From reads to chromosomes: What is the optimal path ?

Clément BIRBES, Andreea Dréau, Camille Eche, Carole Iampietro, Cécile Grohs, Didier Boichard,
Cécile Donnadieu, Christine Gaspin, Denis Milain, Claire Kuchly, Christophe Klopp and Matthias Zytnicki

Introduction

Sequencing advances in the last years led to a significant increase in the number and quality of published de novo genomes, each one of them being a result of a mix of technologies and assembly strategies. The purpose of our study is to identify the intake of each type of reads and their coverage in order to determine an optimal approach, depending on the sequencing cost or the expected assembly quality, for a complete genome assembly of a bos taurus individual.

Methods

- Input data: Oxford Nanopore, Pacific Bioscience HiFi and CLR, 10X Chromium, HiC
- Assembly pipeline steps:

1. contig assembly: we compared wtdbg2, Flye and Shasta (Fig1), and the impact of coverage (Fig2), read length (Fig3) and the influence of the sequence technology (Fig4) on the assembly
2. polishing and splitting: we developed a Nextflow pipeline allowing the polishing of the assembly using long reads with Racon and short reads with Pilon
3. scaffolding with 3d-dna
4. gap filling and final polishing

- We used this pipeline in two separate ways:
  - ONT reads for the contig assembly and the first step of polishing, then 10X Chromium data to polish, then scaffolding with HiC data, and finalize the assembly by gap-filling with ONT reads and polishing with 10X reads.
  - Pacific Bioscience CLR for the contig assembly and HiFi for polishing, then finalize the assembly by scaffolding with HiC and polishing again with HiFi.
- Assembly evaluation: total assembly size, N50, the BUSCO score...

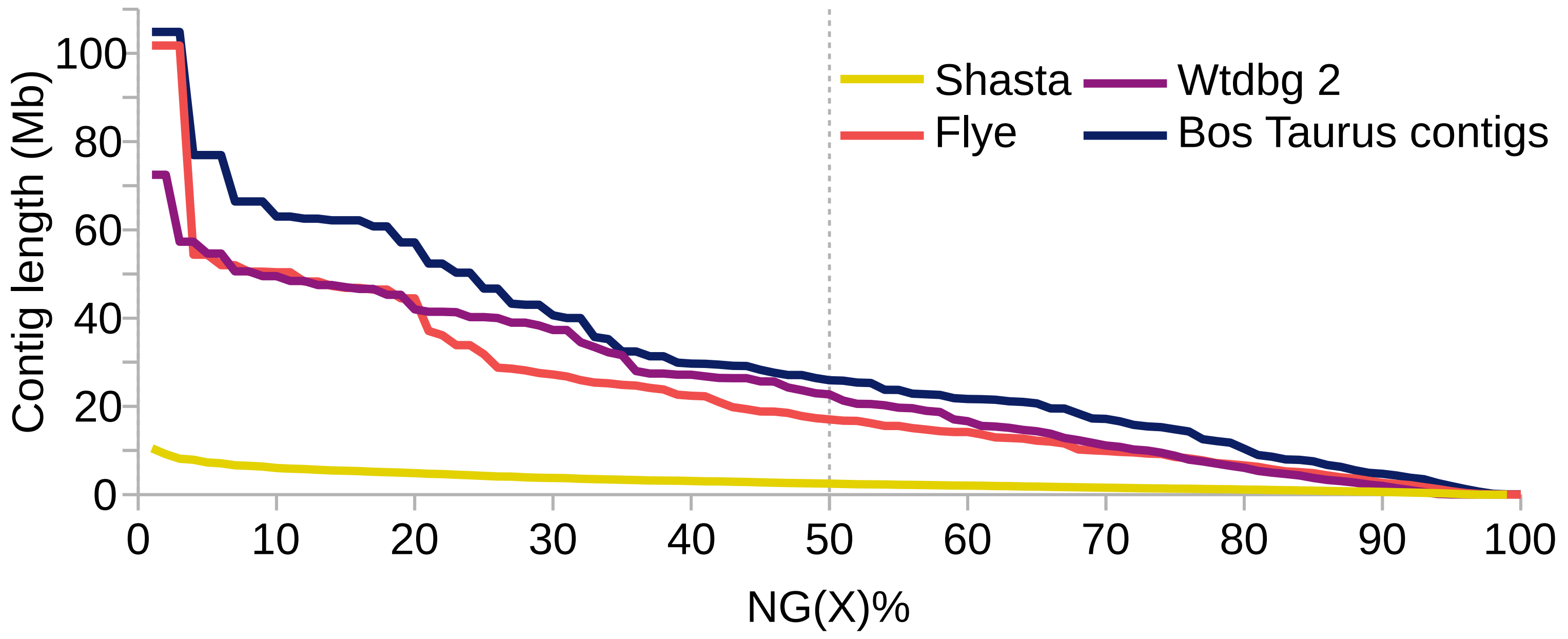

Figure 1: Impact of assemblers on contig lengths for a bovine individual sequenced at 58x ONT coverage

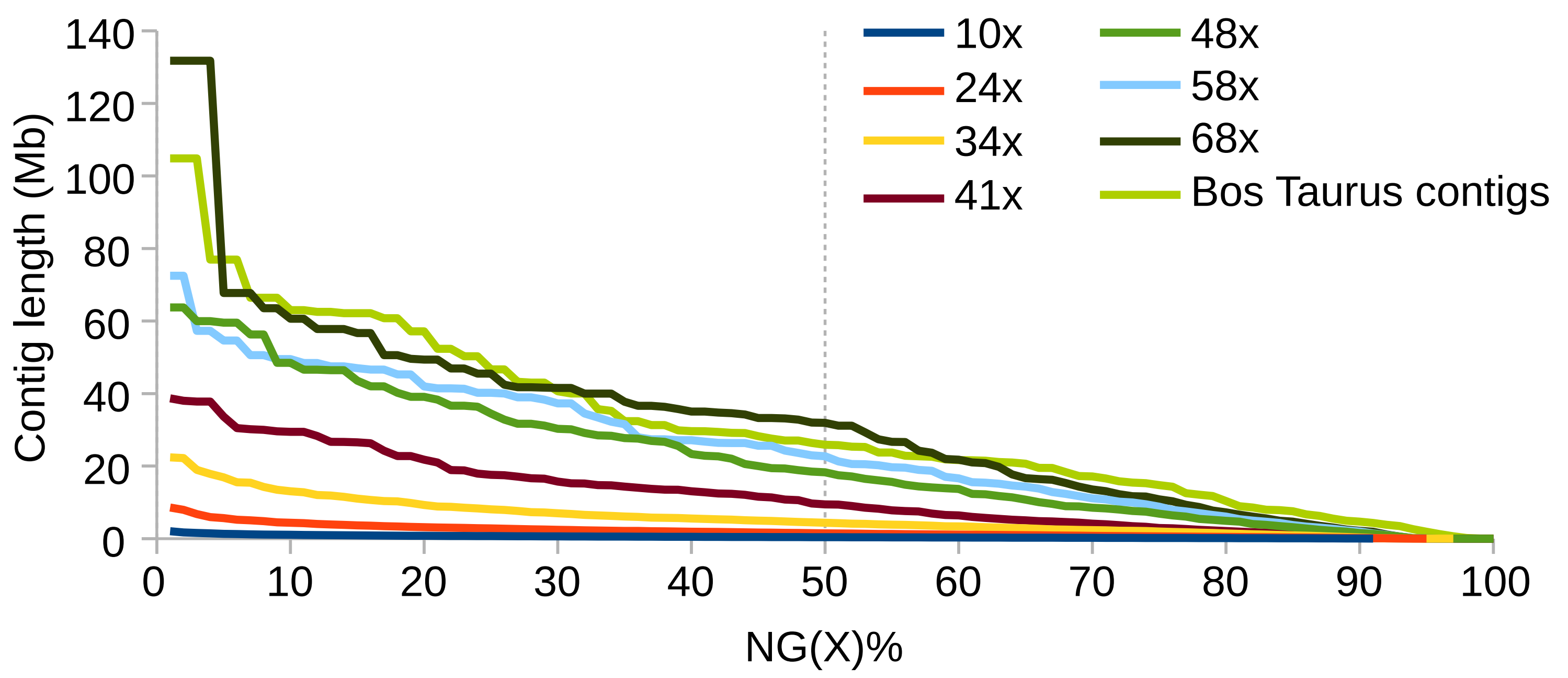

Figure 2:Impact of Oxford Nanopore reads coverage on different Bos taurus assembly with wtdbg2

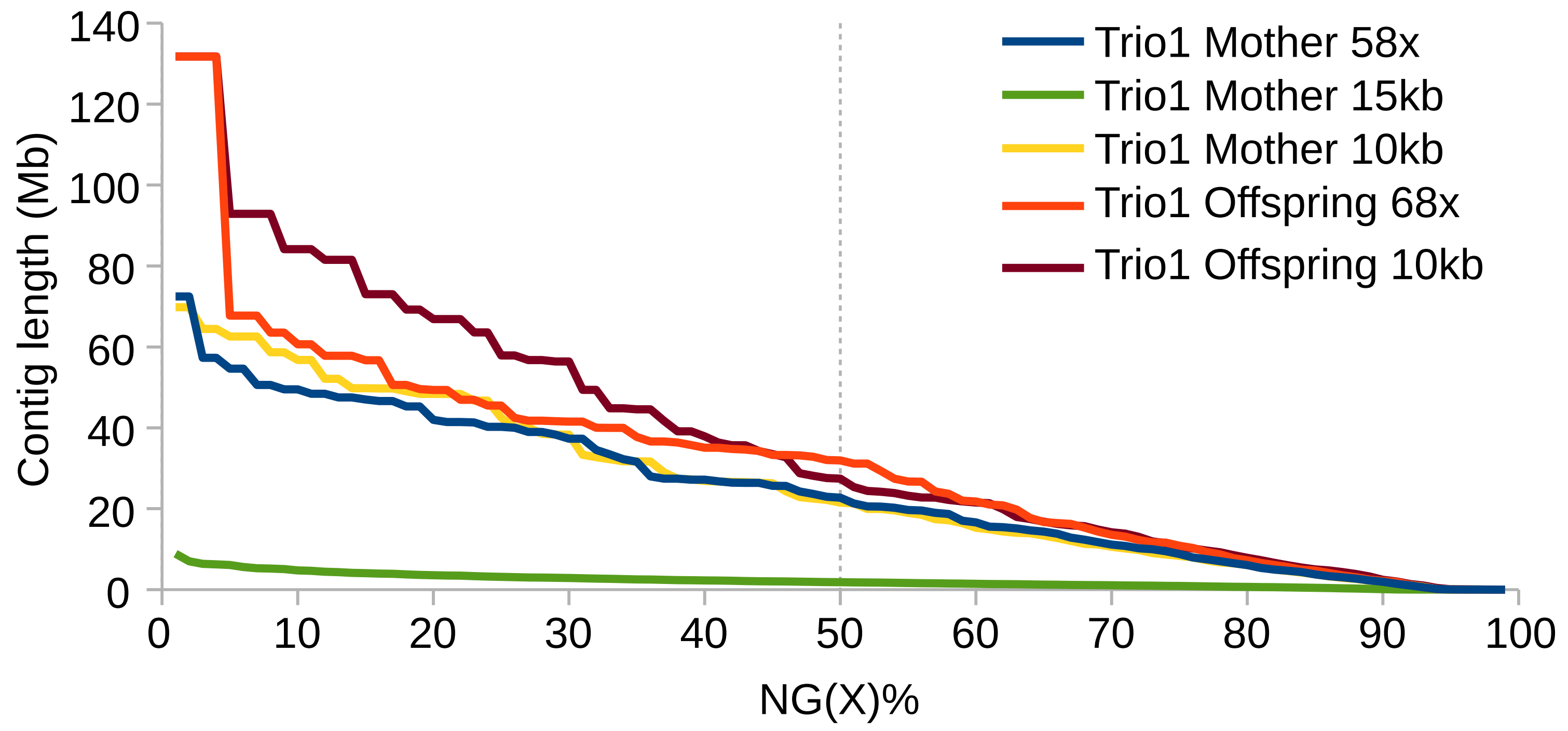

Figure 3: Impact of reads filtering on Bos taurus assembly with wtdbg2

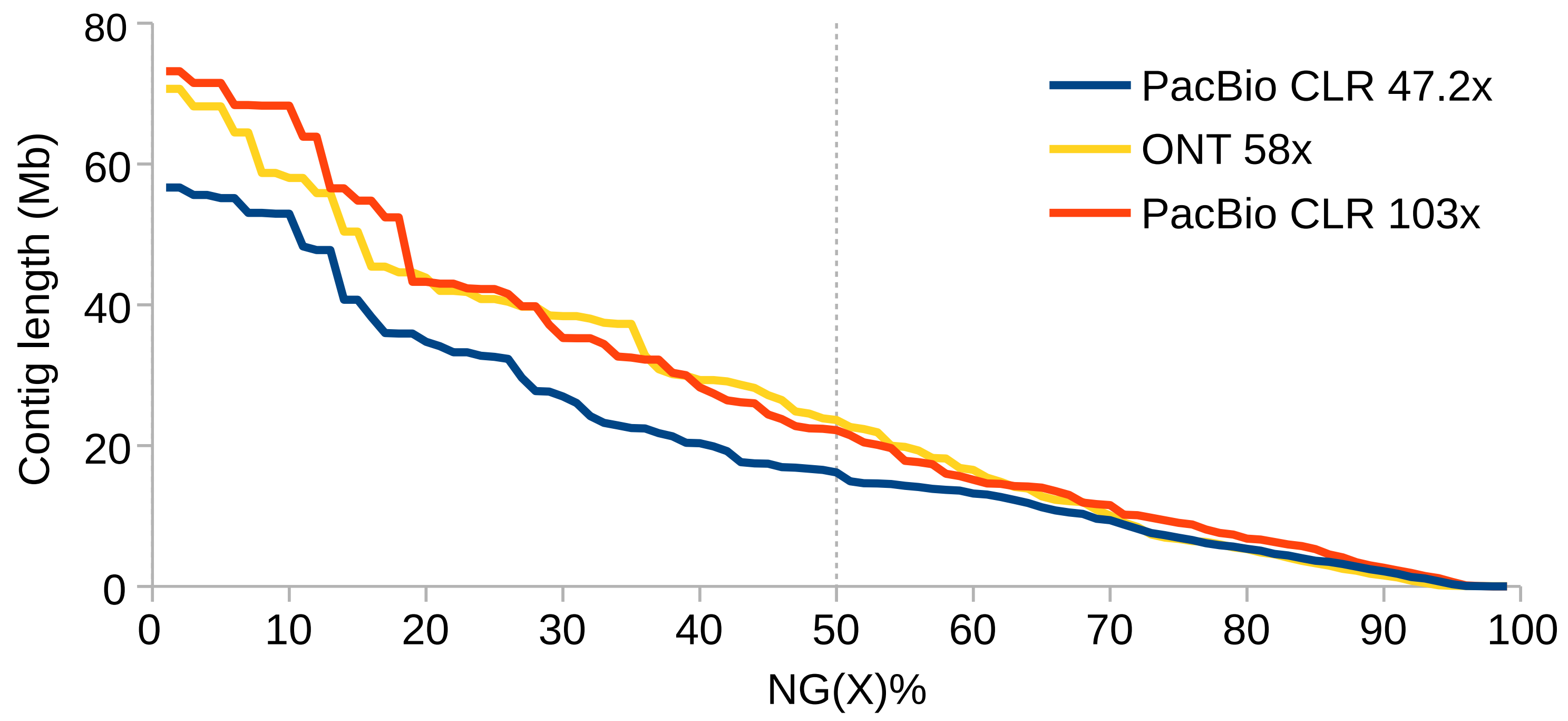

Figure 4: Technology influence on contig lengths Results obtained with wtdbg2 on 5 runs of ONT (58x), 1 run of PacBio CLR (47.2x) and 2 runs of PacBIo CLR (103x)

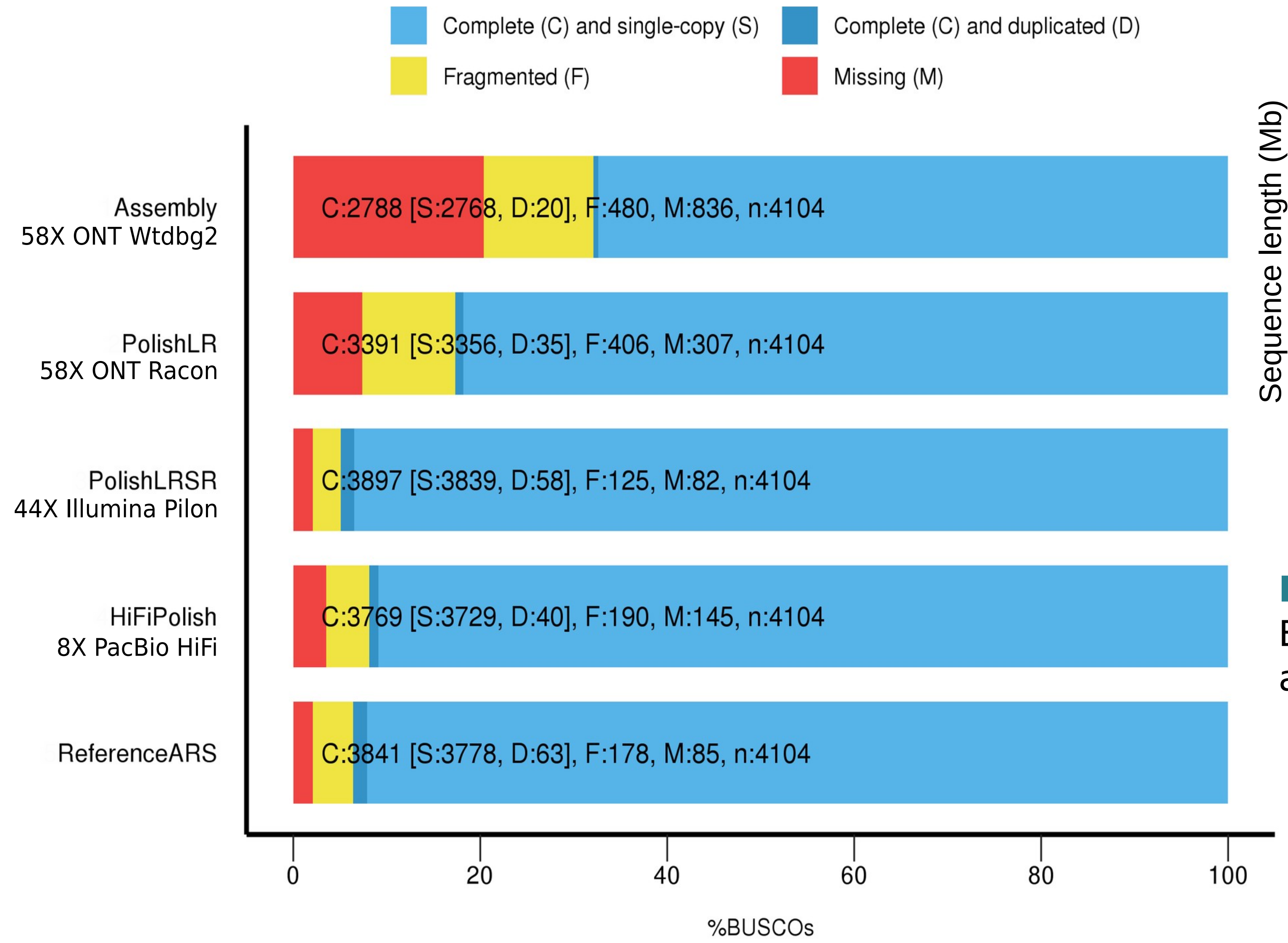

Figure 5: Busco score evolution after pipeline steps
Evolution of the Busco score after different step in the assembly pipeline. HiFiPolish and ReferenceARS are comparison line

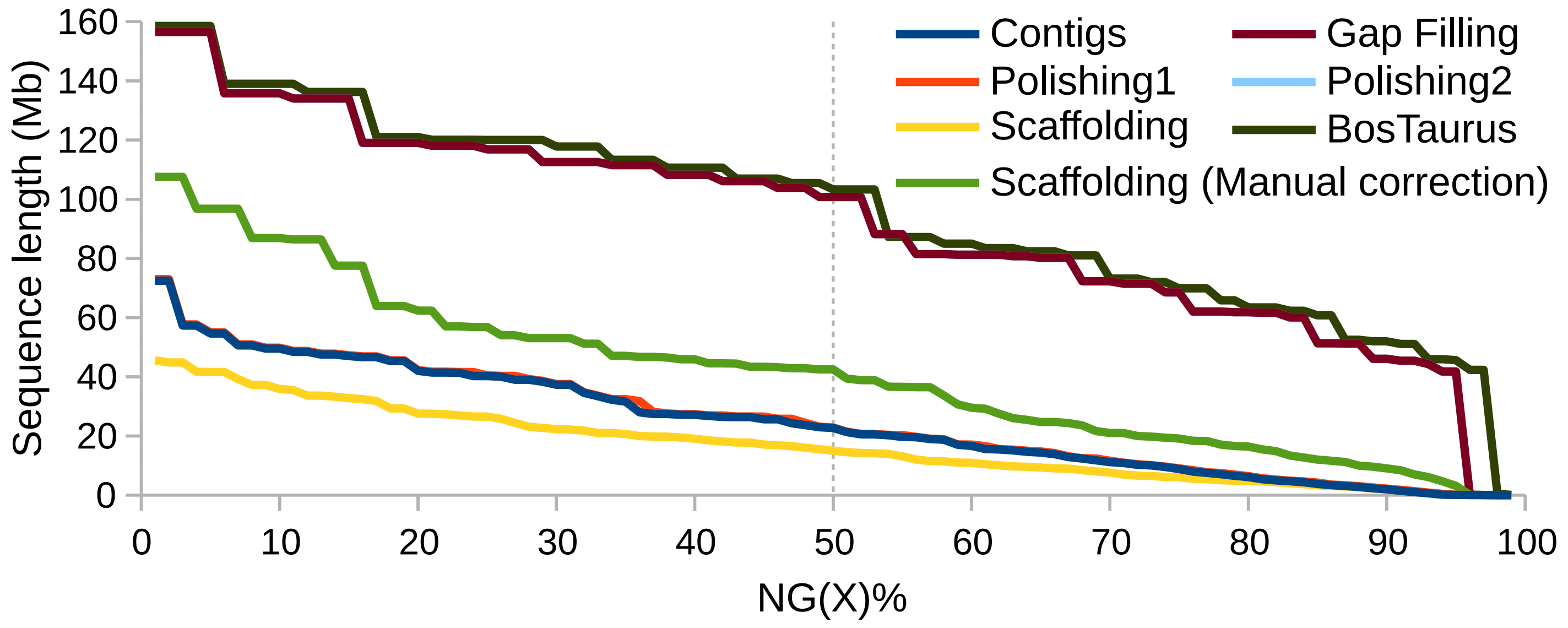

Figure 6: NG(X)% evolution after each pipeline steps
Evolution of the assembly size metrics after different step in the assembly pipeline.

Results

For contig assembly wtdbg2 has the best quality/time ratio (Fig1). The contig quality strongly depends on the coverage (Fig2) and the length of the reads (Fig3). In terms of technology we obtained the best coverage/quality ratio using ONT reads (Fig 4).
For our bovine genome we started from a 67.9% Busco score and 25Mb N50 after contig assembly, improved to 83% Busco score after long reads polishing, to finally reach a 95% Busco score after short reads polishing (Fig5). By using HiFi, we reach a similar quality but way faster, as we don't need two polishing step. That's why if we can get enough data, HiFi are more interesting.
The scaffolding step with 3d-dna allows manual correction which is often helpful in overcoming scaffolder issues.
After scaffolding and final polishing, we obtain a Busco score of 94.9% and an 104Mb N50 which is slightly better than the current bovine reference (Fig 5 and 6).

Conclusion

- for long contigs: wtdbg2 and a minimum of 40x coverage of ONT reads
- for a complete correction of the assembly: we developed of a polishing pipeline which allows the use of ONT, 10x Chromium, PacBio CLR and HiFi data
- for long scaffolds: 3d-dna using high quality contigs and Hi-C reads followed by a manual correction step led to chromosome length scaffolds

UNION EUROPÉENNE
l'Europe s'engage en Occitanie
La Région Occitanie Pyrénées - Méditerranée
Genotoul GeT
Genotoul Bioinfo