Manual Genome Curation using PretextView

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

Course overview

Day 1

Session 1: Manual curation overview

Session 2.1: What to infer from assembly quality metrics?

Session 2.2: Decontaminate your assembly before curation

Day 2

Session 3.1: Beginning manual curation

- How to use PretextView
- Single haplotype curation

Day 3

Session 3.2: How to generate your own PretextView Hi-C maps

- Dual haplotype curation
- Generating the curated fasta file

Day 4

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

Day 5

Session 5: Working on more challenging genomes

Most of the time will be for hands-on

Session 1: Manual curation overview

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

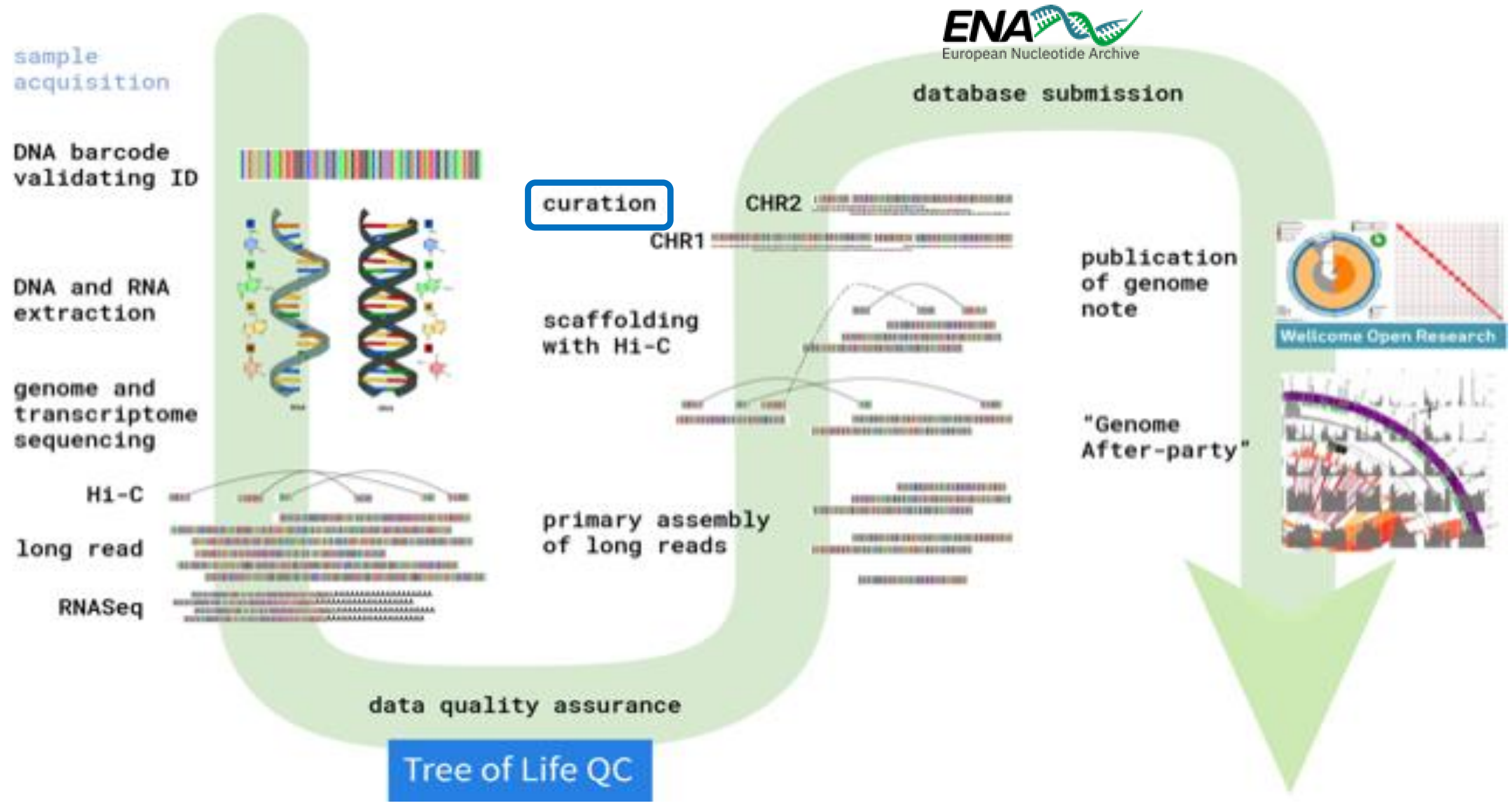
What is genome curation?

“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**”

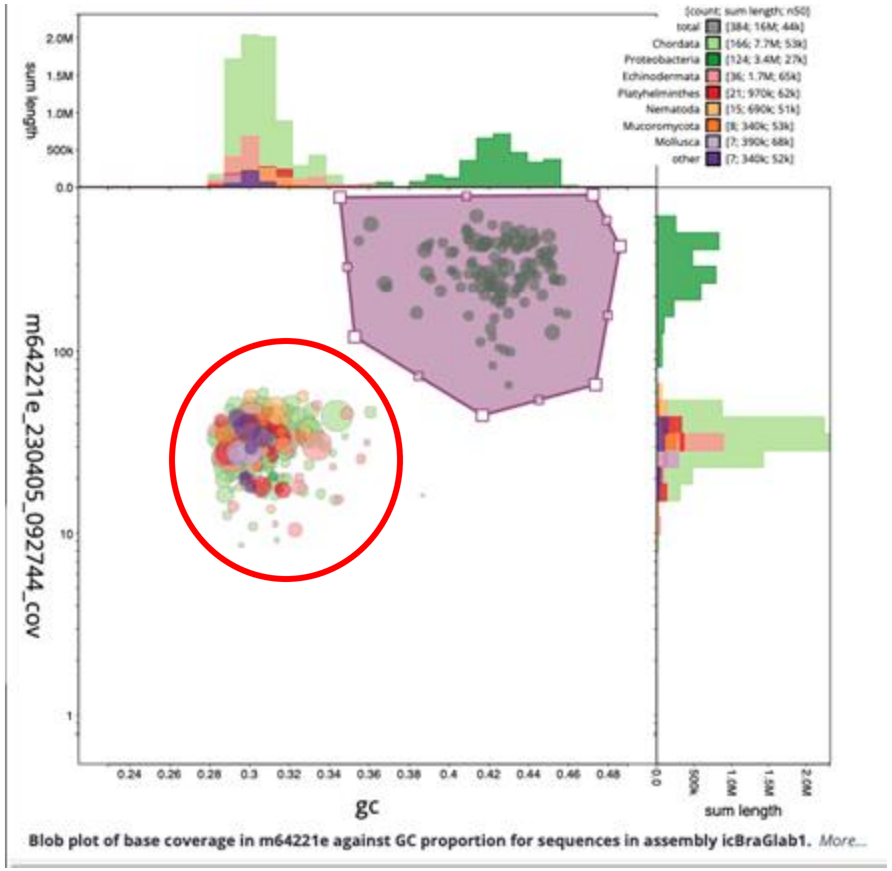
Our experience:

- Darwin Tree of Life Project
- Vertebrate Genome Project
- Aquatic Symbiosis Genomes project
- European Reference Genomes Atlas
- Genome Reference Consortium
- Telomere 2 Telomere Project
- Human Pangenome Reference Consortium

The Tree of Life genome factory

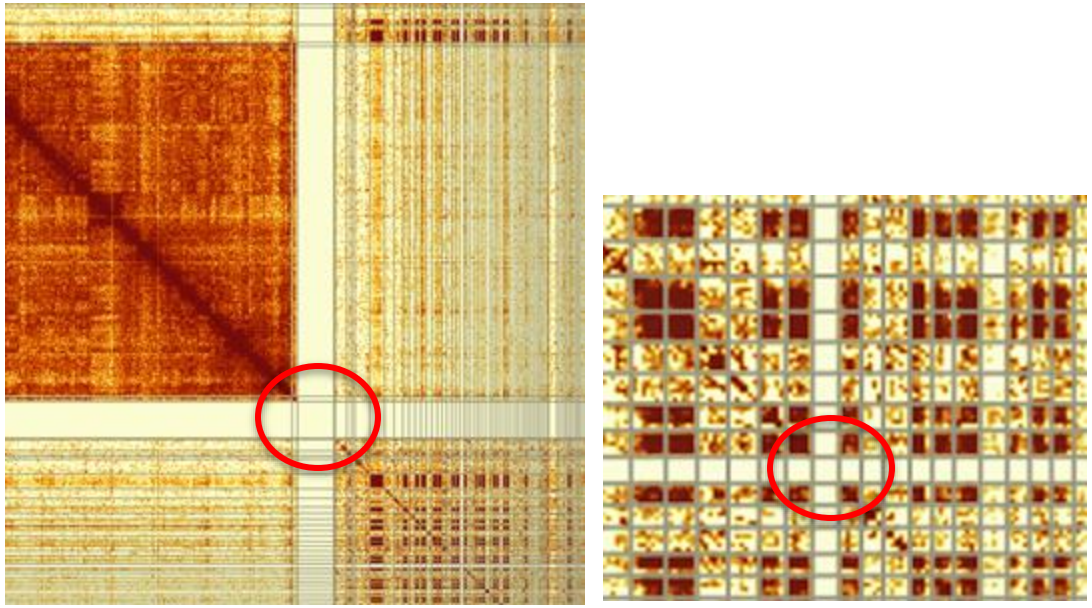


Decontamination examples



Blob plot of base coverage in m64221e against GC proportion for sequences in assembly icBraGlab1. *More...*

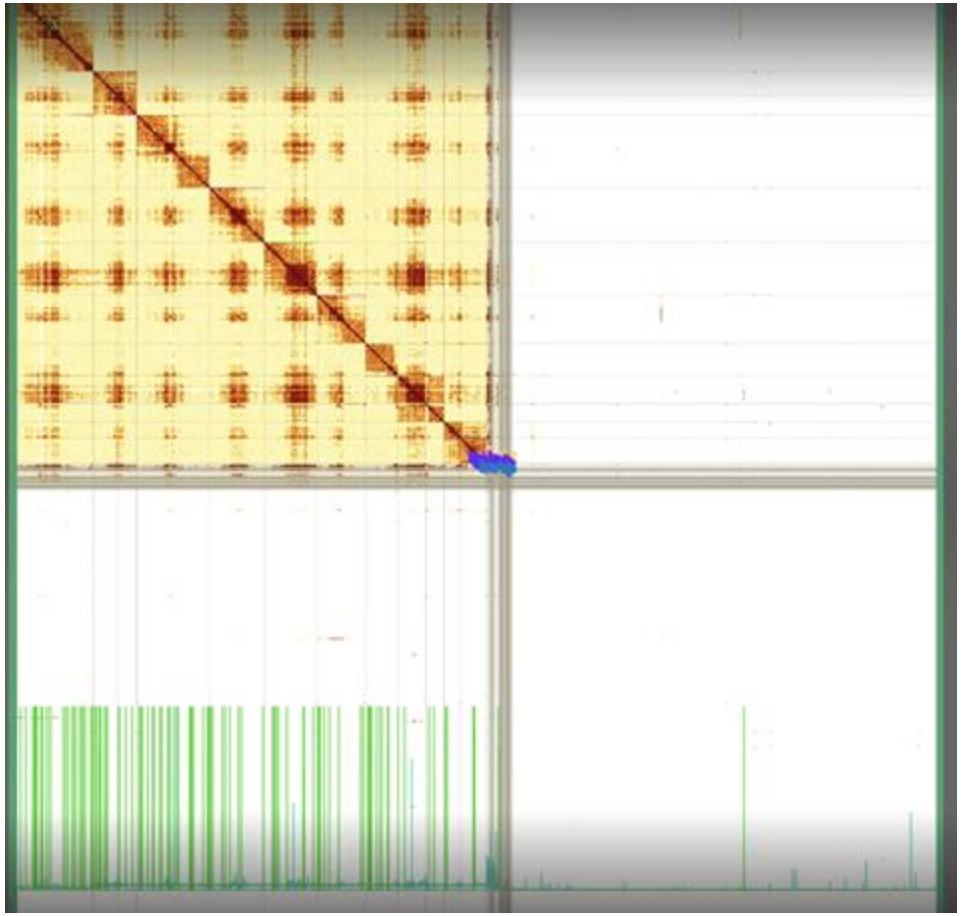
BUSCO hits and GC vs read coverage distribution



HiC contact map

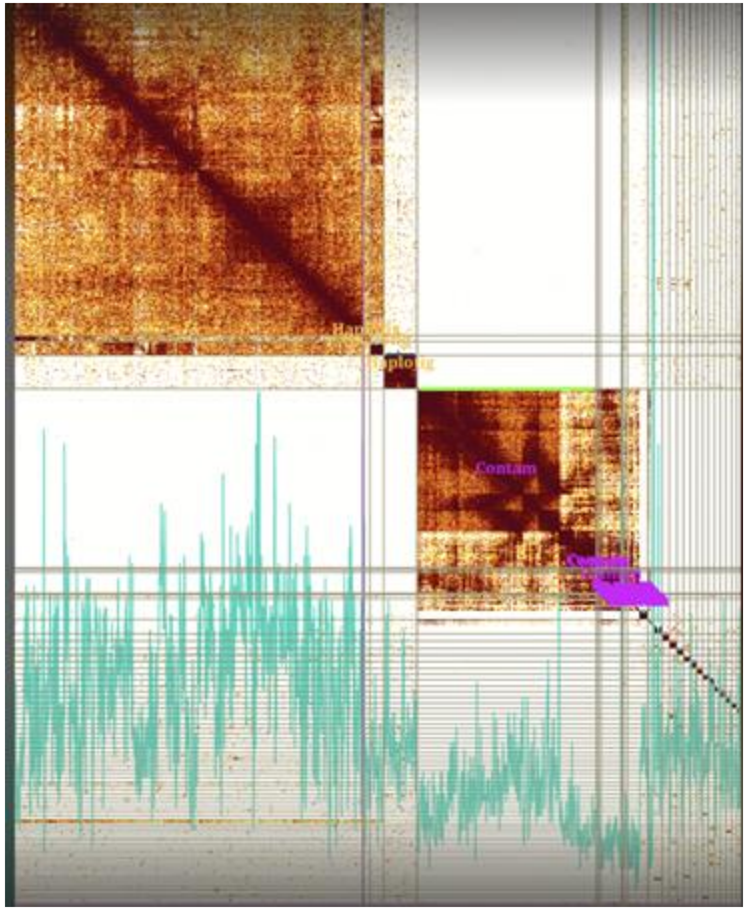
Decontamination examples

HiC - uncontaminated sample
Pacbio - contaminated sample



Diptera genome with fungi contamination

HiC and PacBio from same sample



Worm genome contaminated with bacteria

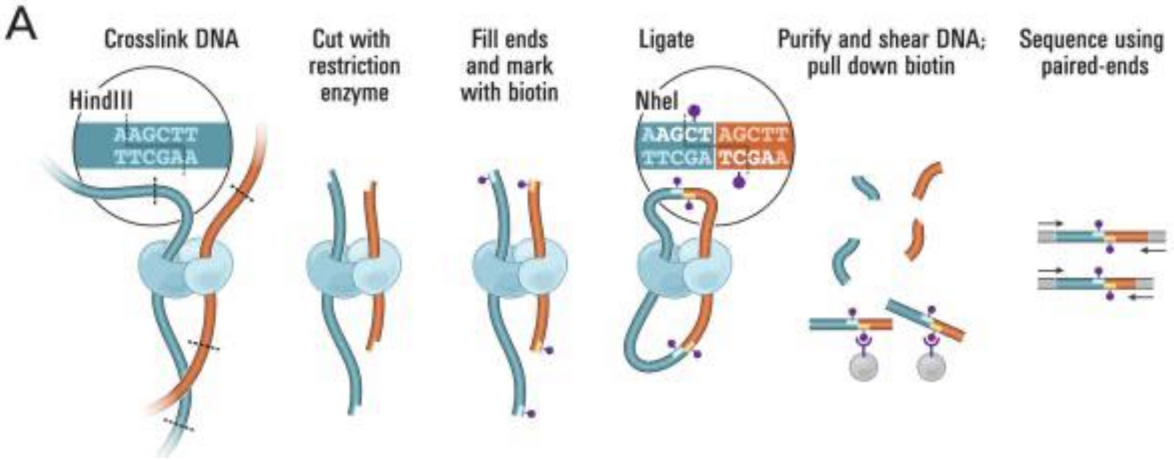
Why do we need curation?

"Curators are the gatekeepers for quality assembly submission"

- Sequence technology and assembly algorithms have come a very long way BUT….they're still far from perfect

- Typical issues:
  - order/orientation problems
  - chromosomes joined over telomeres
  - false duplications
  - genomic quirks – eg bird micro-chromosomes, large volume of repetitive sequence
- Improve assembly strategy and software

HiC data - our No. 1 curation resource

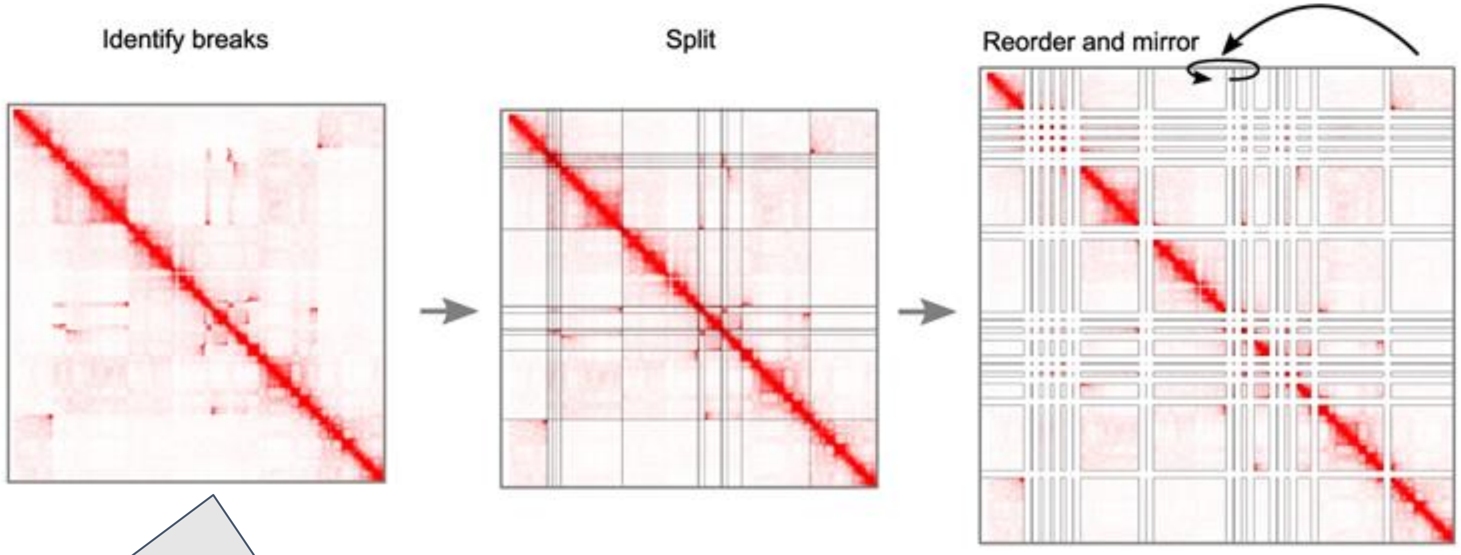

A
Crosslink DNA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends
HindIII
AAGCTT
TTCGAA
NheI
AAGCT AGCTT
TTCGA TCGAA

< 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

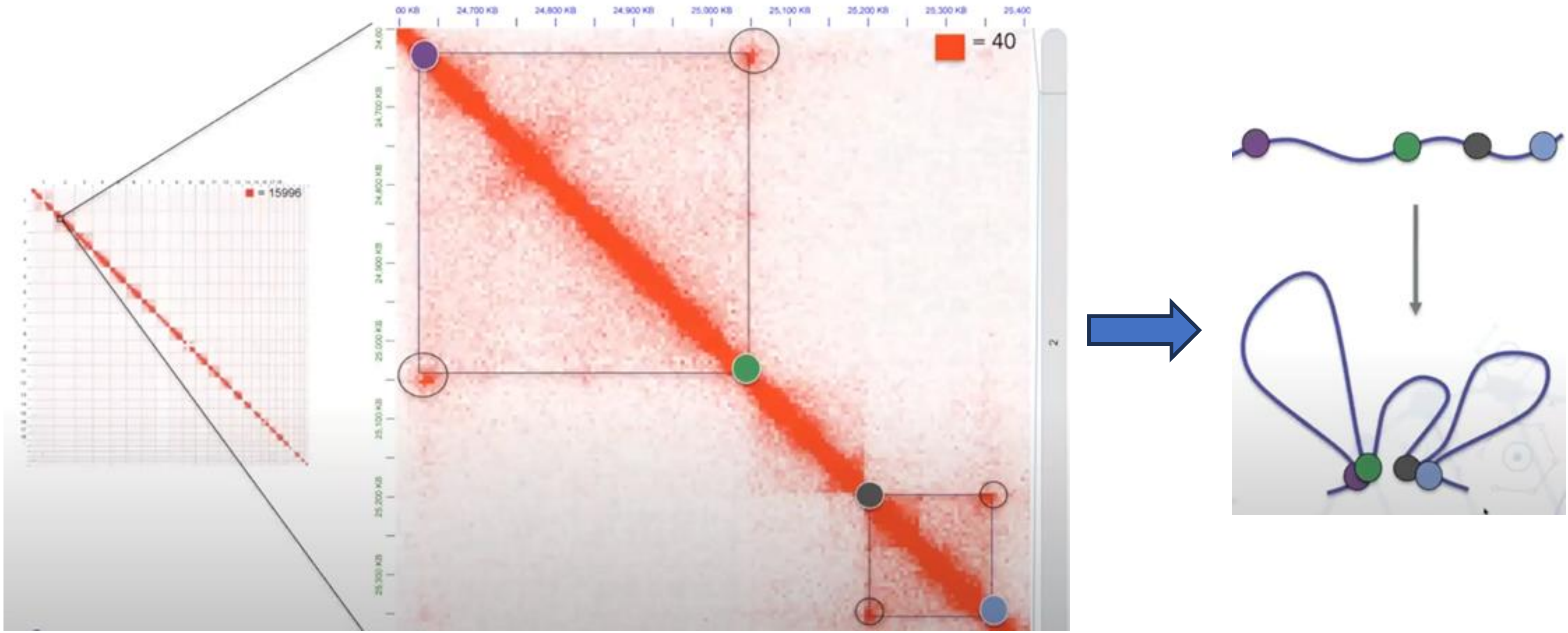
“in-situ” sequencing gives evidence of what sequence belongs next to what sequence.

The result is a contact map



HiC data - our No. 1 curation resource

Chromatin conformation with Hi-C

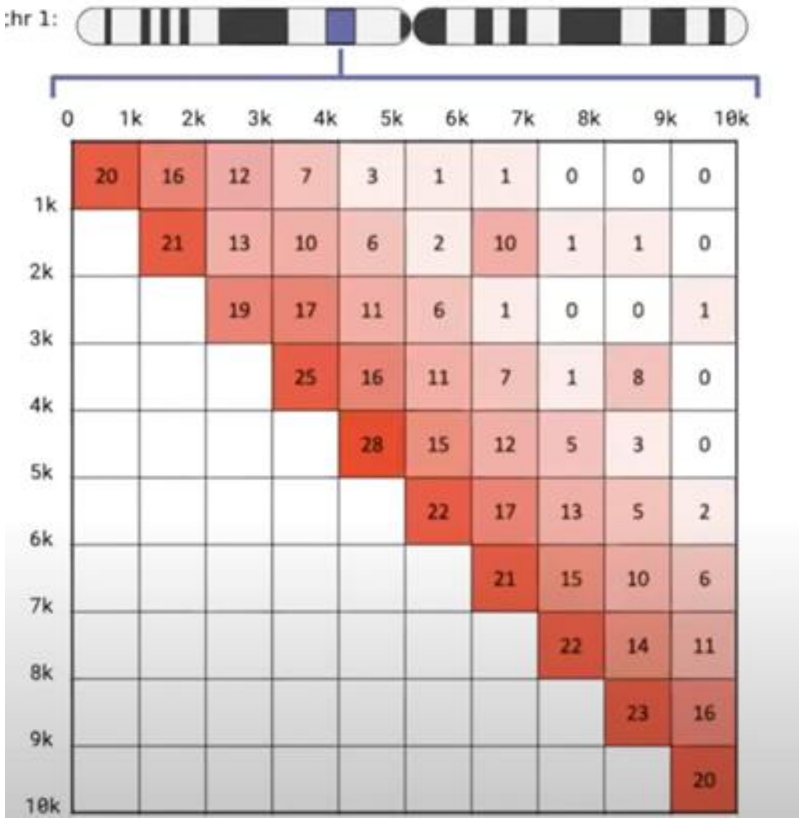


ARIMA GENOMICS

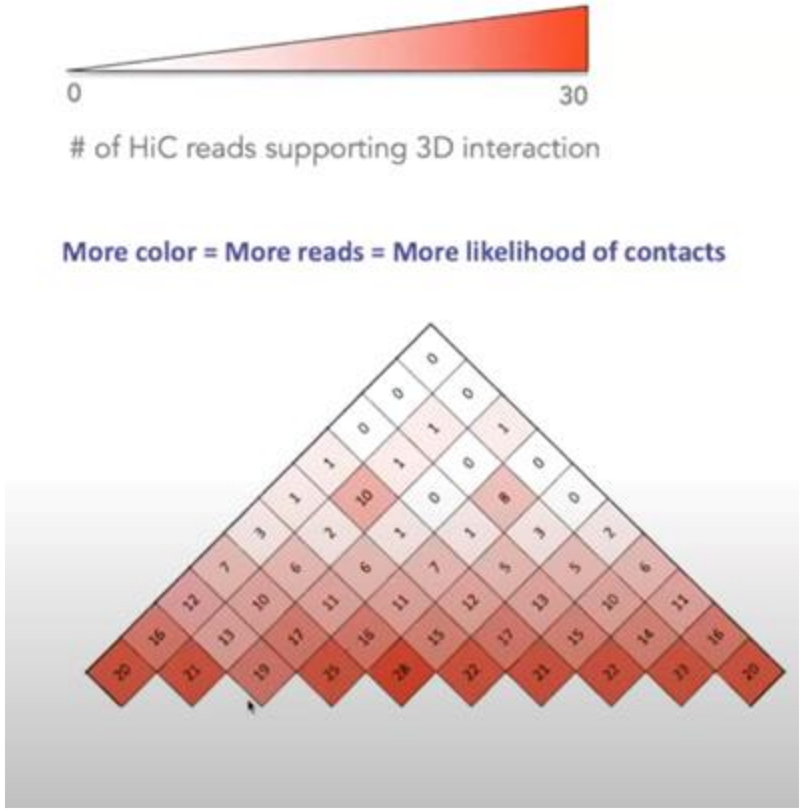
HiC data - our No. 1 curation resource

Visualization

Contact matrix colored based on hic reads counts

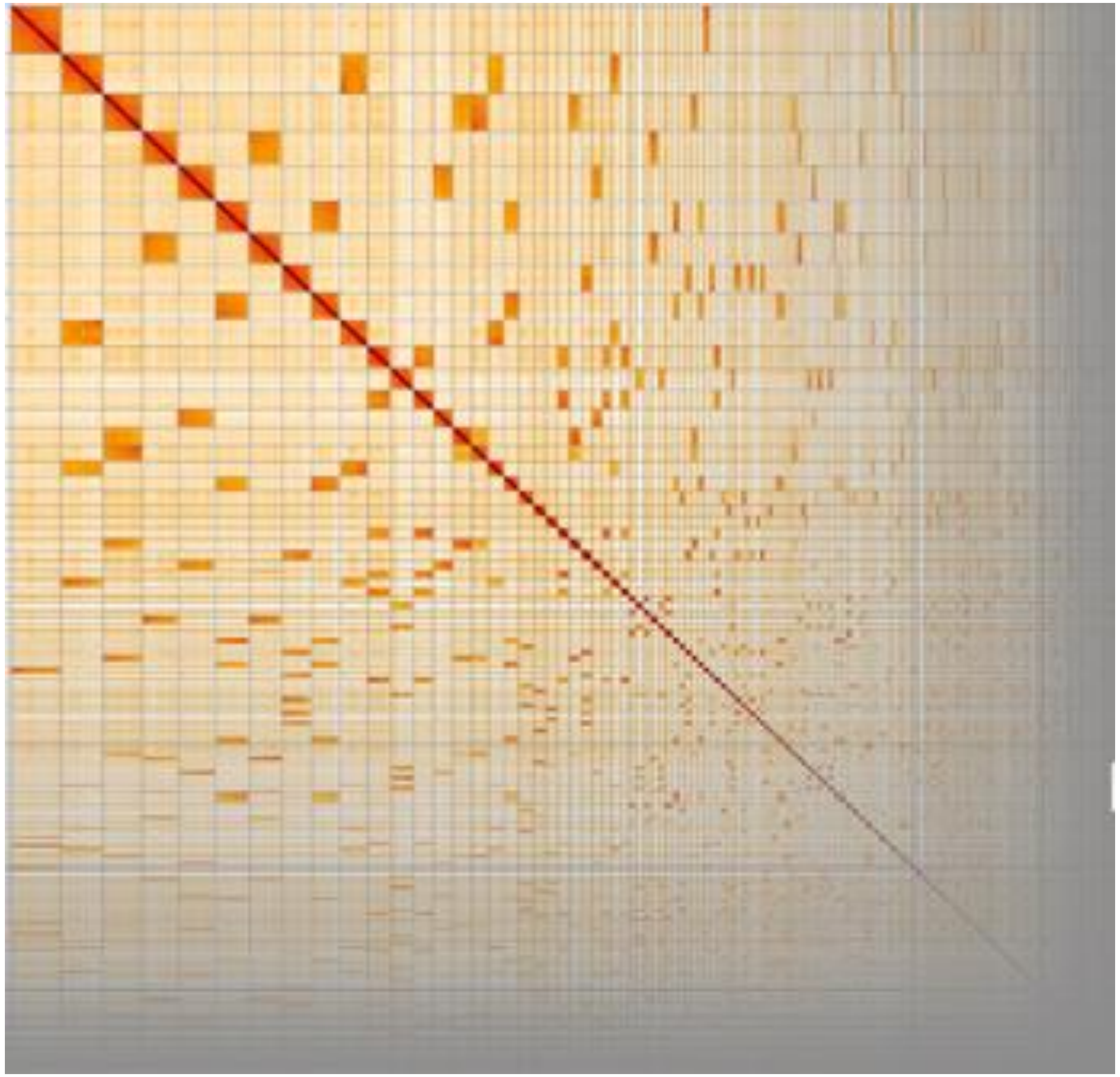

:hr 1:
0
1k
2k
3k
4k
5k
6k
7k
8k
9k
10k

20	16	12	7	3	1	1	0	0	0
	21	13	10	6	2	10	1	1	0
		19	17	11	6	1	0	0	1
			25	16	11	7	1	8	0
				28	15	12	5	3	0
					22	17	13	5	2
						21	15	10	6
							22	14	11
								23	16
									20



More color = More reads = More likelihood of contacts

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



Interpreting a HiC map

Centre diagonal show self matches, eg chr1 vs chr1

Off diagonal show relationship between different chromosomes/scaffolds (eg chr1 vs scaffold52).

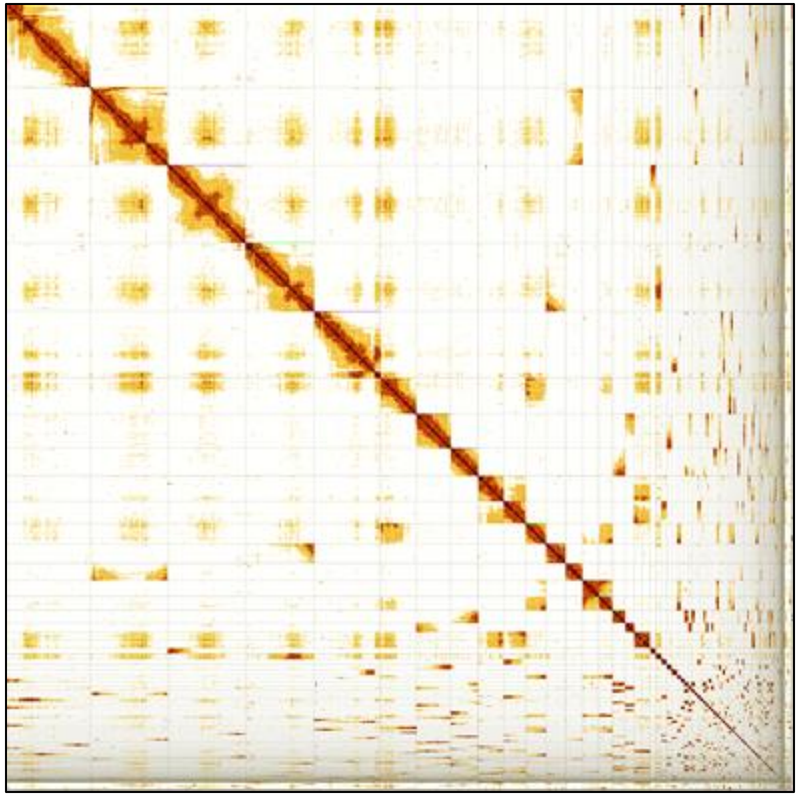
The darker the off-diagonal square, the stronger the relationship between the scaffolds.

Horizontal and vertical lines delineate chromosome/scaffold boundaries.

Evolution of a manually curated assembly

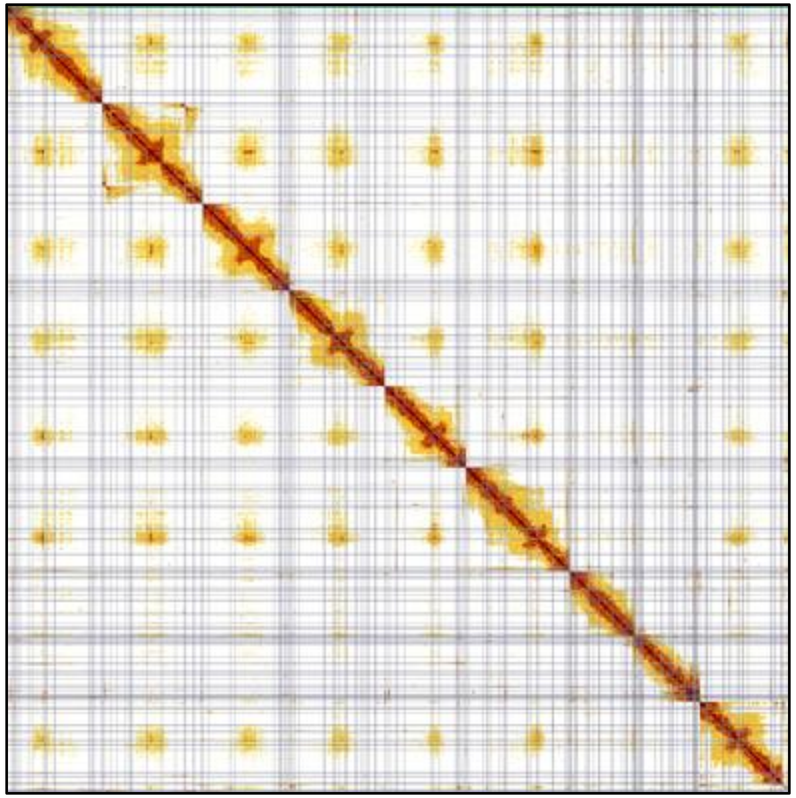
Patella pellucida
Blue-rayed limpet

n = 230
N50 = 33.1Mb



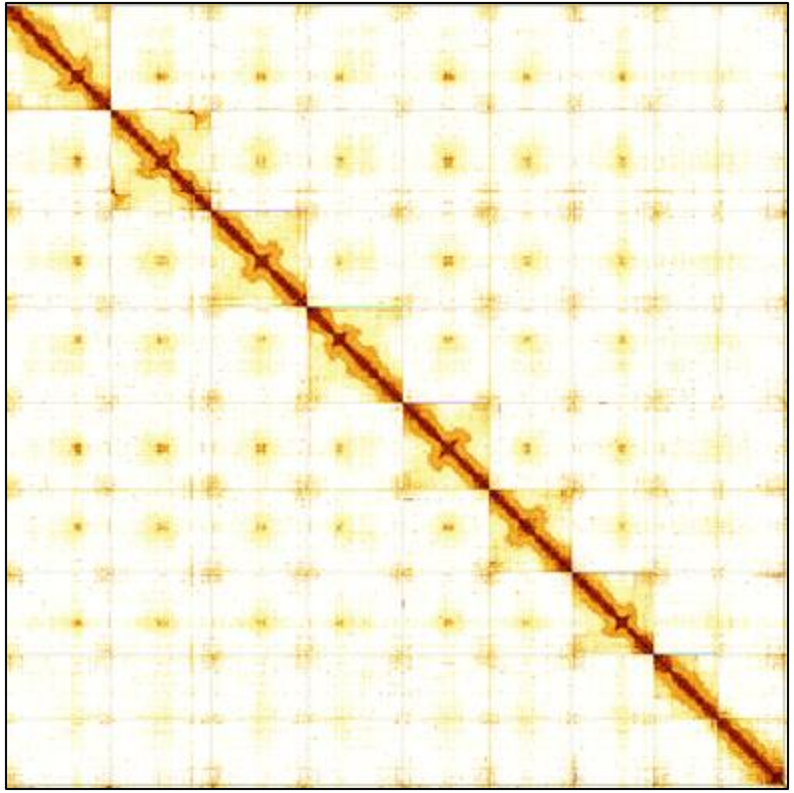
Pre curation assembly

225 joins
84 breaks
29 haplotype removals



after pretext manipulation

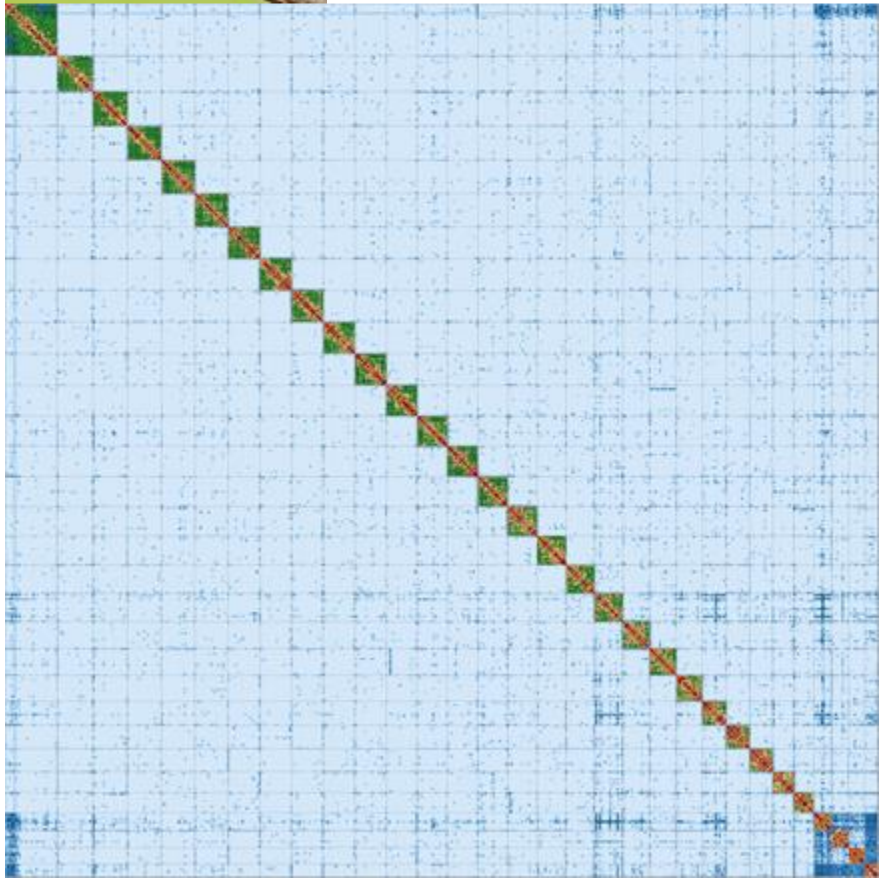
n = 62
N50 = 87.1Mb
99.85% of genome in 9 Chromosomes



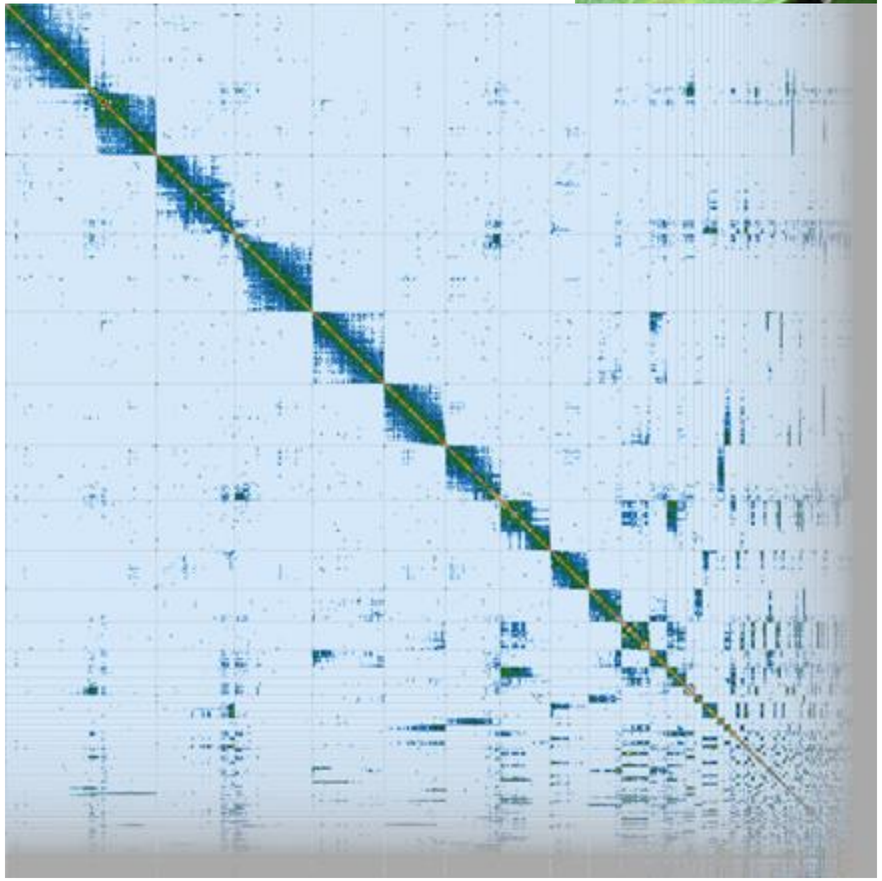
Post curation

Varying chromosome contiguity

Diachrysia chrysitis
n=31 (ZZ) 41 scaffolds



Sarcophaga variegate
n=6 (XY) 480 scaffolds

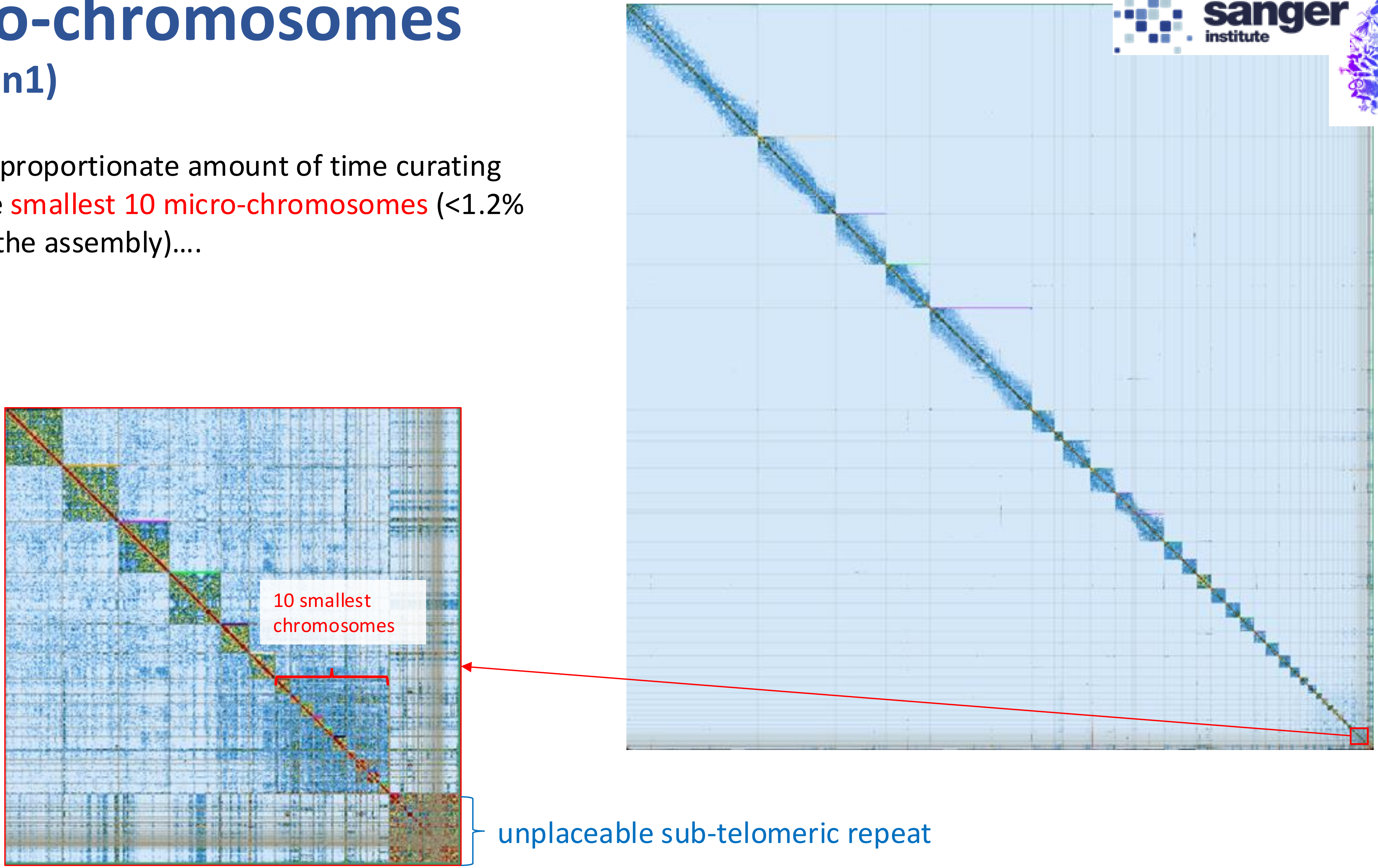


Assemblies from ToLA automated pipeline vary considerably in contiguity

Micro-chromosomes

(bCucCan1)

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



10 smallest chromosomes

unplaceable sub-telomeric repeat

Curation accessory tools

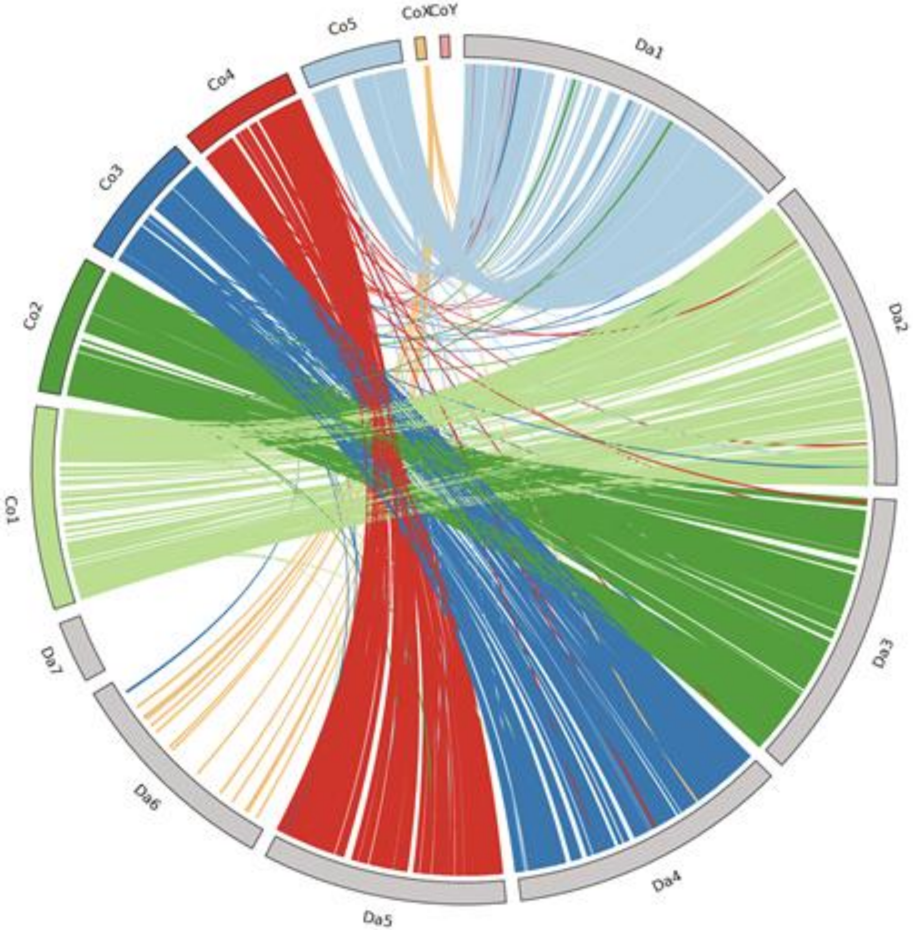
Synteny analysis

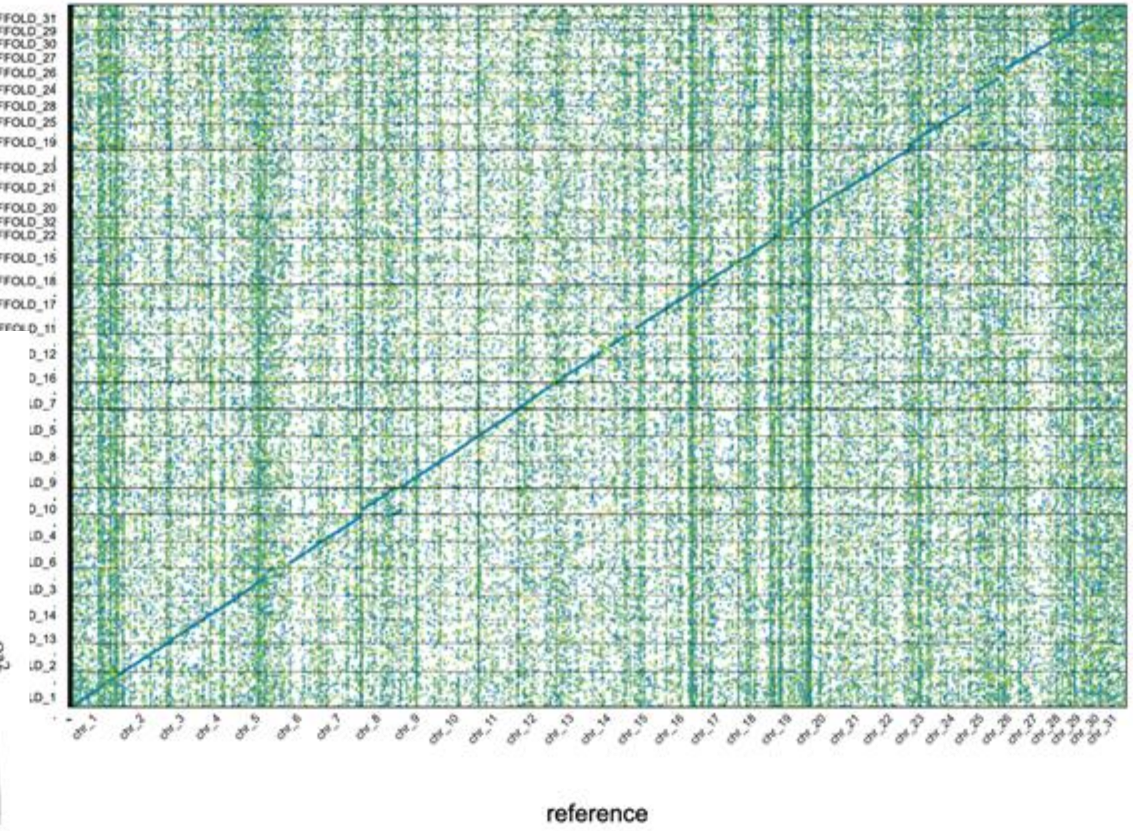
Alignment

(Nucmer)

BUSCO

(TreeVAL)





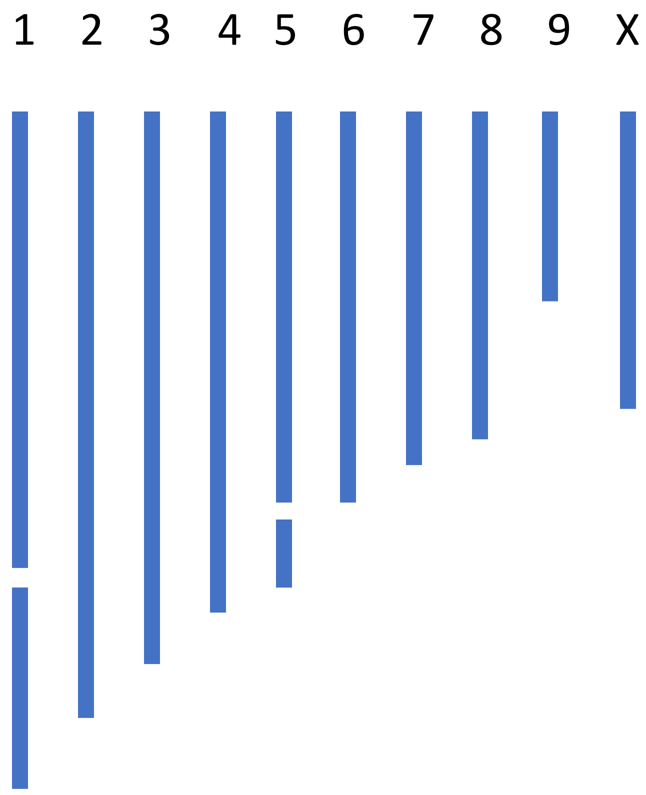
Chromosome naming

By size

- Autosomes large > small

By synteny

- Existing reference
- Genetic map
- Align close relative with sound chromosome naming

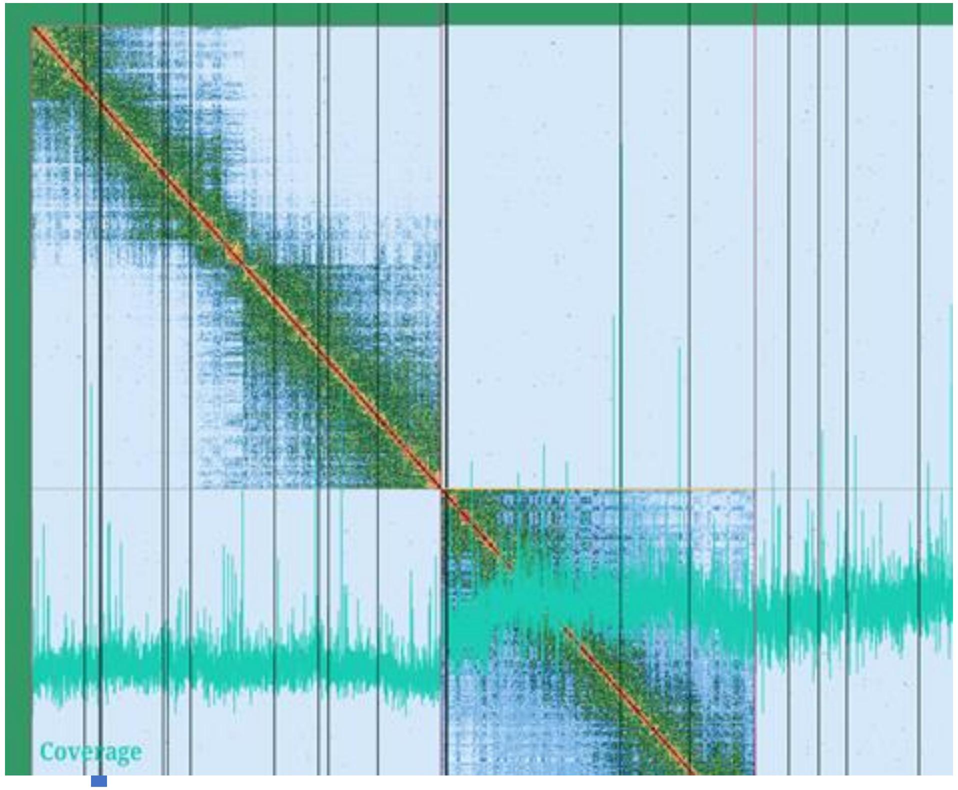


Sex chromosome identification

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)

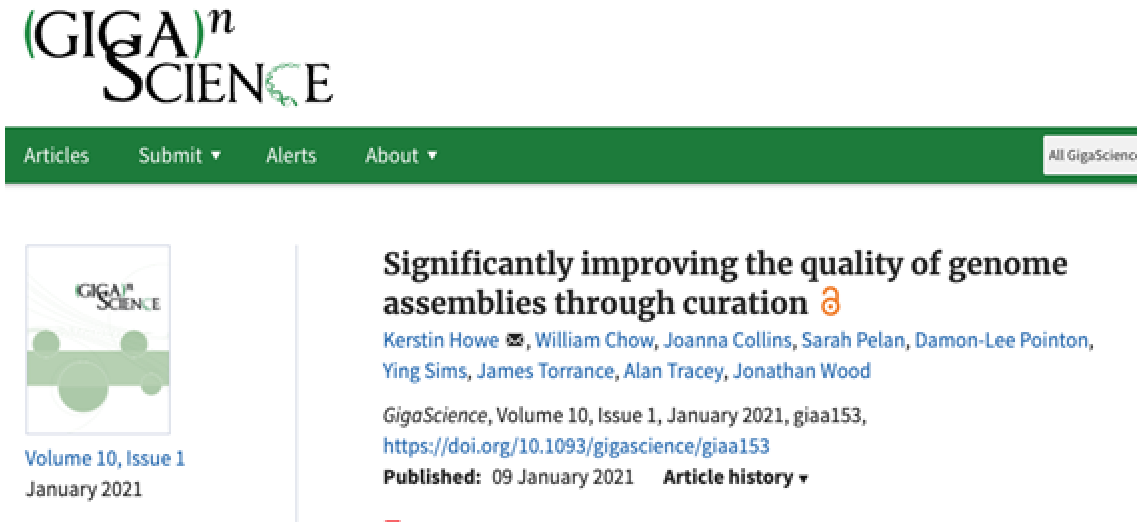
Rapid Curation (distributed)

- Rapid curation tools:
  - https://github.com/sanger-tol/rapid-curation
- Singularity Hi-C maps (PretextView)
  - Feature tracks
- Scripts for manipulating fasta files
  - rapid_split
  - pretext-to-tpf
  - multi_join.py / rapid_join
- Documentation, tutorials, slack channel support:
  - https://assemblycuration.slack.com/

Genome Reference Informatics Team

GRIT Curators:

- Jo Wood
- Jo Collins
- Sarah Pelan
- **Michael Paulini**
- Dominic Absolon
- Tom Mathers
- **Camilla Santos**
- Karen Brooks



Volume 10, Issue 1
January 2021

Significantly improving the quality of genome assemblies through curation

Kerstin Howe, William Chow, Joanna Collins, Sarah Pelan, Damon-Lee Pointon, Ying Sims, James Torrance, Alan Tracey, Jonathan Wood

GigaScience, Volume 10, Issue 1, January 2021, giaa153, https://doi.org/10.1093/gigascience/giaa153

Published: 09 January 2021 **Article history ▾**

https://doi.org/10.1093/gigascience/giaa153

Resources

- https://github.com/sanger-tol/rapid-curation
- Singularity: Hi-C maps and feature creation pipeline

- https://assemblycuration.slack.com
- grit@sanger.ac.uk

- grit@sanger.ac.uk (GRIT team)