final report

Project code: P.PSH.0868

Prepared by: Elizabeth Ross
Queensland Alliance for Agriculture and Food Innovation

Date published: 2nd September 2019

PUBLISHED BY
Meat and Livestock Australia Limited
Locked Bag 1961
NORTH SYDNEY NSW 2059

Characterisation of the Brahman Genome

This is an MLA Donor Company funded project.

Meat & Livestock Australia acknowledges the matching funds provided by the Australian Government to support the research and development detailed in this publication.

Executive summary

Brahman are a *Bos indicus* breed that are used extensively for meat production in northern Australia. Previously, genomic work on Brahman cattle has used the *Bos taurus* genome assembly as a reference for tasks such as genomic prediction and genome wide association studies. Recent advances in sequencing has seen achievable DNA sequencing read lengths on high throughput machines dramatically increase from 300bp up to 300,000bp, with theoretical maximums over 1 million base pairs in length. This presents a new opportunity to develop breed specific reference assemblies that are more appropriate for use in northern Australian cattle research.

Here we have assembled a high quality genome of a Brahman cow. The reference was assembled using 195GB of long read data, 43GB of Chicago short read data, 41GB of Hi-C short read data, and 180GB of short read Illumina whole genome sequence data. Contigs were built using long reads, then scaffolding was performed using the Chicago and Hi-C data. Finally the genome was polished with the short read whole genome sequence and gaps filled with both long read and short read sequences. The final assembly is highly contiguous with a total size of 2.67GB and all chromosomes in 1 to 3 scaffolds, with only 403 gaps in the total assembly. Telomere sequence was identified in high numbers at the ends of eight of the 30 chromosomes, suggesting that these chromosomes have been sequenced and assembled to their very ends.

When data was mapped back to the assembled genome both the autosomes and the X chromosome showed even coverage throughout. Three autosomes displayed slightly elevated coverage and upon investigation it was revealed that this was caused by the presence of a large satellite (1.711a), suggesting that the read length was not long enough to fully identify and characterise the location and number of satellite 1.711a in the genome. There did not appear to be inflated depths across any of the other repetitive elements in the genome indicating that most other highly repetitive elements have been accurately mapped and characterised.

Structural comparison to the *B. taurus* genome revealed that while the autosomes show little large scale variation, there are six putative large structural differences on the X chromosome. These structural differences will cause errors in imputation and difficulty interpreting genome wide association studies on the X chromosome if an inappropriate genome is utilised. The availability of a Brahman reference genome brings the northern beef industry in line with its southern counterpart in terms of available genomic resources.

Additional sequencing of over 200 Brahman, Droughtmaster and Santa Gertrudis animals reveals a rich abundance of SNP distributed throughout the northern beef cattle genome. Analysis of long read data also found 27,482 putative structural variants that are greater than 50bp in size, with the largest approaching 80kb located on the X chromosome. Furthermore, regions of taurine introgression have been identified on chromosomes 8, 12, 14, 23, 26 and 29; the rest of the genome is mostly *B. indicus* origin. The introgression analysis revealed that protein biosynthesis genes are significantly enriched in regions of *B. indicus* origin. Additionally evidence that the Celtic poll allele is segregating in the Australian Brahman population was apparent from the sequencing of four polled bulls.

A high quality genome reference assembly built using long read data, and information from the whole genome sequencing of Northern Beef cattle is now available for use in the Australian Beef research community. This project removes the hurdles faced by researchers when they do not have a reliable reference assembly that is relevant to the animals farmed in Northern Australia.

Table of contents

1 Background

1.1 Brahman Cattle

Domesticated cattle are descendants of the European auroch and are divided into two subspecies: *Bos taurus taurus* and *Bos taurus indicus,* commonly known as *Bos taurus* and *Bos indicus*. The *B. taurus* subspecies has thrived throughout Europe and encompasses breeds such as Holstein and Hereford (Legge 1991; O'Neill *et al.* 2010; Utsunomiya *et al.* 2017). *B. indicus* or Zebu cattle originated in modern day Pakistan and evolved to suit warm tropical climates such as India, China and northern Australia (Hansen 2004; Villalobos-Cortes *et al.* 2015). *B. indicus* are characterized by a hump over their withers and includes the Nelore and Brahman breeds (Hansen 2004).

The Australian beef cattle herd, worth $16.85 billion AUD, is often broken up into a northern and southern herd (Mclean *et al.* 2014; Meat & Livestock Australia 2017; Greenwood *et al.* 2018). The southern herd is largely high yielding *B. taurus* animals on highly productive pastures (Meat & Livestock Australia 2017). In contrast, the northern herd is predominantly *B. indicus* cattle in warmer areas and on pastures that have lower nutritional value (Meat & Livestock Australia 2017). Despite the northern herd producing 58% of the countries beef (Meat & Livestock Australia 2017), the two remain equal in economic value (Meat & Livestock Australia 2017). Australia is the third largest global exporter of beef, exporting just under 1.5 million tonnes carcass weight equivalent in 2018 (Meat & Livestock Australia 2018).

1.2 The history of DNA sequencing

Sequencing technology can be broken into three distinct generations (Leggett and Clark 2017). First generation technology is characterized by the use of chain termination (Sanger *et al.* 1965; Brownlee and Sanger 1967; Dube *et al.* 1968; Min Jou *et al.* 1972), second generation sequencers by high throughput short reads and third generation sequencers by high throughput long reads (Pareek *et al.* 2011; Heather and Chain 2016). Long reads are reads greater than 1000bp and are valuable in the detection of structural variants (SV), including copy number varients (CNVs) (Pollard *et al.* 2018) as well as genome assembly.

Second generation sequencers were released in 2005 (Margulies *et al.* 2005; Shendure *et al.* 2005) and were characterized by parallel reactions that drastically increased throughput compared to first generation "Sanger" sequencing (Heather and Chain 2016). The throughput of parallel sequencers was a huge improvement on Sanger sequencing (Mardis 2008; Khalifa *et al.* 2016). Illumina has maintained the largest market share in the second generation market to date, through their unrivalled throughput: 1.6-1.8 Tb/day (Table 1; Pillai *et al.* 2017). The high throughput and low cost per base of second generation sequencing technology has made it an invaluable tool for SNP identification and quantitative studies of the genome.

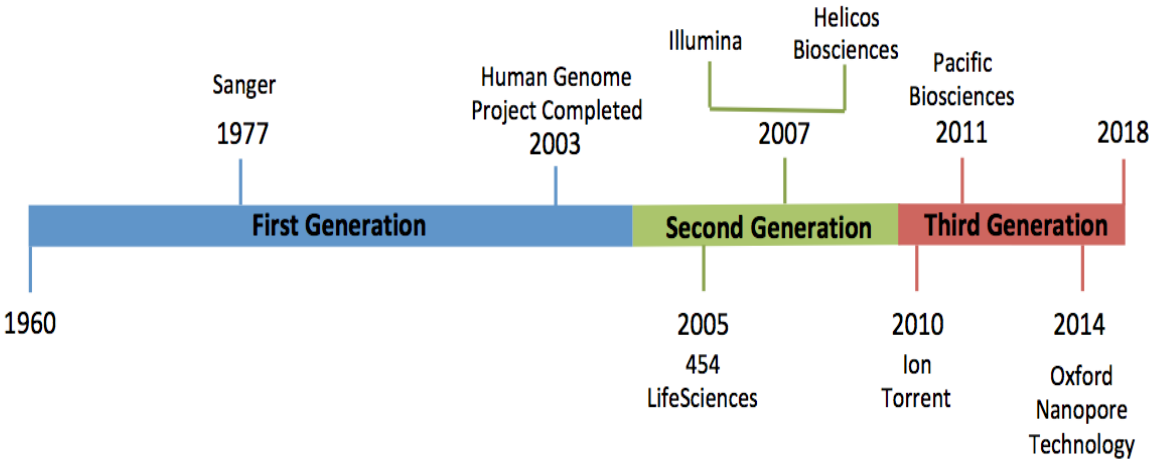


Fig. 1. A timeline fo 1960 through to 2018 illistrating the major commercial releases of genomic sequencing technology (Holt and Jones 2008; Mardis 2008; Morozova and Marra 2008; Schuster 2008; Stangier 2009; Pareek et al. 2011; Liu et al. 2012; van Dijk et al. 2014).

Second-generation sequencer's reliance on DNA amplification steps, such as polymerase chain reaction (PCR), is a major drawback to the technology (Liu *et al.* 2012; Bahassi and Stambrook 2014; Heather and Chain 2016) as it introduces nucleotide biases, a single nucleotide polymorphism (SNP) error profile and alters the relative abundance of DNA templates (Acinas *et al.* 2005; Aird *et al.* 2011; Pareek *et al.* 2011; Kebschull and Zador 2015). Template strands with extreme G + C compositions are often under represented or absent (Aird *et al.* 2011). Third generation sequencers eliminate the amplification step by using single molecule sequencing (Heather and Chain 2016).

1.3 The bovine genome

Until recently the development of a reference genome took years of work by whole teams. This meant that for many species a reference genome was not available. In the case of cattle, the only available reference genome was derived from the DNA of a *B. taurus* animal (Elsik *et al.* 2009). While the taurine genome was an incredible useful tool, it is now possible to make a breed specific genome for Brahman cattle. This genetic differences between taurine and indicine cattle (including Brahman) are well known (2009). Having a relevant reference genome assembly will allow the Northern beef industry (which extensively uses Brahman cattle and crossbreeds) to leverage the knowledge and methods of genomic technologies, including genomic predictions, genome wide association studies, imputation of SNP data and gene expression analysis to achieve practical outcomes for the industry, such as increased accuracies in genomic trait prediction and identification of important genetic loci.

2 Project objectives

A high definition and accurate assembly of the Australian Brahman breed including all the variation within and between Brahman and *B. taurus* breeds. This will include short range variation such as SNPs, insertions and deletions and critically the long range structural variation between the two cattle subspecies.

- Input to content of future SNP chip or Genotype by Sequence technologies to improve prediction accuracies and reduce assay costs.

- Enhanced identification of causal or SNPs closely linked to causal mutations in the Australian Brahman breed enabling higher prediction accuracy in Genomic Selection.

- Ability to identify *B. taurus* and *B. indicus* genome content in cross bred or composite cattle enabling the tracking of key genomic elements responsible for traits such as tropical adaptation, fertility, meat quality etc.

- Training of the next generation of researchers and professionals that will have the capacity to deliver large research projects that impact on productivity of beef production in Northern Australia.

3 Methodology

3.1 Animal selection

3.1.1 Reference animal

To ensure sequencing depth of the X chromosome the reference sequence was based on a female Brahman (**Error! Reference source not found.**). The animal was selected in consultation with the Australian Brahman Breeders Association based on pedigree information.

3.1.2 Bull sequence

In addition to the reference animal, DNA from other animals that are important for the north AustralianBeef herd were sequenced using short read technology. Semen straws were sourced from historically and currently influential bulls.

While sequencing costs have decreased dramatically in recent years, it is still imperative to obtain the most benefit possible from the animals that are sequenced. To this end, the expertise of Dr. Brian Burns was used to identify animals which had the most influence on the population's genetic variation, whilst also balancing relationships with already sequenced animals. It is not favourable to sequence several closely related animals as the genetic information being captured is shared between related individuals. Prof. Hayes used SNP chip data to identify animals that provided the most valuable contribution to the current dataset. A ranked list of bulls was supplied for cross reference with the available DNA samples. DNA was obtained from Dr. Russel Lyons. Sample volumes were estimated and concentrations of DNA within each of the samples was measured using a Qubit, a highly sensitive DNA quantification device that specifically measured double stranded

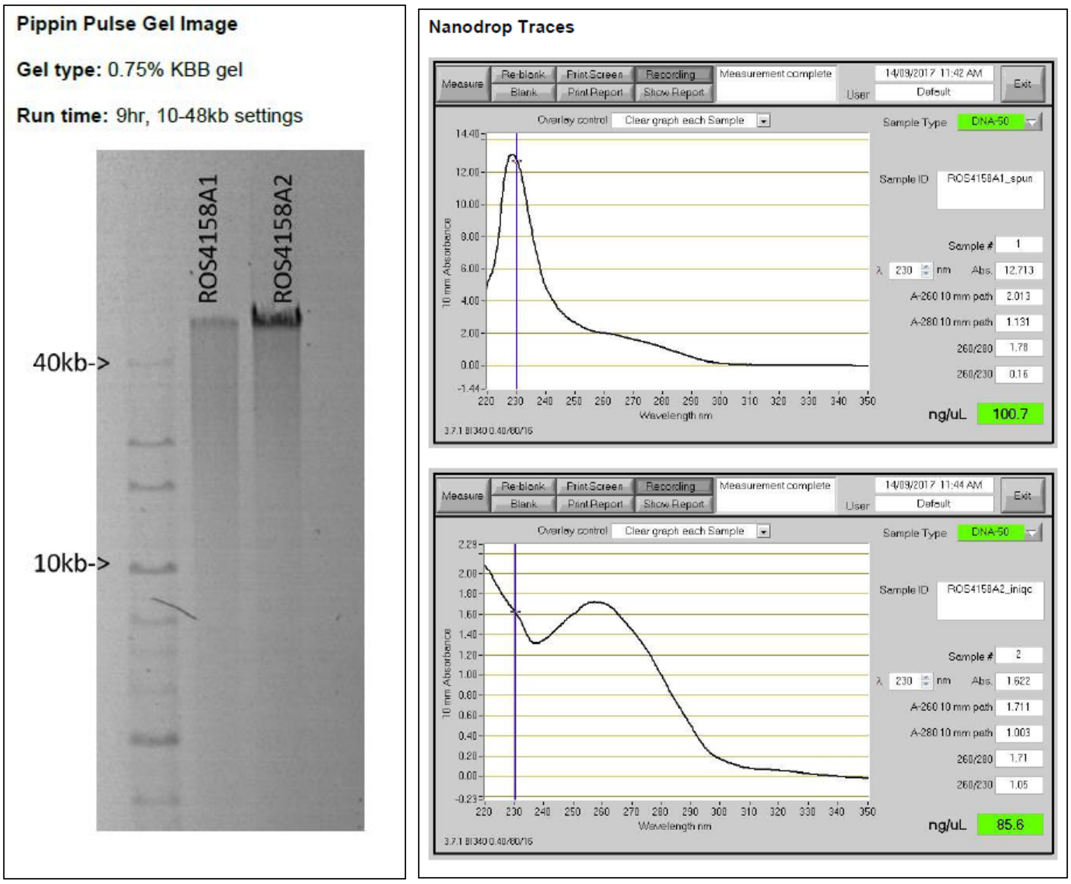
DNA, ensuring that concentrations are not falsely overestimated due to free nucleotides or single stranded DNA and RNA. The volume and concentration data was combined to obtain a total quantity estimate for each sample. Samples with sufficient DNA for sequencing were diluted to a standard concentration.

Fig. 2 Elrose Neomi 3492 was the animal selected as the source of DNA for the reference assembly

3.2 Data generation

3.2.1 Long read data

Long read data was generated on the sequel (Pacific Biosciences; PacBio) sequencing machine, due to its relatively long read length and random error profile. DNA was extracted from the kidney of the reference animal (Fig. 3) and sent for sequencing at the Ramaciotti Centre for Genomics (Sydney, Australia). The prepared library was sequenced using Single Molecule Real-Time (SMRT) sequencing on 28 SMRTcells across 5 separate runs using MagBead loading with a target insert of 20kb.



Pippin Pulse Gel Image

Gel type: 0.75% KBB gel

Run time: 9hr, 10-48kb settings

Nanodrop Traces

Measurement complete
14/09/2017 11:42 AM
User Default
Overlay control: Clear graph each Sample
Sample Type: DNA-50
Sample ID: ROS4158A1_spun
Sample #: 1
λ 230 nm Abs. 12.713
A-260 10 mm path: 2.013
A-280 10 mm path: 1.131
260/280: 1.78
260/230: 0.16
ng/uL 100.7
3.7.1 BI340 0.40/80/16

Measurement complete
14/09/2017 11:44 AM
User Default
Overlay control: Clear graph each Sample
Sample Type: DNA-50
Sample ID: ROS4158A2_iniqc
Sample #: 2
λ 230 nm Abs. 1.622
A-260 10 mm path: 1.711
A-280 10 mm path: 1.003
260/280: 1.71
260/230: 1.05
ng/uL 85.6
3.7.1 BI340 0.40/80/16

Fig. 3 DNA quality control metrics of the DNA samples extracted from the reference animal Elrose Neomi. A) A pulse field electrophoresis gel image showing the size range of DNA extracted from kidney and blood. B) A spectrophotometer trace for the DNA showing purity metrics.

3.2.2 Short read data

DNA was extracted from the semen of the bulls using a modified PureGene (Qiagen) protocol. The DNA was stored at 4°C before being sent for sequencing at the Ramaciotti Centre for Genomics.

A Novaseq run, the largest capacity sequencer currently available, was utilised to obtain the sequences of the final bulls. A single run yielded 3.4Tb of data, which is over 10% more than the advertised output of the machine. For each animal, pair end data was obtained, with each end consisting of 150bp. The quality of the run was high, with scores above 30, which are considered to be extremely high quality data. For comparison, the 1000 Bull Genomes Project (Daetwyler *et al.* 2014) requires data to have an average quality of 20. All of the Brahman sequences are substantially greater than this cut-off. The GC percentage was 43%, in line with other cattle sequencing data.

3.2.3 Scaffolding data

A blood sample that was taken from the reference animal post slaughter was stored at -80°C and shipped on dry ice to Dovetail Genomics (Chicago, USA) for the generation of scaffolding libraries. Both Hi-C and Chicago libraries were made for the sample by Dovetail Genomics and sequenced on a 2x151bp Illumina platform. The Chicago method is able to identify and orientate short range joins, <

1Mb in size, such as those observed on the X chromosome. The Hi-C method is used for larger scale joins, such as those present on the autosomes.

3.2.4 Assembly settings

Contigs were assembled from the PacBio subreads using Falcon. The assembly settings were: pread > 120000, daligned raw read overlap settings '-T16 -k14 -h35 -w7 -e.80 -l1000 -s1000 -M58 -B200' , daligner pread overlap settings '-T16 -k14 -h60 -w6 -e.95 -l500 -s1000 -M58 -B50', falcon consensus calling '--output_multi --min_idt 0.70 --min_cov 4 --max_n_read 200 --n_core 24' and final overlap filtering '--max_diff 140 --max_cov 220 --min_cov 5 --bestn 20 --n_core 24'.

The primary and associated contigs were combined for error correction, called polishing. The subreads bam files (which were directly obtained from the sequencing provider) were aligned to the assembled contigs using pbalign with default parameters. The alignments were then used to polish the contigs. Polishing of each contig was run on 24 cores and output included sequence files (fasta, fastq) and variant files (vfc and gff). The polished contigs were then used in a second round of polishing using the same method and parameters.

3.3 Molecular Protocols

3.3.1 Extraction of high quality achieve quality DNA from semen straws

A protocol for purification of DNA from semen using the Qiagen Puregene Tissue Kit was provided by Mr Brett Mason (DEDJTR). This protocol produced high quality long DNA suitable for sequencing, including long read sequencing.

Protocol

1. Dispense semen into a 1.5 ml microfuge tube.

2. Centrifuge at top speed (~16 000 g) for 30 s. Remove supernatant by pipetting.

3. Add 600 µl Cell Lysis Solution to the pellet and resuspend by pipetting.

4. Add 24 µl 1M dithiothreitol (DTT) and mix by inverting 25 times (DTT should be used in a fume cupboard). Pulse spin.

5. Add 5 µl Proteinase K and mix by inverting 25 times. Pulse spin.

6. Incubate overnight at 55°C, shaking at 250 rpm.

7. After overnight incubation, add 5 µl Proteinase K. Mix by inverting 25 times, pulse spin, and continue incubating for about 4 hrs more.

8. Cool sample to room temperature.

9. Add 200 µl Protein Precipitation Solution, vortex for 20 s, and leave on ice for 5 min.

10. Centrifuge at top speed for 3 min. The supernatant contains the DNA, and the precipitated proteins should form a tight pellet.

11. Dispense 600 µl isopropanol into a 1.5 ml screw-cap tube. Pour the supernatant into the isopropanol.

12. Mix the sample by inverting 50 times.

13. Centrifuge at top speed for 1 min to pellet the DNA.

14. Add 600 µl 70 % ethanol. Centrifuge at top speed for 1 min. Pour off supernatant.

15. Repeat step 14.

16. Centrifuge again at top speed for 1 min, and remove any remaining ethanol by pipetting.

17. Add 50 µl TE Buffer. To resuspend, leave at room temperature overnight and then vortex.

The resulting DNA was visualized on a Pippin Pulse gel electrophoresis machine (Fig. 4) as per the manufactures instructions.

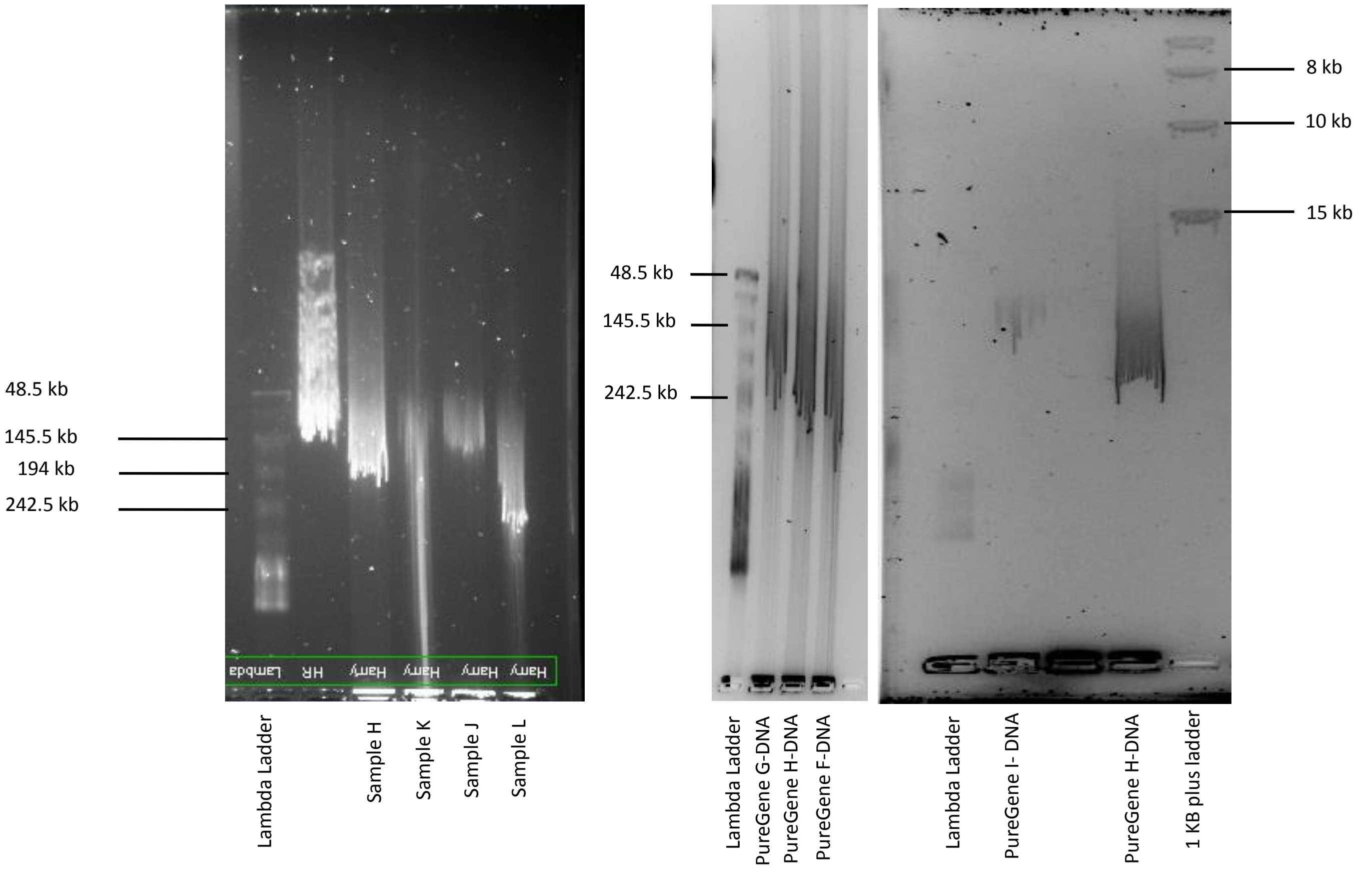


Fig. 4 Pulse Field Gel Electrophoresis images used to examine the length of DNA extracted from the semen samples. 0.75% SeaKem Gold Agarose was used and run for 16hrs using the pippin pulse "5-480 kb" setting. The gel was set and run using 0.5 X KBB buffer, and one KB plus and lambda DNA ladder were used as reference.

3.3.2 Nextera and TruSeq Illumina library preparation

Both Nextera and TruSeq library preparations were used for the whole genome sequencing of bull semen. While both methods were successful, the Nextera protocol is able to produce sequence data from as little as 10ng of DNA, which is 1/100 of the DNA input required for the TruSeq library preparation method. This has the distinct advantage that it can therefore be used to sequence samples where the volume is limited, as may be the case with semen samples from historically important bulls.

3.4 Short reads sequence analysis

3.4.1 Quality control

Short reads were obtained from a range of instruments including the Illumina HiSeq, NovaSeq and BGI sequencer. Reads were obtained in fastq format which provides information not only on the sequence but also the reliability of the sequence data. Data was trimmed to the 1000 Bulls Genome Project standard – briefly, this comprised removing any adapters which are artefacts of the sequencing process, then removing the poor quality regions from the sequence before a final minimum average quality threshold (q>20) and length (l>50) is applied. The percentage of read pairs that pass the quality control is an indication of run quality. Where less than 80% of reads pass the filter the data quality is low - this was not the case for our data.

3.4.2 Alignments

The reads that passed quality control were then aligned to the reference genome using BWA-mem. BWA-mem is a widely used and efficient program for comparing short read data to a reference genome. The output files, called BAM files, contain the alignment position and alignment quality for each of the millions of reads from each sample. Any duplicate reads were then flagged with Picard MarkDuplicates.

3.4.3 Genetic variants identified

The short read analysis pipeline (Fig. 5) was based on the 1000 Bull Genomes Project (Daetwyler *et al.* 2014). Variants were called based on the alignments contained in the BAM files, differences between the reference and the aligned data indicate genetic variants. Variants were called using GATK’s HaplotypeCaller and variants filtered for reliability and quality. Read depths from the filtered reads were also calculated using GATK.

For the analysis of introgression, poll alleles and variant origin additional steps were applied. After alignment and filtering, SNP and indels were filtered using hard filters with thresholds that were based on what is recommended by GATK for removal of variants with poor quality scores.

The applicable settings that were adjusted form the default included QD: the quality by depth which is the quality score normalized by allele depth; FS: the Fisher Strand test to detect strand bias, lower scores indicate more strand bias; MQ: the root mean square of the mapping quality score; ReadPosRankSum: the ranked sum test for the distance of alleles from the end of the reads, as the closer the variant is to the end of the read, the more error prone it is; MQRankSum: the mapping qualities of reads that support the reference allele with those that support the alternative allele; SOR: measures strand bias; InbreedingCoeff: the level of inbreeding in a group of samples by estimating the population allele frequency from the sample genotypes.

SNP were filtered with the following annotations and thresholds:

QD<2.0‖FS>60.0‖MQ<40.0‖ReadPosRankSum<−8.0‖MQRankSum<−12.5‖SOR>3.0QD<2.0‖FS>60.0‖MQ<40.0‖ReadPosRankSum<−8.0‖MQRankSum<−12.5‖SOR>3.0

InDels were filtered with the following annotations and thresholds:
QD<4.0||FS>100.0‖InbreedingCoeff<−0.8||ReadPosRankSum<−10.0||SOR>6.0QD<4.0||FS>100.0‖InbreedingCoeff<−0.8||ReadPosRankSum<−10.0||SOR>6.0

3.4.4 Trait associated variants mapped to new reference

Variants that have previously been associated with important traits were mapped onto the new reference using BLAST. Positions were filtered for specificity to ensure reliability of new positions.

3.4.5 Variant origin and annotation

B. taurus and Gir (Gyr) variants were obtained from the 1000 Bulls Genomes Project (Daetwyler *et al.* 2014) and from the study by Liao *et al.* (2013), respectively. Only SNP data from each study was used and allele frequencies were calculated using in-house scripts. Following this, SNP found in common between *B. taurus* and Brahman and between Gir and Brahman were determined, selected and placed into separate files. This was done with a python script that selected SNP that shared the same genomic location.

To determine introgression of *B. taurus* and Gir genomic DNA in Brahman, a similar method was inspired by the calculation as described in Bolormaa *et al.* (2011a), with the difference being our formula uses 250kb windows. Our method involved calculating *bosind_250* values, where the *bosind_250* statistic estimates how differentiated two populations are based on sequence, allele frequency and SNP calls. In this analysis *bosind_250* values were calculated between *B. taurus* and *B. indicus* SNP that are found in the Brahman animals using the following formula:

bosind_250=Ht−HsHtbosind_250=Ht−HsHt

where:

Hs=PBT(1−PBT)+Pbrai(1−Pbrai)Hs=PBT(1−PBT)+Pbrai(1−Pbrai)

and

Ht=2∗(PBT+Pbrai)2×1−PBT+Pbrai2Ht=2∗(PBT+Pbrai)2×1−PBT+Pbrai2

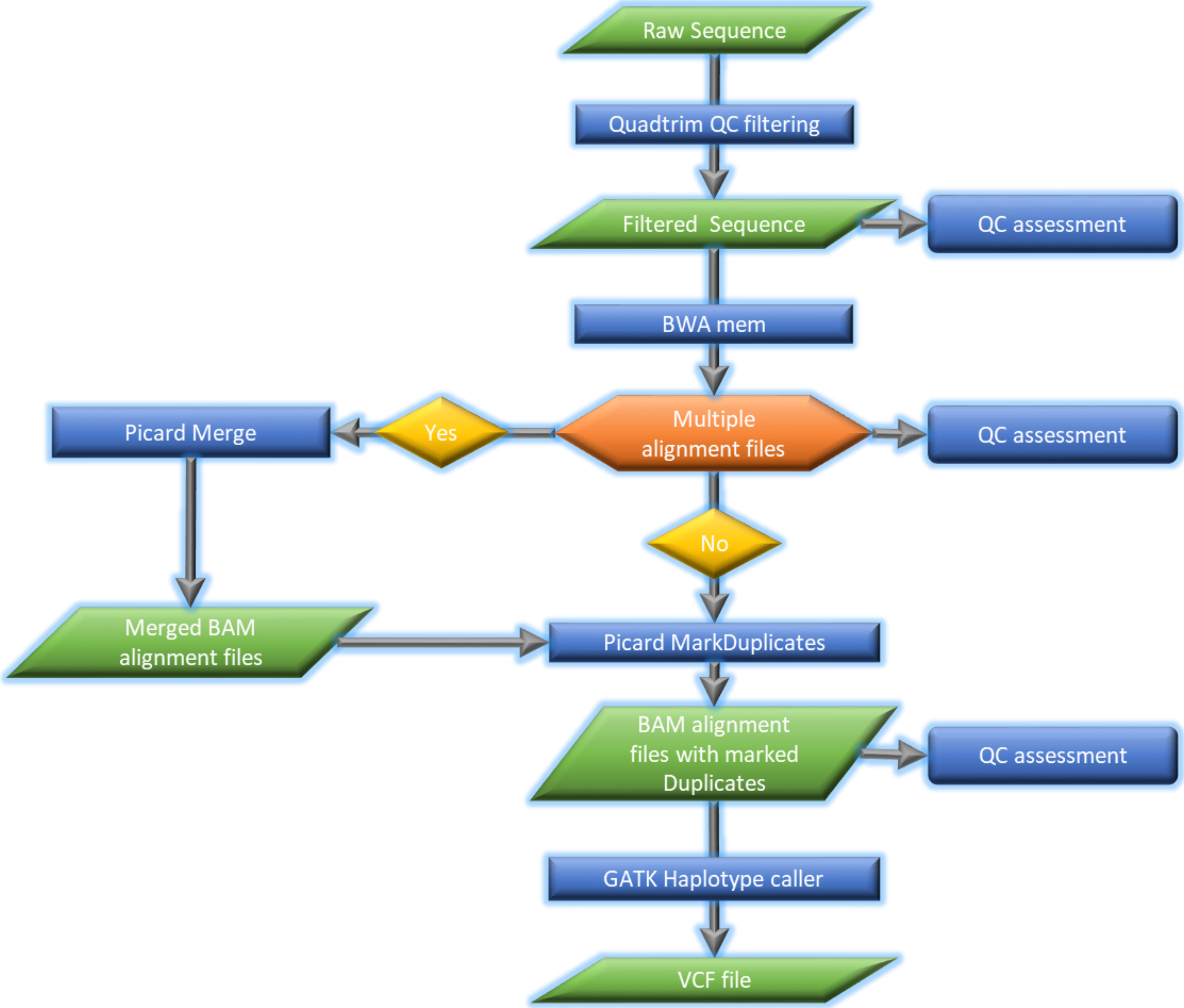


Fig. 5 Alignment pipeline for variant calling of short read bull data against the new Brahman reference.

PBT is the SNP allele frequency of the alternative allele in either *B. taurus* or *B. indicus* and *Pbrai* is the SNP call in that Brahman SNP individual. *Pbrai* is 0 if the SNP call is homozygous for the reference, 0.5 if the SNP call is heterozygous and 1 if the SNP call is homozygous for the alternative allele.

The calculation was performed twice for each Brahman animal across all SNP, the first calculation used common SNP between Brahman and *B. indicus*, the second calculation used common SNP between Brahman and *B. taurus*.

All SNPs were grouped into windows of 250 kilobases (kb) and the average *bosind_250* values (*bosind_250avg*) were calculated for each window We used 250kb windows as this provides enough SNP from sequence data in each window to perform this analysis and obtained robust estimates of the *bosind_250* value. This was done by adding the *bosind_250* values across all SNP within a 250kb window and dividing that number by the total number of SNP found in that window as shown in the following formula:

bosind_250avg=∑(bosind_250fw)nbosind_250avg=∑(bosind_250fw)n

where *bosind_250fw* are the *bosind_250* values for all SNP found in a fixed-size window and *n* is the total number of SNP found in that fixed-size window.

Annotation of the Brahman SNP was carried out using the NGS-SNP tool (Grant *et al.* 2011). CpG Isles annotations were from the study by Su *et al.* (2014). Micro RNA (miRNA) target site annotations were from the Microcosm Database (Griffiths-Jones *et al.* 2007). Long noncoding RNA annotations were from Koufariotis *et al.* (2015).

4 Results

4.1 Data Summary

4.1.1 PacBio

A total of 195Gbp of sequence data was obtained from the PacBio Sequel. Given an expected genome length of 2.8GB, the achieved sequencing depth was 69.6X coverage. For the sequence read, the mean polymerase N50 was 20.5 kbp ± 0.3kbp across 28 SMRT cells, and the mean subread N50 was 17.3 kbp ±0.18 (Table 1).

4.1.2 Illumina

After filtering and alignment the total amount of Illumina data represents 1938X coverage of the northern cattle genome, and includes representatives of Brahman, as well as Droughtmaster, Santa Gertrudis and Red Sindhi.

4.1.3 Scaffolding

In total 43.2GB of paired end sequence data was generated from three Chicago libraries, and 41.3GB of paired end sequence data was generated from three Hi-C libraries. The physical coverage of the genome from the Chicago library was 53.1X (Fig. 6), and the physical coverage from the Hi-C data was 21,814.4X (Fig. 7). This difference is due to the much larger insert size of the Hi-C data.

Table 1. Sequencing statistics of the reference animal Elrose Neomi as reported by the Ramaciotti Centre for Genomics.

Run Number	Cell Number	Total Bases (Gb)	Longest Subread (bp)	Longest Subread N50 (bp)	Polymerase RL (bp)	Polyerase N50 (bp)	Adapter Dimer	Short Insert	Run Start	Run Complete	Primary Analysis Version
1	1	6.797	11449	18250	13580	21750	0.06	0.37	6/10/2017 4:16	9/10/2017 16:37	5.0.0.12545
1	2	9.232	10400	17250	12616	21250	0.06	0.38	6/10/2017 4:16	9/10/2017 16:37	5.0.0.12545
2	1	8.788	11786	18750	14081	23250	0.08	0.36	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	2	7.912	11346	18250	13375	21750	0.09	0.38	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	3	6.937	11233	17750	13162	21750	0.09	0.44	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	4	8.526	11217	18250	13401	22250	0.06	0.37	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	5	8.005	10798	17750	12538	21250	0.16	0.52	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	6	6.738	10795	17750	12232	20250	0.12	0.57	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	7	6.297	10809	17750	12317	20750	0.1	0.63	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
2	8	7.962	10678	17750	12541	21250	0.07	0.34	10/10/2017 6:45	13/10/2017 19:06	5.0.0.12545
3	1	6.918	11685	18250	13812	22250	0.04	0.29	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	2	8.391	10425	17750	12063	20750	0.06	0.52	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	3	7.326	10695	17750	12476	20750	0.06	0.34	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	4	8.642	10918	18250	12909	21750	0.08	0.37	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	6	7.743	10539	17250	12282	20750	0.05	0.42	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	7	6.834	10423	17250	12017	20250	0.08	0.31	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
3	8	8.507	10859	17750	12971	22250	0.06	0.4	16/10/2017 2:36	19/10/2017 14:58	5.0.0.12545
4	1	6.562	10629	17250	12248	20250	0.09	0.43	20/10/2017 2:48	23/10/2017 15:09	5.0.0.12545
4	2	7.691	11189	18250	13314	22250	0.06	0.22	20/10/2017 2:48	23/10/2017 15:09	5.0.0.12545
4	3	6.962	10571	17250	12232	20250	0.12	0.45	20/10/2017 2:48	23/10/2017 15:09	5.0.0.12545
4	4	8.345	10782	17750	12752	21750	0.06	0.3	20/10/2017 2:48	23/10/2017 15:09	5.0.0.12545
5	1	5.334	9400	16250	10428	18250	0.09	0.5	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	2	5.677	9418	16250	10558	18250	0.08	0.46	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	3	5.475	9158	15750	10245	17750	0.08	0.6	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	4	4.8792	9657	16250	10768	18250	0.09	0.53	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	5	4.703	9144	15750	10067	17750	0.11	0.51	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	6	4.553	9052	15250	9908	16750	0.16	0.67	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
5	7	4.750	9079	15750	9987	17250	0.12	0.47	30/10/2017 2:53	2/11/2017 15:13	5.0.0.12545
Mean		7.02	10504.8	17339.3	12174.3	20464.3	0.085	0.434	-	-	-
S.E.M		0.26	153.5	178.3	234.0	334.7	0.006	0.021	-	-	-

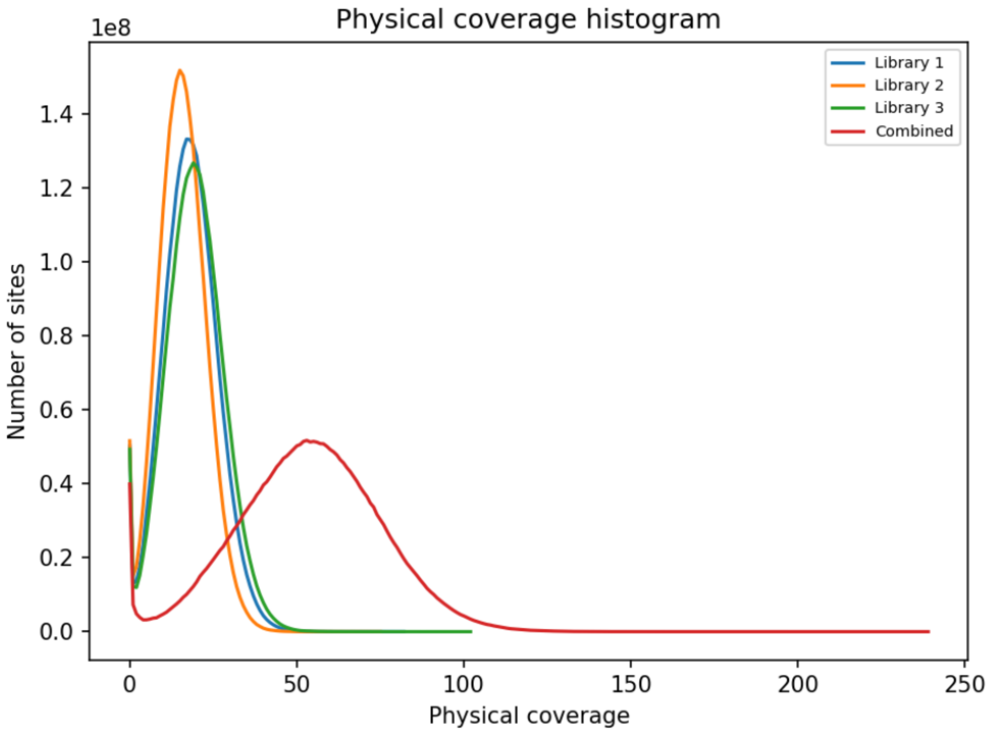


Fig. 6 Histogram of physical coverage over input assembly. Coverage values are calculated as the number of read pairs with inserts between 1 and 100 kb spanning each position in the input assembly.

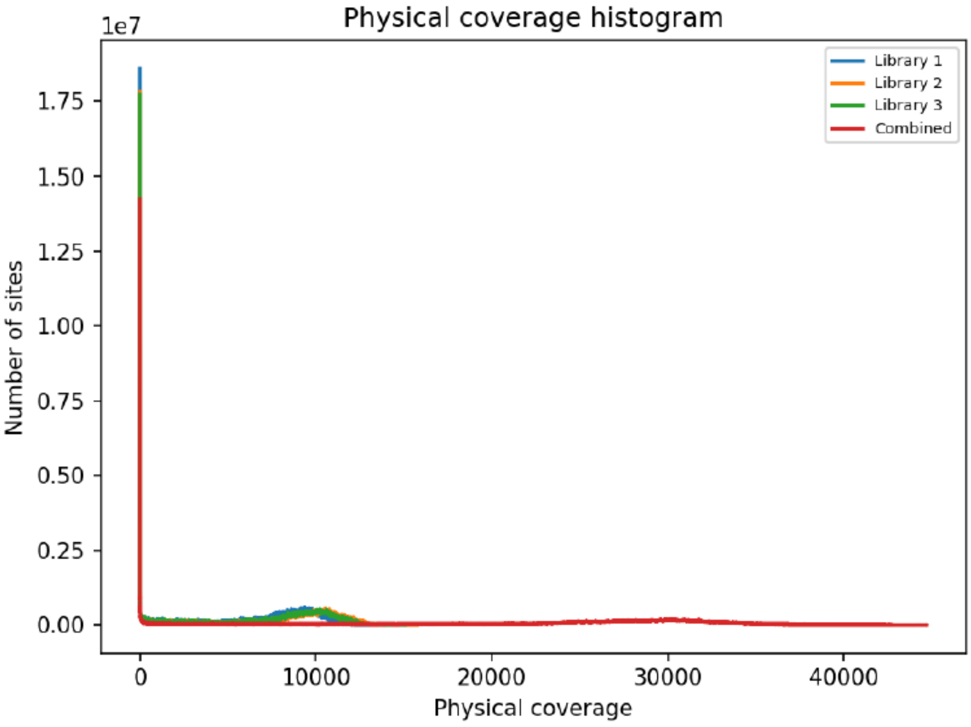


Fig. 7 Histogram of physical coverage over input assembly. Coverage values are calculated as the number of read pairs with inserts between 10 and 10,000 kb spanning each position in the input assembly.

4.2 Contig Assembly

Contigs were assembled in the High Performance computing system at the University of Queensland. After optimisation the final assembly consisted of 2,673,706,688bp in 1867 contigs – which are continuous stretches of sequence that the assembly algorithm has generated. The size of the contigs ranged from 5412bp to 56,814,202bp. The N50, N90 and N95 of the assembly were 11,122,568; 1,510,691; and 1,510,691 respectively. These numbers are the minimum contig length needed to cover 50%, 90% and 95% of the assembled genome, and as a measure of the continuity of the assembly.

The longest contig was homologous to 47% of the *B. taurus* chromosome 1. Additionally 59% of chromosome 28 was covered by a single contig. Chromosomes 2, 3, 4, 5, 6, 8, 9, 10, 11, 16, 19, 20, and 24 were all represented in contigs that covered at least 20% of the chromosome. 93% of BUSCO genes were detected, with 260 of the 303 full length single copies.

4.3 Scaffolding

Both Hi-C and Chicago libraries were made for the same animal by Dovetail Genomics and sequenced on a 2x151bp Illumina platform. The HiRise software (Dovetail Genomics) made 850 and 256 joins from the Chicago and Hi-C data respectively (Fig. 8).

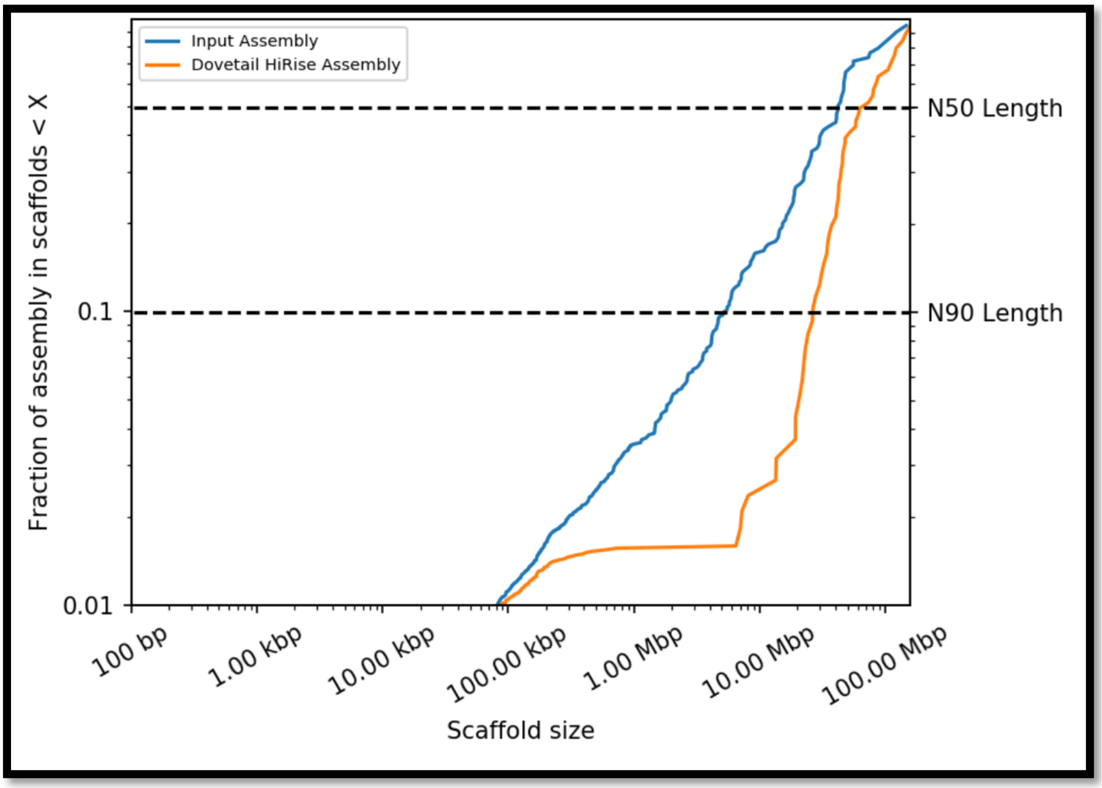


Fig. 8

A comparison of the contiguity of the input assembly and the final HiRise scaffolds. Each curve shows the fraction of the total length of the assembly present in scaffolds of a given length or smaller. The fraction of the assembly is indicated on the Y-axis and the scaffold length in basepairs is given on the X-axis. The two dashed lines mark the N50 and N90 lengths of each assembly.

4.1 Gap filling and error correction

4.1.1 Polishing

Scaffolds were polished with the original PacBio sequence reads then gaps were filled using the same reads by PBJelly (English *et al.* 2012). A total of 5 rounds of polishing and gap filling was performed before no additional gaps were filled. Regions with were less that 1Mb in length and flanked by gaps were reverse complimented and another 2 rounds of PBJelly gap filling performed. Regions for which this did not result in any filled gaps were reverse complimented again to ensure the original orientation was maintained.

4.1.2 Gap filling with additional long reads

Data from a MinION flowcell was obtained by preparing a 1D library from a Brahman bull, producing 4.7GB of data with a read N50 of 14.5kb. The reads were mapped to the reference using minimap2 (Li 2018) with the preset ONT settings. Regions with gaps were examined and three were completely covered by reads, the first was a 8.6kb repeat on chromosome 1, the second was a 5.5kb insert, the third showed an incorrectly inserted 7.7kb region that was replaced with 320bp.

4.1.3 Analysis of remaining gaps

After gap filling 317 gaps remain on the assembled chromosomes (with 86 in unplaced scaffolds). Inspection of the 50kb of sequence either side of the remaining gaps revealed that the gaps tend to be surrounded by highly repetitive or duplicated sequence (Fig. 9). This may indicate that the inability to fill the remaining gaps is due to complex regions of the genome, and the extremely long sequence reads may be required to close the reaming gaps, as the reads must span long repetitive stretches.

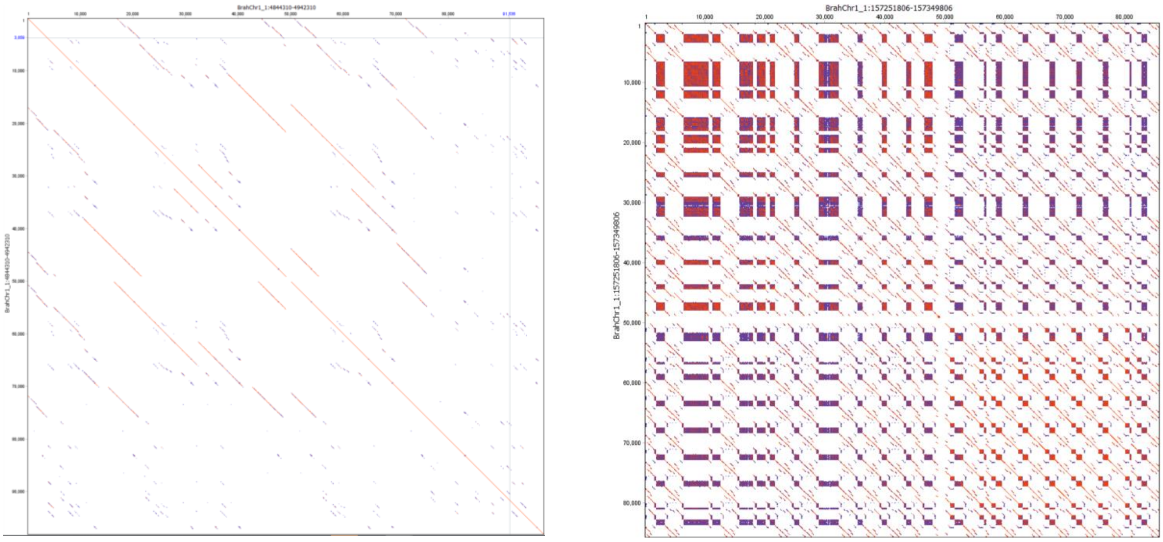


Fig. 9 Dotplots showing repetitive nature of the sequence surrounding the remaining gaps in the genome. The diagonal line from top left to bottom right indicates self-alignment, with the gap in the centre of the sequence. Other lines indicate homology, a sign of a repeated segment. Dotplots were generated in Geneious 11.0.2 using the EMBOSS 6.5.7 tool dottup.

4.1.4 Removal of spurious indels that are replacing SNP

Illumina reads were aligned to the assembled genome using BWA-mem paired alignment. The alignment was then used to call variants with the mpileup command. The output file was then filtered to identify regions which showed two single basepair insertion/deletions within 5bp of each other, with a minimum total depth of 8 reads, a minimum number of reads with the deletion of 4, and ratio of between 0.2 to 0.8 of the reference (i.e. the base in the reference compared to the number of bases with the deletion). The second base position that matched these criteria for each region was removed from the contig – i.e. the reference sequence was changed from including the insertion of the indel, to the deletion. The reads were then realigned to the edited scaffold and the edited positions were inspected for the retention of an indel call – which would indicate that the position was a true indel and not a SNP erroneously called as two side-by-side indels. The positions which exhibited insertions or deletions in at least 2 reads were manually inspected (N=52 of 583,675 edited sites). Of these 52 sites, 27 were assessed as unimproved compared to the original sequence and the base deletion was reversed such that the original sequence was reinstated.

4.2 Assembly Mapping

4.2.1 Depth

Sequence reads from the DNA of the reference animal were mapped back to the assembly. When reads were mapped the normalised (corrected for scaffold length) number of mapped reads were quite stable for the large scaffolds, which represent chromosomes (Table 2). The exceptions were BrahChr18_1, BrahChr23_1 and BrahChr8_1. The depth of the unplaced contigs forms a continuous distribution (Fig. 10), from 1.4 to 11665 reads mapped (after normalisation for scaffold length), compared to an average of 48.2 mapped reads for the normal large scaffolds.

Deeper examination of the depth of scaffold depth revealed the depth variance was isolated to discrete regions of the scaffolds. BrahChr8_1 displayed several high depth peaks when the Illumina data was mapped to the scaffold, however most resolved when the PacBio data was used. Examining the one remaining peak which was located in the 22kb of the scaffold revealed a highly repetitive structure (Fig. 11; Fig. 12). The repeated sequence showed homology to the *B. taurus* 1.711a satellite.

BrahChr18_1 revealed at the excessive depth was limited to the ends of the scaffold (Fig. 13). The first 50kb region displayed a depth in excess of 8000X in the Illumina data (note the axis is limited to 8000X for readability), while the depth observed in the PacBio data was limited to 800X. A dotplot which displays self-homology and can be used to find repetitive elements revealed that this region is highly repetitive (Fig. 14). A single repetitive element was extracted and again found to be homologous to the *B. taurus* 1.711a satellite. Another small peak is visible at the end of the scaffold. This smaller peak does not appear to be repetitive, but does show some homology to *B. taurus* mitochondrial antiviral signalling protein (MAVS).

Examination of the depth across scaffold BrahChr23_1 revealed several peaks when the shorter Illumina data is examined however most of these resolve when the long PacBio data is used (Fig. 15). Further investigation of the remaining peak reveals a 18320bp duplication (Fig. 16). Comparison of the peak sequence to the nt public database using BLAST did not suggest any know repetitive structures, however there was small sections with homology to several proteins including MHCII and reverse transcriptase. The translated protein sequence was compared to the non-redundent (nr) NCBI database using BLASTx, which also resulted in homology to a reverse transcriptase family protein. This scaffold contains four gaps, one of which is just upstream of this high coverage area, at position 1357242 of scaffold BrahChr23_1. This may suggest a local misassembled region that is causing both the depth anomaly and the inability of the gap to be filled.

Taken together, this information demonstrates the value of using long read data for high quality assemblies. When short read data is used there appears to be several areas of the scaffolds which are unable to produce accurate mapping, however the nature of long read data resolves this issue by allowing reads to map to the unique regions flanking repeats and therefor reduce, if not fully resolve, the ambiguity. There does, however, appear to still be a small number of regions that are characterised by repeats which are not fully resolved. It may be that these regions are collapsed regions of long repeats which prevent the reads from mapping with even depth. This anomaly may be resolved by the addition of super long reads, which can span the entire repeat.

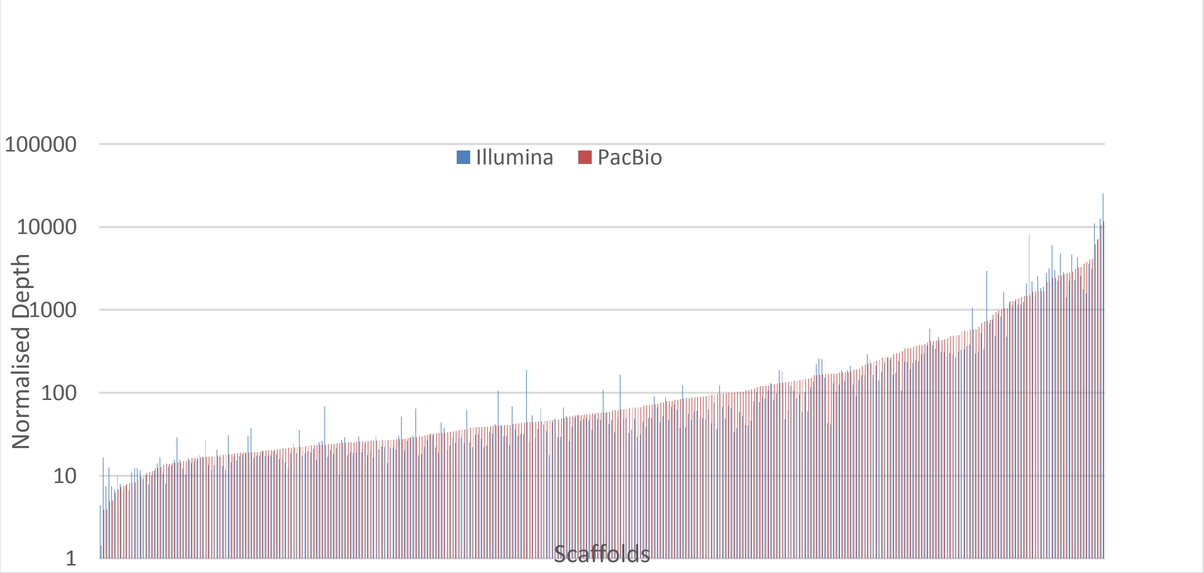


*Fig. 10 Sequencing depth of unplaced scaffolds using both PacBio and Illumina data. Scaffold order is based on PacBio sequence depth. Due to differences in read lengths and data volumes, Illumina normalised depth was calculated as (N/L)*100, while PacBio was (N/L)*10000, where N is the number of mapped reads and L is the length of the contig. Note the large variation in sequencing depth compared to the contigs that are assigned to chromosomes.*

*Table 2. Length and depth metrics for the Brahman genome scaffolds that are assigned to chromosomes. Due to differences in read lengths and data volumes, Illumina normalised depth was calculated as (N/L)*100, while PacBio was (N/L)*10000, where N is the number of mapped reads and L is the length of the contig. Note the large variation in sequencing depth compared to the contigs that are assigned to chromosomes.*

Scaffold Name	Scaffold Length	Illumina mapped reads	Normalised Illumina Depth	PacBio mapped reads	Normalised PacBio Depth
BrahChr1_1	1.57E+08	63149677	40.1	738619	46.9
BrahChr10_1	23164039	9396234	40.6	112155	48.4
BrahChr10_2	79069930	30916222	39.1	371664	47.0
BrahChr11_1	7309453	2801274	38.3	34578	47.3
BrahChr11_2	41948137	16322841	38.9	194126	46.3
BrahChr11_3	57746229	22980414	39.8	278007	48.1
BrahChr12_1	40949876	16057437	39.2	192602	47.0
BrahChr12_2	30957533	12155966	39.3	154882	50.0
BrahChr12_3	14284756	5719115	40.0	69588	48.7
BrahChr13_1	83583526	33128107	39.6	403557	48.3
BrahChr14_1	13615691	5603958	41.2	68689	50.4
BrahChr14_2	44434581	17414790	39.2	206519	46.5
BrahChr14_3	24281917	9368952	38.6	112157	46.2
BrahChr15_1	84677148	34517674	40.8	427205	50.5
BrahChr16_1	7099584	2896242	40.8	36111	50.9
BrahChr16_2	74095649	29152499	39.3	350092	47.2
BrahChr17_1	73050433	28811651	39.4	344982	47.2
BrahChr18_1	13389667	19178439	143.2	235876	176.2
BrahChr18_2	19269050	7646159	39.7	92239	47.9
BrahChr18_3	34434199	13684495	39.7	170851	49.6
BrahChr19_1	58402991	24952132	42.7	297730	51.0
BrahChr19_2	6469098	2489190	38.5	31062	48.0
BrahChr2_1	1.36E+08	53286307	39.1	637362	46.8
BrahChr20_1	71666168	27703544	38.7	333192	46.5
BrahChr21_1	34749500	16025499	46.1	196594	56.6
BrahChr21_2	35757710	14036899	39.3	169528	47.4
BrahChr22_1	61015590	24352219	39.9	297658	48.8
BrahChr23_1	8146469	3995343	49.0	48560	59.6
BrahChr23_2	19253135	7685102	39.9	92352	48.0
BrahChr23_3	26276138	10291152	39.2	125241	47.7
BrahChr24_1	62280714	24379863	39.1	295100	47.4
BrahChr25_1	42514956	17666496	41.6	218174	51.3
BrahChr26_1	51727733	20477531	39.6	248276	48.0
BrahChr27_1	45760196	18022754	39.4	217959	47.6
BrahChr28_1	45536915	21936314	48.2	233832	51.3
BrahChr29_1	21873311	8924905	40.8	110401	50.5
BrahChr29_2	30021527	12037729	40.1	147301	49.1
BrahChr3_1	1.22E+08	47371296	39.0	568433	46.8
BrahChr4_1	1.2E+08	46844916	39.1	557897	46.5
BrahChr5_1	40304107	15844027	39.3	188813	46.8
BrahChr5_2	80691544	31454035	39.0	378021	46.8
BrahChr6_1	1.18E+08	46385303	39.3	558035	47.3
BrahChr7_1	1.11E+08	43573109	39.3	525832	47.5
BrahChr8_1	22485169	15557045	69.2	201929	89.8
BrahChr8_2	47827448	18658915	39.0	221736	46.4
BrahChr8_3	42728095	16516243	38.7	198978	46.6
BrahChr9_1	1.05E+08	42517653	40.6	493939	47.1
BrahChrX_1	1.47E+08	59124873	40.4	715585	48.8

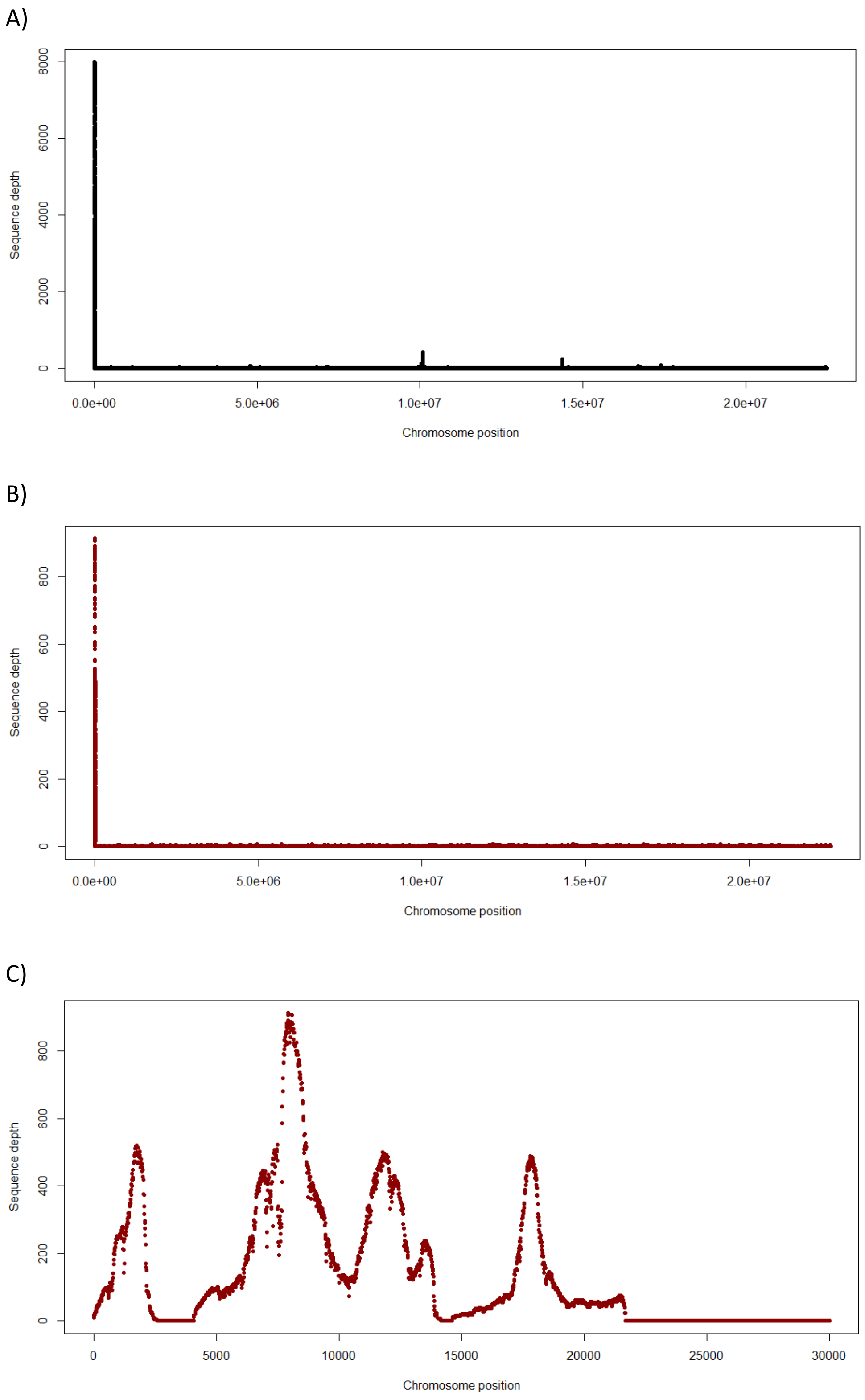


Fig. 11 Sequence depth of scaffold BrahChr8_1. A) Illumina data across the entire scaffold. B) PacBio data across the entire scaffold. C) PacBio data across the high depth region.

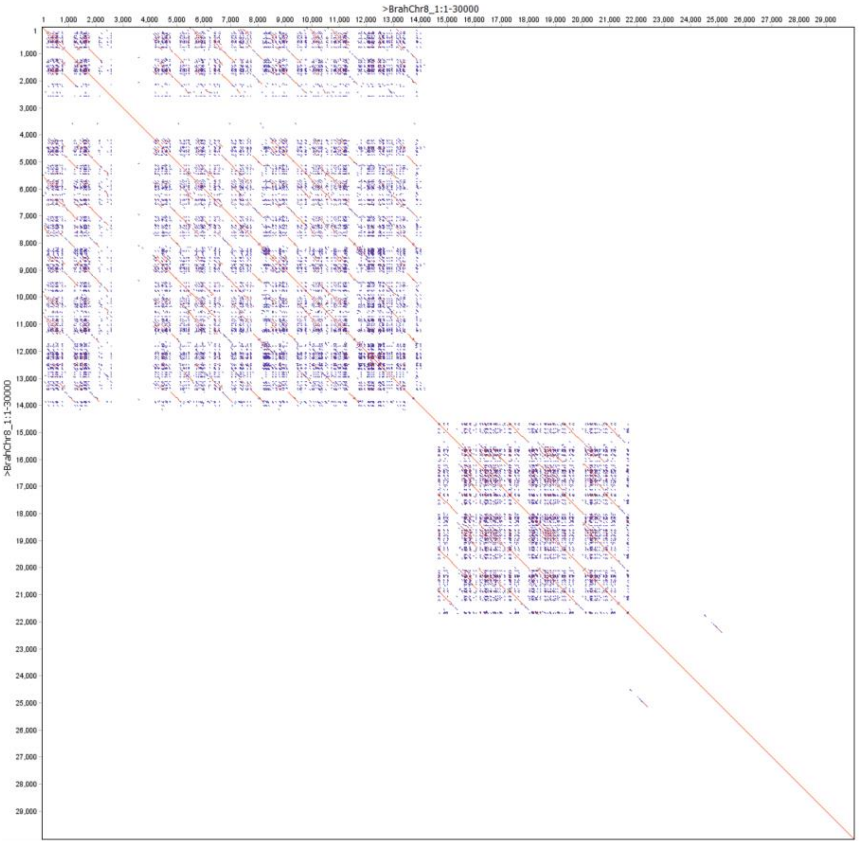


Fig. 12 A self-dotplot of the high depth region at the start of scaffold BrahChr8_1 showing a high level of repetitiveness, as illustrated by the diagonal lines.

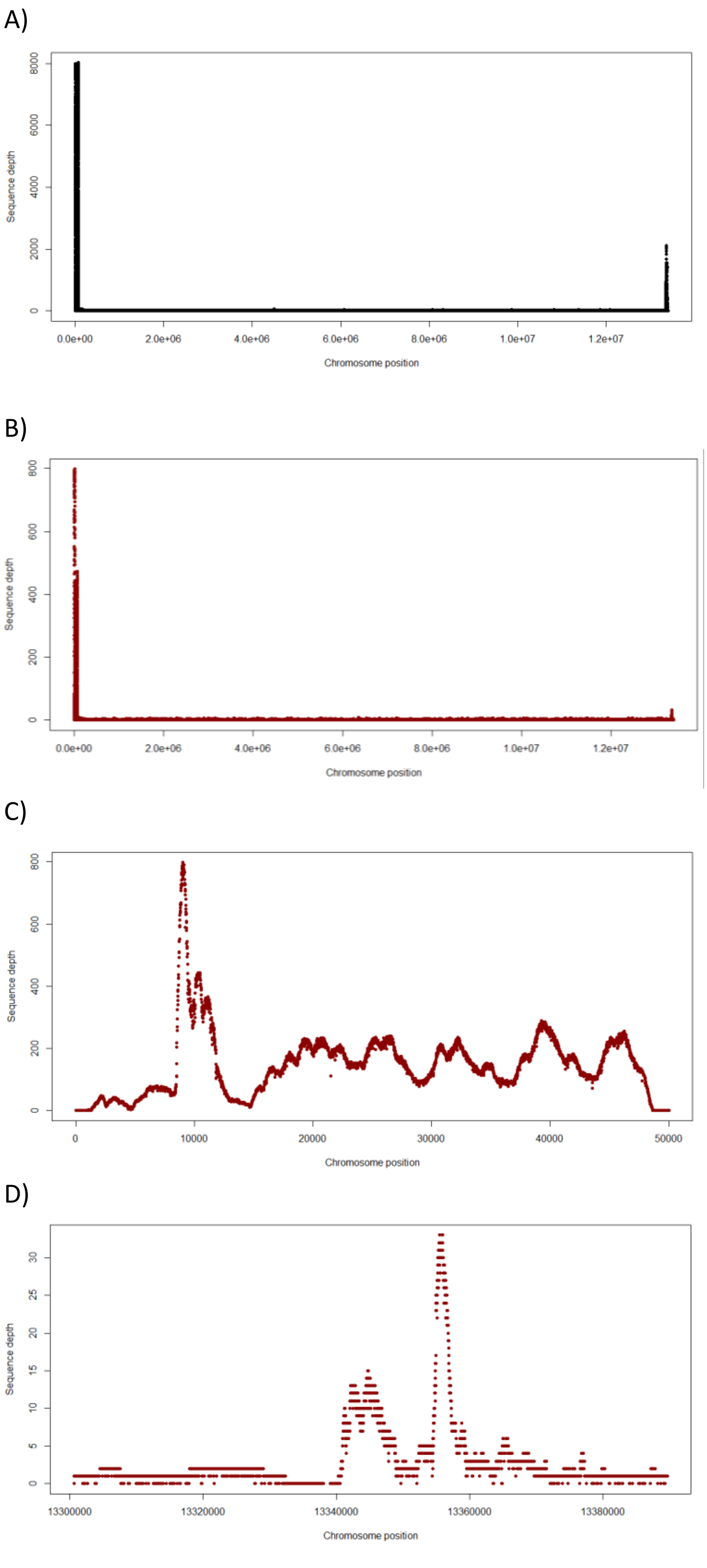


Fig. 13 Sequence depth of scaffold BrahChr18_1. A) Illumina data across the entire scaffold (depth capped at 8000X). B) PacBio data across the entire scaffold. C) PacBio data across the high depth region at the start of the scaffold. D) PacBio data across the high depth region at the end of the scaffold.

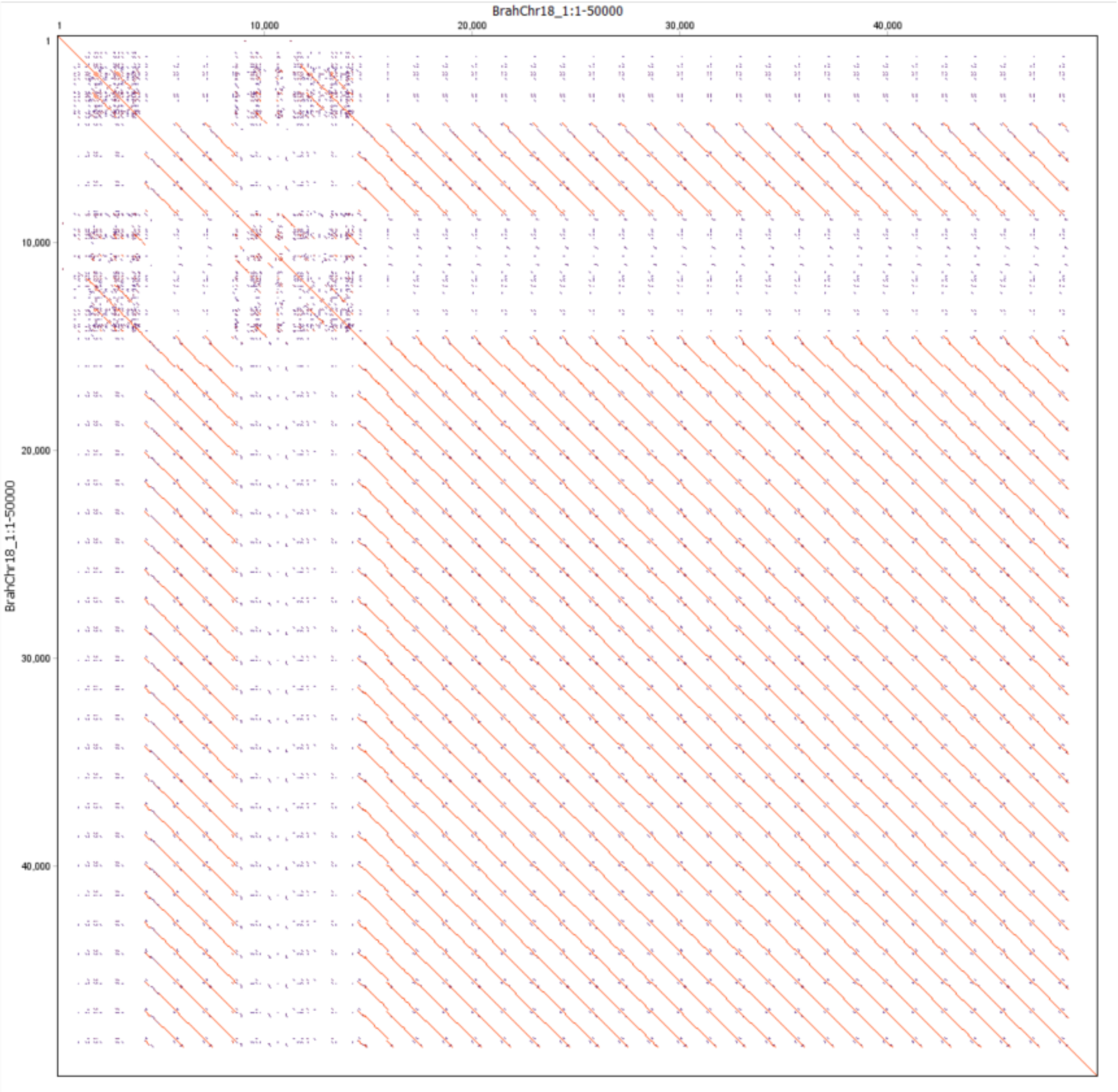


Fig. 14 A self-dotplot of the high depth region at the start of scaffold BrahChr18_1 showing a high level of repetitiveness, as illustrated by the diagonal lines.

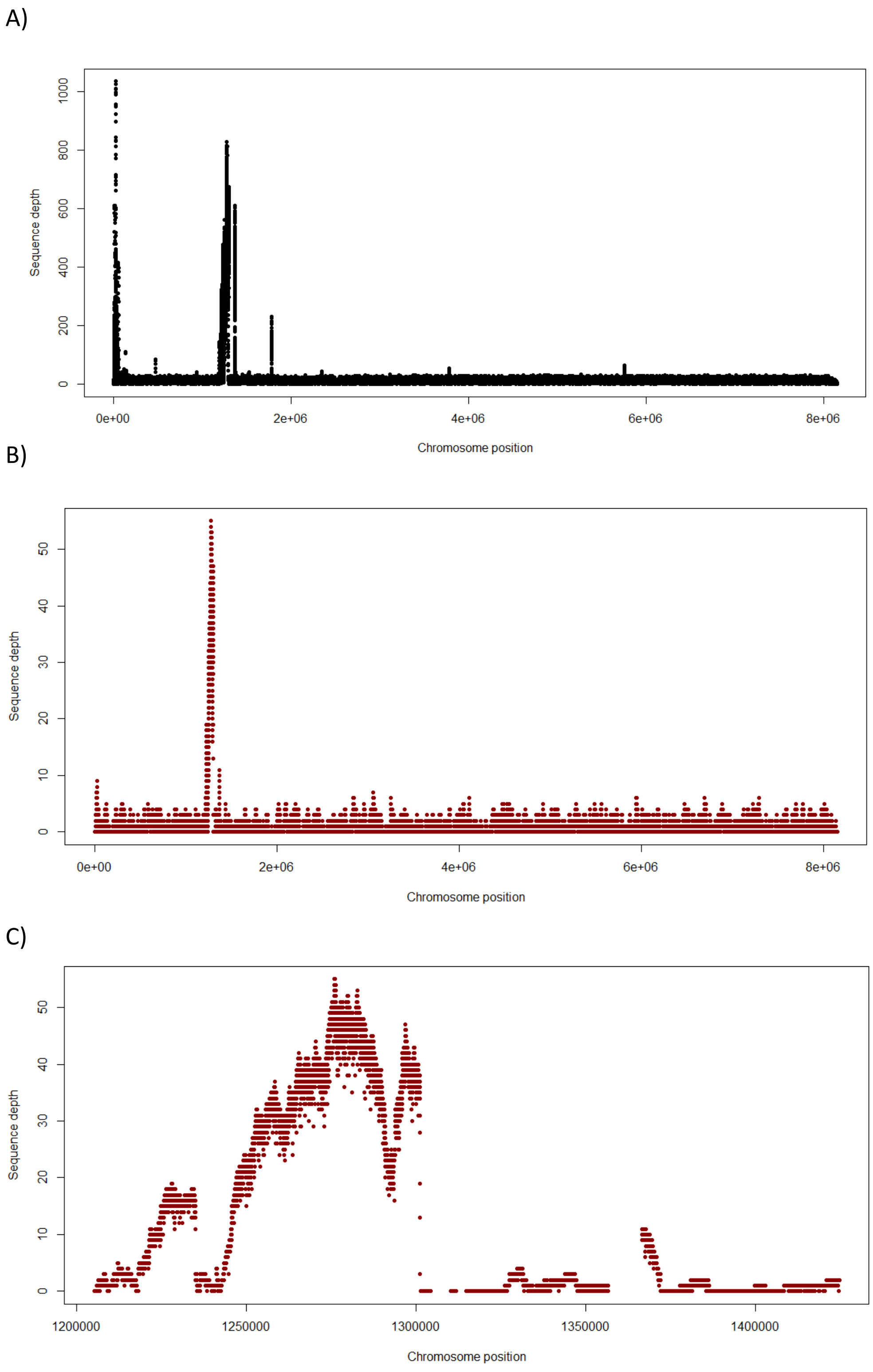


Fig. 15 Sequence depth of scaffold BrahChr23_1. A) Illumina data across the entire scaffold. B) PacBio data across the entire scaffold. C) PacBio data across the high depth region.

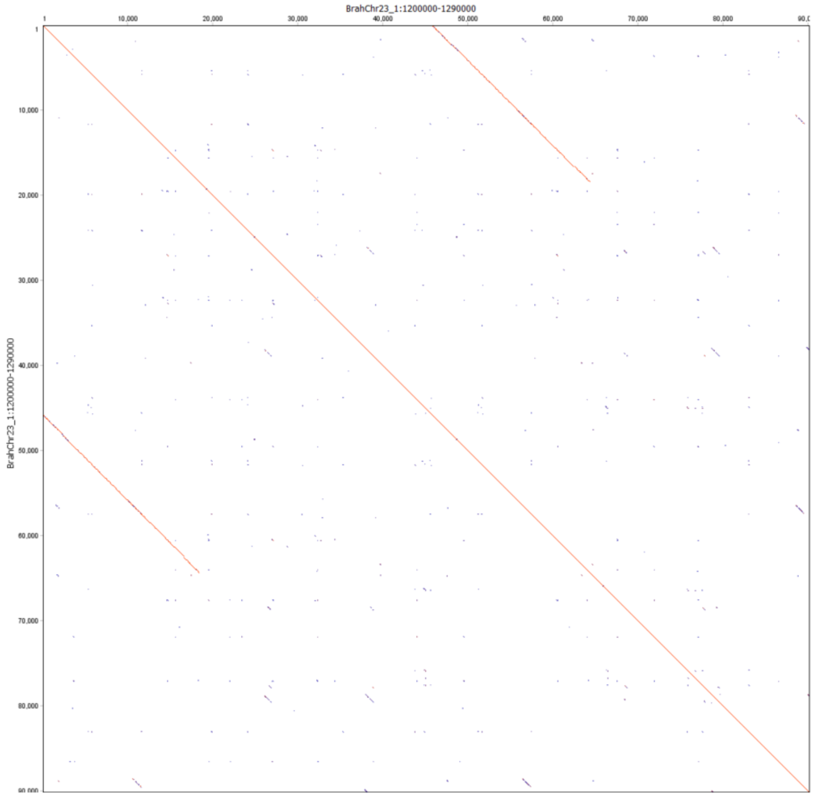


Fig. 16 A self-dotplot of the high depth region in scaffold BrahChr23_1 showing a 18kb duplication, as illustrated by the diagonal lines.

4.3 The X chromosome SNP location

The X chromosome was assembled in a single scaffold and appears to have 6 large structural differences compared to the *B. taurus* X chromosome (Fig. 17). It is important to note that these differences need to be validated using independent data to ensure their reliability. The X chromosome still contains the most gaps of any of the completed chromosomes, indicating regions of large repeats and low sequence coverage. Still, the continuality of the Brahman X chromosome is a substantial improvement on the old *B. taurus* X chromosome. When SNP were allocated to a position on the new Brahman X chromosome there were substantial rearrangements observed. This is a consequence of the structural variations observed between the Brahman and taurine X chromosome. If the structural differences are validated, this may indicate that it is particularly important to use a Brahman derived X chromosome for analyses which rely heavily on the position of SNP, such as imputation and genome wide association studies (GWAS).

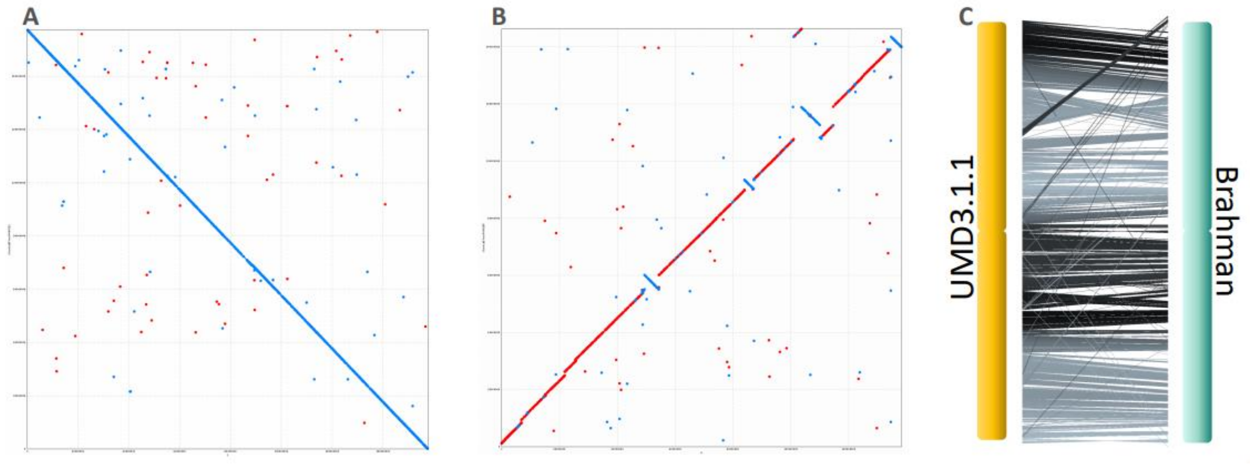


Fig. 17 Comparison of the B. taurus and Brahman genome assemblies for chromosome 1 (A) and chromosome X (B). C) The location of SNP on the X chromosome based on homology between the reference sequence and the molecular probe used.

4.4 Chromosome completeness

The assembled chromosomes are highly complete and contiguous (Table 3). At most, full length chromosomes are fully contained within three scaffolds, which were ordered and orientated to match the *B. taurus* assembly. Generally there is very high structural consensus between the *B. taurus* and Brahman genome assembly.

Additionally, when the scaffolds were examined for the telomere repeat sequence (TTAGGG)n, these were identified as being highly abundant at the end of chromosomes 1, 9, 13, 15, 18, 14, 20 and 27 (Fig. 18). In addition, 9 of the unplaced scaffolds were rich in the telomere repeat sequence and are likely not able to be non-ambiguously placed in the Brahman reference sequence due to their repetitive nature.

The full length mitochondrial genome (Fig. 19) was assembled and annotated using MITOS (Bernt *et al.* 2013). Comparison to publically available bovine mitochondrial genomes indicates the mitochondrial genome of the reference animal is of *B. taurus* origin, concordant with the use of *B. taurus* females being used to grade up the Brahman breed when it was first introduced to Australia.

Table 3. Length in basepairs of each Brahman genome, and if the entire chromosome is contained within a single scaffold. Where the chromosome is split into two or three scaffolds, the number of scaffolds is indicated.

Scaffold name	Length	Whole Chromosome
BrahChr1_1	157337387	Yes
BrahChr2_1	136163192	Yes
BrahChr3_1	121510754	Yes
BrahChr4_1	119929227	Yes
BrahChr5_1	40304107	No – 2 pieces
BrahChr5_2	80691544	-
BrahChr6_1	117939204	Yes
BrahChr7_1	110810751	Yes
BrahChr8_1	22485169	No – 3 pieces
BrahChr8_2	47827448	-
BrahChr8_3	42728095	-
BrahChr9_1	104812478	Yes
BrahChr10_1	23164039	No – 2 pieces
BrahChr10_2	79069930	-
BrahChr11_1	7309453	No – 3 pieces
BrahChr11_2	41948137	-
BrahChr11_3	57746229	-
BrahChr12_1	40949876	No – 3 pieces
BrahChr12_2	30957533	-
BrahChr12_3	14284756	-
BrahChr13_1	83583526	Yes
BrahChr14_1	13615691	No – 3 pieces
BrahChr14_2	44434581	-
BrahChr14_3	24281917	-
BrahChr15_1	84677148	Yes
BrahChr16_1	7099584	No – 2 pieces
BrahChr16_2	74095649	-
BrahChr17_1	73050433	Yes
BrahChr18_1	13389667	No – 3 pieces
BrahChr18_2	19269050	-
BrahChr18_3	34434199	-
BrahChr19_1	58402991	No – 2 pieces
BrahChr19_2	6469098	-
BrahChr20_1	71666168	Yes
BrahChr21_1	34749500	No – 2 pieces
BrahChr21_2	35757710	-
BrahChr22_1	61015590	Yes
BrahChr23_1	8146469	No – 3 pieces
BrahChr23_2	19253135	-
BrahChr23_3	26276138	-
BrahChr24_1	62280714	Yes
BrahChr25_1	42514956	Yes
BrahChr26_1	51727733	Yes
BrahChr27_1	45760196	Yes
BrahChr28_1	45536915	Yes
BrahChr29_1	21873311	No – 2 pieces
BrahChr29_2	30021527	-
BrahChrX_1	146503746	Yes

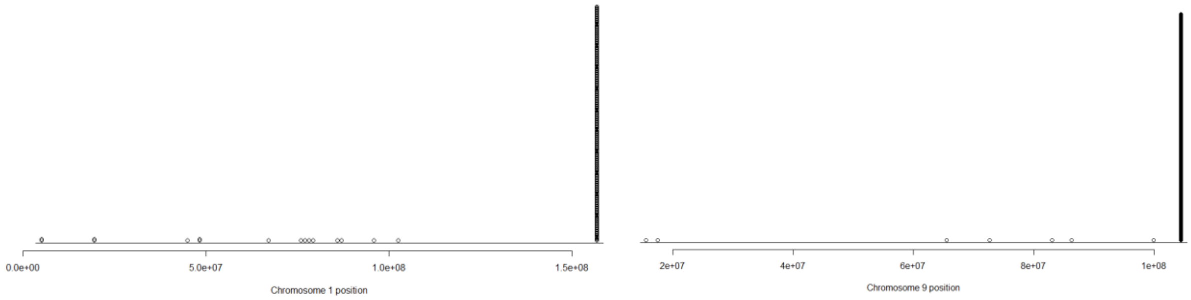


Fig. 18 Stacked dotplots of the locations of telomere sequence across scaffolds. Note that the telomere sequence is only observed in high quantities at the end of the scaffolds.

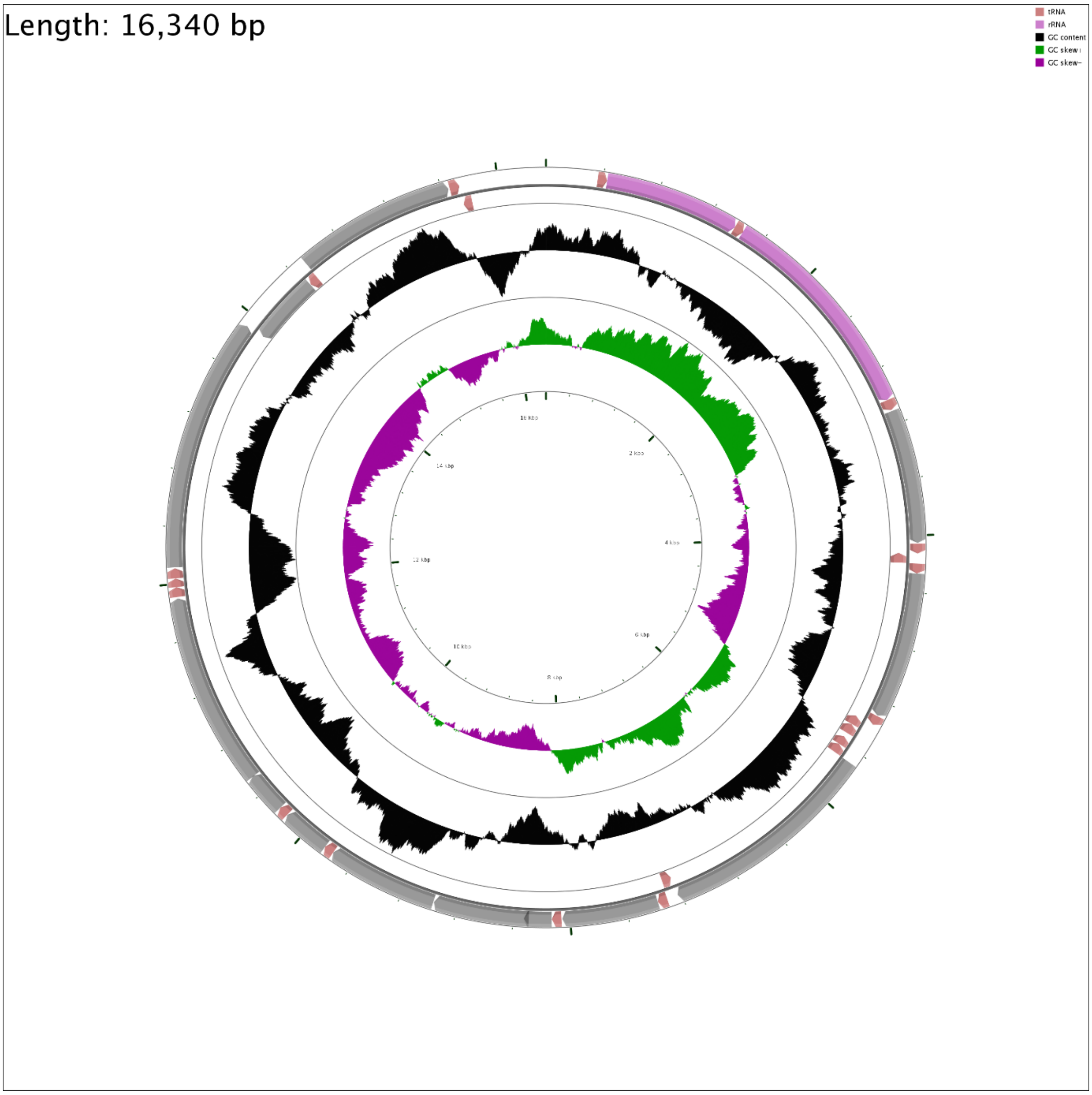


Fig. 19 The annotated mitochondrial genome of the reference Brahman. Image generated by CGView (Grant and Stothard 2008).

4.5 Bull sequence analysis

In total, 210 animals have been sequenced using short read technology (Fig. 20). This is in addition to other animals sequenced by other groups as part of the 1000 Bulls GenomesProject. Together, the sequenced animals give an overview of the genomic diversity of major north Australian beef breeds.

On average 95% of reads passed the QC procedure (Fig. 21). Many (45%) of those that passed had some removal of poor quality bases at the end, while the remaining 55% passed at full length. Of those that failed, the most common reason was that there was no section of the read that was longer than 50bp and of high quality (Fig. 22). Only a small number of reads were removed due to high levels of uncalled bases (N counts) or low average quality.

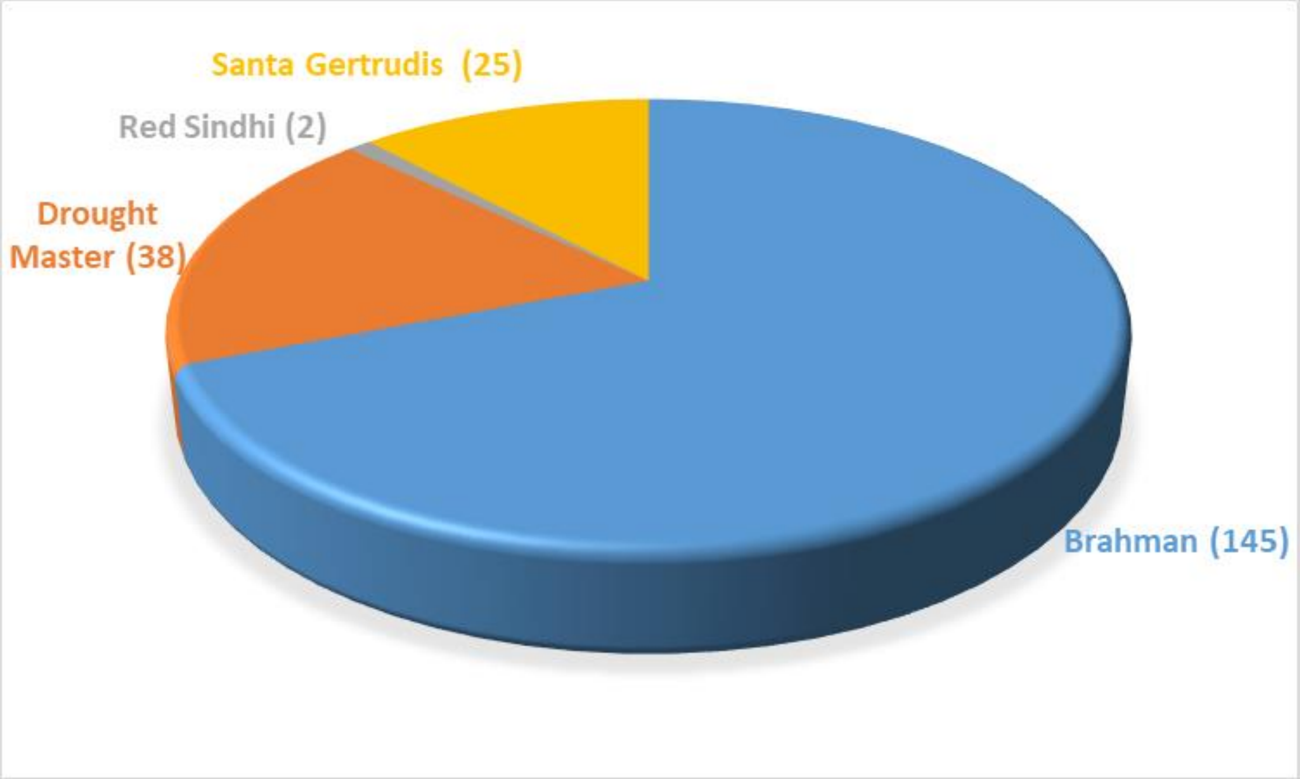


Fig. 20 Number of animal sequenced using short read sequence technology. Number in brackets following breed name indicates the number of animals.

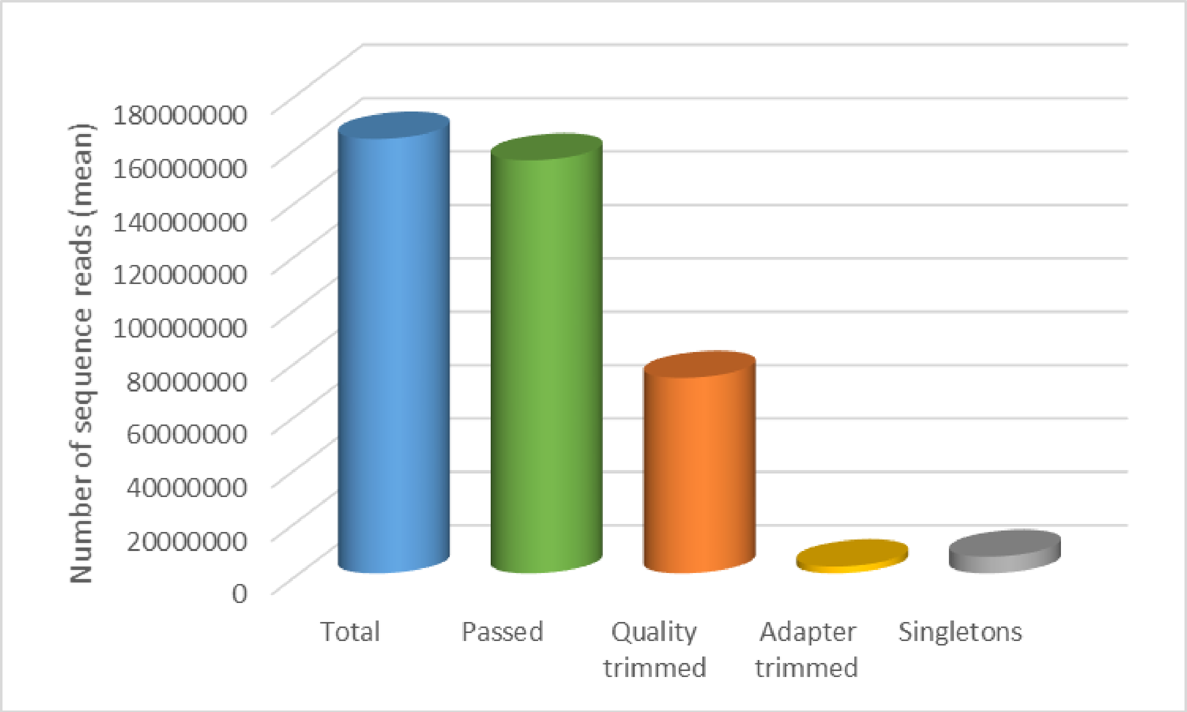


Fig. 21 Reads that passes QC for the 210 animals - quality controls applied.

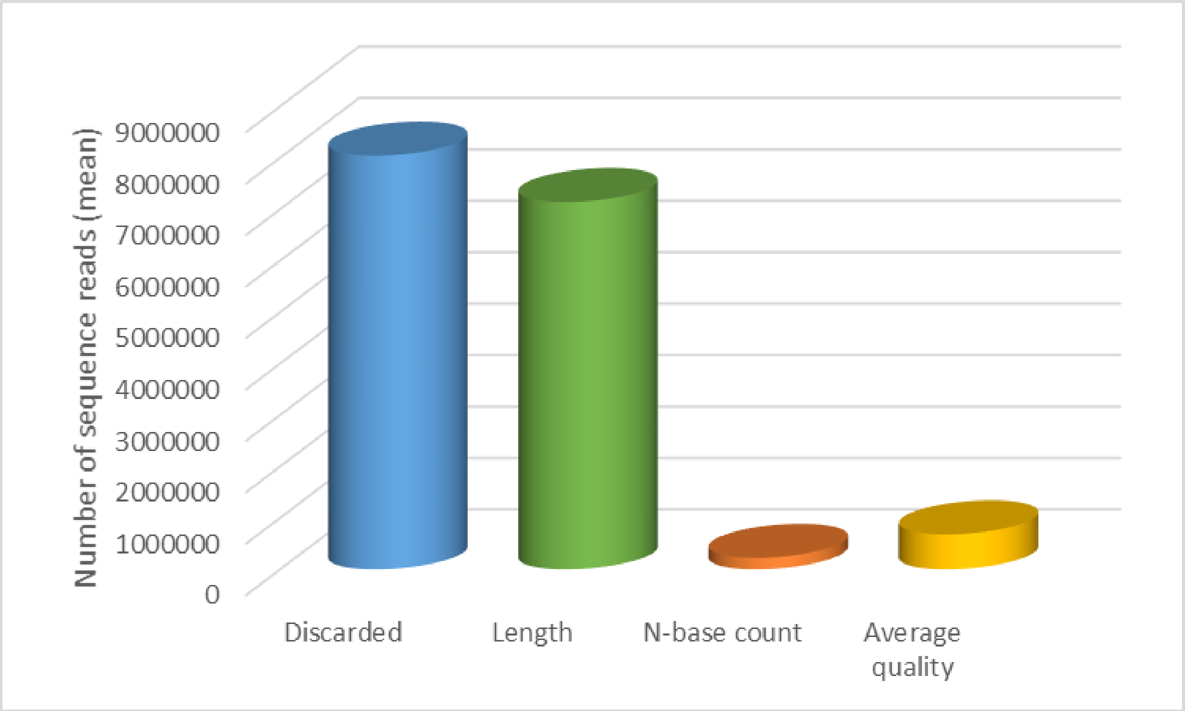


Fig. 22 Reads that were discarded in the quality control procedure.

4.5.1 Short read platform comparison

After the short reads were alignned to the reference genome, the percentage of reads aligning and the error rate sequencing platforms were compared (Fig. 23). Most of the aligned sequences aligned at a rate >90%. Alignment rates were not correlated with the error rate, suggesting that sufficient quality can be achieved on all platforms for mapping back to a reference genome. However, higher error rates may reduce the accuracy of variant calling and therefore, lower error rate platforms should be used where possible.

An assessment of the sequencing technology used was made. Animals have been sequenced on three platform types, the BGI sequencer, Illumina's HiSeq and Illumina's Novaseq. The Novaseq had the lowest error rate of the three platforms, followed by Illumina's HiSeq, then the BGI machine. This information can be used to inform future sequencing of beef animals, as the large capacity of the Novaseq allows for cost effective bull sequencing without compromising the quality of the data.

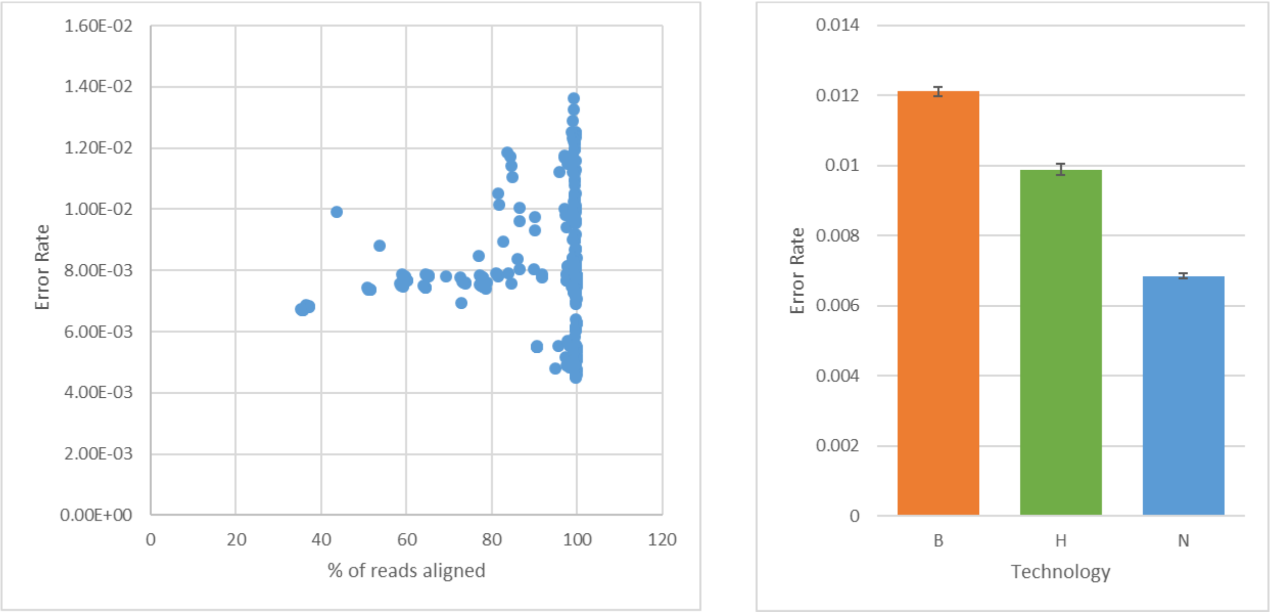


Fig. 23 A) Comparison of alignment rates and error rates in the aligned reads. B) Error rates across the three platforms tested: B- BGI, H- Illumina HiSeq, N- Illumina Novaseq. Error bars show standard error of the mean.

4.5.2 Variant calling

Genetic variants were identified across the genome based on the alignment of the short read data. Variants were filter based on presence in at least 2 animals. In total 51304846 variants were identified across the genome (Table 4). On chromosome 29, for example (Fig. 24), there were slightly over 1 million variants identified in at least 4 animals of the Brahman, Santa Gertaudis, Red Sindhi and Droughtmasters (the animals sequenced on BGI were not included in this analysis as the higher error rate may have influenced the identified variants). There is a decay in the number of SNP as the required number of animals in which the variant is observed is increased (Fig. 25), as is expected. This represents an increase in certainty of the legitimacy of the variant, as well as an increase in variant abundance.

Even when looking at variants that are highly abundant in the population of animals sequenced, there is still good coverage of variants across the chromosomes.

Table 4. Number of variants identified form the short read alignments that are present in at least 2, 5, 10, 50, 100 and 200 animals.

	≥2 Animals	**≥5 Animals**	**≥10 Animals**	**≥50 Animals**	**≥100 Animals**	**≥200 Animals**
Chromosome 1	3173527	2706753	2355556	1434137	836971	34507
Chromosome 2	2594569	2214468	1922985	1172644	698253	25564
Chromosome 3	2267048	1929378	1678496	1011796	594301	24498
Chromosome 4	2419653	2073300	1813672	1130567	677577	28117
Chromosome 5	2268460	1932401	1689033	1070979	675053	47188
Chromosome 6	2326191	1996825	1759146	1073357	630203	27952
Chromosome 7	2119470	1801048	1571753	957761	560332	21222
Chromosome 8	2133439	1816975	1583806	976580	591437	24837
Chromosome 9	2068866	1769616	1542612	930483	546512	20839
Chromosome 10	2077448	1759565	1528936	936386	564900	27737
Chromosome 11	1998235	1705767	1494555	908013	540687	26733
Chromosome 12	2001530	1676339	1438074	849018	494325	18659
Chromosome 13	1544723	1320228	1154290	717522	442121	22189
Chromosome 14	1567453	1337644	1166227	697872	407132	17493
Chromosome 15	1903521	1614238	1409674	859074	504956	19880
Chromosome 16	1585429	1346377	1171813	728727	437011	19969
Chromosome 17	1483408	1271997	1105568	650714	369060	15244
Chromosome 18	1299781	1103132	951827	568390	330342	14071
Chromosome 19	1183140	996753	872050	544459	340568	16403
Chromosome 20	1477910	1269812	1115452	679428	405900	16849
Chromosome 21	1429682	1214674	1048486	633895	372877	14378
Chromosome 22	1159684	995550	868757	526210	309300	13864
Chromosome 23	1322938	1116894	947260	543815	331218	14959
Chromosome 24	1267275	1094519	962838	594572	364017	14531
Chromosome 25	844868	718795	625038	382612	223341	10053
Chromosome 26	1058578	907588	786834	470996	270977	10782
Chromosome 27	1052738	914871	796833	474136	276610	10837
Chromosome 28	1025493	886770	780092	484662	276943	10126
Chromosome 29	1144228	976246	858075	528058	322411	14203
Chromosome X	1505561	1161745	959437	517729	277004	3407
Total	51304846	43630268	37959175	23054592	13672339	587091

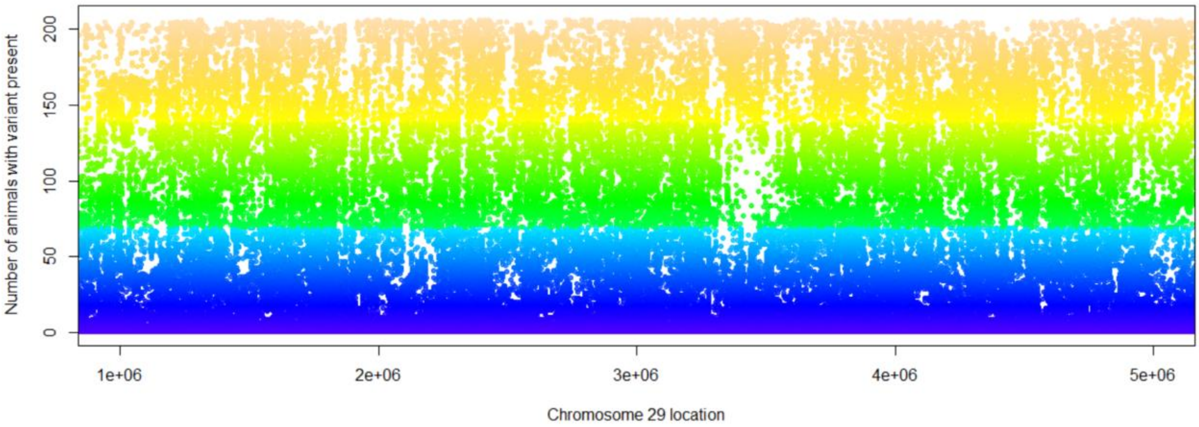


Fig. 24 Position of variants in approximately 10% of chromosome 29.

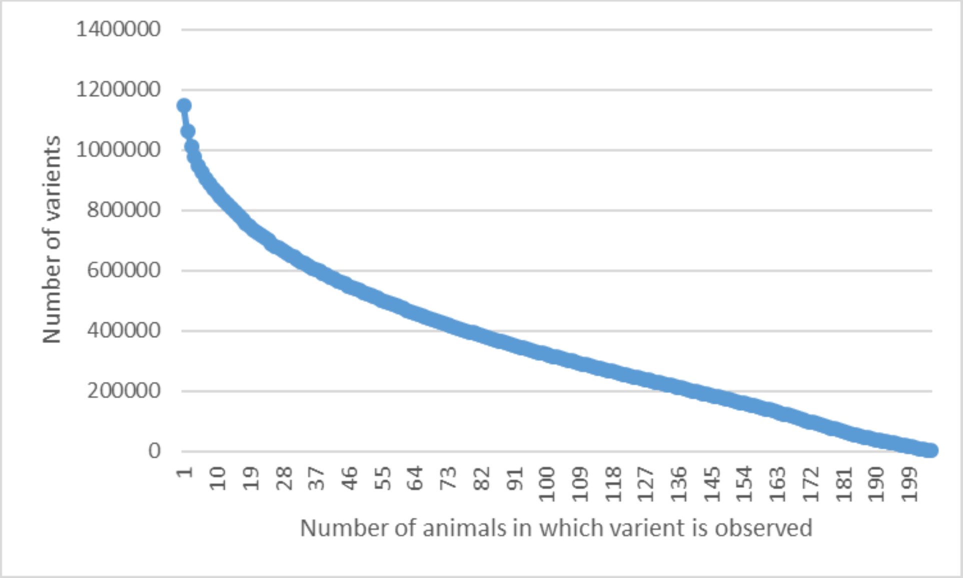


Fig. 25 Identified variants compared to the number of animals (between 2 and 206) that the variant was observed in. Chromosome 29 is shown.

4.5.3 SNP Annotation and origin

The following analysis section has been published and should be cited as Koufariotis, L., et al. (2018). "Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled." Scientific Reports 8(1): 17761.

Although the number of bulls sequenced is only a small representation of the Australian Brahman breed, the animals were chosen to capture the greatest possible amount of genetic variation (Fig. 26). The first 46 bulls selected captured 17% of the genetic variation represented in the pedigree. Variants that were shared or unique between Brahman, *B. taurus* and Gir were examined. Of the 36.1 million filtered Brahman variants, 15.9 million variants were uniquely found in Brahman (Fig. 27). The origin of these variants is possibly one of the other three indicine foundation breeds

(Ongole, Guzerat and Krishna). Brahman and *B. taurus* shared 10.7 million variants that were not present in Gir. There were only 9 million SNP in the Gir dataset, reflecting the small number of Gir animals sequenced, however, 95% of variants from Gir were also identified in the Brahman variants. Overall, a total of 6.7 million variants (18.5% of Brahman variants) were in common between all three datasets.

To identify regions of the genome in each animal that were indicine or taurine in origin, we calculated bosind_250 values (based on the b values (Bolormaa *et al.* 2011a) bosind_250 statistic is an estimate of comparing two populations based on their sequence, allele frequencies and SNP calls to determine how differentiated the two populations are). The global bosind_250 value (across all SNP) for the Brahmans was substantially lower for *B. indicus* (Gir, 0.188) than for *B. taurus* (0.264), reflecting the high indicine content of Brahman genomes. There was substantial variation between the Brahman animals regarding the proportion of the genome that was indicine derived, ranging from 0.26/0.24 indicine/taurine composition to 0.27/0.17 indicine/taurine composition. On average 8.94% of the Brahman genome was taurine derived, slightly lower than the previous estimate of 10% (Bolormaa *et al.* 2011a), possibly due to the inclusion of several USA derived sires.

When the percentage of taurine/indicine was calculated across the chromosomes, only a few segments showed strong *B. taurus* introgression. Chromosomes 8, 12, 14, 23, 26 and 29 all show regions of strong *B. taurus* introgression, with the remainder of the genome largely of *B. indicus* origin.

Genes in the indicine and taurine regions were examined. In the regions of nearly fixed *B. indicus* origin there were 1,609 unique genes. These included a significant enrichment for genes involved in intermediate filament, hormone, protein biosynthesis, cytoplasm and phosphoproteins. We found 61 genes to be enriched for protein biosynthesis, and could be associated with lower protein turnover.

To determine if the level of enrichment for the above functions could occur by chance, we performed a permutation test (randomly selecting 1,609 genes 100 times, the equivalent to the 5% most significant genes in the *B. indicus* regions). Functional annotation of each permutation was performed with DAVID. The keywords for each permutation were recorded and grouped together to create a word cloud plot of the most commonly occurring keywords. Over the 100 randomly selected gene permutations, protein biosynthesis, 4Fe-4S nucleotidyltransferase and biological rhythms were never found in the 100 random gene permutations gene set, which indicates statistical significance (Fig. 28).

Individual regions that were enriched for *B. indicus* content include chromosome 21 positions 7.5–10 Mb with a 6-fold difference in bosind_250 values between *B. taurus* and *B. indicus* (bosind_250 value 0.67 and 0.11 for *B. taurus* and *B. indicus,* respectively). One gene of interest, melanoma antigen family L2 (MAGEL2), is implicated in fertility and evidence indicates the gene may influence testicular size, fertility and growth (Utsunomiya *et al.* 2014). MAGEL2 is also implemented in circadian rhythms and adaptation to new environments in mice (Fountain *et al.* 2017).

The region between 53–55 Mb on chromosome 8 is of strong *B. indicus* origin and contains the genes PRUNE2, GNA14, GNAQ, CEP78 and PSAT1. In Simmental cattle, an association analysis with foreshank weight found a region that includes the genes GNAQ and CEP7824 (Wu *et al.* 2014).

On chromosome 14 at positions 24,250,000-24,500,000 bp, the XK, Kell blood group complex subunit-related family, member 4 (XKR4) gene is found approximately 396 kb upstream of the PLAG1 gene (Fig. 29). The gene is of interest as it has been observed to be associated with subcutaneous rump fat thickness (Bastin *et al.* 2014). In addition, polymorphisms in this region have been found to be associated with rump fat thickness (Porto Neto *et al.* 2012) residual feed intake, and average daily feed intake (Bolormaa *et al.* 2011b; Lindholm-Perry *et al.* 2012) in cattle.

Annotation of fixed variants show relatively similar proportions to the annotations of all genomic Brahman variants (67.32% intergenic and 25.35% intronic). However, there was an increased proportion of missense mutations in SNP that were fixed for the alternative allele in Brahmans versus *B. taurus*. 0.86% of Brahman SNP that are fixed for the alternative allele are annotated as missense, and this value varies from the proportion of Brahman genomic SNPs annotated as missense, which is 0.35% (Table 5). This difference in the proportion of missense annotated SNP represents an 85% increase in the percent difference between fixed variants for the alternative allele that are missense, compared to all missense variants in the WGS dataset (*P* = 0.009).

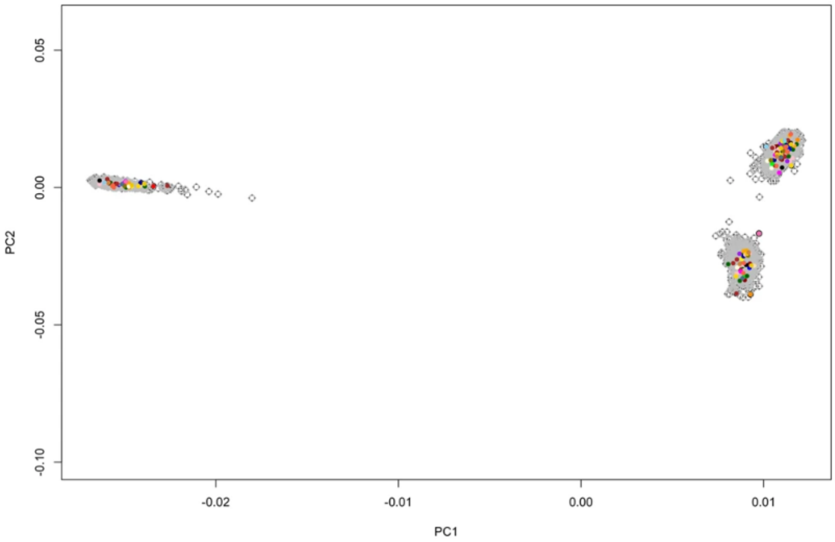


Fig. 26 Principle component analysis of the genomic relationship between studs. Forty-six bulls selected for sequencing (in colour) relative to the diversity of the Brahman population represented by a large sample of genotyped animals with 24 K SNP (1021 animals in grey). PC1 and PC2 are the first and second principle components of genomic relationship matrix among the genotyped animals.

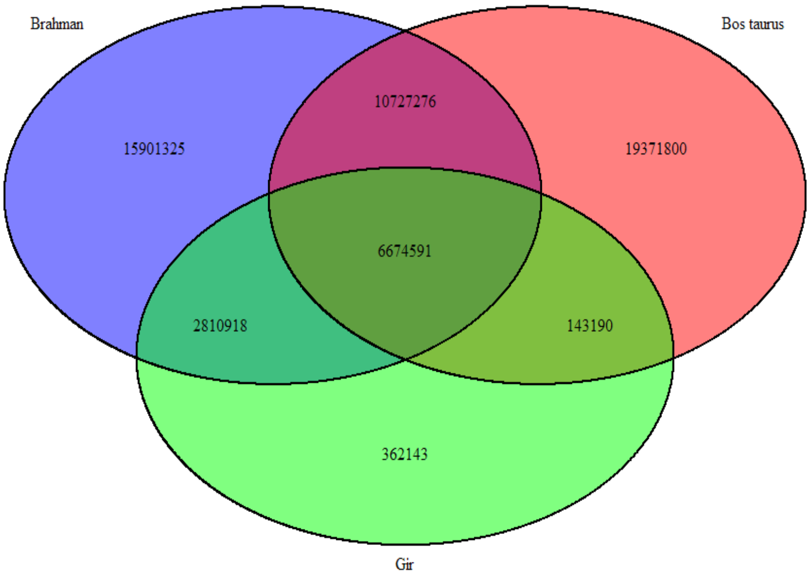


Fig. 27 Common variants between the cattle breeds. The blue circle represents the Brahman variants, red represents B. taurus and green represents all Gir variants. We find that overall, most of the Gir variants are shared with Brahman breeds and only a small proportion of variants are common between all 3 breeds.

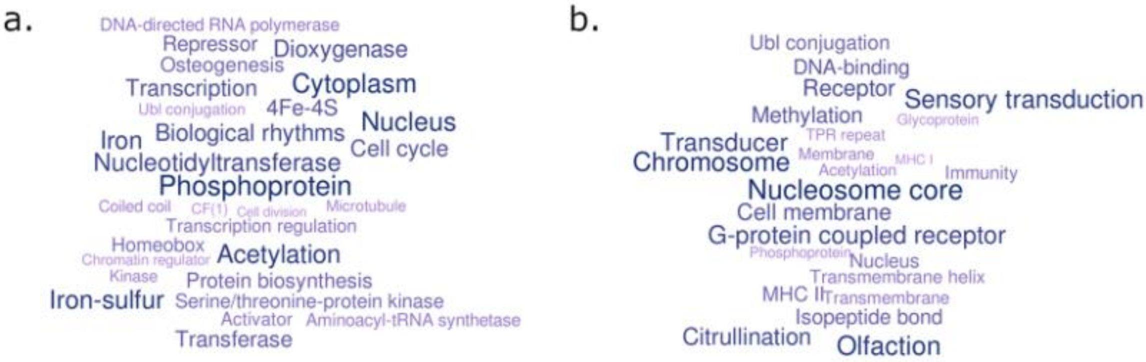


Fig. 28 *Word cloud showing the keywords after functional categorising genes using DAVID. Larger keywords represent those that show the strongest associations for that set of genes (based on the DAVID p-value). (a) Functional categorised genes that are in genomic segments that are homozygous or nearly homozygous in the sequenced bulls and of Bos indicus origin (b). Functional categorised genes located in genomic segments that show strong Bos taurus introgression.*

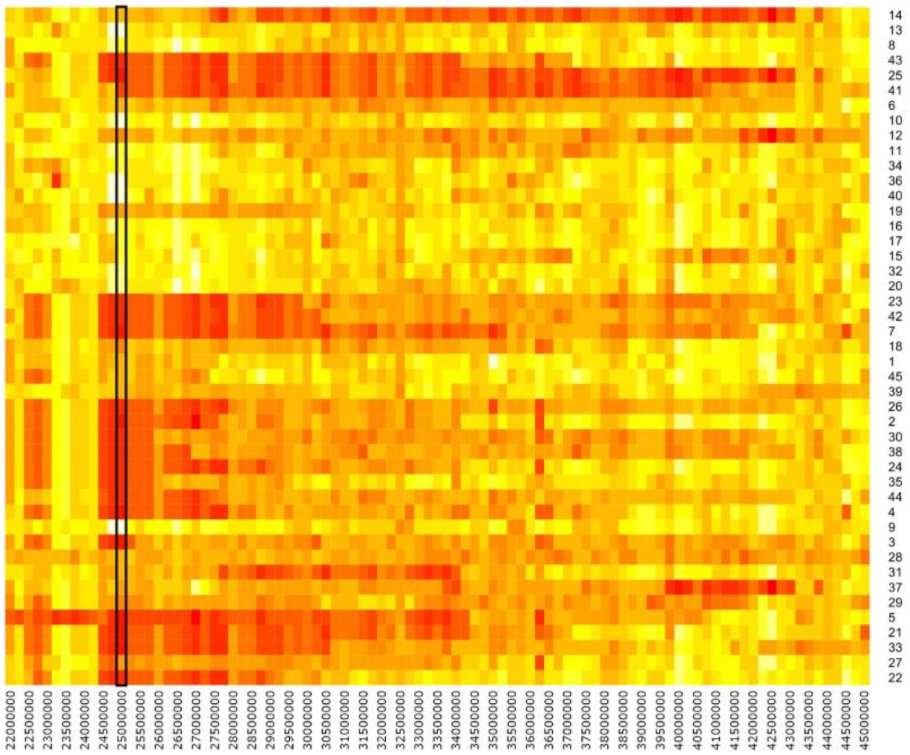
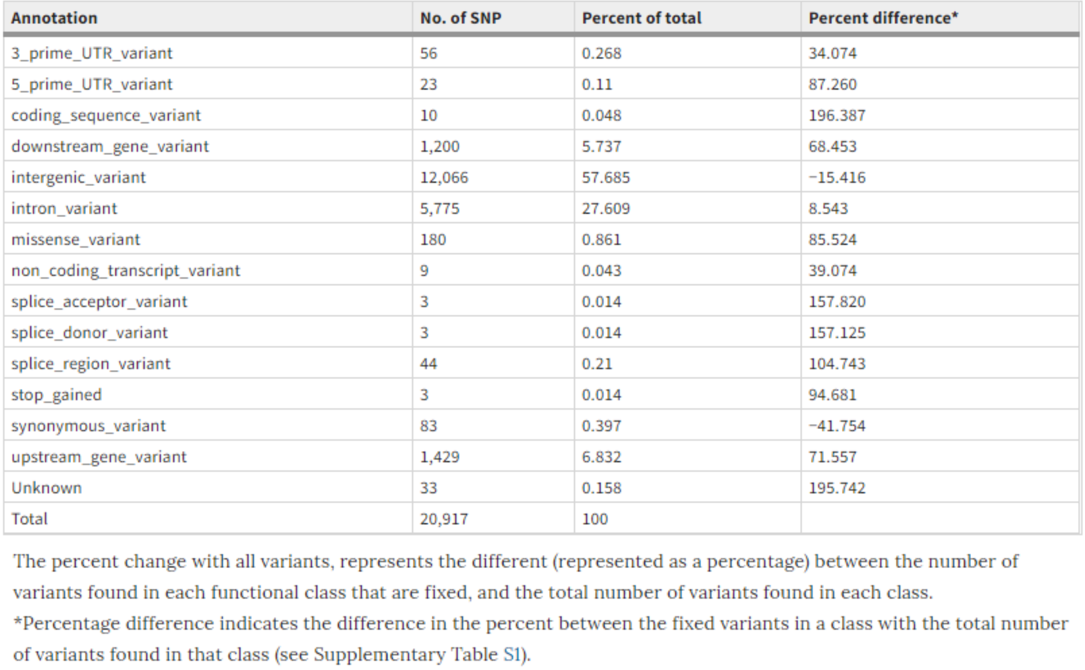


Fig. 29 *Heatmap of the chromosome 14 region with high Bos taurus introgression. This includes the PLAG1 gene found at positions 25,007,291–25,009,296 as indicated by the black box. On the y-axis are all the sequenced animals, ordered by year of birth (with older animals at the top). On the x-axis is the genomic positions, incremented in 250 kb windows. A red colour indicates more Bos taurus introgression, a white colour indicates more regions that are of Bos indicus origin.*

Table 5 Annotation information of the variants that are fixed in Brahmans for the alternative allele.



Annotation	No. of SNP	Percent of total	Percent difference*
3_prime_UTR_variant	56	0.268	34.074
5_prime_UTR_variant	23	0.11	87.260
coding_sequence_variant	10	0.048	196.387
downstream_gene_variant	1,200	5.737	68.453
intergenic_variant	12,066	57.685	−15.416
intron_variant	5,775	27.609	8.543
missense_variant	180	0.861	85.524
non_coding_transcript_variant	9	0.043	39.074
splice_acceptor_variant	3	0.014	157.820
splice_donor_variant	3	0.014	157.125
splice_region_variant	44	0.21	104.743
stop_gained	3	0.014	94.681
synonymous_variant	83	0.397	−41.754
upstream_gene_variant	1,429	6.832	71.557
Unknown	33	0.158	195.742
Total	20,917	100	

The percent change with all variants, represents the different (represented as a percentage) between the number of variants found in each functional class that are fixed, and the total number of variants found in each class.

*Percentage difference indicates the difference in the percent between the fixed variants in a class with the total number of variants found in that class (see Supplementary Table S1).

4.5.4 Brahman polled mutation

The following analysis section has been published and should be cited as Koufariotis, L., et al. (2018). "Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled." Scientific Reports 8(1): 17761.

The polled phenotype (naturally hornless) is an important trait for the northern beef industry due to its effect on animal welfare, handling and meat quality. Animals that are horned are mostly dehorned to reduce damage caused to other animals and the stockmen. However, the dehorning process is both painful and can cause a delay in weight gain.

Four Brahmans that were polled, identified through progeny testing, were sequenced to allow investigation into the poll allele in Brahman cattle. One animal (referred to as animal 47) was heterozygous for the polled allele (*Pp*) and the other three animals (referred to as animals 48, 49, 50) were homozygous for the polled allele (*PP*). The three *PP* animals were compared to the 46 other Brahmans.

The polled allele has been located on chromosome 1 in the positions 1.6–2.2 Mb in cattle (Georges *et al.* 1993; Brenneman *et al.* 1996; Seichter *et al.* 2012). Three putative mutations have been reported. The first polled mutation is a 80kb duplication in the region on 1.91–1.99 Mb that is found in Friesian cattle (Medugorac *et al.* 2012). The second polled mutation is found starting at the position 1.71 Mb and is believed to be of Celtic origin, described as a 212bp duplication in polled animals that replaces a 10bp sequence at the positions 1,706,051–1,706,060 bp (Wiedemar *et al.* 2014). This mutation is referred to as the Celtic mutation. The third polled mutation originates from Mongolian Turano cattle and was recently found to be introgressed in Mongolian Yaks (Medugorac *et al.* 2017). As with the other two polled mutations, this mutation is found on chromosome 1 between positions 1,889,854 bp and 2,010,574 bp. The architecture of this mutation is an 11-bp motif (conserved among Bovidae and duplicated in the Friesian mutation) due to a rearrangement of a 219bp duplication-insertion, a 7bp deletion and a 6bp insertion (Medugorac *et al.* 2017). This third polled mutation was not observed to be segregating in the Brahman animals, and it is unlikely to be segregating as it has only ever been described in Mongolian Turano cattle and Mongolian yaks.

In Brahmans, a single study has mapped the polled locus in 68 genotyped Brahman cattle to the same region as in *B. taurus* cattle (Mariasegaram *et al.* 2012), yet the origin of the Brahman polled mutation remains unknown.

Examining the fixed-size windows in the polled locus region shows slight introgression of *B. taurus* (see data availability, Chromosome 1). Using our sequence data, the total number of reads aligning to the reference were counted in 50bp incrementing windows in and around the polled region and the standard deviation was calculated for each of the 50bp window. We examined the coverage of the reads as an increase in the coverage can indicate a duplication mutation. In this study, the region around 1.65Mb and 1.90Mb stands out as an increase in coverage for the three *PP* animals compared to the 46 original Brahman bulls, indicating the presence of structural variants, such as duplications, in the *PP* animals.

In the positions where the Friesian mutation is on chromosome 1, we find no distinct difference or patterns in read coverage across the entire 80kb region between the three *PP* animals and the 46 original animals (Fig. 30). If the Friesian 80 Kb duplication was present in Brahman, we would expect to find at least a two-fold increase in the coverage where the mutation exists between the homozygous polled and the 46 original Brahmans. Thus, in this study, we did not find evidence that the Friesian mutation is segregating in the three *PP* polled Brahman animals.

At the positions on chromosome 1 where the Celtic mutation is predicted to be, we find that the 3 *PP* animals show an increase in coverage, when compared to the 46 other Brahman (Fig. 31). A tentative duplication is found at this region, approximately where the Celtic mutation is described, as the coverage for the *PP* animals is increased significantly.

Furthermore, the alignment files containing the mapped reads for chromosome 1 were visualized using the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir *et al.* 2013) tool concentrating on the region where the Celtic mutation is found. We find in the three *PP* animals, there is a region between the positions of 1,706,045–1,706,060bp which shows a 10bp deletion event, consistent with the Celtic mutation. This deletion can be unambiguously seen in two of the polled animals (48 and 50), with minimal reads mapping to this region, especially for these two polled animals. Furthermore, the *PP* animal 49 seems to be carrying some variant of the polled mutation, as few reads map to the 10bp deletion. This can possibly indicate that while this animal is homozygous for the polled mutation, the mutation affecting the poll phenotype might not be at the 10bp region where the deletion is located. This needs to be followed up with more research and will provide the basis for future work.

In the 46 original animals that had their genomes sequenced, most do not display this deletion, indicating *pp (*horned) animals with a small chance of *Pp* animals.

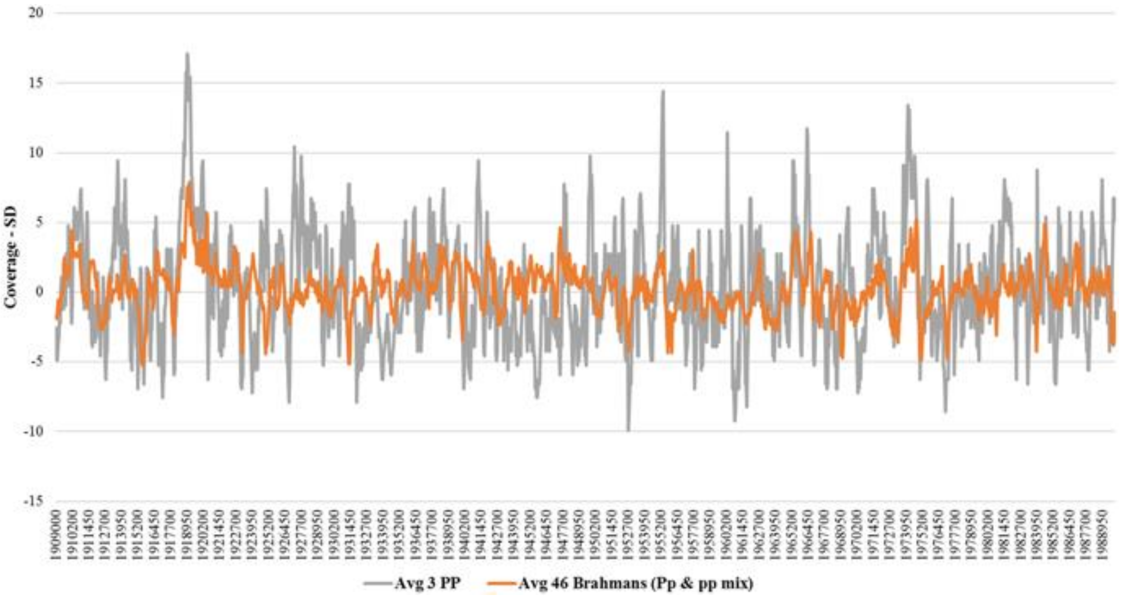


Fig. 30 Holstein-Friesian 80kb polled variation. Figure showing the read coverage (expressed as standard deviations from the mean) between the three homozygous polled Brahmans and 46 horned Brahman in the region where the 80kb Holstein-Friesian polled duplication is found.

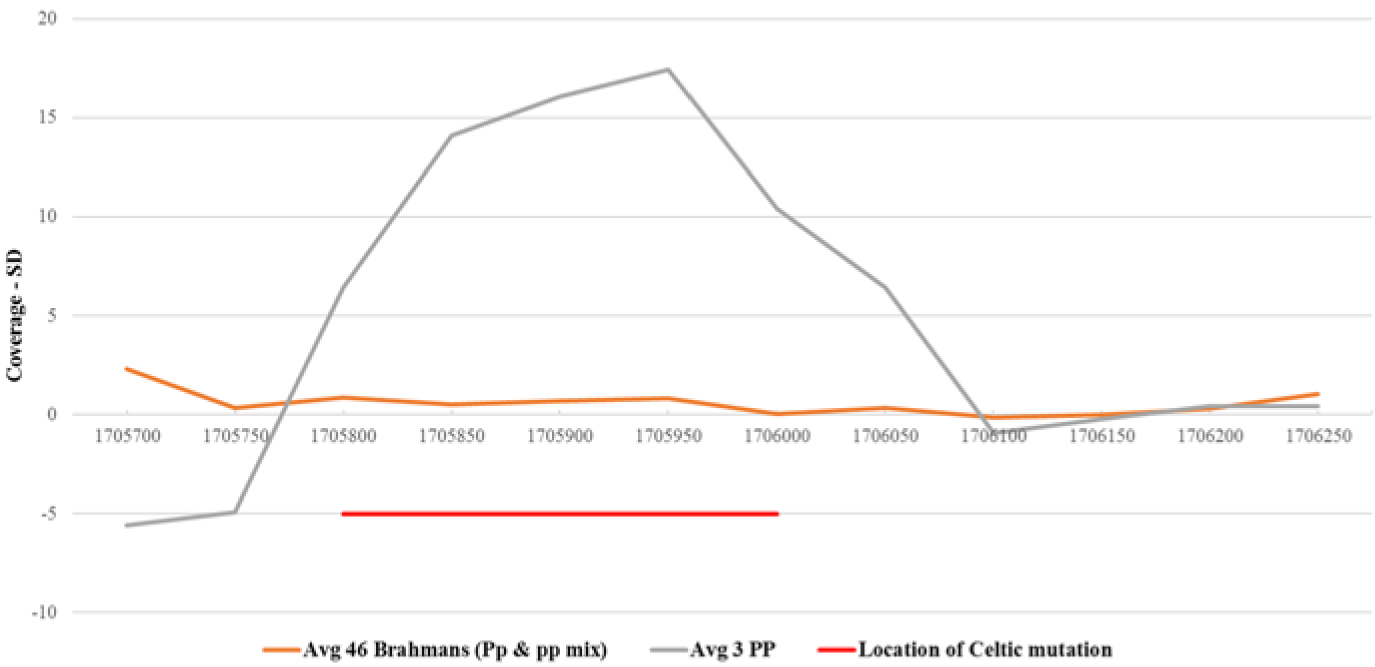


Fig. 31 Plot of the number of reads mapping to the reference homed in on where the Celtic mutation has been described. In the 3 PP animals, we see an increase in the coverage, characteristic of a duplication event, while the 46 original animals show consistent coverage. Here the orange line represents the standard deviation in the coverage for the original 46 animals. The grey line represents the standard deviation in the coverage for the 3 PP animals. The red line represents the location of the Celtic mutation as described by Wiedemar et al. 2014.

4.6 Structural variants

Long read alignment files were generated from Oxford Nanopore sequencing of the reference animal using minimap2 (Li 2018). The alignment files were used to examine structural variants (SV) within the reference genome. The C+ package Svim (version 1.4.3) was used to examine the reads (Heller and Vingron 2019). All default parameters were used except for the minimum mapping quality and minimum SV size. These parameters were changed to account for the increased error rate, as well as the Indel error profile of Nanopore sequencing. The minimum MapQ score was set to 30 to exclude reads which map with less than 99.9% certainty. The minimum SV size was increased from 10bp to 50bp to account for false positives reported as a result of the indel error profile. Structural variants on unplaced scaffolds are not reported at these are likely to be highly repetitive and therefore not informative due to ambiguous mapping of reads. The long read data used averaged 3X fold coverage – therefore structural variants with less than 2 or more than 12 supporting reads were filtered out.

A total of 27,482 insertion/deletions passed the set criteria (Fig. 32; Fig. 33), with variants identified across all 30 of the Brahman chromosomes. Most (87.4%) of the structural variants identified were less than 500 bases in length, however some were very long, the largest being a 79841bp deletion on the X chromosome. Examination of the location of the structural variants indicates there is an increased number near the ends of the autosomes, while the X chromosome does not show the same pattern.

It is important to note that these structural variants are only identified in one animal and at a low coverage. Studies on other cattle breeds have found more structural variants, but have used many more animals and/or deeper coverage (Boussaha *et al.* 2015; Couldrey *et al.* 2017). To our knowledge this is the first time large structural variants have been characterised at the whole genome level in Brahman cattle using long read data.

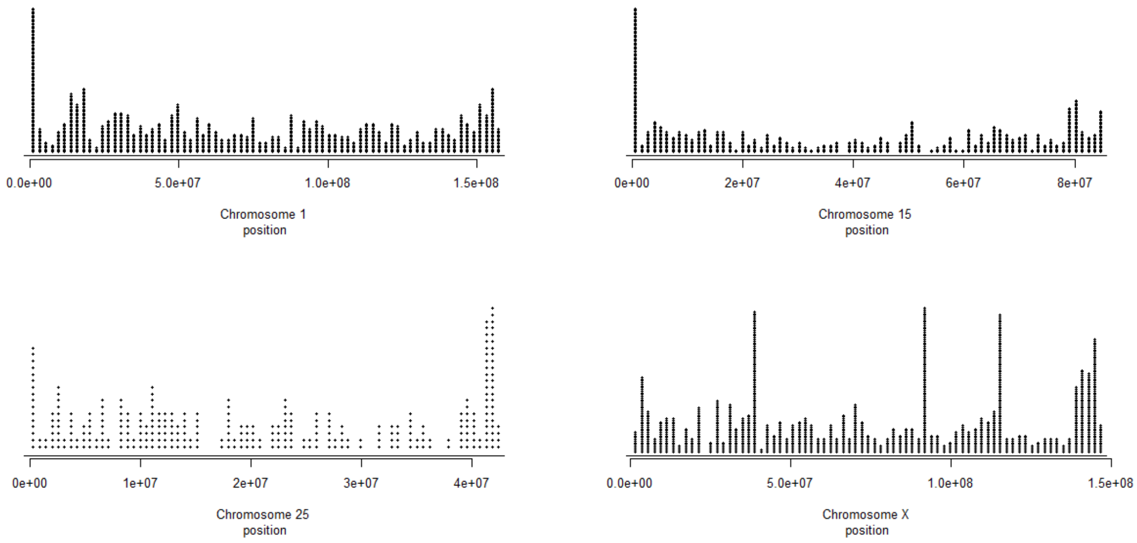


Fig. 32 The location of structural variants identified across autosomes (chromosomes 1, 15 and 25 only are shown) and the X chromosome. Stacked dot plots show the distribution of values in a data set, here they illustrate areas on the chromosomes where many structural variants are identified.

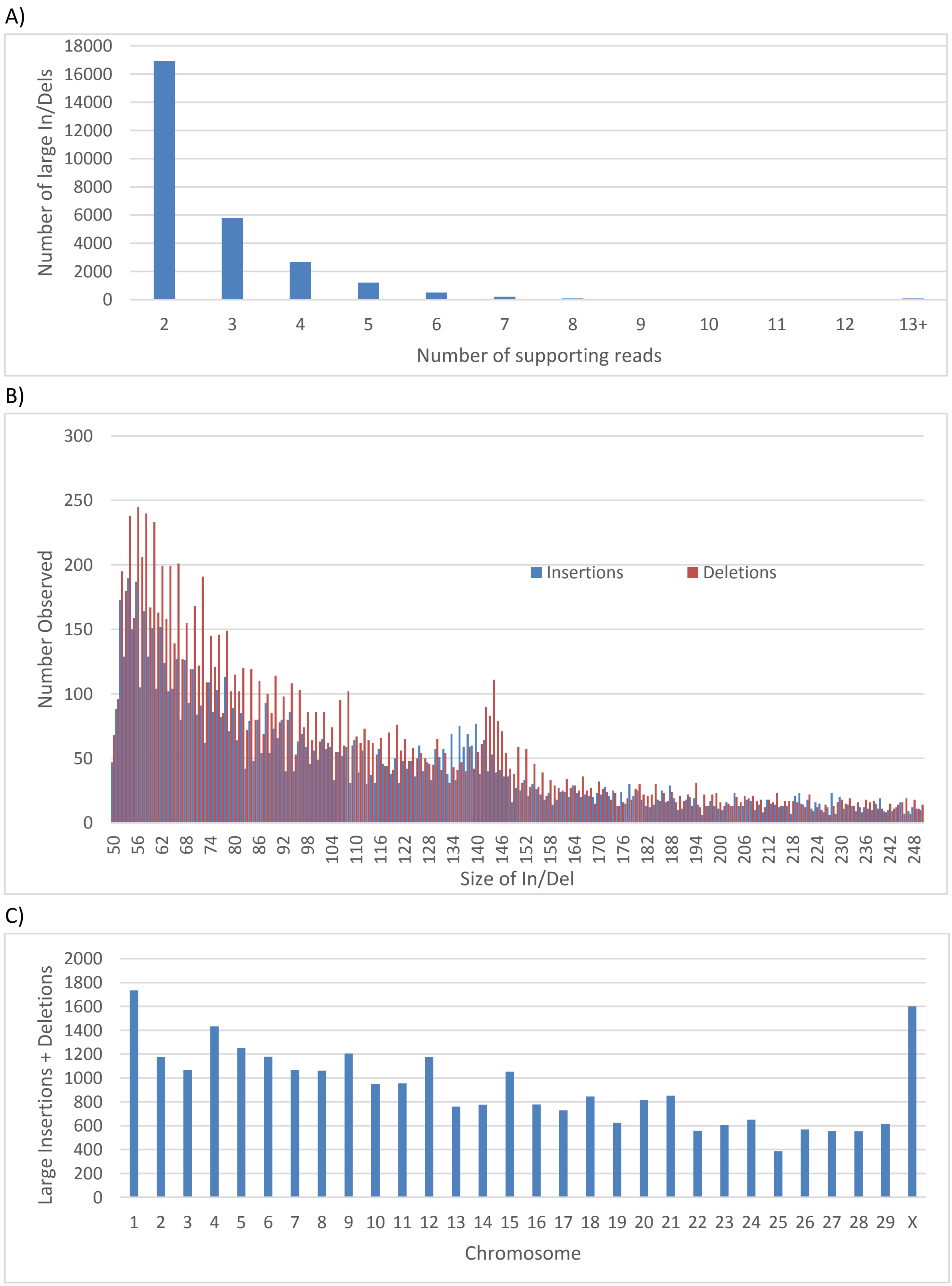


Fig. 33 The number of structural variants (insertions of deletions) of 50bp or greater with different support. Higher numbers of supporting reads suggest more confidence in the structural variant, however extreme depth indicates the presence of a repetitive element likely causing an erroneous variant call. B) Size of structural variants identified. Only those between 50 and 250bp in length are shown.)   C) Number of structural varients on each chromosome.

5 Discussion

5.1 Reference Genome

The Brahman genome that has been assembled for this project is high quality and is now available for public use. The availability of a high quality Australian Brahman genome provides a valuable tool for northern beef research. The limits of having to base genomic investigations of Brahman and Brahman crossbreed cattle on a *B. taurus* reference sequence has now been resolved.

Gap filling reference sequences is a time consuming process. While most gaps can be filled automatically, there are a small number that remain. It appears that the gaps remaining in the Brahman genome assembly are mostly the result of highly repetitive regions in the reference animal. This means that the solution to closing the remaining gaps is not to simply increase the depth of sequencing, but rather to focus on producing long reads that can span the repeat region. It is also possible that these same regions may not be as highly repetitive in other Brahman individuals and so sequencing of more Brahman cattle may lead to the closer of the remaining gaps.

Structural variation is observed in the Brahman X chromosome compared to the *B. taurus* X chromosome. These variations require validation, however, if they are a true reflection of the underlying genome the variation between taurine and indicine will have an effect on the imputation of genotypes. This is because if SNP are not correctly ordered in the reference then any SNP that are imputed may be incorrectly called.

5.2 Whole genome sequencing of Bulls

Brahman cattle are genetically distinct from taurine cattle. Previous studies have used genetic variants that were, for the most part, originally identified in taurine cattle. The specific variants segregating in the Brahman population were largely unknown. The variants discovered here can now be integrated into genomic studies of northern beef cattle. For example, this information will be particularly informative where genotypes are imputed to full sequence level from lower density SNP-chip data. Variants that were not previously characterised were not able to be imputed, and thus were not included in GWAS analyses. It is possible that this set of newly identified SNP will include the causal mutations for important production traits. Future studies will be able to include Brahman specific variants in their datasets.

Methods for the extraction of DNA from semen samples and appropriate sequencing methods were also identified. The included extraction protocol combined with the low input Nextera library preparation will allow the genomes of genetically important bulls can be sequenced in the future, even if the semen has been stored for a long time.

Of particular interest was the identification of the Celtic allele in the Brahman population. While the Friesian allele was not identified in this dataset, its absence from the population cannot be confirmed as it may be present in animals that have not been sequenced. Continuing to sequence Brahman bulls that are important to the industry, and in particular polled animals, will go some way to confirming the presence or absence of other polled alleles.

5.3 Milestone summary

All milestones have been met.

1	DNA from reference animal prepared	15/06/17	**ACHIEVED June 2017**
2	DNA from key ancestors prepared PacBio sequencing of reference animal complete	15/12/17	DNA from key ancestors prepared – **ACHIEVED Feb 2018** PacBio Sequencing Of Reference Animal Complete – **ACHIEVED Nov 2017**
3	Short read sequence from key ancestors complete Dovetail long range mapping complete	15/06/18	Short Read Sequencing – **ACHIEVED Jul 2018** Dovetail long range mapping **ACHIEVED Feb 2018**
4	Reference Sequence complete	15/08/18	**ACHIEVED Aug 2018**
5	All short reads mapped on to reference. All genetic variants in data set identified. All identified trait associated variants mapped onto new reference	*new milestone date* 15/11/18	**ACHIEVED Nov 2018**
6	Final Project Report Submitted	*New report date as requested by MLA and extended due to staff absence* 30/8/18	**COMPLETE Aug 2019**

5.4 Objectives achieved

Objective: A high definition and accurate assembly of the Australian Brahman breed including all the variation within and between Brahman and B. taurus breeds. This will include short range variation such as SNPs, insertions and deletions and critically the long range structural variation between the two cattle subspecies.

A high quality assembly of a Brahman cow has been produced. Variation in the Brahman breed has been identified using short read data, and an analysis compared with the origin of the variation to *B.*

taurus and to the founder breed Gir. We did not find chromosome level variation between breeds with the exception of the X chromosome, where several rearrangements were observed.

Objective: Input to content of future SNP chip or Genotype by Sequence technologies to improve prediction accuracies and reduce assay costs.

The variant data has been included in the 1000 Bull Genomes Project dataset and is now available for use in chip design as well as for imputation. This allows the data to be used in predictions and other genetic analysis.

Objective: Enhanced identification of causal or SNPs closely linked to causal mutations in the Australian Brahman breed enabling higher prediction accuracy in Genomic Selection.

The identification of variants in Brahman cattle has been completed and these variants can now be used to fine map GWAS to identify causal mutations, as well as for use in genomic selection.

Objective: Ability to identify B. taurus and B. indicus genome content in cross bred or composite cattle enabling the tracking of key genomic elements responsible for traits such as tropical adaptation, fertility, meat quality etc.

The *B. taurus* and *B. indicus* content of the genome was examined and regions for taurine descent identified. Studies on tropical adaptation, fertility and meat quality in Brahman and Brahman-crossbred cattle will use this information to interpret results in the context of tracking the origin of causative variation.

Objective: Training of the next generation of researchers and professionals that will have the capacity to deliver large research projects that impact on productivity of beef production in Northern Australia.

This project was largely completed by early career researchers (Dr Ross, Dr Koufariotis, Dr Nguyen). This has allowed them to deepen their knowledge of the beef production in northern Australia and provided them with experience delivering large research projects.

6 Conclusions/recommendations

6.1 Outputs

The Brahman genome is complete and freely available for use in research and industry.

Variants from the sequenced bulls have been included in the 1000 Bull Genomes Project dataset for use in imputation and other studies

6.2 Future Options

Additional long read data may be able to close the remaining gaps in the Brahman genome.

Imputation of variants using the Brahman genome should be compared to the taurine genome.

While the genome is annotated computationally using the UMD3.1.1 annotations, it would be advantageous to identify gene expression levels specific to Brahman cattle, as these can inform GWAS studies on which genes are highly expressed for traits in a given environment. Such information could identify selection targets for northern beef cattle.

7 Key messages

A reference assembly of the Brahman genome is now available for use. Additionally, whole genome sequence of over 200 Brahman and related breeds have been used to identify variants important for the northern beef industry. These resources should be used to increase the accuracy and relevance of cattle research to the northern beef industry.

7.1 Project contributions

This project was completed with contributions From Dr. Elizabeth Ross; Prof Stephen Moore; Prof Ben Hayes; Dr Lambros Koufariotis; Dr Loan Nguyen; Dr Russel Lyons and the late Dr. Brian Burns. The knowledge and passion that Dr. Burns had for the Northern Beef Industry will be greatly missed in the research community and the broader industry.

8 Bibliography

Acinas, SG, Sarma-Rupavtarm, R, Klepac-Ceraj, V, Polz, MF (2005) PCR-induced sequence artifacts and bias: Insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Journal of Applied and Environmental Microbiology* **71**, 8966-8969.

Aird, D, Ross, MG, Chen, WS, Danielsson, M, Fennell, T, Russ, C, Jaffe, DB, Nusbaum, C, Gnirke, A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, R18.

Bahassi, E, Stambrook, PJ (2014) Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis* **29**, 303-310.

Bastin, B, Houser, A, Bagley, C, Ely, K, Payton, R, Saxton, A, Schrick, F, Waller, J, Kojima, C (2014) A polymorphism in XKR 4 is significantly associated with serum prolactin concentrations in beef cows grazing tall fescue. *Animal genetics* **45**, 439-441.

Bernt, M, Donath, A, Jühling, F, Gärtner, F, Florentz, C, Fritzsch, G, Pütz, J, Middendorf, M, Stadler, P (2013) MITOS: Improved de novo Metazoan Mitochondrial Genome Annotation. *Molecular phylogenetics and evolution* **69**, 313-319.

Bolormaa, S, Hayes, B, Hawken, R, Zhang, Y, Reverter, A, Goddard, M (2011a) Detection of chromosome segments of zebu and taurine origin and their effect on beef production and growth. *Journal of Animal Science* **89**, 2050-2060.

Bolormaa, S, Neto, LP, Zhang, Y, Bunch, R, Harrison, B, Goddard, M, Barendse, W (2011b) A genome-wide association study of meat and carcass traits in Australian cattle. *Journal of Animal Science* **89**, 2297-2309.

Boussaha, M, Esquerré, D, Barbieri, J, Djari, A, Pinton, A, Letaief, R, Salin, G, Escudié, F, Roulet, A, Fritz, S, Samson, F, Grohs, C, Bernard, M, Klopp, C, Boichard, D, Rocha, D (2015) Genome-Wide Study of Structural Variants in Bovine Holstein, Montbéliarde and Normande Dairy Breeds. *PloS one* **10**, e0135931.

Brenneman, R, Davis, S, Sanders, J, Burns, B, Wheeler, T, Turner, J, Taylor, J (1996) The polled locus maps to BTA1 in a Bos indicus× Bos taurus cross. *Journal of Heredity* **87**, 156-161.

Brownlee, GG, Sanger, F (1967) Nucleotide sequences from the low molecular weight ribosomal RNA of Escherichia coli. *Journal of Molecular Biology* **23**, 337-53.

Consortium, TBH (2009) Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* **324**, 528.

Couldrey, C, Keehan, M, Johnson, T, Tiplady, K, Winkelman, A, Littlejohn, MD, Scott, A, Kemper, KE, Hayes, B, Davis, SR, Spelman, RJ (2017) Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. *Journal of dairy science* **100**, 5472-5478.

Daetwyler, HD, Capitan, A, Pausch, H, Stothard, P, Van Binsbergen, R, Brøndum, RF, Liao, X, Djari, A, Rodriguez, SC, Grohs, C (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics* **46**, 858.

Dube, SK, Marcker, KA, Clark, BF, Cory, S (1968) Nucleotide sequence of N-formyl-methionyl-transfer RNA. *Nature* **218**, 232-3.

Elsik, CG, Tellam, RL, Worley, KC (2009) The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. *Science* **324**, 522.

English, AC, Richards, S, Han, Y, Wang, M, Vee, V, Qu, J, Qin, X, Muzny, DM, Reid, JG, Worley, KC, Gibbs, RA (2012) Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PloS one* **7**, e47768.

Fountain, MD, Tao, H, Chen, CA, Yin, J, Schaaf, CP (2017) Magel2 knockout mice manifest altered social phenotypes and a deficit in preference for social novelty. *Genes, Brain and Behavior* **16**, 592-600.

Georges, M, Drinkwater, R, King, T, Mishra, A, Moore, SS, Nielsen, D, Sargeant, LS, Sorensen, A, Steele, MR, Zhao, X (1993) Microsatellite mapping of a gene affecting horn development in Bos taurus. *Nature genetics* **4**, 206.

Grant, JR, Arantes, AS, Liao, X, Stothard, P (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* **27**, 2300-2301.

Grant, JR, Stothard, P (2008) The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research* **36**, W181-W184.

Greenwood, PL, Gardner, GE, Ferguson, DM (2018) Current situation and future prospects for the Australian beef industry - A review. *Asian-Australasian Journal of Animal Sciences* **31**, 992-1006.

Griffiths-Jones, S, Saini, HK, van Dongen, S, Enright, AJ (2007) miRBase: tools for microRNA genomics. *Nucleic acids research* **36**, D154-D158.

Hansen, PJ (2004) Physiological and cellular adaptations of zebu cattle to thermal stress. *Journal of Animal Reproduction Science* **82-3**, 349-360.

Heather, JM, Chain, B (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1-8.

Heller, D, Vingron, M (2019) SVIM: structural variant identification using mapped long reads. *Bioinformatics* btz041.

Holt, RA, Jones, SJM (2008) The new paradigm of flow cell sequencing. *Genome Research* **18**, 839-846.

Kebschull, JM, Zador, AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic acids research* **43**, e143.

Khalifa, ME, Varsani, A, Ganley, ARD, Pearson, MN (2016) Comparison of Illumina de novo assembled and Sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus Sclerotinia sclerotiorum. *Journal of Virus Research* **219**, 51-57.

Koufariotis, LT, Chen, Y-PP, Chamberlain, A, Vander Jagt, C, Hayes, BJ (2015) A catalogue of novel bovine long noncoding RNA across 18 tissues. *PloS one* **10**, e0141225.

Legge, AJ (1991) The Walking Larder - Patterns of Domestication, Pastoralism, and Predation - Cluttonbrock,J. *Antiquity* **65**, 147-153.

Leggett, RM, Clark, MD (2017) A world of opportunities with nanopore sequencing. *Journal of Experimental Botany* **68**, 5419-5429.

Li, H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100.

Liao, X, Peng, F, Forni, S, McLaren, D, Plastow, G, Stothard, P (2013) Whole genome sequencing of Gir cattle for identifying polymorphisms and loci under selection. *Genome* **56**, 592-598.

Lindholm-Perry, A, Kuehn, L, Smith, T, Ferrell, C, Jenkins, T, Freetly, H, Snelling, W (2012) A region on BTA14 that includes the positional candidate genes LYPLA1, XKR4 and TMEM68 is associated with feed intake and growth phenotypes in cattle 1. *Animal genetics* **43**, 216-219.

Liu, L, Li, YH, Li, SL, Hu, N, He, YM, Pong, R, Lin, DN, Lu, LH, Law, M (2012) Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* Article ID 251364.

Mardis, ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**, 133-41.

Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA, Berka, J, Braverman, MS, Chen, YJ, Chen, Z, Dewell, SB, Du, L, Fierro, JM, Gomes, XV, Godwin, BC, He, W, Helgesen, S, Ho, CH, Irzyk, GP, Jando, SC, Alenquer, ML, Jarvie, TP, Jirage, KB, Kim, JB, Knight, JR, Lanza, JR, Leamon, JH, Lefkowitz, SM, Lei, M, Li, J, Lohman, KL, Lu, H, Makhijani, VB, McDade, KE, McKenna, MP, Myers, EW, Nickerson, E, Nobile, JR, Plant, R, Puc, BP, Ronan, MT, Roth, GT, Sarkis, GJ, Simons, JF, Simpson, JW, Srinivasan, M, Tartaro, KR, Tomasz, A, Vogt, KA, Volkmer, GA, Wang, SH, Wang, Y, Weiner, MP, Yu, P, Begley, RF, Rothberg, JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-80.

Mariasegaram, M, Harrison, BE, Bolton, JA, Tier, B, Henshall, JM, Barendse, W, Prayaga, KC (2012) Fine-mapping the POLL locus in B rahman cattle yields the diagnostic marker CSAFG29. *Animal genetics* **43**, 683-688.

Mclean, I, Holmes, P, Counsell, D, Bush AgriBusiness Pty Ltd, Holmes & Co. (2014) The Northern beef report: 2013 Northern beef situation analysis. Meat & Livestock Australia Ltd., North Sydney, NSW, 2059.

Meat & Livestock Australia, 2017. Fast Facts: Australia's Beef Industry.

Meat & Livestock Australia, 2018. Fast facts - Australia's beef industry. https://www.mla.com.au/globalassets/mla-corporate/prices--markets/documents/trends--analysis/fast-facts--maps/mla_beef-fast-facts-2018.pdf.

Medugorac, I, Graf, A, Grohs, C, Rothammer, S, Zagdsuren, Y, Gladyr, E, Zinovieva, N, Barbieri, J, Seichter, D, Russ, I (2017) Whole-genome analysis of introgressive hybridization and characterization of the bovine legacy of Mongolian yaks. *Nature genetics* **49**, 470.

Medugorac, I, Seichter, D, Graf, A, Russ, I, Blum, H, Göpel, KH, Rothammer, S, Förster, M, Krebs, S (2012) Bovine polledness–an autosomal dominant trait with allelic heterogeneity. *PloS one* **7**, e39477.

Min Jou, W, Haegeman, G, Ysebaert, M, Fiers, W (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**, 82-8.

Morozova, O, Marra, MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255-64.

O'Neill, CJ, Swain, DL, Kadarmideen, HN (2010) Evolutionary process of Bos taurus cattle in favourable versus unfavourable environments and its implications for genetic selection. *Evolutionary Applications* **3**, 422-433.

Pareek, CS, Smoczynski, R, Tretyn, A (2011) Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52**, 413-35.

Pillai, S, Gopalan, V, Lam, AKY (2017) Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. *Critical Reviews in Oncology Hematology* **116**, 58-67.

Pollard, MO, Gurdasani, D, Mentzer, AJ, Porter, T, Sandhu, MS (2018) Long reads: their purpose and place. *Human Molecular Genetics* **27**, R234-R241.

Porto Neto, L, Bunch, R, Harrison, B, Barendse, W (2012) Variation in the XKR4 gene was significantly associated with subcutaneous rump fat thickness in indicine and composite cattle. *Animal genetics* **43**, 785-789.

Sanger, F, Brownlee, GG, Barrell, BG (1965) A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of Molecular Biology* **13**, 373-98.

Schuster, SC (2008) Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16-8.

Seichter, D, Russ, I, Rothammer, S, Eder, J, Förster, M, Medugorac, I (2012) SNP-based association mapping of the polled gene in divergent cattle breeds. *Animal genetics* **43**, 595-598.

Shendure, J, Porreca, GJ, Reppas, NB, Lin, X, McCutcheon, JP, Rosenbaum, AM, Wang, MD, Zhang, K, Mitra, RD, Church, GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-32.

Stangier, KA (2009) Next-generation sequencing: a short comparison. *Journal of Laboratory Medicine* **33**, 267-270.

Su, J, Wang, Y, Xing, X, Liu, J, Zhang, Y (2014) Genome-wide analysis of DNA methylation in bovine placentas. *BMC genomics* **15**, 12.

Thorvaldsdóttir, H, Robinson, JT, Mesirov, JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192.

Utsunomiya, YT, Carmo, AS, Neves, HH, Carvalheiro, R, Matos, MC, Zavarez, LB, Ito, PK, O'Brien, AMP, Sölkner, J, Porto-Neto, LR (2014) Genome-wide mapping of loci explaining variance in scrotal circumference in Nellore cattle. *PloS one* **9**, e88561.

Utsunomiya, YT, Milanesi, M, Utsunomiya, ATH, Torrecilha, RBP, Kim, ES, Costa, MS, Aguiar, TS, Schroeder, S, do Carmo, AS, Carvalheiro, R, Neves, HHR, Padula, RCM, Sussai, TS, Zavarez, LB, Cipriano, RS, Caminhas, MMT, Hambrecht, G, Colli, L, Eufemi, E, Ajmone-Marsan, P, Cesana, D, Sannazaro, M, Buora, M, Morgante, M, Liu, G, Bickhart, D, Van Tassell, CP, Solkner, J, Sonstegard, TS, Garcia, JF (2017) A PLAG1 mutation contributed to stature recovery in modern cattle. *Scientific Reports* **7**, Article number: 17140

van Dijk, EL, Auger, H, Jaszczyszyn, Y, Thermes, C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics* **30**, 418-426.

Villalobos-Cortes, A, Martinez, A, Vega-Pla, JL, Landi, V, Quiroz, J, Marques, JR, Delgado, JV (2015) Genetic relationships among five zebu breeds naturalized in America accessed with molecular markers. *Italian Journal of Animal Science* **14**, 158-162.

Wiedemar, N, Tetens, J, Jagannathan, V, Menoud, A, Neuenschwander, S, Bruggmann, R, Thaller, G, Drögemüller, C (2014) Independent polled mutations leading to complex gene expression differences in cattle. *PloS one* **9**, e93435.

Wu, Y, Fan, H, Wang, Y, Zhang, L, Gao, X, Chen, Y, Li, J, Ren, H, Gao, H (2014) Genome-wide association studies using haplotypes and individual SNPs in Simmental cattle. *PloS one* **9**, e109330.