SeqOccIn

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

Contig error correction based on linked-read sequencing data

Andreea Dréau, Clément Birbes, Christophe Klopp, Matthias Zytnicki

November 23rd, 2020

SeqBIM

Genotoul GeT

Genotoul Bioinfo

SeqOccIn Project

- Axis 1 - Genome assembly and structural variation detection
- Axis 2 - Epigenetics marks detection
- Axis 3 - Metagenomics analysis
- Axis 4 - High molecular weight DNA extraction and evolution of the IT infrastructure

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)

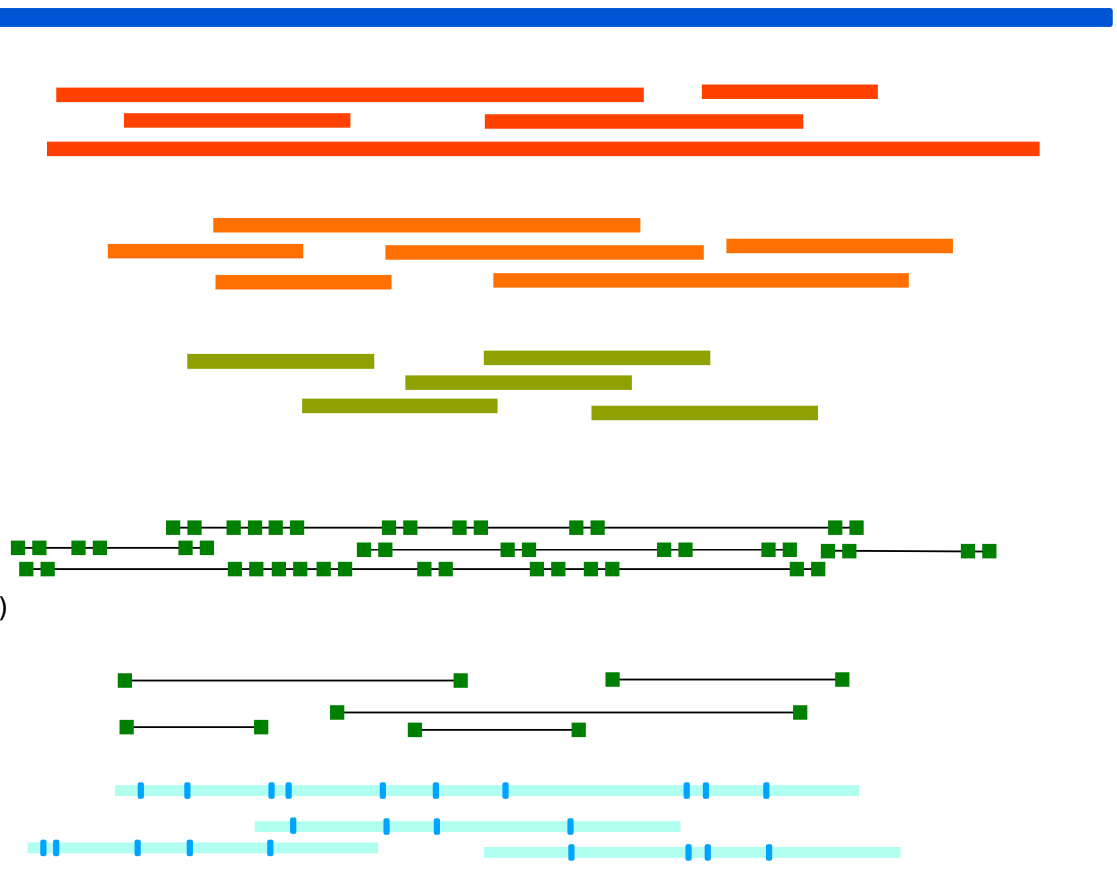
PacBio CLR
~15% errors
50 kb N50 (up to ~200 kb)

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

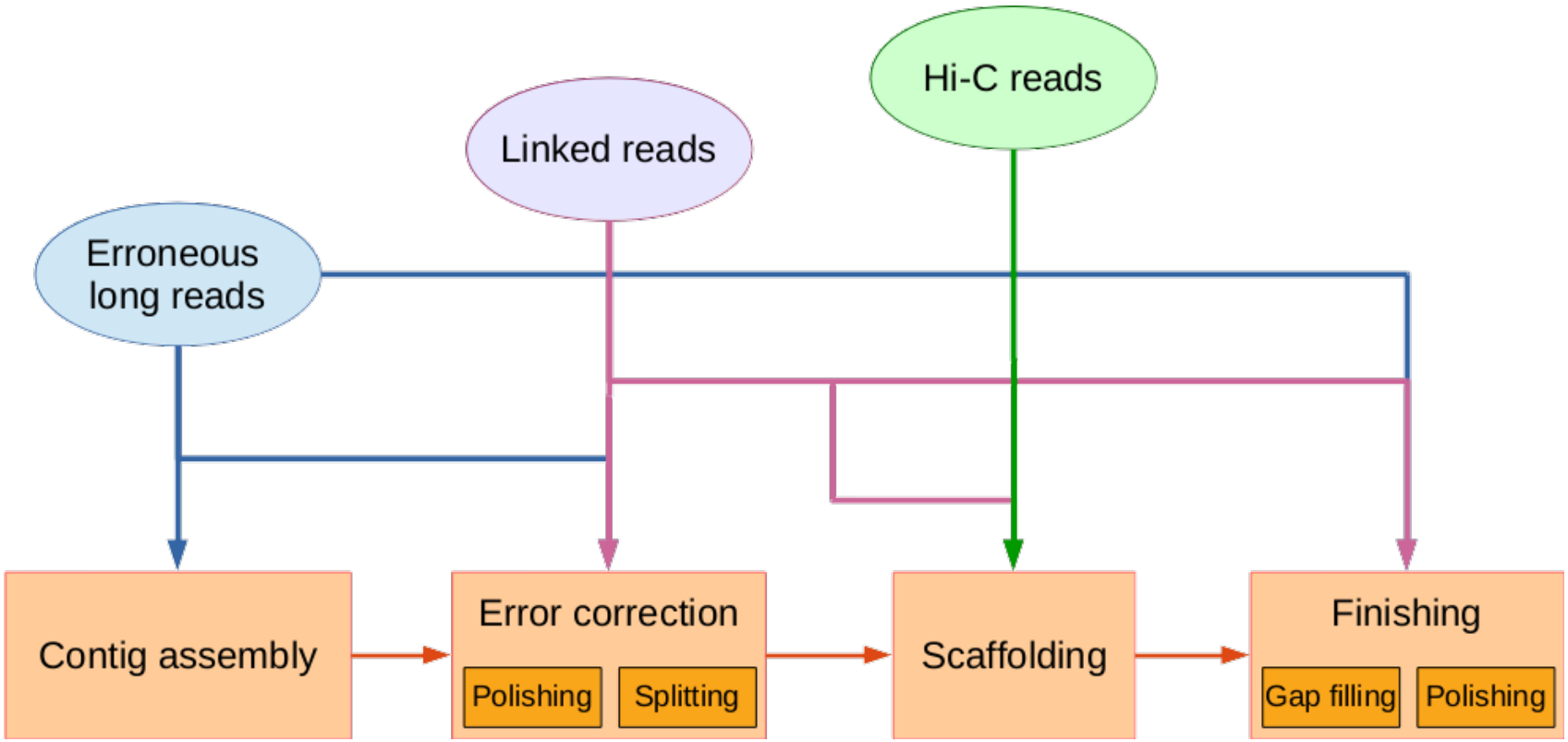
10x Chromium +
Illumina paired ends
~0.2% errors
150bpx2 (molecule length ~80 kb)

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

Bionano optical mapping
~300kb molecule length

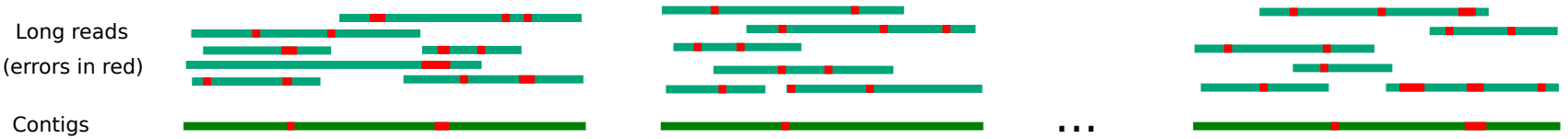


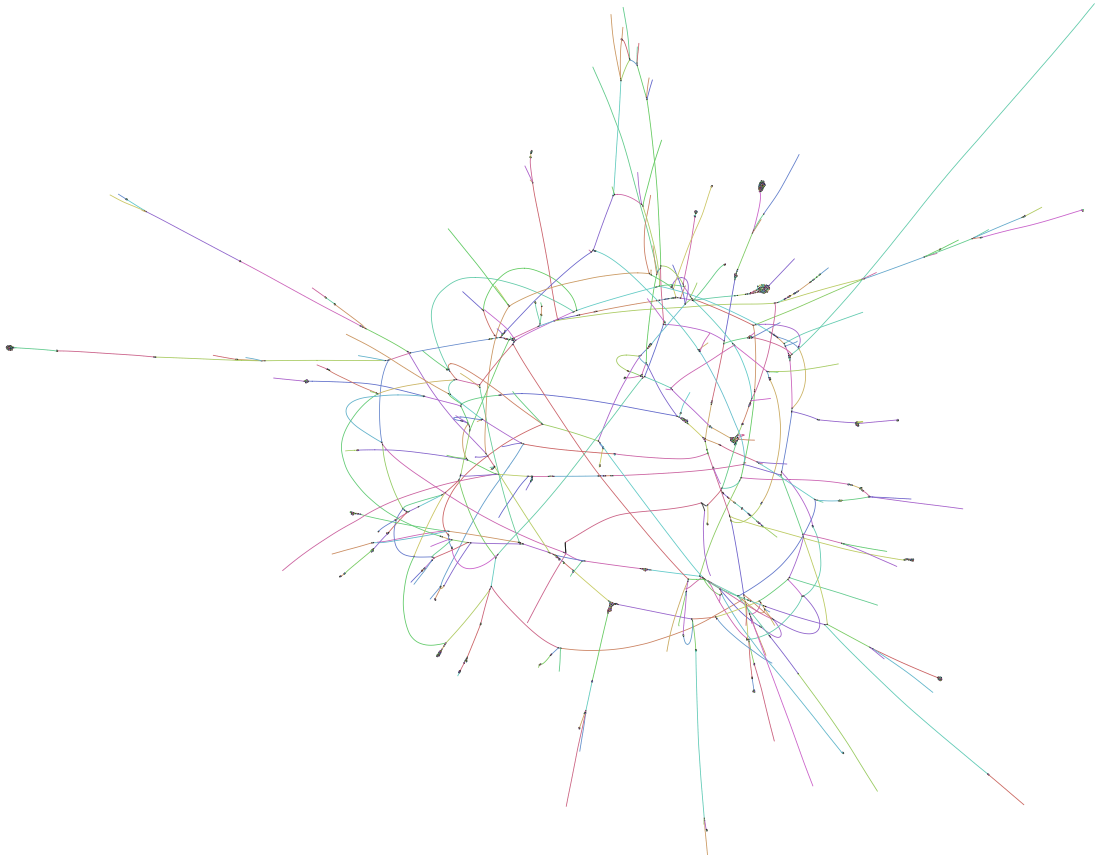
Genome Assembly Pipeline



Contig assembly with erroneous long read data

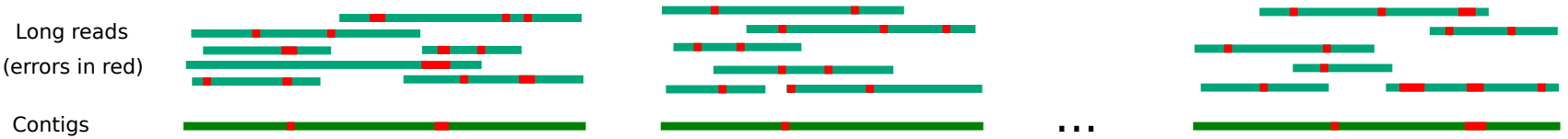
Construction of long chromosomal parts without N's based on overlapping reads and *an assembly graph*

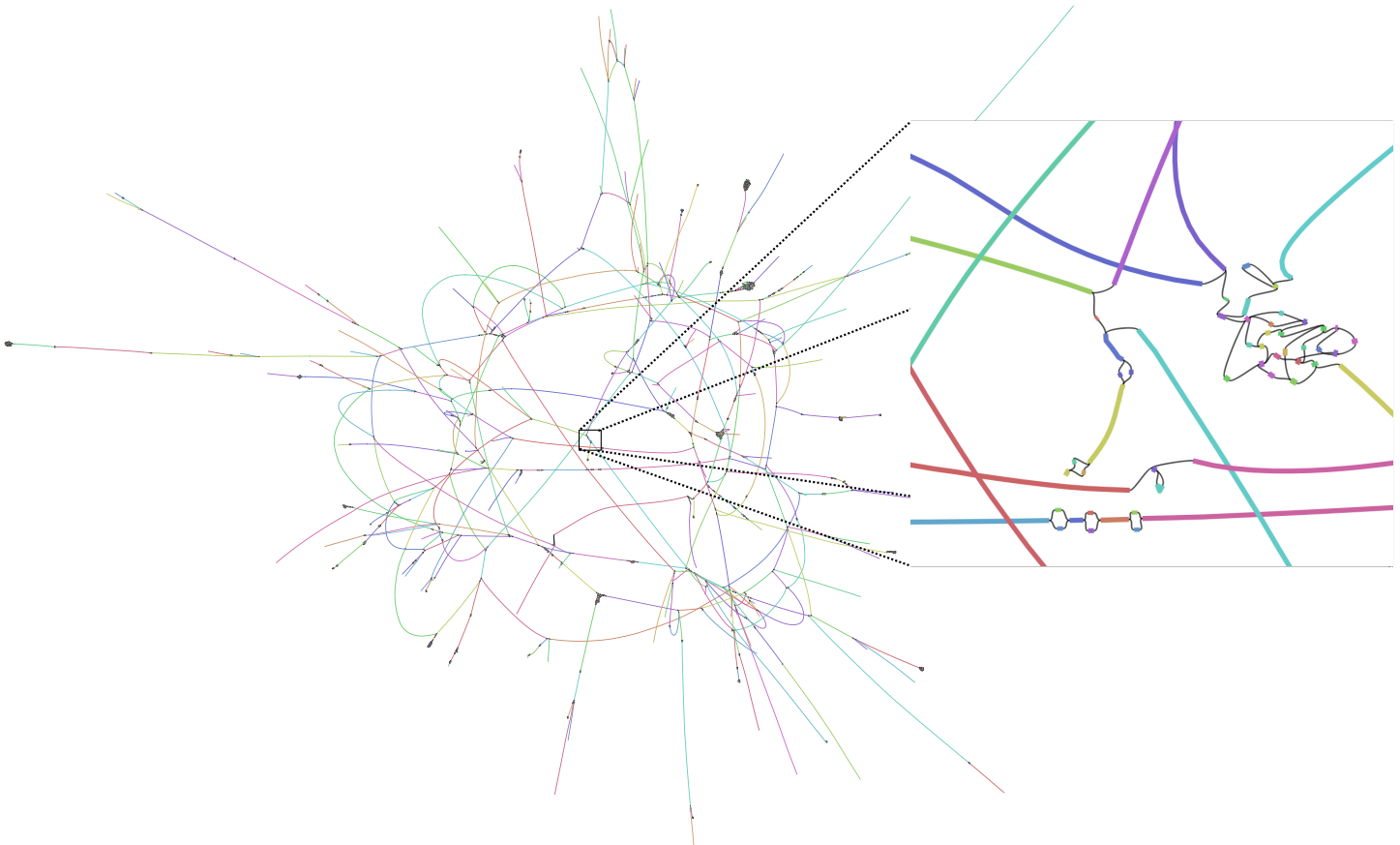




Contig assembly with erroneous long read data

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





Error correction

Polishing

Correction of **sequencing** errors based on the alignment of short and long reads to the contigs

Splitting

Correction of **connection** errors based on coverage drop

Error correction

Polishing

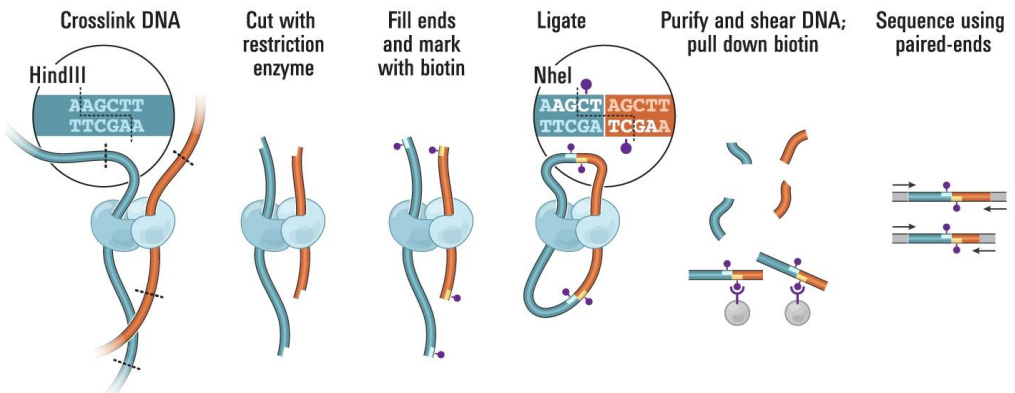
Correction of **sequencing** errors based on the alignment of short and long reads to the contigs

Splitting

Correction of **connection** errors based on coverage drop . . .

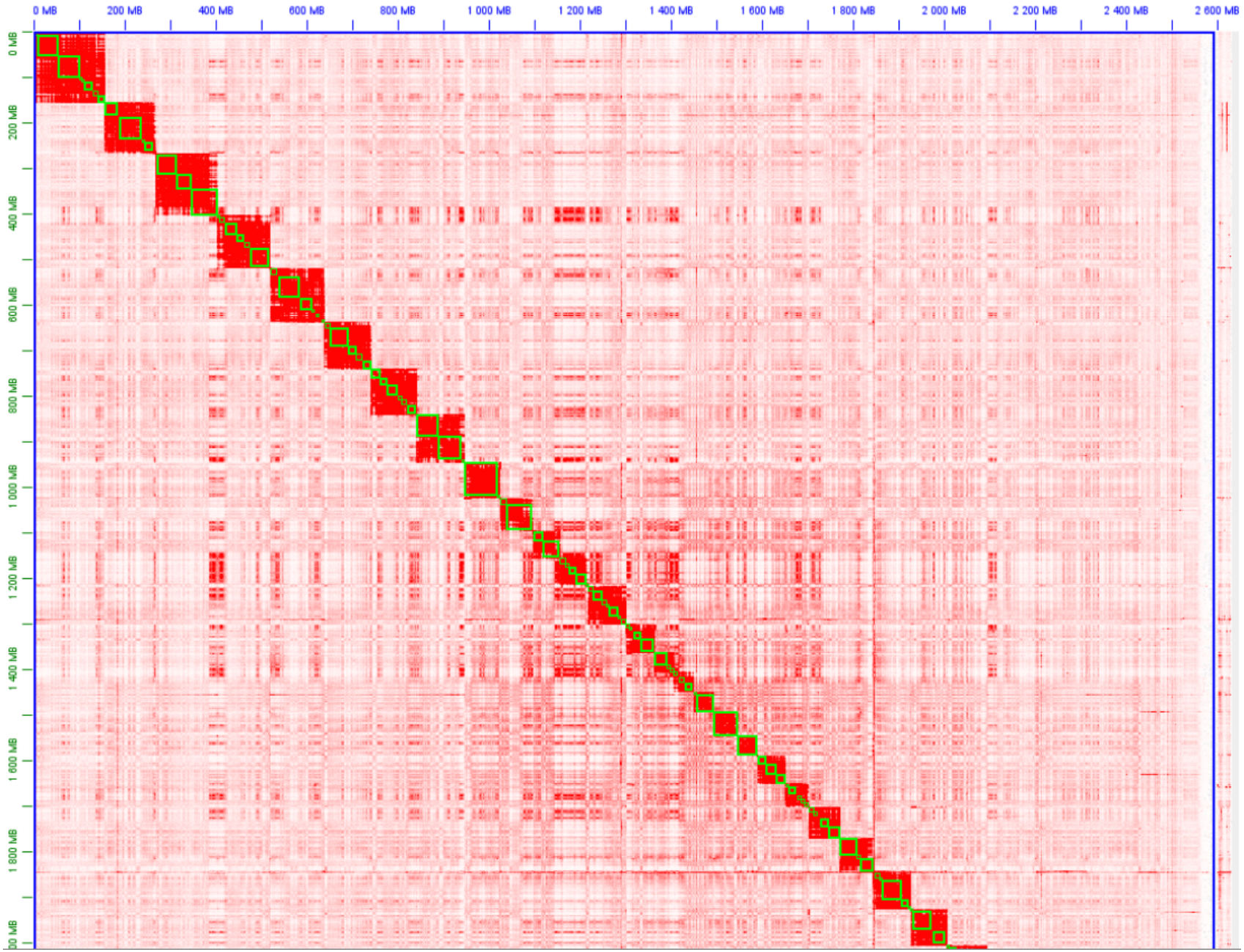
Scaffolding with Hi-C reads

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

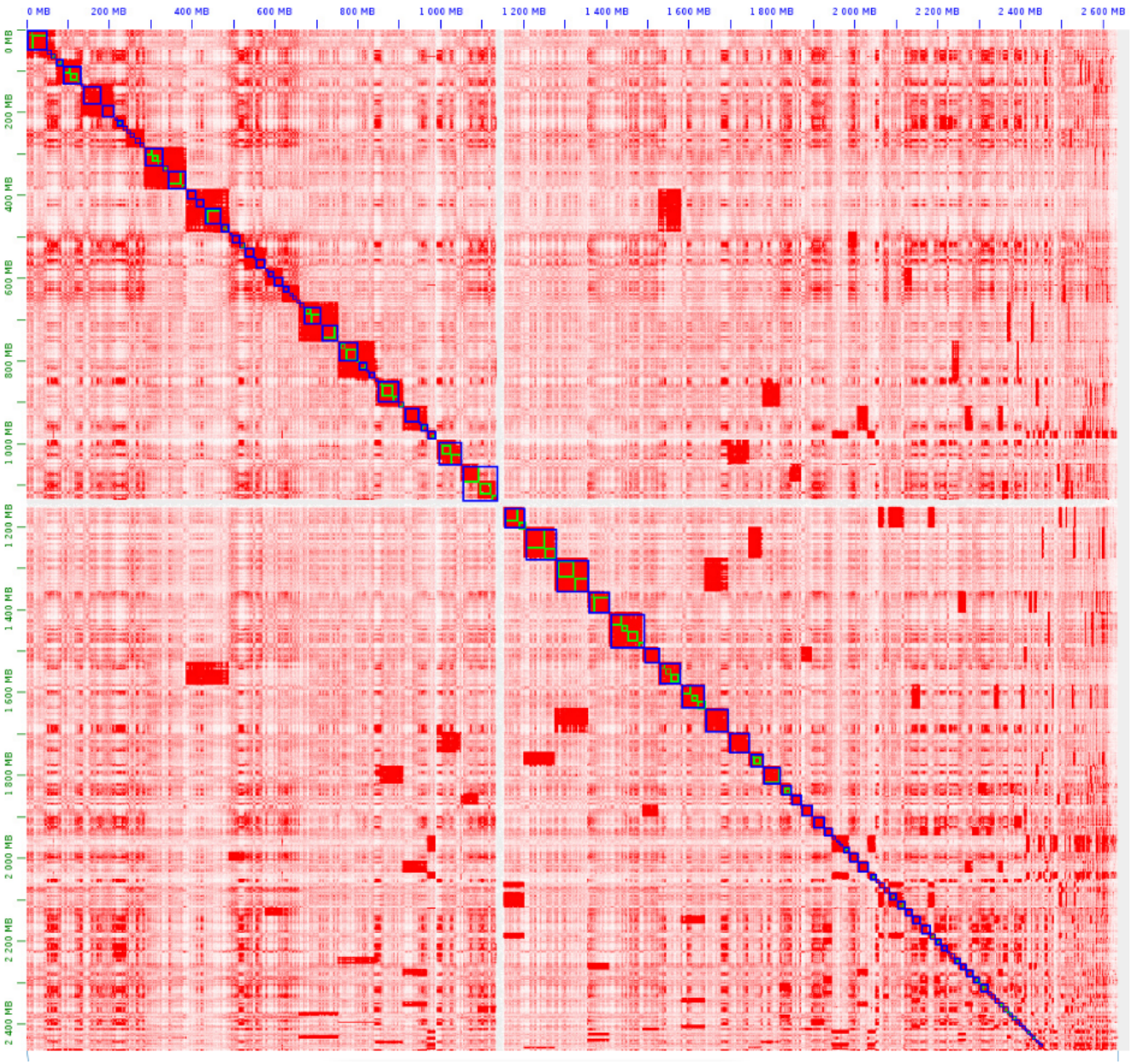


[Lieberman-Aiden et al., Science, 2009]

Hi-C heat map

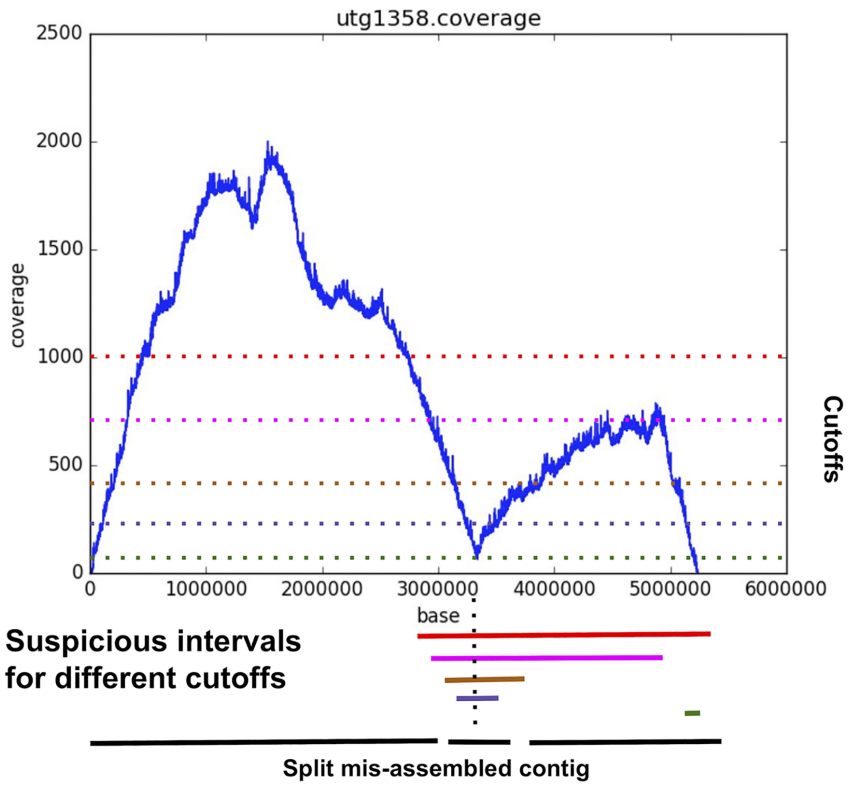

0 MB
200 MB
400 MB
600 MB
800 MB
1 000 MB
1 200 MB
1 400 MB
1 600 MB
1 800 MB
2 000 MB
2 200 MB
2 400 MB
2 600 MB

Hi-C heat map

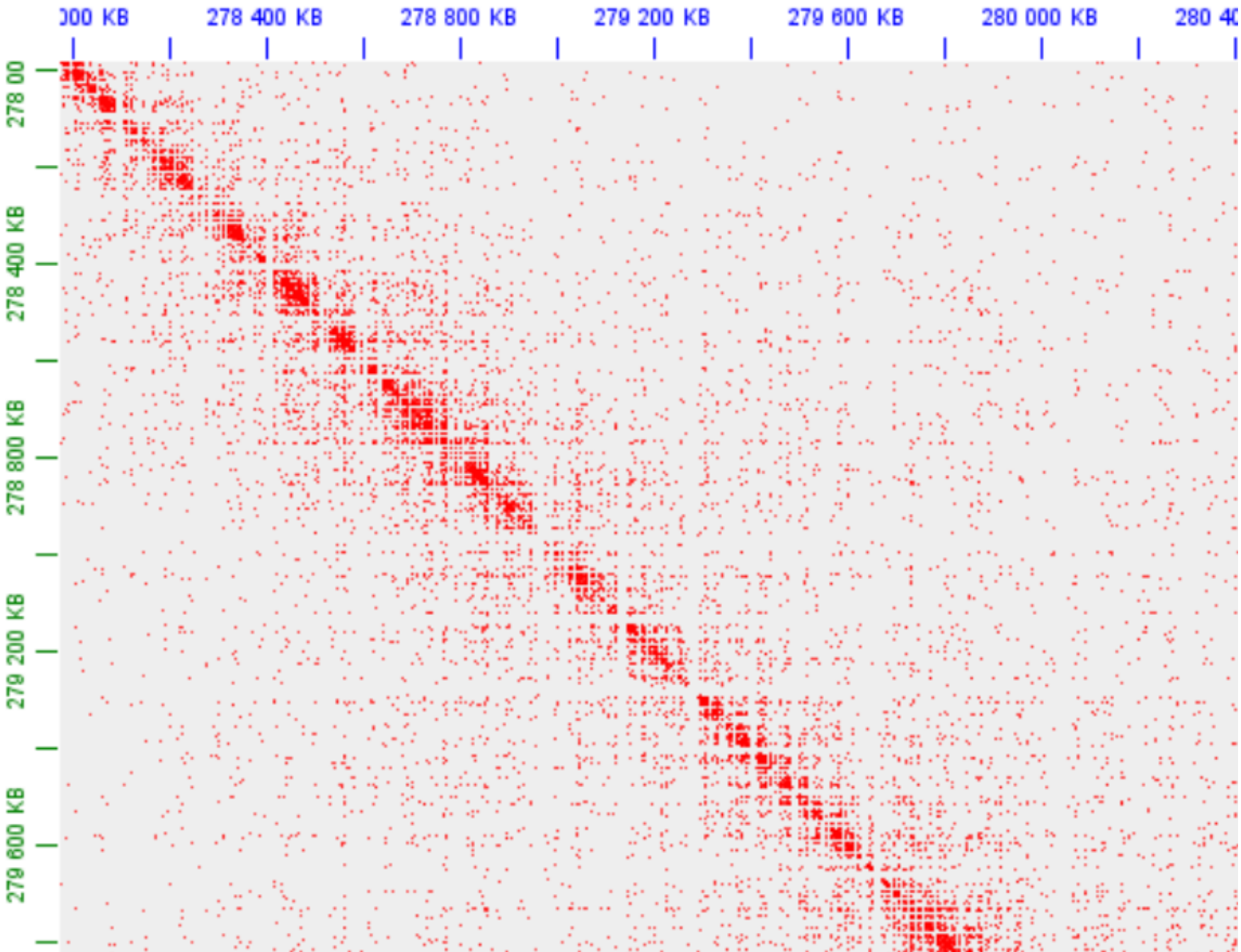

0 MB
200 MB
400 MB
600 MB
800 MB
1 000 MB
1 200 MB
1 400 MB
1 600 MB
1 800 MB
2 000 MB
2 200 MB
2 400 MB
2 600 MB

Hi-C scaffolding methods

- align reads on contigs
- split contigs
- scaffold construction

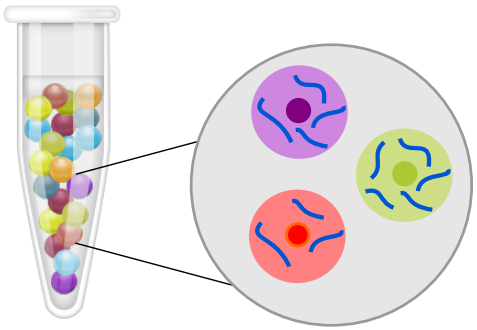


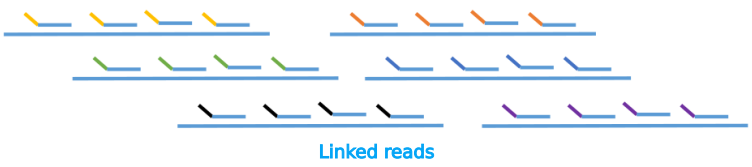
Hi-C heat map (zoom)



Can we use a more homogeneous read coverage?

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

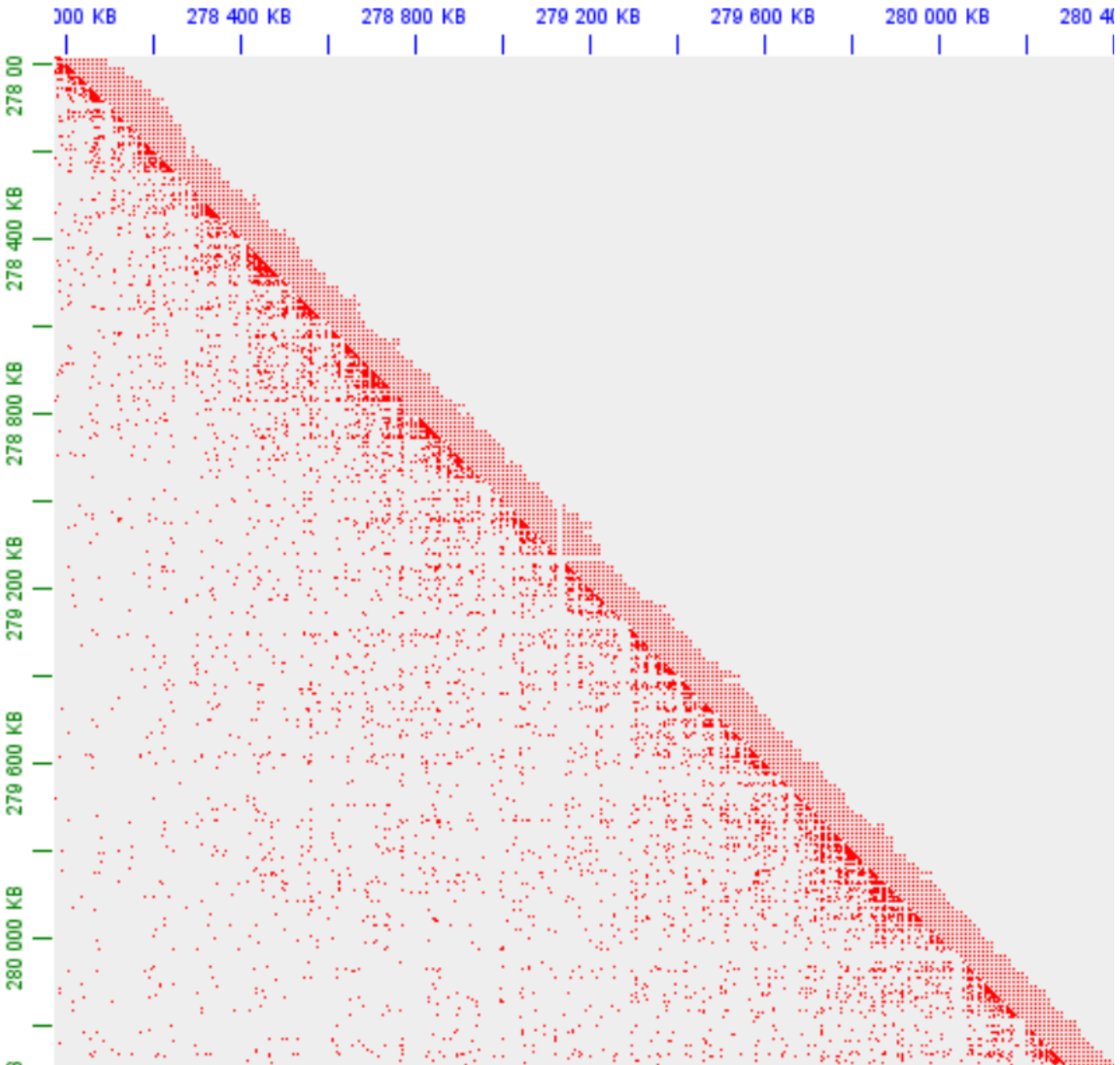




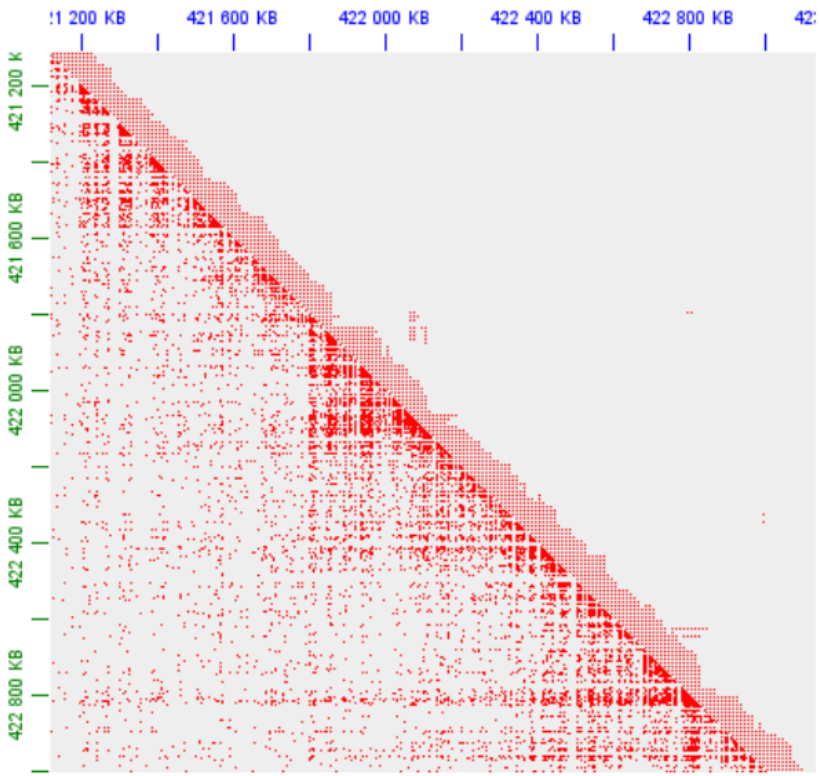
- generated from long DNA molecules
- tagged with a specific barcode
- computational reconstruction of single molecules
- provides low-cost long-range information

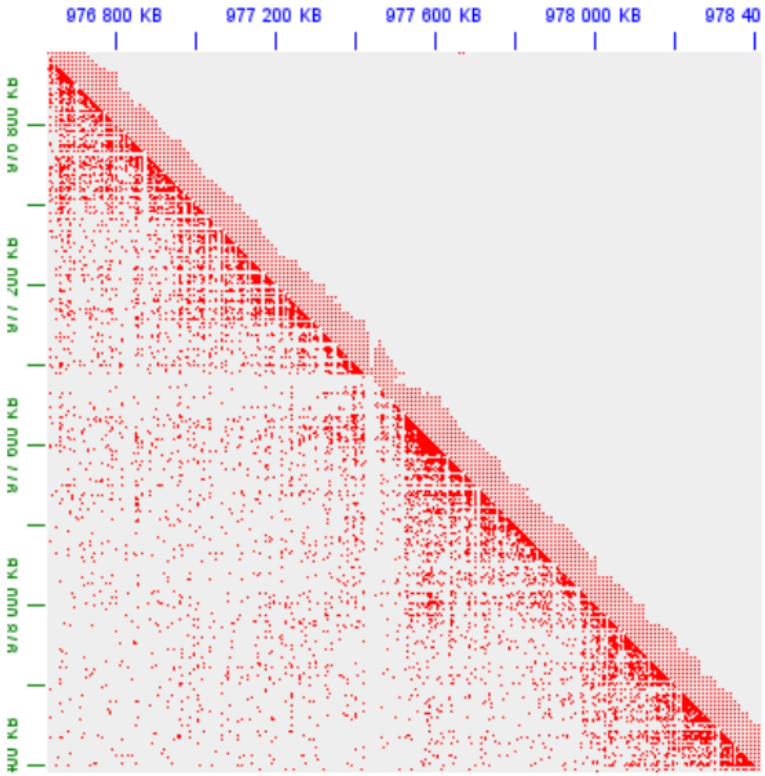
Source: 10x Genomics

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?

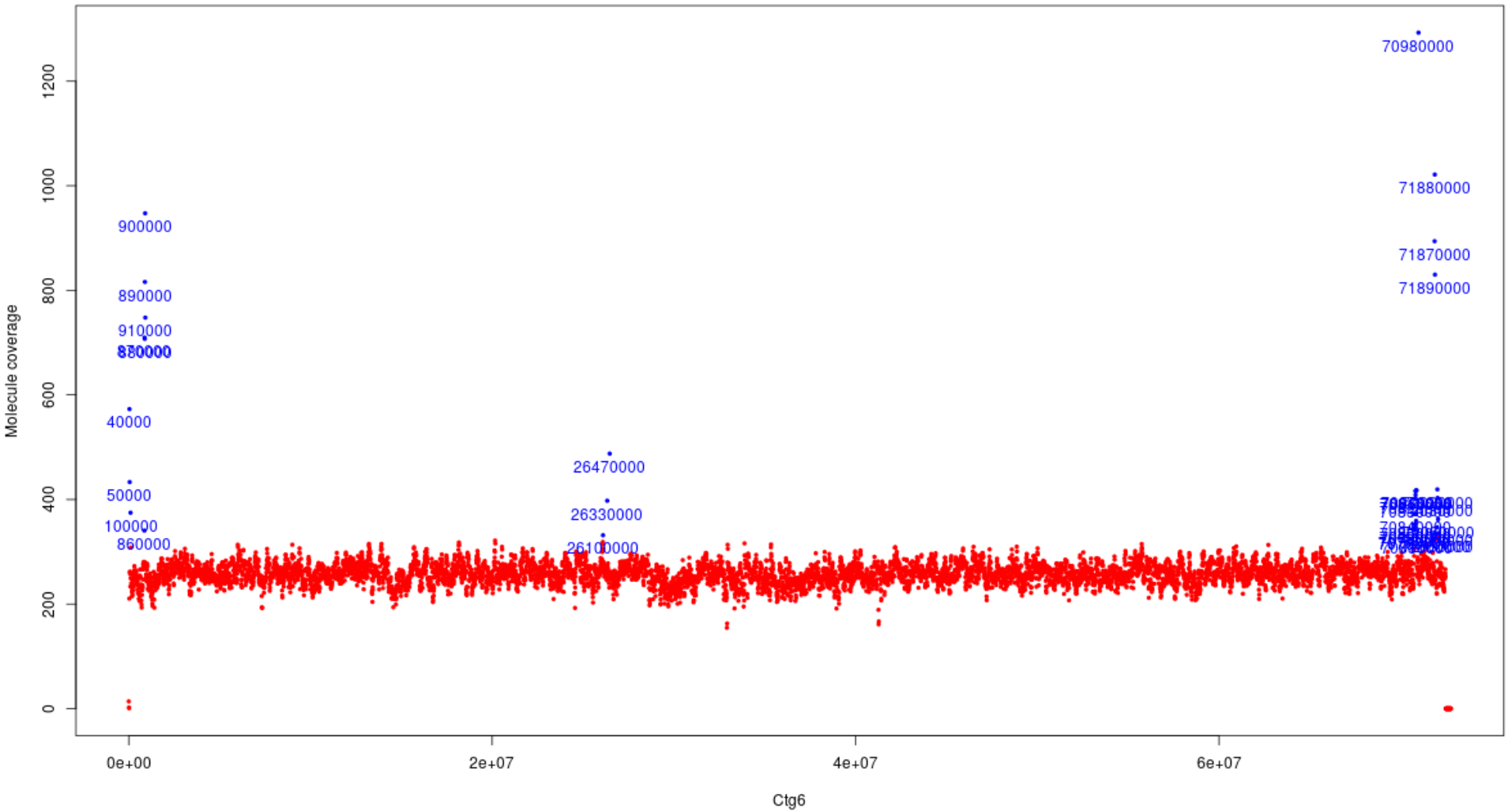
Tigmint [S.D. Jackman et al., BMC Bioinformatics, 2018]

- Checks each window for a minimum number of spanning molecules
- Cuts contigs at windows with insufficient coverage surrounded by windows with sufficient coverage.

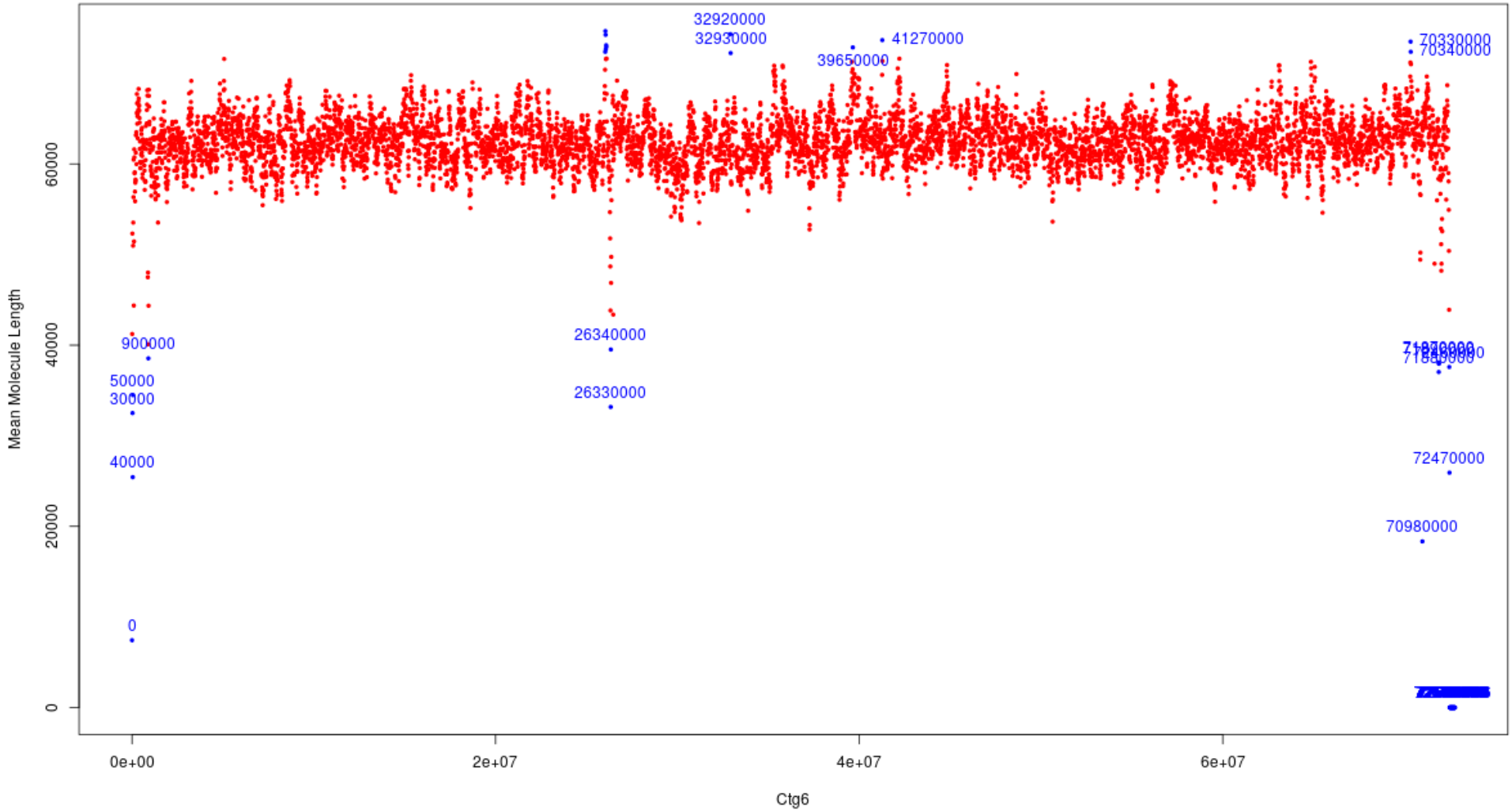
Problem

Not adapted to contigs assemblies based on long reads which are more accurate => no correct cuts in our assemblies

Molecule coverage on a contig (10kb interval)



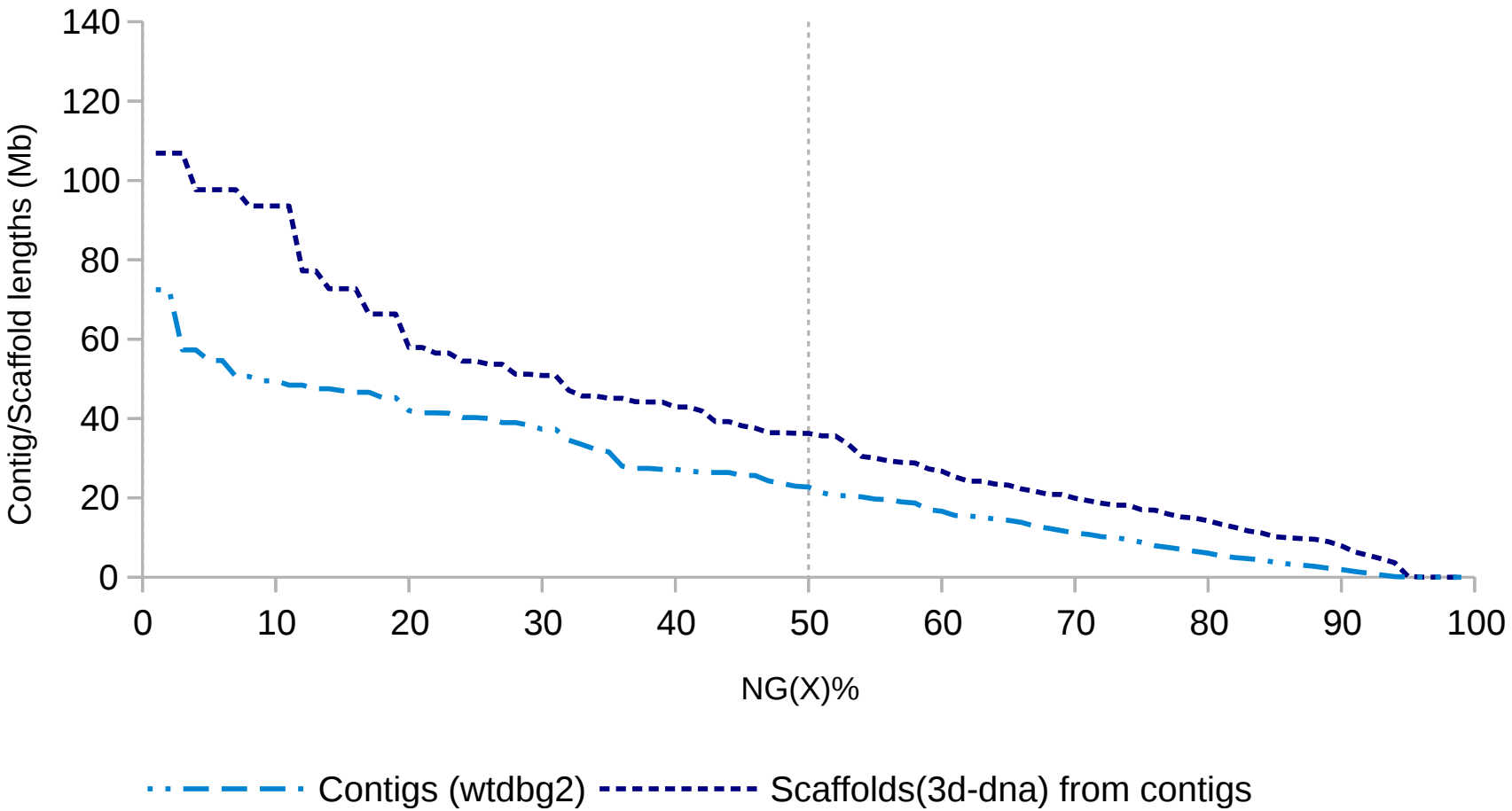
Mean molecule length on a contig (10kb interval)



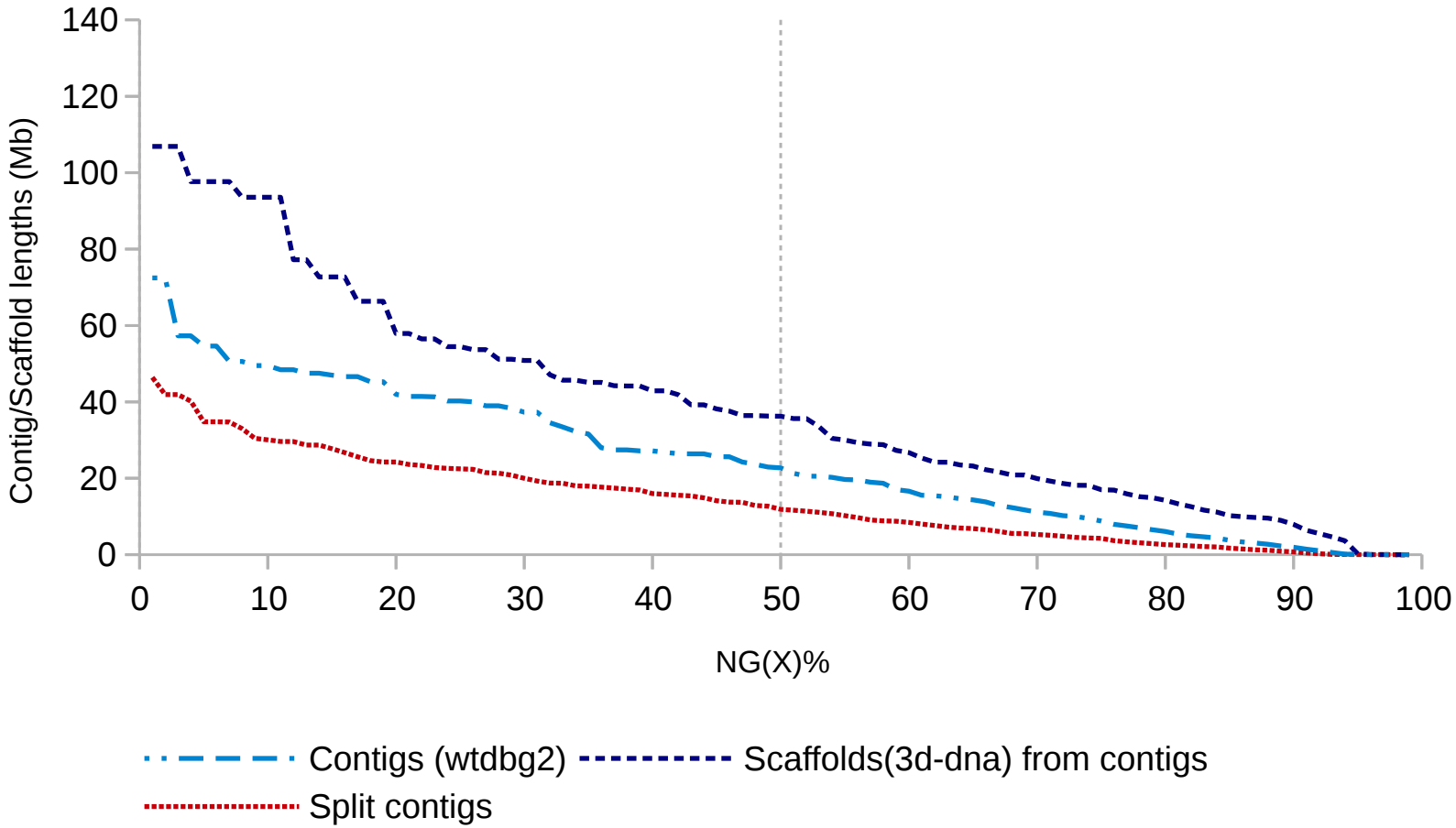
Split contigs with 10x linked reads

- Align linked-reads on contigs
- Identify molecules (barcode, beginning and ending position, number of reads)
- Compute molecule profiles per interval (10kb)
  - Number of starting molecules
  - Number of ending molecules
  - Molecule coverage
  - Mean read density/molecule
  - Mean molecule length
- Identify outliers intervals and split contigs
- Re-connect contigs with Hi-C scaffolding methods

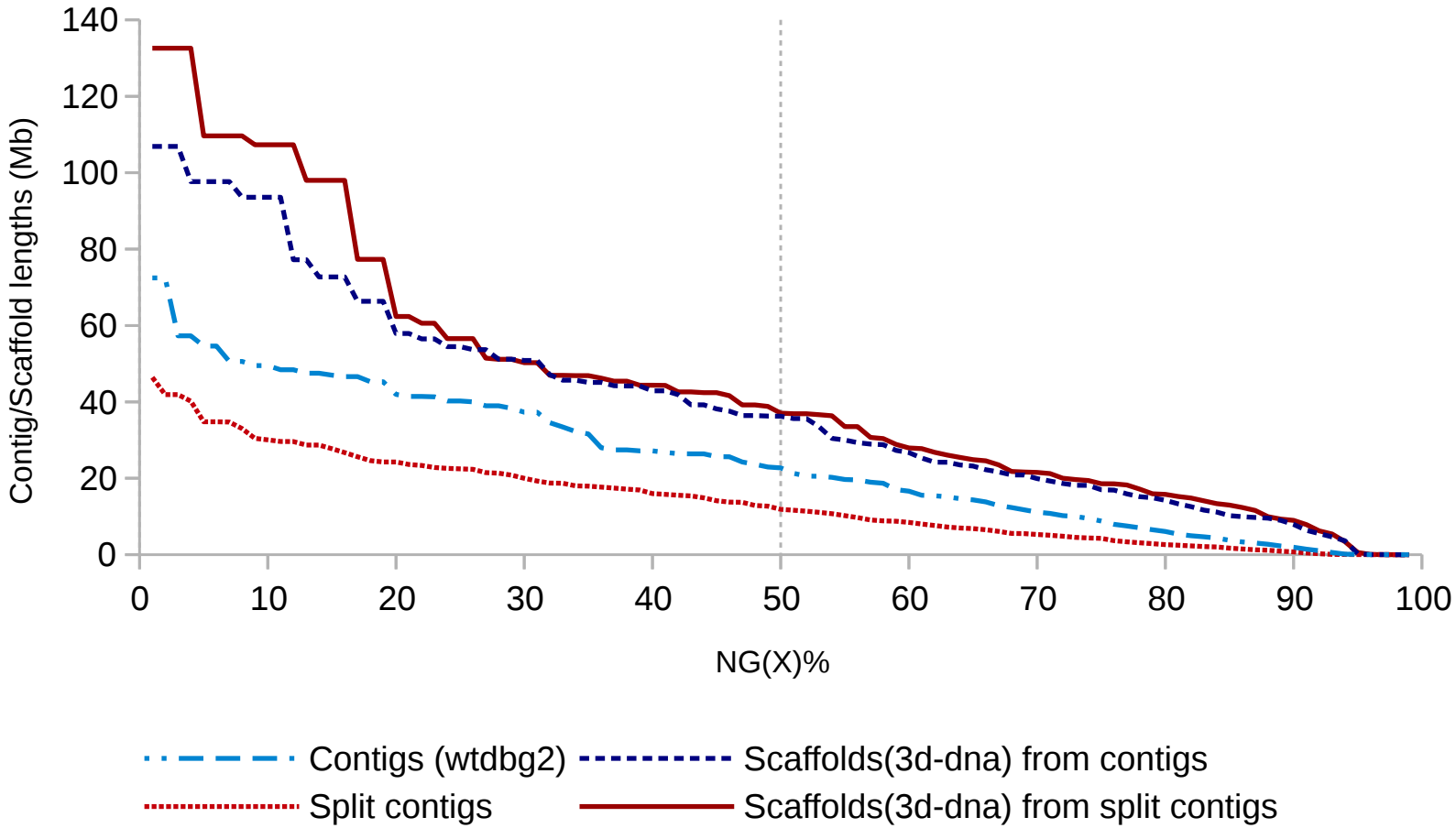
Results without splitting



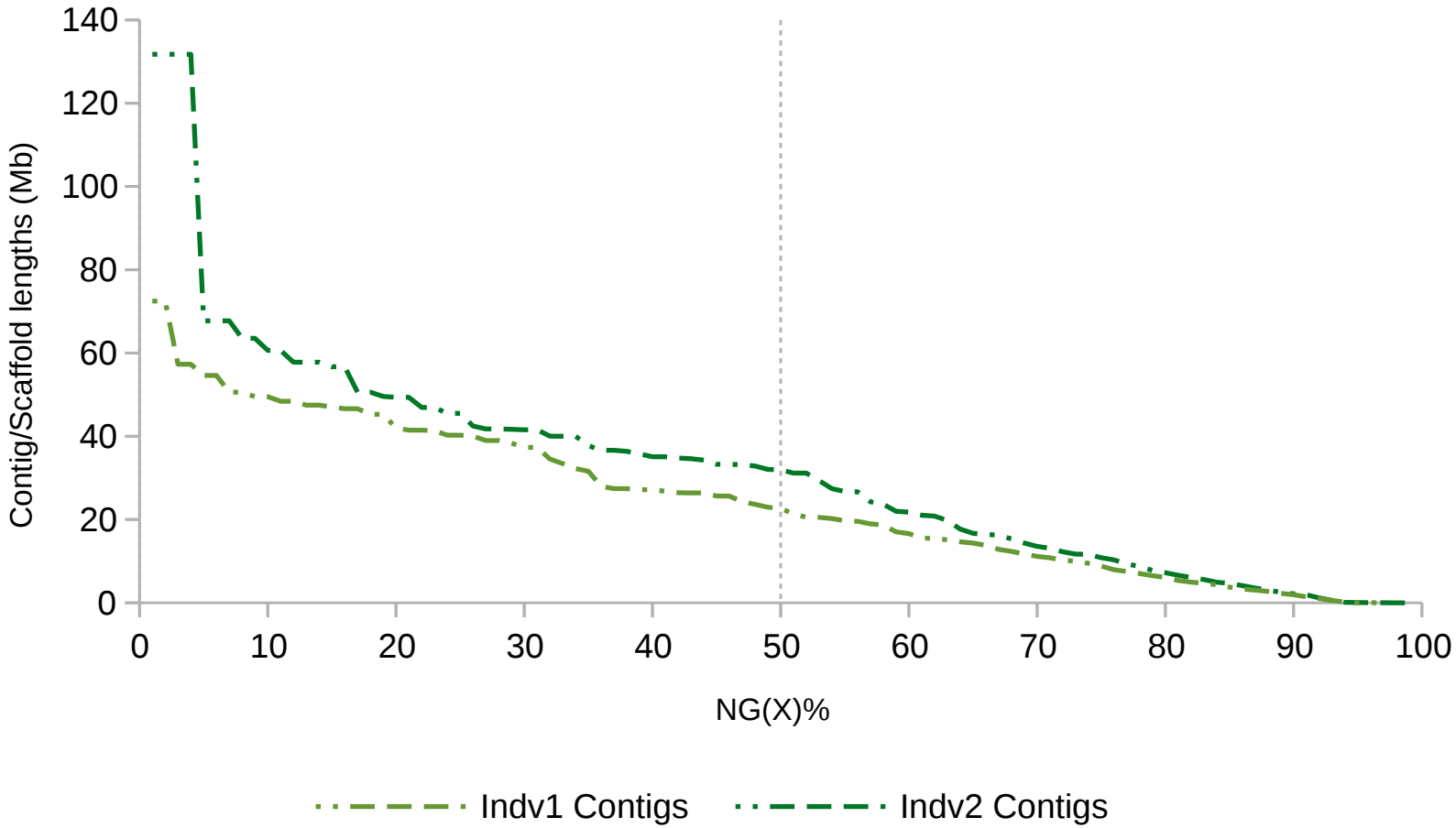
Impact on contig length



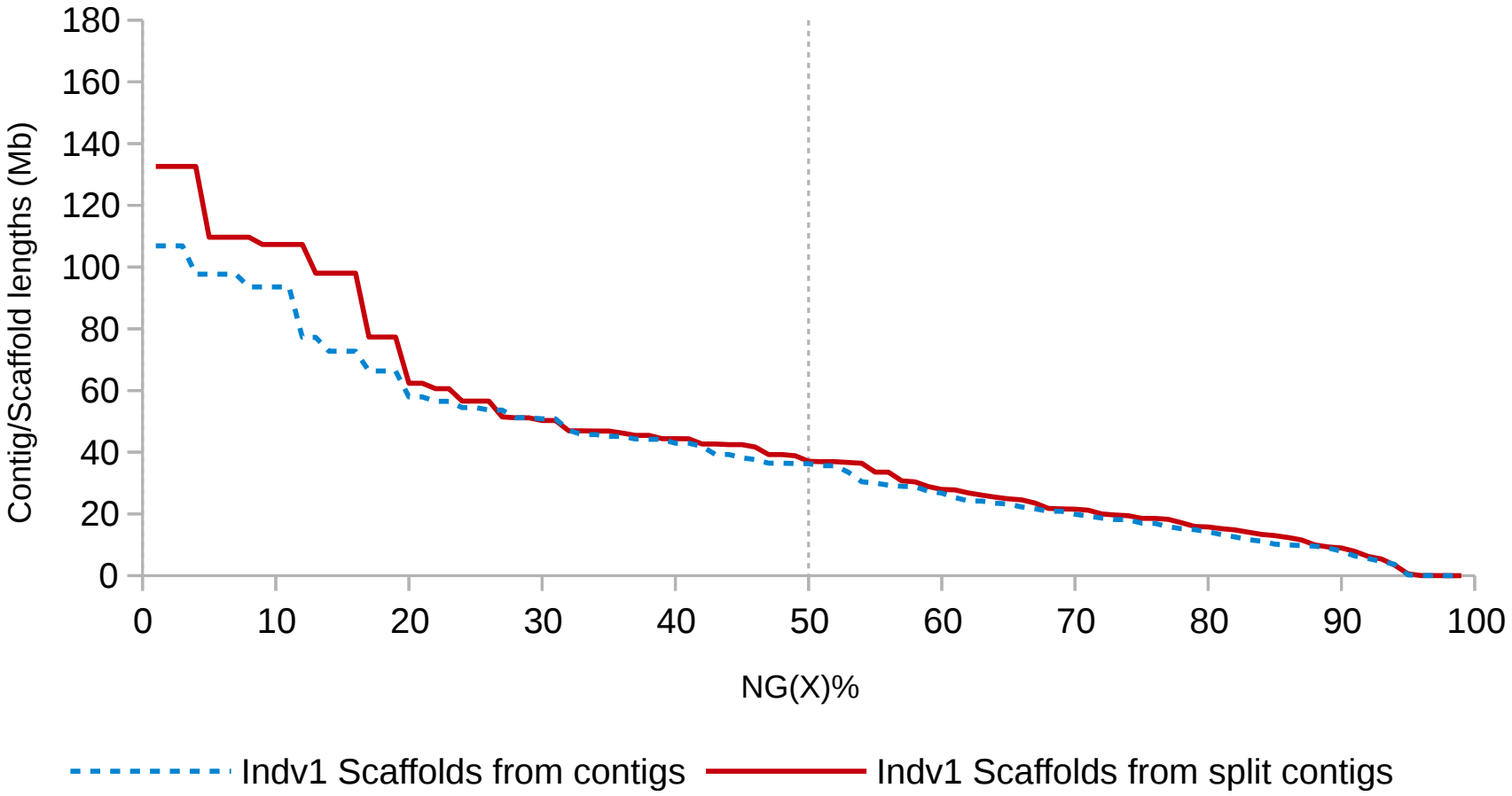
Impact on scaffold length



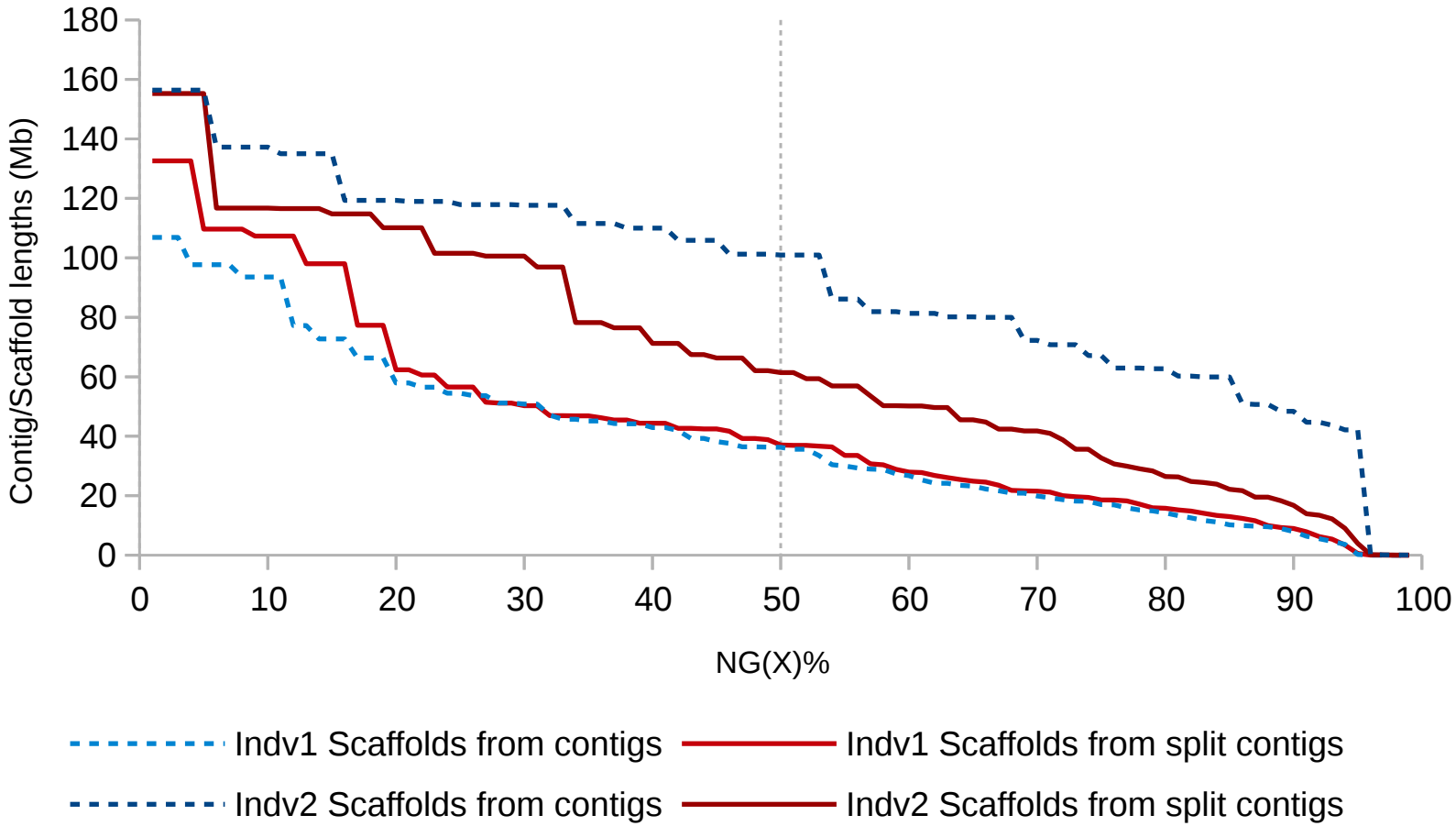
Impact on scaffold length on a high contiguity assembly



Impact on scaffold length on a high contiguity assembly



Impact on scaffold length on a high contiguity assembly



Future improvements

- Stronger constraints for a split
- Adding a scaffolding step based on linked reads
  - build graph based on molecules profiles
  - connect branchless paths of contigs