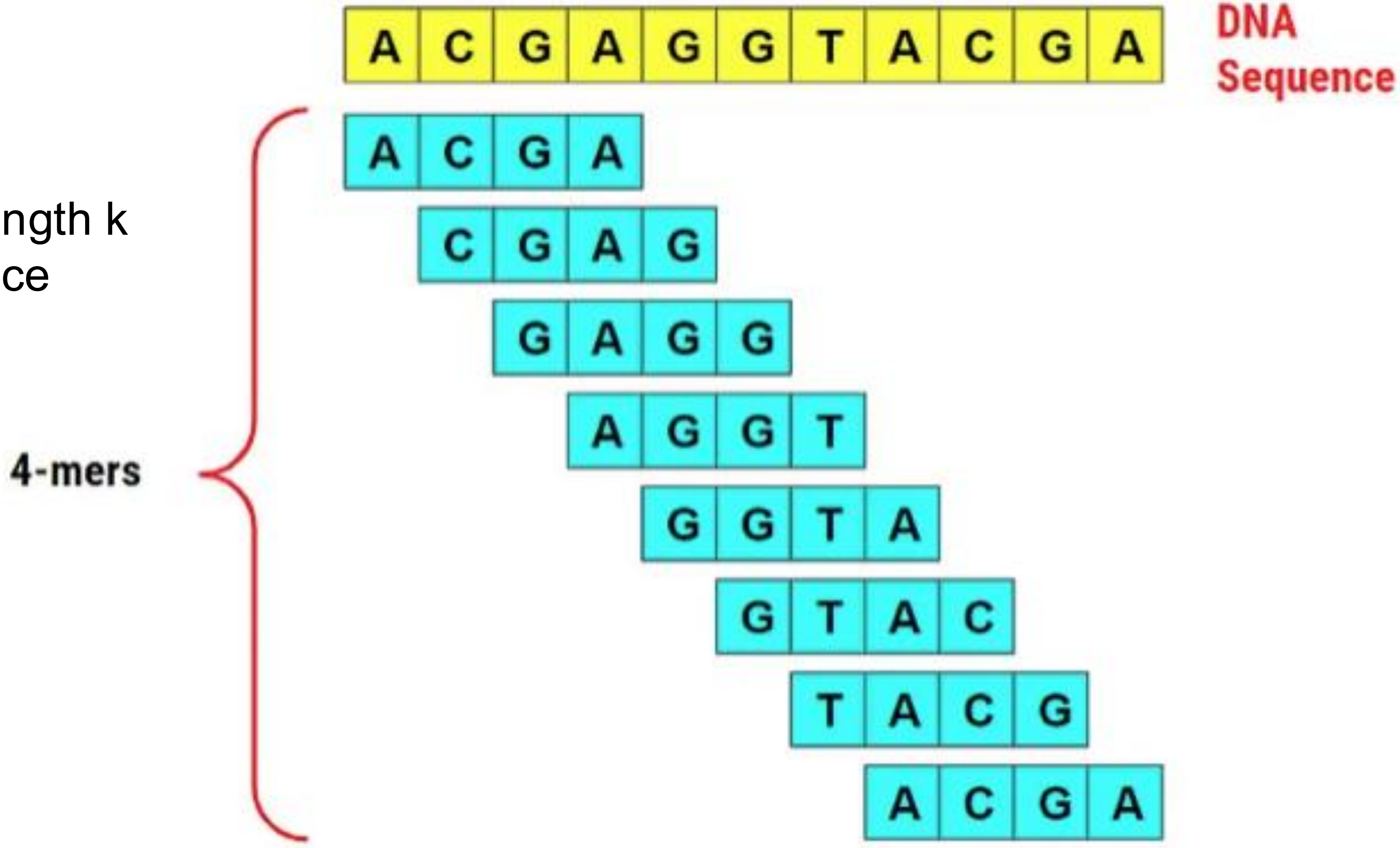
Session 2.1: What to infer from assembly quality metrics?

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

What are kmers?

Substrings of length k of a sequence

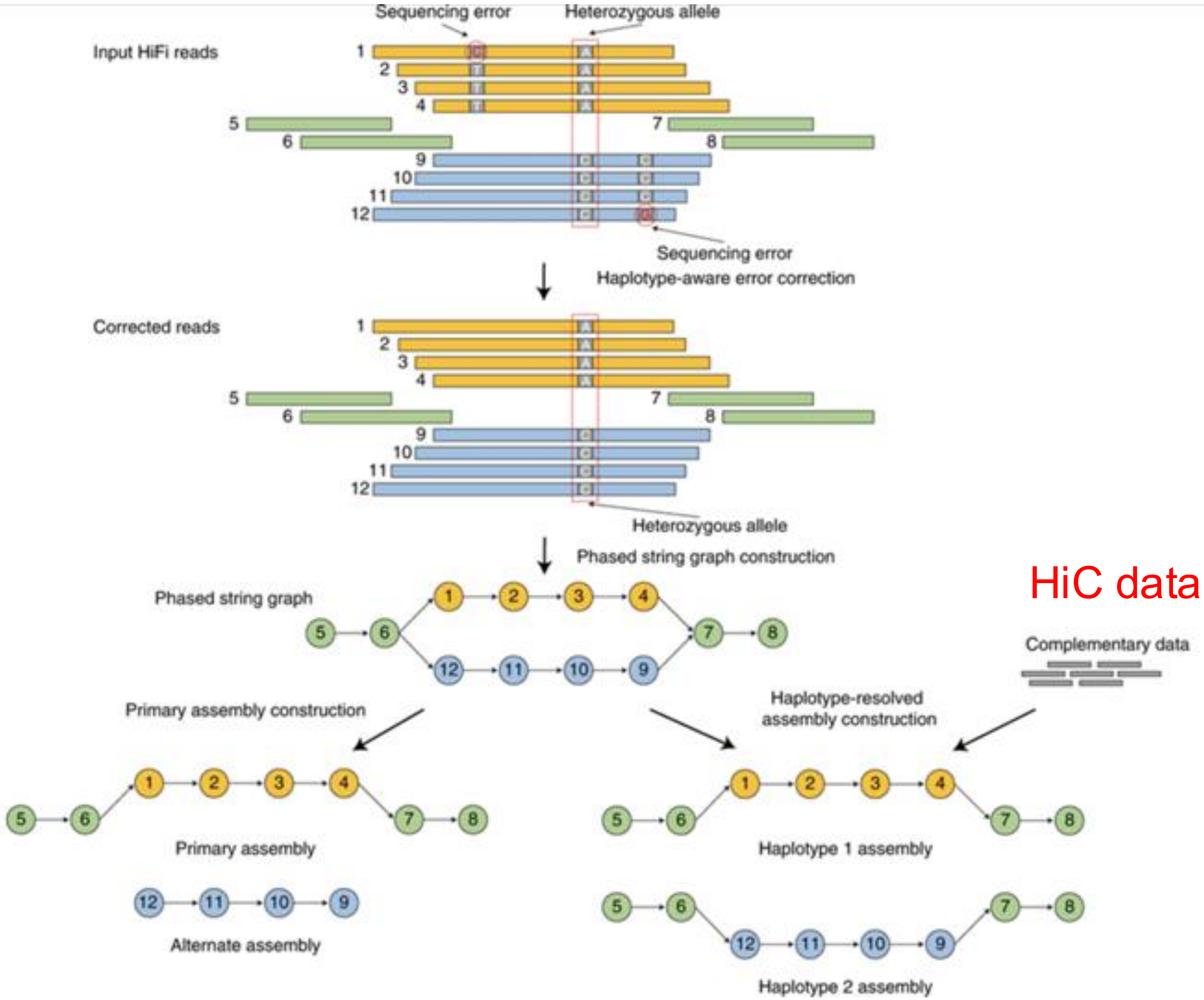


4-mers in the sequence ACGAGGTACGA

https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29

Phased assembly

Heterozygous and repetitive regions



HiC data

Cheng et al. (2021) - https://doi.org/10.1038/s41592-020-01056-5

Primary:
All homozygous regions + 1 copy of each heterozygous region

Alternative:
All that is duplicated in the primary

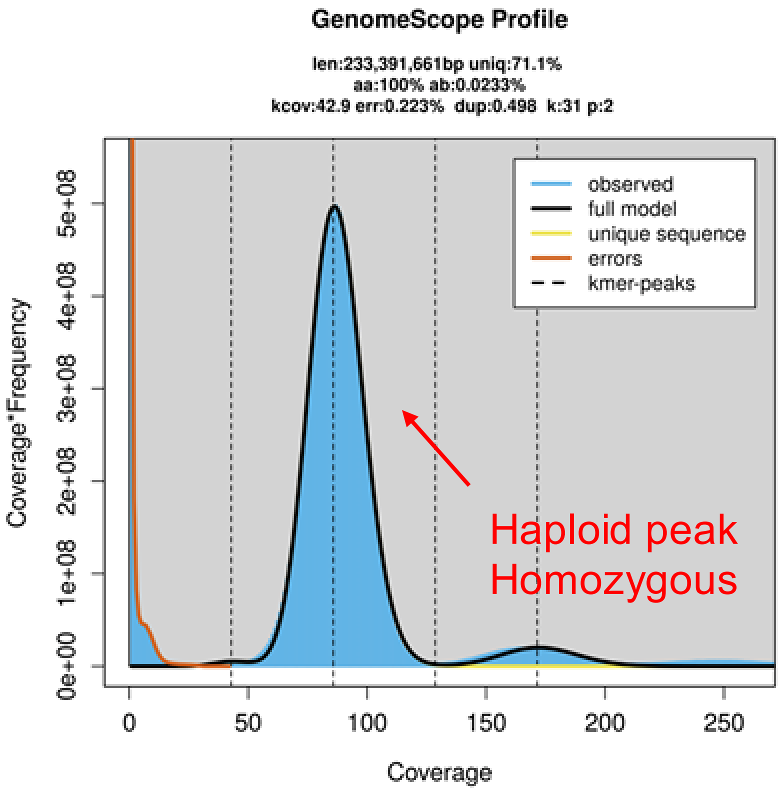
Chromosomes are phased

Haplotype 1
+
Haplotype 2

K-mer distribution

Diploids

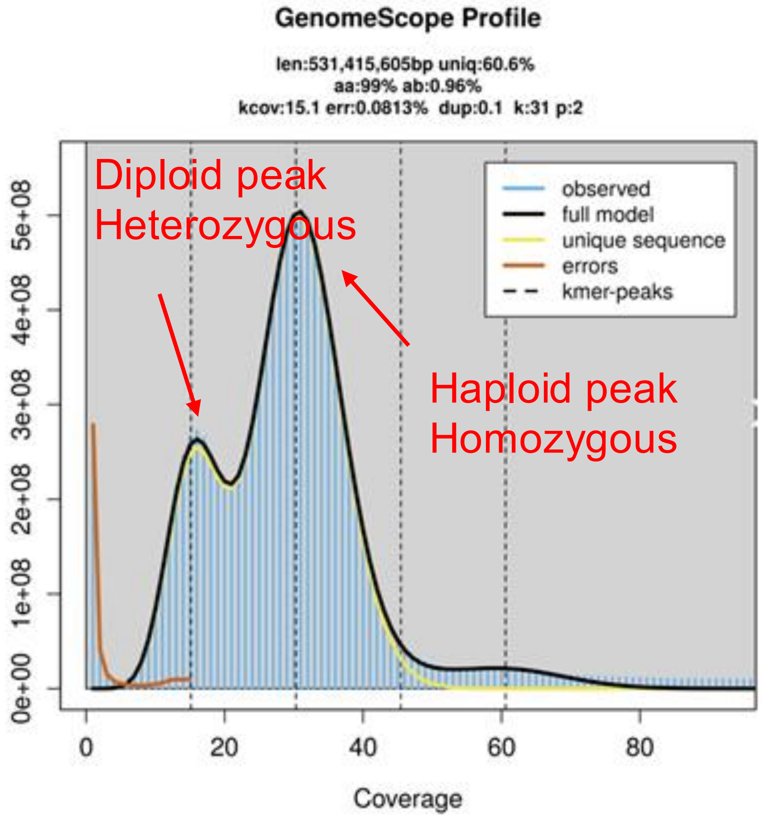
ddCarHirs1



Haploid peak
Homozygous

Super low heterozygosity (0.02%)

ilEreMont1

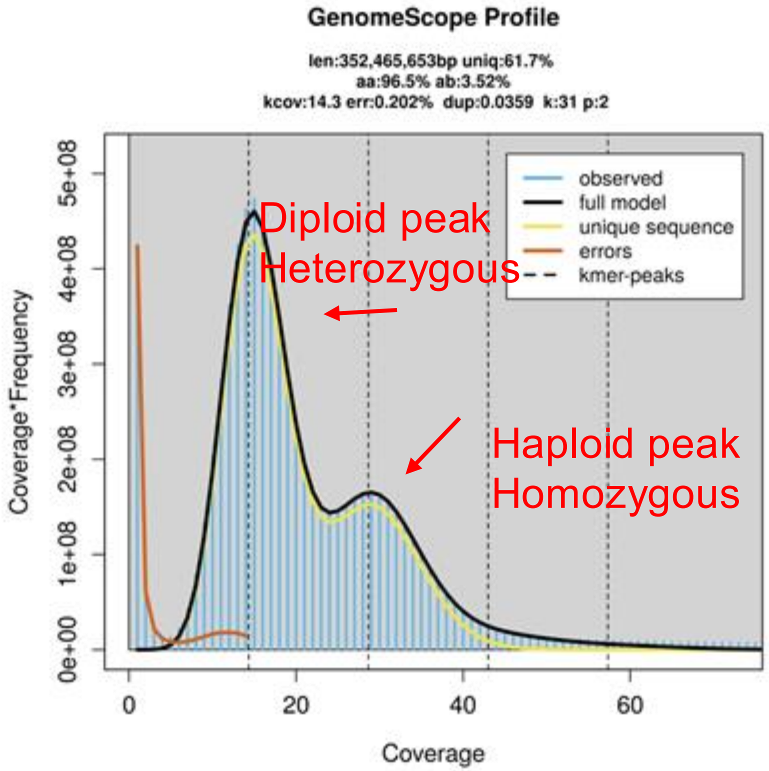


Diploid peak
Heterozygous

Haploid peak
Homozygous

Low - medium heterozygosity (~ 1%)

icHipVari1



Diploid peak
Heterozygous

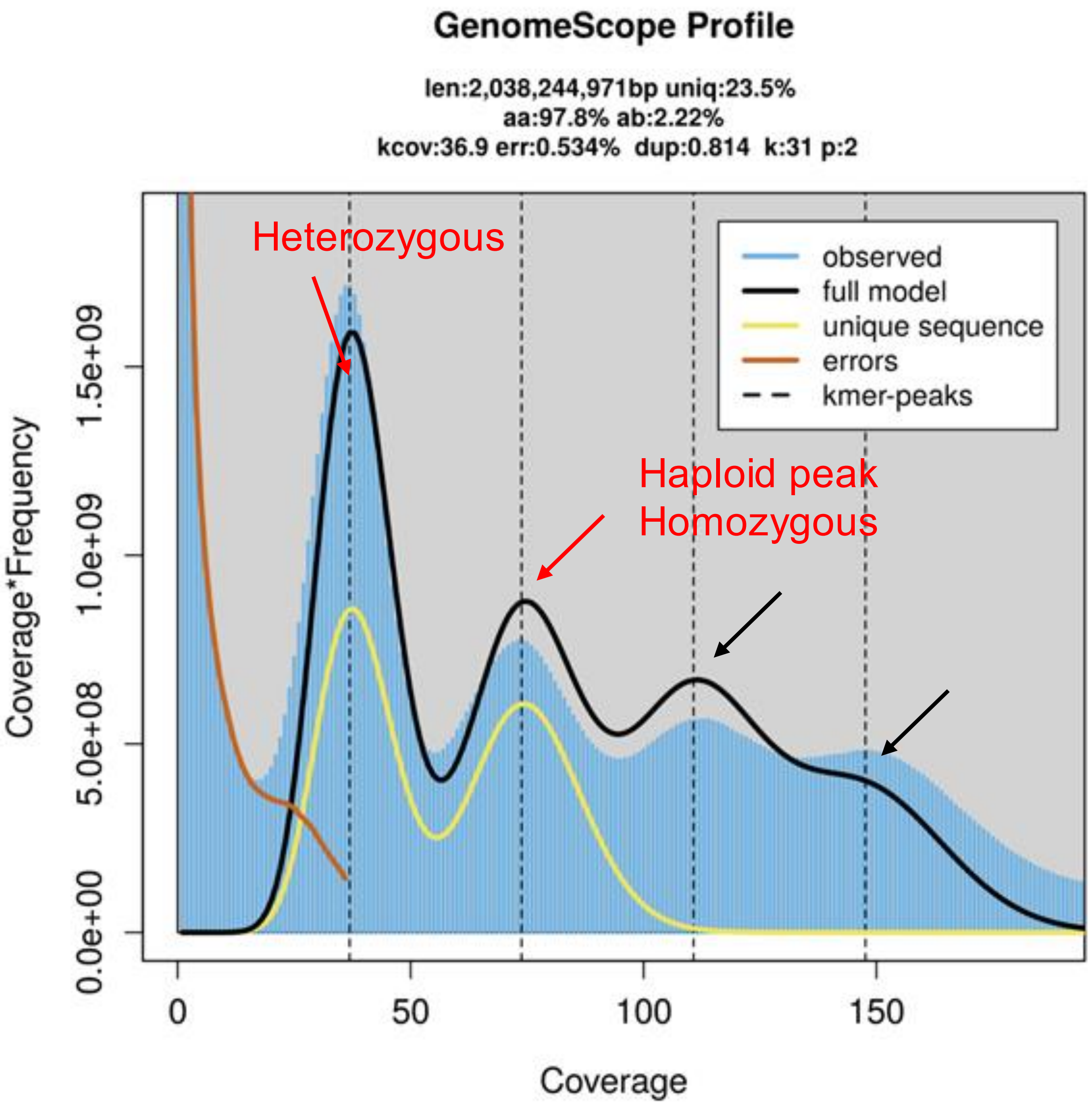
Haploid peak
Homozygous

Medium – high heterozygosity (3.52%)

K-mer distribution

Polyploids

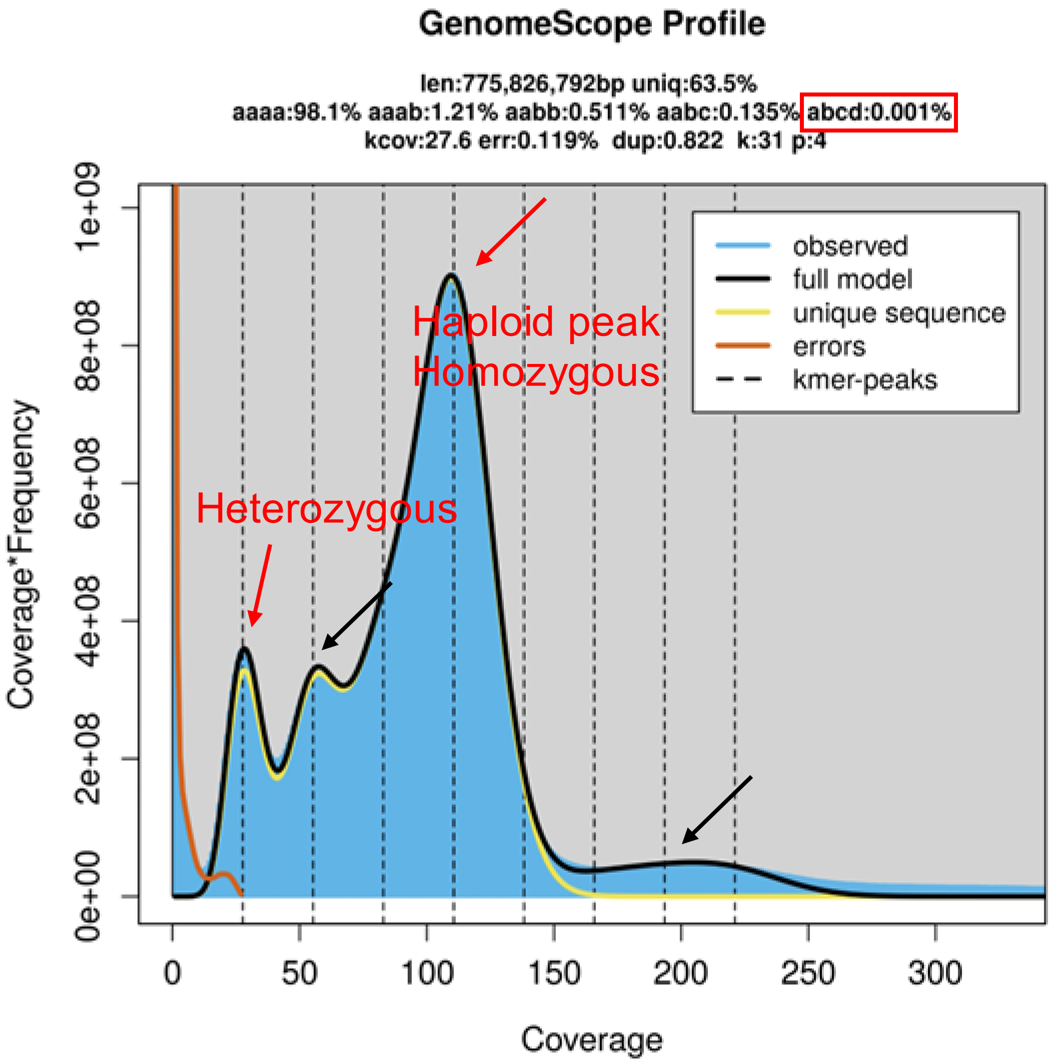
wgTheLage1



Heterozygous

Haploid peak
Homozygous

daMenTrif1



Haploid peak
Homozygous

Heterozygous

K-mer distribution and purging

Low heterozygosity (1.0)

Species	Assembler	Contig N50 (Mbp)	Contigs #	Scaffold N50	Scaffolds #	Length (Mbp)	BUSCO
ilEreMont1	Hifiasm	10,6	187			585,5	C:98.8% [S:95.5%,**D:3.3%],**F:0.5%, M:0.7%,n:1367
ilEreMont1	Hifiasm + purging	10,9	99			557,2	C:98.7% [S:97.7%,**D:1.0%],**F:0.5%, M:0.8%,n:1367
ilEreMont1	Hifiasm + scaffolding	10,9	109	21,6	45	557,2	C:98.7% [S:97.7%,**D:1.0%],**F:0.5%, M:0.8%,n:1367
ilEreMont1	hifiasm-hic.scaffolding_hap1.yahs	7,7	215	21,5	116	530,5	C:92.8%[S:92.3%,**D:0.5%],** F:0.5%,M:6.7%,n:1367
ilEreMont1	hifiasm-hic.scaffolding_hap2.yahs	9,2	196	21,3	95	543,2	C:98.8%[S:98.5%,**D:0.3%],** F:0.7%,M:0.5%,n:1367

Not too much difference in assembly size after purging or phased assembly

K-mer distribution and purging

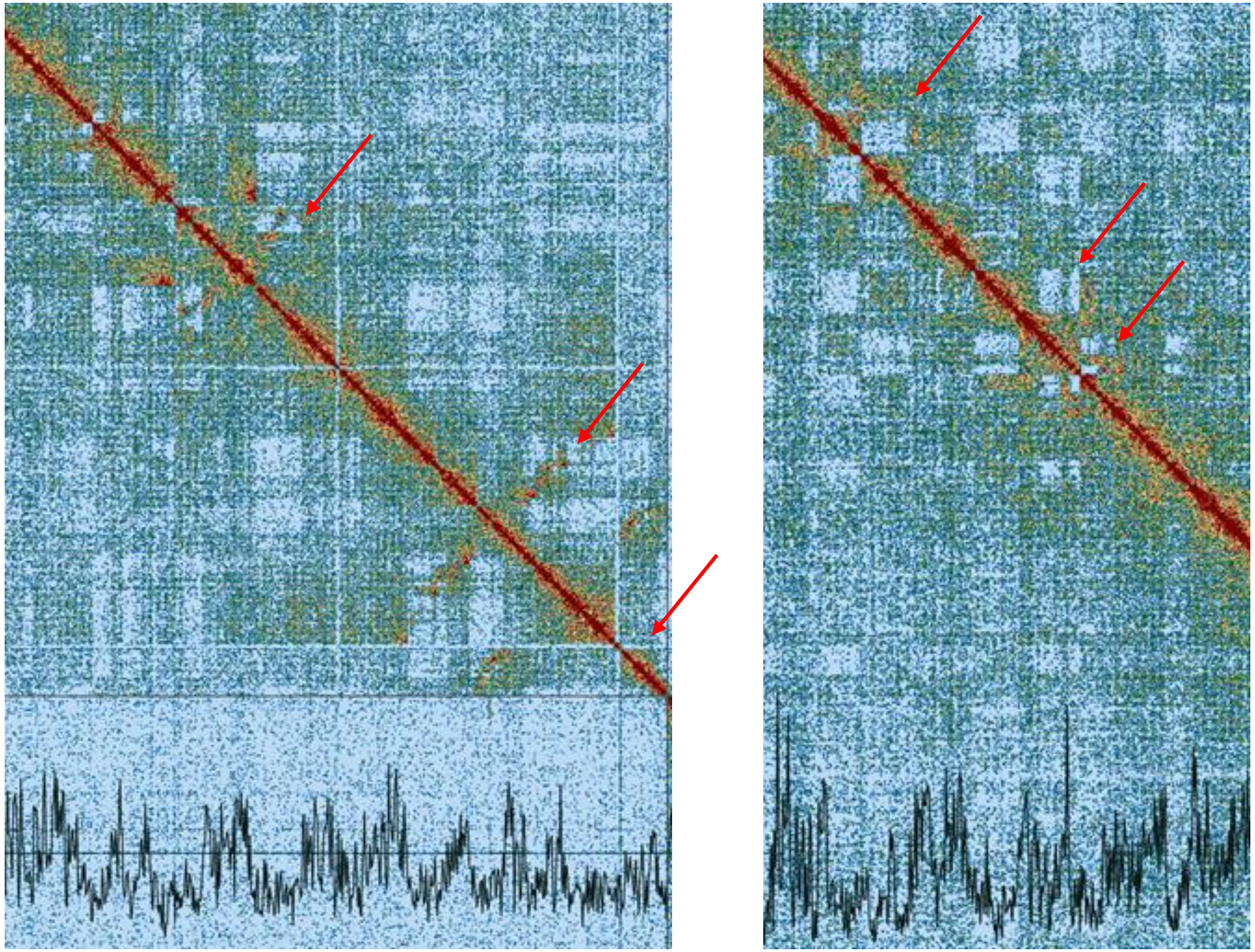
Medium - high heterozygosity (2.43)

Species	Assembler	Contig N50 (Mbp)	Contigs #	Scaffold N50	Scaffolds #	Length (Mbp)	BUSCO
laLemMinu1	Hifiasm (primary)	122	13,752			794	C:98.8%[S:89.2%,**D:9.6%],** F:0.5%,M:0.7%,n:425
laLemMinu1	hifiasm.purging	190	9,196			657,5	C:98.6%[S:93.4%,**D:5.2%],** F:0.5%,M:0.9%,n:425
laLemMinu1	hifiasm-hic.scaffolding_ hap1.yahs	96	11,567	1,932,940	8,015	573,16	C:97.2%[S:90.4%,**D:6.8%**], F:0.9%,M:1.9%,n:425
laLemMinu1	hifiasm-hic.scaffolding_ hap2.yahs	111	8,891	6,207,450	5,623	525,91	C:97.4%[S:93.2%,**D:4.2%],** F:0.7%,M:1.9%,n:425

Difference in assembly size – size is expected to change after purging or phasing

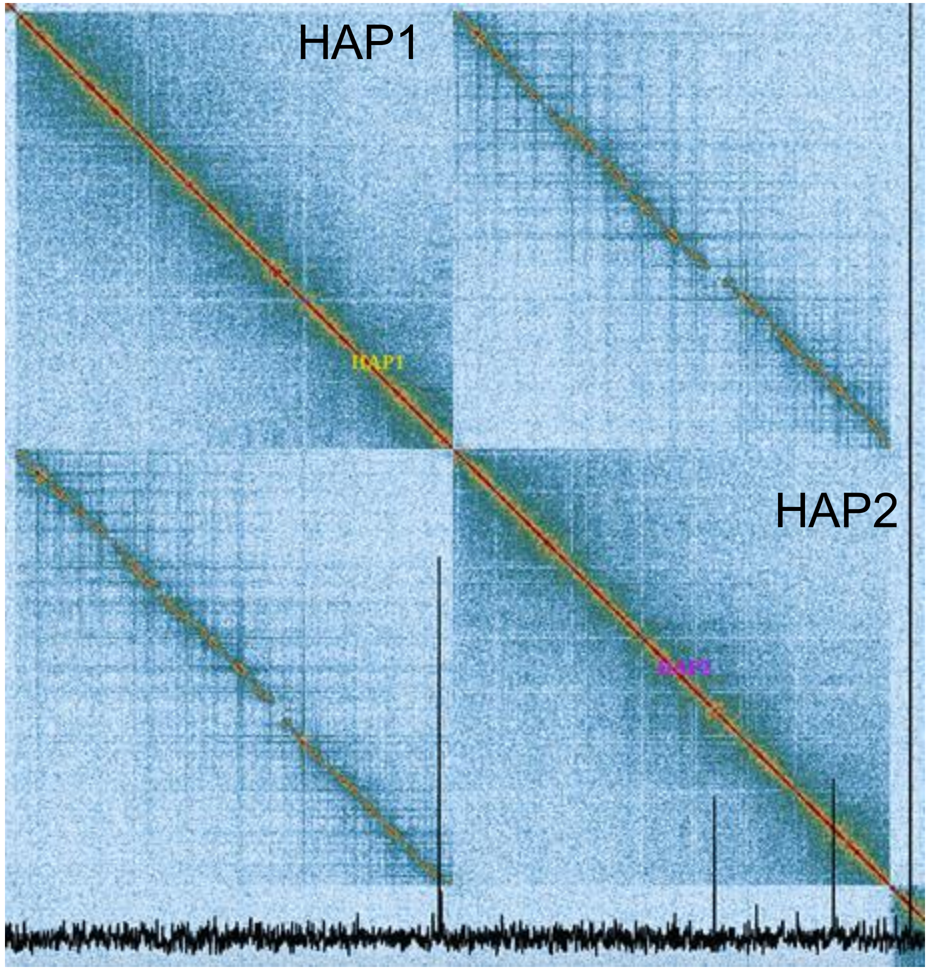
K-mer distribution and purging

laLemMinu1
Hifiasm purged assembly looks like this



Many retained haplotigs

Phased assembly

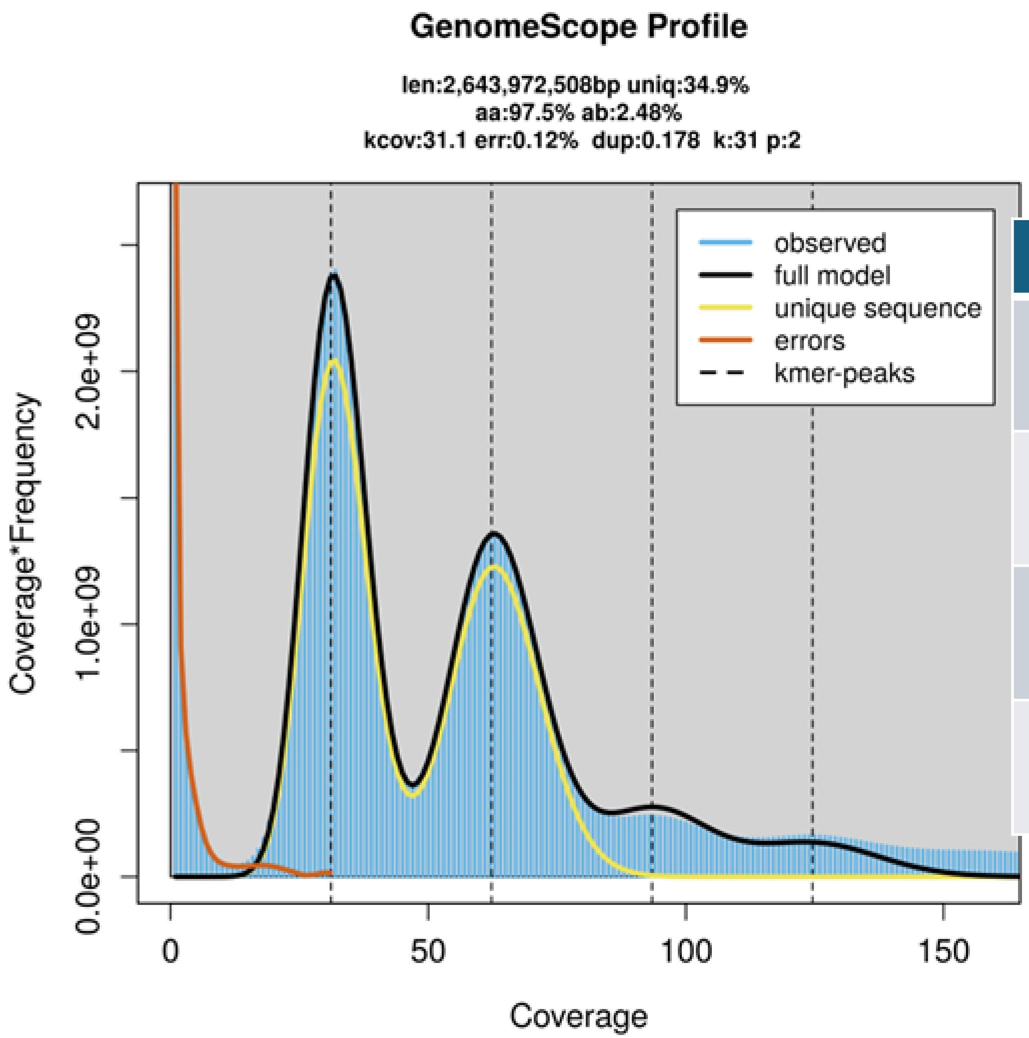


No more haplotigs

Heterozygous and repetitive regions

Retained hap dups and repetitive regions very hard to assemble

pacbio daMatCham1 GenomeScope 2.0 linear plot



Medium to high heterozygosity (2.48%)
Very repetitive genome (65%)

asm	Length	BUSCO
Hifiasm	3,50 Gbp	C:97.4%[S:52.6%,**D:44.8%],**F:0.3%,M:2.3%,n:2326
Hifiasm.purging	1,37 Gbp	C:**77.0%**[S:57.7%,**D:19.3%],**F:0.9%,M:22.1%,n:2326
Hifiasm_hap1	2,58 Gbp	C:96.9%[S:83.1%,**D:13.8%],**F:0.4%,M:2.7%,n:2326
Hifiasm_hap2	2,55 Gbp	C:96.7%[S:90.7%,**D:6.0%],**F:0.4%,M:2.9%,n:2326

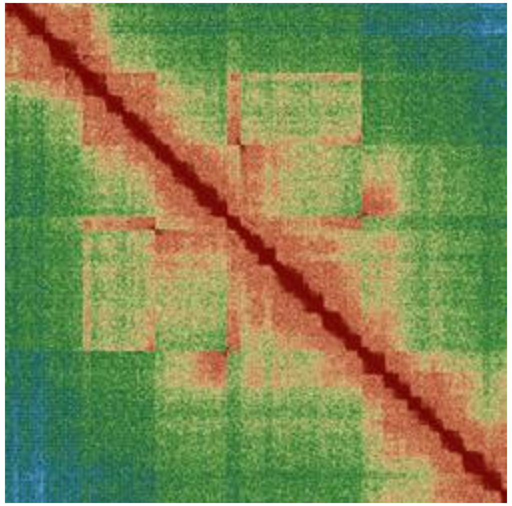
Real genome size close to 5 Gbp

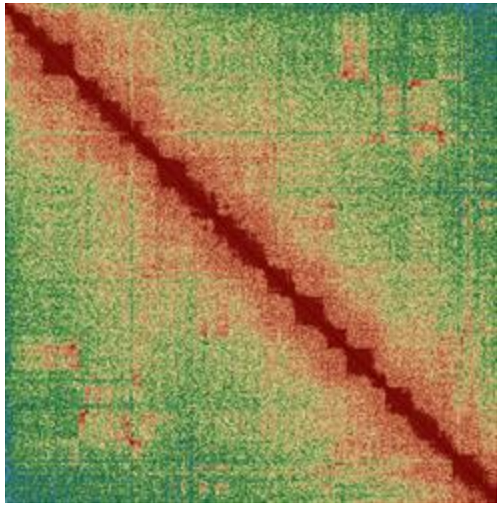
Difference between primary (purged) and merged assemblies for the same high heterozygous genome

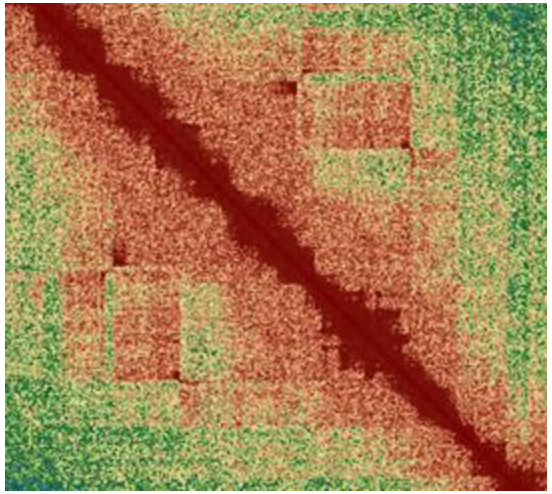
daMatCham1

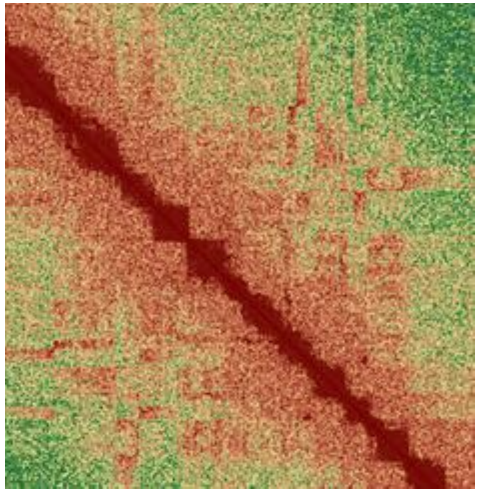
Even purged, there are inversions impossible to solve during curation

Primary assembly

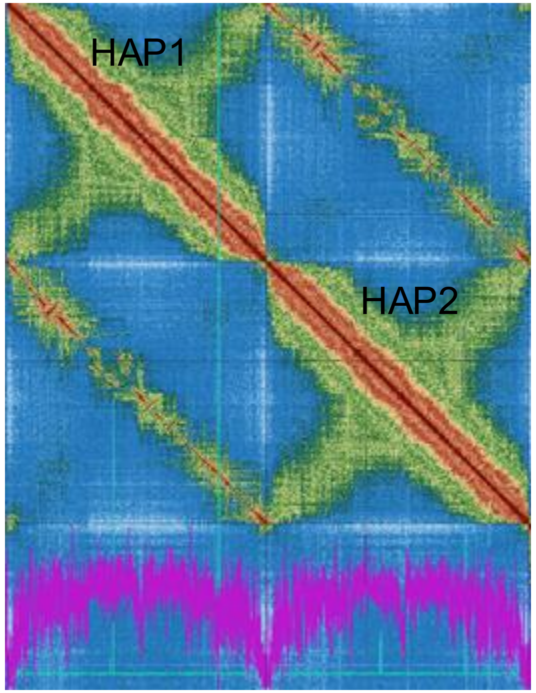


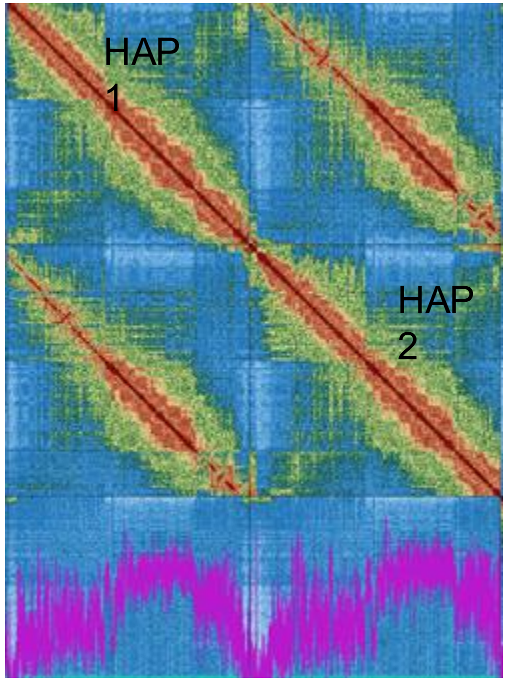


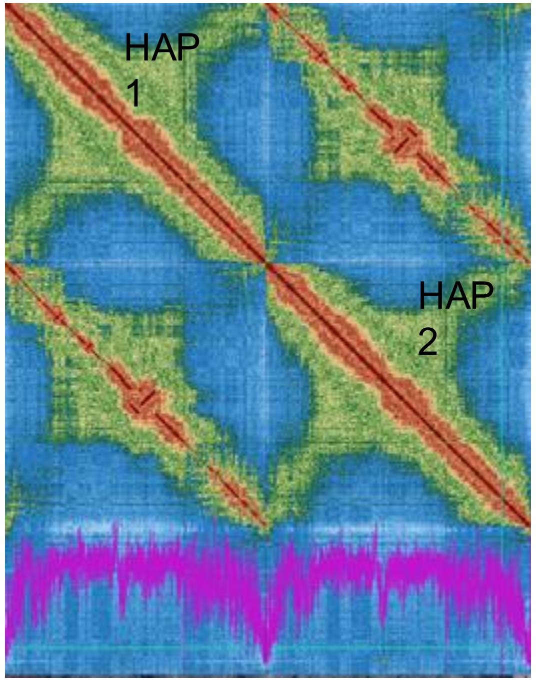




Phased assembly



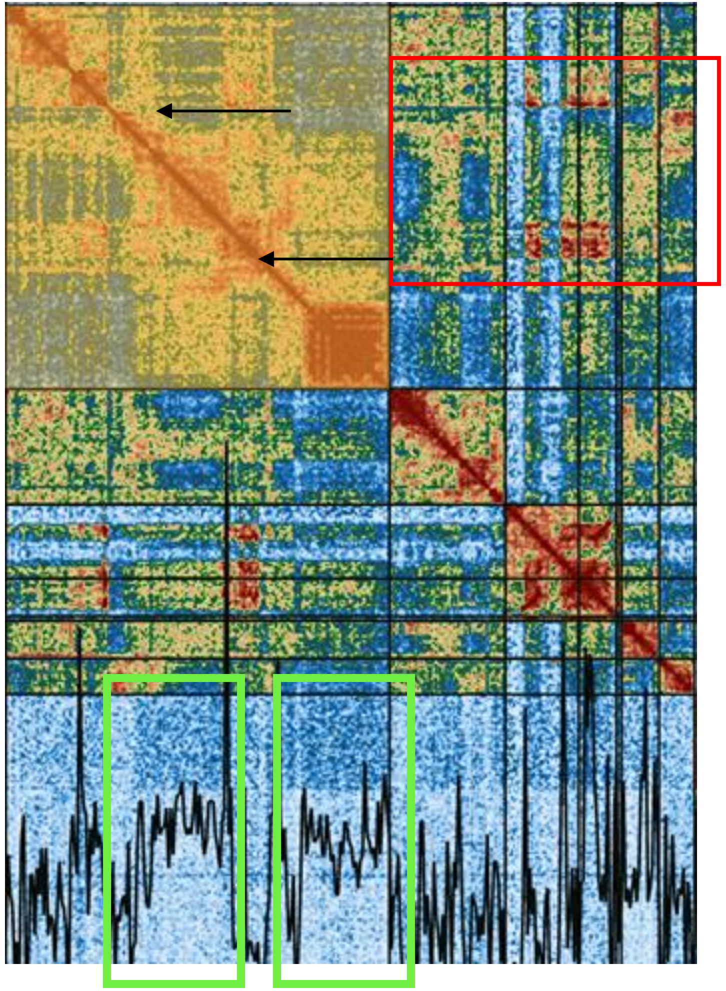




Repeat track - purple

Phased assemblies - Repeats

Repetitive scaffold +
smaller repetitive scaffolds from the
shrapnel



Looks to be
collapsed

Where should they go?

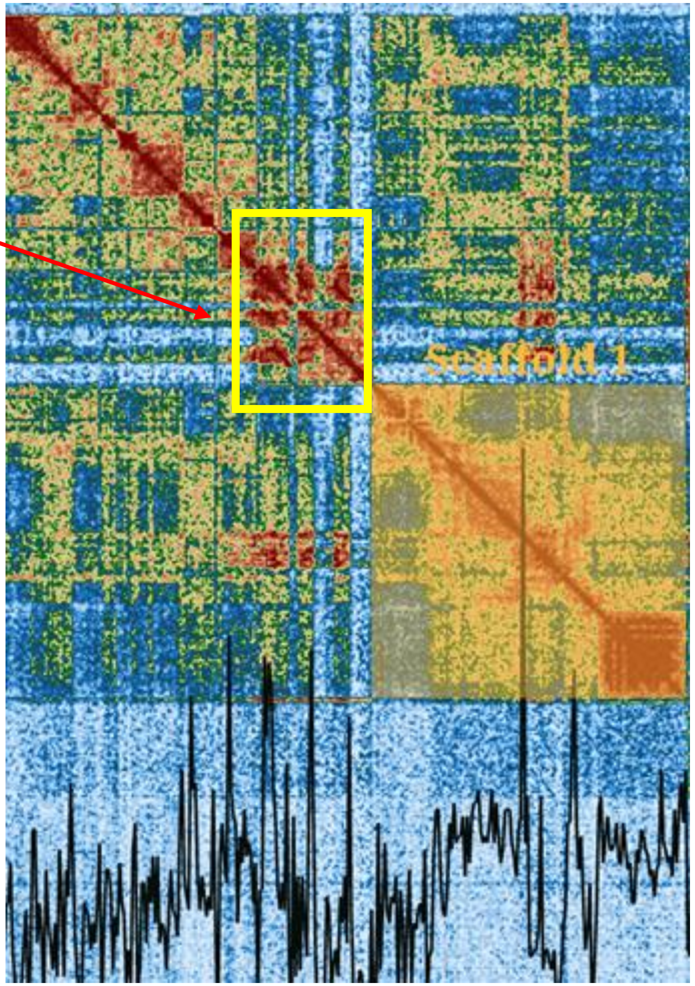
Duplicated
regions

???

sHetFra1

Primary assembly

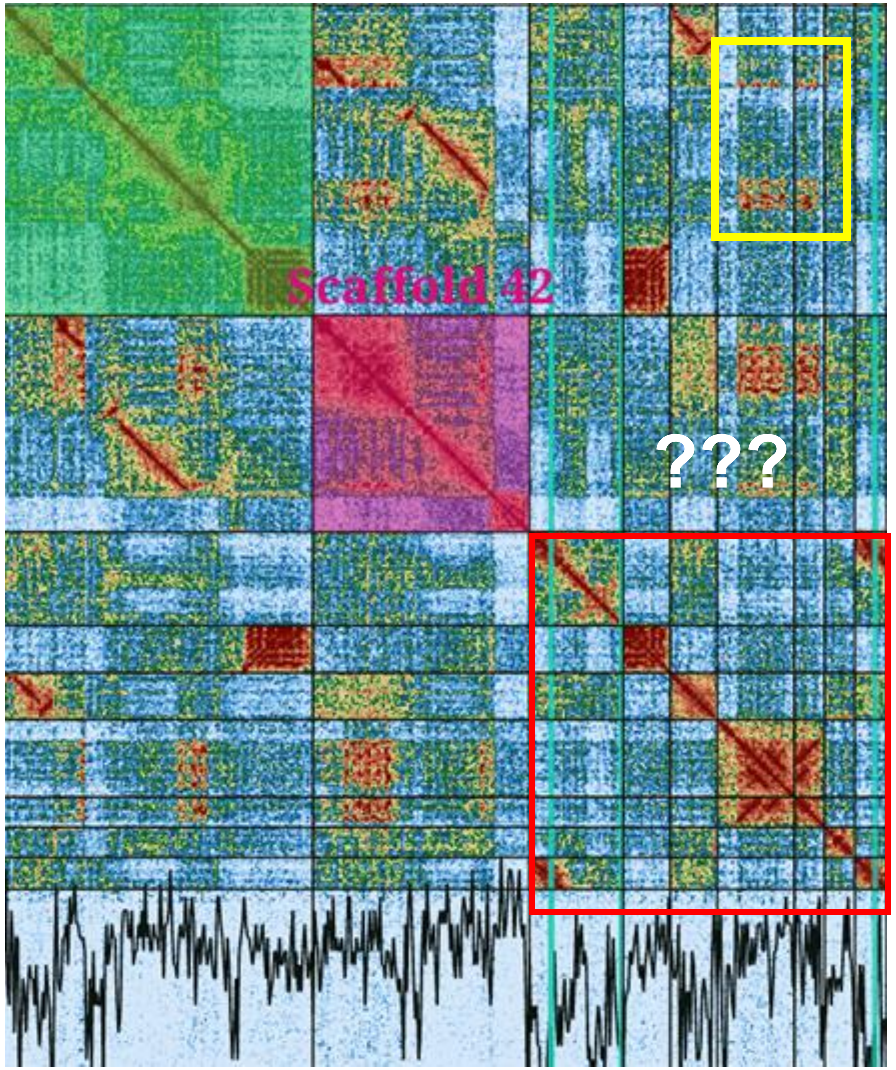
Is this the best representation?



Phased assemblies - Repeats

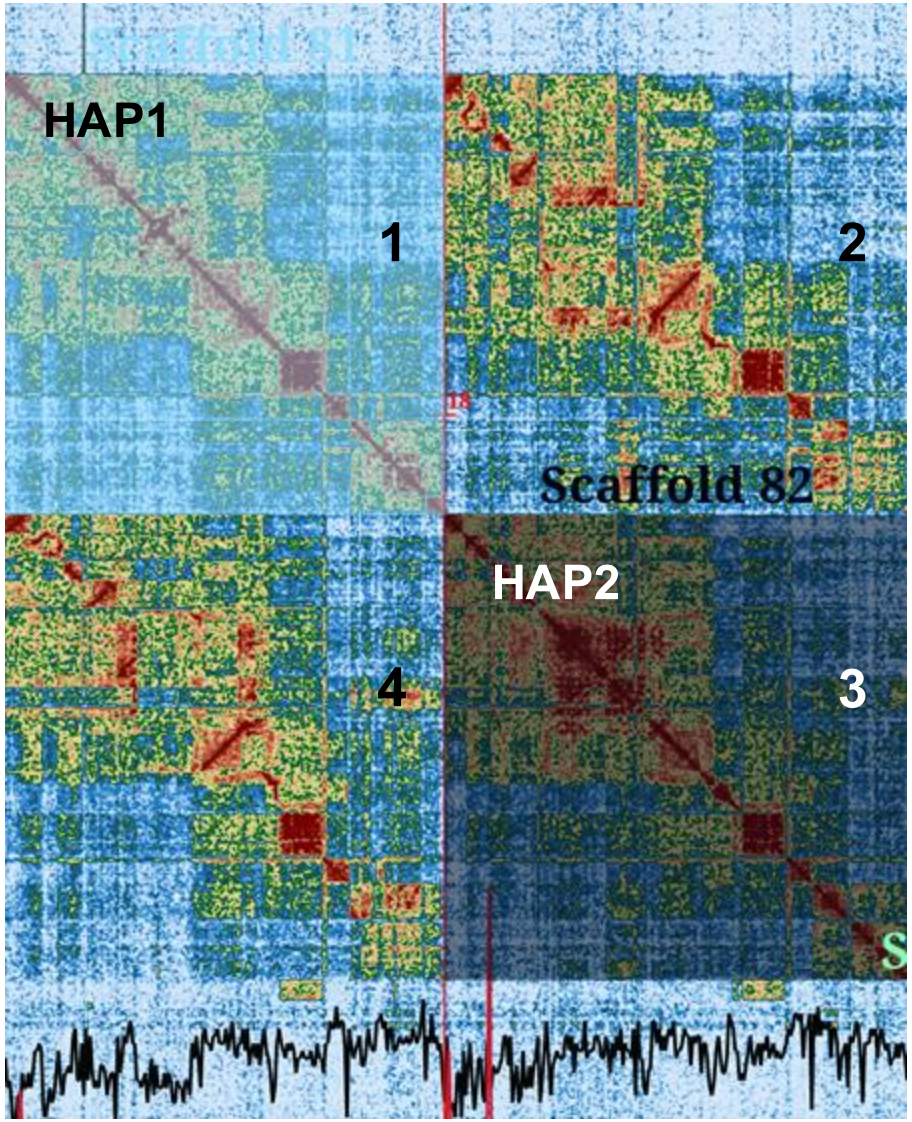
Haplotypes 1 and 2 should be as similar as possible

Primary assembly



They look to go in more than one place

Phased assembly



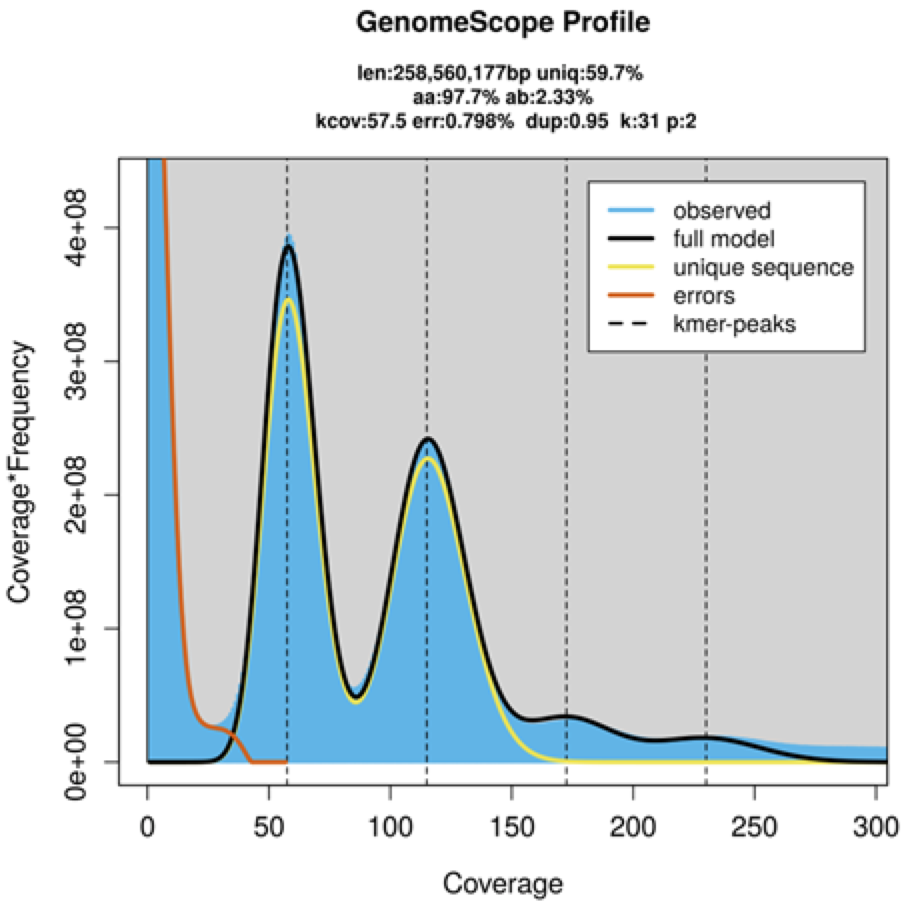
Repeats in one hap slightly assembled helps to assemble repeats in the other hap

sHetFra1

What to expect when purging doesn't work

odCliOrie1

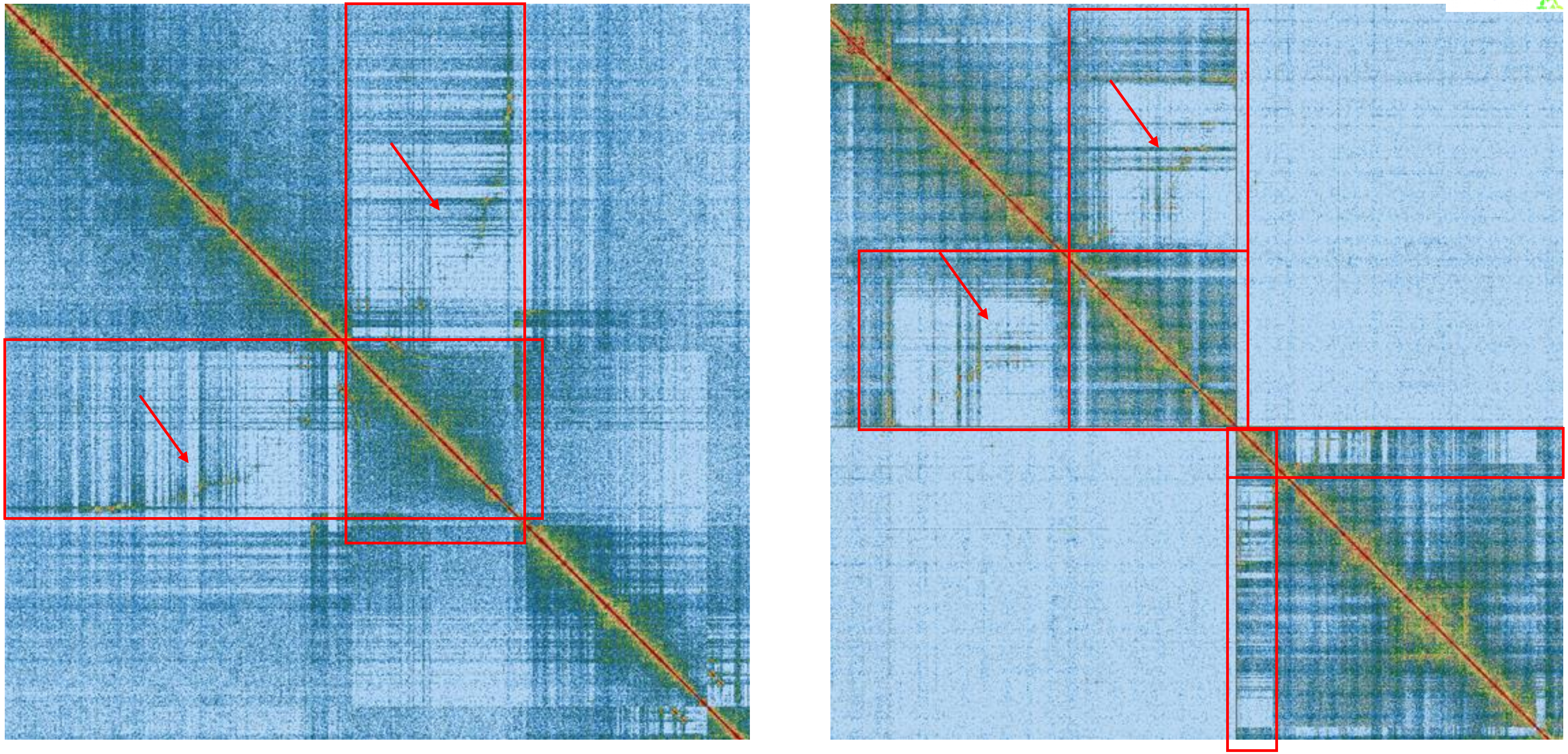
2.35 heterozygozity



asm	Length	BUSCO
Hifiasm	894 Mbp	C:89.0%[S:67.2%,**D:21.8%],**F:5.1%,M:5.9%,n:954
Hifiasm purging	807 Mbp	C:88.6%[S:69.5%,**D:19.1%],**F:5.2%,M:6.2%,n:954

What to expect when purging doesn't work

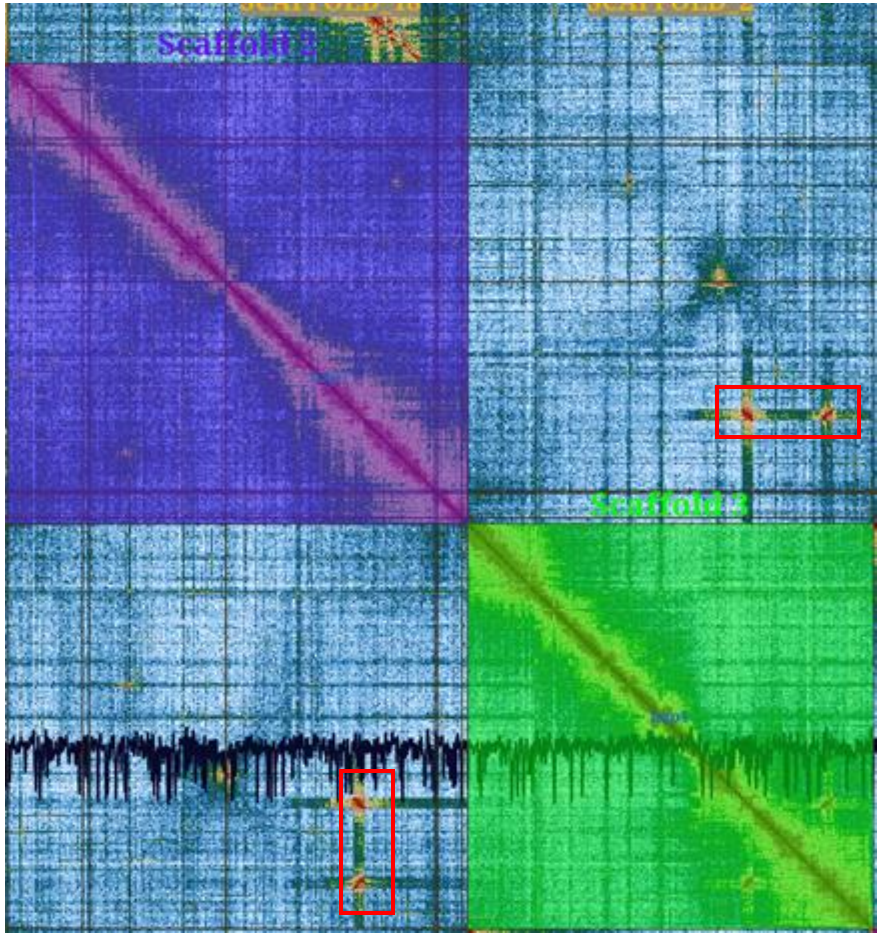
odCliOrie1 2.35 heterozygozity



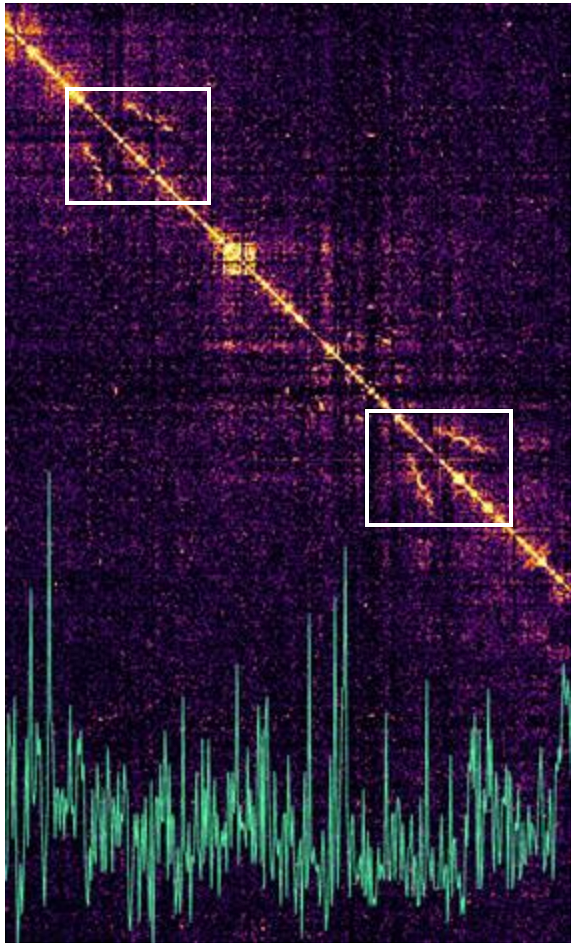
Many remaining haplotigs. Should be removed during curation

Real gene duplication or retained haplotig?

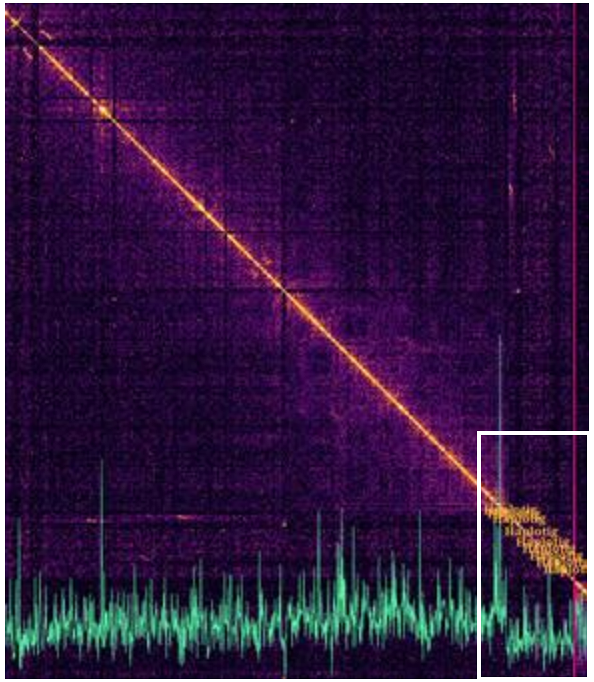
Real gene duplication – even coverage



gfHygCocc2



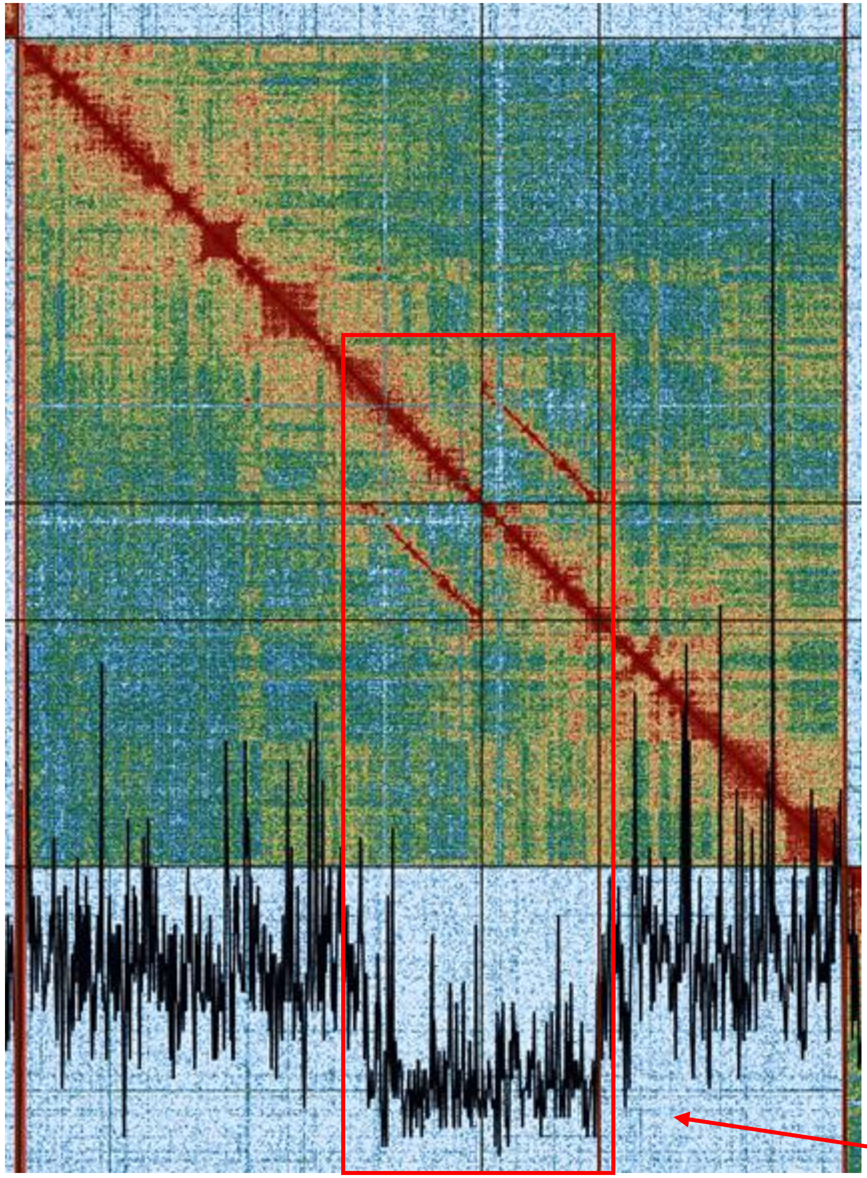
xbLucDiva1



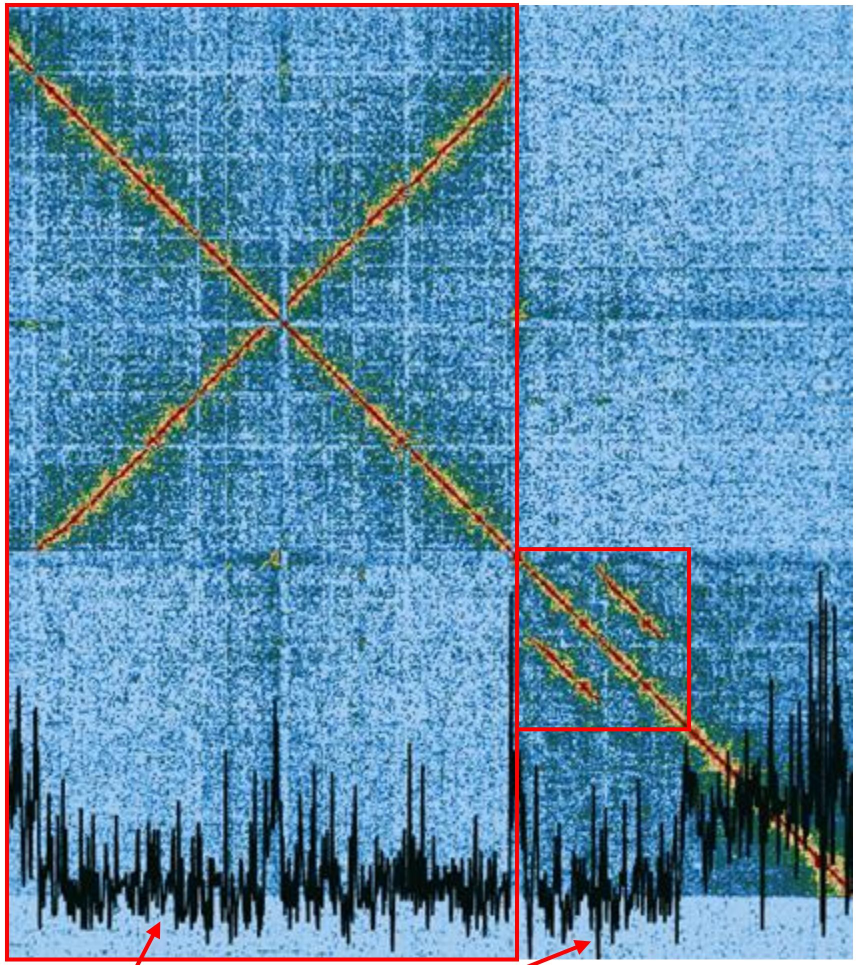
Half coverage haplotigs

Retained haplotigs examples

ilThyBati1



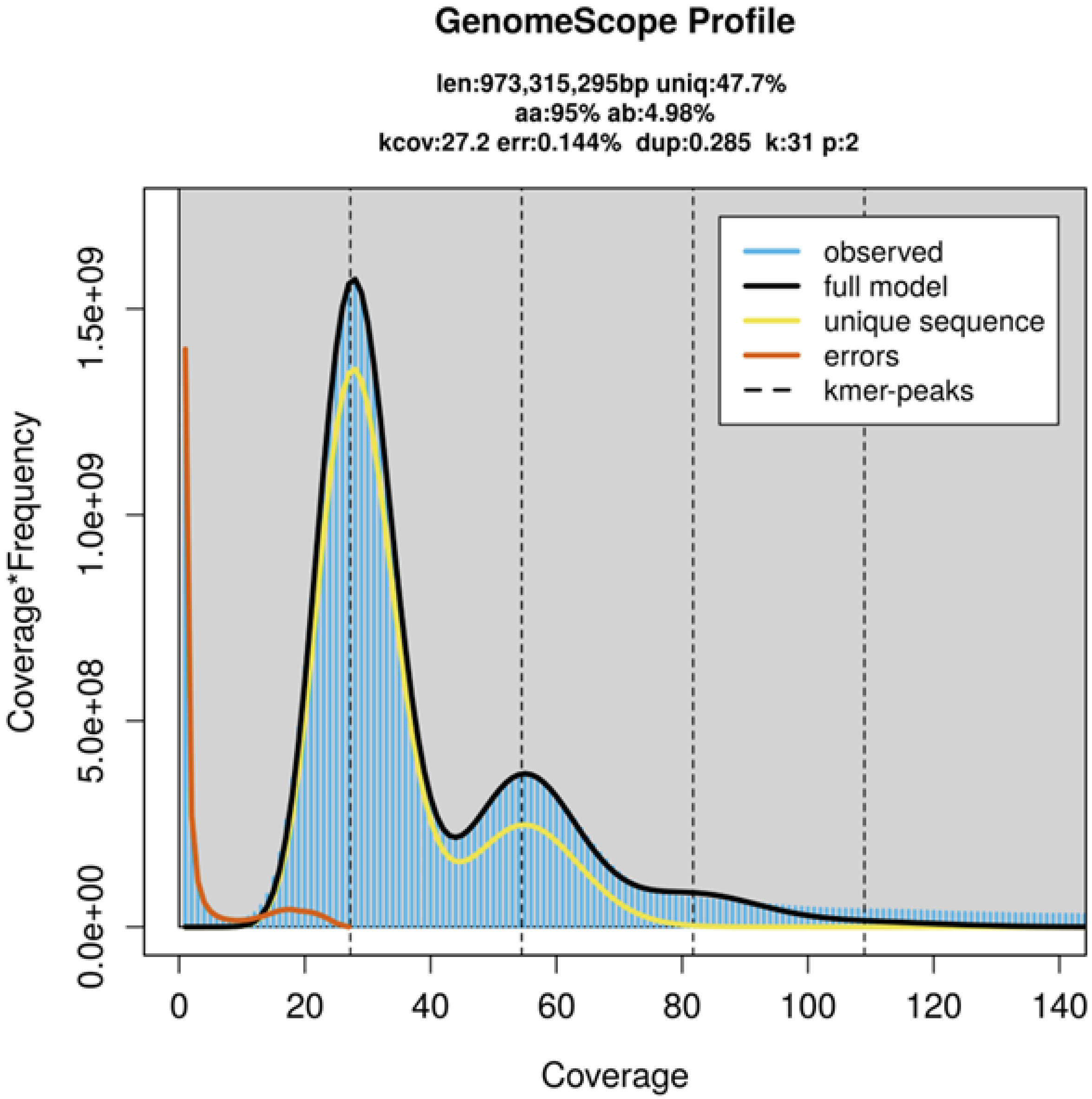
xbTriPhas3



Half coverage

Real haplotigs, should be removed

Phased assemblies



xbArcSenh1

Heretozygozity = 5 %

Alternative to solve medium to high heterozygosity
Inversions
Purging issues
Repetitive regions (in part)

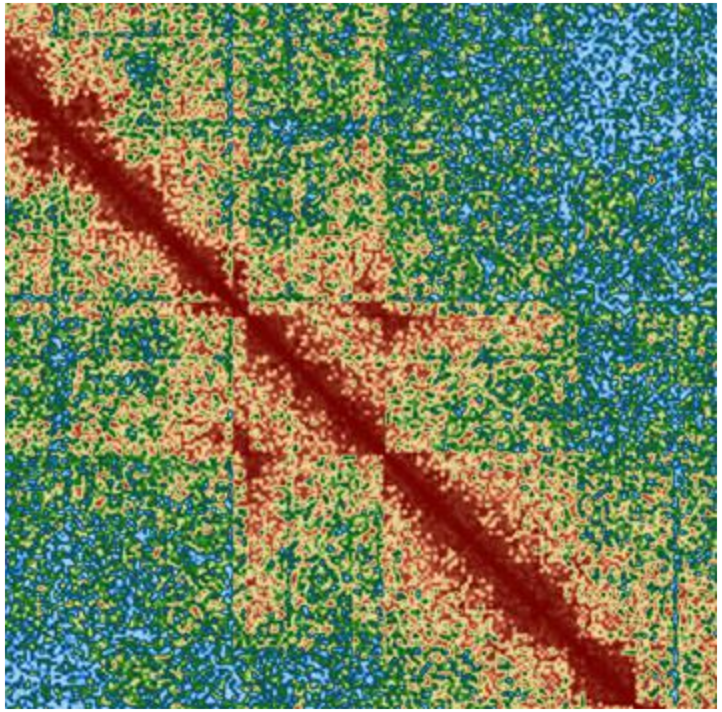
Only possible when PB and HiC data are from the same sample

Phased assemblies - Inversions

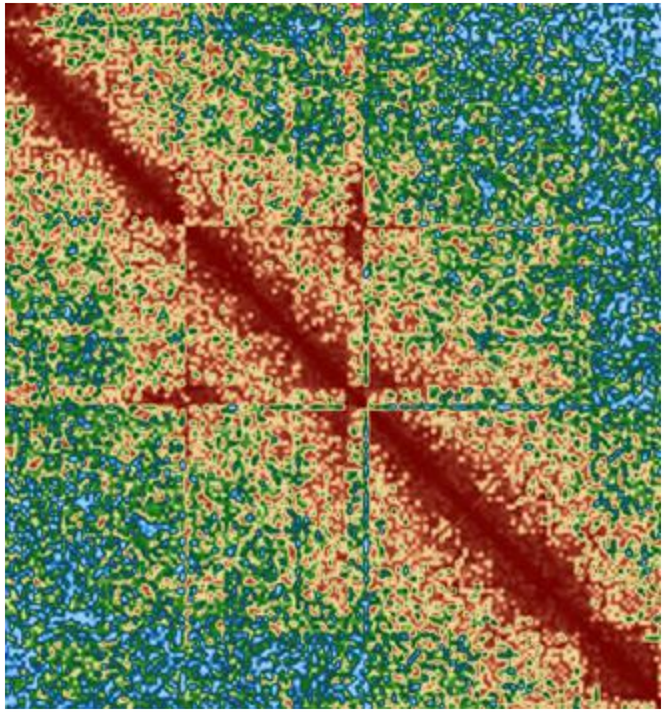
High heterozygosity + inversions between haplotypes
(sister chromatids)

Primary assembly
Inversion
Never looks right

Conformation1

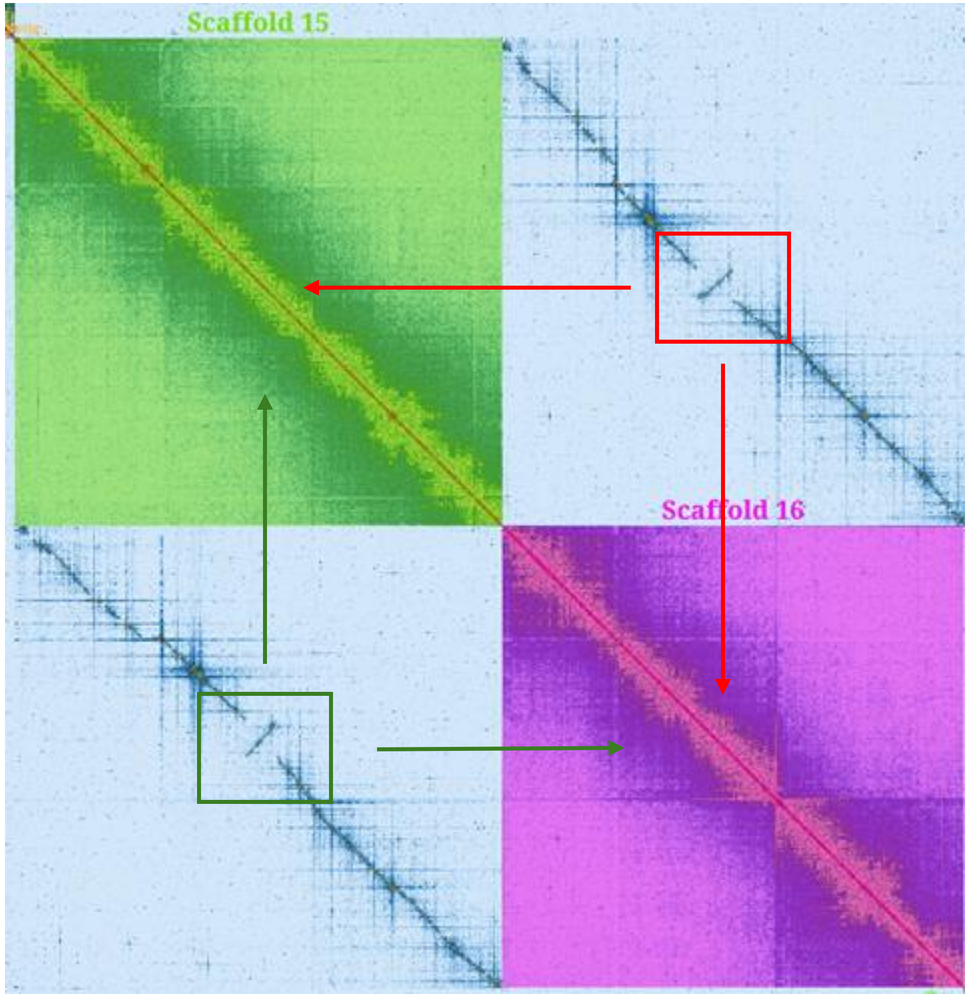


Conformation 2



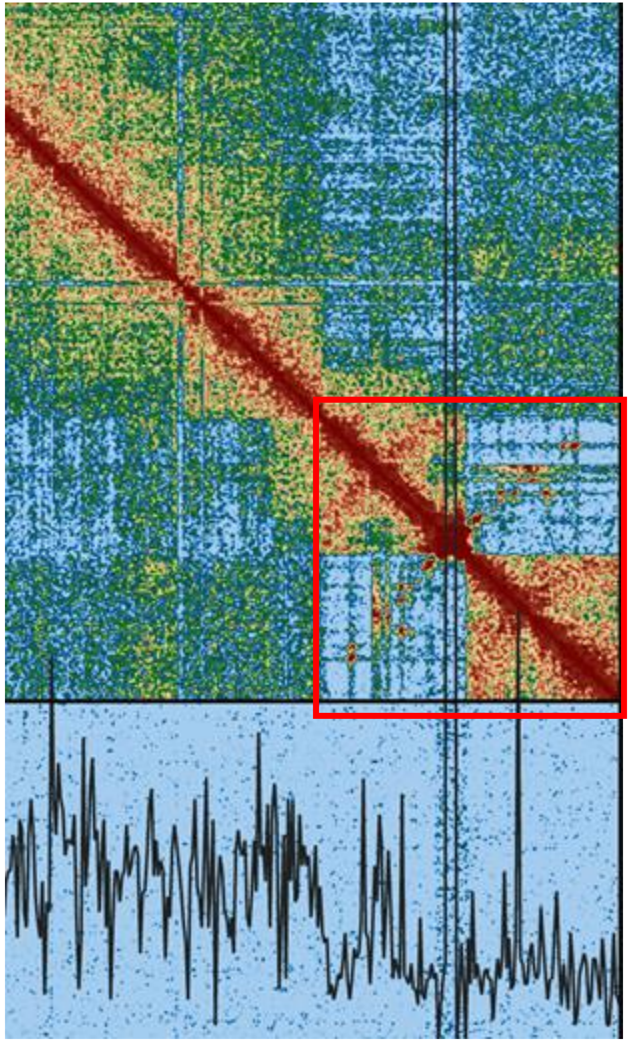
xbArcSenh1

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

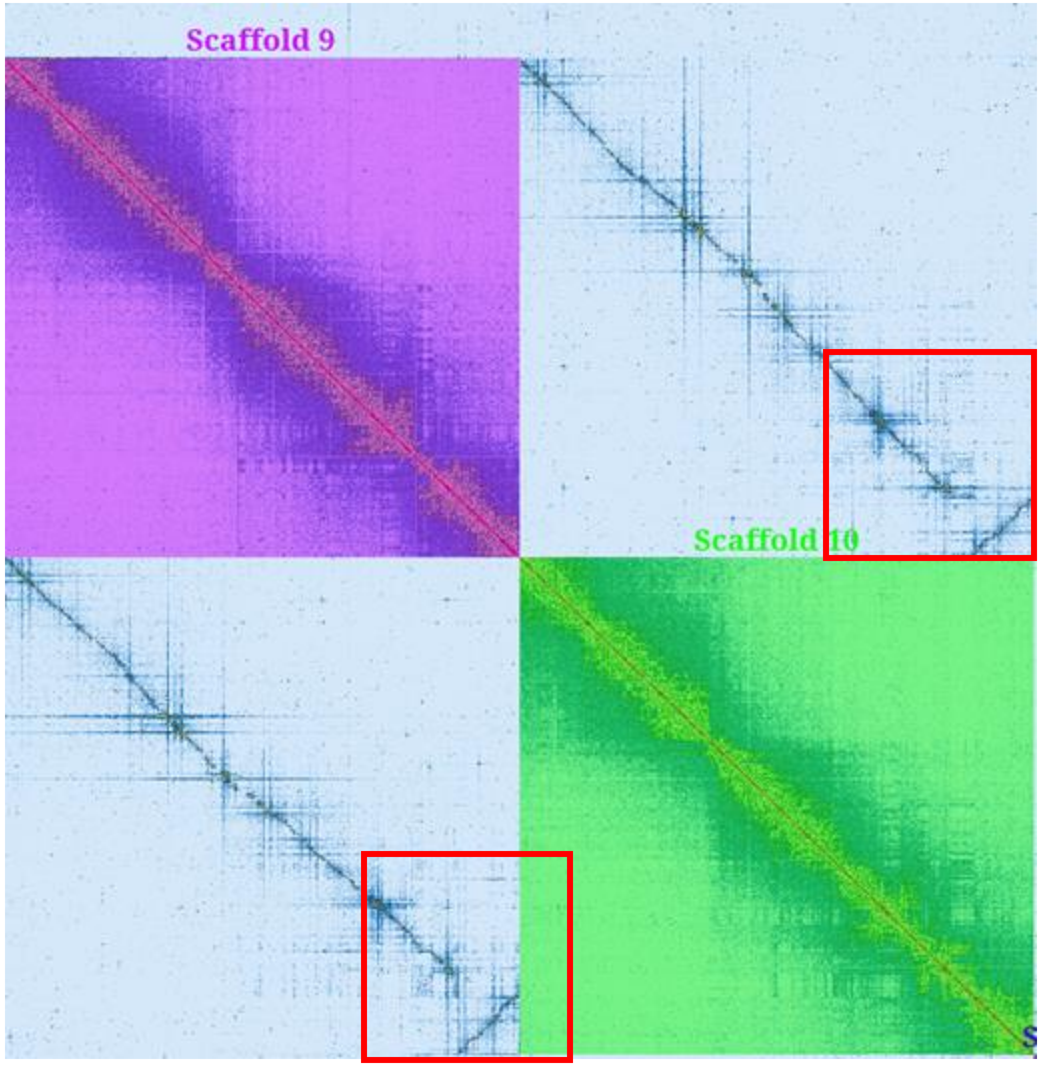


Phased assemblies – Inversions + haplotigs

Primary assembly



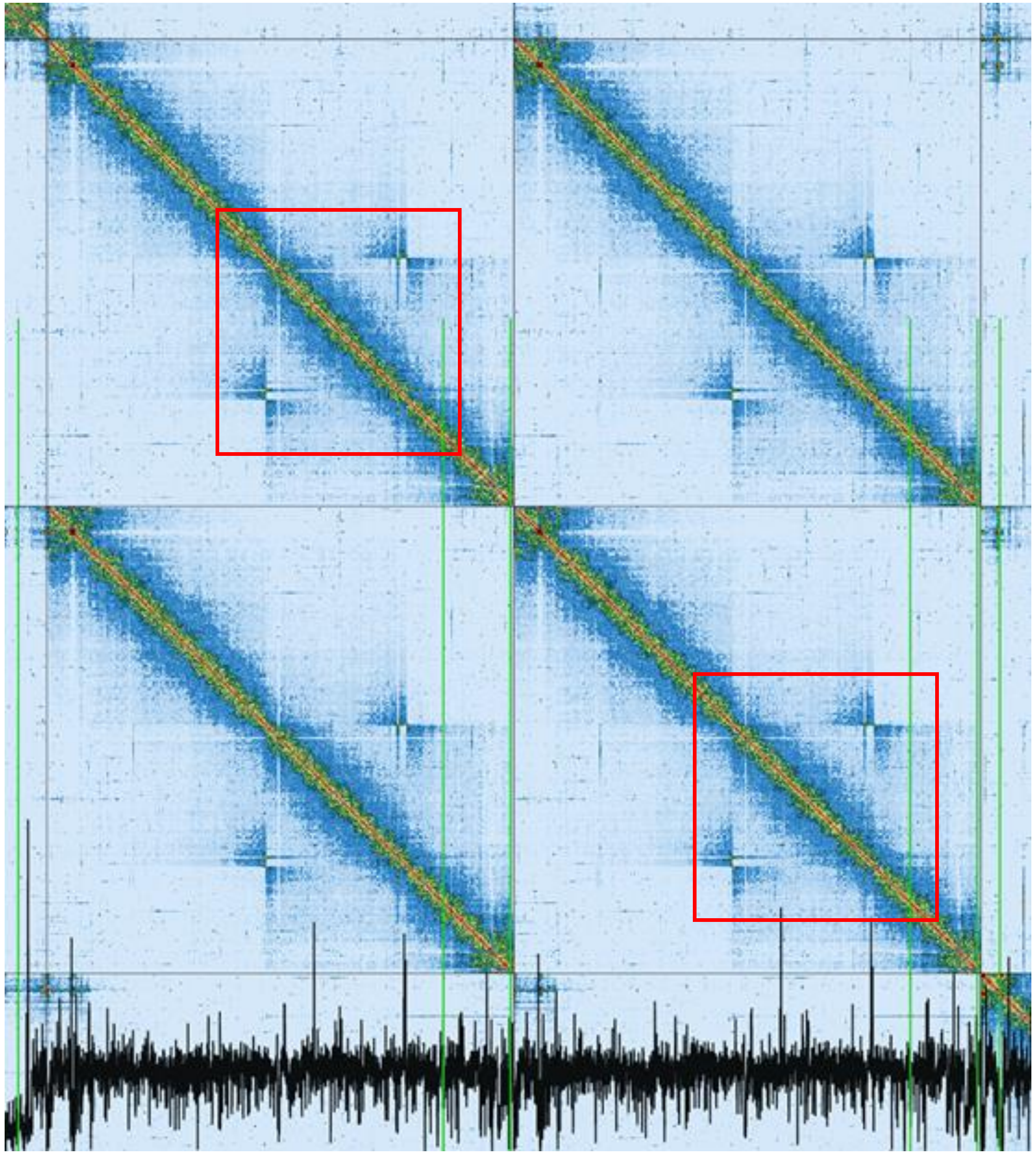
xbArcSenh1



Telomeric region is inverted between haps
Purging failed

Resolved when we have both haps

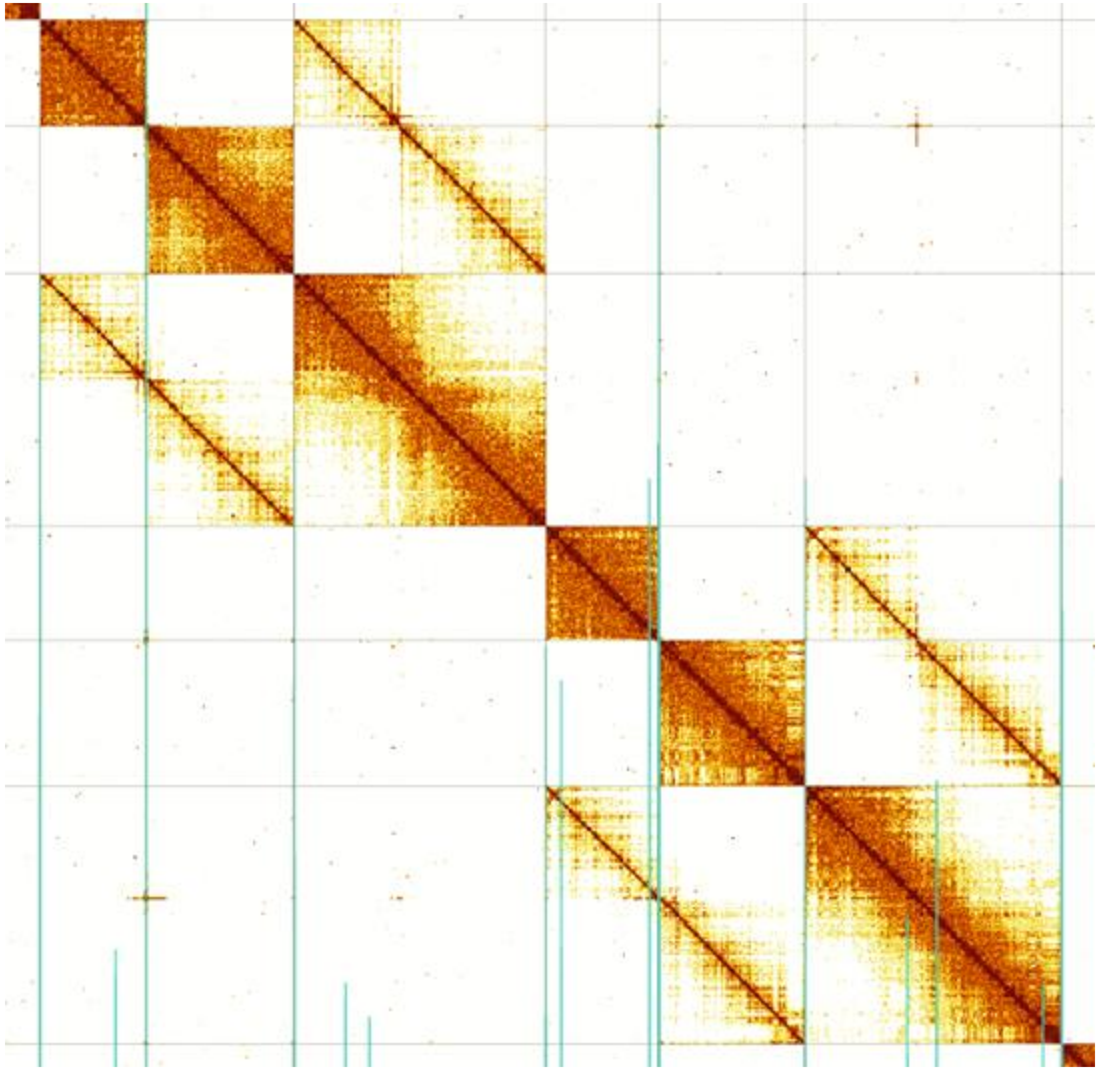
What happens when PB and HiC are from different samples? – Phased assemblies

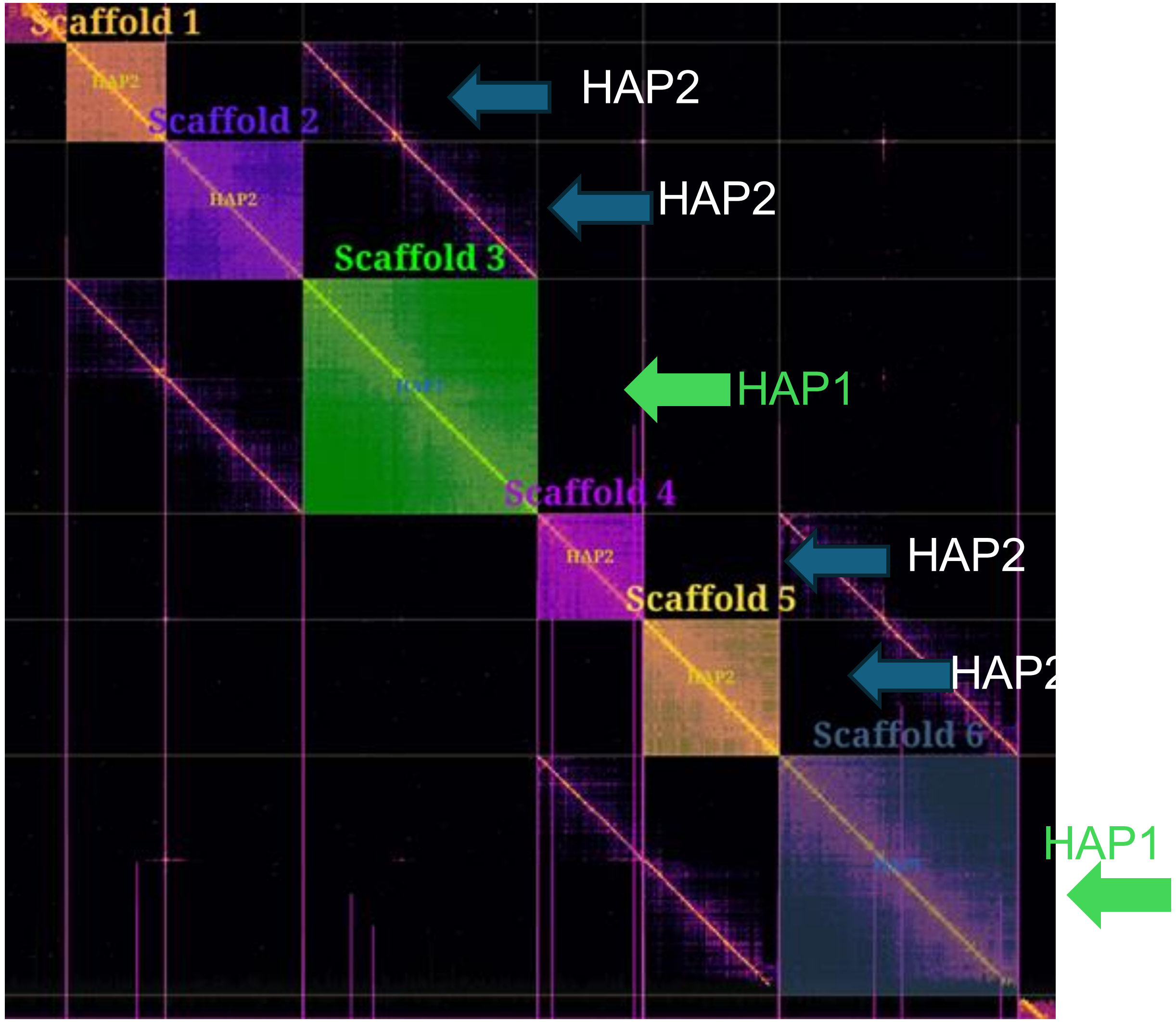


ieBaeAtla2

Phased assemblies

Polymorphism among haplotypes – different chromosome number





HAP2

HAP2

HAP1

HAP2

HAP2

HAP1

Hands-on

https://github.com/epaule/Physalia-Manual-Genome-Curation/blob/main/Session2.1.md