Session 3: Beginning manual curation
Day 3

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

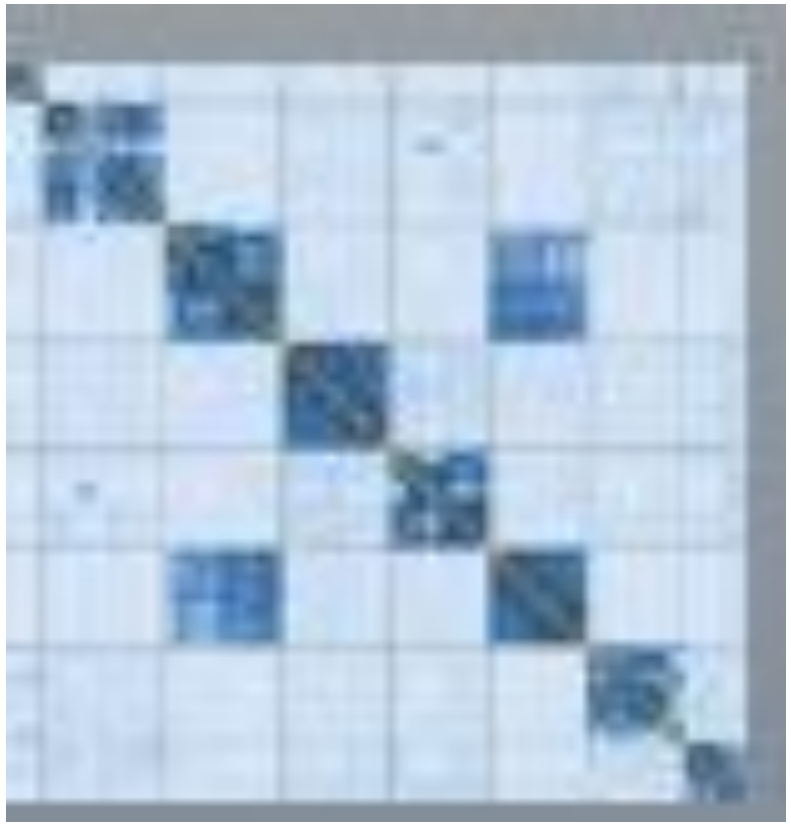
Overview

- **Analysis pipelines**
- How to generate your own PretextView Hi-C maps

- **Some curation tricks**

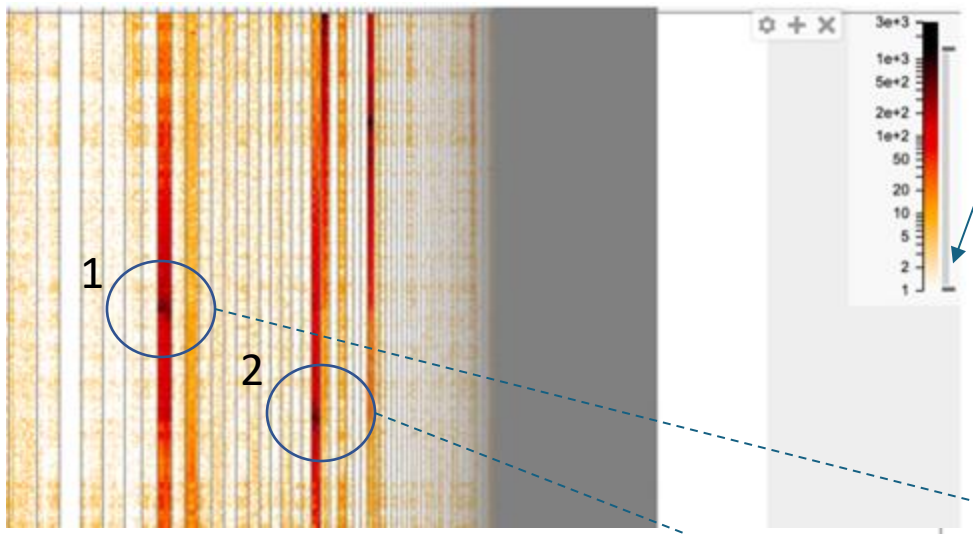
- **Curation tools**
- Rapid curation workflow

  How to produce a curated fasta file



Shrapnel

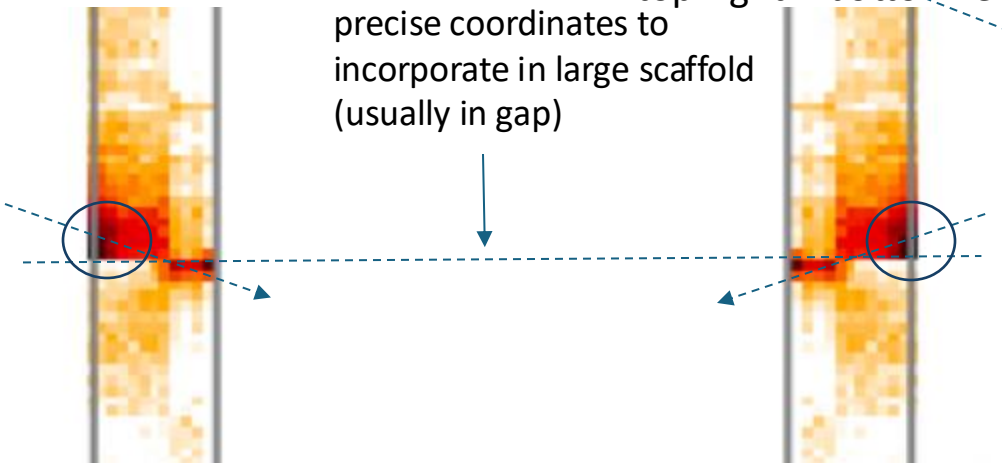
Incorporation of smaller scaffolds into larger ones
Usually in gaps



top left-> bottom right

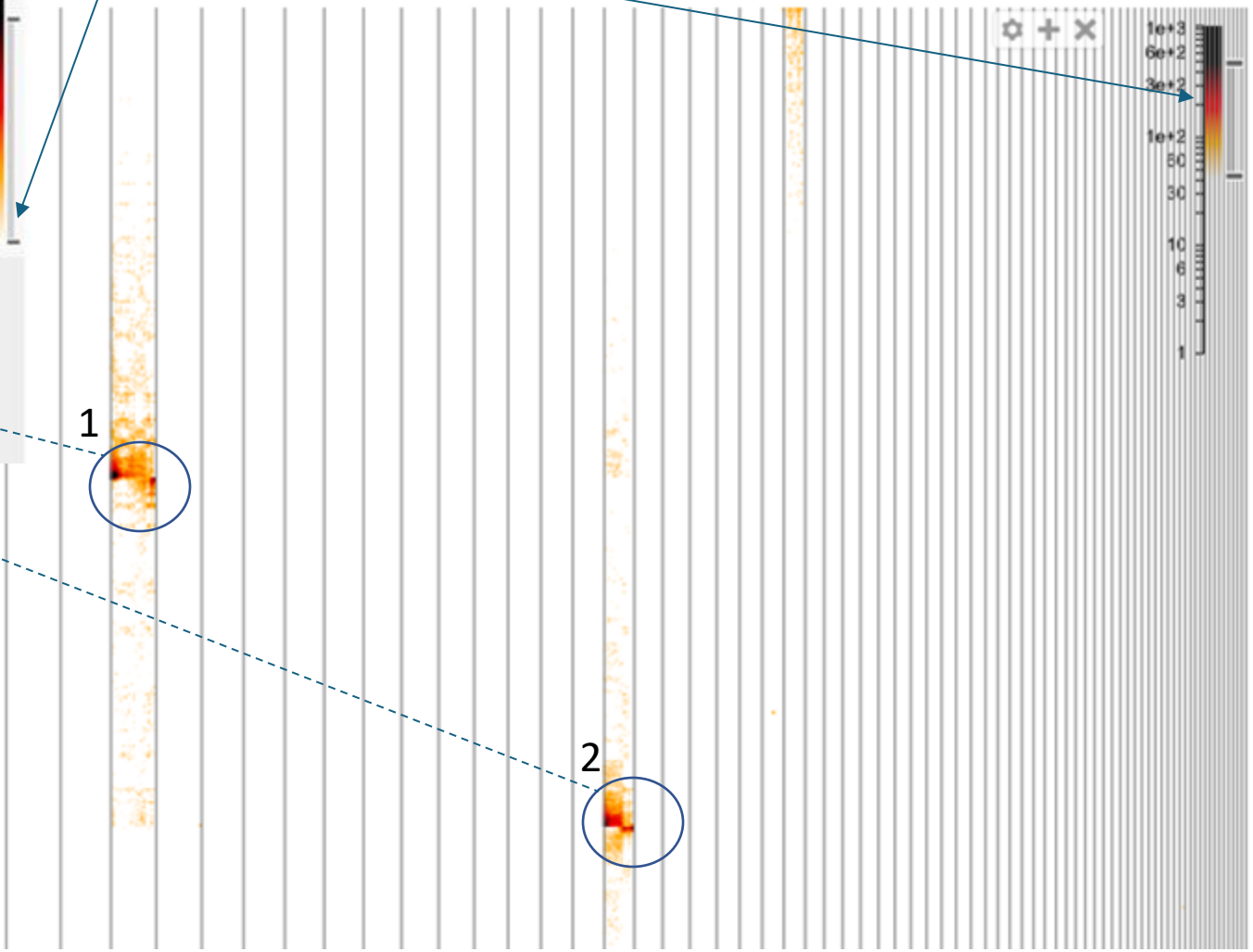
top right > bottom left

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



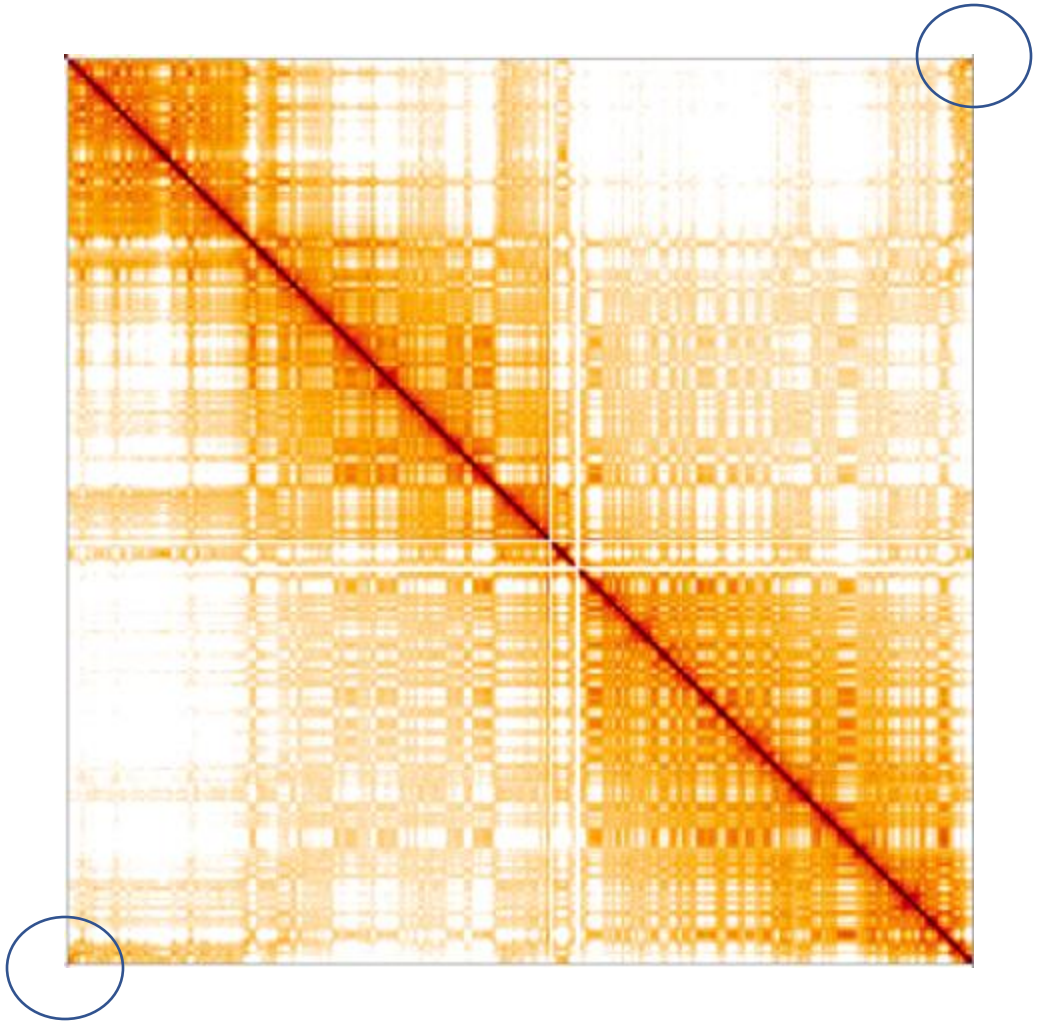
forward orientation

reverse orientation



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

Linking between chromosome ends



mZalCal1 – scaffold4

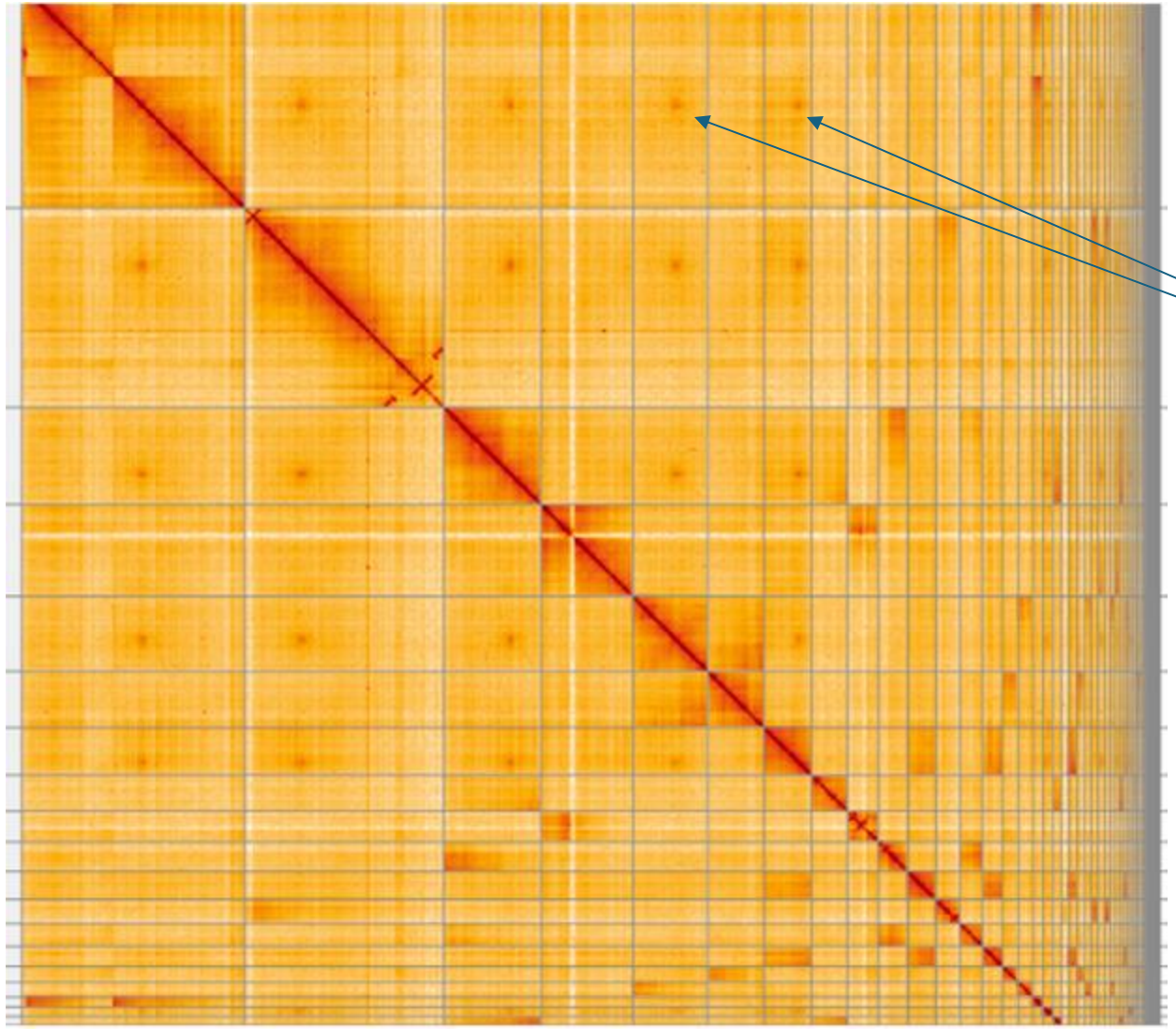
We often see affinity (ie off-diagonal signal at a level higher than we'd expect) between chromosome ends on the same chromosome. All evidence suggests that when we see this the chromosome is assembled correctly.

Telomeres are lighting each other up

↓

Usually chromosome is well assembled

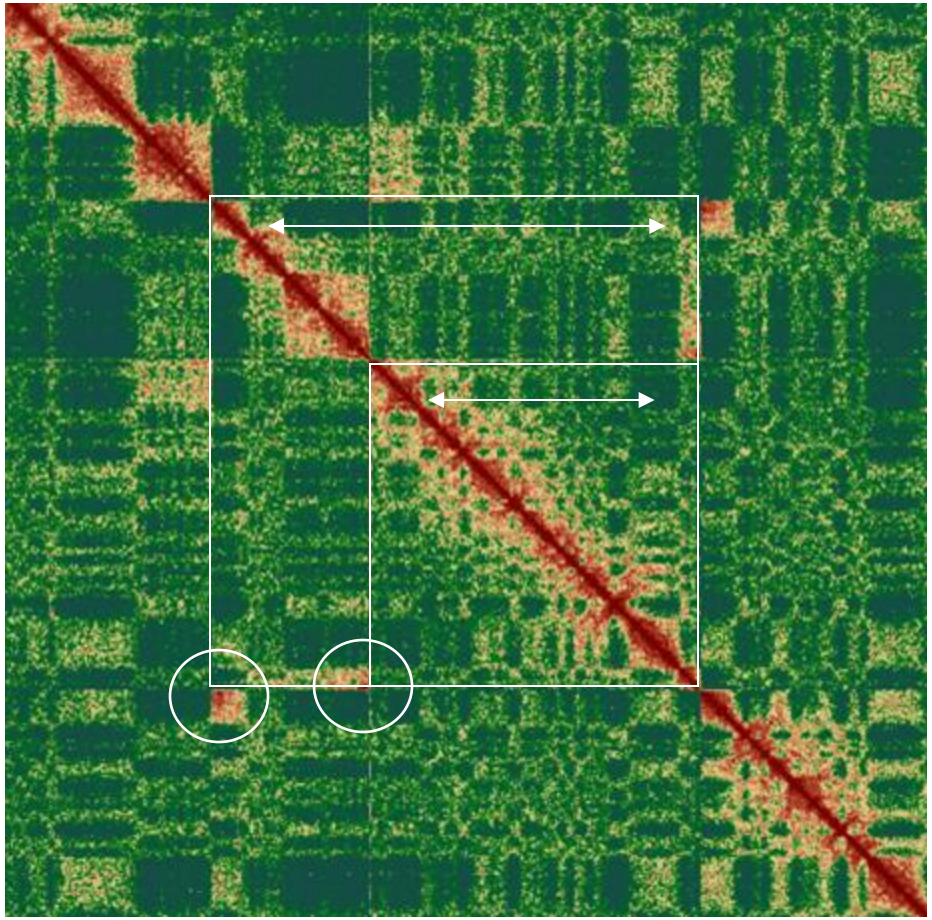
Centromeres also light each other up along the map



iHerIll2

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

Colour schemes



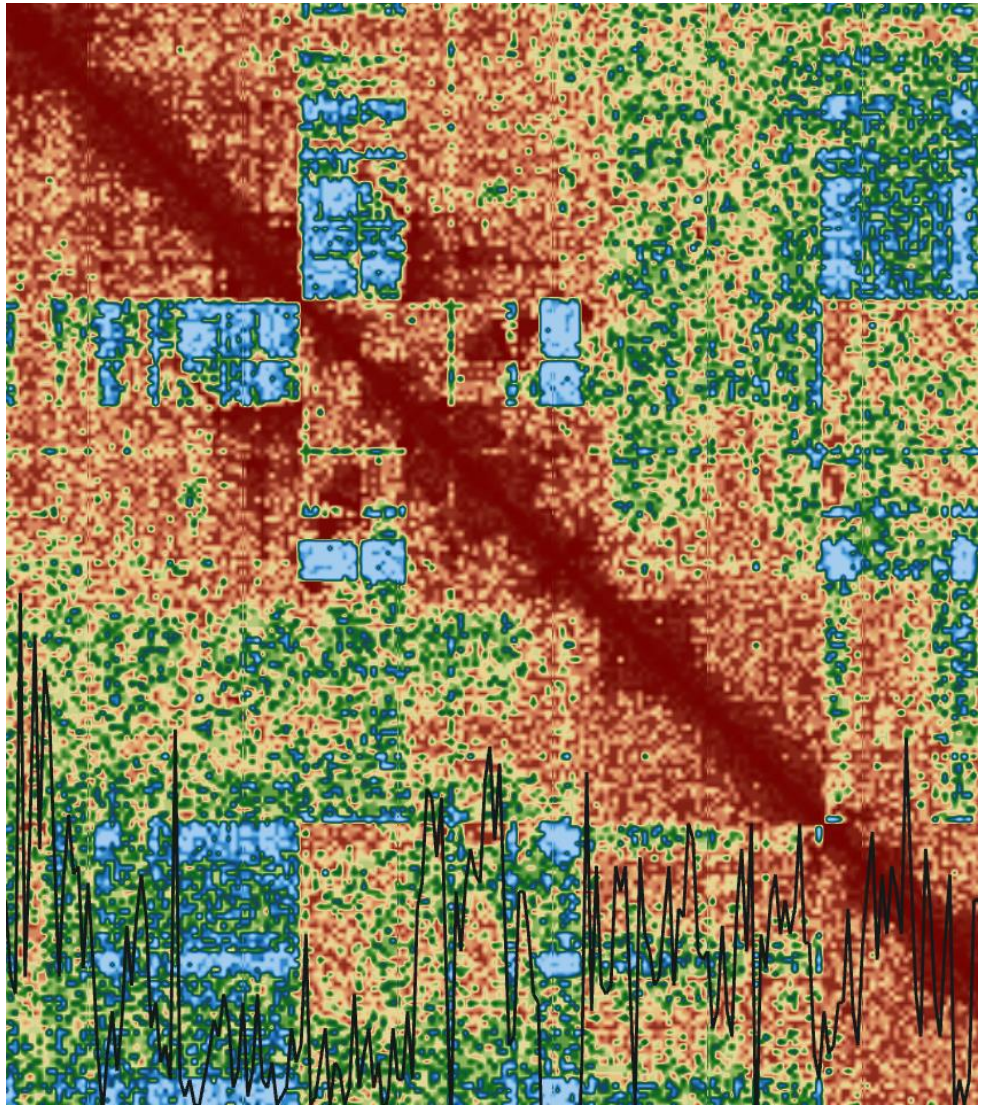
bPteGut1 superscaffold6

Choice of colour schemes is important

2 misassemblies are strongly highlighted in Pretext

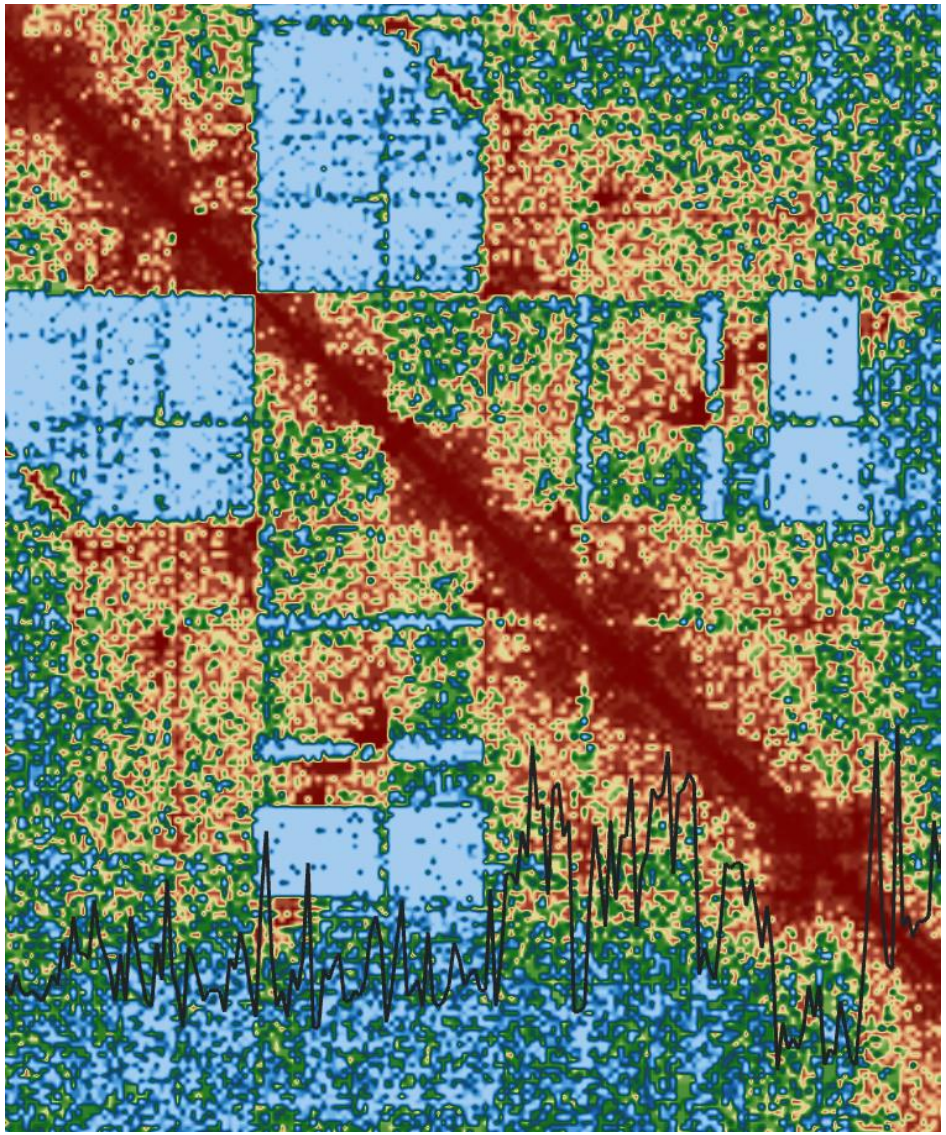
3-way colour scheme called “three wave blue-green-yellow”.

Pretext normal vs. high resolution maps – resolution issues in Pretext



Same zoom level

Works well for haplotigs



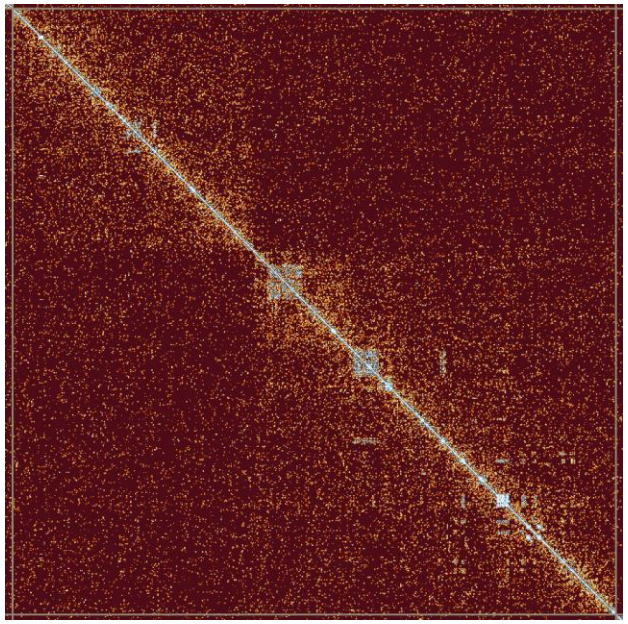
Normal resolution

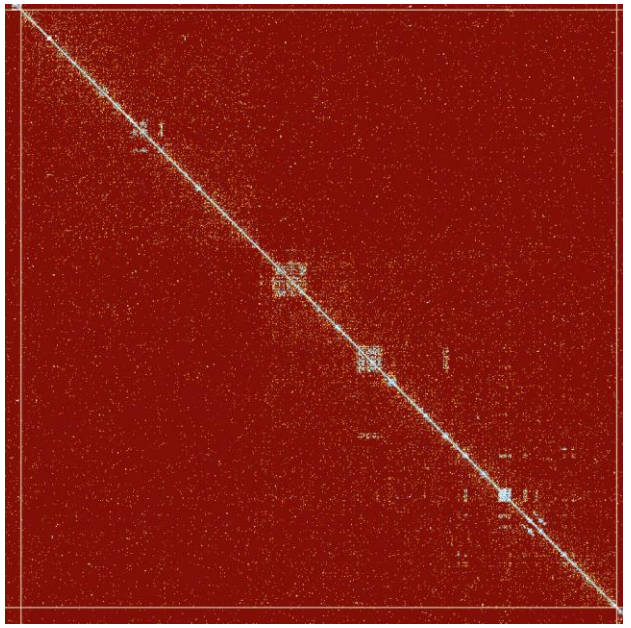
More details when you zoom-in

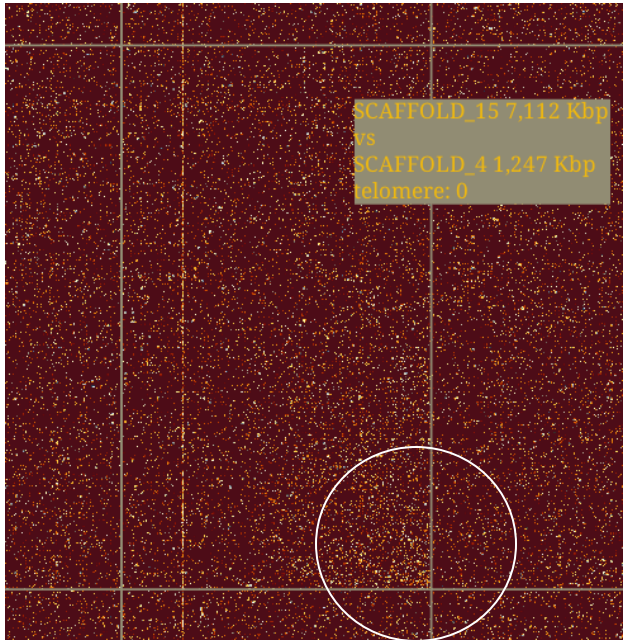
High resolution

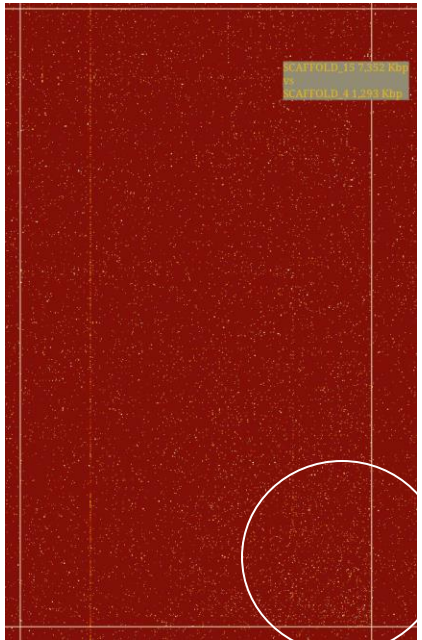
Pretext normal vs. high resolution maps – resolution issues in Pretext

Normal resolution





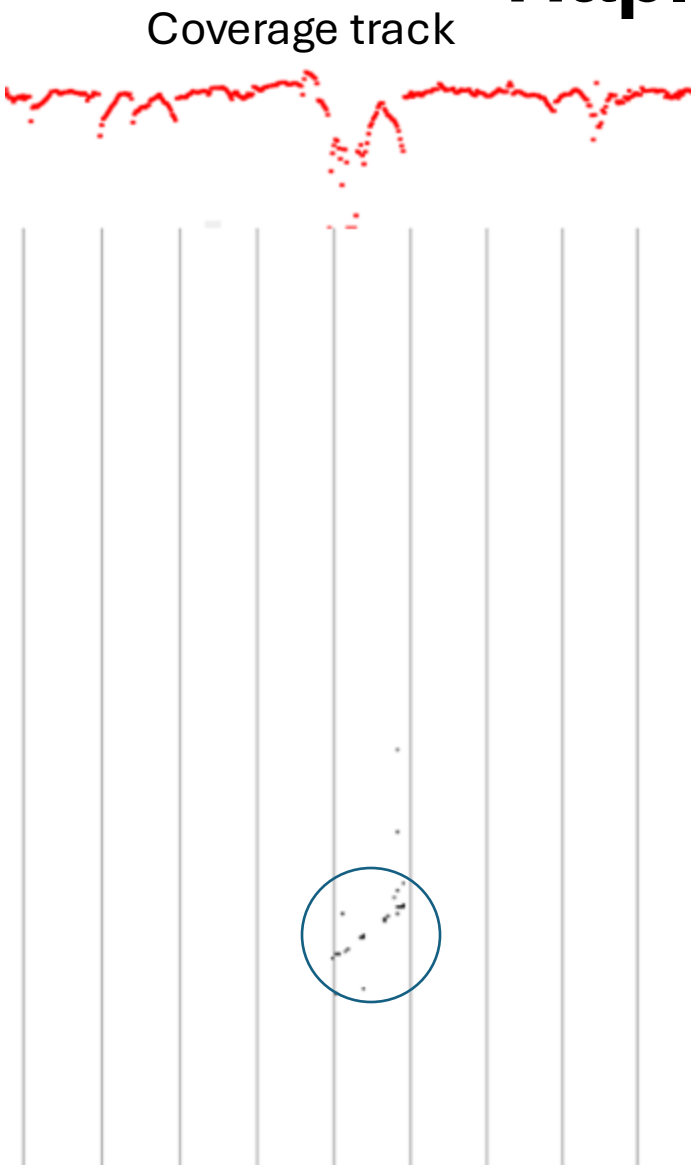


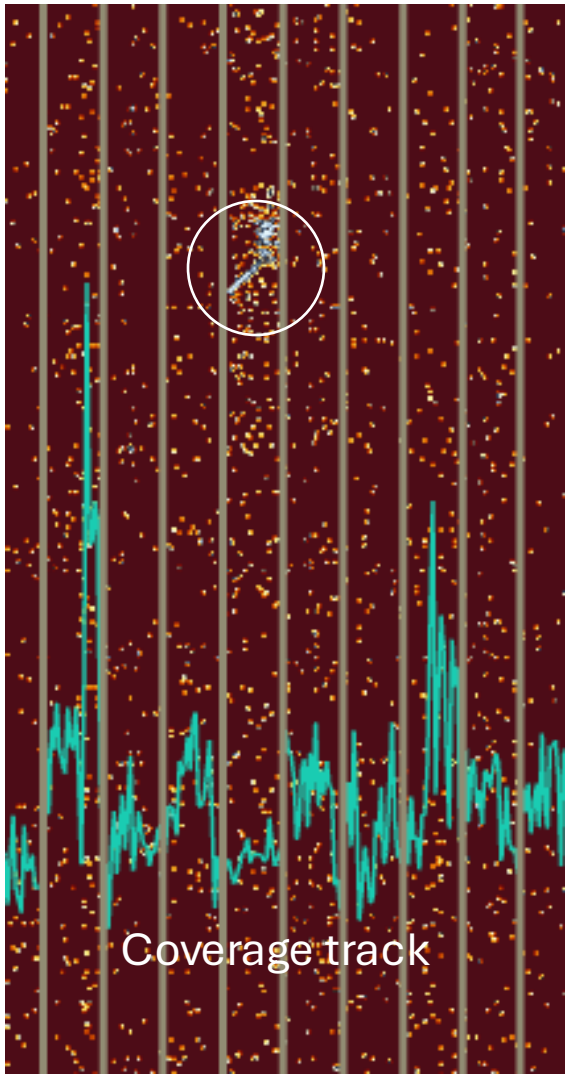


High resolution

Bad for joins
Poor HiC

Haplotypic shrapnel contig

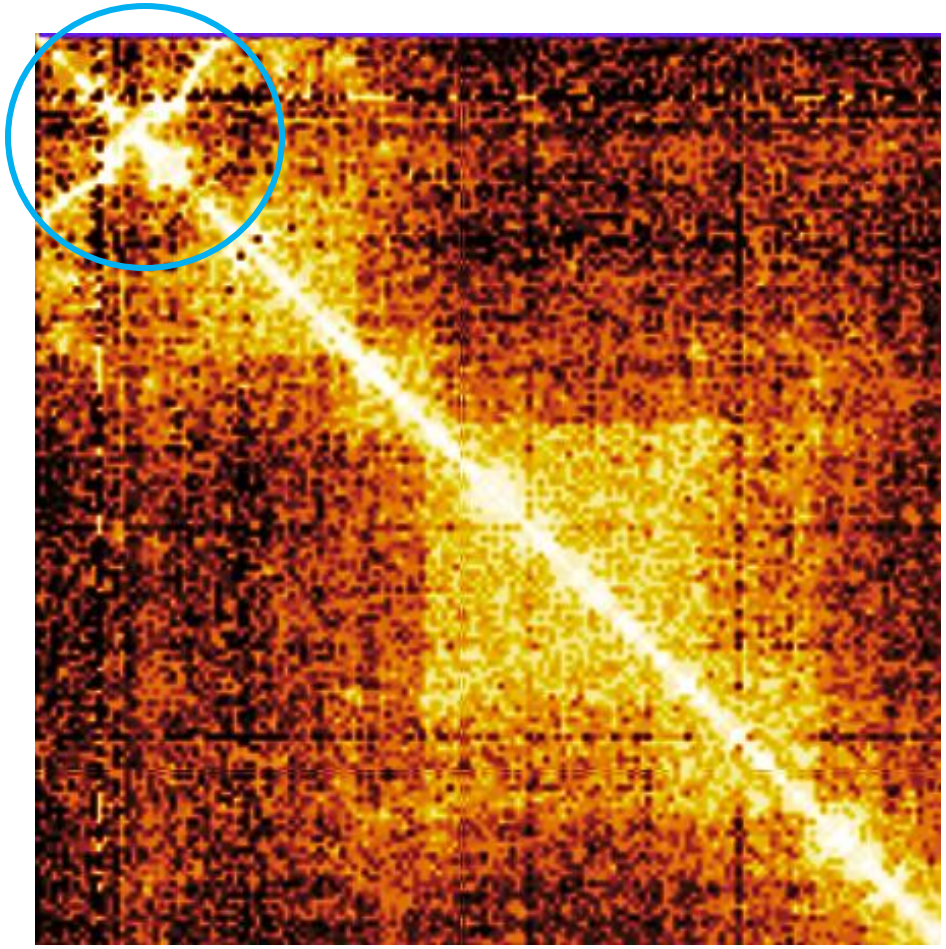




Coverage plot show the contig has half depth and the sporadic contacts are typical of a haplotypic contig. From this plot, you can see that the haplotype is entirely contained in the chromosome in the reverse orientation.

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

Inverted haplotypes

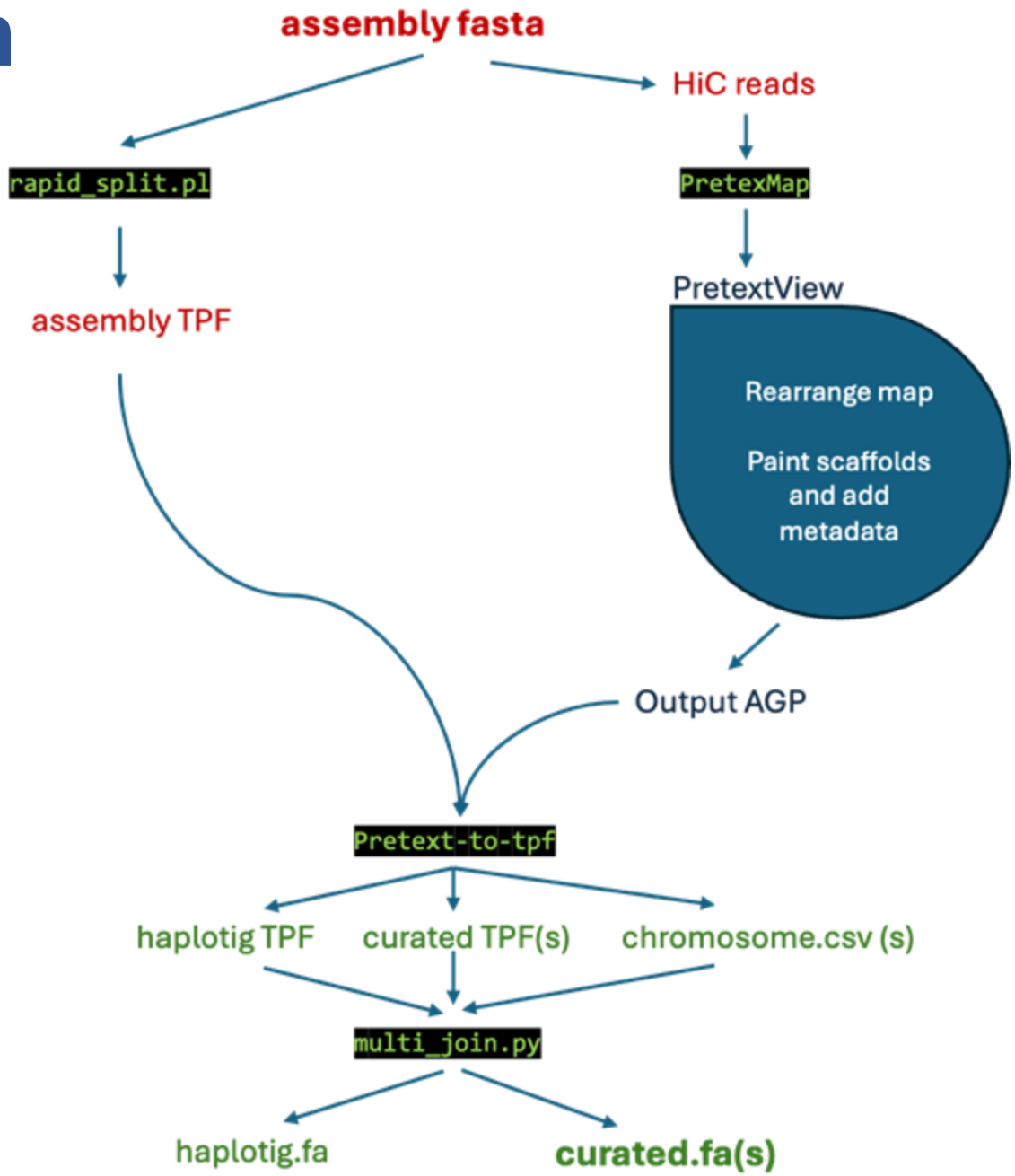


iAphHyp

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.

How to produce your curated fasta file?

Rapid Curation (distributed)



*TPF = Tile Path File

Flat text file that dictates the order and orientation of component sequences making up an assembly. Records both sequence coordinates and gap locations.

*AGP = A Golden Path

Defines the order and orientation of the genome blocks (chromosomes)

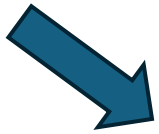
File processing

After curation and:

Adding all relevant metadata tags

Painting chromosomes

AGP and savestate generation



Curated fasta file

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
```

File processing

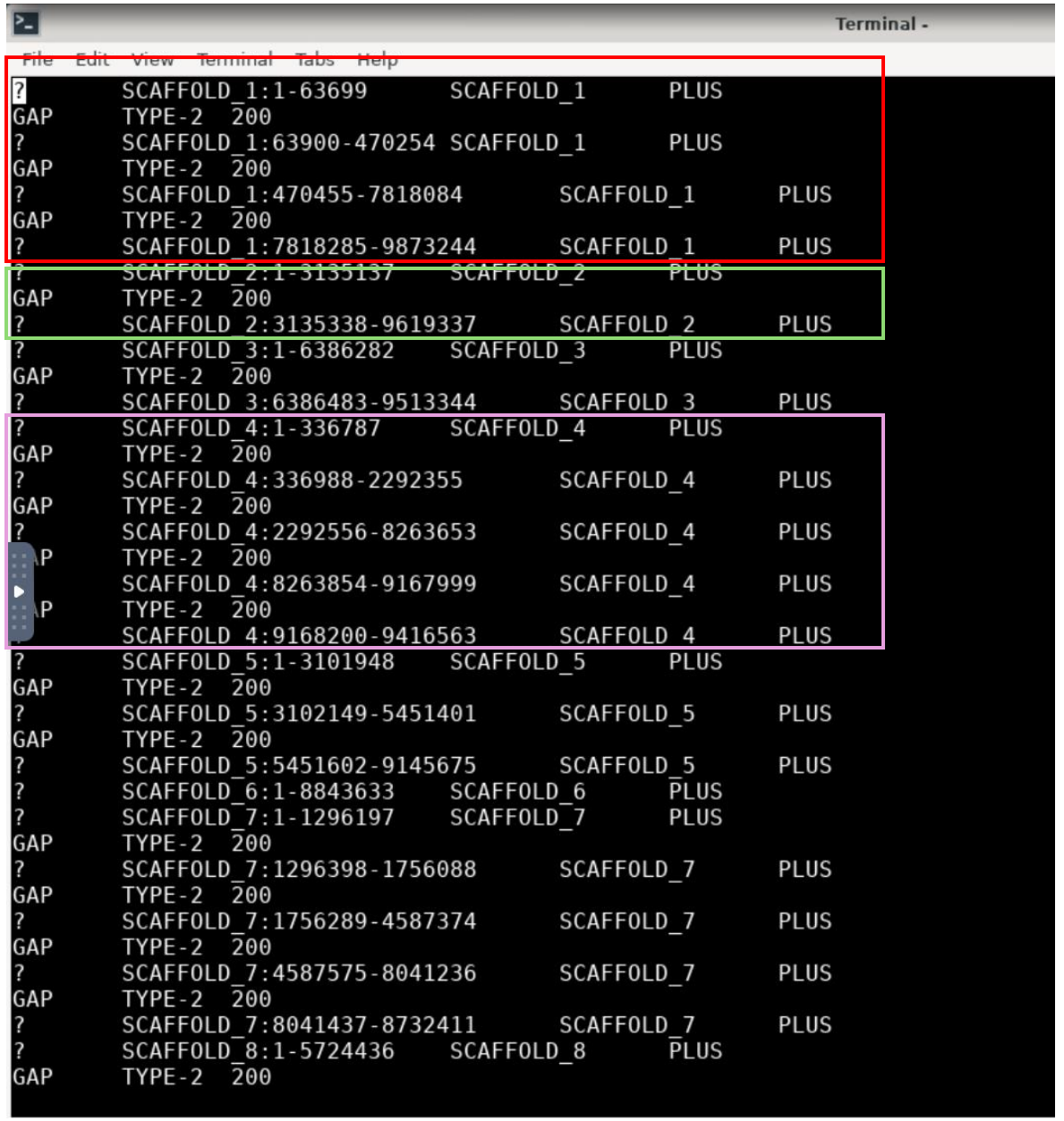
Original decontaminated fasta file

`rapid_split.pl -fa <your_fasta>`

original.tpf

scaffolds.tpf (text file)

Order and orientation of component sequences in an assembly.
Records both **sequence coordinates and gap locations**.



`pretext-to-tpf` script (alias ptt)

Single haplotype curation – Primary assembly AGP

or

Dual haplotype curation – HAP1 and HAP 2 assemblies AGP file

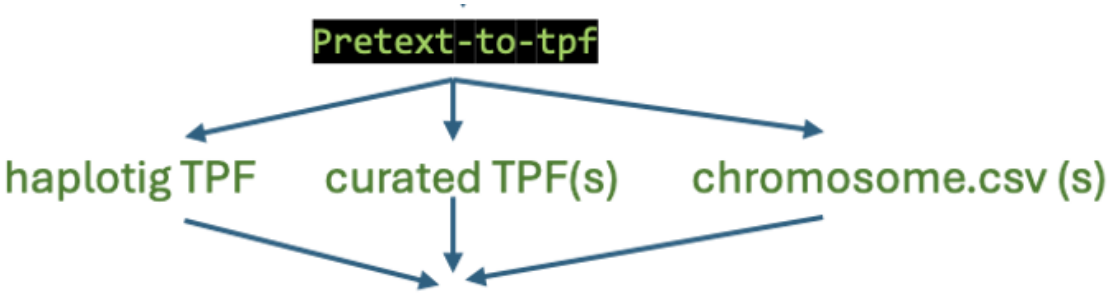
Required:

```
ptt -a original.fa.tpf -p <your_species>.agp_1 -o <output_name>.tpf -w -f
```

-w: overwrite

-f: force to run and overwrite

Outputs:



For dual curation only:

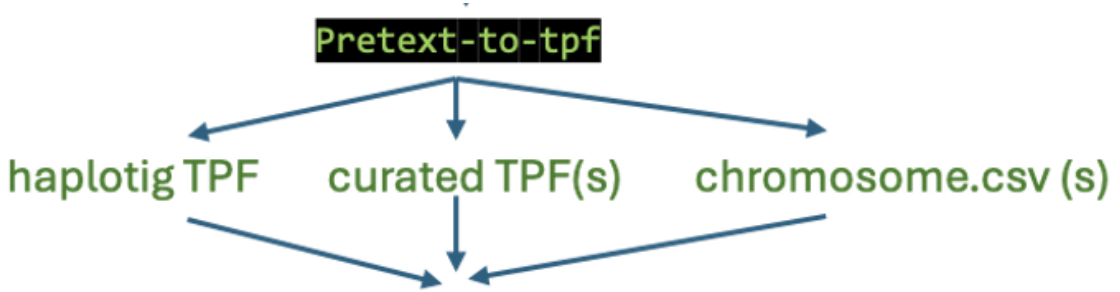
Curated_HAP1.tpf	Chrs_HAP1.csv
Curated_HAP2.tpf	Chrs_HAP2.csv

Curated.log

↓

Curation stats

`pretext-to-tpf` script (alias ppt)



Terminal -

File Edit View Terminal Tabs Help

```
?       SCAFFOLD_23:1-338256    H_1     PLUS
?       SCAFFOLD_4:9168200-9416563      H_2     MINUS
?       SCAFFOLD_3:1215246-1344812      H_3     PLUS
?       SCAFFOLD_9:4677803-4789498      H_4     PLUS
?       SCAFFOLD_3:1103551-1215245      H_5     PLUS
?       SCAFFOLD_1:540606-647833        H_6     PLUS
?       SCAFFOLD_8:5651787-5724436      H_7     PLUS
?       SCAFFOLD_8:906967-973983        H_8     PLUS
?       SCAFFOLD_1:1-63699      H_9     PLUS
?       SCAFFOLD_1:5517753-5580301      H_10    PLUS
?       SCAFFOLD_5:4793967-4856515      H_11    PLUS
?       SCAFFOLD_17:1561800-1617349     H_12    PLUS
?       SCAFFOLD_2:7501463-7555075      H_13    PLUS
?       SCAFFOLD_321:1-47667    H_14    MINUS
?       SCAFFOLD_13:6187925-6232602     H_15    PLUS
?       SCAFFOLD_18:2685158-2725367     H_16    PLUS
?       SCAFFOLD_14:1818401-1849675     H_17    PLUS
?       SCAFFOLD_2088:1-11161   H_18    PLUS
```

Terminal -

File Edit View Terminal Tabs Help

```
?       SCAFFOLD_1352:1-20225   RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1140:1-22893   RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:470455-540605        RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:647834-4043373       RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:366361-470254        RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:63900-366360 RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:4043374-5517752      RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:5580302-7818084      RL_1    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_1:7818285-9873244      RL_1    PLUS
?       SCAFFOLD_475:1-39409    RL_1_unloc_1    MINUS
?       SCAFFOLD_2:1-3135137    RL_2    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_2:3135338-5933258      RL_2    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_2:5933259-6062825      RL_2    MINUS
GAP     TYPE-2  200
?       SCAFFOLD_2:6062826-7501462      RL_2    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_2:7555076-9619337      RL_2    PLUS
?       SCAFFOLD_30:1-201000    RL_2_unloc_1    PLUS
?       SCAFFOLD_3:1-1103550    RL_3    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_3:1344813-6386282      RL_3    PLUS
GAP     TYPE-2  200
?       SCAFFOLD_27:1-221802    RL_3    MINUS
GAP     TYPE-2  200
?       SCAFFOLD_3:6386483-9513344      RL_3    PLUS
?       SCAFFOLD_4:1-336787     RL_4    PLUS
GAP     TYPE-2  200
```

File Edit View Terminal Tabs Help

```
RL_1,RL_1_unloc_1
RL_2,RL_2_unloc_1
RL_3
RL_4
RL_5,RL_5_unloc_1
RL_6
RL_7
RL_8
RL_9
RL_10
RL_11
RL_12
RL_13
RL_14
RL_15
RL_16
RL_17
RL_18
```

`pretext-to-tpf` script (alias ptt)

HAP1

Curated.log

HAP2

```
 GNU nano 6.2                                                curated.log
       530,757,975  bp sequence (minus gaps)
 Autosomes:
   n = 19
        36,678,067  RL_16
               ...  ...
        16,495,084  RL_37
       452,563,291  bp total
 Named:
   n = 2
        34,548,320  W
        41,470,319  Z
        76,018,639  bp total
 Unplaced:
   n = 56
           143,703  HAP1_SCAFFOLD_72
               ...  ...
             6,518  HAP1_SCAFFOLD_188
         2,176,045  bp total

curated_HAP2
       452,233,478  bp sequence (minus gaps)
 Autosomes:
   n = 21
        35,504,589  RL_2
               ...  ...
        15,148,779  RL_12
       450,698,550  bp total
 Unplaced:
   n = 32
           133,814  HAP2_SCAFFOLD_81
               ...  ...
             1,000  HAP2_SCAFFOLD_155
         1,534,928  bp total

curated_Haplotigs
           250,681  bp sequence (minus gaps)
   n = 1
           250,681  H_1

Curation made 6 cuts in contigs, 7 breaks at gaps and 56 joins
```

Curation stats

multi_join.py script

Single haplotype curation

```
multi_join.py -t <curated>.tpf  \
-c chrs.csv \
-d <curated>_Haplotigs.tpf \
-f original.fa -o <TolID>
```

Only if you detected haplotigs in your assembly

Output:

```
<ToLID>.1.primary.curated.fa
<ToLID>.1.additional_haplotigs.curated.fa
<ToLID>.1.chromosome.list.csv
<ToLID>.1.inter.csv
```

multi_join.py script

Dual haplotypes curation

```
multi_join.py -t <curated>_HAP1.tpf -t2 <curated>_HAP2.tpf \
-c chrs_HAP1.csv -c2 chrs_HAP2.csv \
-f original.fa -o <ToLID>
```

Output:

<ToLID>.hap1.1.primary.curated.fa
<ToLID>.hap2.1.primary.curated.fa
<ToLID>.hap1.1.primary.chromosome.list.csv
<ToLID>.hap2.1.primary.chromosome.list.csv
<ToLID>.hap1.1.all_haplotigs.curated.fa
<ToLID>.hap2.1.all_haplotigs.curated.fa
<ToLID>.1.additional_haplotigs.curated.fa
<ToLID>.hap1.1.inter.csv
<ToLID>.hap2.1.inter.csv

Remap curated fasta file to HiC reads

inter.csv file

Scaffolds in original map and chromosomes in curated map match

```
  GNU nano 6.2                                  jaMonPala3.1.inter.csv *

RL_1,1,yes
RL_2,2,yes
RL_3,3,yes
RL_4,4,yes
RL_5,5,yes
RL_7,6,yes
RL_8,7,yes
RL_6,8,yes
RL_9,9,yes
RL_10,10,yes
RL_14,11,yes
RL_11,12,yes
RL_13,13,yes
RL_12,14,yes
RL_4_unloc_1,4,no
RL_12_unloc_1,14,no
RL_14_unloc_1,11,no
```

Chromosomes are size sorted in the curated map