Research article

Long-range linkage disequilibrium events on the genome of dromedary camels as a signal of epistatic and directional positive selection

Hussain Bahbahani

Department of Biological Sciences, Faculty of Science, Kuwait University, Sh. Sabah Al-Salem Campus, Kuwait

A B S T R A C T

The genome of dromedary camels has been subjected to various evolutionary forces, such as genetic admixture, natural positive selection, and epistatic selection. These forces are considered as main factors associated with the formation of long-range linkage disequilibrium (LRLD) events. We have analyzed whole-genome data of 56 dromedary camel samples from different geographical regions across the Arabian Peninsula for two main purposes: first, to assess the level of linkage disequilibrium, and second, to identify autosomal LRLD events. The analysis revealed a mean r^2 value of 0.25 ($\pm$0.028) over the dromedary autosomes, with a continuous decay until reaching a plateau at inter-variant distances >400 kb. A total of 1847 LRLD events were identified within the dromedary autosomes, which harbor 36 prevalent haplotypes. A level of genetic admixture was observed among the dromedary populations analyzed, which might be a source for the observed LRLD events. Four functional interactions were revealed among the genes found within the LRLD events, with some genes overlapping with prevalent haplotypes, indicative of potential epistatic selection. Genes related to renal function, fertility, thermal regulation, bone structure, and insulin regulation were found among the LRLD genes. These genes, along with the defined prevalent haplotypes, can be considered as hotspots for natural positive selection associated with the LRLD distribution on dromedary genomes. In this study, we have for the first time analyzed the genome of dromedary camels for LRLD events possibly influenced by forces including genetic admixture, epistatic and positive selection. The revealed LRLD elements and prevalent haplotypes should be accounted for when designing breeding programmes to conserve the genetic stock of this well-adapted domestic species.

1. Introduction

Linkage disequilibrium (LD) is defined as the non-random assortment of alleles from different loci, which might be associated with different evolutionary forces, such as population admixture, positive selection, and genetic drift [1]. This genetic parameter has been estimated using different statistics, e.g., D, D', λ, and r^2, depending on the study objectives [1]. The square correlation coefficient (r^2), which is extensively used in association analyses, estimates the correlation between markers and genes of interest. This statistic, which ranges between 0 (no-disequilibrium) to 1 (complete disequilibrium) is preferred in animal populations LD analyses due to its low

sensitivity to small sample sizes and low allele frequency markers [2]. Using these different estimators, the LD level and extent of LD decay have been widely investigated in different animal species, e.g., sheep [3,4], horse [5,6], cattle [7,8], and in Bactrian and dromedary camels [9,10]. The estimation of such LD pattern can be used to infer different evolutionary events, such as effective population size, level of genetic drift, and footprints of selection [1].

The strong LD between different markers in a chromosome would result in the construction of haplotype blocks, which are defined as regions with minimal recombination and few observable haplotypes [11]. Haplotype blocks may lead to the formation of long-range linkage disequilibrium (LRLD) events if high LD occurs between pairs of sites within two blocks separated by a large genetic distance, e. g., more than 1 Mb [12]. This genetic phenomenon has been investigated in the genome of Human populations [12,13] and non-human populations [14,15]. In 2013, Koch, Ristroph [12] have analyzed the HapMap 2 data and defined LRLD patches on the 22 autosomes of Yoruba populations from Nigeria. These LRLD events are of great interest as they might result from different factors, such as population admixture, genetic drift, epistatic selection, and positive selection on desired variants. Wilson and Goldstein [13] have investigated the X chromosome of Bantu-Semantic (Lemba) hybrid population for LRLD. The admixed nature of their genome was considered as the main force behind these regions. Epistatic selection is a type of evolutionary force that influences combined allele sets from different loci resulting in an increase in their frequencies and stable LD between them [16]. El Hou, Rocha [15] have investigated the genome of three French beef cattle breeds for LRLD events. Functional gene interactions were observed in some of the identified LRLD events suggesting the possible role of epistatic selection in shaping their LRLD pattern.

Dromedary camels belong to the Camelini tribe within the Camelidae family populating north Africa, the Arabian Peninsula and southeast Asia [17]. The continuous efforts in resequencing the whole genome of different dromedary populations, and the availability of a functionally-annotated chromosome-level dromedary reference genome assembly (CamDro3) [18] helped in deeply analyzing the genome of this species. This is due to the substantial coverage of the genome provided by whole genome sequence data, and the lack of ascertainment bias that usually associated with the commercially available genotyping SNP arrays. Analyzing the genome of dromedary camels has primarily focused on assessing their genetic diversity [19–21], searching for signatures of selection on an autosomal [22–24] and mitochondrial DNA levels [25,26], investigating their demographic history [9,27], and even efforts towards candidate gene analyses [28,29]. Investigating the genetic diversity of dromedary camels from the Arabian Peninsula classified them into three main groups corresponding to their geographical distribution in the Arabian Peninsula. This genetic distinction was associated with a substantial level of genetic admixture among the dromedary camel populations analyzed [19,21]. Such genetic admixture might be related to the historical use of dromedaries in transporting people and trading over long distance journeys [17].

Dromedary camels exhibit superior adaptability to various environmental conditions they populate, which are mainly characterized as deserts and semi-deserts [30]. They employ a heterothermic thermoregulation mechanism to regulate body temperature, especially under water deprivation conditions, to avoid water loss through extensive sweating [31]. They can also tolerate water loss equivalent to ~25 % of their body weight in addition to enduring water deprivation for 20–35 days [32]. The hump of dromedaries is a fat reservoir used as an energy source to allow them to endure longer periods of hunger [30]. These traits highlight the dromedary genome as a focal point for natural selection, driving such adaptabilities. Genes related to fat and energy metabolism, salt tolerance and stress response were found to be under adaptive evolution upon sequencing the whole genome of Bactrian and dromedary camels [33]. Analyzing the whole genome of dromedary camels from the Arabian Peninsula [23,34] and Iran [24] also revealed various genes with biological implications associated with their adaptive physiologies, such as lipid metabolism, long-term memory, kidney physiology, and fertility, showing signs of positive selection. At the mitochondrial DNA level, signals of site-specific positive selection were defined on codons within the mitochondrial protein-coding genes of dromedary camels, potentially contributing to their environmental adaptability [25,26].

Such evolutionary forces shaping the genome of dromedary camels are considered as main factors associated with the formation of LRLD events. In this study, we investigated the genome of dromedary camels from the Arabian Peninsula for LRLD events using whole genome sequence data. The identified LRLD events were functionally characterized and possible functional interactions between genes were assessed to explore the hypothesis that natural positive and epistatic selections were playing a role in shaping the LRLD profile of the dromedary genome.

2. Materials and methods

2.1. Whole genome sequence data

Whole genome sequence data from 56 dromedary camels across various regions of the Arabian Peninsula: the north, center, west, and southwest (Saudi Arabia); east (Kuwait); and southeast (Oman), were included in the study. These data were publicly available as follows: 27 samples from Al Abri, Alfoudari [23], 13 samples from Bahbahani and Almathen [21], and 16 samples from Bahbahani, Alfoudari [34] (Supplementary Table S1). Raw sequence reads were filtered out by Fastp version 0.22.0 [35] if: (1) 10 % or more of the read bases were uncertain; (2) 40 % or more of the read bases were of phred-scaled quality score ≤20; (3) read lengths were shorter than 15 bases; or (4) read complexity was less than 30 %. Bases with a phred-scaled quality score <20 were also filtered out from the reads. The remaining high-quality sequence reads were mapped against the dromedary reference genome assembly (CamDro3) using the *bwa-mem* algorithm in Burrow-Wheeler Aligner (BWA) version 0.7.17 [36]. Subsequently, the reads were sorted based on co-ordinates, and PCR-duplicates were marked and excluded using the *SortSam* and *MarkDuplicates* options in Picard tool version 1.119 (https://broadinstitute.github.io/picard/). Summary statistics calculated for mapped reads included: percentage of reference genome coverage, mean depth of coverage, percentage of mapped reads, and percentage of paired mapped reads via the *coverage* and *flagstat* tools in SAMTools version 1.13 [37].

2.2. Single nucleotide polymorphisms (SNPs) calling and filtering

Autosomal SNPs on the dromedary whole genome sequence were called via the *HaplotypeCaller* algorithm using the GVCF mode implemented in GATK version 4.2.5.0 [38]. *CombineGVCFs* and *GenotypeGVCFs* algorithms were utilized to combine and genotype the variants of the dromedary samples, respectively. This was followed by variant filtration using the *VariantFiltration* algorithm in GATK as conducted in Ref. [34]. SNPs with a depth of coverage ranging between five reads and three standard deviations from the mean depth of coverage across samples were retained. The total 6,143,889 autosomal SNPs were pruned based on various quality control criteria implemented in PLINK version 1.9 [39,40]. A total of 2,061,432 and 889,081 SNPs were excluded due to having a minor allele frequency $\leq$5 % and a genotyping call rate <100 %, respectively. Subsequently, a total of 120,165 SNPs were pruned from the dataset as they significantly departed from the Hardy-Weinberg equilibrium (*P*-value < $1X10^{-6}$), resulting in a retention of 3,073,211 SNPs for the LD analyses. Additionally, SNPs with r^2 higher than 0.5 (2,644,313 SNPs) were further excluded, leaving 428,898 SNPs for the admixture analysis. All the genotyped dromedary samples exhibited a 100 % genotyping call rate and an identity-by-state (IBS) $\leq$ 95 % and hence none were excluded from any further analyses. To avoid any biases in LD related to family structure caused by first-degree relatives, genetic relatedness among the dromedary pairs was assessed using a pairwise relatedness test implemented in the –*relatedness2* function of VCFtools version 1.13 [41]. None of the analyzed dromedary pairs showed first-degree relatedness, i.e., relatedness score (Φ) ranged between 0.177 and 0.354.

2.3. Admixture analysis

The level of genetic admixture among the dromedary camel populations was investigated using ADMIXTURE version 1.23 [42]. Cluster values (*K*) ranging from 1 to 6 were analyzed, with 200 bootstrap iterations for each analysis, to reflect the different geographical locations of the dromedary samples included. The *K* value with the lowest cross-validation error was specified as the optimal number of clusters for the dromedary samples analyzed. The output of the conducted analyses was plotted using ggplot2 [43] in R version 4.1 [44].

2.4. LD and haplotype blocks analyses

To measure LD levels, pairwise r^2 values were estimated for the different SNP pairs on the dromedary autosomes using the –r^2 command in PLINK 1.9. A threshold of 0.099 was set for pairwise r^2 values (–*ld-window-r2 0.099*). The extent of LD decay was investigated in bins spanning set inter-variant distances across the dromedary autosomes. Haplotype blocks on these autosomes were determined using the –*blocks* and –*blocks-max-kb 3000* options in Plink 1.9 [45]. Blocks with a minimum of three SNPs were retained to avoid formation of spurious blocks.

2.5. LRLD events identification

LRLD events between two haplotype blocks were defined upon fulfilling two criteria as outlined by El Hou, Rocha [15]. This included ensuring that at least two SNPs in each block were in high LD ($r^2 \geq 0.6$) with SNPs in the other block, and the genomic distance between the two haplotype blocks was $\geq$1 Mb. Spearman's test was conducted to determine any significant correlation between the number of LRLD events and the autosomal size. The frequency of the haplotypes within haplotype block pairs of the defined LRLD events were calculated using the –*hap-freq* tool in PLINK version 1.7 [39]. The heatmaps of LRLD haplotype blocks were visualized using LDBlockShow software version 1.04 [46].

2.6. Functional characterization of the LRLD events

Genes within each LRLD event's block were retrieved from the dromedary reference genome assembly (CamDro3) gene list. Functional enrichment analysis was conducted using the functional annotation tool implemented in the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics resource version 6.7 [47,48] using *Camelus dromedarius* as the background speices. Significantly enriched terms were defined if their enrichment score was $\geq$1.3. Functional interactions between the genes of each LRLD block pair were investigated using the Search Tool for the Retrieval of Interacting Genes (STRING) webserver [49].

3. Results

3.1. Summary statistics of the mapped sequence reads

On average, 99.7 % of the reads were mapped on the dromedary reference genome. Approximately, 95.8 % of these mapped were properly paired that covered 95.6 % of the reference genome. The mean depth of coverage for the mapped reads were ranged between 8X and 35X, with a mean of 20X (Supplementary Table S2).

3.2. Admixture analysis

The cluster value (*K*) equal to one showed the lowest cross validation error and hence was considered as the optimal number of

clusters describing the genetic structure of the analyzed dromedary populations (Supplementary Table S3). At $K = 2$, a genetic distinction was revealed separating dromedaries from the southeast of the Arabian Peninsula from those in other geographical regions (Fig. 1). At this K value and beyond, a degree of genetic admixture was observed among the dromedary samples from the different geographical regions of the Arabian Peninsula (Fig. 1 and Supplementary Fig. S1).

3.3. LD estimation and haplotype blocks analysis

The mean r^2 value over the dromedary autosomes was estimated to be 0.25 (±0.028). Comparable r^2 values were observed among dromedary autosomes, with chromosome 20 showing the highest r^2 value ($r^2=0.35$) followed by chromosome 16 ($r^2=0.31$), i.e., >2 standard deviations from the mean r^2. Whilst, the lowest r^2 value was found in chromosome 36 ($r^2=0.16$), i.e., <2 standard deviations from the mean r^2 (Supplementary Table S4). The maximum r^2 value was observed between SNP pairs with inter-variants distances ≤25 kb ($r^2=0.518$), while r^2 values > 0.3 were found between pairs at an inter-variants distance ≤100 kb. An inversely proportional relationship was revealed between the r^2 values and the inter-variants distance, with mean r^2 values dropping below 0.2 and reaching a plateau at inter-variants distances >400 kb (Fig. 2). After excluding haplotype blocks defined by only two SNPs, a total of 135,028 haplotype blocks were defined over the dromedary autosomes clustering 2,478,091 SNPs, which is 80.6 % of the total number of SNPs. Chromosome 1 carried the maximum number of blocks (10,494 blocks), while chromosome 34 harbored the least (2279 blocks) (Supplementary Table S5). The total size of the defined haplotype blocks was 1.35 Gb, which covers approximately 69.9 % of the dromedary autosomes. The largest haplotype blocks were found on chromosomes 20 and 14 with sizes equal to 698 kb and 654 kb, respectively (Supplementary Table S6).

3.4. LRLD events characterization

Based on the criteria followed to define LRLD, a total of 712 haplotype blocks were arranged in 1847 LRLD events across the dromedary autosomes, with an average of 61 events per autosome (Supplementary Table S7). The number of LRLD events was moderately correlated with the autosomal size (Spearman's rho = 0.52, P-value = 0.001). The maximum number of LRLD events was found in chromosome 6 (432 LRLD events), while chromosomes 4, 21, 23 and 27 carried the least (single LRLD event each). None of the LRLD events were revealed in chromosomes 18, 25, 26, 28, 31 and 33 (Fig. 3 and Supplementary Table S8). The total size of the LRLD events was 93.4 Mb, covering approximately 4.8 % of the dromedary autosomes. The largest LRLD event was found on chromosome 14 (777.6 kb in size), while the smallest event (18 kb in size) was on chromosome 6. The mean distance between the block pairs in LRLD events was 1.68 Mb (±440 Kb), with a block pair in an LRLD event on chromosome 12 showing the largest inter-block distance (3 Mb) (Supplementary Table S8). The haplotype frequencies of the LRLD block pairs ranged between 0.01 and 0.95 with a mean of 0.14 (±0.17). A total of 36 prevalent haplotypes, i.e., with frequencies ≥0.9, were distributed across 12 autosomes (Supplementary Table S9).

3.5. Functional characterization of the LRLD events

A total of 404 genes were found within the haplotype blocks of the LRLD events (Supplementary Table S10). Based on the DAVID enrichment analysis, 26 functional enrichment clusters were defined, in which two clusters were significantly enriched: chaperones and protein folding (enrichment score = 1.78), and ATP binding and ATPase activity (enrichment score = 1.3) (Supplementary Table S11). A total of four functional interactions were specified by STRING: two in chromosome 6, and an interaction in chromosome 20 and 22, separately (Table 1, Fig. 4 and Supplementary Fig. S2). Following comprehensive literature review, biological pathways

Fig. 1. Admixture analysis plot for cluster value (K) equal to two on dromedary camel. Dromedary samples are from the north (N), center (C), southwest (SW), west (W), east (E) and southeast (SE) of the Arabian Peninsula.
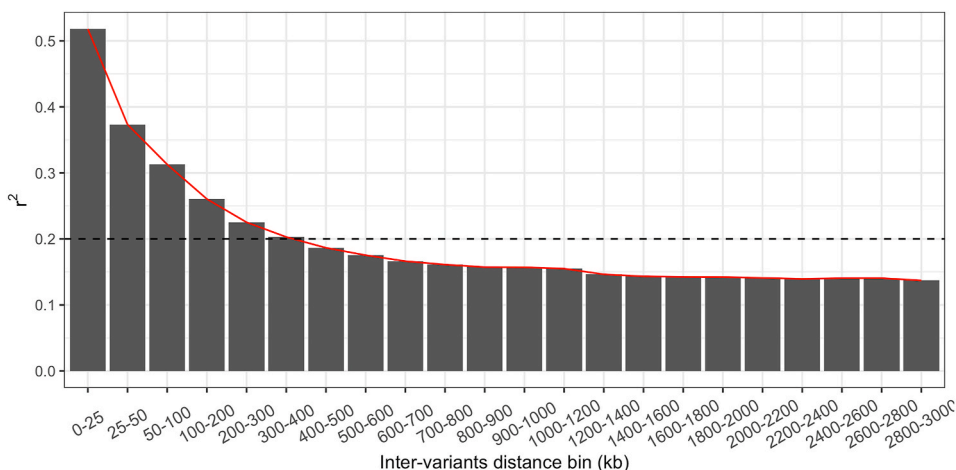
Fig. 2. The extent of LD decay measured by the square correlation coefficient (r2) across the autosomes of dromedary camels. The dashed line threshold is at $r^2 = 0.2$.
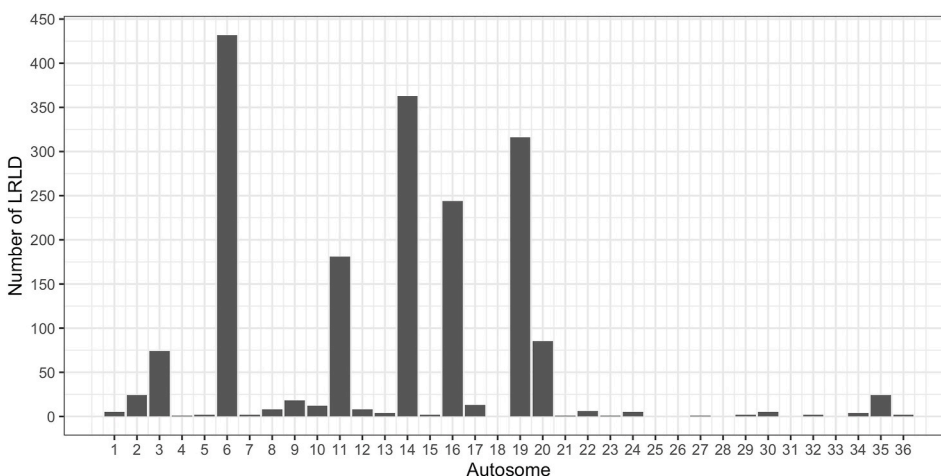

Fig. 3. Number of long-range linkage disequilibrium (LRLD) events on the autosomes of dromedary camels.

related to renal function, fertility, thermal regulation, insulin regulation, and bone structure were defined for the retrieved genes (Table 2).

4. Discussion

The genome of dromedary camels has been investigated here to assess the LD levels and define LRLD events on their autosomes. The mean r^2 value between the autosomal variants was estimated to be 0.25, which continued to decay until it reached a plateau beyond an inter-variant distance of 400 kb. The mean r^2 value estimated in our dromedary samples is contrasted to values obtained in other animal species, e.g., Australian sheep breeds (0.13–0.2) [3], Italian cattle breeds (0.07–0.17) [8], and Polish horse breeds (0.25–0.34)

Table 1
Genes within the haplotype blocks of the LRLD events showing possible functional interactions based on STRING.

Chr	LRDL events (Block_A with Block_B)	Gene A	Block A	Gene B	Block B
6	Block_230 with Block_453	*DENND4A*	Block_230	*RAB8B*	Block_453
	Block_233 with Block_453		Block_233		
6	Block_238 with Block_453	*RAB11A*	Block_238	*RAB8B*	Block_453
	Block_239 with Block_453		Block_239		
20	Block_1520 with Block_1625	*HLA-DRA*	Block_1520	*TNXB*	Block_1625
22	Block_51 with Block_121	*EEF2*	Block_51	*RPS15*	Block_121

Table 2

Functional categories of candidate genes and their respective LRLD haplotype block.

| Functional category	Gene ID	Gene name	Haplotype Block
Thermal regulation	*HSP90AA1*	Heat shock protein 90 alpha family class A	Chr 6: Block_5792
	CKB	Creatine kinase B	Chr 6: Block_5851
Renal function	*KCNK5*	Potassium Two Pore Domain Channel Subfamily K Member	Chr 20: Block_1250
Fertility	*CATSPER2*	Cation Channel Sperm Associated 2	Chr 6: Block_1565
	ARMC2	Armadillo Repeat Containing 2	Chr 8: Block_2185
Insulin regulation	*GLP-1R*	Glucagon-like peptide 1 receptor	Chr 20: Block_1250
Bone structure	*ZFAND3*	Zinc Finger AN1-Type Containing 3	Chr 20: Block_1304
	COL5A2	Collagen Type V Alpha 2 Chain	Chr 5: Block_3367
	MAPK13	Mitogen-Activated Protein Kinase 13	Chr 20: Block_1395
	B4GALNT3	Beta-1,4-N-Acetyl-Galactosaminyltransferase 3	Chr 34: Block_1573
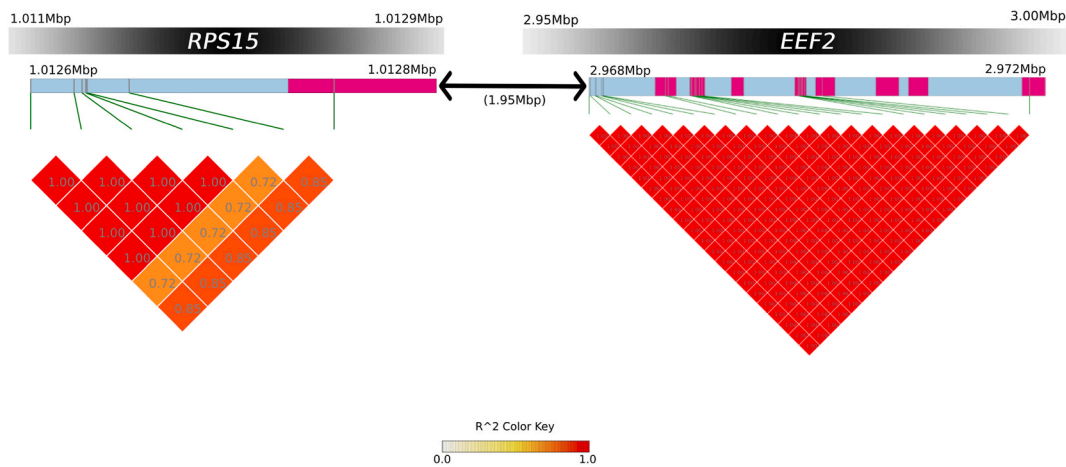
Fig. 4. Heatmaps of LRLD haplotype blocks in chromosome 22 containing *RPS15* and *EEF2* genes that exhibited potential functional interaction based on STRING database. Pink blocks represent coding regions. Double headed arrow shows that distance between the two haplotype blocks.

[6]. This variation might be attributed to different factors, such as sample size, population evolutionary history (directional selection and bottleneck), marker type used (microsatellite or SNPs), and marker density and distribution [50]. In our study, SNPs from whole genome sequence data were analyzed to estimate LD, which lead to higher marker density than genotypes from SNP genotyping arrays used in the other studies. r^2 values > 0.3 and 0.2 were considered as a useful threshold for Quantitative Trait Loci (QTLs) mapping and accurate estimation of genomic breeding values, respectively [51–53]. These values were estimated between SNP pairs separated by ≤ 400 kb and thus hold significant potential for such purposes, provided accurate and standard phenotyping measurements are available. Comparable mean r^2 values were found among the dromedary autosomes, with the exception of chromosomes 16 and 20. The observation of higher r^2 values in these two autosomes is of great interest and hence warrants further investigation, due to the potential QTLs located within them [8].

Approximately 0.5 % of the defined haplotype blocks constructed 1847 LRLD events that cover 4.8 % of the autosomal genome size. A total of 36 prevalent haplotypes were revealed among these events, which need to be further investigated to determine the biological significance and their association with dromedaries' adaptive phenotypes. LRLD events may stem from several evolutionary factors, including genetic admixture, epistatic selection, and positive selection on desired variants. The role of genetic admixture on LRLD events has been investigated by Wilson and Goldstein [13] on Bantu-Semantic hybrid population. The genomes of dromedary camel samples have also demonstrated high level of genetic admixture. This phenomenon was observed previously in dromedary camels from Arabian Peninsula using microsatellite markers [19] and whole genome sequence data [21], southeast of the Arabian Peninsula (Oman) [23], and Sudan [22]. Such evolutionary events might be predominantly linked to the historical use of dromedary camels in cross-continental transportation and trade [17]. Initially domesticated in the southeast of the Arabian Peninsula during the late second millennium [20], dromedaries facilitated trade along the "incense road" connecting the southern and northern regions of the Peninsula. This ancient trading route, which was primarily used to transport spices and perfumes, likely fostered crossbreeding between the dromedary samples from the different geographical locations of this area [54]. This highly adapted species to the harsh desert environment was also utilized to transport people from different areas of the Arabian Peninsula and Africa to Makkah, in the east of the Arabian Peninsula, during the annual pilgrimage of Islam [54]. These lengthy journeys might also enhance such crossbreeding and the associated genetic admixture. Another factor that might be linked to the observed genetic admixture among Arabian Peninsula dromedaries is the common practice of random breeding between different camel populations adopted by the owners, except possibly for camels used in racing competitions. Such breeding practices lack standardized phenotypic measurements and criteria. The impact

of genetic admixture on the defined LRDL events can be further supported upon the availability of genotyping data for the ancestral populations of the current dromedary camels. Such data would help in investigating for the introgressed haplotypes and hence the corresponding LRLD events.

Epistatic selection on alleles from different loci is one of the factors that mediates LRLD events [55]. This type of selection can be investigated upon defining functional interactions between genes from the haplotype block pairs of the LRLD events. In one example, the interaction between Tenascin X gene (*TNXB*) and the class II histocompatibility antigen (*HLA-DRA*) gene in one LRLD event on chromosome 20 is of high interest. *TNXB* is an extracellular matrix glycoprotein involved in collagen fibril deposition in connective tissue [56]. Variants on this gene and in *HLA-DRA* were found to be associated with Rheumatoid Arthritis, which is an autoimmune disease mainly affecting the connective tissue in joints [57]. Another functional interaction was determined between elongation factor 2 (*EEF2*) and 40S ribosomal protein S15 (*RPS15*) in chromosome 22. Two prevalent haplotypes were found within these two genes indicating the possible influence of epistatic selection on the corresponding LRLD event. Many of the genes found on the defined LRDL events were not functionally annotated and hence labeled as unknown loci. Improving the annotation level of the dromedary genome assembly and defining the functional roles of these loci might lead to revealing further possible interactions that may indicate the role of epistatic selection in shaping the LRLD pattern distribution on the genome of dromedary camels.

Several genes found within the LRLD events are associated with the adaptive physiologies of dromedary camels, potentially undergoing positive selection. Elevated temperatures and scarcity of water are two main characteristics of the environments surrounding Arabian Peninsula dromedary camels. Genes involved in tolerating thermal stress, such as heat shock protein 90 alpha family class A (*HSP90AA1*) and creatine kinase B (*CKB*) on chromosome 6, can be considered as hotspots for selection on the dromedary genome. These two genes were found to be associated with thermoresistance in Chinese Holstein cows [58] and adipocytes thermogenesis under cold stress in mice [59], respectively. Interestingly, a prevalent haplotype in one LRLD block was found to overlap the *CKB* gene. This haplotype needs to be functionally investigated to indicate it's role in dromedaries' adaptive physiologies. Heat shock proteins are member of the chaperone gene family involved in maintaining protein folding under heat stress [60]. Based on the DAVID analysis, chaperones and protein folding genes were found to be significantly enriched among the genes found within the LRLD events.

Dromedaries are known for their exceptional water conservation abilities during periods of dehydration, which requires unique renal functions to survive [61]. Investigations into identified haplotype blocks of this study revealed genes with possible associations to renal function. Notably, the Potassium Two Pore Domain Channel Subfamily K Member (*KCNK5*) gene found in chromosome 20, acknowledged for its involvement in renal K^+ transport, acts as a potential "molecular switch" adjusting K^+ flow based on the rate of basolateral HCO_3^- transport, as suggested by Warth, Barrière [62] in knock-out experiments on mice.

In the challenging desert heat, detrimental effects on mammalian reproductive systems can occur after prolonged heat exposure, impacting sperm cells and ovarian activity. Within the haplotype blocks, multiple genes influencing mammalian fertility have been identified. Cation Channel Sperm Associated 2 (*CATSPER2*) gene in chromosome 6 and Armadillo Repeat Containing 2 (*ARMC2*) gene in chromosome 8, shed light on possible adaptations related to fertility. A study on male mice found that *CATSPER2* was associated with driving hyperactivated sperm motility, which is essential for fertilization [63]. The *ARMC2* gene was reported to cause male infertility in humans and mice in cases of biallelic mutations, which induced multiple morphological abnormalities in the sperm flagella [64].

In the challenging environments of arid deserts characterized by scarce food resources, insulin tolerance plays a crucial role for camels. This adaptation allows them to efficiently regulate their energy usage, particularly during periods of food scarcity, ensuring their ability to endure and thrive in demanding conditions. The Glucagon-like peptide 1 receptor (*GLP-1R*) gene, found within an LRLD event in chromosome 20, is a regulator in insulin secretion as it acts as an incretin [65]. Another study using diabetic mice found that overexpression of the *ZFAND3* (Zinc Finger AN1-Type Containing 3) gene in the liver improved insulin resistance [66]. This gene, which is within an LRLD event in chromosome 20, was previously defined as a candidate gene under selection in dromedary camels [34].

The robust bone structure of camels offers vital support during long treks in deserts and enhances overall resilience in treacherous conditions. The Collagen Type V Alpha 2 Chain (*COL5A2*) gene, found in chromosome 5, and Mitogen-Activated Protein Kinase 13 (*MAPK13*) gene, found in chromosome 20, have been previously identified as positively selected genes in cetaceans associated with bone microstructure and osteoclast differentiation, respectively [67]. Another gene involved in bone structures is *B4GALNT3* (Beta-1, 4-N-Acetyl-Galactosaminyltransferase 3) found in an LRLD in chromosome 34. *B4GALNT3* knock-out mice exhibited lower bone mass and were more prone to bone fractures [68].

5. Conclusions

In this study, we have for the first time analyzed the genome of dromedary camels for LRLD events. Increasing the number of dromedary samples from different areas of the Arabian Peninsula is one of the factors that might further improve estimating their LD pattern and the LRLD distribution. Functional interactions were defined among some of the genes within the defined LRLD events, which might link epistatic selection with these events. Several biologically significant genes and prevalent haplotypes were observed among the defined LRLD regions, suggesting a role for positive selection in shaping the LRLD distribution. Improving the assembly, upon mapping the unplaced scaffolds, and the annotation of the current dromedary reference would further enhance the functional characterization of the LRLD events. Moreover, these genetic elements need to be further explored upon the inclusion of phenotypic data related to dromedaries' adaptability and productivity. Upon finding potential associations between haplotypes and/or variants on the LRLD events and the desired phenotypes, these genetic elements can be targeted through breeding programs to ensure their inheritance to the following generations and hence conserved in the species. This will help in maintaining dromedary camels'

adaptability to their local environment.

Data availability

The datasets used and analyzed during the were previously published and available in the European Nucleotide Archive (ENA) (Project numbers: PRJEB47650 (https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB47650), PRJEB66318 (https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB66318) and PRJEB67314 (https://www.ebi.ac.uk/ena/browser/text-search?query=PRJEB67314).

CRediT authorship contribution statement

Hussain Bahbahani: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The author declares that he has no financial competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e34343.

References

[1] S. Qanbari, On the extent of linkage disequilibrium in the genome of Farm animals, Front. Genet. 10 (2020) 1304.
[2] J. Bohmanova, M. Sargolzaei, F.S. Schenkel, Characteristics of linkage disequilibrium in north American Holsteins, BMC Genom. 11 (2010) 421.
[3] H.A. Al-Mamun, S.A. Clark, P. Kwan, C. Gondro, Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep, Genet. Sel. Evol. 47 (2015) 90.
[4] S. Barani, A. Nejati-Javaremi, M.H. Moradi, M. Moradi-Sharbabak, M. Gholizadeh, H. Esfandyari, Genome-wide study of linkage disequilibrium, population structure, and inbreeding in Iranian indigenous sheep breeds, PLoS One 18 (2023) e0286463.
[5] L.J. Corbin, S.C. Blott, J.E. Swinburne, M. Vaudin, S.C. Bishop, J.A. Woolliams, Linkage disequilibrium and historical effective population size in the Thoroughbred horse, Anim. Genet. 41 (Suppl 2) (2010) 8–15.
[6] I. Jasielczuk, A. Gurgul, T. Szmatoła, E. Semik-Gurgul, K. Pawlina-Tyszko, M. Stefaniuk-Szmukier, et al., Linkage disequilibrium, haplotype blocks and historical effective population size in Arabian horses and selected Polish native horse breeds, Livest. Sci. 239 (2020) 104095.
[7] A.M. Pérez O'Brien, G. Mészáros, Y.T. Utsunomiya, T.S. Sonstegard, J.F. Garcia, C.P. Van Tassell, et al., Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions, Livest. Sci. 166 (2014) 121–132.
[8] M.C. Fabbri, C. Dadousis, R. Bozzi, Estimation of linkage disequilibrium and effective population size in three Italian Autochthonous beef breeds, Animals (Basel) 10 (2020) 1034.
[9] M. Bitaraf Sani, J. Zare Harofte, A. Bitaraf, S. Esmaeilkhanian, M.H. Banabazi, N. Salim, et al., Genome-Wide diversity, population structure and demographic history of dromedaries in the central desert of Iran, Genes 11 (2020) 599.
[10] C. Liu, H. Chen, Z. Ren, X. Yang, C. Zhang, Development of genomic resources and identification of genetic diversity and genetic structure of the domestic bactrian camel in China by RAD sequencing, Front. Genet. 11 (2020) 797.
[11] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, et al., The structure of haplotype blocks in the human genome, Science 296 (2002).
[12] E. Koch, M. Ristroph, M. Kirkpatrick, Long range linkage disequilibrium across the human genome, PLoS One 8 (2013) e80754.
[13] J.F. Wilson, D.B. Goldstein, Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population, Am. J. Hum. Genet. 67 (2000) 926–935.
[14] M. Maccaferri, M.C. Sanguineti, E. Noli, R. Tuberosa, Population structure and long-range linkage disequilibrium in a durum wheat elite collection, Mol. Breed. 15 (2005) 271–290.
[15] A. El Hou, D. Rocha, E. Venot, V. Blanquet, R. Philippe, Long-range linkage disequilibrium in French beef cattle breeds, Genet. Sel. Evol. 53 (2021) 63.
[16] S. Id-Lahoucine, A. Molina, A. Cánovas, J. Casellas, Screening for epistatic selection signatures: a simulation study, Sci. Rep. 9 (2019) 1026.
[17] P.A. Burger, E. Ciani, B. Faye, Old World camels in a modern world - a balancing act between conservation and genetic improvement, Anim. Genet. 50 (2019) 598–612.
[18] J.P. Elbers, M.F. Rogers, P.L. Perelman, A.A. Proskuryakova, N.A. Serdyukova, W.E. Johnson, et al., Improving Illumina assemblies with Hi-C and long reads: an example with the North African dromedary, Mol. Ecol. Resour 19 (2019) 1015–1026.
[19] F. Almathen, H. Bahbahani, H. Elbir, M. Alfattah, A. Sheikh, O. Hanotte, Genetic structure of Arabian Peninsula dromedary camels revealed three geographic groups, Saudi J. Biol. Sci. 29 (2022) 1422–1427.

[20] F. Almathen, P. Charruau, E. Mohandesan, J.M. Mwacharo, P. Orozco-terWengel, D. Pitt, et al., Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 6707–6712.

[21] H. Bahbahani, F. Almathen, Homogeneity of Arabian Peninsula dromedary camel populations with signals of geographic distinction based on whole genome sequence data, Sci. Rep. 12 (2022) 130.

[22] H. Bahbahani, H.H. Musa, D. Wragg, E.S. Shuiep, F. Almathen, O. Hanotte, Genome diversity and signatures of selection for production and performance traits in dromedary camels, Front. Genet. 10 (2019) 893.

[23] M. Al Abri, A. Alfoudari, Z. Mohammad, F. Almathen, W. Al-Marzooqi, S. Al-Hajri, et al., Assessing genetic diversity and defining signatures of positive selection on the genome of dromedary camels from the southeast of the Arabian Peninsula, Front. Vet. Sci. 10 (2023) 1296610.

[24] R. Khalkhali-Evrigh, N. Hedayat, L. Ming, Jirimutu, Identification of selection signatures in Iranian dromedary and Bactrian camels using whole genome sequencing data, Sci. Rep. 12 (2022) 9653.

[25] H. Bahbahani, S. Al-Zoubi, F. Ali, A. Afana, M. Dashti, A. Al-Ateeqi, et al., Signatures of purifying selection and site-specific positive selection on the mitochondrial DNA of dromedary camels (Camelus dromedarius), Mitochondrion 69 (2023) 36–42.

[26] E. Mohandesan, R.R. Fitak, J. Corander, A. Yadamsuren, B. Chuluunbat, O. Abdelhadi, et al., Mitogenome sequencing in the genus Camelus reveals evidence for purifying selection and long-term divergence between wild and domestic bactrian camels, Sci. Rep. 7 (2017) 9970.

[27] R.R. Fitak, E. Mohandesan, J. Corander, A. Yadamsuren, B. Chuluunbat, O. Abdelhadi, et al., Genomic signatures of domestication in Old World camels, Commun. Biol. 3 (2020) 316.

[28] F. Almathen, H. Elbir, H. Bahbahani, J. Mwacharo, O. Hanotte, Polymorphisms in MC1R and ASIP genes are associated with coat color variation in the arabian camel, J. Hered. 109 (2018) 700–706.

[29] T. Maraqa, B.H. Alhajeri, H. Alhaddad, FGF5 missense mutation is associated with dromedary hair length variation, Anim. Genet. 52 (2021) 848–856.

[30] A. Ali, B. Baby, R. Vijayan, From Desert to medicine: a review of camel genomics and therapeutic products, Front. Genet. 10 (2019) 17.

[31] A. Tibary, K. El Allali, Dromedary camel: a model of heat resistant livestock animal, Theriogenology 154 (2020) 203–211.

[32] H. Musa, E. Shuiep, I. El-Zubeir, Camel husbandry among pastoralists in darfur, western Sudan, Nomadic Peoples 10 (2006) 101–105.

[33] H. Wu, X. Guang, M.B. Al-Fageeh, J. Cao, S. Pan, H. Zhou, et al., Camelid genomes reveal evolution and adaptation to desert environments, Nat. Commun. 5 (2014) 5188.

[34] H. Bahbahani, A. Alfoudari, A. Al-Ateeqi, M. Al Abri, F. Almathen, Positive selection footprints and haplotype distribution in the genome of dromedary camels, Animal 18 (2024) 101098.

[35] S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.

[36] H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform, Bioinformatics 26 (2010) 589–595.

[37] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993.

[38] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303.

[39] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575.

[40] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, GigaScience 4 (2015), 7-7.

[41] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, et al., The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158.

[42] D.H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals, Genome Res. 19 (2009) 1655–1664.

[43] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2009. New York.

[44] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[45] D. Taliun, J. Gamper, C. Pattaro, Efficient haplotype block recognition of very long and dense genetic sequences, BMC Bioinf. 15 (2014) 10.

[46] S.-S. Dong, W.-M. He, J.-J. Ji, C. Zhang, Y. Guo, T.-L. Yang, LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files, Briefings Bioinf. 22 (2020) bbaa227.

[47] B.T. Sherman, M. Hao, J. Qiu, X. Jiao, M.W. Baseler, H.C. Lane, et al., DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update), Nucleic Acids Res. 50 (2022) w216–w221.

[48] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2009) 44–57.

[49] D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, et al., STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res. 47 (2018) D607–D613.

[50] I. Jasielczuk, A. Gurgul, T. Szmatoła, E. Semik-Gurgul, K. Pawlina-Tyszko, M. Szyndler-Nędza, et al., Comparison of linkage disequilibrium, effective population size and haplotype blocks in Polish Landrace and Polish native pig populations, Livest. Sci. 231 (2020) 103887.

[51] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics 157 (2001) 1819–1829.

[52] K.G. Ardlie, L. Kruglyak, M. Seielstad, Patterns of linkage disequilibrium in the human genome, Nat. Rev. Genet. 3 (2002) 299–309.

[53] F.X. Du, A.C. Clutter, M.M. Lohuis, Characterizing linkage disequilibrium in pig populations, Int. J. Biol. Sci. 3 (2007) 166–178.

[54] R.T. Wilson, Camels the Tropical Agriculturalist, Macmillan, Education, 1998.

[55] R.V. Rohlfs, W.J. Swanson, B.S. Weir, Detecting coevolution through allelic association between physically unlinked loci, Am. J. Hum. Genet. 86 (2010) 674–685.

[56] J.R. Mao, G. Taylor, W.B. Dean, D.R. Wagner, V. Afzal, J.C. Lotz, et al., Tenascin-X deficiency mimics Ehlers-Danlos syndrome in mice through alteration of collagen deposition, Nat. Genet. 30 (2002) 421–425.

[57] W. Zheng, S. Rao, Knowledge-based analysis of genetic associations of rheumatoid arthritis to inform studies searching for pleiotropic genes: a literature review and network analysis, Arthritis Res. Ther. 17 (2015) 202.

[58] T.M. Badri, K.L. Chen, M.A. Alsiddig, L. Li, Y. Cai, G.L. Wang, Genetic polymorphism in Hsp90AA1 gene is associated with the thermotolerance in Chinese Holstein cows, Cell Stress Chaperones 23 (2018) 639–651.

[59] J.F. Rahbani, J. Bunk, T. Lagarde, B. Samborska, A. Roesler, H. Xiao, et al., Parallel control of cold-triggered adipocyte thermogenesis by UCP1 and CKB, Cell Metabol. 36 (2024) 526–540.

[60] E.A. Craig, B.D. Gambill, R.J. Nelson, Heat shock proteins: molecular chaperones of protein biogenesis, Microbiol. Rev. 57 (1993) 402–414.

[61] B.D. Siebert, W.V. Macfarlane, Water turnover and renal function of dromedaries in the desert, Physiol. Zool. 44 (1971) 225–240.

[62] R. Warth, H. Barrière, P. Meneton, M. Bloch, J. Thomas, M. Tauc, et al., Proximal renal tubular acidosis in TASK2 K+ channel-deficient mice reveals a mechanism for stabilizing bicarbonate transport, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 8215–8220.

[63] T.A. Quill, S.A. Sugden, K.L. Rossi, L.K. Doolittle, R.E. Hammer, D.L. Garbers, Hyperactivated sperm motility driven by CatSper2 is required for fertilization, Proc. Natl. Acad. Sci. USA 100 (2003) 14869–14874.

[64] C. Coutton, G. Martinez, Z.E. Kherraf, A. Amiri-Yekta, M. Boguenet, A. Saut, et al., Bi-Allelic mutations in ARMC2 lead to severe astheno-teratozoospermia due to sperm flagellum malformations in humans and mice, Am. J. Hum. Genet. 104 (2019) 331–340.

[65] Y. Zheng, F. Wu, M. Zhang, B. Fang, L. Zhao, L. Dong, et al., Hypoglycemic effect of camel milk powder in type 2 diabetic patients: a randomized, double-blind, placebo-controlled trial, Food Sci. Nutr. 9 (2021) 4461–4472.

[66] K. Shimizu, Y. Ogiya, K. Yoshinaga, H. Kimura, S. Michinaga, M. Ono, et al., ZFAND3 overexpression in the mouse liver improves glucose tolerance and hepatic insulin resistance, Exp. Clin. Endocrinol. Diabetes 130 (2022) 254–261.

[67] D. Sun, X. Zhou, Z. Yu, S. Xu, I. Seim, G. Yang, Accelerated evolution and diversifying selection drove the adaptation of cetacean bone microstructure, BMC Evol. Biol. 19 (2019) 194.

[68] S. Movérare-Skrtic, J. Voelkl, K.H. Nilsson, M. Nethander, T.T.D. Luong, I. Alesutan, et al., B4GALNT3 regulates glycosylation of sclerostin and bone mass, EBioMedicine 91 (2023) 104546.