SeqOccIn

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

Axis 1
Genome assembly

Genotoul GeT

http://get.genotoul.fr
@get_genotoul

http://bioinfo.genotoul.fr

Genotoul Bioinfo

Genome assembly project

Context: New advances in the field of genomic sequencing

Aims:

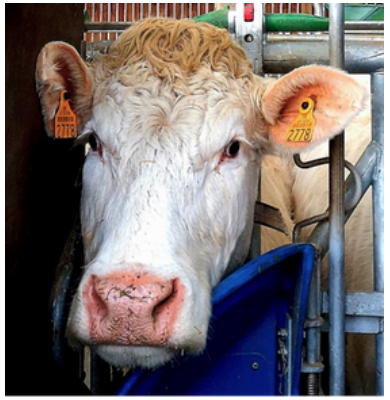
1. Identify the best practice to obtain the highest quality assembly at a specific price
2. Development of new methods for de novo assembly

Data: Oxford Nanopore, Pacific Bioscience, Hi-C, 10x Chromium, Bionano optical mapping …

Species: Cattle, Maize, Pig, Quail, Goat, Sheep, …

SeqOccIn

Genomes specificities



- 2.7 Gb genome length
- 30 Chromosomes
- 40 % Repetitive regions
- 2 Trio (6 Animals)

- 2.4 Gb genome length
- 10 Chromosomes
- 85 % Repetitive regions
- 4 Strains

- 2.8 Gb genome length
- 38 Chromosomes
- 35 % Repetitive regions
- 1 Animal

- 0.93 Gb genome length
- 33 Chromosomes
- 15 % Repetitive regions
- 1 Animal

Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES L'Europe s'engage en Midi-Pyrénées

SeqOccIn

Available data

Chromosome fragment

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

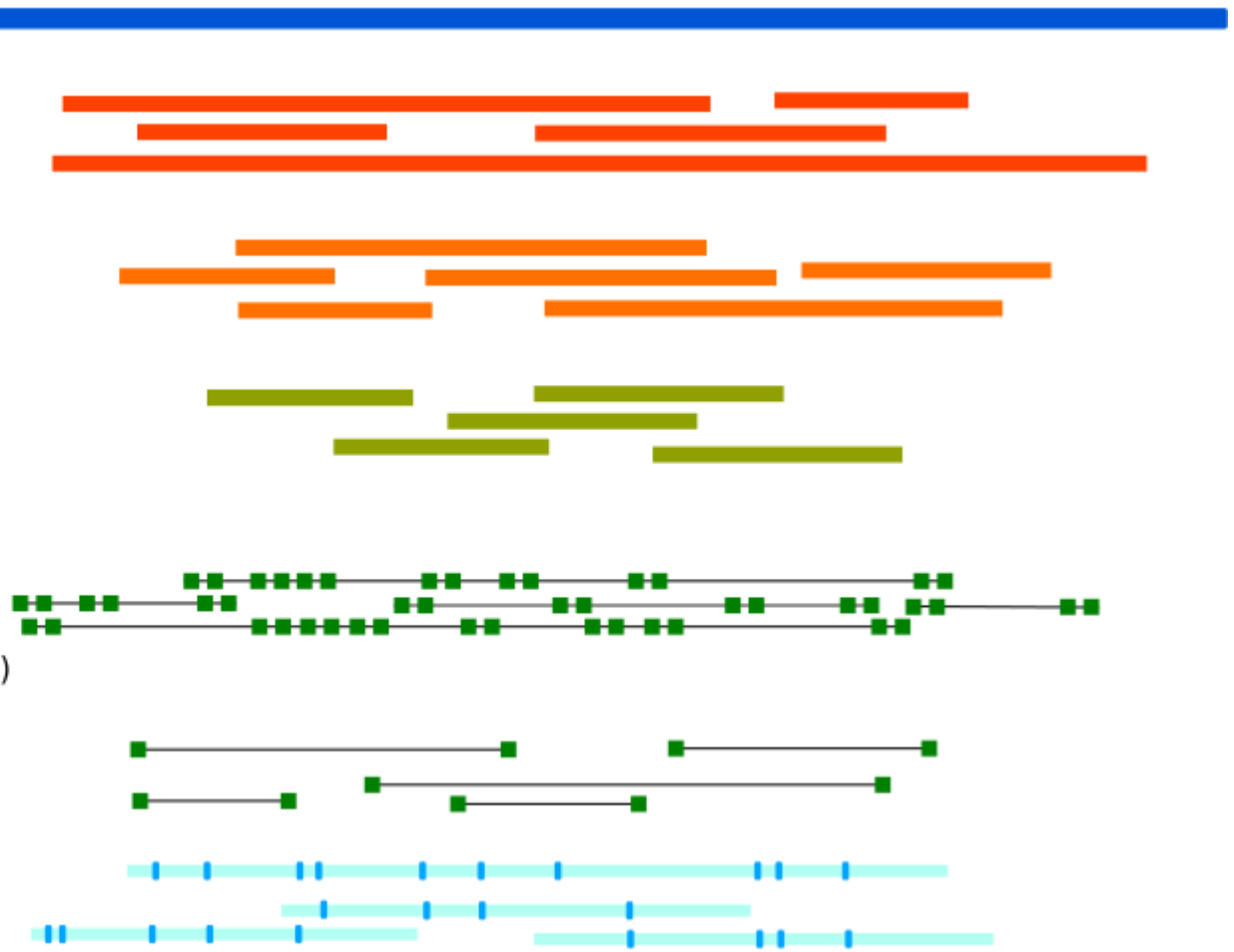
PacBio CLR
~15% errors
50 kb N50 (up to ~200 kb)

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)

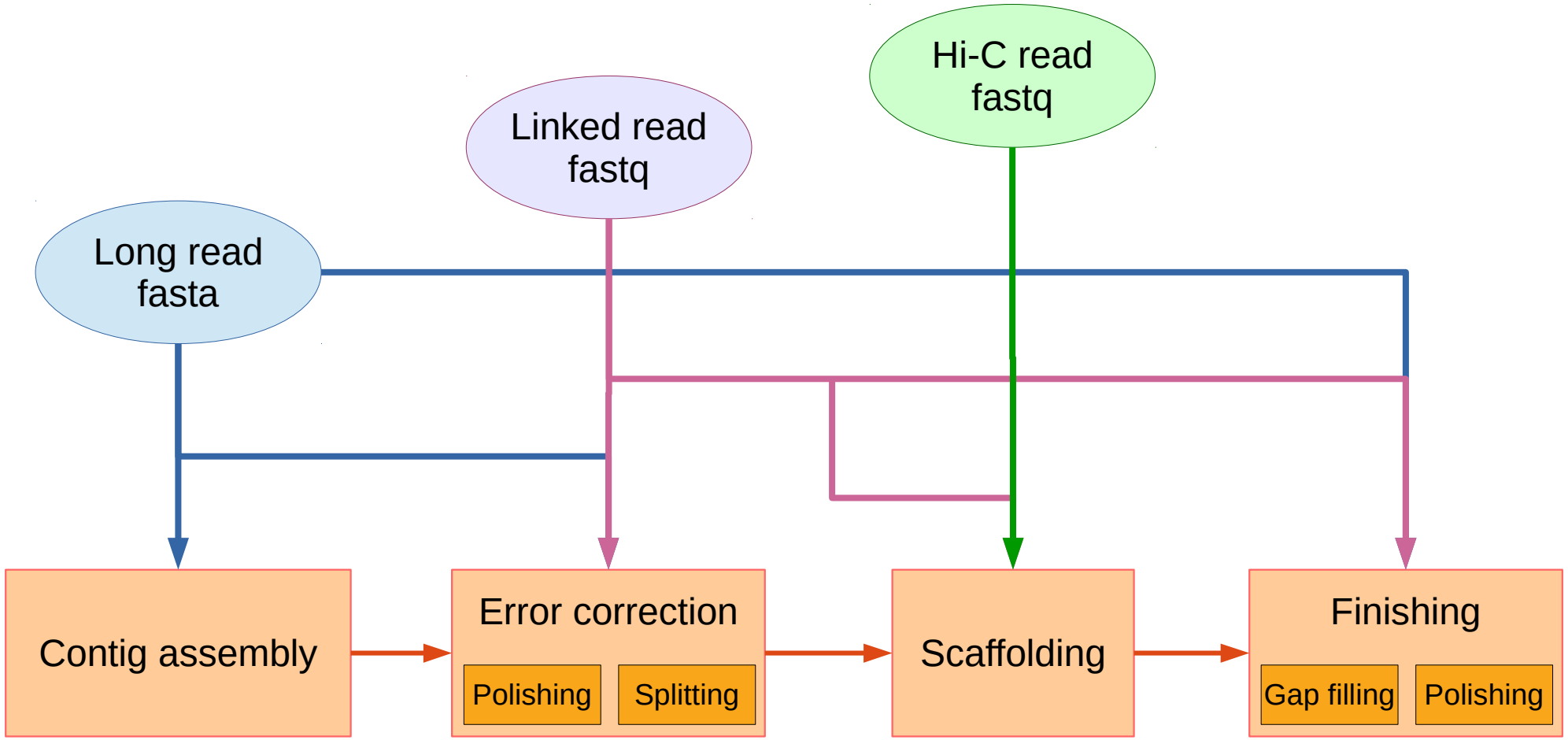
Bionano optical mapping
~300kb molecule length



Genotoul GeT Genotoul Bioinfo

SeqOccIn

Current assembly pipeline

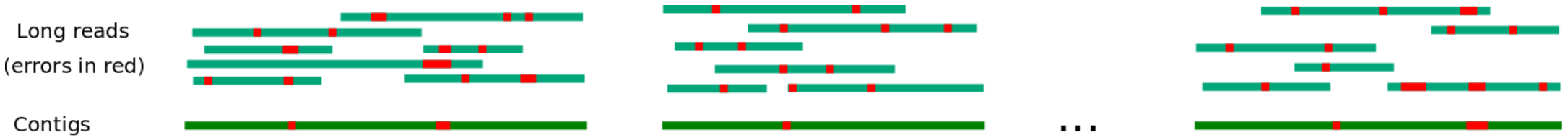


Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES l'Europe s'engage en Midi-Pyrénées

Contig assembly

Construction of long chromosomial parts without N's based on overlapping reads



Assemblers tested: - **wtdbg2** - best ratio between contigs quality, run time and memory required
- flye - similar results to wtdbg2 but longer run time
- shasta - requires very long reads
- canu - requires high coverage and large disk space
- ...

SeqOccIn

Assembly evaluation

N50: defines assembly quality in terms of contiguity. Given a set of contigs, the N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.

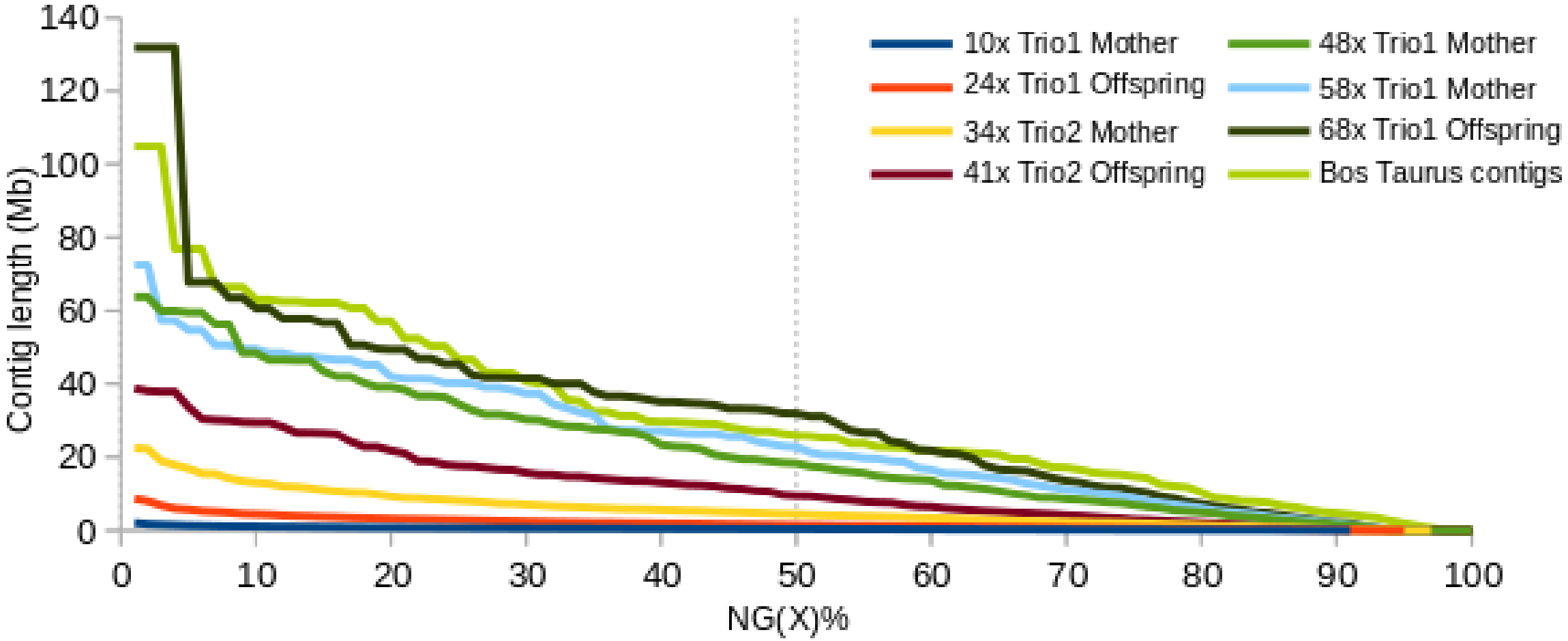
BUSCO score: assesses genome assembly completeness by finding a particular set of near-universal single copy orthologs

C: complete (S : single-copy and D : duplicated)
F: fragmented
M: missing

Contigs N50: 20-25 Mb
Contig's BUSCO score: C:67.9%[S:67.4%,D:0.5%],F:11.7%,M:20.4%

Genotoul GeT Genotoul Bioinfo
INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES L'Europe s'engage en Midi-Pyrénées

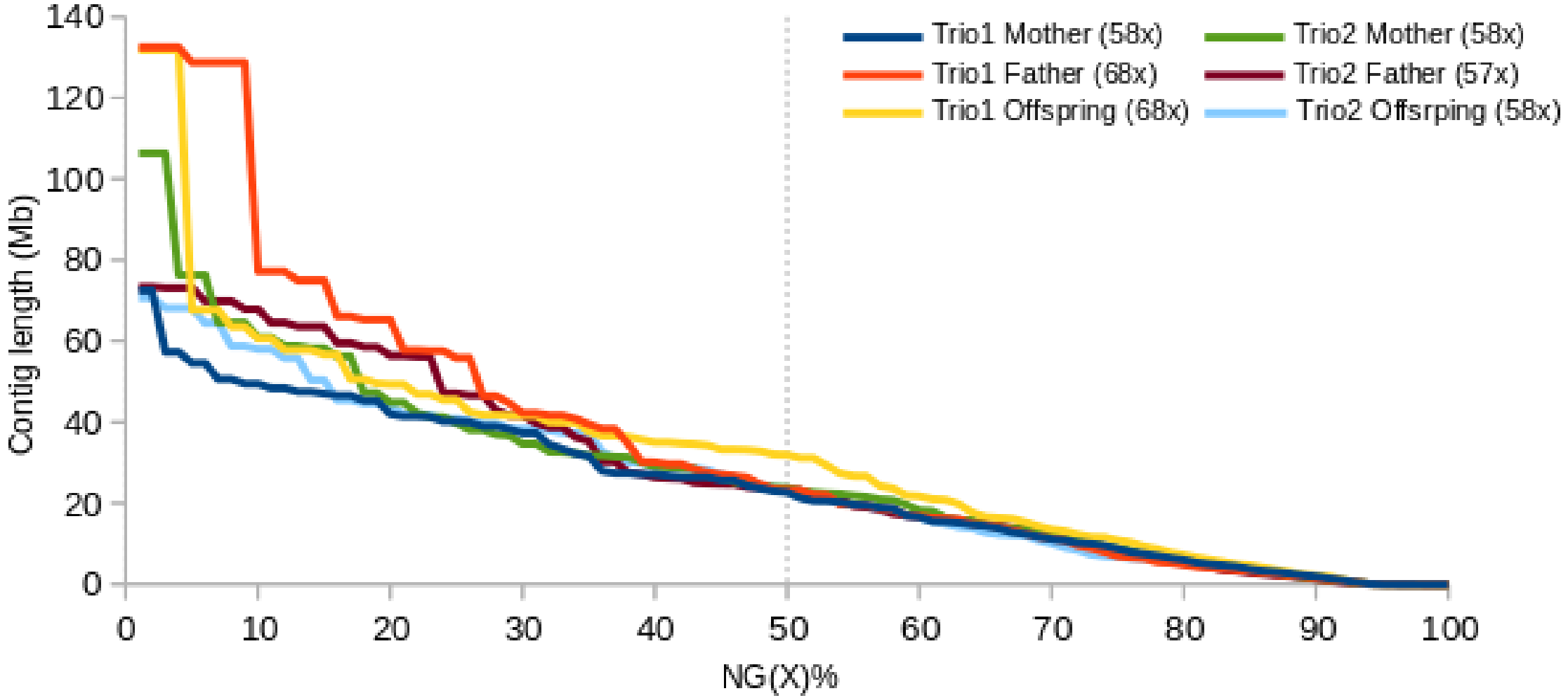
Coverage impact on female bovine assembly



Genotoul GeT
Genotoul Bioinfo
INRA SCIENCE & IMPACT
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE
RÉGION MIDI-PYRÉNÉES
l'Europe s'engage en Midi-Pyrénées

SeqOccIn

Maximum coverage results

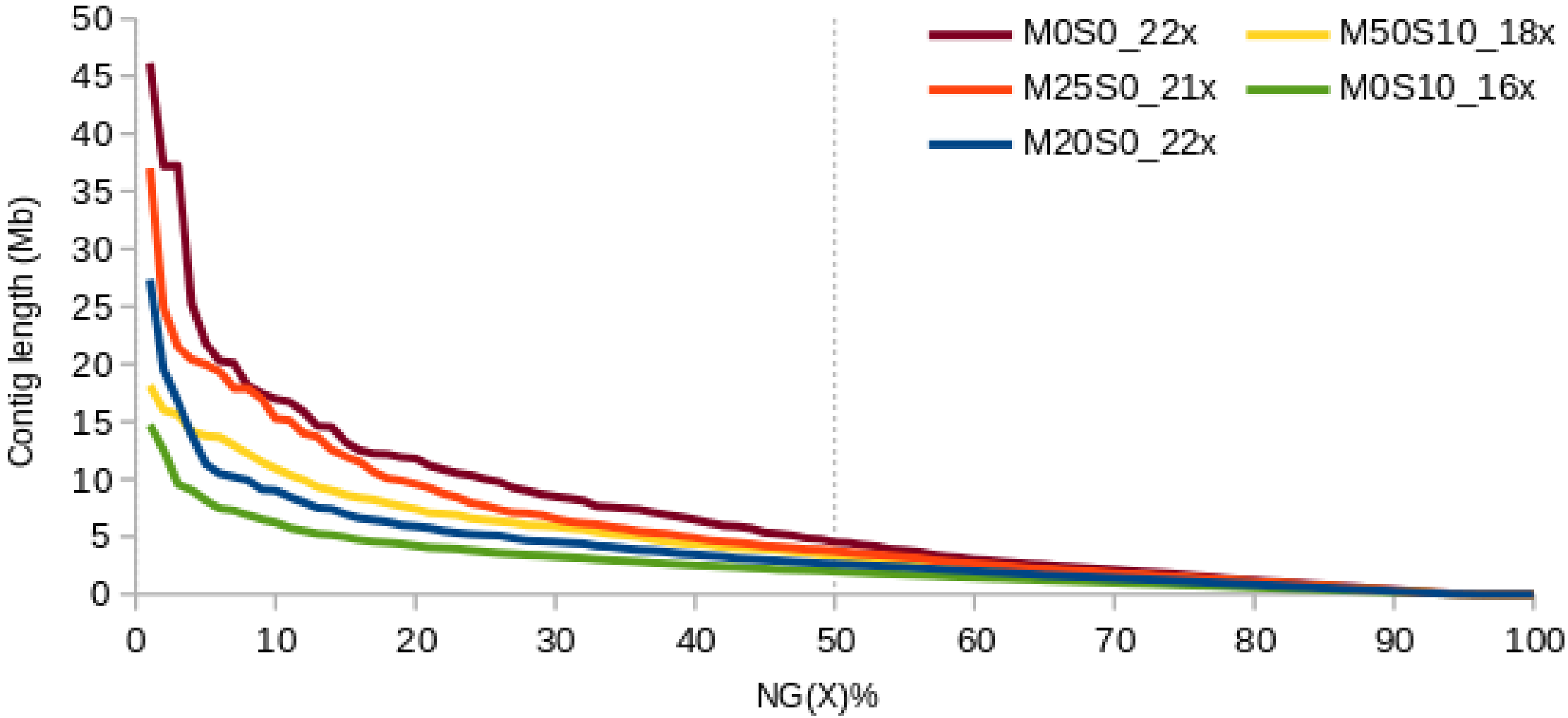


Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES L'Europe s'engage en Midi-Pyrénées

SeqOccIn

Protocol impact on contig length

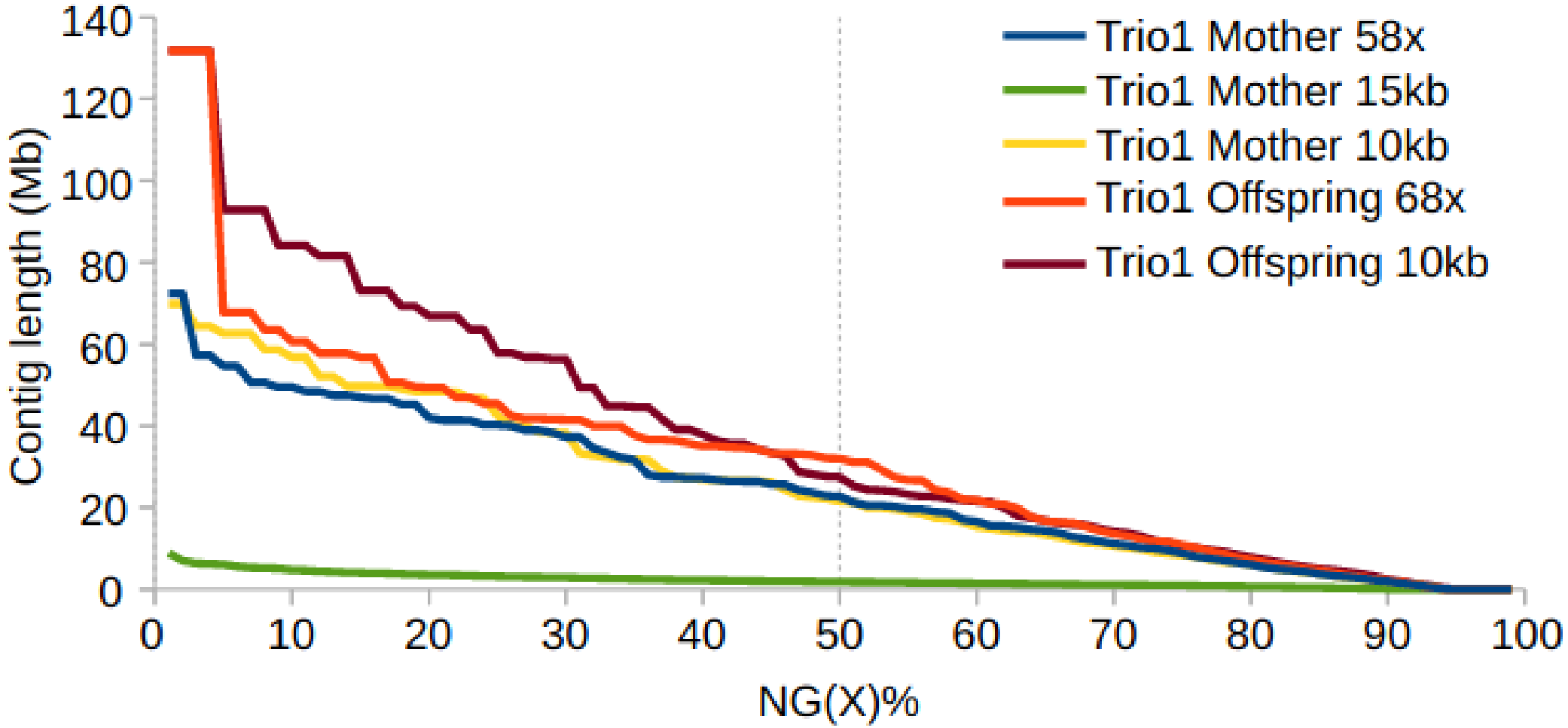


Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES l'Europe s'engage en Midi-Pyrénées

SeqOccIn

Bioinformatic read length filter



Genotoul GeT Genotoul Bioinfo

SeqOccIn

Error correction

Two types of errors in contigs :

- sequence errors due to sequencing => corrected through **polishing**
- connection errors due to assembler choices => corrected through **splitting**

Steps in the error correction pipeline:

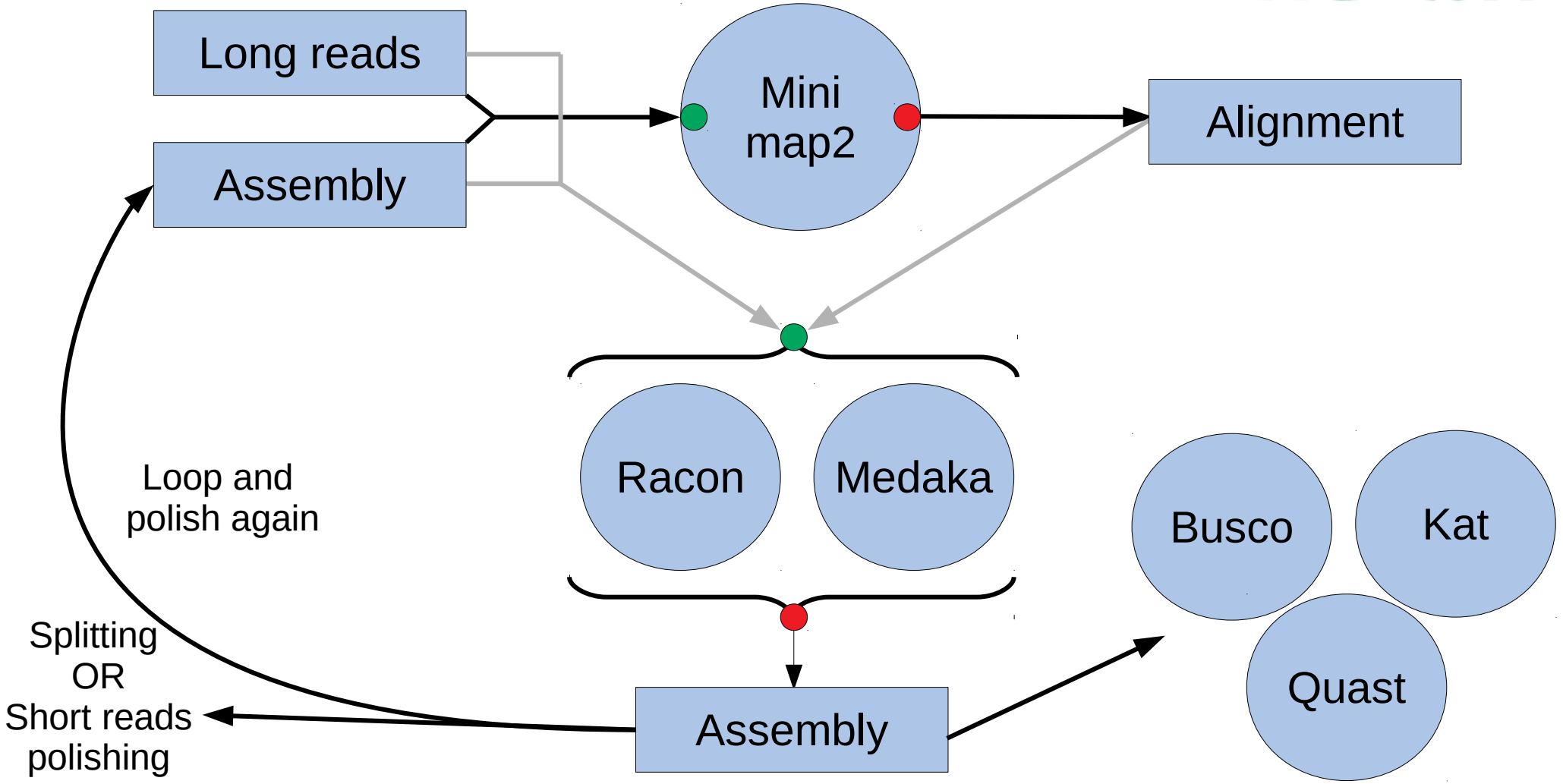
1. **Polishing based on long reads –** align long reads to contigs and build consensus
2. **Splitting based on long reads –** remove contigs regions covered by less than 4 reads

   remove sequence extension errors made by polisher
3. **Polishing based on short reads –** align short reads to contigs and build consensus

Genotoul GeT Genotoul Bioinfo
INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES l'Europe s'engage en Midi-Pyrénées

Long read polishing

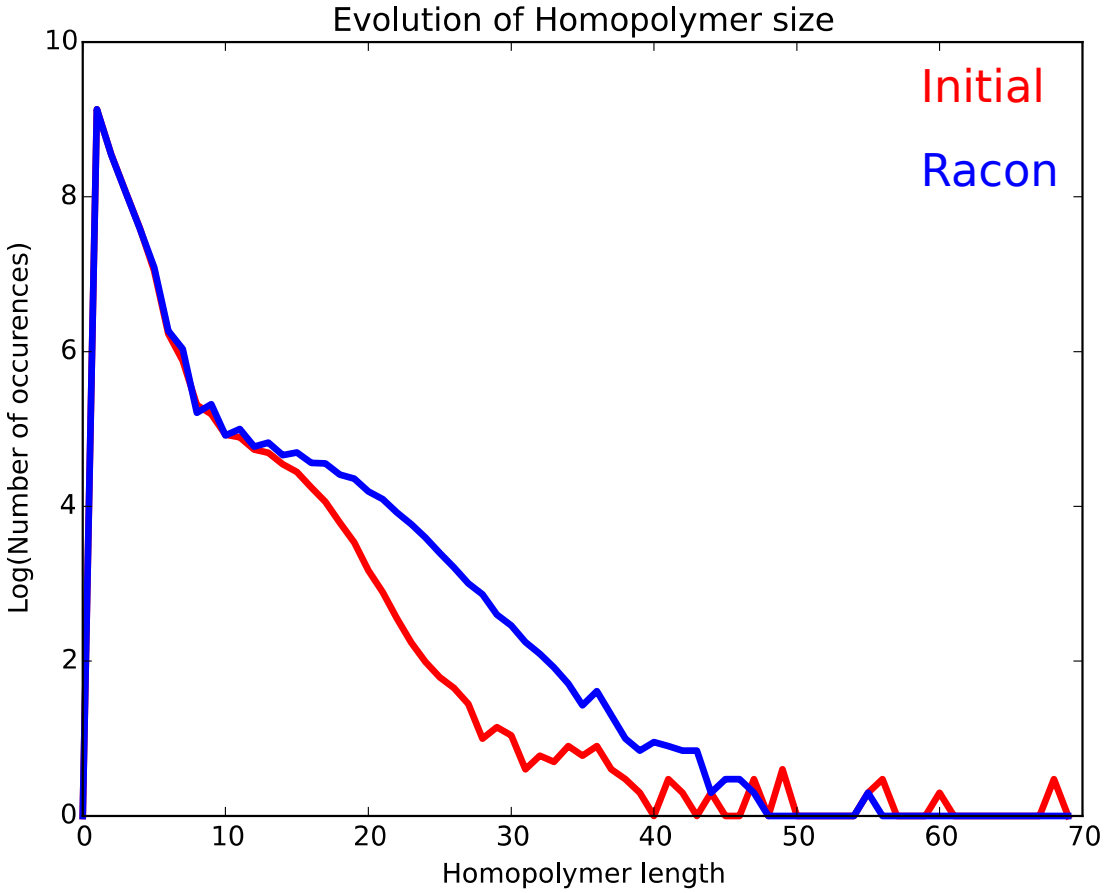
nextflow



Genotoul GeT
Genotoul Bioinfo
INRA SCIENCE & IMPACT
Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE
RÉGION MIDI-PYRÉNÉES
l'Europe s'engage en Midi-Pyrénées

SeqOccIn

Long read polishing



<u>BUSCO score after contig assembly:</u>
C:67.9% [S:67.4%,D:0.5%],F:11.7%,M:20.4%

<u>BUSCO score after long read polishing:</u>
C:83 % [S:82.1%,D:0.9%],F:9.8%,M:7.1%

<u>Contigs N50 after contig assembly:</u>
20 / 25 Mb

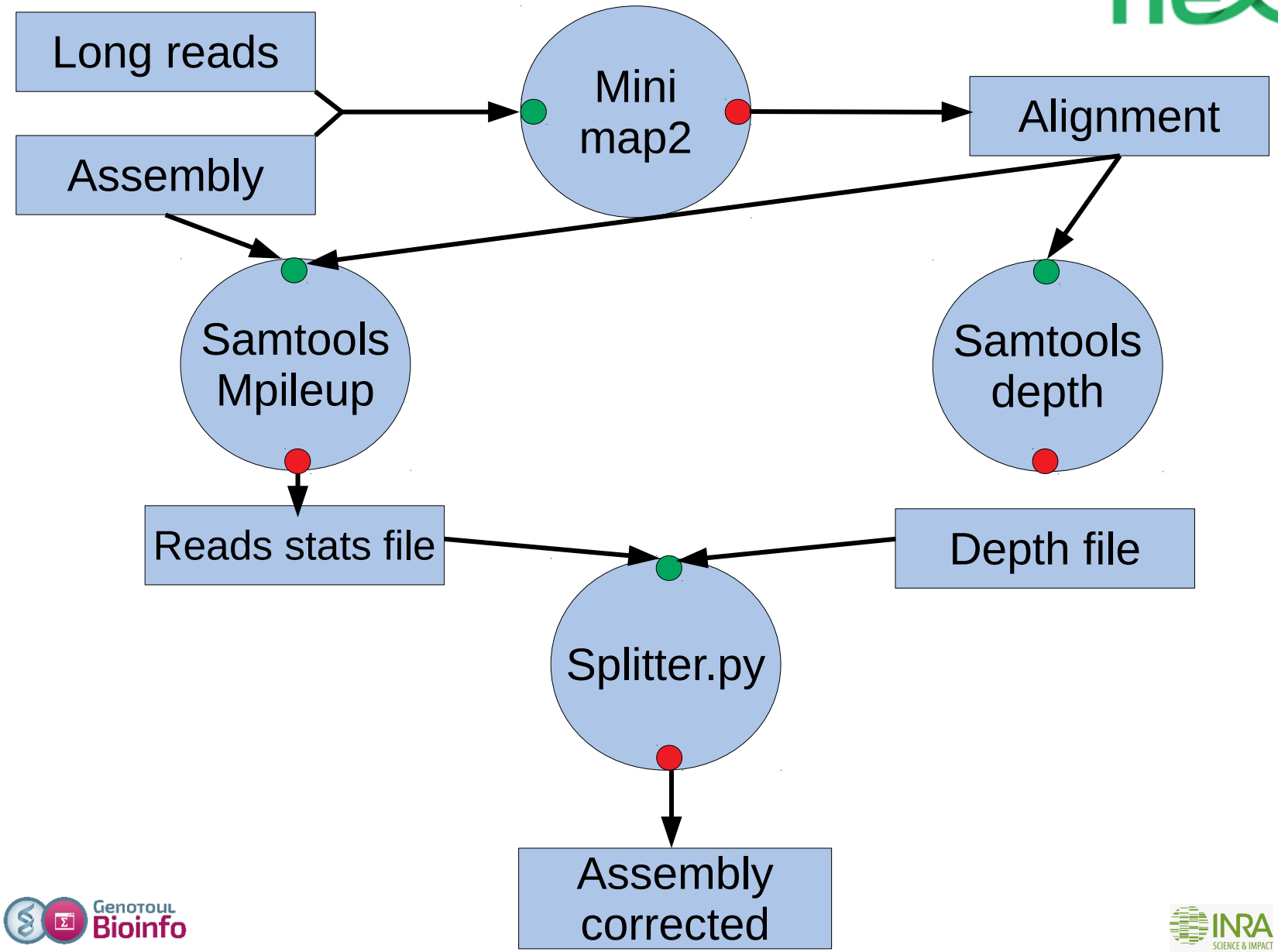
<u>Contigs N50 after long read polishing:</u>
22 / 26 Mb

Genotoul GeT Genotoul Bioinfo

SeqOccIn

Long read splitting

nextflow

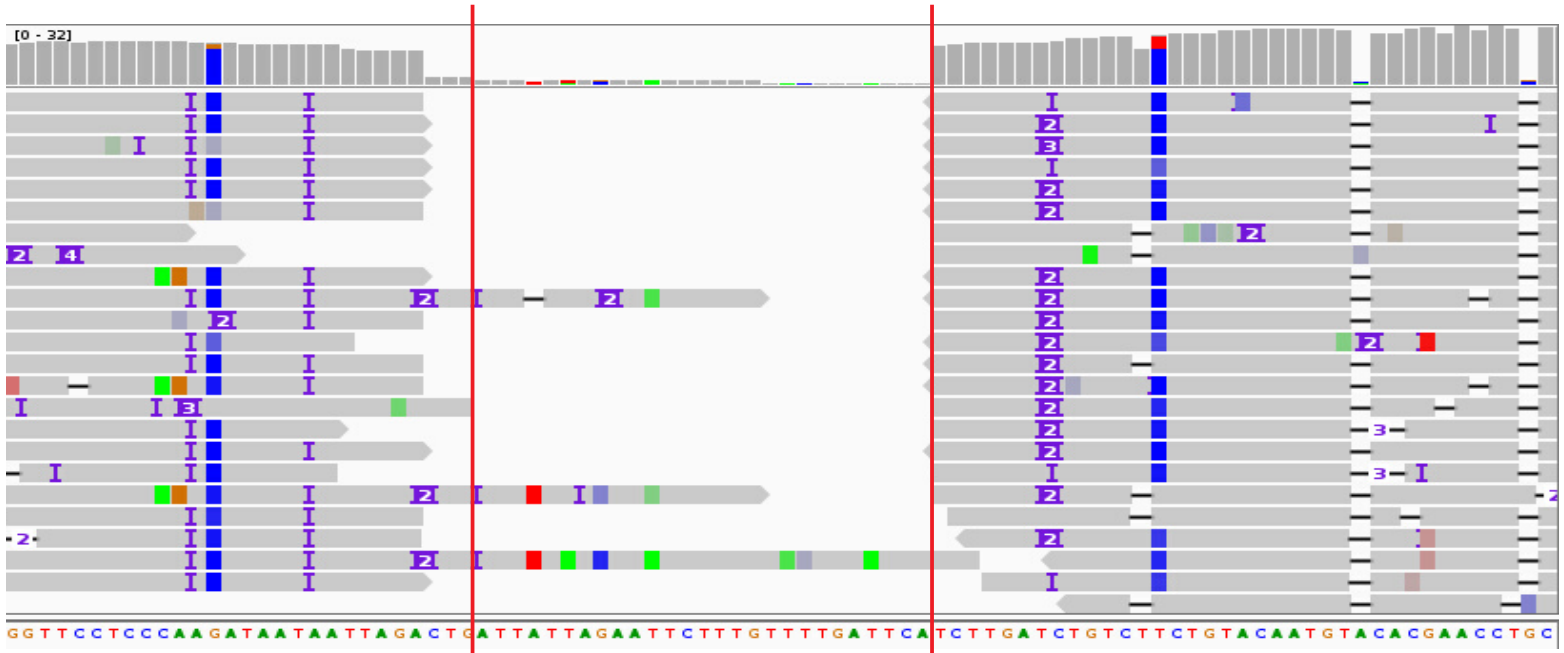


Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES L'Europe s'engage en Midi-Pyrénées

SeqOccIn

Long read splitting



High coverage area
(>4X)

Low coverage area
(<4X)

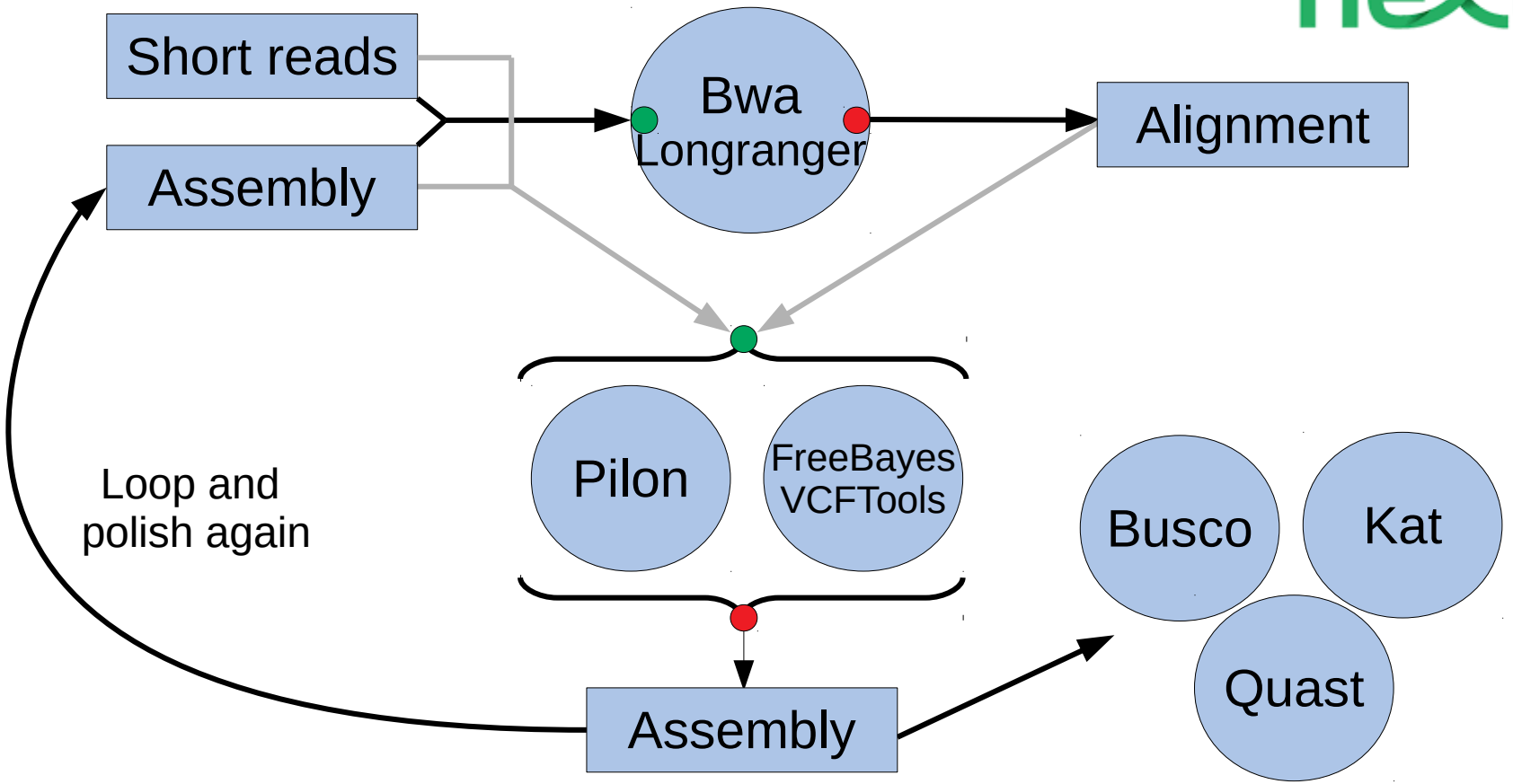
High coverage area
(>4X)

<u>BUSCO score:</u>
C:82.6%[S:81.7%,D:0.9%],F:9.9%,M:7.5%

<u>Contigs N50:</u>
22 / 26 Mb

Genotoul GeT Genotoul Bioinfo

Short read polishing



BUSCO score before short read polishing:

C:82.6%[S:81.7%,D:0.9%],F:9.9%,M:7.5%

BUSCO score after short read polishing:

C:95 %[S:93.9%,D:1.1%],F:3.0%,M:2.1%

SeqOccIn

Scaffolding

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

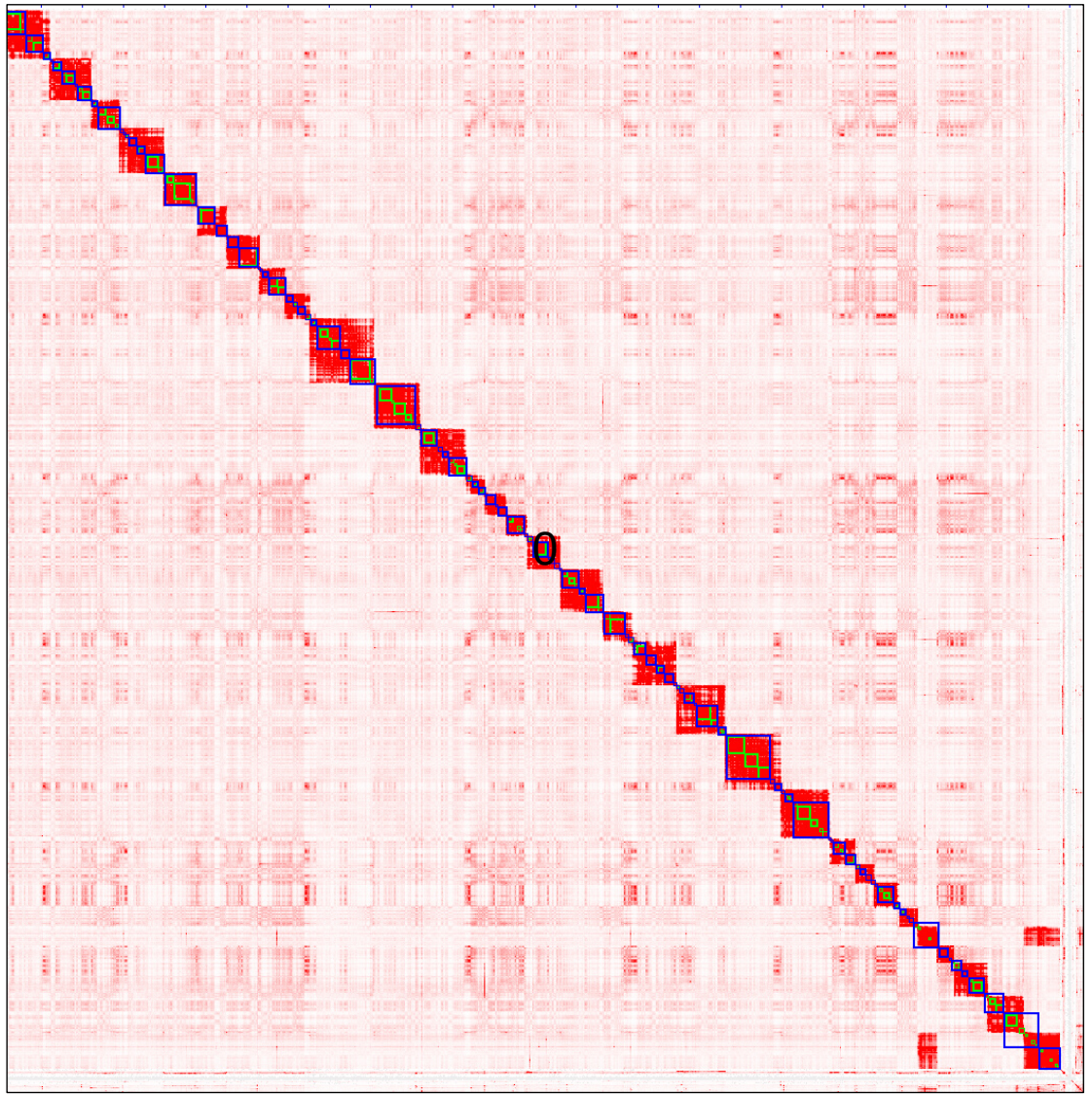
Scaffolders tested: - **3d-dna** - allows easy manual correction of scaffolds)
- Salsa - similar results to 3d-dna on high coverage (>10x) but better results with very low coverage (<7x)
- ...

<u>BUSCO score after scaffolding:</u>
C:95 %[S:93.9%,D:1.1%],F:3.0%,M:2.1%

<u>Scaffold N50:</u>
95 / 110Mb

Genotoul GeT Genotoul Bioinfo

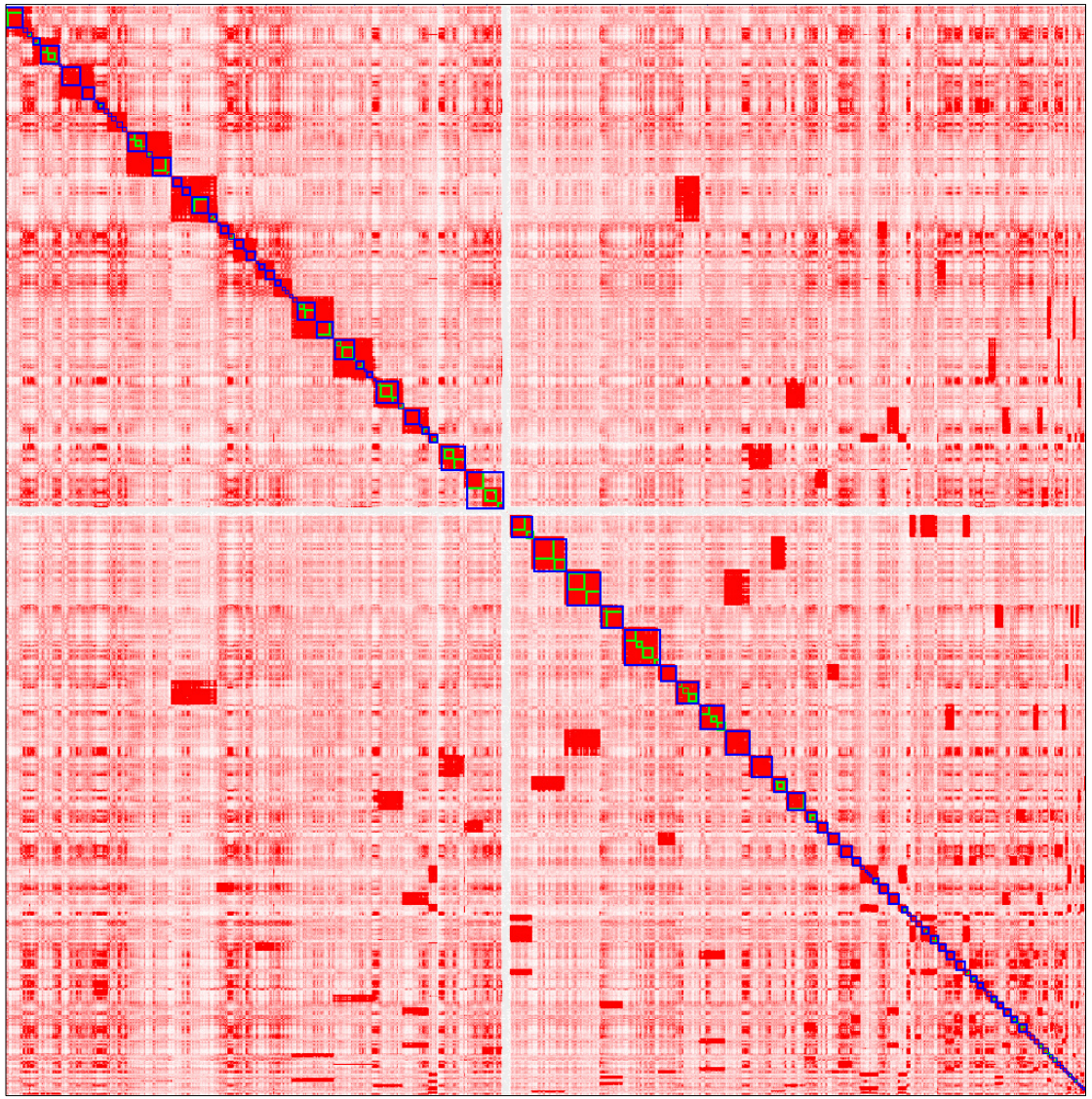
Hi-C contact map



Mother Trio1 – 14x Hi-C coverage

SeqOccIn

Coverage impact on Hi-C maps

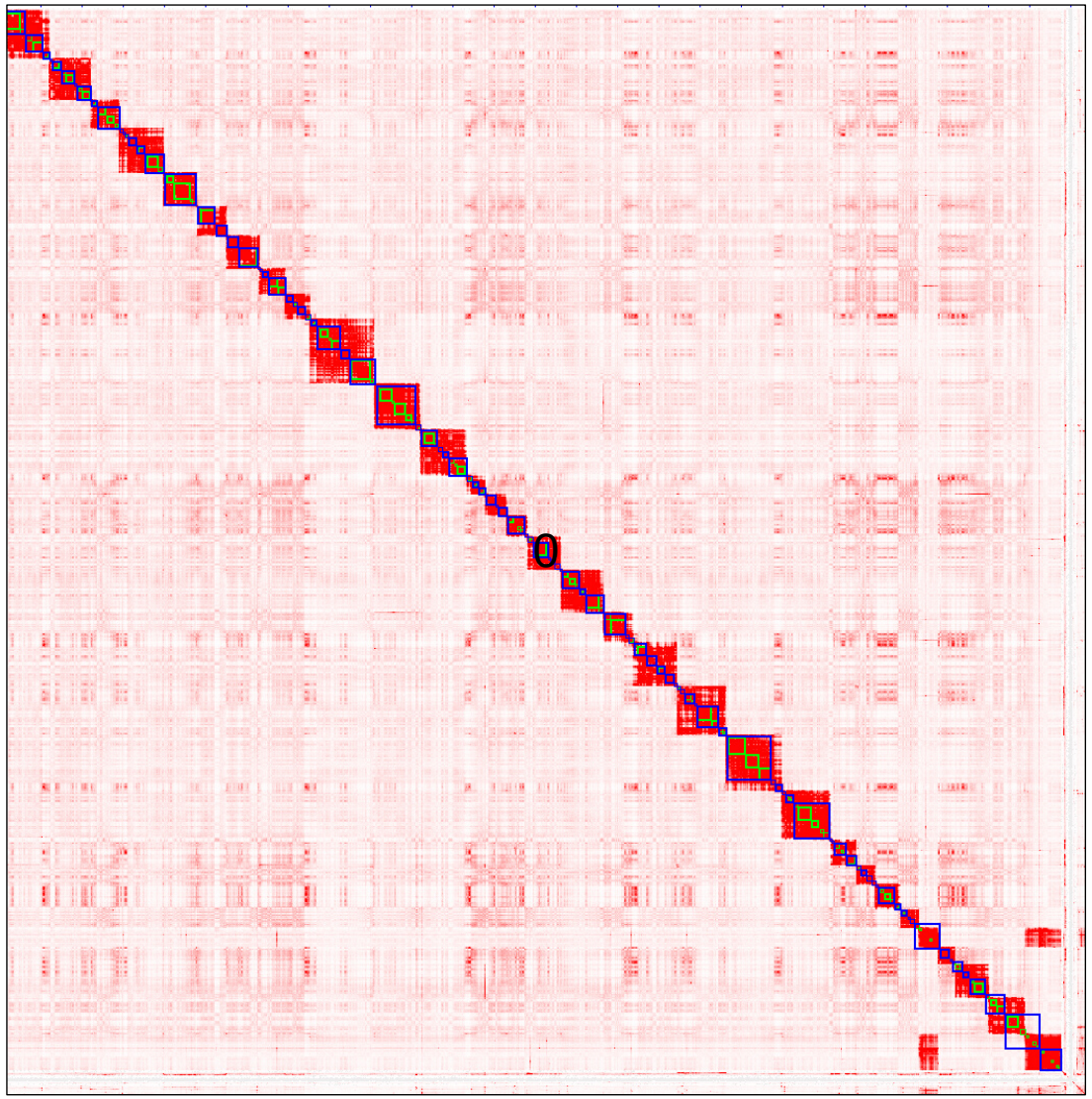


Mother Trio1 – 11x Hi-C coverage

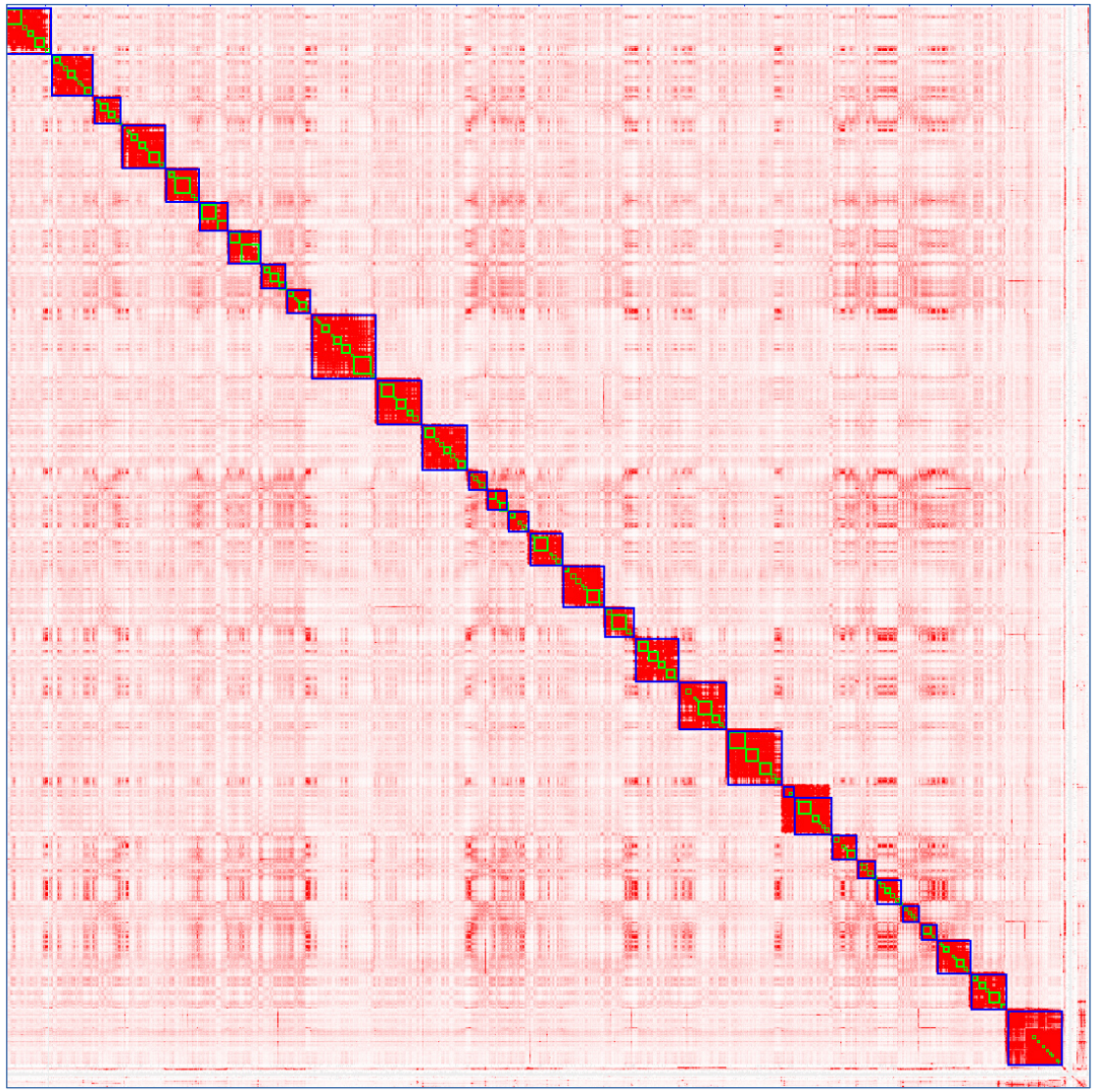
Genotoul GeT Genotoul Bioinfo

SeqOccIn

Manual correction



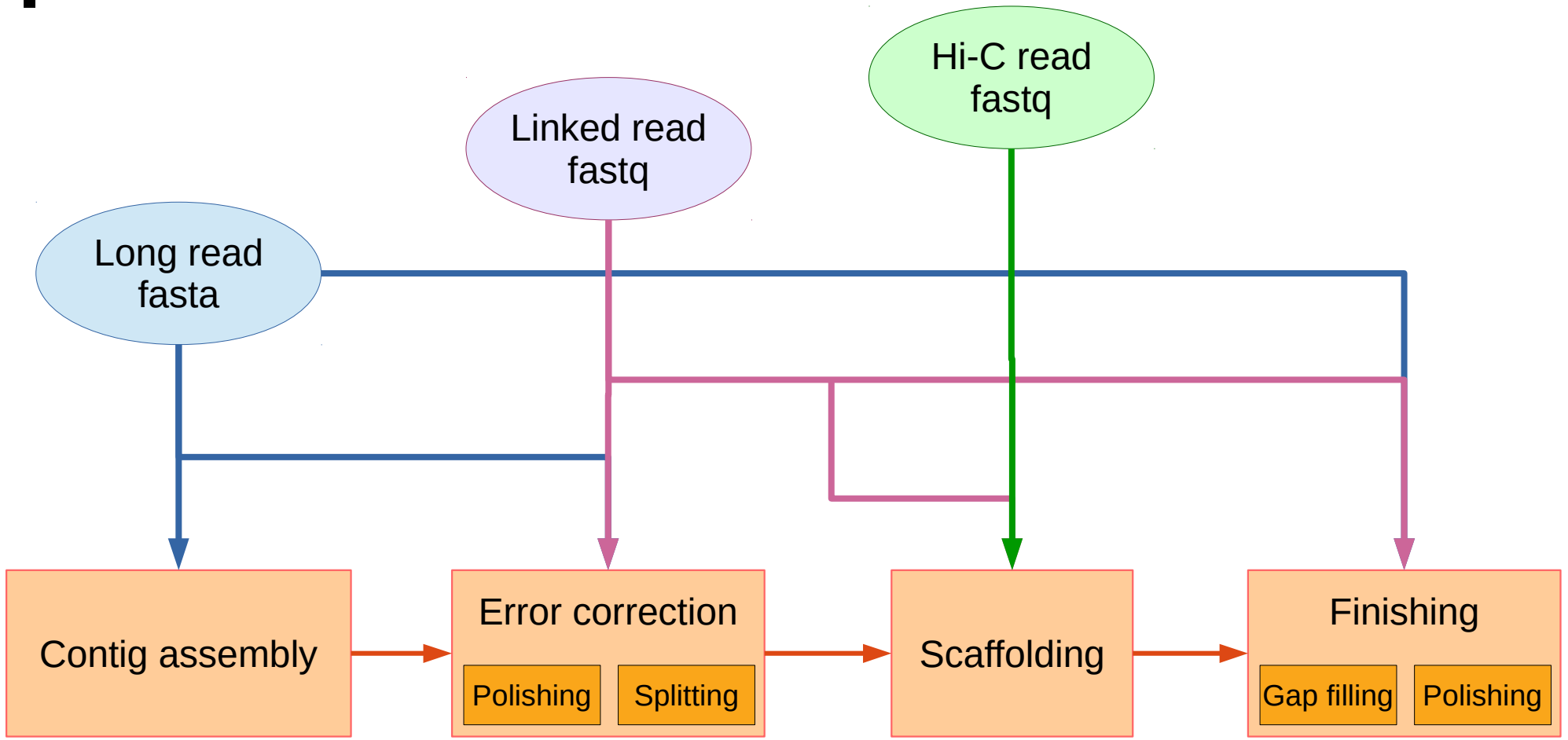
Mother Trio1 – 14x after 3d-dna



Mother Trio1 – 14x after manual correction

SeqOccIn

Completion time

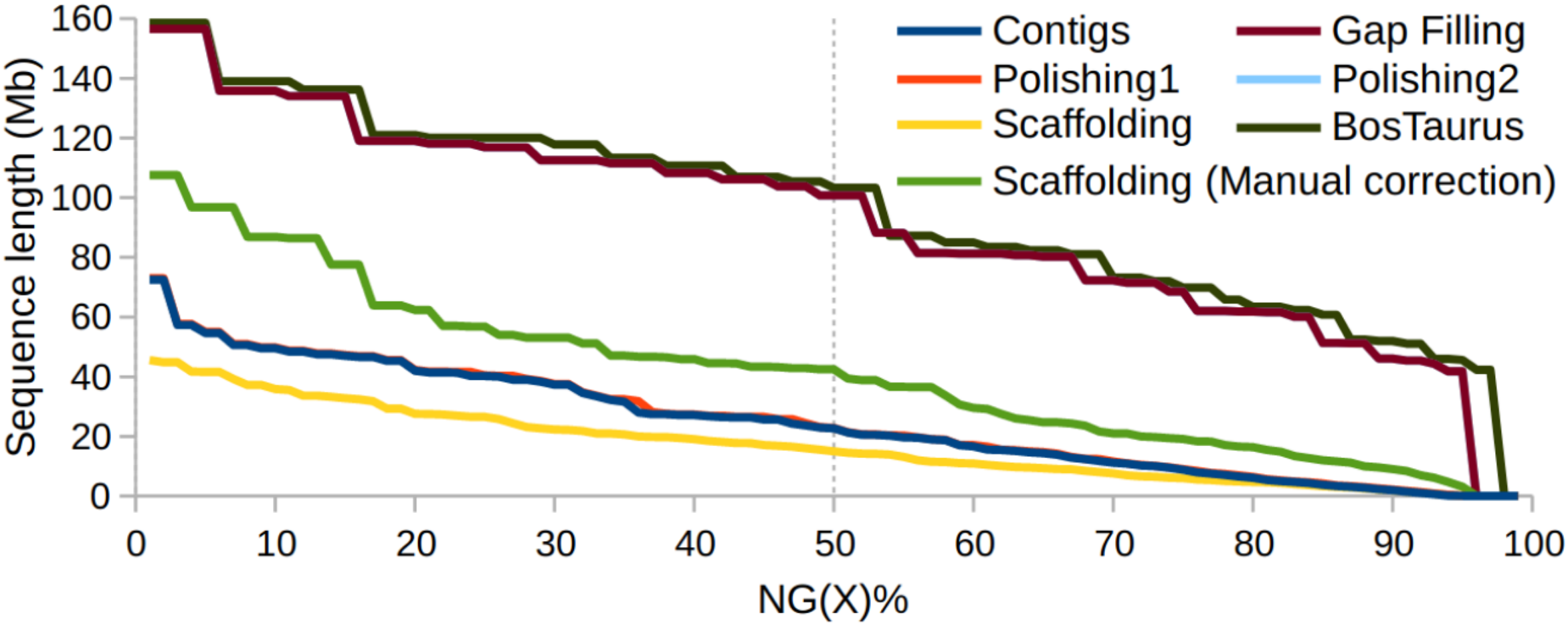


CPU hours :	≈ 2500h	≈ 1600h + 2600h + 50h	≈ 1500h + ≈ 4h manual correction	≈ 2000h + 2600h

Genotoul GeT Genotoul Bioinfo

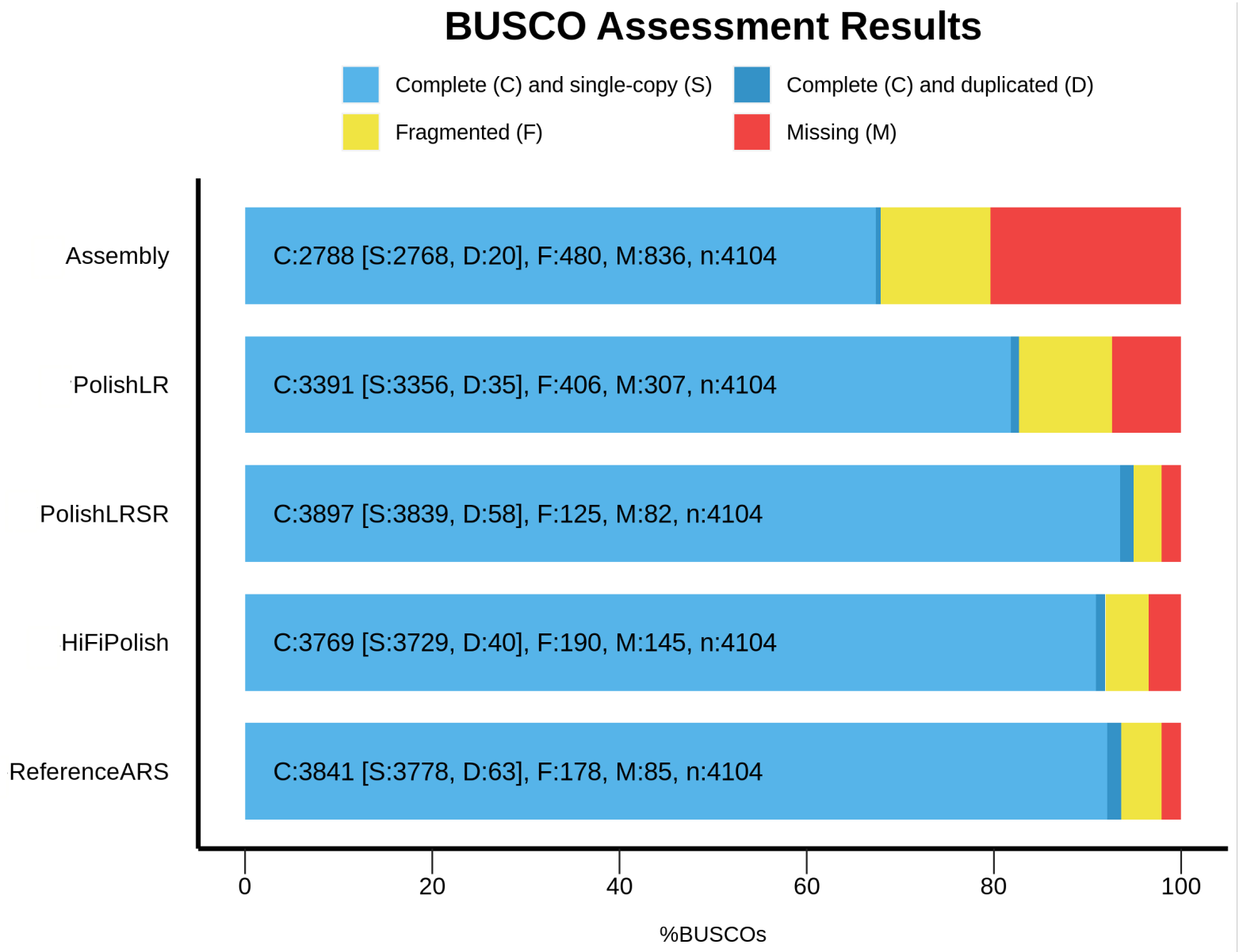
SeqOccIn

N50 evolution



Genotoul GeT Genotoul Bioinfo

Busco score evolution



Supernova – 10x Chromium based pipeline

« Push-button » pipeline that generates phased, whole-genome de novo assemblies (contigs and scaffolds) from a 10x Chromium-prepared library:

- Contigs build with a classical paired-end approach (problem: very short)

- Contigs phased and connected into scaffolds based on barcode information

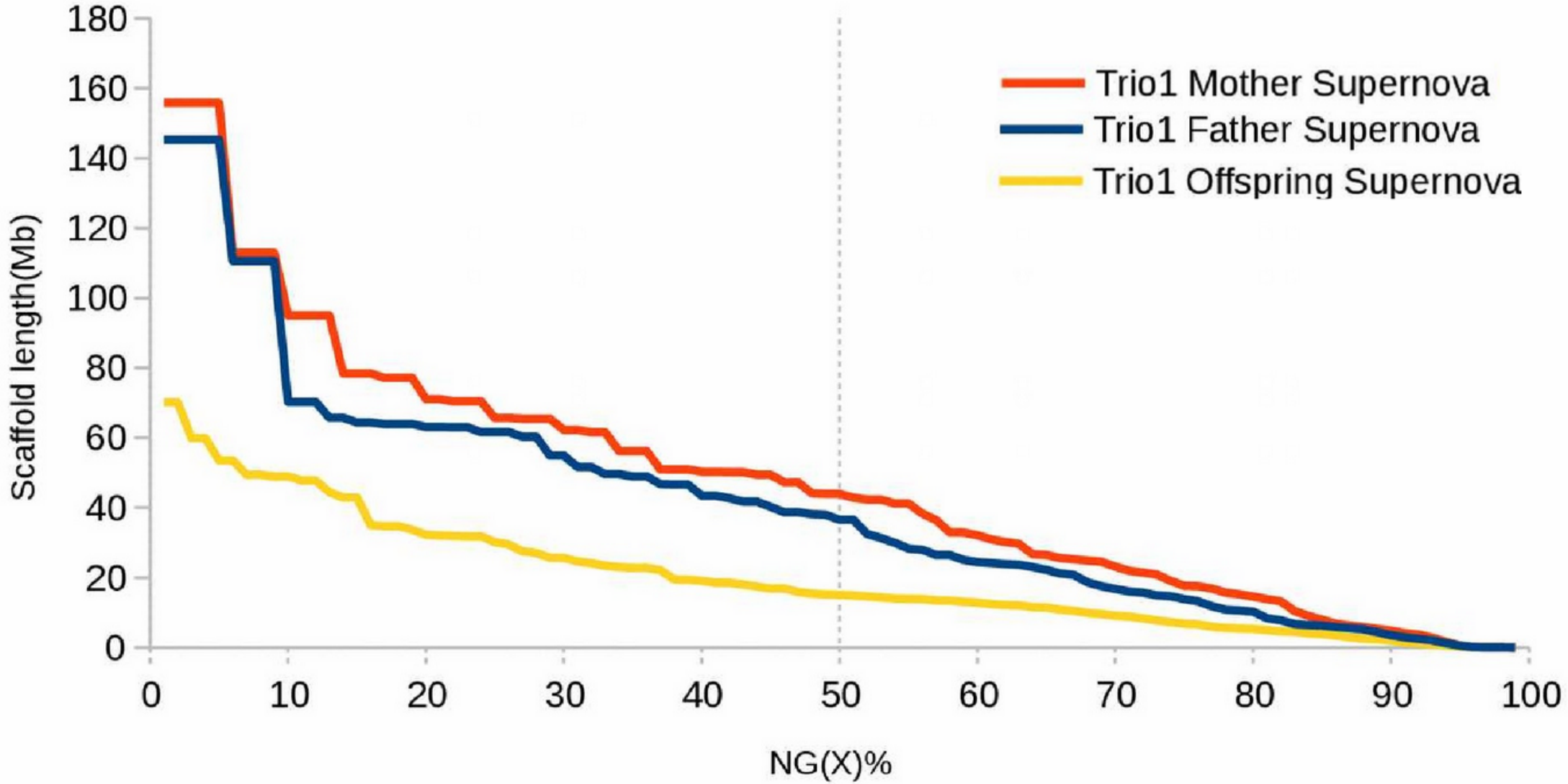
Completion time
≈ 4000h

BUSCO score:
C:95% [S:93.7%,D:1.3%],F:2.9%,M:2.1%

Scaffold N50:
40 / 45Mb

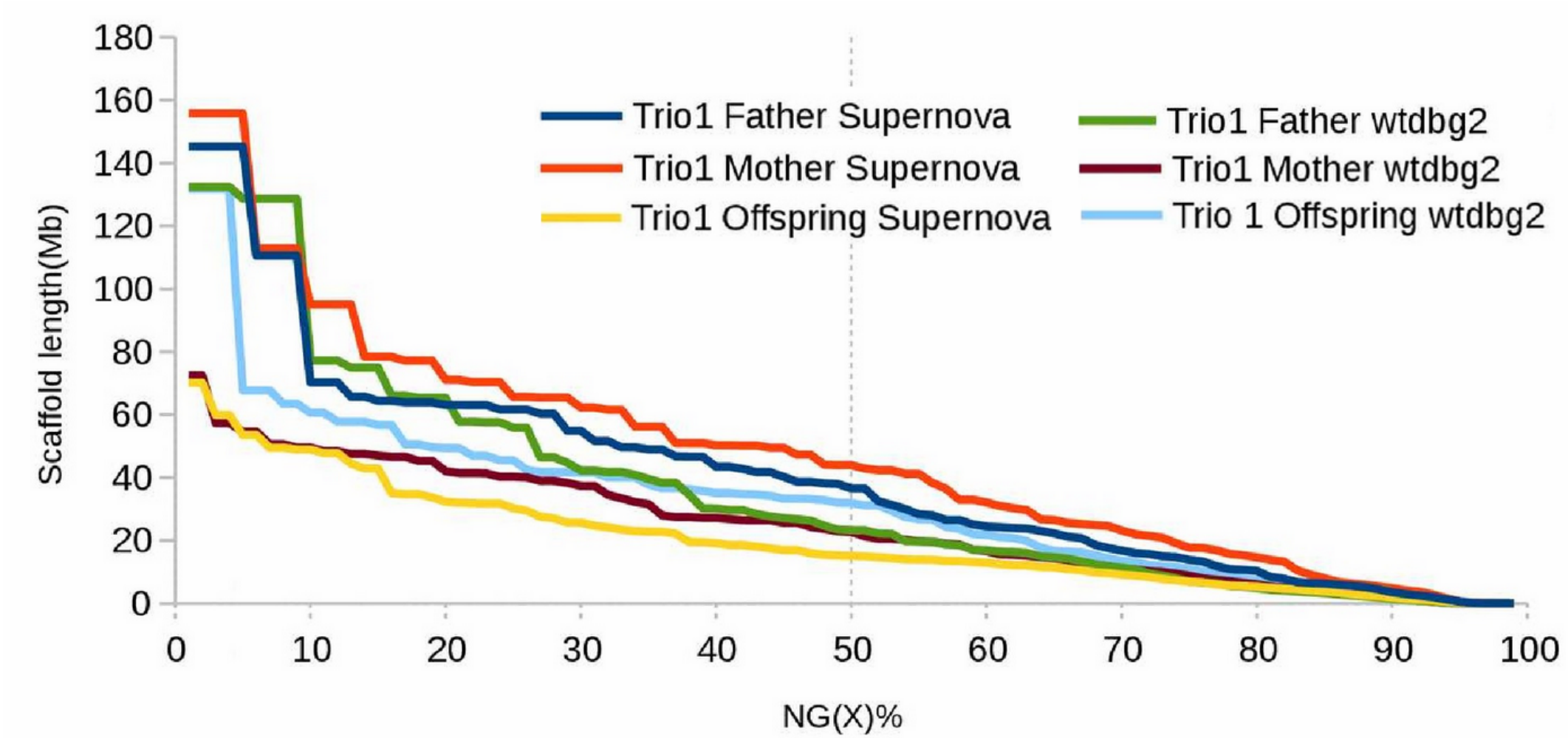
SeqOccIn

Supernova – 10x Chromium based pipeline

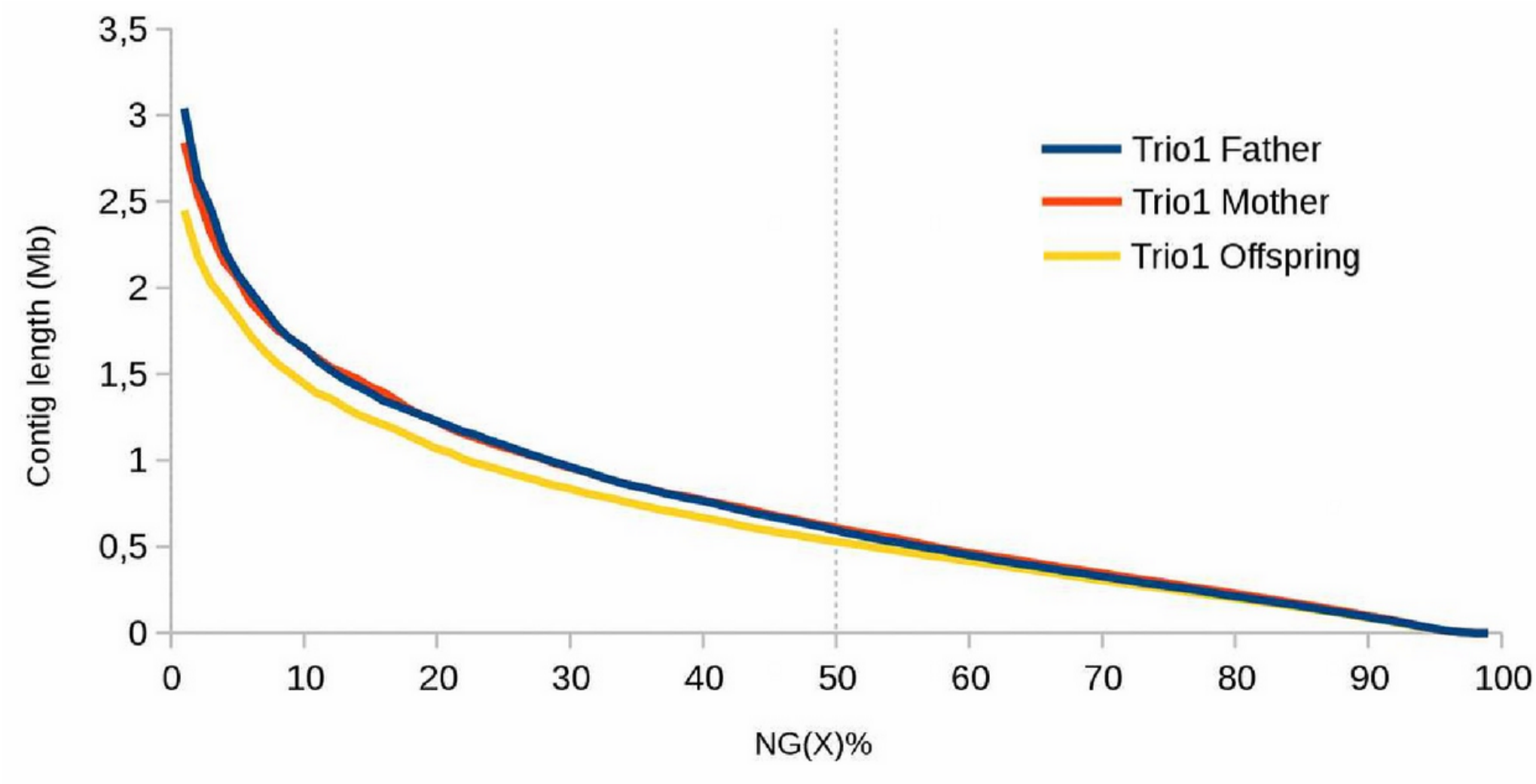


Genotoul GeT Genotoul Bioinfo

Supernova – 10x Chromium based pipeline



Supernova – 10x Chromium based pipeline



SeqOccIn

Work in progress

Individu	Assembly	Polishing LR	Splitting	Polishing SR	Scaffolding	Gap Filling	Final Polishing
Mother1 37165							
Father1 37164							
Offspring1 37163							
Mother2 37162							
Father2 37161							
Offspring2 37160							
Sus crofa							
Coturnix japonica							
Mais F2							
Mais MBS847							
Mais F252							
Zea Mais F4							

Genotoul GeT Genotoul Bioinfo

INRA SCIENCE & IMPACT Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE RÉGION MIDI-PYRÉNÉES l'Europe s'engage en Midi-Pyrénées