Manual Curation of Genome Assemblies

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

Genome Reference Informatics Team (GRIT)

Jo Wood

Danil Zilov

Dominic Absolon

Jo Collins

Camilla Santos

Karen Brooks

Sarah Pelan

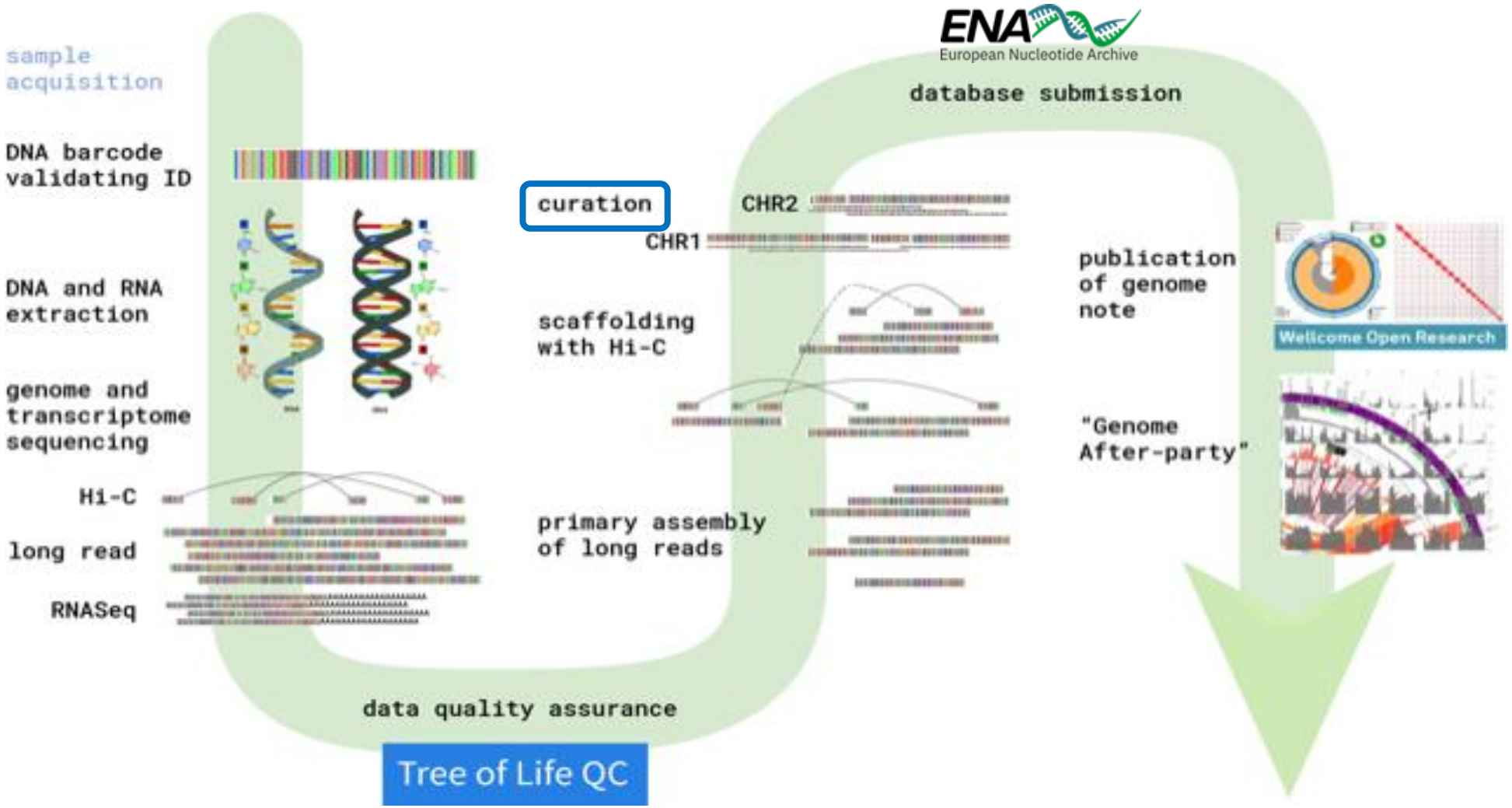
Tom Mathers

Michael Paulini

Karen Houliston

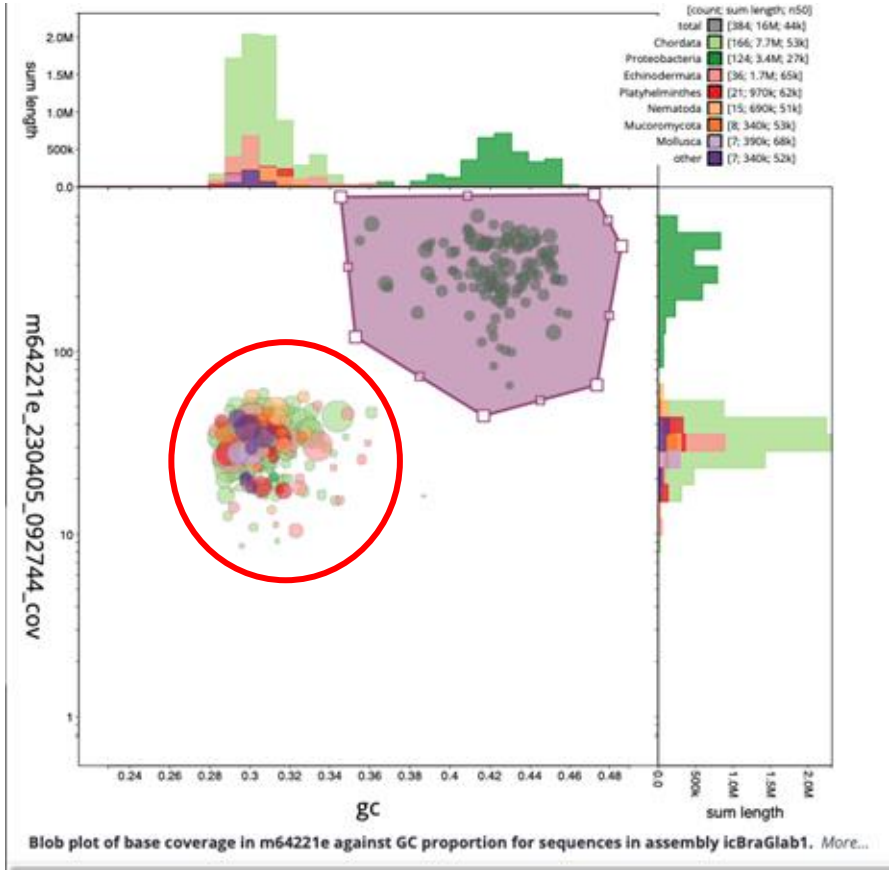
Scientific Publications Editor

The Tree of Life genome factory



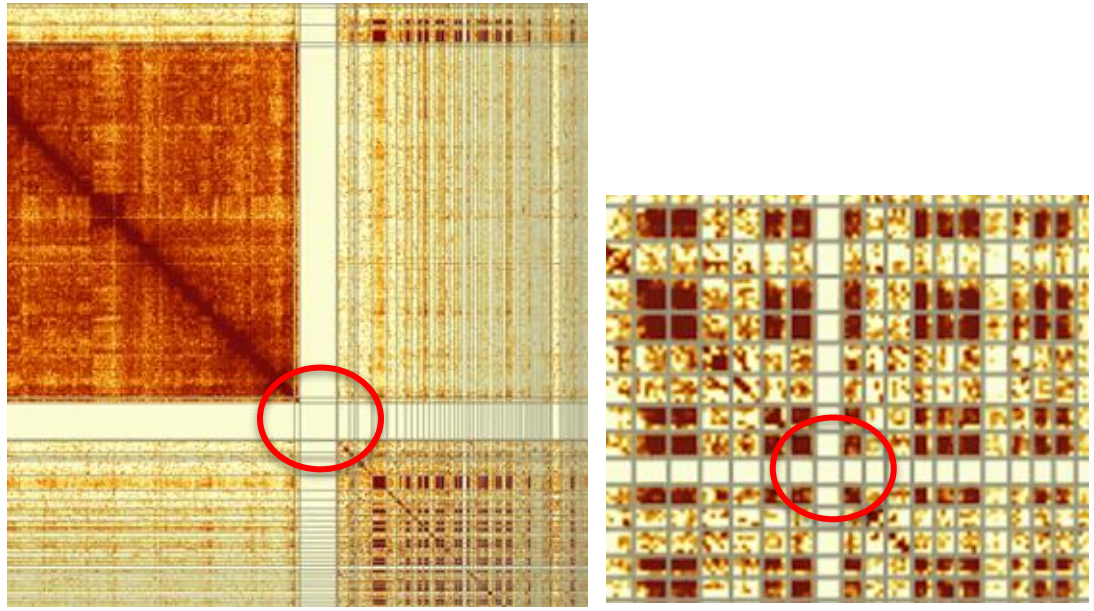
Decontamination examples

Pre-curation



BUSCO hits and GC vs read coverage distribution

Post-curation

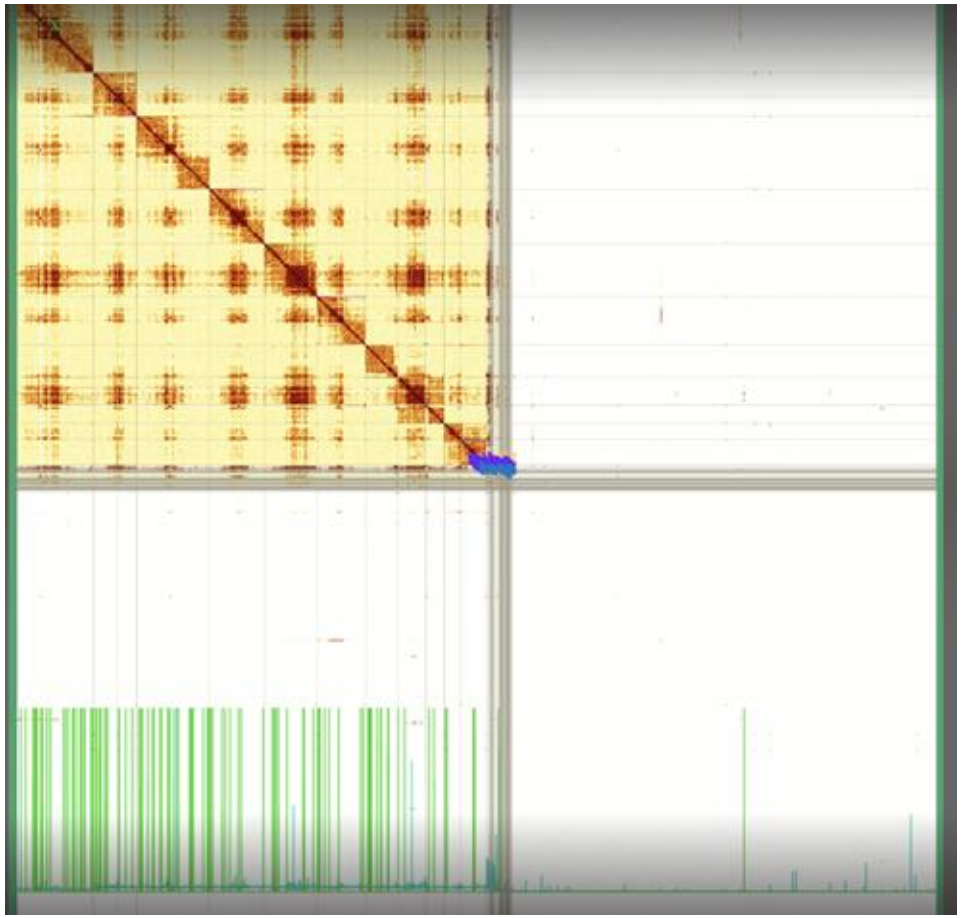


HiC contact map

Decontamination examples

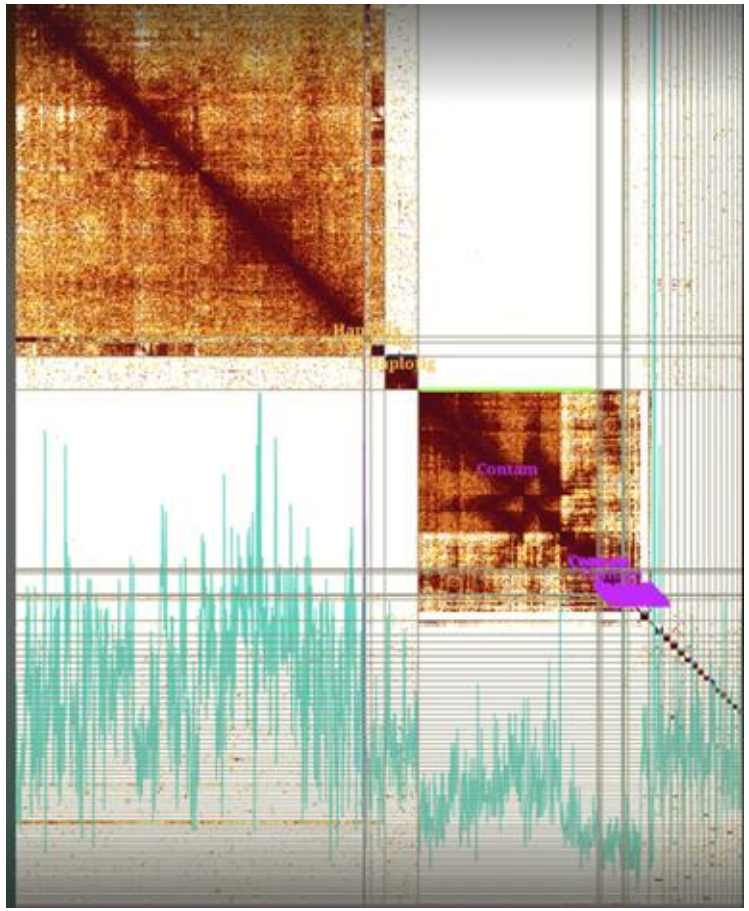
Post-curation

HiC - uncontaminated sample
Pacbio - contaminated sample



Diptera genome with fungi contamination

HiC and PacBio from same sample

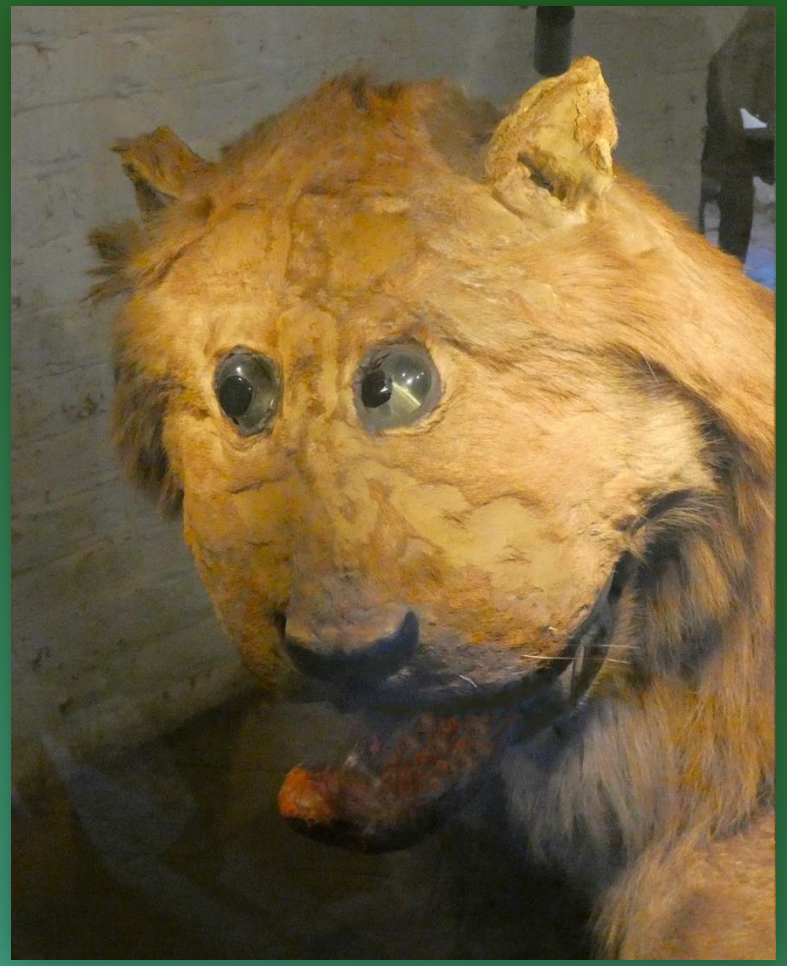


Worm genome contaminated with bacteria

What is genome curation?

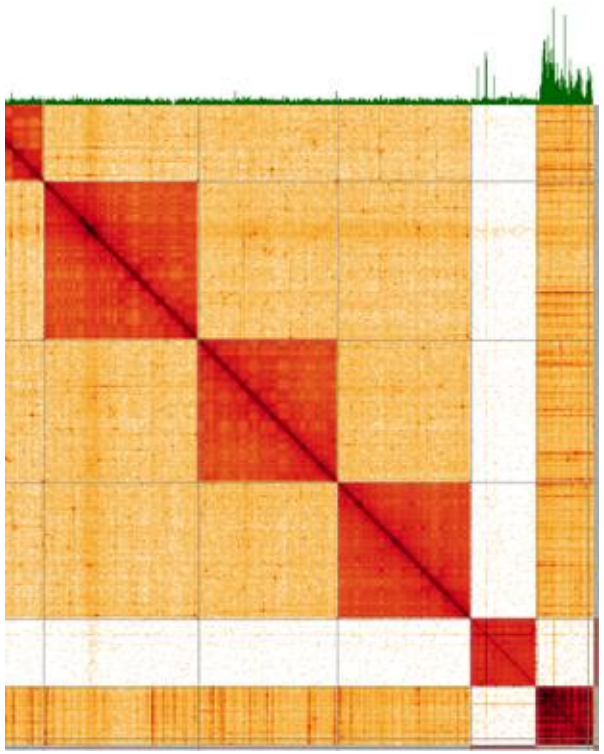
"Assimilating evidences from **all available data types** and using these to **reshape automated assemblies** to get as close as possible to **chromosomally resolved assemblies**, guided by karyotype, fixing misassemblies, removing all contamination and removing haplotypic sequence, **in a reasonable timeframe**"

Why do we need curation?

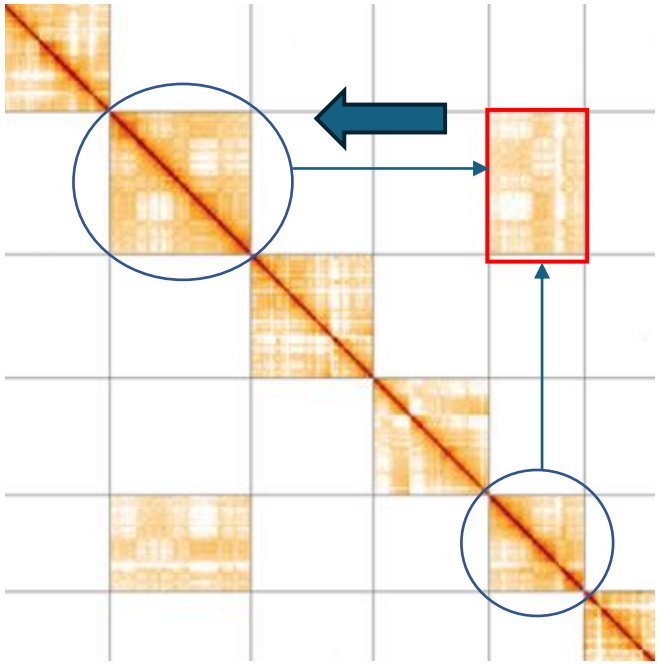


Some of the main issues

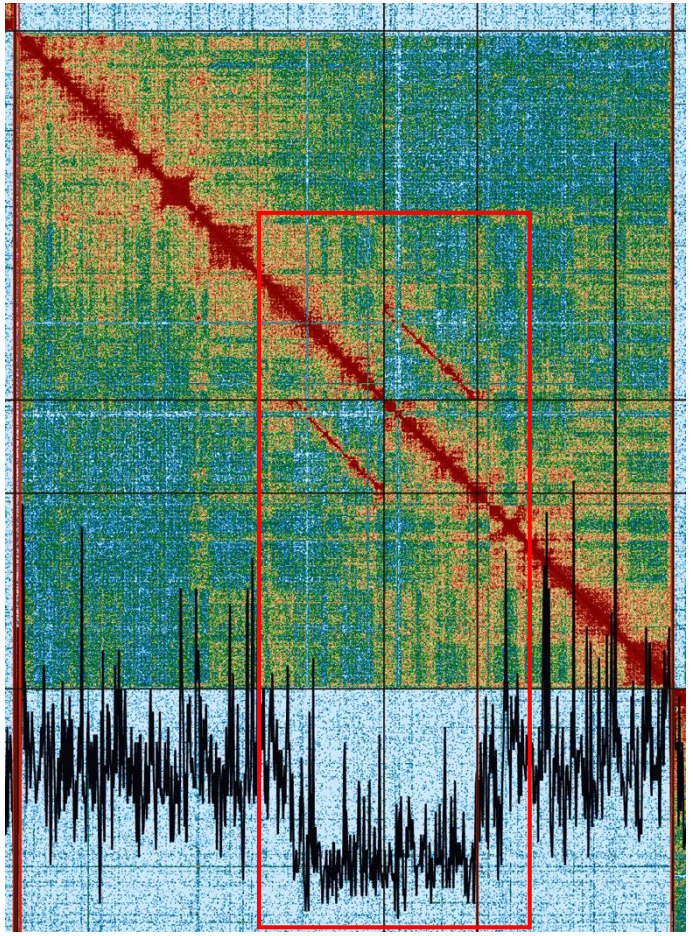
Contamination



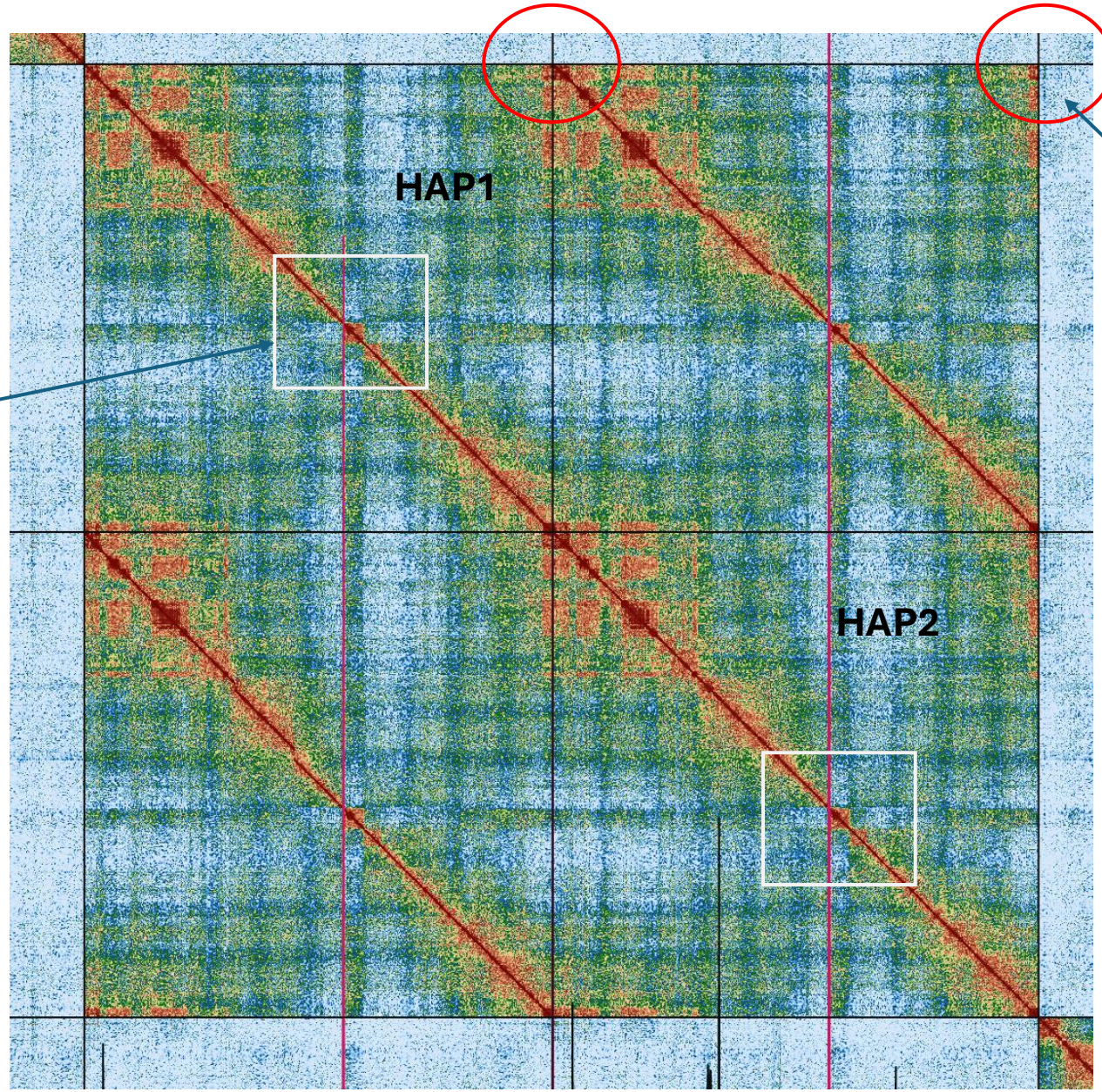
Misassemblies



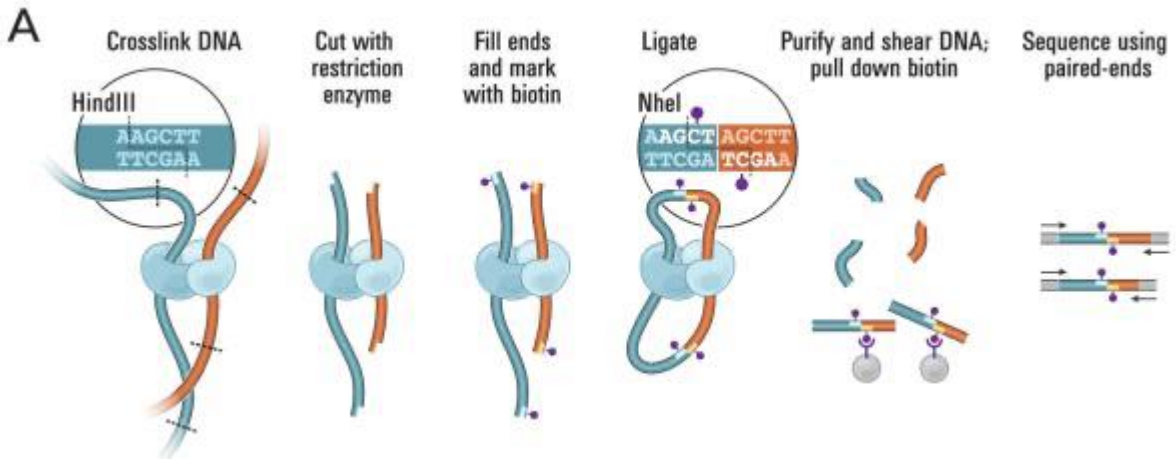
Haplotigs



Joined by the telomeres



HiC data - our No. 1 curation resource

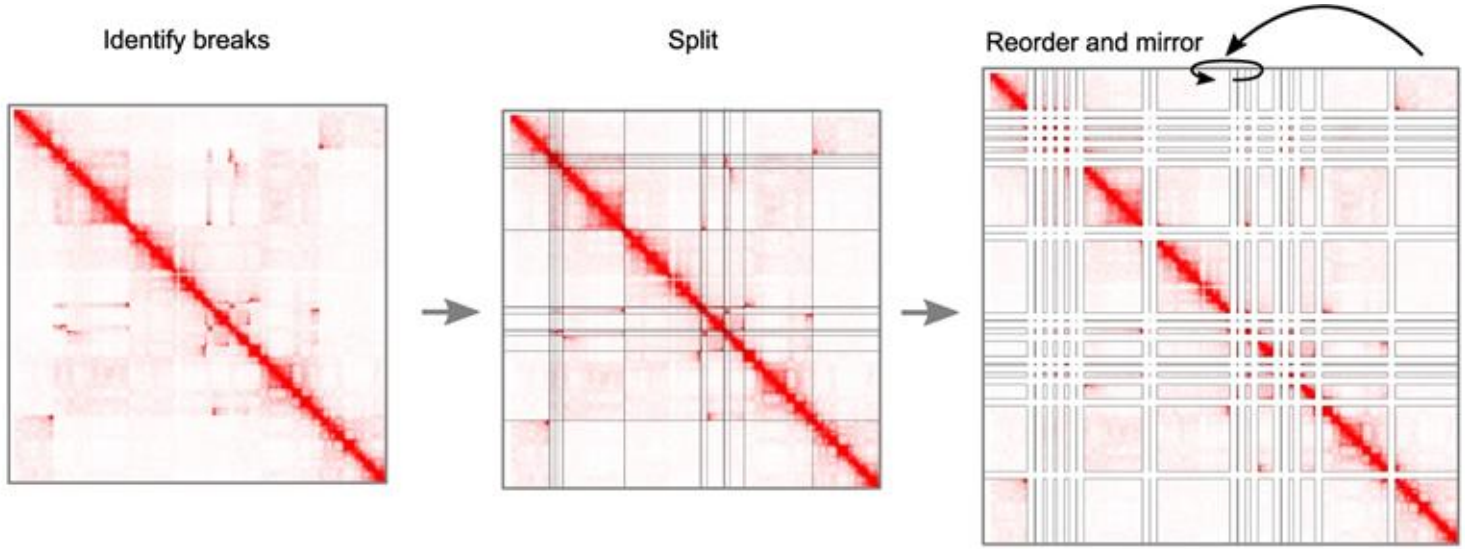


< Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. PMID: 19815776; PMCID: PMC2858594.

Schöpflin, R., Melo, U.S., Moeinzadeh, H. *et al.* Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nat Commun* 13, 6470 (2022). https://doi.org/10.1038/s41467-022-34053-7
v

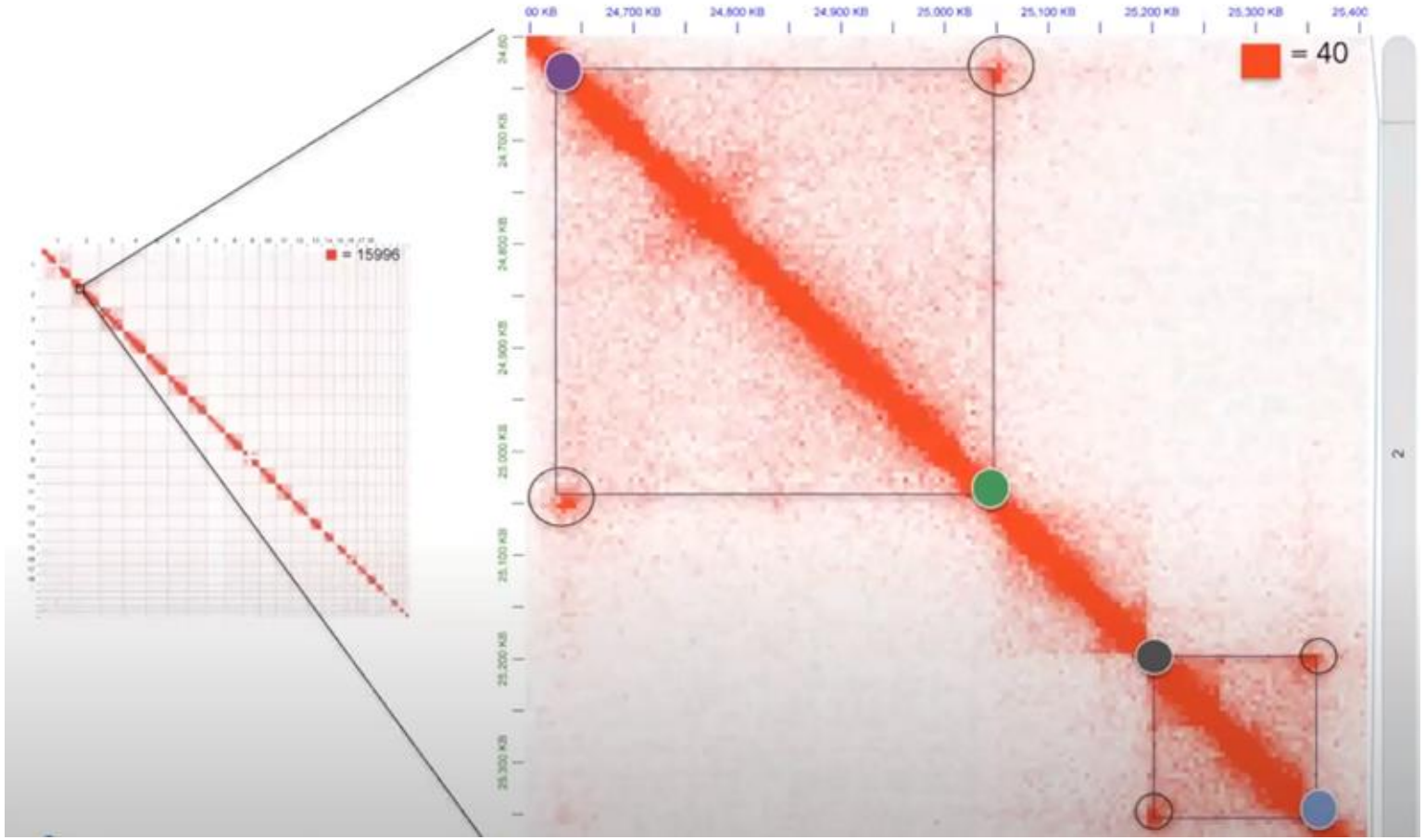
"in-situ" sequencing gives evidence of what sequence belongs next to what sequence.

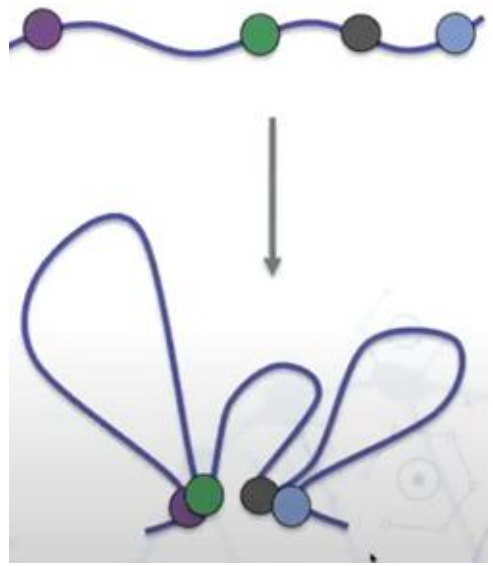
The result is a contact map



HiC data - our No. 1 curation resource

Chromatin conformation with Hi-C

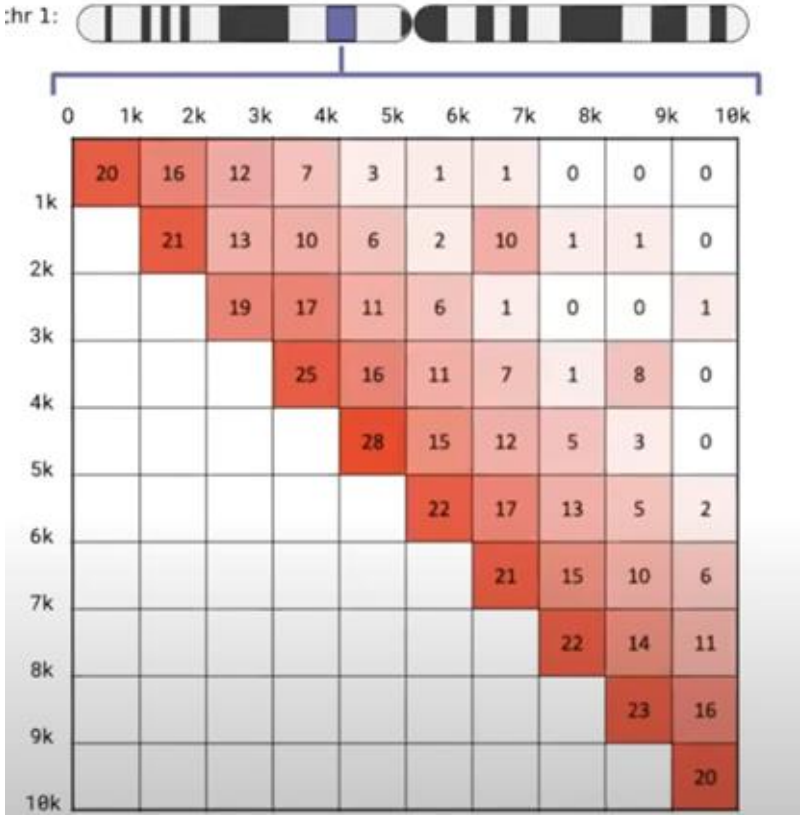




HiC data - our No. 1 curation resource

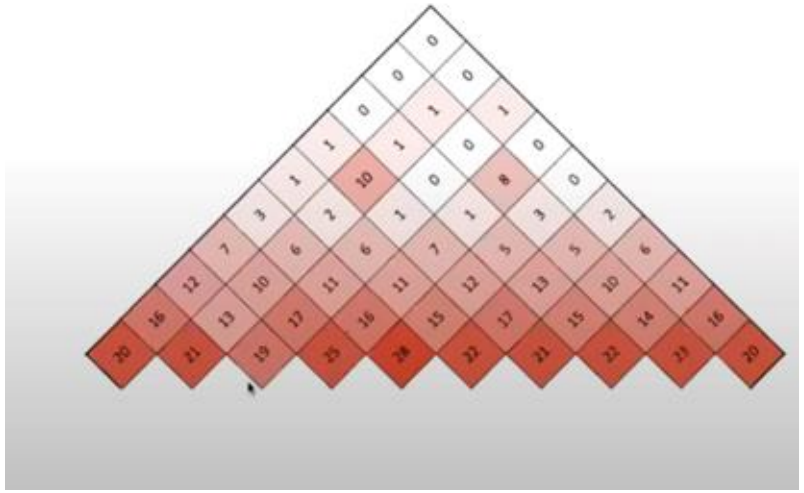
Visualization

Contact matrix colored based on hic reads counts



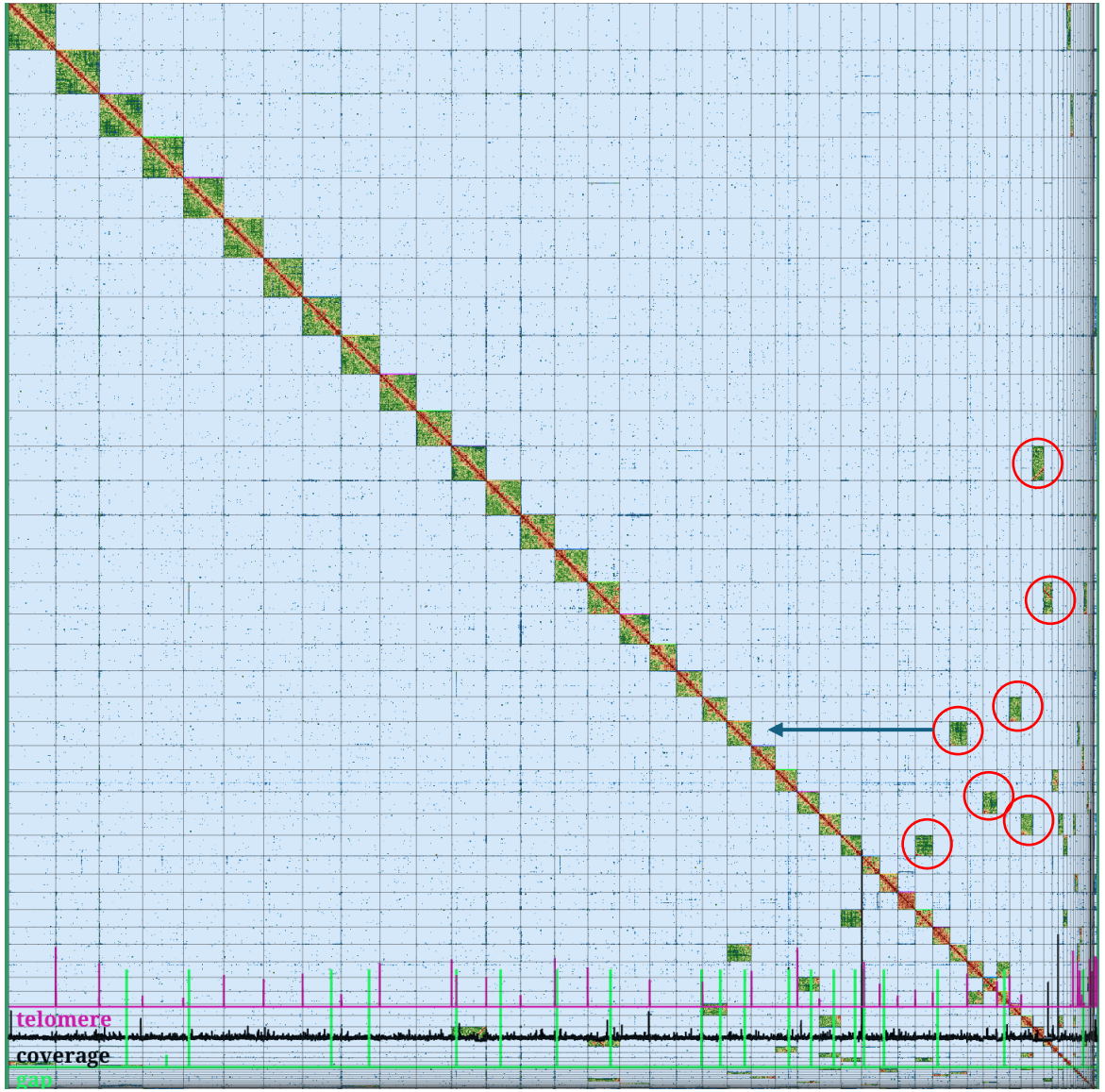
# of HiC reads supporting 3D interaction

More color = More reads = More likelihood of contacts



Interactions within chroms are stronger (self matches) than between chrom

Interpreting a HiC map



tracks

PretextView

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

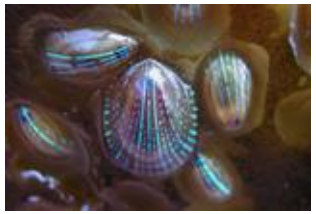
Centre diagonal show self matches, eg chr1 vs chr1
Diagonal mirrors itself

Off diagonal show relationship between different chromosomes/scaffolds.

The darker the off-diagonal square, the stronger the relationship between the scaffolds.

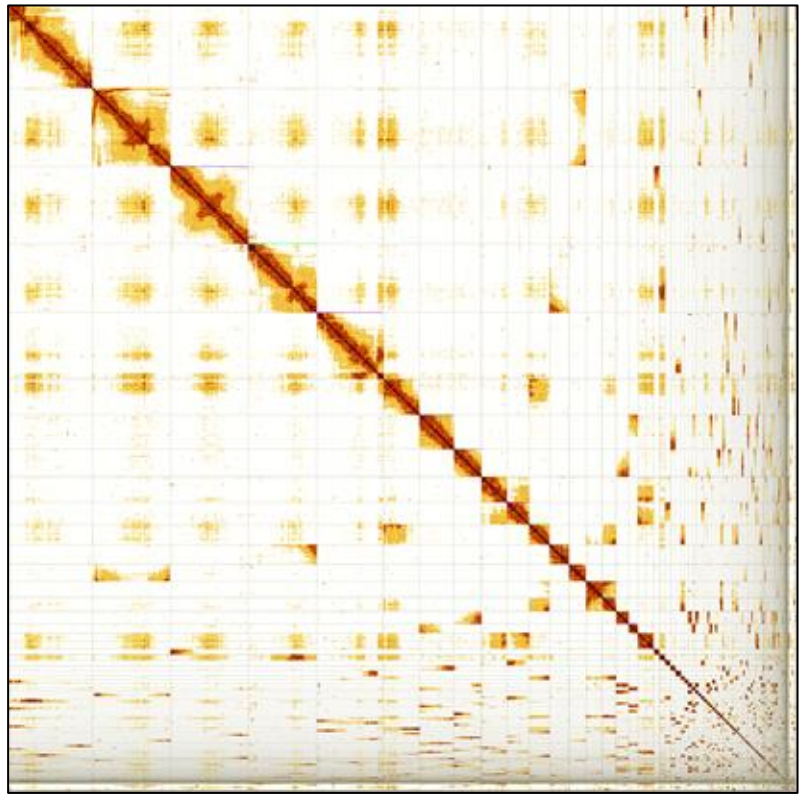
Horizontal and vertical lines delineate chromosome/scaffold boundaries.

Evolution of a manually curated assembly



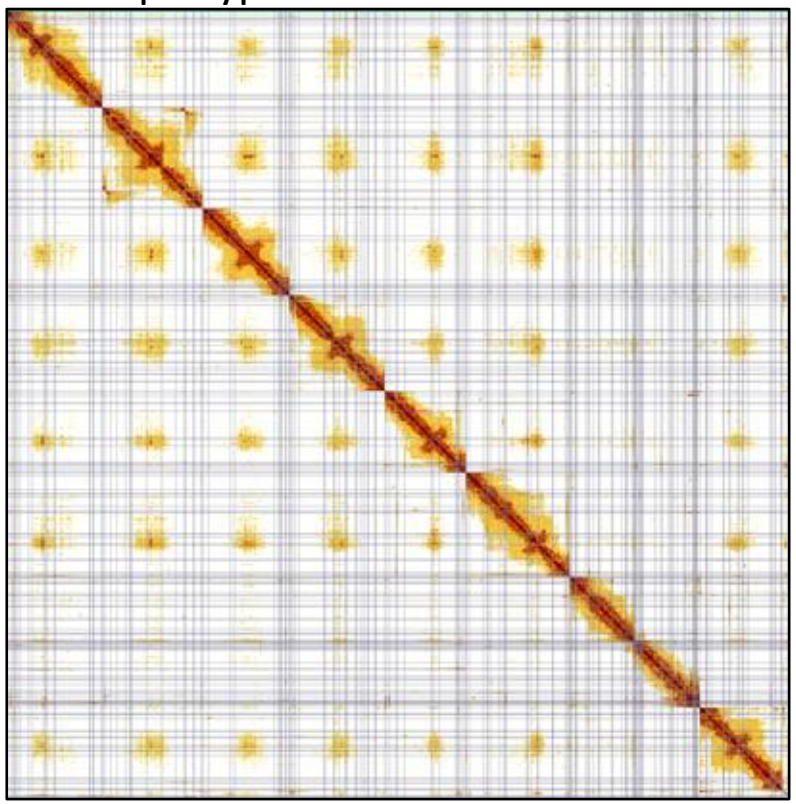
Patella pellucida
Blue-rayed limpet

n = 230
N50 = 33.1Mb



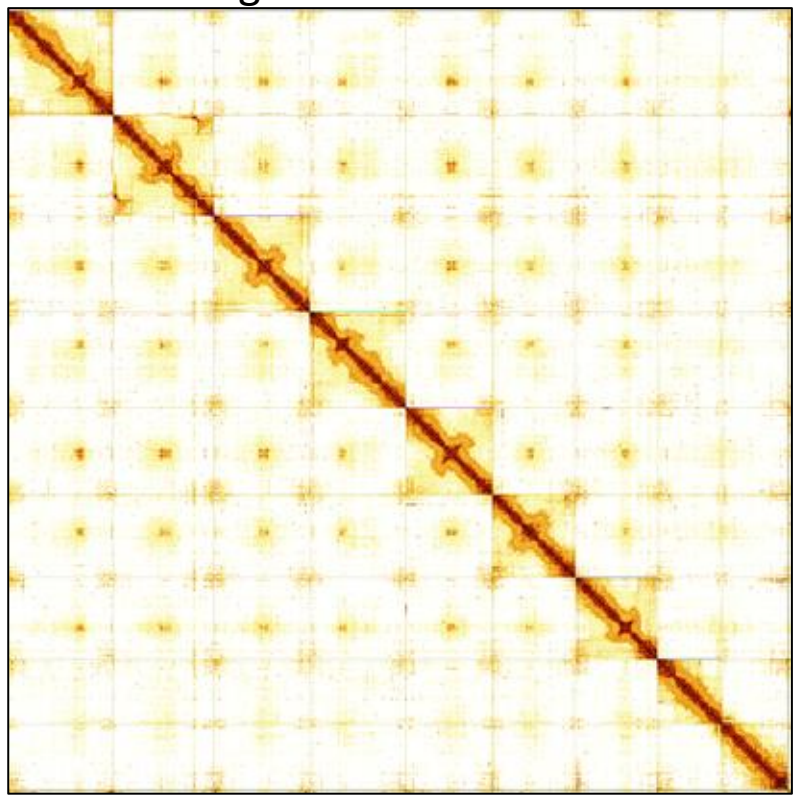
Pre curation assembly

225 joins
84 breaks
29 haplotype removals



after pretext manipulation

n = 62
N50 = 87.1Mb
99.85% of genome in 9 Chromosomes



Post curation

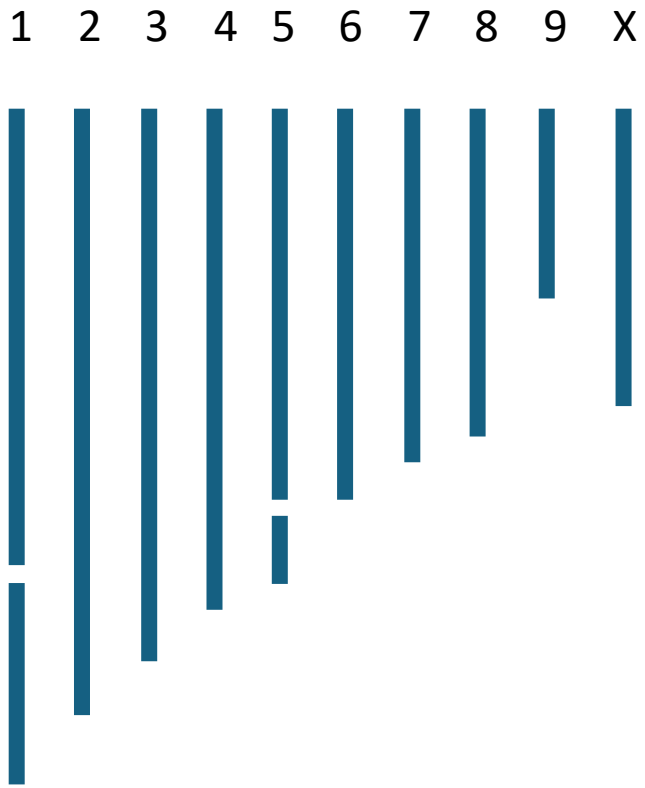
Chromosome naming

By size

- Autosomes large > small

By synteny

- Existing reference



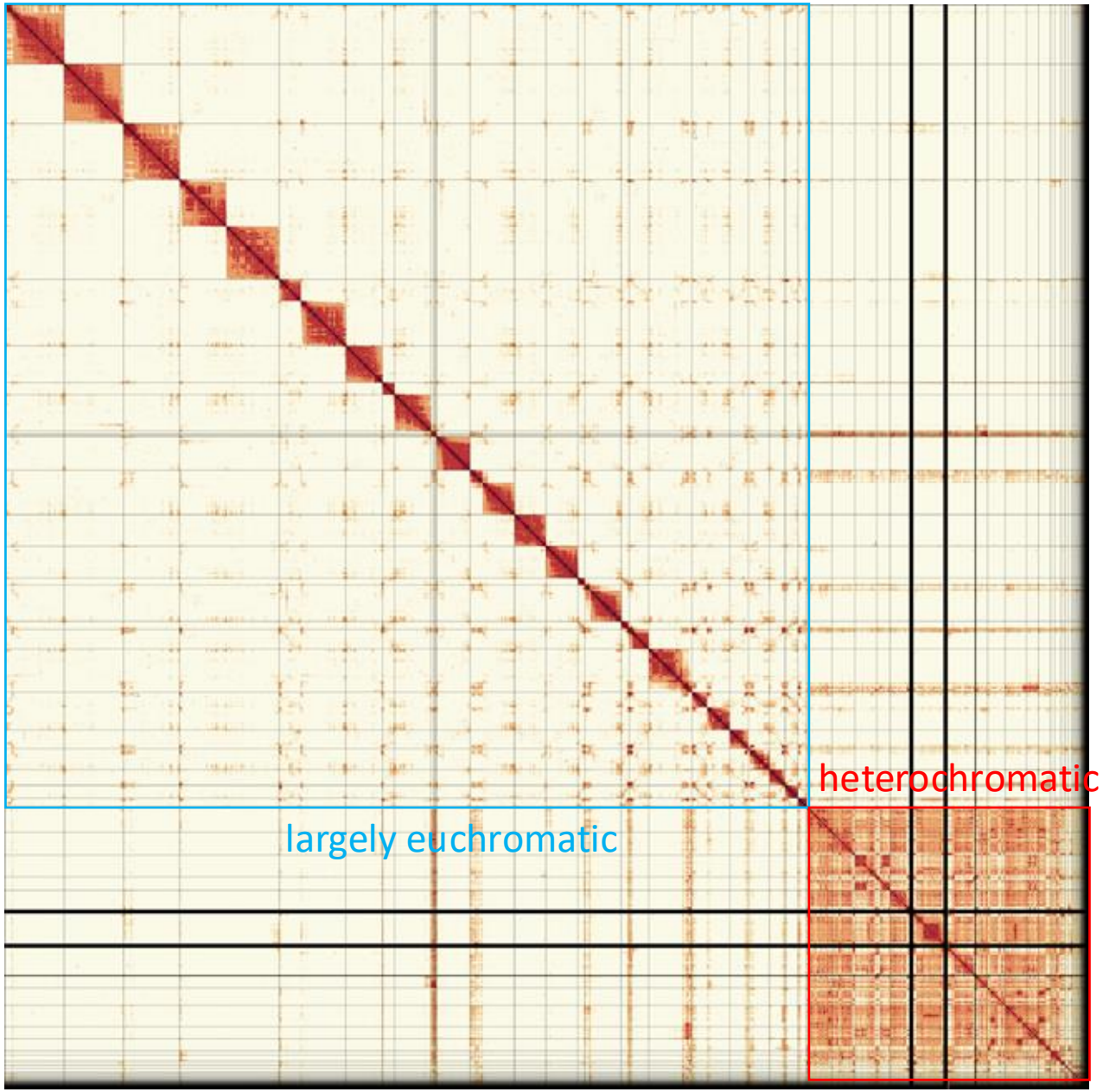
Some of the main challenges...

- Highly repetitive genomes
- Sex chromosomes
- Microchromosomes
- Polyploids
- Bad phasing
- Poor quality Hi-C data

Contrast between euchromatic and heterochromatic portion of the genome

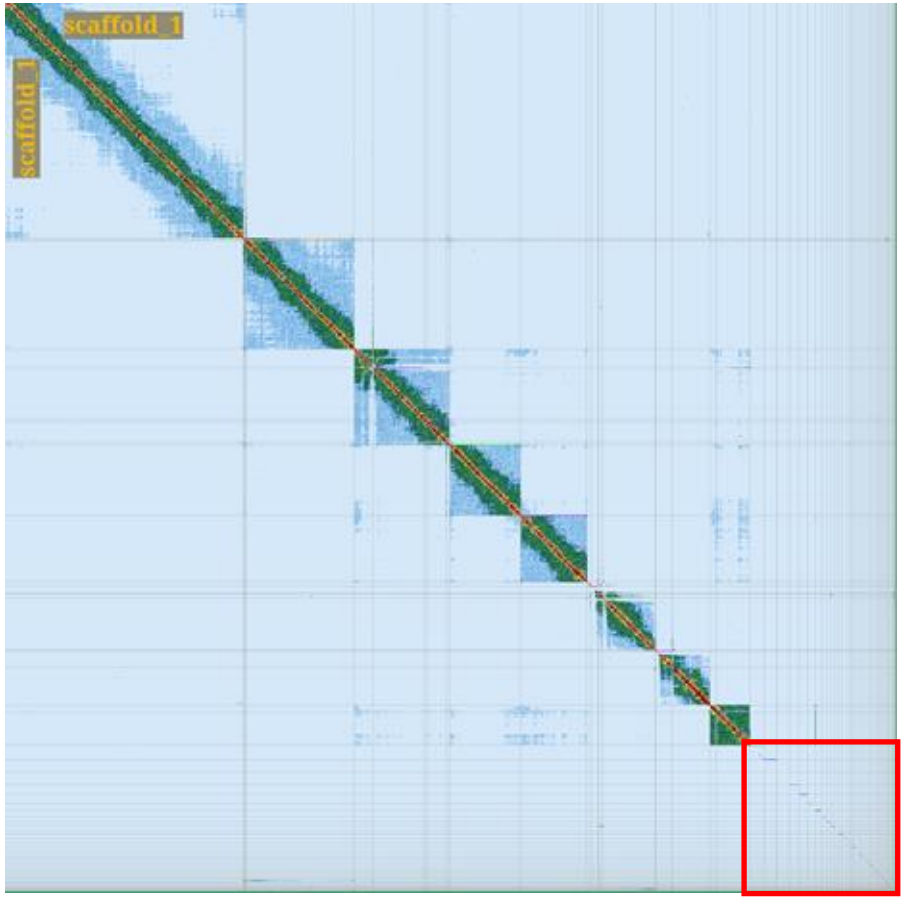
Non-repetitive HiC signal can be seen for 26 chromosomal entities, in stark contrast to the heterochromatic portion of the genome (centromeric and short-arm sequences which in the case of this wasp do not have enough specific association with a particular chromosome to enable them to by placed.

iyNysSpin1_1

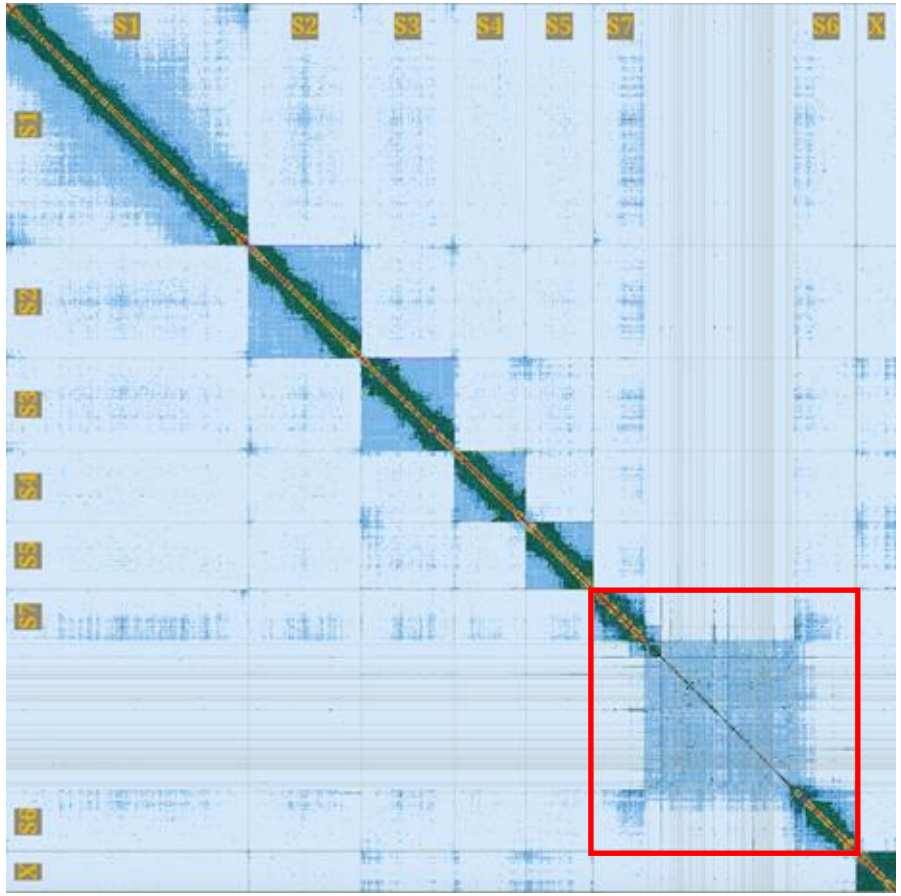


Additional clues (multi-mapping + karyotype)

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

Sex chromosome identification

Identifying sex chromosomes is difficult. We only assign sex chromosomes when we are beyond doubt.

By coverage

Heterogametic sex chromosomes = half read coverage –

By synteny

When allosomes are homomorphic

- Existing reference
- Genetic map

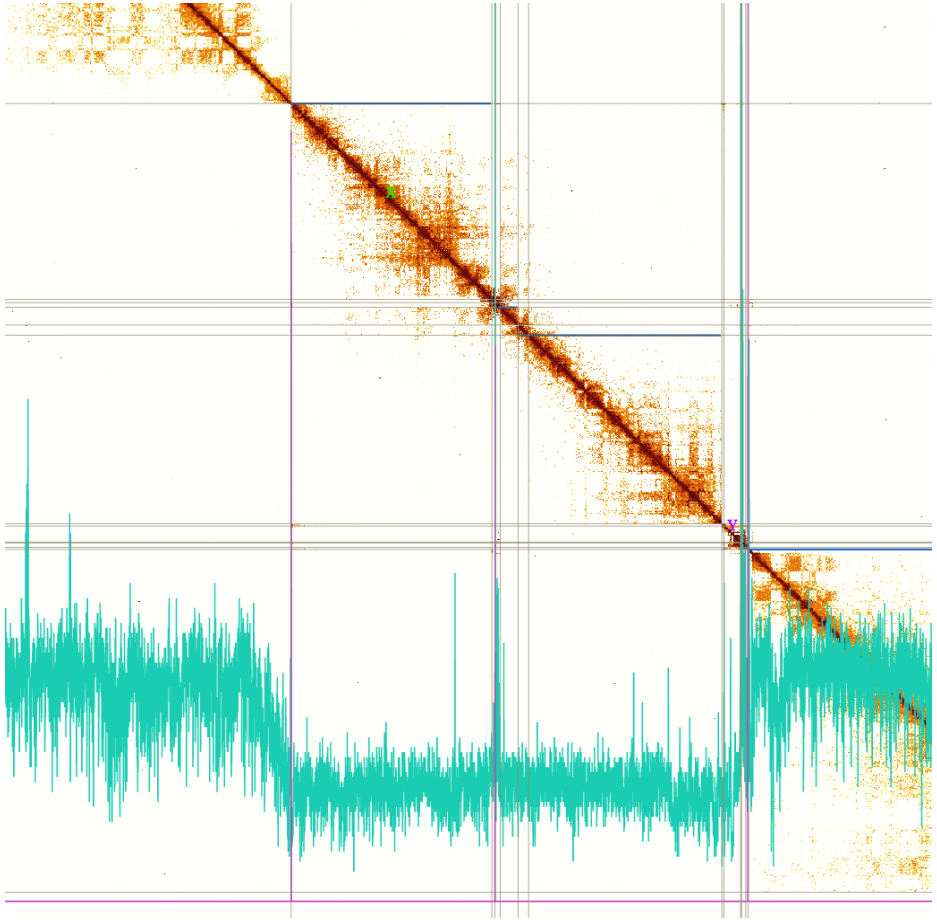
Caution!

Synteny works well for sex chromosome identification in some orders but not in others:

Good examples: Coleoptera, Lepidoptera

Bad examples: Diptera (high sex chrom. turnover rate)

PacBio read half-coverage



Sex chromosome identification

Identifying sex chromosomes is difficult. We only assign sex chromosomes when we are beyond doubt.

By coverage

Heterogametic sex chromosomes = half read coverage –

By synteny

When allosomes are homomorphic

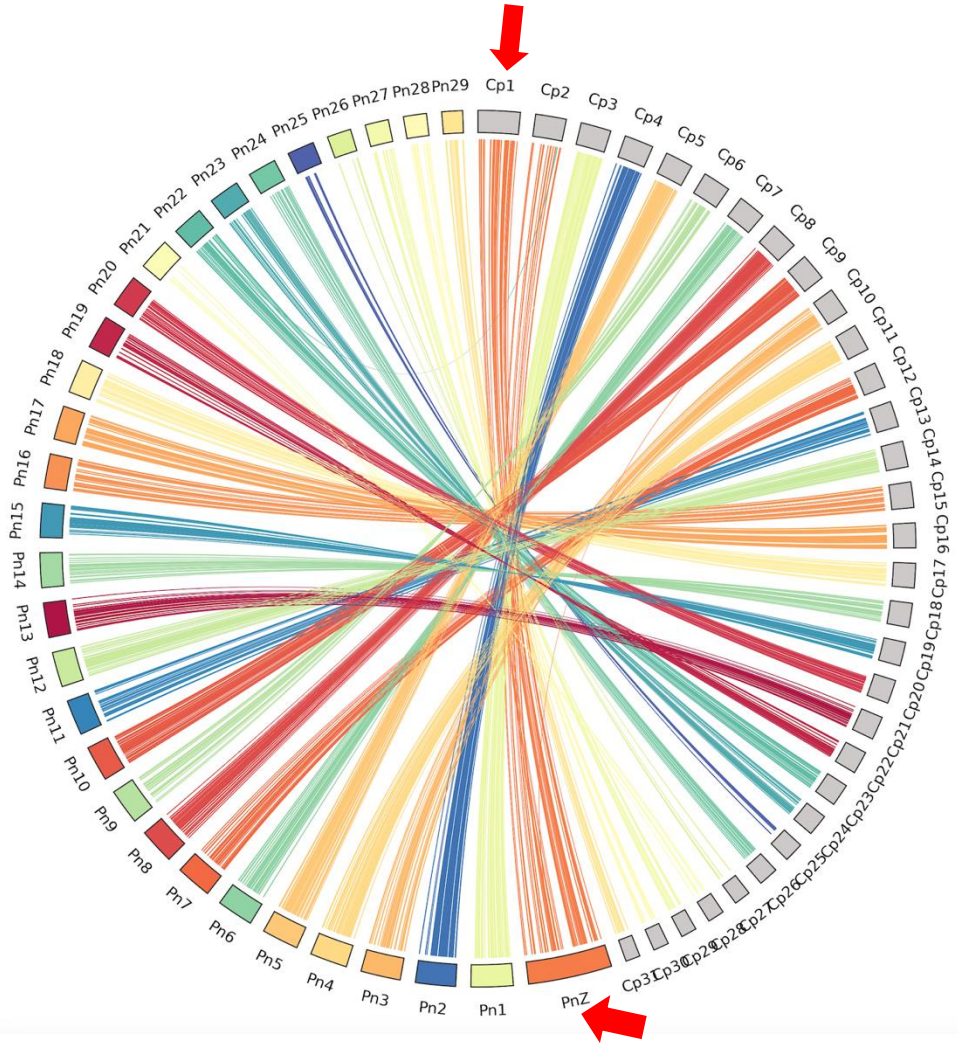
- Existing reference
- Genetic map

Caution!

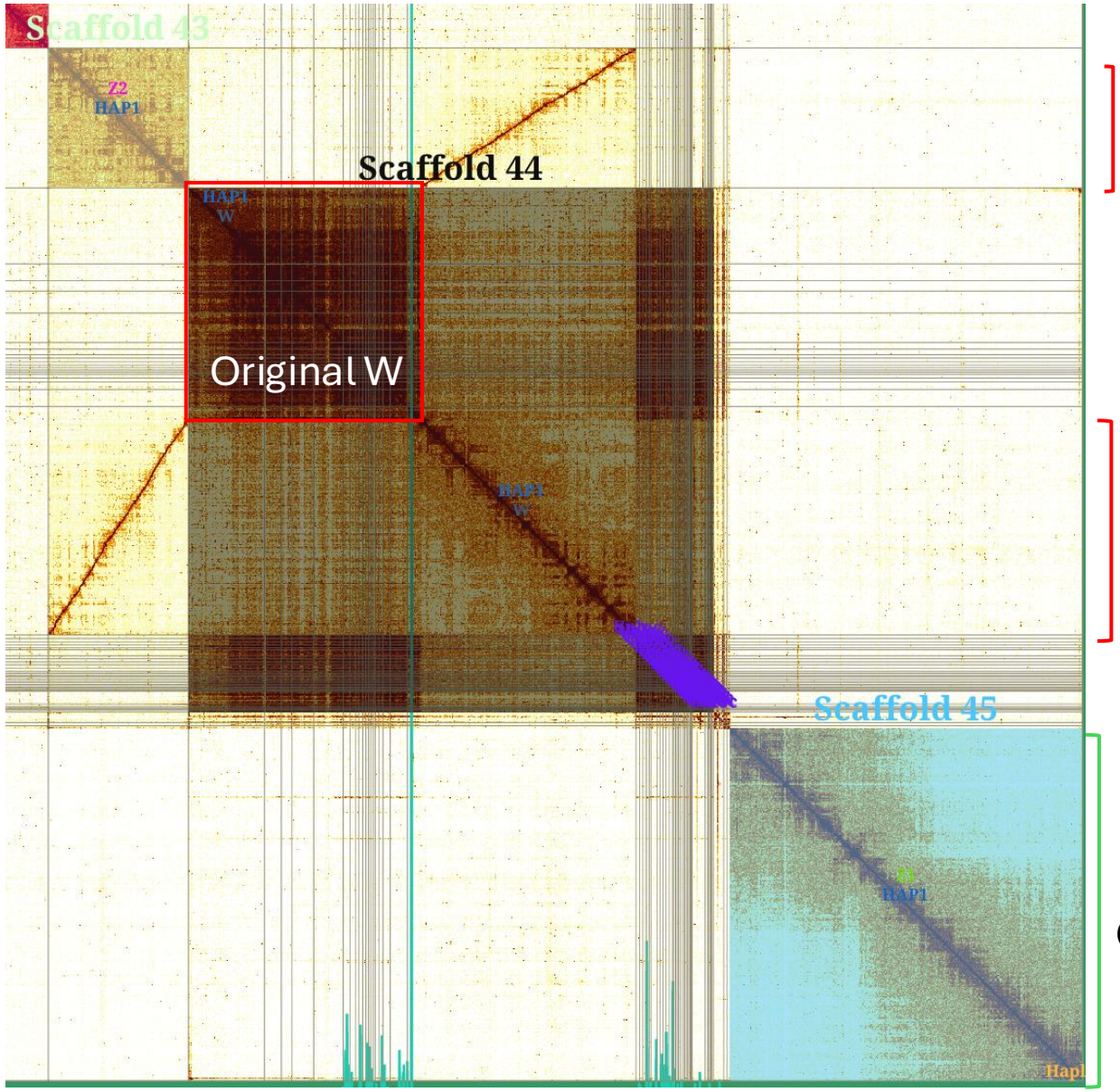
Synteny works well for sex chromosome identification in some orders but not in others:

Good examples: Coleoptera, Lepidoptera

Bad examples: Diptera (high sex chrom. turnover rate)



Autosome + sex chrom fusion = neo sex chroms



Autosome = neo Z (Z2)

W chrom

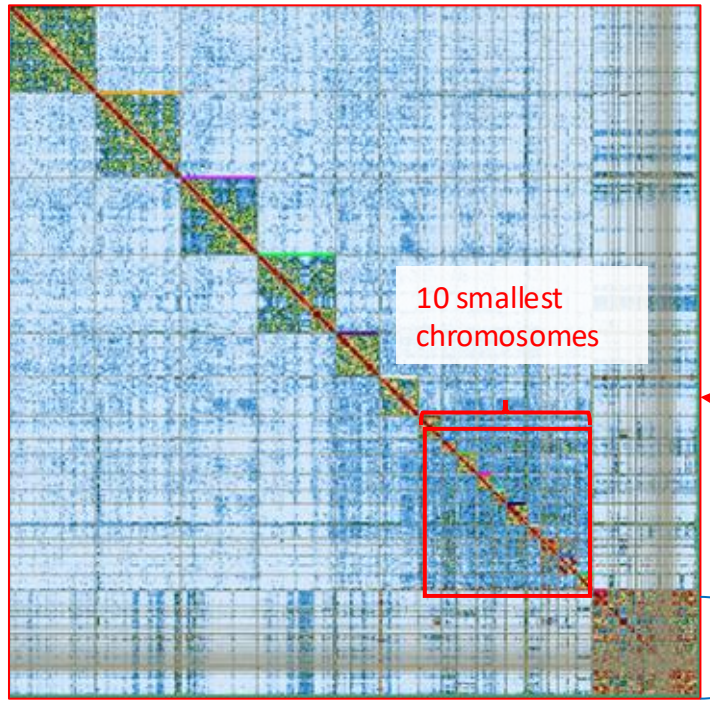
Autosome

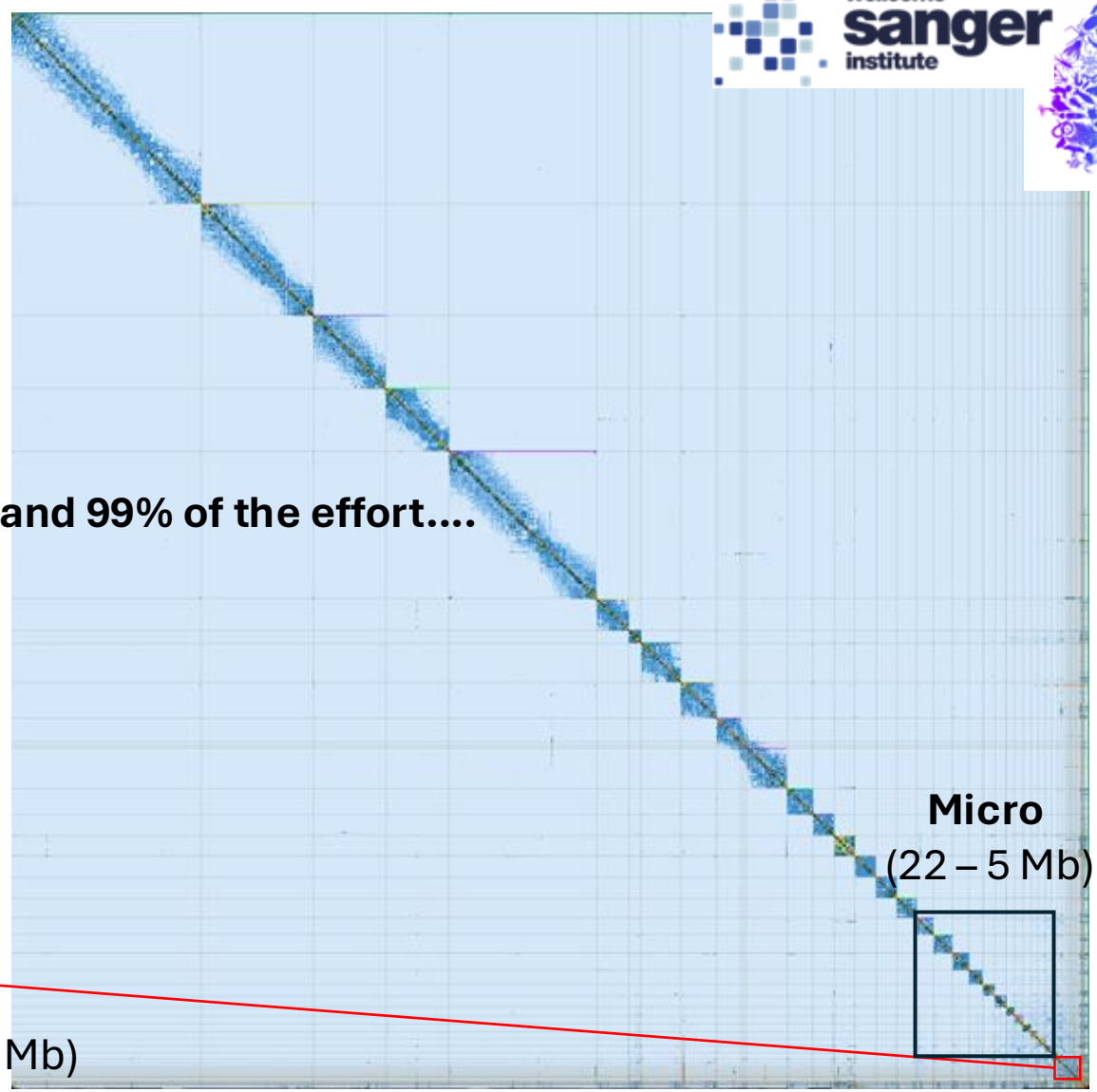
Canonical Z

Micro-chromosomes
(bCucCan1)

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

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



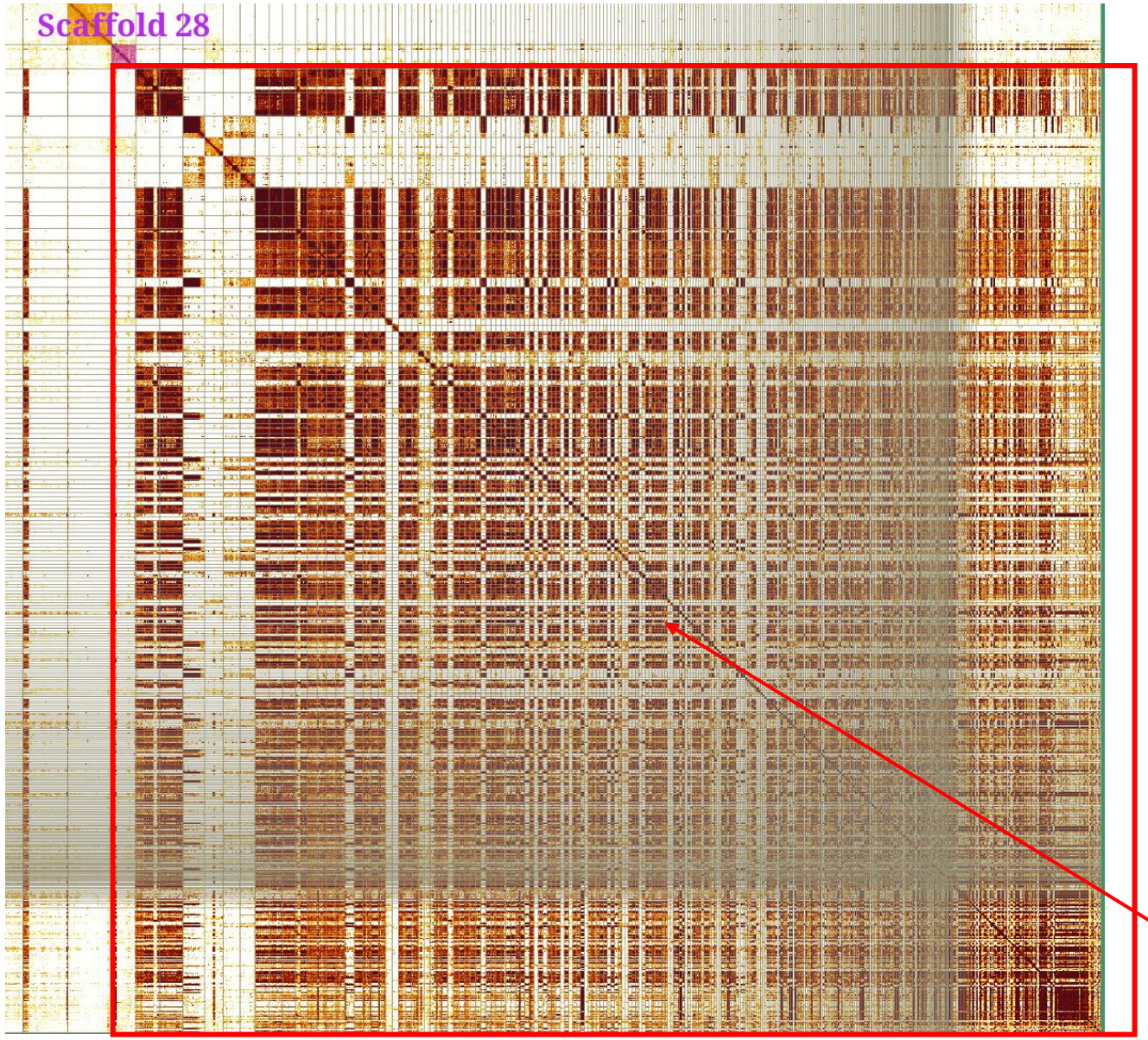


“dot” (< 5 Mb) chromosomes

unplaceable sub-telomeric repeat

Microchromosomes

(By Tom Mathers)



HiFi data

Quick curation of larger scaffolds only recovers 28 chromosomes.

Expected karyotype is 39 autosomes + Z + W

Remaining 13 chromosomes are somewhere in here!

Micros ???

How do we fish out the micros?

Our main approach for birds is

Tom Mathers

Michael Paulini

`MicroFinder` script for birds (HiFi/ ONT)

Miniprot to **map a set of conserved microchromosome-associated proteins** to a draft assembly and then **counts the resulting hits** and **orders the input assembly by the number hits**

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.sh <hap1_fasta> scaffold_length_cutoff

MicroFinder.sh <hap2_fasta> scaffold_length_cutoff
```

`scaffold_length_cutoff (Kbp)`

It will:

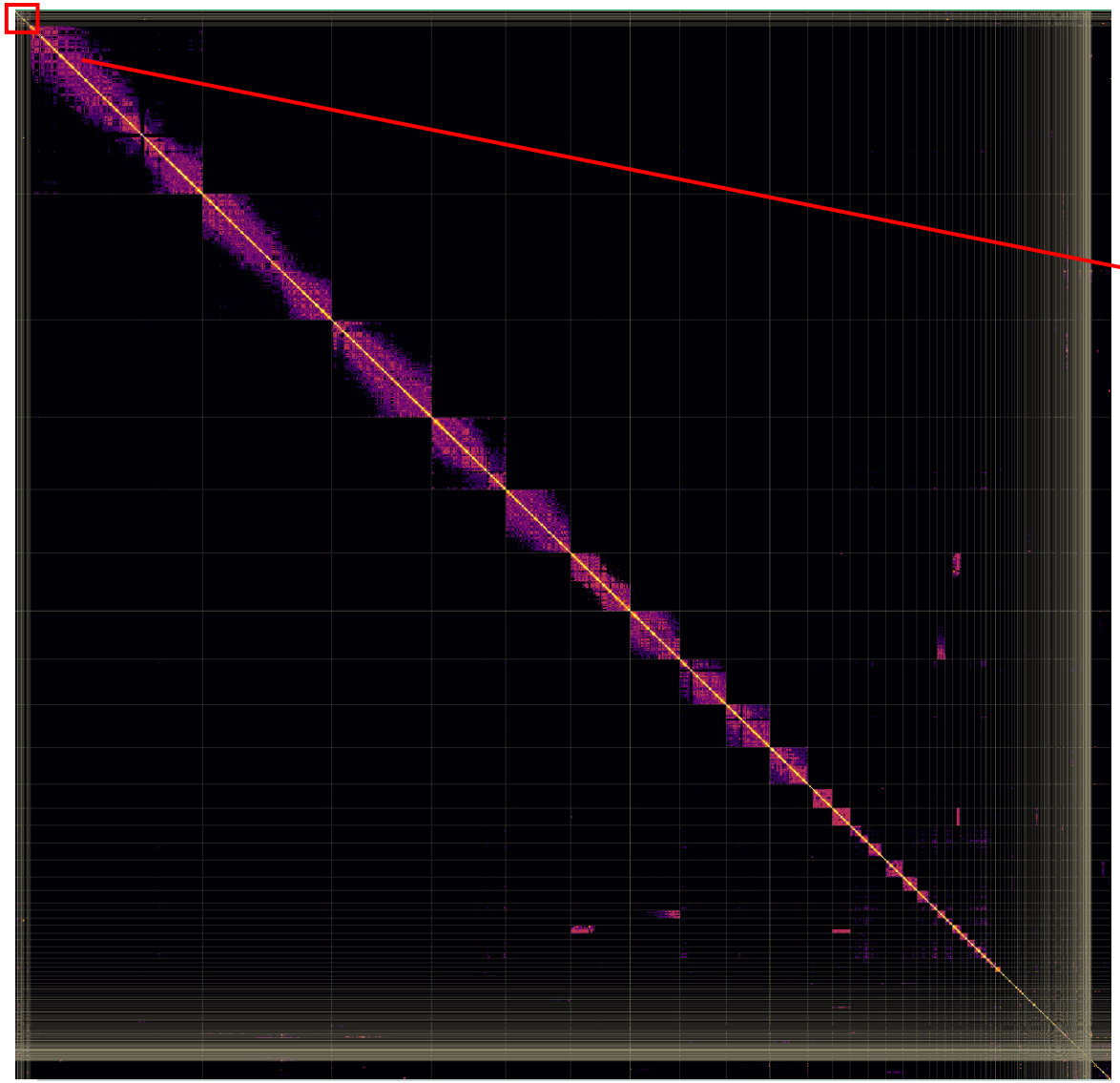
Align your genome to a conserved database of bird microchromosomes and look for gene content

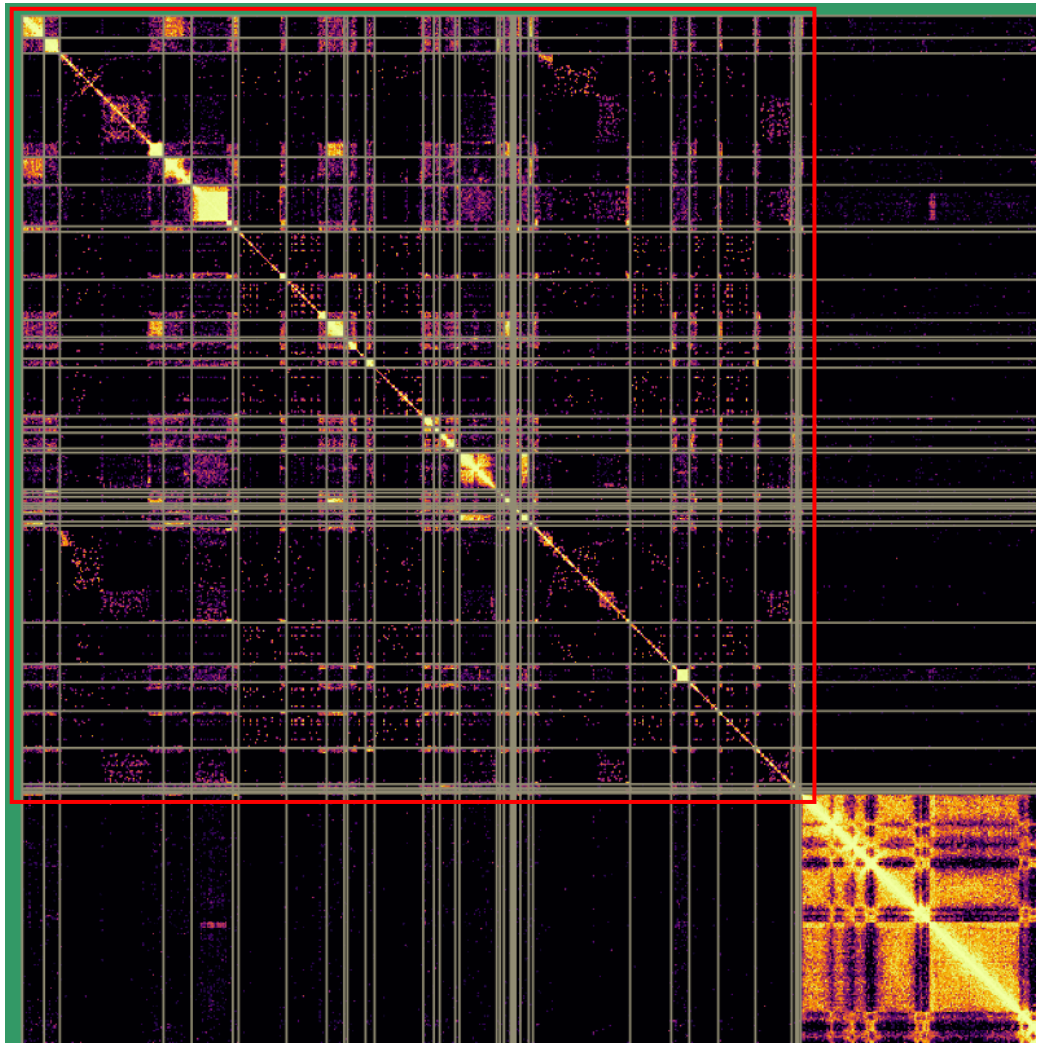
Sort by number of gene hits and then by size (< 5Mbp only) and move them to the beginning of the fasta file

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





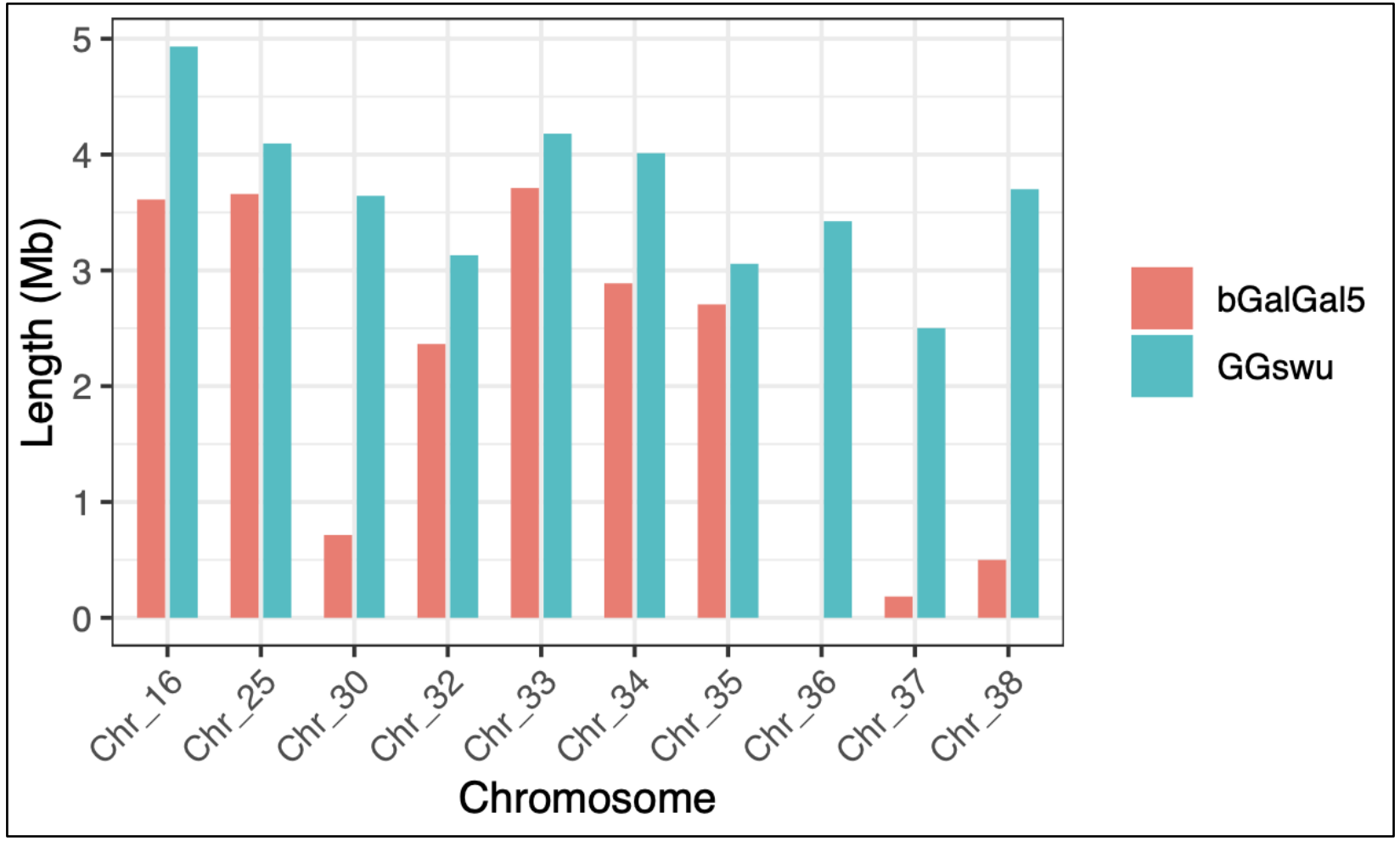
Ordered by gene count

Is it possible to have more complete bird microchromosomes?

Nanopore data looks promising!!!

Bird and fish gene-rich regions

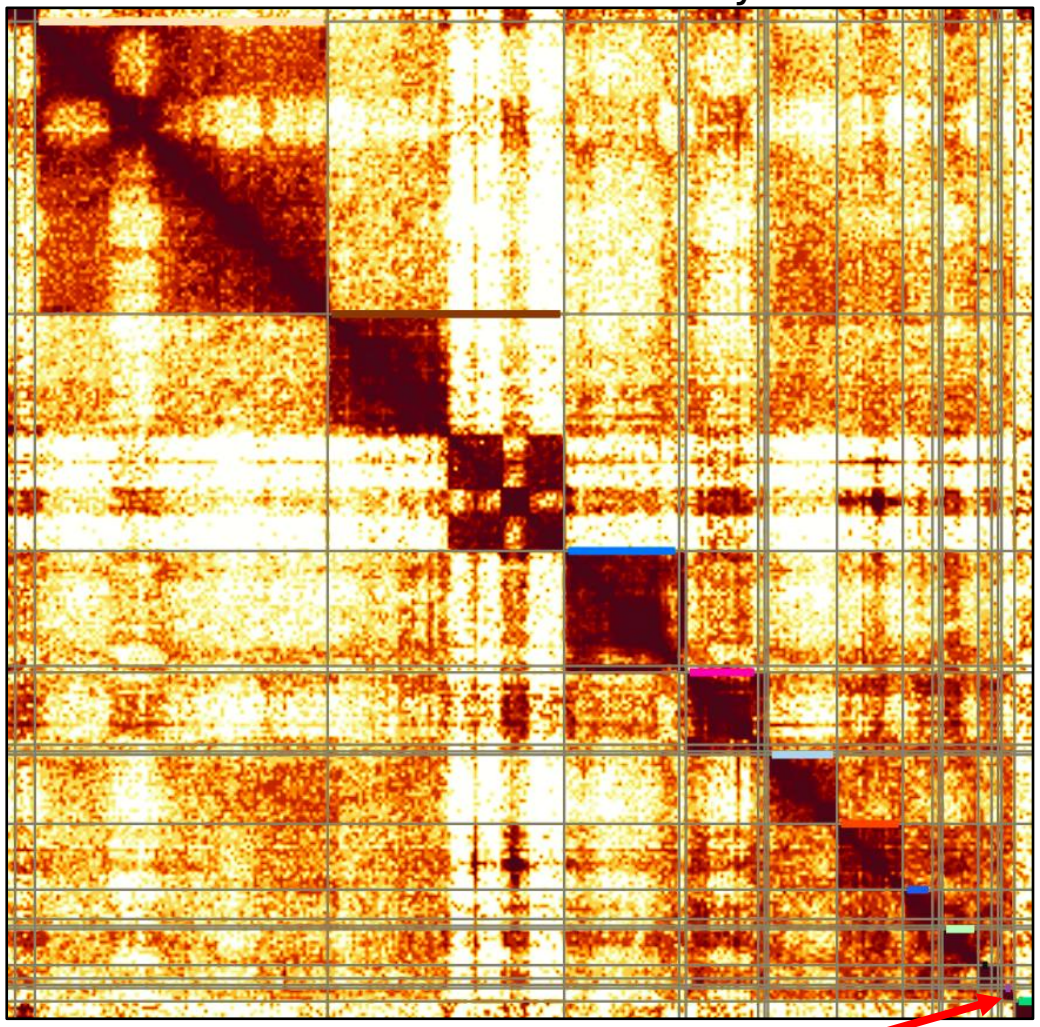
“Dot” chromosomes are substantially shorter in bGalGal5 (HiFi) than Ggswu (HiFi + ONT)



*In contrast, the size difference of the macrochromosomes is < 5%

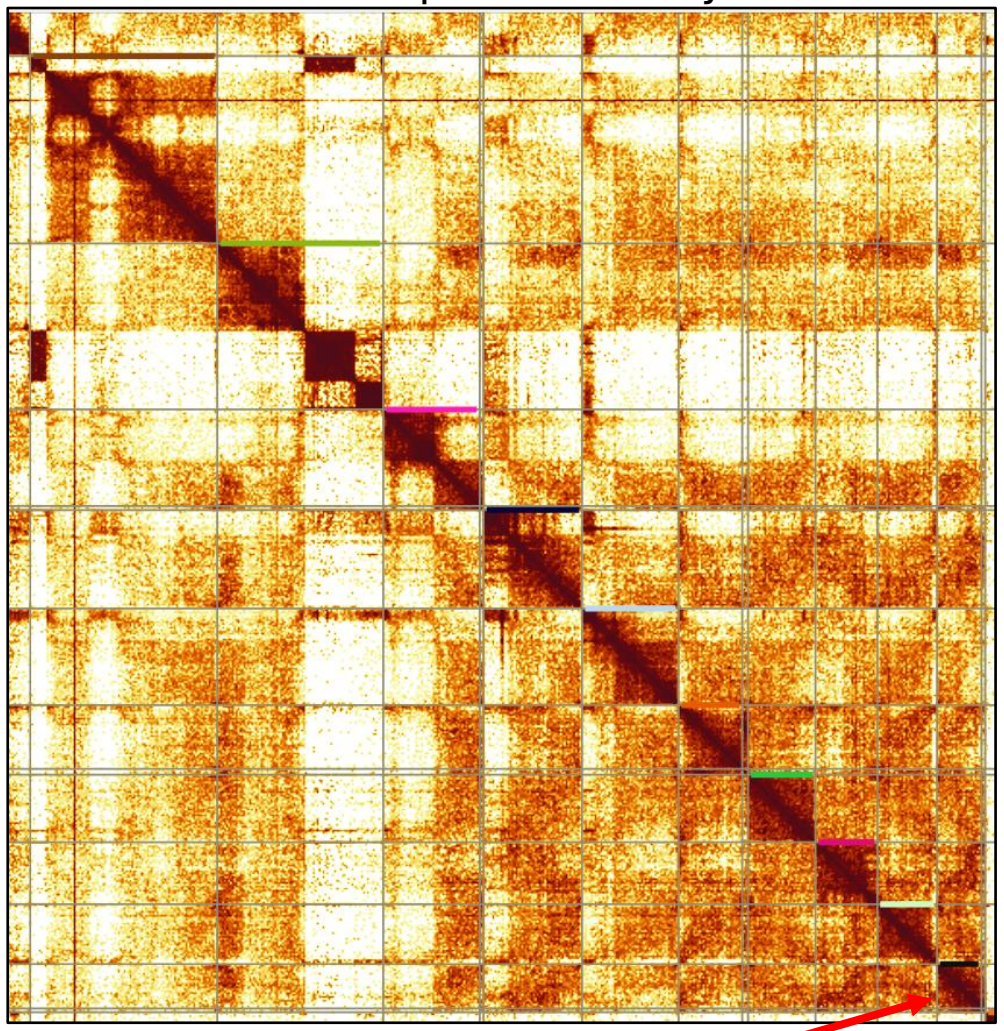
bAytFul3 Aythya fuligula (tufted duck) smallest10 chromosomes

PacBio HiFi assembly



Chr 39 = 125 kb

Nanopore assembly

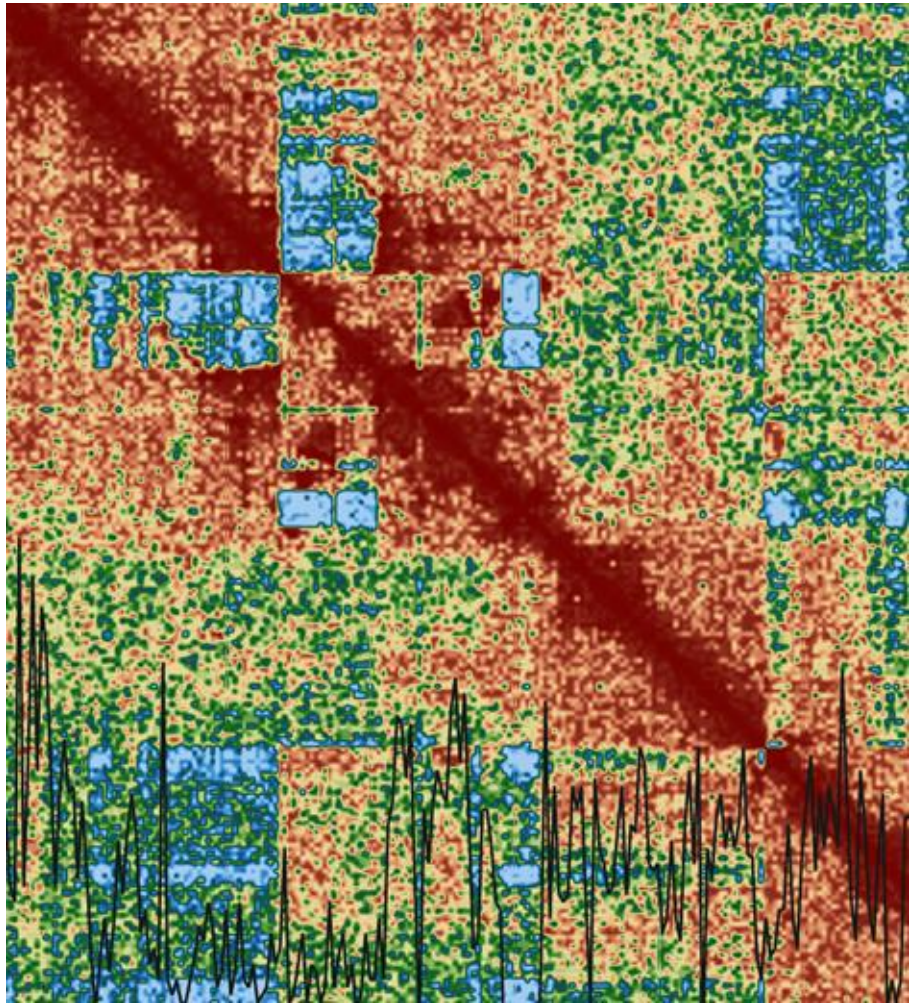


Chr 39 = 1.15Mb

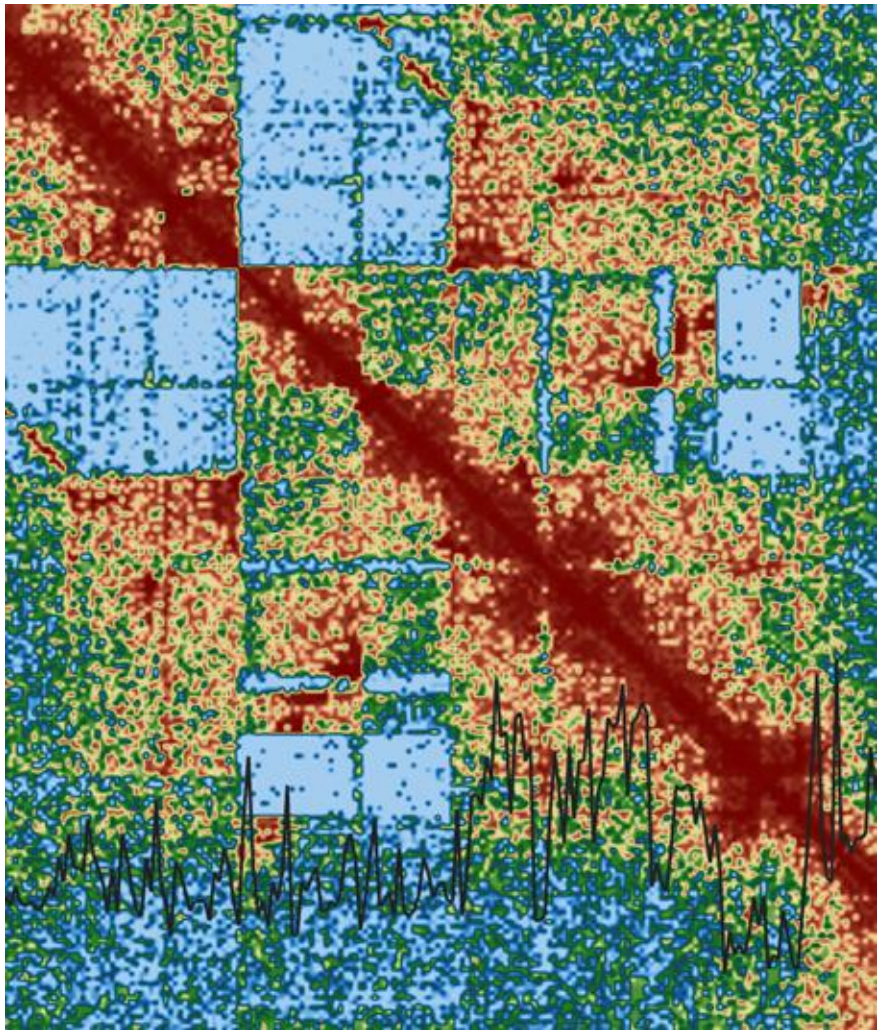
Pretext normal vs. high resolution maps

High res should be the option for all assemblies, except poor HiC signal

Same zoom level



Works well
for haplotigs



Normal resolution

More details when you zoom-in

High resolution

All haplotypes genome assembly curation

and

Polyploid genomes

Standard Pipeline Assembly

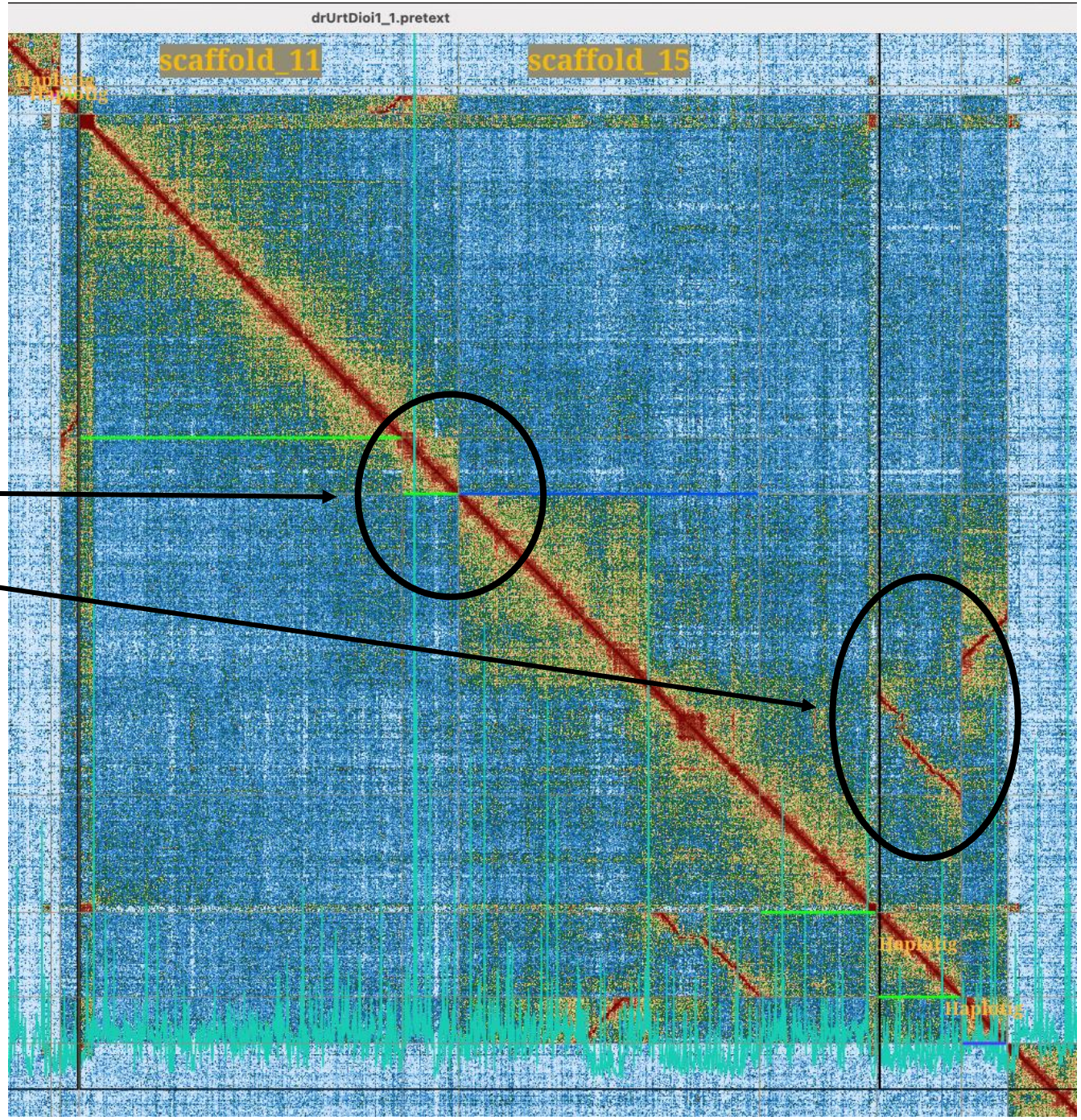
By Dominic Absolon

drUrtDioi1 – tetraploid

Initial “primary” assembly had issues:

- Missing sequence
- Over-represented sequences

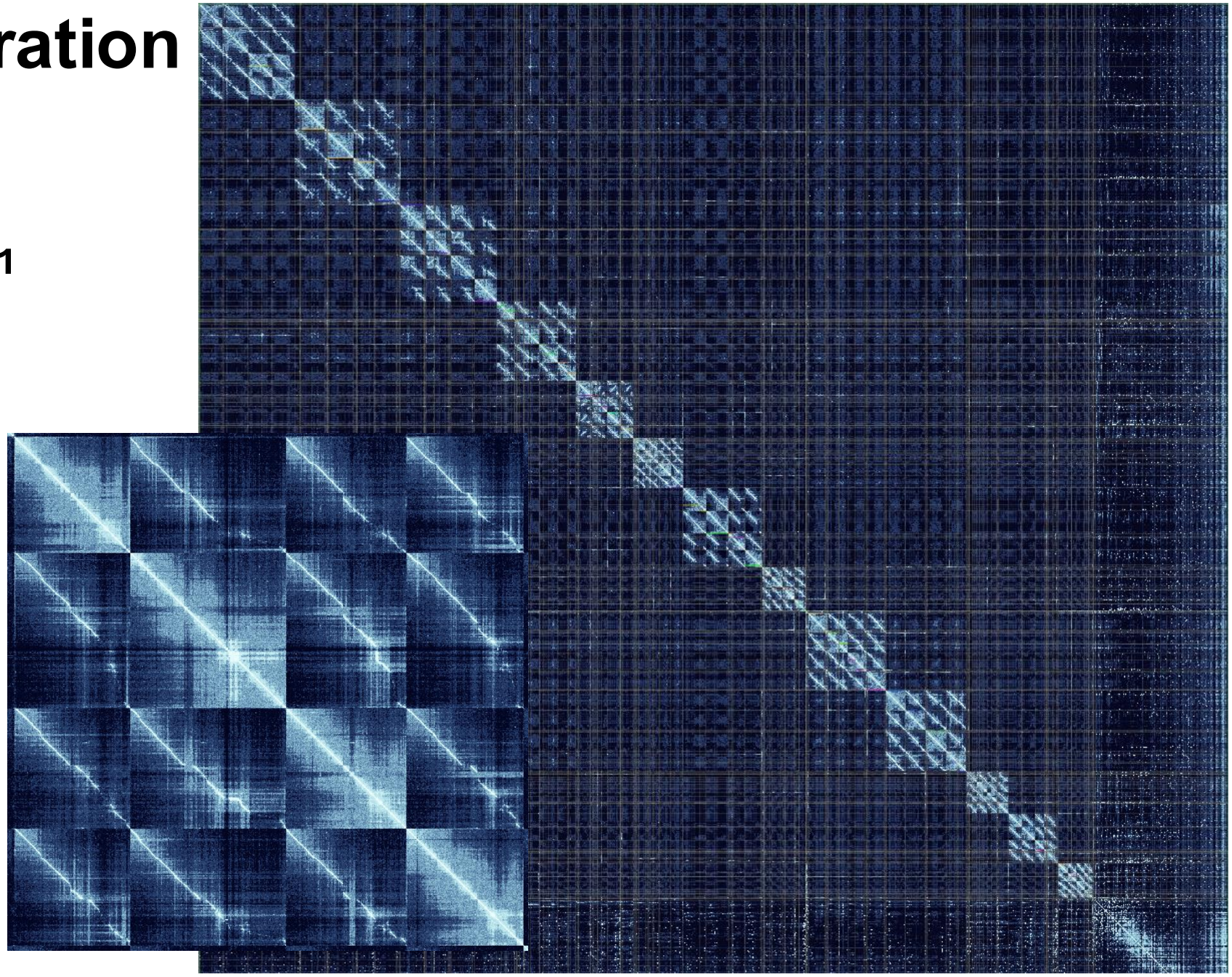
It doesn’t work well as a primary/alternative



All haplotypes assembly and curation

HAP1 file:
Chromosome-level curated HAP1

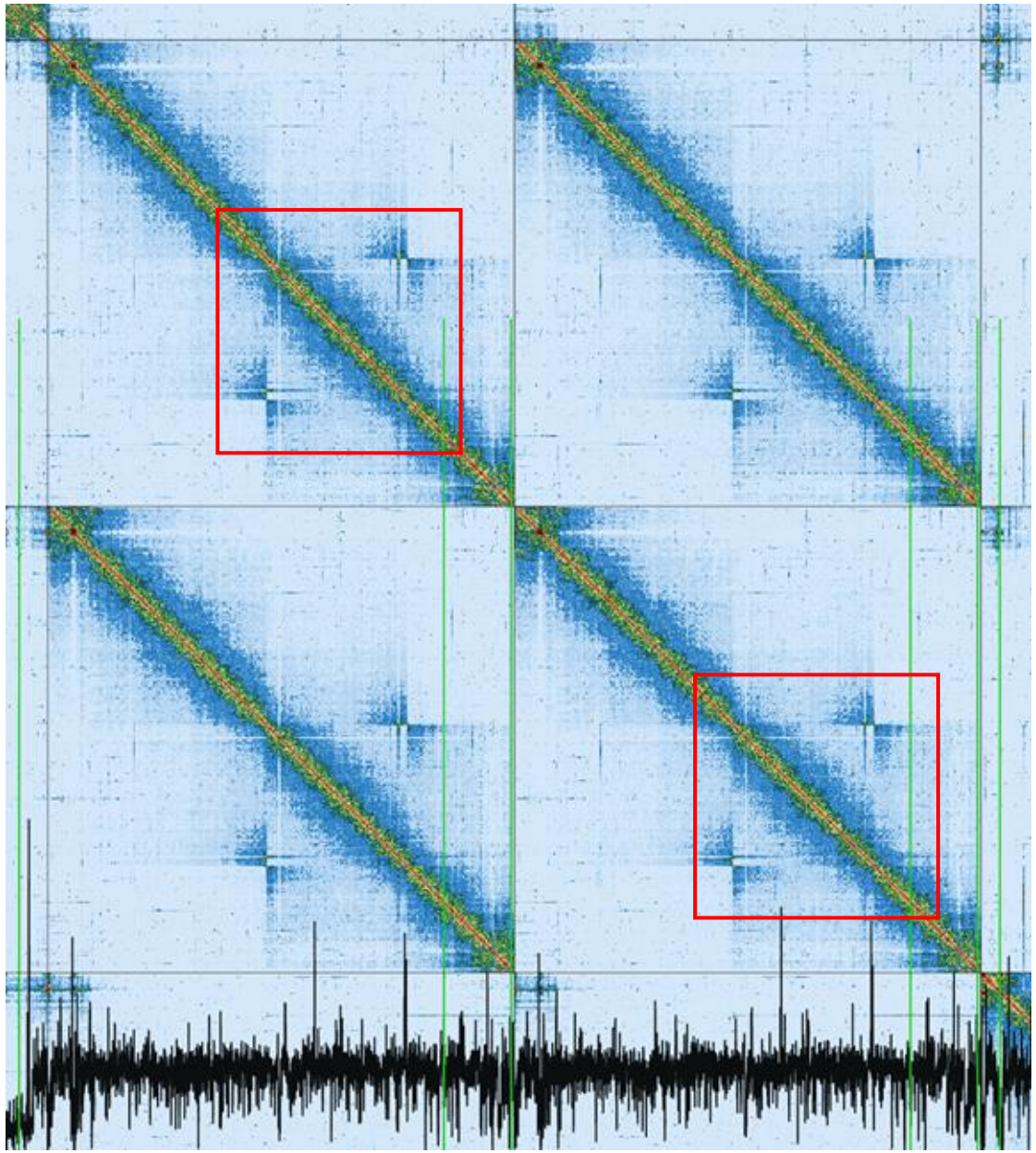
HAP2 file:
Scaffold-level
HAP2, HAP3 and HAP4



All haplotypes/**phased** assemblies can be helpful when:

PacBio and HiC from the same sample

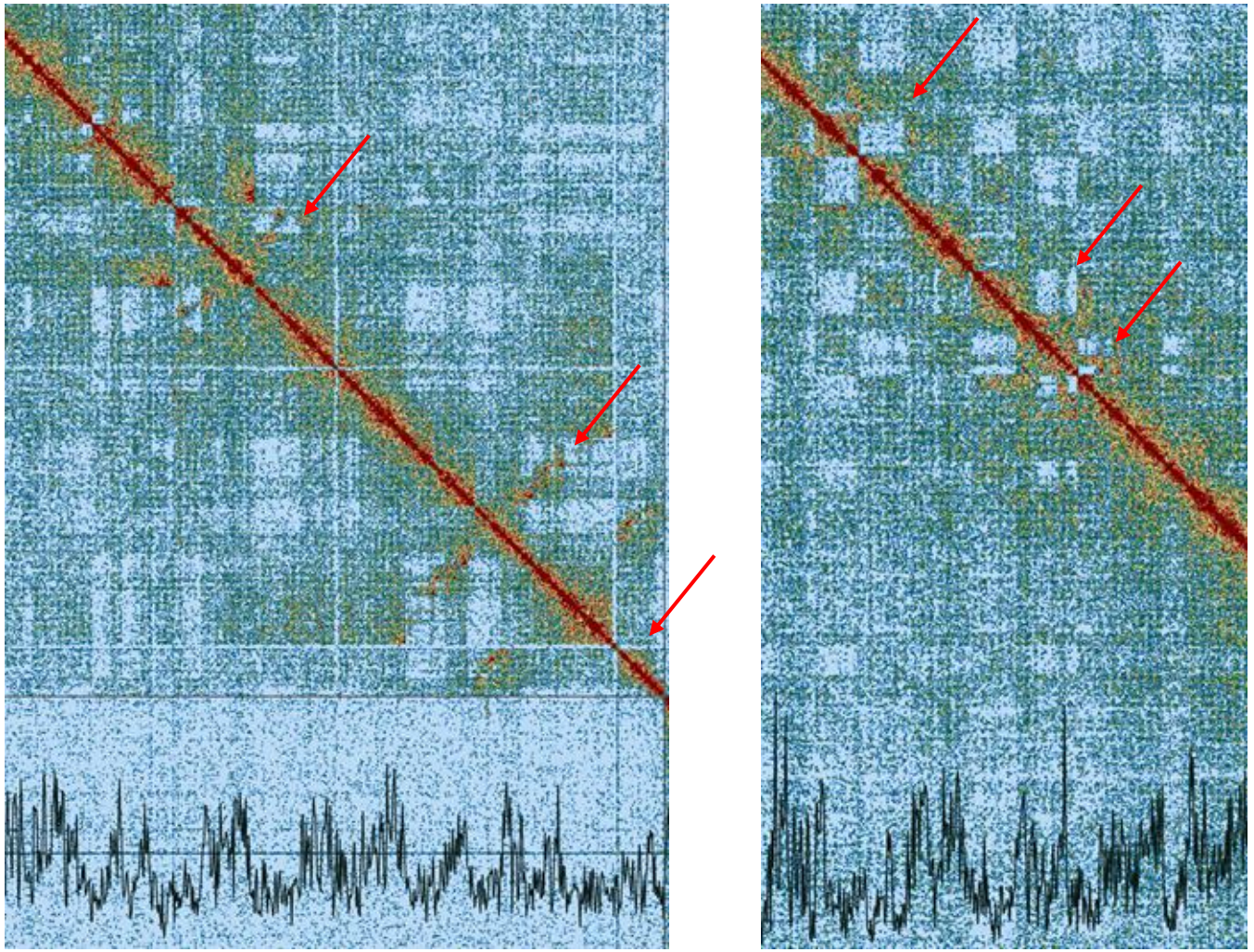
What happens when PB and HiC are from different samples? – Phased assemblies



ieBaeAtla2

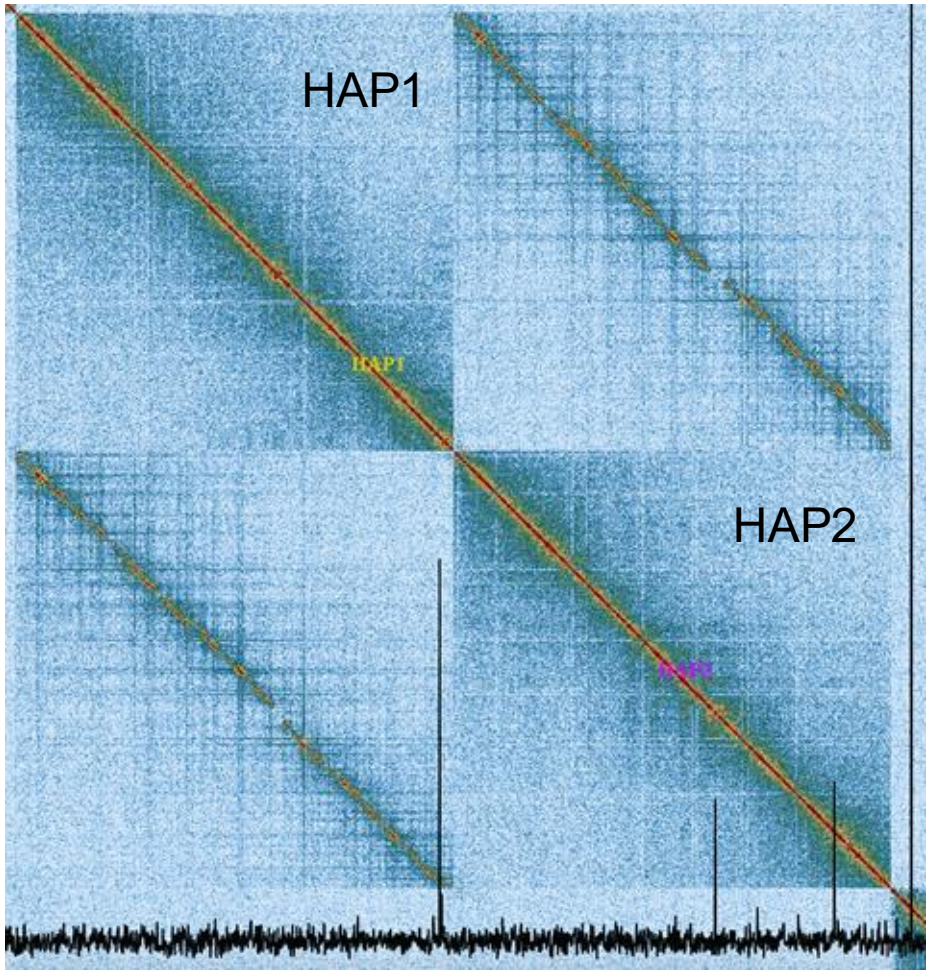
Many retained haplotigs/purging doesn't work

Hifiasm purged assembly looks like this



Many retained haplotigs

Phased assembly



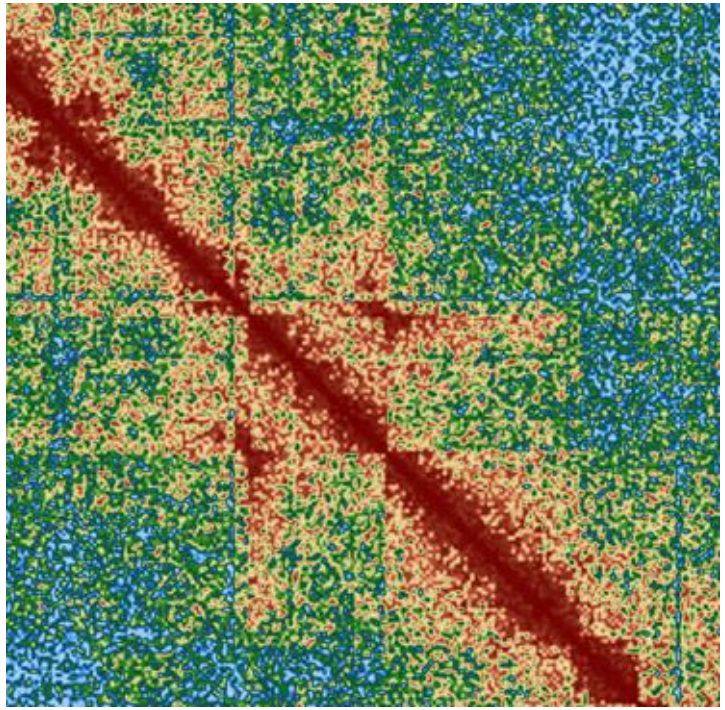
No more haplotigs

Polymorphisms between haplotypes - Inversions

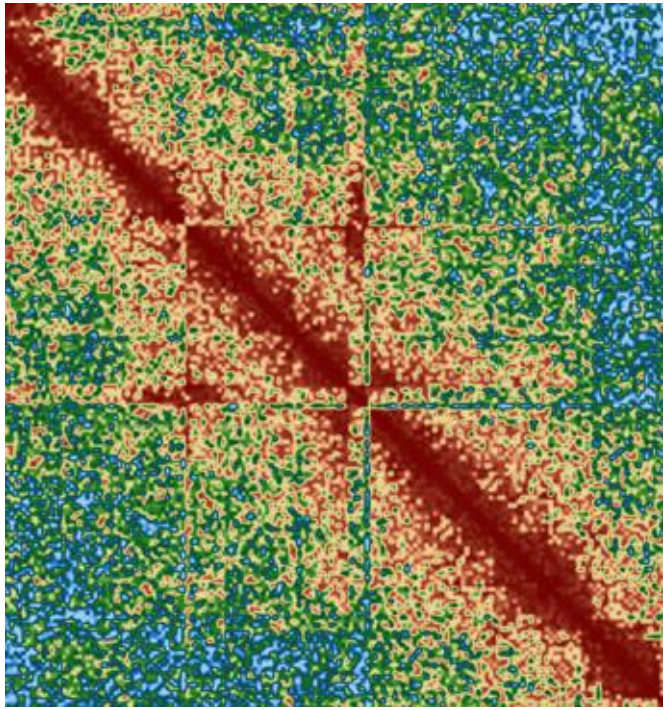
High heterozygosity + inversions between haplotypes
(sister chromatids)

Primary assembly
Inversion
Never looks right

Conformation1

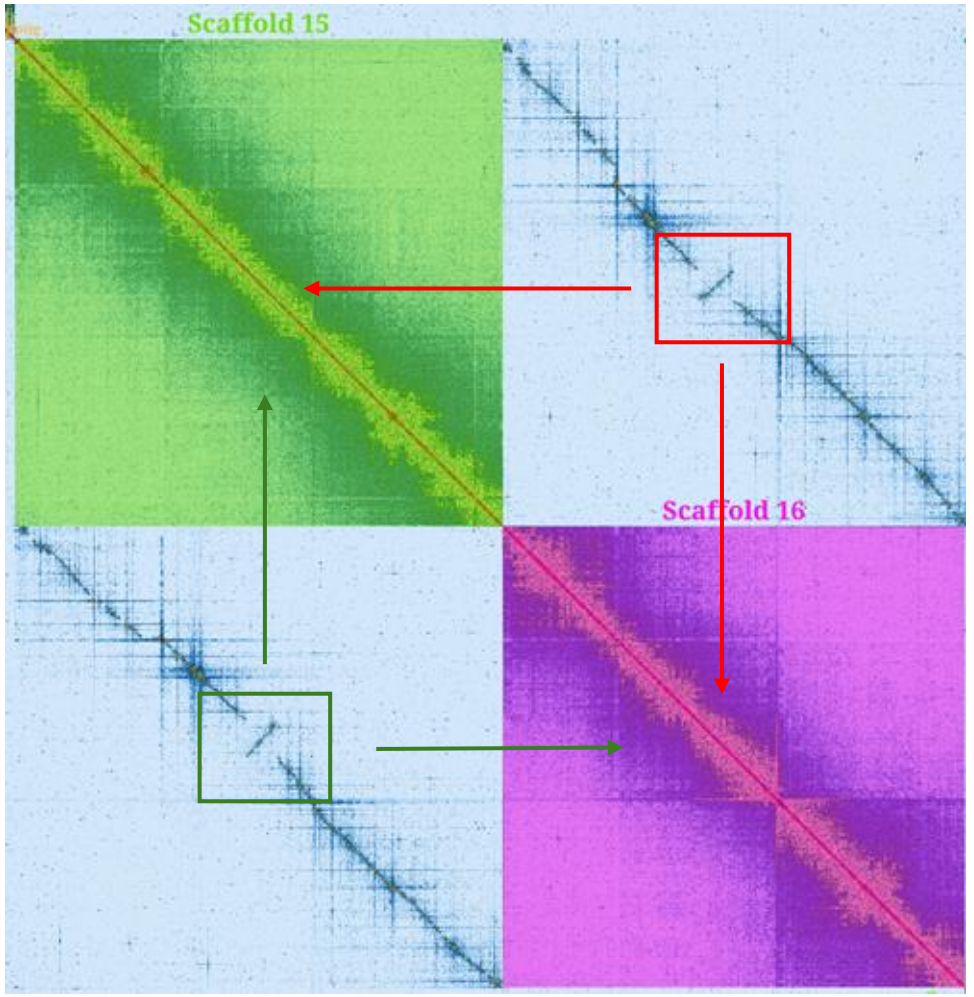


Conformation 2



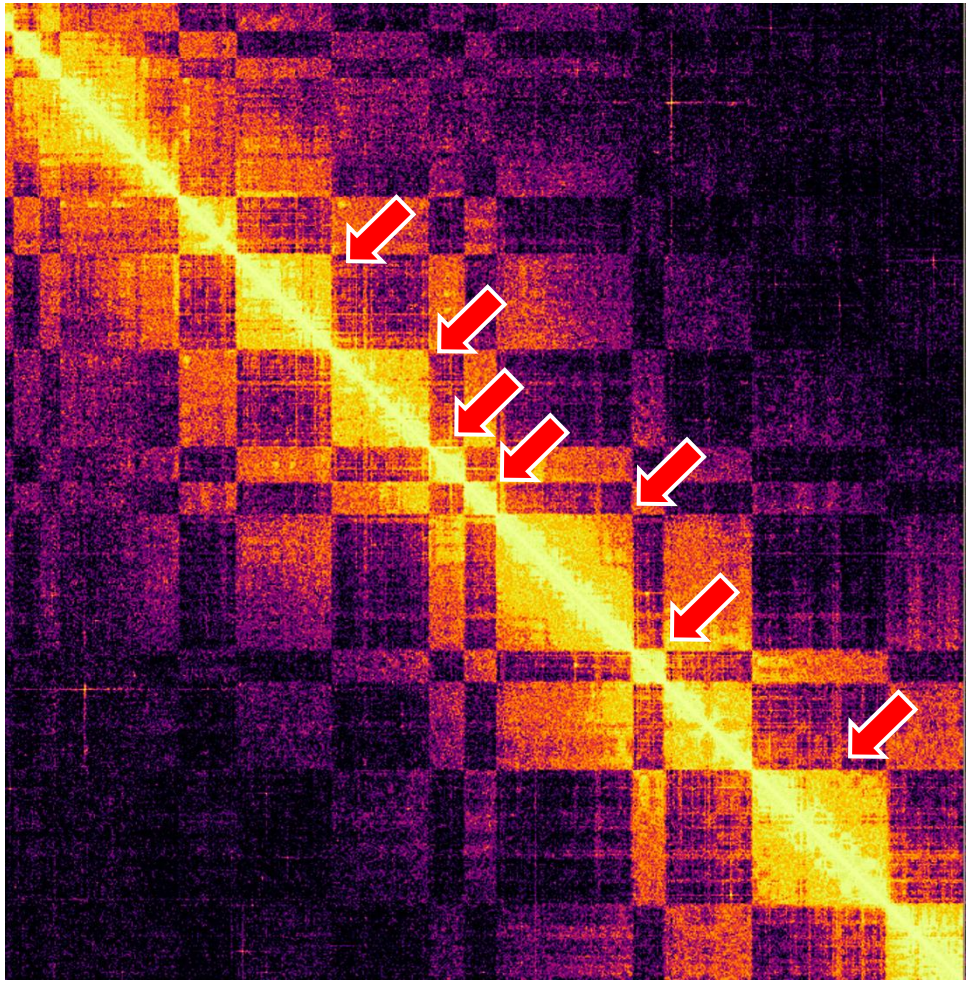
xbArcSenh1

Pri + alt scaffolded together assembly
Inversion
Resolved when 2 haplotypes are available

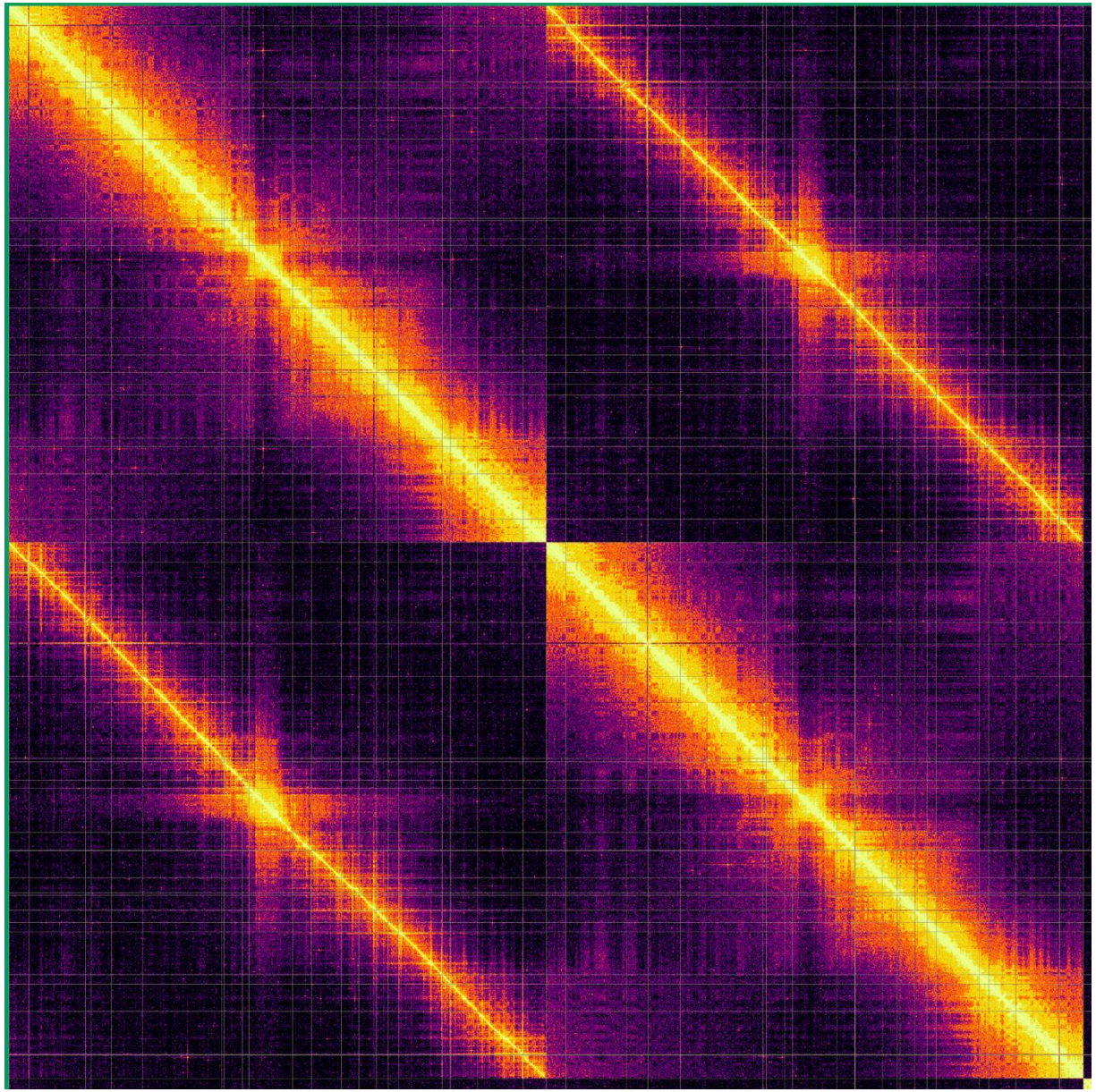


Haplotype bad phasing

When it doesn’t work it is a source of confusion…

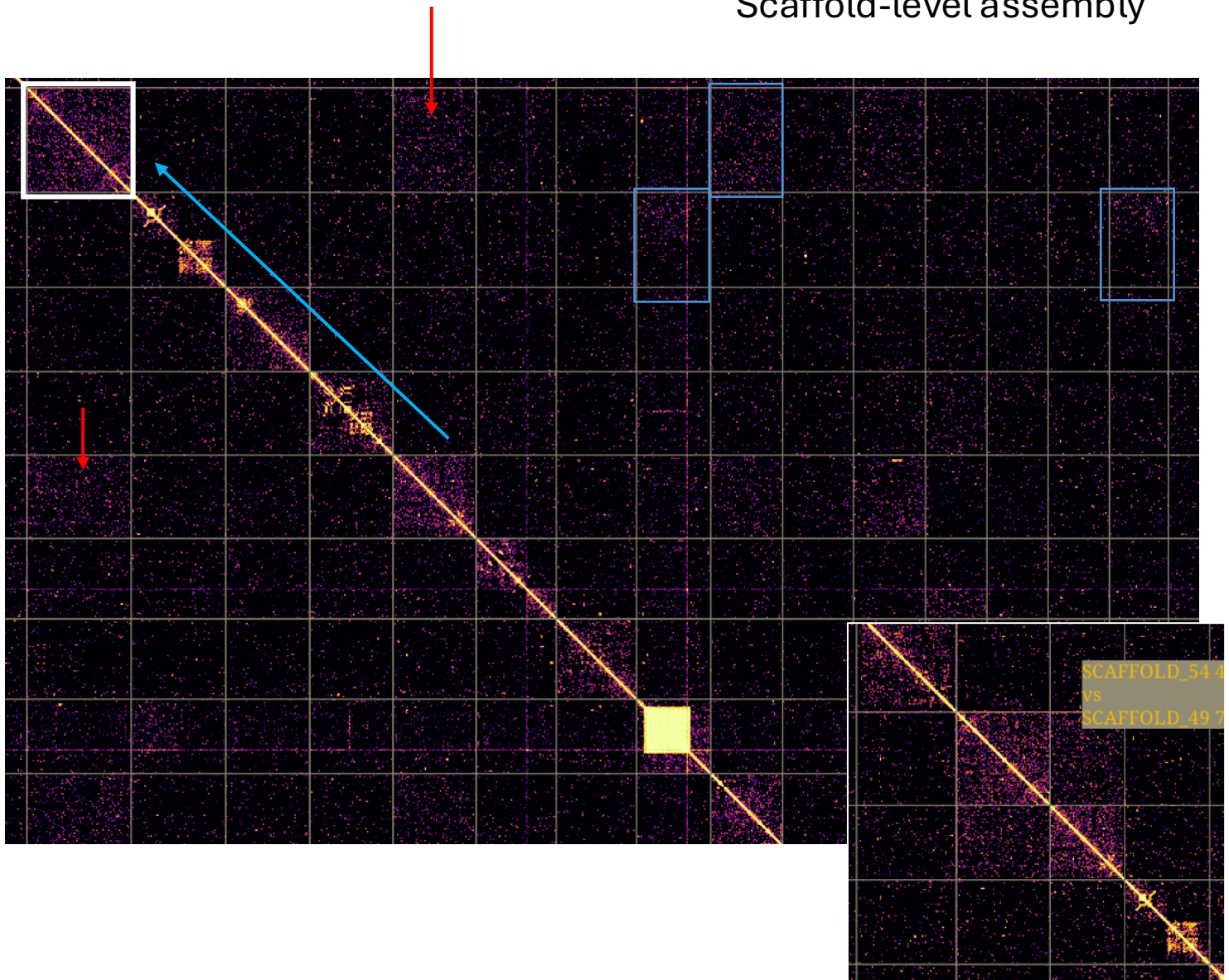


After manual phasing

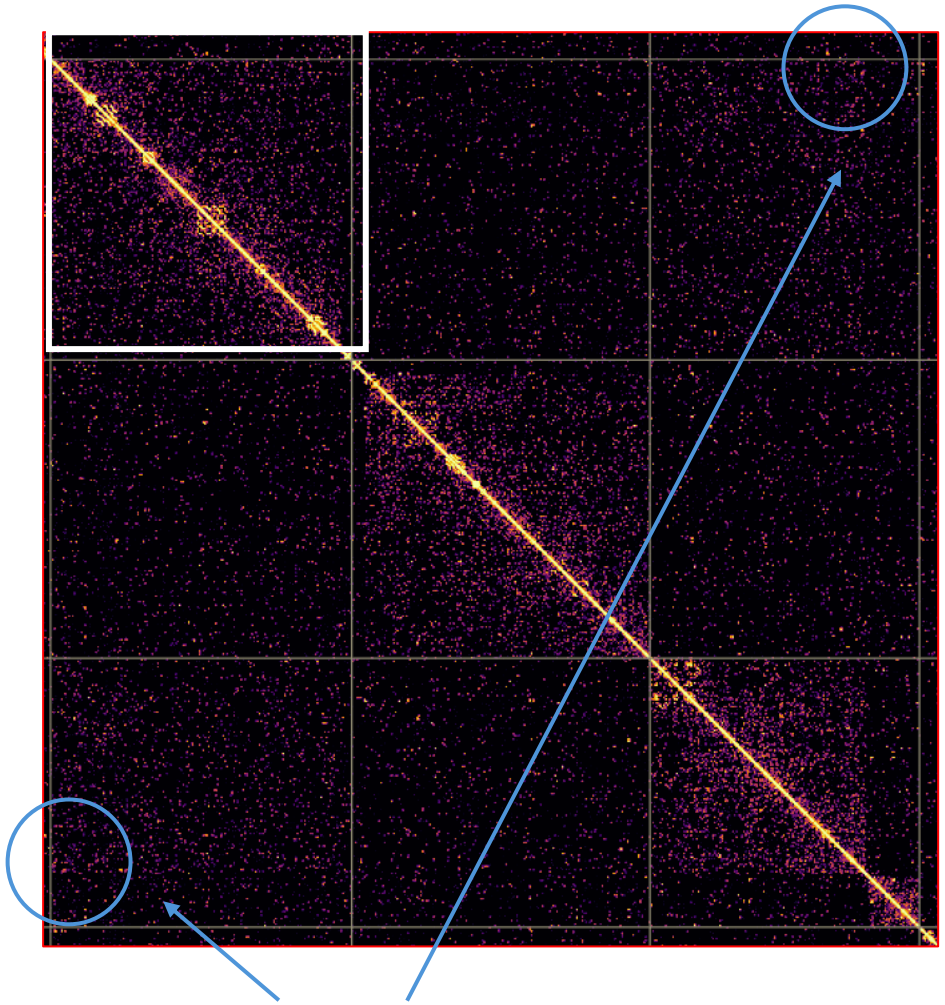


High chromosome number + Bad HiC + no telo information

Scaffold-level assembly



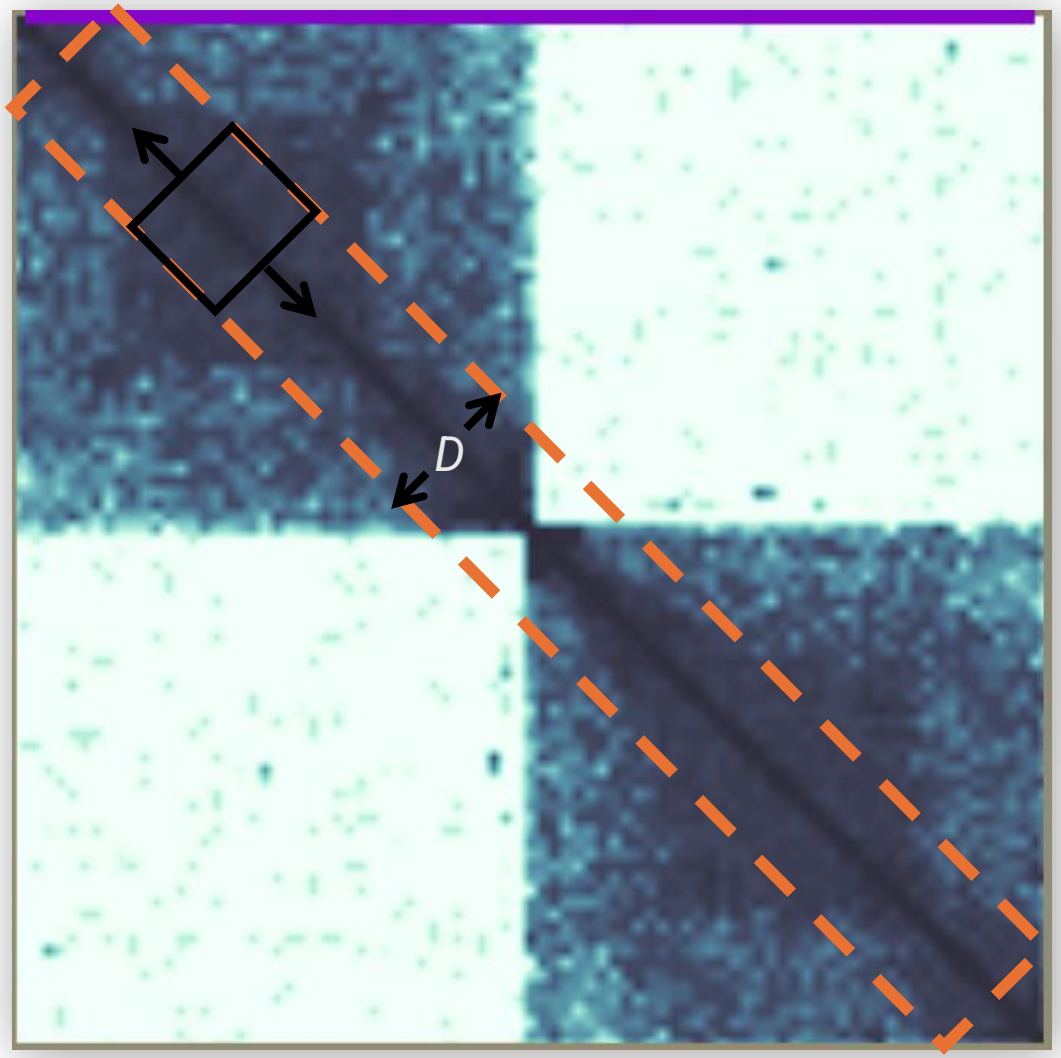
After zoom in

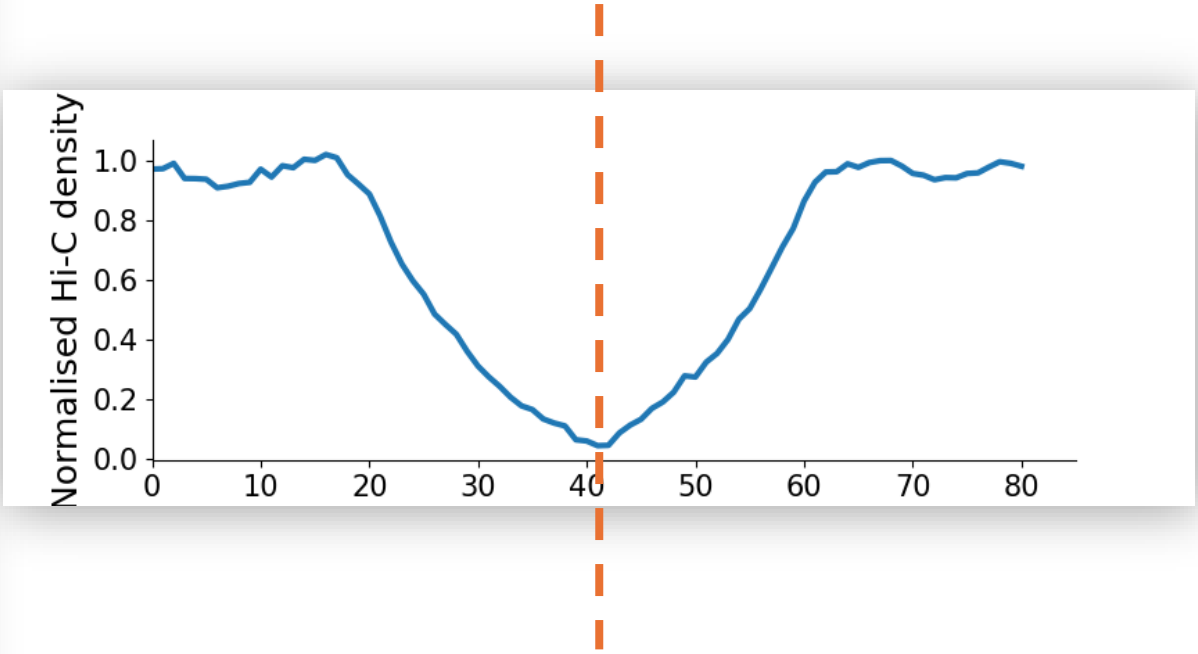


Strongest HiC signal

PretextView AI features

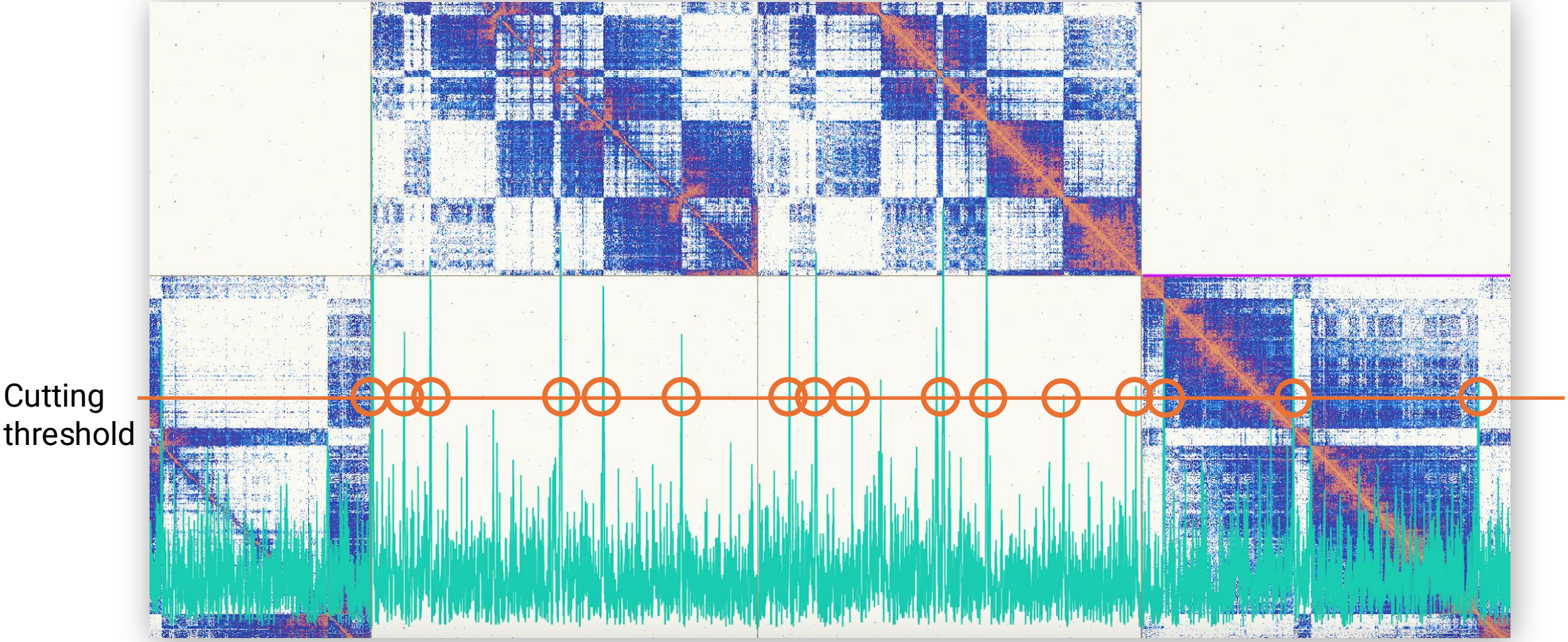
Pixel cut





PretextView AI features

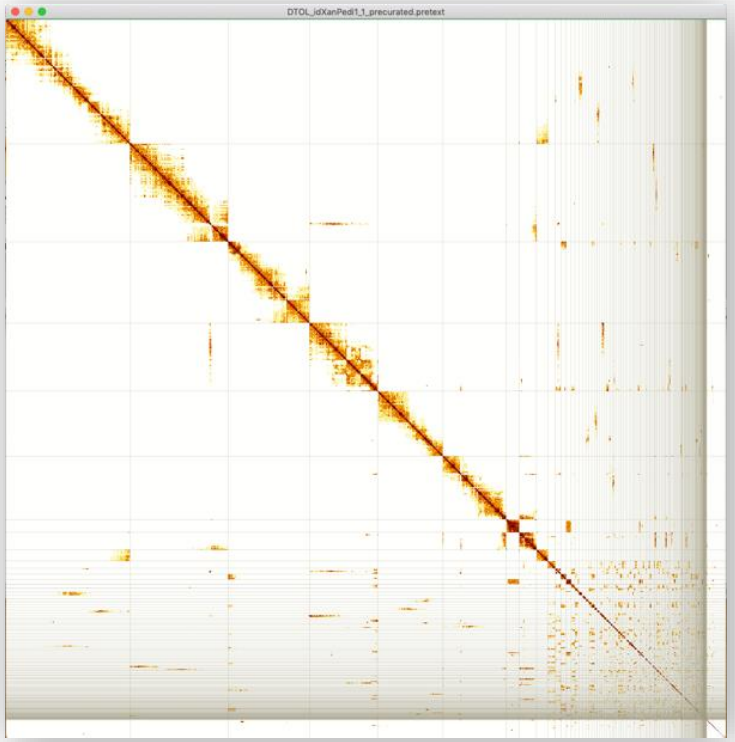
Pixel cut



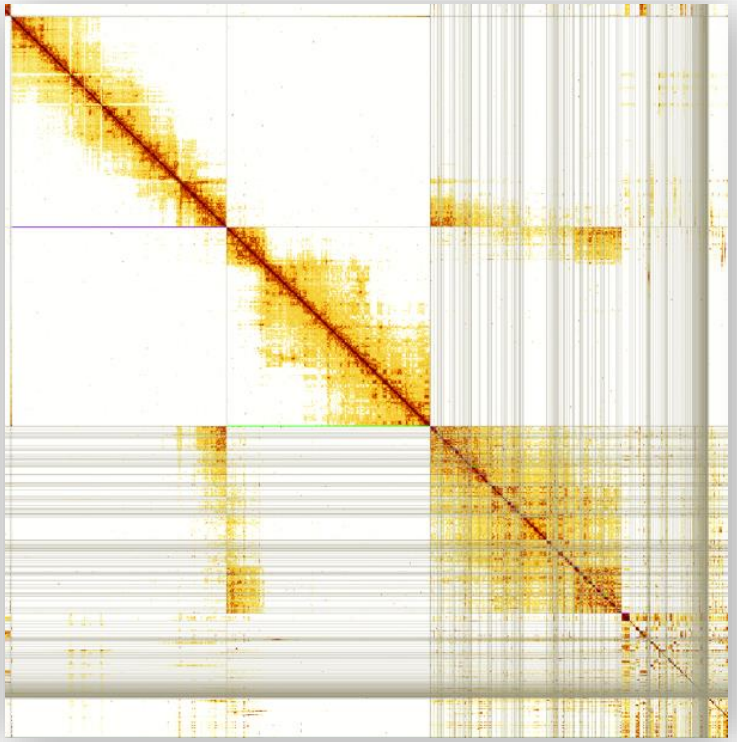
PretextView AI features

Xanthogramma pedissequum (Hoverfly)

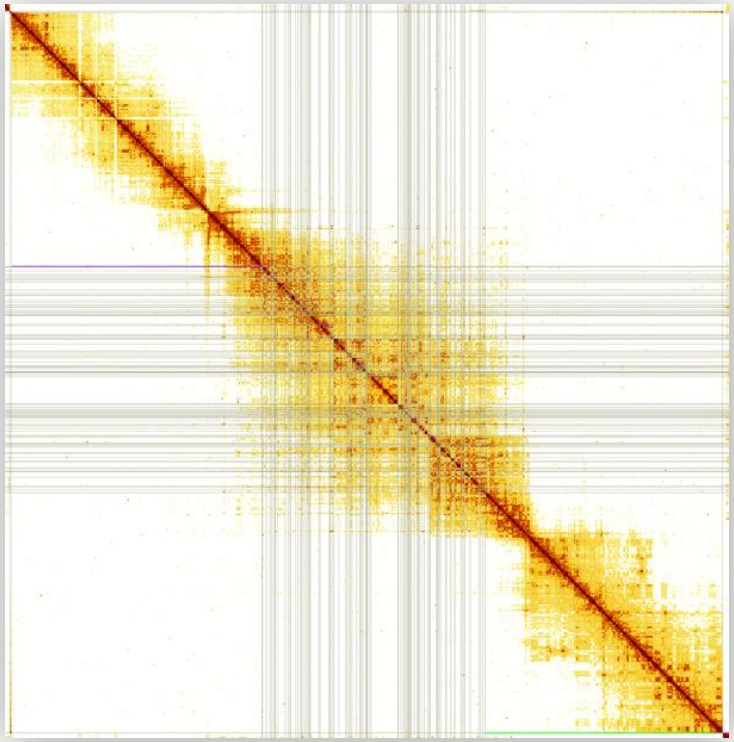
Pixel sort (‘F’ key shortcut)



Pre-curation



Grouped and ordered



Resolved

The finishing process – painting

After curation you should:

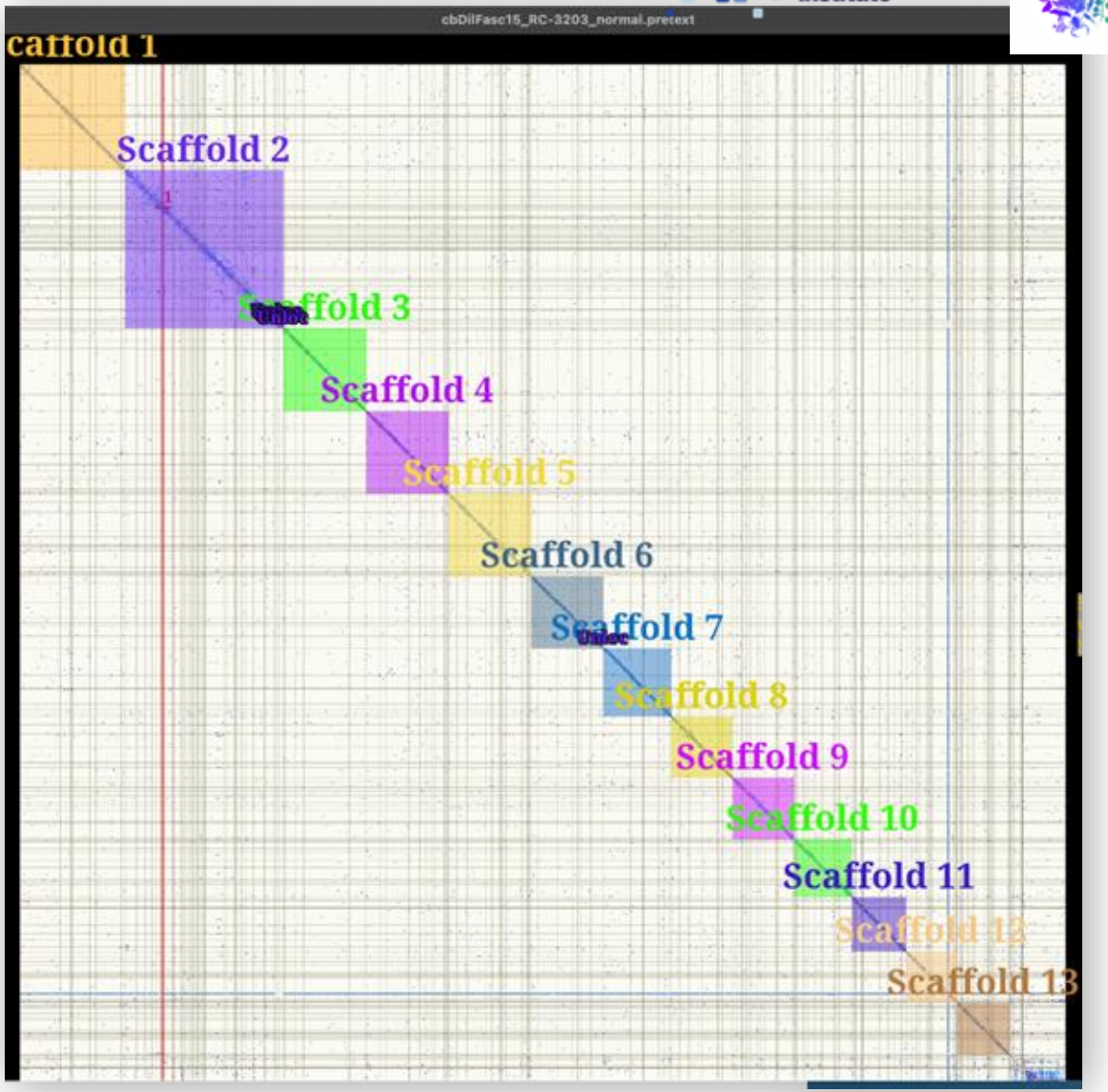
Add all relevant metadata tags

Paint chromosomes

AGP and savestate generation

Curated fasta file

High quality chromosome-level assemblies



AGP generation

https://www.ncbi.nlm.nih.gov/genbank/genome_agp_specification/

```
  GNU nano 6.2                                                    odGeoParv2_1_normal.pretext.agp_1
##agp-version   2.1
# DESCRIPTION: Generated by PretextView Version 0.2.5
# HiC MAP RESOLUTION: 3951.358154 bp/texel
Scaffold_1      1       15805   1       W       SCAFFOLD_8      1880847 1896651 +       Painted Haplotig
Scaffold_1      15806   15905   2       U       100     scaffold        yes     proximity_ligation
Scaffold_1      15906   122591  3       W       SCAFFOLD_34     1       106686  +       Painted
Scaffold_1      122592  122691  4       U       100     scaffold        yes     proximity_ligation
Scaffold_1      122692  3066453 5       W       SCAFFOLD_8      2117928 5061689 -       Painted
Scaffold_1      3066454 3066553 6       U       100     scaffold        yes     proximity_ligation
Scaffold_1      3066554 3141628 7       W       SCAFFOLD_43     1       75075   +       Painted
Scaffold_1      3141629 3141728 8       U       100     scaffold        yes     proximity_ligation
Scaffold_1      3141729 3363004 9       W       SCAFFOLD_8      1896652 2117927 -       Painted
Scaffold_1      3363005 3363104 10      U       100     scaffold        yes     proximity_ligation
Scaffold_1      3363105 4303527 11      W       SCAFFOLD_8      940424  1880846 +       Painted
Scaffold_1      4303528 4303627 12      U       100     scaffold        yes     proximity_ligation
Scaffold_1      4303628 6303014 13      W       SCAFFOLD_1      1       1999387 +       Painted
Scaffold_1      6303015 6303114 14      U       100     scaffold        yes     proximity_ligation
Scaffold_1      6303115 6322870 15      W       SCAFFOLD_83     1       19756   -       Painted
Scaffold_1      6322871 6322970 16      U       100     scaffold        yes     proximity_ligation
Scaffold_1      6322971 15047569        17      W       SCAFFOLD_1      1999388 10723986        +       Painted
Scaffold_2      1       2789658 1       W       SCAFFOLD_23     1       2789658 -       Painted
Scaffold_2      2789659 2789758 2       U       100     scaffold        yes     proximity_ligation
Scaffold_2      2789759 7349626 3       W       SCAFFOLD_1      13944343        18504210        +       Painted
Scaffold_3      1       7558948 1       W       SCAFFOLD_2      1       7558948 +       Painted
Scaffold_3      7558949 7559048 2       U       100     scaffold        yes     proximity_ligation
Scaffold_3      7559049 8037162 3       W       SCAFFOLD_2      7558949 8037062 -       Painted
```

Generating the curated fasta file

```
pretext-to-asm -a <original>.fa -p <output_from_pretextview>.agp -o <assembly_name>.fa
```

pretext-to-asm

https://github.com/sanger-tol/agp-tpf-utils

```
Usage: pretext-to-asm [OPTIONS]

Options:
  -a, --assembly PATH             Assembly before curation, usually a FASTA
                                  file. FASTA files will be indexed, creating
                                  a '.fai' and a '.agp' file alongside the
                                  assembly if they are missing or are older
                                  than the FASTA.  [required]
  -p, --pretext PATH              Assembly file from Pretext, which is usually
                                  an AGP.  [required]
  -o, --output FILE               Output file template, typically:
                                  '<ToLID>.<VERSION>.fa'

                                  e.g. --output mVulVul1.2.fa

                                  for version 2 of the assembly of 'mVulVul1'.
                                  If <VERSION> is not specified, it defaults
                                  to '1'.

                                  The output file type is determined from its
                                  extension. When the outuput is FASTA
                                  ('.fa'), an AGP format file ('.fa.agp') is
                                  also written.

                                  The names of output files created are
                                  printed to STDERR.

                                  If not given, prints to STDOUT in 'STR'
                                  format.
  -c, --autosome-prefix TEXT      Prefix for naming autosomal chromosomes.
                                  [default: SUPER_]
  -f, --clobber / --no-clobber    Overwrite any existing output files.
                                  [default: clobber]
  -l, --log-level [debug|info|warning|error|critical]
                                  Diagnostic messages to show.  [default:
                                  INFO]
  -w, --write-log / -W, --no-write-log
                                  Write messages into a '.log' file alongside
                                  the output file  [default: write-log]
  --help                          Show this message and exit.
```

Pretext-to-asm output files

```
ilSchScha1.1.haplotigs.agp
ilSchScha1.1.haplotigs.fa
ilSchScha1.chr_report.csv
ilSchScha1_hap1.1.curated.pretext.agp_1
ilSchScha1.hap1.1.primary.chromosome.list.csv
ilSchScha1.hap1.1.primary.curated.agp
ilSchScha1.hap1.1.primary.curated.fa
ilSchScha1.hap1.1.primary.curated.fa.agp
ilSchScha1.hap1.1.primary.curated.fa.fai
ilSchScha1.hap2.1.primary.chromosome.list.csv
ilSchScha1.hap2.1.primary.curated.agp
ilSchScha1.hap2.1.primary.curated.fa
ilSchScha1.info.yaml
ilSchScha1.log
```

Pretext-to-asm output files

GNU nano 6.2 ilNeoNubi2.chr_report.csv

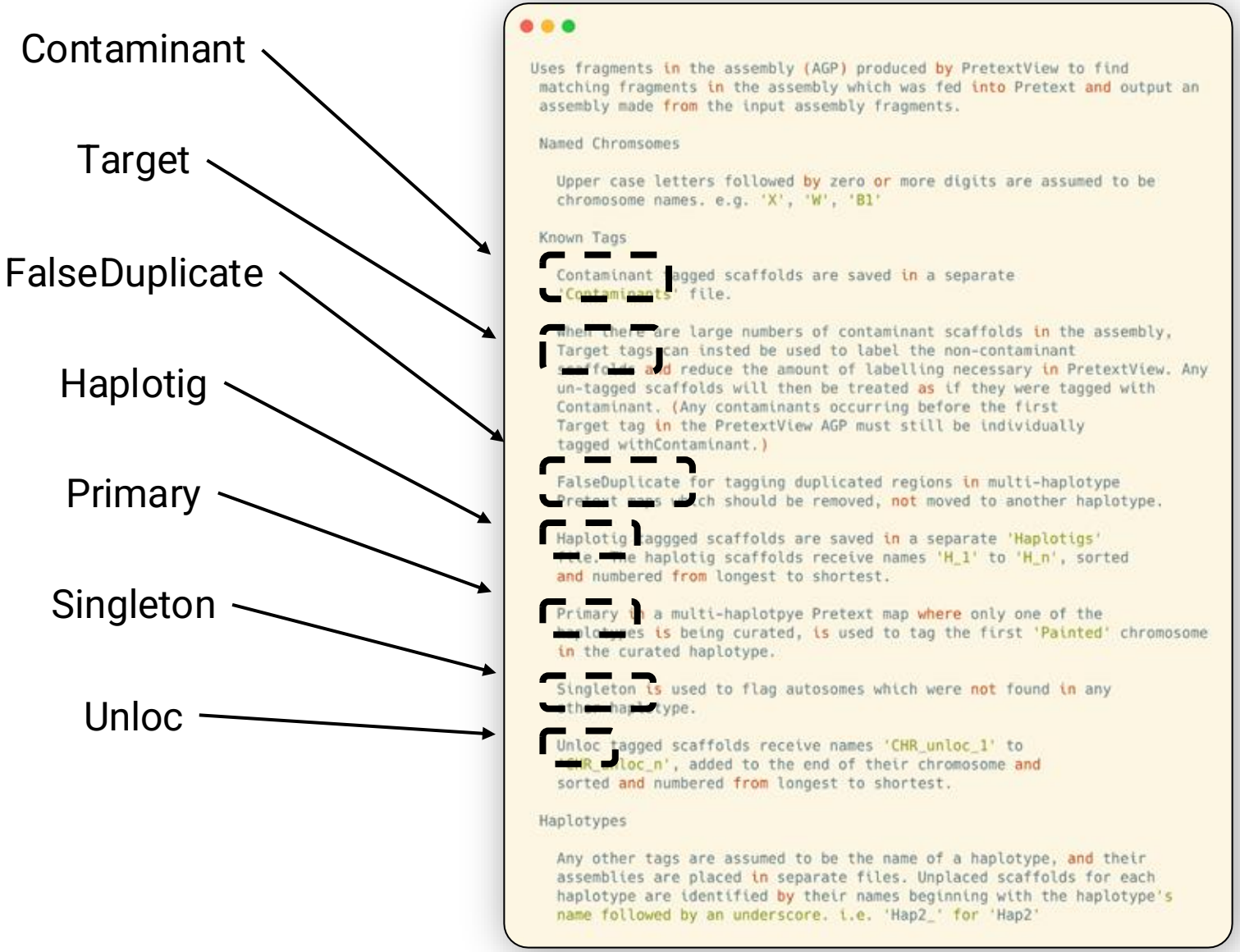
```
"assembly","seq_name","chromosome","localised","pretext_scaffold","length","length_minus_gaps"
"HAP1","SUPER_1","1","true","Scaffold_2",17920404,17920404
"HAP1","SUPER_2","2","true","Scaffold_4",17815506,17815506
"HAP1","SUPER_3","3","true","Scaffold_6",16217648,16217548
"HAP1","SUPER_4","4","true","Scaffold_8",15961867,15961867
"HAP1","SUPER_5","5","true","Scaffold_10",15900027,15900027
"HAP1","SUPER_6","6","true","Scaffold_12",14957033,14957033
"HAP1","SUPER_7","7","true","Scaffold_14",14939051,14939051
"HAP1","SUPER_8","8","true","Scaffold_16",14873331,14873331
"HAP1","SUPER_9","9","true","Scaffold_18",14703592,14703592
"HAP1","SUPER_10","10","true","Scaffold_20",14176904,14176904
"HAP1","SUPER_11","11","true","Scaffold_22",14159098,14159098
"HAP1","SUPER_12","12","true","Scaffold_24",13813620,13813620
"HAP1","SUPER_13","13","true","Scaffold_26",13805808,13805008
"HAP1","SUPER_14","14","true","Scaffold_28",13112795,13112795
"HAP1","SUPER_15","15","true","Scaffold_30",12998824,12998824
"HAP1","SUPER_16","16","true","Scaffold_32",12785512,12785412
"HAP1","SUPER_17","17","true","Scaffold_34",12690657,12690657
"HAP2","SUPER_1","1","true","Scaffold_3",17852375,17852375
"HAP2","SUPER_2","2","true","Scaffold_5",17820748,17820748
"HAP2","SUPER_3","3","true","Scaffold_7",16219065,16219065
"HAP2","SUPER_4","4","true","Scaffold_9",15971563,15971563
"HAP2","SUPER_5","5","true","Scaffold_11",15913097,15913097
"HAP2","SUPER_6","6","true","Scaffold_13",14833091,14833091
"HAP2","SUPER_7","7","true","Scaffold_15",14928166,14928166
"HAP2","SUPER_8","8","true","Scaffold_17",14893242,14893242
"HAP2","SUPER_9","9","true","Scaffold_19",14672243,14672243
"HAP2","SUPER_10","10","true","Scaffold_21",14126870,14126870
"HAP2","SUPER_11","11","true","Scaffold_23",14173908,14173908
"HAP2","SUPER_12","12","true","Scaffold_25",13812745,13812745
"HAP2","SUPER_13","13","true","Scaffold_27",13870117,13869317
"HAP2","SUPER_14","14","true","Scaffold_29",13116826,13116826
"HAP2","SUPER_15","15","true","Scaffold_31",12996534,12996534
"HAP2","SUPER_16","16","true","Scaffold_33",12803231,12803231
```

Chromosome list file

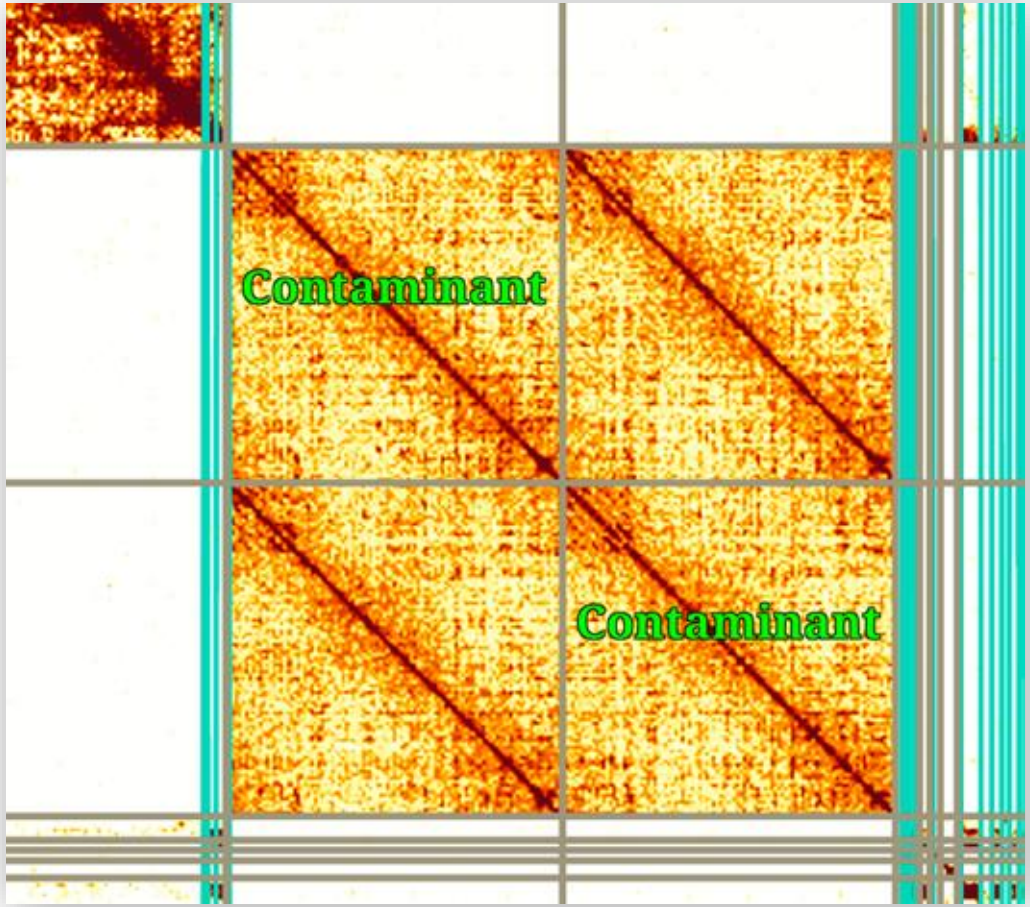
GNU nano 6.2

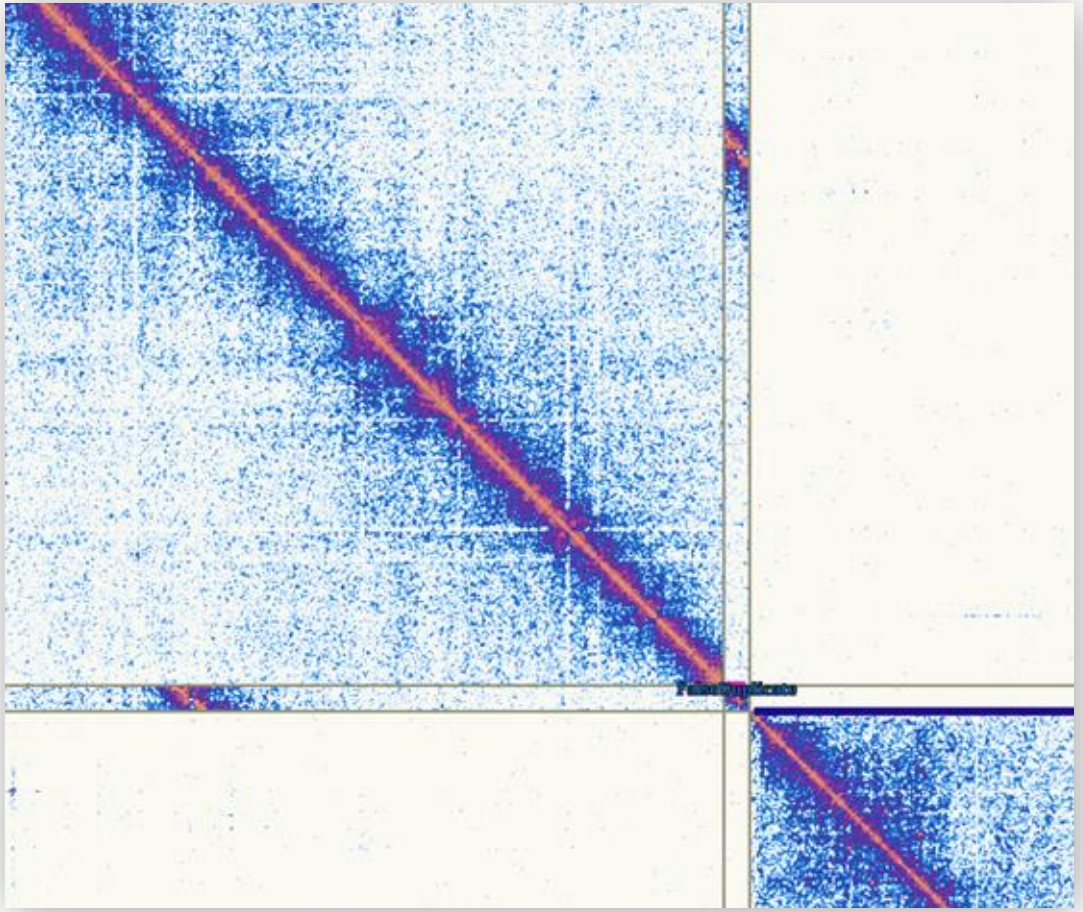
```
SUPER_1,1,yes
SUPER_2,2,yes
SUPER_3,3,yes
SUPER_4,4,yes
SUPER_5,5,yes
SUPER_6,6,yes
SUPER_7,7,yes
SUPER_8,8,yes
SUPER_9,9,yes
SUPER_10,10,yes
SUPER_11,11,yes
SUPER_12,12,yes
SUPER_13,13,yes
SUPER_14,14,yes
SUPER_15,15,yes
SUPER_16,16,yes
SUPER_17,17,yes
SUPER_18,18,yes
SUPER_19,19,yes
SUPER_20,20,yes
SUPER_21,21,yes
SUPER_22,22,yes
SUPER_23,23,yes
SUPER_24,24,yes
SUPER_25,25,yes
SUPER_26,26,yes
SUPER_27,27,yes
SUPER_28,28,yes
SUPER_29,29,yes
SUPER_W,W,yes
SUPER_W_unloc_1,W,no
SUPER_W_unloc_2,W,no
SUPER_W_unloc_3,W,no
SUPER_W_unloc_4,W,no
```

What pretext-to-asm does



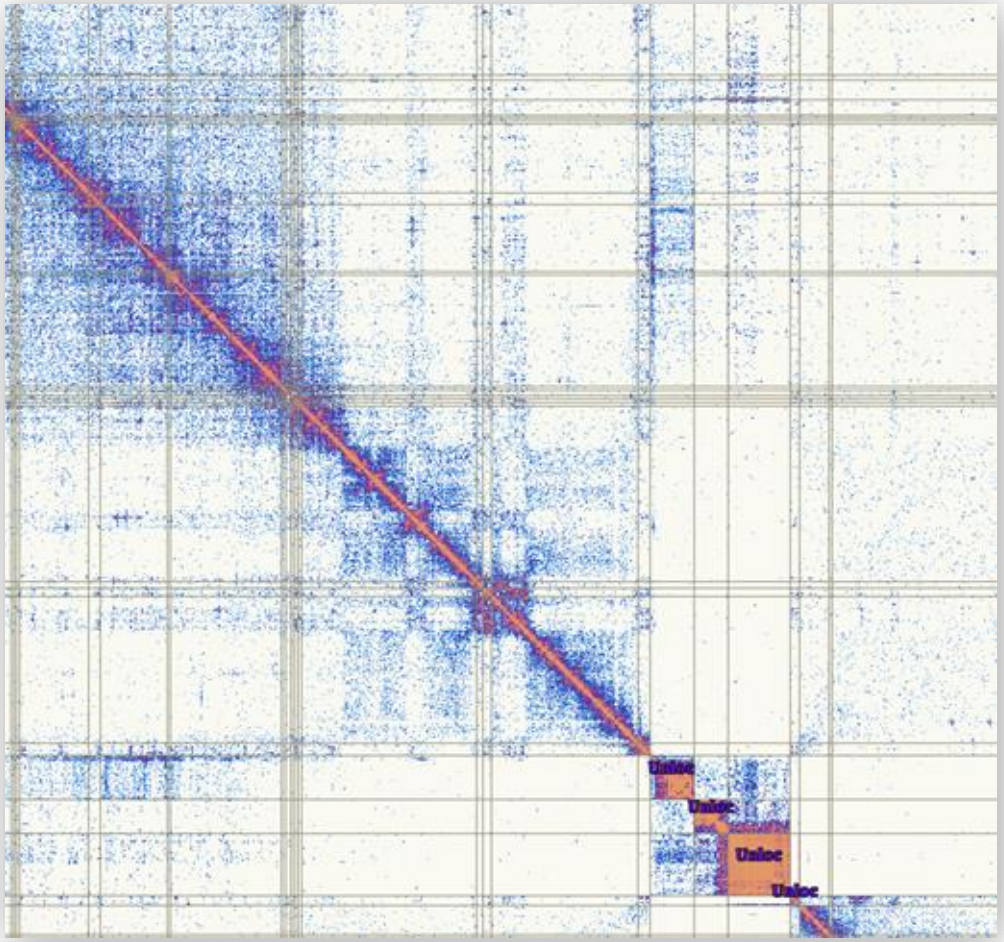
What pretext-to-asm does



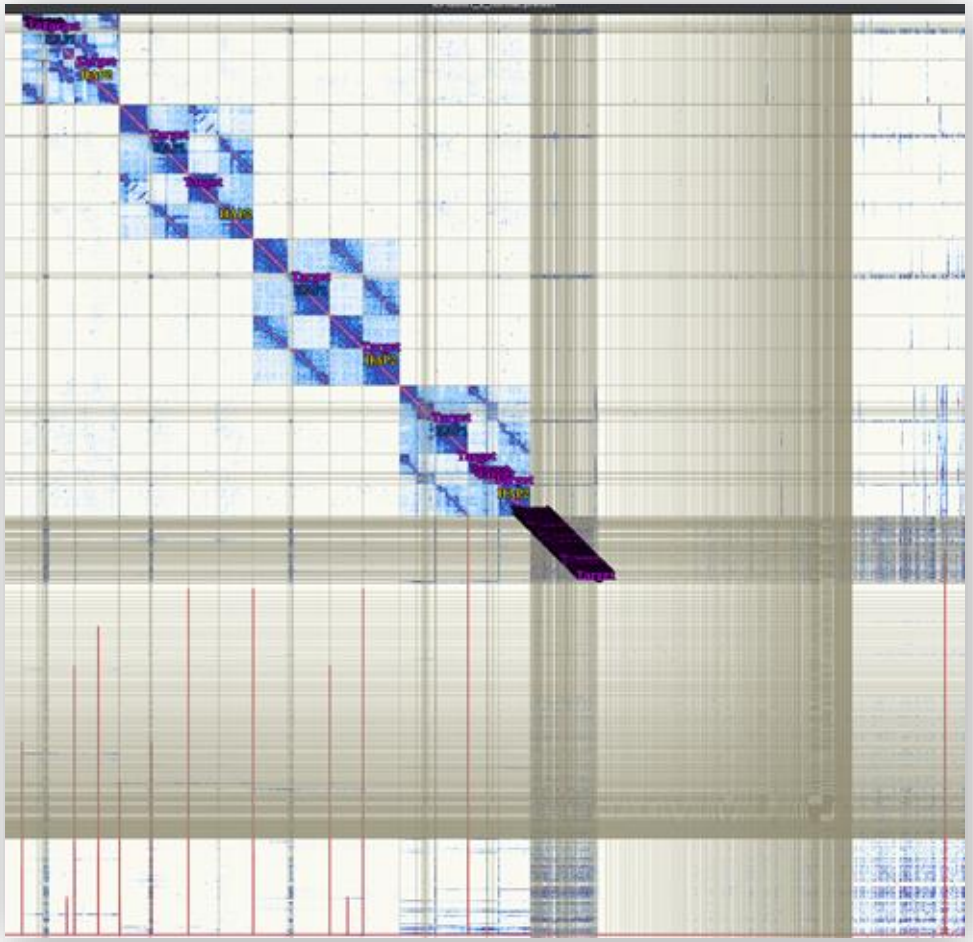


Combined maps
Uneven coverage

What pretext-to-asm does



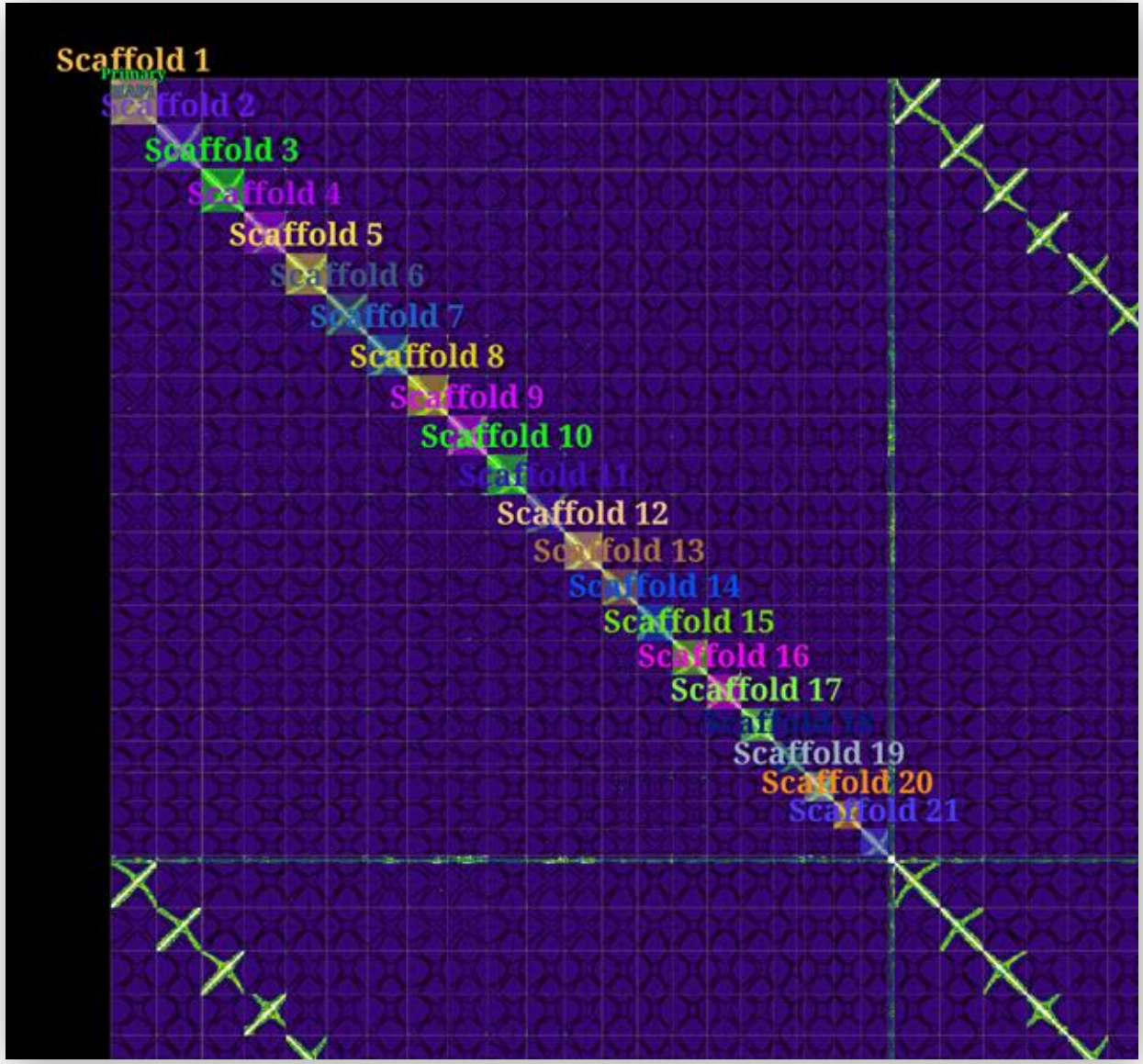
‘Unloc’ tag



‘Target’ tag

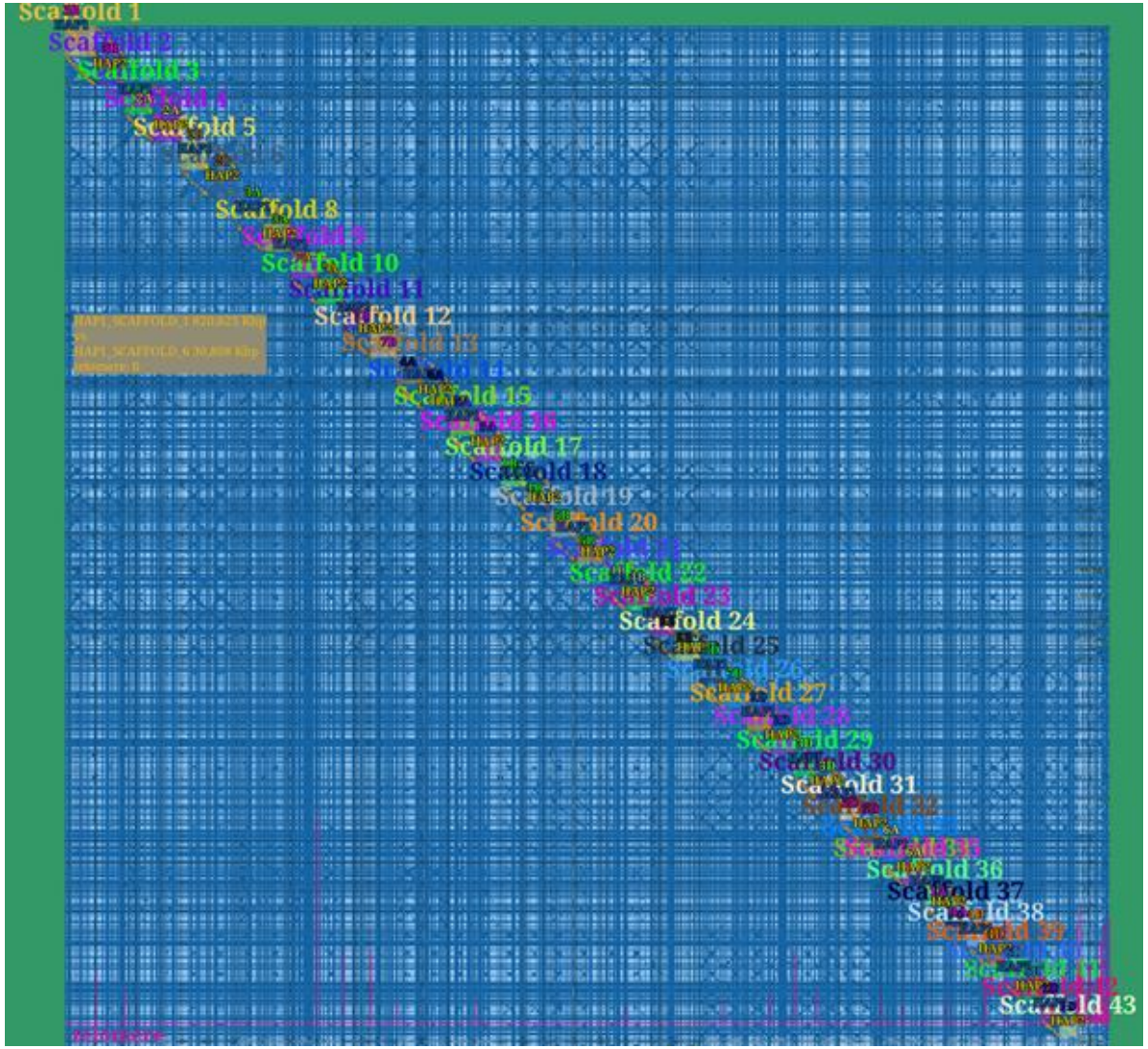
What pretext-to-asm does

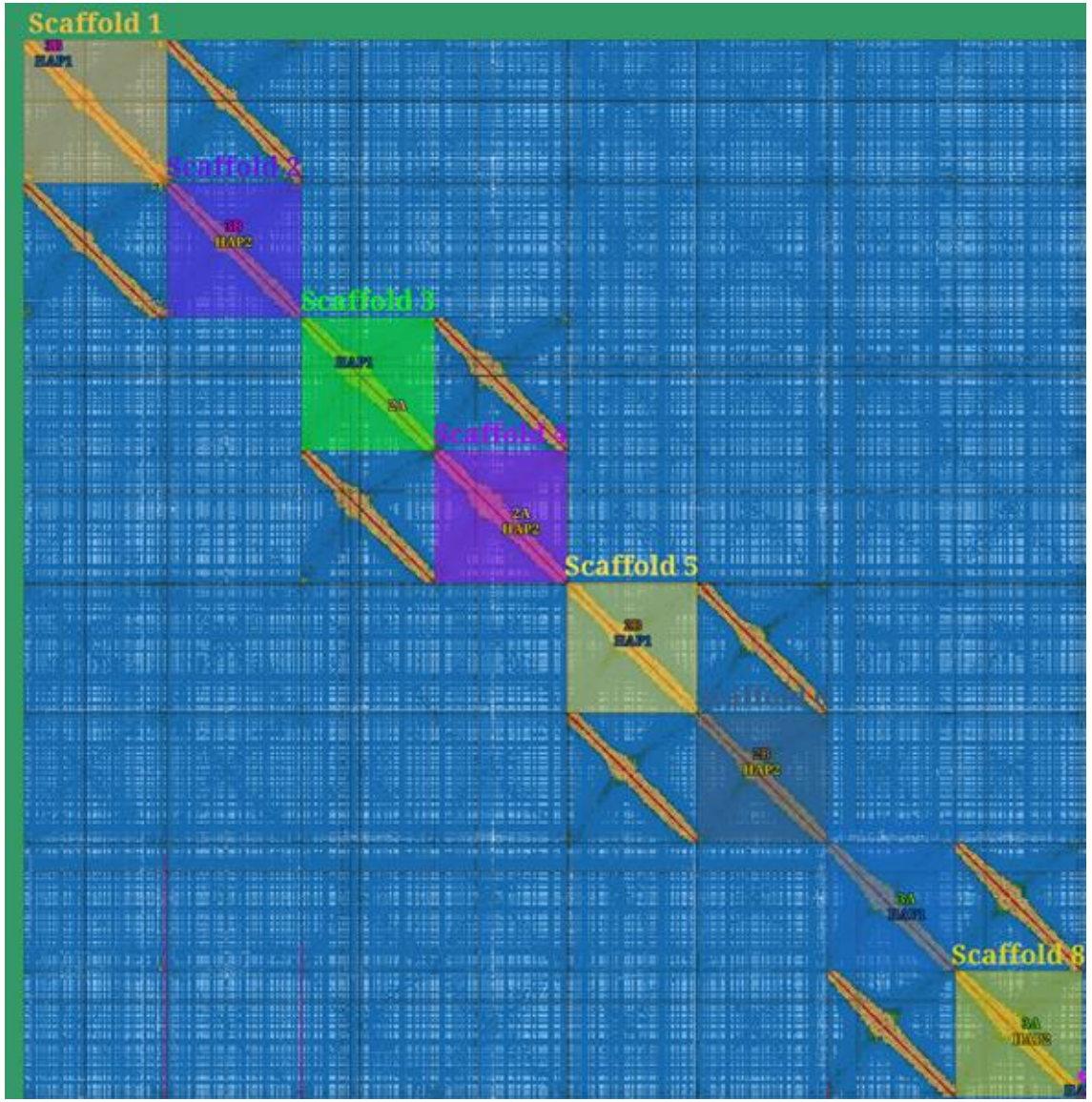
‘Primary’ tag



What pretext-to-asm does

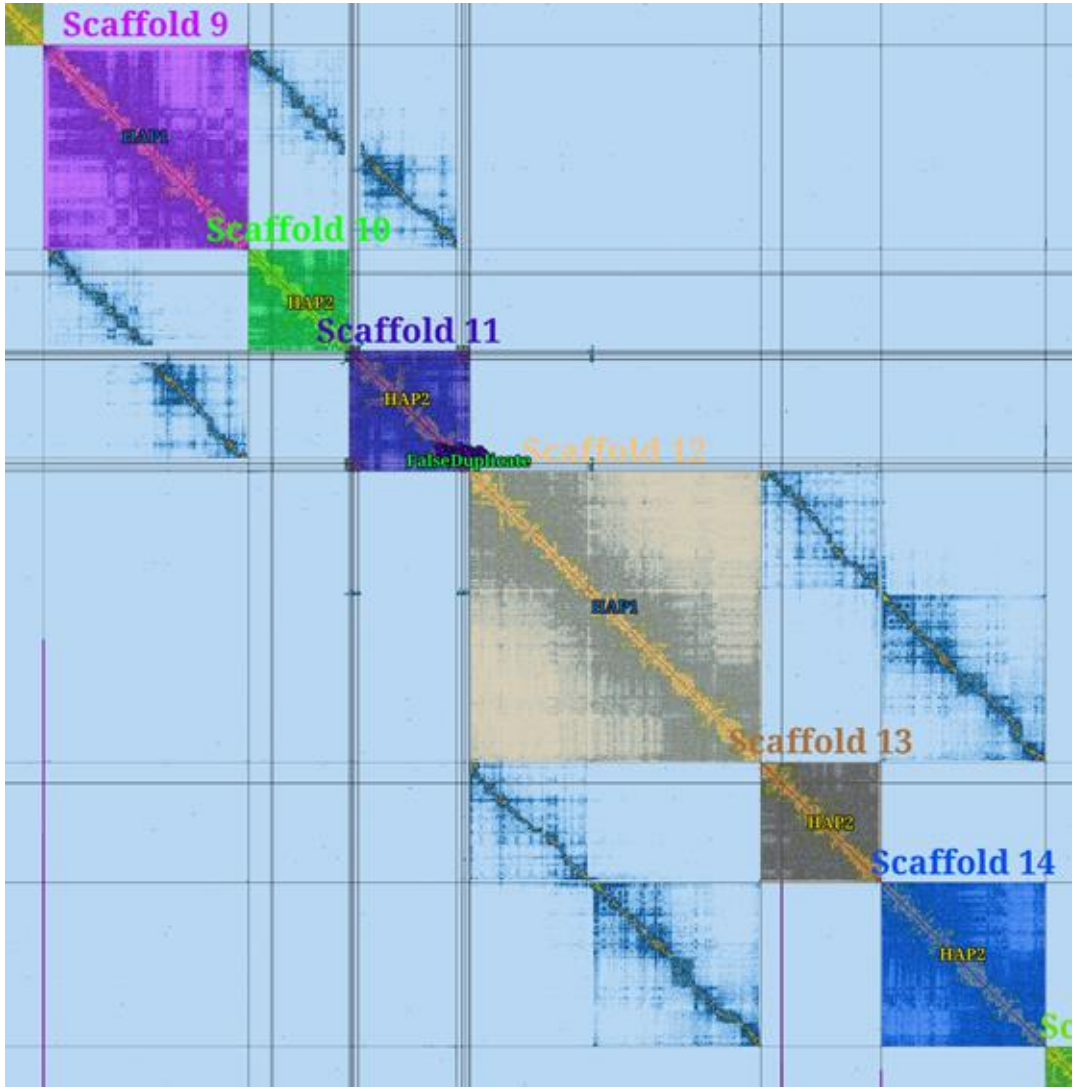
Renaming after a reference



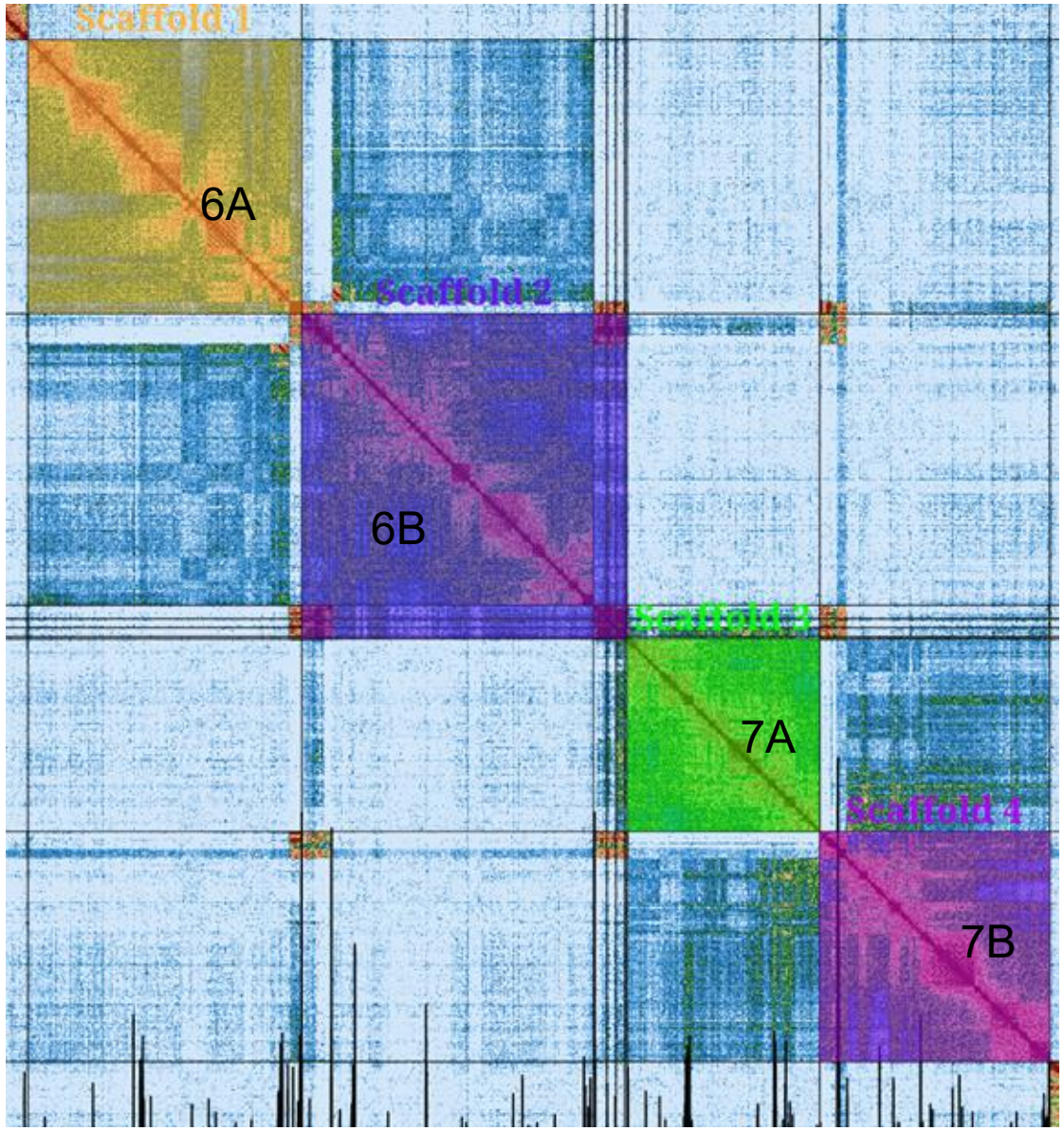


What pretext-to-asm does

Dealing with fusions/fissions



Fissioned chroms in HAP2 file



HAP2 is renamed after HAP1

Generating your own HiC maps with tracks

CurationPretext NextFlow Pipeline

https://pipelines.tol.sanger.ac.uk/curationpretext

```
nextflow run sanger-tol/curationpretext \
  --input { input.fasta } \
  --cram { path/to/hic/cram/ } \
  --reads { path/to/longread/fasta/ } \
  --read_type { default is "hifi" }
  --sample { default is "pretext_rerun" } \
  --teloseq { default is "TTAGGG" } \
  --map_order { default is "unsorted" } \
  --multi_mapping { default is "0" (for no mapping)} \
  --all_output <true/false> \
  --outdir { OUTDIR } \
  -profile <docker/singularity/{institute}>
```

Resources

- https://github.com/sanger-tol/rapid-curation

- Producing the curated fasta file: pretext-to-asm
- https://github.com/sanger-tol/agp-tpf-utils

- Curationpretext: Hi-C maps and feature creation pipeline https://pipelines.tol.sanger.ac.uk/curationpretext

- https://assemblycuration.slack.com
- grit@sanger.ac.uk (GRIT team)

Physalia Manual Genome Curation course – Next November