Session 4:
Challenging genomes to curate and strategies to work with them

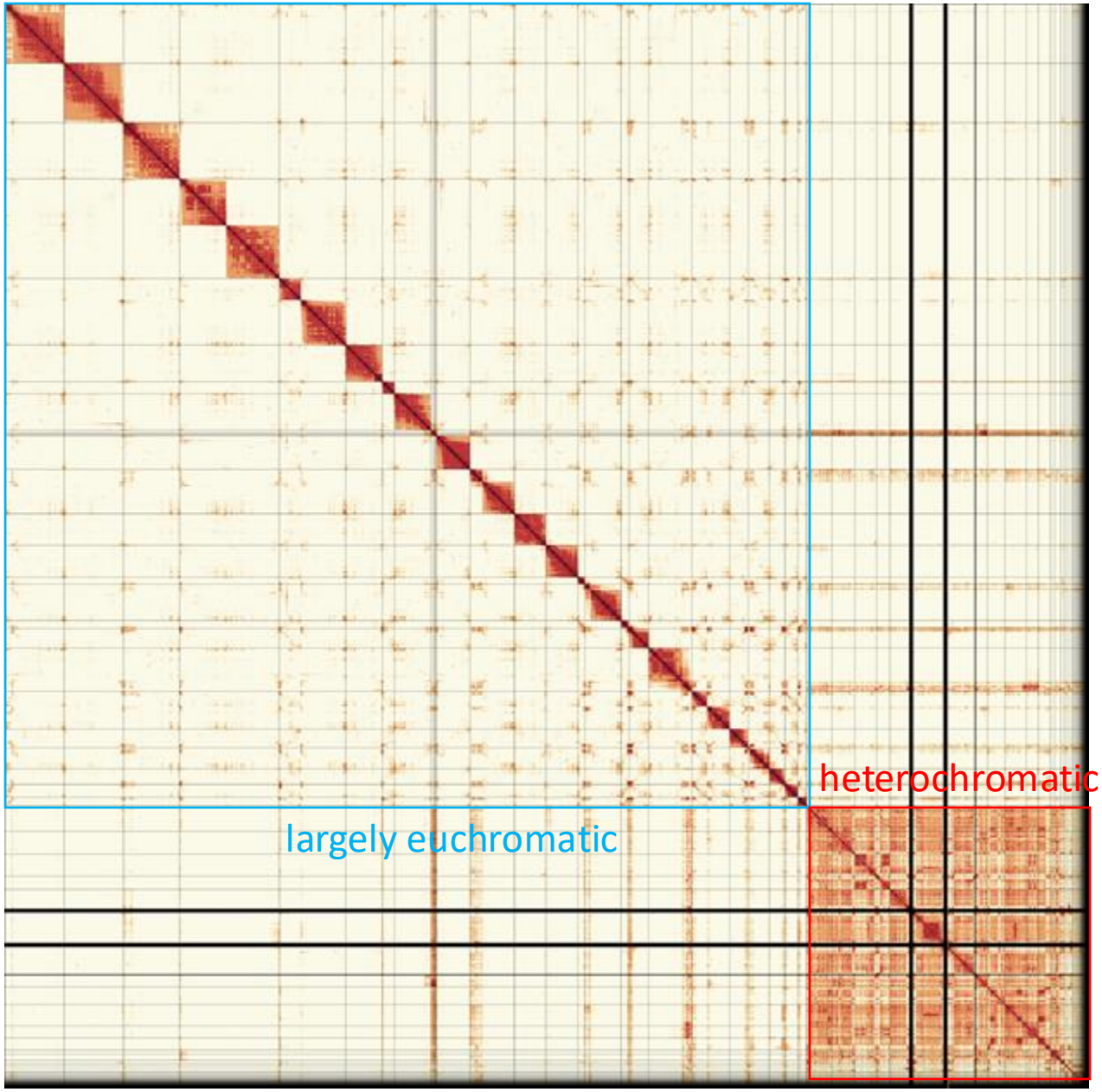
Genome Reference Informatics Team (GRIT)
Wellcome Sanger Institute - Tree of Life

High repeat and heterochromatin content

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



High repeat and heterochromatin content

HiC bias

- More represented:

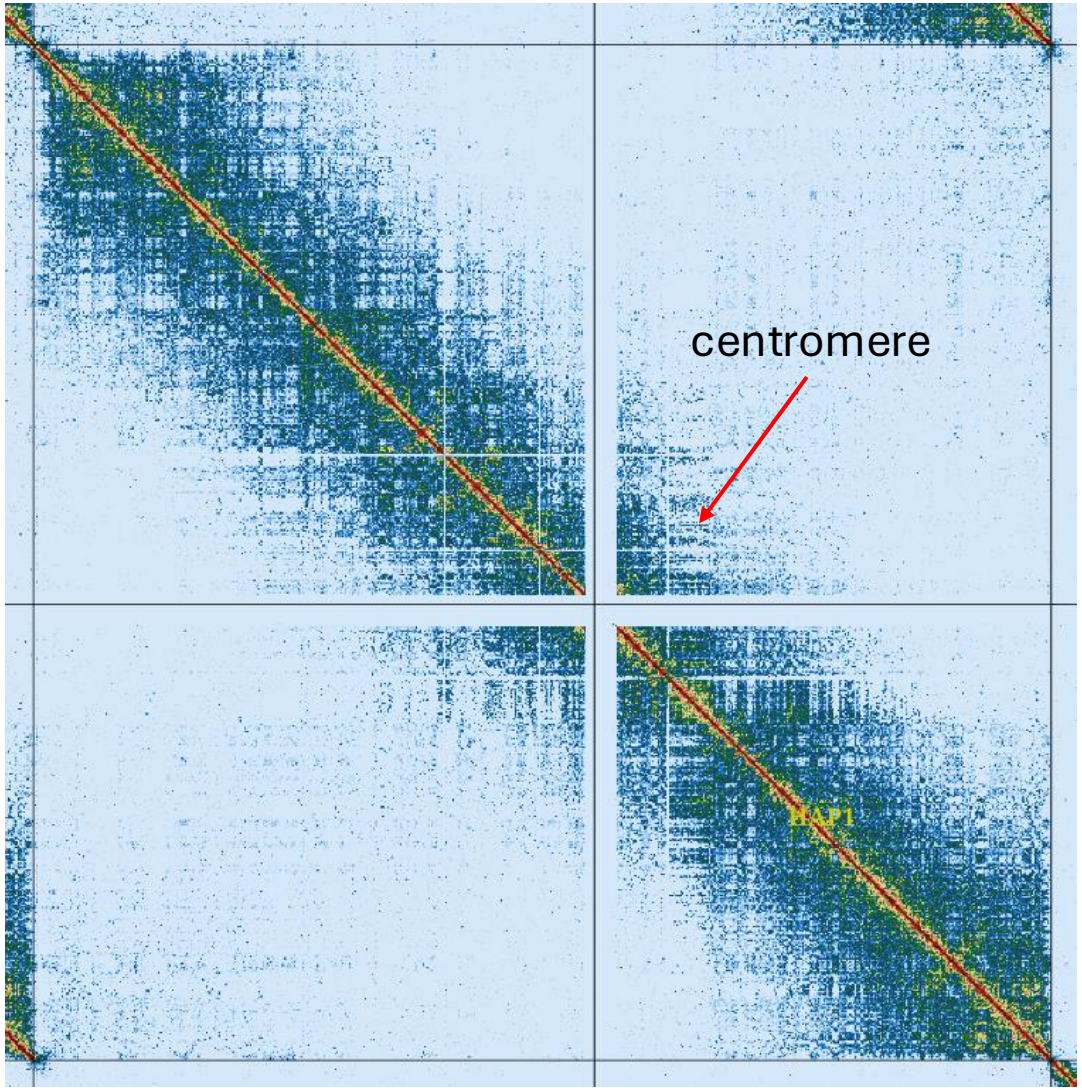
-High GC content: due to PCR and sequencing bias

- Less represented (Low mappability):

Repetitive regions: Centromers, telomeres and repetitive regions

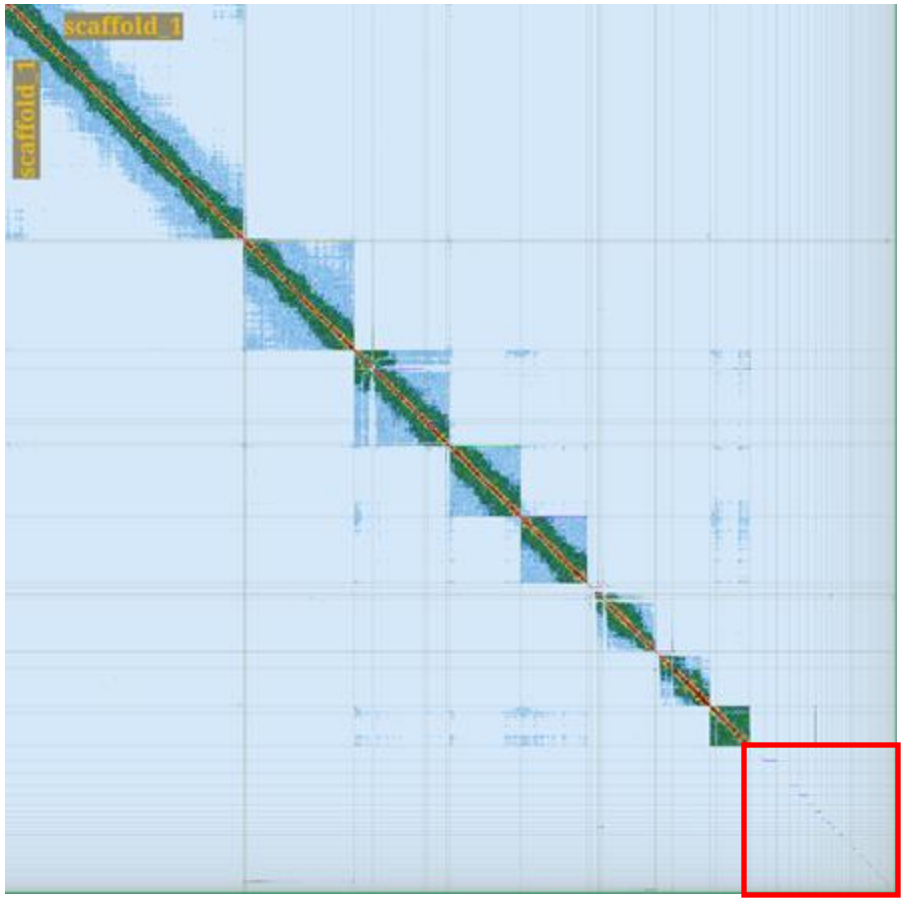
HiC reads align with too many regions

Worse when you have HiC low coverage

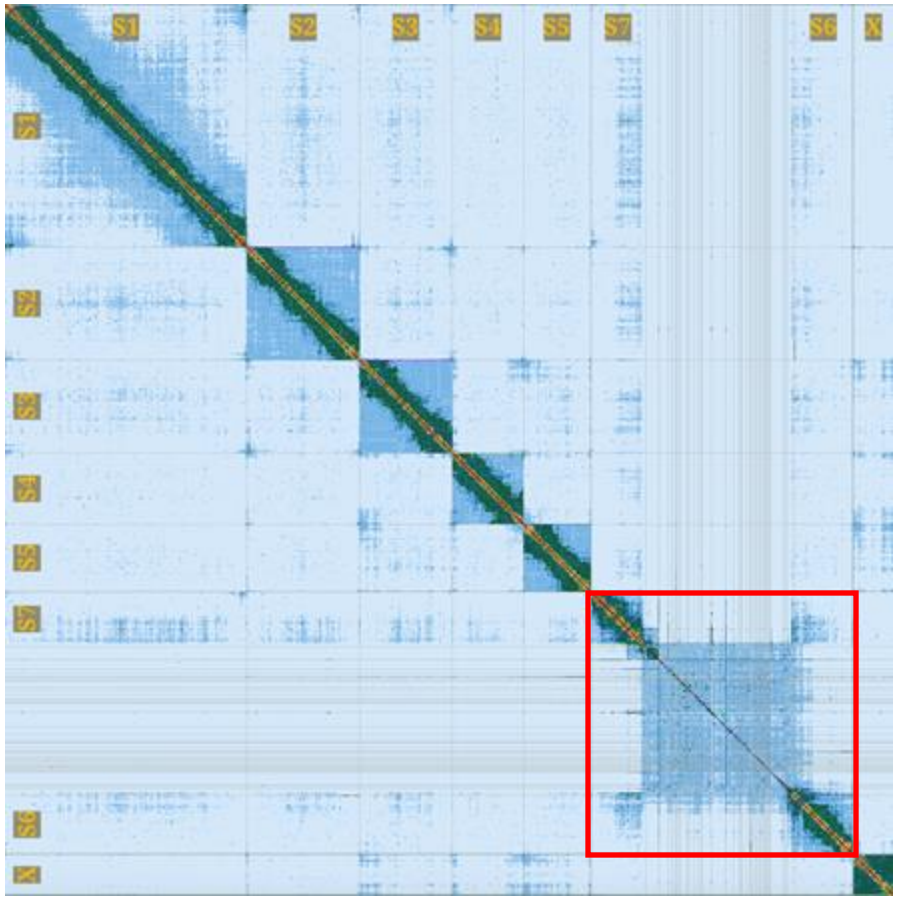


Multi-mapping + karyotype

Multi mapping reads reveal hidden linkage between ‘separate’ chromosome scaffolds and blank repetitive scaffolds



multi-mapping ‘off’



multi-mapping ‘on’

Rhagonycha fulva

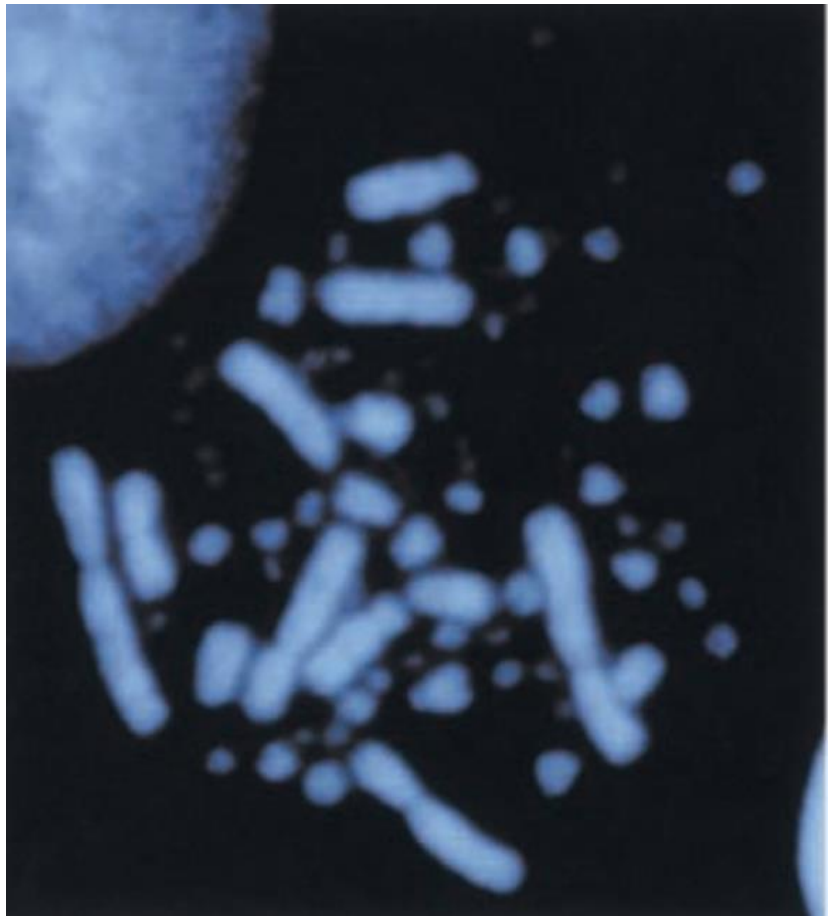
Karyotype image confirms presence of large heterochromatic chromosome

https://doi.org/10.1007/s10709-006-9106-5

Microchromosomes

Birds, sharks and reptiles (only?)

Bird genomes are organized in macro- and micro-chromosomes (By Tom Mathers)



Chicken chromosomes (n = 39)

Masabanda *et. al.* 2004, *Genetics*

The chicken genome is typically divided into 10 **macrochromosomes** (>23 Mb in length) and 29 **microchromosomes** (22 – 2.5 Mb in length).

Microchromosomes can be further divided into **micro** (22 – 5 Mb, n = 19) and "**dot"** (< 5 Mb, n = 10) chromosomes.

Dot chromosomes represent a major challenge for assembly and curation.

* Sizes based on latest chicken near-T2T assembly (Huang *et. al.* 2023, *PNAS*).

In cuckoo, the 10 smallest chromosomes represent 1.2% of the genome assembly but contain 10% of the protein coding genes.

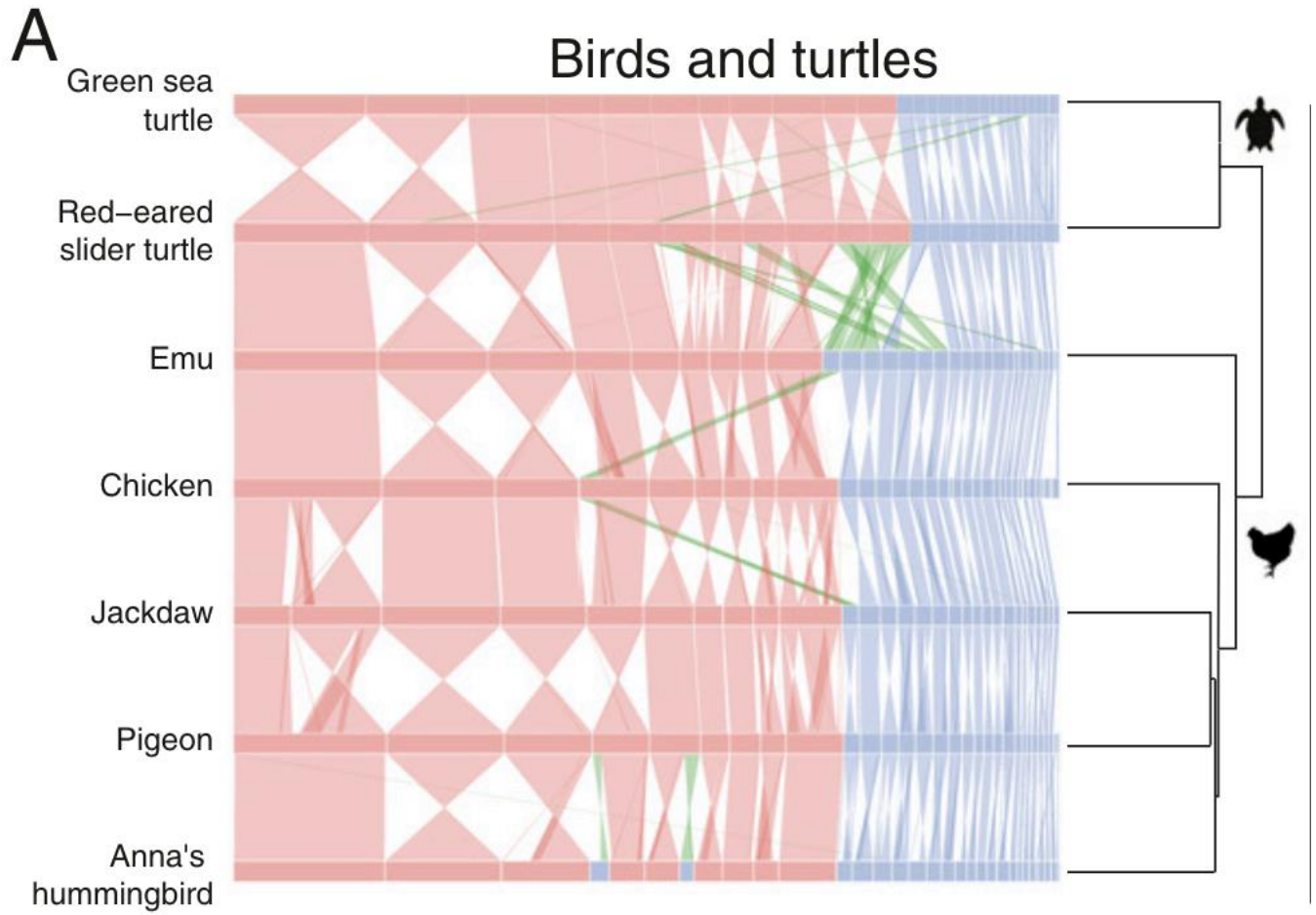
Microchromosomes

Less than 2% of the assembly and 99% of the effort….

Microchromosomes are often highly fragmented and mixed in with repeat scaffolds in assembly “shrapnel”

Considerable gene content

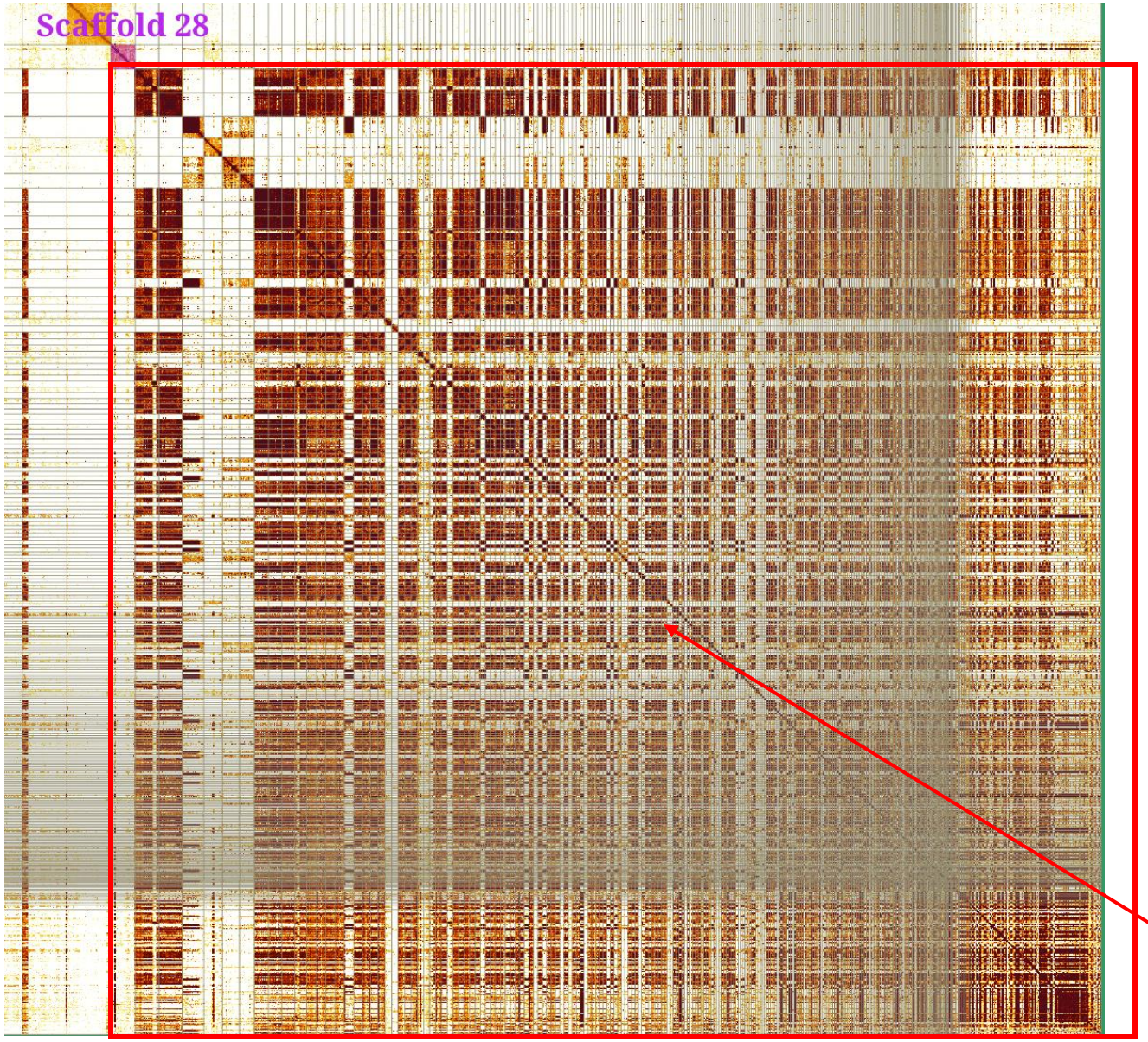
Microchromosomes are highly conserved across birds and reptiles (By Tom Mathers)



Waters *et. al.* 2021, *PNAS*

Microchromosomes

(By Tom Mathers)



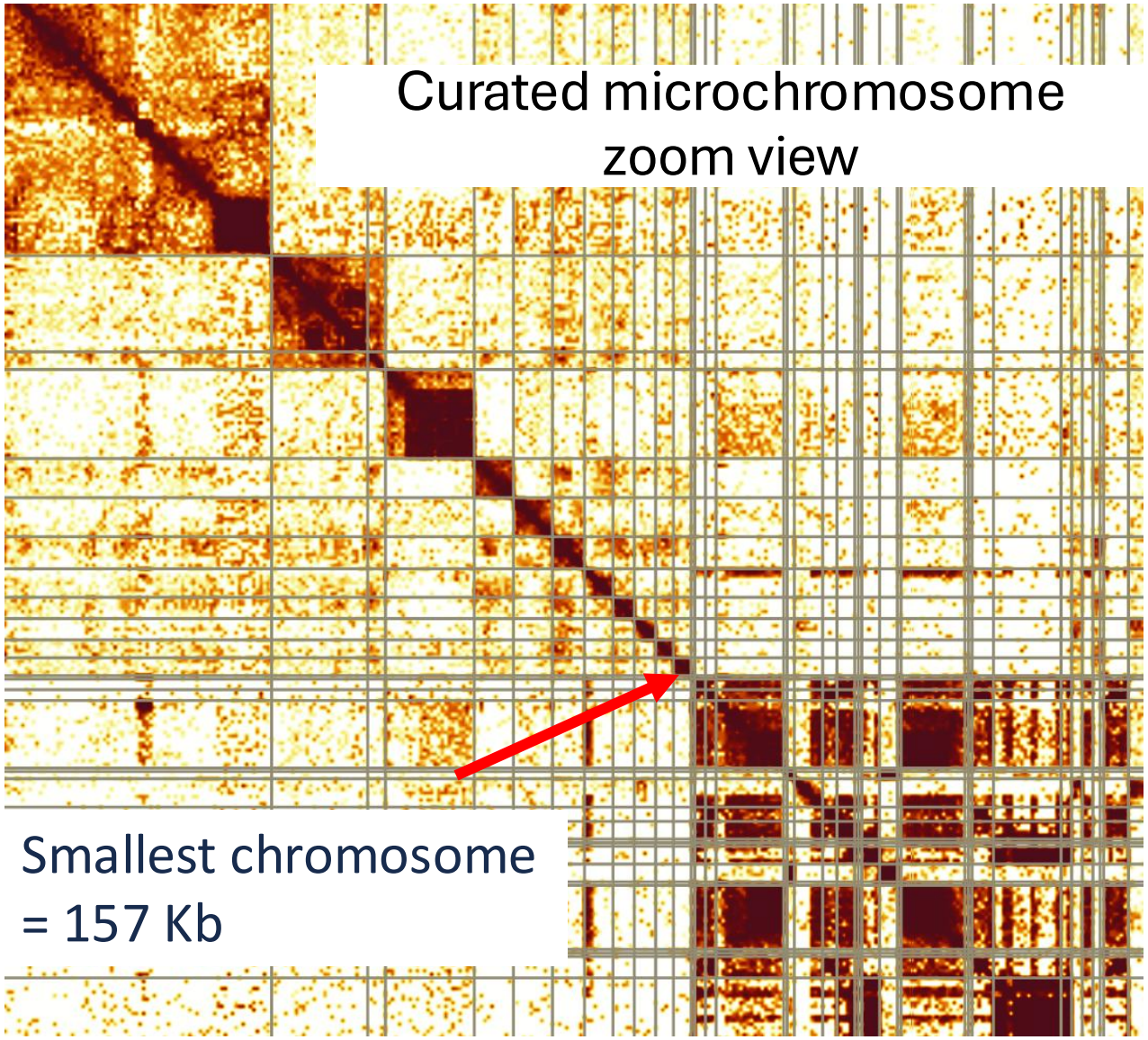
Quick curation of larger scaffolds only recovers 28 chromosomes.

Expected karyotype is 39 autosomes + Z + W

Remaining 13 chromosomes are somewhere in here!

Micros ???

Microchromosomes



(By Tom Mathers)

To find these missing chromosomes we **rely on elevated background HIC signal between micros.**

Additionally,
JBrowse looking for gene-rich small contigs

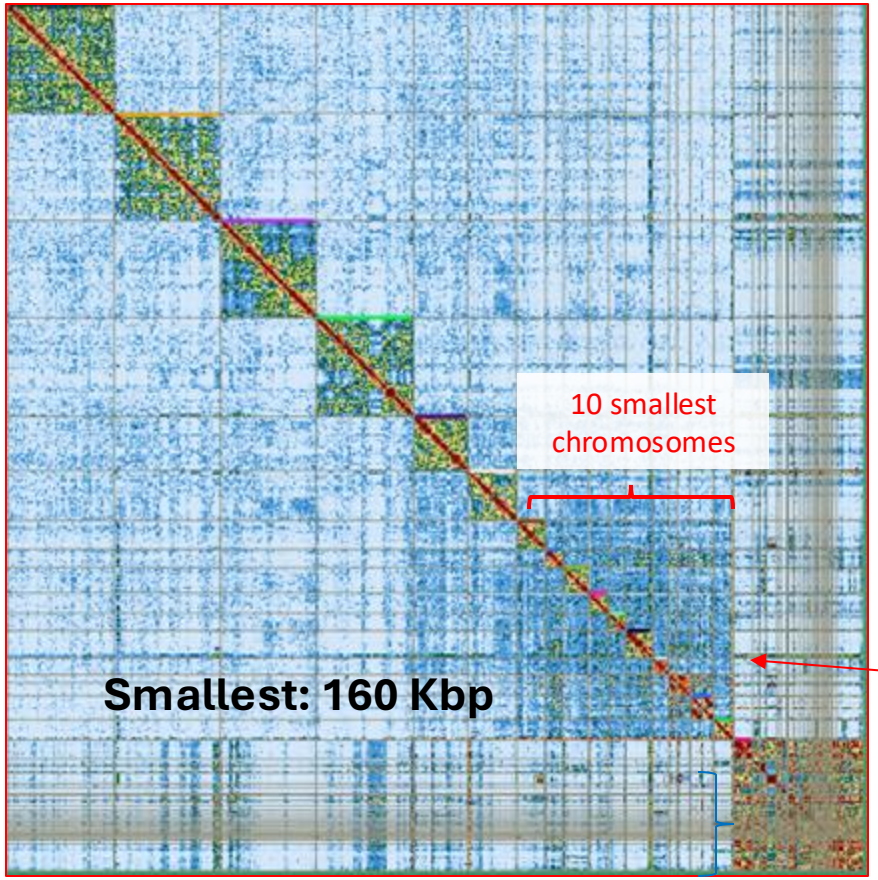
Manually inspect whole genome **alignments** vs well assembled relatives (if present).

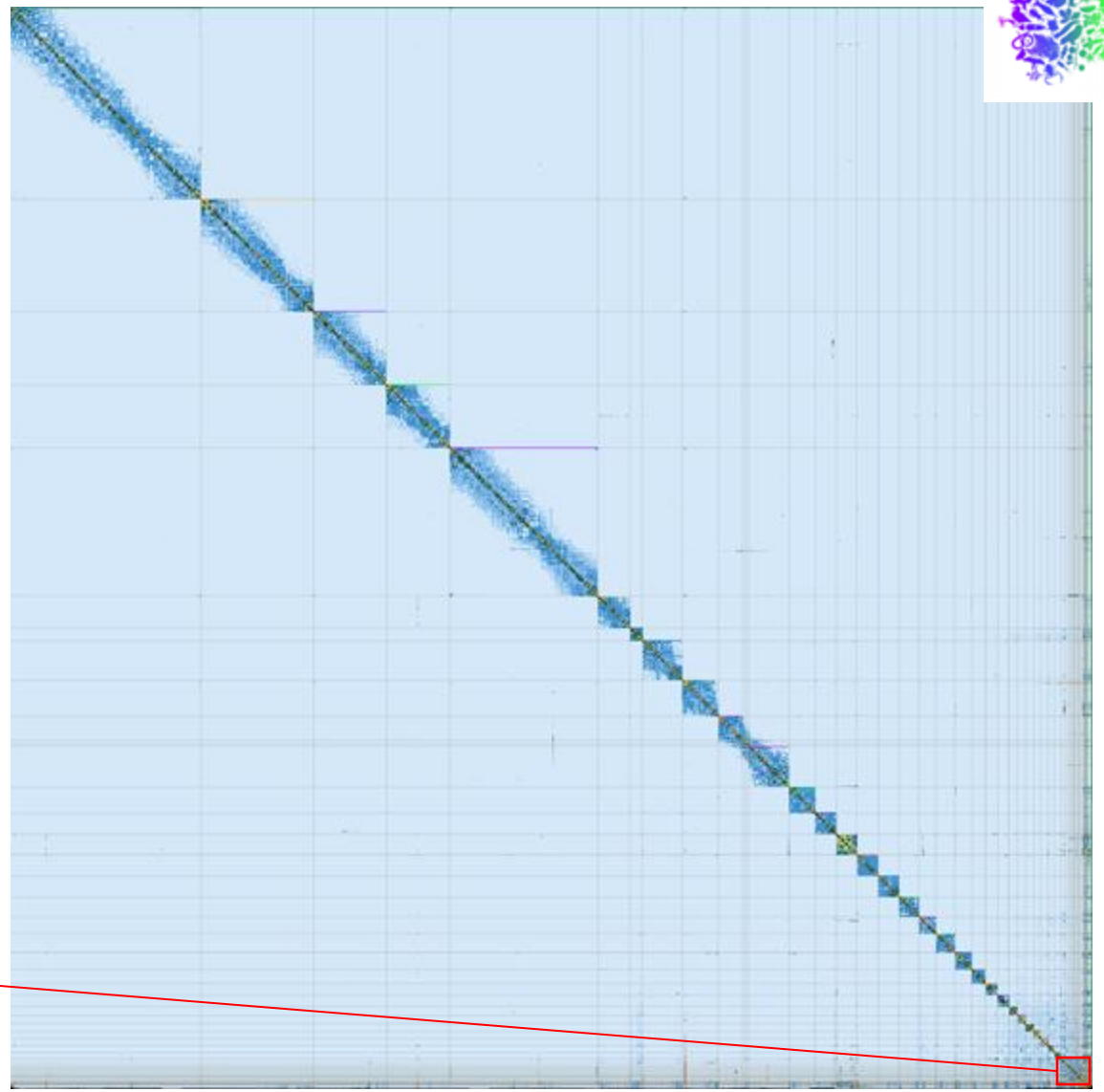
Very time consuming.

Micro-chromosomes - Birds

(bCucCan1)

- Disproportionate amount of time curating the smallest 10 micro-chromosomes (<1.2% of the assembly)….

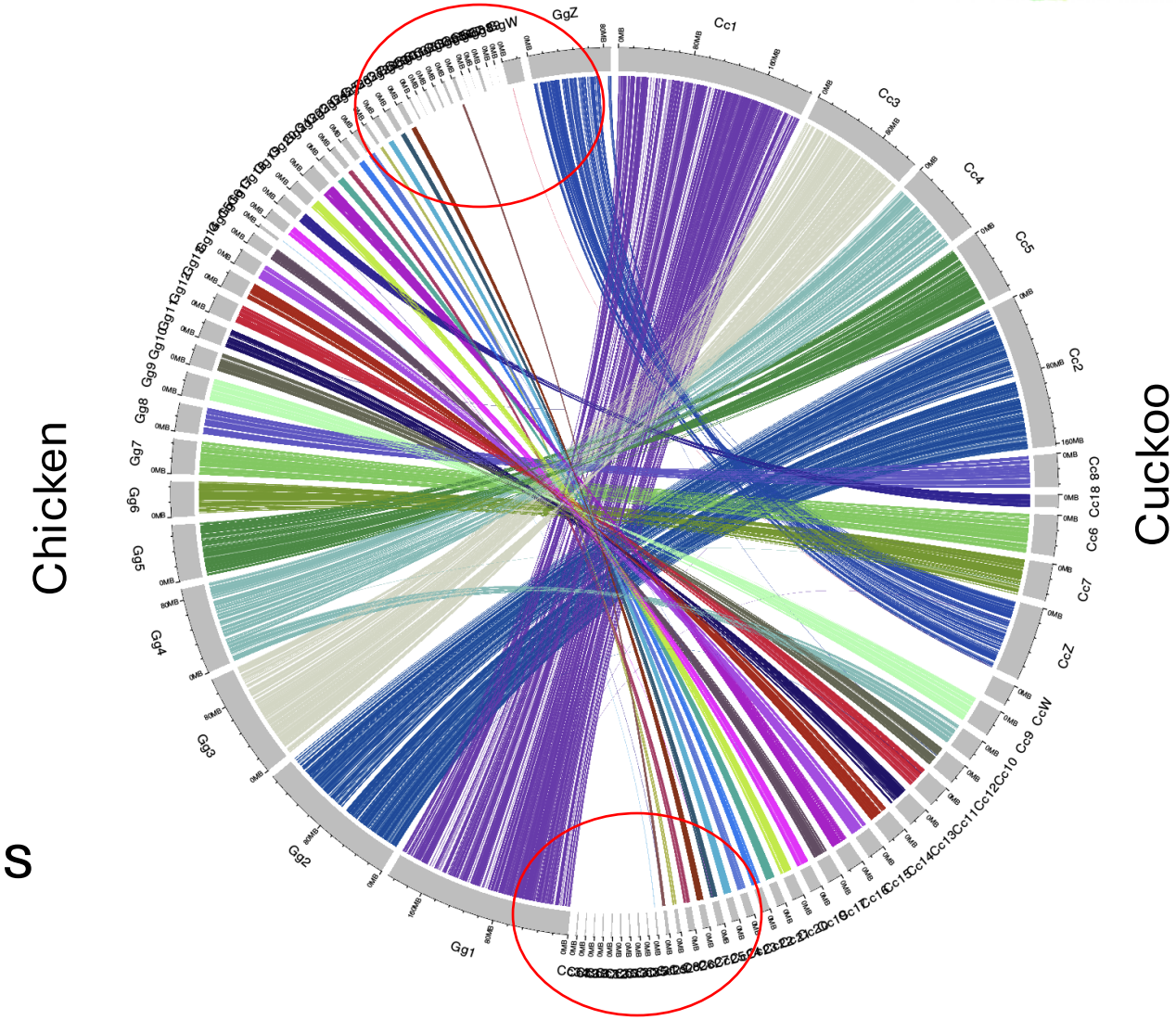




unplaceable sub-telomeric repeat

Most challenging group: birds

- Very time consuming
- Highly repetitive and fragmented (ideal is a hybrid assembly using HiFi and ONT reads)
- Really tiny
- BUSCO Aves ODB 10 gene set does not cover all microchromosomes



(By Tom Mathers)

How do we fish out the micros?

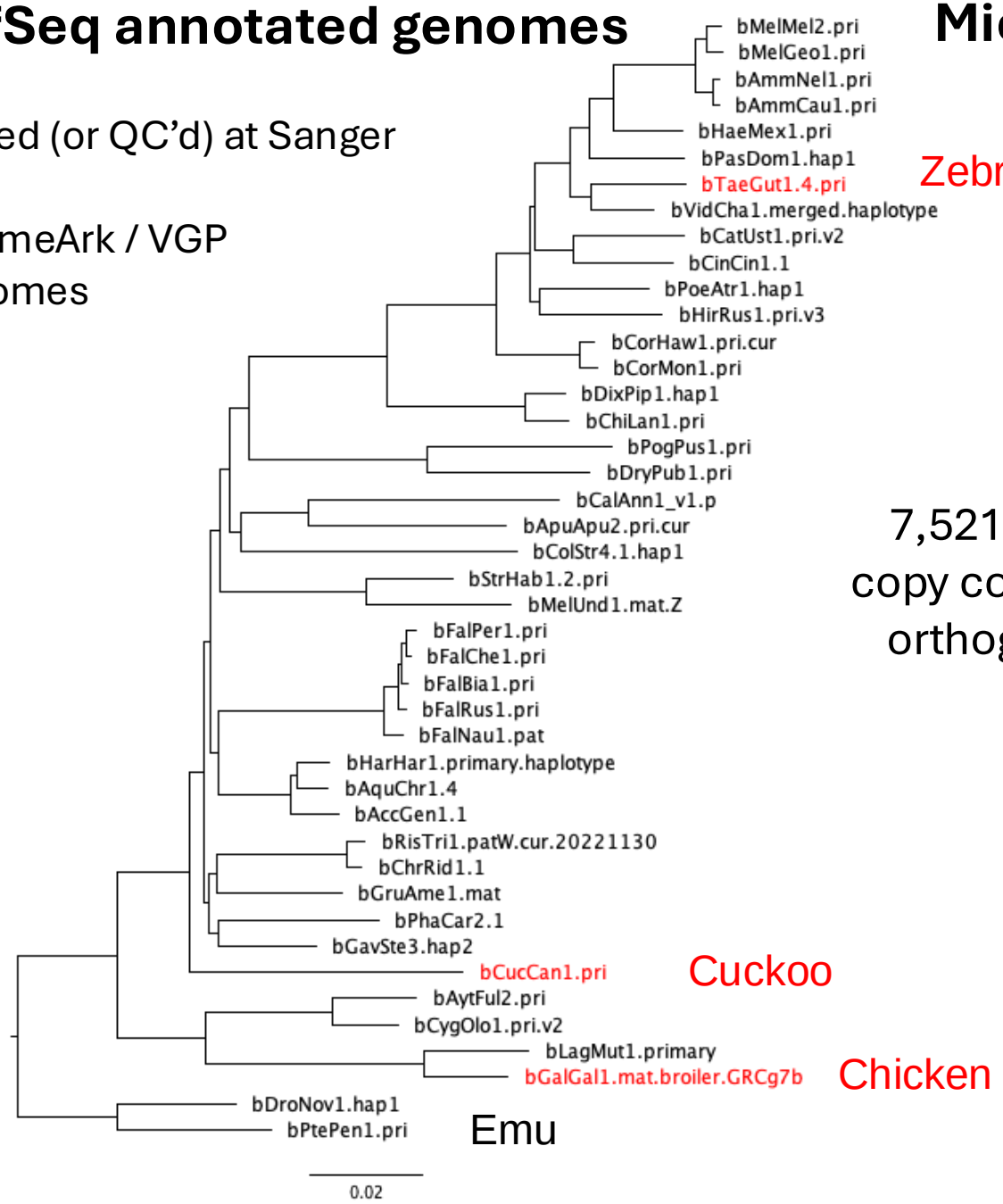
Main approaches we use for birds:

1. `MicroFinder` script for birds (by Tom Mathers and Michael Paulini)
2. Nucmer
3. Gene content in Jbrowse

43 RefSeq annotated genomes

All curated (or QC'd) at Sanger

38 GenomeArk / VGP
125 genomes
4 DToL



Zebra finch

Cuckoo

Chicken

Emu

MicroFinder script (by Tom Mathers)

Reduce protein sets to longest transcript per gene

↓

Cluster all proteins (n = 704,742) with OrthoFinder

↓

18,449 orthogroups

←

7,521 single copy conserved orthogroups

↓

Select "fuzzy" conserved orthogroups with KinFin

↓

14,055 orthogroups present >= 50% of species, max 3 copies

How do we fish out the micros? (Birds)

MicroFinder script for birds:

https://github.com/sanger-tol/MicroFinder

Recommended:
16 cores
24 Gb RAM
Scaffolds > 5Mbp will not be ordered

The script should be run for each haplotype separately:

```
"/MicroFinder.v0.1.sh <hap1_fasta> <output_hap1_name> scaffold_length_cutoff"

"/MicroFinder.v0.1.sh <hap2_fasta> <output_hap2_name> scaffold_length_cutoff"
```

`scaffold_length_cutoff (Kbp)`

It will:

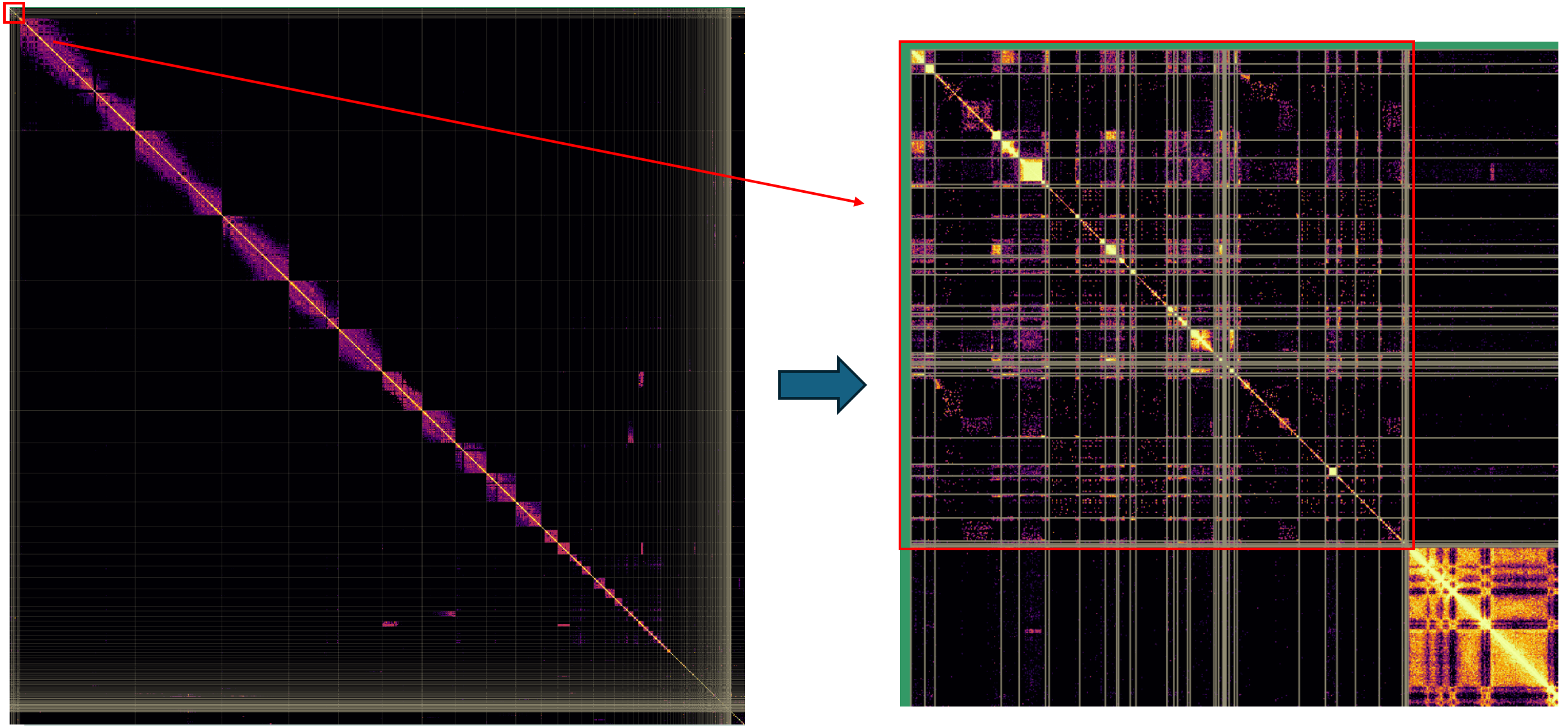
Align your genome to a conserved database of bird microchromosomes and look for gene content

Sort by number of gene hits and then by size (< 5Mbp only) and move them to the beginning of the fasta file

Generate a new fasta file

How do we fish out the micros? (Birds)

Potential micros will appear on the top left of hap1 and hap2 new Pretext maps: **single map curation**



How do we fish out the micros? (Birds, sharks, reptiles)

Nucmer (Dot viewer: https://dot.sandbox.bio)

```
nucmer -p <align_name> <reference_fasta> <query_fasta>
```

Output: align_name.delta

```
DotPrep.py --delta <align_name.delta> --out <output_align_name>
```

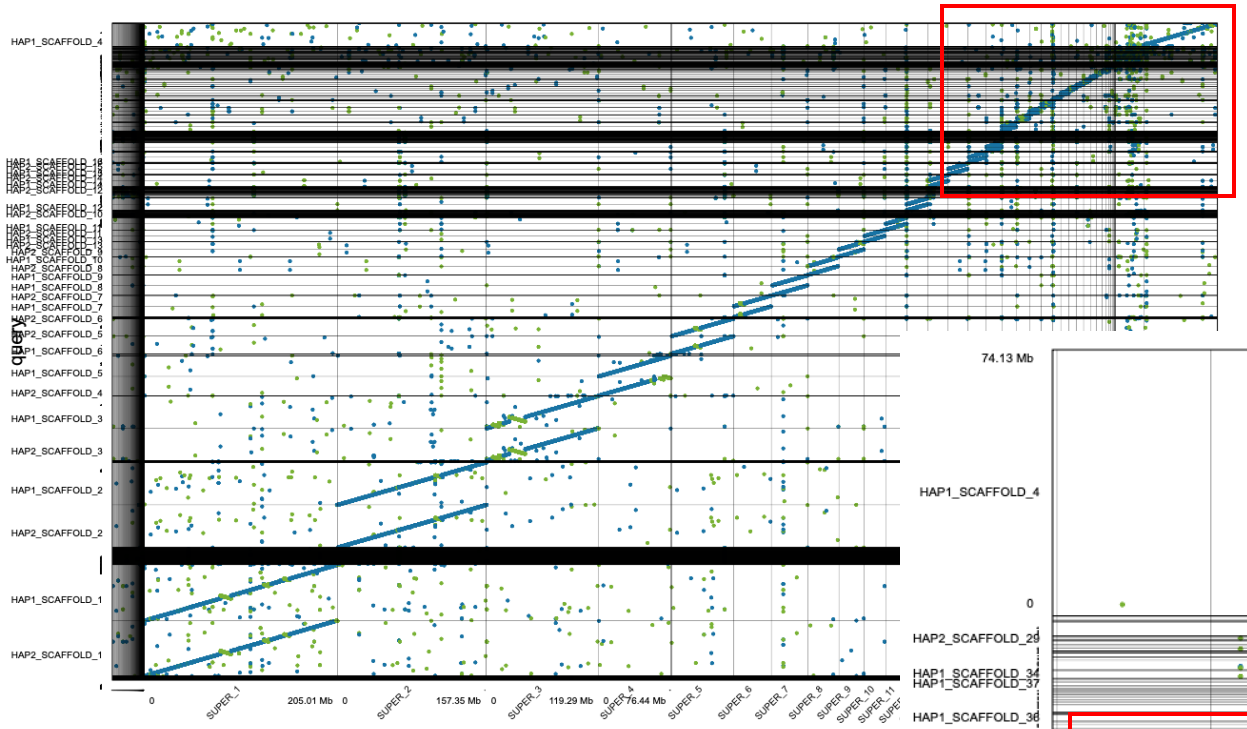
Output:

```
<output_align_name>.coords
<output_align_name>.coords.idx
```

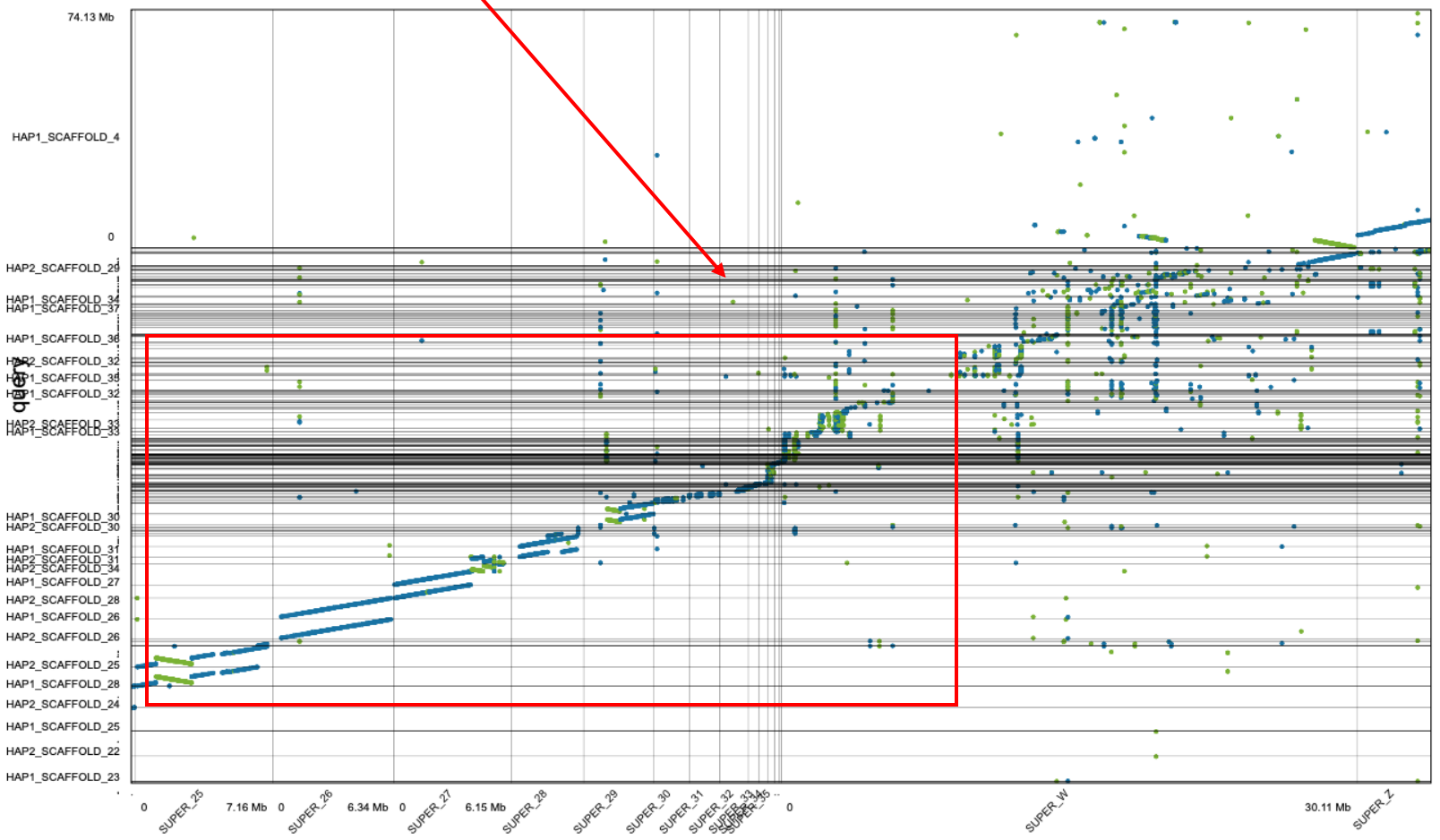
In which order the micros go in the map?

How do we fish out the micros?

(Birds, sharks, reptiles)



Zoom-in: Right click and drag
Zoom-out: double click



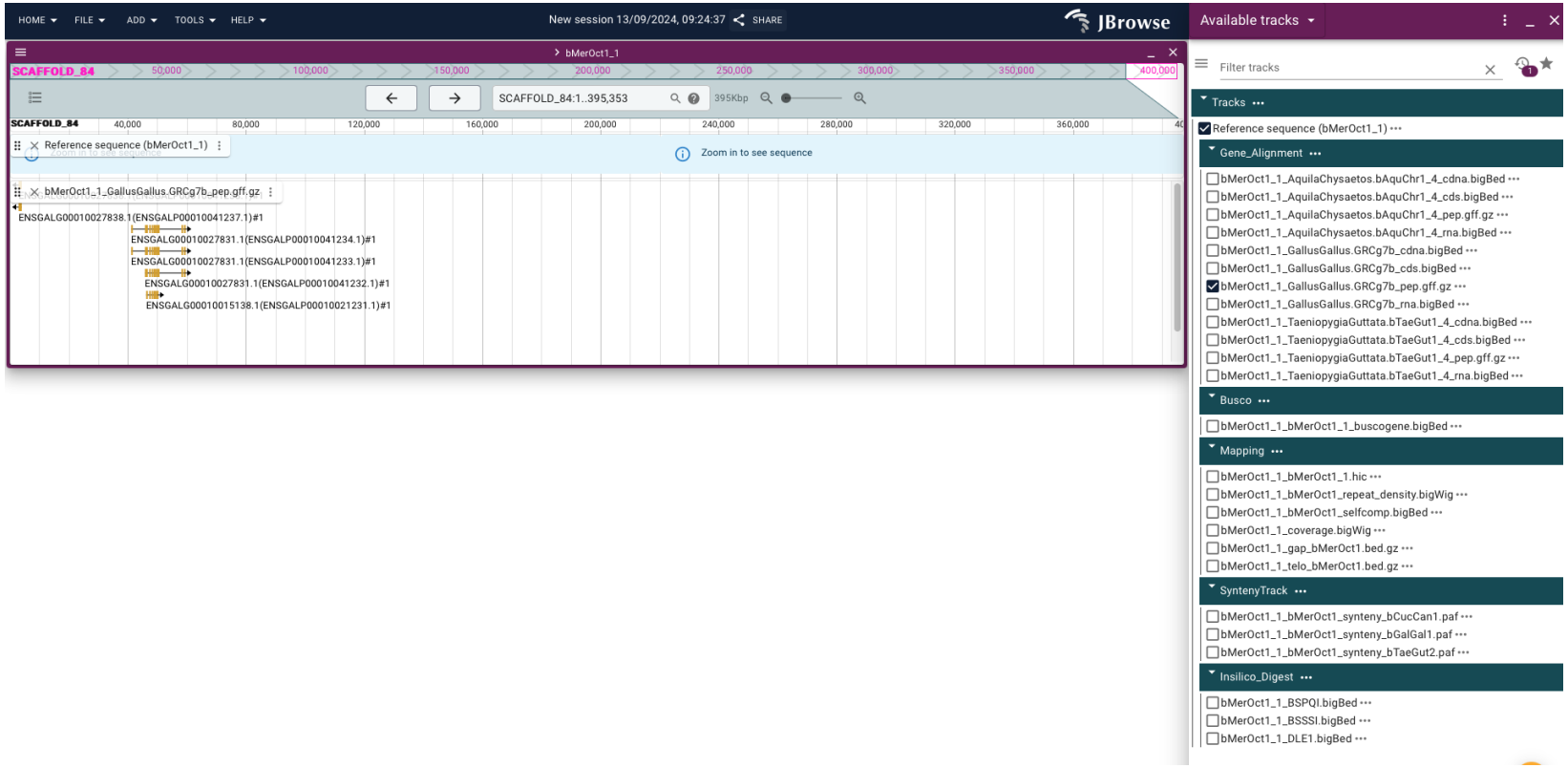
In which order the micros go in the map?

Dot: https://dot.sandbox.bio

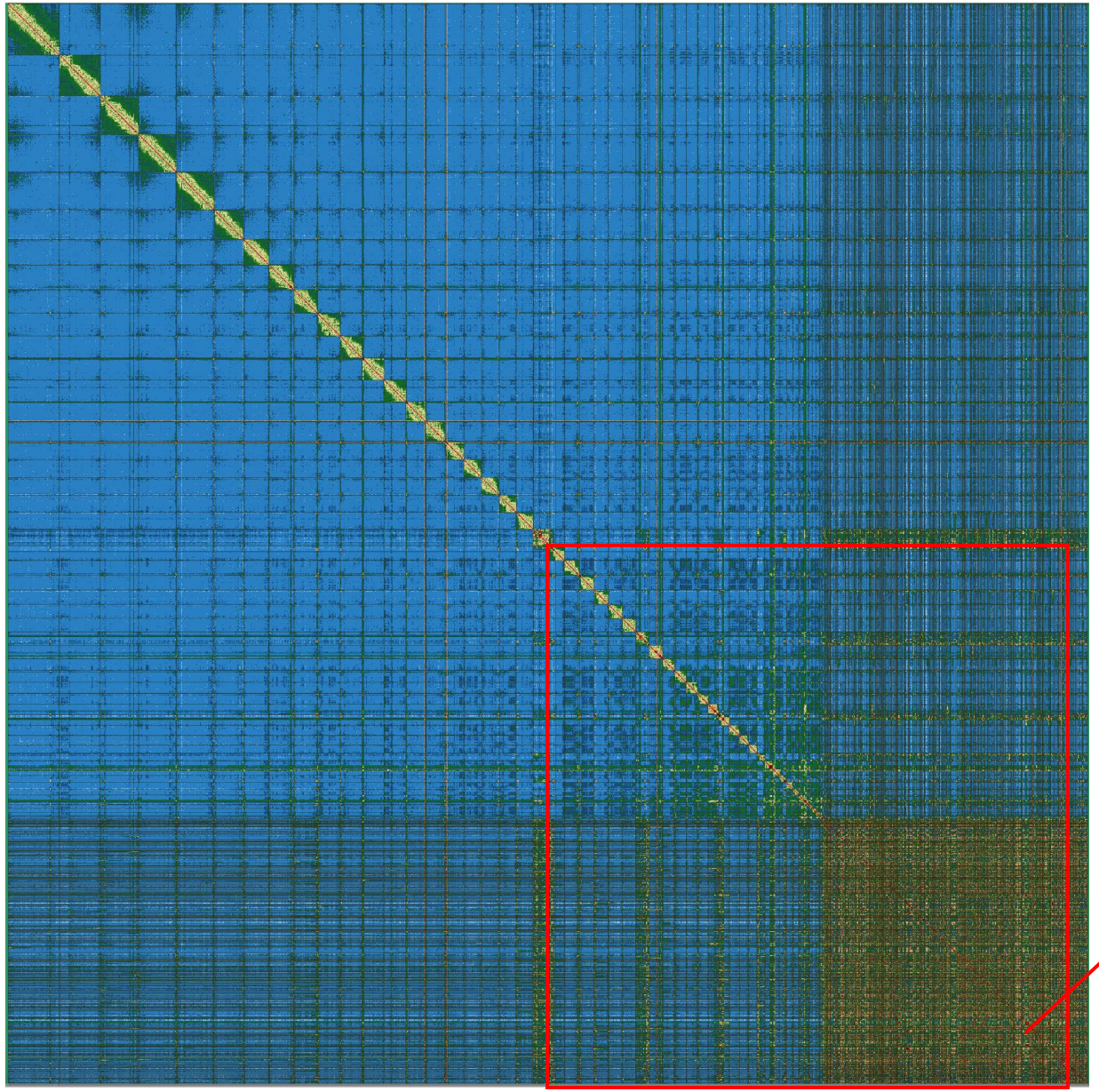
How do we fish out the micros?

(Mostly sharks and reptiles)

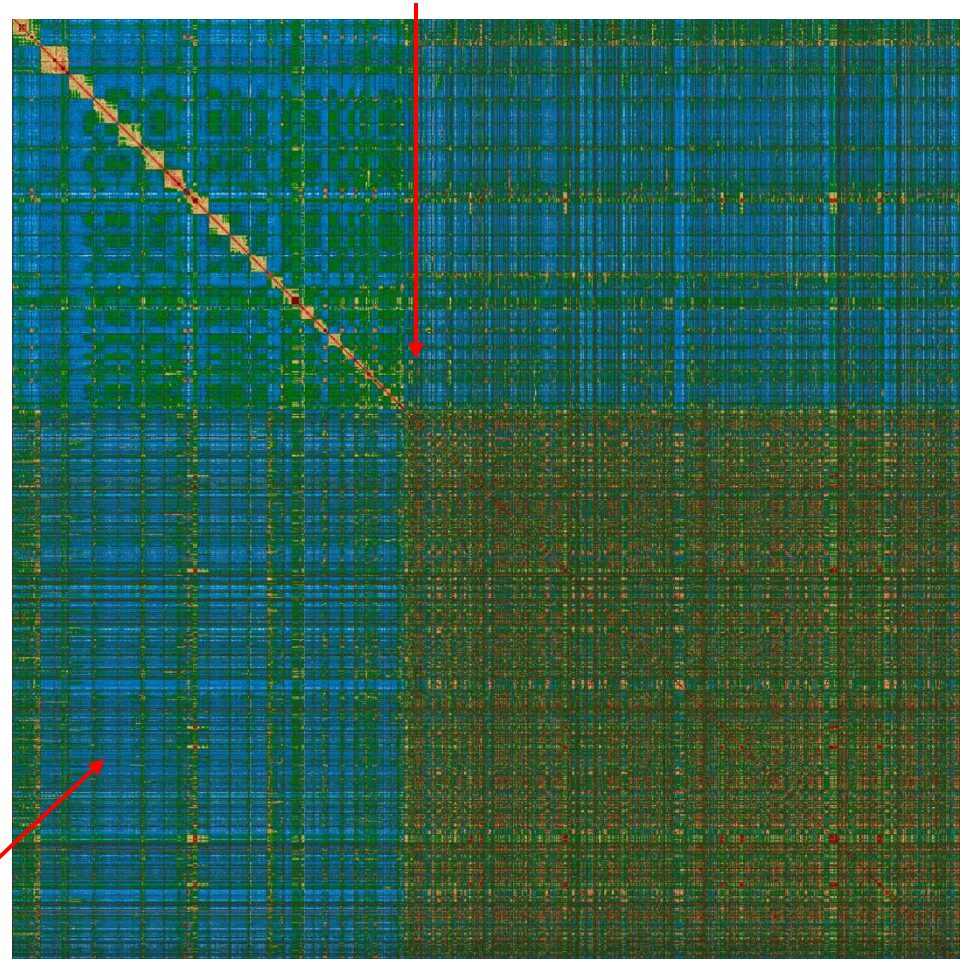
Gene content in Jbrowse



Micro-chromosomes - Sharks



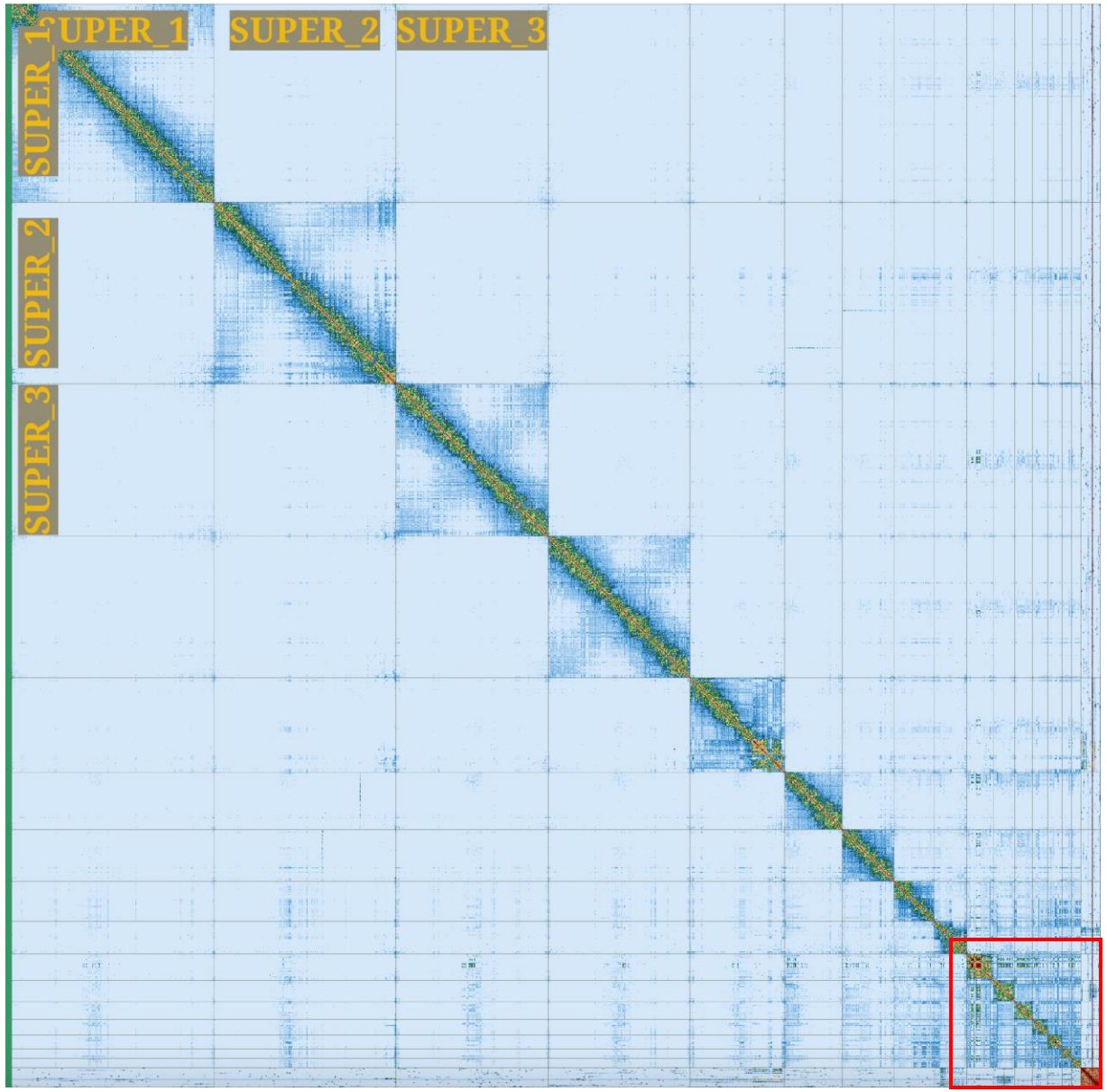
16 micros
Smallest one: < 9 Mbp

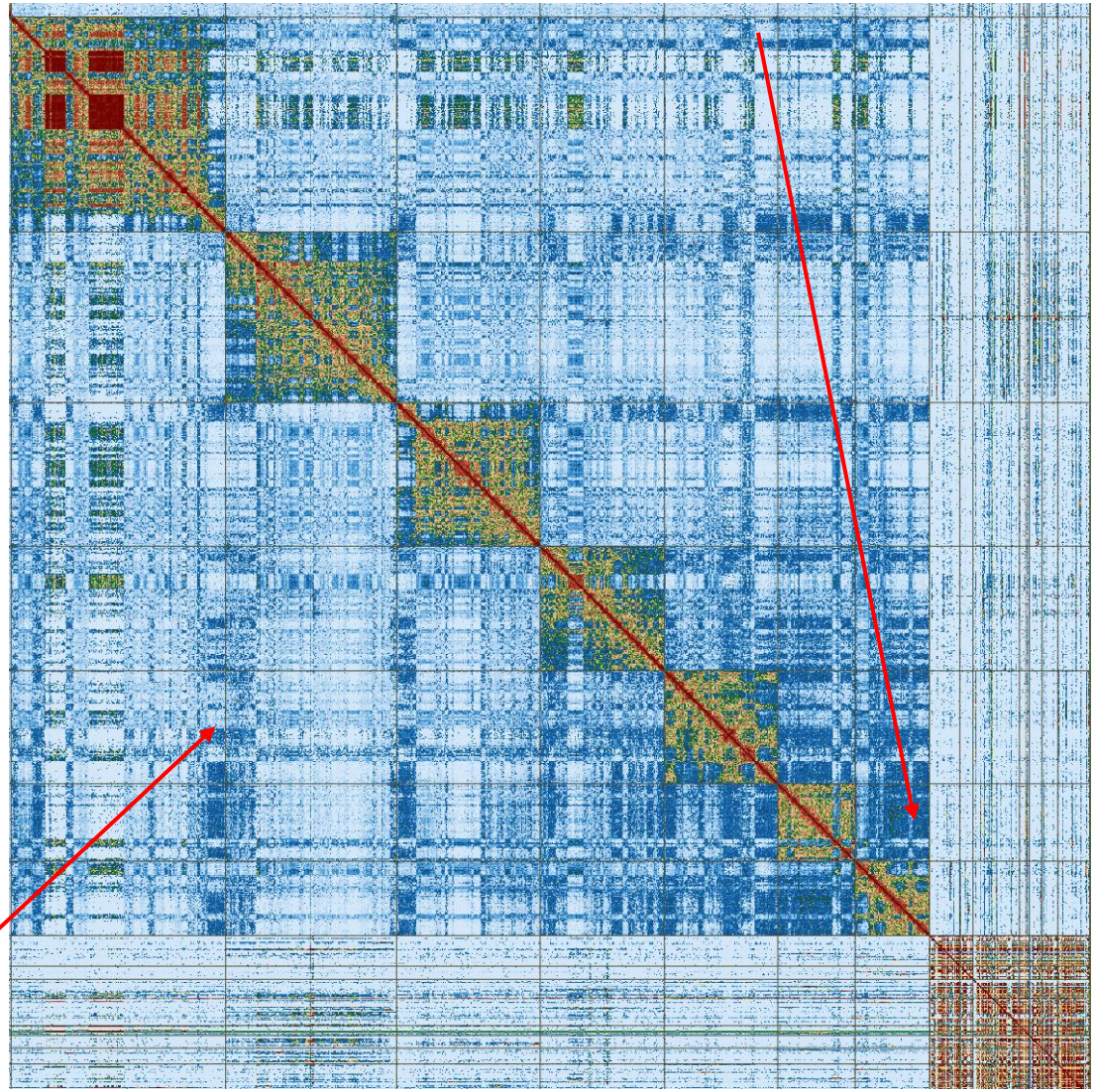


Main challenge:
Usually the most repetitive ones

Micro-chromosomes - Reptiles

7 micros
Smallest one: 12 Mbp





High chromosome number
Poor HiC

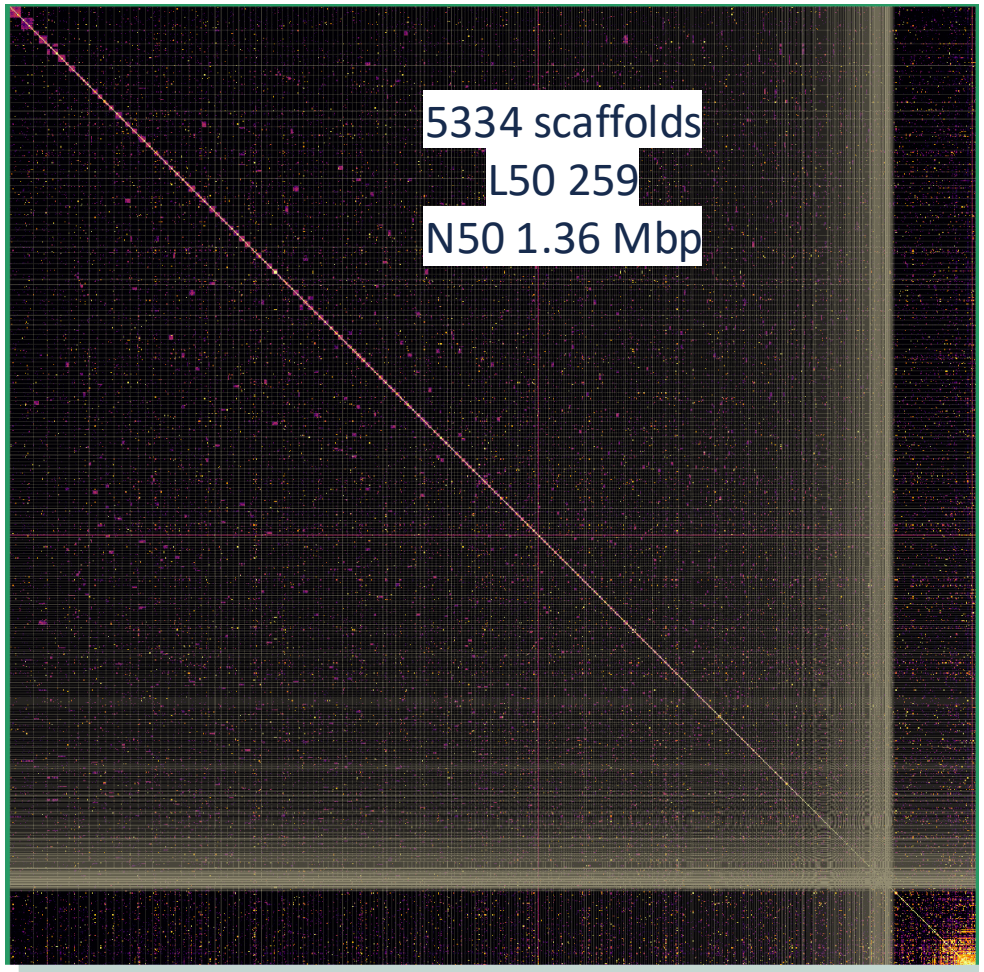
High chromosome number + poor HiC + no telo information

Symbiodinium – 92 chroms protist, no karyotype available or close species with genomic data available

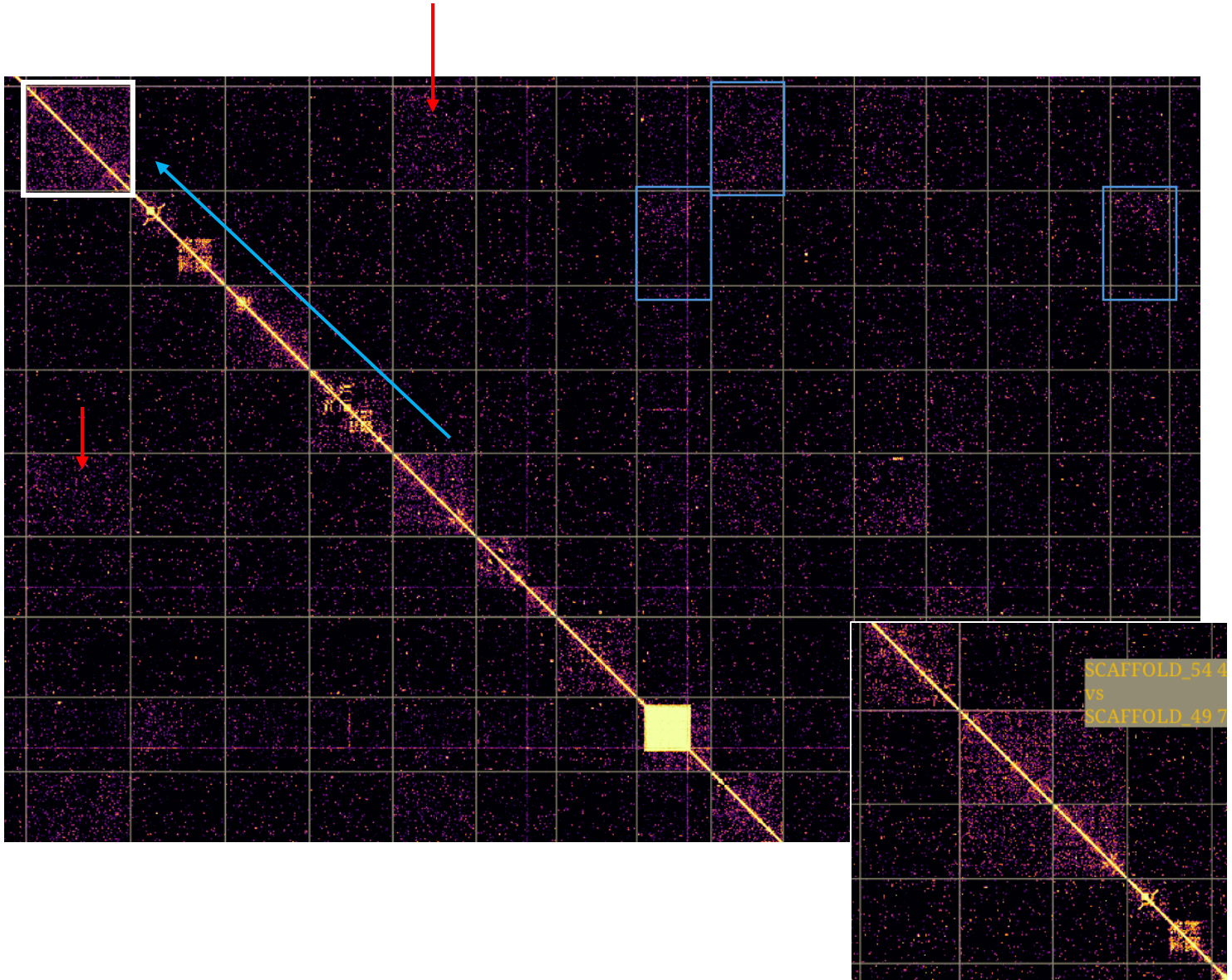
Some approaches:

1. Adjust gamma contrast in Pretext

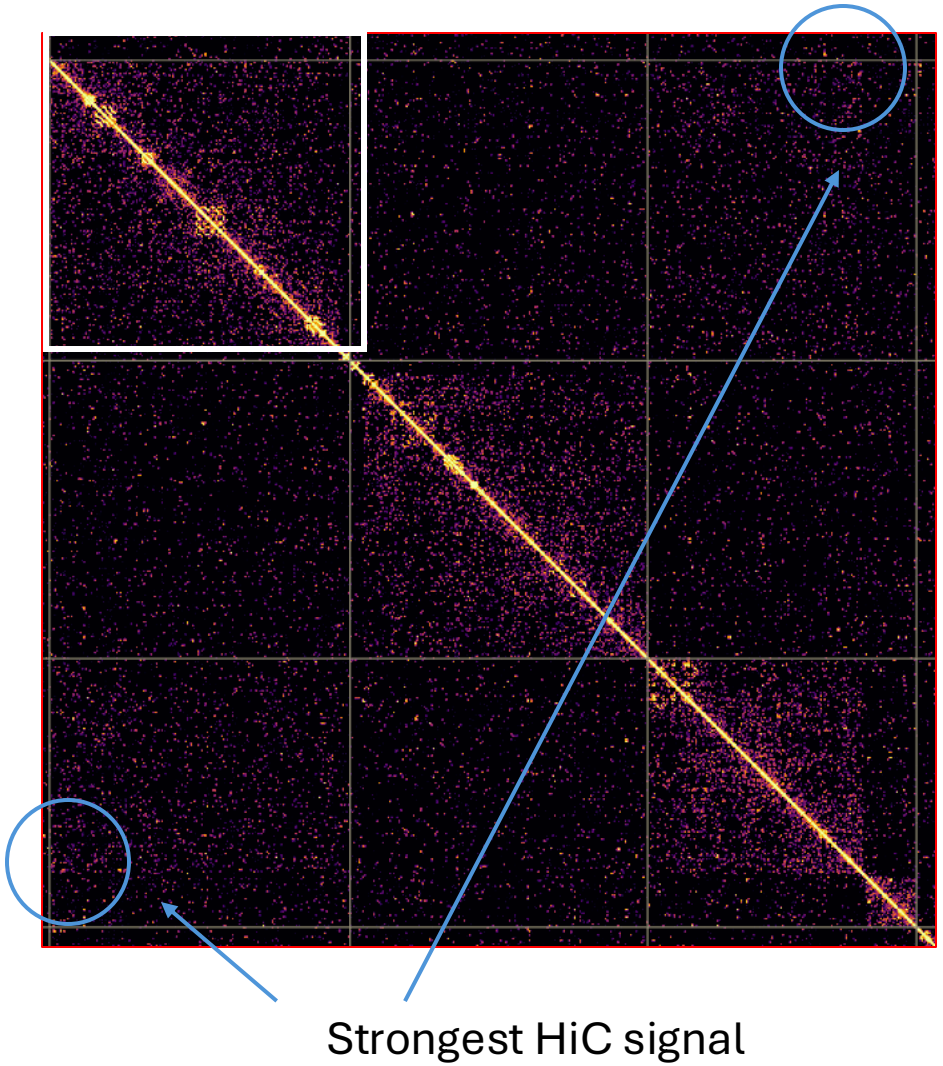
2. Higher possible contrast in Pretext colours, dark background
3. Normal resolution maps
4. Zoom in
5. Use a comparator (when available)
6. Telomere track
7. Top up



High chromosome number + Bad HiC + no telo information



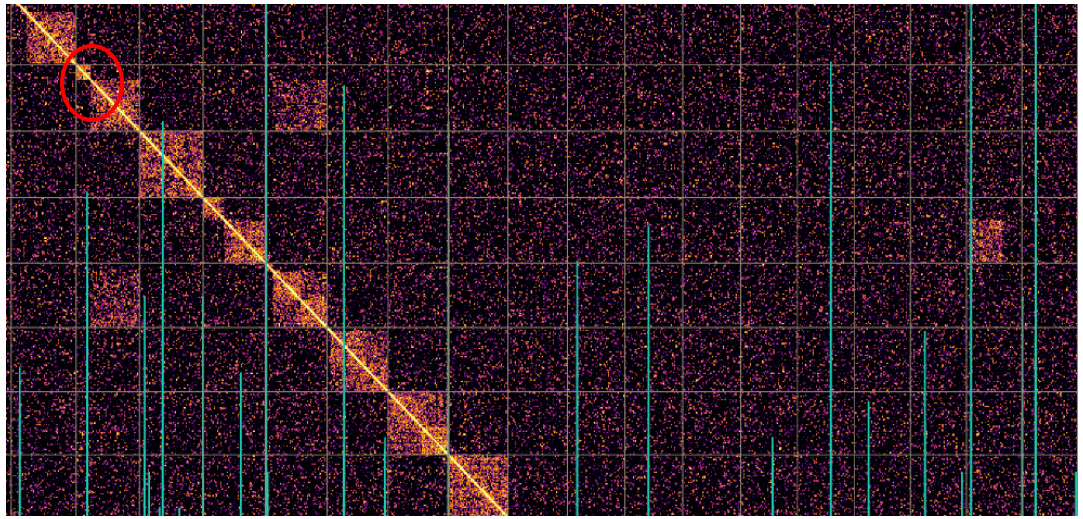
After zoom in



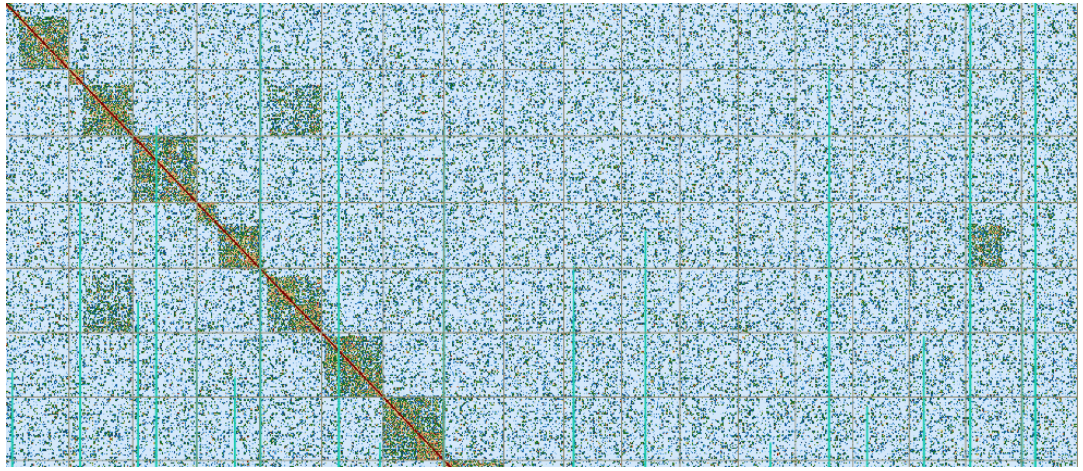
Strongest HiC signal

High chromosome number + Bad HiC + no telo information

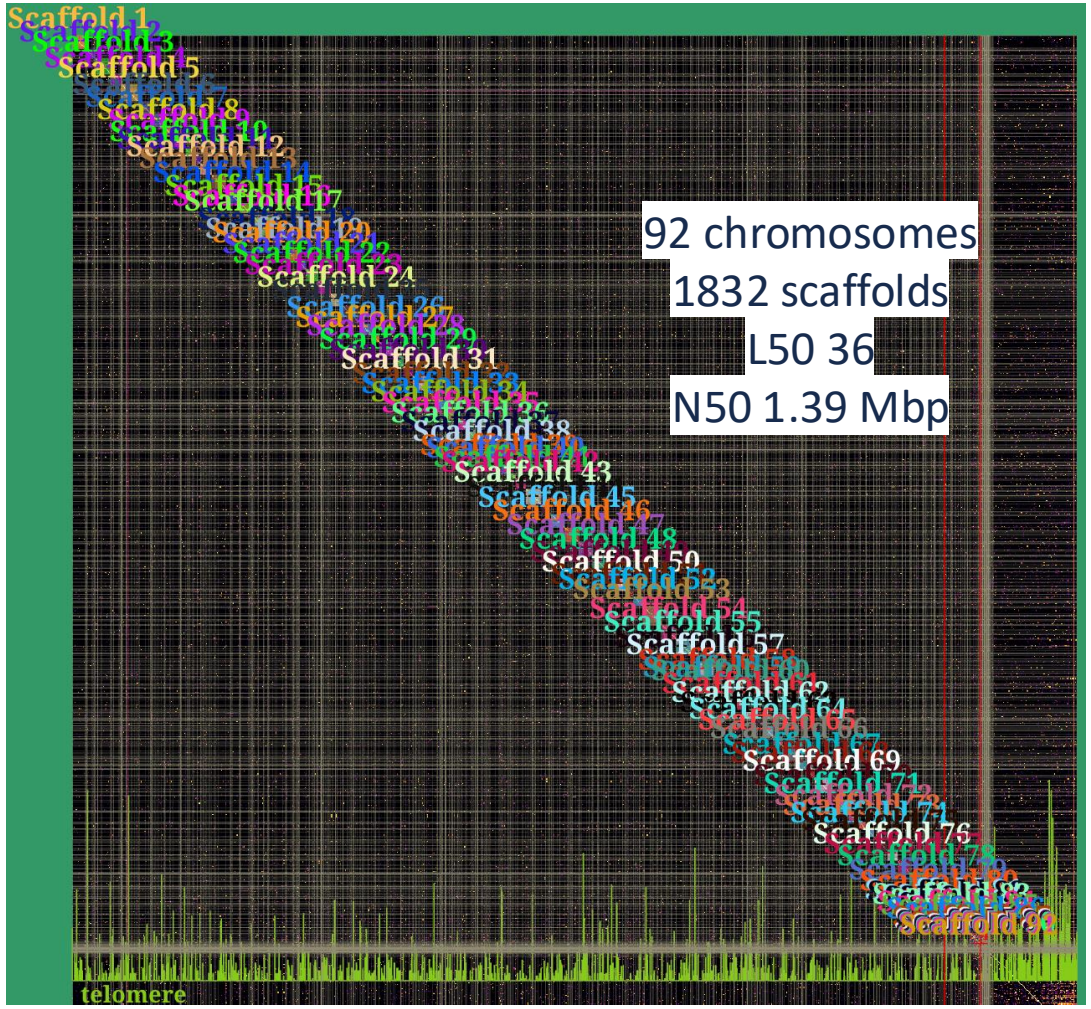
Gap track may also be helpful for breaks and cuts



Difference in the HiC signal visualization



Main source: HiC signal
High contrast colours sets (darker background helps)

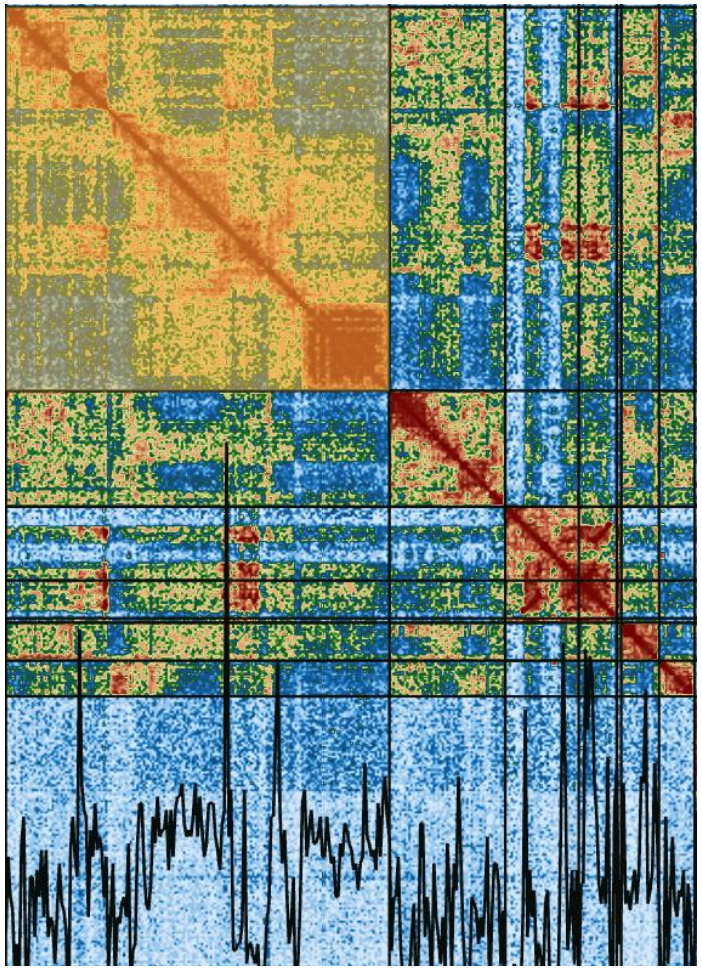


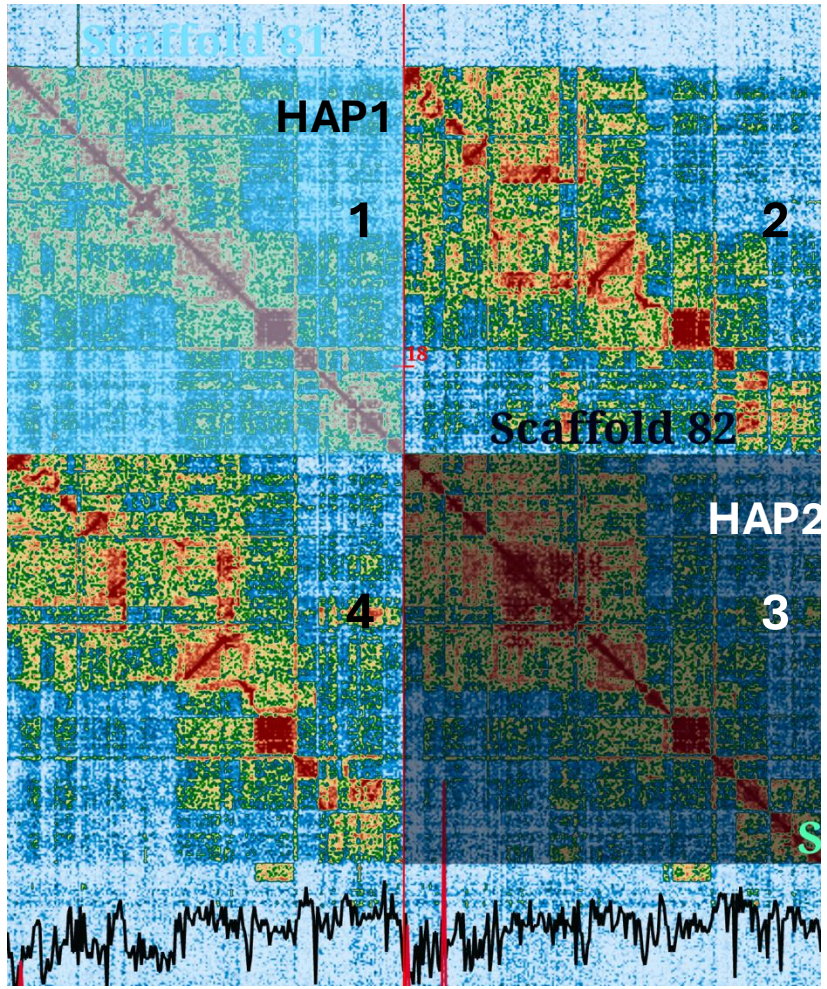
Symbiodinium curated map

Haplotype phasing

Haplotype phasing

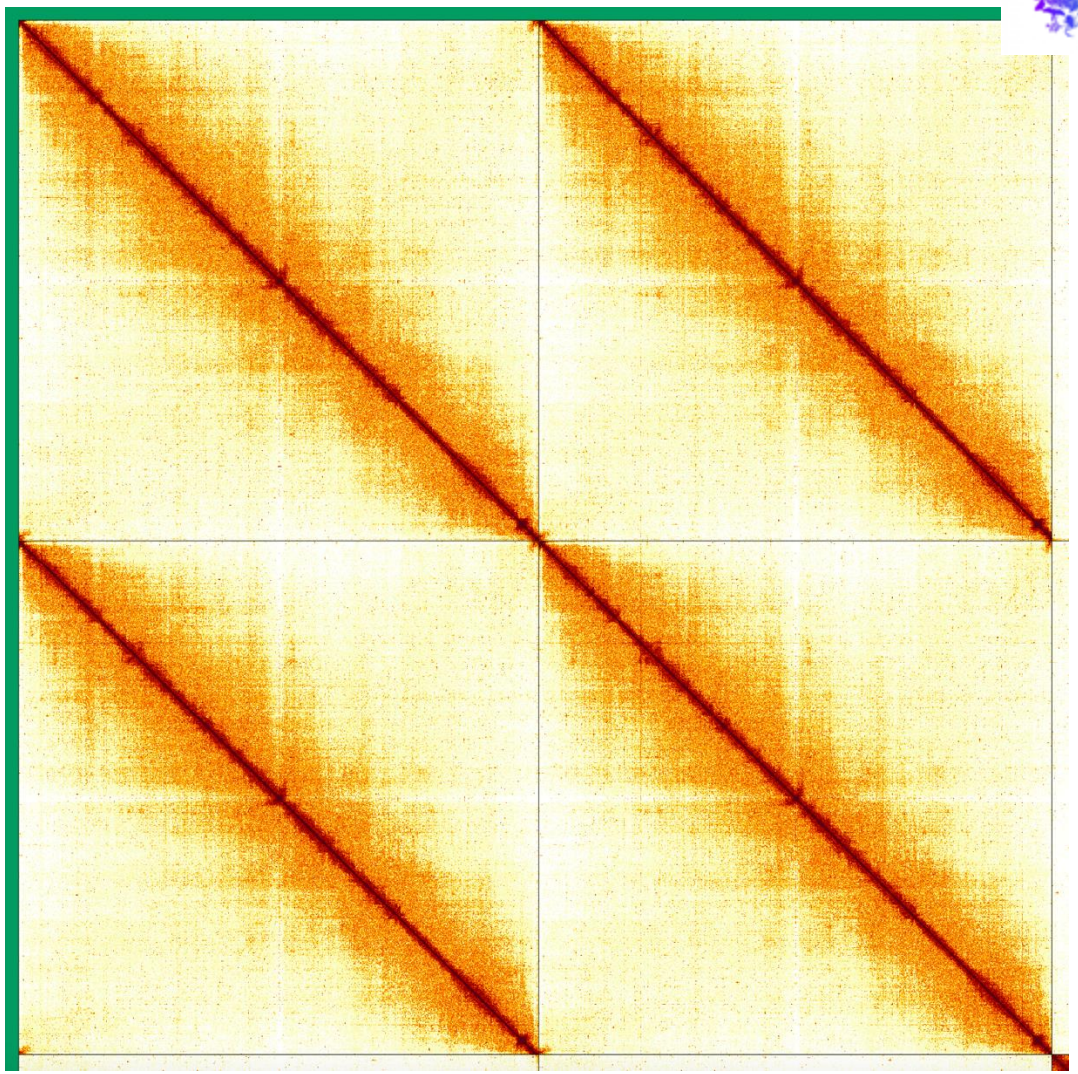
- Repetitive regions
- High amount of retained haplotigs and inversions inversions





Haplotype bad phasing

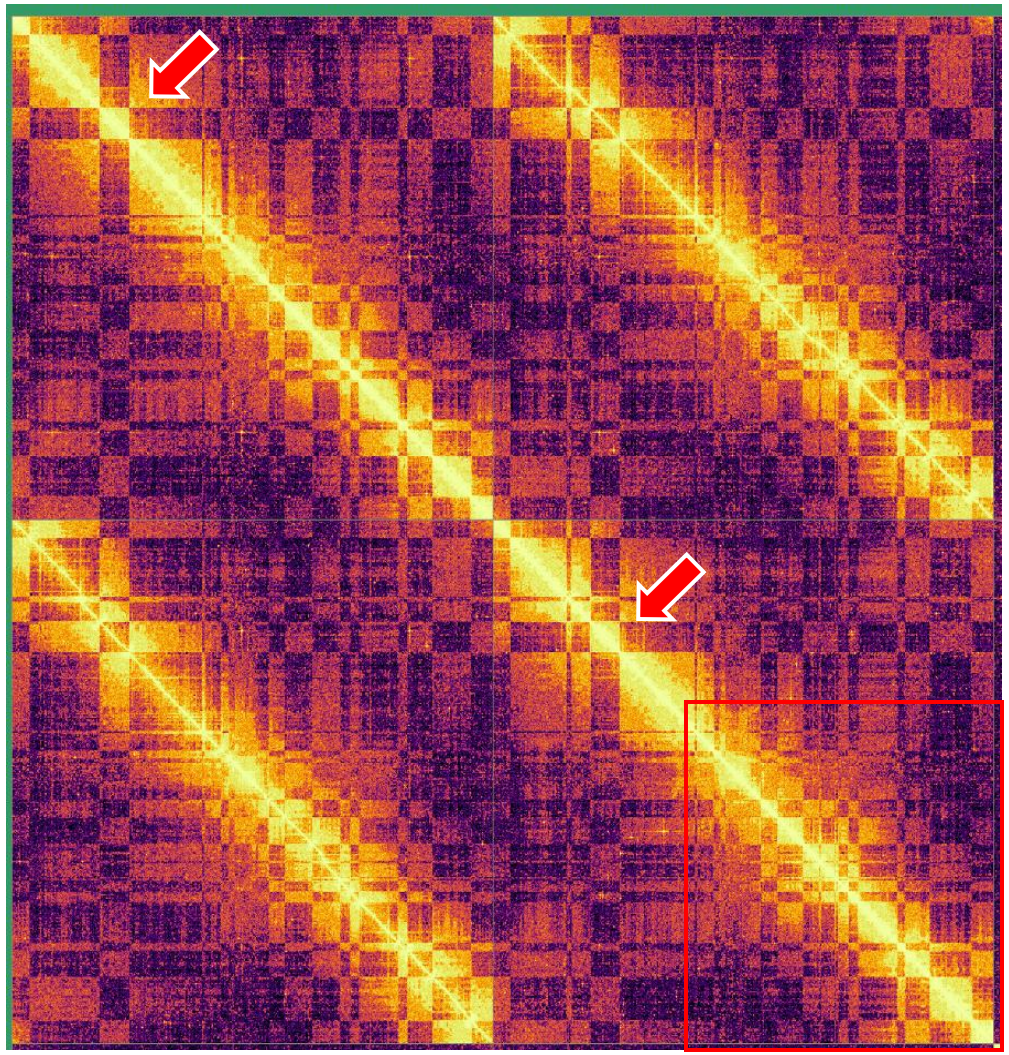
This is how the map should look like when phasing worked well



Contiguous HiC signal
No blanks or weak signal regions

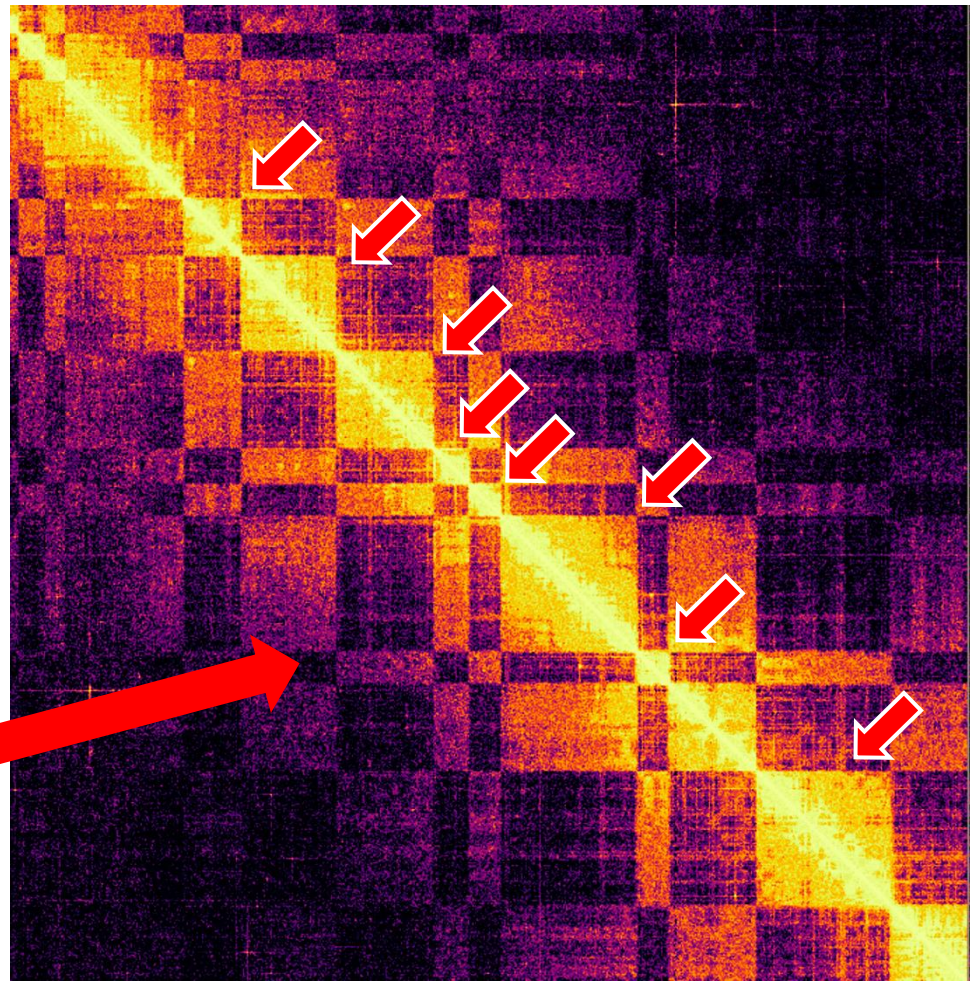
dcOxyDigy1

Haplotype bad phasing



xbMysUnda1

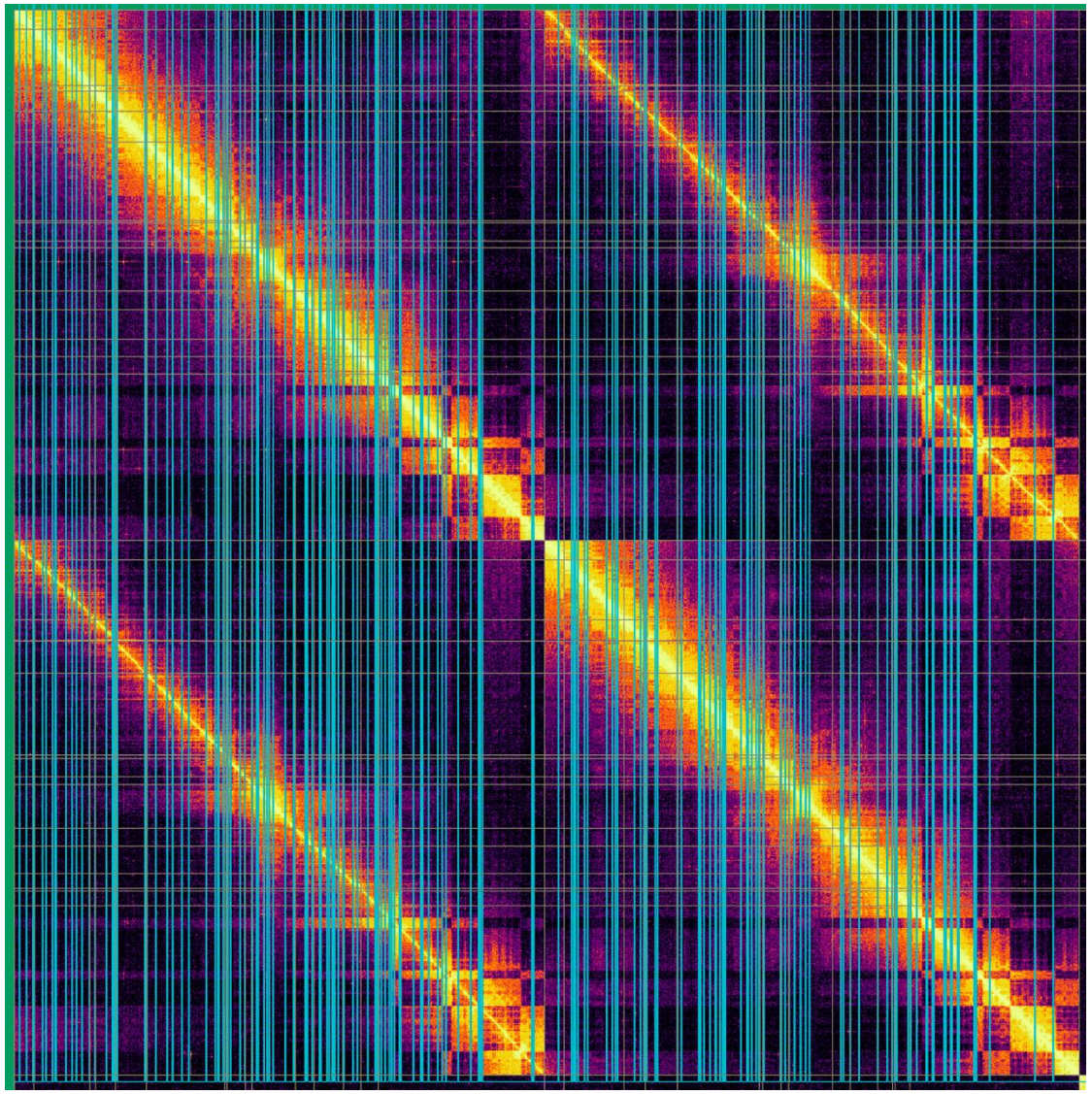
But when it doesn't work is a source of confusion…



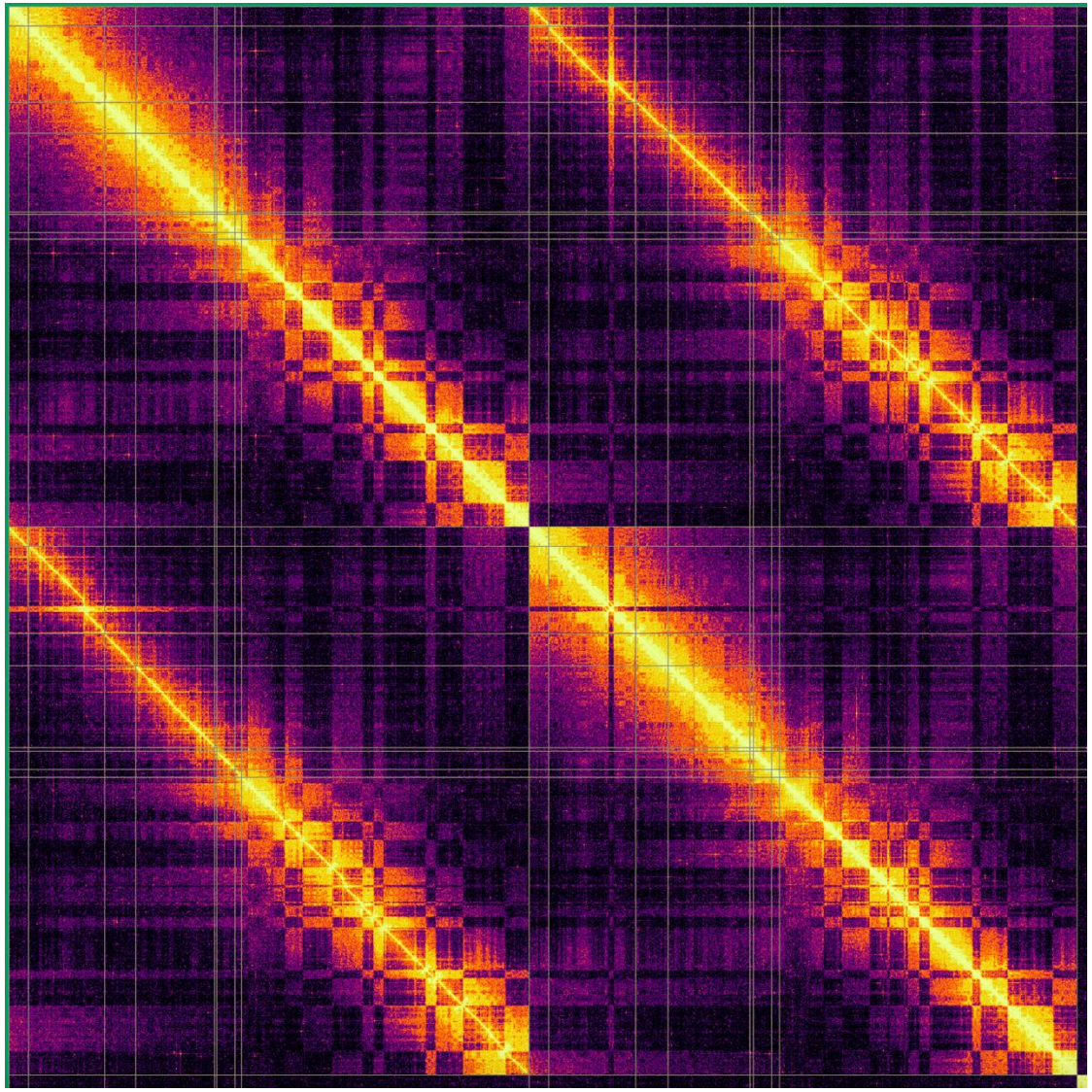
Haplotype bad phasing

Swap bits between the haplotypes

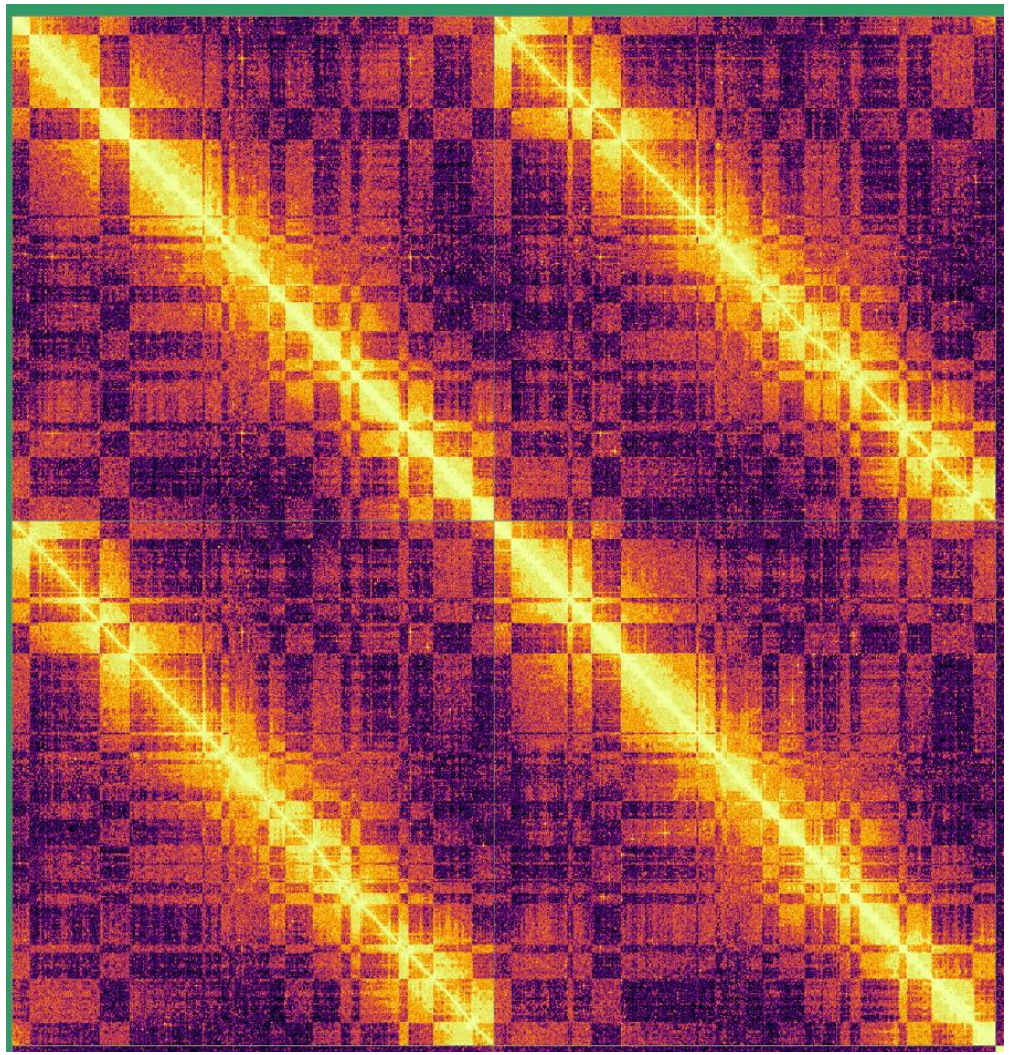
Gap track on



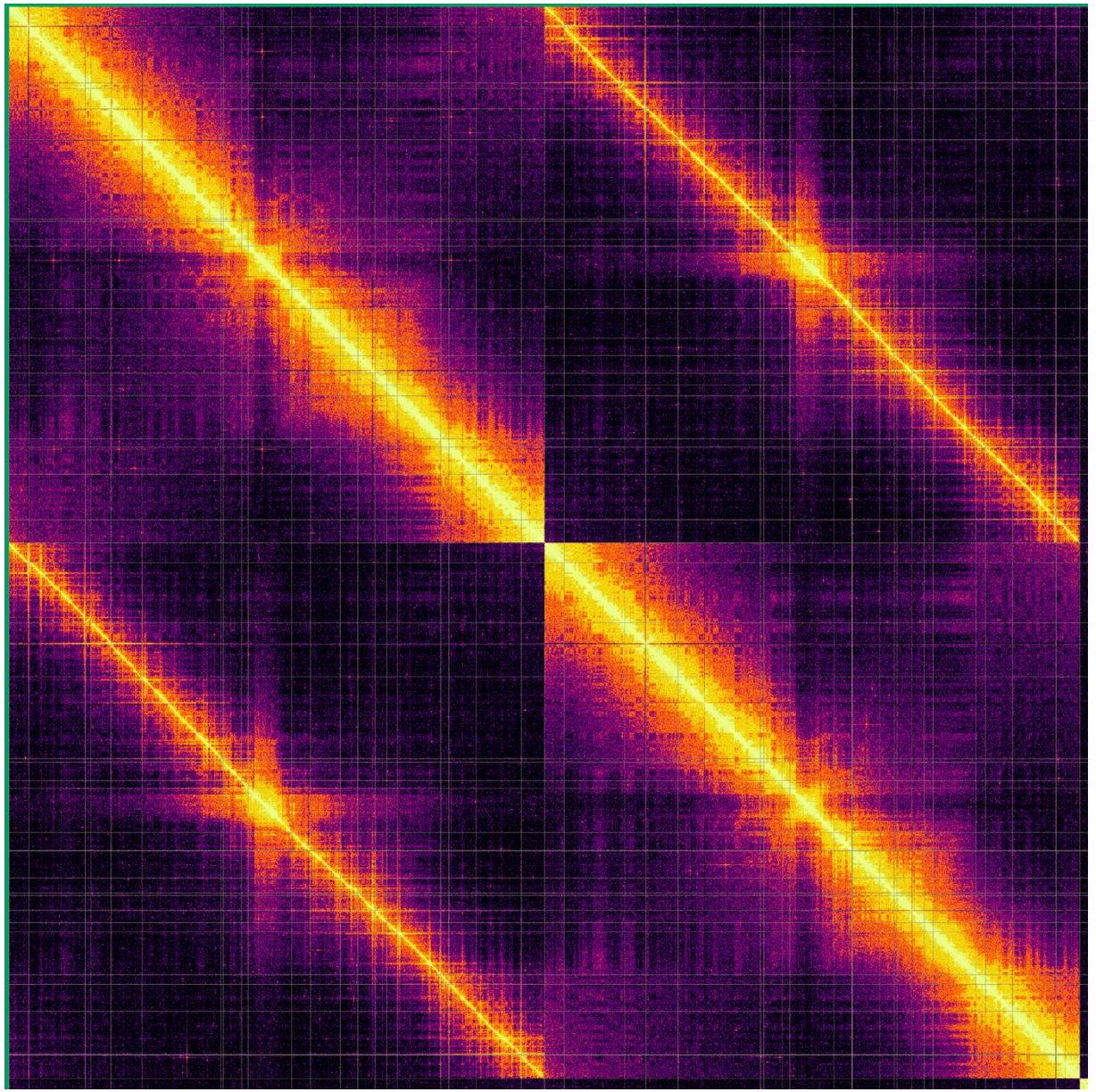
Manual phasing in progress



Haplotype bad phasing

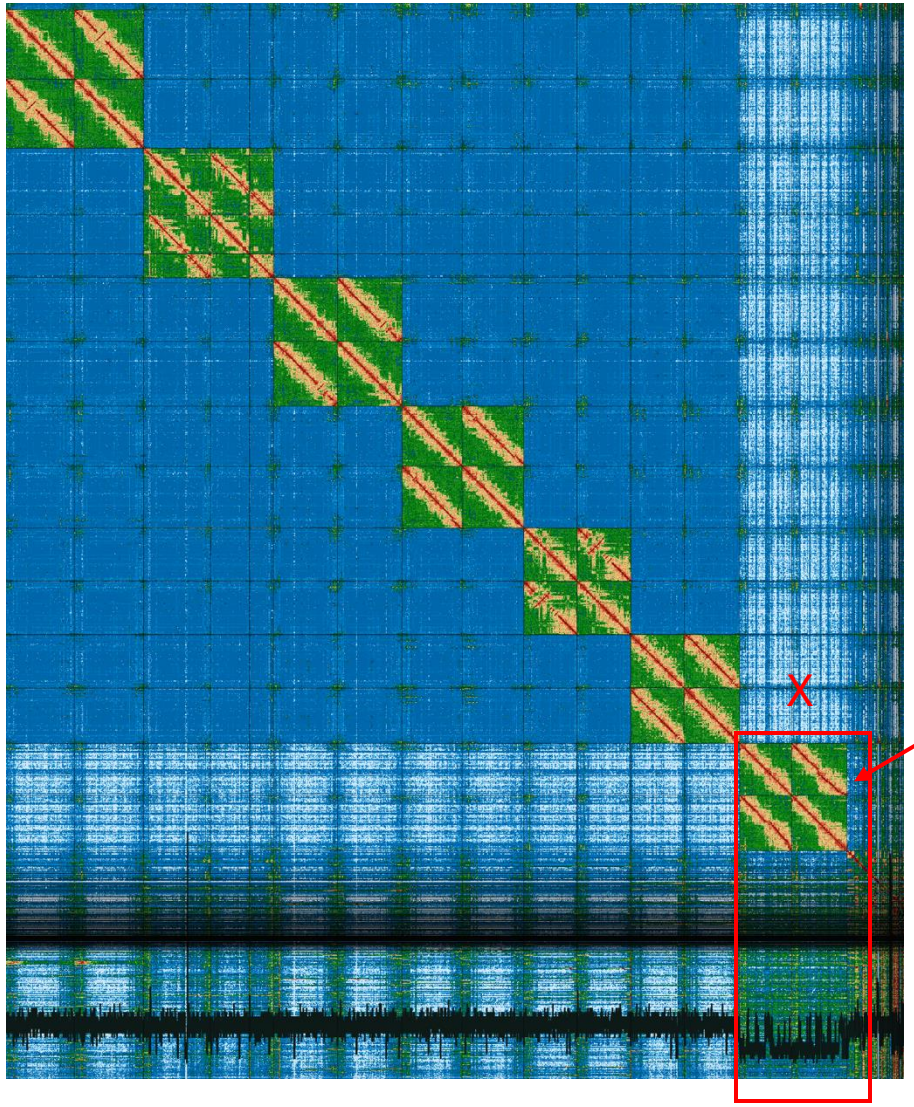


After manual phasing



Haplotype bad phasing – Sex Chromosome

XO species



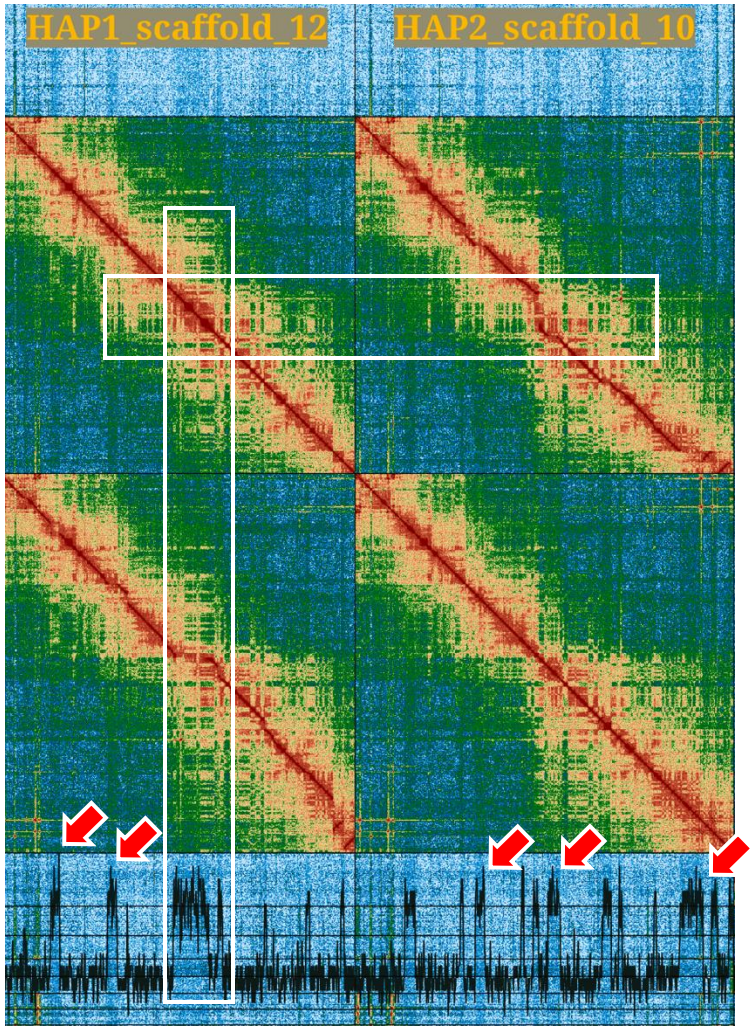
Assembly artifact

Use X from primary assembly

Higher coverage collapsed regions

Regions present only in hap 1 and not in hap2 and vice-versa

ioSymSang1



Should not be half coverage → False duplicated X

All haplotypes assembly curation

Standard Pipeline Assembly

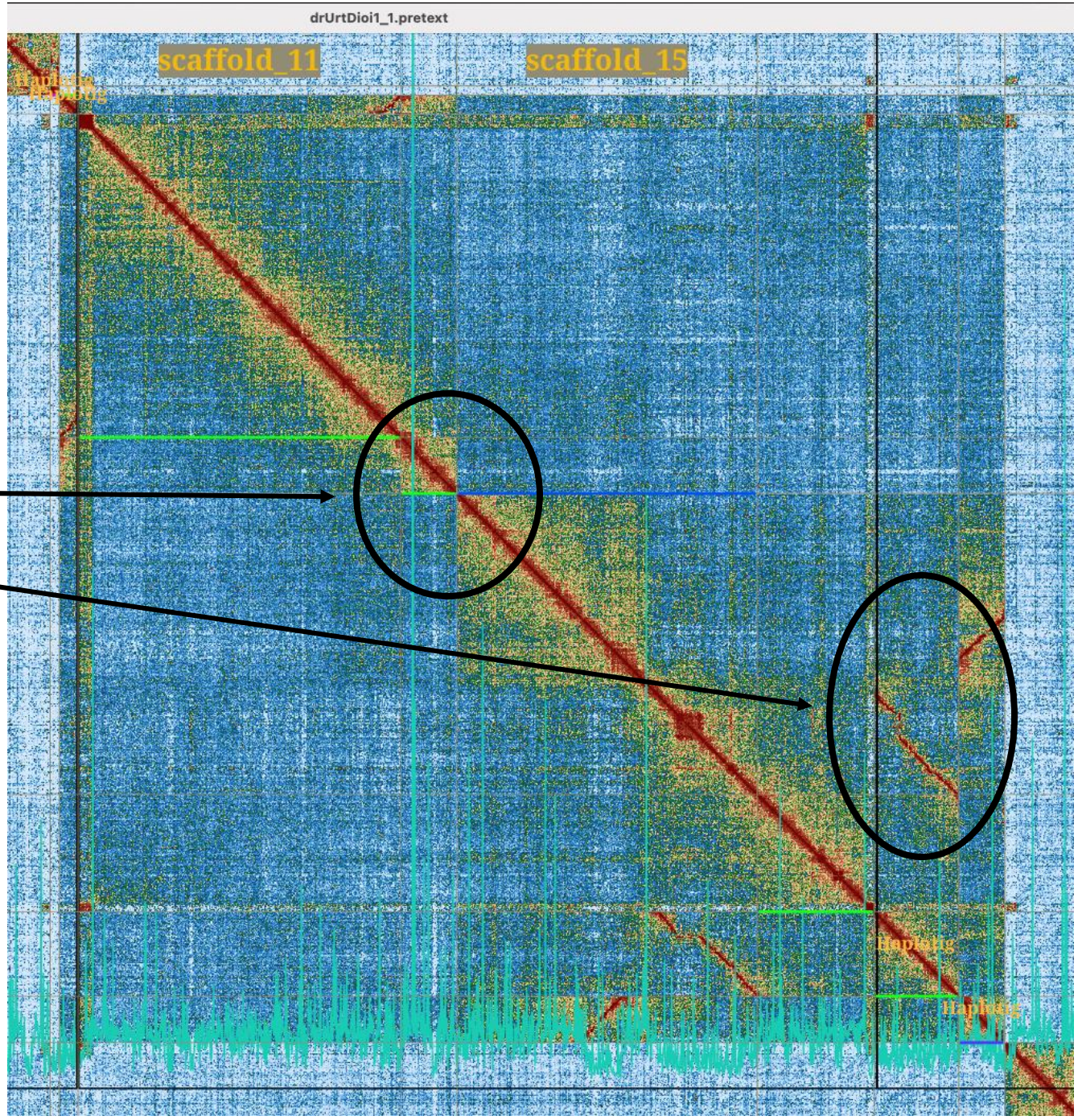
By Dominic Absolon

drUrtDioi1 – tetraploid

Initial “primary” assembly had issues:

- Missing sequence
- Over-represented sequences

Primary assembly:
hifiasm (w/ purging) + purge_dups + hicmapping + yahs

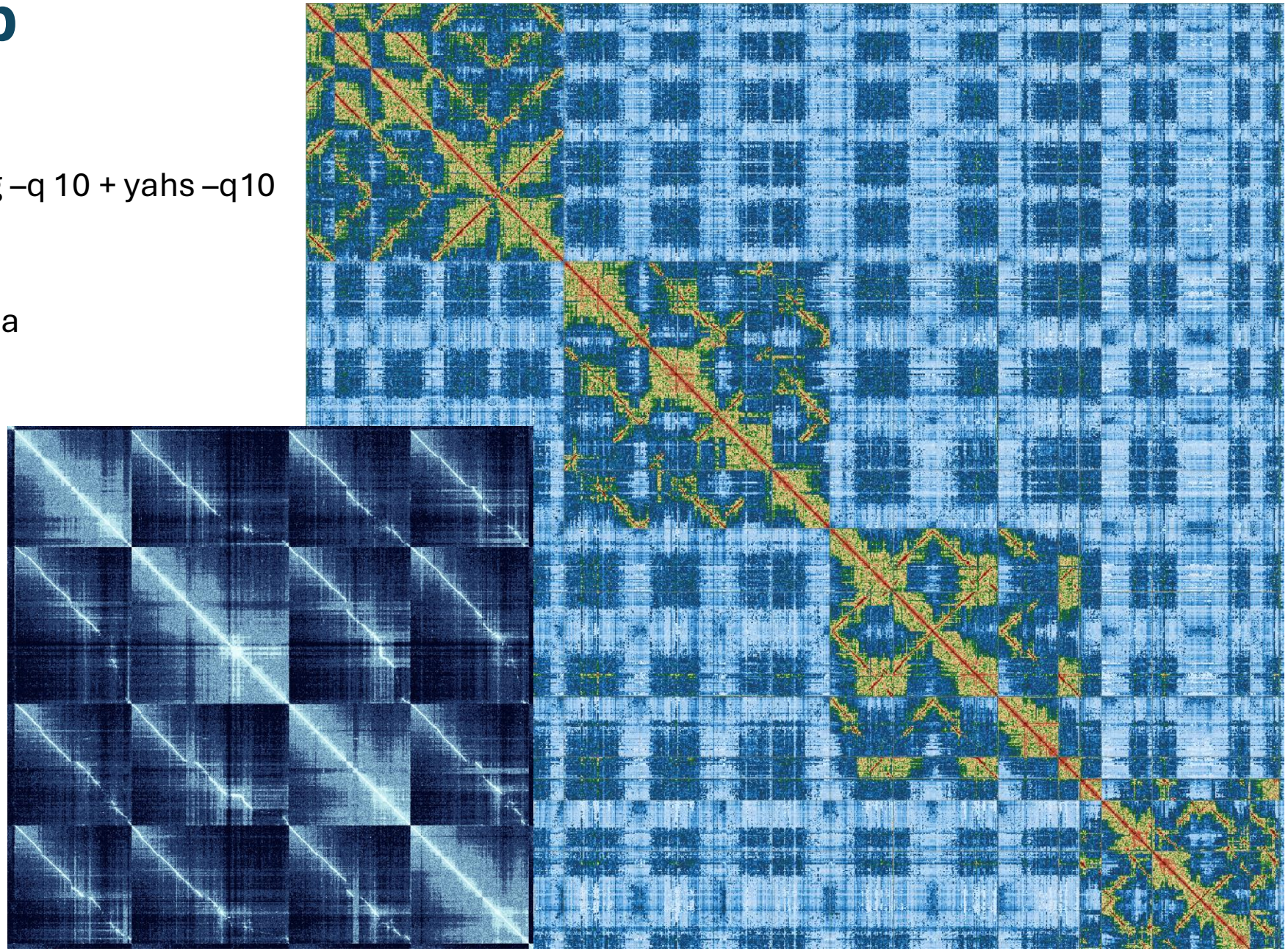


All haplotype map

By Dominic Absolon

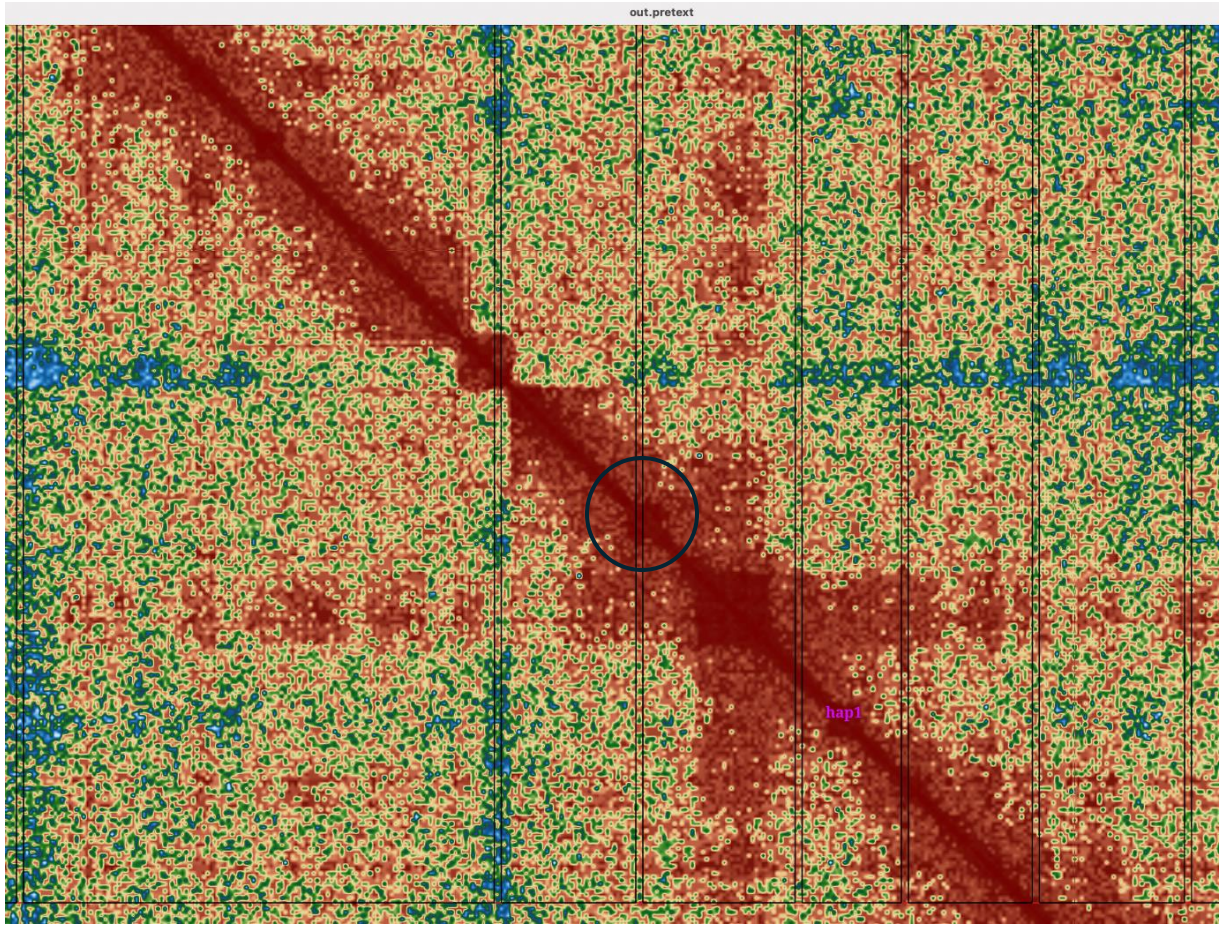
All haplotype assembly:
Hicanu all contigs + hic-mapping –q 10 + yahs –q10

Have the capabilities of outputting a curated fasta for each of the haplotypes

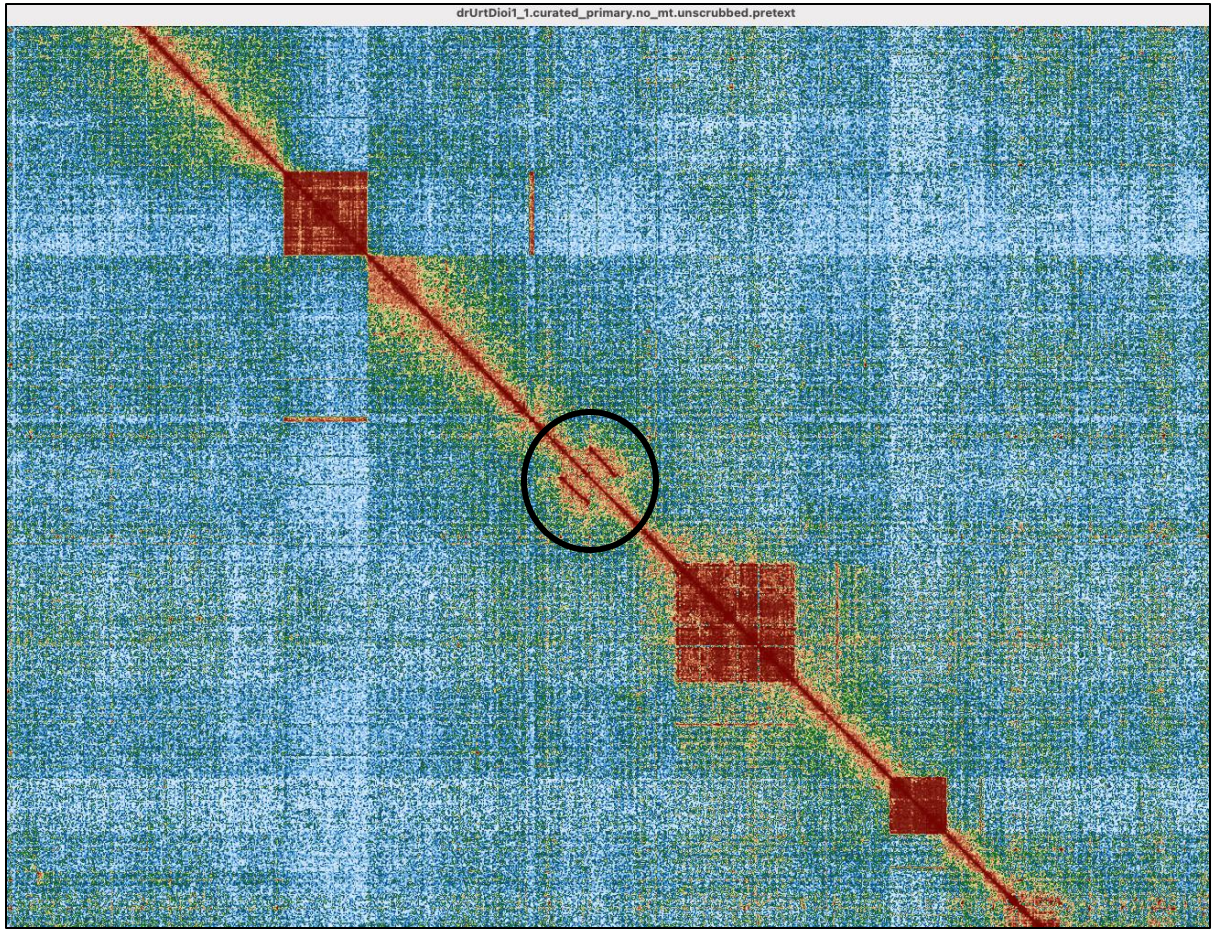


All haplotype X Single haplotype maps

Curating – all 4 haps – hi res



Curated – 1 single hap - hi res



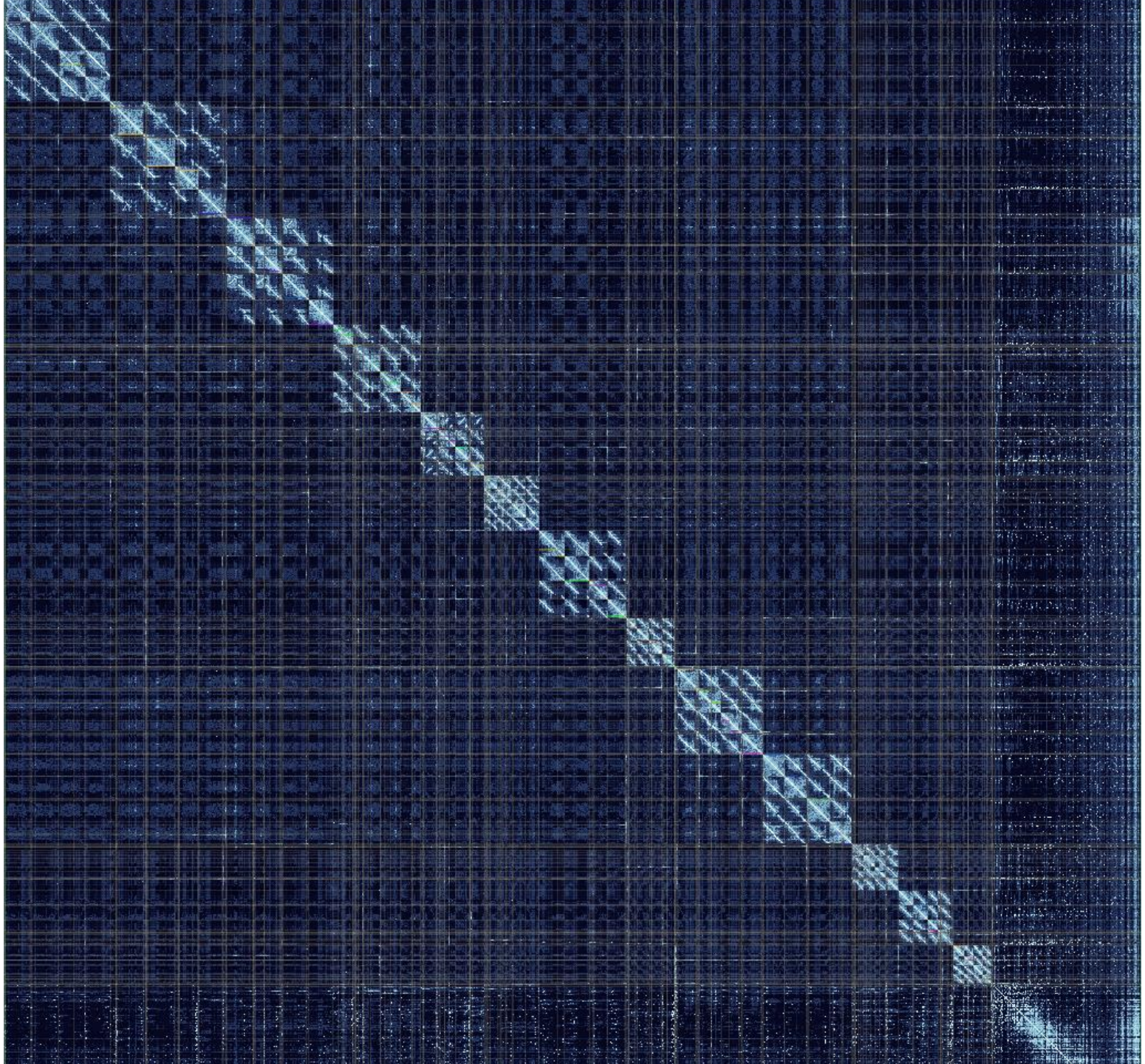
By Dominic Absolon

Resolution split among all haps

1 hap mapped against all HiC dataset

A working approach:

By Dominic Absolon



wellcome sanger institute

Polyploid genomes

Polyploid genomes

- Main issues:

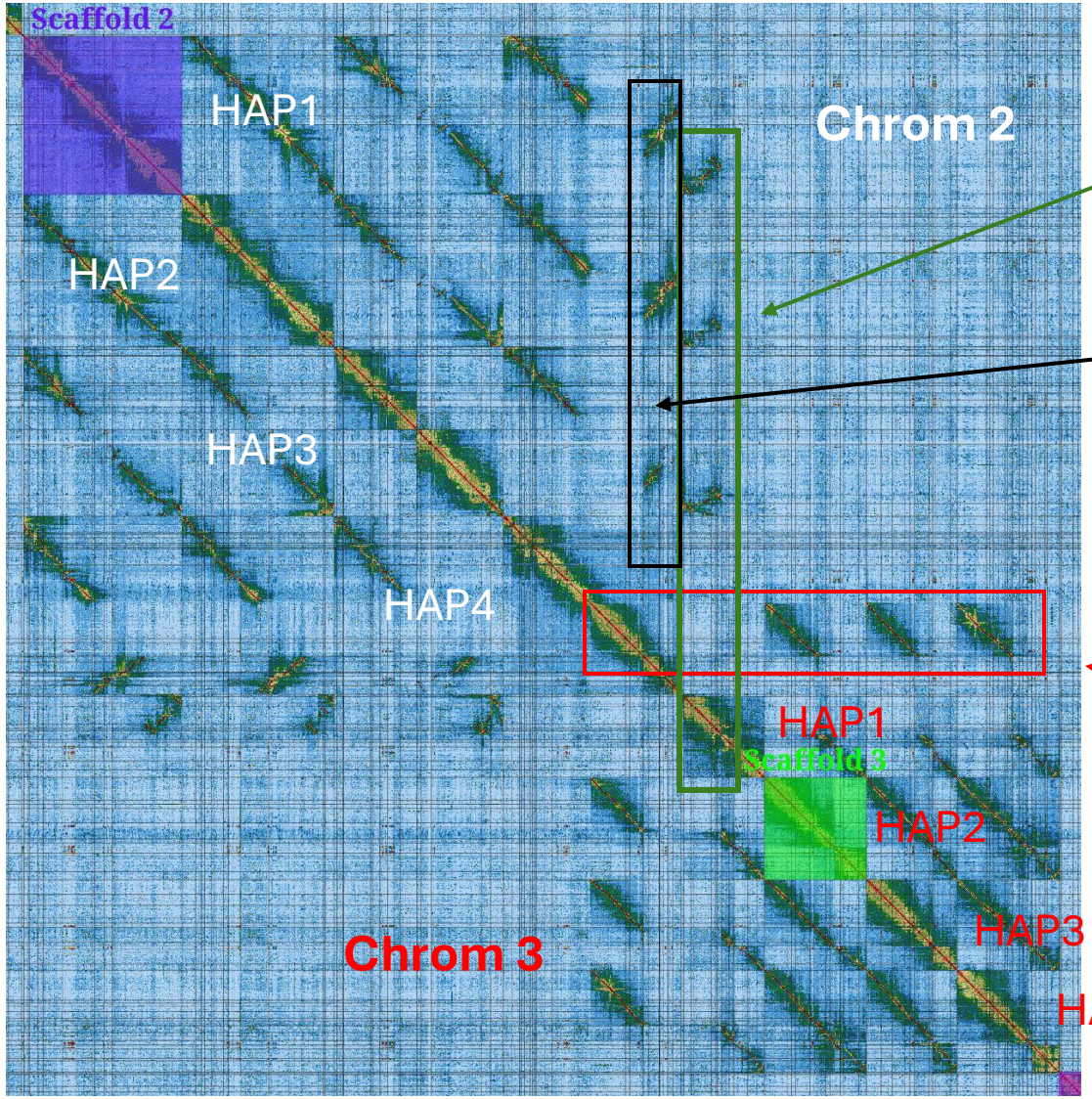
1. Polymorphism among haplotypes
2. Translocations between regions of different chromosomes
3. Curated fasta mapping to the whole HiC dataset in the curated map

hap1 fasta only

Differences (look like errors) in the curated map

Polyploid genomes

Translocations between regions of different chromosomes



Chrom 3 affinity with one region of chrom 2 in 3 haps (except HAP4)

Duplicated region in chrom 2

Chrom 2 affinity with one region of chrom 3 in 3 haps (except HAP1)

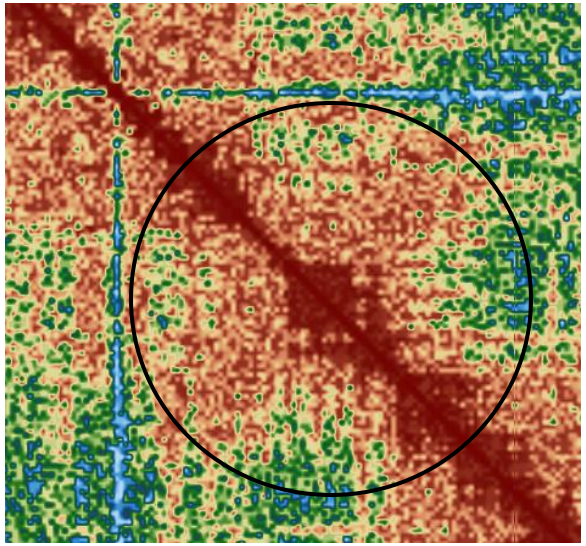
Only 1 hap will be in the final curated fasta
Translocations will not show

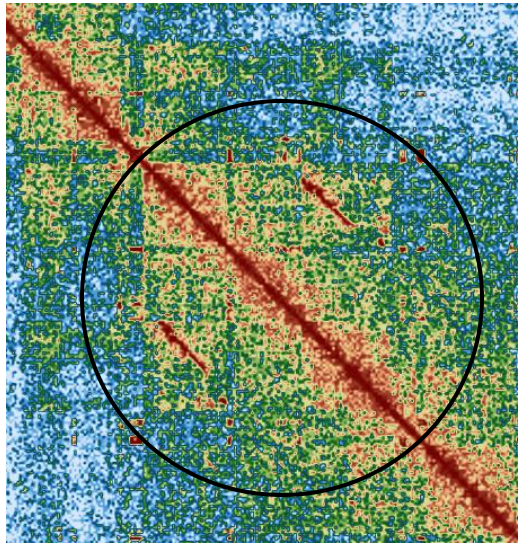
wgTheLage1

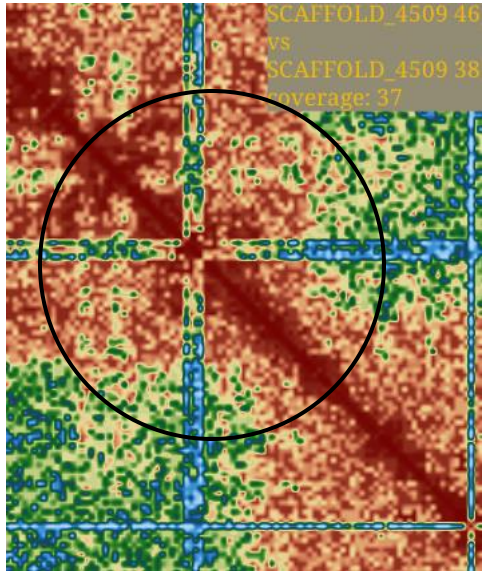
Polyploid genomes

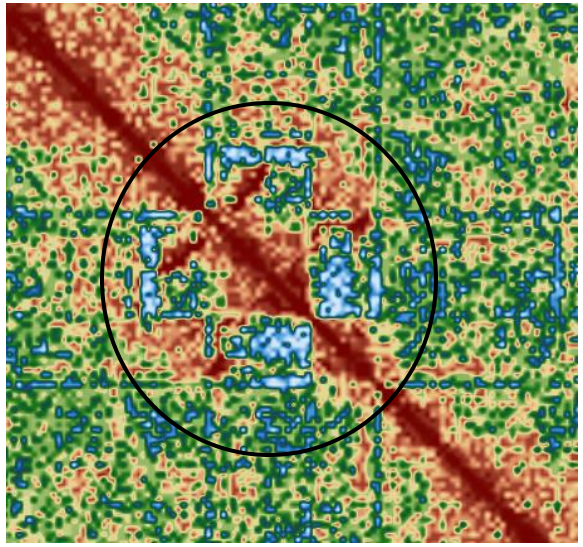
All set of HiC data mapping to one hap fasta only. Differences (look like errors) in the curated map

All haplotypes map









One hap only curated map

Allopolyploid genomes (Polyploids or not)

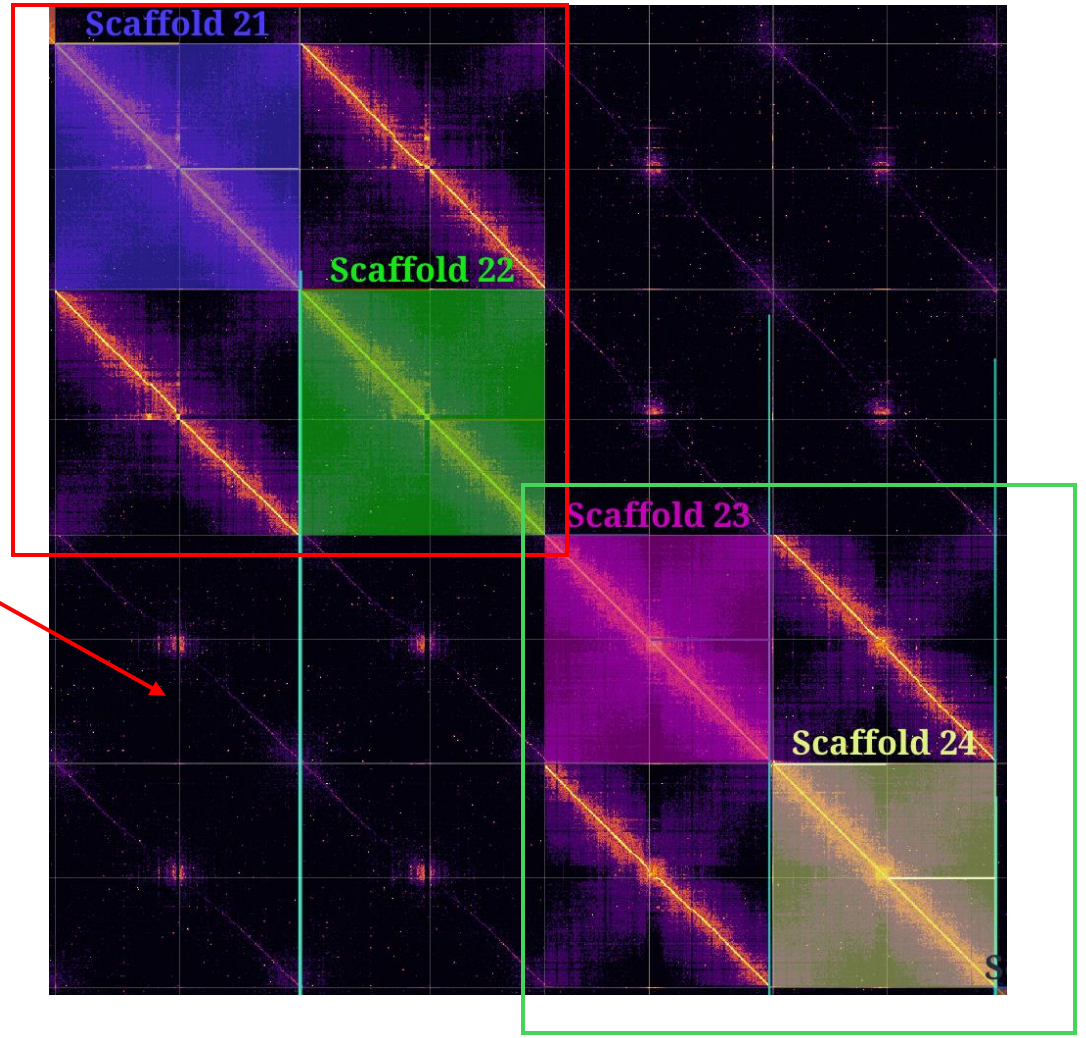
Two or more complete sets of chromosomes from different species

Dual curation

Species A

Chromosomes from different species

Weak parallel lines
Low homology



Species B

2N = 28
N = 14

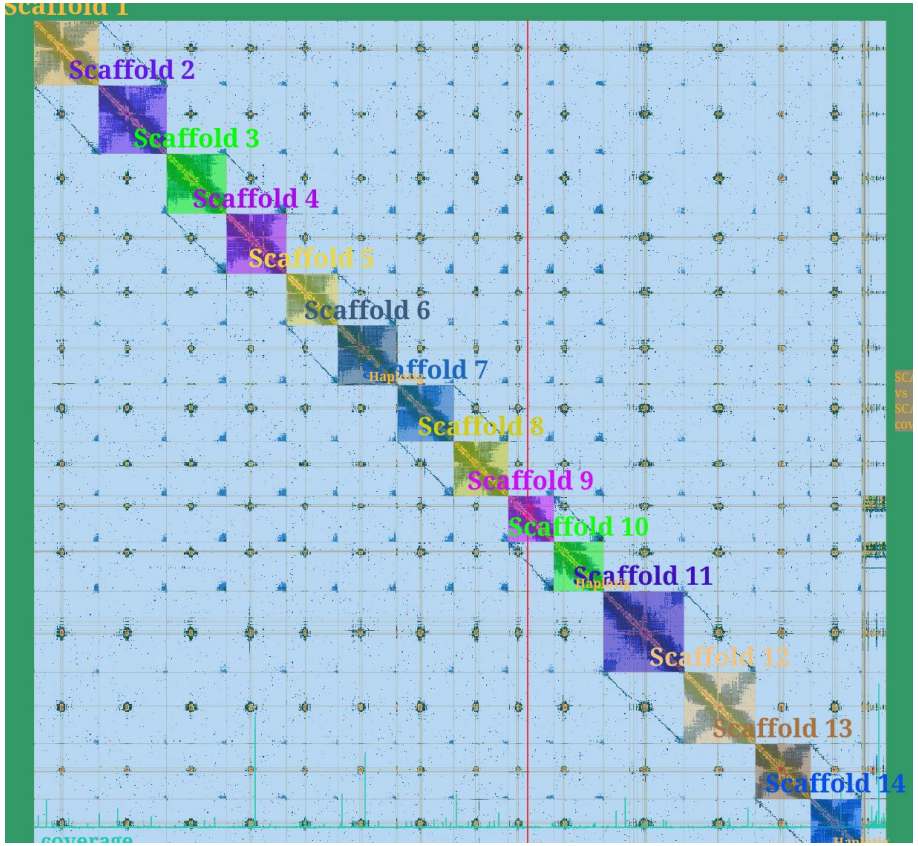
Only 1 hap is submitted

Allopolyploid genomes (Polyploids or not)

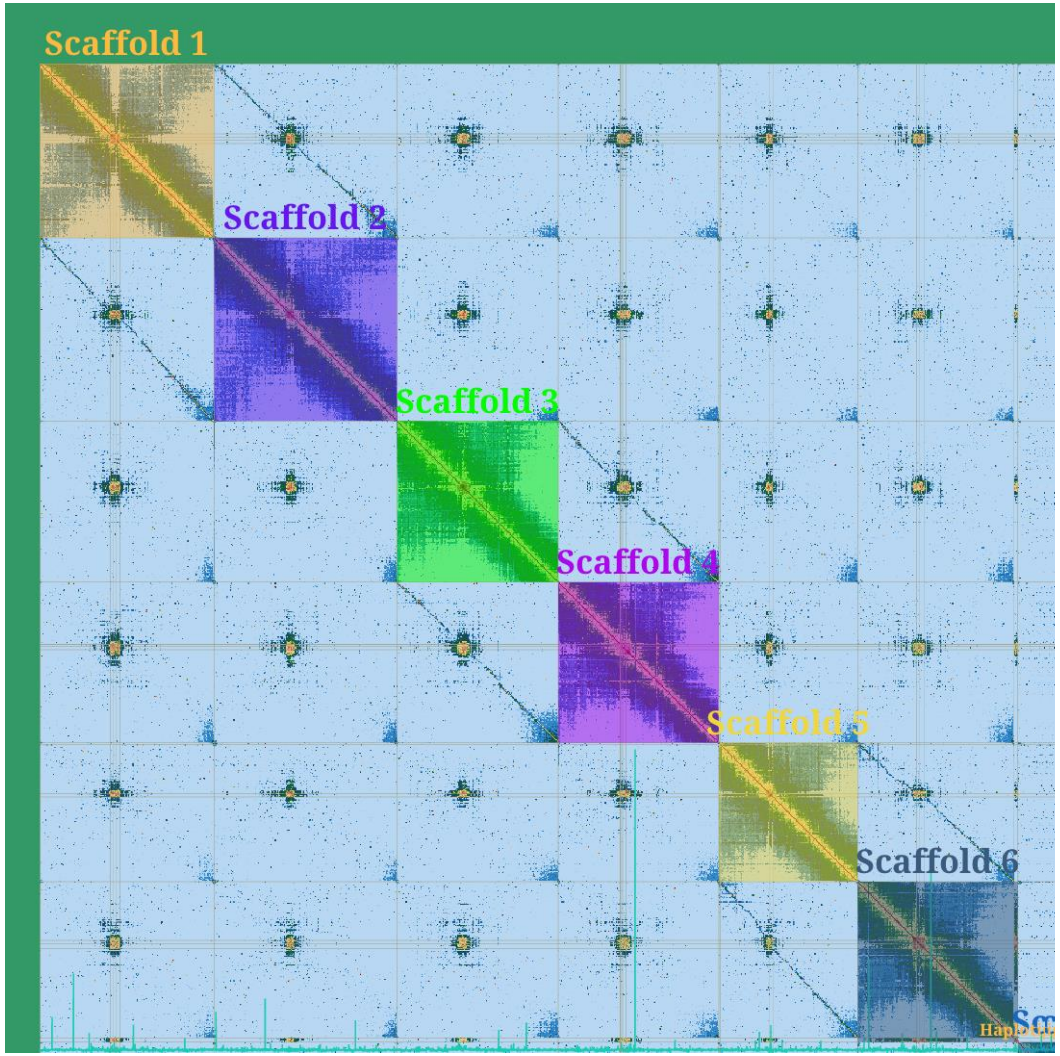
Allopolyploid genomes

Single hap curation
Primary assembly purged

2N = 28
N = 14



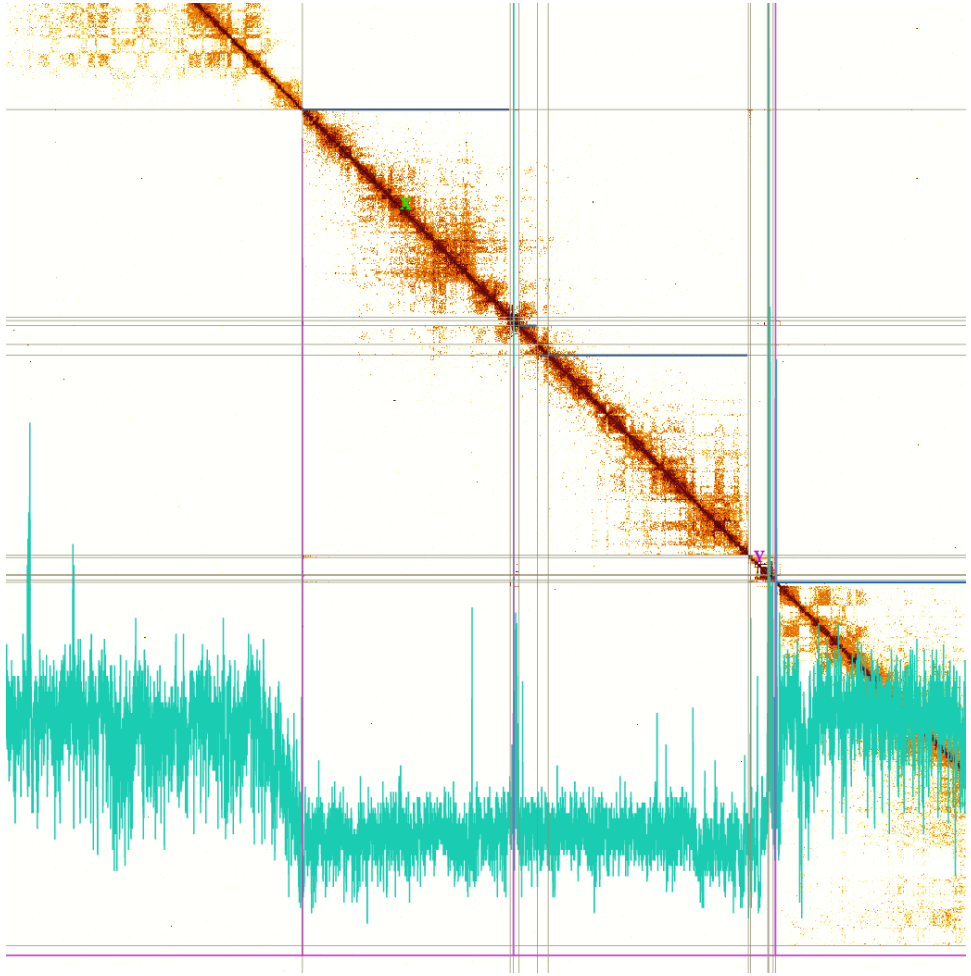
Curated haploid map



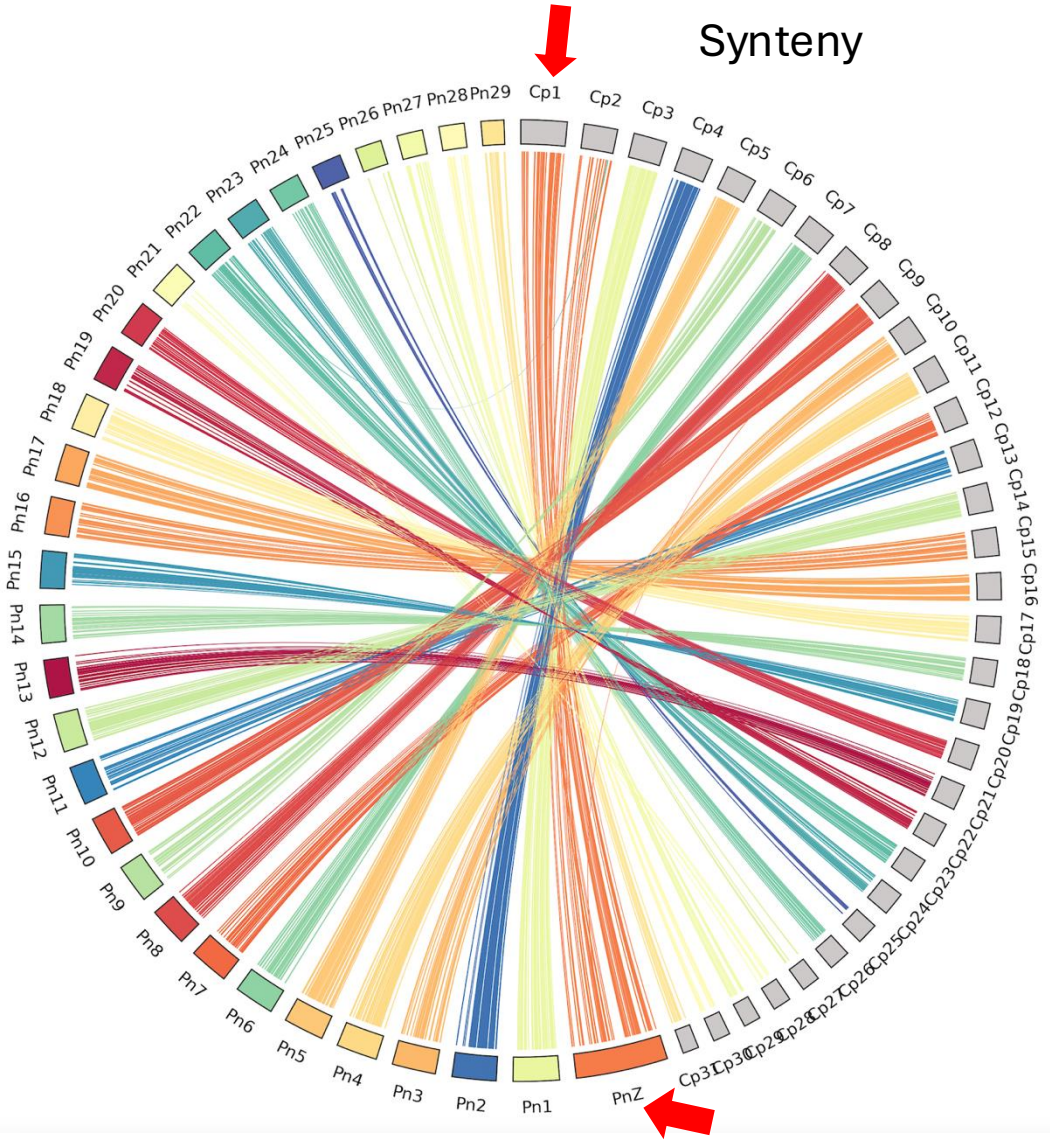
Sex chromosomes

How do we usually identify/assign sex chroms?

PacBio read coverage

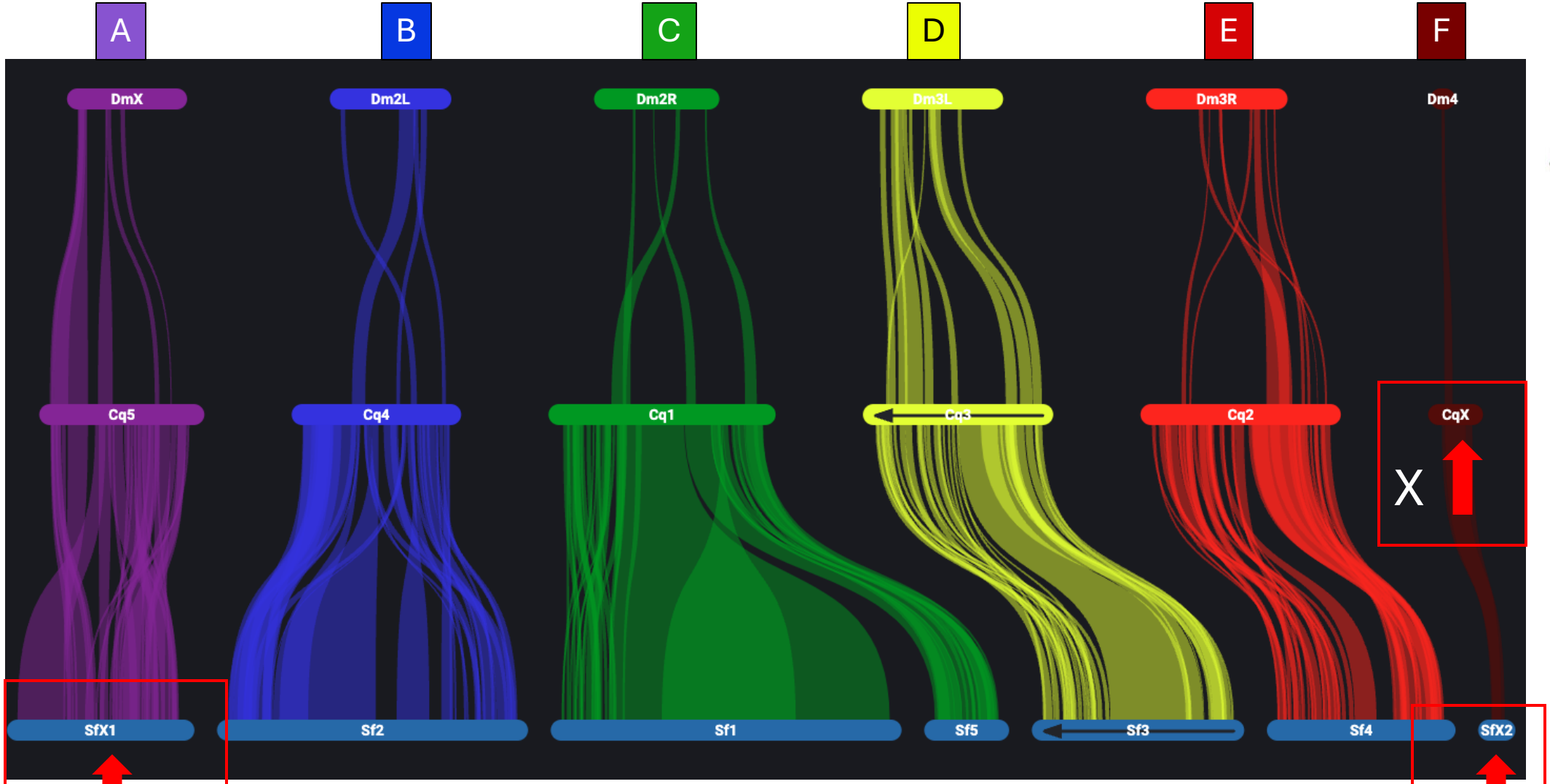


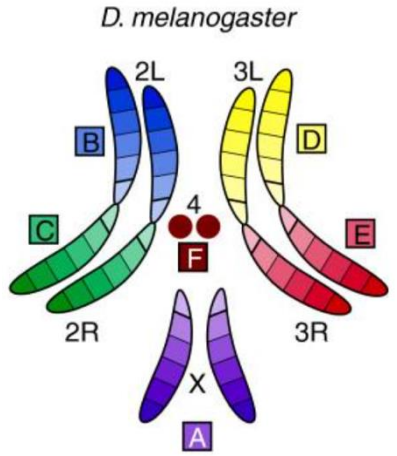
Synteny



Painting Muller elements onto Conopidae with *D. melanogaster*

Turnover is frequently observed in Diptera





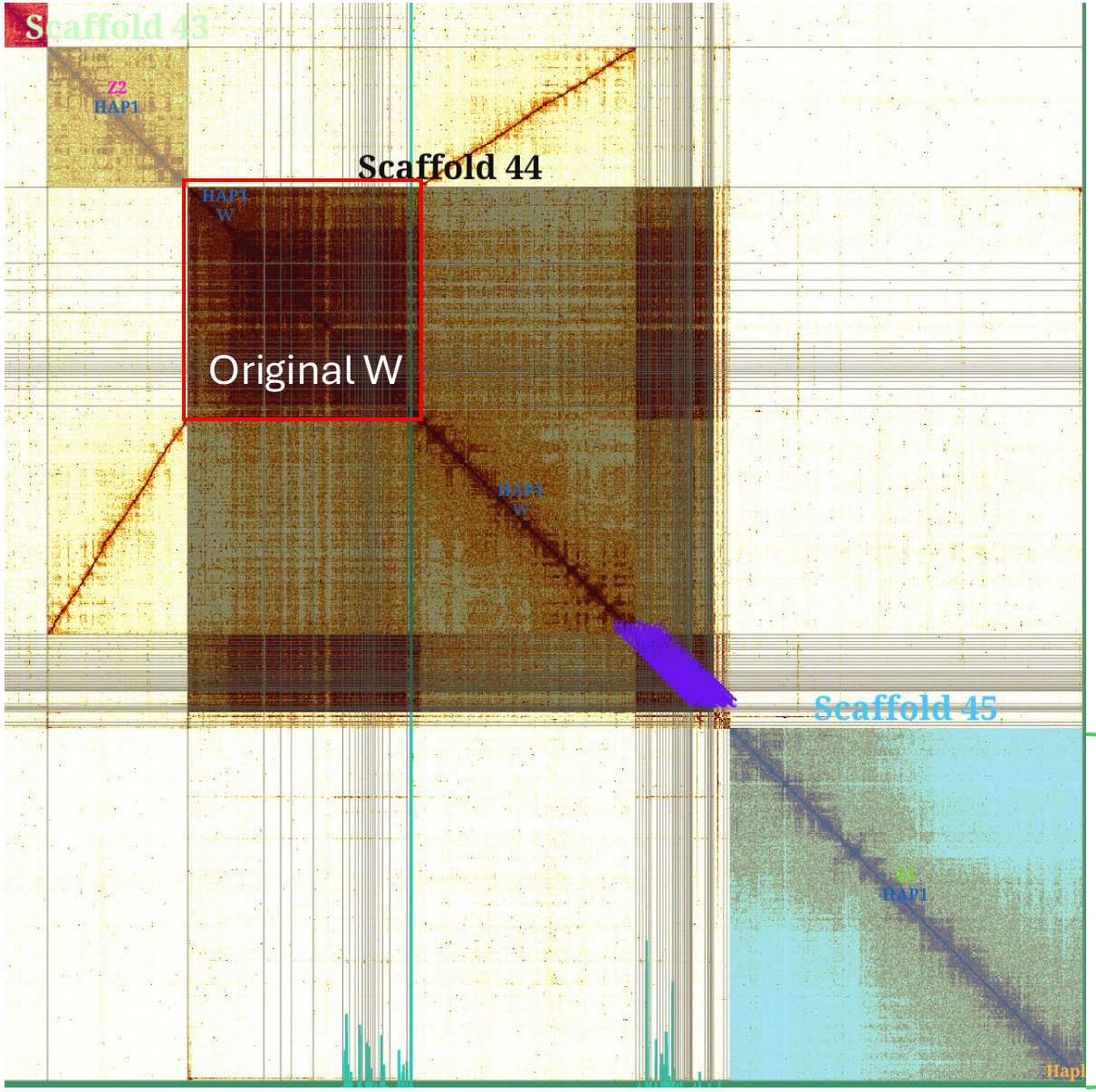
IdConQuad1

Turnover (canonical X)

idSicFerr1 (version 2)

Turnover (neo X) X2

Autosome + sex chrom fusion = neo sex chroms



Autosome = neo Z (Z2)

W chrom

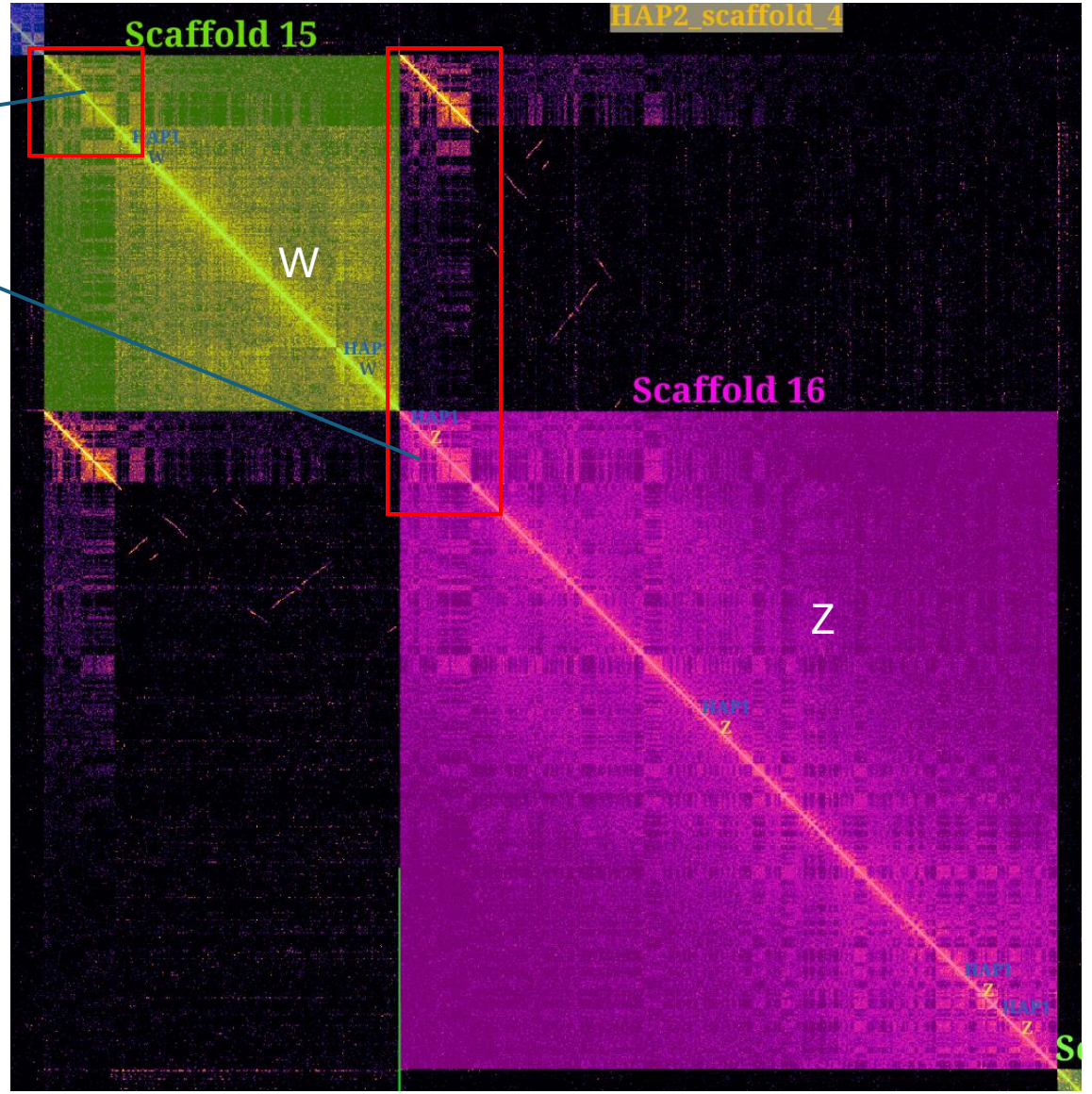
Autosome

Canonical Z

Homology regions between sex chroms = PAR

They should be assembled at the same region in both sex chroms

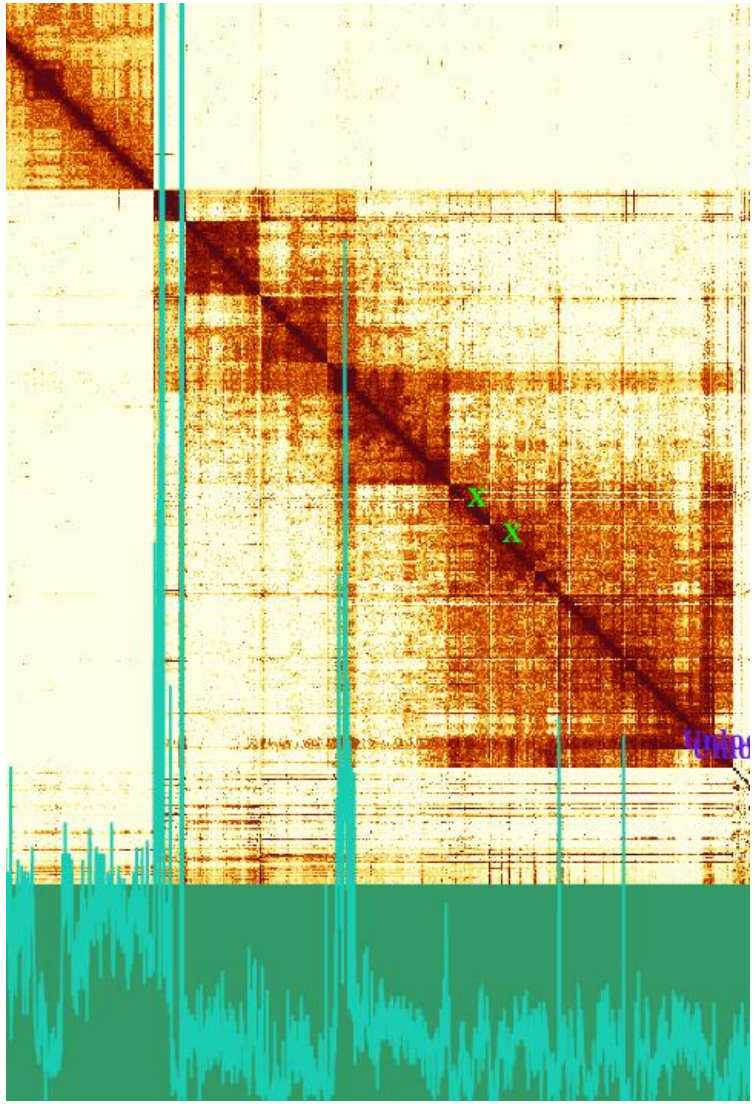
Used to self orientate sex chroms



PARs (Pseudo-Autosomal-Regions) are highly similar sequences usually found at one end of a sex-chromosome pair. The rest of the sex chromosomes are typically highly diverged.

HiC data may be mandatory to identify sex chroms

PacBio from male sample
HiC from unknown sex sample
(potential male)

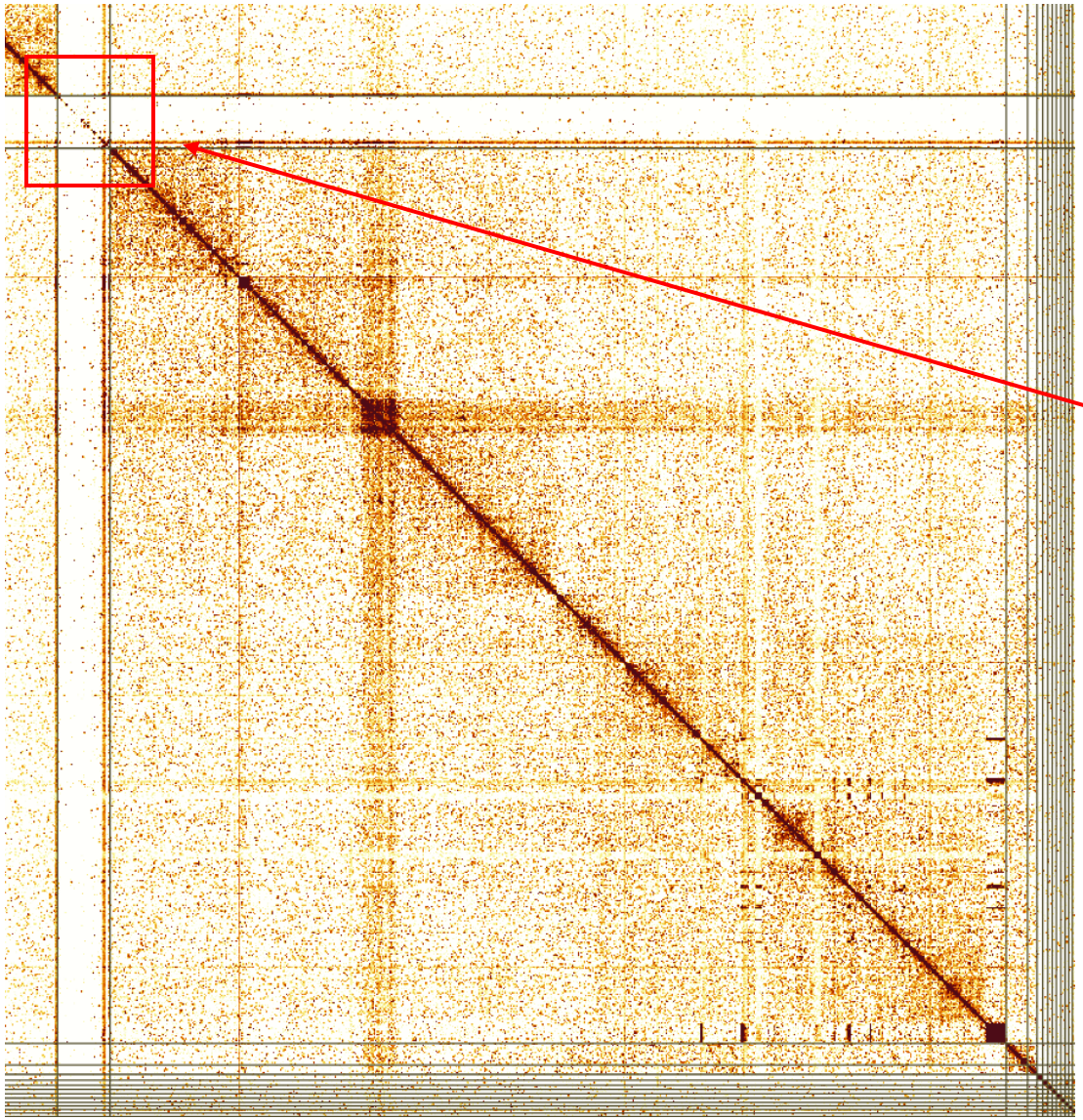


X is half coverage

Where is the Y chromosome?

HiC data may be mandatory to identify sex chroms

HiC from a female sample
No HiC signal for Y



Here is the Y chromosome!