Data Descriptor

First Draft Genome Assembly of the Peruvian Creole Cattle Breed (*Bos taurus*) and its Comparative Genomics among the Bovinae Subfamily

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; genomica@inia.gob.pe (R.E.); ganaderia_sdiee@inia.gob.pe (F.-A.C.); deyanirafigueroa66@gmail.com (D.F.); r_cambioclimatico@inia.gob.pe (W.S.); promegnacional@inia.gob.pe (C.Q.); hvasquez@inia.gob.pe (H.V.V.); jmaicelo@untrm.edu.pe (J.L.M.)
[2] Facultad de Zootecnia, Agronegocios y Biotecnología, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Chachapoyas 01001, Peru; jmaicelo@untrm.edu.pe (J.L.M.)
[3] Laboratorio de Biología Molecular, Universidad Nacional de Frontera, Av. San Hilarión 101, Sullana 20103, Peru; jgonzales@unf.edu.pe
* Correspondence: carbizu@inia.gob.pe; Tel.: +51-986288181

Abstract: The Peruvian creole cattle (PCC) is a neglected breed, and is 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 paired-end 150 strategy on the Illumina HiSeq 2500 platform, obtaining 320 GB of sequencing data. The obtained genome size of the PCC was 2.77 Gb with a contig N50 of 108Mb and 92.59% complete BUSCOs. Also, 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. We downloaded proteomes and genomes of the Bovinae subfamily, and conducted a comparative analysis with our draft genome. Phylogenomic analysis showed that PCC is related to *Bos indicus.* Also, we identified 7,746 family genes shared among the Bovinae subfamily. This first PCC genome is expected to contribute to a better understanding of its genetics to adapt to the tough conditions of the Andean ecosystem, and evolution.

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

Keywords: NGS; Andean; neglected breed; genome

1. Summary

According to Scheu et al [1] cattle domestication started in the 9th millennium BC in Southwest Asia. Similarly, Upadhyay et al [2] referred to the genetic origin and domestication of European cattle 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]. From these, four are domesticated species, *B. mutus*, *B. javanicus*, *B. gaurus*, and *B. primigenius* which is represented by their domestic forms *B. taurus* and *B. indicus primigenius* [3]. The taxonomic status of *B. taurus* and *B. indicus* are controversial [4]. Through a mitochondrial analysis, Hiendleder et al.

 doi:10.20944/preprints202208.0349.v1

[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 the cattle has 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 other breeds.

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 necessity to implement conservation measures. Similarly, Giovambattista et al [7] reported the 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 the American Creole cattle utilizing microsatellite markers, mitochondrial DNA, and Y chromosome information. They sampled 40 creole breeds representing the whole American continent. Latin American countries contemplated, additionally to the ones already considered by Delgado et al. [6], were Bolivia, Chile, Caribe, Suriname, and Venezuela. Ginja et al. concluded that creole cattle possess 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 the 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. Also, 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 the Argentinian cattle, finding similarities in productive and reproductive aspects.

In Peru, bovine creole cattle have received little importance. Through the use of six microsatellite markers, Yalta-Macedo et al. [13] inferred the PCC ancestry and proposed that it descended from cattle from the Iberian peninsula. This study also suggested male-mediated African cattle influenced in the PCC. More recently, Arbizu et al. [14] confirmed these findings by revealing the phylogenetic relationship of the PCC with African cattle (Boran and Arsi). According to M. Rosemberg (UC del Sur, pers. comm.), PCC can be found as a cross-breed with Brown Swiss breeds around 3500 m.a.s.l., mainly in the Peruvian highlands [15]. PCC lacks of a national registration program [16]. They are phenotypically distinguishable by their smaller size when compared to other exotic breeds. Efforts to determine the full potential of muscle growing of the PCC are still limited [17–19].

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

2. Data Description

The whole genome sequencing data was deposited in the Short Read Archive (SRA) database under accession number SRR19664292, Biosample SAMN29095626, Bioproject PRJNA849594. The total of raw pair-reads were 854,737,766 sequences with mean length

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

of 150 pb, 44% GC content, and a total output of 320 GB sequencing data. After the trimming, we retained 88.2% of sequencing data, more 790 million high-quality sequence reads with approximate 281.6 Gb total sequencing data were generated. We did not detect overrepresented sequences and adapters. Also, the average quality per read was 40.

2.1. Genomic Survey

We obtained a low heterozygosity, slightly repetitive (40.22%), and the estimation of the genome (2.67 Gpb) was closed to the reported references of genomes for *B. taurus* (ARS-UCD1.3: 2.71 Gpb, ARS-LIC_NZ_Jersey: 2.64Gpb, Brown Swiss: 2.66Gpb, ARS-LIC_NZ_Holstein-Friesian_1: 2.66 Gpb, Btau_5.0.1: 2.72 Gpb). Based on the draft genome size estimated, subsequent *de novo* assembly and genome annotation were performed with the sequencing depth of approximately 47.44 X coverage (Figure 1).

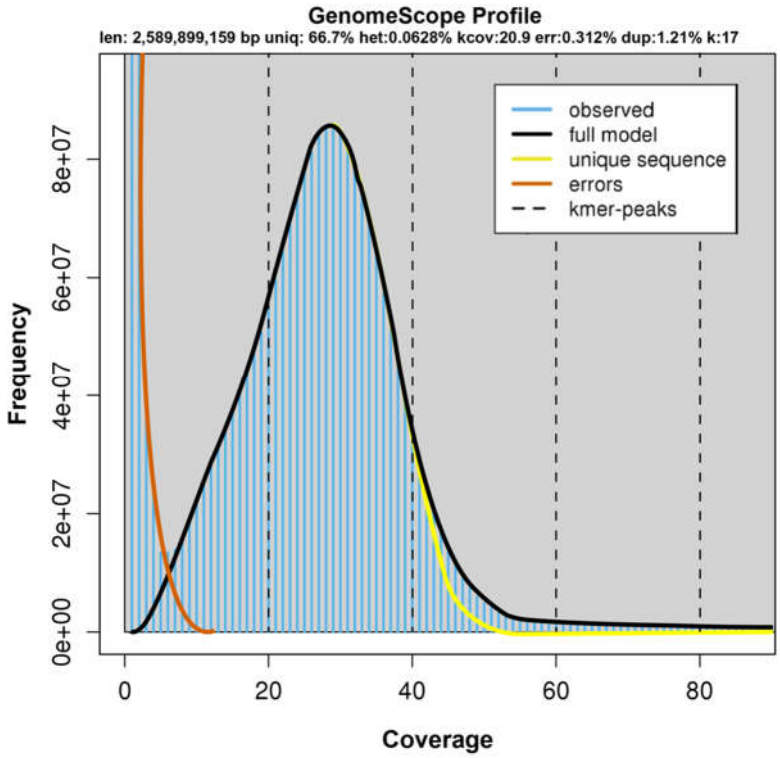

Figure 1. Distribution of K-mers in the draft genome of the Peruvian creole cattle.

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	7,242	74,944,637
L50	63,921	11
L75	133,082	19
Largest contig	109,017	164,677,788
Total length	2,679,899,159	2,771,461,908
GC (%)	41.92	41.87
# contigs (>= 1000 bp)	307,114	10,953
# contigs (>= 5000 bp)	179,627	1,848
# contigs (>= 10000 bp)	92,431	777
# contigs (>= 25000 bp)	14,279	210
# contigs (>= 50000 bp)	726	75

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

2.1. Assembly de novo and validation

We obtained different assemblies from SOAPdenovo2 [20], and Masurca [21] programs. We continued our analysis with Masurca due to a better N50 and longer contigs (Supplementary Data 1). Assembly Masurca genome was scaffolded, and we obtained a total length of 2.77 GB, which had 10,953 contigs (≥ 1,000 bp) with a GC content of 41.92 %. The longest contig was of 16,467,778 pb. 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. Also, with the scaffolding approach, we improved the N50, from 12.84 kb to 108.72 Mb (Table 1), and the number of complete mammalian single-copy BUSCOs (Benchmarking Universal Single Copy) increased from 1,620 to 3,800 complete BUSCOs (Table 2). We obtained 3,744 complete and single-copy BUSCOs (S), 56 complete and duplicated BUSCOs (D), 165 fragmented BUSCOs (F), and 139 missing BUSCOs (M). In comparison with other bovine species with (i) scaffold level assembly such as *B. mutus* (N50 of 1.4 Mb) [22], *B. bison* (N50 of 7.19 Mb) [23], *B. frontalis* (1 Mb) [24], and (ii) chromosome level assembly such as *B. indicus* and *Bubalus bubalis* with an N50 of 106.3 Mb [25] and 111 Mb [26], respectively, our assembly has an N50 of 108.72 Mb, showing good quality. Additionally, our assembly has 92.59% complete BUSCOs (C) (S: 91.23% + D: 1.36%), similar to *B. indicus* with 92.9 % (C) (S: 91.9% + D: 1%) [25], but is far from *B mutus* with 97% (C) (S: 96.5% + D: 0.5%) [22], *B. bubalis* with 98.4% (C) (S: 97.0% + D: 1.4%) [26] and *B. bison* with 95.4% (C) (S: 64.4% + D: 31.1%) [23] (Figure 2).

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

Terms	Contigs	Scaffold
Complete BUSCOs (C)	1,620	3,800
Complete and single-copy BUSCOs (S)	1,580	3,744
Complete and duplicated BUSCOs (D)	40	56
Fragmented BUSCOs (F)	1,573	165
Missing BUSCOs (M)	911	139

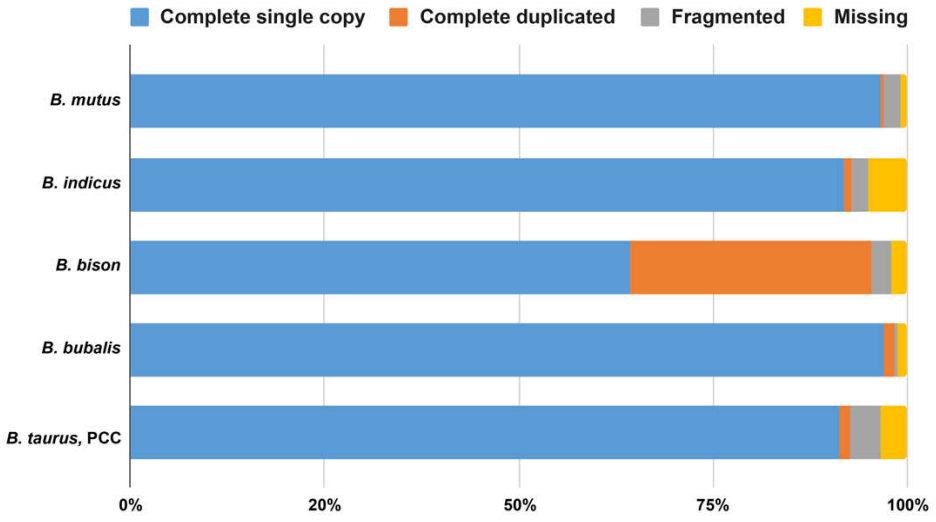

Figure 2. Comparison of BUSCO analysis of the Peruvian creole cattle (PCC) with other Bovinae species.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

2.2. Repeat Annotation

We identified 897.59 Mb of repeated sequences, which represents 32.39% of the assembled PCC. This is less than the repeated sequences of *B. frontalis* (43.66%) [24], *B. grunniens* (43.9%) [27], *Syncerus caffer* (37.21%) [28] and *Bison bonasus* (47.03%) [29]. Similar to the *B. grunniens* assembly [27], LINE elements represented the majority of identified repeats in our assembly with 28.55%, but higher than *S. caffer* (25.57%) [28] and lower than *B. bonasus* (39.84%) [29].

Table 3. Summary of the repetitive DNA of the 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	0.0104%
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	245,870	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%
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 %
Low Complexity	499,594	3,869,690 bp	0.14%

Table 4. Summary of SSR distribution in the Peruvian Creole (PCC) cattle genome and its comparison to other Bovinae species.

Parameter	*B. taurus*, PCC	*B. frontalis*	*B. indicus*	*B. mutus*	*B. bison*
Total size of examined sequences (bp)	2,771,480,930	3,002,445,034	2,673,949,103	2,832,776,395	2,828,031,685
Total number of identified SSRs	990,823	1,114,686	899,003	1,039,426	978,892
Frequency (SSR/Mb)	371,09	371,31	336,2	392,98	346,14
Number of SSRs present in compound formation	89,588	109,541	82,308	109,567	91,260

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

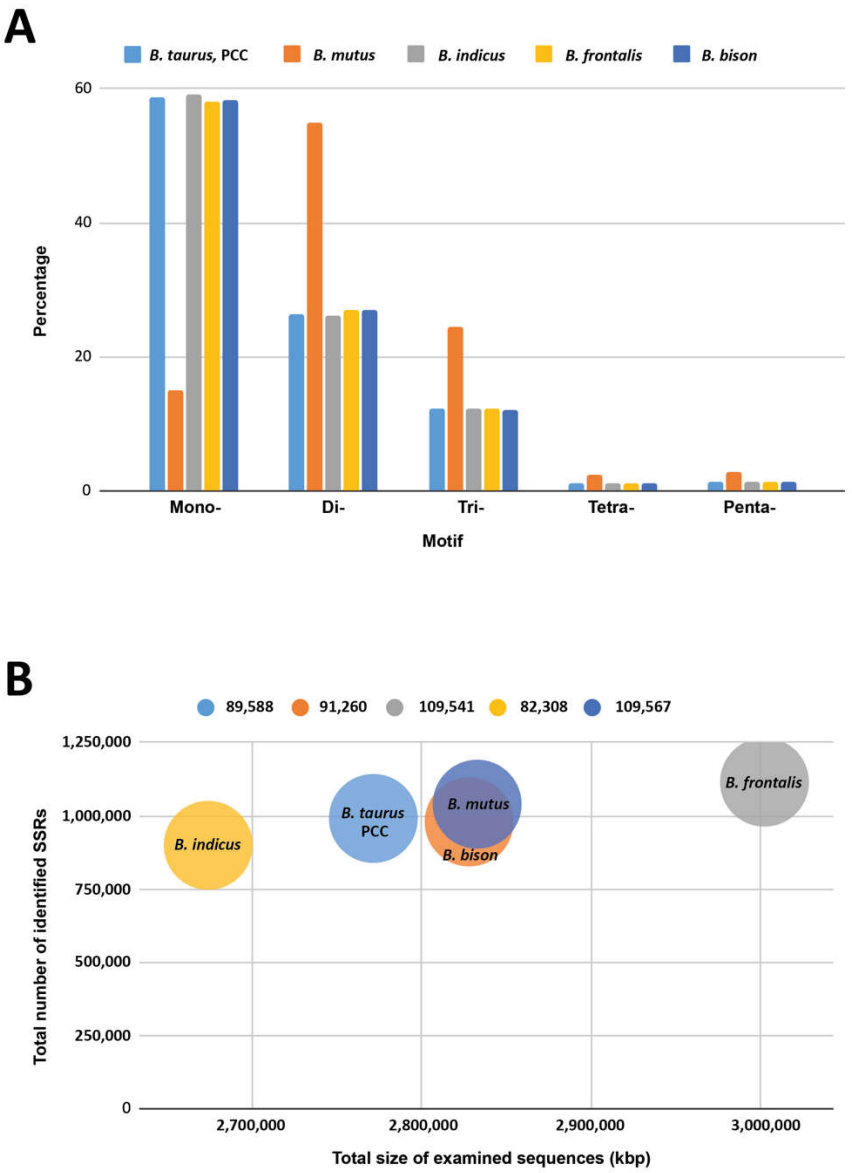

Figure 3. Frequencies of SRRs in subfamily Bovinae. (A) SSR motif percentage of the Peruvian Creole (PCC) with other Bovinae species. (B) Bubbles that represent the total number of SSRs, the total size of the examined sequences (Kbp), and the number of SSRs present in the formation of the compound.

The most abundant microsatellite motif type of PCC that we identified were mononucleotide repeats accounting for 58.7% (582,189) of the total SSRs, followed by dinucleotide repeats 26.4% (261,381), trinucleotide repeats 12.2% (121,213), pentanucleotide repeats 1.4% (14,057), tetranucleotide repeats 1.2% (11,639), and finally hexanucleotide repeats 0.035% (344). This is similar to the microsatellite motif distribution of other bovinae species such as *B. bison*, *B. frontalis* and *B. indicus*, but differs with *B. mutus* in which the most abundant microsatellite motif type were dinucleotide repeats (54.97%), followed by trinucleotide repeats (24.48%) and then mononucleotide repeats (15.02%) (Figure 3a). A total of 990,823 microsatellite loci were identified based on the assembled PCC draft genome sequence, with a frequency of 371.09 SSR/Mb, which is similar to *B. frontalis* (371.31 SSR/Mb), lower than *B. mutus* (392.98 SSR/Mb) but higher than *B. bison* (346.14 SSR/Mb) and *B. indicus* (336.20 SSR/Mb) (Table 4). Also, the number of SSRs present in compound formation of PCC (89,588) was very similar to the other *Bovinae* species studied here. It is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

also emphasized that the Bovinae species' genome sizes and the overall number of discovered SSRs are quite similar (Figure 3b).

2.3. Genomic Annotation and Genomic Comparative

With gene-prediction methods (ab initio prediction, homology-based searching) to annotate protein-coding genes in the draft genome, we obtained 19,803 annotated protein-coding genes. For genomic comparative, we analyzed a total of 489,024 genes from *Bovinae* species, of which 473,948 genes were clustered in orthogroups. Also, we identified 28,946 orthogroups, of which 5,250 were species-specific orthogroups. A total of 27,391 genes were present in species-specific orthogroups (Supplementary Data2). Also, phylogenomic analysis based on a single alignment of single-copy orthologous showed that, as expected, the PCC evolved closely with *B. indicus*, and both share a common ancestor with *B. mutus* and *B. bison*. Our result is in agreement with previous work [14,29,30] (Figure 4.b). Bootstrap values of 100% support the clades in the phylogenomic analysis.

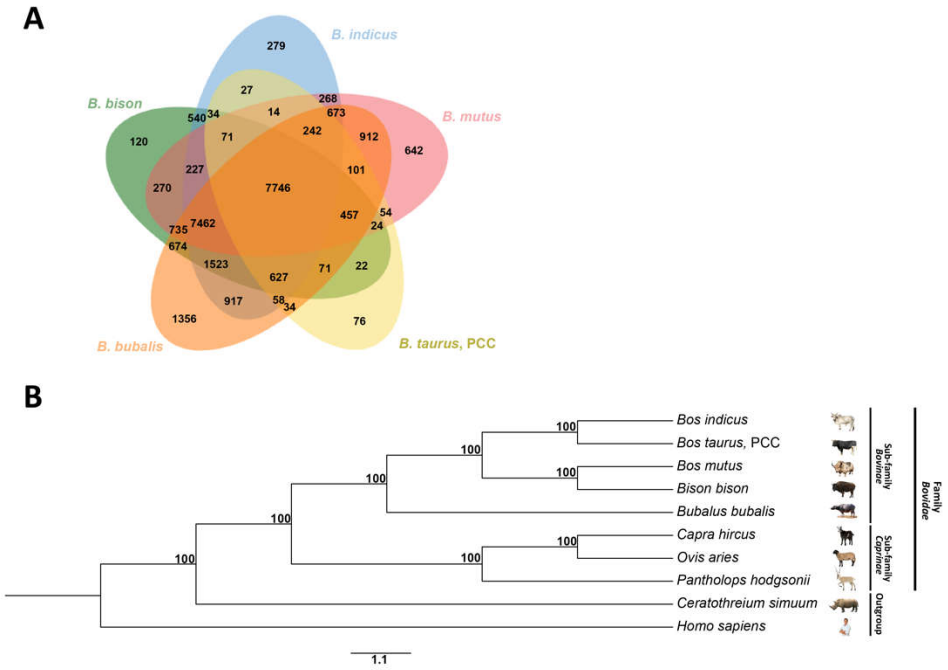

Figure 4. (A) Venn diagram of shared and unique gene families in subfamily Bovinae (*B. taurus* PPC, *B. indicus*, *B.*, *Bubalis bubalis*, and *B. bison*). **(B)** Phylogenomic tree and of orthologous gene clusters of Bovinae subfamily (*B. frontalis*, *B. taurus* PPC, *B. indicus*, *B. mutus*, *B. frontalis*, *Bubalus bubalis*, and *B. bison*) compared to five mammals (*Ovis aries*, *Capra hircus*, *Pantholops hodgsonii*, *Ceratotherium simum simum*, and *Homo sapiens*).

Venn diagram shows that the PCC, *B. indicus*, *B. mutus*, *B. bubalis*, and *B. bison* contain a core set of 7,746 gene families in common. Also, there were 76 gene families containing 515 genes specific to the PCC. Besides, a similar number of gene families specific of Bovinae species (*B. bison*: 129, *B.indicus* 279, *B.mutus*; 642) were identified. *Bubalus bubalis* obtained a greater amount of family of specific genes (1,356) (Figure 4b).

Based on gene ontology analysis, the biological processes of 7,746 shared gene family clusters identified were classified as physiological processes (24%), and cellular processes (15%). Ion binding (25%) and molecular transducer activity (11%) made up the majority of the cellular component organization's involvement in molecular functions. Moreover, the cellular component contained a number of groups having subcellular entities (23%) and membrane (22%) (Figure 5).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

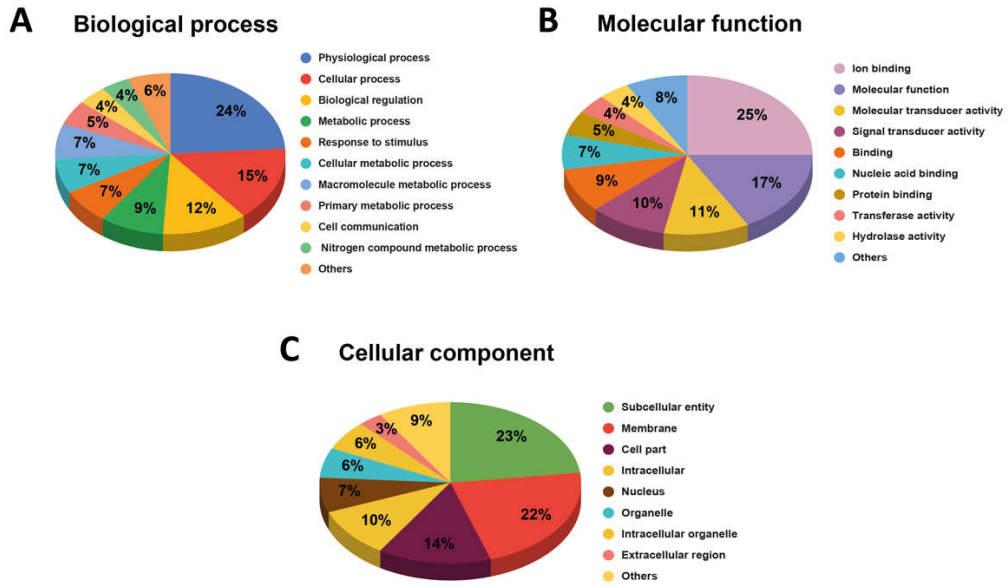

Figure 5. Gene ontology of shared gene family clusters among Bovinae species.

In summary, we report the first draft genome of the PCC. Since there are limited genomic sequence resources for Peruvian cattle, our study expects to provide a reference for animal breeding programs of this important livestock resource of the Peruvian Andean region.

3. Methods

3.1. Sample Collection and DNA Extraction

We collected hair sample from the tail from a single male specimen from Andagua, Arequipa (3574 masl; −15.499548°, −72.359927°). Since this individual possessed most of the classical characteristics of a Peruvian creole cattle, we decided to select if 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. Mitochondrial genome of this individual was recently sequenced [14].

3.2. Genome Sequencing and Genomic survey.

In this study, pair-end DNA sequencing was carried out using Illumina HiSeq 2500 platform. The raw reads were checked by FastQC v.0.11.9 [31]. Also, trimming quality (phred Q > 25) and remotion of adapters were conducted with Trimmomatic v0.36 [32] and TrimGalore v.0.6.7 [33] softwares, respectively. For the genomic survey, we used Jellyfish v.2.0 [33]. Genome Scope v1.0.0 (Cold Spring Harbor Laboratory, Laurel Hollow, US) [34] 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 [20], and Masurca v.4.0.6 [21]. Next, we used QUAST v.5.2.0 [35] for statistics of assemblies. Masurca resulted in improved assembly statistics and was subjected to Samba scaffolder v.1.0 [21] for scaffolding and gap-filling. For the reference-based scaffolding, we used a reference genome of *B.taurus* (Genbank: GCA_002263795.3). Next, we used QUAST with the output of the scaffolding. Validation of assembly was assessed using

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 August 2022 doi:10.20944/preprints202208.0349.v1

three different approaches: (i) filtered PE Illumina reads were remapped to detect errors in the assembly using Bowtie2 v.2.4.2 [36] and SamTools v.1.7 [37] softwares, (ii) the BUSCO v.5.4.2 [38] 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 (4,104 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 [39]. We used jcvi vecscreen (https://github.com/tanghaibao/jcvi), which use Univecdatabase (https://ftp.ncbi.nlm.nih.gov/pub/UniVec/) for detection of vectors, and mapped the scaffolds against to nt/nr NCBI database (https://www.ncbi.nlm.nih.gov/) using BLAST v.2.2.26 [40] 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 JANIWY000000000.

3.4. Repeat Annotation

To identify repetitive elements, we used *de novo* and homolog-based methods. For the *de novo* approach, we used Repeatmodeler v.1.08 [41] to generate a *de novo* PCC repeat library, which is subsequently used in RepeatMasker v4.0.7 [42] to annotate repeats. For the homology-based approaches, we used Repbase v4.0.7 [43], RepeatMasker and RMBlast v2.2.27 [42]. All repeat results were merged. Final genome assembly was repeat-masked using the library repeats using RepeatMasker. The SSRs were identified in the PCC genome using MISA perl script (http://pgrc.ipk-gatersleben.de/misa/) [44] 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), and hexamer (six nucleotides, n > 3). Also, for SSR analysis, we added the genomes of *B. bison* (GenBank: GCA_000754665.1), *B. mutus* (GenBank: GCA_000298355.1), *B. indicus* (GenBank: GCA_000247795.1) and *B. frontalis* (GenBank: GCA_007844835.1). We used BUSCO to examine the quality of assemblies. Afterwards, we used the MISA script with the same parameters for PCC.

3.5. Genome annotation

MAKER v.3.01.03 [45] was run on the repeat-masked genome with SNAP v.2006-07-28 [46] and AUGUSTUS v.2.7 [47] programs. For evidence to guide the annotation process. We retrieved ESTs of *B. taurus* from NCBI database (ftp://ftp.ncbi.nih.gov/repository/dbEST/) and proteomes of Bovidae species: *B. taurus* (GenBank: GCF_002263795.2), *B. indicus* (GenBank: GCF_000247795.1) and *B. mutus* (GenBank: GCF_000298355.1). As MAKER was run iteratively for two times, the predictions were curated against a high-quality protein database of UNIPROT (https://www.uniprot.org/) using BLAST with E-value of 1e-05.

3.6. Comparative genomics

For comparative genomics, we retrieved, from the NCBI database, proteomes of eight species: *B. mutus*, *B. indicus*, *B. bison* (GenBank: GCF_000754665.1), *Bubalus bubalis* (GenBank: GCF_019923935.1), *Ovis aries* (GenBank: GCF_016772045.1), and *Capra hircus* (GenBank: GCF_001704415.2), and as outgroup we added *Homo sapiens* (Refseq: GCF_000001405.40) and *Ceratothreium simuum* (Refseq: GCF_000283155.1). To exclude putative fragmented genes, genes encoding protein sequences shorter than 30 amino acids were filtered out. With these data set, we clustered orthologous groups with OrthoFinder v2.5.2 [48] with default parameters. Single-copy orthologous sequences present in all species were extracted and individually aligned with MAFFT v7.273 [49]. We curated the alignments with Trimal v.3.1.1 [50]. Next, we concatenated the alignments. The maximum likelihood phylogenomic tree was calculated based on a single alignment of single-copy orthologous genes by RAxML v8.2.12 [51] with PROTGAMMAWAG as model of amino

acid change and 1,000 rapid bootstraps for robustness. Also, comparative analysis of the organization of orthologous gene clusters was carried out using genes of subfamily Bovinae: Our assembly of *B. taurus*, *B. mutus*, *B. indicus*, *B. bubalis* and *B. bison* through OrthoVenn2 web tool [52] with E-value of 0.01and inflation value of 1.5. Briefly, to identify orthologous groups, OrthoVenn2 employs the OrthoMCL v.2.0.9 [53] clustering algorithm to annotate and compare ortholog groups. An overall performance is made by the OrthoMCL. By using the InParanoid and the Markov Clustering algorithm [54], the DIAMOND v0.9.24 alignment detects potential orthology and InParalogy relationships [55] and produces disjoint groupings of highly related proteins. The Gene Ontology (GO) terms for biological process, molecular function, and cellular component categories were assigned to the corresponding orthologous cluster (shared among subfamily Bovinae) by identifying similarity to sequences in the Uniprot database.

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*, 1–11, doi:10.1186/S12863-015-0203-2.
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. *Hered. 2017 1182* **2016**, *118*, 169–176, doi:10.1038/hdy.2016.79.
3. Garrick, D.J.; Ruvinsky, A. *The Genetics of Cattle*; CABI, Ed.; 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, doi:10.1159/000118756.
5. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, Weinstock GM, Adelson DL, Eichler EE, Elnitski L, Guigó R, Hamernik DL, Kappes SM, Lewin HA, Lynn DJ, Nicholas FW, Reymond A, Rijnkels M, Skow LC, Zdobno, Z.F. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* **2009**, *324*, 522–528, doi:10.1126/science.1169588.
6. Delgado, J.V.; Martínez, A.M.; Acosta, A.; Lvarez, L.A.A.´; Armstrong, E.; Camacho, E.; Cañ, J.; Corté, O.; Dunner, S.; Landi, V.; et al. Genetic Characterization of Latin-American Creole Cattle Using Microsatellite Markers. *Anim. Genet.* **2012**, *43*, 2–10, doi:10.1111/j.1365-2052.2011.02207.x.
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, doi:10.1016/j.gene.2013.01.002.
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. Reports 2019 91* **2019**, *9*, 1–16, doi:10.1038/s41598-019-47636-0.
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, doi:10.53588/alpa.293402.
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, doi:10.1080/23802359.2019.1704639.
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, doi:10.1017/S2078633613000313.
12. Aracena, M.; Mujica, F. Caracterización Del Bovino Criollo Patagónico Chileno: Un Estudio de Caso. *Agro Sur* **2011**, *39*, 106–115, doi:10.4206/AGROSUR.2011.V39N2-05.
13. 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, doi:10.1093/BIOINFORMATICS/BTV351.
14. Mukherjee, S.; Cai, Z.; Mukherjee, A.; Longkumer, I.; Mech, M.; Vupru, K.; Khate, K.; Rajkhowa, C.; Mitra, A.; Guldbrandtsen, B.; et al. Whole Genome Sequence and de Novo Assembly Revealed Genomic Architecture of Indian Mithun (*Bos Frontalis*). *BMC Genomics* **2019**, *20*, 1–12, doi:10.1186/S12864-019-5980-Y.
15. Zhang, S.; Liu, W.; Liu, X.; Du, X.; Zhang, K.; Zhang, Y.; Song, Y.; Zi, Y.; Qiu, Q.; Lenstra, J.A.; et al. Structural Variants Selected during Yak Domestication Inferred from Long-Read Whole-Genome Sequencing. *Mol. Biol. Evol.* **2021**, *38*, 3676–3680, doi:10.1093/MOLBEV/MSAB134.
16. Glanzmann, B.; Möller, M.; le Roex, N.; Tromp, G.; Hoal, E.G.; van Helden, P.D. The Complete Genome Sequence of the

African Buffalo (Syncerus Caffer). *BMC Genomics* **2016**, *17*, 1–7, doi:10.1186/S12864-016-3364-0/FIGURES/1.
17. Wang, K.; Wang, L.; Lenstra, J.A.; Jian, J.; Yang, Y.; Hu, Q.; Lai, D.; Qiu, Q.; Ma, T.; Du, Z.; et al. The Genome Sequence of the Wisent (*Bison Bonasus*). *Gigascience* **2017**, *6*, 1–5, doi:10.1093/GIGASCIENCE/GIX016.
18. 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).
19. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120, doi:10.1093/BIOINFORMATICS/BTU170.
20. Krueger Trim Galore! Babraham Bioinformatics Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 5 August 2022).
21. 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, doi:10.1093/BIOINFORMATICS/BTX153.
22. 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*, doi:10.1186/2047-217X-1-18/2656146.
23. 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, doi:10.1371/JOURNAL.PCBI.1009860.
24. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* **2013**, *29*, 1072–1075, doi:10.1093/BIOINFORMATICS/BTT086.
25. Langmead, B.; Salzberg, S.L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359, doi:10.1038/nmeth.1923.
26. 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, doi:10.1093/BIOINFORMATICS/BTP352.
27. Wu, T.D.; Watanabe, C.K. GMAP: A Genomic Mapping and Alignment Program for MRNA and EST Sequences. *Bioinformatics* **2005**, *21*, 1859–1875, doi:10.1093/BIOINFORMATICS/BTI310.
28. 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, doi:10.1016/S0022-2836(05)80360-2.
29. Abrusán, G.; Grundmann, N.; Demester, L.; Makalowski, W. TEclass—a Tool for Automated Classification of Unknown Eukaryotic Transposable Elements. *Bioinformatics* **2009**, *25*, 1329–1330, doi:10.1093/BIOINFORMATICS/BTP084.
30. Bedell, J.A.; Korf, I.; Gish, W. MaskerAid: A Performance Enhancement to RepeatMasker. *Bioinformatics* **2000**, *16*, 1040–1041, doi:10.1093/BIOINFORMATICS/16.11.1040.
31. 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, doi:10.1159/000084979.
32. Beier, S.; Thiel, T.; Münch, T.; Scholz, U.; Mascher, M. MISA-Web: A Web Server for Microsatellite Prediction. *Bioinformatics* **2017**, *33*, 2583–2585, doi:10.1093/BIOINFORMATICS/BTX198.
33. 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, doi:10.1002/0471250953.BI0411S48.
34. Korf, I. Gene Finding in Novel Genomes. *BMC Bioinformatics* **2004**, *5*, 1–9, doi:10.1186/1471-2105-5-59/TABLES/4.
35. 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.
36. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* **2019**, *20*, 1–14, doi:10.1186/S13059-019-1832-Y/FIGURES/5.
37. Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.
38. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* **2009**, *25*, 1972–1973, doi:10.1093/BIOINFORMATICS/BTP348.
39. Li, L., Stoeckert, C. J., & Roos, D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* **2003**, *13*, 2178–2189, doi:10.1101/GR.1224503.
40. Östlund, G.; Schmitt, T.; Forslund, K.; Köstler, T.; Messina, D.N.; Roopra, S.; Frings, O.; Sonnhammer, E.L.L. InParanoid 7: New Algorithms and Tools for Eukaryotic Orthology Analysis. *Nucleic Acids Res.* **2010**, *38*, D196–D203, doi:10.1093/NAR/GKP931.
41. Van Dongen, S.M. Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, 2000.