Methods

Genome information. Domestic cattle genome follows the pattern of the Bovidae family : 29 autosomes and 2 sex chromosomes. The autosomes are acrocentrics, which mean their centromeres are terminal. Liu, Y. *et al.*[1] estimated the *Bos_taurus* genome size to be approximately 2.87Gb based on measurement from ESTs and previous genome size. To validate these results we estimated the genome size using PacBio Sequel2 HiFi data using Jellyfish (v2.2.10)[2] and Genomescope2.0[3]. Genomescope2.0 estimated the genome size to be approximately 2,876Mb (Fig1). In addition the heterozygosity was estimated to be 0.289% and the proportion of repeated sequences to 34.8%.

De novo assemblies of consensus bovine charolais genome. Thanks to the wide variety of data available, we have developed two different de novo contigs assembly and polishing pipeline.

First strategy was used to assemble both parent and offspring. This methods combine Oxford Nanopore and 10X chromium reads. Contigs assembly was performed with Oxford Nanopore reads (using Wtdbg2 (v2.3)[4] with recommended option for large genome assembly with ONT data and respectively 58X, 52X and 58X coverage for father, mother, offspring. A first polishing step was performed using Racon (v1.4.10)[5] with default parameters and the same Nanopore reads. Then we polished the assembly with pilon (v1.22)6 and between 84X and 102X coverage of 10X chromium reads (See table1).

The second strategy use only the 31X coverage PacBio HiFi reads to produce an Hifiasm (v0.15.5)[7] assembly of the offspring with default parameters.

For both strategy we scaffolded the produced assembly into chromosomes using the 3D-Dna pipeline (version version 180114)[8] with Hi-C reads first. Then we produced 2 link-heatmap with 3D-Dna. The first map using Hi-C reads and recommended setting. The second map using 10X chromium reads with in-house protocol which proceeds as follow:

- Count the number of tags shared between each contig block
- Generate a 10X merged-nodups.txt file
- Create the .hic file

%todo%

Finally we corrected this combined heatmap manually with Juicebox (v1.9.8)[9] (Fig 2)

Diploid assemblies of bovine charolais genome. Two other assemblies were tested, based on the production of haplotyped assemblies with the use of HiFi reads (31X) combined with: Hi-C reads from the heifer (28X) in the first assembly and Illumina reads (84X-100X) from the parents in the second assembly. Both assemblies were done with Hifiasm (v0.15.5)

Technical validation

Reads validation. On voit avec les bio ce qu'ils veulent mettre / ce qu'on doit rajouter si on doit rajouter ?

HiC evaluation.

We used Juicer(v1.5.6)[10] to analyze the Hi-C reads aligned to the Wtdbg2 polished assembly (See Table2). From around 250 million reads pairs sequenced for each individual, more than 94% are considered as alignable (Normal paired + Chimeric paired)(Fig3). More than 72% are unique and a

minimum of 75,000,000 pairs are useful for the Hi-C map. For the final HiC map elaboration, we got: 151865989 pairs for the offspring, 151680029 pairs for the mother and 78011693 pairs for the father, wich respectively represent 15X, 15X and 8X useful coverage.

Assembly evaluation.

To evaluate the completeness of our 6 produced bovine assemblies (2 Offspring Consensus assemblies, 2 Parental assemblies and 2 phased assemblies) we compared them to those of *Bos_taurus* reference ARS-UCD1.2 (Table X). The size of our assemblies is close to the 2.87Gb estimation, with respectively 2.7Gb for ONT assembly and >3.1Gb for Hifi assemblies. Hifi assemblies producing larger assemblies, due to the fact that the low error rate in the reads makes it possible to better assemble the repeated areas

In addition we produced an alignment of each bovine assemblies against reference and visualized them with Dgenies[11] (Fig4).

The assembled genomes were also subjected to Benchmarking Universal Single-Copy Orthologs (v5.1)[12] which evaluate the genome completeness using the genes in the mammalia release10 dataset (mammalia.odb10). These BUSCO results showed a great completeness of the 4 genomes, from 95.2% to 95.9% of conserved genes, including around 1.9% fragmented genes (Table X).

%todo%

Diploid assembly evaluation.

%todo% In house protocol a expliquer

Code Availability

%todo%

1.Jellyfish v2.2.10 count and hist for Genomescope2.0
zcat -f *.reads.fastq.gz | jellyfish count /dev/fd/0 -C -m 21 -o reads.jf
jellyfish histo -h 10000 reads.jf > reads.histo

2.Wtdbg2 v2.3 assembly
wtdbg2 -g 2.8g -x ont -i input_reads1.fastq.gz input_reads2.fastq.gz -fo assembly_raw
wtpoa-cns -i assembly_raw.ctg.lay.gz -fo assembly_raw.cgt.fa

3.Racon v1.4.10 polishing
racon input_reads.fastq.gz alignement.sam assembly_raw.cgt.fa > assembly_racon.fa

4.Pilon v1.22 polishing
pilon --genome assembly_racon.fa --bam alignement.bam --fix bases,gaps --output assembly_pilon.fa

5.Hifiasm v0.15.5 assembly
hifiasm -o assembly_Hifi.asm input_reads.fastq.gz

6.Hifiasm v0.15.5 diploid Hi-C assembly
hifiasm -o assembly_Hifi_HiC.asm --h1 HiC_reads1.fastq.gz --h2 HiC_reads2.fastq.gz
input_reads.fastq.gz

7.Hifiasm v0.15.5 diploid parental assembly
yak count -b37 -o mother.yak illumina_mother_reads1.fastq.gz illumina_mother_reads2.fastq.gz
yak count -b37 -o father.yak illumina_father_reads1.fastq.gz illumina_father_reads2.fastq.gz
hifiasm -o assembly_Hifi_Parental.asm -1 mother.yak -2 Father.yak input_reads.fastq.gz

8.3D-Dna (version version 180114) heatmap production and scaffold construction
run-asm-pipeline.sh -r 0 assembly.fa juicer_merged_nodups.txt
run-asm-pipeline-post-review.sh --sort-output -r genome.assembly assembly.fa
juicer_merged_nodups.txt

9. Align file with minimap2 (v2.5)against reference for DGenies visualization
minimap2 -cx asm5 reference.fa assembly.fa > map.paf

10.BUSCO (v5.1)
busco -i assembly.fa -o Busco_assembly -l mammalia_odb10 -m geno

Reference

%todo%

1. Liu, Y., Qin, X., Song, XZ.H. *et al. Bos taurus* genome assembly. *BMC Genomics* **10,** 180 (2009).

2. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers
Guillaume Marcais; Carl Kingsford
Bioinformatics (2011) 27(6): 764-770 first published online January 7, 2011
doi:10.1093/bioinformatics/btr011

3.Vurture, GW, Sedlazeck, FJ, Nattestad, M, Underwood, CJ, Fang, H, Gurtowski, J, Schatz, MC (2017) Bioinformatics doi: https://doi.org/10.1093/bioinformatics/btx153

4.Ruan, J. and Li, H. (2019) Fast and accurate long-read assembly with wtdbg2. Nat Methods doi:10.1038/s41592-019-0669-3

5.Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. . 2017;27:737–46. doi:10.1101/gr.214270.116.

6.Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS ONE 9(11): e112963. https://doi.org/10.1371/journal.pone.0112963

7.Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, **18**:170-175. https://doi.org/10.1038/s41592-020-01056-5

8. Je cherche

9.Neva C. Durand*, James T. Robinson*, Muhammad S. Shamim, Ido Machol, Jill P. Mesirov, Eric S. Lander, and Erez Lieberman Aiden. "Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom." Cell Systems 3(1), 2016.

10. JUicer

11.Cabanettes F, Klopp C. (2018) D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6:e4958 https://doi.org/10.7717/peerj.4958

12.Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210-2. doi: 10.1093/bioinformatics/btv351. Epub 2015 Jun 9. PMID: 26059717.

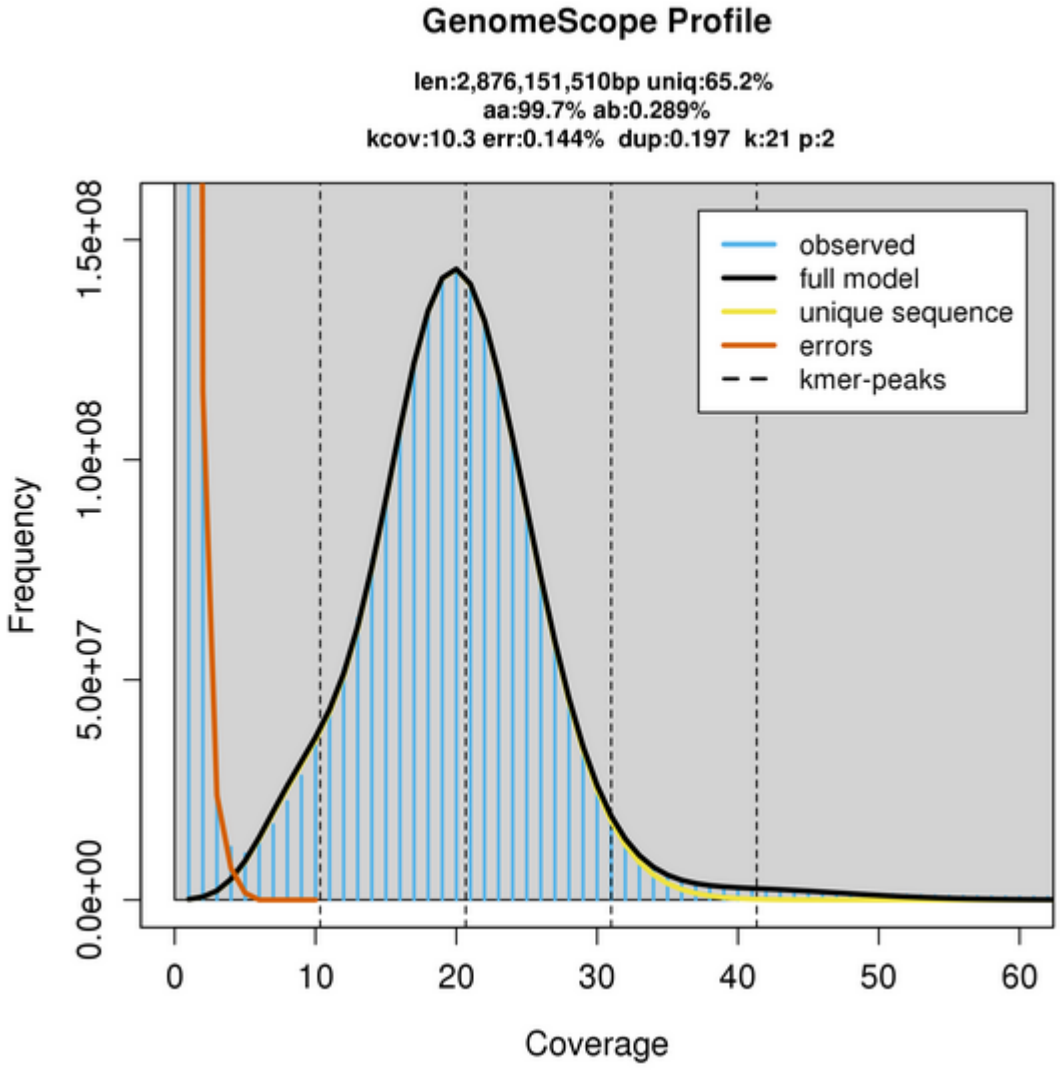


Fig1

Fig2 : Map HiC

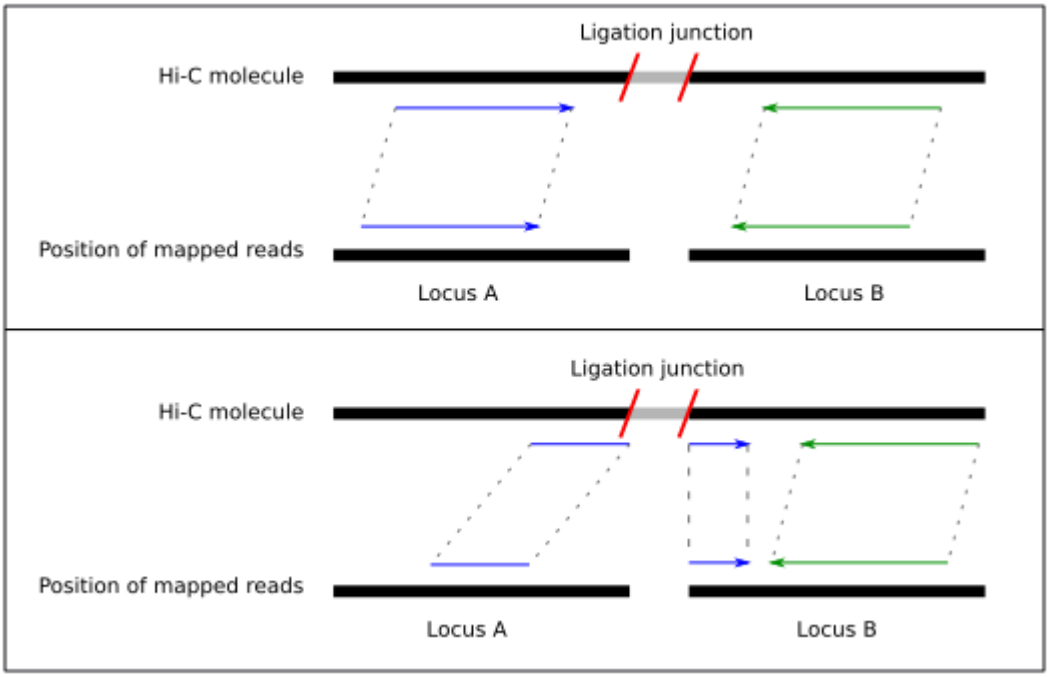
Fig3 : Dgenies

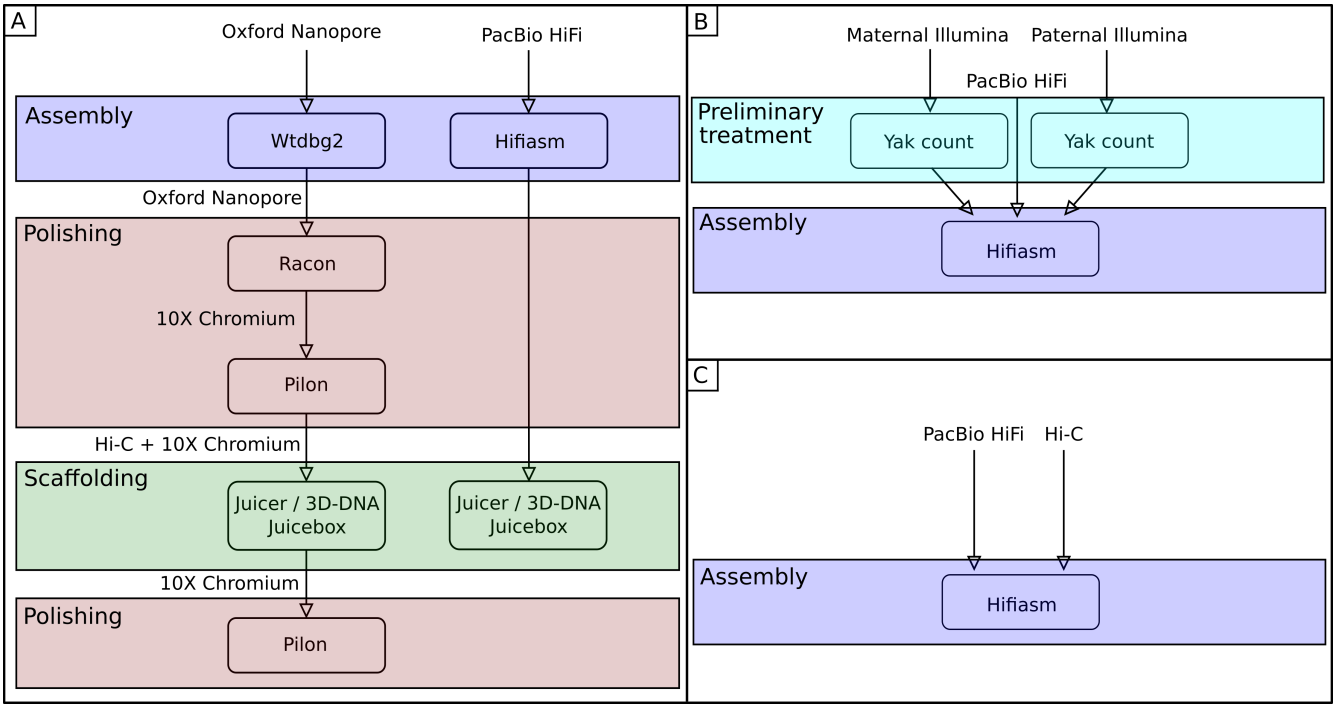
Table1 :

		37162 Mother2	37161 Father2	37160 Offspring2	37160 Offspring2	37160 Offspring2		37160 Offspring2	
Contig assembly	**Data type**	ONT	ONT	ONT	CCS	CCS + parents illumina		CCS + HiC	
	Quantity	52X	58X	58X	31X	31X + (84X + 100X)		31X + 28X	
	Assembler	Wtdbg2	Wtdbg2	Wtdbg2	Hifiasm	Hifiasm		Hifiasm	
	Number of contigs	7 755	6 763	7 226	1 444	2 871	2 300	2 658	2 136
	Total size	2 717 660 923	2 677 825 133	2 701 288 401	3 244 632 679	3 156 028 877	3 113 483 345	3 077 978 241	3 184 033 110
	Longest contigs	106 352 596	73 447 159	70 660 114	158 432 916	158 635 831	159 192 497	159 802 830	159 189 228
	N50 contigs length	23 061 806	23 343 626	23 641 545	84 059 894	71 619 842	69 165 538	80 106 842	71 644 334
	BUSCO	C:72.5% [S:71.5%,D:1.0%] F:7.0%,M:20.5%	C:70.4% [S:69.4%,D:1.0%] F:7.0%,M:22.6%	C:70.2% [S:69.1%,D:1.1%] F:7.3%,M:22.5%	C:95.9% [S:94.0%,D:1.9%] F:1.2%,M:2.9%	C:95.8% [S:93.8%,D:2.0%] F:1.2%,M:3.0%	C:95.3% [S:93.4%,D:1.9%] F:1.2%,M:3.5%	C:95.7% [S:93.8%,D:1.9%] F:1.2%,M:3.1%	C:95.8% [S:93.9%,D:1.9%] F:1.2%,M:3.0%
Polishing	**Data type**	ONT / 10X	ONT / 10X	ONT / 10X					
	Quantity	52X / 100X	58X / 84X	58X / 102X					
	Polisher	Racon / Pilon	Racon / Pilon	Racon / Pilon					
	Number of contigs	6 136	5 758	5 783					
	Total size	2 713 741 594	2 682 719 214	2 700 580 867					
	Longest contigs	106 883 913	73 822 673	71 054 616					
	N50 contigs length	23 877 165	23 478 552	23 984 524					
	BUSCO	C:95.1% [S:93.3%,D:1.8%] F:1.4%,M:3.5%	C:94.1% [S:92.4%,D:1.7%] F:1.4%,M:4.5%	C:95.2% [S:93.5%,D:1.7%] F:1.4%,M:3.4%					
Final	**Data type**	HiC / 10X	HiC / 10X	HiC / 10X	HiC / 10X				
	Quantity	30X/100X	15X / 84X	28X / 102X	28X / 102X				
	Scaffolder	3D-Dna	3D-DNA	3D-DNA	3D-Dna				
	Numberof scaffolds	419	418	4 600	1 391				
	Total size	2 715 671 474	2 683 991 743	2 705 347 253	3 244 660 179				
	Longest scaffolds	157 620 661	157 098 391	156 401 018	175 340 775				
	N50 scaffolds length	101 781 768	104 841 493	100 959 810	87 697 707				
	BUSCO	C:95.1% [S:93.3%,D:1.8%] F:1.4%,M:3.5%	C:94.1% [S:92.4%,D:1.7%] F:1.4%,M:4.5%	C:95.2% [S:93.5%,D:1.7%] F:1.4%,M:3.4%	C:95.8% [S:93.9%,D:1.9%] F:1.2%,M:3.0%				

Table2 :

	Offspring	Mother	Father
Sequenced reads pair	249 106 334	273 714 576	228 056 189
Normal paired	144 928 674 (58.18%)	162 123 850 (59.23%)	165 280 136 (72.47%)
Chimeric Paired	89 927 036 (36.10%)	96 249 728 (35.16%)	49 971 564 (21.91%)
Chimeric Ambiguous	10 611 105 (4.26%)	10 654 890 (3.89%)	9 008 267 (3.95%)
Unmapped	3 639 519 (1.46%)	4 686 108 (1.71%)	3 796 222 (1.66%)
Alignable (Normal+Chimeric)	234 855 710 (94.28%)	258 373 578 (94.40%)	215 251 700 (94.39%)
Unique Reads	191 959 319 (77.06%)	197 408 456 (72.12%)	185 631 137(81.40%)
Intra-fragment Reads	19 673 272 (7.90%)	24 134 852 (8.82%)	82 312 082 (36.10%)
Below MAPQ Threshold	20 420 058 (8.20%)	21 593 575 (8.89%)	25 307 362 (11.10%)
Inter-chromosomal	85 424 851 (34.29%)	79 253 824 (28.95%)	46 609 959 (20.44%)
Intra-chromosomal	66 441 138 (26.67%)	72 426 205 (26.46%)	31 401 734 (13.77%)
Short Range (<20Kb)	25 101 061 (10.08%)	29 162 137 (10.65%)	13 035 849 (5.72%)
Long Range (>20Kb)	41 340 050 (16.60%)	43 264 001 (15.80%)	18 365 834 (8.05%)




A
Oxford Nanopore
PacBio HiFi
Assembly
Wtdbg2
Hifiasm
Oxford Nanopore
Polishing
Racon
10X Chromium
Pilon
Hi-C + 10X Chromium
Scaffolding
Juicer / 3D-DNA
Juicebox
Juicer / 3D-DNA
Juicebox
10X Chromium
Polishing
Pilon
B
Maternal Illumina
Paternal Illumina
PacBio HiFi
Preliminary
treatment
Yak count
Yak count
Assembly
Hifiasm
C
PacBio HiFi
Hi-C
Assembly
Hifiasm