SeqOccIn

INRAE
la science pour la vie, l'humain, la terre

Contig error correction based on linked reads

Andreea Dréau

October 1st, 2020

Genotoul GeT

Genotoul Bioinfo

SeqOccIn Project

Axis 1 - bioinfo

Claire Kuchly & Christophe Klopp
Thomas Faraut & Matthias Zytnicki
Clément Birbes, Arnaud Di-Franco & Andreea Dréau

SeqOccIn - Genome assembly project

- **Context:** New advances in the field of genomic sequencing
- **Aims:**
  1. Identify the best practice to obtain the highest quality assembly at a specific price
  2. Development of new methods for de novo assembly
- **Species:** Cattle, Maize, Pig, Quail, Goat, Sheep
- **Data:** Oxford Nanopore, Pacific Bioscience, 10x Chromium, Hi-C, Bionano optical mapping

Available data

Chromosome fragment

Oxford Nanopore
~16% errors
30 kb N50 (up to ~1 Mb)

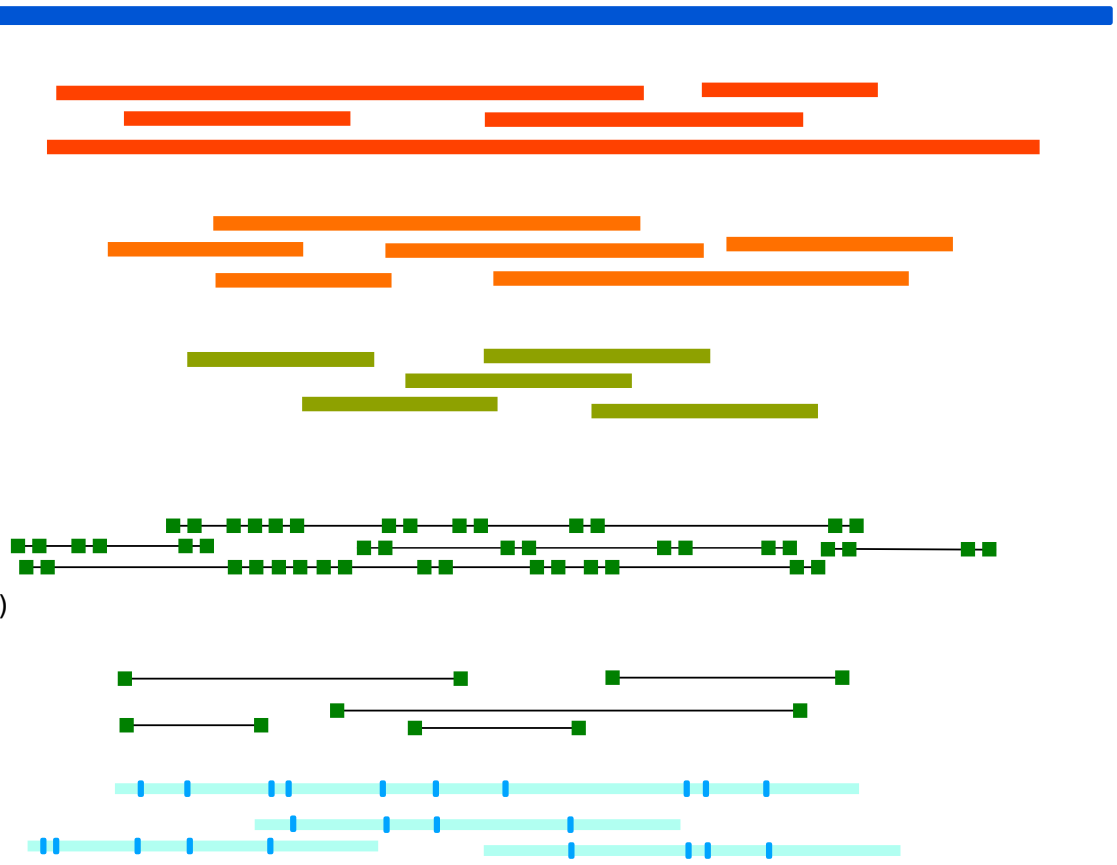
PacBio CLR
~15% errors
50 kb N50 (up to ~200 kb)

PacBio HiFi
~1% errors
15 kb N50 (up to ~40 kb)

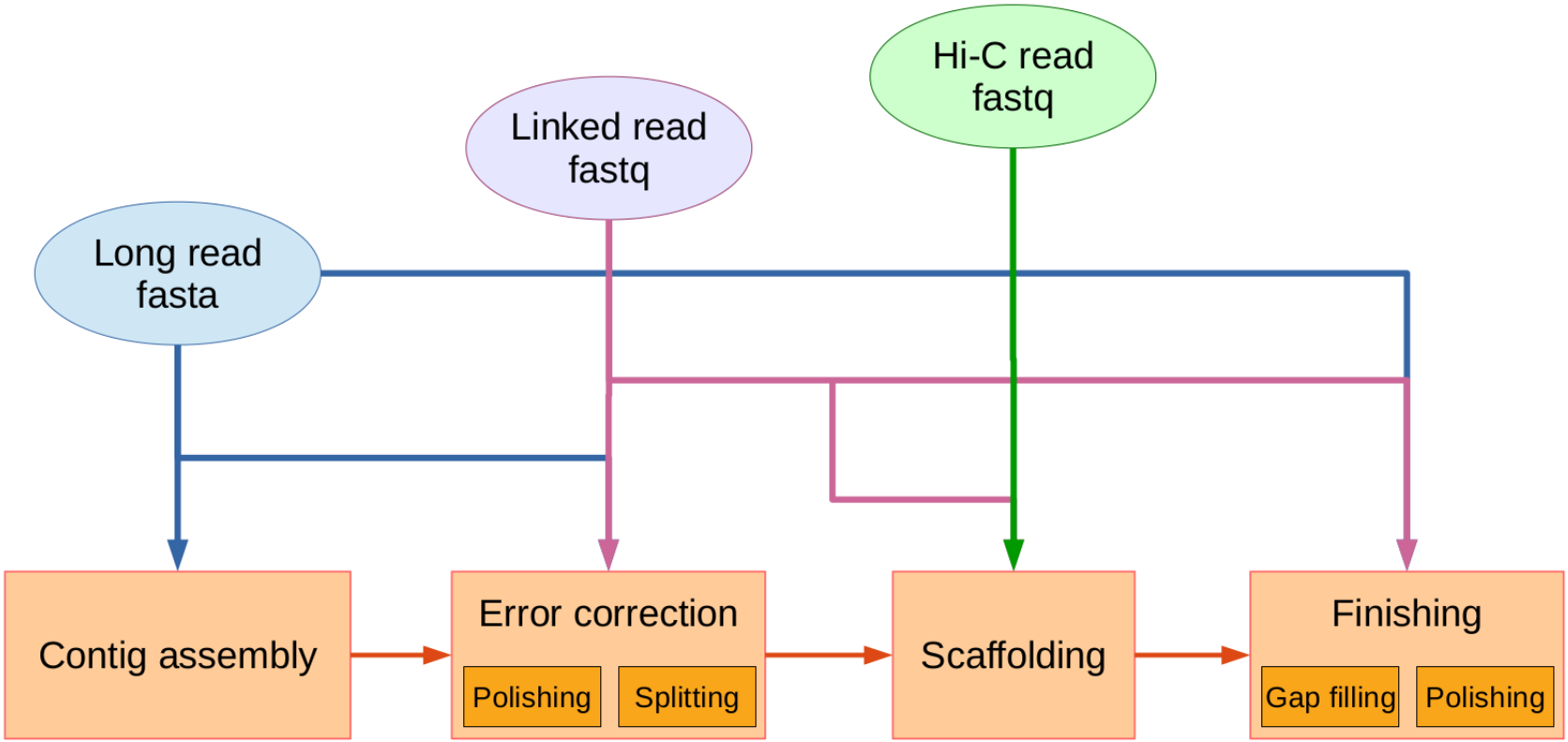
10x Chromium +
Illumina paired ends
~0.2% errors
150bpx2 (molecule length ~80 kb)

Illumina Hi-C
~0.2% errors
150bpx2 (contact length ~20 kb)

Bionano optical mapping
~300kb molecule length



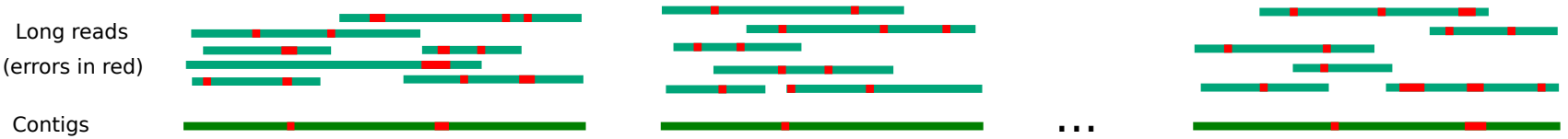
Genome Assembly Pipeline

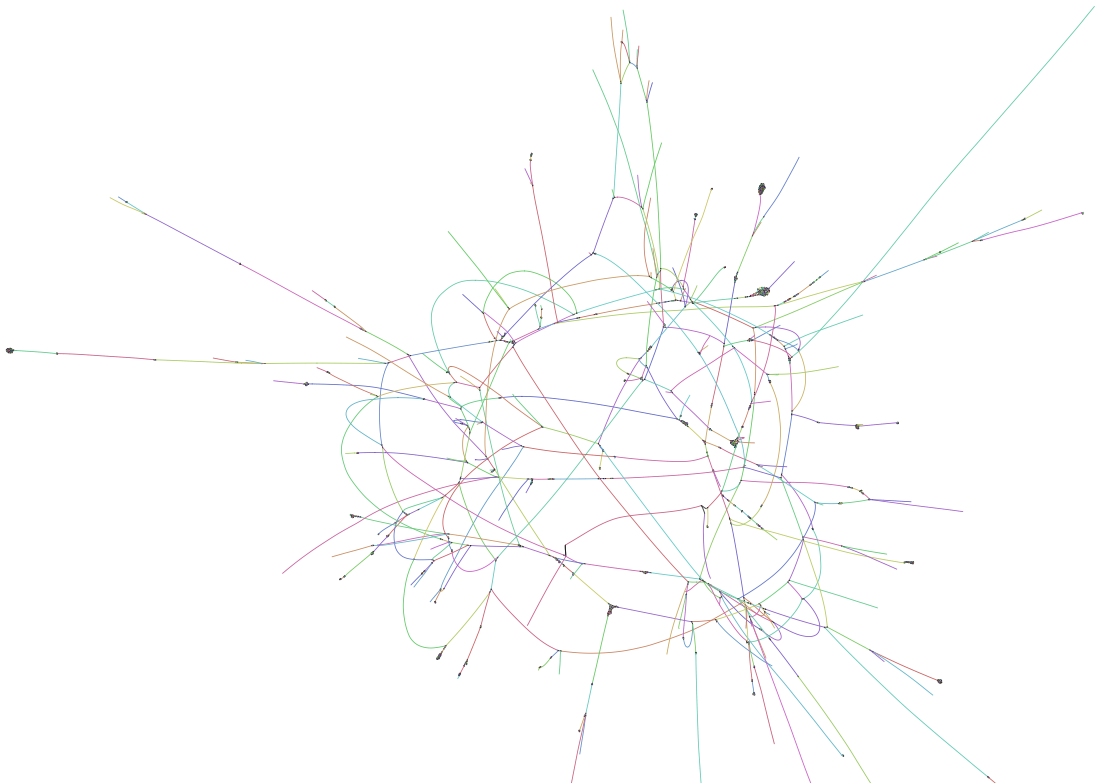


How do we connect the reads?

Contig assembly

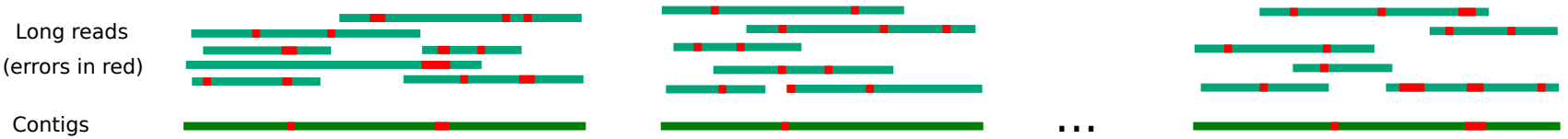
Construction of long chromosomal parts without N's based on overlapping reads and *an assembly graph*

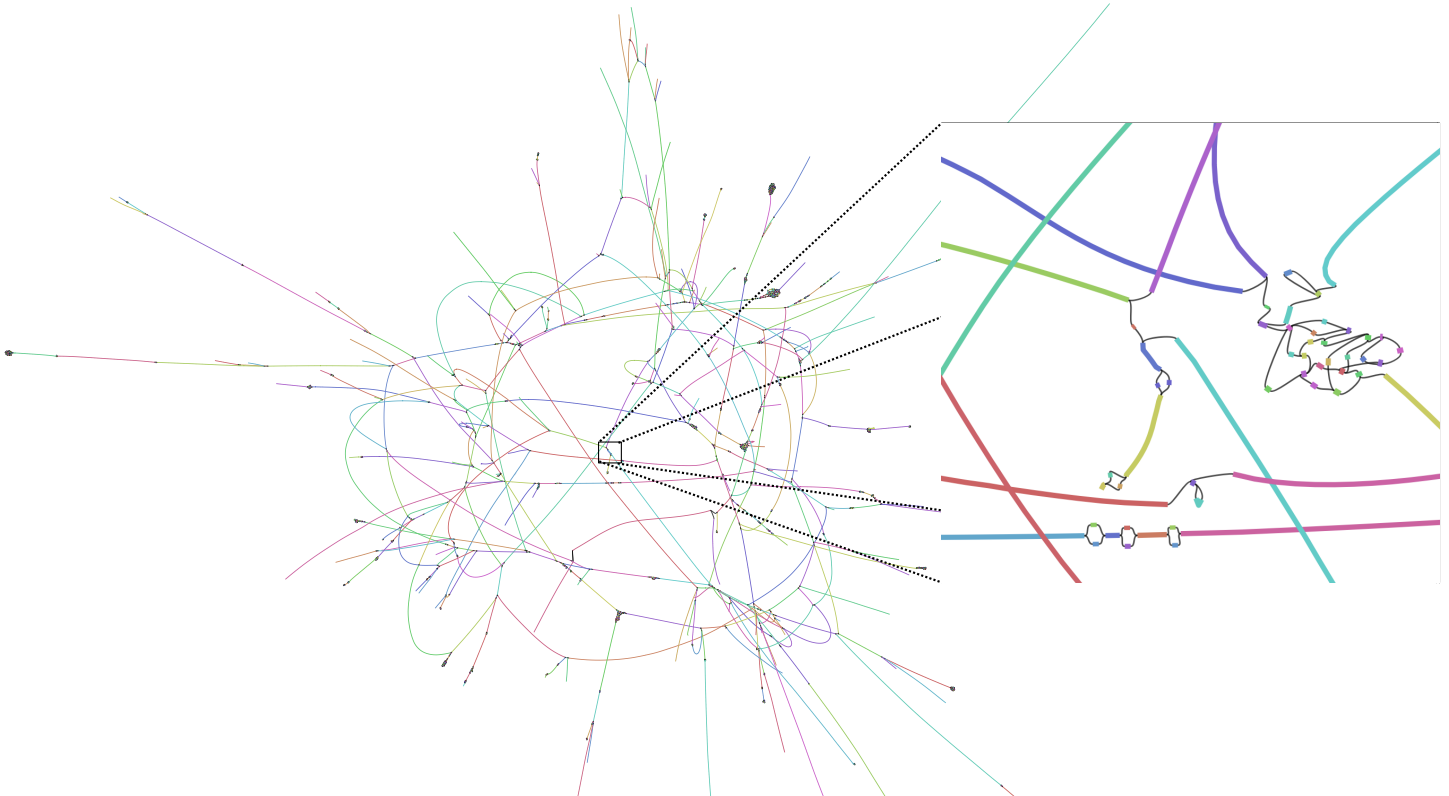




Contig assembly

Construction of long chromosomal parts without N's based on overlapping reads and *an assembly graph*

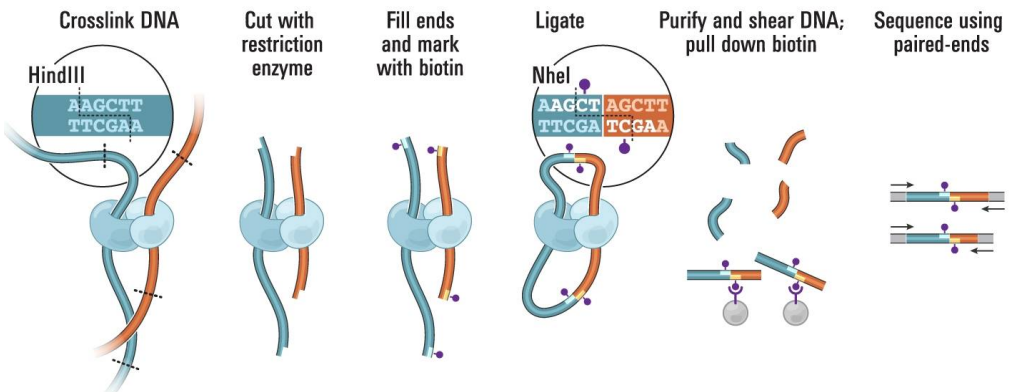




How do we connect the contigs?

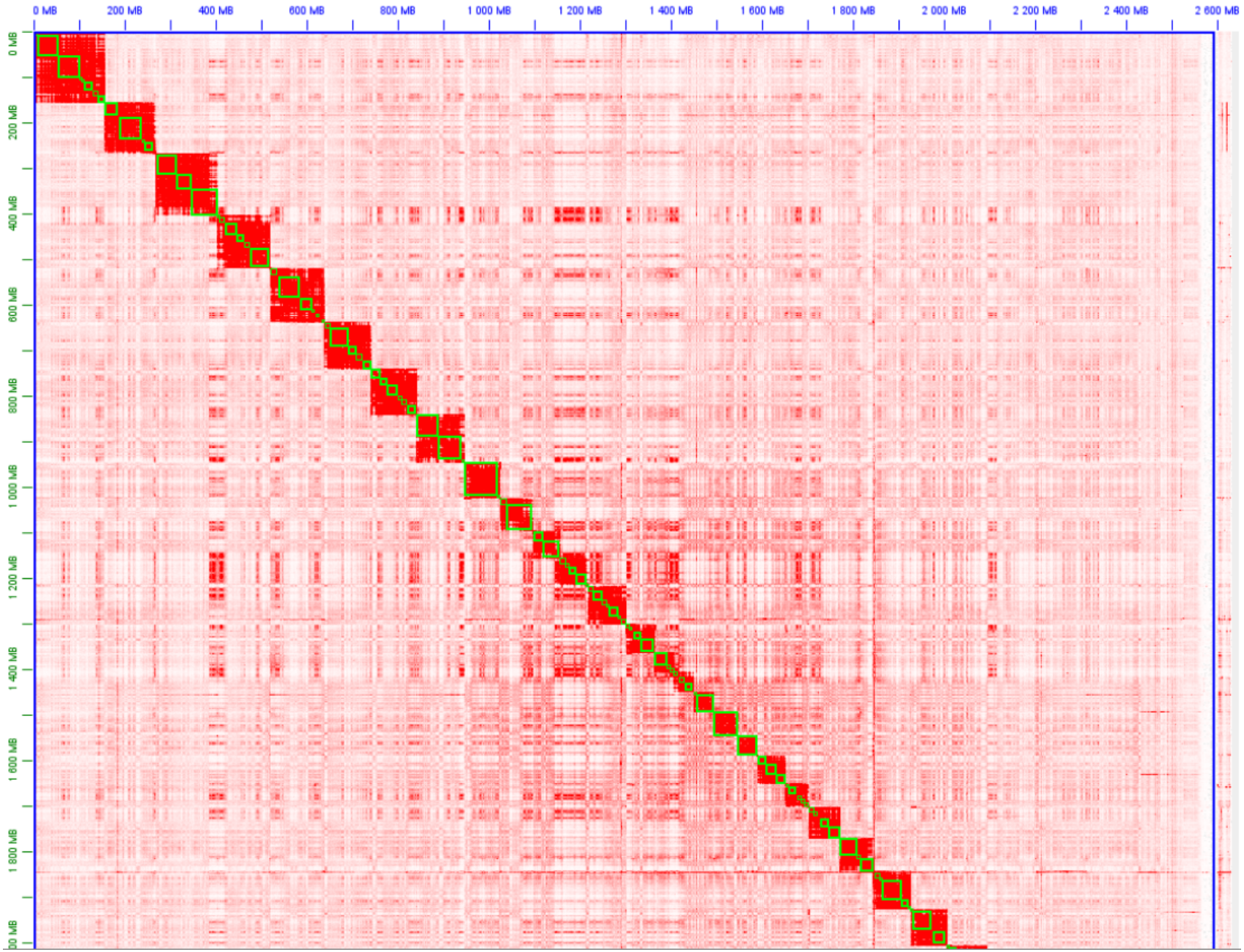
Scaffolding with Hi-C reads

Align Hi-C reads and connect contigs into scaffolds/chromosomes based on contacts



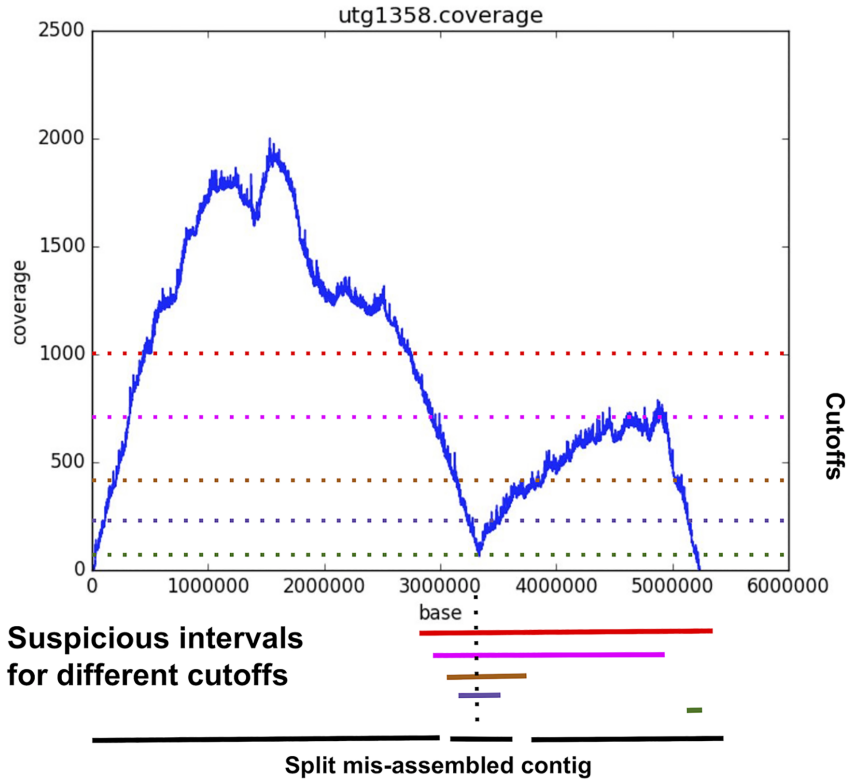
[Lieberman-Aiden et al., Science, 2009]

Hi-C heat map

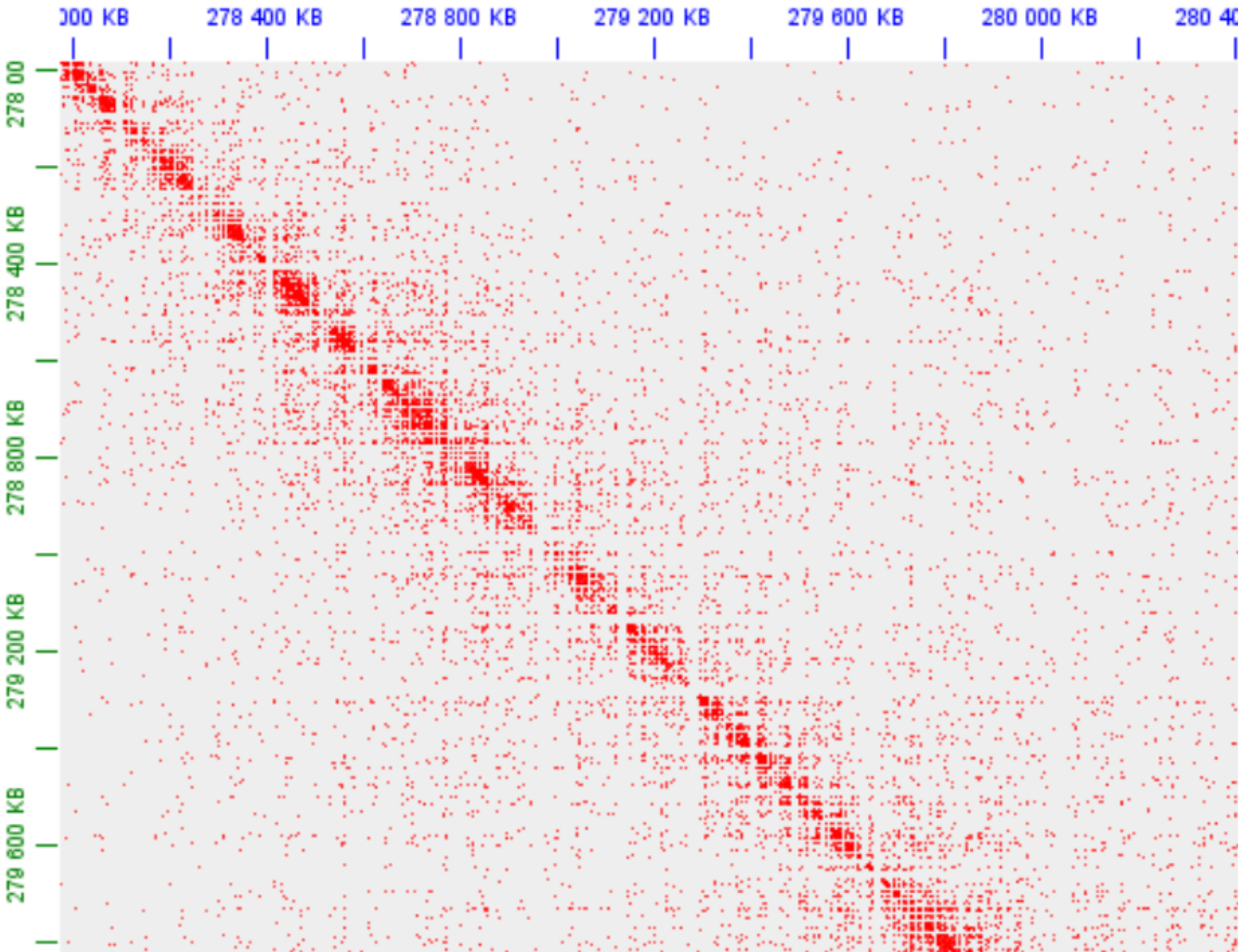

0 MB
200 MB
400 MB
600 MB
800 MB
1 000 MB
1 200 MB
1 400 MB
1 600 MB
1 800 MB
2 000 MB
2 200 MB
2 400 MB
2 600 MB

Hi-C scaffolding methods

- align reads on contigs
- split contigs
- scaffold construction

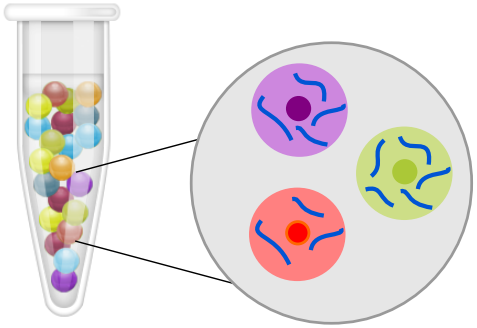


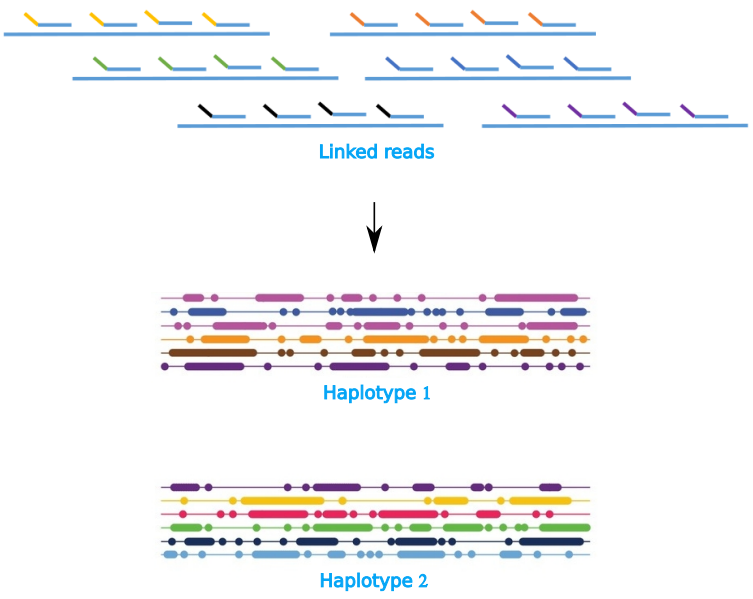
Hi-C heat map (zoom)



Can we use a more homogeneous read coverage?

Long-range "linked-read" sequencing using 10x Genomics

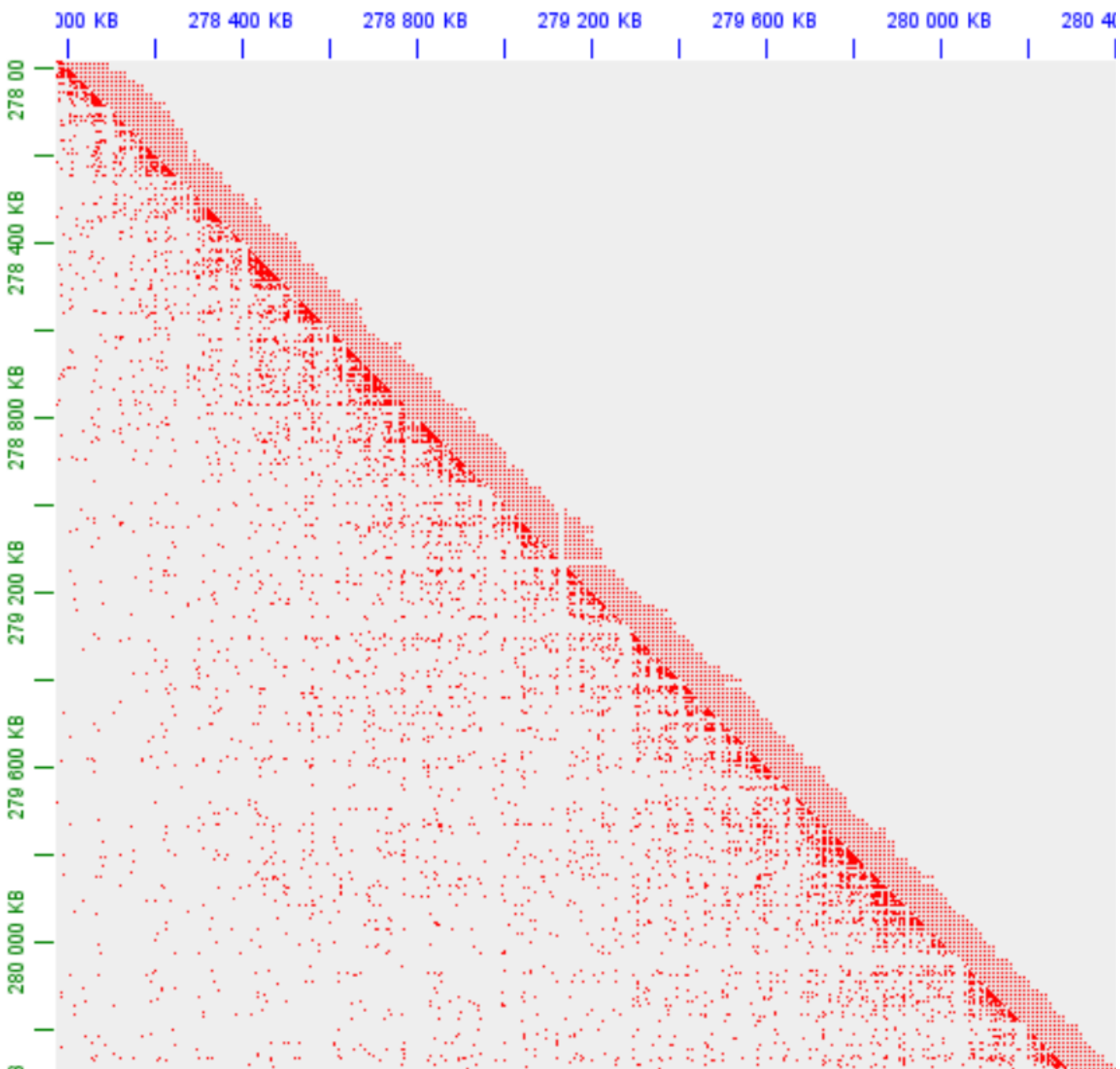




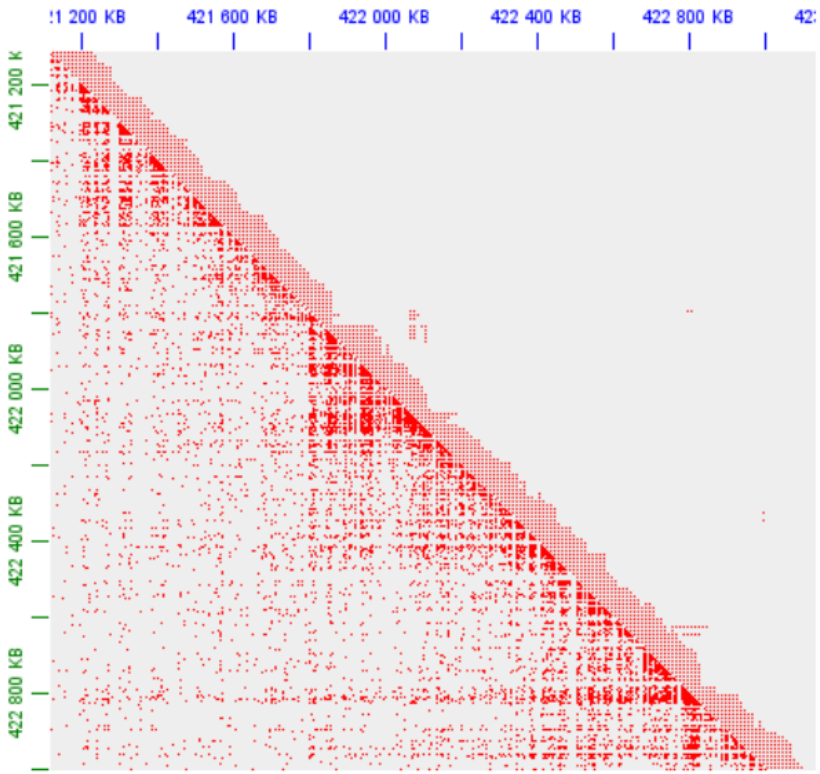
- generated from long DNA molecules
- tagged with a specific barcode
- computational reconstruction of single molecules
- provides low-cost long-range information

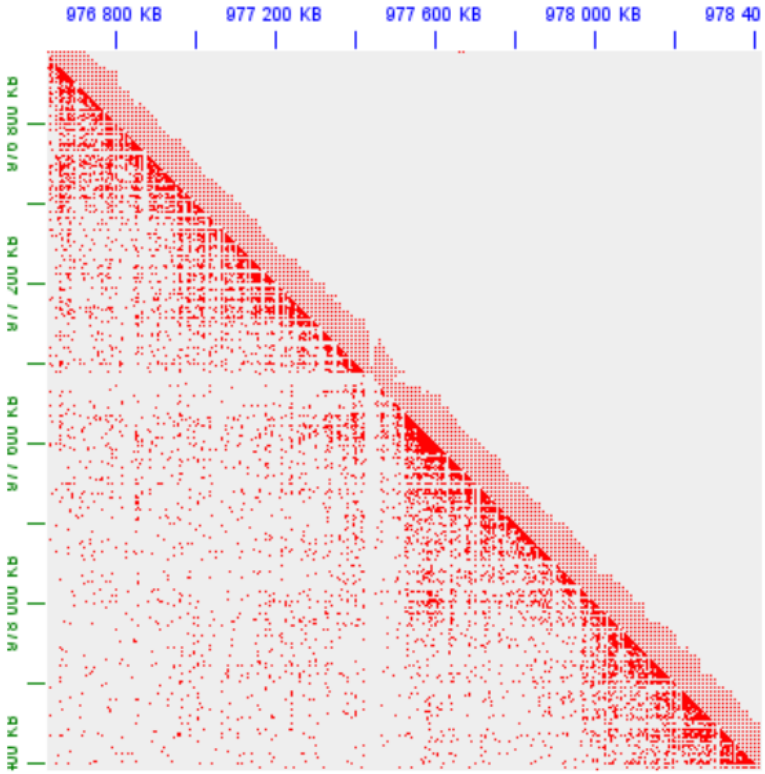
Source: 10x Genomics

Hi-C vs 10x Chromium heat map: Contig assembly errors



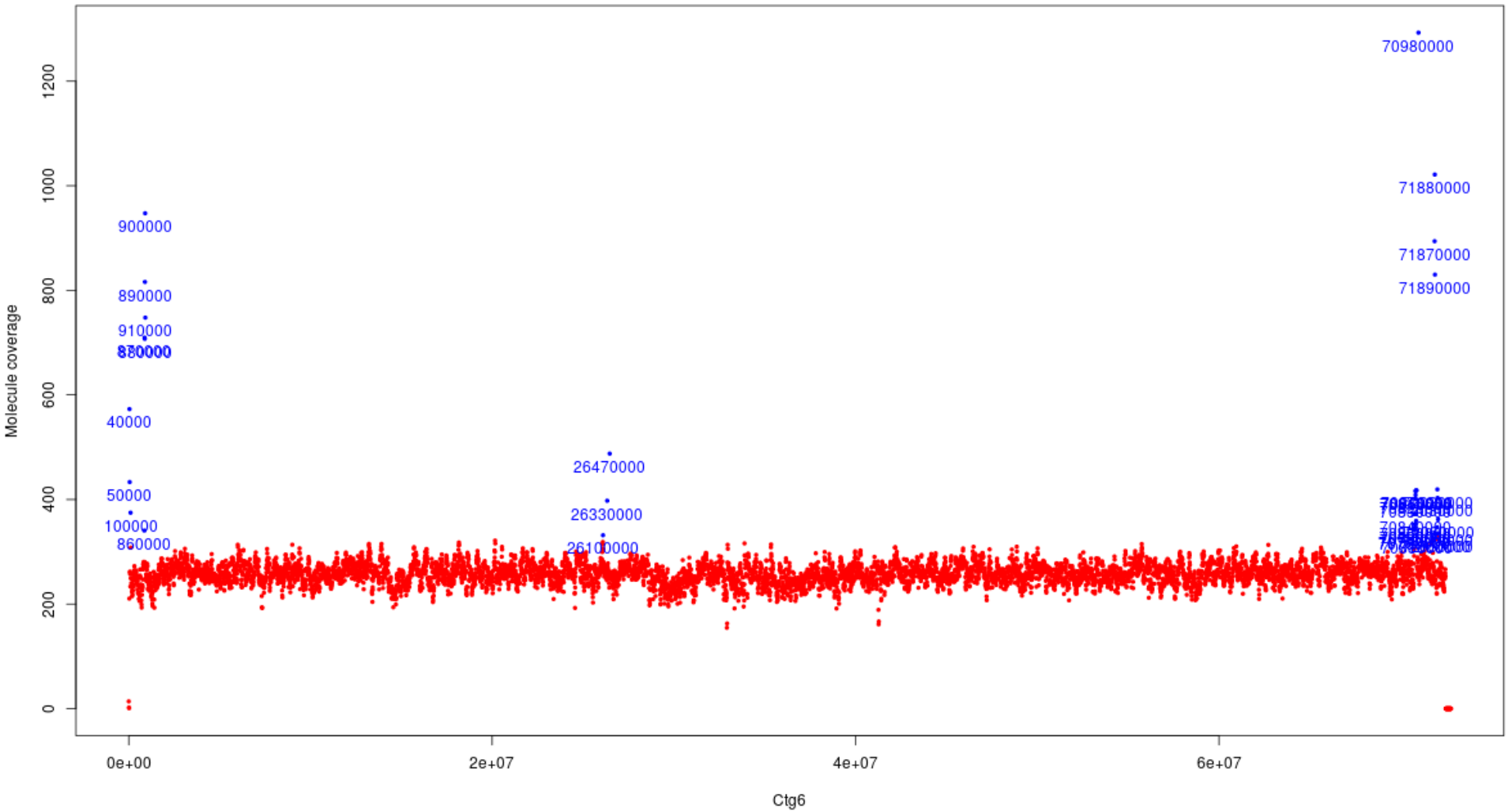
Hi-C vs 10x Chromium heat map: Contig assembly errors



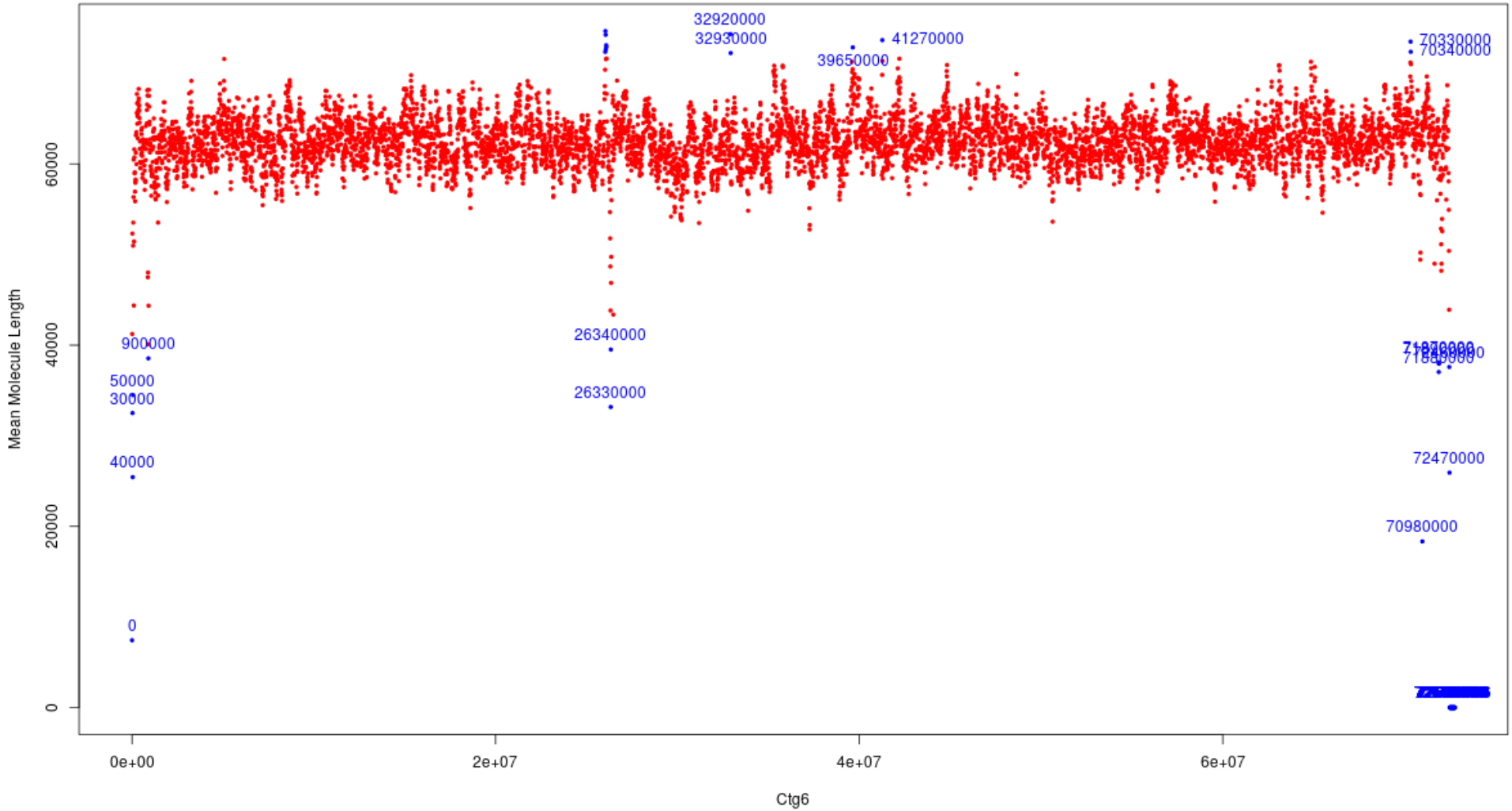


Can we find contig assembly errors with 10x molecules?

Molecule coverage on a contig (10kb interval)



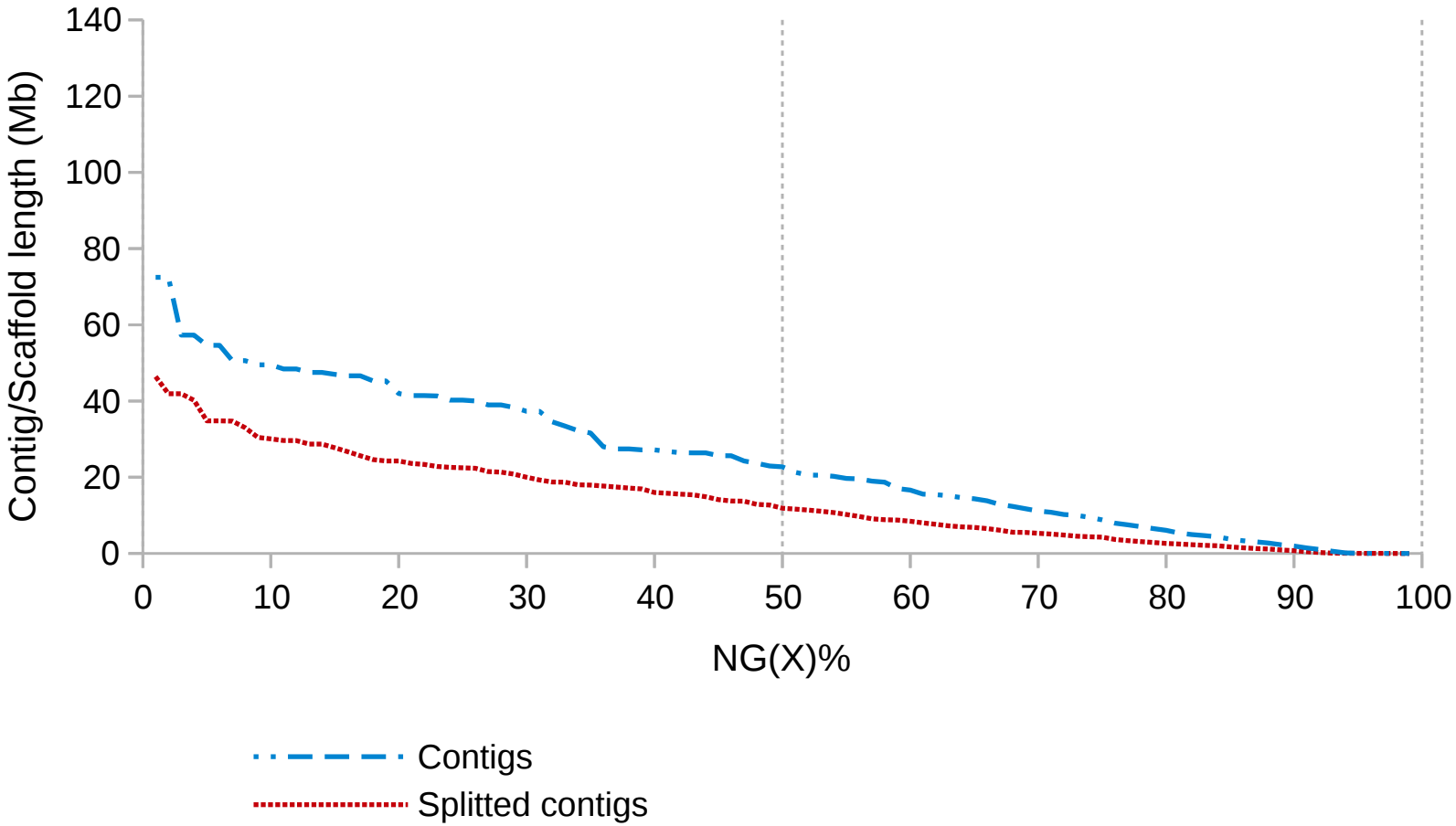
Mean molecule length on a contig (10kb interval)



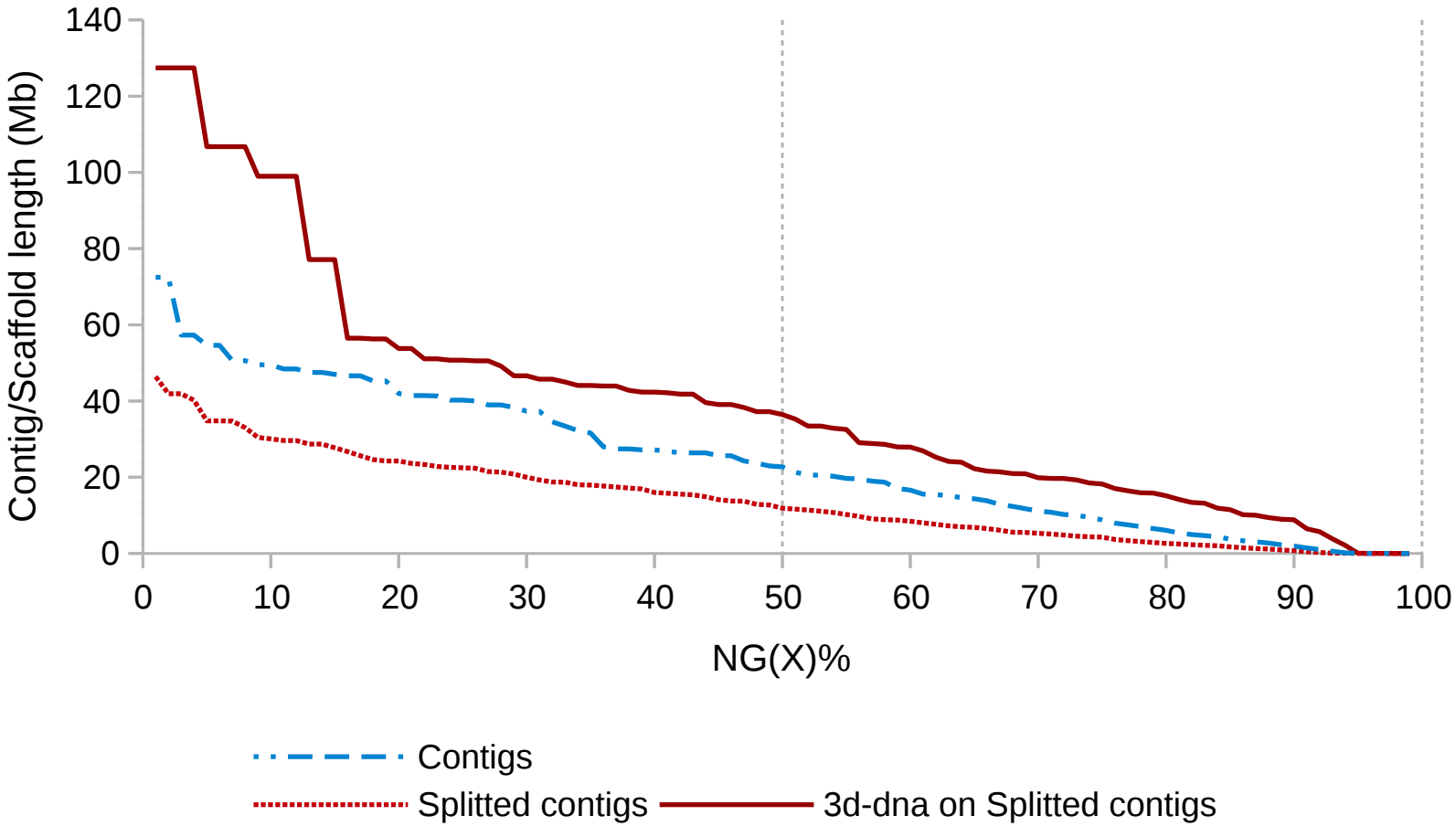
Split contigs with 10x linked reads

- Align linked-reads on contigs: Long Ranger align/bwa mem
- Identify molecules: Long Ranger reportMolecules
- Compute molecule profiles per interval (10kb)
  - Molecule coverage
  - Mean read density/molecule
  - Mean molecule length
- Identify outliers intervals
- Split contig if an interval is an outlier for at least two profiles
- Re-connect with Hi-C scaffolding methods

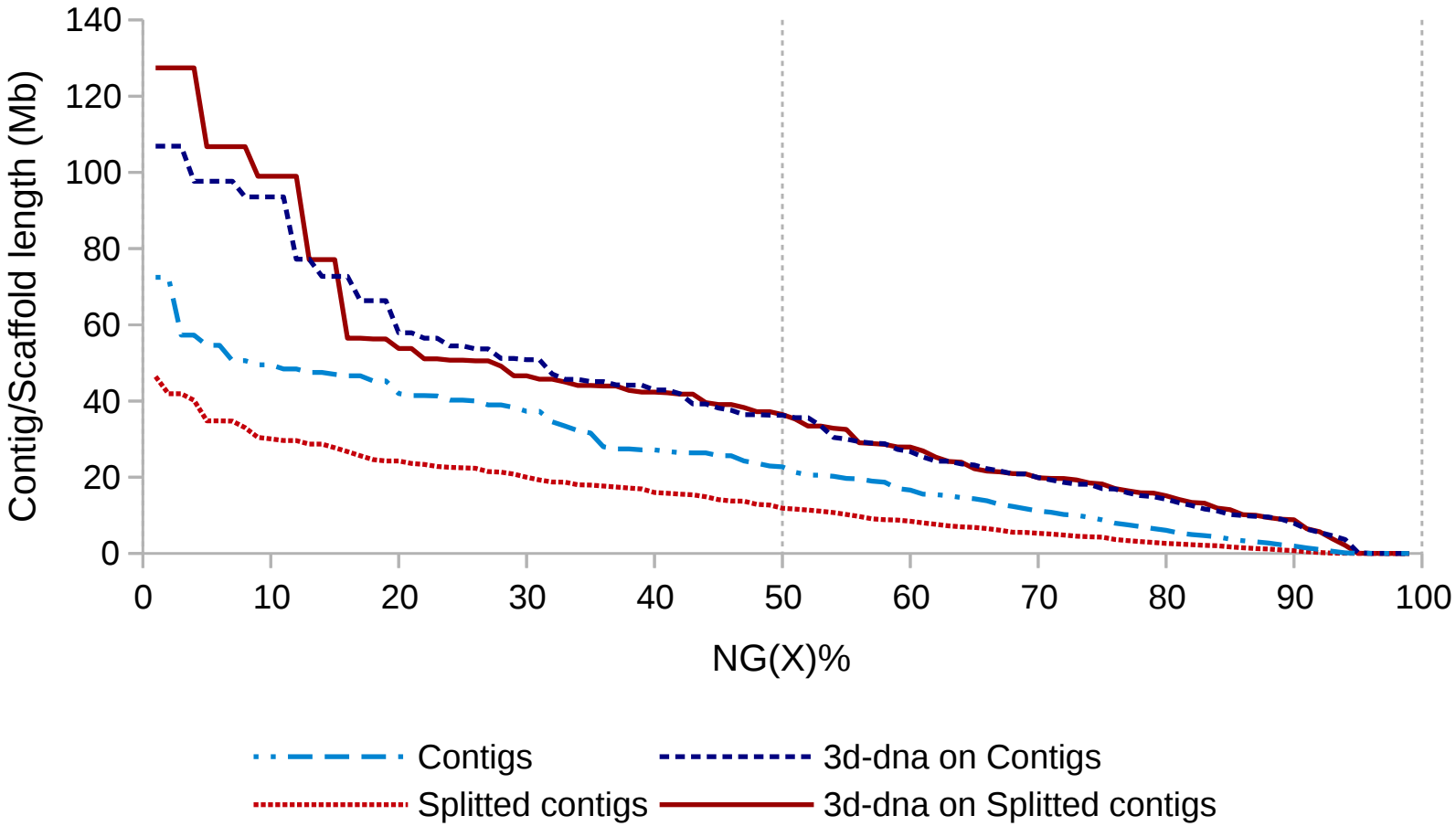
Impact on contig length



Scaffolding with 3d-dna



Scaffolding with 3d-dna



Future plans

- Stronger constraints for a split
- Tests with different assemblers and Hi-C scaffolders
- Scaffold with linked reads first
  - build graph based on 10x links
  - connect branchless paths of contigs