Signatures of purifying selection and site-specific positive selection on the mitochondrial DNA of dromedary camels (Camelus dromedarius)

Signatures of purifying selection and site-specific positive selection on the mitochondrial DNA of dromedary camels (*Camelus dromedarius*)

Hussain Bahbahani[1*], Sanaa Al-Zoubi[1], Fatima Ali[1], Arwa Afana[1], Mohammed Dashti[2], Abdulaziz Al-Ateeqi[3], David Wragg[4], Suzanne Al-Bustan[1], Faisal Almathen[5,6]

[1] Department of Biological Sciences, Faculty of Science, Kuwait University, Sh. Sabah Al-Salem campus, Kuwait.

[2] Genetics and Bioinformatics Department, Dasman Diabetes Institute, Kuwait City, Kuwait

[3] Environment and Life Sciences research center, Kuwait Institute for Scientific Research, Kuwait City, Kuwait.

[4] The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK.

[5] Department of Veterinary Public Health and Animal Husbandry, College of Veterinary Medicine, King Faisal University, 400 Al-Ahsa, Kingdom of Saudi Arabia.

[6] Camel Research Center, King Faisal University, 400 Al-Ahsa, Saudi Arabia.

*Corresponding author: Hussain Bahbahani (hussain.bahbahani@ku.edu.kw)

Running title: Purifying and positive selections on camel mtDNA

Key words:
adaptive divergence, mtDNA protein-coding genes, signatures of selection, PAML

Abstract

The two species of the Old World Camelini tribe, dromedary and Bactrian camels, show superior adaptability to the different environmental conditions they populate, e.g. desert, mountains and coastal areas, which might be associated with adaptive variations on their mitochondrial DNA. Here, we investigate signatures of natural selection in the 13-mitochondrial protein-coding genes of different dromedary camel populations from the Arabian Peninsula, Africa and southwest Asia. The full mitogenome sequences of 42 dromedaries, 38 domestic Bactrian, 29 wild Bactrian camels and 31 samples representing the New World Lamini tribe reveal species-wise genetic distinction among Camelidae family species, with no evidence of geographic distinction among dromedary camels. We observe gene-wide signals of adaptive divergence between the Old World and New World camels, with evidence of purifying selection among Old World camel species. Upon comparing the different Camelidae tribes, 27 amino acid substitutions across ten mtDNA protein-coding genes were found to be under positive selection, in which, 24 codons were defined to be under positive adaptive divergence between Old World and New World camels. Seven codons belonging to three genes demonstrated positive selection in dromedary lineage. A total of 89 codons were found to be under positive selection in Camelidae family based on investigating the impact of amino acid replacement on the physiochemical properties of proteins, including equilibrium constant and surrounding hydrophobicity. These mtDNA variants under positive selection in the Camelidae family might be associated with their adaptation to their contrasting environments.

Introduction

The Camelidae family is divided into two tribes: Old World camels (Camelini) and New World camels (Lamini). The Camelini tribe is composed of two domestic species: single-humped dromedary camels (*Camelus dromedarius*) and two-humped Bactrian camels (*Camelus bactrianus*), and a two-humped wild Bactrian camel species (*Camelus ferus*). The Lamini tribe, which diverged from the Camelini tribe approximately 16.3 million years ago, is composed of two domestic species: Llama (*Lama glama*) and Alpaca (*Vicugna pacos*), and two wild species: Guanaco (*Lama guanicoe*) and Vicuna (*Vicugna Vicugna*) (Wu et al., 2014, Burger et al., 2019).

The camel species of these two tribes populate contrasting environmental niches. Dromedary camels, also known as Arabian camels, are mainly distributed throughout hot arid regions of Africa, Arabian Peninsula and southwest Asia, while some populations inhabit the mountains and coast of the west, southwest and southeast of the Arabian Peninsula (Almathen et al., 2018). Domestic and wild Bactrian camels are widespread in central Asia, including Mongolia, China and Kazakhstan, throughout cold plains and rocky mountains reaching altitudes up to 4000m (Burger et al., 2019). Camels of the Lamini tribe populate the high-altitude mountains of South America which can reach 7000m in height (Wu et al., 2014).

These contrasting environmental conditions might be associated with different adaptations associated with energy production and thermoregulation. For example, under water stress, dromedary camels show adaptive heterothermy, during which their body temperature fluctuates by 6^{o}C – 8^{o}C, between 34^{o}C to more than 40^{o}C, to avoid water loss through perspiration during high temperatures (Tibary and El Allali, 2020). While the environmental

niche occupied by Lamini species is mainly characterized by high altitude hypoxia, for which Lamini camels exhibit superior adaptability (Beall, 2007).

The mitochondria are involved in producing ATP molecules as an energy source and generating heat to maintain body temperature (Blier et al., 2001). This organelle carries its own genetic material comprising 13 protein-coding genes (*ND1, ND2, ND3, ND4, ND4L, ND5, ND6, CYTB, COX-1, COX-2, COX-3, ATP6* and *ATP8*), 22 tRNA and 2 rRNA genes. The encoded proteins are part of the electron transport chain (Complex I – Complex V) embedded in the inner-membrane of each mitochondrion, which have been found to be targets of selective pressures to meet these two primary physiological functions (Blier et al., 2001).

The analysis of mtDNA protein-encoding genes has been shown to be effective at identifying footprints of purifying and adaptive (positive) selection in a range of species. Ruiz-Pesini *et al*. (2004) report that in humans living within the arctic zone, several mitochondrial genes exhibit signatures of adaptive selection, including *ND2*, *ND4* and *ATP6*. Signatures of positive selection in arctic dwelling hare have also been identified in mitochondrial genes, including *ATP8*, *CYTB* and *ND4* (Melo-Ferreira *et al*., 2014). To date, analyses of the mitogenomes of Old World camels has largely concerned diversity analyses of partial mtDNA sequences, which indicates limited phylogeographic structuring among the dromedary camels in Asia and Africa (Almathen et al., 2016, Alaqeely et al., 2021). A study by Mohandesan et al. (2017) analyzed the full mitogenome of the three Old World camel species, revealing signals of gene-wide purifying selection including positive selection in 18 sites or "codons" among the three species. However, only a limited number of samples (≤ 10) were analysed for each species.

Here, we investigated the mtDNA protein-coding genes of dromedary camels (n = 42) distributed along Africa, Arabian Peninsula and southwest Asia for signatures of selection at gene-wide and site-specific levels. Publicly available full mitogenome sequences of camels representing domestic Bactrian (n = 38), wild Bactrian (n = 29) and New World Lamini (n = 31) were also included for comparison.

Materials and Methods

Whole genome sequence data from the Camelidae family

Genomic DNA of seven dromedary camels, which included two from Kuwait and five from Saudi Arabia, was extracted from 5ml blood samples using the DNeasy® Blood and Tissue kit (Qiagen). Whole-genome sequencing of DNA was performed on the Illumina Hiseq 2000 platform using 150 bp paired-end libraries with 270 bp insert size. Sequencing of the Kuwaiti samples was performed by the BGI in China, and at Macrogen in South Korea for the Saudi samples (Table S1).

Whole-genome sequencing data for 102 samples was retrieved from the European Nucleotide Archive. This included: 21 dromedary samples from the Arabian Peninsula (Bioproject accession number: PRJEB47650) (Bahbahani and Almathen, 2022); four dromedaries from Iran, 31 domestic Bactrian camels, and 19 wild Bactrian camels from Central Asia (Bioproject accession number: PRJNA383081) (Ming et al., 2020); and 27 camel samples belong to the Lamini tribe from South America: Llama (n = 7), Vicuna (n = 6), Guanaco (n = 7), and Alpaca (n = 7) (Bioproject accession number: PRJNA612032) (Fan et al., 2020) (Table S2).

Raw sequence read processing and full mtDNA sequence generation

The raw sequencing reads were processed using Fastp v0.22 (Chen et al., 2018). This included trimming sequencing adapters, and discarding reads if: (1) 50% or more of the read bases were uncertain; (2) 40% or more of the read bases were low quality (base $Q_{phred} \leq 20$); (3) reads length was shorter than 15 bases; or (4) read complexity was less than 30%. The remaining high-quality reads of the dromedary, domestic Bactrian, wild Bactrian, and Lamini camels were aligned to reference mtDNA sequences for their respective species: dromedary (NC_009849.1); domestic Bactrian (NC_009628.2); wild Bactrian (NC_009629.2); and Alpaca (NC_002504.1). Reads were aligned using the *bwa-mem* algorithm implemented in Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010). The resulting alignments were coordinate-sorted using the *SortSam* option, and PCR-duplicates were removed using the *MarkDuplicates* and *REMOVE_DUPLICATES=true* options in the Picard tools v1.119 (http://broadinstitue.github.io/picard/index.html).

Single Nucleotide Polymorphisms (SNPs) were called using the *HaplotypeCaller* algorithm implemented in Genome Analysis Toolkit (GATK) v4.1 (McKenna et al., 2010). Quality control filtering criteria were applied on the per-sample variants using the *VariantFilteration* algorithm in GATK. This included: (1) excluding variants with low quality by depth (QD) (QD < 2); (2) excluding variants with root mean square of mapping quality for all reads of a site less than 40 (MQ < 40.0); (3) excluding variants with base quality score less than 30 (QUAL < 30); (4) excluding variants with high probability of allele-specific strand bias between forward and reverse strand (FS > 60); (5) excluding variants with bias in mapping quality between the reads supporting the reference and alternative alleles (MQRankSum < -12.5); and (6) excluding variants with bias in the position of the alternative allele towards the ends of the reads (ReadPosRankSum <-8). SNPs with depth of coverage ranging between two reads and three

153 standard deviation (SD) from the mean depth of coverage across samples were retained. Fasta
154 files of the full mtDNA sequences of each sample were generated using
155 *FastaAlternateReferenceMaker* algorithm in GATK v4.1. (File S1)
156
157 **mtDNA diversity statistics and molecular phylogeny**
158 The generated mtDNA sequences were aligned with the full mtDNA sequences downloaded
159 from NCBI's Genbank using the *Clustal W* algorithm implemented in MEGA v10 (Kumar et
160 al., 2018). The NCBI sequences included: dromedary camels (n = 10), domestic Bactrian
161 camels (n = 7), wild Bactrian camels (n = 10), Llama (n = 1), Vicuna (n = 1), Guanaco (n = 1),
162 and Alpaca (n = 1) (Table S3). The number of variable sites, parsimony-informative and
163 singletons, were calculated with DnaSP v6 (Librado and Rozas, 2009) for the 140 mtDNA
164 sequences. Haplotype (H_d) and nucleotide (π) diversities were calculated for dromedaries,
165 domestic Bactrian, wild Bactrian, and Lamini samples, separately, using DnaSP. Tajima's D
166 neutrality index and Fu and Li's F statistics were also calculated with DnaSP to determine
167 historical population expansions and contractions of the different Camelidae species.
168
169 We identified the best-fitting DNA evolution model according to the Bayesian Information
170 Criterion (BIC) implemented in MEGA software which was the GTR (General Time
171 Reversible) model with gamma correction parameter = 0.56. A neighbor-joining distance tree
172 was constructed using MEGA, employing 1000 bootstrap replications to assess the mtDNA
173 relationship among the camel species. ggtree package (Yu, 2020) for R software version 4.2
174 (R Core Team, 2022) was used to plot the tree.
175
176
177

178 **Gene-wide signatures of selection analyses**

179 Signatures of selection were investigated on the 13 mtDNA protein-coding genes at a gene-
180 wide level using three approaches: McDonald-Kreitman (MK); neutrality index (NI); and ω
181 ratio tests using DnaSP. The MK test measures the ratio of nonsynonymous (Pn) to
182 synonymous (Ps) substitutions within species as compared to the ratio of nonsynonymous (Kn)
183 to synonymous (Ks) substitutions between species. A Kn/Ks > Pn/Ps indicates a signal of
184 positive selection, while a Kn/Ks < Pn/Ps and Kn/Ks = Pn/Ps are signals of purifying selection
185 and neutrality, respectively (McDonald and Kreitman, 1991). NI calculates the ratio of
186 polymorphic sites within species to the fixed sites between species at nonsynonymous and
187 synonymous substitutions (NI=(Pn * Ks)/(Kn * Ps)). Whereby NI = 1 equates no selection, NI
188 > 1 indicates purifying selection, while NI < 1 indicates positive selection. The ω ratio is the
189 ratio of non-synonymous substitutions per non-synonymous sites (Dn) to the synonymous
190 substitutions per synonymous sites (Ds). A Dn/Ds ratio > 1 indicates positive selection, while
191 a Dn/Ds ratio < 1 indicates purifying selection, and Dn/Ds = 1 indicates no selection. Two-
192 tailed Fisher's exact test *P*-values were computed to define statistically significant signals for
193 the MK and NI tests. Each of these analyses were conducted at three levels: (1) comparing
194 dromedary camels (n = 42) to domestic and wild Bactrian (n = 67); (2) comparing the
195 dromedary camels (n = 42) to the New World Lamini camels (n = 31); and (3) comparing Old
196 World camels (dromedary and Bactrian) (n = 109) to the New World Lamini camels (n = 31).

197

198 **Site-specific signatures of selection analyses**

199 Signatures of selection at the site-specific level were identified along the 13 mtDNA protein-
200 coding genes using the site models implemented in the CODEML package of Phylogenetic
201 Analysis by Maximum Likelihood (PAML) v4.9 (Yang, 1997, Yang, 2007, Yang et al., 2000).
202 Several different models implemented in the CODEML package were evaluated. The one ratio

(M0) model allows a single ω for all codons. The nearly neutral model (M1a) assumes two classes of codons: one with $0 \leq \omega_0 < 1$ and a proportion of codons p_0, while the second class assumes $\omega_1 = 1$ and a proportion of codons $p_1 = 1 - p_0$. Model M2a (positive selection) is an extension of the M1a model with a third class that allows for $\omega_2 > 1$ and a proportion of codons $p_2 = 1 - p_1 - p_0$. Model M3 (discrete) uses by default three discrete classes to model the heterogeneity of ω between codons. The M7 and M8 models assume a beta distribution of ω over codons with two beta function parameters (p and q). The M7 model does not allow for codons under positive selection by constraining ω to be in the interval (0, 1). In contrast, the M8 model allows for codons with $\omega_1 > 1$ with a proportion of codons p_1. After calculating the log likelihood value (L) of each model fitted to our data, likelihood ratio tests (LRTs) were conducted between the M1a–M2a and M7–M8 models to test for positive selection, and between the M0-M3 models to test for variable selection pressures among codons. The statistic for each LRT is defined as twice the log likelihood difference between two models ($2\Delta L$). This statistic was compared to a χ^2 distribution with a degree of freedom (d.f.) equal to the difference in the number of parameters between the two models (d.f. = 2 for M1a–M2a and M7–M8, and d.f. = 4 for M0-M3). A Bayes Empirical Bayes (BEB) approach was used to identify codons under positive selection. Sites with BEB posterior probability > 0.5 were considered to be under positive selection, with a value > 0.95 considered as extreme selection (Yang et al., 2005). The site model analyses were conducted at the three different levels described previously (1: dromedary with domestic and wild Bactrian; 2: dromedary with New World Lamini; 3: Old World camels with New World Lamini).

Branch-site models of positive selection, implemented in the CODEML package (Zhang et al., 2005), were also tested on the mtDNA protein-coding genes to detect codons under positive selection in a specific lineage, the "foreground lineage", but which remained neutral or under

purifying selection in the other lineages, the "background lineages". The dromedary camels were set as the foreground lineage, while the other camel populations were set as background lineages. On the branch-site models, Model A assumes positive selection by defining four classes of codons. Class 0 includes codons with $0 \leq \omega_0 < 1$ throughout the tree. Class 1 includes codons with $\omega_1 = 1$ throughout the tree. Class 2a includes codons with $0 \leq \omega_0 < 1$ in the background lineages, and $\omega_2 \geq 1$ in the foreground lineage. Class 2b includes codons with $\omega_1 = 1$ in the background lineages, and $\omega_2 \geq 1$ in the foreground lineages. The null model, which does not assume positive selection, is the same as Model A, but fixes $\omega_2 = 1$. As in the site models, the log likelihood value (L) was calculated for each model fitted to our data, and LRTs were conducted between Model A and the null model to test for positive selection. The calculated log likelihood statistic for the LRT was compared to a χ^2 distribution (d.f. = 1). A BEB approach was used to identify codons under positive selection as in the site models.

Codons under positive selection were also determined using TreeSAAP v3.2 (Woolley et al., 2003). TreeSAAP identifies selective influences based on 31 structural and biochemical properties. It accounts for the impact magnitude of amino acid changes on physiochemical properties of the protein, and tests if the observed degree of amino acid substitution deviates from the neutral expectation. If the change in magnitude ≥ 6 and the goodness-of-fit test *P*-value < 0.001, it is considered a strong indication of positive selection in the physiochemical property tested. For estimating the significance of changes, the software outputs two main values: category value and statistical z-score value (equivalent to *P*-value). The category values were numbered from (1 to 8); 1 is the most conservative amino acid category and 8 is the most radical value. Category values ≥ 6 and z-score value ≥ 3.09 are considered as significant signals of positive selection. A phylogenetic tree for each mitochondrial protein-coding gene in the

252 different species of the Camelidae family was constructed using MEGA and used as an input
253 for TreeSAAP.

254

255 **Results**

256 **Camelidae full mtDNA diversity statistics and molecular phylogeny**

257 The alignment of the 140 full mtDNA sequences returned 3,520 variable sites, of which 191
258 were singletons and 3,329 were parsimony-informative. Higher nucleotide and haplotype
259 diversities were calculated for Lamini samples (H_d = 0.998 and π = 0.0254) than for those of
260 the Camelini tribe (Table 1). Among the Camelini tribe, relatively lower nucleotide diversity
261 and higher haplotype diversity values were observed in dromedaries (π = 0.00109 and H_d =
262 0.997) than in domestic Bactrian (π = 0.00124 and H_d = 0.989) and wild Bactrian camels (π =
263 0.00115 and H_d = 0.781). Negative Tajima's D and Fu and Li's F statistics were calculated for
264 dromedary camels (Tajima's D = -2.24, Fu and Li's F = -2.87) and domestic Bactrian camels
265 (Tajima's D = -1.223, Fu and Li's F = -2.42), which deviated significantly from neutrality in
266 dromedaries. Positive Tajima's D and Fu and Li's F statistics were calculated for wild Bactrian
267 camels (Tajima's D = 2.655, Fu and Li's F = 2.067) and Lamini camels (Tajima's D = 0.167,
268 Fu and Li's F = 0.524), which deviated significantly from neutrality in the wild Bactrian
269 camels.

270

271

272

273

274

275

Table 1: Haplotype and nucleotide diversities, Tajima's D and Fu and Li's F statistics for dromedary, domestic Bactrian, wild Bactrian and Lamini camels included in the study.

	C. dromedarius	*C. bactrianus*	*C. ferus*	*Lamini*
Haplotype diversity (Hd)	0.997	0.989	0.781	0.998
Nucleotide diversity (π)	0.00109	0.00124	0.00124	0.0254
Tajima's D	-2.24*	-1.22	2.655*	0.167
Fu and Li's F	-2.87*	-2.42	2.067*	0.524

*Significant values with *P*-value < 0.05

The neighbour-joining phylogenetic tree revealed species-wise mtDNA distinction among the Camelidae family, separate dromedary, Bactrian and Lamini camels from each other. Within the Lamini tribe, the Vicuna camels were differentiated from the other New World camels, while among the two-humped Old World camels we observed differentiation between the domestic and wild Bactrian camels. No geographic-genetic distinction was observed among the different dromedary camel populations (Fig. 1).

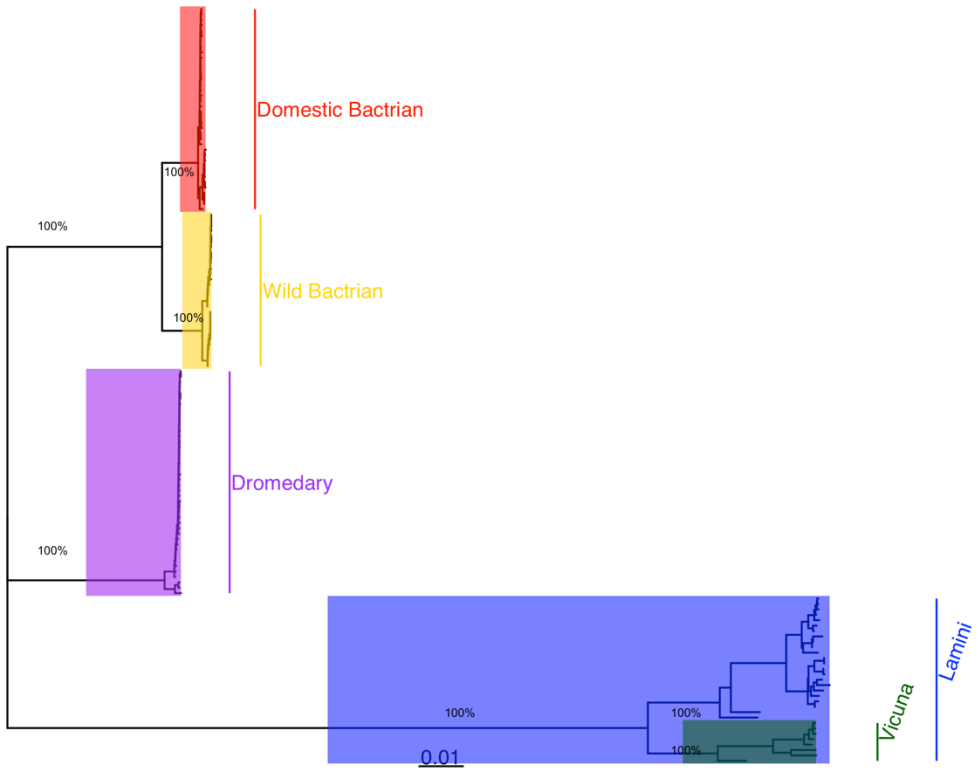

Fig. 1. Neighbor joining phylogenetic tree of the Camelidae family based on the full mitogenome.

Gene-wide level signatures of selection

Based on the McDonald-Kreitman (MK) test and Neutrality index (NI), significant signals of purifying selection were obtained in six genes; *ND1, COX1, ND4, ND5, ND6*, and *CYTB,* upon comparing dromedary with domestic and wild Bactrian camels. While non-significant signals of purifying and positive selection were found across the remaining genes (Table 2 and Table S4). Comparing dromedary with Lamini camels revealed significant signals of positive selection in *ND5, ATP6* and *COX3* (Table 2 and Table S4). Comparing Old World camels with Lamini camels showed significant signals of positive selection in *ND1, ND4L*, *ND5, ATP6 and COX3* (Table 2 and Table S4).

Table 2: Types of selection signals defined on the mitochondrial protein-coding genes based on McDonald-Kreitman test and Neutrality index on the three different analyses levels.

	Dromedary with domestic and wild Bactrian		Dromedary with Lamini		Old World with Lamini	
Gene	**Type of selection**	***P*-value[1]**	**Type of selection**	***P*-value[1]**	**Type of selection**	***P*-value[1]**
ND1	Purifying	0.019*	Positive	0.061	Positive	0.022*
ND2	Positive	0.471	Positive	0.531	Positive	0.748
COX1	Purifying	0.044*	Positive	1	Positive	0.390
COX2	Positive	1	Positive	0.327	Positive	0.070
ATP8	Purifying	1	Positive	0.158	Positive	0.603
ATP6	Positive	0.706	Positive	0.017*	Positive	0.041*
COX3	Purifying	0.302	Positive	0.015*	Positive	0.02*
ND3	Purifying	0.637	Positive	0.783	Positive	0.619
ND4L	Purifying	0.492	Positive	0.166	Positive	0.012*
ND4	Purifying	0.043*	Positive	0.577	Positive	0.295
ND5	Purifying	0.015*	Positive	0.038*	Positive	0.031*
ND6	Purifying	0.009**	Purifying	0.101	Positive	1
CYTB	Purifying	0.0003***	Purifying	0.532	Purifying	0.378

[1] The *P*-values are the significant values of the two-tailed fisher's exact test. *values with asterisk are significant with significant level of *P<0.05, **P<0.01, ***P<0.001.*

The ratio of non-synonymous substitutions per non-synonymous sites (Dn) to the synonymous substitutions per synonymous sites (Ds) (ω ratio) was less than one, indicating purifying selection, on all the mtDNA protein-coding genes at the different analyses. Among the thirteen genes, *ATP8* showed the highest ω ratio, whilst, *COX1* and *COX2* showed the lowest ω ratios (Fig. S1 and Table S5).

Site-specific signatures of selection

The LRT results from the M0-M3 models, which tests for variable selection pressures among codons, returned significant signals of selection both for dromedary with Lamini, and Old World with New World Lamini camels, in which a proportion of these sites had evolved under positive selection with ω value larger than 1 (Table S6). Twenty-seven amino acid substitutions in ten mtDNA protein-coding genes were identified to be under positive selection by Bayes Empirical Bayes (BEB) analyses. This included seven codons among dromedary with domestic and wild Bactrian, 21 codons among dromedary with Lamini, and 24 codons among Old World camels with New World Lamini camels (Table S7). Two of these codons are in genes for which positive selection models (M8 and M2a) fit the data significantly better than neutral models (Table 3). These are *COX2* (dromedary with Lamini, M8 and M2a) and *ND6* (Old World with New World Lamini camels, M8) (Table S7). Of particular note is codon 138 in *COX2* which, among dromedary with Lamini and Old World camels with New World Lamini, is under extreme selection based on the BEB posterior probability value.

Table 3: Positively selected codons identified by the site-models based on the posterior probability of the Bayes Empirical Bayes (BEB).

Gene	Codon	Amino acids in Dromedary	Amino acids in Bactrian	Amino acids in Lamini	Level of analysis
COX2	138	V	V	W	Dromedary with Lamini (BEB=0.996) (*P*-value <0.05)
					Old World with New World (BEB=0.996) (P-value <0.05)
ND6	102	I	S,A	L	Old World with New World (BEB=0.921) (*P*-value <0.01)

P-values are the significant values of the M8-M7 and M1a-M2a Likelihood ratio tests.

Based on the BEB posterior probability values of the branch-site model with dromedary camels as the foreground lineage, a total of seven codons within three genes were found to be under

positive selection. These include codons 34 and 66 in *ATP8*, codons 1 and 51 in *ATP6,* and codons 101, 116 and 133 in *ND6*. The three codons identified in *ND6* exhibit intra-species variation among the dromedaries. None of the genes carrying these codons have a significantly better fit to the positive selection model (Model A) than the neutral null model (Table 4 & Table S8).

Table 4: Sites identified by branch-site models, based on Bayes Empirical Bayes (BEB), as potentially under selection.

Gene	Codon	Amino acid in Foreground lineage	Amino acid in Background lineage	BEB Posterior probability	*P*-value
ATP8	34* 66	P L	L , Y Q	0.795 0.542	$P > 0.05$
ATP6	1* 51	V Q	M R	0.513 0.52	$P > 0.05$
ND6	101	I (V in Omani and Kenya samples)	I, M	0.833	$P > 0.05$
	116	I (V in Omani and Kenya samples)	I	0.909	
	133	S (F in Omani and Kenya samples)	S	0.901	

P-values are the significant values of the positive selection alternative and neutral null models likelihood ratio test of the branch-site models.
* Sites identified in Mohandesan et al., (2017) to be under positive selection.

We next sought to investigate the magnitude of the impact of amino acid replacement on the physiochemical properties of the proteins. Using TreeSAAP we identified 89 codons in nine genes to be under positive selection. These included eight codons in *ND1*, eleven codons in *ND2*, two codons in *ATP6*, five codons in *COX3*, seven codons in *ND3*, seven codons in *ND4*, twenty-seven codons in *ND5*, ten codons in *ND6*, and twelve codons in *CYTB* genes. Of these codons, 71 were found to be under positive selection between dromedaries and Lamini. Positive selection was also observed in 26 and 25 codons between dromedaries and domestic Bactrian, and dromedaries and wild Bactrian camels, respectively. In particular, two amino acid replacements (codon 9 in *ND4* and codon 101 in *ND6*) were found to be polymorphic among dromedaries, differentiating the two Omani camels and a single Kenyan dromedary from the other dromedaries. Out of the 31 physiochemical properties considered by TreeSAAP, we identified significant changes in six properties: the equilibrium constant (ionization COOH) (pK′); power to be at the middle of alpha helix (α_m); long range non-bonded energy (E_l); solvent accessible reduction ratio (R_a); surrounding hydrophobicity (H_p); and alpha helical tendencies (P_α) (Table S9).

Discussion

By analyzing the full mitogenomes of the Camelidae family species we have identified signatures of positive adaptive divergence between Camelini and Lamini tribes. Within Camelini species, signals of gene-wide purifying selection and site-specific positive selection were also observed that might be associated with their adaptation to the different environmental conditions they are populating.

Signals of past population expansion and/or positive selection in dromedary camels can be observed based on the significant negative Tajima's D and Fi and Lu's F statistics. These signals might relate to historical population growth and the continental-wise distribution of dromedaries throughout Asia and Africa. While significant positive values in wild Bactrian and non-significant negative values in domestic Bactrian camels potentially relates to the restricted distribution of wild Bactrian camels, being found only in a protected area in the Mongolian Gobi Desert, and the limited distribution of the domestic Bactrian camels in Central Asia (Burger et al., 2019). These results are in agreement with those reported in Camelini species by Mohandesan et al. (2017). Positive Tajima's D and Fu and Li's F statistics were observed for New World Lamini species, which might also be related to the confined distribution of these camels in South America (Fan et al., 2020).

The phylogenetic analysis revealed inter-species genetic distinction between Camelini and Lamini tribes. Within the Camelini tribe, dromedary camels separated from Bactrian camels, which also show genetic distinction between domestic and the wild-type. However, neither continental-wise nor population-wise genetic distinction was observed among the different dromedary camel populations included in the study. The observed inter-species genetic distinction between the Camelini and Lamini tribes is likely resulted from their divergence about 16.3 Mya, which was followed by the later migration of the Camelini species to Eurasia, about 6.5 – 7.5 Mya, and the Lamini tribe to South America about 3 Mya (Burger et al., 2019). The later divergence between the dromedary and Bactrian camels, about 4.4 Mya, and the different geographical distribution of these two species may explain the mitochondrial genetic distinction observed between them. Such separation has also been observed previously at autosomal level using whole genome sequence data (Ming et al., 2020, Bahbahani and Almathen, 2022). The genetic distinction between domestic and wild Bactrian camels can

likely be attributed to the 1.1 Mya divergence between them (Burger et al., 2019). Genetic separation is also observed between the wild vicuna species and the other Lamini species, with *Lama* and *Vicugna* diverging around 2-3 Mya (Fan et al., 2020).

The lack of genetic distinction among the different dromedary camel populations may be due to the historical purpose of this well-adapted species in transporting goods and people between Asia and Africa, and throughout the Arabian Peninsula. Dromedary camels were initially used in transportation along the "incense road" connecting the southern to the northern parts of Arabian Peninsula. Dromedaries were used to transport goods between Africa and Asia during the 1st millennium Before Common era (BC) via the Islands of Socotra. In parallel to transporting goods, dromedaries were used to transfer people from Africa and different parts of the Arabian Peninsula to Makkah during the annual Pilgrimage (Wilson, 1998). All of these practices were associated with high interbreeding and gene flow between the dromedary camels along the region, which have also been observed using whole genome sequence data (Bahbahani and Almathen, 2022). Current breeding practices of local camel owners, which rely on random mating between camels, further enhance gene flow between dromedary populations.

Although the ω ratios indicate signals of purifying selection on all mtDNA protein-coding genes at the varying levels of analysis, significant signals of positive selection were identified on *ND1, ND4L*, *ND5, ATP6 and COX3* upon comparing Old World camels with New World Lamini. This indicates adaptive divergence between these two tribes, which might be attributed to the different environmental conditions they inhabit. The various species of Camelini populate the desert and mountainous areas of Asia and Africa, adapting to arid and semi-arid

416 conditions including tolerance to high temperature, and scarcity of food and water. In contrast,
417 Lamini species inhabit the high altitudes of south America, reaching more than 7000 m, and
418 thus face environmental challenges characterized with high altitude such as hypoxia (Beall,
419 2007, Wu et al., 2014). These environmental challenges present significant pressure on mtDNA
420 genes, for example to adapt to low oxygen concentration in the environment given the role of
421 mitochondria in oxygen consumption and production of ATP molecules. Upon comparing the
422 two domestic Old World camel species (dromedary and Bactrian), significant signals of
423 purifying selection were defined in six mtDNA protein-coding genes: *ND1, COX1, ND4, ND5,*
424 *ND6*, and *CYTB*, among which *ND1*, *ND5* and *ND6* have previously been identified to be under
425 purifying selection (Mohandesan et al., 2017). Here, purifying selection would act to eliminate
426 deleterious alleles that negatively affect the functionality of mitochondria.

427

428 The MK, NI and ω ratio analyses are considered conservative approaches to detect signatures
429 of selection given that mutations affecting specific codons in a gene that can be masked by
430 gene-wide purifying selection pressures (Crandall et al., 1999). The site-model BEB analysis
431 identified a total of 21 and 24 codons to be under positive selection, potentially associated with
432 adaptive divergence between dromedaries or Old World camels and New World camel species,
433 respectively. Seven codons under positive selection were identified when comparing
434 dromedary with domestic and wild Bactrian, indicative of adaptive divergence between these
435 two Old World camel species, potentially related to the different environmental niches they
436 occupy. Interestingly, seven codons identified by the branch-site model BEB analysis were
437 found to be under positive selection in dromedaries, indicating that these codons might be
438 associated with adaptation to desert environment in dromedaries. Among these sites, codon 1
439 in *ATP6* was found to be under positive selection in dromedaries using the branch-site models
440 by Mohandesan et al. (2017). Although all of these codons were defined to be under positive

441 selection based on BEB posterior probability, none of the genes fit the signature of selection
442 alternative model (Model A) better than the neutral null model, which was also observed by
443 Mohandesan et al. (2017).
444
445 Several codon substitutions were found to be under positive selection by the TreeSAAP
446 software between dromedary and Lamini, and between dromedary and Bactrain. This approach
447 is considered more sensitive than ω ratio-based approaches since a single biochemically
448 adaptive physiochemical change can be too weak to be identified by gene-based approaches
449 where conservation scores are high (Hughes, 2007, McClellan, 2013). We observed magnitude
450 impacts on six physiochemical properties of proteins: the equilibrium constant (ionization
451 COOH), power to be at the middle of alpha helix, long range non-bonded energy, solvent
452 accessible reduction ratio, surrounding hydrophobicity, and alpha helical tendencies. Changes
453 in these properties can impact protein functions in several ways. An increase in alpha helical
454 tendency (P_α) may lead to long and rigid alpha helix, which renders interactions with amino
455 acid motifs more difficult. Decreasing this property makes the amino acid more flexible to an
456 open alpha helix, increasing the likelihood of amino acid interactions within a protein (Burkin
457 et al., 2000). Changes of the equilibrium constant (ionization COOH) (pK′) and solvent
458 accessible reduction ratio (R_a) impacts amino acid water solubility, rending it more hydrophilic
459 or hydrophobic, and is considered an important change when dehydrated, increasing longevity
460 by reducing reactive oxygen species (ROS) production (Beckstead et al., 2009). Increasing the
461 surrounding hydrophobicity (H_p) makes the surrounding area of the amino acid site
462 hydrophobic (Mohandesan et al., 2017). While changes in long-range non-bonded energy (E_l)
463 impacts amino acid interactions and may contribute of protein structure change.
464

The signals of selections characterized here were investigated based on gene-wide and site-specific analyses. These analyses have taken the advantage of including large number of samples in this study from the different species of the Camelidae family. These analyses can be further improved upon including camel samples from more diverse geographical locations with specific phenotypic traits, such as racing performance and milk production. Such phenotypes may explain the polymorphic codons within the dromedary lineage, defined by the branch-site models and TreeSAAP, in *ND6* and *ND4* genes.

Conclusion

In this study signatures of selection were investigated in the mitochondrial protein-coding genes of different dromedary camel populations by comparing to Bactrian and Lamini camels, revealing signals of adaptive divergence between Old World and New World camel species. Gene-level purifying selection was identified among the Old World camels, and signals of positive selection were identified at specific codons in Old World camels. These codons provide a link between the mtDNA evolution in such camel species and their adaptation to the diverse environmental conditions they populate.

Funding sources

This work was supported by the college of graduate studies at Kuwait university through MSc. grants of Mrs. Sanaa Al-Zoubi and Ms. Arwa Afana.

Data Availability

We have deposited the whole genome sequence data underlying these analyses at the European Nucleotide Archive (Bioproject accession number: PRJEB53955).

Acknowledgment

The authors would like to thank the camel owners who supplied camel blood samples from the Arabian Peninsula.

Conflict of interest

The authors declare not competing of interest.

References

ALAQEELY, R., ALHAJERI, B. H., ALMATHEN, F. & ALHADDAD, H. 2021. Mitochondrial Sequence Variation, Haplotype Diversity, and Relationships Among Dromedary Camel-Types. *Frontiers in Genetics,* 12.

ALMATHEN, F., CHARRUAU, P., MOHANDESAN, E., MWACHARO, J. M., OROZCO-TERWENGEL, P., PITT, D., ABDUSSAMAD, A. M., UERPMANN, M., UERPMANN, H. P., DE CUPERE, B., MAGEE, P., ALNAQEEB, M. A., SALIM, B., RAZIQ, A., DESSIE, T., ABDELHADI, O. M., BANABAZI, M. H., AL-EKNAH, M., WALZER, C., FAYE, B., HOFREITER, M., PETERS, J., HANOTTE, O. & BURGER, P. A. 2016. Ancient and modern DNA reveal dynamics of domestication and cross-continental dispersal of the dromedary. *Proc Natl Acad Sci U S A,* 113**,** 6707-12.

ALMATHEN, F., ELBIR, H., BAHBAHANI, H., MWACHARO, J. & HANOTTE, O. 2018. Polymorphisms in MC1R and ASIP Genes are Associated with Coat Color Variation in the Arabian Camel. *J Hered,* 109**,** 700-706.

BAHBAHANI, H. & ALMATHEN, F. 2022. Homogeneity of Arabian Peninsula dromedary camel populations with signals of geographic distinction based on whole genome sequence data. *Scientific Reports,* 12**,** 130.

BEALL, C. M. 2007. Two routes to functional adaptation: Tibetan and Andean high-altitude natives. *Proceedings of the National Academy of Sciences,* 104**,** 8655-8660.

BECKSTEAD, W. A., EBBERT, M. T., ROWE, M. J. & MCCLELLAN, D. A. 2009. Evolutionary pressure on mitochondrial cytochrome b is consistent with a role of CytbI7T affecting longevity during caloric restriction. *PLoS One,* 4**,** e5836.

BLIER, P. U., DUFRESNE, F. & BURTON, R. S. 2001. Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. *Trends Genet,* 17**,** 400-6.

BURGER, P. A., CIANI, E. & FAYE, B. 2019. Old World camels in a modern world - a balancing act between conservation and genetic improvement. *Animal genetics,* 50**,** 598-612.

BURKIN, D. J., KIM, J. E., GU, M. & KAUFMAN, S. J. 2000. Laminin and alpha7beta1 integrin regulate agrin-induced clustering of acetylcholine receptors. *J Cell Sci,* 113 (Pt 16)**,** 2877-86.

CHEN, S., ZHOU, Y., CHEN, Y. & GU, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics,* 34**,** i884-i890.

CRANDALL, K. A., KELSEY, C. R., IMAMICHI, H., LANE, H. C. & SALZMAN, N. P. 1999. Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol,* 16**,** 372-82.

FAN, R., GU, Z., GUANG, X., MARÍN, J. C., VARAS, V., GONZÁLEZ, B. A., WHEELER, J. C., HU, Y., LI, E., SUN, X., YANG, X., ZHANG, C., GAO, W., HE, J., MUNCH, K., CORBETT-DETIG, R., BARBATO, M., PAN, S., ZHAN, X., BRUFORD, M. W. & DONG, C. 2020. Genomic analysis of the domestication and post-Spanish conquest evolution of the llama and alpaca. *Genome Biology,* 21**,** 159.

HUGHES, A. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99**,** 364–373.

KUMAR, S., STECHER, G., LI, M., KNYAZ, C. & TAMURA, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol,* 35**,** 1547-1549.

LI, H. & DURBIN, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26**,** 589-95.

LIBRADO, P. & ROZAS, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics,* 25**,** 1451-2.

MCCLELLAN, D. A. 2013. Directional Darwinian Selection in proteins. *BMC Bioinformatics,* 14 Suppl 13**,** S6.

MCDONALD, J. H. & KREITMAN, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature,* 351**,** 652-654.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.

MING, L., YUAN, L., YI, L., DING, G., HASI, S., CHEN, G., JAMBL, T., HEDAYAT-EVRIGHT, N., BATMUNKH, M., BADMAEVNA, G. K., GAN-ERDENE, T., TS, B., ZHANG, W., ZULIPIKAER, A., HOSBLIG, ERDEMT, NATYROV, A., MAMAY, P., NARENBATU, MENG, G., NARANGEREL, C., KHONGORZUL, O., HE, J., HAI, L., LIN, W., SIRENDALAI, SARENTUYA, AIYISI, LI, Y., WANG, Z. & JIRIMUTU 2020. Whole-genome sequencing of 128 camels across Asia reveals origin and migration of domestic Bactrian camels. *Communications Biology,* 3**,** 1.

MOHANDESAN, E., FITAK, R. R., CORANDER, J., YADAMSUREN, A., CHULUUNBAT, B., ABDELHADI, O., RAZIQ, A., NAGY, P., STALDER, G., WALZER, C., FAYE, B. & BURGER, P. A. 2017. Mitogenome Sequencing in the Genus Camelus Reveals Evidence for Purifying Selection and Long-term Divergence between Wild and Domestic Bactrian Camels. *Scientific Reports,* 7**,** 9970.

R Core Team. 2022. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

TIBARY, A. & EL ALLALI, K. 2020. Dromedary camel: A model of heat resistant livestock animal. *Theriogenology,* 154**,** 203-211.

WILSON, R. T. 1998. *Camels The tropical agriculturalist*, Macmillan Education.

WOOLLEY, S., JOHNSON, J., SMITH, M. J., CRANDALL, K. A. & MCCLELLAN, D. A. 2003. TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. *Bioinformatics,* 19**,** 671-672.

WU, H., GUANG, X., AL-FAGEEH, M. B., CAO, J., PAN, S., ZHOU, H., ZHANG, L., ABUTARBOUSH, M. H., XING, Y., XIE, Z., ALSHANQEETI, A. S., ZHANG, Y., YAO, Q., AL-SHOMRANI, B. M., ZHANG, D., LI, J., MANEE, M. M., YANG, Z., YANG, L., LIU, Y., ZHANG, J., ALTAMMAMI, M. A., WANG, S., YU, L., ZHANG, W., LIU, S., BA, L., LIU, C., YANG, X., MENG, F., WANG, S., LI, L., LI, E., LI, X., WU, K., ZHANG, S., WANG, J., YIN, Y., YANG, H., AL-SWAILEM, A. M. & WANG, J. 2014. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun,* 5**,** 5188.

YANG, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci,* 13**,** 555-6.

YANG, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

YANG, Z., NIELSEN, R., GOLDMAN, N. & PEDERSEN, A. M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 155**,** 431-49.

YANG, Z., WONG, W. S. & NIELSEN, R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol,* 22**,** 1107-18.

YU, G. 2020. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics,* 69**,** e96.

597 ZHANG, J., NIELSEN, R. & YANG, Z. 2005. Evaluation of an improved branch-site
598 likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol,*
599 22**,** 2472-9.
600
601
602
603
604
605
606
607
608

Supporting information

Fig. S1. The ω ratios for the 13 mtDNA protein-coding genes at three different analyses levels. 1) Dromedary camels (n=42) compared to Bactrian camels (n=67), 2) Dromedary camels (n=42) compared to Lamini (n=31), 3) Dromedary and Bactrian camels (n=109) compared to Lamini (n=31).

File S1. Detailed script for mapping raw reads, calling variants and retrieving mtDNA sequences

Table S1. The seven dromedary samples with whole genome sequences from this study.

Table S2. Whole genome sequence data of the Camelidae family species included in the study from European Nucleotide Archive.

Table S3. Full mtDNA sequences of Camelidae family species included in the study from NCBI Genebank.

Table S4. McDonald-Kreitman (MK) test and Neutrality index (NI) values on the three different analyses levels for the mtDNA protein-coding genes. The values with asterisk are significant with Fisher exact P-value *<0.05, **<0.01, ***<0.001.

Table S5. The ω ratio (Dn/Ds) analyses for the 13 mtDNA protein-coding genes at three different analyses levels. A) Dromedary camels (n=42) compared to Bactrian camels (n=67), B) Dromedary camels (n=42) compared to Lamini (n=31), C) Dromedary and Bactrian camels (n=109) compared to Lamini (n=31).

Table S6. Likelihood values and parameter estimates of the different site models implemented in CODEML package for the mtDNA protein-coding genes in all three data analysis levels. n.s = non-significant.

Table S7. Potential positively selected sites identified using the site-models based on Bayes Empirical Bayes (BEB) in genes at the three different data analyses levels. P-values are the significant values of the M8-M7 Likelihood ratio test. n.s = non-significant.

Table S8. Likelihood values and parameter estimates of the Branch-site models implemented in CODEML package for the mtDNA protein-coding genes upon specifying dromedary camels as foreground.

Table S9. TreeSAAP positive selection sites based on the impact on amino acids physiochemical properties changes. Pairwise comparisons between dromedary and Bactrian, dromedary and wild Bactrian and dromedary and Lamini camels. The equilibrium constant (ionization COOH) (pK′), power to be at the middle of alpha helix (αm), long range non-bonded energy (E1), solvent accessible reduction ratio (R_a), surrounding hydrophobicity (Hp), and alpha helical tendencies (Pα).