



HAL
open science

De novo assembly of bovine genome using PacBio applications

Camille Eché, Clément Birbes, Carole Iampietro, Andreea Dréau, Claire Kuchly, Christophe Klopp, Arnaud Di Franco, Thomas Faraut, Matthias Zytnicki, Erwan Denis, et al.

► **To cite this version:**

Camille Eché, Clément Birbes, Carole Iampietro, Andreea Dréau, Claire Kuchly, et al.. De novo assembly of bovine genome using PacBio applications. SMRT Leiden 2021 – Young Investigator Virtual Conference, May 2021, virtuel, Netherlands. <hal-03541224>

HAL Id: hal-03541224

<https://hal.science/hal-03541224v1>

Submitted on 24 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Eché, Camille¹ ; Birbes, Clement² ; Iampietro, Carole¹ ; Dréau, Andreea² ; Kuchly, Claire¹ ; Klopp, Christophe² ; Di Franco, Arnaud² ; Faraut, Thomas⁵ ; Zytnicki, Matthias² ; Denis, Erwan¹ ; Fritz, Sébastien³⁻⁴ ; Boussaha, Mekki³ ; Grohs, Cécile³ ; Boichard, Didier³ ; Gaspin, Christine² ; Milan, Denis¹⁻⁵ ; Donnadiou, Cécile¹

¹ INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France ² Plateforme Bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRAE, Castanet-Tolosan, France. ³ Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France. ⁴ Allice, 75012 Paris, France ⁵ GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet-Tolosan Cedex, F-31326, France.

Background and objectives

GeT-PlaGe is a genomic core facility which provides technologies and expertise in genome sequencing to academic and private research teams. The SeqOccIn (Sequencing Occitanie Innovation) project, supported by Get-PlaGe and Genotoul Bioinfo platforms, was selected by the Occitanie Region as part of the call for projects "Regional Research and Innovation Platforms". The main objective is to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes in agronomical field: from SNP and structural variations detection, to the production of a high quality assemblies.

Therefore, to obtain the best genome using PacBio datas, several *de novo* assembly strategies on a Charolais bovine will be presented : CLR datas only, CCS datas only and trio binning approach. In this poster, we highlight biological and bioinformatic aspects taking into account the scientific purpose and different constraints (time, budget, samples qualities).

For this work, we sequenced the same individual (offspring bovine) with different applications: PacBio, ONT and Hi-C, and parents with classical short reads (2x150 pb Illumina).

Workflow for PacBio library preparation

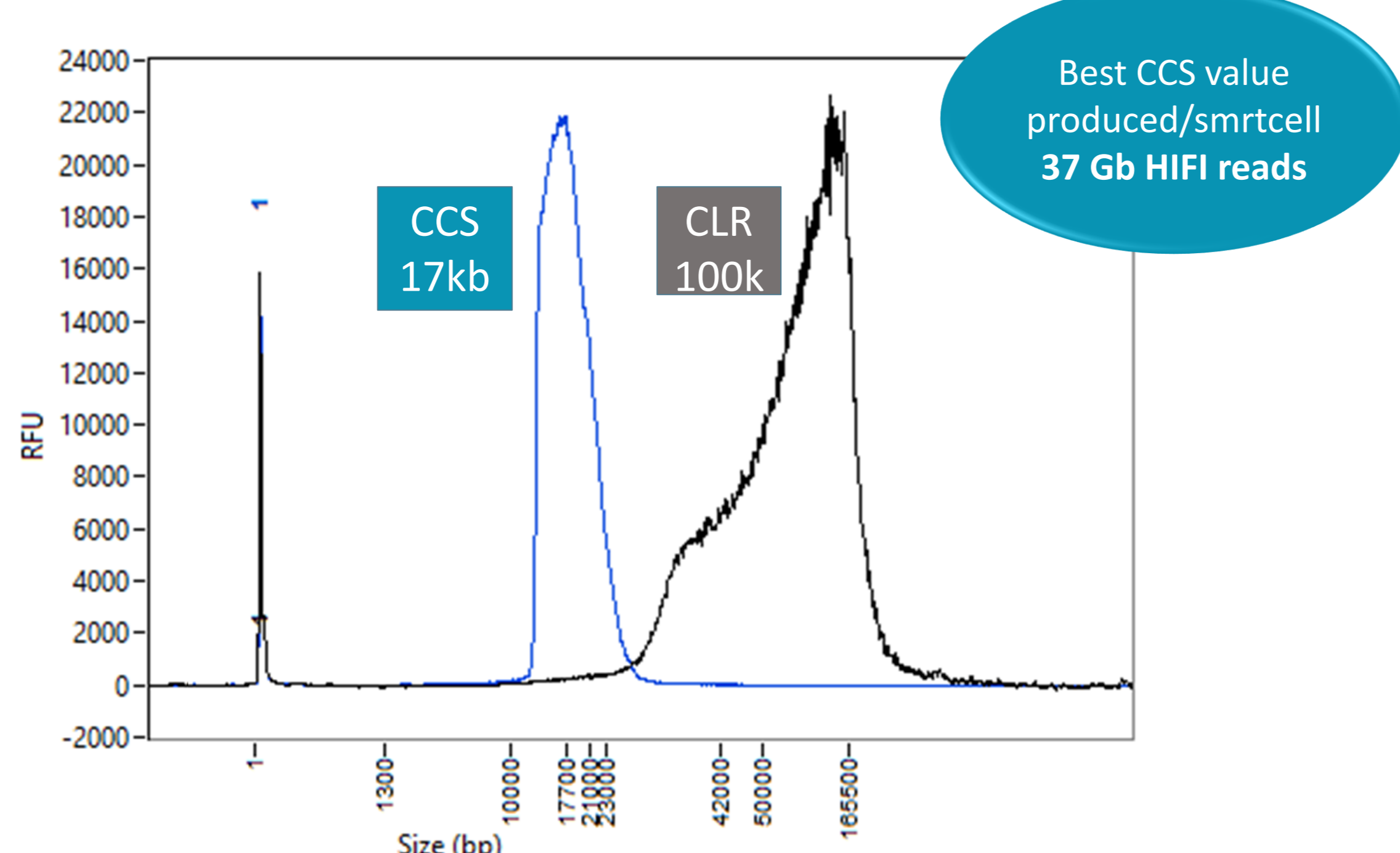
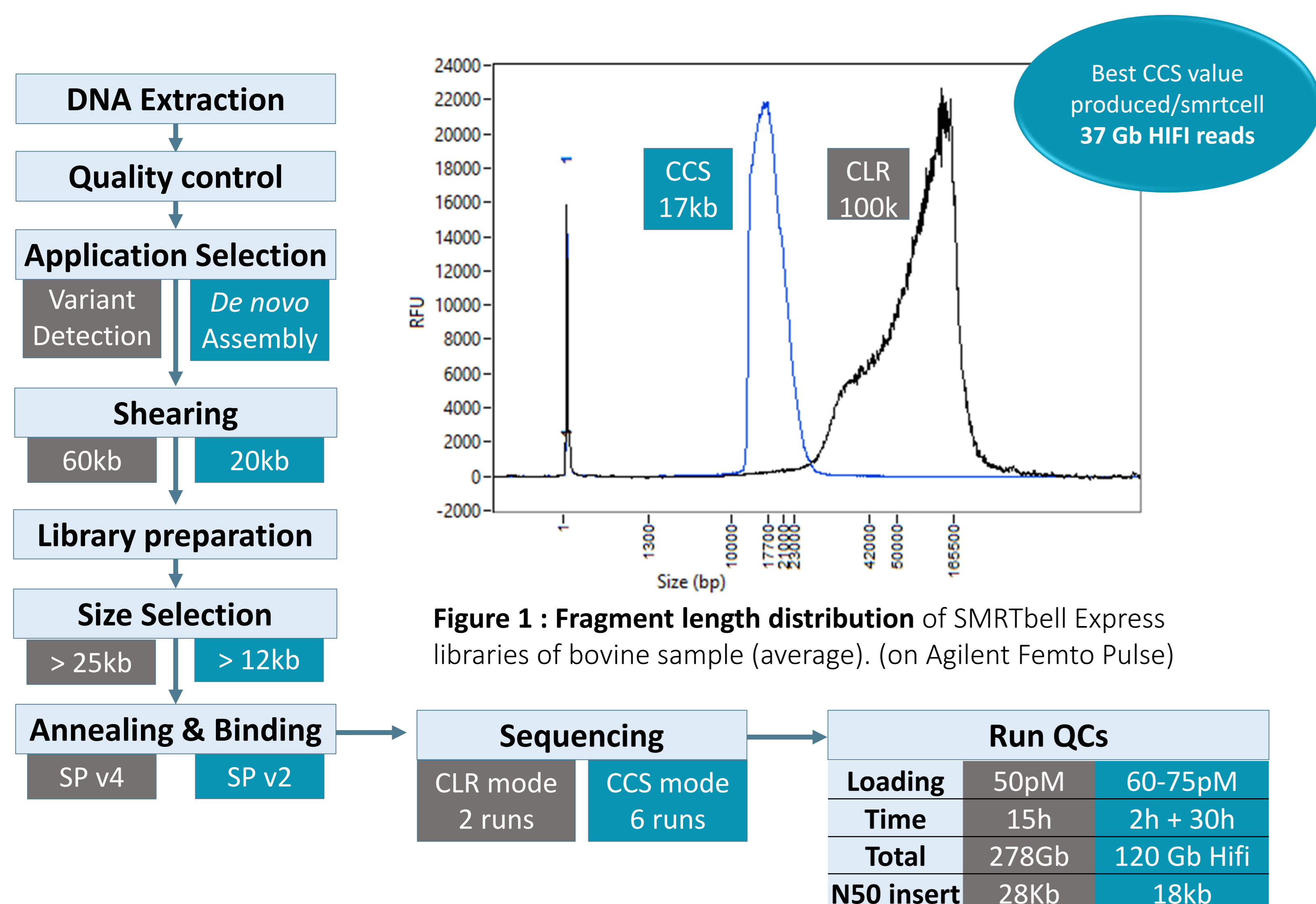


Figure 1 : Fragment length distribution of SMRTbell Express libraries of bovine sample (average). (on Agilent Femto Pulse)

Complementary informations: DNA Extraction kit: Promega; Quality control : Femto-Qubit-Nanodrop ; Shearing: Megaruptor 3 ; Library preparation : SMRTbell Express Template Prep Kit 2.0 + Nuclease Treatment of SMRTbell Library, Size Selection: BluePippin

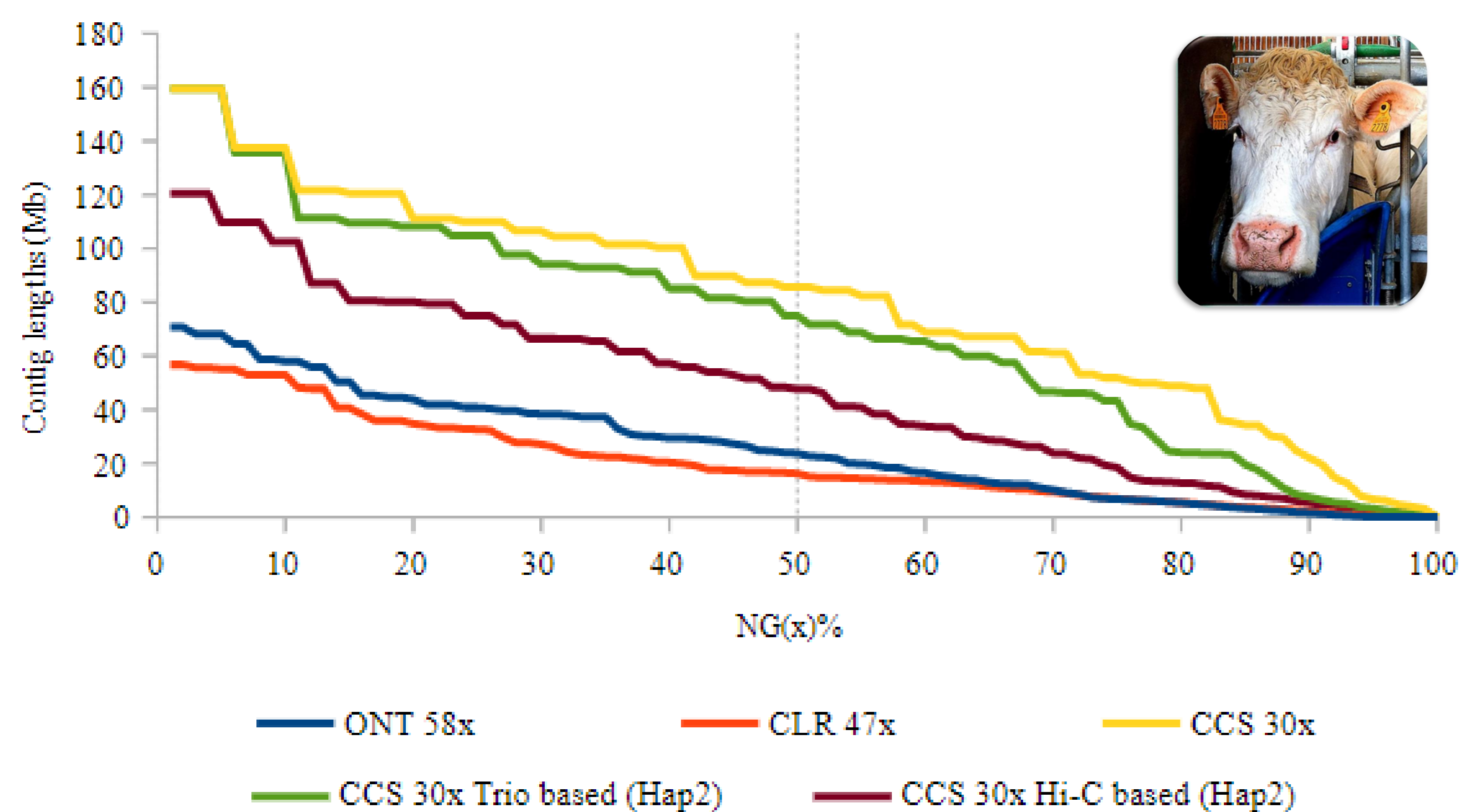


Figure 2 : NG(X) % values for ONT, CLR and CCS assemblies of the offspring bovine.

As showed in Table 1, assemblies based on long erroneous reads (PacBio CLR and ONT) produce shorter contigs (NG(50) lower) than CCS assemblies. Between CCS based assemblies, **the longest contigs were obtained with CCS stand-alone method and with the trio binning approach**. Phasing is more challenging using Hi-C reads than short reads sets with uniform coverage of the individual's parents.

De novo assembly

Data type	Offspring datas only						+ parental datas	
	ONT	CLR	CLR	CCS	CCS Hi-C based (Hap1)	CCS Hi-C based (Hap2)	CCS Trio based (Hap1)	CCS Trio based (Hap2)
Coverage	58X	103X	47X	30X	30X	31X	30X	30X
Assembler	Wtdbg2	Wtdbg2	Wtdbg2	HifiAsm	HifiAsm	HifiAsm	HifiAsm	HifiAsm
Number of contigs	7 226	2 937	2 857	2 175	4 418	3 553	3 943	4 056
Total size (Gb)	2.7	2.64	2.63	3.14	3.14	3.12	2.93	3.2
N50 (Mb)	23.6	22.39	16.54	82.31	30.99	38.25	49.19	66.5

Table 1: Summary of bovine assembly statistics.

Different primary bovine assemblies (no scaffolding) were produced from Oxford Nanopore Technology (ONT - SQK-LSK109, Guppy version 3.0.6), PacBio CLR and PacBio CCS reads obtained from the same individual. The same assembler (WTDG2 V2.3) was used for long erroneous reads (ONT and CLR) while CCS reads were assembled using HifiAsm V0.15 in three different ways: based only on the CCS information (CCS column), using the trio binning approach where the two haplotypes are constructed based on short reads sets of the parents (CCS Trio based columns) and a diploid assembly using Hi-C reads (CCS Hi-C based columns). While the assembly size was slightly larger than the expected size (2.8Gb) **in all HifiAsm assemblies, the gain in contiguity is significant compared with the long erroneous reads based assemblies**. When comparing the two long erroneous sequencing techniques though, the assembler requires a lower coverage with ONT reads than with CLR reads for similar results due to the longer size of ONT reads.

BUSCO Assessment Results

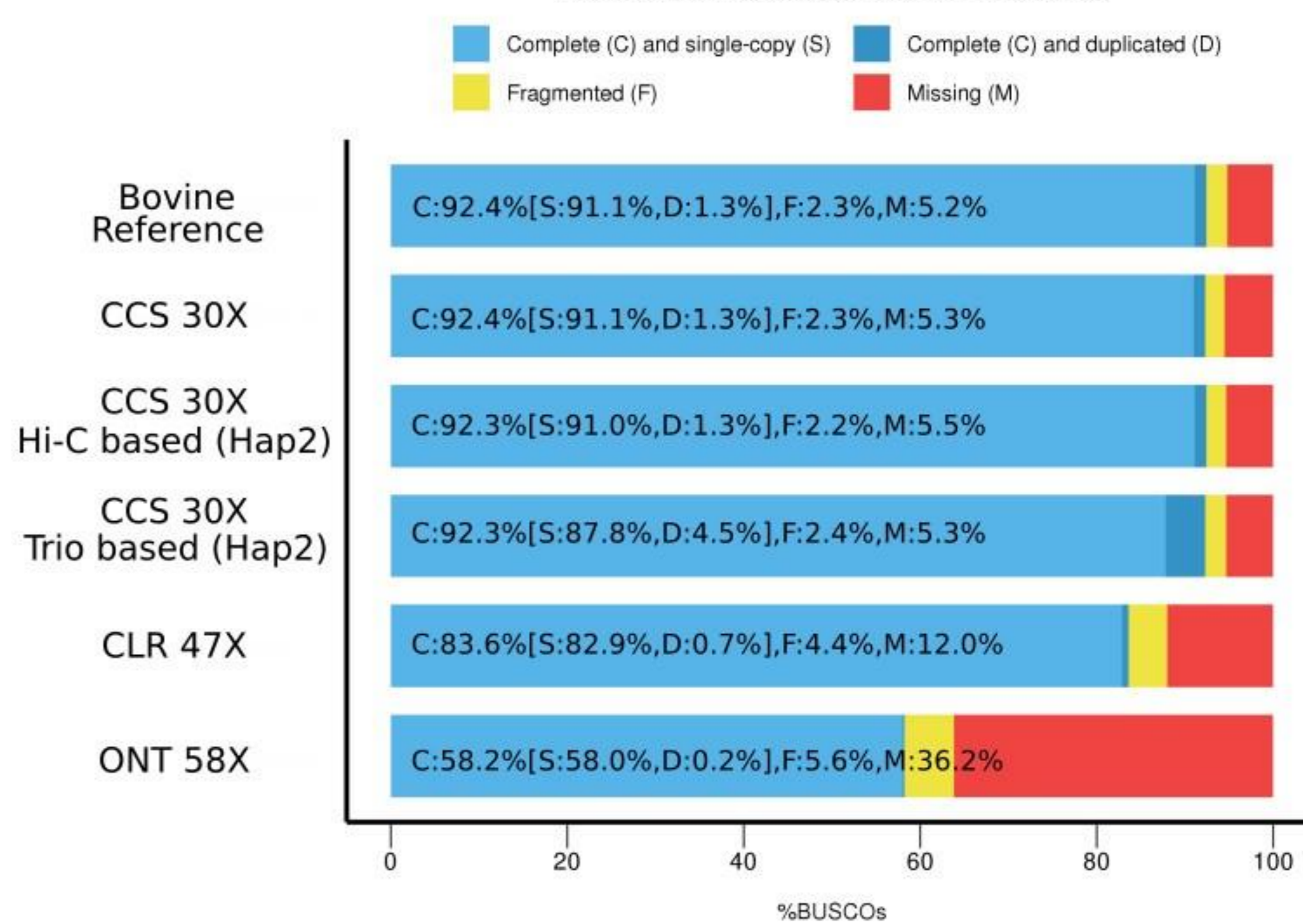


Figure 3 : Busco statistics of the offspring Bovine assembly versus reference sequence.

The primary assembly produced with ONT reads contains more errors and therefore fewer genes are found. The CLR approach decreases the number of residual errors and produces a better assembly with higher number of correct genes. Due to the **high quality of CCS data, the assembly contains fewer errors and the busco scores are therefore much higher and similar to the refined bovine reference (ARS-UCD1.2)**, but with significant higher costs.

	Required Coverage	Extraction type	Min Sample Size	Min Quantity	Lab Time	Required SMRTcells	Cost	Busco Score (C)	N50	Assembly Total size
CLR	50X	HMW	50kb	10µg	3 days +++	1 to 2*	+	83,6%	16,5 Mb	2,6 Gb
CCS	30X (15x/haplotype)	Simple	20 kb	20-30µg	5 days +++++	3 to 6*	+++	92,4%	82 Mb	3,1 Gb

Table 2: Summary of requirements and expected results for CCS and CLR based assemblies.

* Depending the sample quality

Thanks to our results we can propose a sequencing strategy to research teams that will fulfill their goal and that will be adapted to their constraints.