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

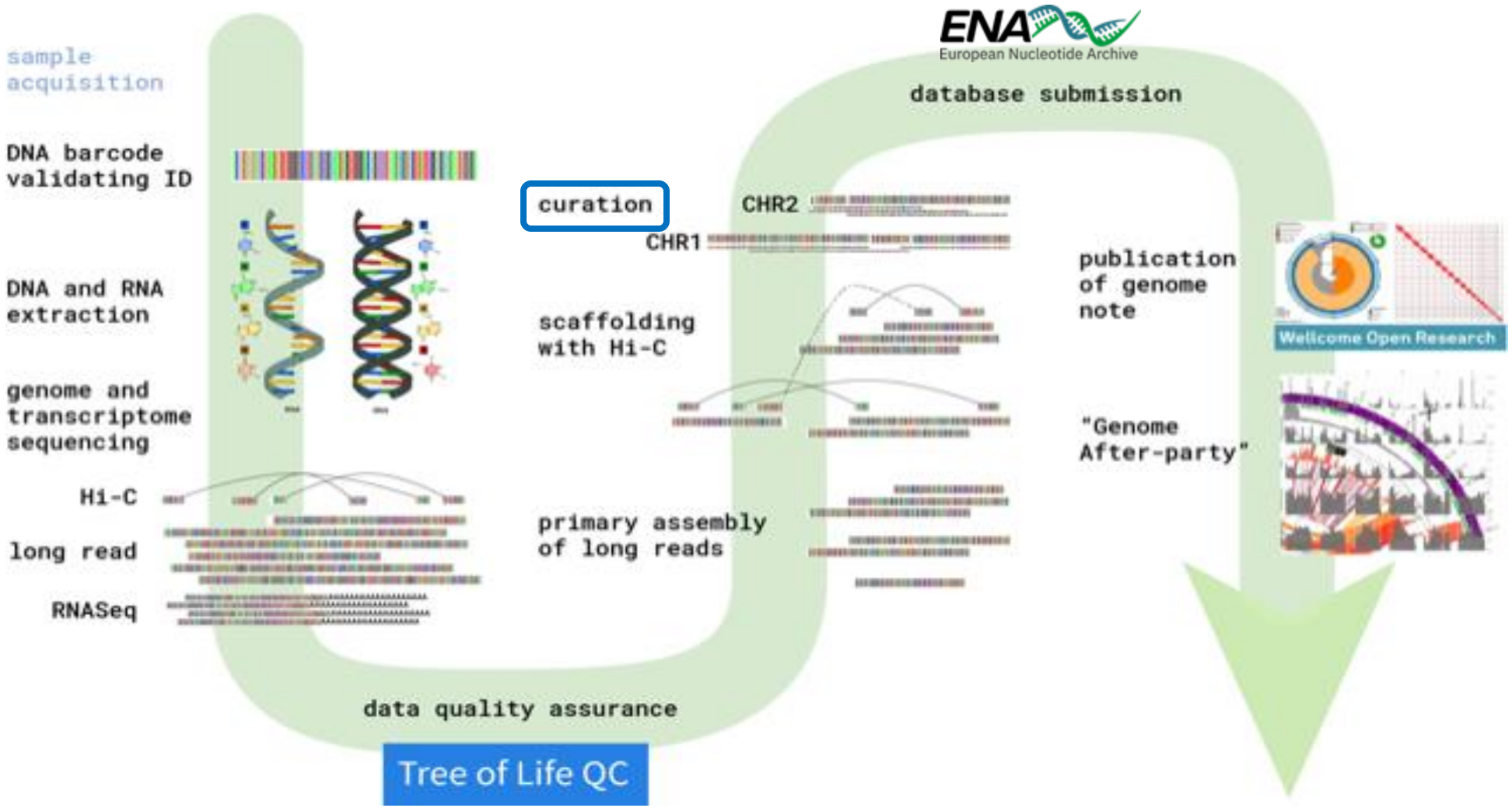
What is genome curation?

"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**"

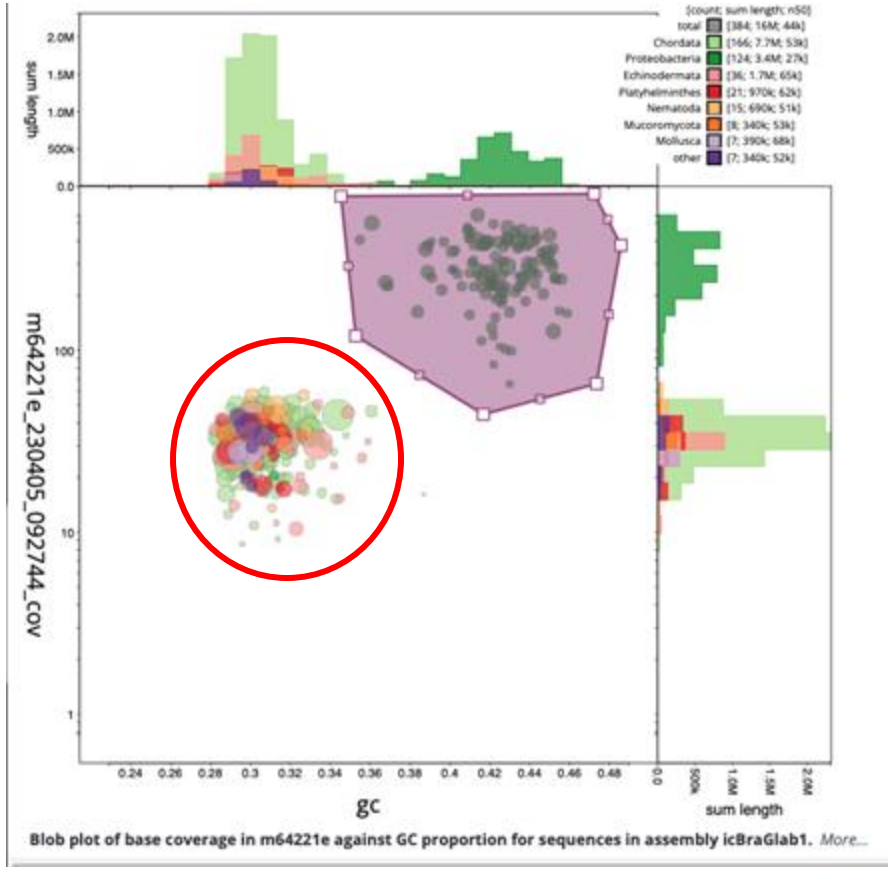
Our experience:

- 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

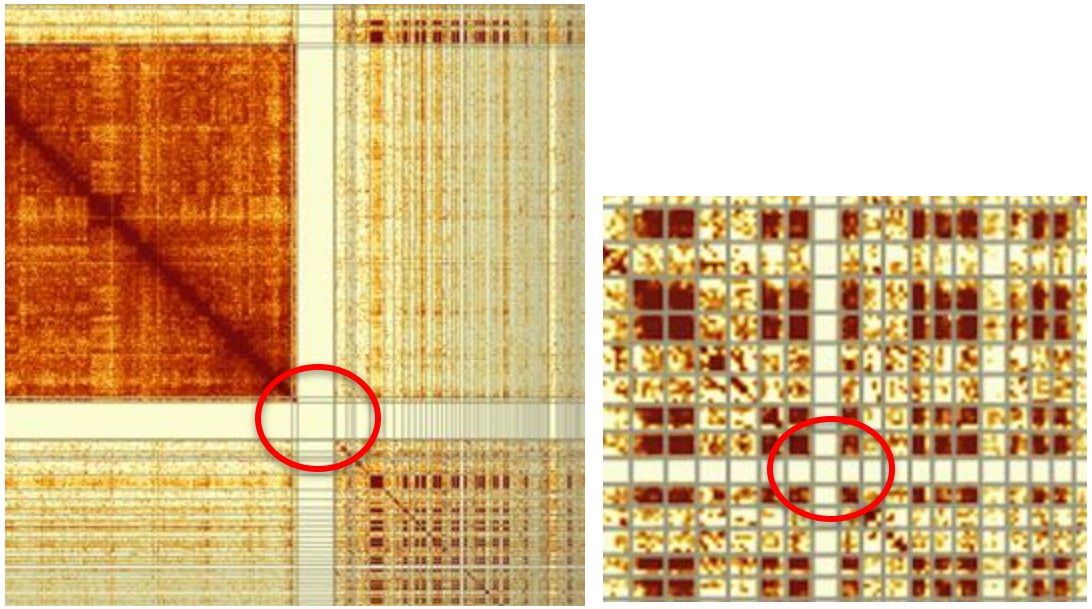
The Tree of Life genome factory



Decontamination examples



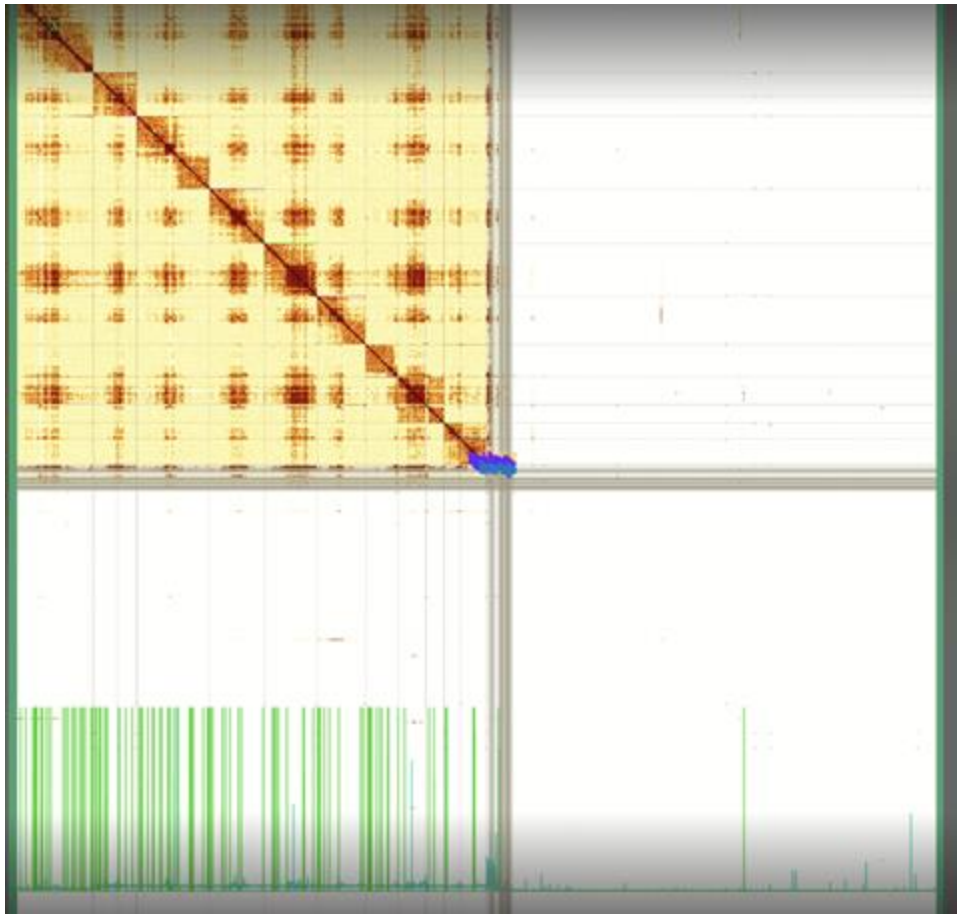
BUSCO hits and GC vs read coverage distribution



HiC contact map

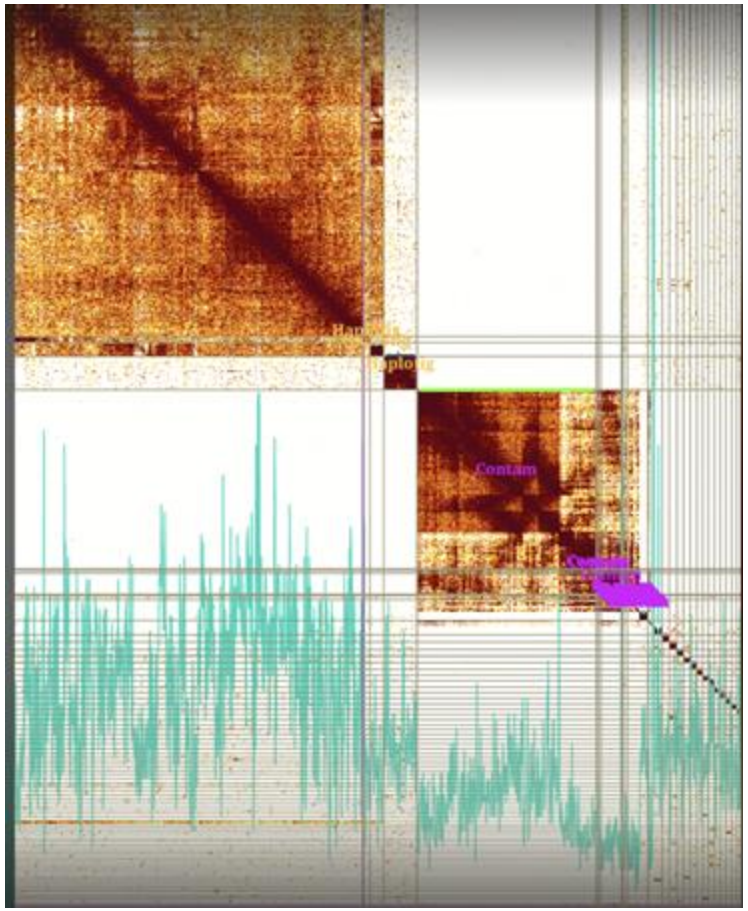
Decontamination examples

HiC - uncontaminated sample
Pacbio - contaminated sample



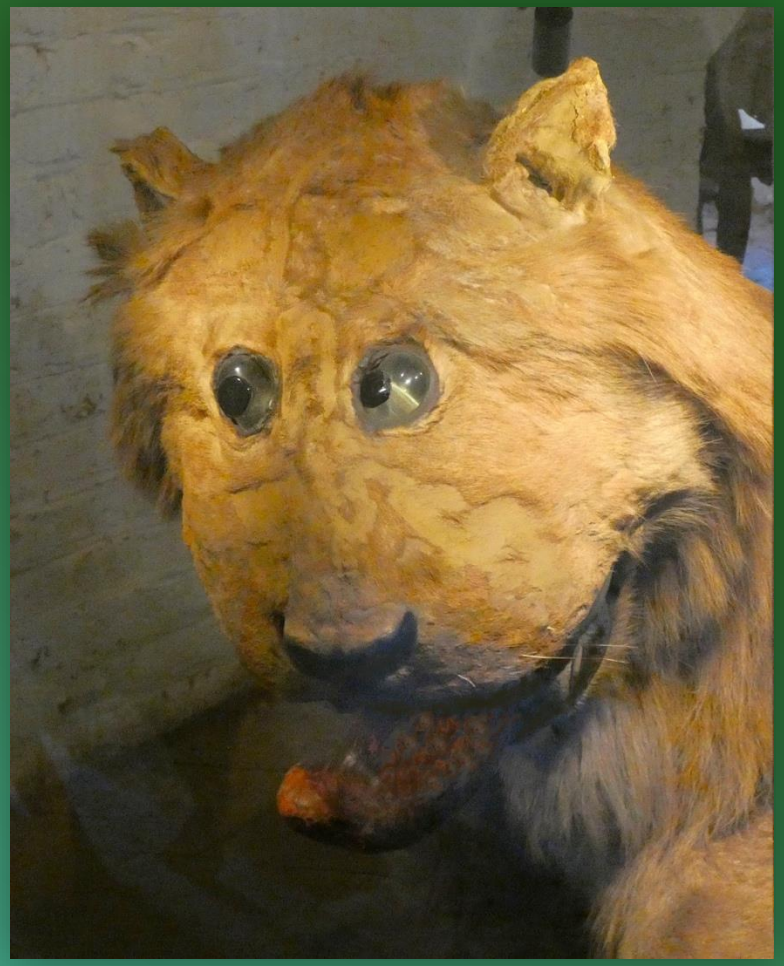
Diptera genome with fungi contamination

HiC and PacBio from same sample



Worm genome contaminated with bacteria

Why do we need curation?



Why do we need curation?

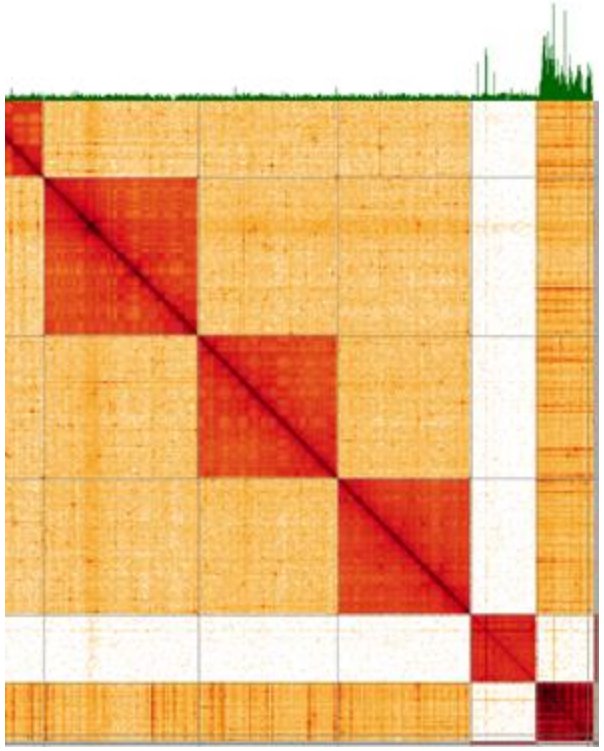
"Curators are the gatekeepers for quality assembly submission"

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

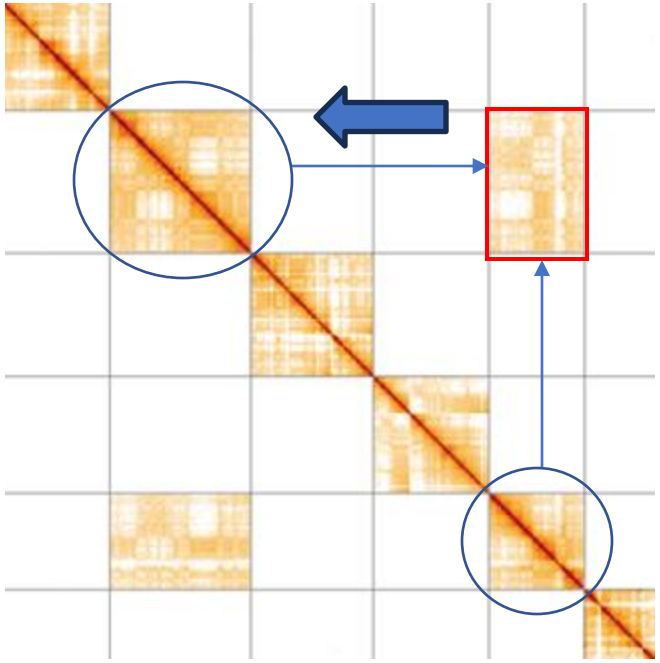
- Typical issues:
  - order/orientation problems
  - chromosomes joined over telomeres
  - false duplications
  - genomic quirks – eg bird micro-chromosomes, large volume of repetitive sequence
- Improve assembly strategy and software

Main issues during curation

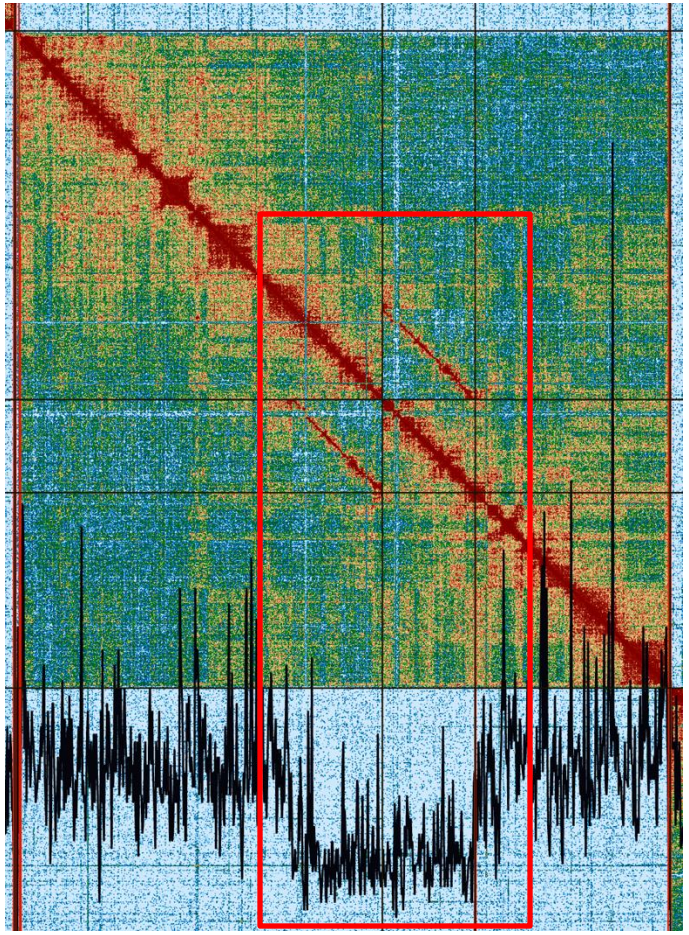
Contamination



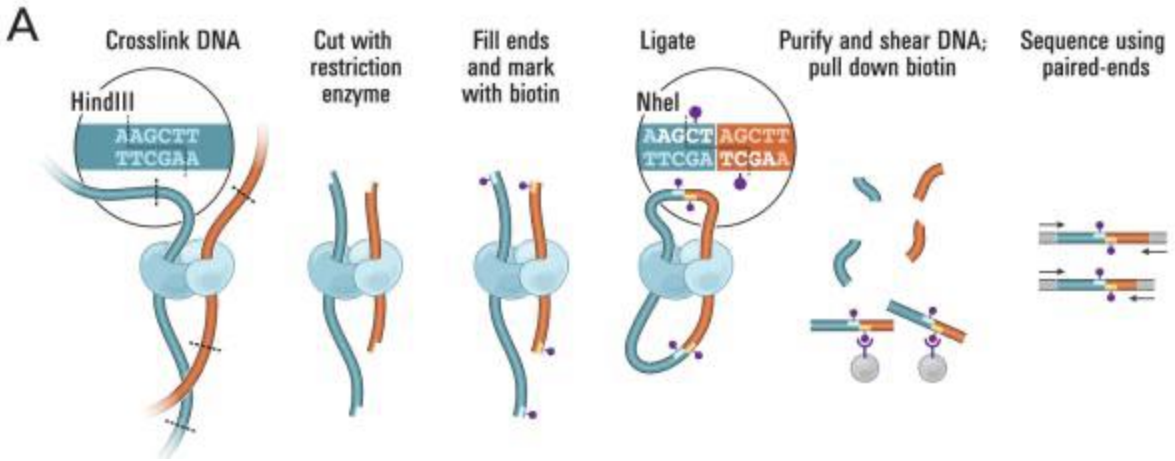
Misassemblies



Haplotigs



HiC data - our No. 1 curation resource

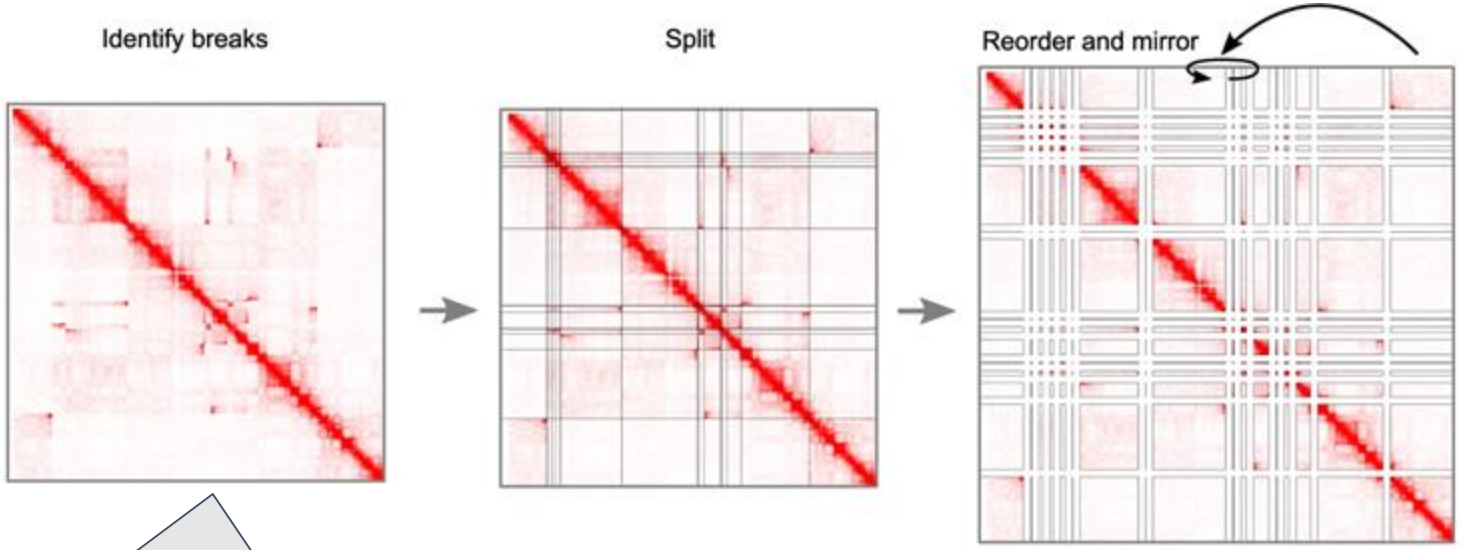


< Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles 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

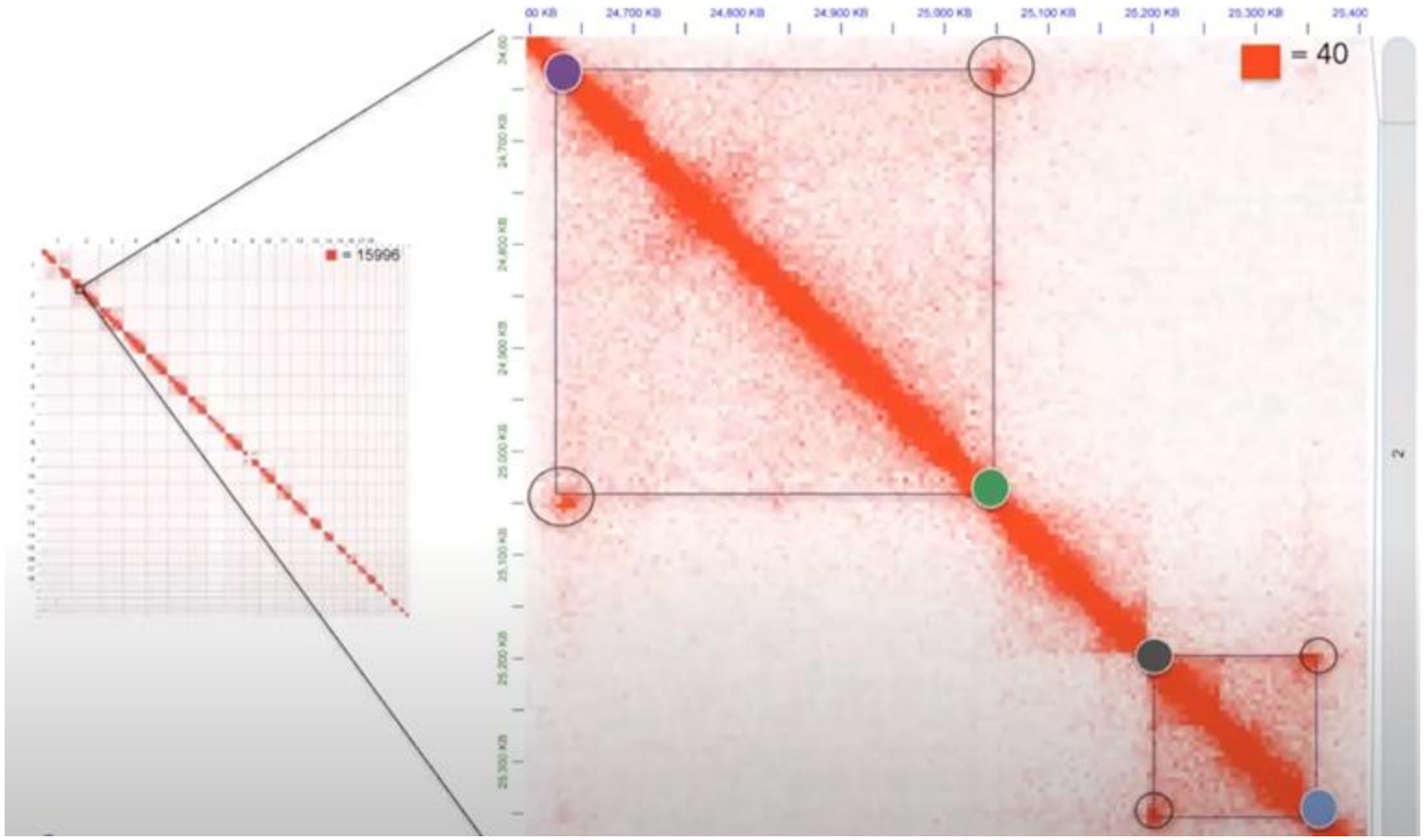
"in-situ" sequencing gives evidence of what sequence belongs next to what sequence.

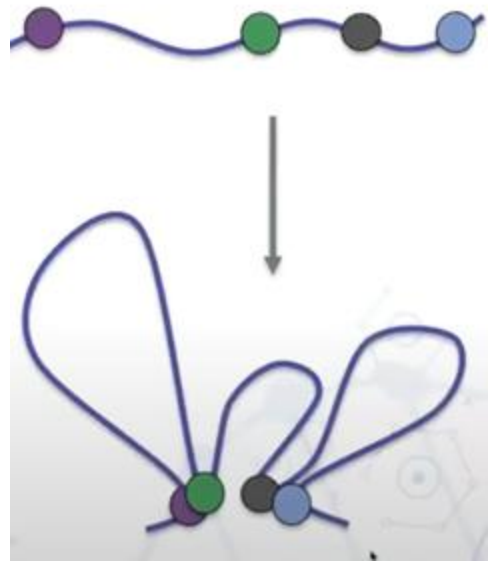
The result is a contact map



HiC data - our No. 1 curation resource

Chromatin conformation with Hi-C

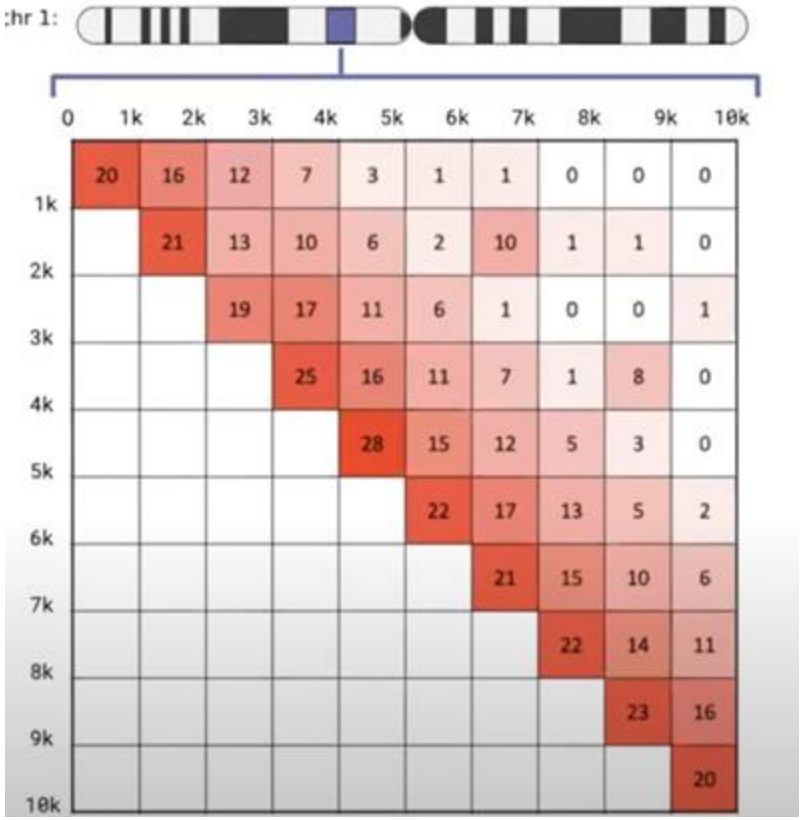




HiC data - our No. 1 curation resource

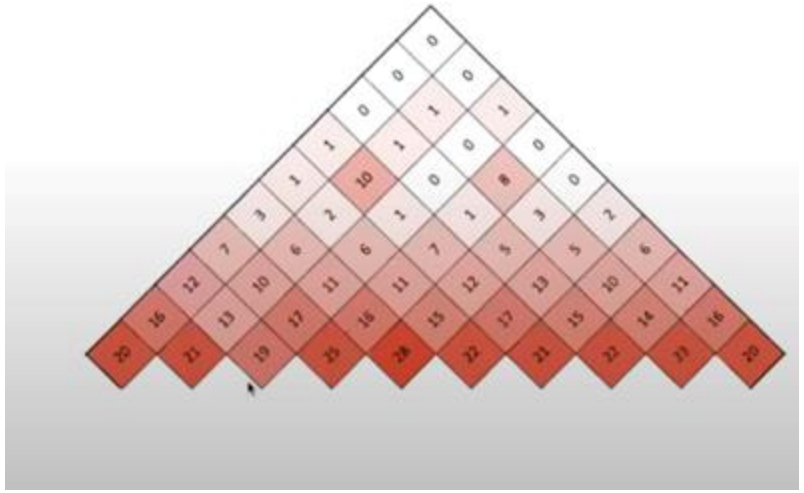
Visualization

Contact matrix colored based on hic reads counts



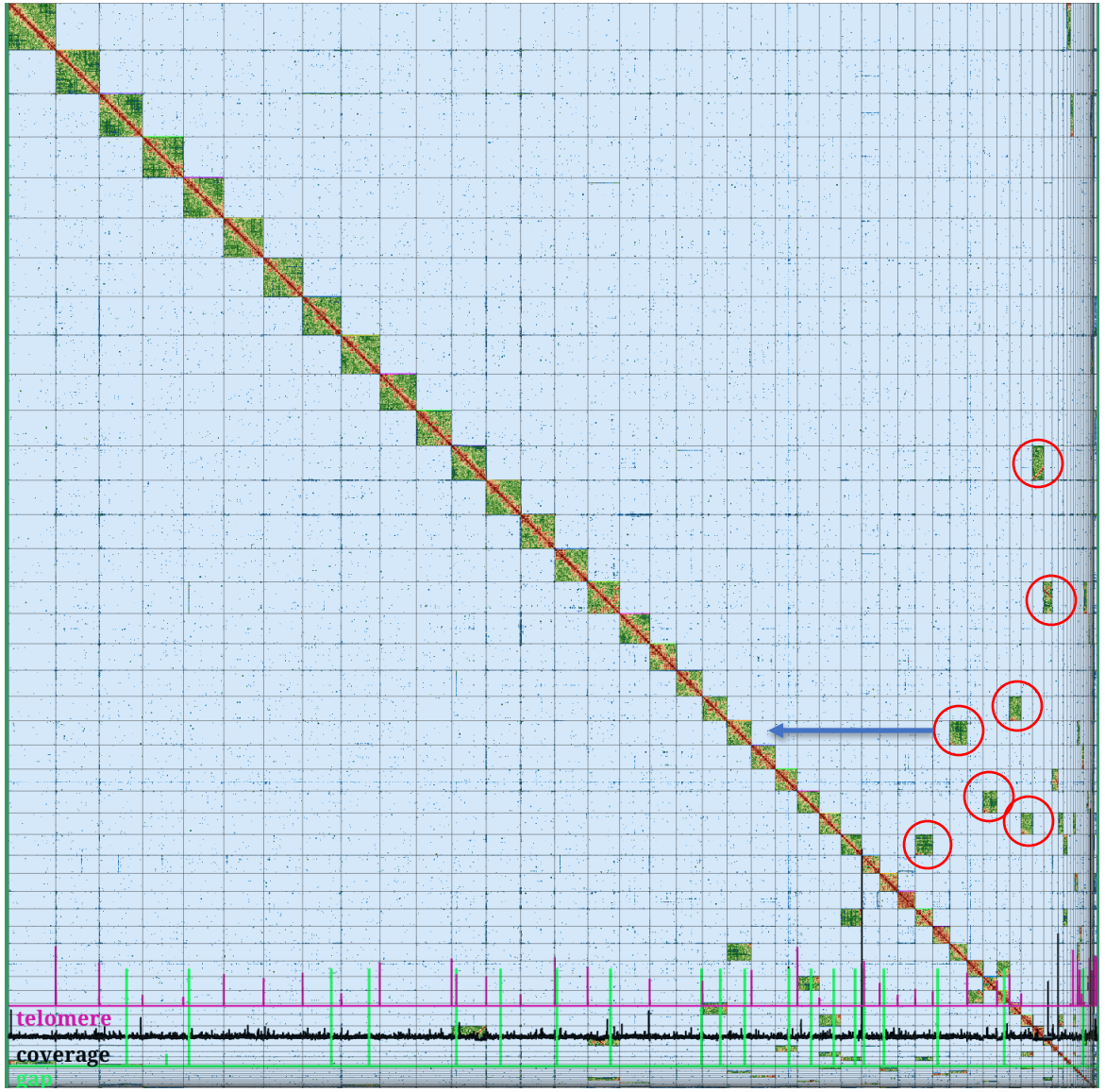
# of HiC reads supporting 3D interaction

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

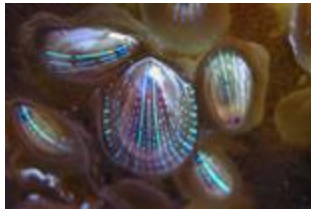
Diagonal mirrors itself

Off diagonal show relationship between different chromosomes/scaffolds.

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

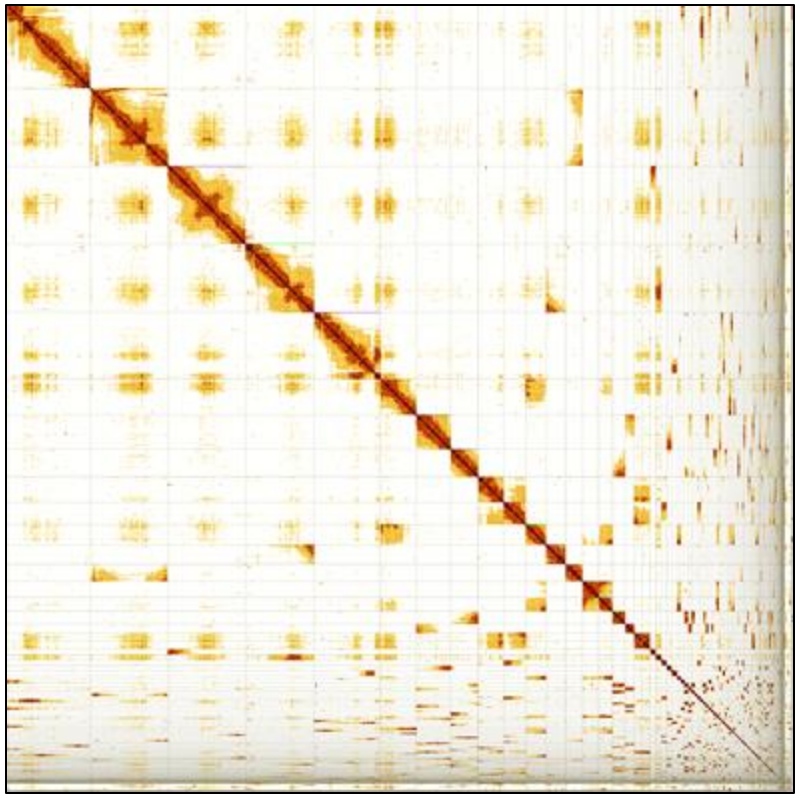
Horizontal and vertical lines delineate chromosome/scaffold boundaries.

Evolution of a manually curated assembly



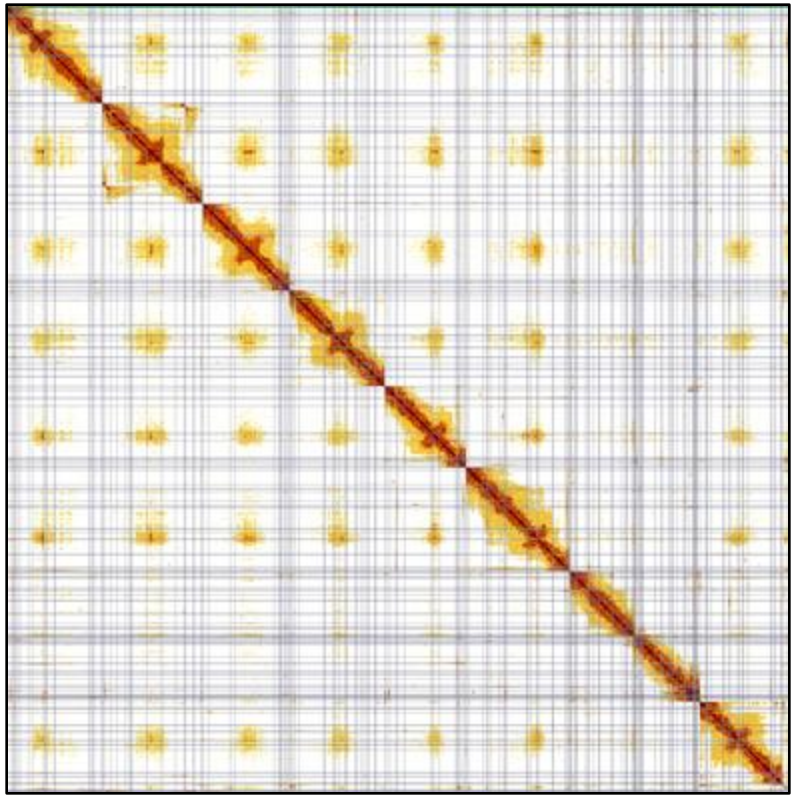
Patella pellucida
Blue-rayed limpet

n = 230
N50 = 33.1Mb



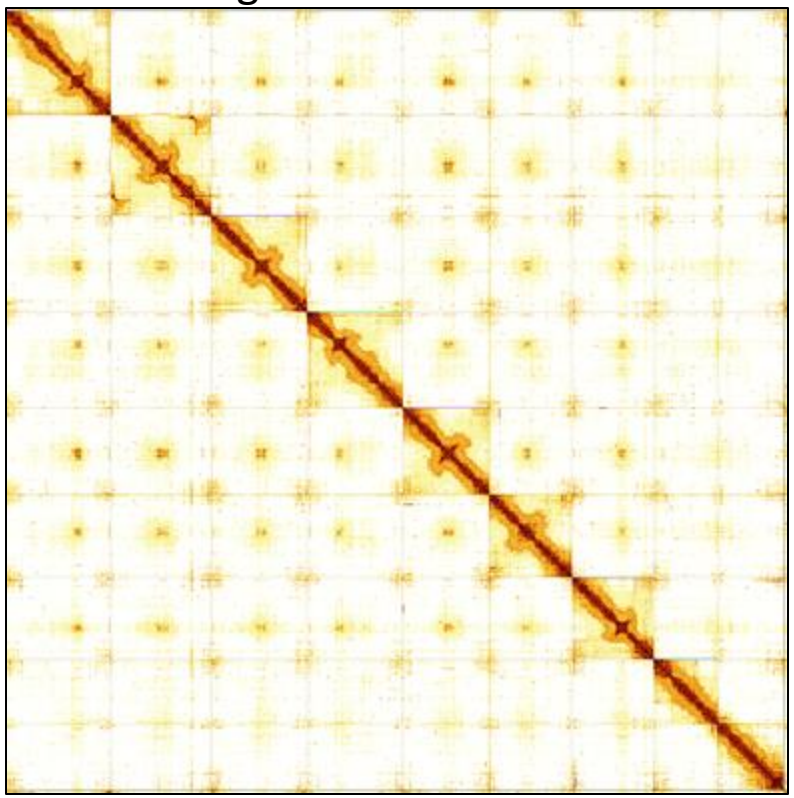
Pre curation assembly

225 joins
84 breaks
29 haplotype removals



after pretext manipulation

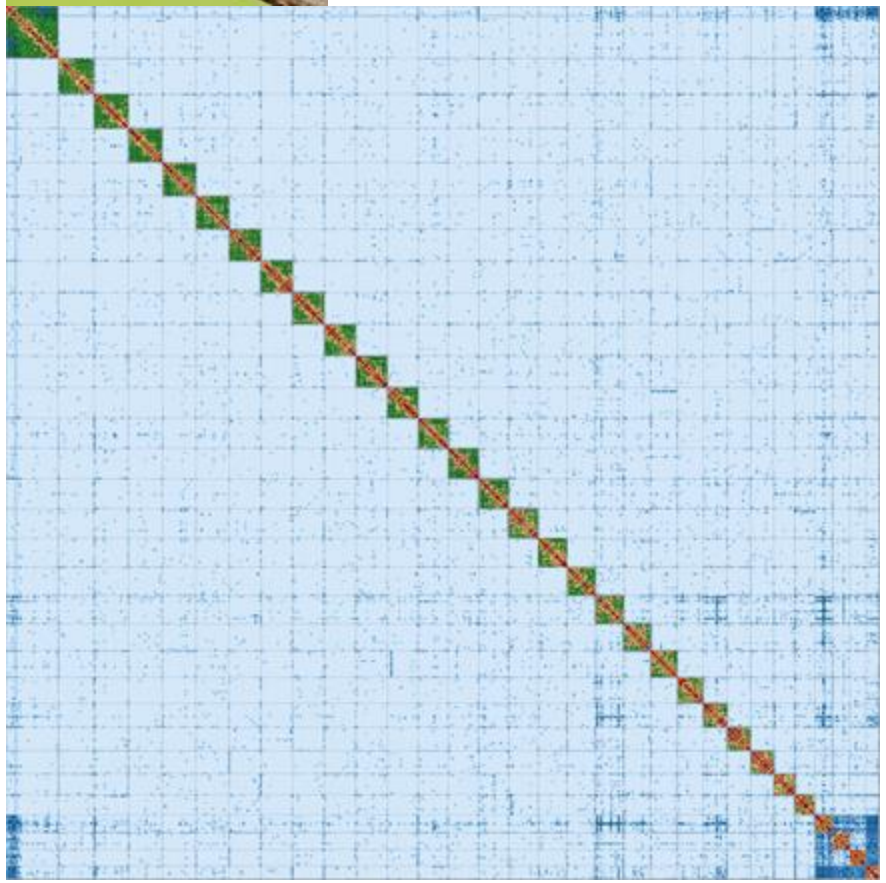
n = 62
N50 = 87.1Mb
99.85% of genome in 9 Chromosomes



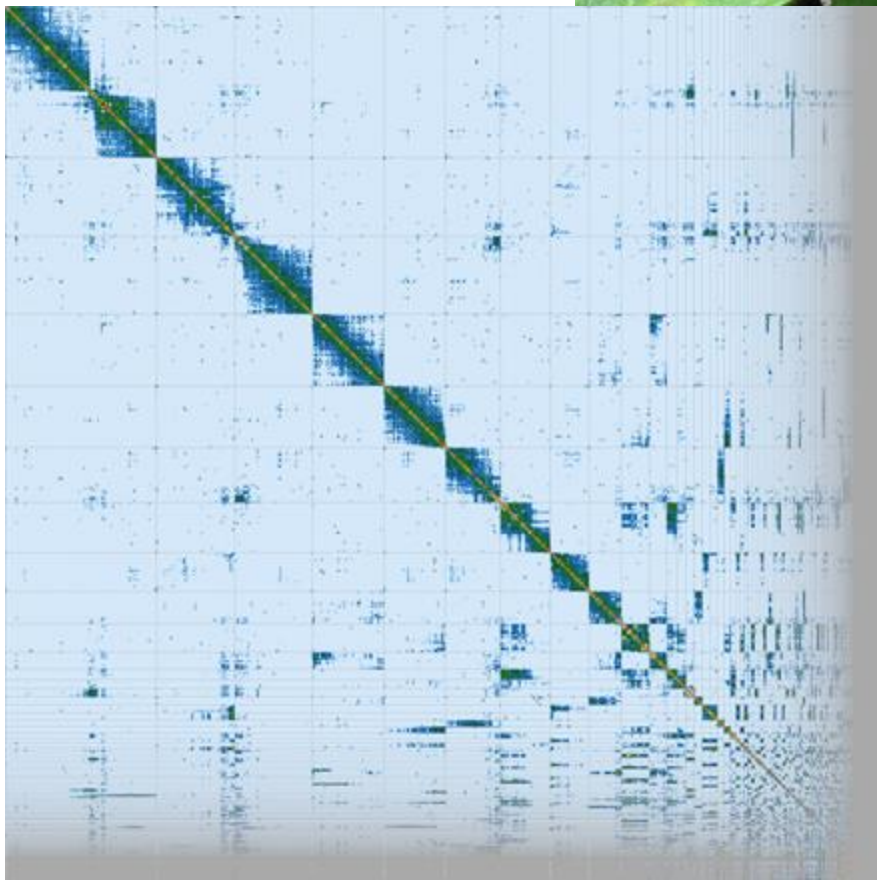
Post curation

Varying chromosome contiguity

Diachrysia chrysitis
n=31 (ZZ) 41 scaffolds



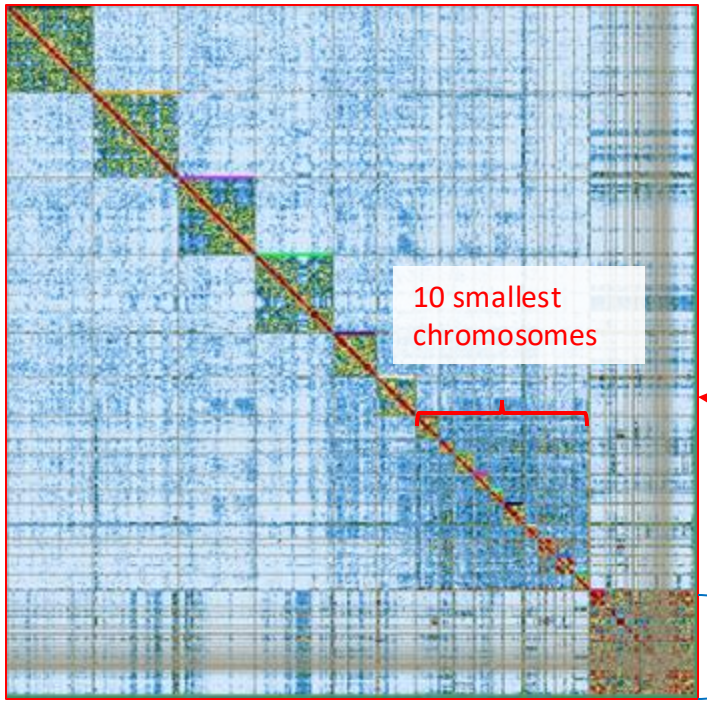
Sarcophaga variegate
n=6 (XY) 480 scaffolds

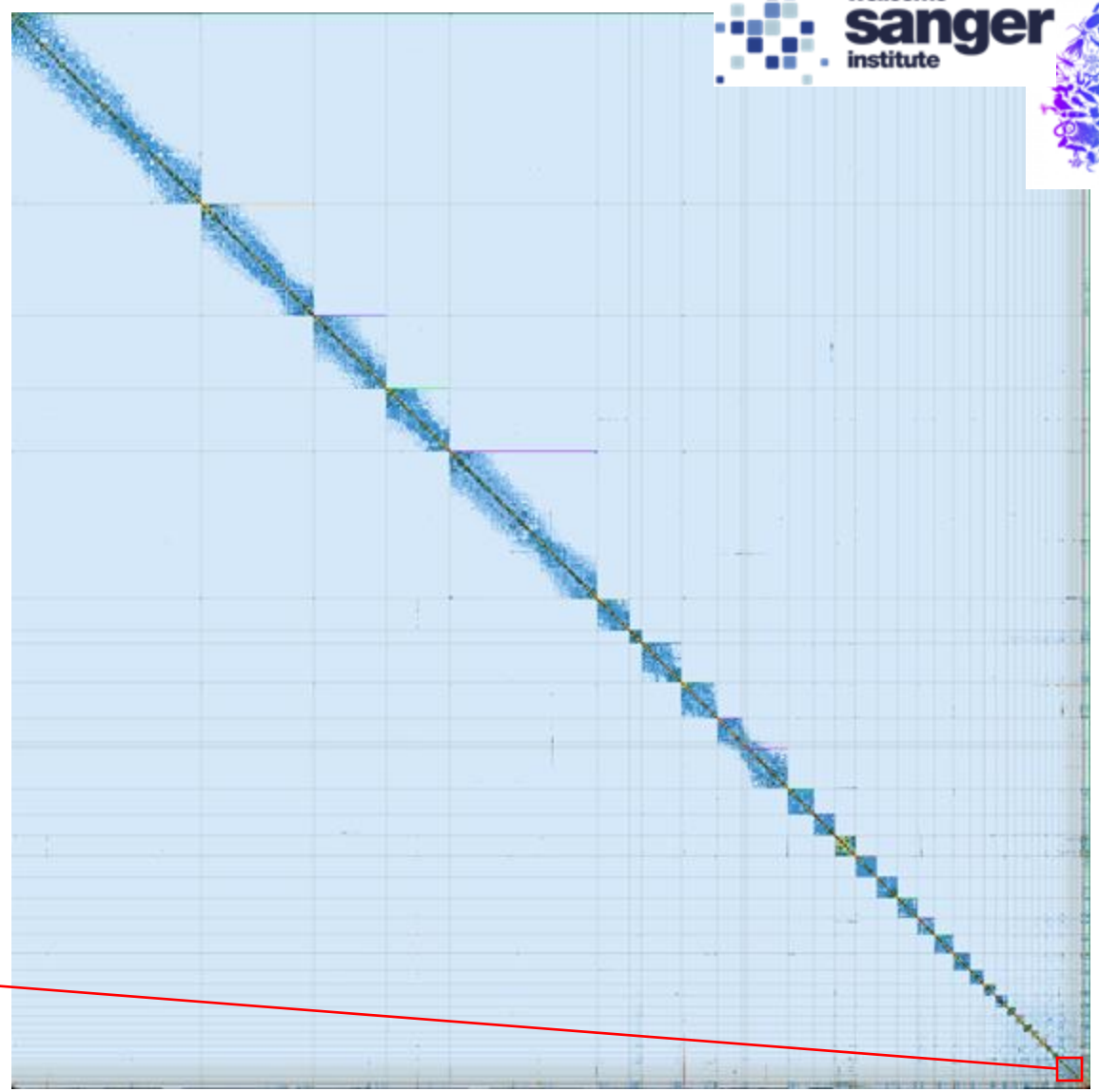


Assemblies from ToLA automated pipeline vary considerably in contiguity

Micro-chromosomes
(bCucCan1)

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

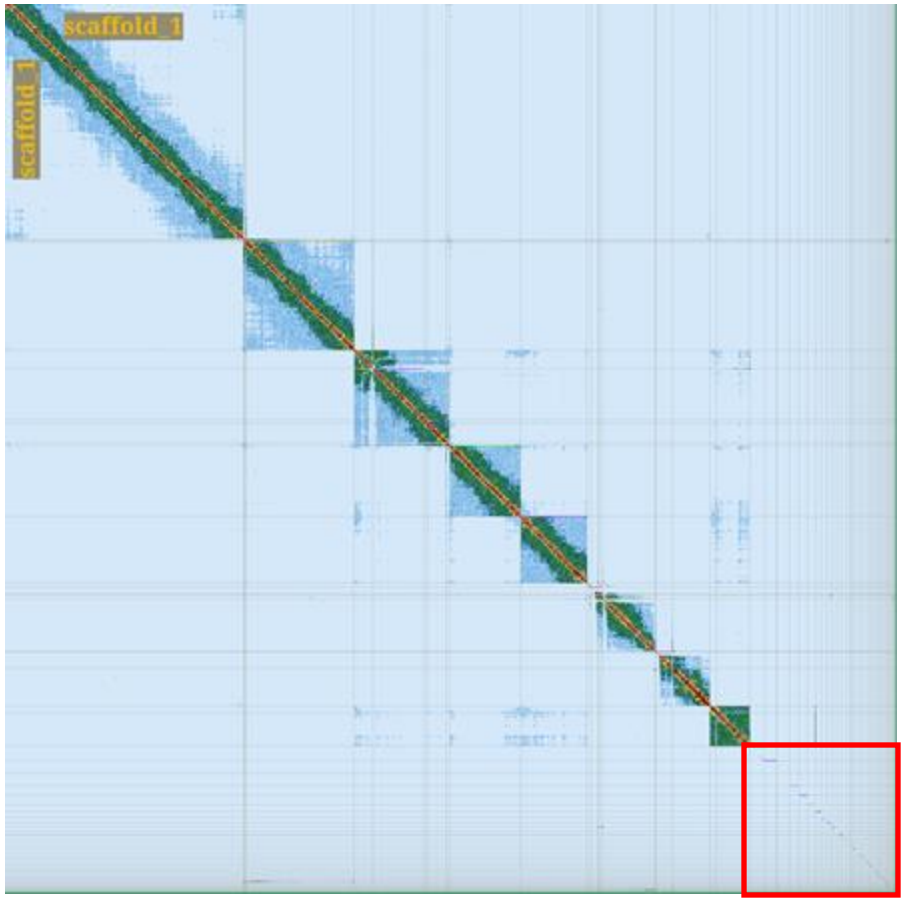




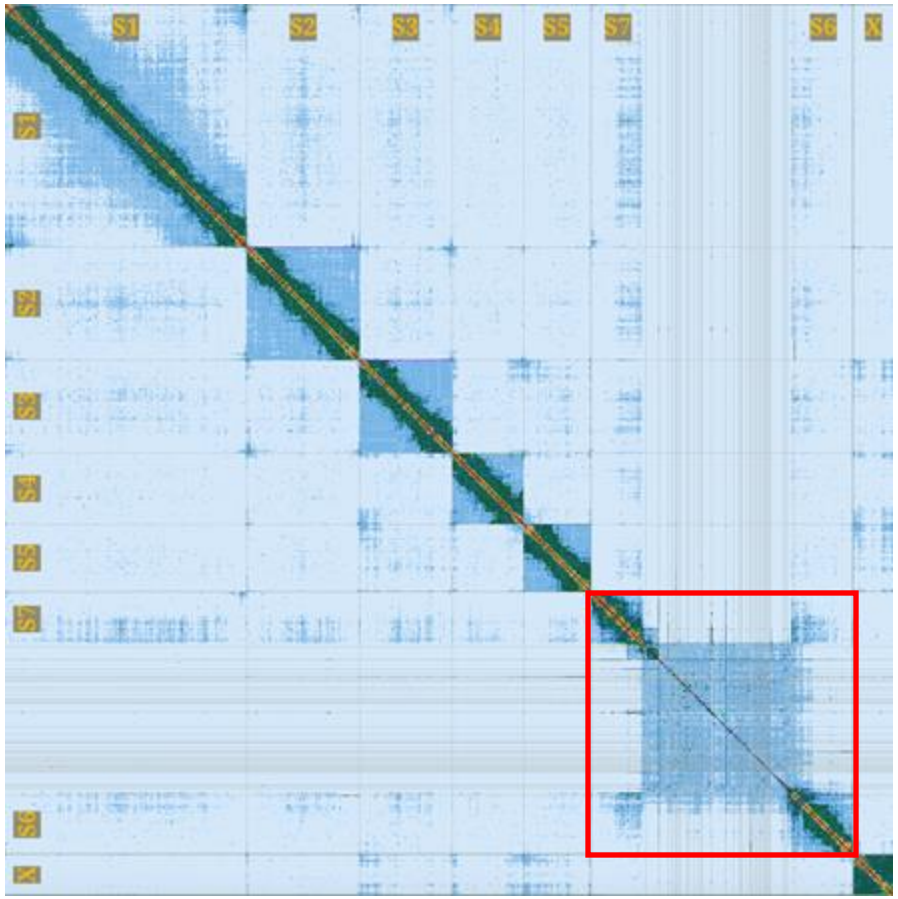
unplaceable sub-telomeric repeat

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

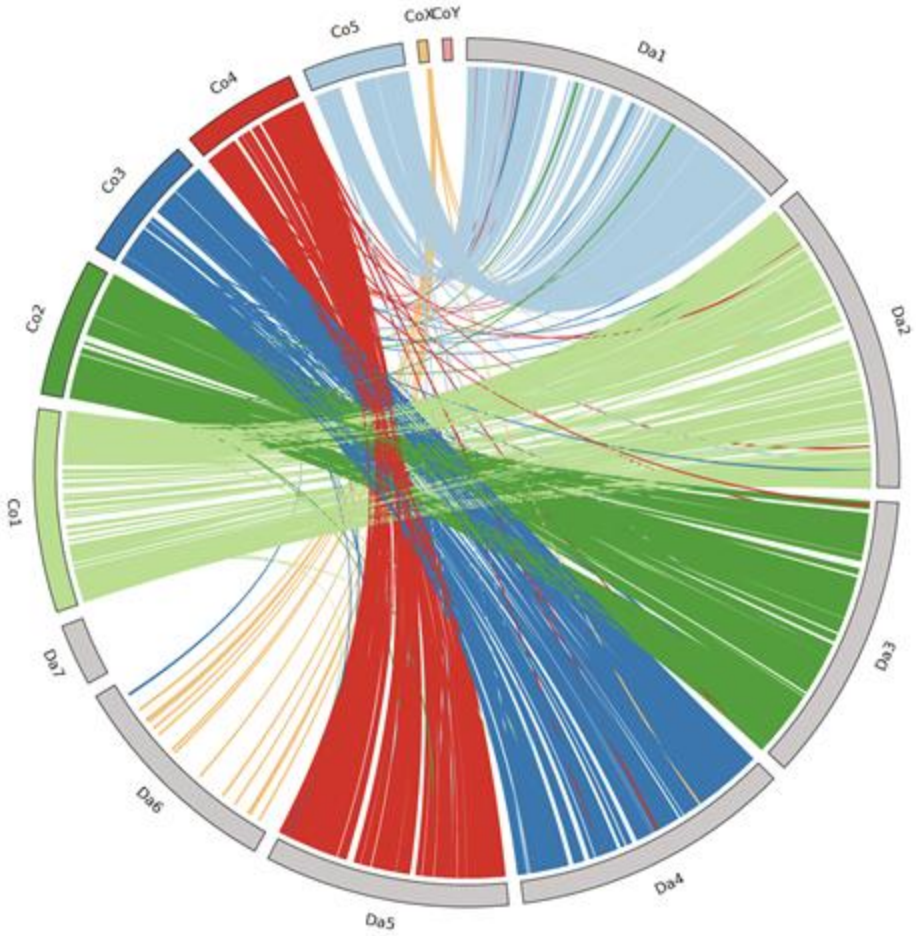
Curation accessory tools

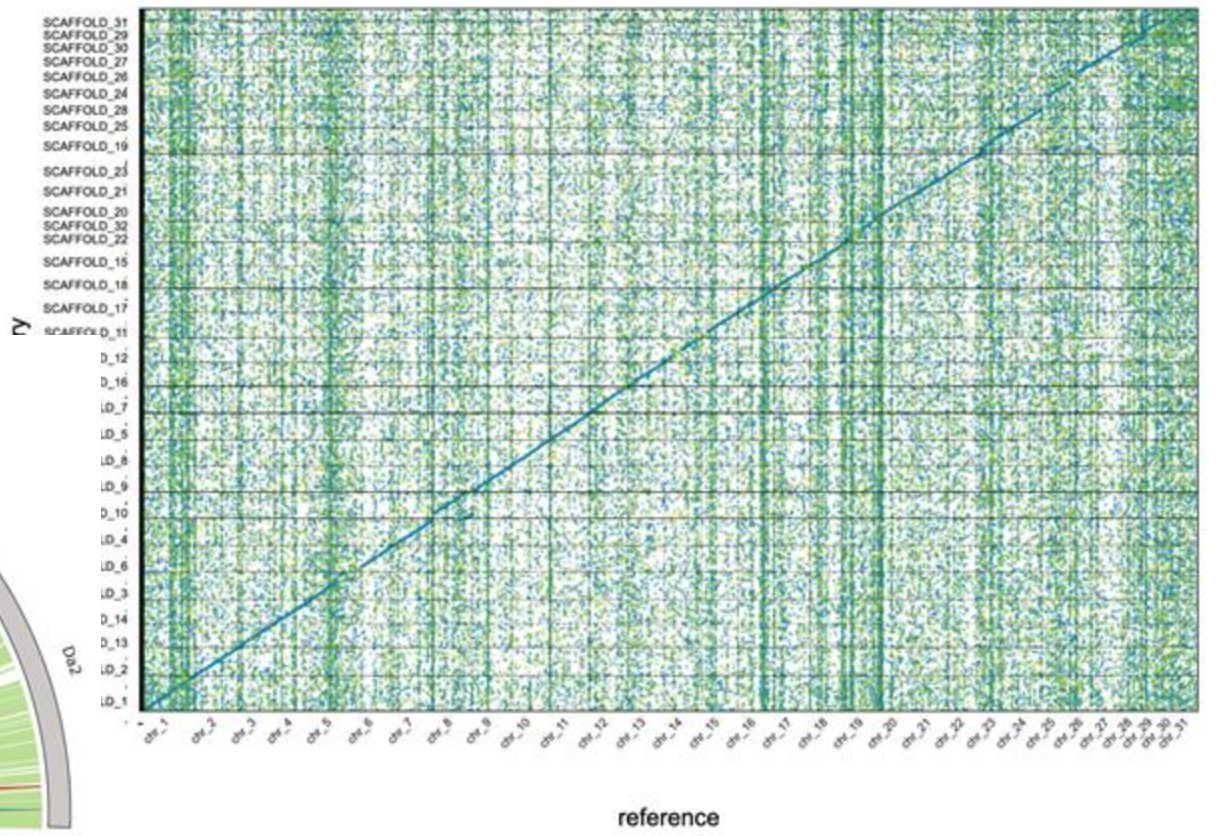
Synteny analysis

Alignment (Nucmer)

BUSCO

(TreeVAL)





Sex chromosome identification

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

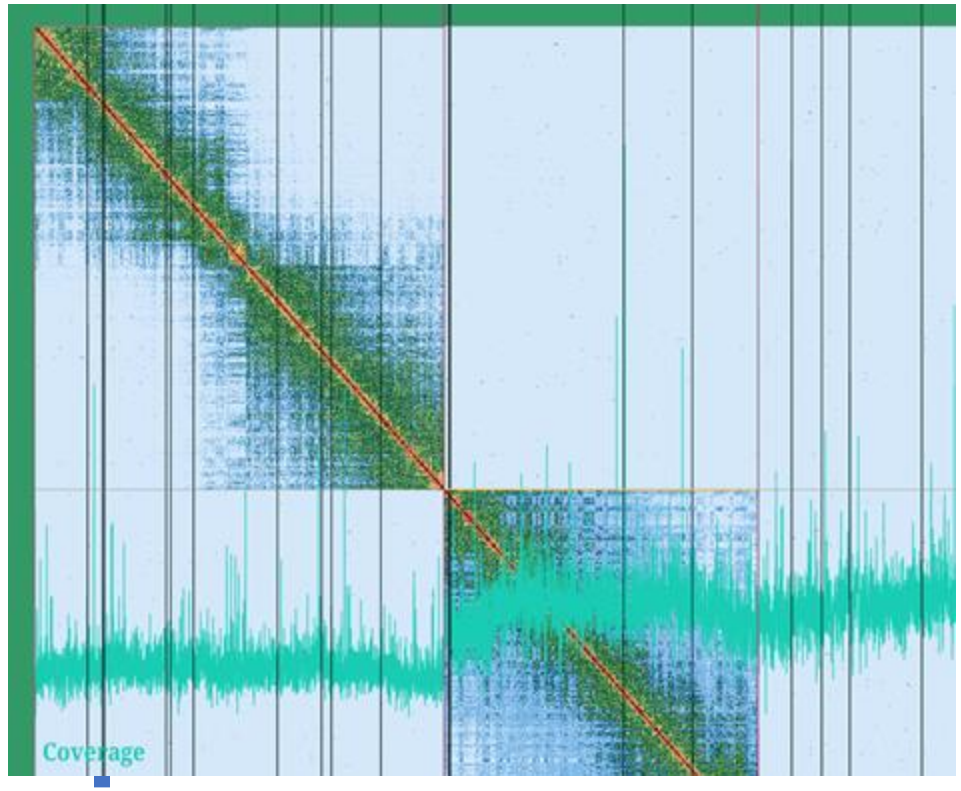
- Existing reference
- Genetic map

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)



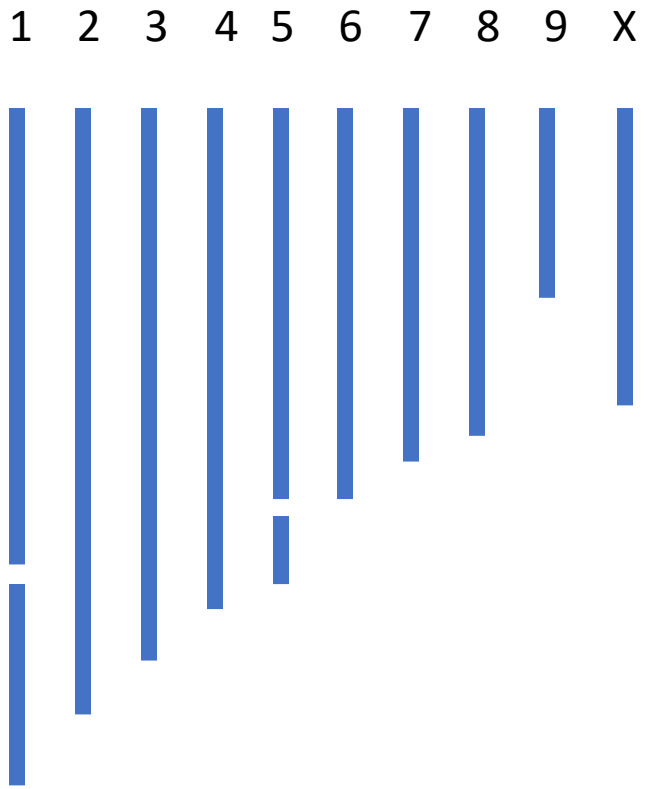
Chromosome naming

By size

- Autosomes large > small

By synteny

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

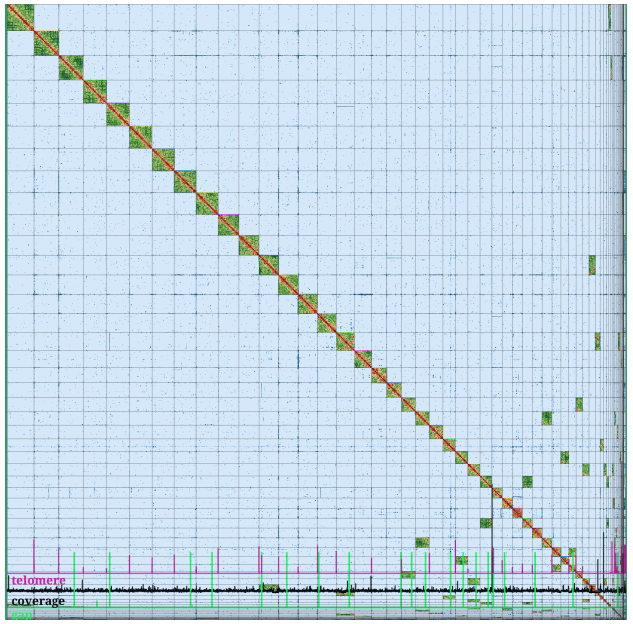


Our input

Automated (most of them) → Scaffolded → Decontaminated assemblies

Auto decontamination

Decontaminated assemblies → Assembly fasta file → Assembly mapped to HiC data → HiC map



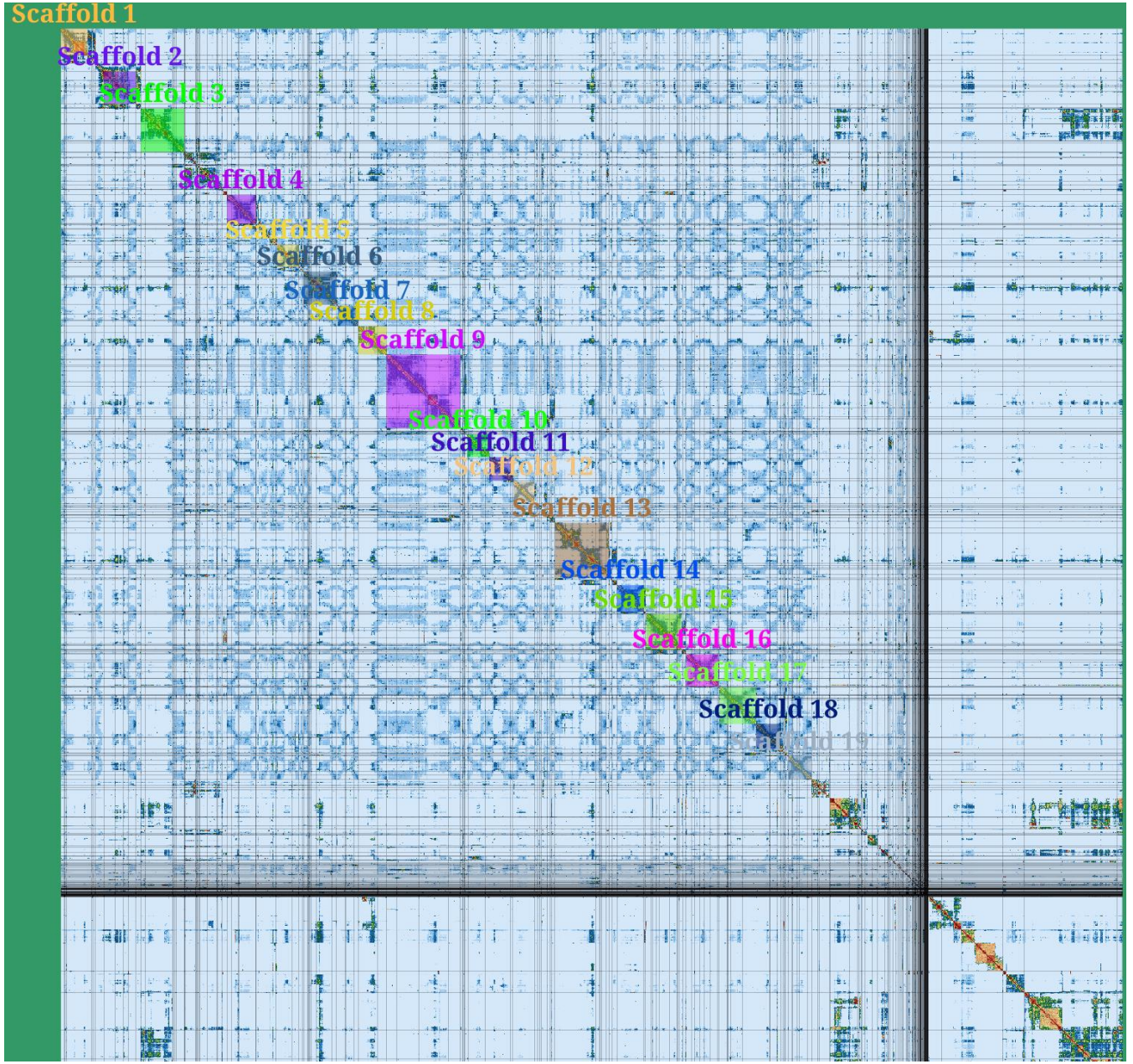
During curation

Misassemblies correction

Remove:
Remaining contamination
Haplotypic duplications

Assign sex chromosomes

Identify/confirm chromosome number



Our output

199 breaks
146 joins
135 haplotigs removed

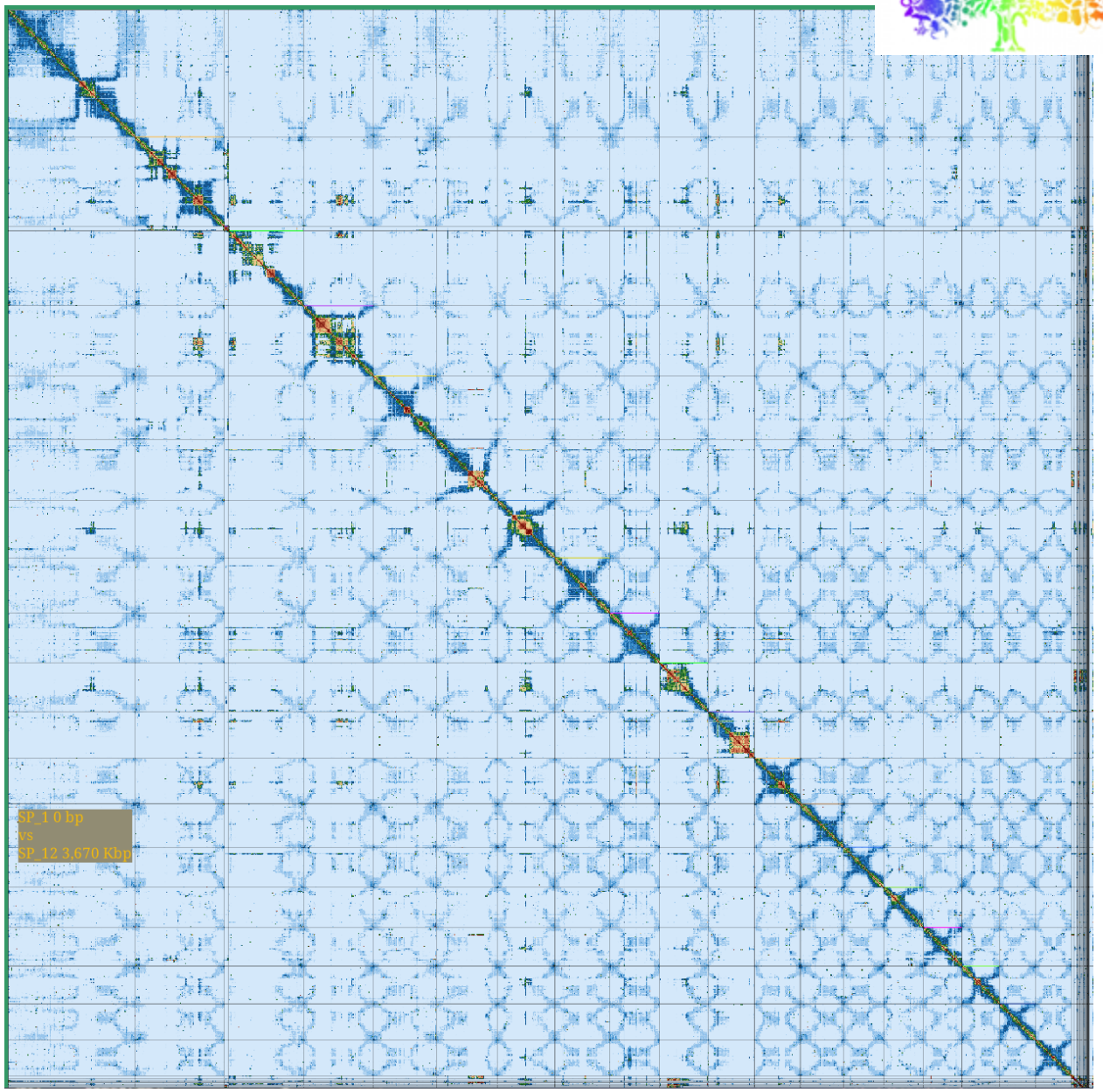
19 chromosomes

Increased contiguity

Curated and resolved

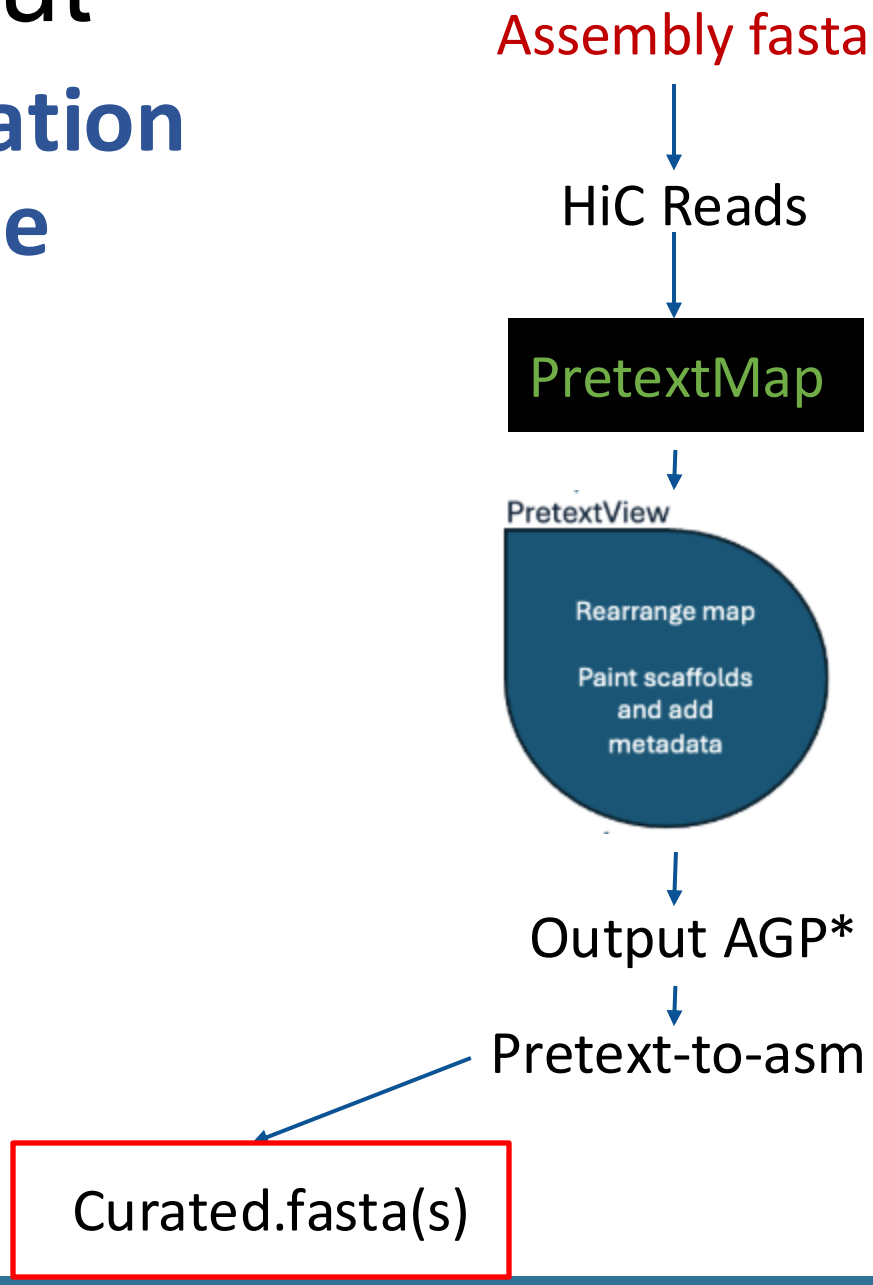
High quality chromosome-level assemblies

Curated fasta file



Our output

Rapid Curation pipeline



*AGP = A Golden Path

Defines the order and orientation of the genome blocks (chromosomes). This is taken from the edited pretext map

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

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)