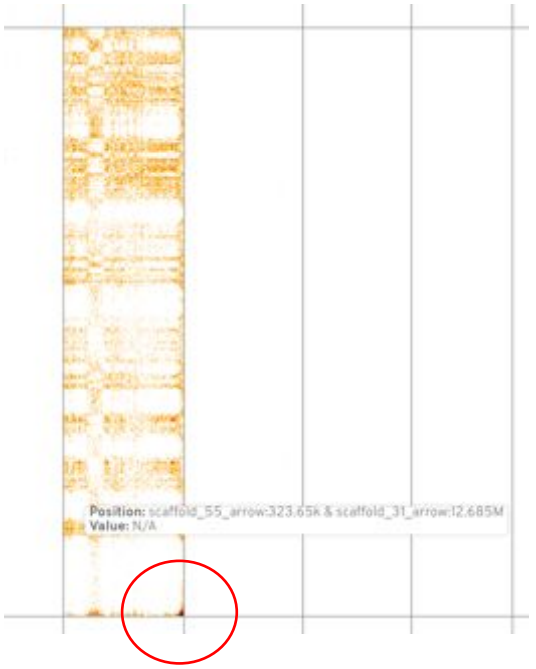
This is wrong but looks right (based on gross affinity from Pretext view)

Correct assembly



Bionano map 135 supports the join made between scaffold 31 (+ve) and scaffold 55 (-ve), Pretext view looks superficially worse despite it being correct on detailed inspection.

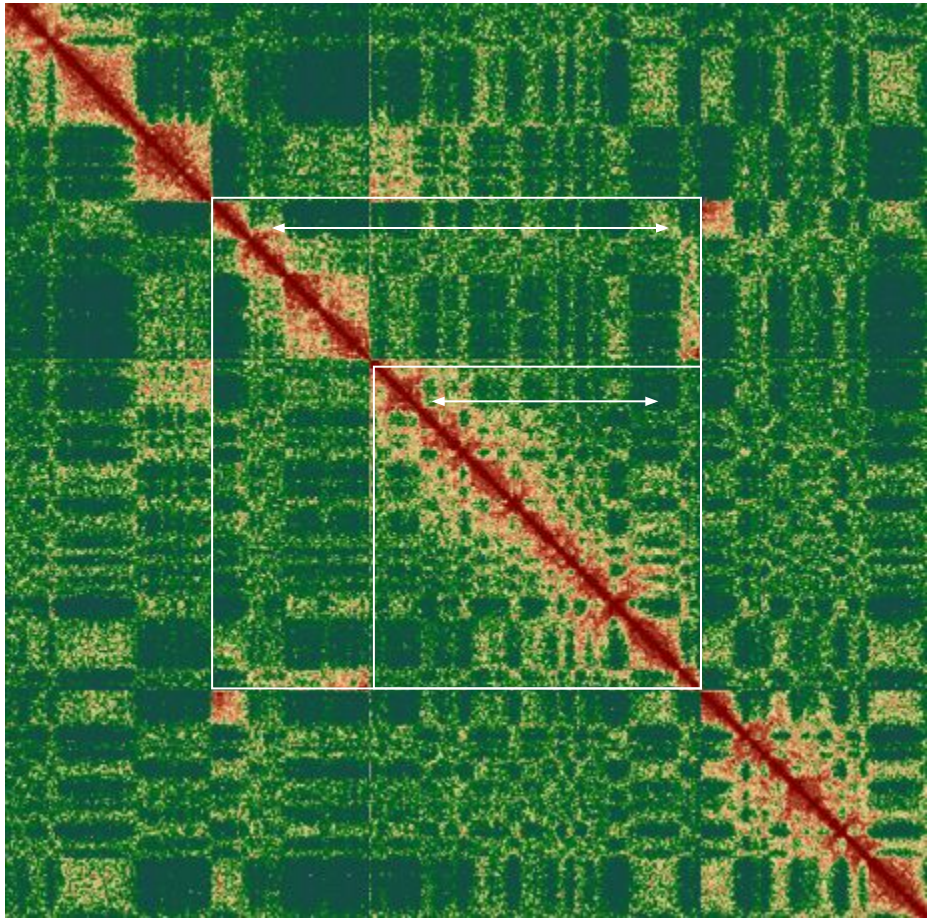




Detailed inspection of HiC in HiGlass confirms the Bionano join – although general signal is weak, at the actual join point there is very strong signal which could be missed in Pretext due to lower resolution.

bCarCri1

Colour schemes



Choice of colour schemes is important. Here 2 misassemblies are strongly highlighted in Pretext using a 3-way colour scheme called “three wave blue-green-yellow”. 2 inversions (highlighted by the white boxes) need to be made to correct the assembly.

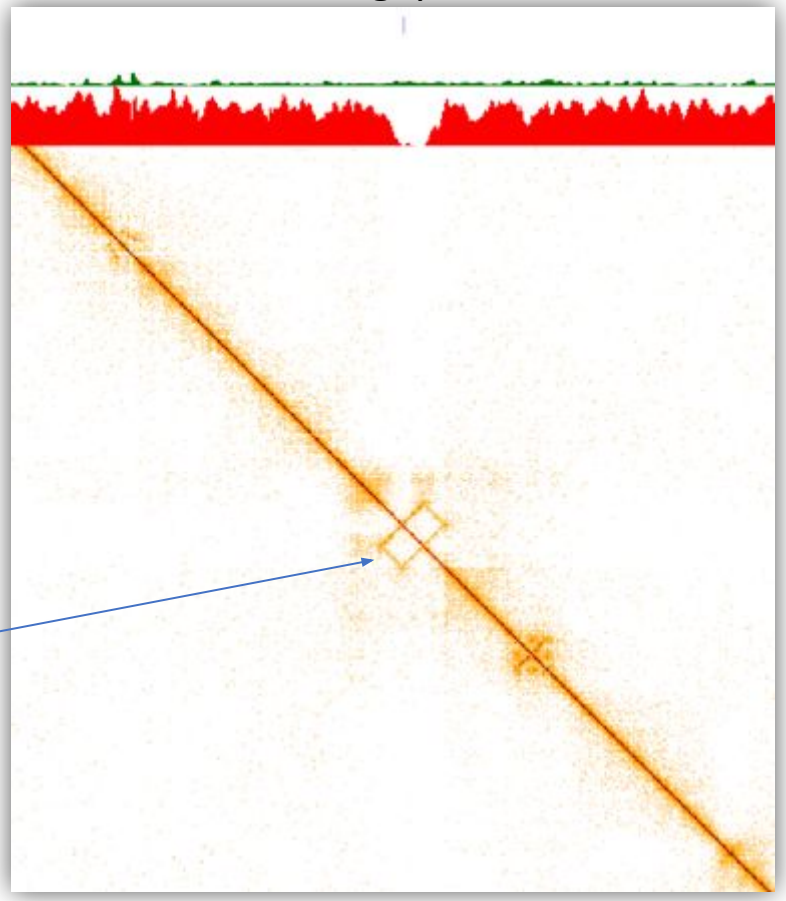
bPteGut1 superscaffold6

HiGlass vs Pretext – resolution issues in Pretext

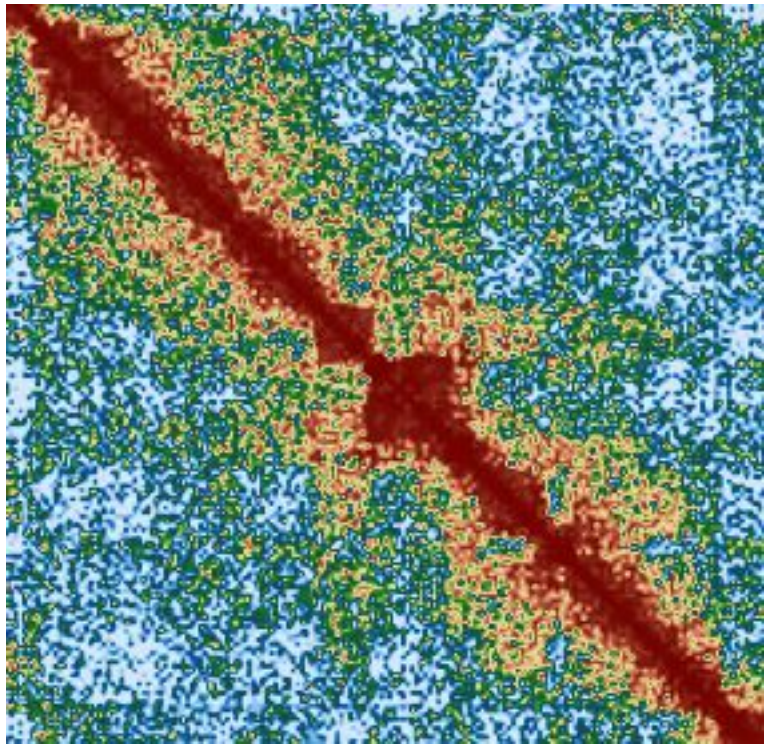
gap

repeat density
read coverage

HiGlass



false triplication of region



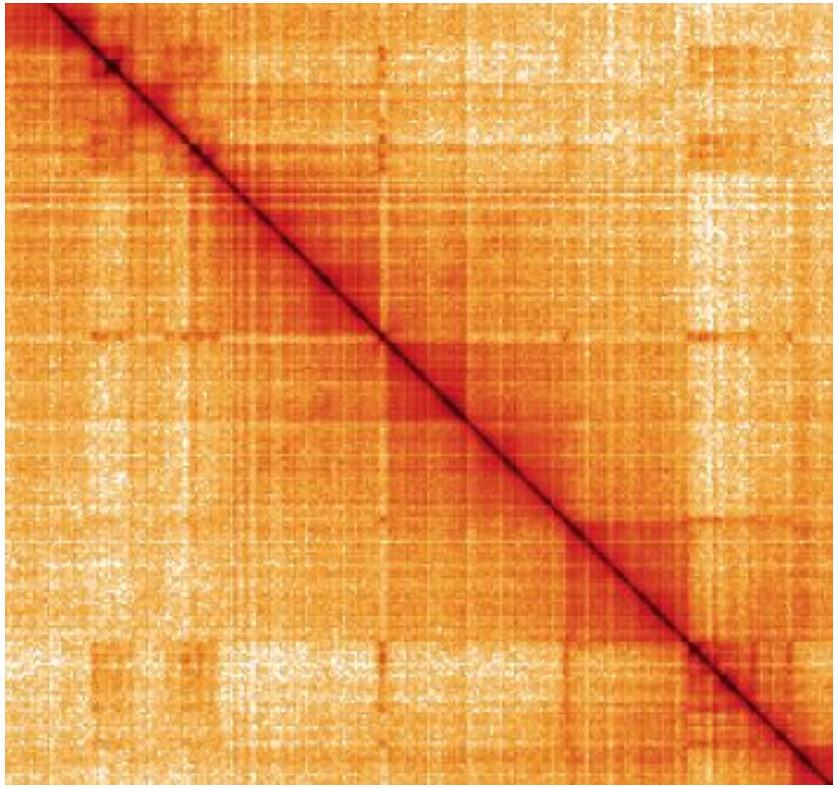
Pretext

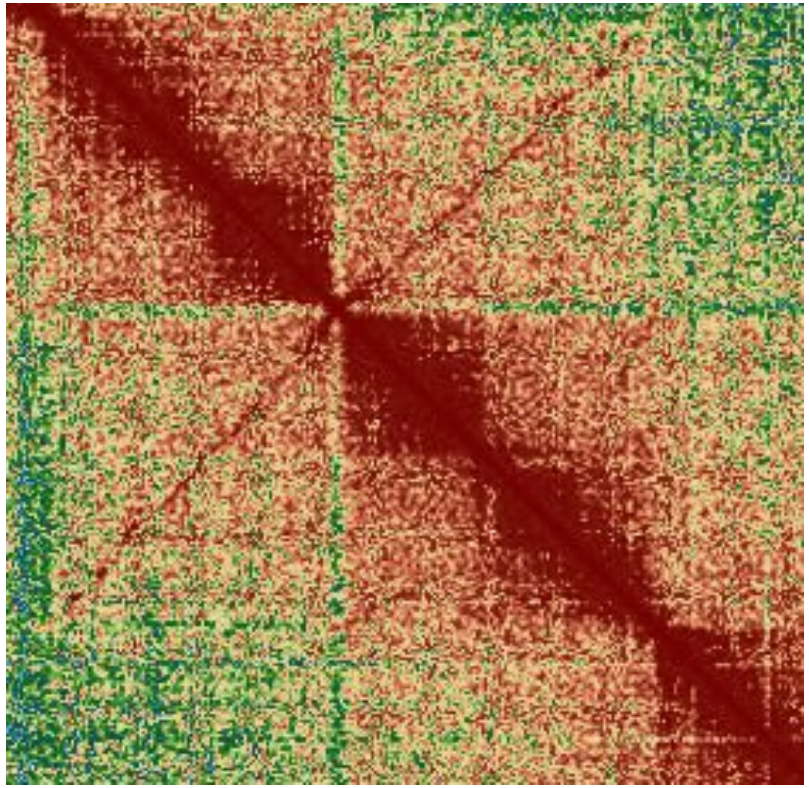
Coverage and repeat pattern in higlass show that this is an erroneously expanded region, likely due to polymerase skipping in PacBio reads. This information is completely lost in Pretext due to low resolution (ie fixed pixel number). Resolution in Pretext is unlikely to improve since memory usage quadrupels as resolution doubles

fCheRos1_1 scaffold_3 28.8mb

HiGlass vs Pretext – over-saturation in Pretext

HiGlass





Pretext

Pretext images often ”over-saturate” such that a very faint signal can appear to be more significant than it really is. Here the inversion signal is virtually background in the HiGlass, but very prominent in Pretext. No curation action is required here.

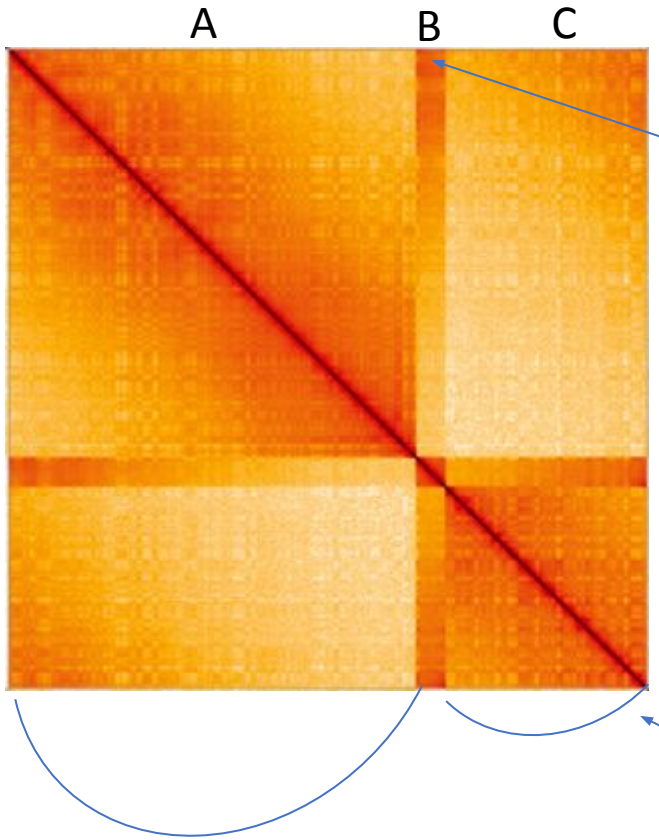
Misassemblies and artefacts

Covers:

- Rearrangements
- Imposter contigs
- Collapsed repeat (tandem, direct, inverted)
- Systematic assembly artefacts
- Scaffolding quirks
- Contamination

Rearrangement scenario 1

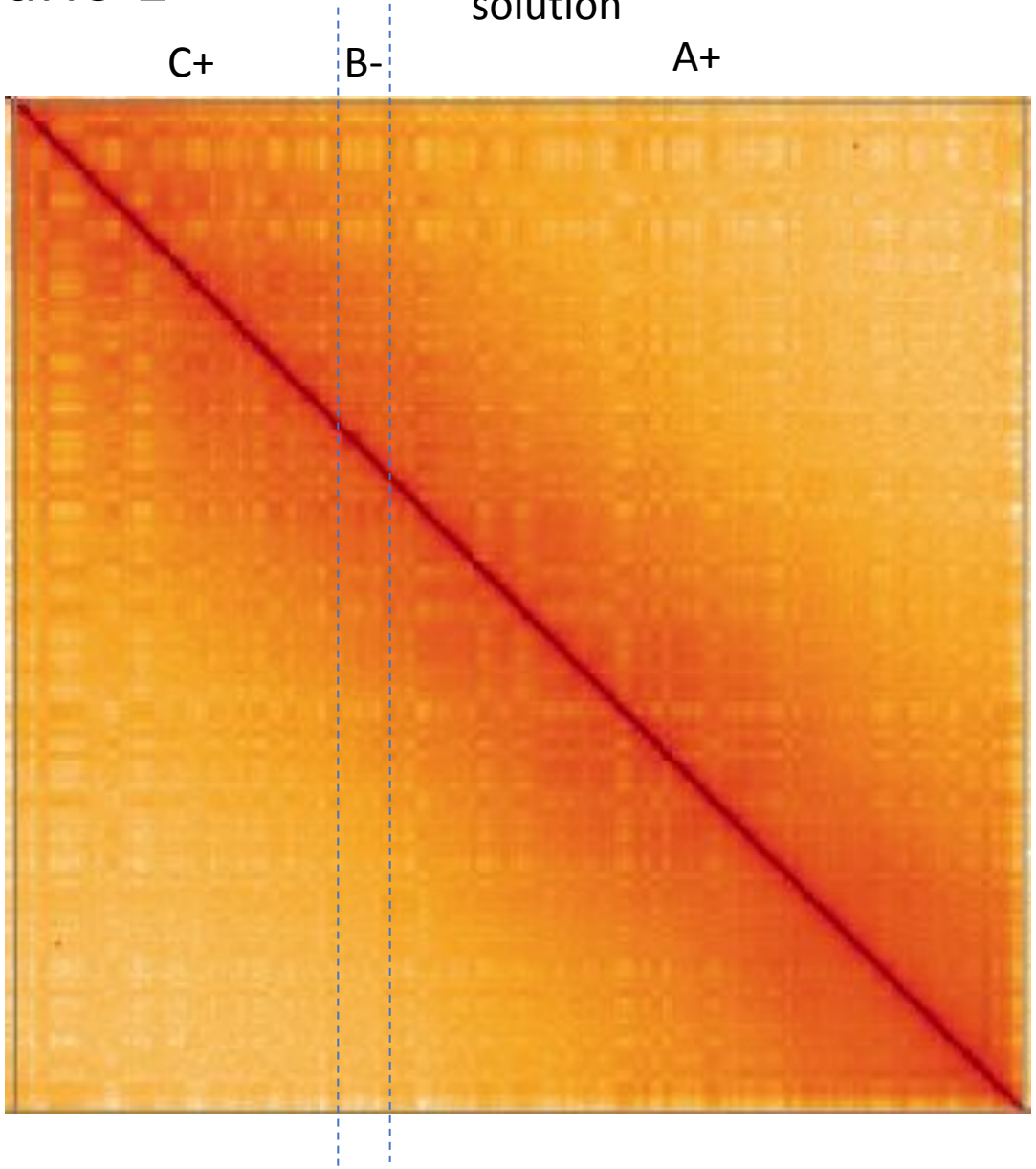
Assembly problem



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

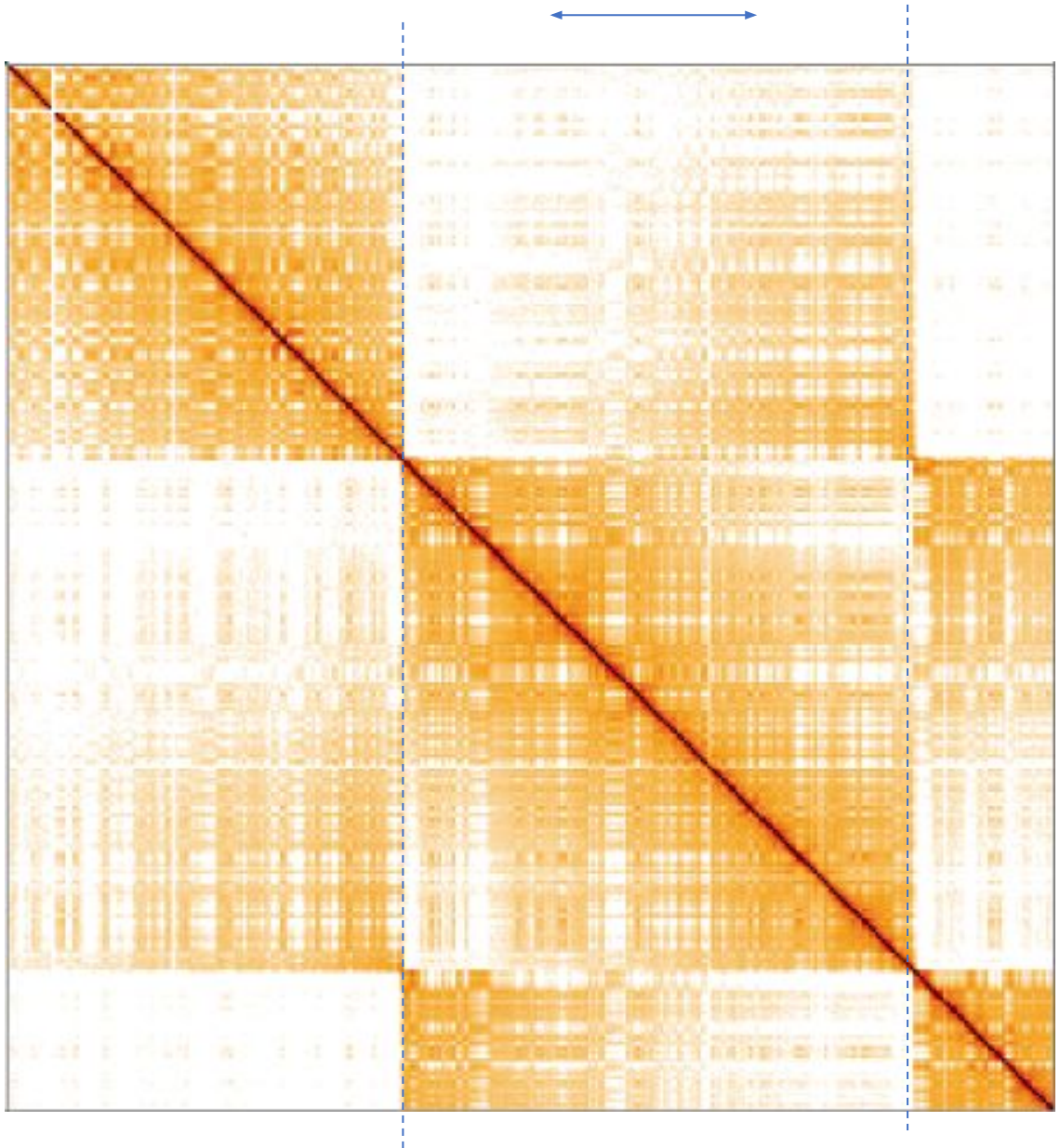
strongest affinity (blue links) between B/r and C/r and B/l and A/l, leading to the solution on the right.

solution



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

Rearrangement scenario 2

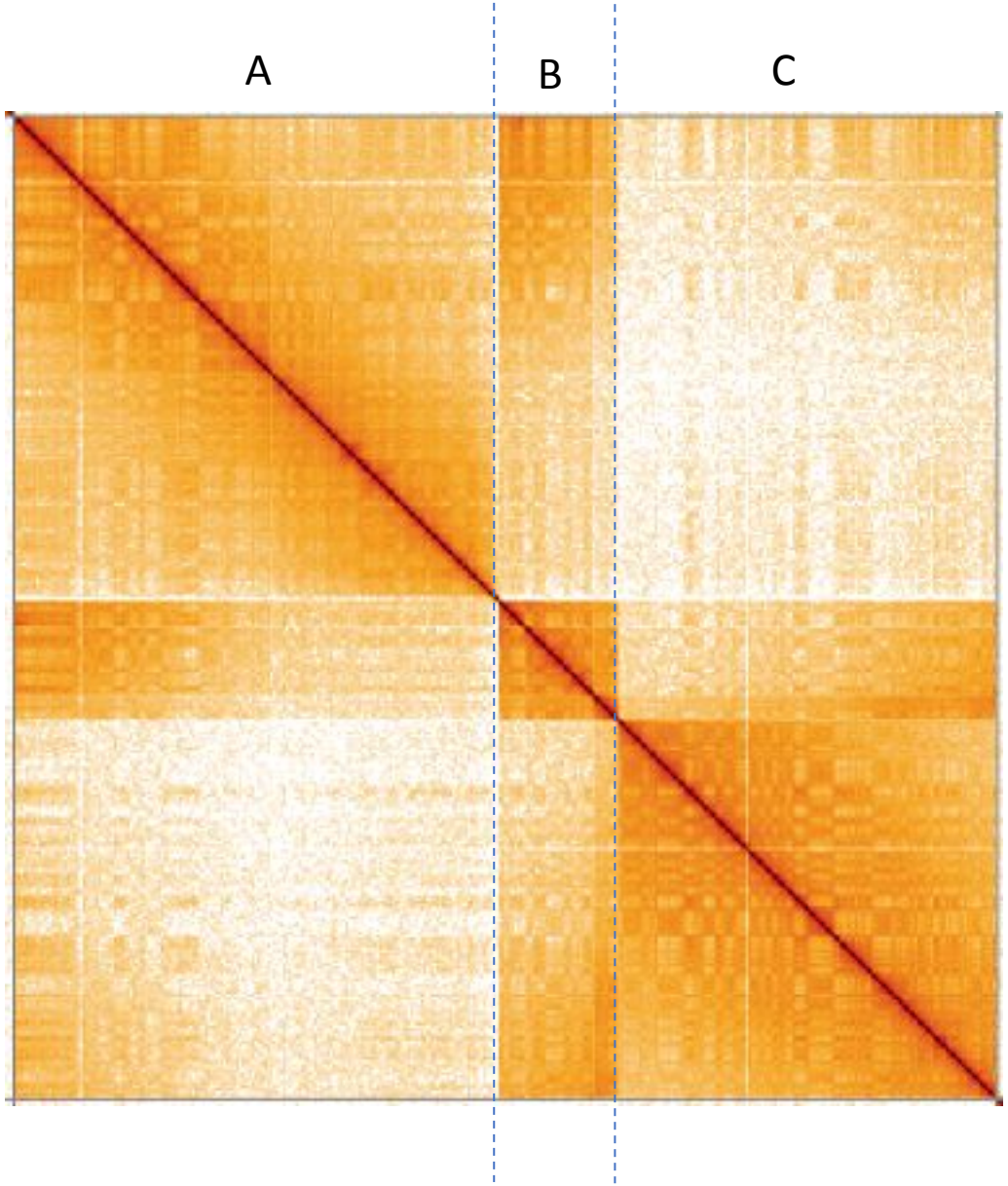


Three distinct sections are visible. The 3 sections are in the correct order, but the 2nd section needs to be inverted.

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.

fThaEle1 scaffold2

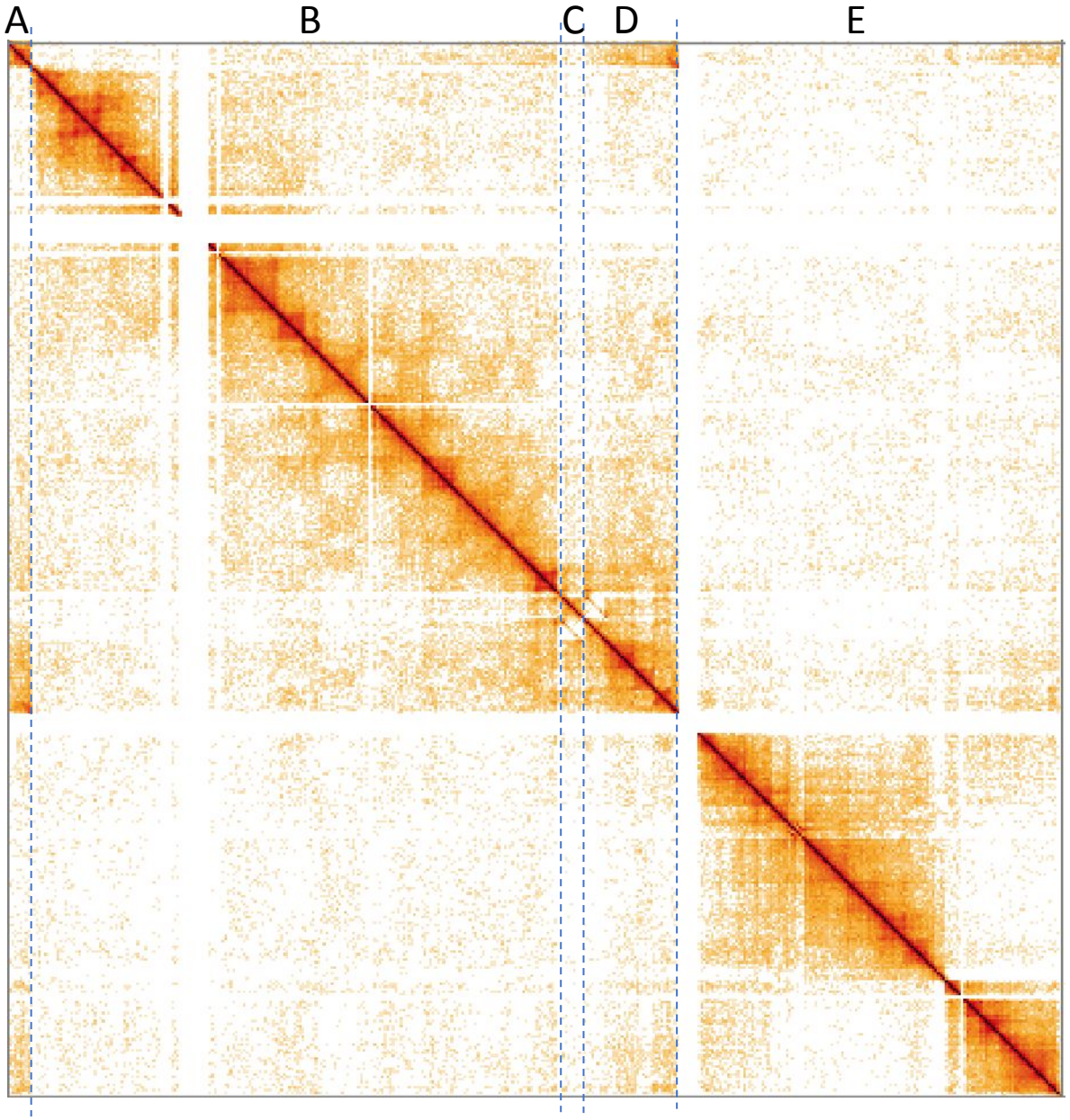
Rearrangement scenario 3



Solution is C+, B-, A+

fAciRut3_p scaffold6

Rearrangement scenario 4



Solution is E+ A+ D- B- (remove C as haplotig)

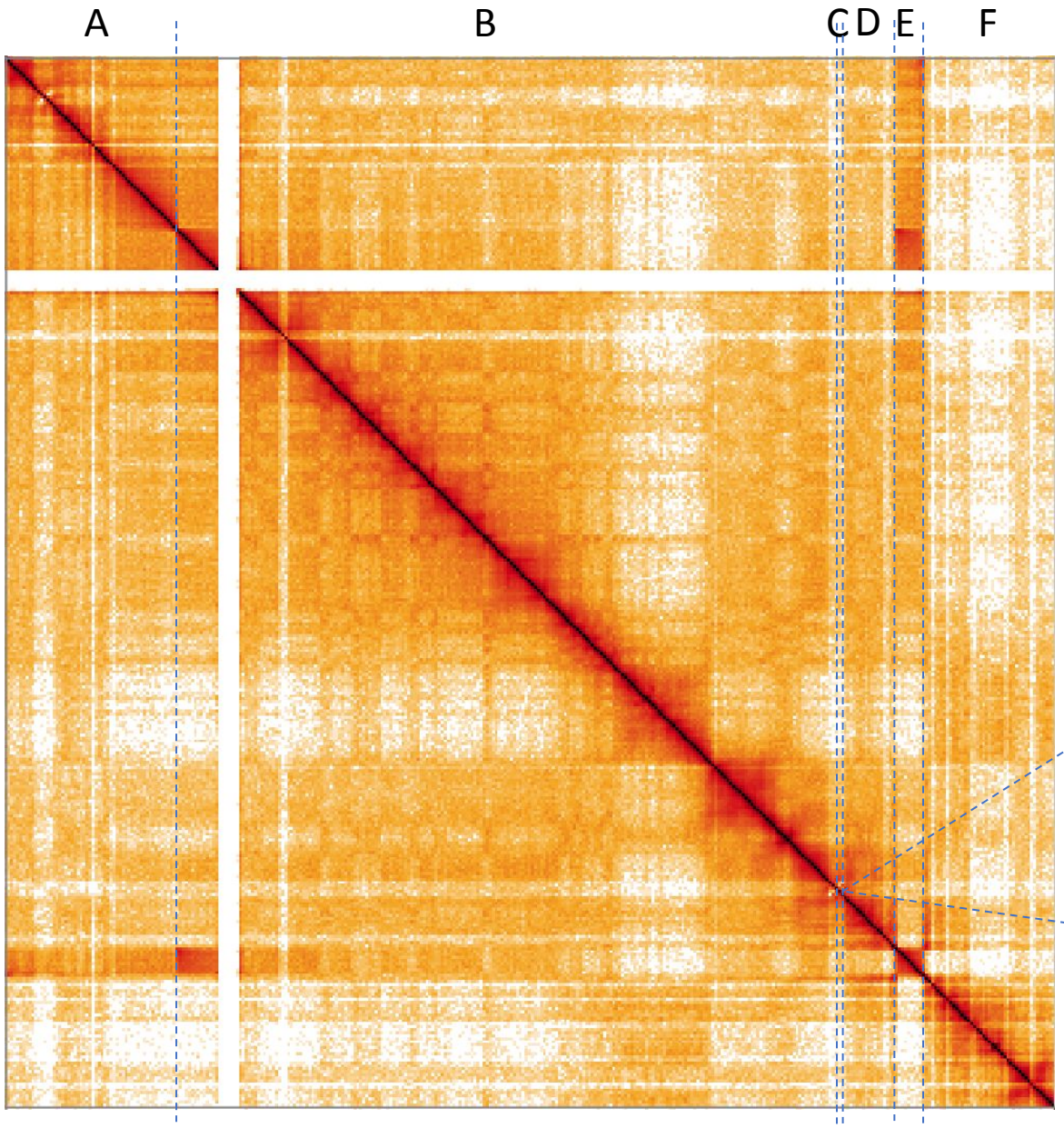
Evidence that C is a haplotig:

1) Lower intensity in the map
2) 2 parallel lines either side of the centre diagonal (the 2 copies of the haplotype which mean that this region is over-expanded). We only want to remove half of the haplotypic duplication, ie one copy.
3) Further evidence (eg self-comp matches) can usually be seen in gEval.

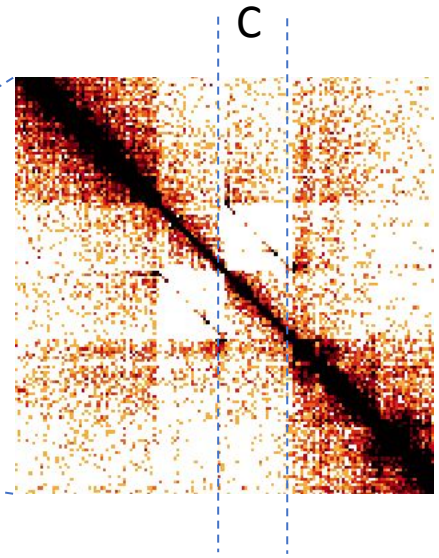
(See section on haplotypes for a fuller description of haplotypes and their resolution)

bGeoTri1 scaffold28

Rearrangement scenario 5



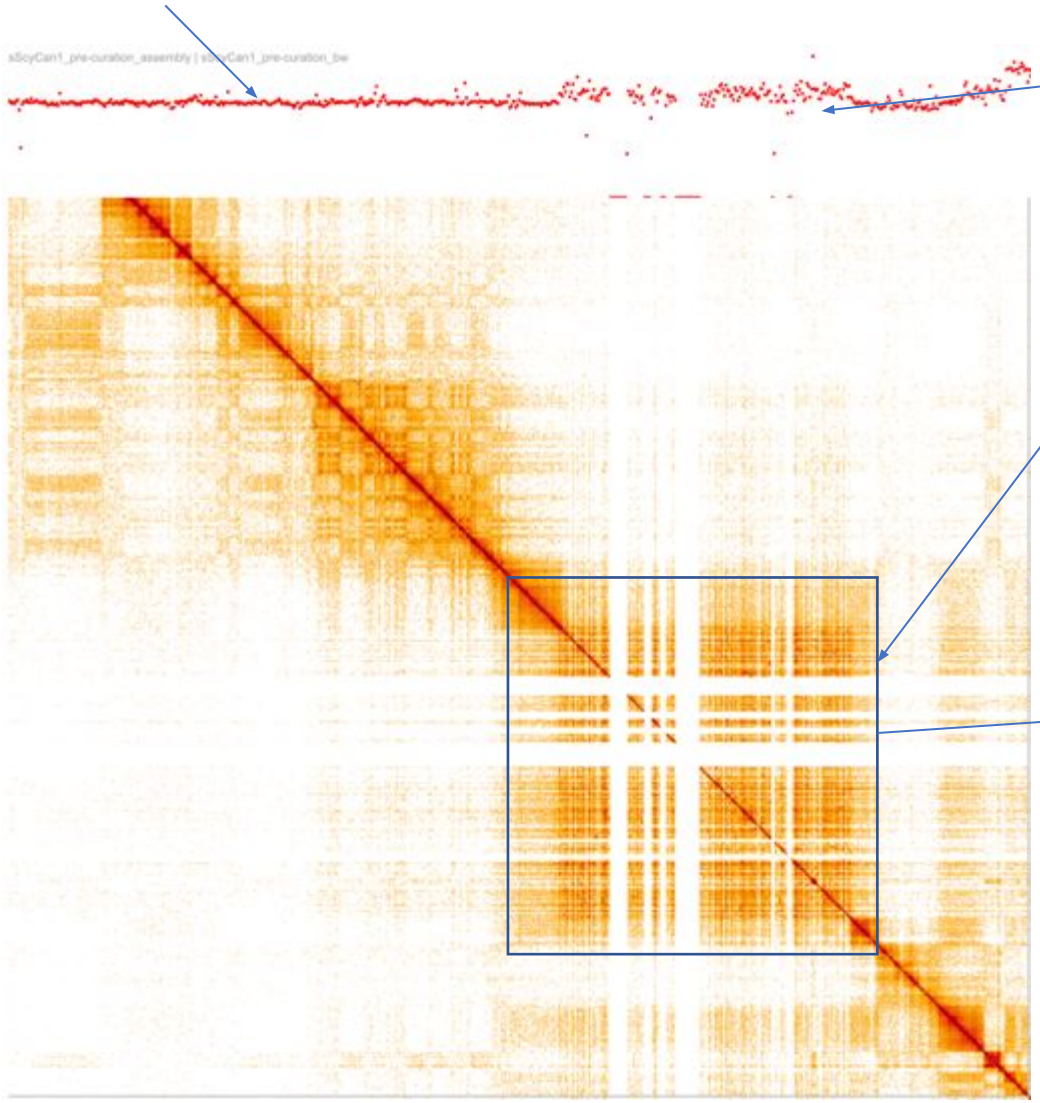
Solution is A- E- B+ D+ F+
remove C as haplotig



bGeoTri1 scaffold12

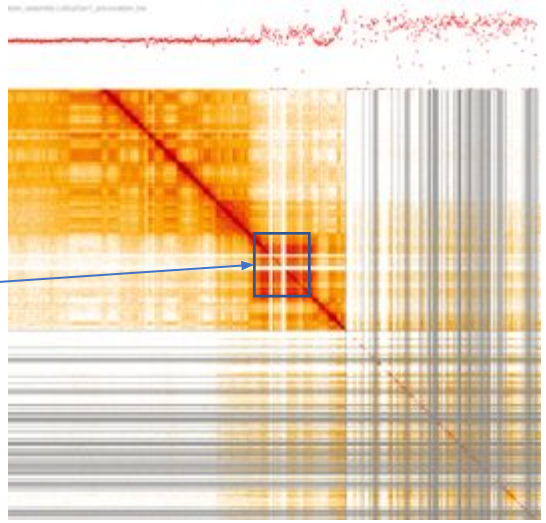
Region of low complexity/tandem repeat

Normal coverage outside the repeat



Elevated coverage due to collapsed assembly in tandem repeat

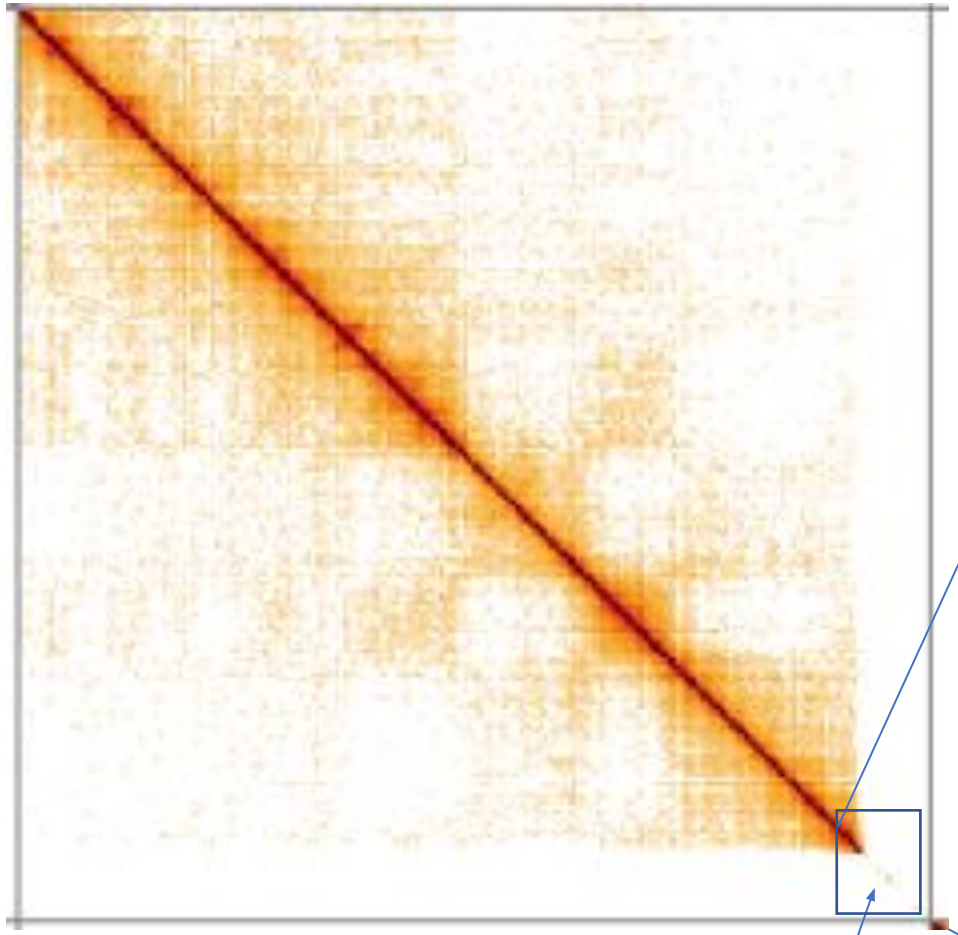
Distinctive HiC pattern due to this tandem repeat – high affinity between all parts of the repeat (due to many reads and their potential to map readpairs all over the repeat), and reduced affinity with the rest of the chromosome



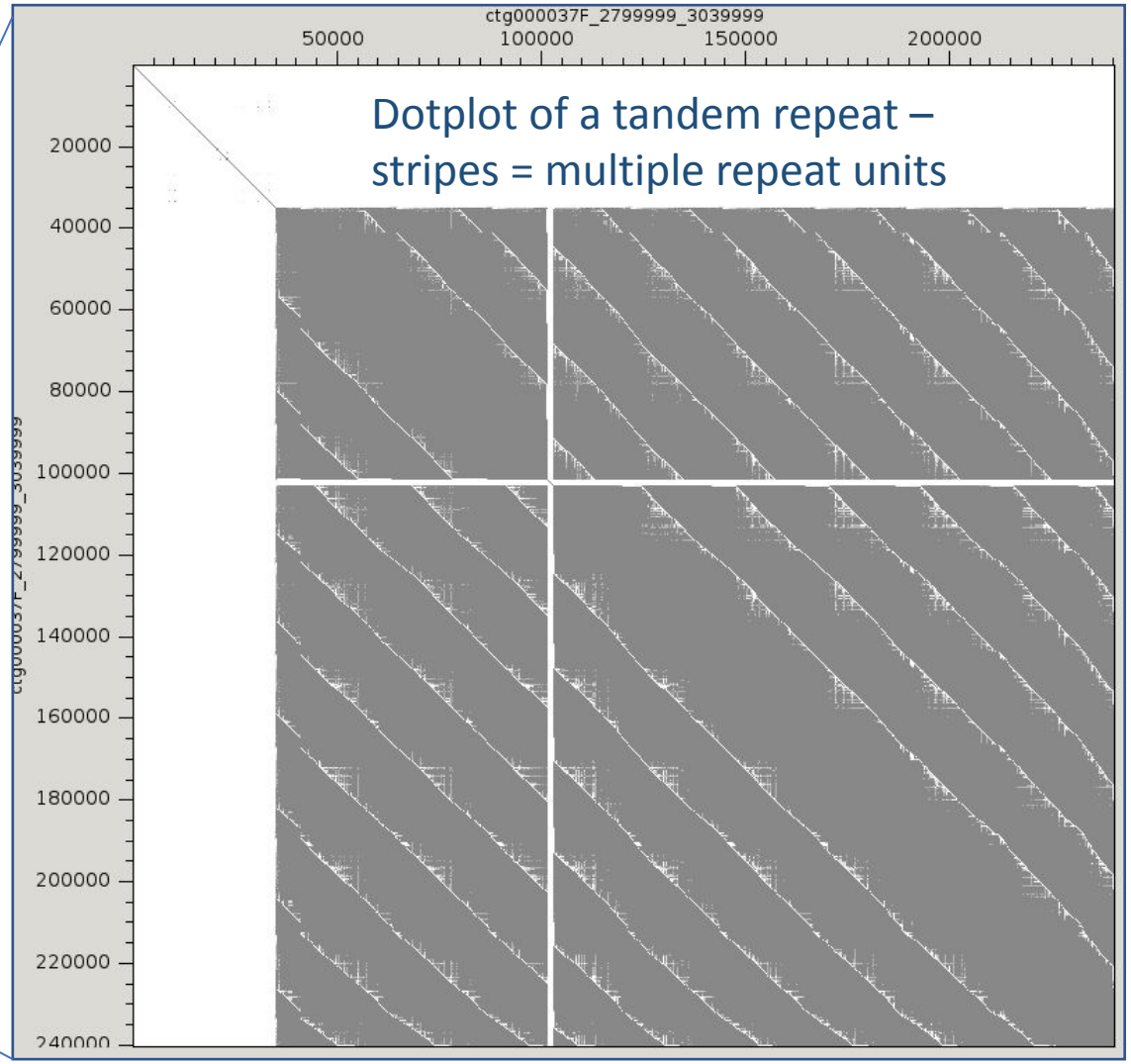
Zooming out a little, many shrapnel contigs have affinity with this repeat and also have high coverage. This looks like a large highly repetitive region of the genome that is resistant to assembly

sScyCan1 scaff31

Tandem repeats cause low HiC signal (if multi-mapping reads are filtered out)

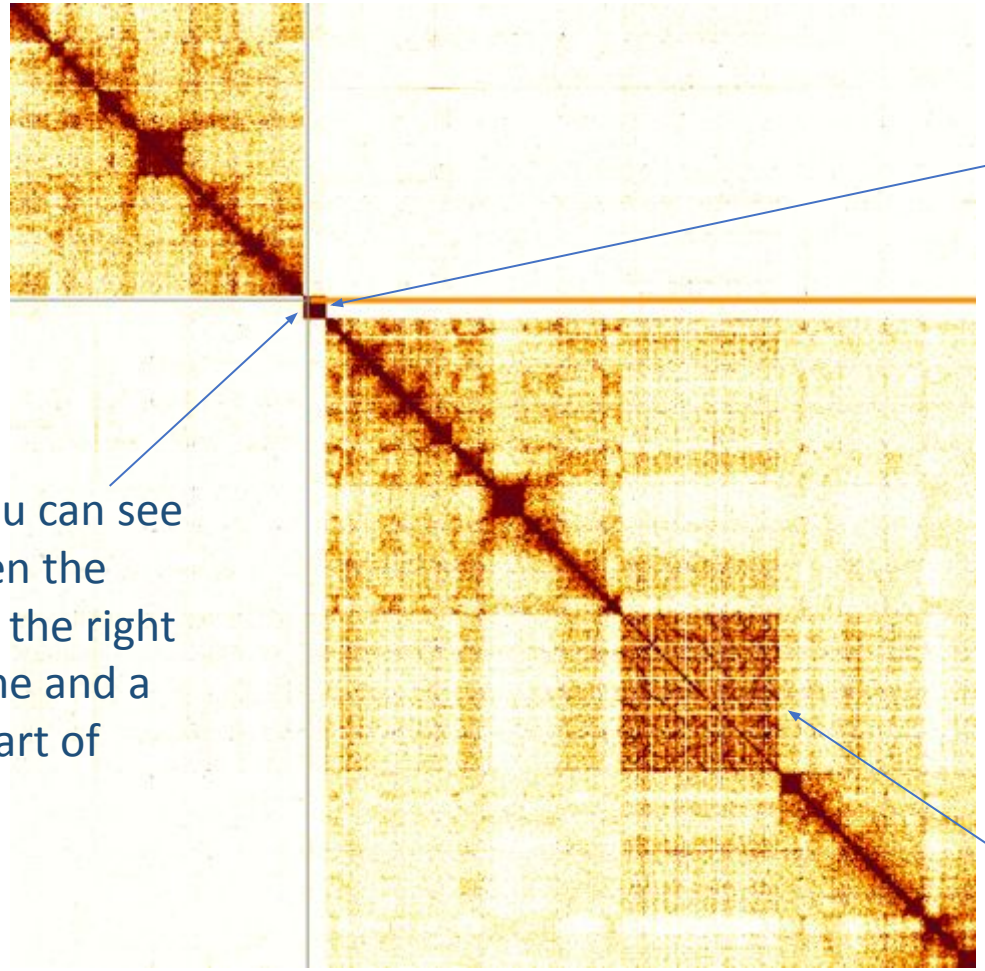


fChaCha 000037F (hifi data)



Here contact is extremely thin… this is explained in this case by tandem repeat reducing ability to map Illumina reads

Tandem repeats with **multi-mapping reads** switched on



ilVanCard2_1 SUPER_26 into SUPER_3

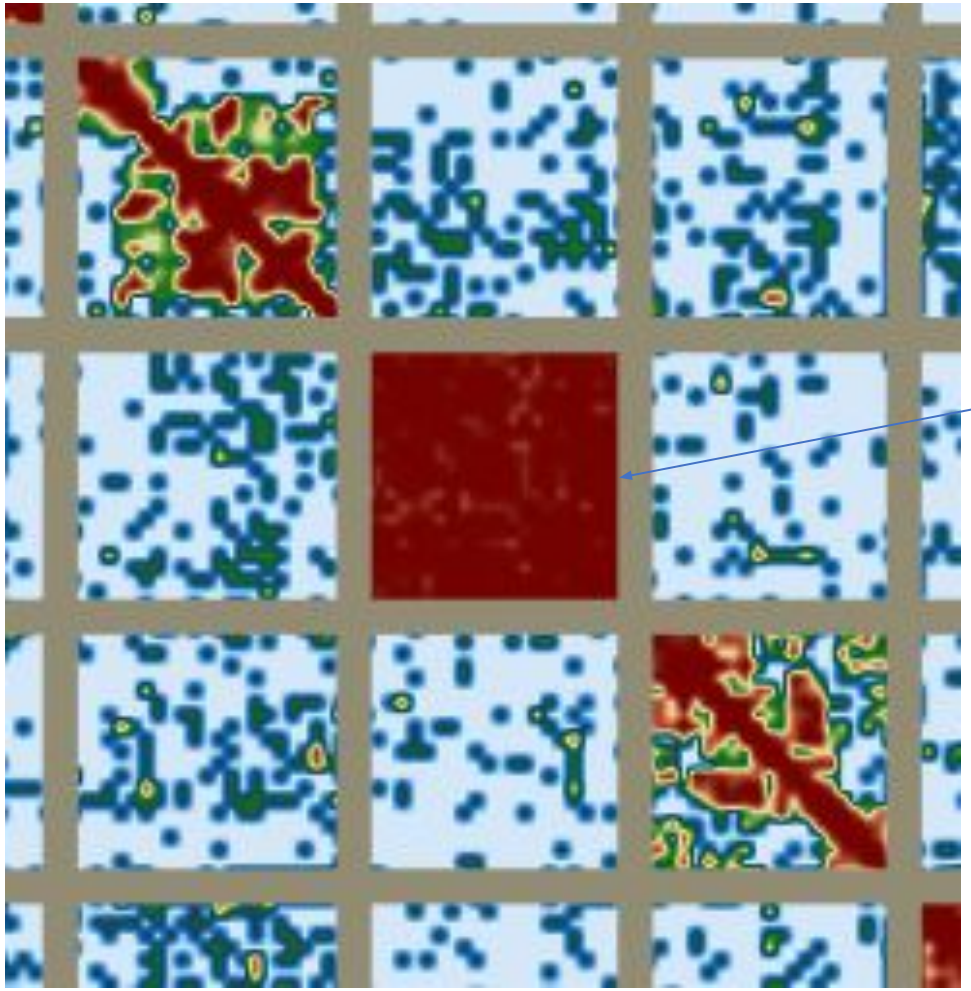
Subtelomeric repeats seeming to have no interaction with the rest of the chromosome. Hicanu and hifiasm assemblies of this region give same result.

It is very common to see subtelomeric regions seeming to have very little affinity with the rest of the chromosome in this way. Relying on HiC alone, it would be very easy to incorrectly remove these regions.

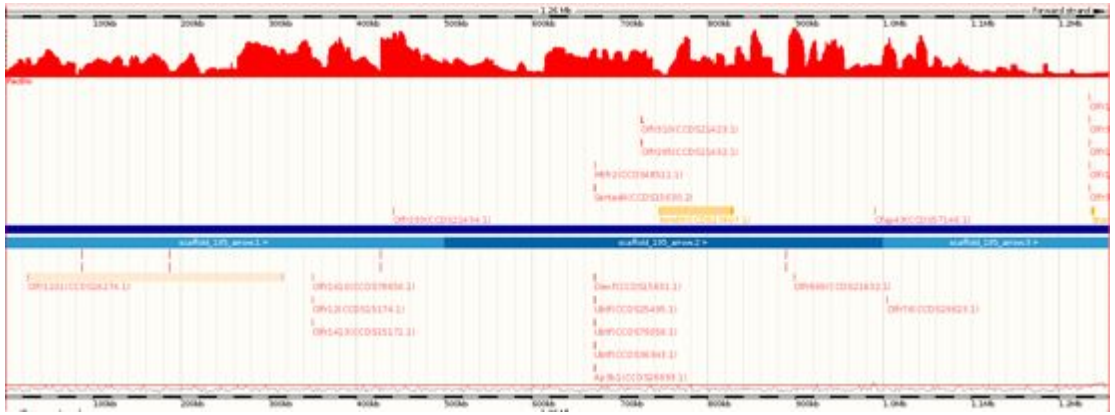
If you look carefully, you can see the association between the subtelomeric repeat at the right end of one chromosome and a similar repeat at the start of another

Another tandem repeat whose structure can be seen much more clearly with multi-mapping reads switched on

Highly repetitive collapsed sequence



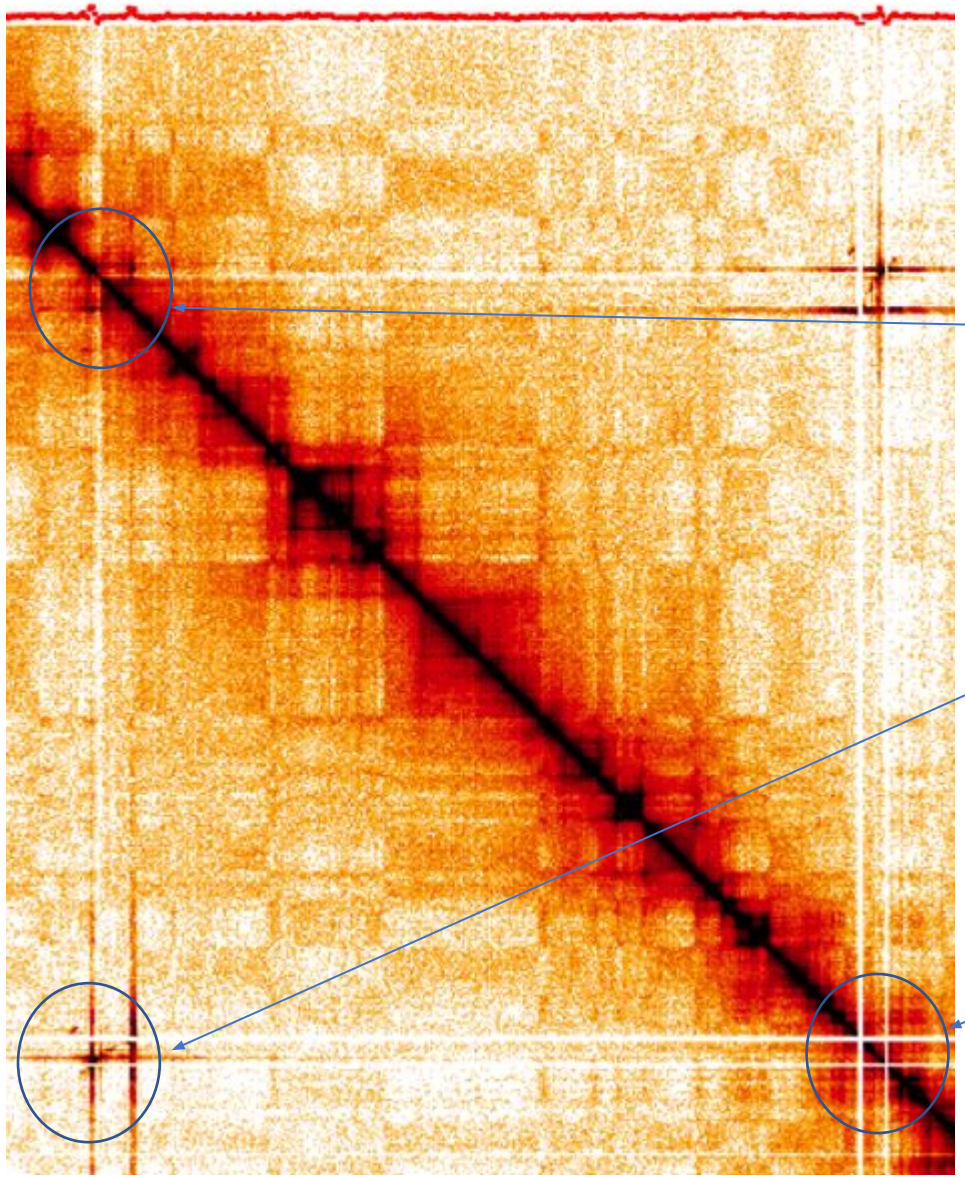
Heavy contact across the entire window, indicative of highly collapsed repetitive sequence as evidenced in the gEval window by coverage and by many genes of the same type mapping (albeit poorly) across the scaffold. This scaffold is around 1.8Mb.



(A couple of other scaffolds have been included in the above image to show how distinctive the repetitive scaffold is)

http://vgp-geval.sanger.ac.uk/VGP_mSciCar1_1/Share/dca3aaefda470d7c0a32cd2012650a62260095
https://vgp-geval.sanger.ac.uk/VGP_mSciCar1_1/Location/View?r=scaffold_195_arrow:-744498-1255501

Gaps due to collapsed repeat



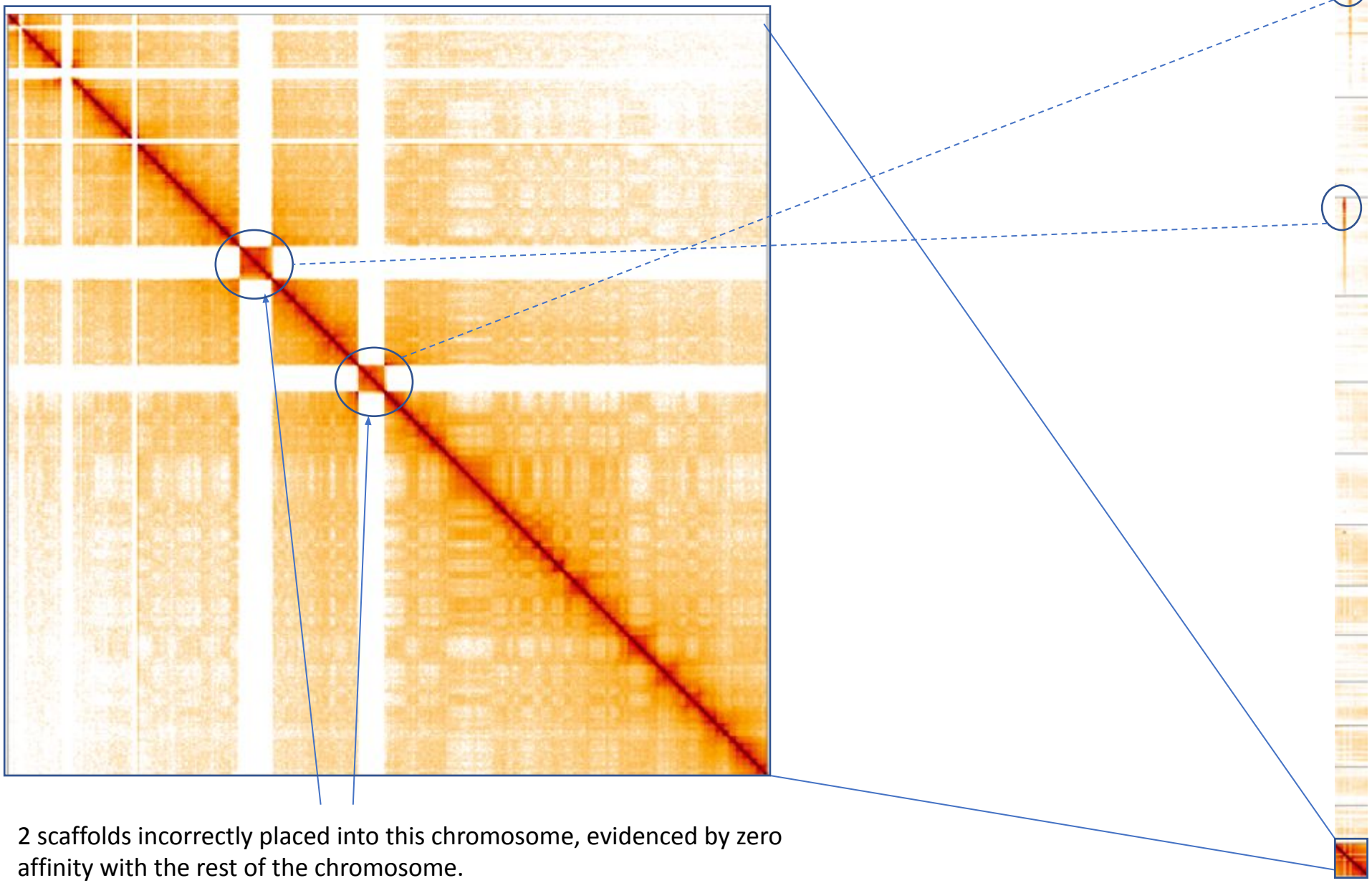
2 collapsed repeats (2 corresponding peaks visible in red coverage histogram at the top of the plot)

Off-diagonal contact usually signifies a misassembly, but it can also occur due to the same large repeat occurring in 2 diverse locations as in this example. The gaps in one repeat could be filled using the collapsed sequence from the corresponding repeat

2 gaps caused by the collapsed repeats (ie reads are missing from this repeat because they are assembled at the collapsed repeat location)

mLemCat1 scaffold1

Imposter sequences



scaffold1

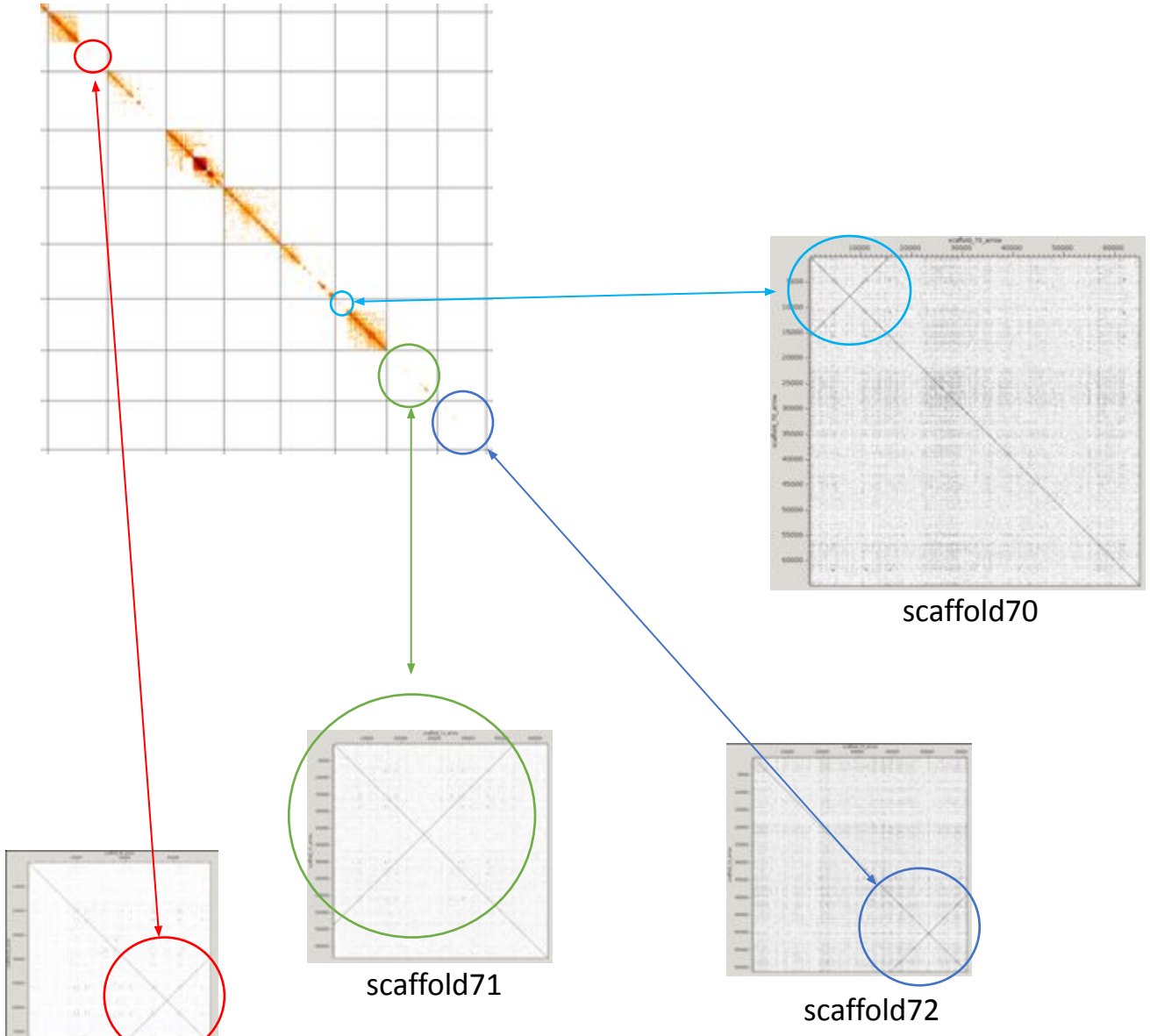
scaffold3

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

2 scaffolds incorrectly placed into this chromosome, evidenced by zero affinity with the rest of the chromosome.

sAmbRad1_2 scaffold14

Systematic sequence artefacts (often specific to a particular assembly)



Regions with zero HiC reads mapping could be sequence artefacts.

In each case investigated in mLemCat1, these "empty" regions appear to be perfect inverted repeats – this is suspicious. Also, they don't have double depth coverage as would be expected with genuine partially resolved inverted repeats.

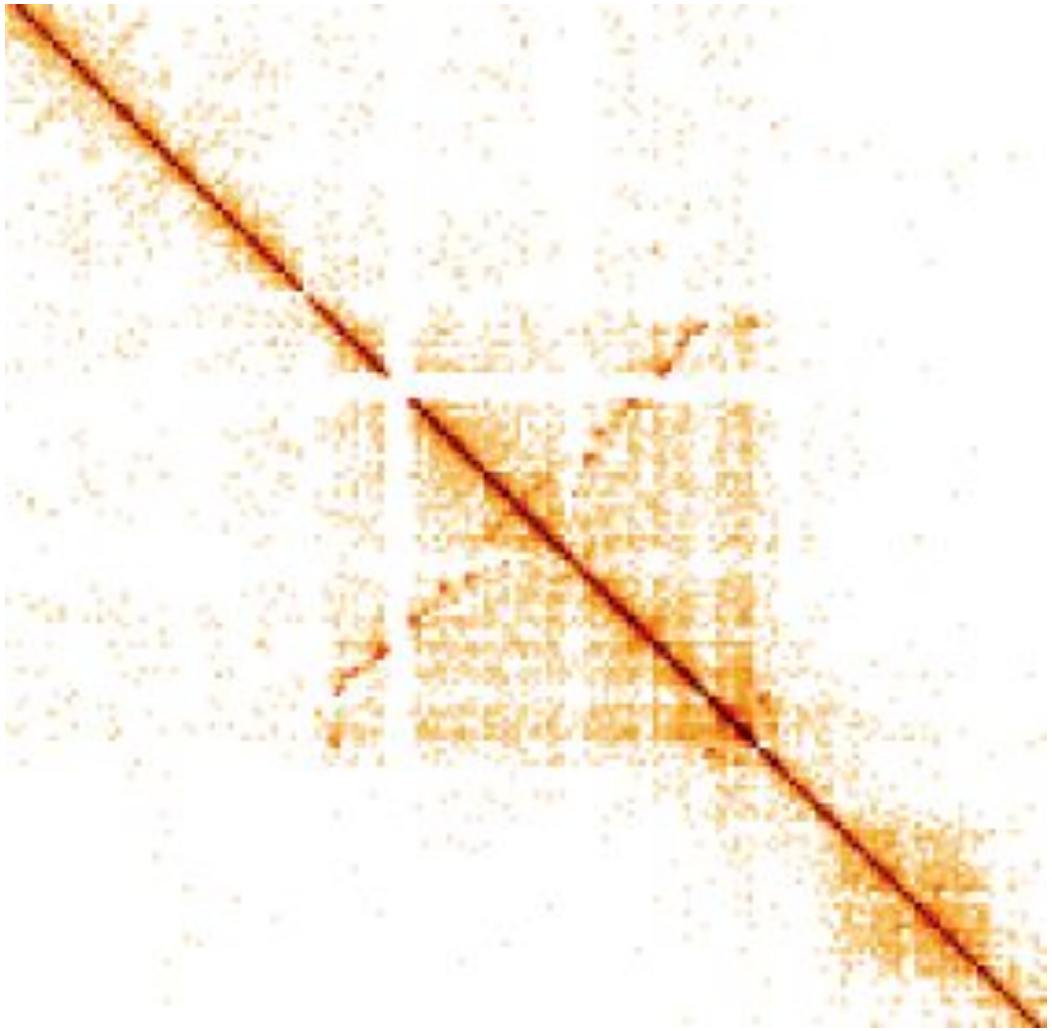
Illumina reads in these regions would be equally able to map in 2 locations so would have a mapping Q of zero and would be filtered away explaining the zero coverage.

If these are sequence artefacts it would explain why they weren't joined to other scaffolds. Removing the false 2nd copy of the sequence in each case prior to scaffolding should lead to better scaffolding.

In this case this might be caused by polymerase switching in PacBio reads.

mLemCat1

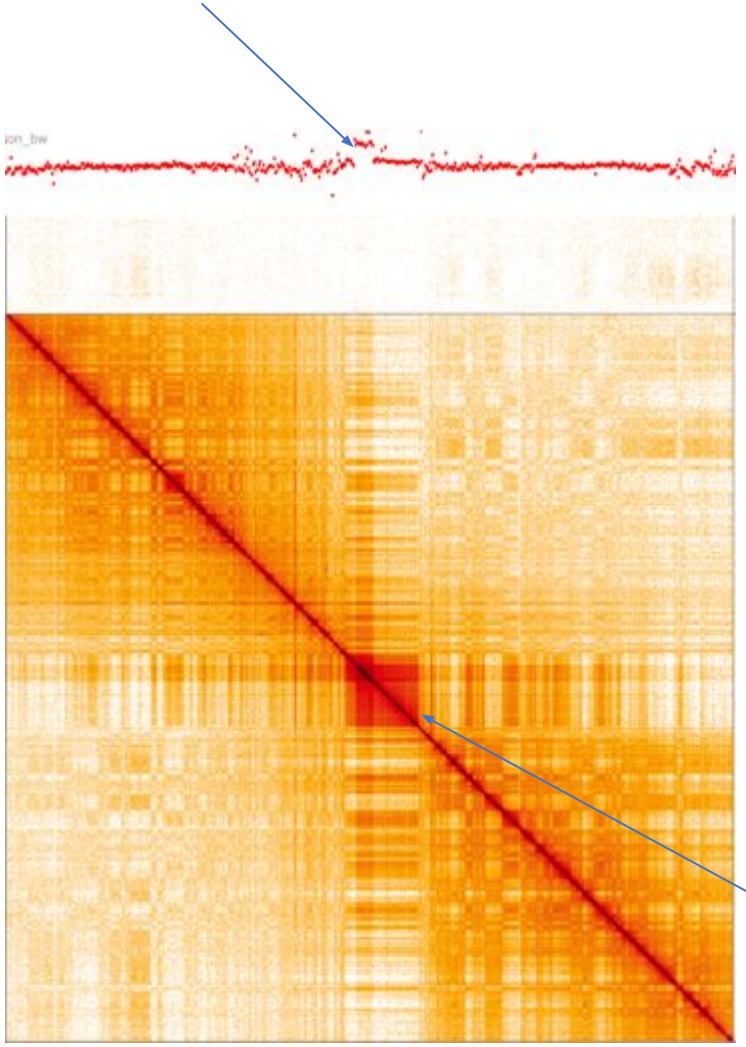
Inverted repeat



This inverted repeat has 2 arms of 500Mb each. They are similar enough that some reads mismap to the other copy giving a distinctive opposing diagonal line. There is nothing to suggest that this is assembled incorrectly. Indeed the fact that the scaffolding has been able to enter and exit the inverted repeat rather than the scaffold terminating in the inverted repeat suggests it has been resolved correctly.
(The breaks in the line are caused by sequence gaps – these may represent the most homologous sections of the inversion where assemblers would struggle to apportion reads to both copies of the repeat correctly)

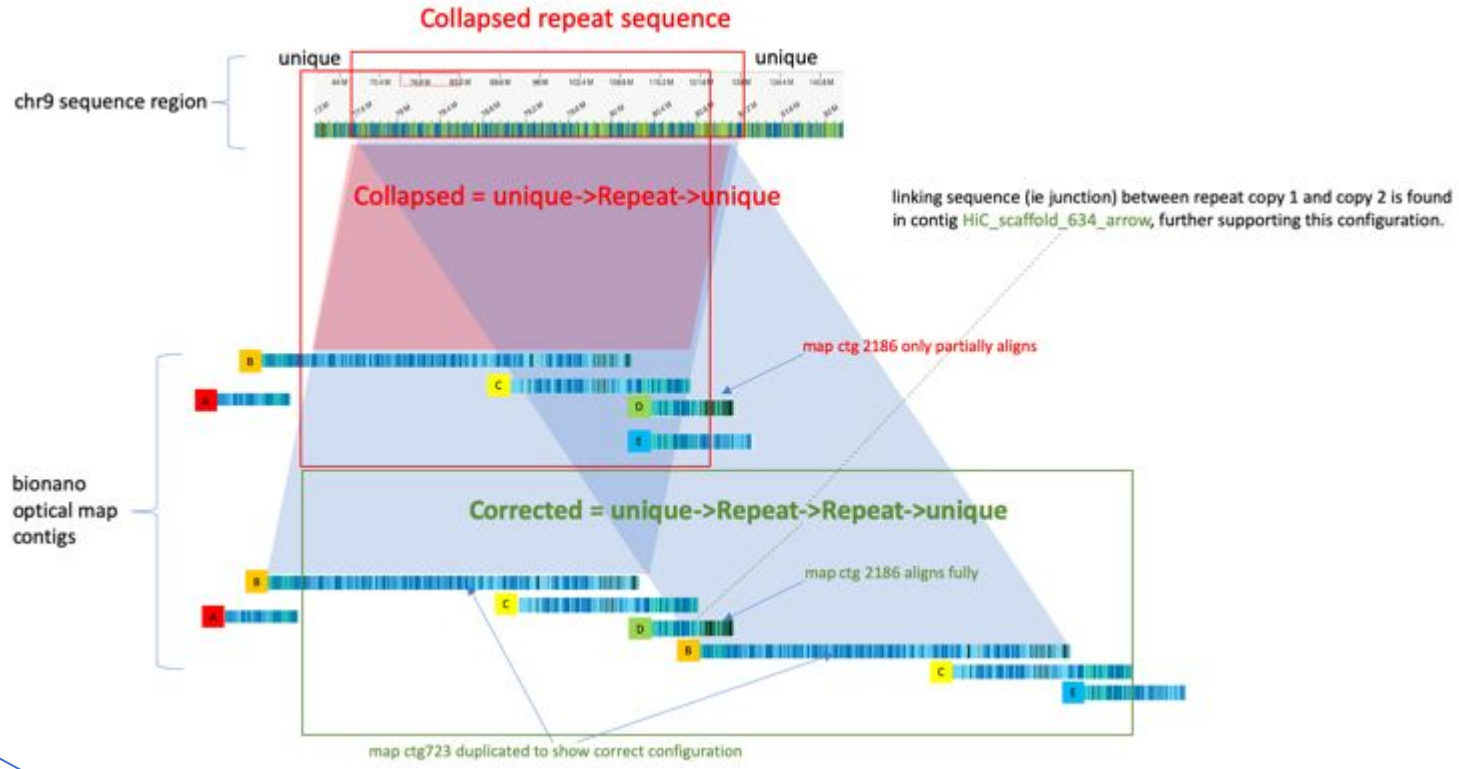
Massive direct repeat

double depth coverage



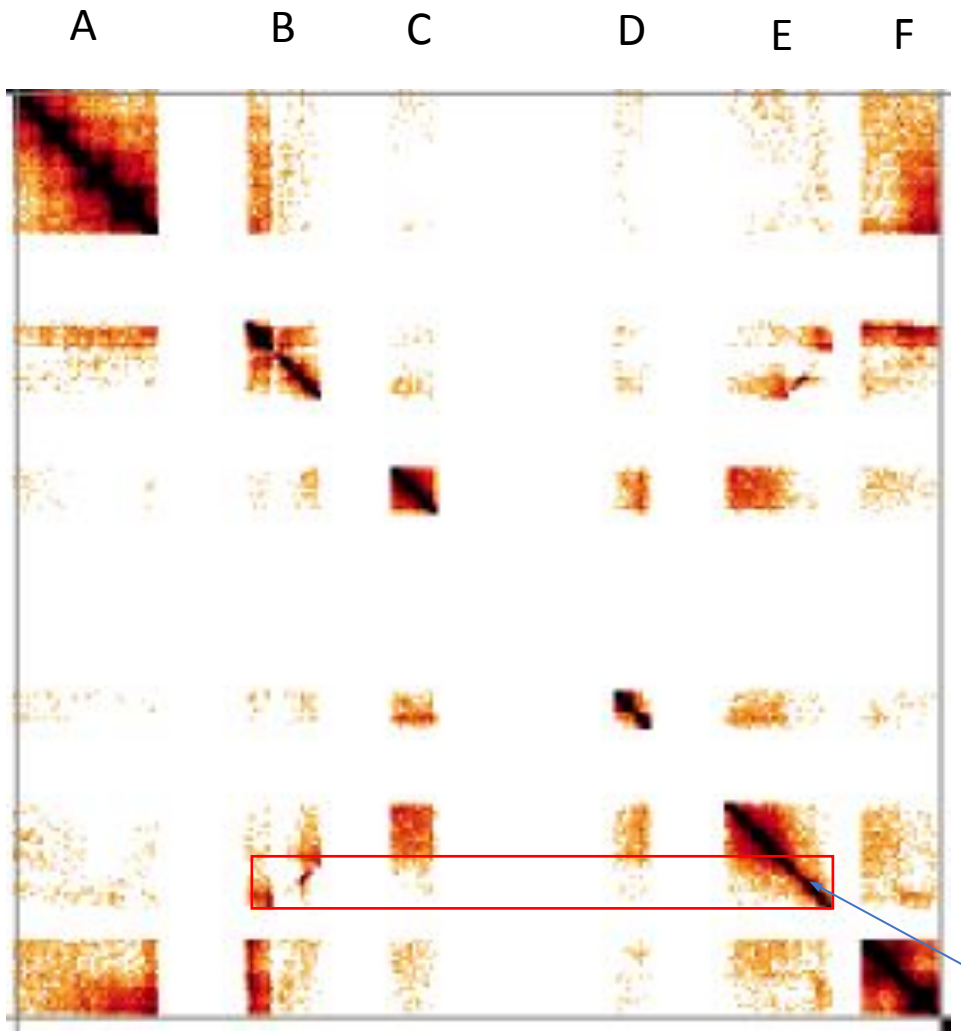
sScyCan1 – scaffold9

This repeat currently accounts for 2.5% of catshark chr9, but should really account for 5%. Double height sequence indicates the 2 copies must be very similar as they've failed to unzip. Bionano data supports a massive 2 copy direct repeat.



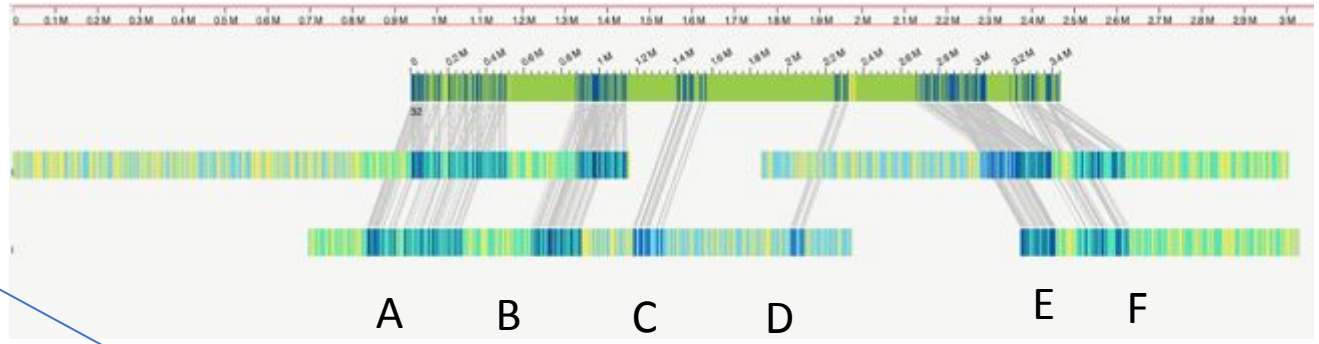
Interestingly, a wider pattern of high contact can be seen surrounding the collapsed repeat and totaling 20Mb. This is a region of high GC and possibly low complexity containing the centromere.

Scaffolding with Bionano first then further scaffolding with 10X can result in scaffolding errors



2 separate Bionano maps (probably representing alternative haplotypes) have scaffolded contigs together into 2 phased blocks (ie certain contigs map better to one map than the other). These blocks have then been subsequently scaffolded together incorrectly by 10X data (ie the blocks have been assembled back to back in the reverse orientation). To fix this, we have to manually fix the scaffolding, interleaving the contigs.

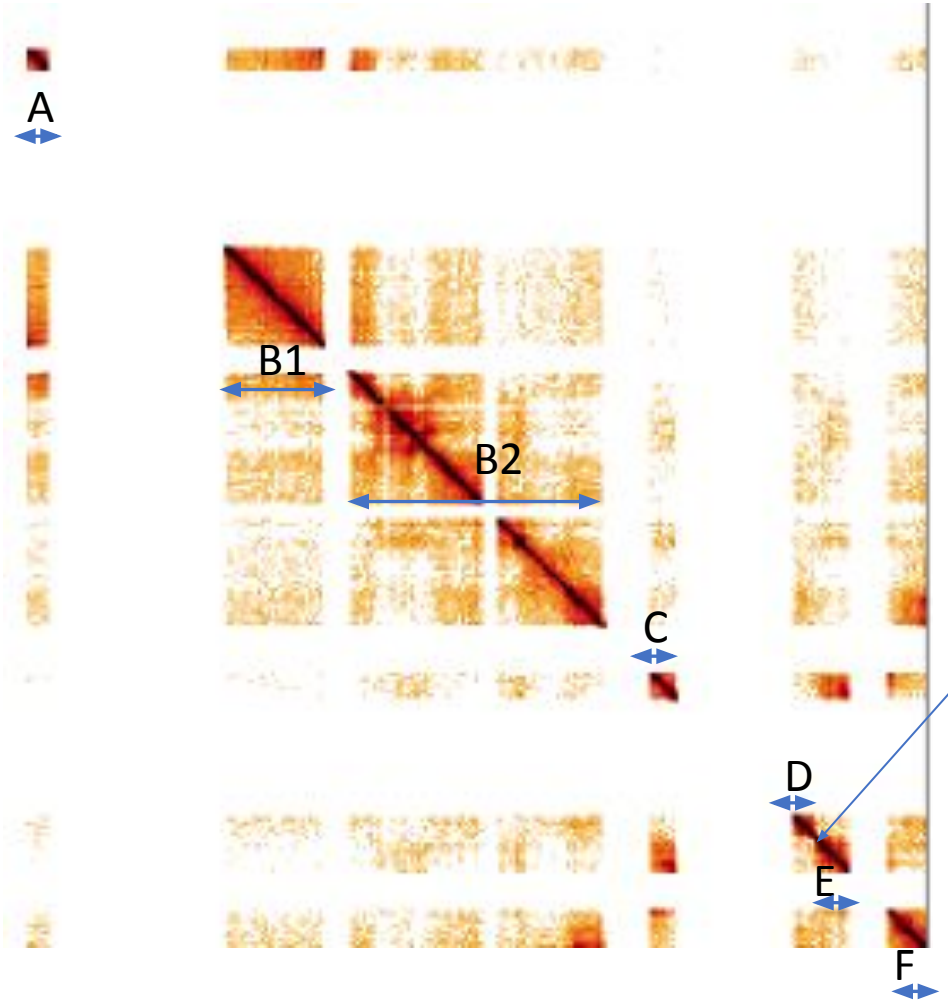
Solution = A+, F-, B+, E-, C+, D+



Remove 1 haplotype to avoid duplication

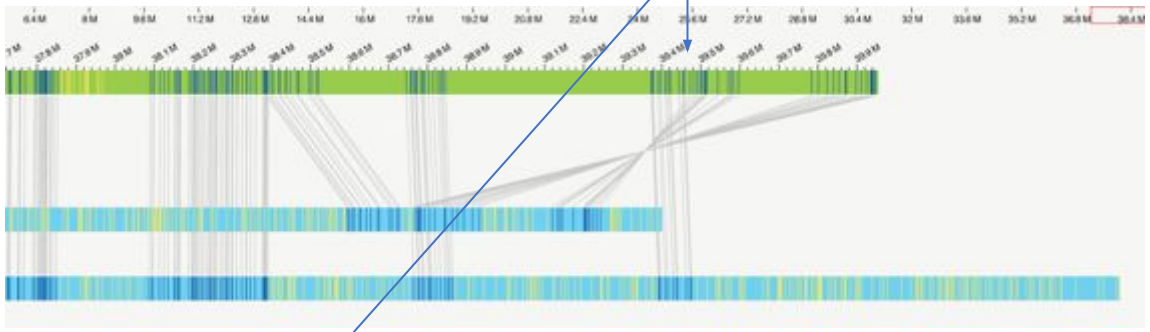
fNotCel1_2 – scaffold32

Misassembly due to haplotypes (most prevalent at chromosome ends)



Solution = B1+, A+, B2+, F-, C+, E-, D+

Mis-join

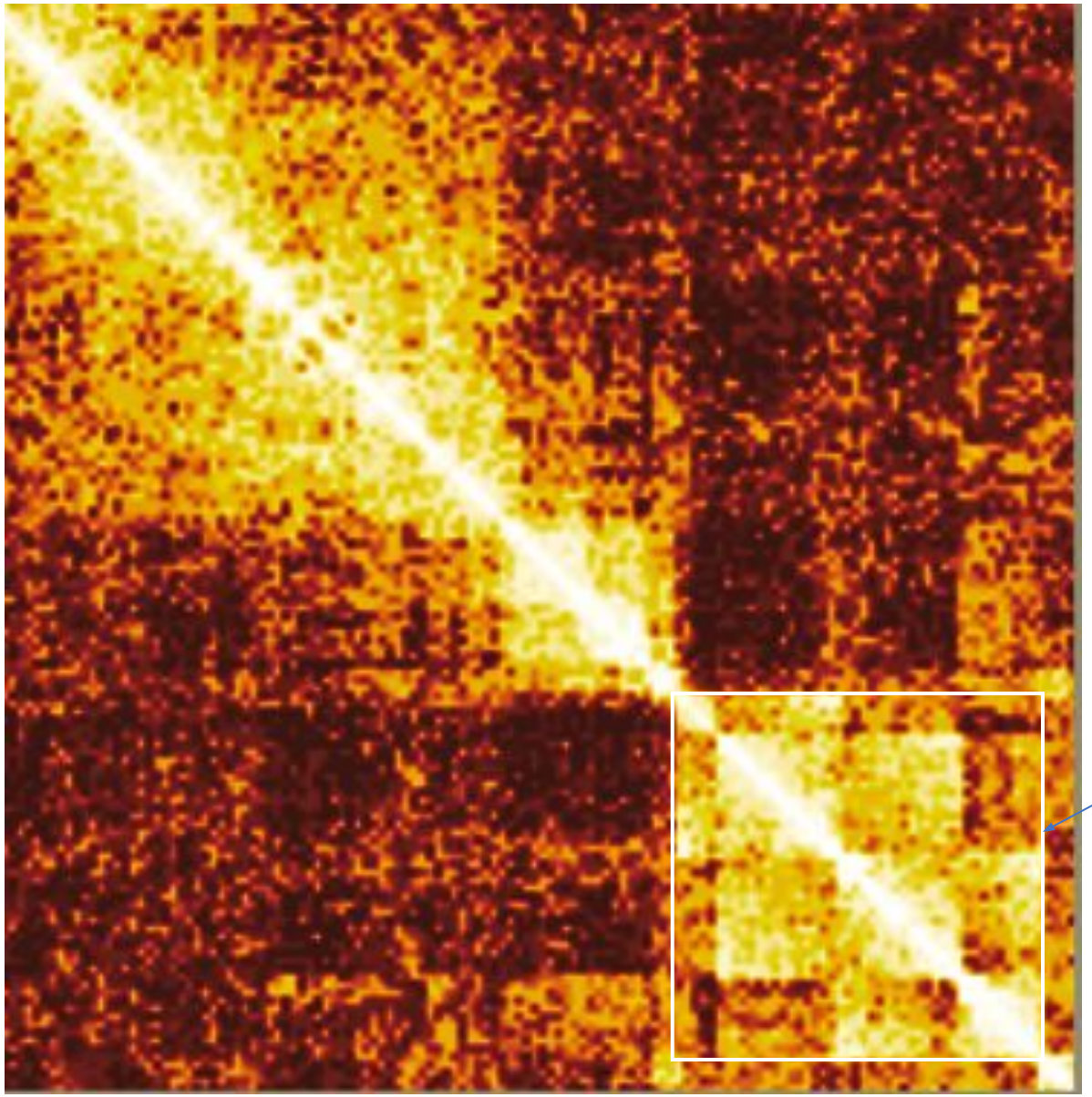


Bionano map above shows chromosome end represented by 2 map contigs which themselves represent different haplotypes.

Where Bionano maps represent haplotypes in a significantly different manner, separate scaffolds with large gaps can arise as the assembly forks. Additionally, in this case, part of the scaffolding has gone wrong (a misjoin between D and E) and the chromosome end is in the wrong orientation. Resolving these requires a lot of manual interventions and can be time-consuming.

fNotCel1_2 scaffold2

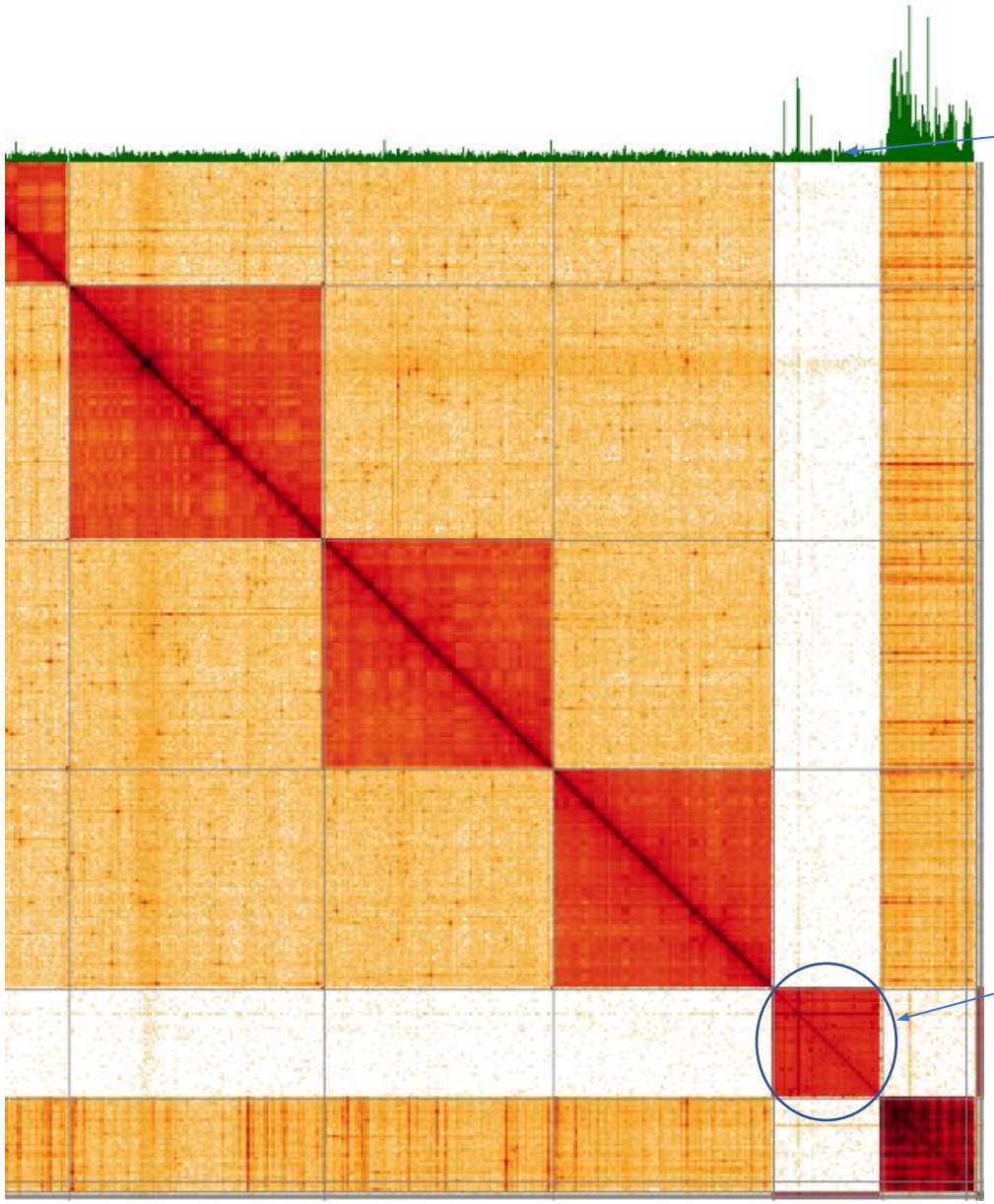
Pretext user error – components in right order, but wrong orientation



4 consecutive sequence blocks all in the wrong orientation.
(this particular error resulted from incorrect manual editing of the assembly tpf)

bGeoTri1 superscaffold28

Contamination



coverage does not suggest a sex chromosome – it is broadly consistent with autosomes

scaffold_32 stands out like a sore thumb due to a very low level of HiC affinity with the rest of the genome. Heterogametic sex chromosomes can sometimes look a bit like this, but they would have half coverage – here we see coverage at autosomal levels or even slightly higher.

Our assemblies have contamination removed before they enter curation, but in this case, a contaminant scaffold made it into our assembly. We were able to identify this as contamination with a blast search.

scaffold32 has below background association with rest of the genome

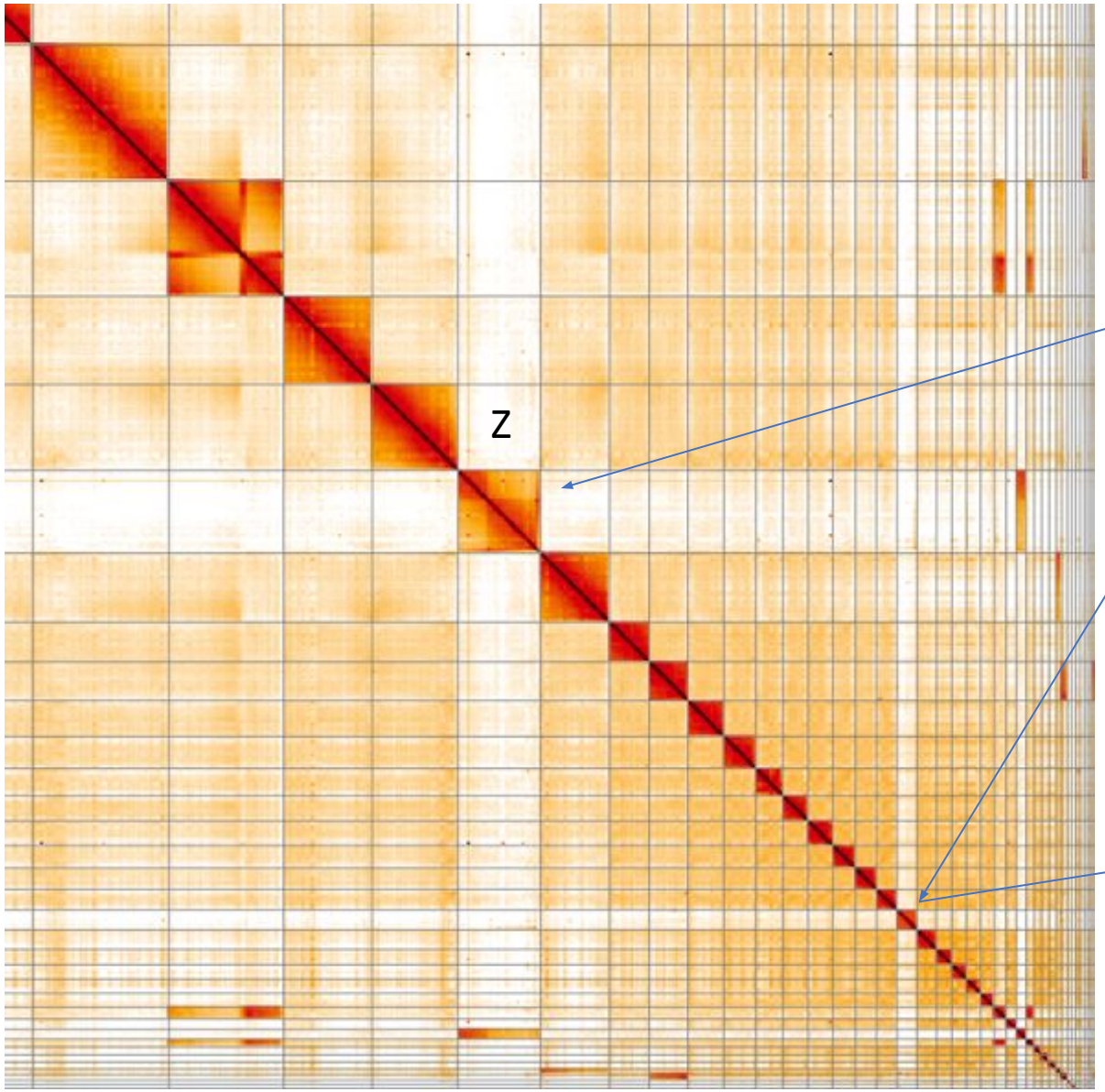
ilColCroc2_1 scaffold32

Biology

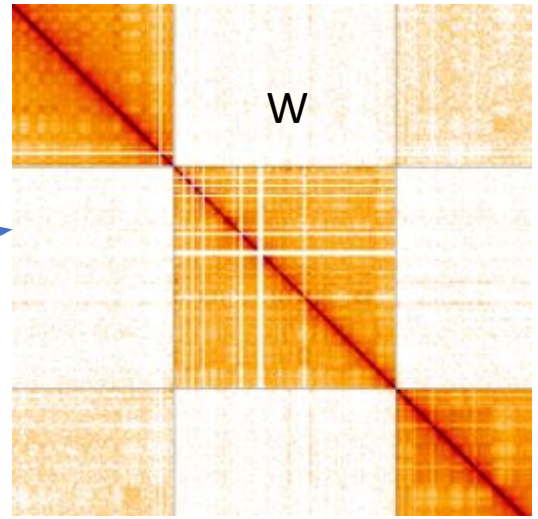
Includes:

- Sex chromosomes
- Mapping to sex-specific genome with different sex reads
- Structural differences between haplotypes
- Centromeres
- Satellite repeat
- Low complexity sequence causing HiC ambiguity
- Chromosomal rearrangements
- Haplotypes
- Short arm repeats
- Whole-genome duplication

Sex chromosomes



Sex chromosomes have half depth coverage which means they have half the level of background HiC signal so stand out as being pale in a HiC map. They (especially the more heterochromatic W and Y) have more gaps than autosomes due to lower coverage. Due to these distinctive properties, sex chromosome pieces usually stand out clearly in the shrapnel, even if their order and orientation can't always be deduced.

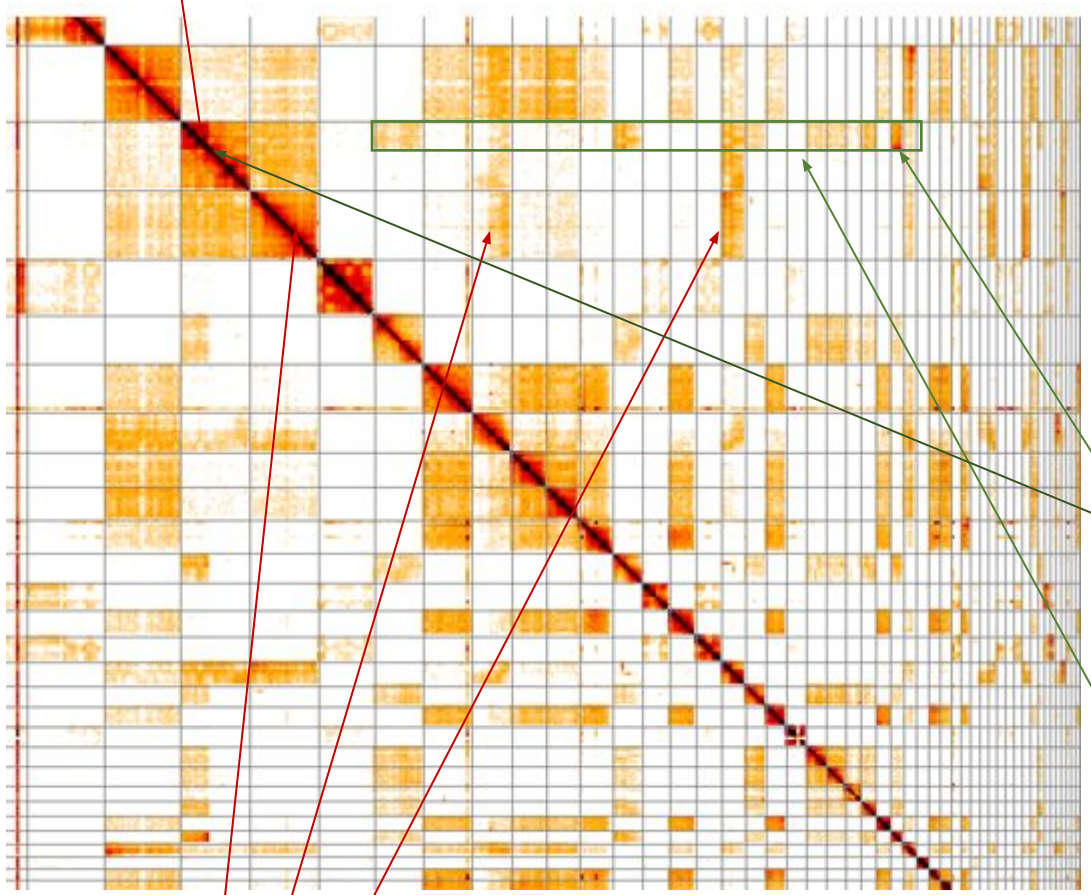


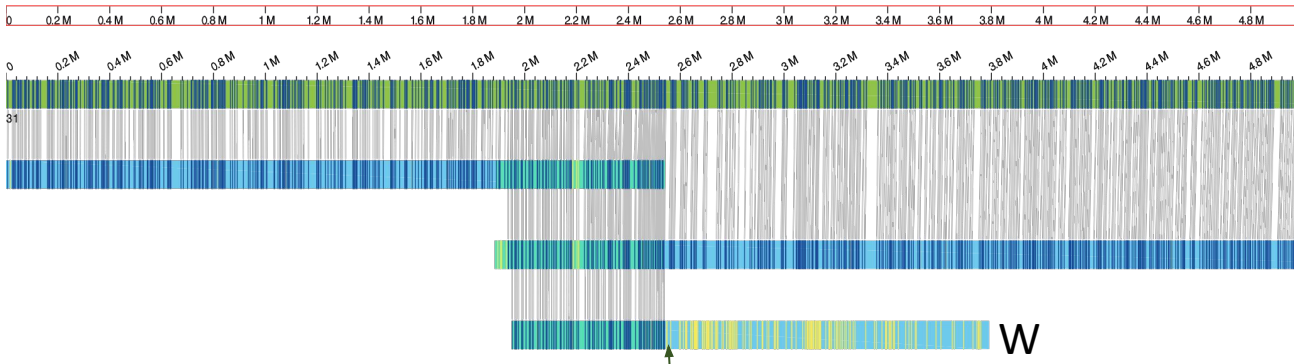
Many white stripes due to many gaps in W chromosome

bGeoTri1

PAR junctions into sex-specific regions (Z/W or X/Y)

PAR junction into Z

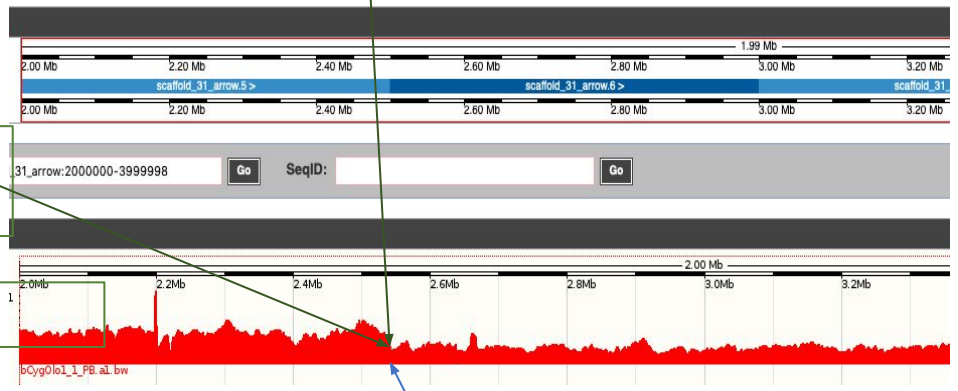




Divergent Bionano map shows the Z map aligned with the Z sequence, with the W chromosome map diverging from the PAR onwards.

contig with W-specific junction into PAR

other W-specific pieces



Some Z-specific pieces

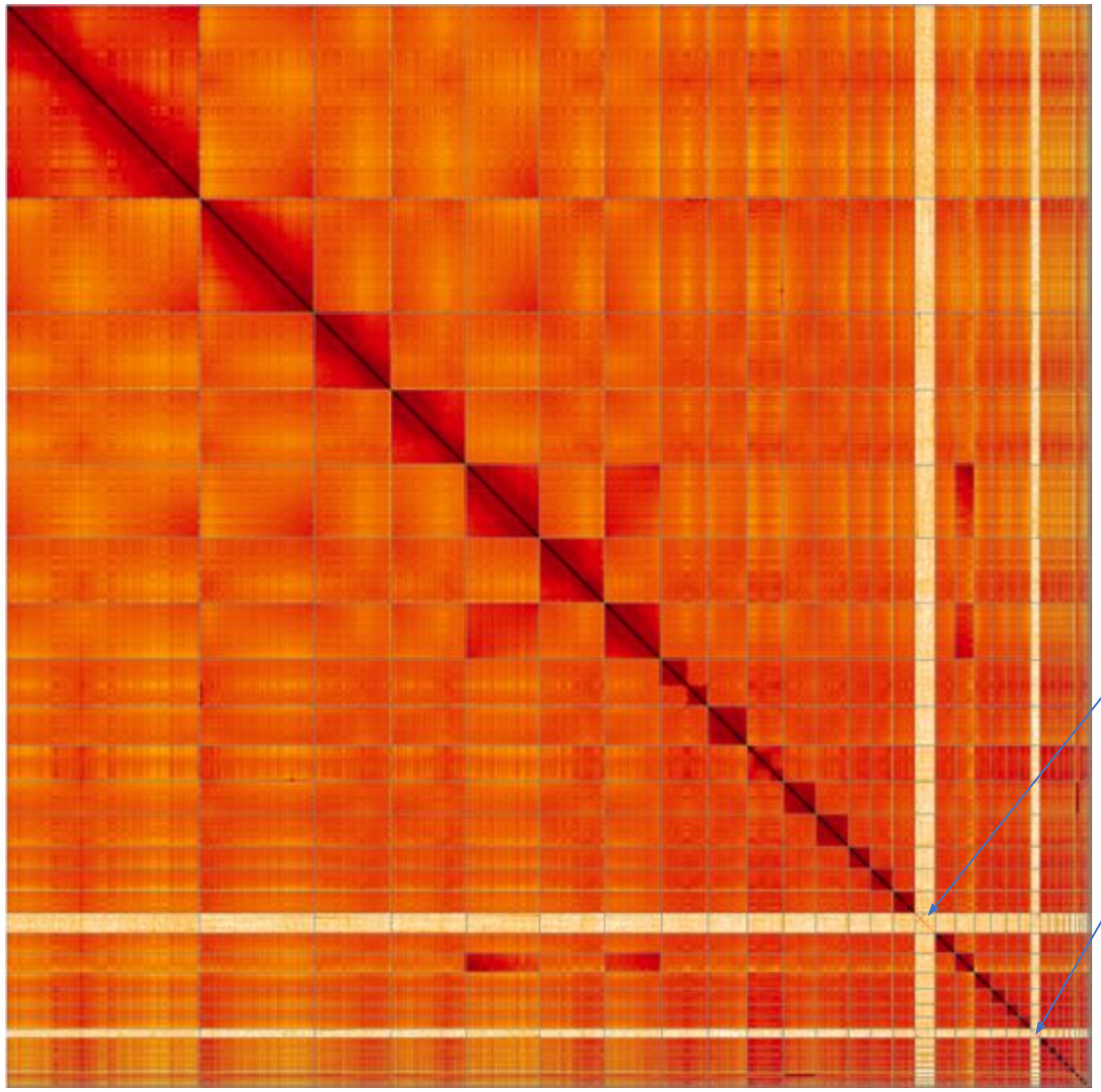
Zero affinity between Z-specific and W-specific scaffolds.

Coverage halves at junction from PAR into Z

(bCygOlo1, scaffold_31_arrow.5 = PAR->Z, scaffold_53_arrow.2 = PAR->W. BioNano anchor 31)

http://vgp-geval.sanger.ac.uk/bCygOlo1_1/Share/6cb6736b9786a211d3799679bb69378a260095
https://vgp-geval.sanger.ac.uk/bCygOlo1_1/Location/View?r=scaffold_31_arrow:2030817-3030814

Female genome visualized using only male HiC reads

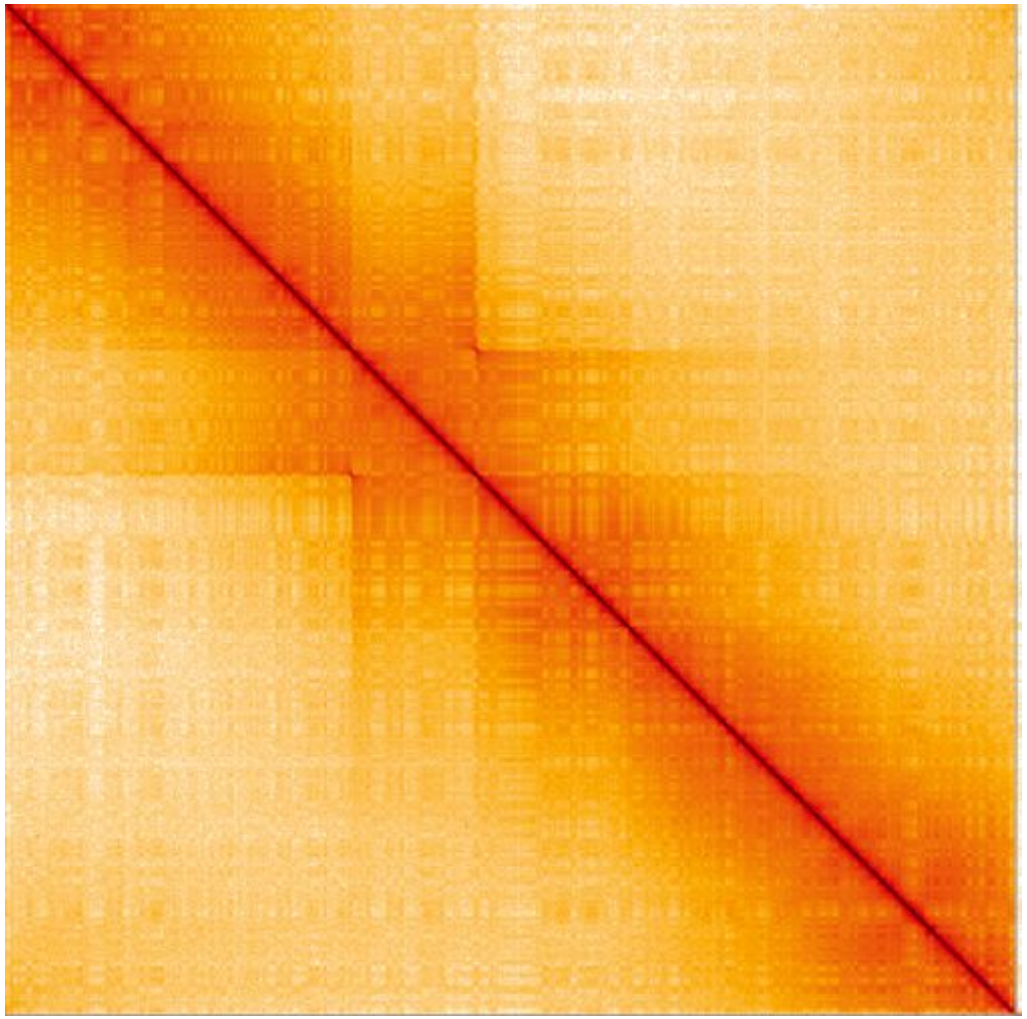


2 scaffolds here have no signal – these are W scaffolds with zero male reads mapping to them (as would be expected)

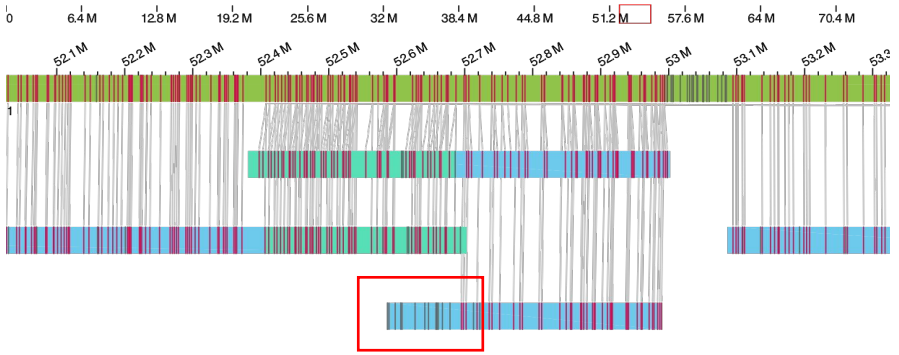
bPhoRub1

HiC inversion between sister chromatids 1

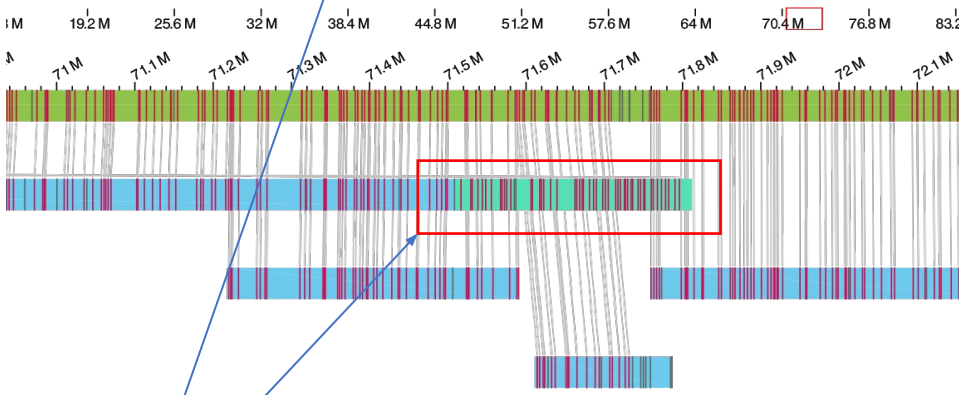
bGeoTri1_1 - Largest chromosome (153Mb). Portion of sister chromatids inverted (52.7-71.7Mb), therefore HiC always looks correct (from the centre diagonal) and incorrect (off-diagonal signal) whichever orientation. Assumed would separate correctly and not be seen in trio assembly HiC map.



Left – Bionano maps agreeing and diverging



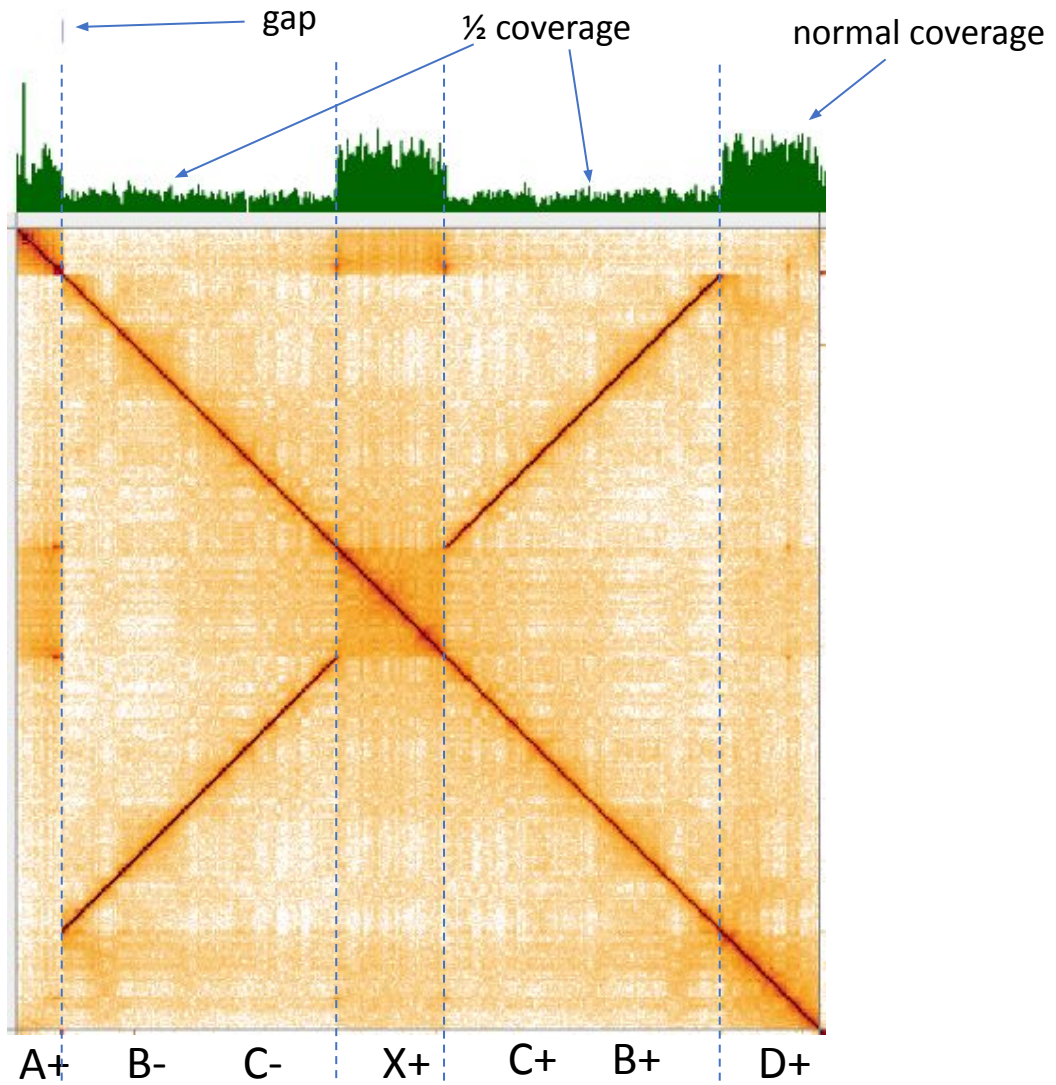
Right – Bionano maps agreeing and diverging



Divergent Bionano maps from the other sister chromatid

HiC inversion between sister chromatids 2

(multi-mapping reads are not filtered in this HiC map)



An inversion between sister chromatids can also present like so. We would assume that this 2nd presentation is likely due to elevated heterozygosity and subsequent failed purging.

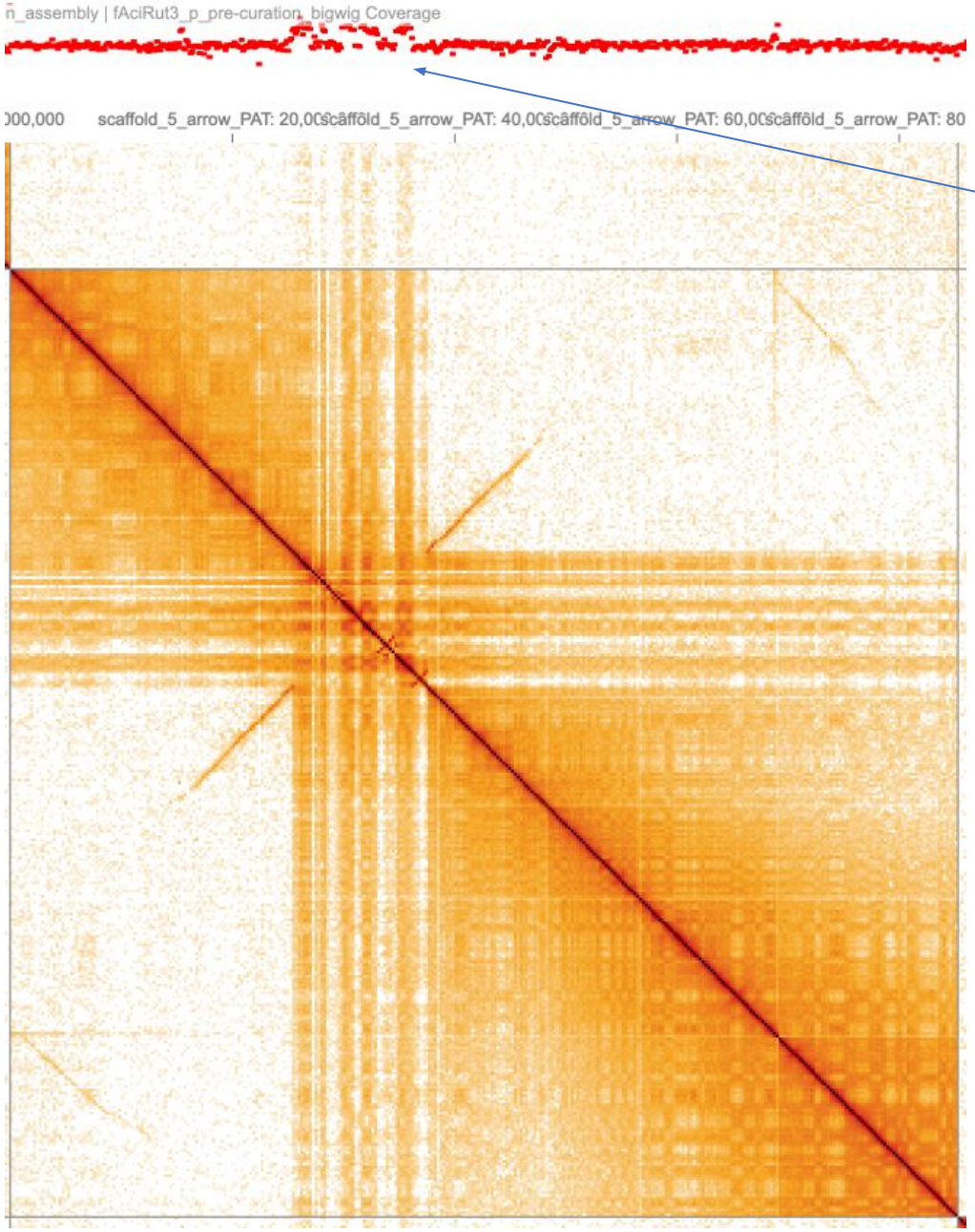
The map shows 2 solutions representing each haplotype. Note that A can't join to B hence the presence of a gap. A must always join to X (but in 2 orientations) hence the presence of 2 contact points between A and L/R of X. The only region that is actually inverted between the haplotypes is X

haplotype 1 = A+ $\overrightarrow{X+}$ C+ B+ D+

haplotype 2 = A+ $\overleftarrow{X-}$ C+ B+ D+

fThuMac1_1 scaffold_1 6.8mb retained haplotypic duplication

Chromosomal rearrangements (possibly due to partial tetraploidy) confounding correct assembly

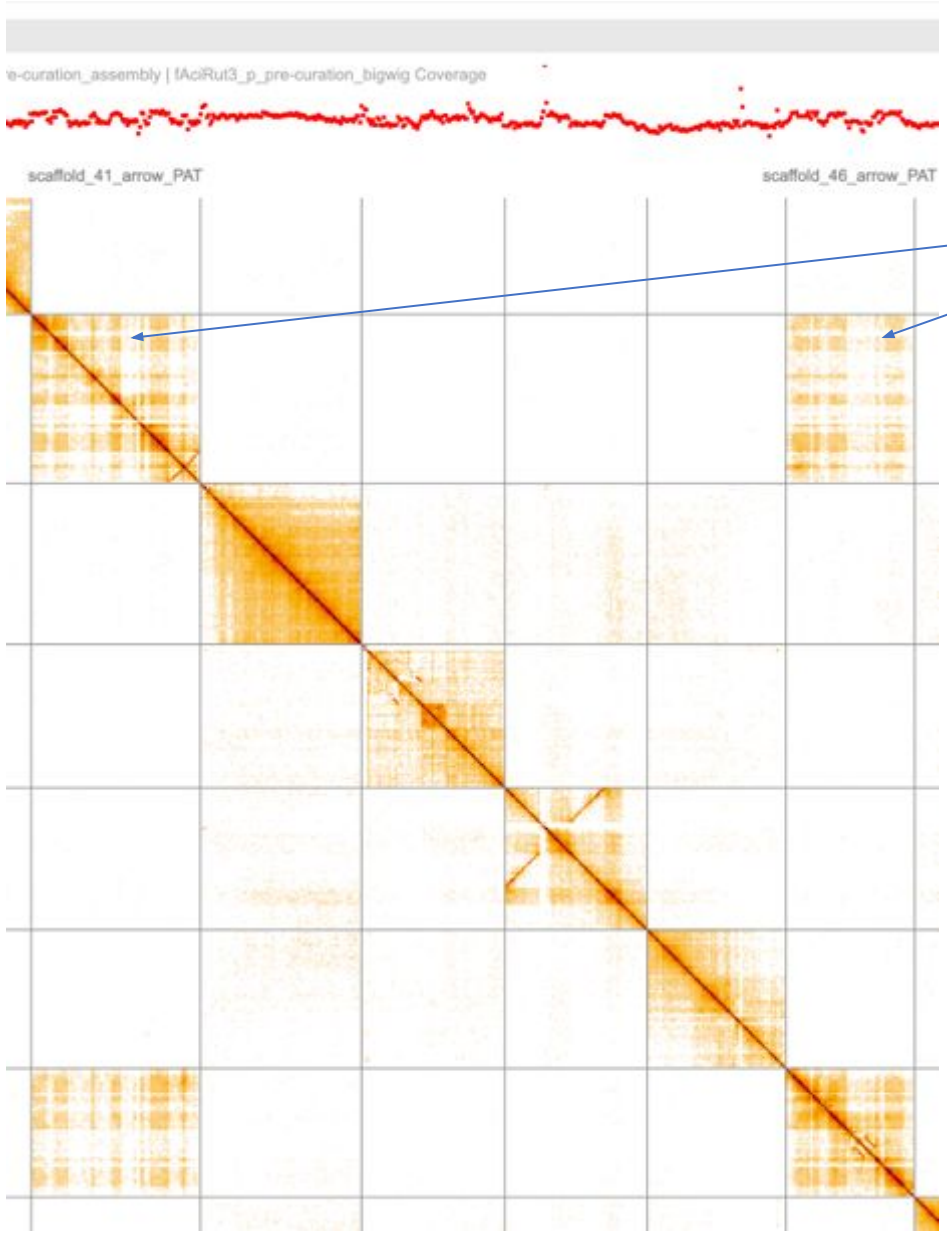


double depth region where assembly is currently collapsed. For a correct representation of the genome, this region would have to be duplicated as the same sequence exists on 2 different chromosomes.

This should be assembled into 2 chrs. The region that is shared should be split so that copies can be placed on both chrs. The HiC duplication signal is indicative of a chromosomal duplication event – a proportion of reads now map to the wrong repeat copies and so an association can be seen between the duplicated regions which currently exist in opposing orientations.

fAciRut3_2_PAT scaffold5

Chromosomal rearrangements confounding correct assembly continued…

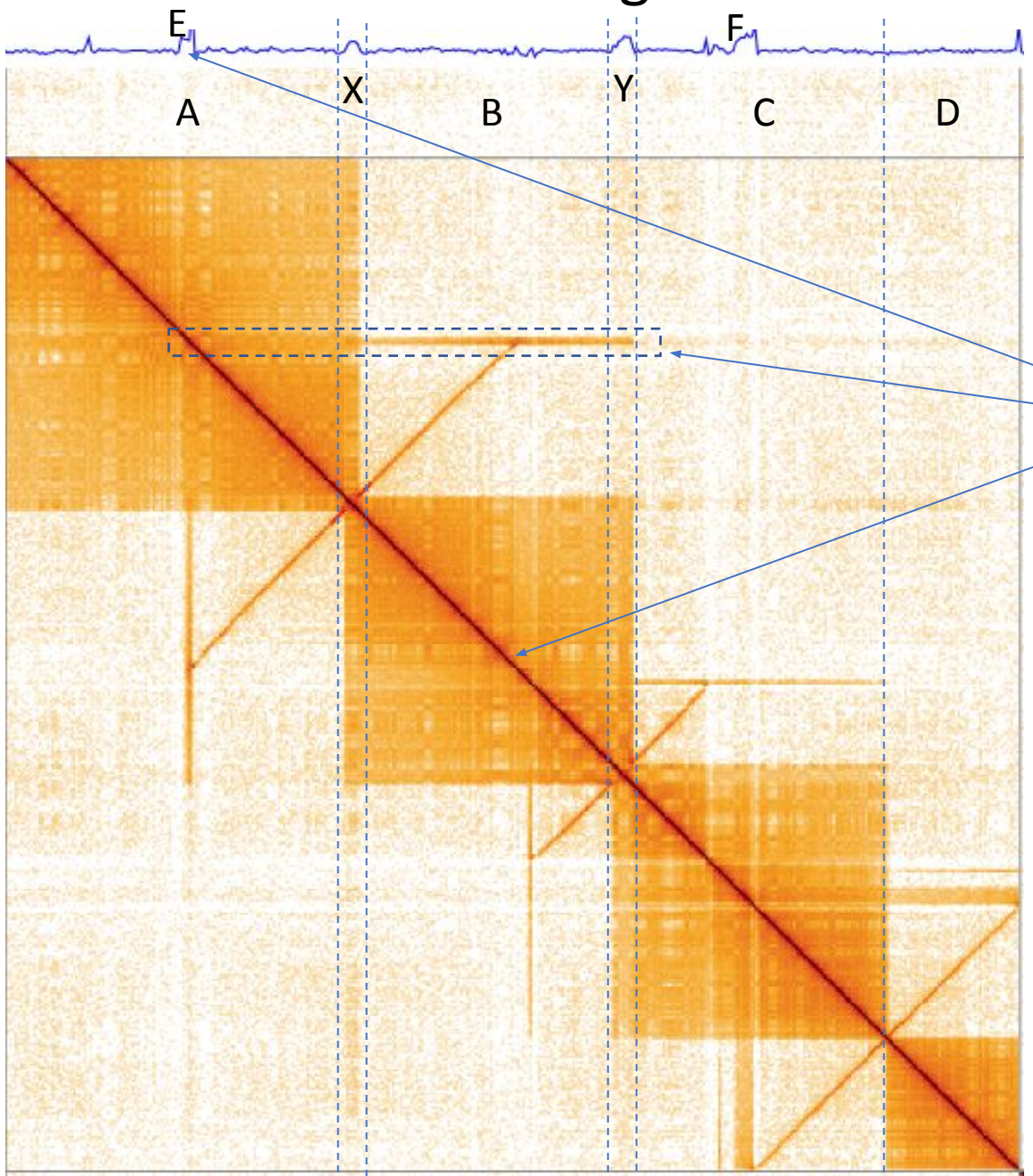


2 scaffolds are actually a jumble of 2 different chromosomes, evidenced by the coverage plot which switches between normal coverage and double coverage (red histogram)

There is so much missassembly, and collapsing of data that it has not been possible to tease these apart

fAciRut3_2_PAT scaffold5

Chromosomal rearrangements confounding correct assembly complex scenario 1



Here a single scaffold actually contains 4 chromosomal pieces (A-D). Collapsed sequences X and Y would need to be duplicated in order to resolve this fully. Chromosome B would be composed entirely of sequence found in chromosomes 1 and 3.

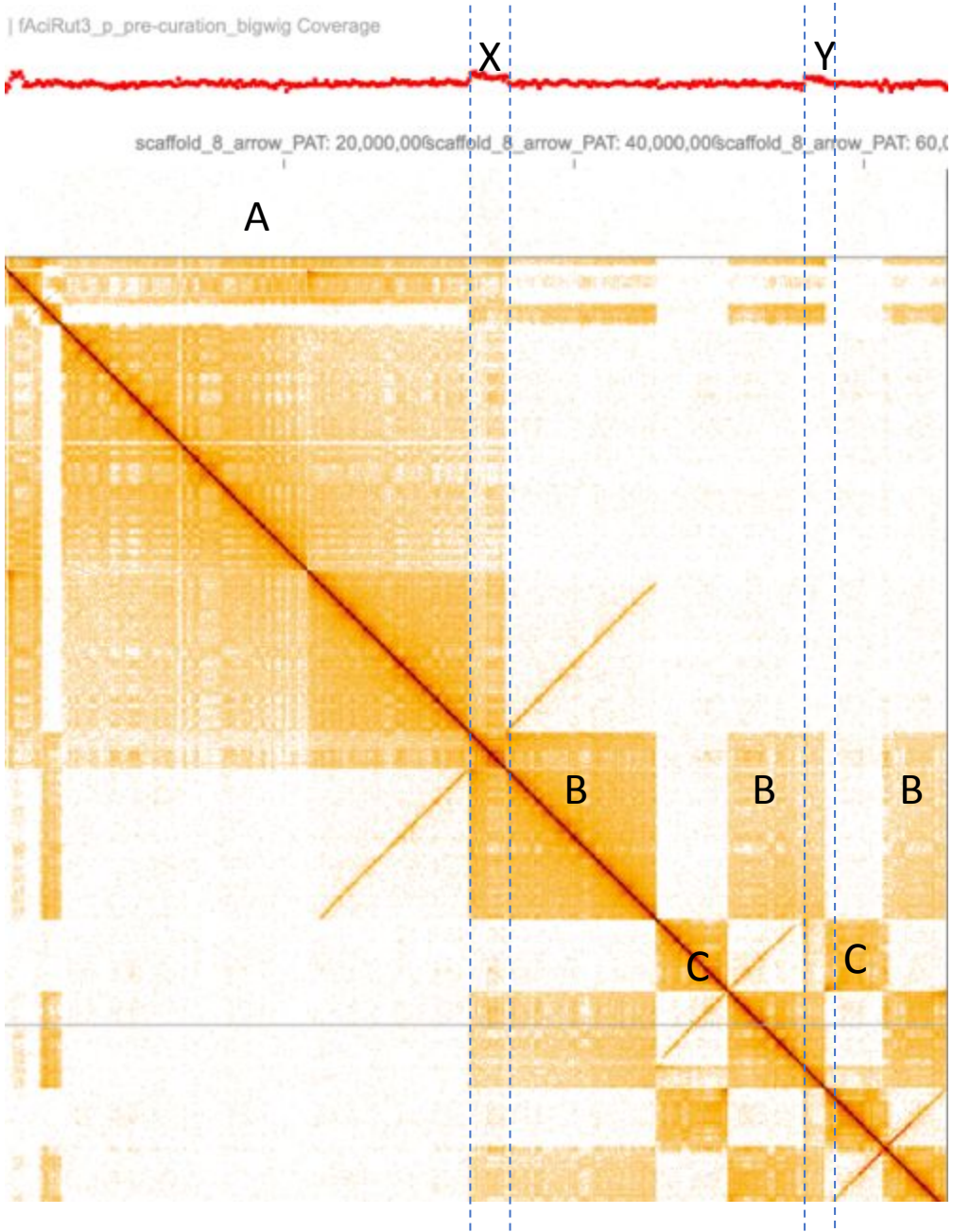
Line due to subtelomeric repeat which needs inserting here – currently collapsed into A.

= Break needed

Chr1 = A+X
Chr2 = X+B1+E+B2+Y
Chr3 = Y+C
Chr4 = D+F-

fAciRut3_m scaffold4

Chromosomal rearrangements confounding correct assembly complex scenario 2



Multiple misassembles within the same scaffold, this separates into 3 Chromosomes (A-C).

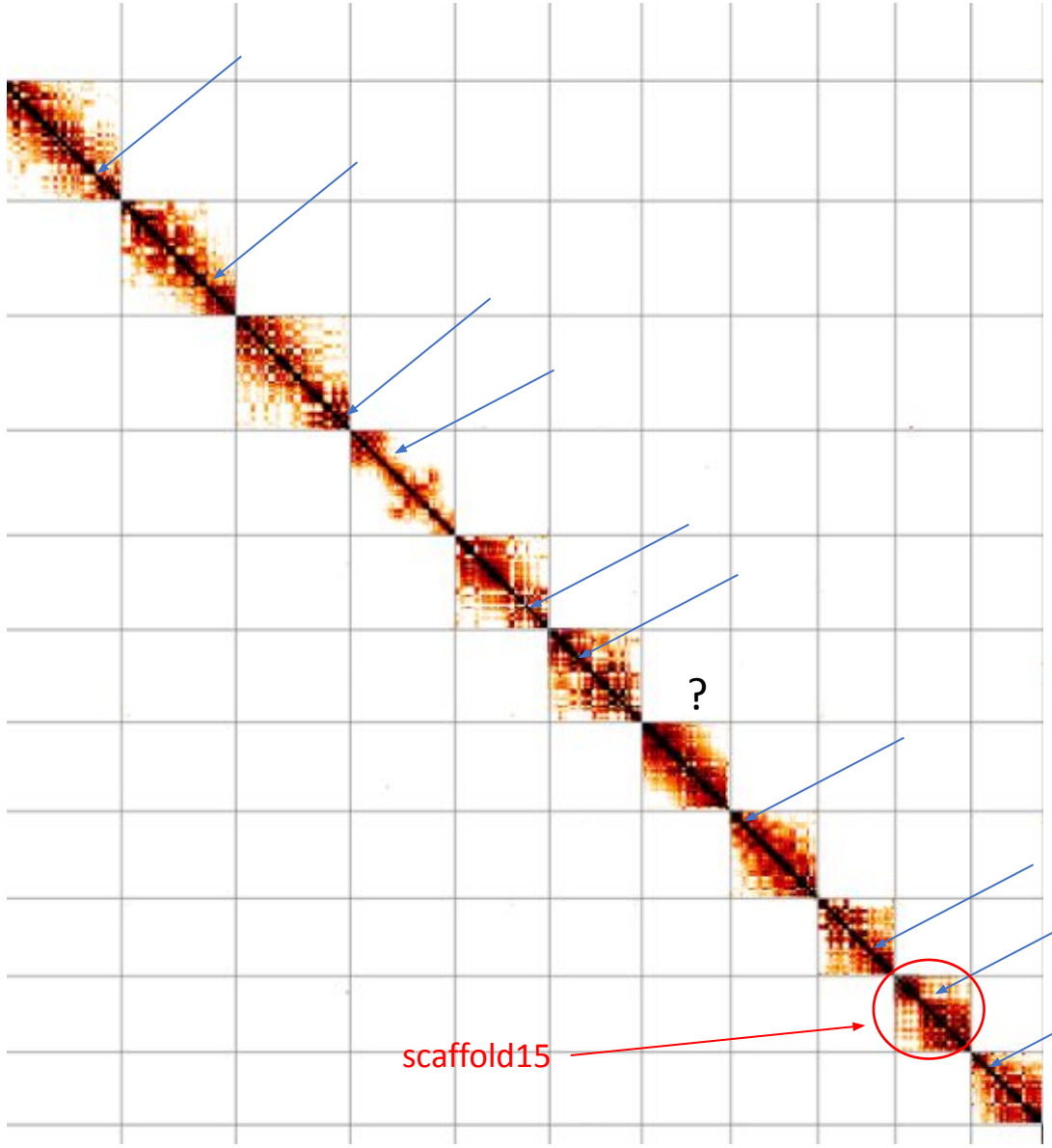
2 regions of collapsed duplication can be seen from read coverage (X and Y) and these have a darker area of contact in addition to the PacBio reads stacking.

Part of chrm B is contained in chrm A (ie they share sequence) and chrm C is entirely composed of elements found in chrm B.

= Break needed

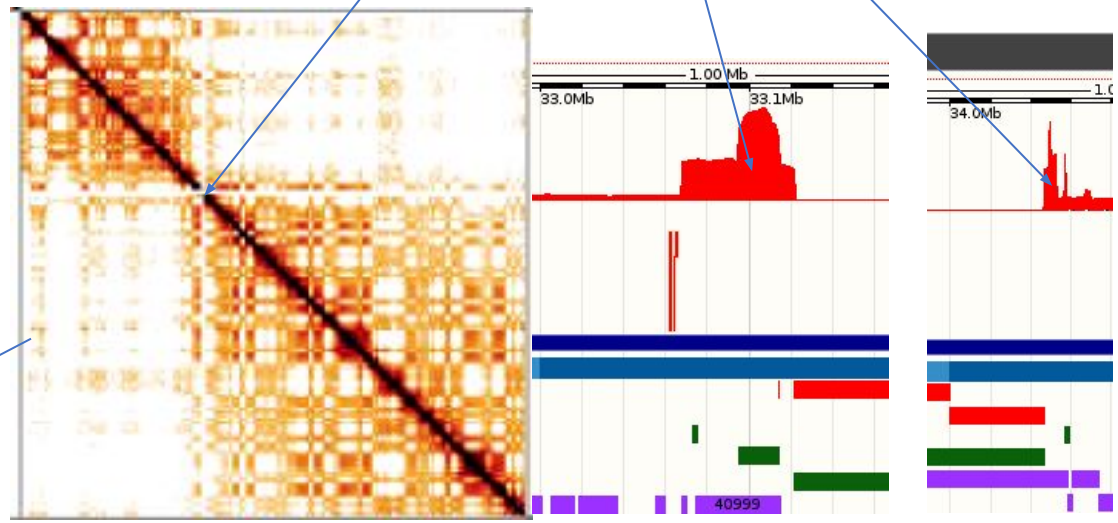
fAciRut3_p scaffold8

Centromeres1



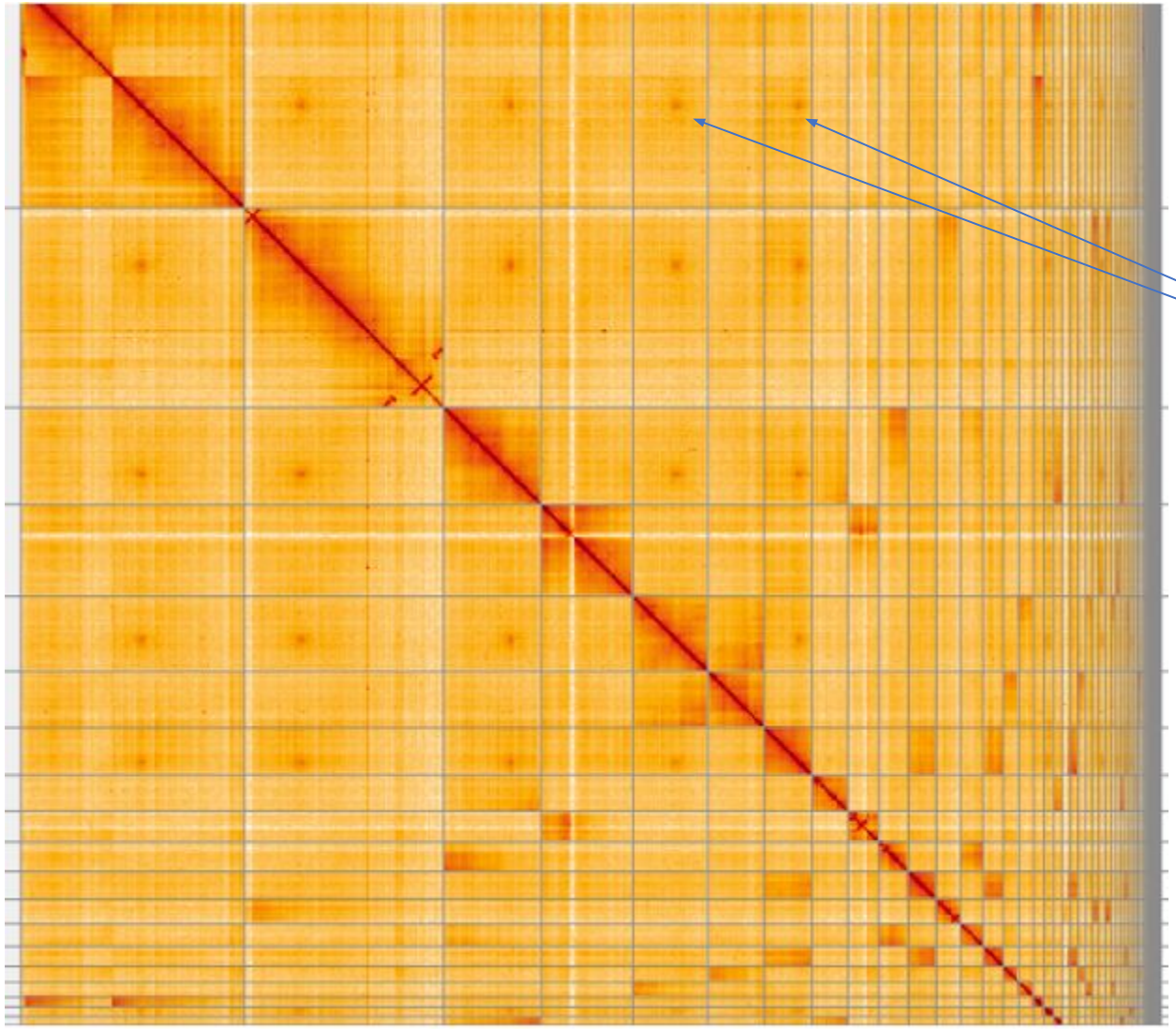
Centromeres and therefore the separation between the p- and q-arm are often visible in the HiC maps as a separation of contact. This is not always obvious to see in the HiC map. Below centromeres have been marked on the plot for a subset of chromosomes. They are most obvious in the HiC data when they're near metacentric and less obvious when near telocentric.

Zoom in on centromere in scaffold15. We can see extreme coverage (these repeats are typically the largest single repeat type in the genome)



mZalCal1 – scaffolds6-16

Centromeres2



iHerIll2

Centromeres have been observed to be highlighted by “hot-spotting” as in these (and all the other) cases in this image.

(other issues can also be seen in this map; misassembly, joins needed and haplotypic duplications)

Centromeres 3 - the importance of additional data sources to aid interpretation

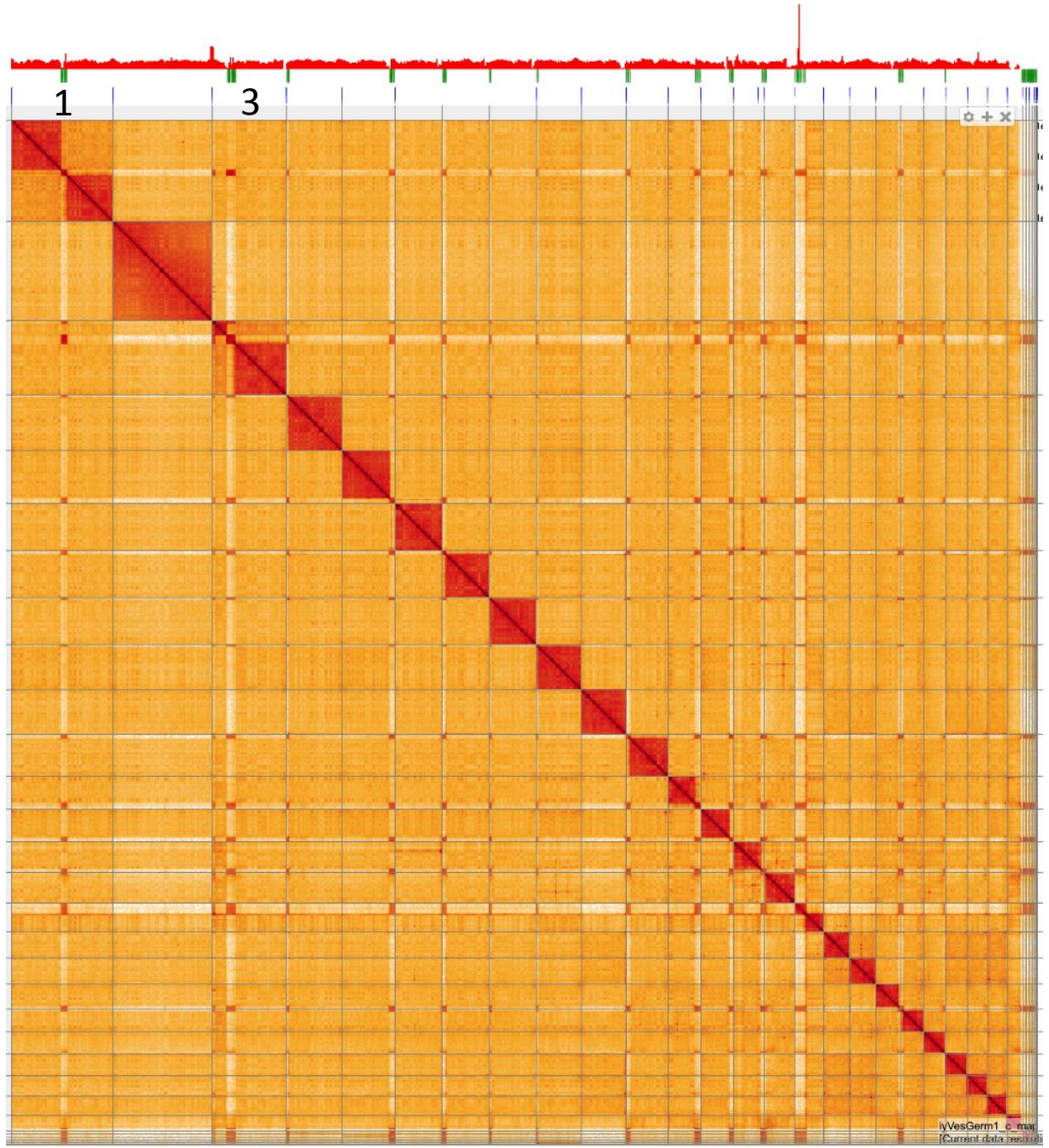
2 weakly associating (via HiC) scaffold pairs (1 and 3 in the image)

Genomic repeat analysis revealed candidate sequences for telomeres and centromeres

This observation allowed the weakly associated scaffold pairs to be confidently joined on centromere (the centromeric repeat explaining the weak association)

Coverage noticeably dips in the centromeres

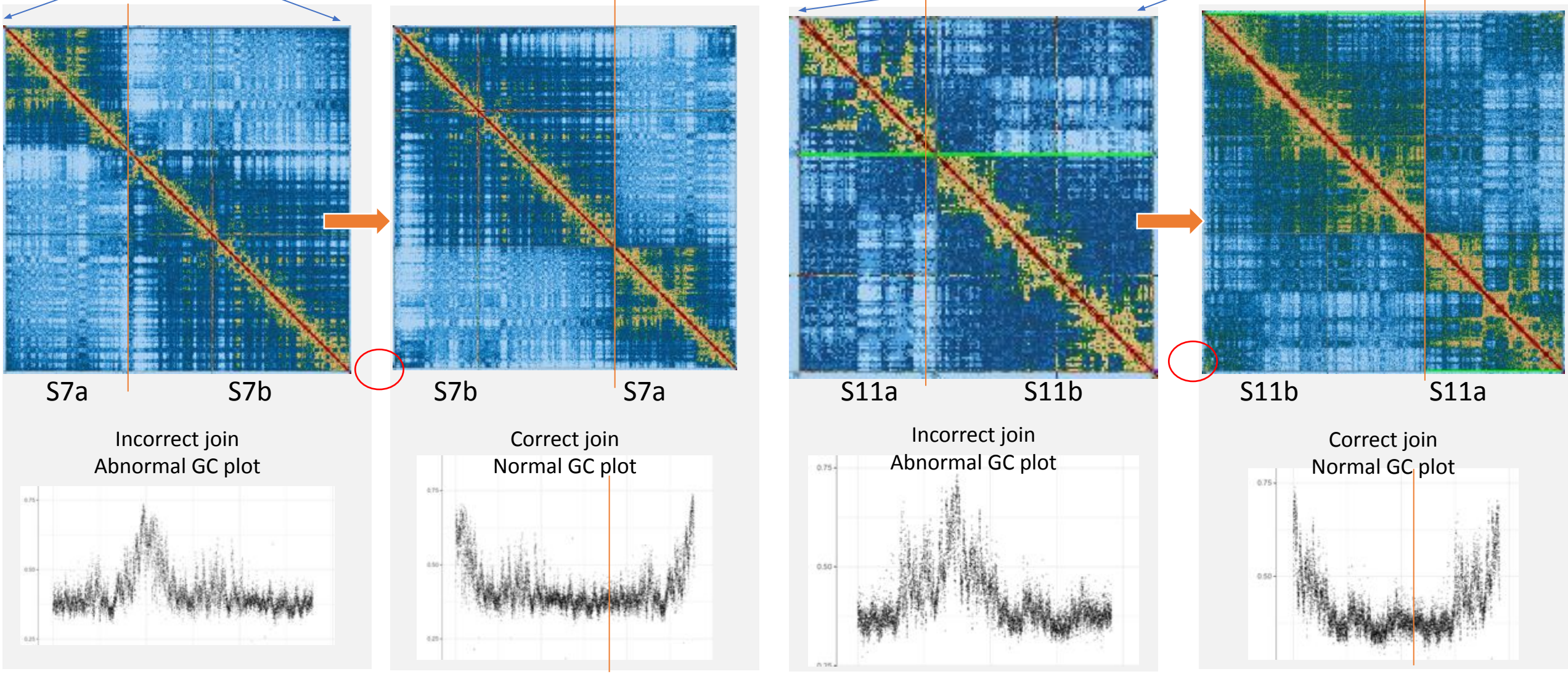
iyVesGerm1_1



Centromeres can cause ambiguous HiC

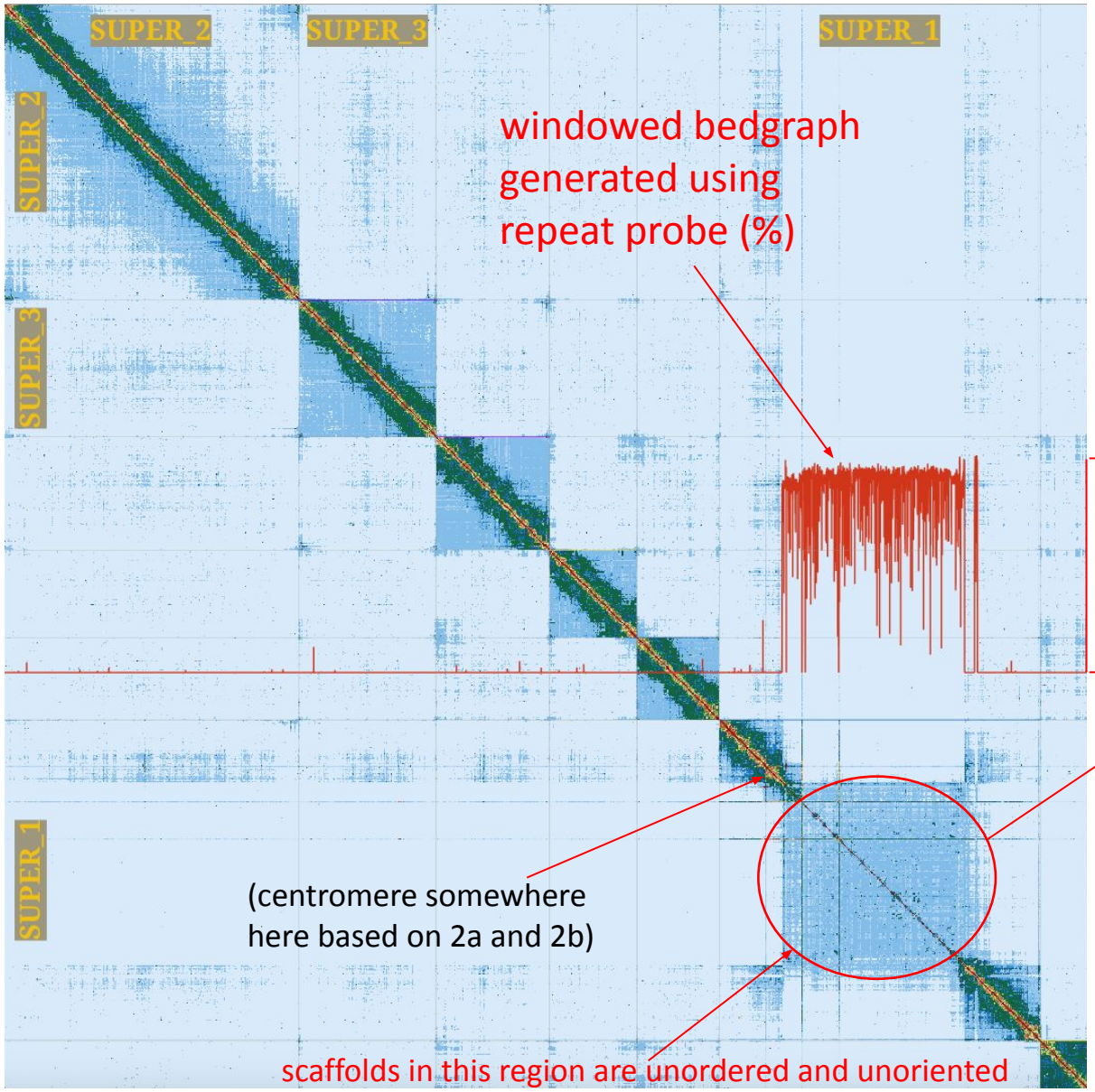
460bp unit tandem repeat sequence in common (likely centromeric)

250bp unit tandem repeat sequence in common (likely centromeric)



In the absence of the huge centromeric sequence (and it's associated signal), the join can incorrectly be made on the 2nd best signal (from the subtelomeres) highlighted by the red circles. The lesson to learn from this is that the HiC maps don't look especially wrong in the incorrect assembly and don't look particularly correct in the fixed assembly

Massive satellite repeat (not centromere)



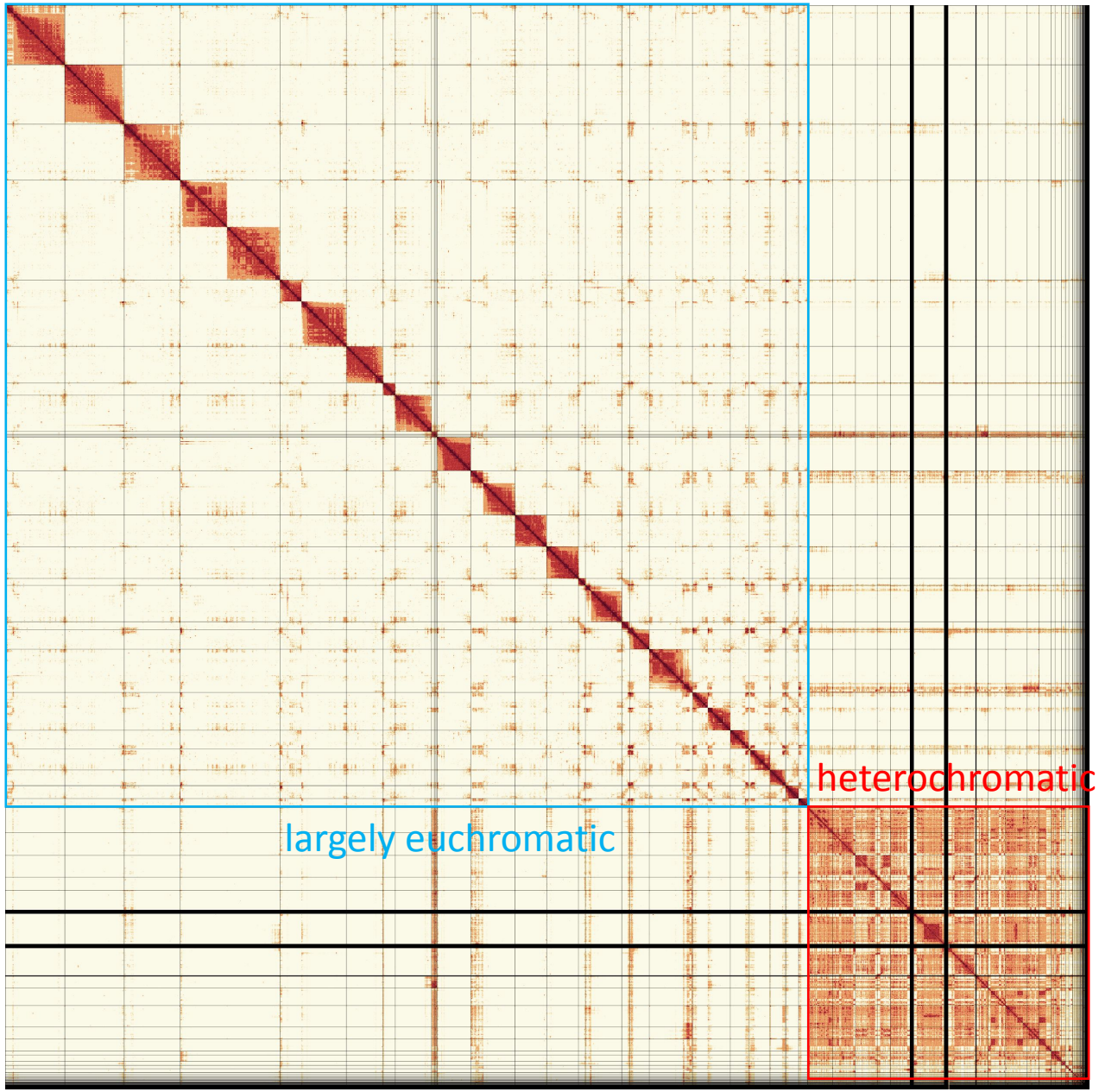
Map produced with multi-mapping reads turned on. This gives equal signal across the whole repeat. Note the association in the corners between euchromatic chromosome ends

icRhaFulv1_1

Contrast between euchromatic and heterochromatic portion of the genome

Non-repetitive HiC signal can be seen for 26 chromosomal entities, in stark contrast to the heterochromatic portion of the genome (centromeric and short-arm sequences which in the case of this wasp do not have enough specific association with a particular chromosome to enable them to by placed.

iyNysSpin1_1

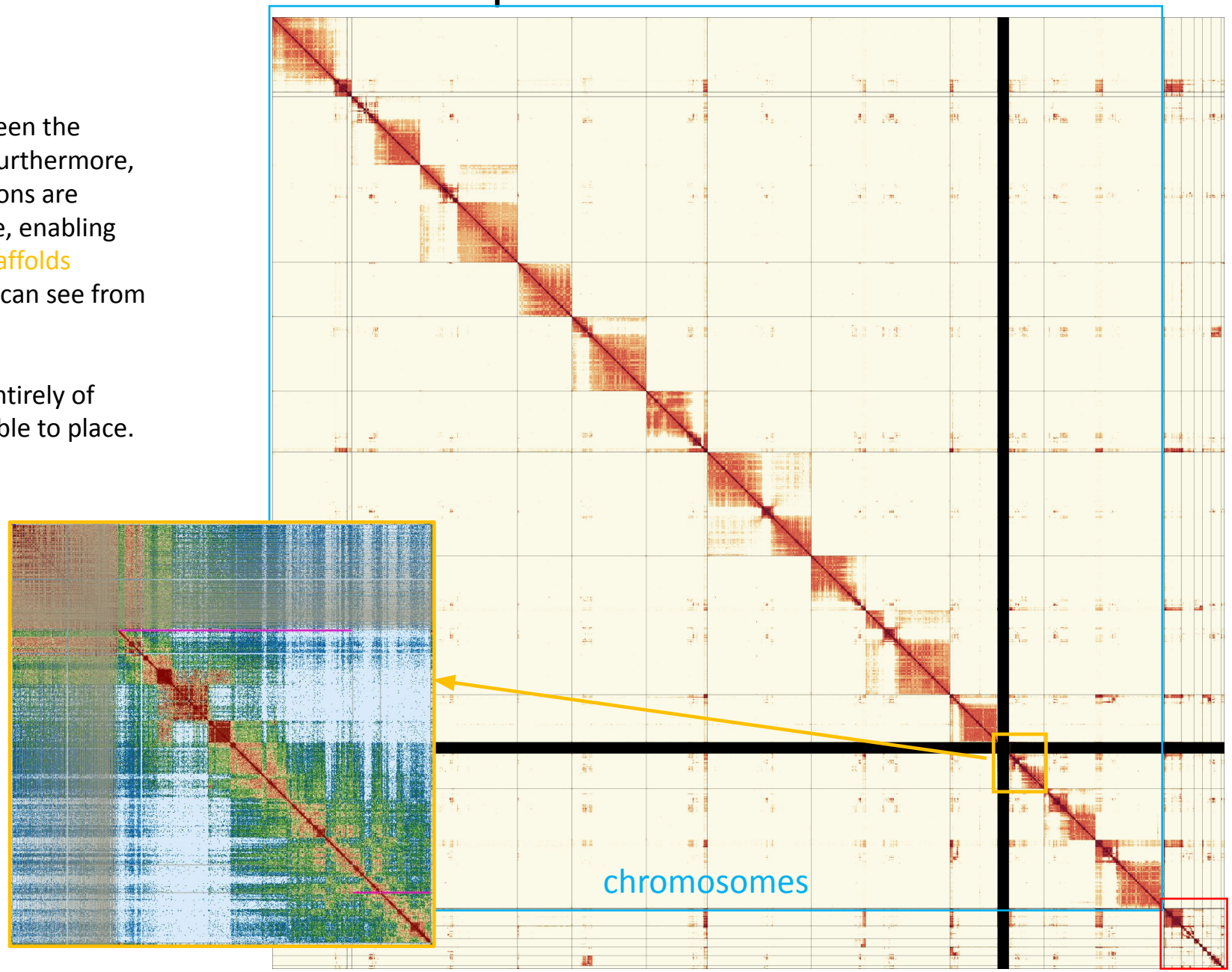


Satellite sequence has weak association with particular chromosome

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.

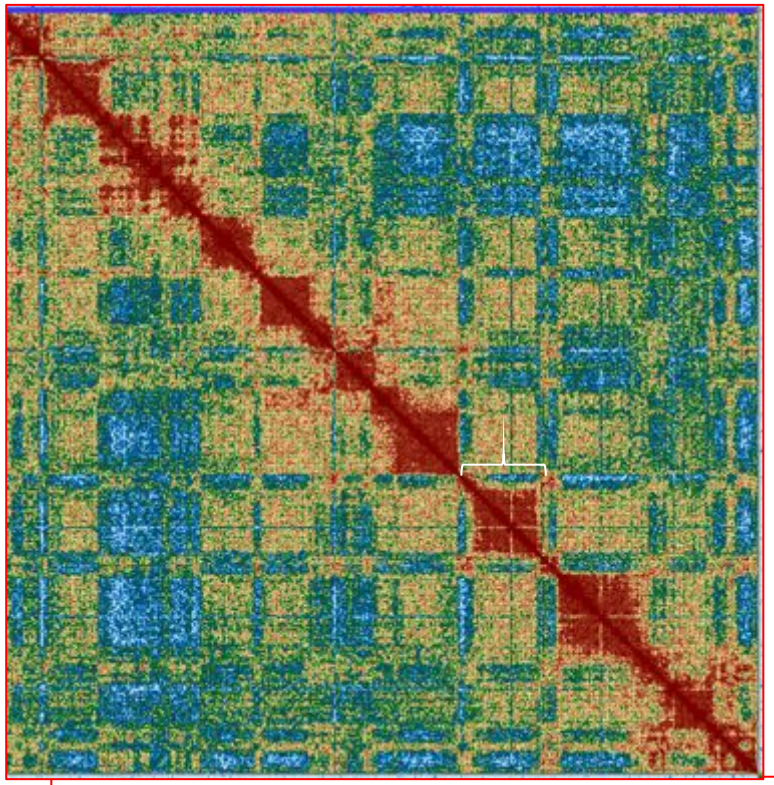
There remain several scaffolds composed entirely of satellite sequence which we have been unable to place.

iyEctLitu1_1

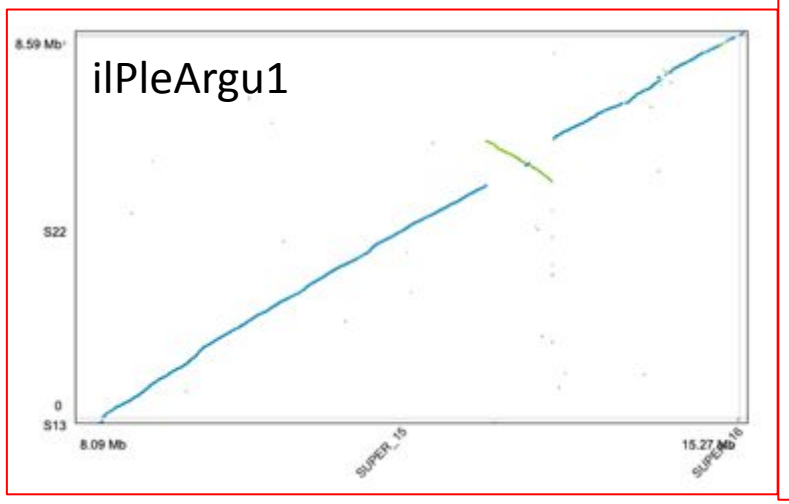


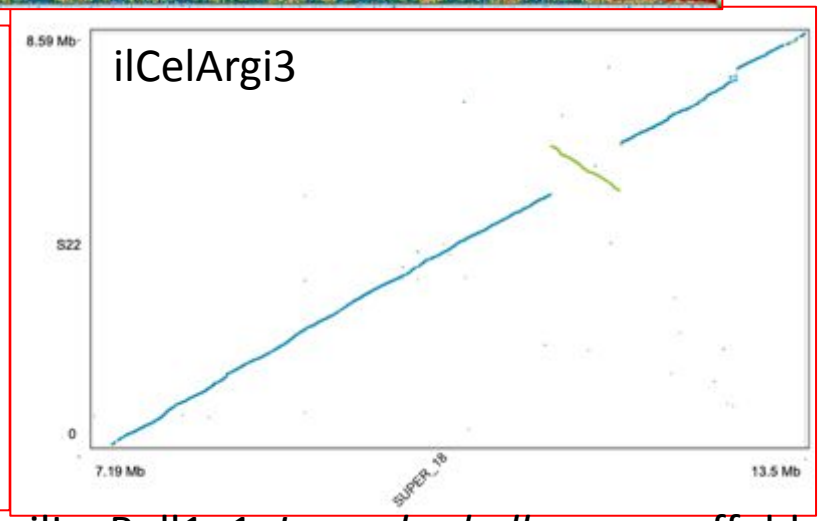
Lepidopteran inversion hidden in HiC map but visible by comparison with 3 related species

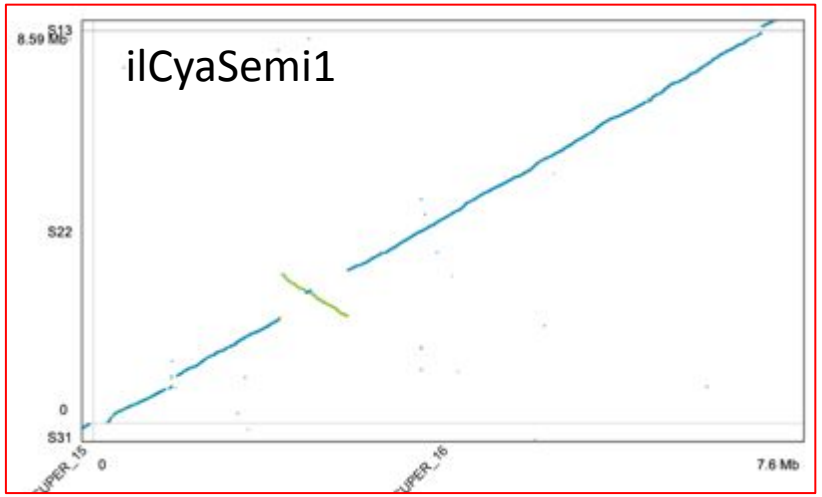
Before flipping
scaff 113 and 227



boundaries of this and similar regions are often low complexity and very long



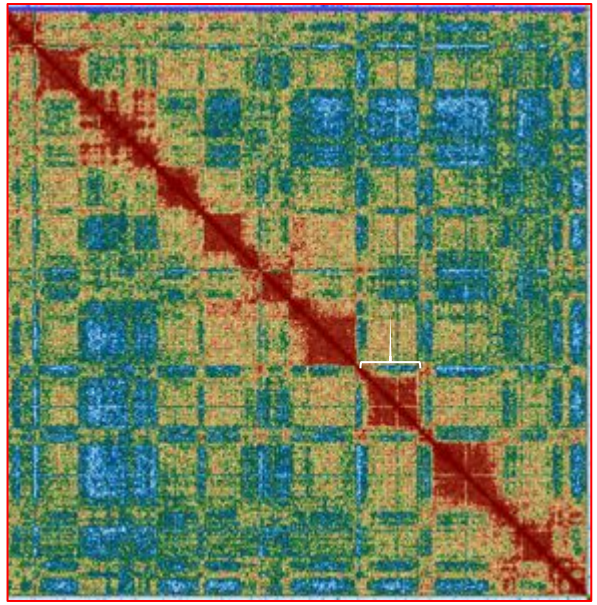




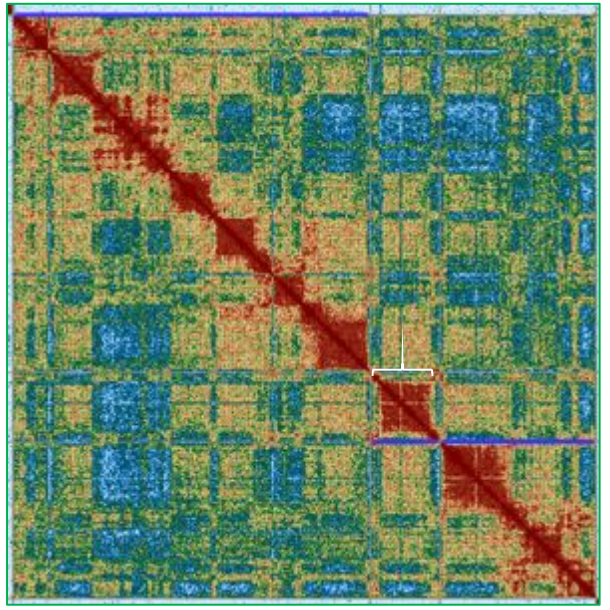
ilLysBell1_1 *Lysandra bellargus* scaffold_22

Lepidopteran inversion fixed – **HiC map looks almost identical between correct and incorrect**

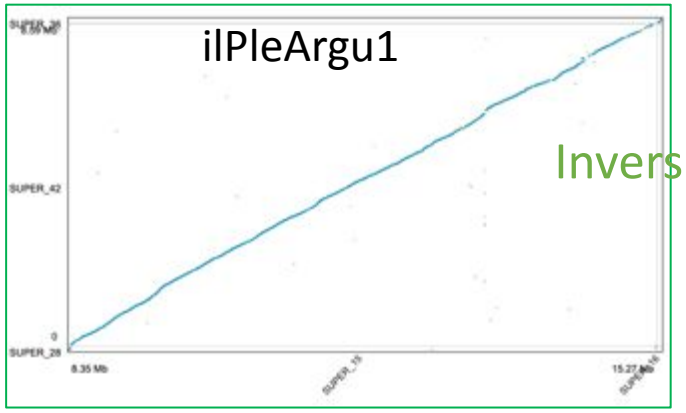
Before flipping scaff 113 and 227

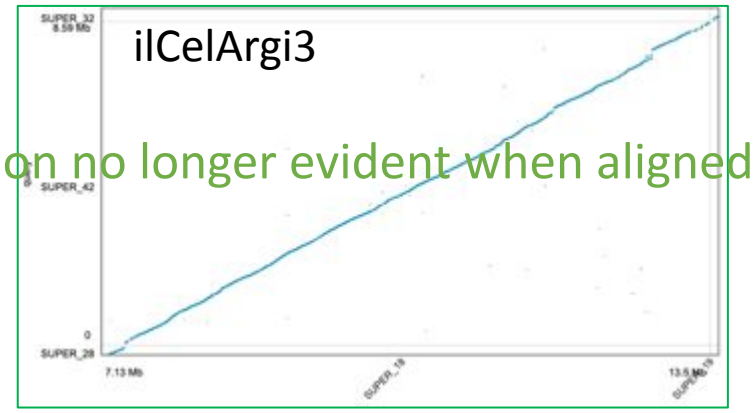


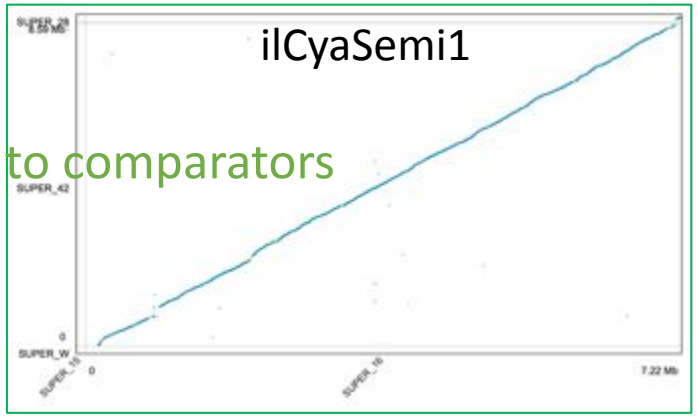
Correct vs incorrect HiC map – almost identical



After flipping scaff 113 and 227



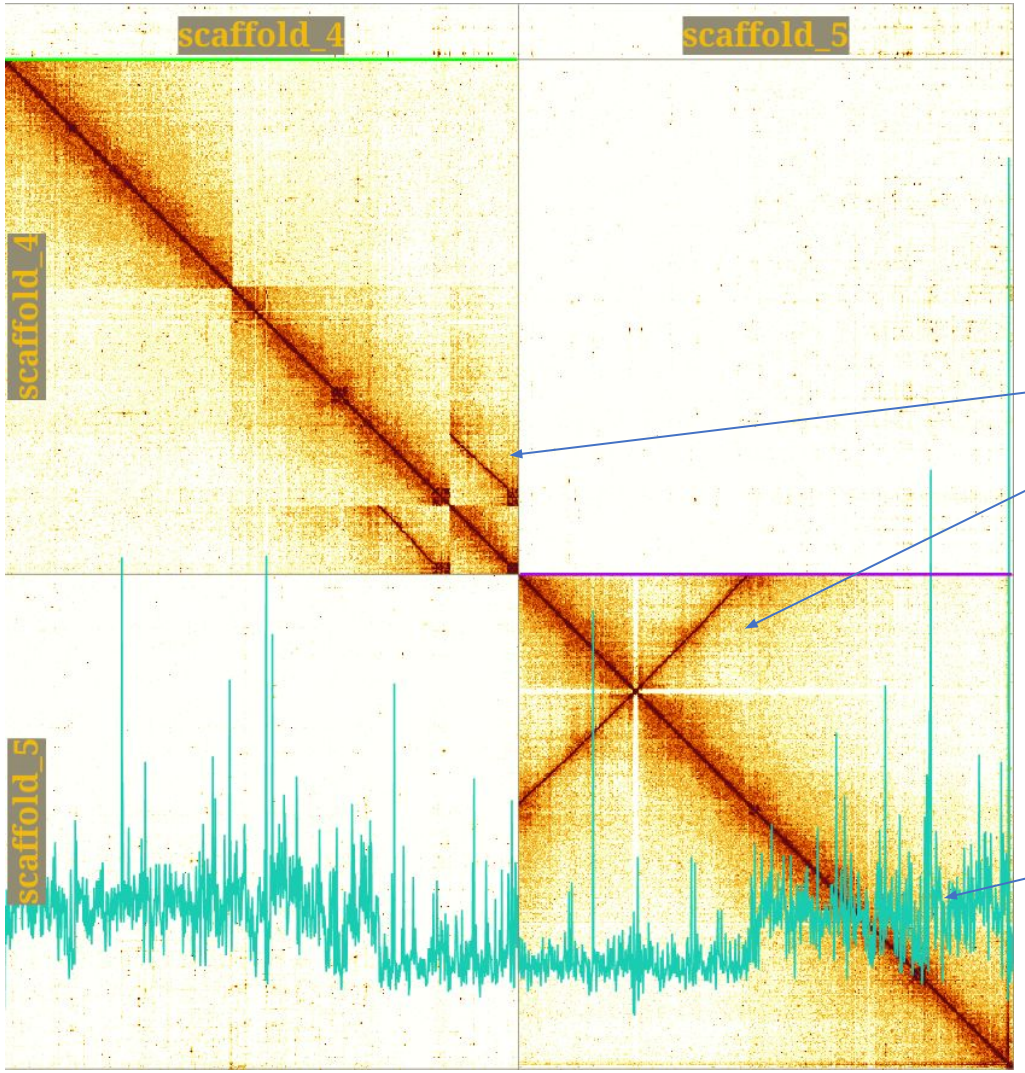




Inversion no longer evident when aligned to comparators

ilLysBell1_1 *Lysandra bellargus* scaffold_22

Haplotypic duplication that needs to be excised

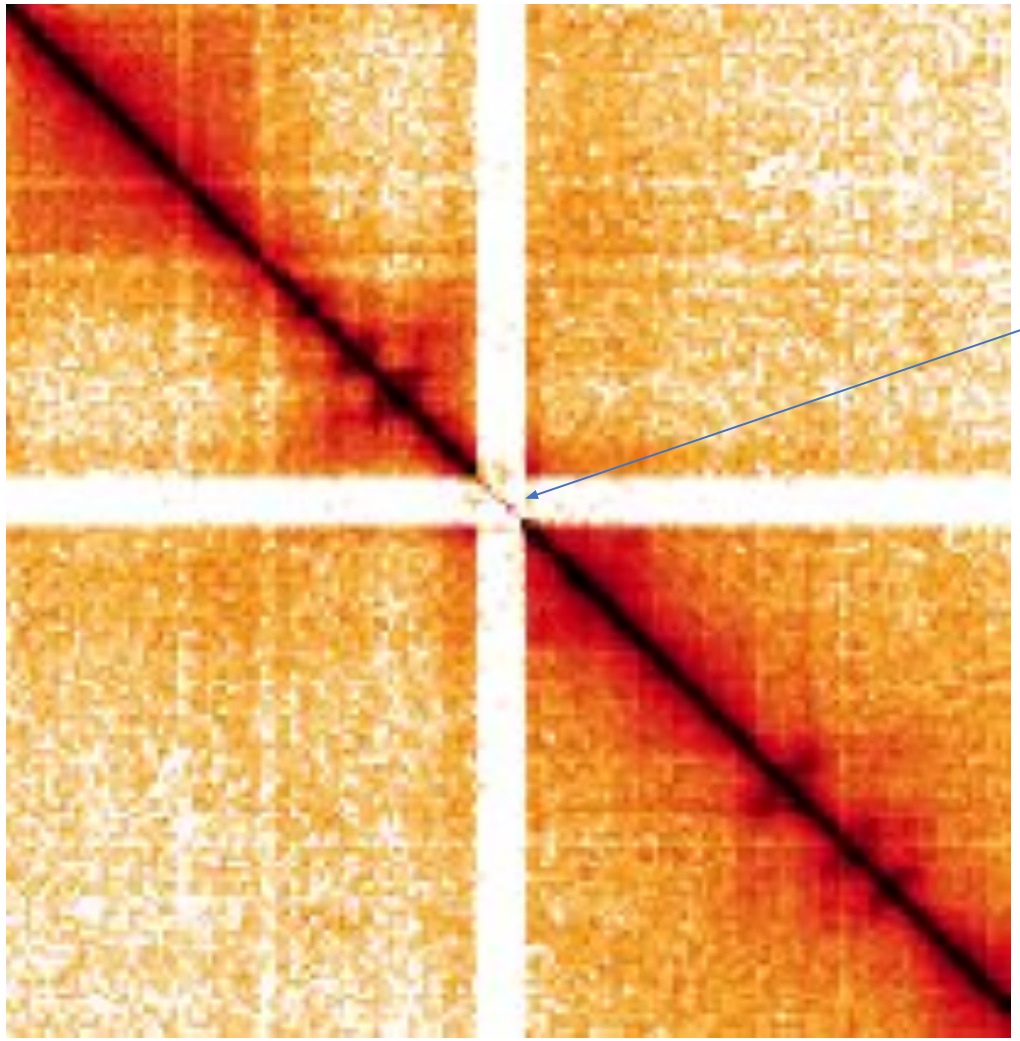


Duplicated sequences - these have half diploid coverage depth and the inflexion point coincides with a gap – the signatures we normally see with haplotypic duplication

Coverage track

qxSacCarc1_1

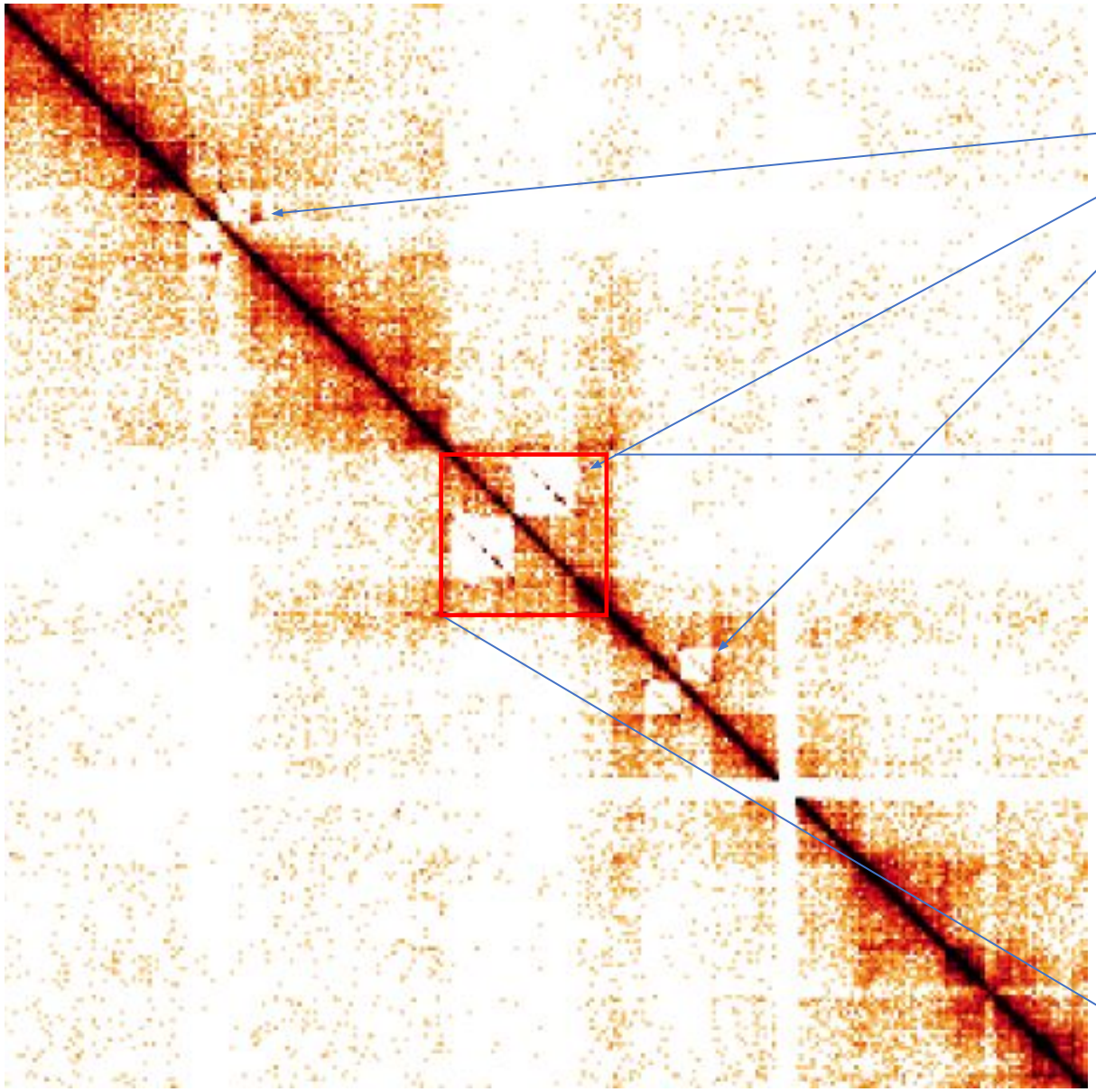
Haplotypes 1



Loss of signal (ie there is sequence here, it is not a gap) due to haplotype scaffold retained in the shrapnel as well as the alternative allele residing in the chromosome. The haplotype in shrapnel should have been removed by Purgedups

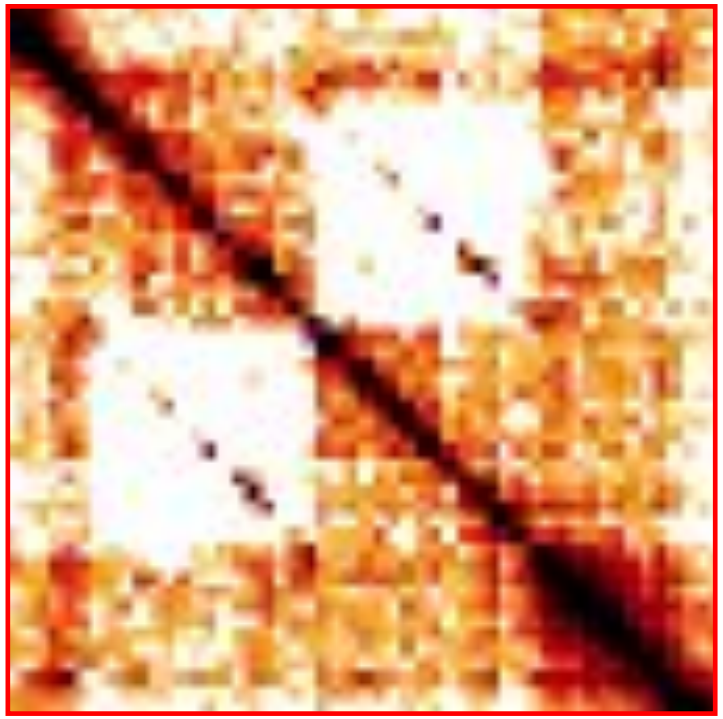
fCycLum1 superscaffold19

Haplotypes 2



Retained haplotypes. Signal on the diagonal is halved. Parallel lines indicate mismapping of reads in parts of the sequence that are most similar between the 2 haplotypes.

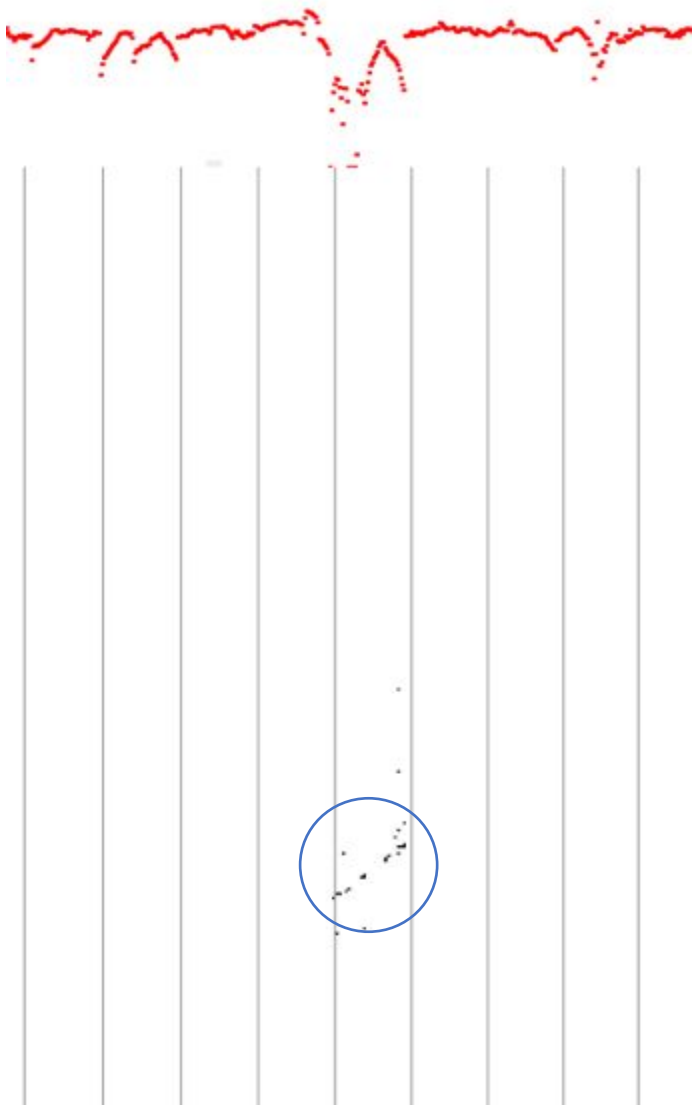
haplotypic duplication



bGeoTri1 superscaffold21

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

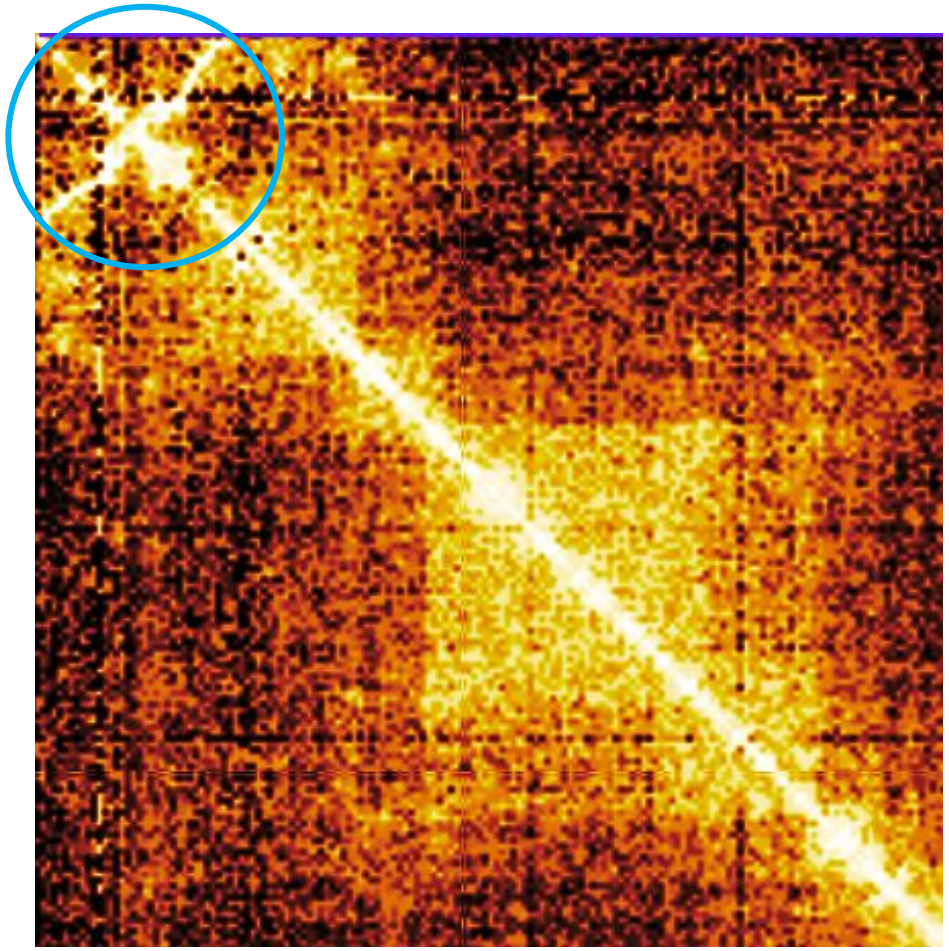
Haplotypic shrapnel contig



Coverage plot show the contig has half depth and the sporadic contacts are typical of a haplotypic contig. From this plot, you can see that the haplotype is entirely contained in the chromosome in the reverse orientation.

(Remember – top right-> bottom left is always reverse orientation and top left-> bottom right is always forward orientation)

Haplotypes 3

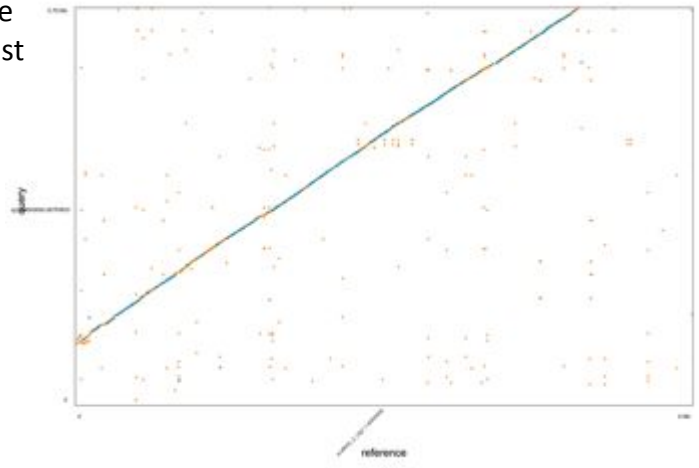


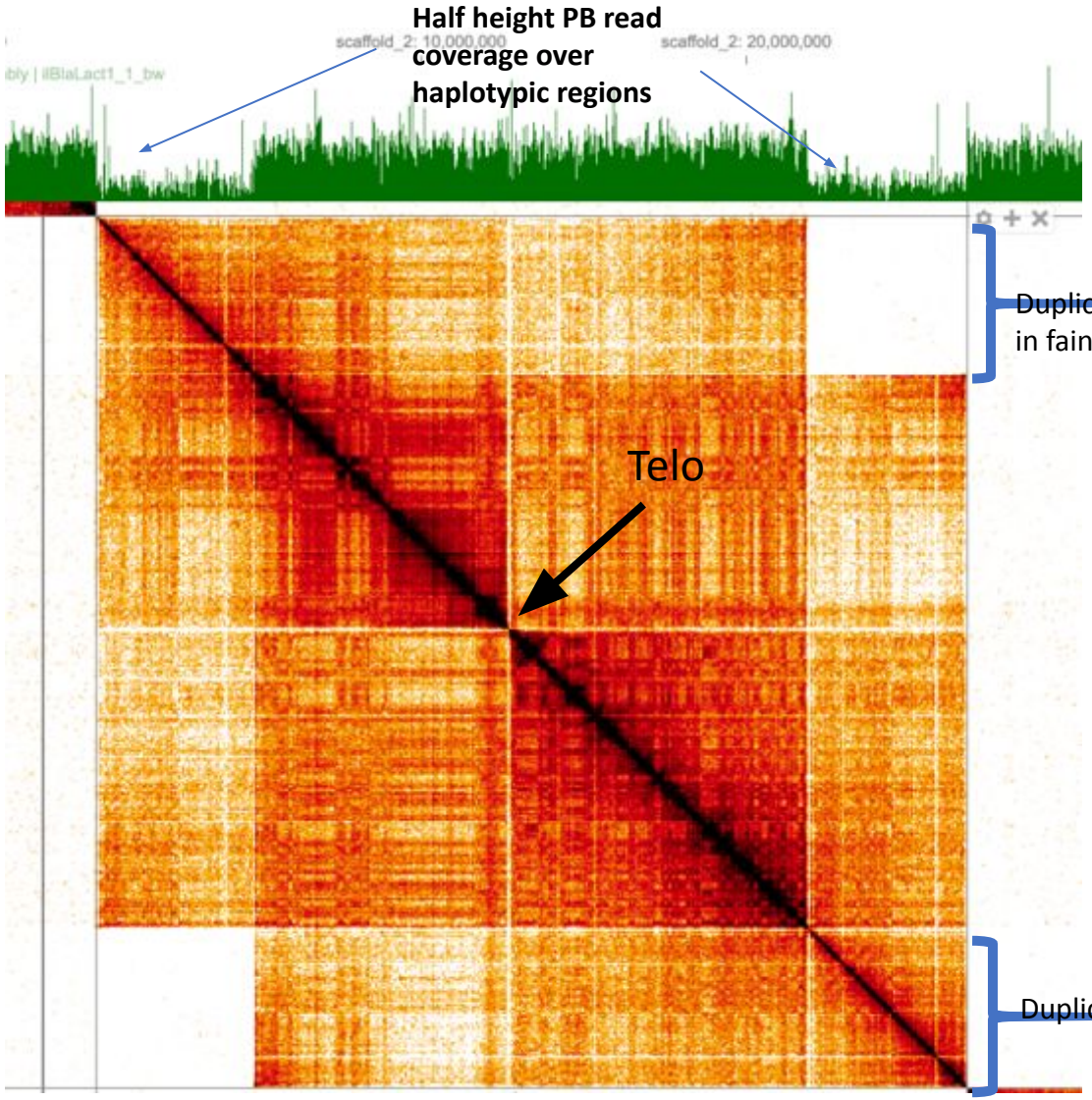
iAphHyp

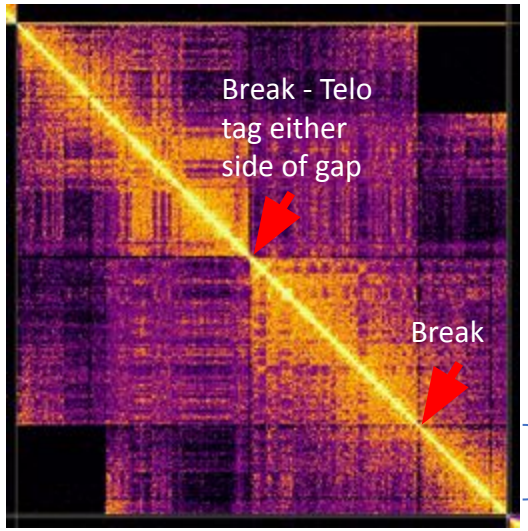
Here we have a haplotypic duplication giving rise to an unusual HiC signal suggestive of an inverted repeat. When we inspect the read coverage, it's clear that this is half what it should be for most of this region.

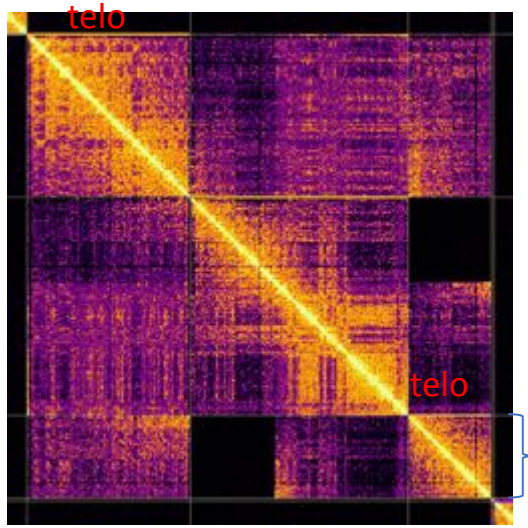
Haplotypes 4

iLBlaLact1_1 has a high level of heterozygosity at 2.8%. This was apparent when looking at the HiC as in some regions the two haplotypes are very divergent from each other thereby confounding the assembler. In this example the scaffold is misassembled as there is a unique section which is then incorrectly joined to 2 different haplotypes of the same region. A nucmer plot pf the 2 suspected duplicates against each other confirms they are haplotypes









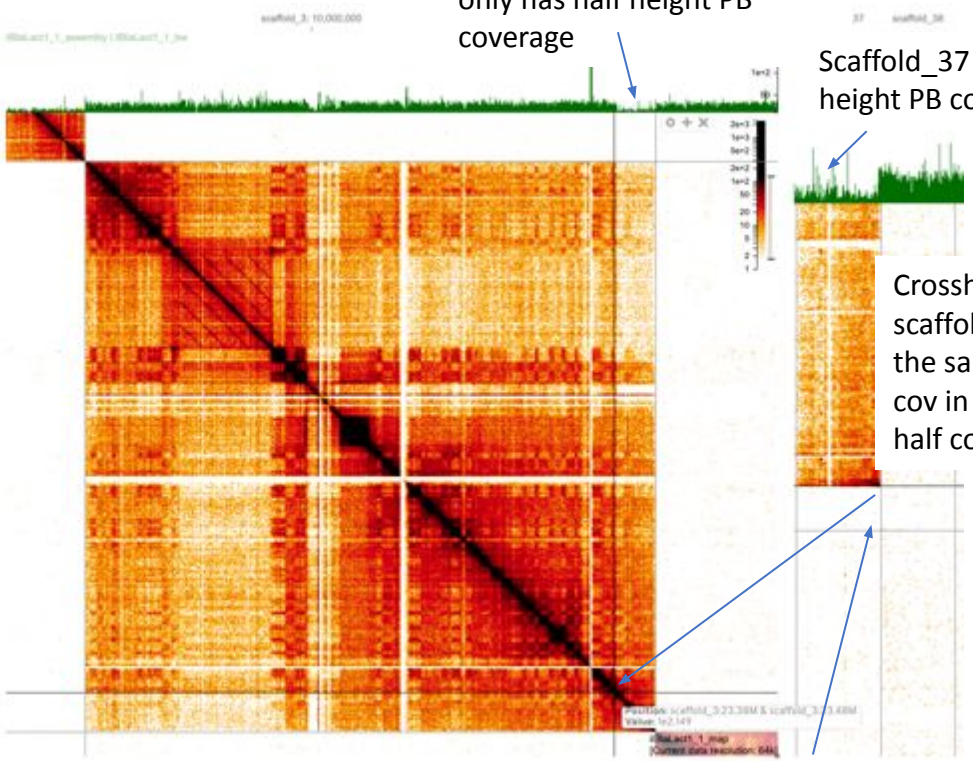
This scaffold was resolved first by removal of one of the haplotypes and then rearrangement. The decision of which haplotype to keep was based on the quality of the existing assembly over the region. The fact that some of the sequence was tagged with the telo repeat helped with the rearrangement of the scaffold following the removal of the haplotypic sequence.

Haplotypes 5

In this example from iLBlacLact2_1 the HiC map indicates that one of the shrapnel scaffolds (scaff_37) wants to join to scaff_3 but in a position where there is already data. Secondly there is no HiC signal between scaff_37 and the last 1.8Mb of scaff_3, this identifies that shrapnel as a haplotype of the end of scaff_3.

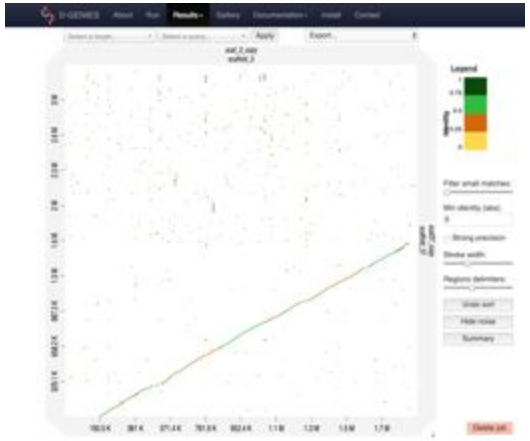
Scaffold 3 has an end that only has half height PB coverage

Scaffold_37 is all half height PB coverage



Crosshairs indicate that scaffold_37 wants to join at the same point that the PB cov in scaff _3 becomes half coverage.

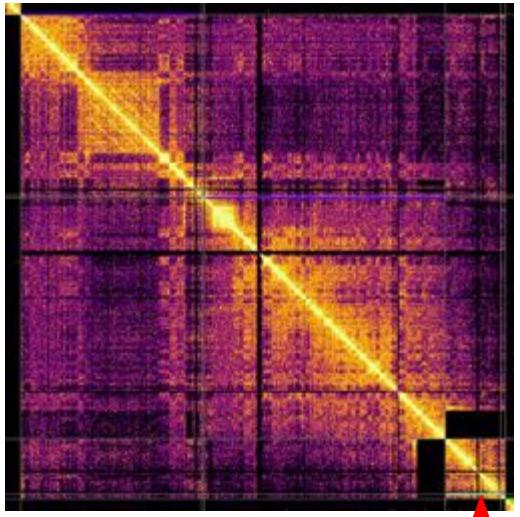
This area of where there is no HiC signal indicates this haplotype is different from the haplotype already incorporated into the primary scaffold.



Dot plot confirms the duplication between the end of scaff_3 and scaff_37 but also shows that the scaff_37 haplotype is longer.

Rearrangement of scaffold_3

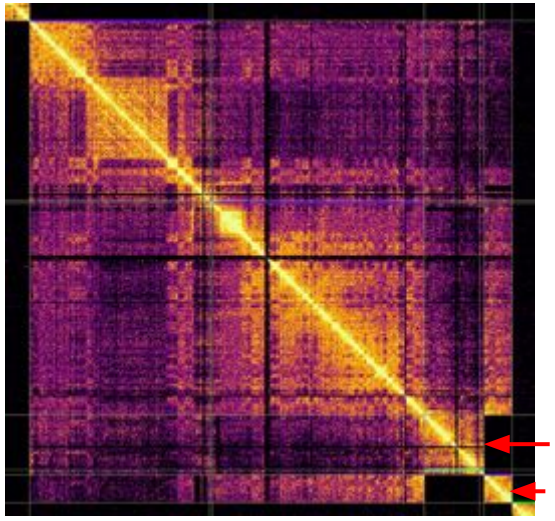
Scaff_3



Scaff_37

In this example a decision had to be made as to which haplotype should be kept in the primary assembly. Either the current end of scaff_3 or replace this with scaff_37.

In this instance it was decided that scaff_37 should be kept as this represented the longest haplotype.

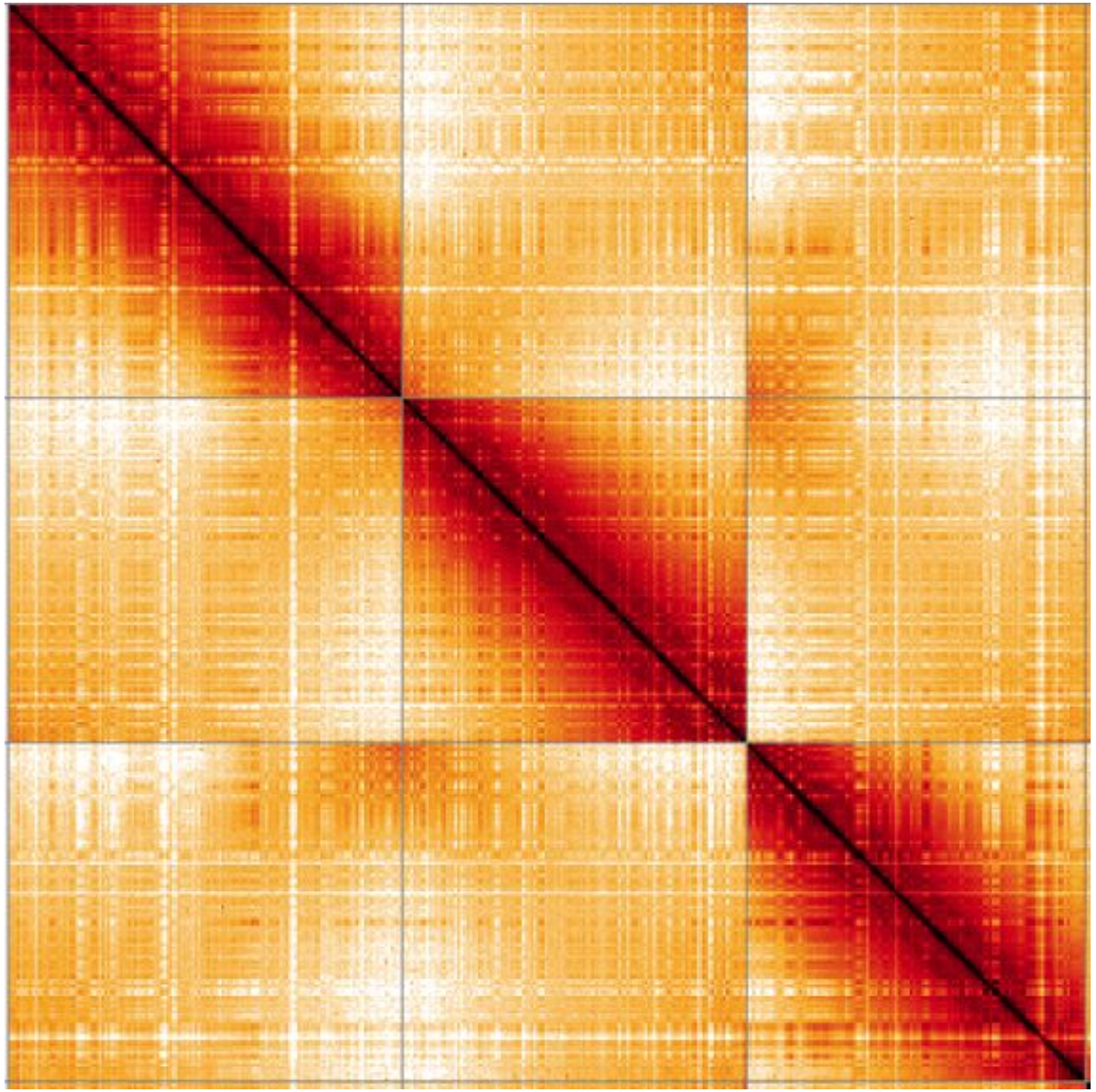


Was Scaff_37

End of 3

End of scaff_3 moved to H, scaff_37 now incorporated into primary to replace end of scaff_3.

Mystery!



rLacAgi1 1st 3 superscaffolds. Arima HiC

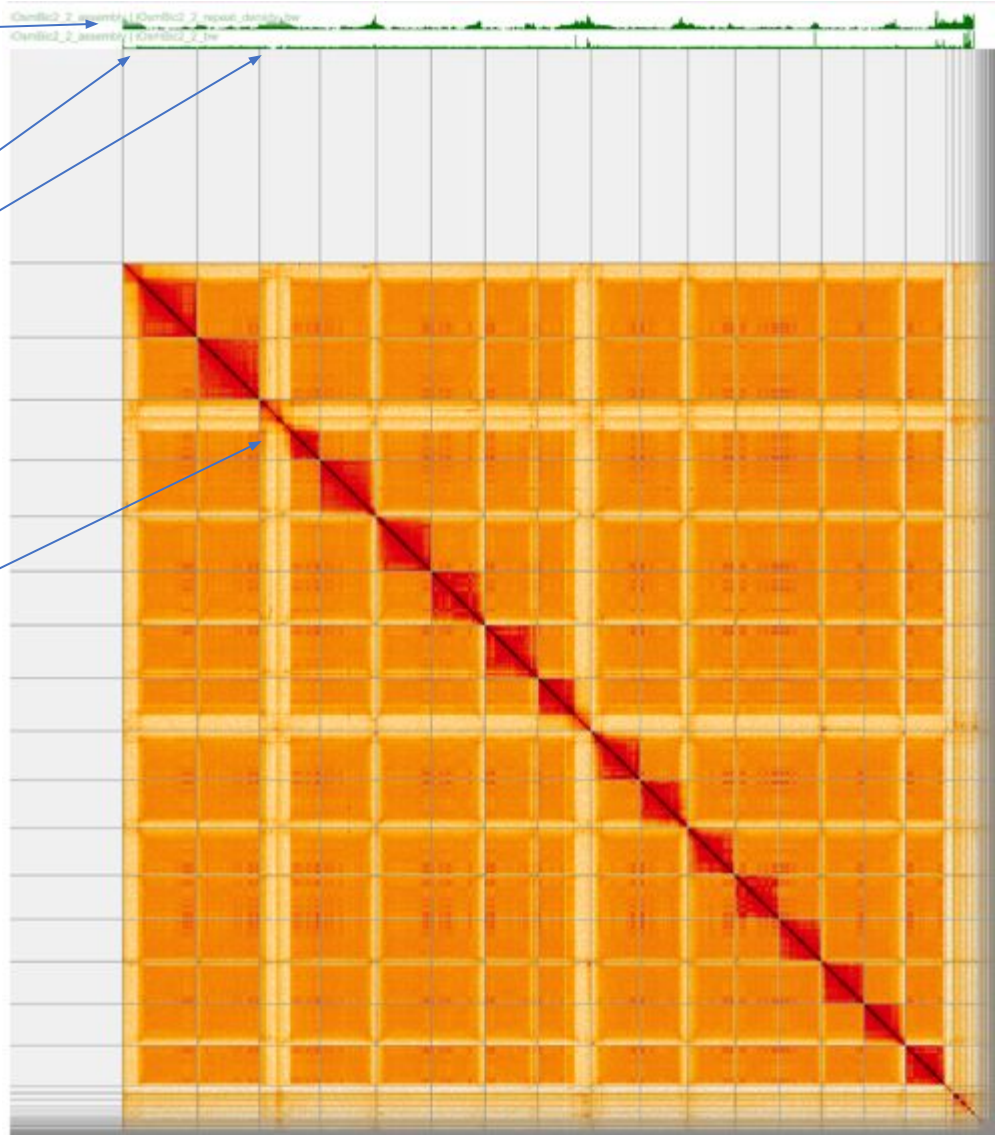
What are the off diagonal parallel stripes (ie the 2 parallel lines just visible either side of the centre diagonal)? In this case moving away from the centre diagonal doesn't result in a gradual and continuous decay. This would be nice to understand!! We have seen this in a number of HiC maps. Although only the first 3 chromosomes are shown, this pattern can be seen in the rest of the chromosomes.

Heterochromatin characterized by high repeat density and pale HiC

repeat density track
correlates with pale banding in HiC. TEs were found to be over-represented in these pale HiC regions

read coverage drops slightly in some of the more repetitive regions due to repetitive mapping

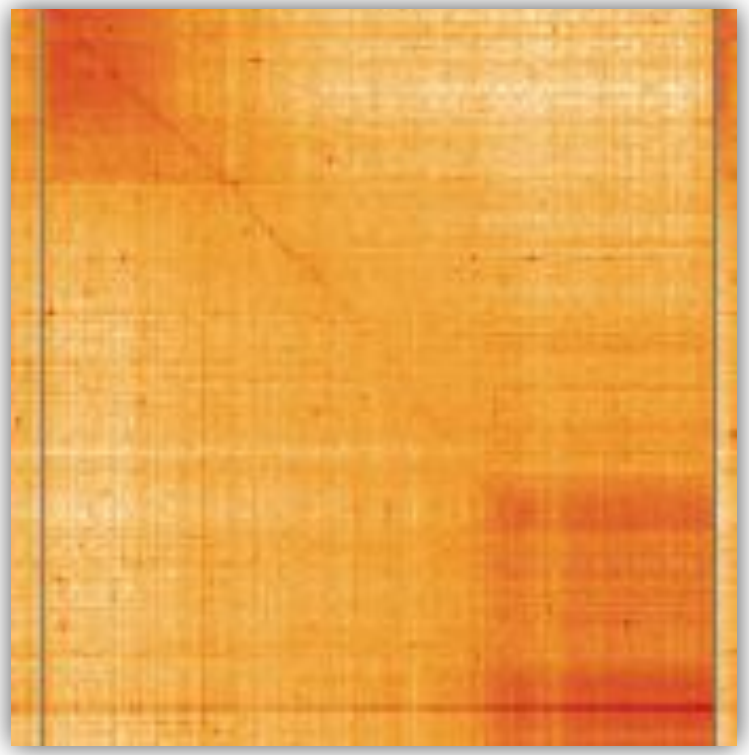
In curation, we decided to chop off the 1st half of scaffold16 as the region has low HiC affinity with this scaffold (and indeed any scaffold). We left this as an island of unassigned repeat.

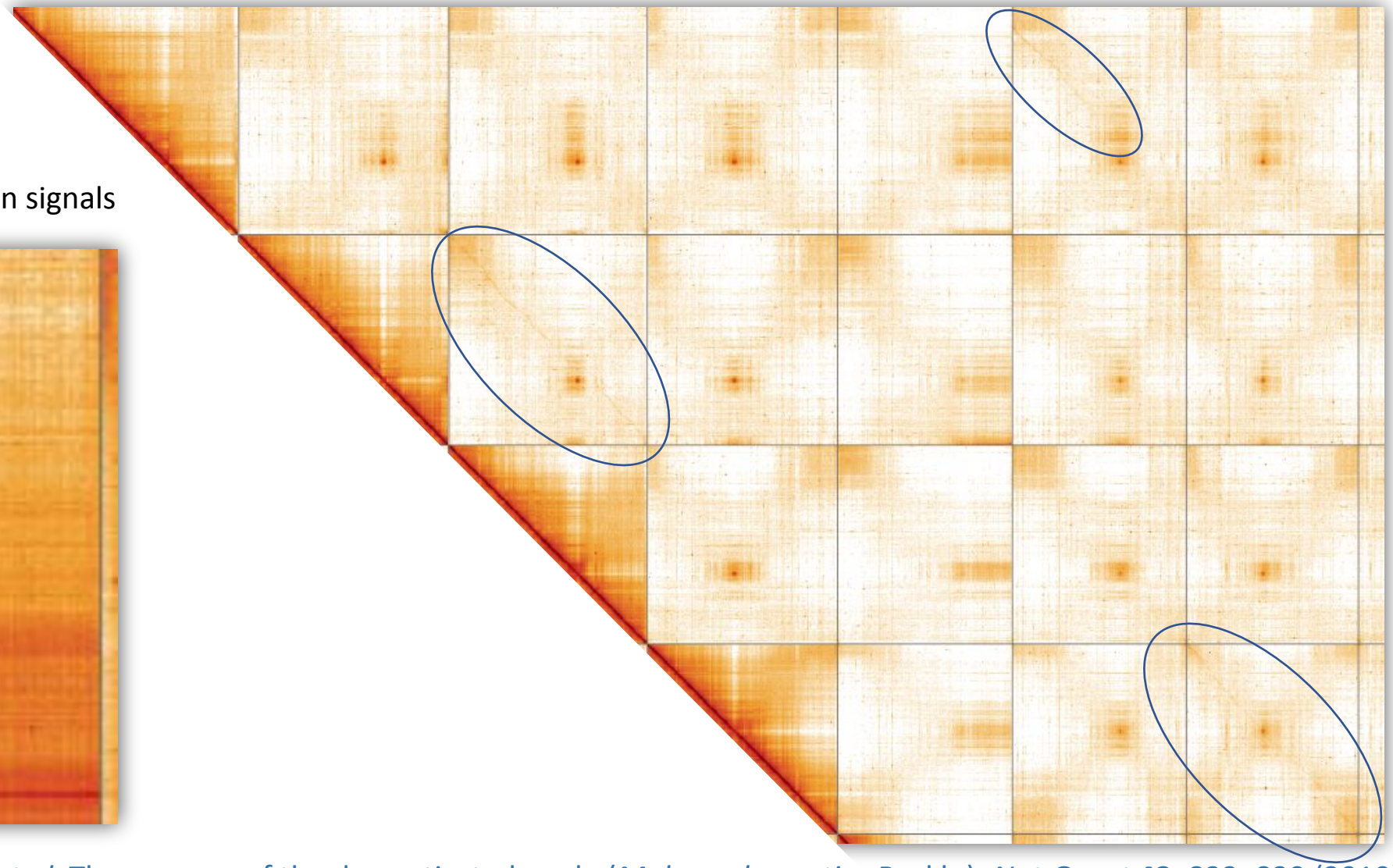


iOsmBic2_2 – whole genome. HiC produced at Baylor

Remnants of genome-wide duplication are clearly visible in the Hi-C and agree with published data

Zoom in on one of the duplication signals





Velasco, R., Zharkikh, A., Affourtit, J. *et al.* The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat Genet* **42,** 833–839 (2010). https://doi.org/10.1038/ng.654

drMalDome11