Azam *et al.*
Journal of Animal Science and Biotechnology (2025) 16:21
https://doi.org/10.1186/s40104-024-01133-1

Journal of Animal Science and Biotechnology

RESEARCH

Open Access

Advancing the Indian cattle pangenome: characterizing non-reference sequences in *Bos indicus*

Sarwar Azam[1,2], Abhisek Sahu[1], Naveen Kumar Pandey[1], Mahesh Neupane[3], Curtis P Van Tassell[3], Benjamin D Rosen[3*], Ravi Kumar Gandham[1,4*], Subha Narayan Rath[2] and Subeer S Majumdar[1]

Abstract

Background India harbors the world's largest cattle population, encompassing over 50 distinct *Bos indicus* breeds. This rich genetic diversity underscores the inadequacy of a single reference genome to fully capture the genomic landscape of Indian cattle. To comprehensively characterize the genomic variation within *Bos indicus* and, specifically, dairy breeds, we aim to identify non-reference sequences and construct a comprehensive pangenome.

Results Five representative genomes of prominent dairy breeds, including Gir, Kankrej, Tharparkar, Sahiwal, and Red Sindhi, were sequenced using 10X Genomics 'linked-read' technology. Assemblies generated from these linked-reads ranged from 2.70 Gb to 2.77 Gb, comparable to the *Bos indicus* Brahman reference genome. A pangenome of *Bos indicus* cattle was constructed by comparing the newly assembled genomes with the reference using alignment and graph-based methods, revealing 8 Mb and 17.7 Mb of novel sequence respectively. A confident set of 6,844 Non-reference Unique Insertions (NUIs) spanning 7.57 Mb was identified through both methods, representing the pangenome of Indian *Bos indicus* breeds. Comparative analysis with previously published pangenomes unveiled 2.8 Mb (37%) commonality with the Chinese indicine pangenome and only 1% commonality with the *Bos taurus* pangenome. Among these, 2,312 NUIs encompassing ~ 2 Mb, were commonly found in 98 samples of the 5 breeds and designated as *Bos indicus* Common Insertions (BICIs) in the population. Furthermore, 926 BICIs were identified within 682 protein-coding genes, 54 long non-coding RNAs (lncRNA), and 18 pseudogenes. These protein-coding genes were enriched for functions such as chemical synaptic transmission, cell junction organization, cell-cell adhesion, and cell morphogenesis. The protein-coding genes were found in various prominent quantitative trait locus (QTL) regions, suggesting potential roles of BICIs in traits related to milk production, reproduction, exterior, health, meat, and carcass. Notably, 63.21% of the bases within the BICIs call set contained interspersed repeats, predominantly Long Interspersed Nuclear Elements (LINEs). Additionally, 70.28% of BICIs are shared with other domesticated and wild species, highlighting their evolutionary significance.

Conclusions This is the first report unveiling a robust set of NUIs defining the pangenome of *Bos indicus* breeds of India. The analyses contribute valuable insights into the genomic landscape of *desi* cattle breeds.

Keywords BICIs, *Bos indicus*, Cattle, Genome assembly, Linked-reads, NUIs, Pangenome

*Correspondence:
Benjamin D Rosen
ben.rosen@usda.gov
Ravi Kumar Gandham
gandham71@gmail.com
Full list of author information is available at the end of the article

Background

Cattle hold incredible importance in global agriculture as one of the most pivotal livestock species. They contribute significantly to human nutrition, the economy, and agricultural practices by providing essential resources like milk, meat, hide, and drought power [1, 2]. In the Indian context, cattle are predominantly reared for milk production and draught purposes. Cattle can be broadly categorized into two primary types: *Bos taurus* and *Bos indicus*, originating from distinct domestication events [3]. Approximately 10,000 years ago, the Fertile Crescent witnessed the emergence of humpless taurine cattle (*Bos taurus taurus*) [4], while about 8,000 years ago, the Indus Valley gave rise to humped indicine cattle (*Bos taurus indicus*) [5, 6]. Genetic studies indicate that these two lineages diverged from a common ancestor approximately 210,000–350,000 years ago, well before the domestication processes took place [5]. This deep genetic divergence underscores the inherent distinctiveness of the ancestral aurochs populations from which both lineages originated. Moreover, multiple migration waves [7], interbreeding, and introgressions with other bovids, such as yak and banteng, have substantially augmented the genetic diversity within the Bovinae group [3, 8]. Furthermore, the continuous selection and adaptation to diverse climates and environmental pressures, including factors such as altitude and endemic diseases, have further molded and diversified the cattle genome [9]. This has resulted in exceptionally high levels of genetic diversity in cattle populations worldwide.

These genetic differences manifest as unique sequences within the genomes of various cattle breeds. This genomic variation is not limited to single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) but extends to encompass other structural variations (SVs) [10]. To explore this diversity, initiatives like the 1000 Bull Genomes Project have examined numerous *Bos taurus* breeds, revealing variations that are not universally present in all individuals of the species [11]. Comparisons among various genome assemblies have revealed specific sequences not universally present in all individuals of the species. While a single reference genome was initially considered sufficient to represent the entire species, it later became evident that there are sequences specific to individual breeds within a species [12]. This realization led to the concept of the "pangenome" initially applied to bacterial genomes [13] and later extended to fungi [14] and plant genomes [15]. Surprisingly, this approach had not been extensively explored in large eukaryotes, especially mammals, until the African Pan-genome Project revealed sequence variability in African human genomes compared to the human reference genome [16]. Subsequent reports on the human pangenome established the existence of specific sequences in populations, augmenting the reference genome to construct a comprehensive pangenome [17]. Similar efforts to explore pangenomes were initiated in other large animals, including cattle [12, 18, 19].

Initially, a read-depth based approach was applied to identify SVs, followed by pangenome construction from these SVs. However, with the availability of multiple genome assemblies, direct sequence comparisons have become more common than mapping reads to the reference genome to identify unique insertions [20]. The latter approach offers a more faithful representation of complex regions and SVs, given the intricacies of calling and representing nested variations. A graph-based method has evolved to compare genome assemblies for the identification of unique insertions, which are then integrated with the reference genome [21]. However, in the context of cattle, only a few efforts have been made to explore the pangenome, primarily within the *Bos taurus* species [12, 18, 19].

For instance, Zhou et al. [12] reported SVs in the ARS-UCD1.2 *Bos taurus* reference genome [22], using 898 individuals, identifying 83 Mb of sequence not found in the reference genome. Leonard et al. [18] used a trio-binning approach to assemble six genomes, including three *Bos taurus taurus*, one *Bos gaurus*, one *Bos taurus indicus* from hybrid progenies, resulting in the discovery of 90 thousand structural variants, including 931 overlapping with coding sequences. Crysnanto et al. [23] revealed a bovine pangenome using the Hereford-based *Bos taurus* reference genome and five reference-quality assemblies from three taurine cattle breeds (Angus, Highland, and Original Braunvieh) and their close relatives Brahman (*Bos taurus indicus*) and yak (*Bos grunniens*). The pangenome contained an additional 70,329,827 bases compared to the *Bos taurus* reference genome. Their multi-assembly approach unveiled 30 and 10.1 million bases unique to yak and indicine cattle, respectively, as well as between 3.3 and 4.4 million bases unique to each taurine assembly.

While the concept of a pangenome is not limited to specific groups of individuals, populations, or species, there are efforts to create larger pangenomes that encompass multiple species. These expanded pangenomes, sometimes referred to as 'super pangenomes' [24] have been developed for crops like soybeans [25] and tomatoes [26], which include both wild and domesticated species. Similar endeavors are being undertaken for cattle by the Bovine Pangenome Consortium [27], which aims to include many wild relatives of *Bos*. However, it's important to note that in human and other species, most experiments focus on creating species-specific pangenomes. These species-specific pangenomes are essential and have

been instrumental in studying population-specific traits, including disease susceptibility and adaptation.

Linked-read technology [28] has been successfully employed for developing pangenomes in humans [29]. Wong et al. [29] published a Non-reference Unique Insertions (NUI) discovery pipeline, identifying missing sequences in the human reference genome from in silico, phased, de novo human genome assemblies generated using linked-reads. They coined the term 'NUI' to describe unique sequences not present in the reference genome but found in other individual's genomes. In this study, we present an effort to construct a pangenome specific to Indian *Bos indicus* cattle. In fact, there are more than 50 registered indigenous breeds of *Bos indicus* cattle in India, of which only a few are dairy breeds. The prominent dairy breeds include Gir, Kankrej, Tharparkar, Sahiwal, and Red Sindhi [30, 31]. Representative genomes of these five *Bos indicus* breeds were sequenced using linked-reads. The assemblies generated were compared with the Brahman genome assembly as currently recognized as the *Bos indicus* reference sequence [32]. Apart from applying the NUI pipeline, we used a graph-based pangenome method to identify unique insertions across *Bos indicus* breeds. Importantly, our effort has led to the development of the first pangenome specific for *Bos indicus* dairy breeds found in India.

Methods

Genome sequencing of cattle breeds

Blood samples were collected from representative individuals of the Gir, Sahiwal, Tharparkar, Red Sindhi, and Kankrej breeds, in compliance with the guidelines of the Committee for Control and Supervision of Experiments on Animals (CCSEA), India. High molecular weight (HMW) DNA was extracted and 10X Genomics Chromium technology [33] libraries were prepared by AgriGenome Pvt. Ltd. (Bangalore, India). Sequencing was subsequently performed on an Illumina HiSeq X system.

Genome assembly

Raw data for each breed were meticulously processed for genome assembly employing the 10X Genomics Supernova assembler v2.1 (RRID: SCR_016756) [33]. The assembly process was conducted on a high-performance Dell PowerEdge R740 server equipped with 754 GB of RAM and 96 threads (48 CPUs) within a CentOS 7 Linux environment. Notably, the 'supernova mkoutput' command was executed twice, first with the '--style = pseudohap1' option, producing genome assemblies representing a single haplotype for each breed. Subsequently, the command was run with the '--style = pseudohap2' option, which generated two assembly files representing both haplotypes within the genome of each breed [34]. These de novo assemblies serve as the foundation for downstream analyses to identify NUI as part of pangenome specific to these indigenous dairy cattle breeds.

NUI discovery pipeline for linked-reads

The NUI discovery pipeline, as developed by Wong et al. [29], was employed in this study for the identification of NUIs in cattle. This pipeline is specifically designed for the analysis of linked-read data and has been validated using human linked-read datasets. For its execution, the NUI pipeline necessitates several prerequisites, including a reference genome, information on segmental duplications within the reference genome, and pseudo-haplotypes derived from linked-read sequencing data. Additionally, the pipeline requires an alignment file in BAM format containing linked-reads aligned to the reference genome. BAM files for each linked-read dataset were generated against the Brahman reference genome using Long Ranger v2.2 (RRID: SCR_018925) [34]. Segmental duplication information for the Brahman reference genome was generated using BISER v1.4 [35].

The NUI pipeline is scripted in Bash shell and incorporates a suite of essential bioinformatics tools, including Samtools v1.2 (RRID: SCR_002105) [36], SAMBAMBA (RRID: SCR_024328) [37], BWA v0.7.15 (RRID: SCR_010910) [38], LASTZ v1.04 (RRID: SCR_018556) [39], RepeatMasker v4.1.0 (RRID: SCR_012954) [40], BEDTools v2.17.0 (RRID: SCR_006646) [41], Dustmasker [42], SAMBLASTER v0.1.24 (RRID: SCR_000468) [43], R v3.6.3 (RRID: SCR_001905) [44] and BLAST v2.9.0+ (RRID: SCR_004870) [42]. In essence, the pipeline commences by extracting unaligned reads from the BAM file using Samtools, SAMBLASTER, and SAMBAMBA, followed by trimming and quality filtering using the FASTX-Toolkit v0.0.14 (RRID: SCR_005534) [45]. These reads are then mapped individually to both pseudohaplotypes using the BWA tool. True alignments are filtered using SAMBAMBA, and genomewide coverage is calculated from the aligned files. Read clusters with coverage levels between 8X and 100X are identified using BEDTools. The pipeline proceeds to identify read clusters and extends the pseudo-haplotype sequences, expanding them by 7,000 bp on each end, or until the ends of the assembled sequences, using BEDTools. These extended contigs are then aligned to the reference genome using LASTZ. The pipeline computes precise breakpoints and identifies insertions within each contig by pinpointing where the sequence alignment diverged and subsequently realigning it. Only insertional sequences with gap sizes $\geq$ 50 bp are retained. The outputs from the two pseudo-haplotypes are combined, and unaligned breakpoint-to-breakpoint sequences from all the contigs are extracted for subsequent analysis. These sequences are

then subjected to RepeatMasker and Dustmasker analysis, with sequences containing ≥ 50 unique bases retained as NUIs. These NUIs were realigned with the reference genome using BLAST and further filtered to exclude all alignments with ≥ 95% identity and 100% coverage, while also excluding those with breakpoints overlapping assembly gaps and segmental duplications within the reference genome. Finally, NUIs identified across the five individuals are merged to create a unified and non-redundant call set, a process facilitated by the "combine_metaNUI.R" script within the pipeline, executed using Rscript. A visualization of the pipeline workflow is provided in Fig. 1.

Construction of multi-assembly graph and insertion discovery

The minigraph tool was employed to construct a multi-assembly graph [21], which facilitated the discovery of insertions. The input for minigraph included the genome assembly of each breed and the Brahman reference genome. The Brahman reference genome served as the backbone, and the base alignment option was enabled ('-cxggs') to ensure alignment consistency. Mash v2.3 (RRID: SCR_019135) [46] was used to determine the genetic distance between the Brahman genome and other cattle genomes. Based on the Mash distance calculations, the assemblies were incrementally provided to minigraph, commencing with the closest distance to the Brahman genome [23]. The output generated by minigraph was obtained in the '.gfa' format, which was subsequently converted to a fasta file using the gfatools v0.5 [47]. Contigs exceeding a length of 50 bp were retained as non-reference sequences, facilitating the identification of insertions within the multi-assembly graph. A visualization is provided in the flowchart in Fig. 1.

Contamination screening of insertion sequences

In this study, we used an in-house, stand-alone tool known as Fasta2Lineage [48], designed to identify the lineage of contigs. Fasta2Lineage utilizes the non-redundant nucleotide (nt) database of NCBI in conjunction with BLASTN for its functionality. The tool takes contig sequences in fasta format as input and conducts a similarity search against all the sequences in the database. It then identifies the best alignments and annotates their complete lineage. Contigs that do not align with any sequence in the nt database are considered novel. Fasta2lineage proves invaluable in identifying and segregating potentially contaminated sequences. For cattle assemblies, any contigs that are associated with microorganisms such as archaea, bacteria, viruses, or plants, as well as non-chordate sequences, are classified as potential contaminations.

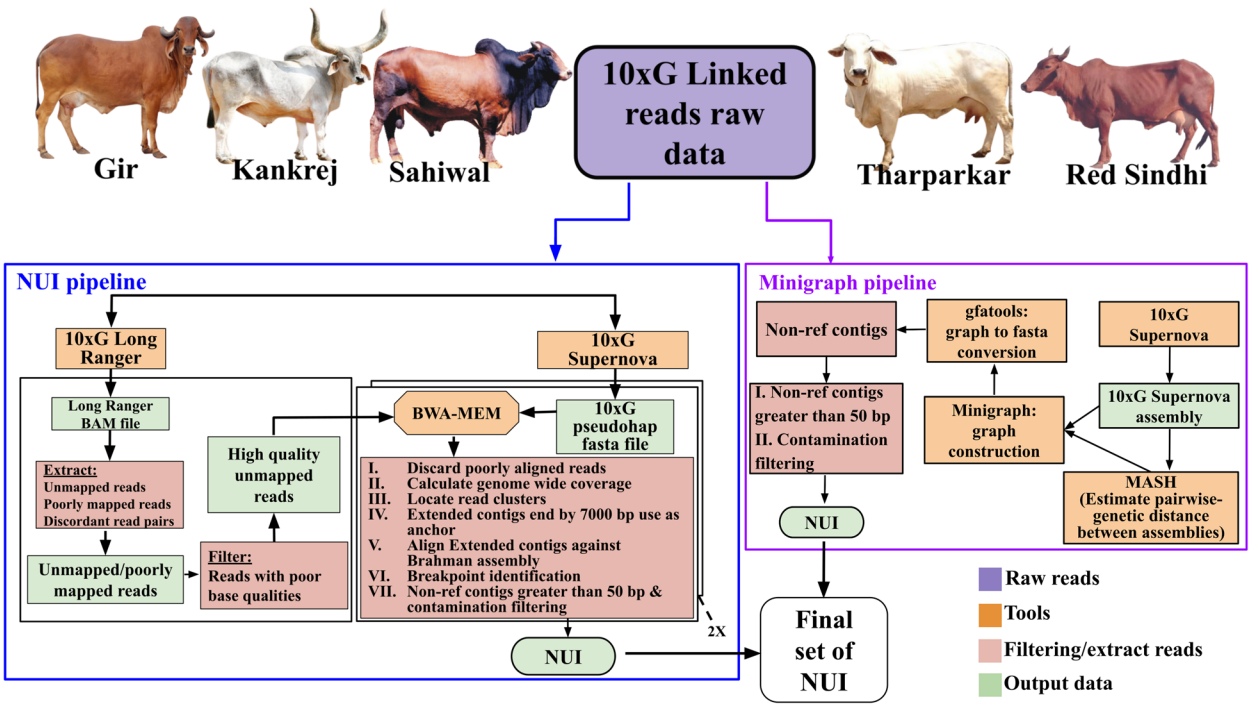

Fig. 1 Identification of Non-reference Unique Insertions (NUIs) in *Bos indicus*. The flowchart illustrates the systematic process for identifying the final set of NUIs. The diagram outlines the sequential steps involved in the selection and refinement of NUIs

In our study, NUIs previously identified using the NUI discovery pipeline and all insertions identified via minigraph were subjected to screening using the Fasta-2Lineage tool. Sequences that were deemed novel or associated with Chordata were selected as decontaminated sequences for both datasets.

Selection of common set of NUI and comparison with published cattle pangenomes

Cleaned contigs from NUI discovery pipeline and minigraph analysis were cross-referenced using the best bidirectional hits method with BLAST, applying stringent criteria of 95% identity and 95% query coverage. Among the selected contigs, NUIs that overlapped with insertion sequences in the minigraph dataset were selected as common and final set of NUIs.

To identify NUIs originating from *Bos taurus* introgression in the Brahman genome, we obtained the coordinates of introgressed regions from the study by Naji et al. [49]. Subsequently, we mapped the coordinates of NUIs onto these introgressed regions. Any NUI that overlapped with the introgressed region coordinates was considered as having originated from these regions.

The final set of NUIs were compared with two published pangenomes, developed by Zhou et al. [12] and Dai et al. [19]. The pangenome by Zhou et al. [12] encompasses 22,324 contigs, collectively spanning a substantial 94 Mb, and was derived from genomic data of 898 cattle across 57 breeds. In contrast, the recently published pangenome of *Bos indicus* cattle by Dai et al. [19] was constructed from genome assemblies of 10 Chinese breeds. This recently established pangenome comprises 74,907 sequences spanning 124.4 Mb. Both sets of pangenome sequences were downloaded and subjected to comparison with the NUIs using the BLAST [42]. NUIs were considered matching when their sequences aligned with a minimum of 95% coverage and 95% identity to the published non-reference sequence.

Assessing the impact of NUIs on the transcriptome of *desi* cattle

To assess the impact of NUIs on the transcriptome, we retrieved RNA-seq data from 46 samples representing four *desi* cattle breeds: Gir, Tharparkar, Sahiwal, and Hariana. The reads were aligned to the Brahman reference genome using STAR aligner v2.7.10 (RRID) [50]. The genome was indexed with the following command: "STAR --runThreadN 32 --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles ./GCF_003369695.1_UOA_Brahman_1_genomic.fna –sjdbGTFfile GCF_003369695.1_UOA_Brahman_1_genomic.gtf --sjdbOverhang 100". All RNA-seq reads were then aligned to the indexed genome, processing one sample at a time. The alignment was performed using the command: "STAR --runThreadN 40 --outReadsUnmapped Fastx --genomeDir ./assembly/ --outFileNamePrefix ./dis_04R --outFilterMultimapNmax 100000 --outSAMunmapped Within KeepPairs --limitOutSAMoneReadBytes 1000000 --readFilesCommand zcat --readFilesIn read1.fq.gz read2.fq.gz". Unmapped reads from each sample were subsequently extracted and mapped onto the NUIs. Since the size of the NUIs began at 50 bp, we included a 300 bp flanking sequence around each NUI. Any reads that aligned to or overlapped with the actual NUIs were considered successfully mapped. Finally, the mapping percentages for each sample on both the Brahman genome and the pangenome (consisting of the Brahman genome plus NUIs) were plotted using R.

To assess the impact on differential gene expression, we utilized a publicly available dataset for heat stress in *desi* cattle published by Sajjanar et al. [51]. This dataset includes raw sequence reads from 3 control and 3 heat-stressed peripheral blood mononuclear cell (PBMC) samples of the Hariana breed. The data were preprocessed using Fastp and aligned to the pangenome using STAR aligner, following the same commands described previously. Potential transcripts assembled on NUI using StringTie [52] were merged with the Brahman reference gene annotation file, and StringTie was re-run again on each sample with the -e option. A read count matrix was then generated using the prepDE script. The read count matrix was provided as input to EdgeR [53], and genes with expression levels below 1 CPM were discarded. Generalized linear model (GLM) fitting was applied to the normalized dataset, and genes with a Benjamini-Hochberg-corrected FDR ≤ 0.05 were considered differentially expressed between heat stress (HS) and control (CN) samples. Differentially expressed genes identified from NUIs were annotated using BLASTx against the non-redundant (nr) database of NCBI, restricted to chordates with an e-value threshold of 1e-5.

Identification of common insertions

To screen the presence of NUIs within the population, we performed genotyping using short-read sequence data with >30X coverage. This data was obtained by sequencing 98 individuals from five cattle breeds: Gir (20), Kankrej (19), Tharparkar (20), Sahiwal (20), and Red Sindhi (19), utilizing the Illumina HiSeqX platform. Each sample underwent preprocessing using fastp v0.23.3 (RRID: SCR_016962) [54]. Further, clean reads of each sample were mapped onto both pseudo haplotypes of each genome using BWA-MEM2 v2.2.1 (RRID: SCR_022192) [55]. The NUI discovery pipeline provided NUI fasta sequences and their coordinates on the representative pseudo haplotypes. With this information, aligned

BAM files for each sample were processed to determine the depth and consensus bases at each NUI coordinate on the pseudo haplotypes using Samtools. Additionally, in-house Perl scripts [56] were employed to discern the presence or absence of NUIs within the sample. NUIs covered with at least 80% coverage and 90% identity were considered 'present,' while those not meeting these criteria were categorized as 'absent'. Subsequently, a genotyping matrix was generated by consolidating data from all the samples. Furthermore, NUIs with minor allele frequency (MAF) < 5% across all samples were excluded, and the remaining NUIs were considered as the final set of *Bos indicus* Common Insertions (BICIs) variably present in *desi* breeds.

Principal component analysis

The NUI discovery pipeline provided information for each BICI and its occurrence in the pseudohaplotypes of each genome. From the NUI discovery pipeline output, we selectively extracted the final set of NUIs and their corresponding occurrence information. This information was used to create a BICIs occurrence matrix, which served as input for the principal component analysis (PCA). The BICIs occurrence matrix was generated using a Perl script [56], with matrix elements designated as 0, 1, and 2. In this matrix, 0 represents the absence of BICI, whereas a value of 1 was assigned if the BICI was present in either of the pseudohaplotypes, and a value of 2 was allocated if the BICI was present in both pseudohaplotypes. The PCA was calculated using R, and a scatterplot of PC1 vs. PC2 was generated using SRplot [57].

Repeat and transposable element analysis of BICIs

The fasta sequences of BICIs were extracted and analyzed to determine the composition of repeats and transposable elements (TEs) using RepeatMasker v4.1.0 (RRID: SCR_012954) [40]. RepeatMasker was executed with --species cow, -xsmall, and -nolow. Within each BICI, the TE element constituting the highest percentage of that sequence was designated as the major TE.

Furthermore, we extended the BICI sequences by 300 base pairs on each side to include flanking sequences. We then employed RepeatMasker with the same parameters to assess the composition of TEs in flanking sequences. This analysis aimed to identify TEs spanning the breakpoints on either side of the BICI and to identify major TEs present in the flanking sequences.

Identifying transcriptionally potent BICIs

We employed the gene prediction tool Augustus v3.4.0 (RRID: SCR_008417) [58] to predict protein-coding genes within the BICIs, using the options '--singlestrand = true --genemodel = complete'. The resulting predicted protein sequences were then subjected to functional and pathway annotation.

Additionally, we conducted a thorough examination of the breakpoint locations of BICIs within the gene sequences of the Brahman reference genome. Annotated genes that coincided with BICI breakpoints were identified and extracted for analysis. For BICIs with breakpoints outside of genes, we conducted a search for the nearest gene within a 20-kilobase (20 kb) range.

To further investigate the transcriptional potential of BICIs, we mapped RNA-seq data from 47 samples representing various indigenous breeds, including Gir, Tharparkar, Sahiwal, and Hariana, onto the Brahman reference genome. The genome was indexed using the STAR aligner v2.7.10 (RRID: SCR_004463) [50] and all dataset were aligned to this indexed genome, one sample at a time, using the commands previously mentioned in "Assessing the impact of NUIs on the transcriptome of *desi* cattle" section. This process generated separate R1 and R2 files in FASTQ format for the unmapped reads. Subsequently, we collected the unmapped reads from all 47 samples and merged these FASTQ files into single read1 and read2 files. We then realigned these merged unmapped reads to the Brahman reference genome without supplying the GTF file. Again, we collected all the unmapped reads in Fastq format and aligned them to BICI with flanking sequences extended by 300 bp on both ends. We generated an index of flanked BICIs and aligned the unmapped reads to these indexed BICIs using parameters as previously mentioned, with the exception of not providing a GTF file. We filtered the resulting alignment file in SAM format based on the following criteria: (a) both reads in a read pair were mapped; (b) read pairs were mapped in the correct orientation and with the correct insert size; (c) reads had alignment scores ≥ 140, equivalent to four total mismatches for a read pair; and (d) reads had a mapping quality score of 255, indicative of unique mapping according to the STAR scoring scheme.

The filtered alignment file was converted into a coordinate-sorted BAM file and processed using Stringtie v2.1.1 (RRID: SCR_016323) [52] to identify novel transcripts within the BICI sequences. The command 'stringtie -o nui_mapped_STAR.gtf filtered_sorted_STAR.bam -p 10' was used. Transcripts that fully resided within the 300 bp flanking sequences were discarded. This analysis enabled us to identify all novel transcripts and their corresponding transcribed BICIs.

Functional analysis of genes with BICIs

Gene Ontology (GO) enrichment analysis was performed on non-redundant set of genes having BICIs using the clusterProfiler package (RRID: SCR_016884) [59] from

the Bioconductor in R (RRID: SCR_006442) [60]. A significance threshold of P-value ≤ 0.05 (FDR by Benjamini–Hochberg) was applied to identify significant enriched terms. The resulting GO IDs from the clusterProfiler package, along with their respective P-values, were then subjected to REVIGO (RRID: SCR_005825) [61] to eliminate redundant GO categories from all enriched terms. Finally, the results of all GO categories, in conjunction with the clusterProfiler package, were visualized using a Python-based tool CirGO [62].

To establish a connection between BICIs and major QTL, non-redundant genes featuring BICIs were aligned with the publicly accessible cattle QTL database [63]. The mapping process involved extracting non-redundant genes with BICIs and conducting a BLAST search against the CDS annotated in ARS_UCD 1.2 assembly to identify orthologous genes. Subsequently, these orthologous genes were cross-referenced with the QTL database, allowing for the characterization of their associations with specific traits.

Evolutionary analysis of the BICIs

To investigate the presence of BICIs in related species, we first aligned BICIs to the genomes of sister species within the *Bos* genus, which includes *Bos taurus* (taurine cattle) [22], *Bos gaurus* (gaur) [64], *Bos grunniens* (domestic yak) [65], *Bos frontalis* (gayal) [66], and *Bos mutus* (wild yak) [65]. In a subsequent analysis, we extended our alignment to species of Bovidae family which includes *Bubalus bubalis* (buffalo) [67], *Tragelaphus oryx* (eland) [68], *Bison bison* (American bison) [69], *Capra hircus* (goat) [70], and *Ovis aries* (sheep) [71]. Reference genome sequences for each of these species were obtained from NCBI, and BICIs were aligned to their respective reference genomes using BLAST. Alignments with a minimum of 95% identity and 95% query coverage were considered genuine matches. Distribution of BICIs across species were plotted using jvenn (RRID: SCR_016343) [72].

Results

Sequencing of *desi* cattle using 10X linked-reads and genome assembly

Genome sequencing of five indigenous cattle breeds, namely Gir, Tharparkar, Kankrej, Sahiwal, and Red Sindhi, was carried out using the 10X Chromium technology [33]. Each genome was sequenced at approximately 100X coverage with 150 bp paired-end reads. Assemblies ranged in size from 2.70 Gb to 2.77 Gb for different breeds (Table 1). Notably, the Sahiwal and Red Sindhi assemblies displayed fewer contigs and larger N50 values compared to Gir, Kankrej, and Tharparkar. In particular, the Sahiwal assembly featured the largest contig, measuring 156 Mb, while the largest contigs in the Red Sindhi, Kankrej, Gir, and Tharparkar assemblies were 134 Mb, 11.4 Mb, 7.5 Mb, and 6.2 Mb, respectively. Sahiwal and Red Sindhi assemblies demonstrated higher contiguity, with 90% of the genome assembled into 62 and 100 contigs, respectively. In contrast, Kankrej, Gir, and Tharparkar assemblies were less contiguous, featuring 5,472, 5,202, and 3,848 sequences, respectively, for L90. Finally, two pseudo-haplotypes were generated for each diploid genome to enable comparisons with the reference genome (Table S1).

Iterative mapping based NUI discovery in *Bos indicus*

Upon completion, the NUI pipeline (Fig. 1) generated a FASTA file and a table representing the sequences of 9,270 non-redundant NUIs and their distribution in each sample, respectively (Table S2, Additional file 3). In further screening of these NUIs for contaminated sequences with stringent criteria, we identified and excluded 122 contaminated contigs, resulting in a clean set of 9,148 NUIs spanning 8 Mb. These NUIs ranged in size from 51 bp to 98.1 kb. Among the five individuals, Gir had the fewest NUIs, totaling 1,972, while Red Sindhi had the highest number of NUIs, with 3,901 (Fig. S1). Notably, most NUIs were unique to each individual, with only 123 NUIs found to be common across all the individuals.

Table 1 Data statistics of five *Bos indicus* de novo genome assemblies

Sample name	No. of contigs	Genome assembly size, bp	Largest contigs, bp	Smallest contigs, bp	N50	L50	L90
Gir	40,451	2,769,658,259	7,549,386	1,000	839,301	920	5,472
Kankrej	33,833	2,731,156,349	11,443,009	1,000	1,184,380	584	3,848
Tharparkar	42,943	2,778,475,415	6,244,210	1,000	851,598	874	5,202
Sahiwal	22,010	2,713,497,564	156,099,758	1,000	59,483,679	17	62
Red Sindhi	20,635	2,700,638,361	134,326,022	1,000	38,294,242	21	100

Graph base NUI discovery in *Bos indicus*

A *Bos indicus* multi-assembly graph was constructed with the Brahman reference genome [32] as backbone. It contained a total of 2,728,215,813 bp across 153,597 nodes, connected by 218,354 edges. Specifically, there are 133,985 edges connecting two reference nodes, 83,089 edges connecting reference nodes to non-reference nodes, and finally, 1,280 edges connecting non-reference to non-reference nodes.

In the multi-assembly graph, a total of 2,709,654,022 bp consists of 111,178 reference nodes, which originated from the Brahman genome. This forms the backbone of the graph. With incremental integration of Kankrej, Gir, Sahiwal, Red Sindhi, and Tharparkar, the graph further expanded. Kankrej added 16,174 nodes with 7,369,711 bp, Gir contributed 9,306 nodes with 3,711,106 bp, Sahiwal accounted for 7,286 nodes with 3,124,476 bp, Red Sindhi introduced 5,536 nodes with 2,629,123 bp, and, lastly, Tharparkar added 4,117 nodes with 1,727,375 bp. Consequently, the multi-assembly graph comprises 42,419 non-reference nodes containing 18,561,791 bp (Table 2). The non-reference nodes in the assembly graph are most abundant in Red Sindhi with 19,527 nodes featuring 8,667,230 bp. Gir features the least with 17,848 nodes encompassing 7,755,164 bp (Table 3).

The core genome of the multi-assembly graph, primarily the nodes shared by all assemblies, has a total count of 52,298 nodes featuring 2,399,340,052 bp. This constitutes 87.95% of the pangenome. Meanwhile, a total of 101,299 nodes, which constitute 12.05% of the pangenome, represent 328,875,761 bp and are considered flexible. This means these nodes do not appear in all assemblies. The flexible content of the genome is subdivided into two categories. Nodes that are present in at least two assemblies count for 76,322 with a total of 313,331,623 bp. Conversely, 24,977 nodes representing 15,544,138 bp are only found in a single assembly.

Table 2 Nodes and edges statistics from minigraph

Graph parameters	Count	Length, bp
All nodes	153,597	2,728,215,813
Reference nodes	111,178	2,709,654,022
Non-reference nodes	42,419	18,561,791
Added from Kankrej	16,174	7,369,711
Added from Gir	9,306	3,711,106
Added from Sahiwal	7,286	3,124,476
Added from Red Sindhi	5,536	2,629,123
Added from Tharparkar	4,117	1,727,375
Total Edges	218,354	-
Edge Ref-Ref	133,985	-
Edge Ref-Nonref	83,089	-
Edge Nonref-Nonref	1,280	-

Table 3 Nodes summary of five *Bos indicus* breeds from minigraph

Non-ref sequences in breed	Node count	Total sequence length, bp
Gir	17,848	7,755,164
Kankrej	18,252	8,506,737
Tharparkar	18,042	7,909,798
Sahiwal	19,195	8,477,068
Red Sindhi	19,527	8,667,230

In the end, the non-reference nodes consisting of 18,561,791 bp were filtered to include only those greater than 50 bp. Subsequently, these selected nodes underwent a screening for contaminated sequences, leading to the identification of 29,477 clean nodes. These 29,477 clean nodes, spanning 17.73 Mb with the largest node of 98.1 kb, were retained as graph-based insertion sequences for further downstream analysis (Fig. S2, Additional file 4).

Selection of common set of NUI and comparison with published pangenomes

The NUI discovery pipeline and graph-based method ultimately provided clean datasets of 9,148 and 29,477 insertions respectively. A comprehensive comparison of these two datasets revealed 6,844 NUIs within the graph-based insertion sequences, establishing them as a confident and final set of NUIs spanning 7.57 Mb (Additional file 5). All these NUIs are breakpoint-resolved and precisely located the chromosomes of the Brahman reference genome (Fig. 2). This invaluable breakpoint information facilitates downstream analysis in the population.

The introgression of *Bos taurus* genetic material into *Bos indicus* breeds, including Brahman, is well-documented. Previous studies by Naji et al. [49] identified 100 introgressed regions in the Brahman reference genome. By examining NUI breakpoints within these introgressed regions, our analysis revealed 934 NUIs distributed across 86 of these regions (Table S3).

Subsequently, these NUIs were compared with two published cattle pangenomes. When compared with the comprehensive cattle pangenome constructed by Zhou et al. [12], consisting of 22,324 contigs compiled from 898 cattle of 57 breeds, only 258 NUIs (~4%) spanning 76.92 kb (~1%), were found to be in common. This may be attributed to the fact that the dataset primarily

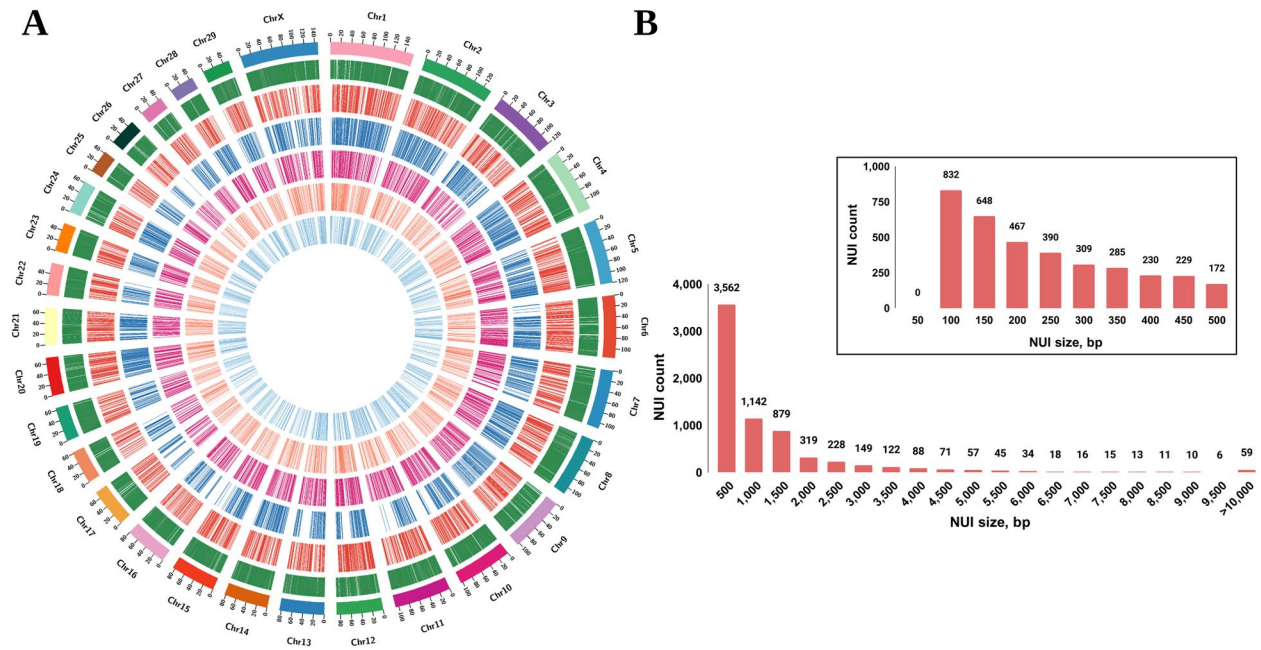

Fig. 2 Overview of Non-reference Unique Insertion (NUI) final set and their distribution. **A** Circos plot of *Bos indicus* pangenome. From the outer to inner track: Chromosome, Gene track, Red Sindhi, Sahiwal, Tharparkar, Kankrej, Gir. **B** Size distribution of NUIs with bin size of 500 bp and 50 bp in zoomed area

included European and African cattle breeds, with relatively limited inclusion of Indian breeds. Specifically, only the Gir breed from our study was featured, alongside other *Bos indicus* breeds like Brahman and Nellore.

In contrast, the comparison with the recently published pangenome of *Bos indicus* cattle [19], developed from genome assemblies of 10 Chinese indicine breeds, demonstrated a more substantial overlap. In fact, when the NUIs identified in this study were compared with 74,907 Chinese pangenome sequences spanning 124.4 Mb, it was observed that 2.80 Mb (~37%), consisting of 3,712 NUIs (~54%), were in common. Smaller contigs were more likely to overlap with 58% of contigs less than 1 kb matching (Table S4).

Impact of NUIs on transcriptomic profiles and differential gene expression

To assess the impact of NUIs and the pangenome on transcriptomic analysis, we compared the mapping efficiency of RNA-seq data aligned to both the Brahman reference genome and the pangenome. Across all samples, a higher proportion of reads mapped to the pangenome, demonstrating improved transcriptome coverage with the inclusion of NUIs. This resulted in an overall increase in the mapping rate by 0.80% (Fig. S3A), equivalent to approximately 274,000 additional reads in a paired-end RNA-seq dataset containing 32,091,193 reads. Breed-specific improvements in read alignment were also observed, with Hariana showing the greatest increase (0.89%), followed by Tharparkar (0.85%), Sahiwal (0.77%), and Gir (0.71%) (Fig. S3B, Table S5). These findings indicate that the incorporation of NUIs into the pangenome enhances transcriptomic mapping efficiency.

To evaluate the impact of NUIs on differential gene expression, we compared RNA-seq data from heat-stressed and control samples of the Hariana breed, as published by Sajjanar et al. [51]. We analyzed six samples in total, consisting of three heat-stressed and 3 control samples. The raw dataset, containing between 108 and 119 million paired-end reads, was preprocessed, resulting in 102–114 million high-quality reads per sample. These high-quality reads were then mapped onto the pangenome, yielding an average of 105 million mapped reads per sample. Subsequently, mapped reads were assembled using StringTie, identifying 132 genes within the NUIs. These genes were appended to the Brahman reference annotation, resulting in a total of 28,555 genes. For differential gene expression analysis, we filtered genes with expression levels of ≥ 1 count per million (CPM) in at least three samples, leading to the detection of 13,536 genes, including 84 genes located within the NUIs. Further analysis revealed 3,912 differentially expressed genes (FDR ≤ 0.05) between heat-stressed and control samples, including 15 genes from the NUIs (Fig. S4).

When comparing our results with those of Sajjanar et al. [51], we observed that several of the top

differentially expressed genes (DEGs) identified in their study were also among the top DEGs in our analysis. However, it is important to note that Sajjanar et al. used the ARS-UCD1.2 reference genome, which represents *Bos taurus*, while our study employed the Brahman reference genome, representing *Bos indicus* in the analysis. Despite these differences in reference genomes, the transcript abundance and expression levels (log_2 fold-change) of these genes were highly consistent across both studies (Table S6). Among the 15 differentially expressed genes from NUIs, 9 were upregulated and 6 were downregulated in the PBMCs of heat-stressed Hariana cattle. Upon annotation, many of these novel genes displayed strong homology with evolutionary closely related species such as *Bubalus kerabu*, *Bubalus bubalis*, *Bos javanicus*, *Bos mutus*, *Bos taurus*, *Capra hircus*, and *Camelus bactrianus* (Table S7). Notable annotated genes include mediator of RNA polymerase II transcription subunit 13, dynein light chain Tctex-type 1, serine hydrolase-like protein 2, *PHLDA1*, and *hnRNP A1*.

Screening for NUIs in the population and identification of BICIs

The evaluation of NUIs within a population of the same cattle breeds is vital for understanding their distribution, variability, and validation. In this study, we conducted genotyping for a total of 6,844 NUIs across 98 individuals, encompassing 20 Gir, 19 Kankrej, 20 Tharparkar, 20 Sahiwal, and 19 Red Sindhi cattle. This genotyping effort revealed the presence of both rare and common NUIs within the population. In fact, 2,789 NUIs were not called in any samples whereas 1,091 NUIs were present across all samples. To focus our analysis on the common NUIs and exclude the rare ones, we applied a stringent filtering criterion based on a MAF threshold of < 5%. As a result, we established a final NUIs call set comprising 2,312 NUIs, collectively spanning ~ 2 Mb of genomic sequence referred to as BICIs (Additional file 6). The BICIs were distributed across all the chromosomes (Fig. 3A). Among these, the largest BICI measured 47.6 kb, whereas the smallest were as short as 52 bp. It's noteworthy that 25% of BICIs (578) were equal to or less than 130 bp in length, and 50% of BICIs (1,156) were 286 bp or shorter (Fig. 3B). This refined BICIs dataset represents the common NUIs in the population and lays the foundation for all subsequent analyses in this study.

Principal component analysis and hierarchical clustering of BICIs

The distribution of the 2,312 BICIs was assessed across the five genome assemblies (Table S8). Notably, a large number of unique BICIs were identified in Red Sindhi (638 BICIs), followed by Sahiwal (591 BICIs), while the fewest BICIs were found in Tharparkar assembly (97 BICIs). Additionally, 16 BICIs were present in all five assemblies. PCA applied to these 2,312 BICIs across the five distinct assemblies effectively differentiated the Gir, Kankrej, and Tharparkar breeds from the Sahiwal and Red Sindhi breeds. PC1 explained 40.8% of the variation, while PC2 accounted for 35.1%. A PCA1 vs. PCA2 plot in (Fig. 3C) provided a clear separation of animals from different geographical regions, underscoring the utility of BICIs in discerning breed differences.

The genotyping results, depicting the presence and absence of BICIs in the population, underwent hierarchical clustering analysis. This analysis resulted in the formation of major clusters representing different breeds. Notably, Red Sindhi and Gir each formed distinct clusters, as did Kankrej and Sahiwal. Tharparkar individuals were an exception and could not be grouped into a single cluster. In fact, they clustered into two groups, with the larger group having one Sahiwal individual and four Kankrej individuals that did not group with their own cluster. The resulting cladogram visually illustrates the distribution of 2,312 BICIs in the population (Fig. S5).

Identifying the major transposable element in BICIs

In our investigation of repeats and the composition of TEs within BICIs, we discovered that approximately 63.21% of the bases within the BICI call set contained interspersed repeats (Table S9). This proportion is notably higher than the total interspersed repeats found in the entire genome of the Brahman reference sequence (46.71%). Further analysis revealed that 11.32% of the BICIs were short interspersed nuclear elements (SINEs), while long interspersed nuclear elements (LINEs) constituted 43.87% of the overall BICI dataset. This contrasts with the genome-wide repeat content of SINEs and LINEs in the Brahman reference genome, which accounted for 11.75% and 27.94%, respectively. It is evident that LINEs are significantly enriched within the BICI dataset, highlighting the prevalence of this particular repetitive element in these insertions. We further explored the distribution of major TE types within BICIs based on their sizes. Across all size categories, LINEs were the most prevalent TE within BICIs. The proportion of LINEs increased from 40.91% in BICIs under 200 bp category to a substantial 69% in longer sequences exceeding 1,000 bp. Conversely, BICIs under 200 bp were predominantly composed of non-repetitive unique sequences, while the occurrence of unique sequences declined as BICI size increased. Long terminal repeats (LTRs) and DNA transposons were observed at lower frequencies within all BICI size ranges, suggesting their relatively limited representation within the BICI dataset regardless of BICI (Fig. 4A). These findings underscore

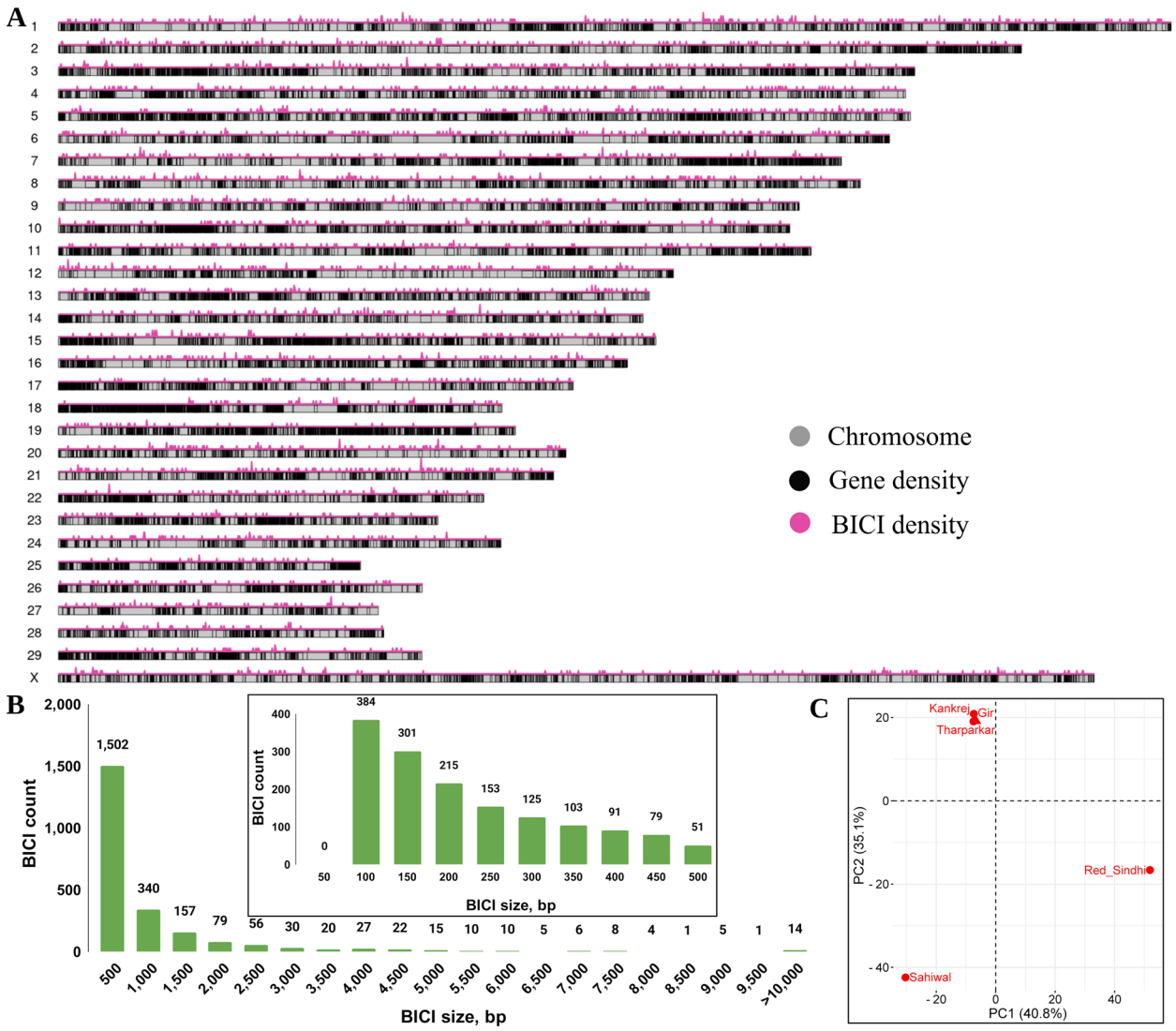

Fig. 3 Overview of *Bos indicus* Common Insertions (BICIs) and their distribution patterns. **A** This ideogram depicts BICIs occurrences across Brahman genome chromosomes. Pink histograms above each chromosome illustrate BICI density using a 100 kb window size. The black line within chromosomes represents gene distribution. **B** The plot illustrates the size distribution of BICIs with a bin size of 500 bp, providing an overview and 50 bp in the zoomed area. **C** The first two principal components are based on the BICI occurrence matrix

the dominance of LINEs and the influence of BICI size on repeat content.

We additionally characterized the repetitive sequences in the flanking regions of BICIs and found that approximately 70% of the BICIs were flanked by a TE on at least one end. This underscores the association of BICIs with TE sequences in their vicinity. Furthermore, we found that around 50% of BICI breakpoint crossing over sequences were encompassed by TEs, emphasizing the substantial impact of TEs on BICI breakpoints (Fig. 4B). This dual observation underscores the intricate interplay between BICIs and TEs in shaping the cattle genome, shedding light on their interdependencies and potential functional roles.

Gene coding and transcriptional potential of BICIs

The genomic distribution and classification of BICIs were examined by pinpointing their breakpoints within the Brahman reference genome annotation. Out of the 2,312 BICIs in our dataset, 926 BICIs were positioned within genic regions, while the remaining 1,386 BICIs were situated outside of annotated genes (Table S10). A detailed analysis of these BICIs revealed that 26 of them were in exons, highlighting their potential to influence protein-coding sequences. The presence of BICIs was

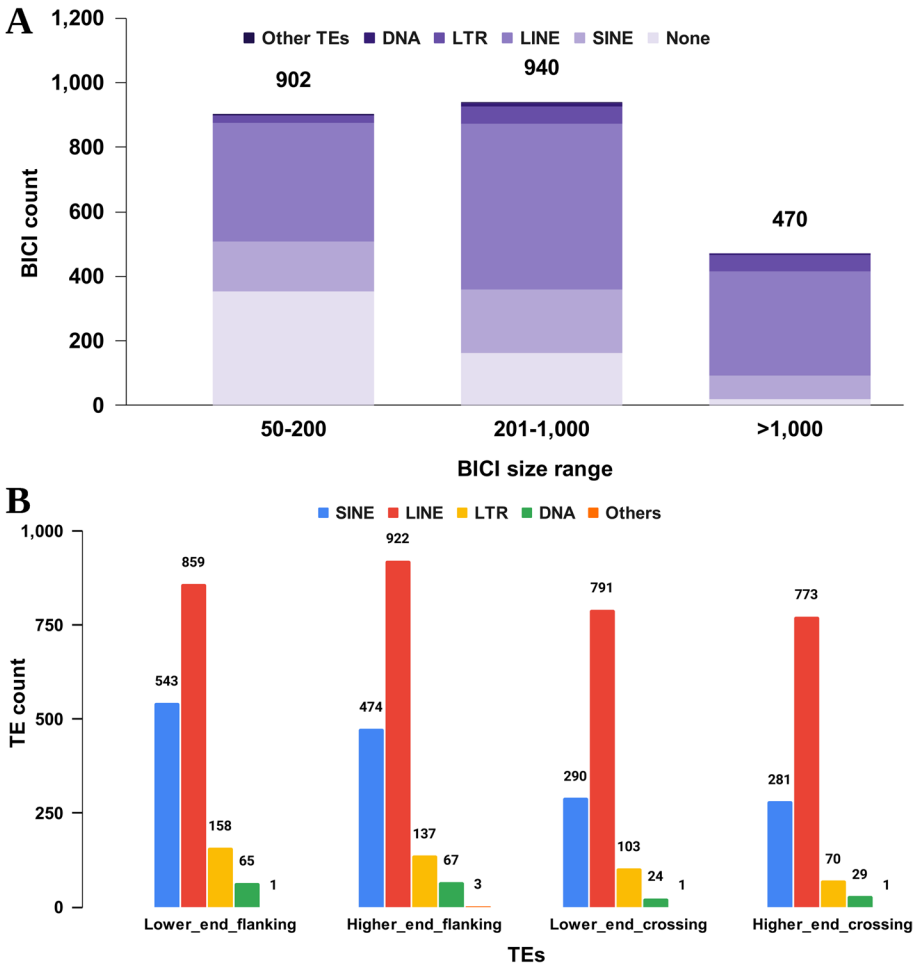

Fig. 4 Distribution of transposable elements on *Bos indicus* Common Insertions (BICIs). **A** Stacked bars represent the total number of BICIs split by three different size ranges. The major transposable elements (TEs) are categorized as SINE (Short Interspersed Nuclear Element), LTR (Long Terminal Repeat), LINE (Long Interspersed Nuclear Element), DNA (DNA Transposon), and NONE (No Interspersed Repeat Detected). "Other TEs" encompasses various minor classes. **B** Bar plot showing the number of TEs flanking and crossing BICIs

associated with a total of 754 genes, demonstrating the diversity of genic regions impacted by these insertions. Remarkably, 121 genes within this group contained two or more BICIs within their span, underscoring the presence of multiple BICIs within specific genomic loci. The majority of genes linked with BICIs were classified as protein-coding genes, strengthening the notion that BICIs may have a functional impact on protein-coding sequences. Additionally, BICIs were found within 54 long non-coding RNA (lncRNA), while few BICIs were associated with 18 pseudogenes (Table 4). These findings provide a comprehensive perspective on the distribution and potential implications of BICIs in various genomic contexts, including protein-coding regions, non-coding sequences, and pseudogenes.

Table 4 Statistics of BICIs annotated in genes

Features	Number
BICIs in genes	926
BICIs in exons	26
BICIs in introns	900
BICIs Intergenic	1,386
Total non-redundant genes	754
Protein coding genes	682
LncRNAs	54
Pseudogenes	18

While the number of BICIs situated within exons is vanishingly small, approximately 0.02%, it remains plausible that some may represent previously unseen exons or regulatory elements with transcriptional potential. To address this, we employed both evidence-based and ab initio methods to assess the transcriptional potential of the BICI dataset. Utilizing high-quality RNA-seq

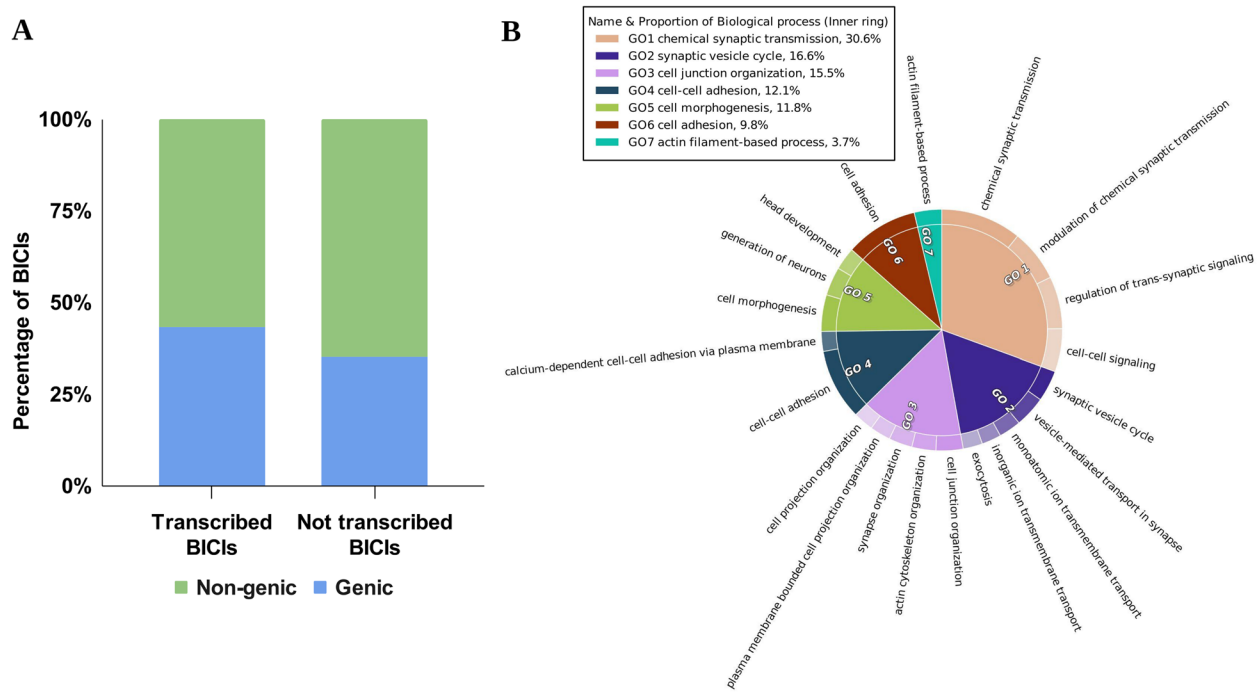

Fig. 5 Characterization of *Bos indicus* Common Insertions (BICIs) in the transcriptome. **A** Stacked bars depict the distribution of BICIs across genic and non-genic regions of the genome. **B** Gene Ontology (GO) annotation analysis for the biological processes associated with BICIs

data from 47 in-house samples derived from PBMCs for evidence-based annotation, we identified 1,631 transcripts originating from 1,395 BICIs. Among these, 662 transcripts from 602 BICIs were located within genic regions, while 969 transcripts were associated with 793 BICIs located outside the genic regions of the reference genome (Fig. 5A, Table S11). Furthermore, our ab initio analysis revealed the transcription potential of 148 genes from 124 BICIs. Of these, 137 were supported by transcripts, with 37 genes from 29 BICIs situated within genic regions and 100 transcripts from 85 BICIs having breakpoints outside the genic regions (Table S12). Of the 11 genes associated with 10 BICIs that lacked transcript support, 7 BICIs were situated in non-genic regions and 3 BICIs were located within genic regions of the Brahman reference genome.

Impact of BICIs in functional genome

The functional analysis of protein coding genes with BICI unveiled enrichment in genes predicted to be associated with key biological processes. Notable enriched terms included chemical synaptic transmission, cell junction organization, cell-cell adhesion, and cell morphogenesis (Fig. 5B). Simultaneously, the examination of major cellular components highlighted involvement in the synapse, plasma membrane region, and monatomic ion channel complex (Fig. S6A). Furthermore, the GO enrichment analysis for molecular functions indicated enrichment in ion channel activity, glutamate receptor activity, cell adhesion-mediated activity, and 3′,5′-cyclic-AMP phosphodiesterase activity (Fig. S6B).

QTL analysis was conducted on 682 protein coding genes with BICIs, revealing approximately 3,368 QTLs associated with 539 genes. Five genes lacked a corresponding location in the ARS-UCD1.2 reference genome [22] and 138 genes did not exhibit enrichment with the cattle QTL database. As expected, several QTLs were associated with individual genes. These QTLs were categorized into six main classes: milk production, reproduction, exterior traits, health, meat production, and carcass traits. Notably, milk production traits (28.7%) were the most enriched category, followed by exterior traits (17.5%) (Fig. 6A). Enrichment analysis further revealed that top QTLs predominantly influenced milk traits (milk yield, milk fat yield, milk fat percentage, milk protein yield, and milk protein percentage) and production traits (body weight, and body weight gain) (Fig. 6B). The integration of functional analysis with QTL mapping revealed a strong link between genes with BICIs and economically important traits in *Bos indicus* cattle. Our analysis of common genes within enriched GO terms further revealed overlap with enriched QTLs. Notably, genes like *CTNNA3* and *CTNNA2*, enriched in GO terms for 'cell adhesion' and 'cell migration' respectively,

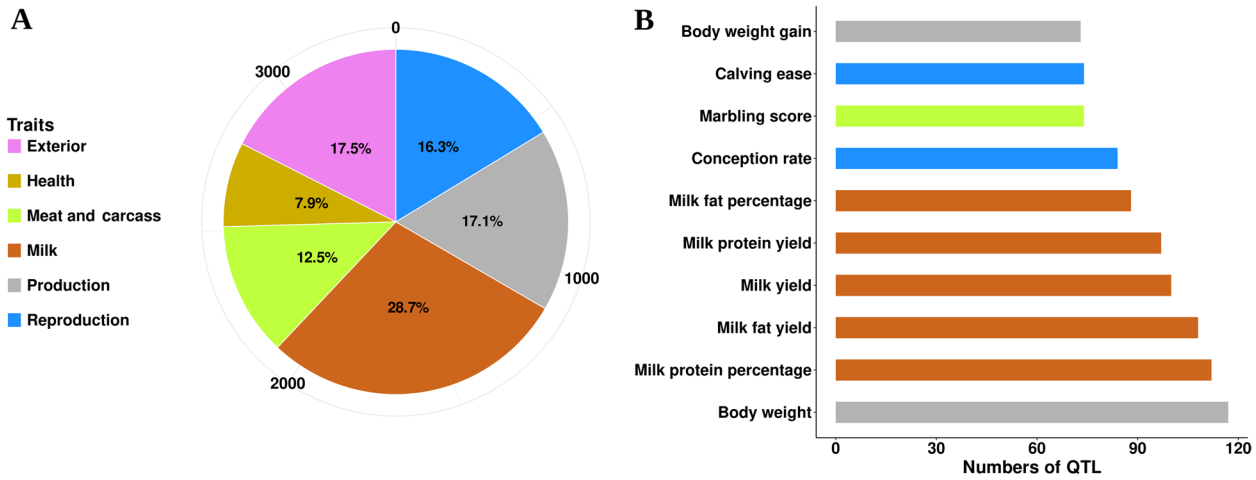

Fig. 6 Trait enrichment analysis of BICIs in known cattle QTL regions. **A** Percentage of QTL type (pie chart) associated with BICIs. **B** Top 10 enriched QTL traits (bar plots) associated with BICIs

reside within the milk QTL region. Similarly, the *ROBO1* gene, enriched in GO terms related to 'cell adhesion' and 'migration', is also present in the reproduction trait QTL. Furthermore, major genes like *ITPR1* and *ITPR2*, enriched in the GO terms related to 'cell morphogenesis', 'calcium ion transmembrane transport', and 'response to hypoxia', are identified within the body weight QTL region.

Evolutionary analysis of the BICIs

To elucidate the evolutionary origins of the BICIs, we conducted alignment analyses with sister species of the *Bos* genus and other evolutionarily related species within the Bovidae family. Specifically, we aligned BICIs to five *Bos* species, which included *Bos taurus* (taurine cattle), *Bos gaurus* (gaur), *Bos frontalis* (gayal), *Bos grunniens* (domestic yak), and *Bos mutus* (wild yak). Out of the total 2,312 BICIs, a substantial proportion, 1,625 (70.28%), was identified across the *Bos* sister species. Each sister species exhibited alignment with over 40% of the BICIs. Specifically, *Bos taurus* aligned with 1,002 BICIs (43.4%), *Bos gaurus* with 1,172 BICIs (50.7%), *Bos frontalis* with 950 BICIs (41.1%), *Bos grunniens* with 1,152 BICIs (49.8%), and *Bos mutus* with 1,156 BICIs (50%) (Fig. 7A). Notably, most BICIs were not exclusive to a single sister species but were shared across multiple sister species. Specifically, 467 (20.2%) BICIs were common to all five sister species, while 415 BICIs were shared with at least four sister species. Additionally, there were a few BICIs that were unique to specific sister species, including *Bos taurus* (218), *Bos gaurus* (51), *Bos frontalis* (28), *Bos grunniens* (13), and *Bos mutus* (3). This distribution of unique BICIs aligns with the phylogenetic tree and genetic introgression observed in the *Bos* genus as previously described by Wu et al. [73].

Similarly, when BICIs were aligned with more distantly related species within the Bovidae family, such as American bison (*Bison bison*), eland (*Tragelaphus oryx*), buffalo (*Bubalus bubalis*), sheep (*Ovis aries*), and goat (*Capra hircus*), only 1,121 BICIs aligned to any of these genomes, which accounted for less than 50% of the BICIs. Most BICIs matched with the American bison (992), followed by buffalo (680), eland (176), sheep (145), and goat (144) (Fig. 7B). Regarding the distribution of matched BICIs, the lowest number of BICIs (51) were shared across all five genomes, followed by 77 shared across four genomes, 102 shared across three genomes, 377 shared across two genomes, and 514 were unique to specific species. Particularly noteworthy is the prevalence of unique BICIs in the bison genome (411), followed by buffalo (101) and eland (2). Goat and sheep did not exhibit any unique BICIs within their genomes. This distinctive BICI distribution across these species highlights their evolutionary divergence and lineage-specific enrichment of BICIs.

Discussion

The present study is dedicated to uncovering the presence and characteristics of NUIs, representing a substantial segment of the cattle pangenome that is absent in the *Bos indicus* reference genome. To unveil these NUIs, the genomes of five *desi* cattle breeds, Gir, Tharparkar, Kankrej, Sahiwal, and Red Sindhi, were sequenced utilizing 10X Chromium technology [33]. The length of these genome assemblies were comparable to the published Brahman reference genome and other cattle genomes assembled using long-reads [19, 22, 32, 74]. However, it was evident that the contiguity of scaffolds

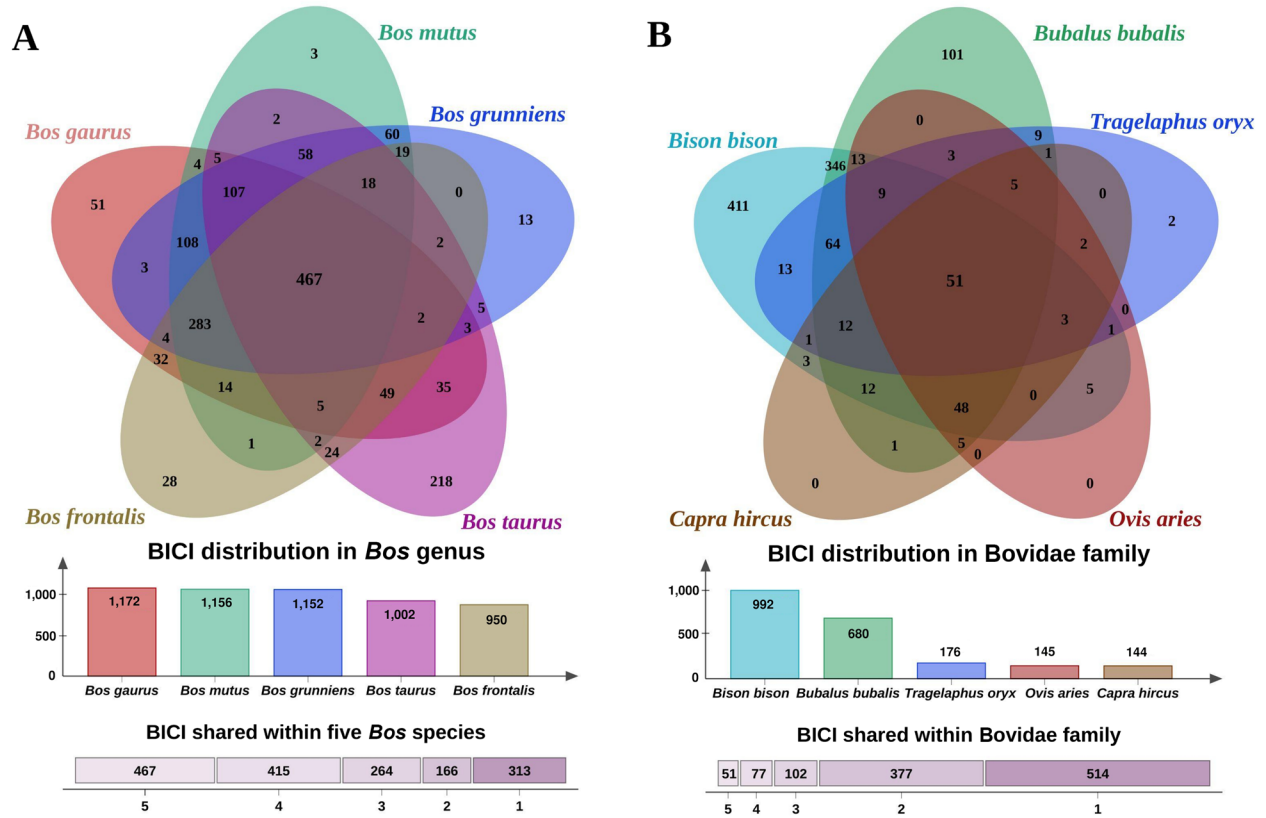

Fig. 7 Evolutionary analysis of *Bos indicus* Common Insertions (BICIs). **A** Venn diagram illustrating the number of BICIs shared within the *Bos* genus. The overlapping regions reveal the extent of shared BICIs among species within the *Bos* genus. **B** Venn diagram showcasing the number of BICIs shared within the Bovidae family. The intersecting areas depict the shared BICIs among species within the broader Bovidae family

in the assemblies varied. This variance can be attributed to the intricacies of linked-read library preparations [33, 34]. Notably, the assembly size was found to be contingent upon the molecular size selected for linked-read preparation.

The evolution of methods for comparing genomes with reference sequences to identify missing segments has been noteworthy. Initially, direct alignment-based methods were employed for NUI detection; however, graph-based methods have gained prominence for their purported comprehensiveness. In our study, we utilized a pipeline specifically tailored for linked-reads, developed by Wong et al. [29] for diploid species and applied to humans. Concurrently, we constructed the pangenome using a graph-based approach, revealing a more extensive set of NUIs spanning 17.7 Mb compared to the 8 Mb identified through the NUI discovery pipeline. Although the pipeline relies on the indirect comparison of five genomes by aligning short reads, it is noteworthy that, out of the 8 Mb identified, 7.5 Mb was also corroborated by the graph-based approach. This observation underscores the stringency of the pipeline, ensuring the identification of a confident set of NUIs. The advantage of the pipeline lies in its ability to precisely report all NUIs for which breakpoints in the reference genome are identified. It accurately pinpoints the location of these NUIs, enhancing confidence in their identification. While alignment methods have traditionally dominated NUI detection, the prevalent trend in pangenome approaches involves the utilization of graph-based methods. This study contributes to the growing body of evidence supporting the effectiveness of both alignment and graph-based methods, emphasizing their complementary roles in constructing a comprehensive understanding of the cattle pangenome.

It is imperative to underscore the distinctive focus of our study, which centers on the Brahman reference genome and marks the pioneering effort in constructing a pangenome specific to *desi* cattle of the subspecies *Bos indicus*. This exclusive emphasis on a specific subspecies serves as a departure from other recent cattle studies, such as those conducted by Zhou et al. [12] and Crysnanto et al. [23], which reported larger pangenome sizes of 83 Mb and 70 Mb respectively. The contrasting sizes reported in these studies can be attributed to their broader inclusion of diverse breeds, encompassing not

only different subspecies but also wild relatives. Notably, the inclusion of a greater number of distant breeds tends to augment the overall pangenome size by capturing a more extensive array of genetic variations.

The identification of NUIs and the construction of a pangenome significantly enhance transcriptomic studies. Our findings demonstrate that the inclusion of NUIs in the reference genome leads to improved RNA-seq mapping efficiency, highlighting the presence of novel, expressed genes within these sequences. This observation aligns with results from previous studies [23, 75], further emphasizing the importance of pangenome-based analyses. Moreover, genes within NUIs can influence differential gene expression outcomes. In our study of heat-stressed Hariana cattle, we identified 15 DEGs associated with NUIs that were absent from the Brahman reference genome. These DEGs include critical stress response genes, such as Mediator of RNA polymerase II transcription subunit 13 [76], *hnRNP A1* [77] and *PHLDA1*, the latter of which has been linked to heat-induced cell death in spermatogenic cells [78], indicating a potential negative effect of heat stress on male fertility in cattle. Interestingly, many of the DEGs associated with NUIs exhibited strong homology to genes in closely related bovine species, such as *Bos javanicus*, *Bubalus bubalis*, and *Bubalus kerabau*, all of which are native to tropical environments and known for their heat tolerance. This suggests that these genes may play a role in the adaptation of *desi* cattle to heat stress. Our pangenome approach uncovered functionally active and biologically relevant genomic features that were missing from the Brahman reference genome. By incorporating NUIs into future transcriptomic studies, we can gain a more comprehensive understanding of gene expression and its role in complex traits.

NUIs identified in the study were subjected to a comprehensive characterization based on the presence and absence variations in a population of 5 *desi* breeds. The focus was on identifying NUIs that displayed frequent and common occurrences, as such variants often play a more substantial role in influencing traits and diseases. Common variants not only exhibit more reliable associations in genetic studies but also possess the potential for broader implications, given their prevalence in diverse populations [79–82]. Consequently, 2,312 NUIs, spanning approximately 2 Mb, were selected as common NUIs and referred to as BICIs, with a MAF > 0.05. It is noteworthy that the majority of the identified NUIs in our study were not classified as common, a pattern that aligns with findings in the African pangenome, where a significant portion of variants was reported as private to individuals. Additionally, in another study focusing on the *Bos indicus* pangenome using short reads (unpublished), a substantial number of NUIs were identified as private to specific samples. Importantly, this filtering process also led to the exclusion of NUIs that were present across all samples. These are the NUIs that are only absent in the Brahman genome, suggesting that the Brahman, an indicus breed from the USA, exhibits genetic distinctions from other Indian indicus breeds. Furthermore, this observation raises the possibility that some NUIs may be present in the Brahman genome but not captured during the assembly process.

The identification of a higher number of NUIs in Red Sindhi and Sahiwal assemblies underscores the significance of a contiguous genome for accurate NUI detection. The presence of BICIs within assemblies facilitates a clear differentiation between breeds, as demonstrated through PCA where animals were separated based on their geographical origin. Genotyping these BICIs in the population revealed a clustering of Red Sindhi, Gir, and Sahiwal, pointing to a genetic distinctiveness among these breeds. However, Tharparkar and Kankrej exhibited a more mixed clustering pattern, suggesting potential genetic similarities or variations between these breeds, which is noteworthy given their geographical proximity. This proximity might contribute to shared genetic traits or interbreed variations, influencing their genomic profiles. The unique clustering pattern observed not only emphasizes the utility of BICIs as markers for capturing and characterizing genetic diversity between cattle breeds but also highlights their practical application for breed differentiation. The regional clustering aligns with findings from other studies employing different marker systems, providing further validation for the consistency of breed-specific genomic signatures.

The exploration into repeats and TEs within BICIs reveals intriguing insights into the cattle genome. In contrast to humans, where SINEs dominate non-reference insertions [29], cattle exhibit a prevalence of LINEs in BICIs. The presence of TEs in the flanking regions of BICIs mirrors human patterns [29], with approximately 70% of BICIs flanked by TEs at least on one end. This shared dynamic emphasizes the role of TEs in both BICIs and their adjacent sequences. The significance of the enrichment of TEs in non-reference insertions has been noted in various reports [83–85], indicating their potential role in BICIs genesis. The notable enrichment of LINEs in cattle prompts further exploration of their functional implications in shaping the cattle genome, building on previous suggestions regarding the significance of TEs in BICIs genesis.

The exploration of transcriptional potential through both ab initio [86, 87] and evidence-based methods [52, 88] establishes that BICIs exhibit transcriptional activity. The predominant annotation of BICIs with

protein-coding genes underscores their potential functional roles. Concurrently, the annotation of BICIs within lncRNA genes signifies their possible regulatory roles in the genome [89, 90]. The presence of BICIs associated with pseudogenes further hints at their diverse functional repertoire [91]. Further, the identification of BICIs within intronic regions, displaying transcriptional potential, suggests the possibility of contributing additional sequences to existing transcripts. The prevalence of BICIs within genes is a common phenomenon, and the observed variation in their presence may contribute to alterations in gene length within a population [16]. In summary, the transcriptional profiling of BICIs provides compelling evidence for their functional significance, with potential roles in both coding and non-coding genomic elements. The diverse genomic locations and associations with different gene types underscore the intricate and multifaceted impact of BICIs on the regulatory landscape of the genome.

Genes found in the top enriched GO Biological Process (BP) categories also exhibit significant enrichment within major QTL regions, supporting a functional link between biological processes and genetic variation. *CTNNA3* plays a crucial role in the formation of cell-cell adhesion complexes, potentially influencing milk production, milk protein content percentage, milk protein yield, milk fat content percentage, and milk fat yield through its role in mammary gland development [92]. *CTNNA2*, found to be positively selected in the Gir dairy breed [93], plays a role in the development of the nervous system and has also shown association with climate adaptation in Mediterranean cattle [94]. The *ROBO1* gene is enriched in GO terms related to 'cell adhesion' and 'migration', potentially impacting fertility and ovarian health [95]. *ITPR1* has been reported to be associated with environmental high-altitude adaptation in the yak [96], potentially influencing body weight regulation. Additionally, *ITPR2* has been linked to heat stress response in US Holsteins [97]. These findings suggest a potential role for these genes in mediating various phenotypic traits through their involvement in crucial biological processes.

The comprehensive alignment analyses aimed at unraveling the evolutionary origins of BICIs provided valuable insights into their distribution among sister species within the *Bos* genus and other evolutionarily related species in the Bovidae family. Notably, a substantial proportion (70.28%) of BICIs was identified across the *Bos* sister species, including *Bos taurus* (exotic cattle), *Bos gaurus* (Gaur), *Bos frontalis* (Gayal), *Bos grunniens* (Domestic Yak), and *Bos mutus* (wild Yak). Interestingly, the observation that the number of BICIs shared solely with *Bos taurus* was not the highest of challenges expectations. The lower total number of BICIs found in *Bos taurus* compared to other wild relatives suggests that *Bos taurus* may have undergone artificial selection [98], resulting in the loss of many BICIs. This is particularly evident in the subset of 283 BICIs shared with all other sister species. Further examination of distantly related species within the Bovidae family, such as American bison, eland, buffalo, sheep, and goat, revealed a varied distribution of matched BICIs, aligning with the inferred evolutionary divergence times based on phylogenetics [73]. This trend of shared and unique BICIs across species is reminiscent of findings in human genomics when NUIs of humans were compared with other species like chimpanzee, gorilla, orangutan, and bonobo [29]. Such comparative analyses not only highlight the evolutionary dynamics within the Bovidae family but also draw parallels with similar studies in different species, providing a broader understanding of the genomic changes accompanying evolutionary divergence.

Conclusions

This study addresses the crucial need to explore and comprehend the genomic diversity within the *Bos indicus* population, with a specific focus on dairy breeds in India. The construction of the *Bos indicus* pangenome, employing alignment and graph-based methods, revealed significant differences in size, emphasizing the importance of diverse approaches. A robust set of NUIs spanning 7.8 Mb which are common to both methods, defines the pangenome of Indian *Bos indicus* breeds. Comparative analyses showcased distinctions with other pangenomes, highlighting the unique genomic landscape of these dairy breeds. The identification of BICIs, particularly within protein-coding genes enriched for specific functions, provided insights into potential roles in various traits related to milk production, reproduction, and health. A substantial proportion of BICIs is shared with both domesticated and wild species, underlining their origin and evolutionary significance in present day cattle. In summary, our study provides valuable new resources, encompassing linked-reads, de novo assemblies, NUIs, BICIs and pangenomic analyses, for future bovine pangenome research.

Abbreviations

BAM	Binary alignment map
BICI	*Bos indicus* Common Insertion
bp	Base pair
CDS	Coding sequence
DEG	Differentially expressed gene
FDR	False discovery rate
GB	Gigabyte
Gb	Gigabase
gfa	Graphical fragment assembly
GO	Gene Ontology
kb	Kilobase
MAF	Minor allele frequency

Mb	Megabase
NCBI	National Center for Biotechnology Information
NUI	Non-reference Unique Insertion
PCA	Principal component analysis
SNP	Single nucleotide polymorphism
SV	Structural variation
TE	Transposable element
QTL	Quantitative trait locus

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40104-024-01133-1.

Additional file 1: Table S1. Data statistics of five *Bos indicus* de novo pseudohaplotype assemblies. Table S2. NUI occurrence matrix from NUI discovery pipeline. Table S3. NUI breakpoints in the UOA_Brahman_1 reference genome overlapping with *B. taurus* introgressed segments. Table S4. NUI in common with published pangenomes. Table S5. Transcriptome read mapping (%) to reference vs. pangenome. Table S6. Comparison of top gene expression under heat stress as reported by Sajjanar et al. (2023) and identified in the pangenome. Table S7. BLASTx results of NUI novel genes against the NCBI non-redundant (nr) chordata database. Table S8. BICIs occurrence matrix. Table S9. Overall repeat content in BICIs. Table S10. BICIs annotation in genic and non-genic region. Table S11. BICIs supported by transcript. Table S12. BICIs supported by AUGUSTUS and transcript. Table S13. Assembly accessions ID. Table S14. WGS accessions included in the study. Table S15. Transcriptome accessions included in the study.

Additional file 2: Fig S1. Upset plot for the shared NUIs among five *Bos indicus* breeds from NUI discovery pipeline. Fig S2. Upset plot for the shared NUIs among five *Bos indicus* breeds from minigraph pipeline. Fig S3. Mapping rate of transcriptome sequencing reads to Brahman reference and pangenome. Fig S4. Volcano plot of differentially expressed NUI genes. Fig S5. Cladogram of 98 samples using presence and absence of NUIs. Fig S6. GO annotation of BICIs.

Additional file 3: NUI discovery pipeline fasta file.

Additional file 4: Minigraph pipeline fasta file.

Additional file 5: Final NUI set fasta file.

Additional file 6: BICI fasta file.

Acknowledgements
The authors sincerely express their gratitude to the Department of Biotechnology (DBT), Ministry of Science, New Delhi, India for providing financial support. Additionally, the authors extend their appreciation to the National Institute of Animal Biotechnology (NIAB) for offering invaluable support throughout the execution of the study. SA specifically acknowledges and appreciates the support received from Dr. G. Taru Sharma, Director, NIAB. MN, CPVT, and BDR were supported by the appropriated project 8042-31000-112-000-D, "Accelerating Genetic Improvement of Ruminants Through Enhanced Genome Assembly, Annotation, and Selection" of the USDA Agricultural Research Service. Any mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Authors' contributions
SA, BDR, SNR and SSM designed the study. SSM, RKG and SA facilitated sample collections and sequencing. AS and NKP assembled the genomes and established the contamination screening pipeline. SA and AS analyzed the data for NUIs identification through NUI discovery pipeline, performed minigraph analysis, genotyping of NUIs in the population, PCA, TE analysis, GO analysis and evolutionary analysis. MN performed QTL analysis. RKG, CPVT, BDR and SA drafted the manuscript. SSM, SNR and CPVT edited the manuscript. All authors reviewed the final manuscript before submission.

Funding
The linked-read data were generated in the project "Genomics for Conservation of Indigenous Cattle Breeds and for Enhancing Milk Yield, Phase-I" [BT/PR26466/AAQ/1/704/2017], funded by the Department of Biotechnology (DBT), India. Transcriptomics data were generated in the project "Identification of key molecular factors involved in resistance/susceptibility to paratuberculosis infection in indigenous breeds of cows" [BT/PR32758/AAQ/1/760/2019], which was also funded by Department of Biotechnology (DBT), India.

Data availability
The linked-read data generated and reported in this article are accessible from the Indian Biological Data Centre (IBDC) with the INSDC accession number mentioned in the (Table S13) Additionally, the Supernova assemblies generated from the raw linked-reads are also available on IBDC. The Illumina sequencing data used in this study have been submitted to IBDC, and all accession numbers are detailed in Table S14. RNAseq data with all accession numbers detailed in Table S15. Given that IBDC is part of the INSDC, all the data can be accessed from the NCBI as well.

Declarations

Ethics approval and consent to participate
Cattle samples were obtained in accordance with the guidelines set forth by the Committee for the Purpose of Control and Supervision on Experiments on Animals (CPCSEA), India with the Institutional Animal Ethics Committee (IAEC) approval.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]National Institute of Animal Biotechnology, Hyderabad, India. [2]Indian Institute of Technology Hyderabad, Sangareddy, India. [3]Animal Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, USA. [4]Animal Biotechnology, ICAR-NBAGR, Karnal, Haryana, India.

Received: 15 August 2024 Accepted: 26 November 2024
Published online: 07 February 2025

References

1. Taye M, Lee W, Jeon S, Yoon J, Dessie T, Hanotte O, et al. Exploring evidence of positive selection signatures in cattle breeds selected for different traits. Mamm Genome. 2017;28:528–41.
2. Gilbert M, Nicolas G, Cinardi G, Van Boeckel TP, Vanwambeke SO, Wint GRW, et al. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Sci Data. 2018;5:180227.
3. Pitt D, Sevane N, Nicolazzi EL, MacHugh DE, Park SDE, Colli L, et al. Domestication of cattle: two or three events? Evol Appl. 2019;12:123–36.
4. Ajmone-Marsan P, Garcia JF, Lenstra JA. On the origin of cattle: how aurochs became cattle and colonized the world. Evol Anthropol. 2010;19:148–57.
5. Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P. Evidence for two independent domestications of cattle. Proc Natl Acad Sci U S A. 1994;91:2757–61.
6. Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci U S A. 2014;111:6139–46.
7. Papachristou D, Koutsouli P, Laliotis GP, Kunz E, Upadhyay M, Seichter D, et al. Genomic diversity and population structure of the indigenous Greek and Cypriot cattle populations. Genet Sel Evol. 2020;52:43.
8. Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. Nat Commun. 2018;9:2337.

9. Bolormaa S, Pryce JE, Kemper KE, Hayes BJ, Zhang Y, Tier B, et al. Detection of quantitative trait loci in *Bos indicus* and *Bos taurus* cattle using genome-wide association studies. Genet Sel Evol. 2013;45:43.
10. Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet. 2007;39:S7–15.
11. Hayes BJ, Daetwyler HD. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. Annu Rev Anim Biosci. 2019;7:89–102.
12. Zhou Y, Yang L, Han X, Han J, Hu Y, Li F, et al. Assembly of a pangenome for global cattle reveals missing sequences and novel structural variations, providing new insights into their diversity and evolutionary history. Genome Res. 2022;32:1585–601.
13. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol. 2008;11:472–7.
14. McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. Microb Genom. 2019;5:5. https://doi.org/10.1099/mgen.0.000243.
15. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32:1045–52.
16. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet. 2019;51:30–5.
17. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 2019;20:149.
18. Leonard AS, Crysnanto D, Fang ZH, Heaton MP, Vander Ley BL, Herrera C, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nat Commun. 2022;13:3012.
19. Dai X, Bian P, Hu D, Luo F, Huang Y, Jiao S, et al. A Chinese indicine pangenome reveals a wealth of novel structural variants introgressed from other *Bos* species. Genome Res. 2023;33:1284–98.
20. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018;36:875–9.
21. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 2020;21:265.
22. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9:9. https://doi.org/10.1093/gigascience/giaa021.
23. Crysnanto D, Leonard AS, Fang ZH, Pausch H. Novel functional sequences uncovered through a bovine multiassembly graph. Proc Natl Acad Sci USA. 2021;118:e2101056118. https://doi.org/10.1073/pnas.2101056118.
24. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci. 2020;25:148–58.
25. Zhuang Y, Wang X, Li X, Hu J, Fan L, Landis JB, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat Plants. 2022;8:233–44.
26. Li N, He Q, Wang J, Wang B, Zhao J, Huang S, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet. 2023;55:852–60.
27. Smith TPL, Bickhart DM, Boichard D, Chamberlain AJ, Djikeng A, Jiang Y, et al. The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biol. 2023;24:139.
28. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016;34:303–11.
29. Wong KHY, Levy-Sakin M, Kwok P-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nat Commun. 2018;9:3040.
30. Ahmad S, Kour G, Singh A, Gulzar M. Animal genetic resources of India - an overview. Int J Livest Res. 2019;9:1–12. https://doi.org/10.5455/ijlr.20181025013931.
31. Masharing N, Sodhi M, Chanda D, Singh I, Vivek P, Tiwari M, et al. ddRAD sequencing based genotyping of six indigenous dairy cattle breeds of India to infer existing genetic diversity and population structure. Sci Rep. 2023;13:9379.
32. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018. https://doi.org/10.1038/nbt.4277.
33. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67.
34. Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, et al. Resolving the full spectrum of human genome variation using linked-reads. Genome Res. 2019;29:635–45.
35. Šerić H, Alkan C, Hach F, Numanagić I. Fast characterization of segmental duplication structure in multiple genome assemblies. Algorithms Mol Biol. 2022;17:4.
36. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:10. https://doi.org/10.1093/gigascience/giab008.
37. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015;31:2032–4.
38. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. https://doi.org/10.48550/arXiv.1303.3997.
39. Harris B. Improved pairwise alignment of genomic DNA. The Pennsylvania State University; 2007.
40. RepeatMasker Open-4.0. http://www.repeatmasker.org. Accessed 20 Apr 2024.
41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. https://doi.org/10.1186/1471-2105-10-421.
43. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503–5.
44. Ripley BD. The R project in statistical computing. MSOR connect. Educational development unit. Univ Greenwich. 2001;1:23–5.
45. FASTX-Toolkit. FASTQ/a short-reads pre-processing tools, http://hannonlab.cshl.edu/fastx_toolkit/index.html. Accessed 13 Mar 2024.
46. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
47. gfatools. Tools for manipulating sequence graphs in the GFA and rGFA formats, https://github.com/lh3/gfatools. Accessed 30 Mar 2024.
48. Fasta2Lineage. https://github.com/gcbl-niab/fasta2lineage. Accessed 11 Apr 2024.
49. Naji MM, Utsunomiya YT, Sölkner J, Rosen BD, Mészáros G. Assessing *Bos taurus* introgression in the UOA *Bos indicus* assembly. Genet Sel Evol. 2021;53:96.
50. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
51. Sajjanar B, Aalam MT, Khan O, Tanuj GN, Sahoo AP, Manjunathareddy GB, et al. Genome-wide expression analysis reveals different heat shock responses in indigenous (*Bos indicus*) and crossbred (*Bos indicus* X *Bos taurus*) cattle. Genes Environ. 2023;45:17.
52. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
53. Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using rsubread and the edgeR quasi-likelihood pipeline. F1000Res. 2016;5:1438.
54. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–890.
55. Vasimuddin M, Misra S, Li H, Aluru S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Rio de Janeiro; 2019. p. 314–24. https://doi.org/10.1109/IPDPS.2019.00041.
56. Bos_indicus_pangenome_10X. https://github.com/gcbl-niab/bos_indicus_pangenome_10x. Accessed 20 June 2024.
57. Tang D, Chen M, Huang X, Zhang G, Zeng L, Zhang G, et al. SRplot: a free online platform for data visualization and graphing. PLoS ONE. 2023;18:e0294236.
58. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.

59. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. The Innovation (Camb). 2021;2:100141.
60. Bioconductor. https://www.bioconductor.org/. Accessed 22 May 2024.
61. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE. 2011;6:e21800.
62. Kuznetsova I, Lugmayr A, Siira SJ, Rackham O, Filipovska A. CirGO: an alternative circular way of visualising gene ontology terms. BMC Bioinformatics. 2019;20:84.
63. Cattle QTL Database. https://www.animalgenome.org/cgi-bin/QTLdb/BT/index. Accessed 20 May 2024.
64. Low WY, Rosen BD, Ren Y, Bickhart DM, To T-H, Martin FJ, et al. Gaur genome reveals expansion of sperm odorant receptors in domesticated cattle. BMC Genomics. 2022;23:344.
65. Gao X, Wang S, Wang Y-F, Li S, Wu S-X, Yan R-G, et al. Long read genome assemblies complemented by single cell RNA-sequencing reveal genetic and cellular mechanisms underlying the adaptive evolution of yak. Nat Commun. 2022;13:4887.
66. Mukherjee S, Cai Z, Mukherjee A, Longkumer I, Mech M, Vupru K, et al. Whole genome sequence and de novo assembly revealed genomic architecture of Indian Mithun (*Bos frontalis*). BMC Genomics. 2019;20:617.
67. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, et al. Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat Commun. 2019;10:260.
68. Chen L, Qiu Q, Jiang Y, Wang K, Lin Z, Li Z, et al. Large-scale ruminant genome sequencing provides insights into their evolution and distinct traits. Science. 2019;364:364. https://doi.org/10.1126/science.aav6202.
69. Bison bison bison genome assembly Bison_UMD1. https://www.ncbi.nlm.nih.gov/data-hub/assembly/gcf_000754665.1/. Accessed 13 Mar 2024.
70. Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, Chan S, et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat Genet. 2017;49:643–50.
71. Davenport KM, Bickhart DM, Worley K, Murali SC, Salavati M, Clark EL, et al. An improved ovine reference genome assembly to facilitate in-depth functional annotation of the sheep genome. Gigascience. 2022;11:11. https://doi.org/10.1093/gigascience/giab096.
72. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. Jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15:293.
73. Wu D-D, Ding X-D, Wang S, Wójcik JM, Zhang Y, Tokarska M, et al. Pervasive introgression facilitated domestication and adaptation in the *Bos* species complex. Nat Ecol Evol. 2018;2:1139–45.
74. Canavez FC, Luche DD, Stothard P, Leite KRM, Sousa-Canavez JM, Plastow G, et al. Genome sequence and assembly of *Bos indicus*. J Hered. 2012;103:342–8.
75. Li R, Fu W, Su R, Tian X, Du D, Zhao Y, et al. Towards the complete goat pan-genome by recovering missing genomic segments from the reference genome. Front Genet. 2019;10:1169. https://doi.org/10.3389/fgene.2019.01169.
76. Sarkar D, Zhu ZI, Knoll ER, Paul E, Landsman D, Morse RH. Mediator dynamics during heat shock in budding yeast. Genome Res. 2022;32:111–23.
77. Guil S, Long JC, Cáceres JF. hnRNP A1 relocalization to the stress granules reflects a role in the stress response. Mol Cell Biol. 2006;26:5744–58.
78. Janus P, Mrowiec K, Vydra N, Widłak P, Toma-Jonik A, Korfanty J, et al. *PHLDA1* does not contribute directly to heat shock-induced apoptosis of spermatocytes. Int J Mol Sci. 2019;21:267.
79. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
80. Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet. 2009;10:381–91.
81. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–53.
82. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
83. Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, et al. Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genet. 2007;3:1939–49.
84. Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. Am J Hum Genet. 2006;79:41–53.
85. Britten RJ. Transposable element insertions have strongly affected human evolution. Proc Natl Acad Sci U S A. 2010;107:19945–8.
86. Hoff KJ, Stanke M. Predicting genes in single genomes with AUGUSTUS. Curr Protoc Bioinf. 2019;65:e57.
87. Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics. 2020;21:293.
88. Yue J-X, Liti G. Long-read sequencing data analysis for yeasts. Nat Protoc. 2018;13:1213–31.
89. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10:155–9.
90. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
91. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. Pseudogenes: pseudo-functional or key regulators in health and disease? RNA. 2011;17:792–8.
92. Ristanic M, Zorc M, Glavinic U, Stevanovic J, Blagojevic J, Maletic M, et al. Genome-wide analysis of milk production traits and selection signatures in Serbian Holstein-Friesian cattle. Animals (Basel). 2024;14:14. https://doi.org/10.3390/ani14050669.
93. Maiorano AM, Lourenco DL, Tsuruta S, Ospina AMT, Stafuzza NB, Masuda Y, et al. Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information. PLoS ONE. 2018;13:e0200694.
94. Flori L, Moazami-Goudarzi K, Alary V, Araba A, Boujenane I, Boushaba N, et al. A genomic map of climate adaptation in Mediterranean cattle breeds. Mol Ecol. 2019;28:1009–29.
95. Ooi E, Xiang R, Chamberlain AJ, Goddard ME. Archetypal clustering reveals physiological mechanisms linking milk yield and fertility in dairy cattle. J Dairy Sci. 2024;107:4726–42. https://doi.org/10.3168/jds.2023-23699.
96. Basang EG-X, Zhu W-D. Whole-genome analysis identifying candidate genes of altitude adaptive ecological thresholds in yak populations. J Anim Breed Genet. 2019;136:371–7.
97. Sigdel A, Abdollahi-Arpanahi R, Aguilar I, Peñagaricano F. Whole genome mapping reveals novel genes and pathways involved in milk production under heat stress in US Holstein cows. Front Genet. 2019;10:928.
98. Cattle breeds: an encyclopedia. 1995. https://www.cabdirect.org/cabdirect/abstract/19950109449. Accessed 20 June 2024.