data

MDPI

Data Descriptor

Reference-Guided Draft Genome Assembly, Annotation and SSR Mining Data of the Peruvian Creole Cattle (*Bos taurus*)

Richard Estrada [1], Flor-Anita Corredor [1], Deyanira Figueroa [1], Wilian Salazar [1], Carlos Quilcate [1], Héctor V. Vásquez [1], Jorge L. Maicelo [1,2], Jhony Gonzales [3] and Carlos I. Arbizu [1,*]

[1] Dirección de Desarrollo Tecnológico Agrario, Instituto Nacional de Innovación Agraria (INIA), Av. La Molina 1981, Lima 15024, Peru
[2] Facultad de Zootecnia, Agronegocios y Biotecnología, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Chachapoyas 01001, Peru
[3] Laboratorio de Biología Molecular, Universidad Nacional de Frontera, Av. San Hilarión 101, Sullana 20103, Peru
* Correspondence: carbizu@inia.gob.pe; Tel.: +51-986288181

Abstract: The Peruvian creole cattle (PCC) is a neglected breed and an essential livestock resource in the Andean region of Peru. To develop a modern breeding program and conservation strategies for the PCC, a better understanding of the genetics of this breed is needed. We sequenced the whole genome of the PCC using a de novo assembly approach with a paired-end 150 strategy on the Illumina HiSeq 2500 platform, obtaining 320 GB of sequencing data. A reference scaffolding was used to improve the draft genome. The obtained genome size of the PCC was 2.81 Gb with a contig N50 of 108 Mb and 92.59% complete BUSCOs. This genome size is similar to the genome references of *Bos taurus* and *B. indicus*. In addition, we identified 40.22% of repetitive DNA of the genome assembly, of which retroelements occupy 32.39% of the total genome. A total of 19,803 protein-coding genes were annotated in the PCC genome. For SSR data mining, we detected similar statistics in comparison with other breeds. The PCC genome will contribute to a better understanding of the genetics of this species and its adaptation to tough conditions in the Andean ecosystem.

Dataset: The genome sequence is openly available in the Genbank of NCBI under the accession number JANIWY000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JANIWY000000000.1 accessed on 4 October 2022). The associated Bioproject, Biosample, and Sequence Read Archive (SRA) numbers are PRJNA849594, SAMN29095626, and SRS13407845, respectively.

Keywords: NGS; neglected breed; genome; reference scaffolding; microsatellites

Citation: Estrada, R.; Corredor, F.-A.; Figueroa, D.; Salazar, W.; Quilcate, C.; Vásquez, H.V.; Maicelo, J.L.; Gonzales, J.; Arbizu, C.I. Reference-Guided Draft Genome Assembly, Annotation and SSR Mining Data of the Peruvian Creole Cattle (*Bos taurus*). *Data* **2022**, *7*, 155. https://doi.org/10.3390/data7110155

Academic Editor: Pufeng Du

Received: 17 August 2022
Accepted: 1 November 2022
Published: 9 November 2022

1. Summary

According to Scheu et al. [1], cattle domestication started in the ninth millennium BC in Southwest Asia. Similarly, Upadhyay et al. [2] referred to European cattle's genetic origin and domestication to start around 10,000 years ago in the Near East. Over the years, its use has been extended worldwide, where cattle species have been distributed and adapted to various climates. The genus *Bos* is divided into six species: *B. gaurus*, *B. javanicus*, *B. mutus*, *B. bison*, *B. sauveli*, and *B. primigenius* [3]. Of these, four are domesticated species: *B. mutus*, *B. javanicus*, *B. gaurus*, and *B. primigenius*, which are represented by their domestic forms *B. taurus* and *B. indicus* [3]. The taxonomic status of *B. taurus* and *B. indicus* are controversial [4]. Through a mitochondrial analysis, Hiendleder et al. [4] determined that *B. taurus* and *B. indicus* lineages diverged 1.7–2.0 million years ago, suggesting these species deserved a subspecies status for taurine and zebuine cattle. The genomics of cattle have been fully studied, with its genome being completely sequenced by 2009; *B. taurus* is one

of the most studied species in the livestock area [5]. This project was developed by more than 300 scientists from different countries. Similar efforts are being performed by other institutes to broaden the knowledge of the Hereford reference breed.

Genetic characterization studies of the creole cattle from Latin America are still limited. Delgado et al. [6] characterized Latin-American creole cattle from 10 countries using 19 microsatellite markers, which included 26 creole cattle breeds. Their results indicated high genetic diversity among creole cattle, suggesting the implementation of conservation measures. Similarly, Giovambattista et al. [7] reported bovine MHC *DRB3* gene diversity in Bolivian "Yacumeño" cattle and Colombian "Hartón del Valle" cattle. The authors' results suggested a high level of genetic diversity for these breeds that could be explained tentatively by multiple origins of creole germplasm (European, African, and Indicus). In a comprehensive study, Ginja et al. [8] evaluated the genetic ancestry of American creole cattle utilizing microsatellite markers, mitochondrial DNA, and Y chromosome information. They sampled 40 creole breeds representing the whole American continent. In addition to those already considered by Delgado et al. [6], cattle from the Latin American and Caribbean countries of Bolivia, Chile, Suriname, and Venezuela were sampled. Ginja et al. concluded that creole cattle have a mixed ancestry where African cattle have played a role in its development. Unfortunately, none of these studies included samples or information from Peruvian cattle. Recently, Raschia and Poli [9] employed a medium-density SNP array to characterize Argentinian creole cattle. They concluded that the genetic relationships showed a close relationship among four groups of creole animals from Argentina. Liu et al. [10] studied the mitochondrial genome of Uruguayan native cattle and demonstrated that it clustered with Korean breeds. In addition, Aguirre Riofrio et al. [11] performed a microsatellite analysis for the genetic characterization of the creole cattle in the southern region of Ecuador. They concluded that the bovines studied are genetically distant from zebuine breeds and their ancestral origin must be related to the Iberic populations. Aracena and Mujica [12] reported the morphological and reproductive characterization of the Chilean Patagonian bovine and indicated that brown is the color base for its hair. They also compared the Chilean Patagonian bovine to Argentinian cattle, finding similarities in productive and reproductive aspects.

There is a significant source of genetic variation in cattle breeds. There are more than 700 breeds of cattle worldwide [3]. Generating assemblies from short reads in large genomes such as bovine is challenging. This is largely due to the repetitive sequences that these genomes contain. However, possession of more complete sequences of draft genomes is very important for future biological applications. One approach by which contigs can be scaffolded is to use references of the same or related species. This strategy yielded much larger contigs and improved assembly parameters [13]. In this type of strategy, an order of genetic markers very similar to the reference genome has been found, given that the components of genomic rearrangements between them are very rare [14]. However, many structural errors are introduced into the final assembly when compared to direct assembly approaches that do not use a reference genome [15].

There are important initiatives to study genetic variation through the study of bovine pangenomes [16,17]. The first reference to the *B. taurus* genome was obtained in 2004 based on a Hereford individual. The latest update of this genome was done in the same cow using continuous long-reading sequencing from Pacific Biosciences (ARS-UCD1.2) [18]. On the other hand, it has been possible to carry out efforts such as that of the Bovine Pangenome Consortium (https://njdbickhart.github.io/ accessed on 25 July 2022), where information on alleles and haplotypes of more than 600 different known breeds of cattle is being obtained worldwide. With the decrease in sequencing costs, genomic studies can be carried out in other regions that have not had precedents, such as the South American region.

In Peru, bovine creole cattle have received a limited amount of research attention. Through the use of six microsatellite markers, Yalta-Macedo et al. [19] inferred the PCC ancestry and proposed that it descended from cattle from the Iberian peninsula. This study also suggested that male-mediated African cattle influenced the PCC. More recently,

Arbizu et al. [20] confirmed these findings, by constructing a phylogenetic tree and revealing the phylogenetic relationships with the African cattle breeds Boran and Arsi. According to M. Rosemberg (UC del Sur, pers. comm.), PCC can be found as a crossbreed with Brown Swiss breeds around 3500 meters above sea level (MASL). PCC is mainly distributed in the Peruvian highlands [21]. In Peru, there is no strong national registration program to record pedigree and productivity for PCC as in other countries [22]. PCC are phenotypically distinguishable by their smaller size when compared to other exotic breeds. Comprehensive efforts to determine the full potential of muscle growth of the PCC have been conducted [23–25].

Therefore, further studies about PCC genomics will be beneficial for conservation programs and future selection activities. For this, additional sampling of bovine creole individuals is being performed by the National Institute of Agrarian Innovation (Spanish acronym INIA), a Peruvian government research institution. The goal of this study is to contribute to the understanding of the PCC molecular characterization, as well as its comparative genomics among the Bovinae subfamily.

2. Data Description

The whole genome sequencing data was deposited in the SRA database under code number SRS13407845, Biosample SAMN29095626, and Bioproject PRJNA849594. The total number of raw pair reads was 854,737,766 sequences, with an average length of 150 bp, a GC content of 44%, and a total sequencing data output of 320 GB. After the trimming, we retained 88.2% of sequencing data, and more than 790 million high-quality sequence reads with approximately 281.6 GB of total sequencing data were generated. No over-represented sequences and adapters were found. In addition, the average quality per read was 40.

2.1. Genomic Survey

We obtained low heterozygosity, moderate repetition (33.3%), and the genome size estimate (2.58 Gb) was close to the reported references of the genome of *B. taurus* (ARS-UCD1.3: 2.71 Gb, ARS-LIC_NZ_Jersey: 2.64 Gb, ARS-LIC_NZ_Holstein-Friesian_1: 2.66 Gb, Brown Swiss: 2.66 Gb, and Btau_5.0.1: 2.72 Gb). Based on the estimated draft genome size, subsequent de novo assembly and genome annotation were performed with a sequencing depth of approximately 47.44 X coverage (Figure 1).

2.2. Assembly De Novo, Reference-Assisted Scaffolding, and Validation

Different assemblies from the SOAPdenovo2 [26] and MaSuRCA [27] programs were obtained. We continued our analysis with MaSuRCA due to a better N50 and longer contigs (Table S1). MaSuRCA assembly was scaffolded with reference-guided scaffolding, and we obtained a total length of 2.77 Gb, which had 10,953 contigs ($\geq$1000 bp) with a GC content of 41.87 %. The longest contig was 164,677,778 bp. In addition, we found that 99.21% of the raw paired-end reads generated from the small insertion libraries were aligned to our final assembled genome. With the scaffolding approach, we improved the N50 from 12.84 kb to 108.72 Mb (Table 1). In addition, the number of complete mammalian single-copy BUSCOs (Benchmarking Universal Single-Copy) increased from 1620 to 3800 complete BUSCOs (Table 2). We obtained 3744 complete and single-copy BUSCOs (S), 56 complete and duplicated BUSCOs (D), 165 fragmented BUSCOs, and 139 missing BUSCOs.

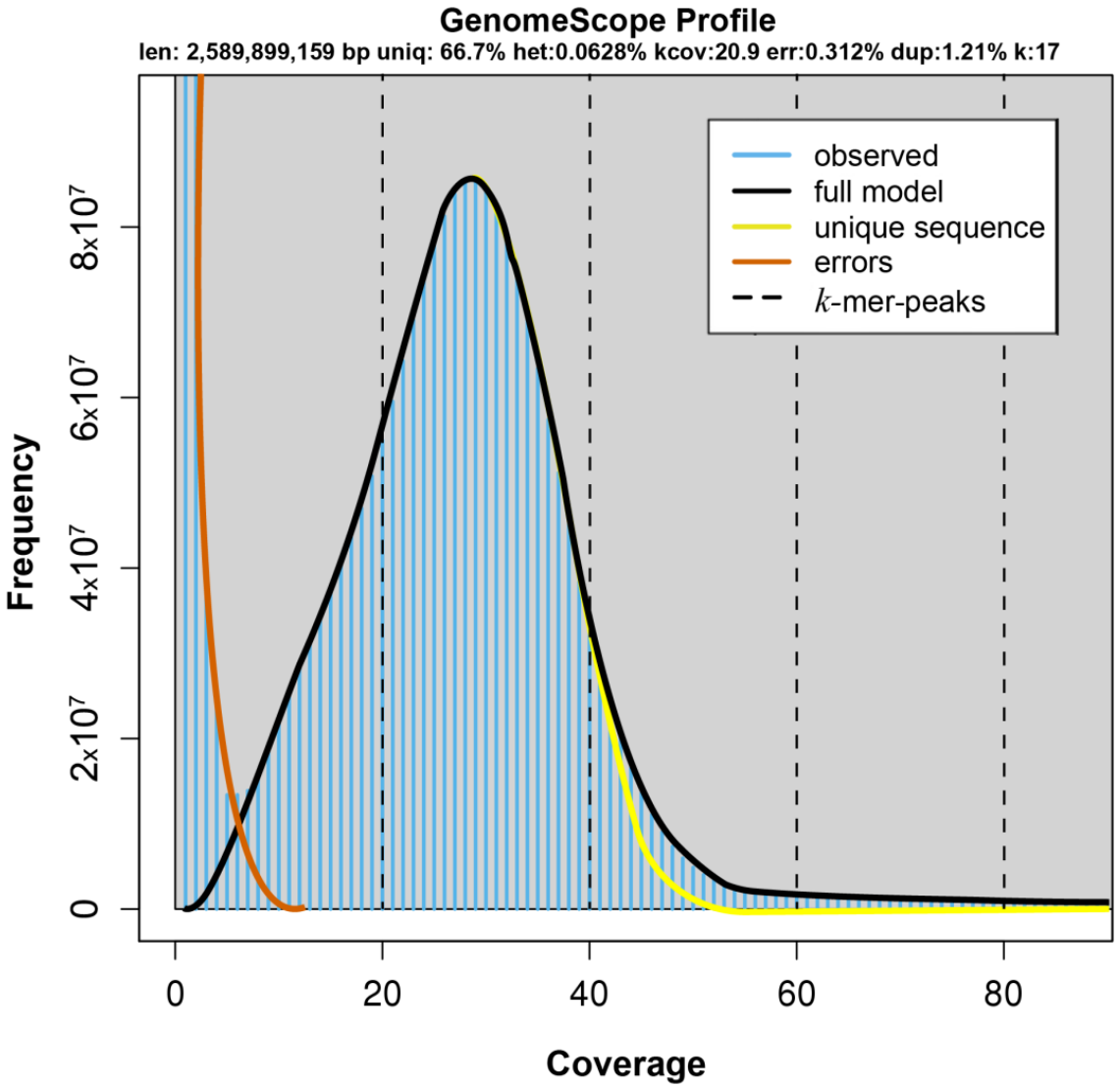

Figure 1. Distribution of *k*-mers in the draft genome of Peruvian creole cattle.

Table 1. Statistics of the completeness of the de novo assembly of the Peruvian creole cattle genome.

Statistic	Contigs	Scaffolds
N50	12,843	108,727,214
N75	7242	74,944,637
L50	63,921	11
L75	133,082	19
Largest contig	109,017	164,677,788
Total length	2,679,899,159	2,814,362,078
GC (%)	41.92	41.87
# contigs (≥1000 bp)	307,114	10,953
# contigs (≥5000 bp)	179,627	1848
# contigs (≥10,000 bp)	92,431	777
# contigs (≥25,000 bp)	14,279	210
# contigs (≥50,000 bp)	726	75
# N's per 100 kbp	0.0	5710.08

These correspond to "number of".

Table 2. Summary of the BUSCO approach in in the Peruvian creole cattle assembly (contigs and scaffolds).

Terms	Contigs	Scaffold
Complete BUSCOs	1620	3800
Complete and single-copy BUSCOs	1580	3744
Complete and duplicated BUSCOs	40	56
Fragmented BUSCOs	1573	165
Missing BUSCOs	911	139

Compared to other cattle breeds' scaffold level assembly (Brown Swiss 26.03 Mb N50) and chromosome level assembly (Hereford: 103.31 Mb, Jersey: 104.07 Mb, Holstein: 100.96 Mb, N'Dama: 104.85 Mb, Nelore: 106.31 Mb, Gyr: 104.3 Mb, and Ankole: 84.48 Mb) our assembly has 108.72 MB N50, showing a high level of scaffolding (Table 3). Additionally, our assembly has 92.59% complete BUSCOs (C) (S: 91.2% + D:1.4%), similar to the Nelore breed with 92.9% C (S: 91.9% + D:1%) and the Hereford breed with 91.7% C (S: 90.6% + D:1.1%) (Figure 2).

Table 3. Comparison of Peruvian creole cattle (PCC) assembly with other *B. taurus* and *B. indicus* species.

	B. taurus					***B. indicus***			
Breed	**PCC**	**Hereford**	**Jersey**	**Holstein**	**Brown Swiss**	**N'Dama**	**Nelore**	**Gyr**	**Ankole**
Level Assembly	Scaffold	Chromosome	Chromosome	Chromosome	Scaffold	Chromosome	Chromosome	Chromosome	Chromosome
Total sequence length	2,814,362,078	2,711,209,831	2,641,777,256	2,665,549,695	2,658,221,619	2,766,829,411	2,673,965,444	2,740,330,345	2,921,040,163
Total ungapped length	2,653,670,481	2,711,181,669	2,641,709,256	2,665,138,195	2,635,427,799	2,708,415,641	2,475,828,999	2,695,917,733	2,834,561,153
Gaps between scaffolds	0	0	0	0	0	0	0	0	0
Number of scaffolds	12,639	1957	229	2306	14,725	1210	32	216,409	7581
Scaffold N50	110,880,623	103,308,737	104,068,235	100,964,413	26,027,505	104,847,410	106,310,653	104,295,553	84,476,814
Scaffold L50	11	12	11	12	29	12	11	12	13
Number of contigs	761,086	2343	365	3129	34,351	3601	253,770	337,292	8473
Contig N50	6381	25,896,116	50,551,513	8,737,306	268,406	11,058,985	28,375	64,498	18,716,610
Contig L50	126,374	32	17	90	2856	71	25,227	11,998	49
Total number of chromo-somes and plasmids	0	31	32	32	0	32	32	31	30
Number of component sequences (WGS or clone)	12,639	1957	229	2306	14,725	1210	253,770	216,409	7581

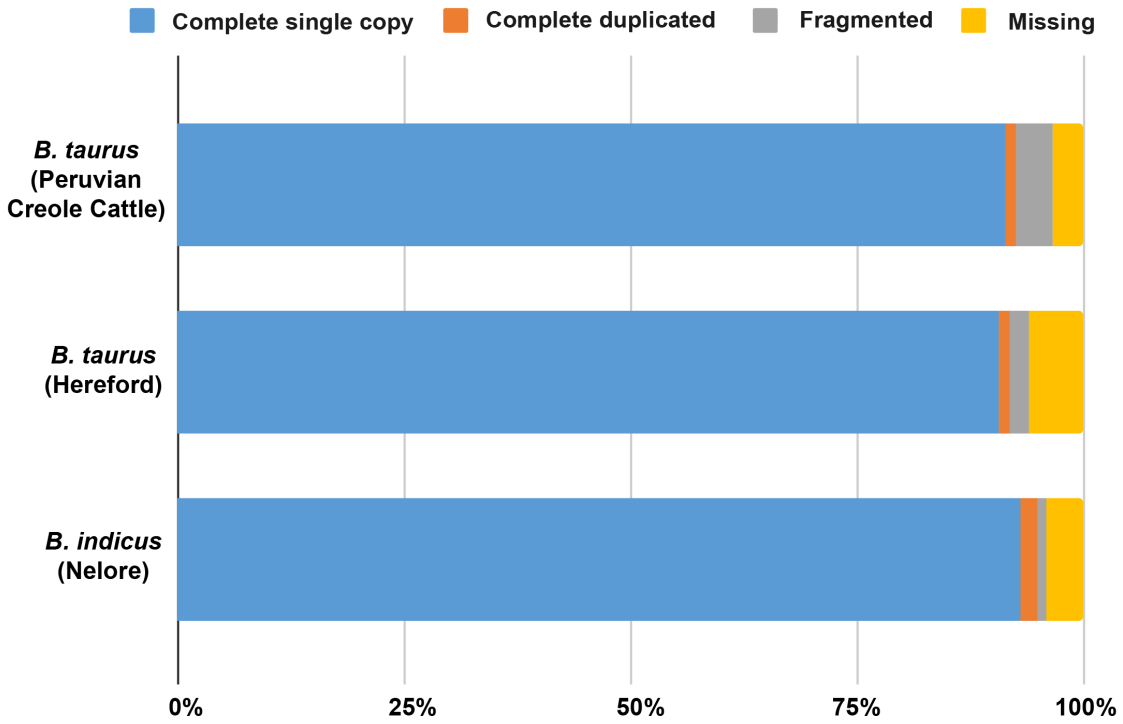

Figure 2. Comparison of BUSCO analysis of the Peruvian creole cattle with the references of *B. taurus* and *B. indicus*.

2.3. Genomic Annotation

We used gene prediction methods (ab initio prediction and homology-based search) to annotate the protein-coding genes in the draft genome, resulting in 19,803 annotated protein-coding genes. We found that our assembly and annotation (19,803 protein-coding genes) is not as complete as that of the *B. taurus* reference genome (ARS-UCD1-HerefordGenbank, GCF_002263795.1), which had 21,039 protein-coding genes. On the other hand, we identified 897,585,367 bp retroelements, corresponding to 32.29% of the assembled PCC draft genome. The most abundant repetition was the LINES type, which represented 28.55% of the total assembly. In addition, we identified other types of repetitive DNAs: the RTE/Bov-B type (16.32%) and L1/CIN4 (10.51%). Notably, 3.48% of repetitive unclassified DNA was found, and 37.38% of the total assembled genome has been classified as a total interspersed repeat (Table 4).

Table 4. Summary of the repetitive DNA of Peruvian creole cattle.

Repetitive DNA	Number of Elements	Length Occupied	Percentage of Sequence
Retroelements	3,484,900	897,585,367 bp	32.39%
SINEs	256.918	28,733,533 bp	104%
LINEs	2,890,366	791,282,631 bp	28.55%
L2/CR1/REX	173.451	19,529,331 bp	0.70%
RTE/Bov-B	1,426,552	452,420,074 bp	16.32%
L1/CIN4	1,111,156	291,303,229 bp	10.51%
LTR elements	33.7616	77,569,203 bp	2.80%
Retroviral	337.127	77,499,189 bp	2.80%
DNA transposons	245.87	41,992,077 bp	1.52%
hobo-Activator	84.758	27,282,703 bp	0.98%
Tc1-IS630-Pogo	60.623	14,480,775 bp	0.52%

Table 4. *Cont.*

Repetitive DNA	Number of Elements	Length Occupied	Percentage of Sequence
Unclassified	665.577	96,490,946 bp	3.48%
Total interspersed repeats		1,036,068,390 bp	37.38%
Small RNA	161.025	17,146,359 bp	0.62%
Satellites	700	416,318 bp	0.02%
Simple repeats	499.594	20,282,441 bp	0.73 %

2.4. SSR Data Mining

The most abundant microsatellite motif type of the PCC were mononucleotide repeats, accounting for 59% (593,627) of the total SSRs, followed by dinucleotide repeats (26.3% or 264,341), trinucleotide repeats (12.1% or 121,761), tetranucleotide repeats (1.2% or 11,665), pentanucleotide repeats (1.4% or 13,824), and, finally, hexanucleotide repeats (0.034% or 346). This is similar to the microsatellite motif distribution of other breeds such as Hereford, Jersey, Holstein, Brown Swiss, N'Dama, Nelore, Gyr, and Ankole (Figure 3A, Table S2).

A total of 1,005,564 microsatellite loci were identified based on the assembled PCC draft genome sequence, with a frequency of 376.62 SSR/Mb, which is almost the same as N'Dama (376.89 SSR/Mb), lower than Jersey (379.76 SSR/Mb), but higher than Holstein (376.01 SSR/Mb), Brown Swiss (375.63 SSR/Mb), Ankole (374.38 SSR/Mb), Hereford (372.55 SSR/Mb), Gyr (360.12 SSR/Mb), and Nelore (336.2 SSR/Mb) (Table 5). In addition, the number of SSRs present in the compound formation of PCC (92,354) was very similar to the other breeds. In addition, it is highlighted that the size of the genomes of the other *B. taurus* breeds is very similar to the Peruvian creole cattle assembly (Figure 3B).

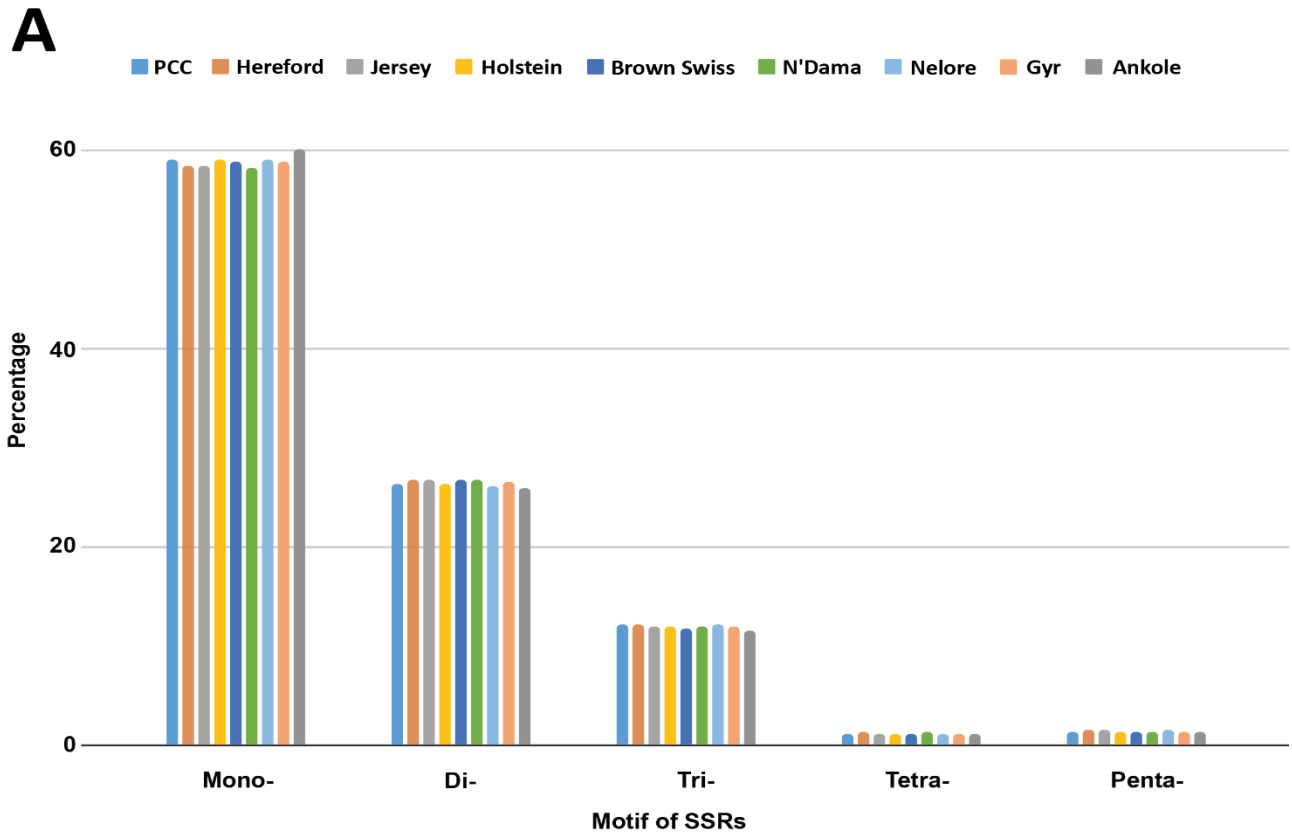

Figure 3. *Cont.*

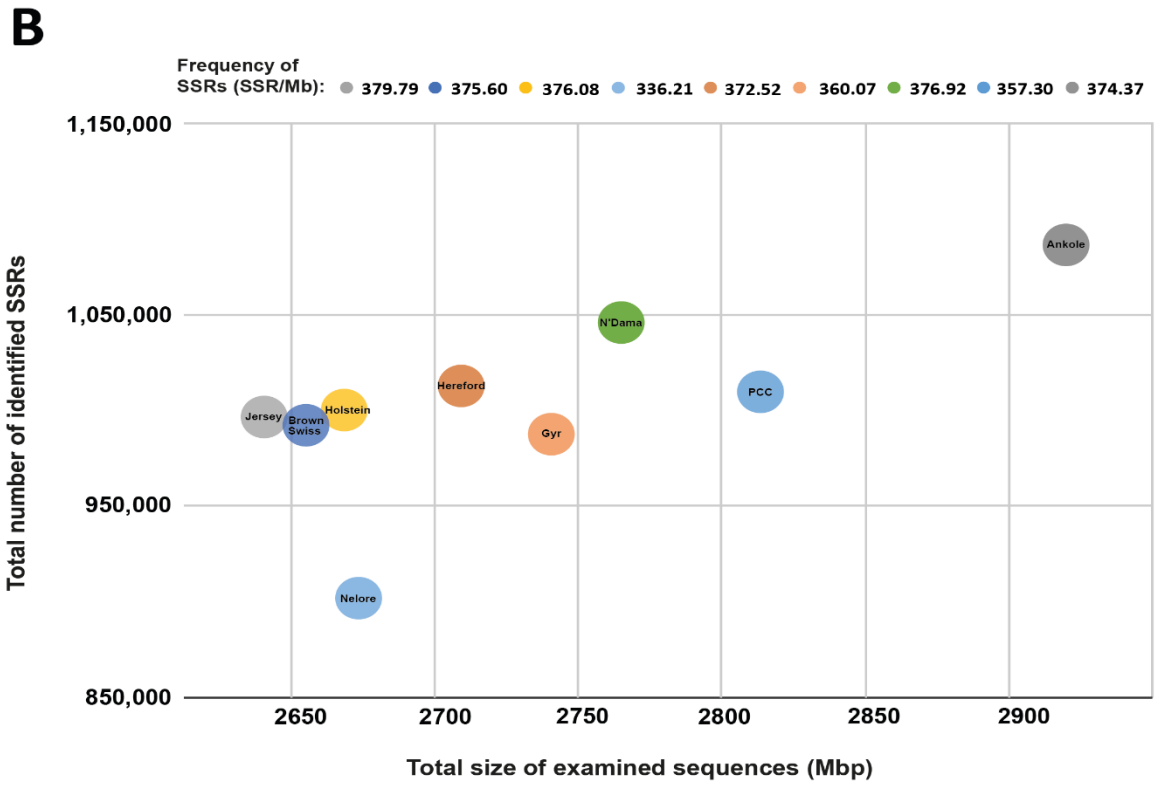

Figure 3. Distribution of SSRs in breeds. (**A**) Percentage of SSR per motif in the Peruvian creole (PCC) in comparison with other breeds. (**B**) Bubbles that represents the total number of identified SSRs, the total size of the examined sequences (Kbp), and the frequency of SSRs (SSR/Mb).

Table 5. Summary of SSR distribution in Peruvian creole cattle (PCC) and other *B. taurus* and *B. indicus* species.

	B. taurus						*B. indicus*		
Breed	**PCC**	**Hereford**	**Jersey**	**Holstein**	**Brown Swiss**	**N'Dama**	**Nelore**	**Gyr**	**Ankole**
Total size of examined sequences (Gbp, with gaps)	2.8144	2.7112	2.6418	2.666	2.6582	2.7668	2.67380	2.7403	2.9210
Total number of identified SSRs	1,005,564	1,009,980	1,003,327	1,002,450	998,430	1,042,868	899,003	986,718	1,093,552
Frequency (SSR/Mb)	357.30	372.52	379.79	376.08	375.60	376.92	336.21	360.07	374.37
Number of SSRs present in compound formation	92,354	104,313	104,005	104,288	98,838	116,622	82,308	93,628	132,842

2.5. Concluding Comments

In summary, we reported the first draft genome of the Peruvian creole cattle. Our draft genome is very similar to other reported draft genomes of *B. taurus*. In addition, we generated information about SSR which will be employed in the near future for population genetics and diversity studies. Since there are limited genomic sequence resources for the PCC, our study hopes to provide a reference for animal improvement programs for this important livestock resource.

3. Methods

3.1. Sample Collection and DNA Extraction

We collected hair samples from the tail of a single male specimen from Andagua, Arequipa (3574 MASL; −15.499548°, −72.359927°). Since this individual possessed most of the classical characteristics of Peruvian creole cattle, we decided to select it for this study.

We extracted genomic DNA with the Wizard Genomic DNA Purification Kit (Fitchburg, WI, USA) following the manufacturer's instructions. The quality and quantity of genomic DNA were assessed using agarose gel electrophoresis and a Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA, USA), respectively. The mitochondrial genome of this individual was recently sequenced [20].

3.2. Genome Sequencing and Genomic Survey

We used a library with 300 bp insert size and paired-end-tag DNA sequencing using the Illumina HiSeq 2500 platform, generating around 2x151 bp reads. A reference-scaffolding approach was used to improve the draft genome, and the raw reads were checked by FastQC v.0.11.9 [28]. In addition, trimming quality (Phred Q > 25) and remotion of adapters were conducted with Trimmomatic v0.36 [29] and TrimGalore software [30], respectively. For the genomic survey, we used Jellyfish v.2.0 [31]. Genome Scope v1.0.0 (Cold Spring Harbor Laboratory, Laurel Hollow, NY, USA) [32] was employed to estimate the features of the genome, including genome size, repeat content, and heterozygosity rate, using the output of Jellyfish and the number of 17-mer for *k*-mer analysis. K-depth was estimated to identify a common single-peak pattern in the *k*-mer frequency distribution analysis.

3.3. De Novo Assembly and Validation

De novo assembly was performed with two assembly algorithms: SOAPdenovo2 v.2.04 [26] and MaSuRCA v.4.0.6 [27]. Next, we used QUAST v.5.2.0 [33] for statistics of assemblies. MaSuRCA resulted in improved assembly statistics and was subjected to Samba scaffolder v.1.0 [34] for scaffolding and gap-filling. For the reference-based scaffolding, we used a reference genome of *B.taurus* (Genbank: GCA_002263795.3) since this is the last updated genomic tool for this species. Next, we used QUAST with the output of the scaffolding. Validation of assembly was assessed using three different approaches: (i) filtered PE Illumina reads were remapped to detect errors in the assembly using Bowtie2 v.2.4.2 [35] and SamTools v.1.7 [36] software, (ii) the BUSCO [37] strategy was used to test the completeness of the genome assembly and gene space, using the mammalian-specific profile—this approach makes use of single-copy genes expected to be present in mammals (4104 genes)—and (iii) available *B. taurus* genomic resources such as CDS (coding DNA sequences) and PacBio transcriptomes data were used to map back to the draft genome using GMAP v.2021.08.25 [38]. We used JCVI VecScreen (https://github.com/tanghaibao/jcvi accessed on 5 July 2022), which uses Univecdatabase (https://ftp.ncbi.nlm.nih.gov/pub/UniVec/ accessed on 20 July 2022) to detect vectors, and mapped the scaffolds against the nt/nr NCBI database (https://www.ncbi.nlm.nih.gov/ accessed on 10 July 2022) using BLAST v.2.2.26 [39] for identifying contamination. The mitochondrial sequences were separated after BLAST searches against databases of mitochondrial sequences. Finally, we removed contaminated scaffolds and vectors to submit to the NCBI database. This assembly has been deposited at DDBJ/ENA/GenBank under the accession number JANIWY000000000.

3.4. Genome Annotation

To identify repetitive elements, we used de novo and homolog-based methods. For the de novo approach, we used Repeatmodeler [40] to generate a de novo PCC repeat library, which is subsequently used in RepeatMasker v4.0.7 [41] to annotate repeats. For the homology-based approaches, we used Repbase v4.0.7 [42], RepeatMasker, and RM-Blastv2.2.27 [43]. All repeat results were merged. Final genome assembly was repeat-masked using the library repeats using RepeatMasker. MAKER [44] was run on the repeat-masked genome with SNAP [45] and AUGUSTUS [46]. For evidence to guide the annotation process, we retrieved ESTs of *B. taurus* from the NCBI database (ftp://ftp.ncbi.nih.gov/repository/dbEST/ accessed on 13 July 2022) and proteomes of the Bovidae species *B.taurus* (Refseq: GCF_002263795.2), *B. indicus* (Refseq: GCF_000247795.1), and *B. mutus* (GCF_000298355.1). MAKER software was run iteratively two times; the predictions were

curated against a high-quality protein database of UNIPROT (https://www.uniprot.org/ accessed on 15 July 2022) using BLAST with an E-value of 1×10^{-5}.

3.5. Identification of Simple Sequence Repeats

The SSRs were identified in the PCC genome using MISA Perl script (http://pgrc.ipk-gatersleben.de/misa/ accessed on 30 July 2022) [47] with the specific settings: monomer (one nucleotide, $n > 12$), dimer (two nucleotides, $n > 6$), trimer (three nucleotides, $n > 4$), tetramer (four nucleotides, $n > 3$), pentamer (five nucleotides, $n > 3$), hexamer (six nucleotides, $n > 3$). In addition, for SSR analysis, we added the genomes of *B. taurus* breeds: Hereford (GenBank: GCA_002263795.3), Jersey (GenBank: GCA_021234555.1), Holstein (GenBank: GCA_021347905.1), Brown Swiss (GenBank: GCA_914753205.1), N'Dama (GenBank: GCA_905123515.1), Nelore (GenBank: GCA_000247795.2), Gyr (GenBank: GCA_002933975.1), Ankole (GenBank: GCA_905123885.1). We used BUSCO to examine the quality of assemblies. Subsequently, we used the MISA script with the same parameters for PCC.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/data7110155/s1, Table S1: Statistics of the completeness of the assemblies from SOAPdenovo2 and MaSuRCA of the Peruvian creole cattle genome. Table S2: Motif percentage of SSR in cattle breeds.

Author Contributions: Conceptualization, C.I.A., C.Q. and F.-A.C.; methodology, R.E., C.I.A., F.-A.C. and D.F.; software, R.E.; validation, R.E., W.S., and D.F.; formal analysis, R.E., F.-A.C. and D.F.; investigation, C.I.A., H.V.V., J.L.M. and J.G.; resources, C.Q., H.V.V., J.L.M., J.G. and C.I.A.; data curation, R.E. and D.F.; writing—original draft preparation, R.E., F.-A.C. and D.F.; writing—review and editing, R.E., F.-A.C., D.F. and C.I.A.; supervision, C.Q., H.V.V., J.L.M., J.G. and C.I.A.; funding acquisition, C.Q. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the following three projects: (1) "Creación del servicio de agricultura de precisión en los Departamentos de Lambayeque, Huancavelica, Ucayali y San Martín 4 Departamentos", (2) "Mejoramiento de la disponibilidad de material genético de ganado bovino con alto valor a nivel nacional 7 departamentos", and (3) "Creación del Servicio de laboratorio de Biología Molecular para la Investigación en la Universidad Nacional de Frontera – Distrito de Sullana" of the Ministry of Agrarian Development and Irrigation (MIDAGRI) of the Peruvian Government, with grant numbers CUI 2449640, CUI 2432072 and CUI 2437731, respectively. C.I.A. was supported by PP0068 "Reducción de la vulnerabilidad y atención de emergencias por desastres".

Institutional Review Board Statement: The sample collection from the cattle specimen was conducted in accordance with the Peruvian National Law No. 30407: "Animal Protection and Welfare".

Informed Consent Statement: Written informed consent was obtained from the owner of the bull studied here.

Data Availability Statement: All data generated during this study are included in this published article.

Acknowledgments: The authors thank Family Ojeda from Arequipa who kindly provided hair samples of their bull for this work, and we also thank Jerry Valdeiglesias and Fernando Gomez for the logistic support in Arequipa. We appreciate the comments of the anonymous reviewers that greatly improved our manuscript. In addition, we thank Ivan Ucharima, Maria Angélica Puyo, Cristina Aybar, and Erick Rodriguez for supporting the logistic activities in the laboratory. Finally, the authors thank the Bioinformatics High-Performance Computing server of Universidad Nacional Agraria la Molina for providing resources for data analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Scheu, A.; Powell, A.; Bollongino, R.; Vigne, J.D.; Tresset, A.; Çakirlar, C.; Benecke, N.; Burger, J. The Genetic Prehistory of Domesticated Cattle from Their Origin to the Spread across Europe. *BMC Genet.* **2015**, *16*, 54. [CrossRef]
2. Upadhyay, M.R.; Chen, W.; Lenstra, J.A.; Goderie, C.R.J.; Machugh, D.E.; Park, S.D.E.; Magee, D.A.; Matassino, D.; Ciani, F.; Megens, H.J.; et al. Genetic Origin, Admixture and Population History of Aurochs (*Bos primigenius*) and Primitive European Cattle. *Heredity* **2016**, *118*, 169–176. [CrossRef] [PubMed]

3. Garrick, D.J.; Ruvinsky, A. *The Genetics of Cattle*; CABI: London, UK, 2014; ISBN 9781119130536.
4. Hiendleder, S.; Lewalski, H.; Janke, A. Complete Mitochondrial Genomes of *Bos taurus* and *Bos indicus* Provide New Insights into Intra-Species Variation, Taxonomy and Domestication. *Cytogenet. Genome Res.* **2008**, *120*, 150–156. [CrossRef] [PubMed]
5. Bovine Genome Sequencing and Analysis Consortium; Elsik, C.G.; Tellam, R.L.; Worley, K.C.; Gibbs, R.A.; Muzny, D.M.; Weinstock, G.M.; Adelson, D.L.; Eichler, E.E.; Einitski, L.; et al. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* **2009**, *324*, 522–528. [CrossRef] [PubMed]
6. Delgado, J.V.; Martínez, A.M.; Acosta, A.; Álvarez, L.A.; Armstrong, E.; Camacho, E.; Cañón, J.; Cortés, O.; Dunner, S.; Landi, V.; et al. Genetic Characterization of Latin-American Creole Cattle Using Microsatellite Markers. *Anim. Genet.* **2012**, *43*, 2–10. [CrossRef]
7. Giovambattista, G.; Takeshima, S.-N.; Ripoli, M.V.; Matsumoto, Y.; Angela, L.; Franco, A.; Saito, H.; Onuma, M.; Aida, Y. Characterization of Bovine MHC DRB3 Diversity in Latin American Creole Cattle Breeds. *Gene* **2013**, *519*, 150–158. [CrossRef]
8. Ginja, C.; Gama, L.T.; Cortés, O.; Burriel, I.M.; Vega-Pla, J.L.; Penedo, C.; Sponenberg, P.; Cañón, J.; Sanz, A.; do Egito, A.A.; et al. The Genetic Ancestry of American Creole Cattle Inferred from Uniparental and Autosomal Genetic Markers. *Sci. Rep.* **2019**, *9*, 11486. [CrossRef]
9. Raschia, M.A.; Poli, M. Phylogenetic Relationships of Argentinean Creole with Other Latin American Creole Cattle as Revealed by a Medium Density Single Nucleotide Polymorphism Microarray. *Arch. Latinoam. Prod. Anim.* **2021**, *29*, 91–100. [CrossRef]
10. Liu, S.J.; Lv, J.Z.; Tan, Z.Y.; Ge, X.Y. The Complete Mitochondrial Genome of Uruguayan Native Cattle (*Bos taurus*). *Mitochondrial DNA Part B Resour.* **2020**, *5*, 443–444. [CrossRef]
11. Aguirre Riofrio, L.; Apolo, G.; Chalco, L.; Martínez, A. Caracterización Genética de La Población Bovina Criolla de La Región Sur Del Ecuador y Su Relación Genética Con Otras Razas Bovinas. *Anim. Genet. Resour. Génétiques Anim. Genéticos Anim.* **2014**, *54*, 93–101. [CrossRef]
12. Aracena, M.; Mujica, F. Caracterización Del Bovino Criollo Patagónico Chileno: Un Estudio de Caso. *Agro Sur* **2011**, *39*, 106–115. [CrossRef]
13. Shieh, Y.-K.; Liu, S.-C.; Lung, L.C. Scaffolding Contigs Using Multiple Reference Genomes. In *Computational Biology and Chemistry*; Behzadi, P., Bernabò, N., Eds.; IntechOpen: London, UK, 2020. [CrossRef]
14. Fertin, G.; Labarre, A.; Rusu, I.; Vialette, S.; Tannier, E. *Combinatorics of Genome Rearrangements*; MIT Press: London, UK, 2009.
15. Kolmogorov, M.; Raney, B.; Paten, B.; Pham, S. Ragout—A Reference-Assisted Assembly Tool for Bacterial Genomes. *Bioinformatics* **2014**, *30*, i302–i309. [CrossRef] [PubMed]
16. Zhou, Y.; Yang, L.; Han, X.; Han, J.; Hu, Y.; Li, F.; Xia, H.; Peng, L.; Boschiero, C.; Rosen, B.D.; et al. Assembly of a Pangenome for Global Cattle Reveals Missing Sequences and Novel Structural Variations, Providing New Insights into Their Diversity and Evolutionary History. *Genome Res.* **2022**, *32*, 1585–1601. [CrossRef] [PubMed]
17. Leonard, A.S.; Crysnanto, D.; Fang, Z.-H.; Heaton, M.P.; Ley, B.L.V.; Herrera, C.; Bollwein, H.; Bickhart, D.M.; Kuhn, K.L.; Smith, T.P.L.; et al. Structural Variant-Based Pangenome Construction Has Low Sensitivity to Variability of Haplotype-Resolved Bovine Assemblies. *Nat. Commun.* **2022**, *13*, 3012. [CrossRef]
18. Rosen, B.D.; Bickhart, D.M.; Schnabel, R.D.; Koren, S.; Elsik, C.G.; Tseng, E.; Rowan, T.N.; Low, W.Y.; Zimin, A.; Couldrey, C.; et al. De Novo Assembly of the Cattle Reference Genome with Single-Molecule Sequencing. *Gigascience* **2020**, *9*, giaa021. [CrossRef]
19. Yalta-Macedo, C.E.; Veli, E.A.; Díaz, G.R.; Vallejo-Trujillo, A. Paternal Ancestry of Peruvian Creole Cattle Inferred from Y-Chromosome Analysis. *Livest. Sci.* **2021**, *244*, 104376. [CrossRef]
20. Arbizu, C.I.; Ferro-Mauricio, R.D.; Chávez-Galarza, J.C.; Vásquez, H.V.; Maicelo, J.L.; Poemape, C.; Gonzales, J.; Quilcate, C.; Corredor, F.-A. The Complete Mitochondrial Genome of a Neglected Breed, the Peruvian Creole Cattle (*Bos taurus*), and Its Phylogenetic Analysis. *Data* **2022**, *7*, 76. [CrossRef]
21. Instituto Nacional de Estadística e Informática IV Censo Nacional Agropecuario. 2012. Available online: http://censos.inei.gob.pe/Cenagro/redatam/# (accessed on 14 March 2022).
22. Mapiye, C.; Chikwanha, O.C.; Chimonyo, M.; Dzama, K. Strategies for Sustainable Use of Indigenous Cattle Genetic Resources in Southern Africa. *Diversity* **2019**, *11*, 214. [CrossRef]
23. Ruiz, R.E.; Saucedo-uriarte, J.A.; Portocarrero-villegas, S.M.; Quispe-ccasa, H.A.; Cayo-colca, I.S. Zoometric Characterization of Creole Cows from the Southern Amazon Region of Peru. *Diversity* **2021**, *13*, 510. [CrossRef]
24. Espinoza, R.; Urviola, G. Biometría y Constantes Clínicas Del Bovino Criollo En El Centro de Investigación y Producción Chuquibambilla de Puno (Perú). *Arch. Zootec.* **2005**, *54*, 233–236.
25. Dipas Vargas, E.S. *Zoometría e Índices Corporales Del Vacuno Criollo En El Matadero de Quicapata de La Provincia de Huamanga, A 2720 Msnm Ayacucho—2014*; Universidad Nacional San Cristóbal de Huamanga: Ayacucho, Peru, 2015.
26. Luo, R.; Liu, B.; Xie, Y.; Li, Z.; Huang, W.; Yuan, J.; He, G.; Chen, Y.; Pan, Q.; Liu, Y.; et al. SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler. *Gigascience* **2012**, *1*, 2047-217X-1-18. [CrossRef] [PubMed]
27. Zimin, A.V.; Marçais, G.; Puiu, D.; Roberts, M.; Salzberg, S.L.; Yorke, J.A. The MaSuRCA Genome Assembler. *Bioinformatics* **2013**, *29*, 2669–2677. [CrossRef] [PubMed]
28. Andrews, S. FastQC A Quality Control Tool for High Throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 5 August 2022).
29. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]

30. Krueger Trim Galore! Babraham Bioinformatics. Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 5 August 2022).
31. Marçais, G.; Kingsford, C. A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of k-Mers. *Bioinformatics* **2011**, *27*, 764–770. [CrossRef] [PubMed]
32. Vurture, G.W.; Sedlazeck, F.J.; Nattestad, M.; Underwood, C.J.; Fang, H.; Gurtowski, J.; Schatz, M.C. GenomeScope: Fast Reference-Free Genome Profiling from Short Reads. *Bioinformatics* **2017**, *33*, 2202–2204. [CrossRef]
33. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
34. Zimin, A.V.; Salzberg, S.L. The SAMBA Tool Uses Long Reads to Improve the Contiguity of Genome Assemblies. *PLoS Comput. Biol.* **2022**, *18*, e1009860. [CrossRef]
35. Langmead, B.; Salzberg, S.L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
36. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
37. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
38. Wu, T.D.; Watanabe, C.K. GMAP: A Genomic Mapping and Alignment Program for MRNA and EST Sequences. *Bioinformatics* **2005**, *21*, 1859–1875. [CrossRef] [PubMed]
39. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
40. Bao, Z.; Eddy, S. Automated de Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res.* **2002**, *12*, 1269–1276. [CrossRef] [PubMed]
41. Tarailo-Graovac, M.; Chen, N. Using Repeat Masker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* **2009**, *25*, 4–10. [CrossRef] [PubMed]
42. Jurka, J.; Kapitonov, V.V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J. Repbase Update, a Database of Eukaryotic Repetitive Elements. *Cytogenet. Genome Res.* **2005**, *110*, 462–467. [CrossRef]
43. Bedell, J.A.; Korf, I.; Gish, W. MaskerAid: A Performance Enhancement to RepeatMasker. *Bioinformatics* **2000**, *16*, 1040–1041. [CrossRef]
44. Campbell, M.S.; Holt, C.; Moore, B.; Yandell, M. Genome Annotation and Curation Using MAKER and MAKER-P. *Curr. Protoc. Bioinforma.* **2014**, *48*, 4.11.1–4.11.39. [CrossRef]
45. Korf, I. Gene Finding in Novel Genomes. *BMC Bioinform.* **2004**, *5*, 59. [CrossRef]
46. Stanke, M.; Diekhans, M.; Baertsch, R.; Haussler, D. Using Native and Syntenically Mapped CDNA Alignments to Improve de Novo Gene Finding. *Bioinformatics* **2008**, *24*, 637–644. [CrossRef]
47. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-Web: A Web Server for Microsatellite Prediction. *Bioinformatics* **2017**, *33*, 2583–2585. [CrossRef]