Naji *et al. BMC Genomics* (2021) 22:108
https://doi.org/10.1186/s12864-021-07412-9

BMC Genomics

RESEARCH ARTICLE Open Access

Investigation of ancestral alleles in the Bovinae subfamily

Maulana M. Naji[1], Yuri T. Utsunomiya[2,3,4,5], Johann Sölkner[1], Benjamin D. Rosen[6*] and Gábor Mészáros[1]

Abstract

Background: In evolutionary theory, divergence and speciation can arise from long periods of reproductive isolation, genetic mutation, selection and environmental adaptation. After divergence, alleles can either persist in their initial state (ancestral allele - AA), co-exist or be replaced by a mutated state (derived alleles -DA). In this study, we aligned whole genome sequences of individuals from the Bovinae subfamily to the cattle reference genome (ARS.UCD-1.2) for defining ancestral alleles necessary for selection signatures study.

Results: Accommodating independent divergent of each lineage from the initial ancestral state, AA were defined based on fixed alleles on at least two groups of yak, bison and gayal-gaur-banteng resulting in ~ 32.4 million variants. Using non-overlapping scanning windows of 10 Kb, we counted the AA observed within taurine and zebu cattle. We focused on the extreme points, regions with top 0. 1% (high count) and regions without any occurrence of AA (null count). High count regions preserved gene functions from ancestral states that are still beneficial in the current condition, while null counts regions were linked to mutated ones. For both cattle, high count regions were associated with basal lipid metabolism, essential for survival of various environmental pressures. Mutated regions were associated to productive traits in taurine, i.e. higher metabolism, cell development and behaviors and in immune response domain for zebu.

Conclusions: Our findings suggest that retaining and losing AA in some regions are varied and made it species-specific with possibility of overlapping as it depends on the selective pressure they had to experience.

Keywords: Ancestral allele, Bovinae, Gene ontology, Whole genome sequences

Background

Divergence and speciation result from long periods of adaptation, selection, and genetic drift after separation of subpopulations. Separation forces individuals to adapt within the current isolated environment and gradually differ from the initial population. Various methodologies and theories have been proposed in efforts for deciphering this process since nineteenth century [1].

Recently, the availability of whole genome sequences (WGS) has become of increasing importance in genetic studies [2]. In cattle studies for example, WGS data of various breeds have been used for inference of demographic history, identification of production traits, calculation of effective population size, estimation of genetic relationships, and population structure analysis [3–5].

In evolutionary analysis, synteny blocks can be inferred as conserved relationships of genomic regions in different species anchored by sets of orthologues genes. With varying size, these blocks can be co-localized in different karyotypes of modern species' respective genomes. Moreover, synteny blocks can be clustered into lineage-specific ones, such as to primates, Rodentia, Felidae, Camelidae, Chiroptera and Bovidae as suggested in a study of syntenic analysis using 87 mammalian genomes [6]. However, orthologous genes within these lineage-specific synteny blocks may present allele variations due

* Correspondence: ben.rosen@usda.gov
[6]Agricultural Research Service USDA, Beltsville, MD, USA
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

to independent evolutionary event after the speciation [7].

Alleles having diverged through mutation are called derived alleles (DA), while alleles that persist in their initial state are termed ancestral alleles (AA) [8]. A reasonable method to assess AA is by comparing shared polymorphic sites of closely related species. Alleles that are still intact and shared by all the related species are most likely the ancestral allele [9]. Another method consists of verifying the allelic state of the last common ancestor (LCA) or the allele within current populations that least differs from the LCA [10].

In a study of autosomal single nucleotide polymorphisms (SNP) in pig, ancestral and derived allelic states of SNP were inferred using four Sus species (*Sus celebensis, Sus barbatus, Sus cebifrons,* and *Sus verrucosus*) and one outgroup species of African warthog for focal species of *Sus scrofa* [11]. In human studies, the outgroup species for inferring AA are primates, namely orangutan (*Pongo sp.*), macaques (*Macaca sp.*), gorilla (*Gorilla sp.*), and bonobos (*Pan paniscus*) [12]. In a cattle study of Utsunomiya et al. (2013) using HD-SNP, Gaur (*Bos gaurus*), water buffalo (*Bubalus bubalis*) and Yak (*Bos grunniens*) were utilized as focal species for cattle.

Defining the ancestral and derived states at polymorphic nucleotide sites is required to test proposed hypotheses regarding molecular evolution processes, such as estimation of allele ages, formation of linkage disequilibrium (LD) patterns and genomic signatures as a result of selection pressures [13, 14]. Human WGS studies benefit from AA database for population analysis, but such a database is lacking in cattle. Consequently, each study repeatedly generates its own putative AA list [5, 12, 15].

Therefore, the goal of this study is to fill this gap and to determine a fixed set of AA in cattle by using outgroup species in the *Bovinae* subfamily, namely gaur, yak, bison, wisent, banteng, and gayal sequences. In addition, we scanned the list of AA for physical regions linked to conserved and mutated traits in taurine and zebu cattle.

Results

Read alignments and principal component analysis

We evaluated alignment results of different species within the Bovinae subfamily against the latest cattle reference sequence ARS-UCD1.2 [16]. On average, the genome was covered by ~5x for banteng, taurine cattle, European bison, gayal, and yak, ~4x for American bison and zebu cattle, and ~ 3x for aurochs. Principle component analysis (PCA) formed clusters and separation of individuals among these nine groups (Fig. 1). Four principal components (PC) explained 36.7, 24.9, 20.5, and 17.7% of the variance for first, second, third, and fourth PC, respectively. Projected by the PC1 and PC2, these Bovinae individuals are clustered together with its closest relatives evidencing genetic relatedness within its sub-species. PC1 explains divergence of cattle (aurochs, zebu, and taurine), from the rest. PC2 gives divergence between cluster containing gayal-gaur-banteng (gagaba) from clusters containing yak and bison. Thus, we can group these individuals into four, namely cattle-aurochs cluster, gagaba cluster, bison cluster, and yak cluster. Outlier individuals, i.e. two gayals and the American

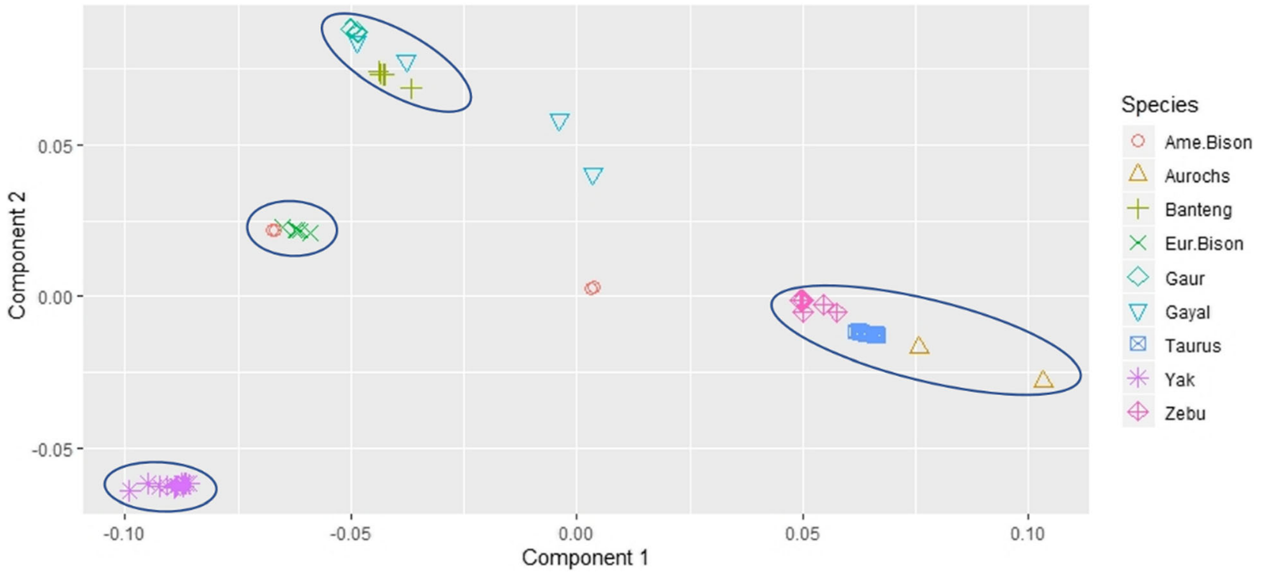

Fig. 1 Principle component analysis

bison, may indicate individuals carrying introgression from cattle.

Phylogenetic trees

Maximum Likelihood phylogenetic trees were constructed for each chromosome [see Additional file 1]. Inferred trees were all similar with Fig. 2 below displaying the tree from chromosome one. In concordance with the principal component analysis, 13 yak individuals are situated together in the top clade of the tree. European bison and American bison have the same node of ancestor, with American bison perceived to be more ancestral. This is in line with a previous study where sister relationships were indicated between American bison and European bison and also between bison clade and yak [17]. Banteng-gaur-gayal share a clade together, however, variations in the order within these three species exist in trees inferred from different chromosomes [see Additional file 1]. Zebu cattle reside on the same upper node with the taurine cattle group. Each breed of taurine

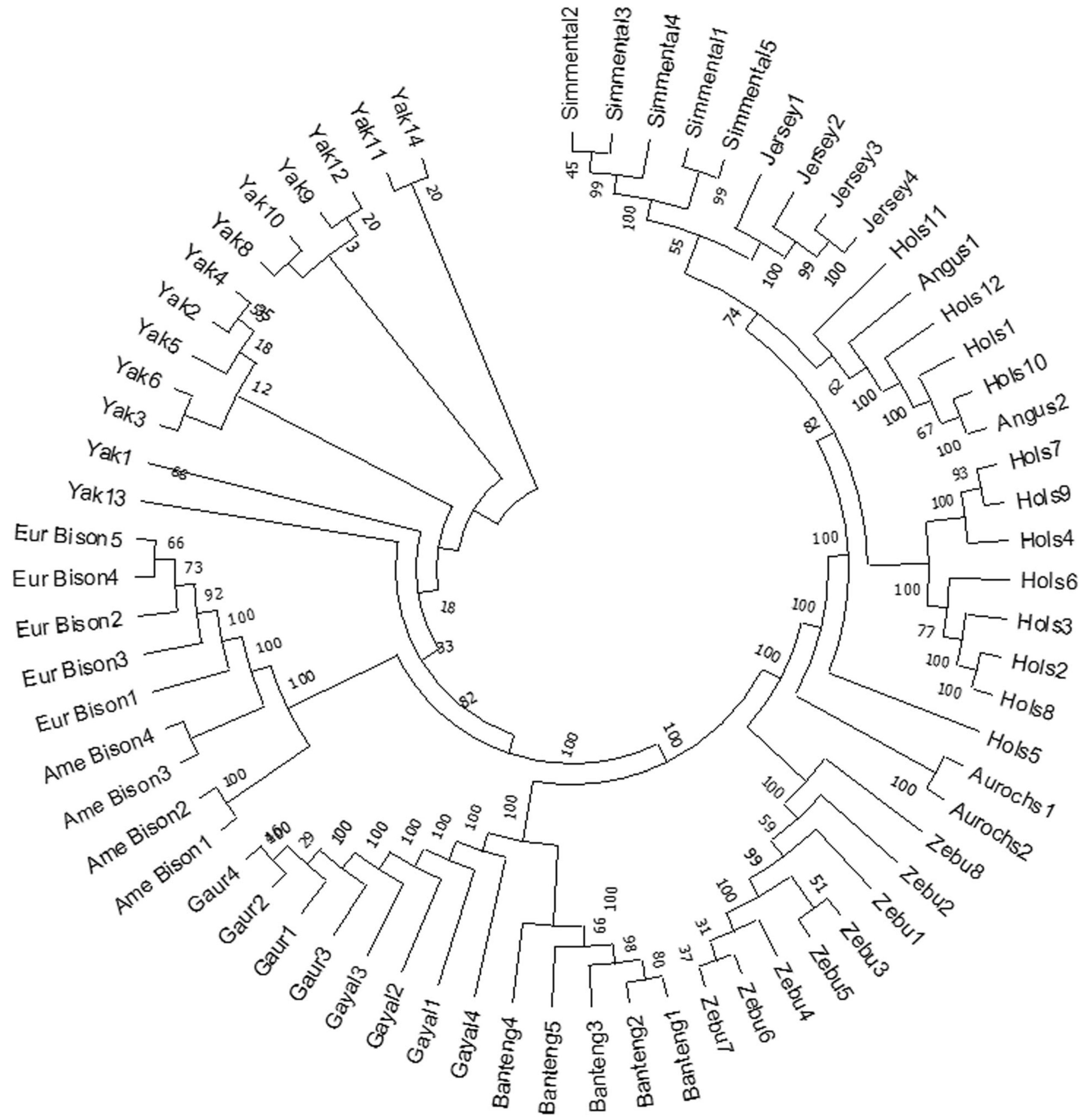

Fig. 2 Phylogenetic tree based on chromosome 1

cattle is well clustered together except for several Holstein individuals. Based on all trees, we defined yak as the most distant relative as it is positioned on the furthest node from cattle.

Inferring ancestral allelic states

The main output of this paper is a list of defined ancestral alleles for cattle, available at https://tinyurl.com/cattle-aa . This list is necessary for several tools used for studying selection signature such as iSAFE, iHS, xp-EHH, EHHST, and hapFLK [18–23] which were built for human population genetics study. We provide this dataset as a foundation for future comparisons of selection signatures in various cattle breeds. It is stored in a simple format of .txt and comprised of 6 columns of chromosome, position, number of alleles, defined ancestral allele, frequency, and which groups agree on the defined ancestral allele. AA were determined as alleles that are fixed in two of three outgroup lineages. Using allele frequency over all individuals in outgroup, we defined ~ 32.4 million variants that are fixed across 29 chromosomes as AA corresponding to 1.2% of the total genome. As shown in Figs. 3, 3.75 million alleles were defined as ancestral from all three lineages of bison, yak, and gayal-gaur-banteng (gagaba). GC content percentage of ancestral alleles is 58%, which is higher than the GC content of the reference genome (~ 42%). Yet, it is worth noting that 22% of these AA are within active transcript regions.

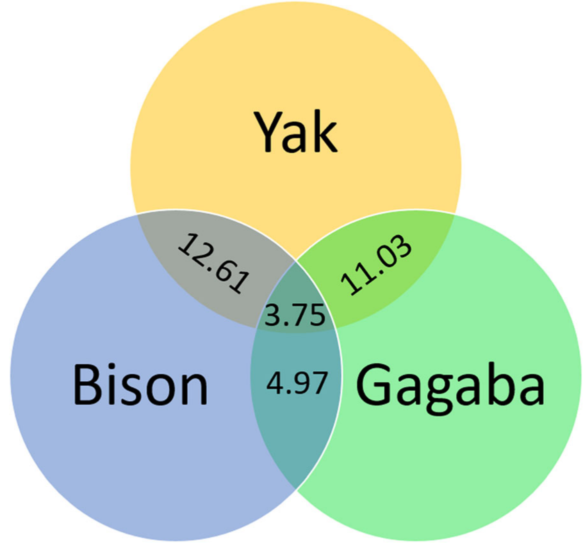

Fig. 3 Intersection of defined ancestral alleles (in millions) from three lineages; bison, yak, and gayal-gaur-banteng (Gagaba)

Windows with high ancestral allele counts in taurine and zebu cattle

We counted AA by non-overlapping windows of 10 Kb in taurine and zebu cattle separately. Figures 4 and 5 present the distribution of AA on chromosome 27 for taurine and zebu, respectively (The distribution of AA for all chromosomes can be found in Additional file 2). For taurine cattle, ancestral allele counts arguably tend to decrease towards the end of chromosome, as demonstrated by the fitted red trend lines. In zebu cattle, ancestral counts are relatively flat throughout the chromosome. Yet, the amplitude pattern is stable for taurine, but more variable for zebu cattle (blue trend line). Peaks of high ancestral alleles count regions in contrast with background averages number of ancestral alleles are clearly distinguished in chromosome 1, 4, 5, 7, 10, 12, 13, 14, 15, 18, 27, 29 in taurine cattle and 1, 2, 3, 4, 6, 10, 12, 13, 14, 15, 18, 23, 27 in zebu cattle [see Additional file 2].

Ancestral counts for the top 0.1% are beyond the mean plus three standard deviations. For taurine cattle, the lowest chromosome specific threshold for ancestral count was 122 on chromosome 25 while the highest was 302 on chromosome 14, while for zebu cattle, it was 102 in chromosome 1 while the highest 200 on chromosome 12. The trends for both groups were similar as shown in Fig. 6. Taurine cattle has mostly higher thresholds implying there are more windows with higher counts of AA compared to zebu cattle.

Windows without the occurrence of ancestral alleles

We found 3306 windows without AA in taurine and 2189 windows in zebu. The highest ratio of windows with null AA counts to total windows was 2.9% on chromosome 29 in taurine and the lowest is 0.14% in chromosome 25 of zebu cattle (Fig. 7). Overall, taurine has more windows without AA except for chromosome 1, 8, 10, and 27. Windows without AA could be explained by a lack of defined AA from outgroups, meaning, there were no fixed alleles that can be found in at least two lineages. Another reason could be that derived alleles are now the major alleles on polymorphic sites, therefore we could not find AA within these windows. In taurine cattle, 65% of windows without AA are due to the latter reason, while in zebu it is 46%.

Annotation of scanning windows with high number of ancestral alleles

We annotated each scanning window passing the respective threshold of top 0.1%, corresponding to 255 regions in taurine and 258 regions in zebu across 29 chromosomes. These regions contained 20 genes in taurine and 40 genes in Zebu. Both groups retained genes functioning in arachidonic acid secretion (GO:0050482),

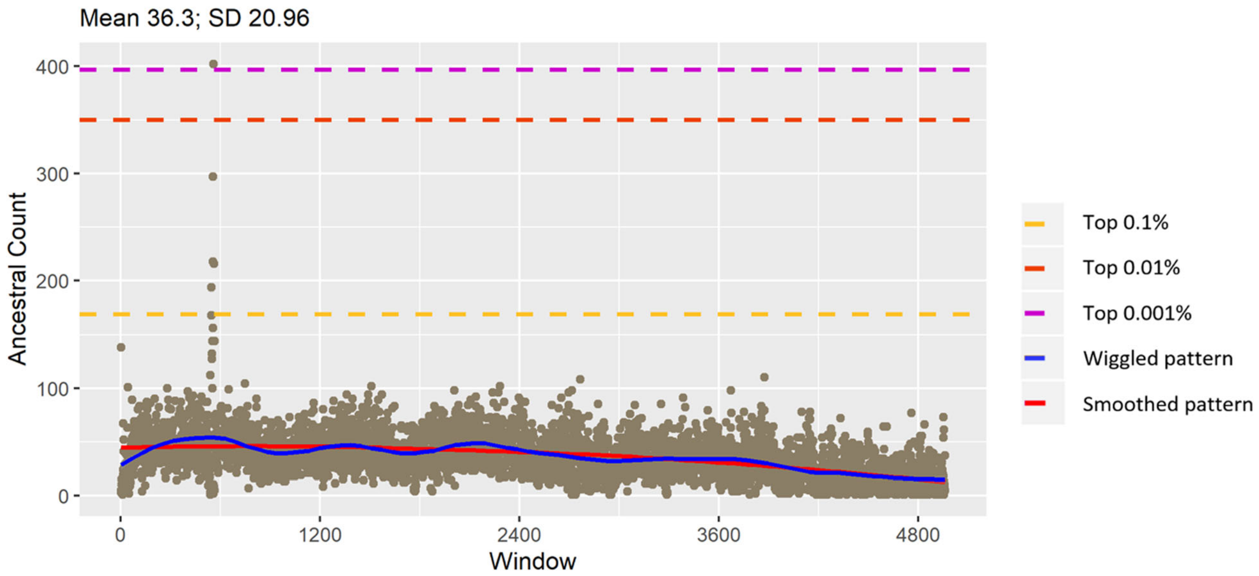

Fig. 4 Distribution of ancestral count in taurine chr 29

phospholipid metabolic process (GO:0006644), and lipid catabolic process (GO:0016042) indicated by LOC100125947 and PLAG2A, as shown in Table 1. These three terms are mainly functioning in primary metabolic process of lipid. Function of defense response to bacterium (GO:0042742) was exclusive to taurine. DEFB genes family in GO:004742 were secreted by leukocytes and epithelial tissues. It is known for its function similar to antimicrobial defense by penetration to microbial's cell membrane and cause microbial death [24]. While calcium ion imports (GO:0070509), represented by SLC8A1 and CACNA1D, was exclusive to zebu defined as function of maintaining and transporting cellular entity in a specific location.

Annotation of scanning windows without ancestral alleles

There were 713 windows in taurine with protein coding genes, while in zebu 121 windows were found. GO terms of regions within scanning windows without AA are attached [see Additional file 3]. There are 42 GO terms defined for taurine and 7 GO terms for zebu. Among those, three terms were found in both, i.e. two antigen processing terms (GO: GO:0002474 and GO:0019882)

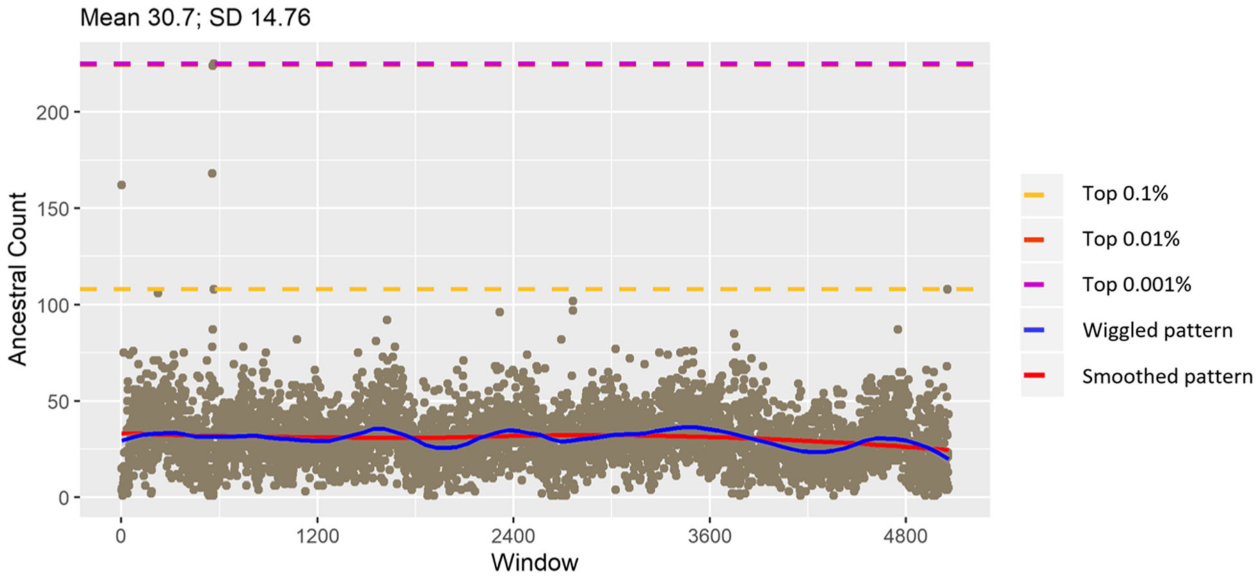

Fig. 5 Distribution of ancestral count in zebu chr 29

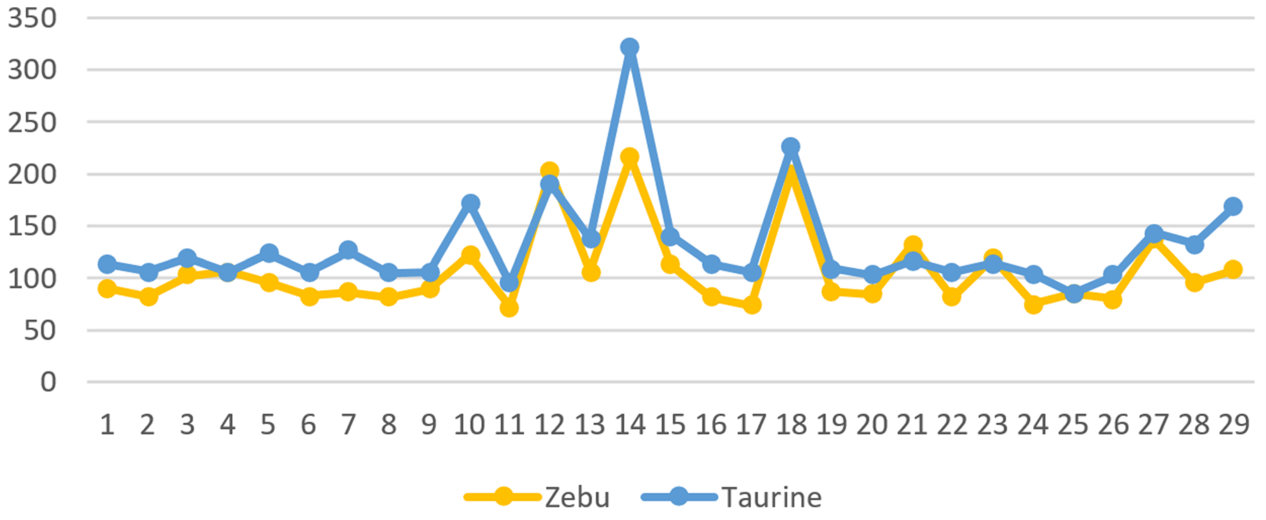

Fig. 6 Threshold of 0.1% top ancestral count

and negative regulation of endopeptidase activity (GO:0010951).

In taurine cattle, apart from terms related to immune system process and cellular function, there are GO terms exclusive to taurine cattle that are related to production traits. For example, GO:0008654, GO:0043410, GO:0045725, GO:0060048, GO:0008016, are related to metabolic process of phospholipid, protein, glycogen, and regulation of muscle and heart contraction. GO:0007613 and GO:0035176 are related to mental information processing systems and is part of learning or memory abilities which can affect cognition and behavior as indicated by CRTC1, TH, ITPR3, DBH, SORCS3 genes. ITPR3 is known as well for process of sensory perception of taste. CRTC1 gene in human has highest transcript expression in brain compared to other tissues and is known for affecting eating behavior [25].

GO:0009611, GO:0071364, GO:0071560 and GO:0008286 are related to response of stimulus such as stress from wounding and transforming growth factor.

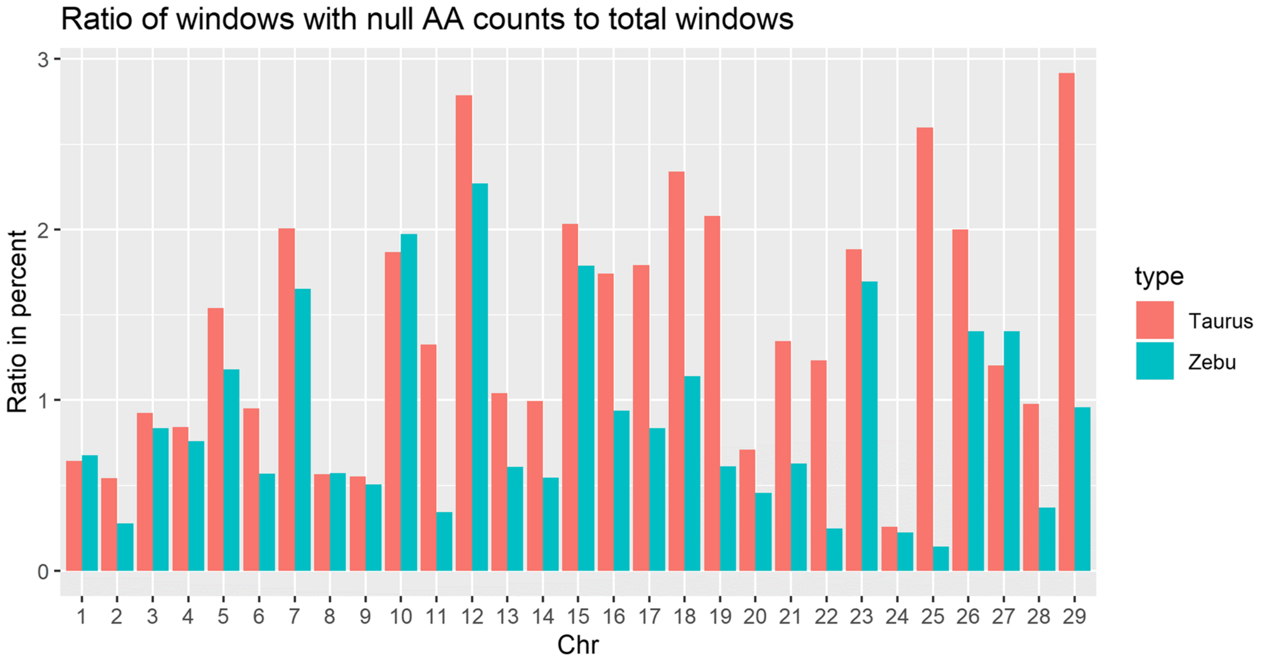

Fig. 7 Ratio of windows with null AA counts to total windows

Table 1 GO terms of genes indicated by high count ancestral alleles

GOTerm	Function	Count	PValue	Genes	Fold Enrichment	Bonferroni
Taurine						
GO:0050482	Arachidonic acid secretion	3	5.0E-04	LOC100125947, PLA2G2A	84.12	0.02
GO:0006644	Phospholipid metabolic process	3	7.7E-04	LOC100125947, PLA2G2A	67.76	0.03
GO:0016042	Lipid catabolic process	3	2.9E-03	LOC100125947, PLA2G2A	34.85	0.10
GO:0042742	Defense response to bacterium	2	9.0E-02	DEFB7, DEFB3	20.08	0.97
Zebu						
GO:0050482	Arachidonic acid secretion	3	8.7E-04	LOC100125947, PLA2G2A	65.00	0.06
GO:0006644	Phospholipid metabolic process	3	1.3E-03	LOC100125947, PLA2G2A	52.36	0.10
GO:0016042	Lipid catabolic process	3	5.0E-03	LOC100125947, PLA2G2A	26.93	0.32
GO:0070509	Calcium ion import	2	2.4E-02	SLC8A1, CACNA1D	78.55	0.85

GO:0048469, GO:0010976, GO:0060425, GO:0002062, are terms related to development of cell, neuron, lung morphogenesis and chondrocyte differentiation in cartilage outgrowth as part of skeletal system and animal organ development as pointed by PTH1R, COL2A1, COL11A2, WNT7A, RUNX3, SOX10, GATA2, PTH1R, and SOX18 genes.

Regions without AA in zebu were mainly related to 5 GO terms in domain of immune response and one term related to cellular process of transmembrane transport. Figure 8 represented distribution of terms found in regions without AA. It is dominated by metabolism terms in taurine and immune response in zebu.

Discussion

We forced mapping short read sequences of different species within Bovinae subfamily into the latest cattle RefSeq ARS-UCD1.2 irrespective of their actual genome structure. Phylogenetic trees were built based on the SNP variants in autosomes. We used subsets of all variants per chromosome to comply with maximum 50,000 markers/sequences per output of the analysis as directed by the software [26]. Despite an unequal number of individuals representing each group, we could infer relationships based on variant similarity and defined four lineages of yak, bison, gagaba and cattle. Even though still related, none of outgroups were in ancestor-descendant relationships apparently.

Defining AA by only a single lineage was not an option since any of the current lineages could have undergone independent evolutionary events and might have diverged from the initial ancestral state. Alleles were set to be ancestral strictly if they are fixed and shared by at least two lineages of yak, bison and gagaba, complying with other similar studies [9, 15]. Using the same dataset, we infered the ancestral alleles several times resulting in the same list of alleles as we strictly considered only variants with fixed allele (100% frequency) in each species. Although, we used the best dataset available in terms size, sequence read quality, and coverage for the outgroup species, additional re-sequencing data of the outgroup species might have slightly modified the defined ancestral alleles as the frequency for those fixed alleles might be changed by new individuals. However, as

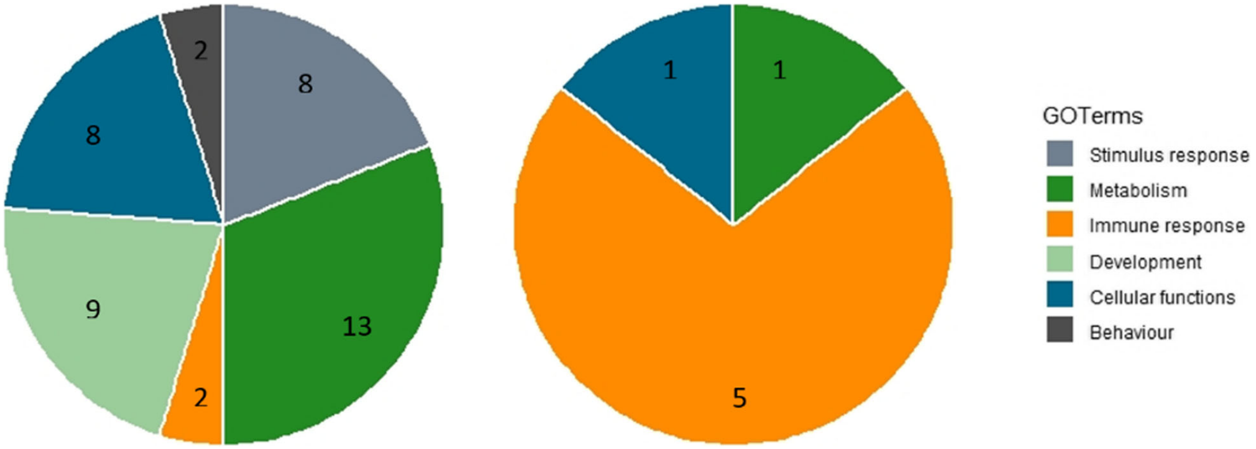

Fig. 8 GO terms for regions without ancestral alleles (taurine-left; zebu-right)

a rigid solution, we defined fixed alleles as ancestral only if they are fixed and shared by at least two different lineages.

Scanning windows of 10Kb were chosen after a preliminary comparison between 1Kb, 10Kb, and 50 Kb windows and considering the average gap between high density markers of 4Kb in identifying different types of selection in a previous study [27]. Ancestral allele counts within scanning windows in taurine and zebu cattle varied in the genome. We took two extreme ends of the occurrence distribution; one is windows with the top 0.1% highest count and second is windows without ancestral allele count. Based on the knowledge that mutation occurs across autosomes with different rates on different scales [28], we expected ancestral allele frequency to be changing as the mutations emerge. Thus, we assumed windows with highest count of AA are the conserved ones while windows without AA are the ones containing relevant mutations, considering important traits or genes that were retained along evolutionary process [7, 8].

Regions with high ancestral counts have GO terms related to primary metabolic process of lipid in both cattle. Genes within these GO terms are likely retained in ancestral states because their basic function are still beneficial. Despite different environments, both cattle need to store energy efficiently in form of lipids. Although cattle diet usually contains two to 4 % lipid, it contributes up to 50% of fat in milk and the most concentrated source of energy. In contrast to human, where liver is the primary site, fatty acid synthesis occurs at adipose tissue in ruminants [29, 30]. Adipose tissue acts as reservoir for efficient energy storage in allowing cattle and mammals in general for surviving adversities such as food shortages during severe winter for taurine or drought for zebu [31]. Defense response to bacteria (GO:0042742) was detected from regions with high ancestral counts in taurine, but found in regions without AA in zebu. In taurine high count regions, DEFB7 and DEFB3 are within this term, while regions without AA in zebu are DEFB6, LOC781146, DEFB1, DEFB3.

For regions without AA where expected mutation occurs, GO terms may have correlated and not necessarily independent from each other as pointed by its function. For grouping, we used the prevalent terms within ancestor charts in quickGO. In taurine, terms are related to behavior, cellular functions, tissue development, immune system, metabolism, and stimulus response. These are in line with suggestion from previous study for likelihood of genes function without AA and positive selection [32]. Within this scope, more GO terms found in taurine cattle compared to zebu possibly due to more intensive selection for production traits. Aiming for higher growth rate, carcass quality, feed efficiency, calving interval, milk production and body conformity has directed animals to be more efficient with higher metabolism rates [33–35]. These selection events might not only be affecting a narrow-region of genome. Instead, it altered several regions simultaneously as production traits are complex involving many QTLs or regions across chromosome with small contribution by each for the expression [36, 37].

In zebu, mutated regions were mainly linked to GO terms of immune response and little to cellular functions and metabolism. Concordance to suggested previously where zebu has been bred to adapt with more marginal production environments compared to taurine [38, 39]. Evidences showed different in relative importance on innate and adaptive immune response towards cattle tick *Rhipicephalus microplus* infestation between zebu and taurine. Skin inflammatory response by high secretion of granulocytes and T-lymphocytes in taurine is not necessary could cease tick invasion. But, an earlier inflammatory response and secretion of an alternate non-volatile T-cell in zebu were more efficient in repel this tick invasion [40, 41].

Nevertheless, not all genes within previously mentioned GO terms can be linked directly to positive selection. As mentioned in previous study, BOLA gene families, which we found also in regions without AA, are a result of balancing selection aiming for preserving genetic diversity as heterozygous animals have more advantage than the homozygous ones [27]. Similarly, we cannot confirm whether genes here are main targets of selection or as hitchhiking effect from genes of interests. For example, genes within GO:0007613, related to behavior memory and taste preferences, could be intended for selection because breeder preferences of tame, good mothering ability and non-picky animals in terms of feed and housing. Alternatively, it could be indirectly selected because animals have to cope with commercial environment as suggested that behavioral patterns were altered for animals in pasture and confinement cases [42, 43].

Our findings suggest that retaining and losing AA in some genes or regions are varied and made it species-specific with possibility of overlapping as it depends on the selective pressure they had to experience. Future work in finding overlapped domains detected by different tools for selection signatures would confirm specific regions/functions peculiar for each various cattle breeds.

Conclusions

We inferred ancestral alleles by combining fixed alleles in three lineages of cattle outgroups. Regions conserving more primitive functions indicated by high count ancestral alleles were linked to lipid metabolism in taurine and zebu. Meanwhile, regions undergone mutation indicated by no preserved ancestral alleles were found more on taurine than zebu. These regions were linked to production traits in taurine and robustness traits in zebu.

Methods

Dataset

WGS of different (sub)species were obtained from NCBI BioProject in fastq format as listed in Table 2, please refer to ‘Availability of Data and Materials’ section for the accession numbers. Taurine cattle group was represented by several commercial breeds, i.e. Holstein, Angus, Jersey, and Simmental. Workflow of the ancestral analysis pipeline is shown in Fig. 9.

Alignment and variant calling

Following Best Practice procedure by Genome Analysis Tool Kit [49–51], single interleaved data sets of FASTQ from each individual were not trimmed based on phred score, because GATK tool takes care of these low quality reads on later step during recalibration process. Datasets were mapped against the cattle reference sequence ARS.UCD-1.2 [16] using BWA-MEM [52] with default parameters. The raw mapped reads were sorted by chromosome position using SortSAM function. Sorted BAM files then underwent duplicates marking using Picard MarkDuplicates. Base Quality Score Recalibration (BQSR) was carried out to adjust the base scores towards various possibly systematic errors. BQSR required supporting files, such as known variant sites in vcf format [44], index and dict files of reference sequence created by using Samtools [53]. Report file in table form was needed for the next step of ApplyBQSR with an output of analysis ready BAM files. Analysis ready BAM files were individually called for variants using HaplotypeCaller with GVCF mode for preparation in cohort analysis workflow. Individual VCFs then combined using CombineGVCFs and went through joint-call cohort for GenotypeGVCFs. SplitVCFs tool was used to split SNPs and Indel variants from cohort VCF file. SNP variants were filtered out for parameter of mapping quality less than 40, QUAL less than 30 and quality by depth less than 30. Header editing of vcf files and splitting by each chromosome were done using bcftools and vcftools.

Principal component analysis

Multisample VCF file was converted to binary plink format using VCFtools. The indep algorithm in PLINK [54] was used with default parameters of 50 variants window size units shifting for every 5 variants with pairwise r^2 threshold of 0.7. This step selected a set of independent variants for reducing redundancy. Then, we set four components to reduce dimension of the whole independent variants and plotted the species based on the first two components.

Phylogenetic trees

We constructed phylogenetic trees from autosomes of our species similar to other studies, so called phylogenomes [55, 56]. SNPhylo [26] processed original multisample VCF files of chromosome 1 to 29 separately to reduce redundant variants based on LD. Parameters were set to 0.1 Low Coverage Samples (PCLS), depth coverage of two, 0.9 LD threshold, 0.1 minor allele frequency and 0.1 missing rate. These parameters were set to meet the maximum variants output by the program and roughly reduce the variants to 10% in output fasta. MEGA X built initial tree using Maximum Parsimony method and inferred final phylogenetic trees for each chromosome by using Maximum Likelihood method and Jukes-Cantor model with 200 bootstraps [57, 58].

Table 2 List of whole genome sequences data

Name	Species	N	Avg. Mbases	Avg. read length	Mapped reads (%)	Clean reads[a] (%)	Coverage	BioProject	Reference
Taurine cattle	*Bos taurus*	23	18,913	248	98.37	78.61	5x	PRJNA238491, PRJNA277147	[44, 45]
Banteng	*Bos javanicus*	5	16,596	250	98.36	84.30	5x	PRJNA427536	[46]
Gaur	*Bos gaurus*	4	18,428	300	98.50	71.61	5x	PRJNA427536	[46]
Yak	*Bos grunniens*	13	22,177	201	98.51	79.47	4x	PRJNA285834	[47]
American Bison	*Bison bison*	4	13,364	200	98.51	82.81	4x	PRJNA427536	[46]
European Bison	*Bison bonasus*	5	18,113	250	98.51	89.67	5x	PRJNA427536	[46]
Gayal	*Bos frontalis*	4	18,610	250	98.49	86.71	5x	PRJNA427536	[46]
Aurochs	*Bos taurus primigenius*	2	17,105	62	98.51	45.17	3x	PRJNA294709	[32]
Zebu cattle	*Bos taurus indicus*	8	10,863	428	98.60	76.98	4x	PRJNA507259, PRJNA427256	[48]

[a] Reads remaining after base quality score recalibration process and used for calling variants

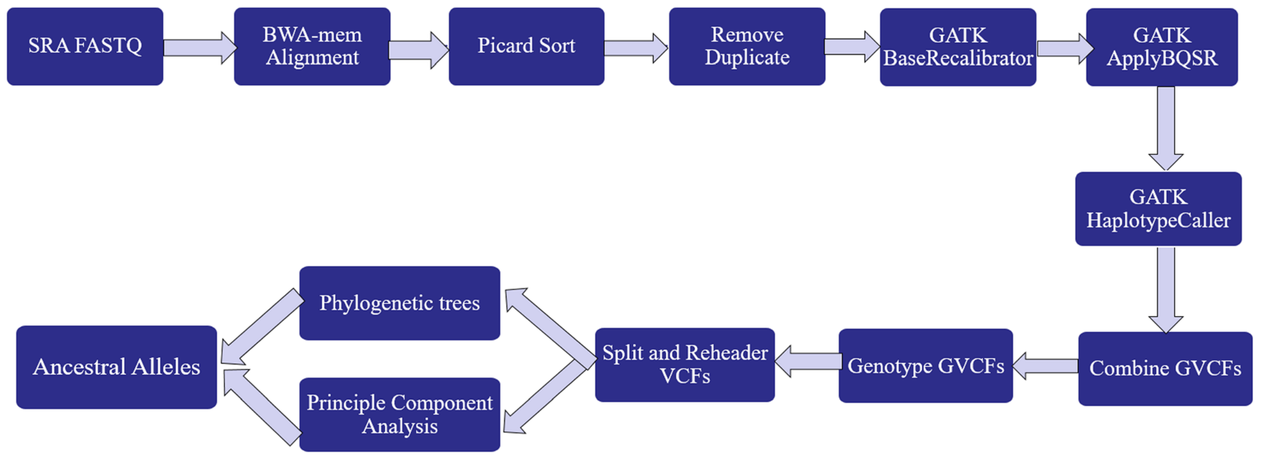

Fig. 9 Workflow of the ancestral allele analysis

Inferring ancestral allelic states

VCFtool was used to call allele frequency spectrum from un-pruned VCF files. Considering branches in phylogenetic trees and clusters of PCA, we defined three lineages of cattle outgroup, i.e. Yak, Bison (American bison and European bison), and Gagaba (Gayal-Gaur-Banteng). For each site, frequency of two alleles of A and a represented by p(A) and q(a) frequency. If p(A) frequency of 1 and found in at least two lineages, we defined "A" allele as ancestral for that site.

We used R [59] to create list of these defined AA for all autosomes. Following packages in R were used to support data analysis and visualization: dplyr [60], ggplot2 [61], and stringr [62]. R functions for calling the ancestral allele in this study are provided in https://github.com/mas-agis/ances-al with an example run for all the scripts provided in [63].

Comparison to cattle groups

A custom script was used to compute summary statistics of allele frequencies and to compare which AA are still intact in zebu and taurine cattle. Notation 1 below, defining how we calculated ϑ, the changing frequency of ancestral allele compared to cattle group:

$$(\text{Notation } 1) : \vartheta = x - p(A_{AA}),$$

where x is the frequency of same allele A in cattle as the ancestral $p(A_{AA})$.

Given ancestral allele denotes as $p(A_{AA})$ with frequency of 1 for A allele, ϑ is calculated by subtract $p(A_{AA})$ from x. Where x can be both major $p(A_{cattle})$ or minor $q(A_{cattle})$ allele in cattle with condition that x must represent the same allele A as the ancestral one. We assigned ϑ for each site of SNP data across the autosome. For example, if major allele in cattle is A matching to $p(A_{AA})$, thus

$$\vartheta = p_{cattle} - p(A_{AA}) = 100\% - 100\% = 0$$

while if minor allele A in cattle matching $p(A_{AA})$, then

$$\vartheta = q_{cattle} - p(A_{AA}) = 30\% - 100\% = -0.7$$

We filtered ϑ with value of 0 meaning ancestral allele persist in cattle groups. To count how many sites persisting with AA, we assigned f(ϑ) score is 1 for every ϑ equal to zero, otherwise we assigned zero to the f(ϑ) as notation 2 below. We used non-overlapping windows of 10 Kb to sum up sites that have value of 1. By this scanning windows, autosomes were divided into regions and total counts were reported. We selected two extreme conditions of windows with highest count and null count of AA. Indicated regions from both conditions were used for further analysis.

(Notation 2):

$$\boldsymbol{T}(\boldsymbol{\vartheta}) = \sum_{t=1}^{n} \sum_{i=10000(t-1)}^{10000t} \boldsymbol{f}(\boldsymbol{\vartheta}_i), \boldsymbol{where}\ \boldsymbol{f}(\boldsymbol{\vartheta}) = \begin{cases} \mathbf{1}, \boldsymbol{\vartheta} = \mathbf{0} \\ \mathbf{0}, \boldsymbol{\vartheta} \neq \mathbf{0} \end{cases}$$

Annotation region of interest

Physical regions indicated by previous step were taken as input for ANNOVAR [64]. We then excluded regions that are fall in the intergenic, downstream and upstream of known genes, leaving only regions that overlapping with functional genes. We filtered out genes defined by highest count regions if were found also in regions without ancestral counts. We used this list of genes for GO analysis using DAVID 6.8 [65, 66]. We report GO of biological process with the Bonferroni corrected *P*-values. Definition and supporting information related to GO were retrieved from database of European Bioinformatics Institute [67].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07412-9.

Additional file 1. Phylogenetic trees from each chromosome

Additional file 2. Distribution of ancestral allele in all chromosomes of taurine and zebu

Additional file 3. Annotation of regions without ancestral alleles

Additional file 4. Accession numbers of individual sequencing reads

Abbreviations

AA: Ancestral allele; DA: Derived allele; BWA: Burrows-Wheeler Aligner; GO: Gene Ontology; NCBI: National Center for Biotechnology; SNP: Single-nucleotide polymorphism

Acknowledgments

Computation was carried out partly in usegalaxy.org platform and Vienna Scientific Cluster (VSC3).

Authors' contributions

GM conceived and designed the study. BDR and MMN coordinated the input dataset. MMN run the analysis and drafted the manuscript. YTU, BDR, JS, and GM interpreted the analysis results and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Funding

MMN is currently receiving the Ernst Mach Grant – ASEA UNINET (Austria).

Availability of data and materials

Accession numbers for each individual used in this study are provided in 'Additional file 4' and can be retrieved from NCBI repository https://www.ncbi.nlm.nih.gov/sra/.
Cattle reference sequences ARS_UCD1.2 is available at NCBI repository https://www.ncbi.nlm.nih.gov/genome/?term=cattle.
Defined cattle ancestral alleles is available at https://tinyurl.com/cattle-aa.
Custom R functions are available https://github.com/mas-agis/ances-al.
Example of scripts for this paper including application of custom R functions is provided in [63].

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interest.

Author details

[1]University of Natural Resources and Life Sciences (BOKU), Vienna, Austria. [2]São Paulo State University (Unesp), School of Veterinary Medicine, Department of Production and Animal Health, Araçatuba, São Paulo, Brazil. [3]International Atomic Energy Agency (IAEA) Collaborating Centre on Animal Genomics and Bioinformatics, Araçatuba, São Paulo, Brazil. [4]AgroPartners Consulting. R. Floriano Peixoto, 120-Sala 43A-Centro, Araçatuba, SP 16010-220, Brazil. [5]Personal-PEC. R. Sebastiao Lima, 1336-Centro, Campo Grande, MS 79004-600, Brazil. [6]Agricultural Research Service USDA, Beltsville, MD, USA.

Received: 25 September 2020 Accepted: 27 January 2021
Published online: 08 February 2021

References

1. Fitzpatrick BM, Turelli M. The geography of mammalian speciation: mixed signals from phylogenies and range maps. Evolution. 2007;60:601–15.
2. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B. A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet. 2012;131:1541–54.
3. Daetwyler HD, Capitan A, Pausch H, Stothard P, Binsbergen R, Brøndum R, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46:858–65.
4. Weldenegodguad M, Popov R, Pokharel K, Ammosov I, Ming Y, Ivanova Z, et al. Whole-genome sequencing of three native cattle breeds originating from the northernmost cattle farming regions. Front Genet. 2019;9:728.
5. Tijjani A, Utsunomiya YT, Ezekwe AG, Nashiru O, Hanotte O. Genome sequence analysis reveals selection signatures in endangered trypanotolerant West African Muturu Cattle. Front Genet. 2019;10:442.
6. Zhao T, Schranz ME. Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. Proc Natl Acad Sci U S A. 2019;116:2165–74.
7. Keightley PD, Campos JL, Booker TR, Charlesworth B. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of <em>Drosophila melanogaster</em>. Genetics. 2016;203:975.
8. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. 2000;155:1405–13.
9. Rocha D, Billerey C, Samson F, Boichard D, Boussaha M. Identification of the putative ancestral allele of bovine single-nucleotide polymorphisms. J Anim Breed Genet. 2014;131:483–6.
10. Rogers AR, Wooding S, Huff CD, Batzer MA, Jorde LB. Ancestral alleles and population origins: inferences depend on mutation rate. Mol Biol Evol. 2007;24:990–7.
11. Bianco E, Nevado B, Ramos-Onsins SE, Pérez-Enciso M. A deep catalog of autosomal single nucleotide variation in the pig. PLoS One. 2015;10:e0118867.
12. Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. eLife. 2017;6:e24284.
13. Matsumoto T, Akashi H, Yang Z. Evaluation of ancestral sequence reconstruction methods to infer nonstationary patterns of nucleotide substitution. Genetics. 2015;200:873.
14. Park L. Ancestral alleles in the human genome based on population sequencing data. PLoS One. 2015;10:e0128186.
15. Utsunomiya YT, Pérez O'Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, Mészáros G, et al. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PLoS ONE. 2013;8:e64280.
16. 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:giaa021. https://doi.org/10.1093/gigascience/giaa021.
17. Wang K, Wang L, Lenstra JA, Jian J, Yang Y, Hu Q, et al. The genome sequence of the wisent (Bison bonasus). Gigascience. 2017;6:1–5.
18. Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, et al. Identifying the favored mutation in a positive selective sweep. Nat Methods. 2018;15:279–82.
19. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72. https://doi.org/10.1371/journal.pbio.0040072.
20. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009;19:826–37.
21. Zhong M, Zhang Y, Lange K, Fan R. A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. Stat Interf. 2011;4:51–63.
22. Vatsiou AI, Bazin E, Gaggiotti OE. Detection of selective sweeps in structured populations: a comparison of recent methods. Mol Ecol. 2016;25:89–103.
23. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193:929–41.
24. Lee C. Chapter 10 - structural genomic variation in the human genome. In: Ginsburg GS, Willard HF, editors. Genomic and personalized medicine (Second Edition). Cambridge: Academic Press; 2013. p. 123–32. https://doi.org/10.1016/B978-0-12-382227-7.00010-0.
25. Rohde K, Keller M, la Cour Poulsen L, Ronningen T, Stumvoll M, Tonjes A, et al. (Epi) genetic regulation of CRTC1 in human eating behaviour and fat distribution. EBioMedicine. 2019;44:476–88.
26. Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 2014;15:162.

27. Porto-Neto LR, Sonstegard TS, Liu GE, Bickhart DM, Da Silva MV, Machado MA, et al. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. BMC Genomics. 2013;14:876.
28. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011;12:756–66.
29. Nafikov RA, Beitz DC. Carbohydrate and lipid metabolism in farm animals. J Nutr. 2007;137:702–5.
30. Wattiaux MA, Grummer RR. Lipid metabolism in dairy cows. In: Lipid metabolism in dairy cows. Madison: University of Wisconsin; 2000. https://federated.kb.wisc.edu/images/group226/52745/4.LipidMetabolisminDairycows.pdf.
31. Young RA. Fat, energy and mammalian survival. Am Zool. 2015;16:699–710.
32. Park SDE, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, et al. Genome sequencing of the extinct Eurasian wild aurochs, Bos primigenius, illuminates the phylogeography and evolution of cattle. Genome Biol. 2015;16:234.
33. Neeteson-van Nieuwenhoven A-M, Knap P, Avendaño S. The role of sustainable commercial pig and poultry breeding for food security. Anim Front. 2013;3:52–7.
34. Hietala P, Juga J. Impact of including growth, carcass and feed efficiency traits in the breeding goal for combined milk and beef production systems. Animal. 2017;11:564–73.
35. Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. A 100-year review: identification and genetic selection of economically important traits in dairy cattle. J Dairy Sci. 2017;100:10251–71.
36. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, et al. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics. 2011;12:408.
37. van den Berg I, Hayes BJ, Chamberlain AJ, Goddard ME. Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. BMC Genomics. 2019;20:291.
38. Chan EKF, Nagaraj SH, Reverter A. The evolution of tropical adaptation: comparing taurine and zebu cattle. Anim Genet. 2010;41:467–77.
39. Pérez O'Brien AM, Utsunomiya YT, Mészáros G, Bickhart DM, Liu GE, Van Tassell CP, et al. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. Genet Sel Evol. 2014;46:19.
40. Franzin AM, Maruyama SR, Garcia GR, Oliveira RP, Ribeiro JMC, Bishop R, et al. Immune and biochemical responses in skin differ between bovine hosts genetically susceptible and resistant to the cattle tick Rhipicephalus microplus. Parasit Vectors. 2017;10:51.
41. Jonsson NN, Piper EK, Constantinoiu CC. Host resistance in cattle to infestation with the cattle tick Rhipicephalus microplus. Parasite Immunol. 2014;36:553–9.
42. Charlton G, Rutter S. The behaviour of housed dairy cattle with and without pasture access: a review. Appl Anim Behav Sci. 2017;192:2–9.
43. O'Connel J, Giller PS, Meaney W. A comparison of dairy cattle behavioural patterns at pasture and during confinement. Ir J Agric Res. 1989;28:65–72.
44. 1000 Bull Genomes project. 2018. http://www.1000bullgenomes.com/.
45. Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, et al. Analysis of copy number variations among diverse cattle breeds. Genome Res. 2010;20:693–703.
46. 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.
47. Qiu Q, Wang L, Wang K, Yang Y, Ma T, Wang Z, et al. Yak whole-genome resequencing reveals domestication signatures and prehistoric population expansions. Nat Commun. 2015;6:10283.
48. Stafuzza NB, de Oliveira Silva RM, Peripolli E, Bezerra LAF, Lôbo RB, de Ulhoa Magnabosco C, et al. Genome-wide association study provides insights into genes related with horn development in Nelore beef cattle. PLoS ONE. 2018;13:e0202978.
49. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
50. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
51. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1–33.
52. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26:589–95.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
54. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
55. Jarvis E, Mirarab S, Aberer A, Li B, Houde P, Li C, et al. Whole-genome analyses resove early branches in the tree of life of modern birds. Science. 2014;346:1320–31.
56. Tsuda K, Kawahara-Miki R, Sano S, Imai M, Noguchi T, Inayoshi Y, et al. Abundant sequence divergence in the native Japanese cattle Mishima-Ushi (Bos taurus) detected using whole-genome sequencing. Genomics. 2013;102:372–8.
57. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.
58. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969. p. 21–132.
59. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2019. https://www.R-project.org/
60. Wickham H, Francois R, Henry L, Müller K. dplyr: a grammar of data manipulation. R package version 0.8.3. 2019. https://CRAN.R-project.org/package=dplyr.
61. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
62. Wickham H. stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. 2019. https://CRAN.R-project.org/package=stringr.
63. Naji M. Protocol - investigation of ancestral alleles in the Bovinae subfamily; 2020. https://doi.org/10.17504/protocols.io.bh99j996.
64. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
65. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
66. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
67. QuickGO. EMBL-EBI; 2020. https://www.ebi.ac.uk/QuickGO/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC