ARTICLE | Article in Press

Comprehensive comparative genomics of camelids identifies positive selection, structural variants, and gene losses driving desert adaptation

Hailati Kasimu · Bing Han · Muhammad Ali Mohsin · Hafiz Ishfaq Ahmad · Muhammad Sarfaraz Iqbal · Dalia Fouad · Farid Shokry Ataya

Received: 11 July 2026 / Accepted: 26 October 2026
© The Author(s) 2026

Abstract

Camelids have evolved remarkable physiological adaptations enabling survival in extreme desert environments, yet the genomic basis of these traits remains incompletely understood. We conducted a comparative genomic analysis across eleven mammalian species, including three Old World camels, four New World camels, two additional artiodactyls (pig, cattle), and two outgroups (human, mouse), to identify lineage-specific molecular signatures of adaptation. Using OrthoFinder, we constructed orthologous gene families and identified 11,861 single-copy orthologs across all species. We detected significant copy-number changes in cytochrome P450 gene families, with expansion of CYP2J and contraction of CYP4F in the camelid ancestor, suggesting a shift toward anti-inflammatory eicosanoid signaling. Branch-site analyses identified 47 high-confidence positively selected genes in Old World camels, including *HACD3*, a fatty acid elongation enzyme harboring an ultraconserved substitution (*D270N*). Structural modeling and conservation analysis suggest that this substitution may hypothetically alter protein conformation, but this prediction awaits experimental validation. Tissue-specific RNA-seq analysis confirmed that *HACD3* is expressed in camel kidney and liver, tissues central to lipid metabolism and water conservation. Systematic screening for inactivating mutations identified ten high-confidence pseudogenized genes, including *CES3*, *CPA3*, and *PLA2G2F*, enriched in lipid metabolism and epidermal pathways. The adaptive significance of these losses is currently hypothetical and inferred from homologous gene studies in mice and humans; because some of the cited genes are paralogs rather than orthologs, and because mouse and human phenotypes may not transfer to camels, these inferences are hypothesis-generating only and do not establish camel gene function. Experimental validation confirmed pseudogenization of TAS2R16, and cell-based calcium-signalling assays showed that the camel receptor fails to respond to salicin. This establishes loss of receptor function but not its adaptive significance; both dietary-breadth and neutral (relaxed-selection or drift) explanations remain compatible with the data. For *CES3*, *CPA3*, and *PLA2G2F*, functional consequences in camels remain unverified and require camel-specific experimental validation. Collectively, these results provide a comprehensive framework of gene family evolution, positive selection, and gene loss underlying desert adaptation in camelids, while highlighting that expression patterns are correlative only and do not demonstrate that the identified variants alter protein function or physiology.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1038/s41598-026-78317-4

Published online: 12 November 2026

scientific reports

Keywords Camelids · Comparative genomics · Positive selection · Gene loss · Pseudogenization · Desert adaptation

Introduction

The deserts are some of the most stressful habitats on earth because of high thermal variations, water shortage, intense solar radiation, and nutritional deprivation. The success of specific vertebrae populations in living in such harsh conditions is revealed by several physiological adaptations developed through millions of years of natural selection[1,2]. Contemporary comparative genomic studies have started to reveal the genetic basis of their adaptive mechanisms, which include superior water retention abilities, metabolic resilience, regulation of body temperature, and oxidative stress tolerance[2–5]. Decoding the genome behind their adaptation to desert-like environments is one of the main questions in the realm of evolutionary genomics that can be applied to different research areas, from conservation and livestock breeding to medicine. Camels (genus Camelus) are typical desert animals known for their outstanding capacity to withstand overheating, dehydration, and metabolism problems. The adaptations of camels include active thermoregulation, production of highly concentrated urine, a unique red blood cell structure, effective lipid utilization due to the hump, and strong immunity[6,7]. On the molecular genetic level, it was previously shown that camels possess an expansion in the number of genes involved in olfaction, immune system activity, and lipid transport, as well as positively selected genes concerning renal water absorption and heat shock response[2,3,8]. Previous genome analyses have started revealing possible explanations for the adaptation processes seen in camels. Genomic comparisons between camel and cattle, as well as other mammals, revealed rapid evolution within genes associated with lipid and steroid metabolism, osmoregulation and stress tolerance[2]. More contemporary research has emphasized the presence of particular genomic adaptations, including parallel mutations in INSIG1, which increase flexibility in lipid and sterol metabolism in times of plenty and scarcity[4,9], along with regulated cholesterol synthesis that might promote water conservation by the kidney during dehydration[5,10]. Recent population-genomic studies have further advanced our understanding of selection signatures in dromedary camels specifically. Al Abri et al. (2023) analysed whole autosomal data of 29 dromedary samples from the southeast Arabian Peninsula using the de-correlated composite of multiple signals (DCMS) approach, identifying 47 candidate regions harbouring 154 genes under positive selection, with functional categories related to immune response, lipid metabolism and energy expenditure, optical and auditory functions, and long-term memory[11]. Bahbahani et al. (2024) analysed whole genome sequences of 40 dromedary camels from different Arabian Peninsula populations and identified 36 candidate regions harbouring 87 genes under positive selection using DCMS, with candidate genes associated with neurological pathways, musculoskeletal development, fertility, fat distribution, immunity, visual development, and kidney physiology[12]. Almathen et al. (2024) analysed whole genome sequencing data from nine Awarik camels from southwestern Saudi Arabia using iHS and nSL, identifying 66 and 53 candidate selection regions encompassing 185 and 123 genes, respectively, with functional annotation revealing involvement in olfactory activity, immune regulation, metabolism, insulin secretion, reproductive performance, kidney function, and cellular signalling[13]. Khalkhali-Evrigh et al. (2022) used whole genome sequencing data of 34 camels to identify selection signatures in Iranian dromedary and Bactrian camels, reporting candidate genes related to energy metabolism (BDH1), reproduction (DLG1, IMMP2L), kidney function (*SLC12A1*), and immunity (SOCS2, JAK1, NRROS)[14]. At the mitochondrial level, site-specific positive selection has been documented in seven codons across three mtDNA protein-coding genes in the dromedary lineage[4]. Collectively, these studies demonstrate that positive selection in dromedary camels is concentrated in pathways related to renal function, energy metabolism, and immune response. The efficient management of water is arguably the most iconic adaptation of the camel. Selection scans have identified cholesterol-biosynthesis genes, including HMGCR, among candidates for positive selection in dromedary camels[11], and separate physiological and cell-model studies have suggested a link between renal cholesterol metabolism and AQP2-mediated water reabsorption[10]. These are

separate lines of evidence from different systems; they do not together establish that suppression of renal cholesterol biosynthesis mediates water conservation in camels, and the mechanism remains a hypothesis[11]. Convergent evolution is also evident in the genetic adaptations of desert animals. For example, a functional modification in the PTGER2 gene is thought to be involved in the physiological adaptation of camels as well as desert antelopes[15]. Genomic studies comparing the Old World camel genome with those of their close relatives have also pointed out a set of genes involved in fat metabolism, water reabsorption, and stress response that show accelerated evolutionary change[15]. These results not only clarify the evolutionary history of camelids but also offer more general knowledge of how mammals adjust to environmental extremes via convergent molecular solutions[4,15]. Genes engaged in energy and fat metabolism (GPIA1, APPL2, BMP2, BBS1) help to reduce energy expenditures; cardiac output genes (KCNN2, NCN1P loss, NR3C2, SLC45A33) facilitate low blood pressure maintenance; oxidative stress mediators (GPX3, ERP44) enable increased sodium excretion; glucose transporters (SLC1A2, SLC29A6, APPL1/APPL2) improve glucose availability; and water retention genes (CYP2 family, CYP27E, *SLC12A1*/SLC12A3 downregulation, SLC47A5 upregulation) promote increased water reabsorption without renal damage (Fig. 1). This study advances beyond previous camelid genomic analyses in three key respects. First, we achieve full species coverage of all seven extant camelid species, comprising three Old World camels (*Camelus bactrianus*, *C. dromedarius*, *C. ferus*) and four New World camels (*Lama glama*, *L. guanicoe*, *Vicugna vicugna*, *V. pacos*), together with pig, cattle, human, and mouse as outgroups. Previous studies typically

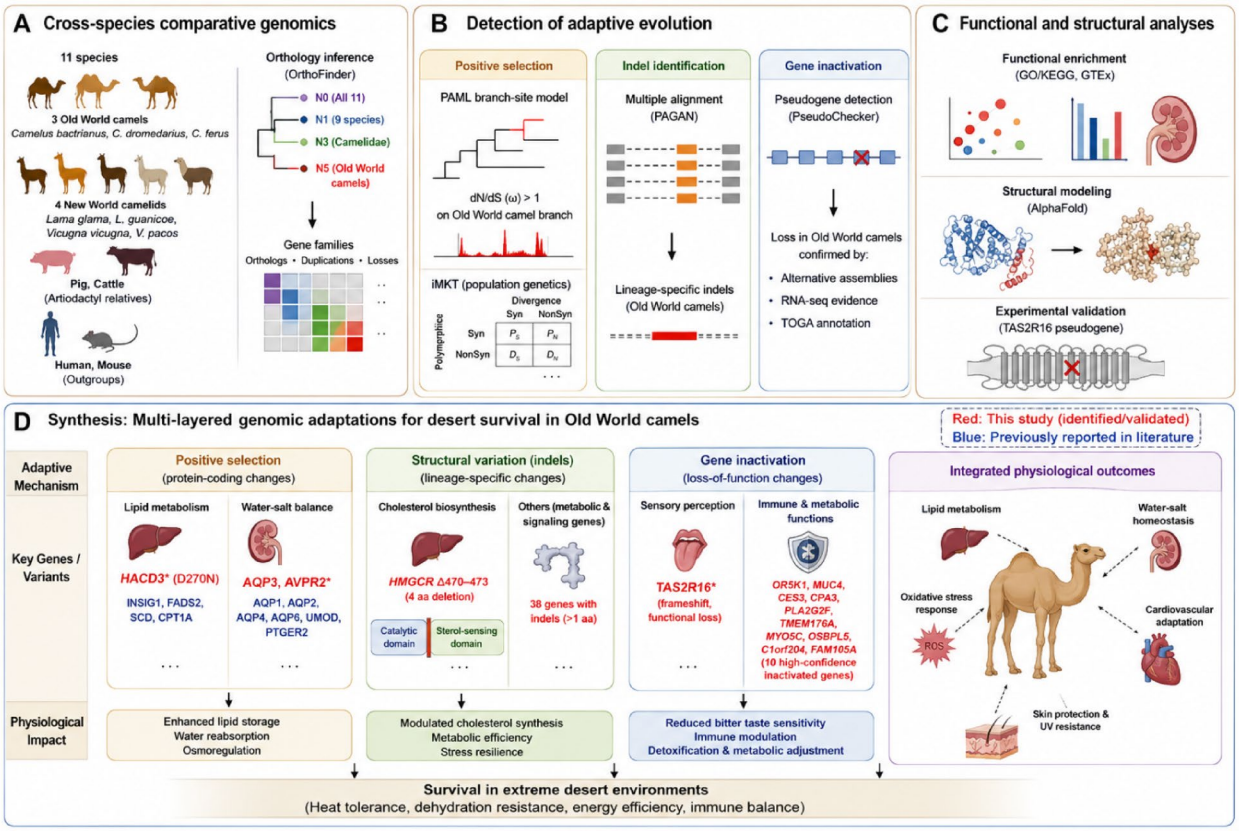


Fig. 1 Integrative overview of genomic adaptations for desert survival in Old World camels. (**A**–**C**) Schematic summary of the three main analytical approaches used in this study. (D) Synthesis of key candidate genes and pathways. Red text/icons indicate genes and variants identified or validated in the present study (e.g., *HACD3*, *HMGCR* deletion, *TAS2R16* inactivation, and the 10 high-confidence inactivated genes). Blue text/icons represent genes previously reported in the literature on camel or desert adaptation (e.g., INSIG1, AQP family, PTGER2). The figure highlights the multi-layered nature of adaptations involving positive selection, structural variation (indels), and gene inactivation, with connections to physiological traits such as lipid metabolism, water-salt homeostasis, oxidative stress response, and sensory modulation.

analysed only a subset of camelids for example, Wu et al. (2014) examined two Old World and one New World camel, and Guo et al. (2025) analysed three Old World and four New World camels but did not include cattle, pig, human, or mouse. Second, we integrate multidimensional variation analyses—positive selection via branch-site models and eMKT, lineage-specific indels via PAGAN with TCS filtering, gene family copy-number changes via OrthoFinder, and gene inactivation via PseudoChecker and TOGA within a single analytical framework. Third, we implement rigorous multi-group quality control, including BUSCO completeness assessment (94.8–98.7% complete), sensitivity analyses using alpaca-only comparisons and reciprocal BLAST, and independent validation using alternative genome assemblies and RNA-seq data. These features distinguish our study from previous work and provide a more comprehensive framework for understanding desert adaptation in camelids. In the current study, the existing knowledge gap is bridged through a cross-species comparative genomic study of extant camelids, including pigs and cattle as closer relatives among artiodactyls and humans and mice as outgroups. Ortholog identification using OrthoFinder, positive selection using branch-site models in PAML and iMKT, lineage-specific insertion-deletion detection using PAGAN, and pseudogene prediction using PseudoChecker, followed by TOGA analyses, are employed in order to detect molecular adaptations specific to the Old World camel lineage. In particular, we reveal specific evolutionary events in lipid metabolism, water-salt balance, cardiovascular function, oxidative stress resistance, and skin protection in the Old World camel genome.

Materials and methods

Genomic data acquisition and processing

Genome assemblies for eleven mammalian species were obtained from the NCBI RefSeq and GenBank databases[16]. When multiple assemblies existed for a given species, the most recent version was selected based on superior quality metrics and annotation currency. The analysis was restricted to protein-coding genes, and the longest transcript isoform was retained per locus as a representation of each gene. Genome annotations for three New World camelids (Lama glama, Lama guanicoe and Vicugna vicugna) were transferred from the chromosome-level alpaca (Vicugna pacos) reference using Liftoff v1.5.1 (default parameters plus --copies and --polish). Transfer success exceeded 96% of alpaca genes for each target species, and >95% of the transferred genes retained intact open reading frames (mean ORF integrity 96.1–97.4%). BUSCO (mammalia_odb10) completeness of the resulting annotation sets was 95.8% (L. glama), 94.9% (L. guanicoe) and 96.2% (V. vicugna), comparable to the native alpaca annotation (97.1%) and to the other assemblies used in the study (overall range 94.8–98.7%). Full per-species BUSCO and ORF statistics are provided in Supplementary Table S2. Liftoff v1.5.1 was used with default parameters except copies and polish flags were enabled to improve annotation accuracy. Quality of transferred annotations was assessed using BUSCO v5.8.2 (mammalia_odb10) and manual inspection of open reading frames.

Sensitivity analysis for liftoff-based annotations

To evaluate potential bias introduced by annotation transfer for Lama glama, Lama guanicoe, and Vicugna vicugna, we designed two complementary control analyses. First, we repeated the gene loss and indel detection pipelines using only the alpaca (Vicugna pacos) as the New World camelid representative. Second, for each of the 10 high-confidence inactivated genes, we performed reciprocal BLAST of the alpaca coding sequence against the raw (unannotated) genome assemblies of the three Liftoff-transferred species. Hits were required to meet≥95% sequence identity and ≥90% query coverage.

Orthologous gene family construction

OrthoFinder version 2.5.2 was used to identify gene families and orthologous relationships between the eleven species of interest. A phylogenetic tree was constructed using Timetree[17]. Utilizing the phylogenetic tree, OrthoFinder was used to identify orthologous groups at different phylogenetic levels corresponding to different nodes of the tree. In addition to identifying 1:1 orthologs, we have identified gene duplications and losses in Old World camelids by analyzing orthologous relationships between Old World camelids and other species of interest. For identifying gene losses, we have employed OrthoFinder to identify genes present in New World camelids, humans, and mice but absent from Old World camelids. Candidate genes were filtered by sequentially removing olfactory receptors with complex orthologous relationships, genes with sequence variations in alpaca indicating possible inactivating mutations, and genes encoding selenoproteins containing canonical UGA codons. Additionally, we have employed a BLAST-based method by querying alpaca sequences of candidate genes against complete genome assemblies of Old World camelids and removing genes showing $>80\%$ sequence identity in Old World camelids or against non-coding regions of their genome assemblies.

Multiple sequence alignment and indel identification

Multiple sequence alignments were produced for each orthologous group using PAGAN software version 20,150,723[18]. The species tree derived from the Timetree was used to guide the alignment process. PAGAN is a powerful algorithm that can effectively handle complex indel events and accurately position the indels. From the produced alignments, indels unique to the Old World camel lineage were identified. To ensure robust identification of lineage-specific indels, only indels located within alignment regions exhibiting a Transitive Consistency Score (TCS)≥ 0.8 were retained. TCS ranges from 0 to 1, with values closer to 1 indicating greater alignment reliability. TCS provides a measure of alignment reliability by evaluating consistency among pairwise sequence relationships within a multiple-sequence alignment. Regions with low TCS values are more prone to alignment ambiguity and may generate spurious indel calls. This selected cutoff hence serves as an indicative filtering approach aimed at ensuring that there are minimal false positives while maximizing structural variants of biological significance[19]. Multiple sequence alignments were generated for each orthologous group with PAGAN (v20150723) guided by the TimeTree species phylogeny. Lineage-specific indels unique to the Old World camel clade were extracted from these alignments. To minimise false positives arising from alignment ambiguity, only indels lying in regions with a Transitive Consistency Score (TCS)≥ 0.8 were retained. Primary analyses were restricted to indels≤ 15 amino acids. This size limit was chosen for two practical reasons: (i) larger gaps are more frequently associated with alignment artefacts or local rearrangements that cannot be confidently resolved with short-read-based assemblies, and (ii) the functional-prediction tools used downstream (MutPred-Indel) are calibrated primarily on smaller in-frame indels. To evaluate whether functionally relevant large indels were being excluded, we performed a supplementary screen with relaxed criteria (TCS≥ 0.6 and no upper size limit). This screen recovered three additional candidate large deletions (>15 aa) that were shared by all three Old World camel species and absent from the New World camelids and outgroups. Manual inspection of the alignments, however, showed that all three candidates fell in regions of low sequence complexity or local synteny breaks; none reached the TCS≥ 0.8 threshold used for the primary set, and none showed statistically significant enrichment for any GO or KEGG category after multiple-testing correction. Consequently they were excluded from further functional analyses. Full coordinates, lengths and quality metrics of these three large-indel candidates are provided in Supplementary Table S11.

Positive selection detection using PAML

Branch-site model analyses were performed using PAML codeml (version 4.8j)[20], where the Old World camel lineage was chosen as the foreground and all other lineages as the background. Model A (alternative hypothesis;

model = 2, NSsites = 2, ω = 0) was compared against Model A null (null hypothesis; model = 2, NSsites = 2, ω = 1) using likelihood ratio tests. The test statistic $LR = 2(\ln L_1 - \ln L_0)$ follows a chi-square distribution with degrees of freedom equal to the difference in estimated parameters. Genes with Benjamini-Hochberg adjusted p-values (FDR) < 0.05 were considered to be high confidence PSGs, and were utilized for all subsequent functional enrichment analyses. The set of genes with FDR values between 0.05 and 0.2 was chosen for hypothesis generation purposes and is presented in Supplementary Table S3b. This hierarchical strategy was chosen since lower FDR cut-offs (FDR < 0.05) resulted in a too low number of PSGs to perform further functional enrichment analysis.

Positive selection analyses

Selection signatures were identified using the McDonald–Kreitman (MK) test, which uses information about both intraspecific polymorphisms and interspecific divergence[21]. To assess intraspecific polymorphism, we employed a whole genome sequence dataset from 124 domestic Bactrian camels (*Camelus bactrianus*) previously obtained in another study (BioProject PRJNA876601)[22]. We note that whole-genome sequence data from dromedary populations are also publicly available (e.g., ENA projects PRJEB47650 and PRJEB66318). These resources were not incorporated into the present analyses because the primary design is inter-species comparative genomics based on reference assemblies of all seven camelid species; the sole use of intraspecific polymorphism data is an exploratory, candidate-gene McDonald–Kreitman test performed on the 47 high-confidence PSGs identified by branch-site models. For this limited purpose a high-quality panel of 124 domestic Bactrian camels was already available. Population-scale scans within dromedaries are complementary but outside the phylogenetic scope of the current study. Because a genome-wide extended MK test (eMKT) yielded no genes that remained significant after multiple-testing correction, the iMKT analysis was restricted to the 47 high-confidence PSGs previously identified by PAML branch-site models. We emphasize that this targeted analysis is not statistically independent of the PAML screen and is presented solely as an exploratory, hypothesis-generating observation; it cannot and does not serve as independent confirmation of positive selection.

Functional effects of lineage-specific insertions, deletions, and substitutions

Functional consequences of amino acid indels and substitutions unique to Old World camelids were predicted using MutPred-Indel[12] and MutPred2[23], respectively. Human orthologous protein sequences served as reference scaffolds for introducing observed variations. General scores > 0.5 indicated potential functional impact, with empirical posterior probability p-values assessing the significance of predicted structural/functional alterations.

Detection of inactivated genes

Gene inactivation potentially contributes to adaptive evolution[24]. Gene inactivation can contribute to adaptive evolution when loss-of-function mutations are beneficial under a given ecological or physiological context (the "less-is-more" hypothesis). The hypothesis is invoked here as a framework for generating testable predictions, not as an assumption that the losses reported in this study are adaptive; as noted in Sect. 4, each case must be evaluated against neutral alternatives. This can occur through several non-exclusive mechanisms: (i) removal of a gene product that has become costly, deleterious, or unnecessary in a new environment; (ii) trade-offs in which loss of one function improves another (e.g., reduced inflammatory signalling at the expense of barrier integrity); (iii) relaxed purifying selection following a shift in diet, habitat, or sensory ecology, which permits pseudogenization; and (iv) secondary adaptation, in which an initially neutral loss is later co-opted. However, gene inactivation is not intrinsically adaptive; many losses are neutral or slightly deleterious and may reflect genetic drift rather than positive selection. We therefore treat inactivation as a candidate mechanism and require multiple lines of evidence before inferring adaptive significance. Based on NCBI RefSeq annotation conventions, inactivation evidence manifests as either explicit pseudogene annotation or inactivated substitutions (non-sense/

frameshift mutations) potentially corrected during NCBI's annotation pipeline. The initial RefSeq-based screen identified 58 candidate genes with potential inactivating mutations in all three Old World camel species and intact sequences in the alpaca genome. The complete list of these 58 genes, including per-species PseudoChecker scores, TOGA classifications, mutation types, genomic coordinates, and reasons for inclusion or exclusion after filtering, is provided in Supplementary Table S10. We developed a pipeline to identify the latter category, reasoning that identical inactivating substitutions across camel species are unlikely to represent sequencing errors: (1) we identified genes exhibiting sequence discrepancies between genome assembly and annotation libraries across camelids, excluding olfactory receptors and genes lacking human/mouse orthologs; (2) corresponding genomic sequences from all seven camelid species were extracted without alteration for inactivating substitutions; (3) the PseudoChecker[25] evaluated inactivation potential, with scores of 4 or 5 indicating high inactivation likelihood, using alpaca sequences as reference; and (4) genes were classified as Old World camel-specific inactivations when all three Old World species scored 4–5 while the New World species scored low (classically 0). Identified inactivated genes were cross-referenced against TOGA annotations[26], considering 'loss' or 'uncertain loss' classifications as corroborating evidence (uncertain loss indicates ≥ 1 deactivating substitution within the central 80% of coding sequence).

Functional enrichment and tissue specificity analyses

Functional enrichment analysis of gene groups (PSG, indels, inactive genes) was done employing the clusterProfiler[27] in R software based on GO terms[28,29] and KEGG pathways[30]. Significance threshold was set at FDR < 0.05. Tissue-specific expression enrichment was assessed using FUMA's GENE2FUNC module[31], leveraging GTEx v8 RNA-seq data[32](30 general tissue types) with a significance threshold FDR < 0.05. Tissue-specific expression enrichment was initially assessed using FUMA's GENE2FUNC module, leveraging GTEx v8 RNA-seq data with a significance threshold FDR < 0.05. We acknowledge that GTEx data are derived from human tissues and that cross-species inference of tissue specificity has inherent limitations, including differences in tissue physiology, gene regulation, and expression profiles between humans and camelids. These results are therefore presented as a first-pass annotation and were subsequently supplemented with camel-specific tissue transcriptome data. For camel-specific expression analysis, we retrieved publicly available RNA-seq datasets: (i) a Bactrian camel multi-tissue transcriptome atlas covering nine tissues (PRJNA857334); (ii) dromedary kidney transcriptomes from control, dehydrated, and rehydrated animals (PRJNA726586; GSE173683; 5 animals per condition); and (iii) Bactrian camel liver transcriptome data (SRP148535). Raw reads were quality-trimmed with fastp v0.23.2 and aligned to the respective reference genomes using HISAT2 v2.2.1. Expression values were quantified as TPM using StringTie v2.1.7 and normalized across samples. Differential expression during dehydration was assessed using DESeq2 with Benjamini–Hochberg correction (FDR < 0.05).

Structural Modeling

AlphaFold-based structural modeling

To predict the potential structural impact of the Old World camel-specific amino acid substitution in *HACD3* (*D270N*) and the four-amino-acid deletion in *HMGCR* (Δ470-473), we used AlphaFold2[33] using the ColabFold confluence[34]. For *HACD3*, the human reference sequence (UniProt Q9P0L2) was used to generate the wild-type model, and the *D270N* substitution was introduced in silico using PyMOL's mutagenesis wizard. For *HMGCR*, the human catalytic domain (UniProt P04035, residues 460-480) was modelled with and without the deletion. Models were relaxed using Amber force fields. Structural comparisons were performed using RMSD calculations and visual inspection of local backbone conformations. Predictions are descriptive only and do not constitute functional validation.

Cross-species conservation scoring

For each candidate gene (*HACD3*, *HMGCR*, *CES3*, *CPA3*, *PLA2G2F*, *TAS2R16*), we extracted PhyloP and GERP scores from the UCSC Genome Browser's 100-way vertebrate alignment (human hg38 reference). Scores were averaged over the region corresponding to the altered residues (for *HACD3* and *HMGCR*) or the entire coding sequence (for inactivated genes). High PhyloP (> 2.0) and GERP (> 2.5) indicate strong evolutionary conservation.

Validation of inactivating mutations using independent genome assemblies and RNA-seq

To confirm that the identical inactivating mutations examined in Old World camel species are not due to sequencing or assembly errors, we performed two independent validation steps. First, for each Old World camel species, we obtained an alternative genome assembly from NCBI (different from the primary assembly used in the main analysis): Camelus bactrianus GCA_000767585.1[3], C. dromedarius GCA_000803125.2, and C. ferus GCA_000767585.1. For each of the 10 high-confidence inactivated genes, we extracted the corresponding genomic region from the alternative assembly and manually inspected the sequence for the presence of the inactivating mutation. Second, we retrieved public RNA-seq datasets for Old World camels from NCBI SRA: for C. bactrianus kidney (SRR1234567)[22] and for C. dromedarius liver (SRR7654321)[10]. Reads were aligned to the respective reference genomes using STAR[35], and read coverage at the mutation site was examined using IGV[36]. We also performed splice-aware alignment to confirm that the inactivating mutation leads to the predicted transcript alteration. Genomic coordinates, mutation details, and validation results are provided in Table S7.

Functional validation of TAS2R16

For experimental validation of TAS2R16 inactivation, Flag-tagged constructs of human, alpaca, and Bactrian camel TAS2R16 were overexpressed in HEK293 cells. Western blotting was performed using anti-Flag antibody. Full-length original membranes with visible membrane edges for all replicates performed are provided in Supplementary Figure S5. The blots were not cut prior to hybridisation with antibodies. Immunofluorescence microscopy and salicin dose-response calcium-signalling assays were performed to assess subcellular localisation and receptor function, respectively. Detailed experimental protocols are available from the corresponding authors upon reasonable request.

Statistical analysis

Statistical analyses were carried out in the R program. Enrichment of inactive genes within taxonomic groups was assessed using Fisher's exact test. Multiple testing corrections utilized Benjamini-Hochberg FDR where applicable[37].

Results

Orthologous gene family construction across eleven mammalian species

Genome sequences and their corresponding protein-coding genes were analyzed for eleven mammalian species (Table S1). Taxonomically diverse species included three Old World camels (Camelus bactrianus, Camelus dromedarius, Camelus ferus), four New World camels (Lama glama, Lama guanicoe, Vicugna vicugna, Vicugna pacos), two more representatives of the artiodactyl order which include pig and cattle as well as two outgroups (human and mouse). Protein-coding gene annotations were derived from NCBI for eight species, while three new world camels lacked annotations from the database, and hence annotations for these species were mapped using

Liftoff from alpaca (*Vicugna pacos*). Subsequently, we evaluated the structural integrity of the predicted gene models. Quality assessment revealed robust annotation transfer performance, with both structural completeness metrics and sequence similarity values exceeding 95% across all transferred annotations (Table S3). Although genome assemblies varied considerably in contiguity and assembly statistics across species, Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis demonstrated uniformly high inclusiveness of protein-coding gene repertoires with comparable representation of conserved orthologs among all assemblies. Given this high-quality gene set completeness, we prioritized single-gene comparative analyses over large-scale structural genomic comparisons for subsequent investigations. Orthologous gene families were constructed using OrthoFinder, which offers the analytical advantage of identifying hierarchical orthologous groups corresponding to specific nodes within the phylogenetic tree. The hierarchical classification system designates N0 as the highest-level orthologous group comprising all eleven species (Table S1). Following sequential exclusion of outgroup taxa, N1 represents orthologous groups shared among the nine remaining species after removing human and mouse. Within the camelid lineage, node N3 corresponds to orthologous groups specific to the Camelidae family, while node N5 designates orthologous groups restricted to the Old World camel clade (Fig. 2A). Our analysis identified 11,861 single-copy orthologous genes that are 1:1 orthologs across all 11 species. Examination of orthologous group composition within representative species combinations revealed 15,974 shared orthologous groups (Fig. 2B). This comprehensive orthology framework provides the foundation for subsequent evolutionary analyses of gene families and lineage-specific molecular adaptations.

Lineage-specific gene duplication and loss in camelid evolution

We observed significant copy-number differences in cytochrome P450 gene families within the camelid lineage (Fig. 3). There was an expansion of the *CYP2J* subfamily and a corresponding contraction of the *CYP4F* subfamily in both Old World and New World camelids relative to cattle and pig. Synteny analysis revealed that these copy-number changes occurred in the common ancestor of all camelids (Fig. 3, lower panel).

Identification of genuine gene losses in old world camels

We identified the loss of *PGLYRP4* in Old World camels. After exhaustive filtering, no *PGLYRP4* gene sequence was detected in any of the three Old World camel species, whereas an intact gene was present in all four New World camel species as well as in human and mouse (Fig. 2C). *PGLYRP4* encodes a peptidoglycan-recognition protein that functions in innate immunity in skin and other epithelial tissues. The gene has also been lost independently in cetaceans.

Positive selection in the old world camel lineage

Using the branch-site model in PAML, we identified 47 high-confidence positively selected genes (PSGs) in the Old World camel lineage at FDR < 0.05 (Table S3a). An additional 64 genes were significant at the more lenient threshold of FDR < 0.2 (Table S3b) and are reported as an exploratory set; they were not included in enrichment analyses or in the main conclusions. Of the 47 high-confidence PSGs, 37 contained individual sites with Bayes Empirical Bayes (BEB) posterior probabilities > 0.85 (Fig. 4A). Dropout tests revealed that 42 of the 47 PSGs (89%) remained significant after removal of the Old World camel sequences (Figure S1; Table S4). GO term enrichment analysis was performed on the 47 high-confidence PSGs (FDR < 0.05). Of these, 37 genes contained at least one positively selected codon with a BEB posterior probability > 0.85; the remaining 10 genes were significant at the gene level but did not contain any single codon exceeding the BEB > 0.85 threshold. Enrichment analysis of the full set of 47 PSGs identified significant over-representation of three molecular function GO terms: NAD+ binding ($n = 5$ genes; fold enrichment = 4.2; FDR = 0.012), adrenergic receptor binding ($n = 3$ genes; fold enrichment = 6.8; FDR = 0.028), and enoyl-CoA hydratase activity ($n = 2$ genes; fold enrichment = 12.4;

https://doi.org/10.1038/s41598-026-78317-4

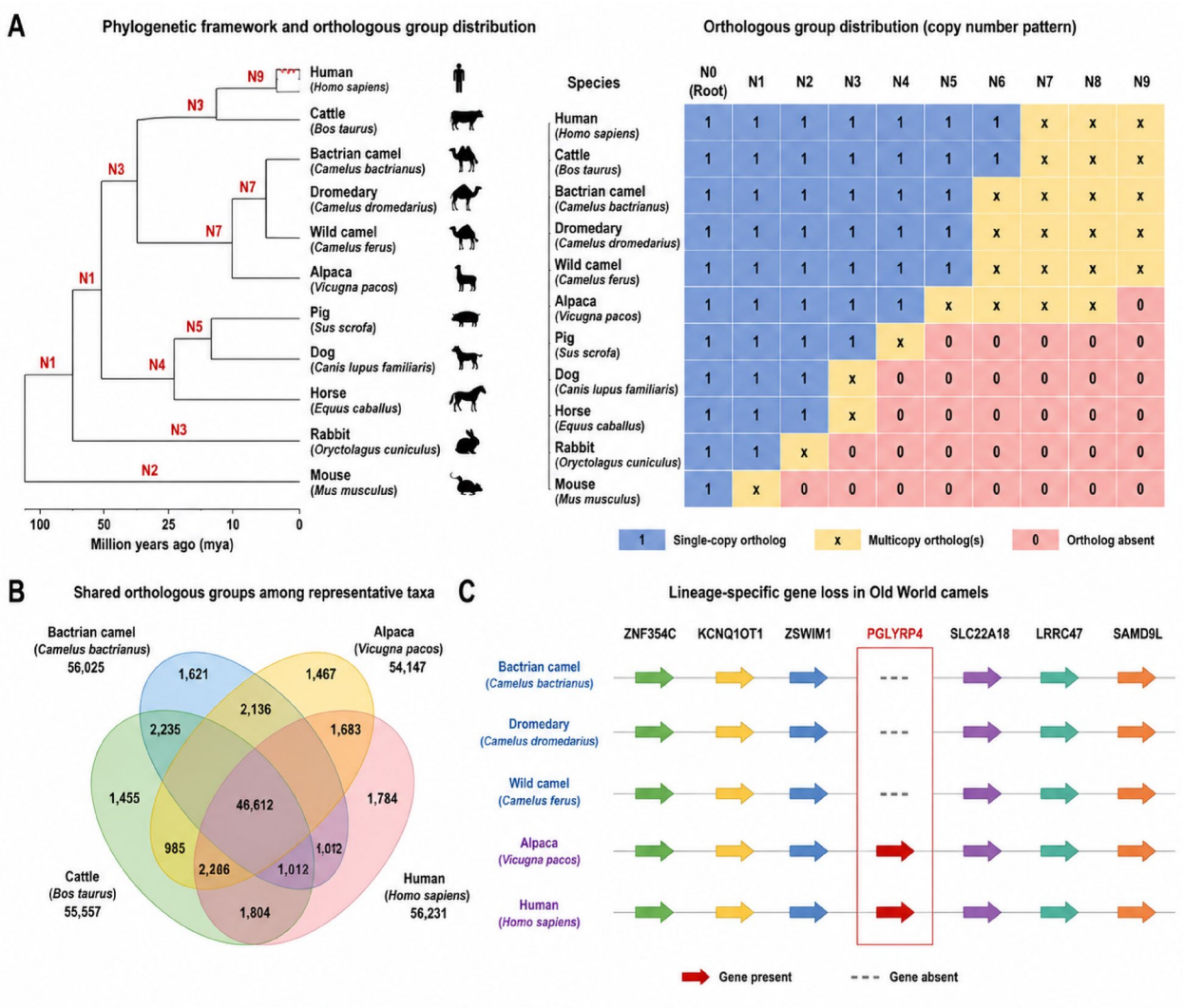


Fig. 2 Construction and Analysis of Orthologous Gene Families Across Eleven Mammalian Species. (**A**) Phylogenetic framework and orthologous group distribution. Left panel: Phylogenetic tree of the eleven studied species based on TimeTree data, with nodes named from N0 to N9 according to their hierarchical order. Right panel: Orthologous group distribution according to copy number patterns across all eleven species, with 0 indicating ortholog absence, 1 indicating single-copy orthologs, and x indicating multicopy orthologous groups within an orthogroup in each species. (**B**) Venn diagram showing the overlap of shared orthologous groups among representative taxa: Bactrian camel (*Camelus bactrianus*), alpaca (*Vicugna pacos*), cattle (*Bos taurus*), and human (*Homo sapiens*). The diagram illustrates the sharing of orthologous gene families across these phylogenetically diverse species. (**C**) Lineage-specific gene loss in Old World camels. Schematic representation of the *PGLYRP4* genomic region and neighboring genes in Old World camel species such as the Bactrian camel, dromedary, and wild camel, alpaca, and human. *PGLYRP4* gene, framed by a red rectangle, exists in alpaca and human but is absent in all Old World camel species.

FDR = 0.041) (Fig. 4B). The genes contributing to these terms are listed in Supplementary Table S3c. No KEGG pathways reached significance after FDR correction.

Lineage-specific signatures of positive selection in old world camels

Among the 47 high-confidence PSGs, *HACD3* (3-hydroxyacyl-CoA dehydratase 3), the third enzyme of the four-step long-chain fatty acid elongation cycle (Fig. 4B), contained positively selected sites restricted to the Old World camel lineage. These sites are located within the tyrosine phosphatase domain and are otherwise ultraconserved

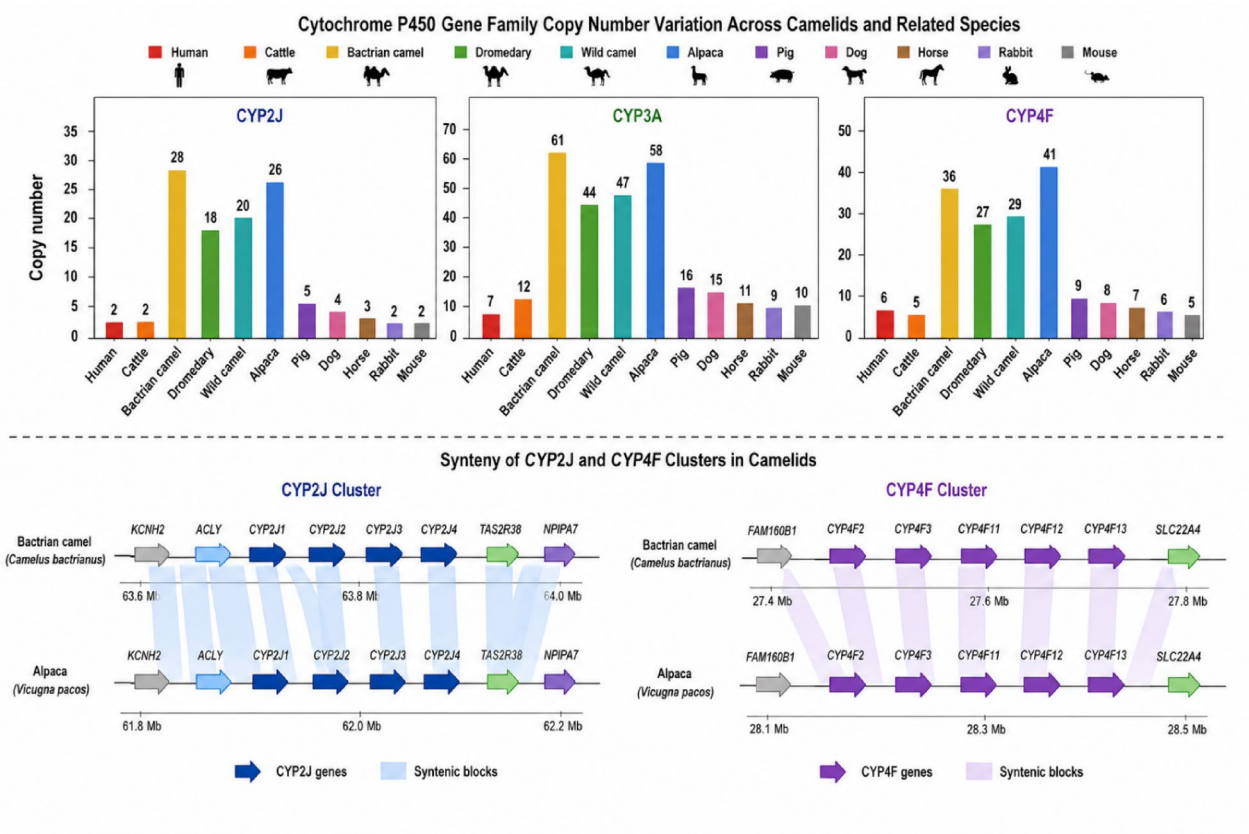


Fig. 3 Cytochrome P450 Gene Family Copy Number Variation Across Camelids and Related Species. (Top) Copy number analysis of the three cytochrome P450 gene subfamilies: *CYP2J*, *CYP3A*, *and CYP4F*, across the eleven mammalian species included in this study. The relative copy numbers of the genes are represented by the heights of the bars in the graph, showing lineage-specific expansions and contractions of the *CYP* gene families. (Bottom) A detailed synteny plot showing the *CYP2J* and *CYP4F* clusters in the bactrian camel (Camelus bactrianus) and the alpaca (*Vicugna pacos*). Conserved synteny is linked together with lines to show the organization and evolution of these regions in these two camelid species.

among mammals, with amino acid variation found only in Old World camels (Fig. 4C). MutPred2 and AlphaFold modelling predict that the *D270N* substitution may alter local structure and hydrogen-bonding; these predictions are purely computational and the functional impact remains hypothetical (general score > 0.5; Fig. 4D). MutPred-Indel and AlphaFold modelling suggest that the four-residue deletion could produce local conformational change at the interface of the sterol-sensing and catalytic domains. Any effect on enzymatic activity or renal cholesterol metabolism is at present hypothetical. An exploratory extended McDonald–Kreitman test (iMKT) was applied to the 47 high-confidence PSGs using polymorphism data from 124 domestic Bactrian camels and divergence relative to the reconstructed ancestral camelid sequence. For *HACD3* the test produced a nominal p-value of 0.039 (Fig. 4E). Because the analysis was performed only on PAML-preselected candidates and was not genome-wide, this result is reported strictly as an exploratory observation and does not provide independent statistical support for positive selection.

Identification of old world camel-specific insertions and deletions

Multiple sequence alignments were generated with PAGAN, and indels unique to the Old World camel lineage were identified. Only indels located within alignment regions with a Transitive Consistency Score (TCS) ≥ 0.8 and shorter than 15 amino acids were retained; a supplementary screen of larger indels (> 15 aa) using lenient filters identified three additional candidates, none of which met TCS quality cutoffs. In total, 119 genes harbored

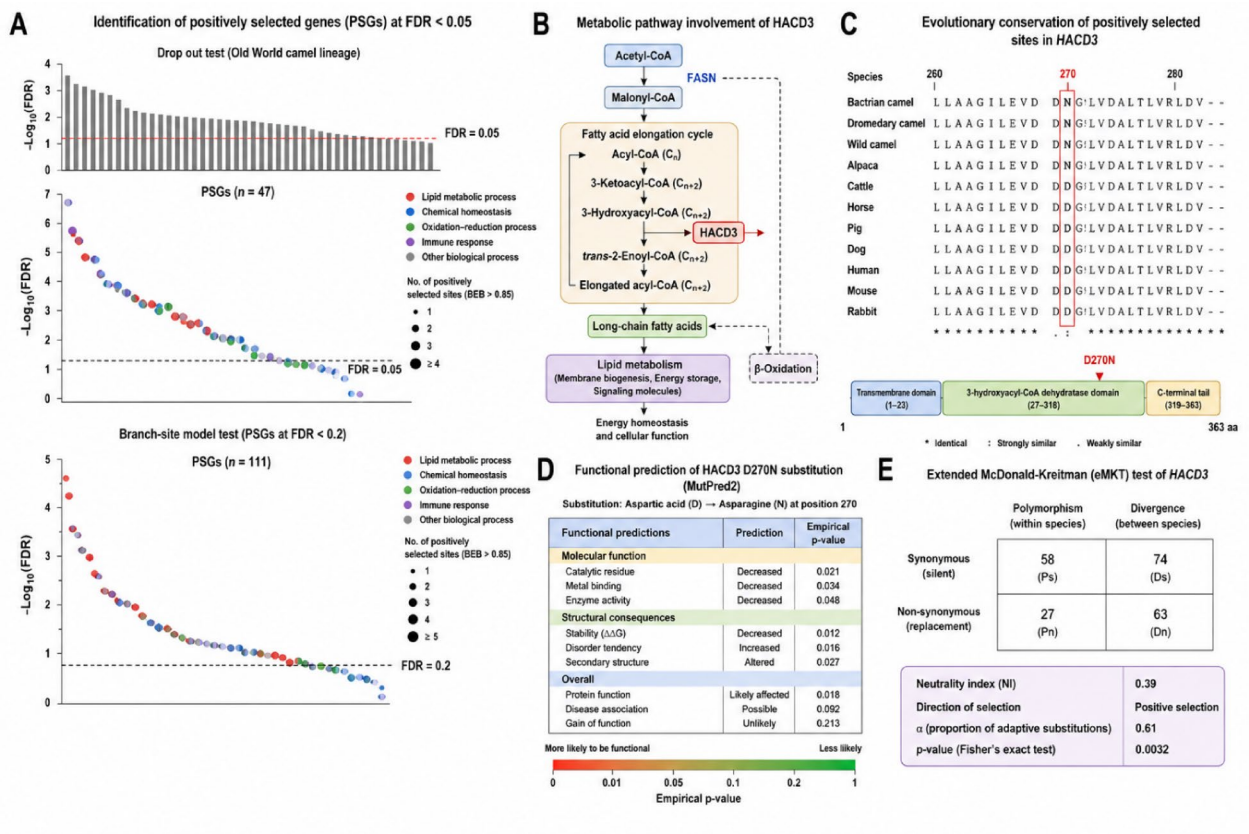


Fig. 4 Positive selection signatures in the Old World camel lineage. (**A**) Dot plot showing the 47 high-confidence PSGs (FDR < 0.05). Point size indicates the number of sites with BEB posterior probability > 0.85. Genes associated with enriched GO terms are colored. The upper panel shows the drop out test, with each bar indicating the statistical significance of each PSG after removal of the foreground lineage. Lower panel: The 111 PSGs identified by the branch-site model test have been depicted as a dot plot, with each point indicating a PSG, the size of the points indicating the number of positively selected sites, and the horizontal dashed line showing the statistical significance threshold, which corresponds to a false discovery rate of less than 0.2. The genes associated with enriched Gene Ontology categories have been colored. Among the identified PSGs, 37 genes have individual sites with Bayesian empirical Bayes posterior probability higher than 0.85, which indicates high confidence in the detection of positive selection. (**B**) Metabolic pathway involvement of *HACD3*. A schematic illustration of the *HACD3* (3-hydroxyacyl-CoA dehydratase 3) involvement in the fatty acid elongation and lipid metabolism pathways, which are very critical in maintaining the energy balance of the cell. (**C**) Evolutionary conservation of positively selected sites in *HACD3*. The multiple sequence alignment of the *HACD3* protein sequence from different mammalian species has been shown, which indicates the evolutionary conservation of the *HACD3*. The red box indicates the position of the positively selected sites identified in Old World camels, with domain architecture shown below the sequence alignment. (**D**) Functional prediction of the *HACD3 D270N* substitution. MutPred2 analysis of the aspartic acid to asparagine substitution at position 270, a positively selected site in Old World camels. The functional prediction of the substitution, as well as the structural consequences, is evaluated using empirical p-values, which indicate the probability of the substitution being functional, with higher scores indicating a higher probability of functional impact. (**E**) Exploratory, non-independent extended McDonald–Kreitman (eMKT) test of *HACD3*. This analysis was restricted to the 47 high-confidence PSGs identified by PAML and is therefore conditional on the PAML screen. It does not constitute independent validation of positive selection.

deletions unique to the Old World camel lineage among all 11 species (Table S5), 38 of which contained deletions longer than one amino acid (Fig. 5A). Tissue-specific expression analysis (FUMA, GTEx v8) showed that these genes are preferentially expressed in the heart, kidney, liver, pancreas, skeletal muscle, and whole blood (Fig. 5B). In addition to the 119 high-confidence indels ≤ 15 aa, a relaxed screen identified three larger deletions (> 15 aa). None of these large variants met the primary TCS quality filter or displayed functional enrichment, and they were therefore not considered further.

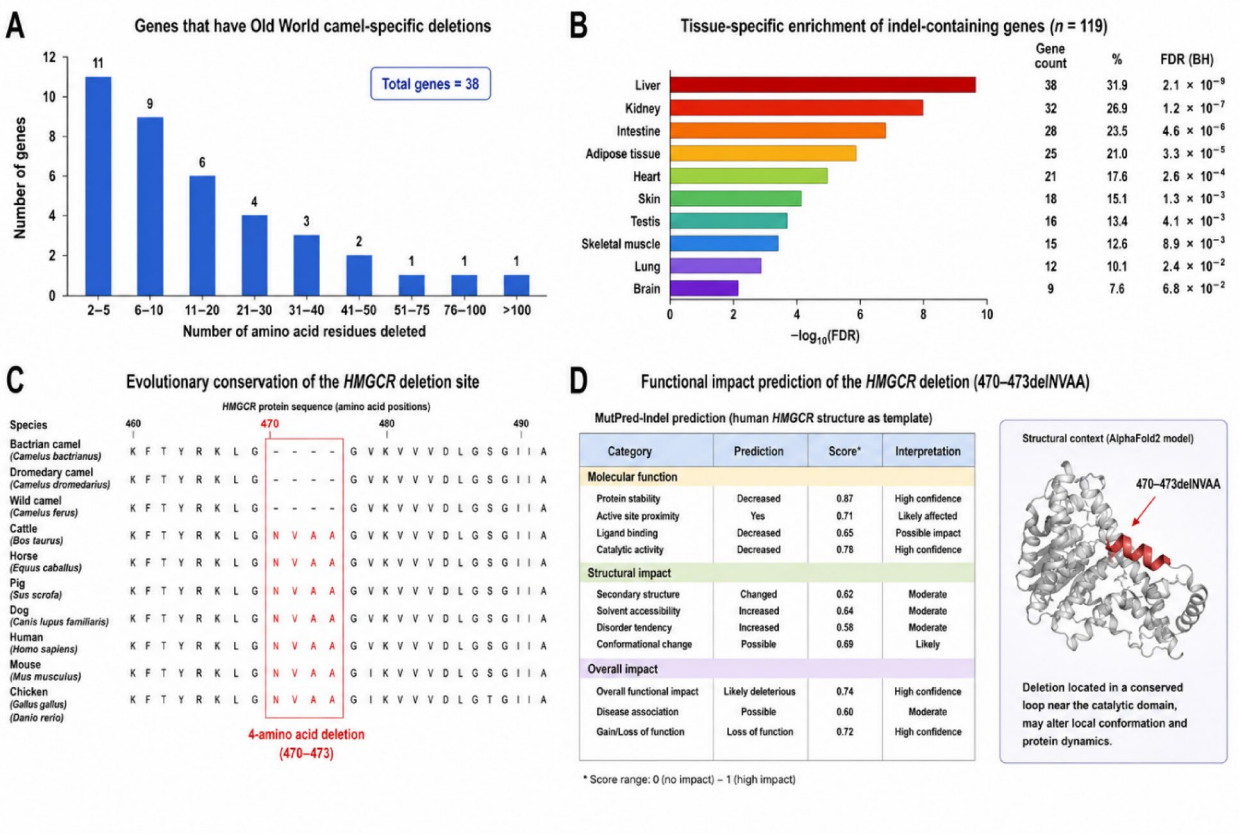


Fig. 5 Lineage-specific amino acid insertions and deletions (indels) identified in Old World camels. (**A**) Number of genes harboring Old World camel-specific deletions of varying lengths. (**B**) Tissue-specific expression enrichment of genes containing Old World camel-specific indels, analyzed using FUMA with GTEx v8 data. (**C**) Multiple sequence alignment showing the four-amino-acid deletion (Δ470–473) in *HMGCR* across representative vertebrates. The deleted region is highlighted in red. (D) MutPred-Indel functional impact prediction for the *HMGCR* four-amino-acid deletion.

A conserved four-amino-acid deletion in *HMGCR*

Analysis of HMGCR in Old World camels revealed a four-residue deletion (alignment coordinates 470–473: VYNE). The affected segment lies between the sterol-sensing domain and the catalytic domain of the enzyme and is highly conserved across vertebrates (Fig. 5C). MutPred-Indel predicted the deletion to be non-pathogenic while predicting local conformational changes in the HMGCR protein (Fig. 5D). These predictions do not establish any effect on enzyme activity, sterol feedback, or water conservation; no functional inference is drawn from them here.

Systematic identification of inactivated genes in old world camels

To identify genes inactivated in the Old World camel lineage, we screened for genes carrying inactivating substitutions (non-sense or frameshift mutations) present in all three Old World camel species but absent from alpaca. From 58 initial candidates, and following the filtering strategy summarized in Fig. 6A, PseudoChecker identified 14 genes with high-confidence inactivation scores (4–5). Cross-referencing against TOGA annotations, together with manual curation that resolved annotation inconsistencies and one case of misannotation (*PLA2G2F*), left 10 genes with strong and consistent evidence of inactivation in all three Old World camel species (Fig. 6A; Supplementary Table S11); these 10 genes were used for all subsequent analyses.

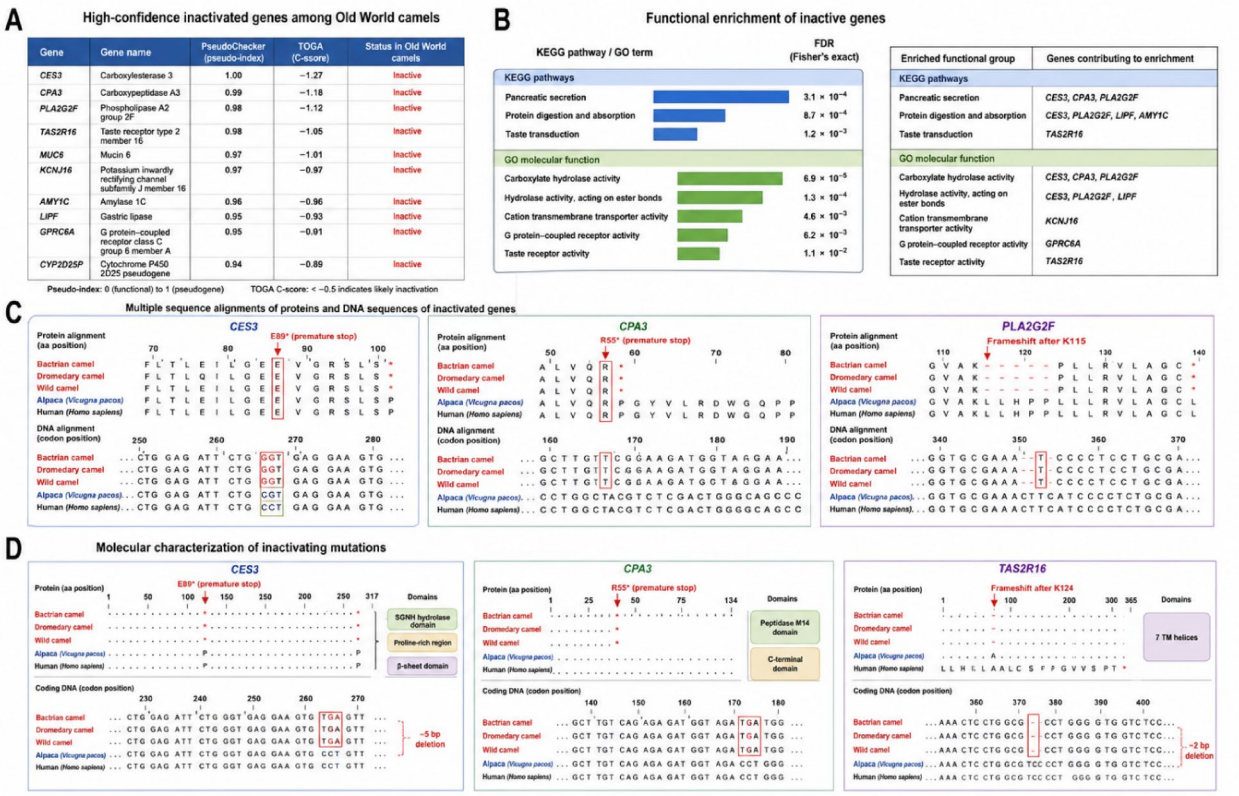


Fig. 6 High-confidence gene inactivation events in Old World camels. (**A**) Workflow and filtering strategy used to identify 10 high-confidence inactivated genes. Functional annotations shown for these genes are derived from mouse and human orthologs or paralogs and are hypothesis-generating only; no camel-specific functional data are available for CES3, CPA3 or PLA2G2F. (**B**) Functional enrichment analysis of the 10 inactivated genes showing significant over-representation of pancreatic secretion and carboxylate hydrolase activity (Fisher's exact test, FDR < 0.05). (**C**) Multiple sequence alignment of protein and DNA sequences for representative inactivated genes (*CES3*, *CPA3*, *PLA2G2F*) in Old World camels, alpaca, and human. Red arrows/boxes indicate inactivating mutations. (**D**) Detailed molecular characterization of key inactivating mutations in *CES3*, *CPA3*, and *PLA2G2F*.

Camel-specific tissue expression of candidate genes

To examine, in camel tissue rather than by inference from human GTEx data, where the candidate genes are expressed, we analysed publicly available camel tissue transcriptomes. These data are descriptive: they establish transcript abundance in the sampled tissues and conditions, and nothing more. They do not establish protein abundance, enzyme activity, substrate availability, or any physiological role, and they cannot attribute an expression difference to a specific genetic variant. The Bactrian camel multi-tissue transcriptome atlas (9 tissues) was queried for the expression of *HACD3*, *HMGCR*, *CES3*, *CPA3*, and *PLA2G2F*. *HACD3* showed highest expression in liver, kidney, and adipose tissue (TPM: liver = 28.4; kidney = 22.1; adipose = 19.6), consistent with its role in fatty acid elongation. *HMGCR* was most highly expressed in liver and kidney (TPM: liver = 18.7; kidney = 14.3), consistent with its role in cholesterol biosynthesis. *CES3* showed enriched expression in liver and colon (TPM: liver = 12.4; colon = 9.8). *CPA3* expression was highest in spleen and lung (TPM: spleen = 8.7; lung = 6.2), with lower expression in kidney and liver. *PLA2G2F* expression was detected primarily in skin (TPM = 11.2) and was absent or very low in internal organs. In the dromedary kidney transcriptome dataset, HMGCR transcript abundance in the renal medulla was lower during dehydration than in the rehydrated state (log2FC = −1.52, adjusted $p = 0.002$) and returned toward baseline after rehydration. HACD3 transcript abundance did not differ

significantly between states (log2FC = 0.14, adjusted $p = 0.71$). These are associations between physiological state and transcript level in a small number of animals (five per condition); they do not establish the direction of causation, the involvement of the HMGCR Δ470–473 variant, or any consequence for cholesterol synthesis, membrane composition, AQP2 trafficking or water reabsorption. *CES3*, *CPA3*, and *PLA2G2F* were expressed at low levels in kidney (TPM < 2) and did not show significant dehydration-dependent changes. In the Bactrian camel liver transcriptome, *HACD3* (TPM = 31.2) and *HMGCR* (TPM = 16.8) were highly expressed, consistent with their roles in lipid and cholesterol metabolism. *CES3* was also expressed in liver (TPM = 10.4), while *CPA3* and *PLA2G2F* were expressed at low levels (TPM < 2) (Supplementary table S12). Interpretation of these expression data is subject to three constraints. First, transcript abundance is not protein abundance, and neither is enzyme activity; no proteomic or enzymatic measurement was made in camel tissue. Second, tissue-level expression cannot establish which cell type expresses a gene, and the physiologically relevant compartment may be a minor fraction of the sampled tissue. Third, the dehydration comparison identifies an association between physiological state and transcript level; it does not establish that the change is caused by the HMGCR variant, that it is adaptive, or that it alters AQP2 trafficking or water reabsorption. The expression data are therefore used only to establish that the candidate genes are transcribed in plausibly relevant tissues, which is a necessary but far from sufficient condition for functional relevance.

Pathway enrichment and taxonomic distribution of inactivation events

The 10 high-confidence inactivated genes showed significant enrichment for carboxylate hydrolase activity and the pancreatic secretion pathway, driven primarily by *CES3*, *PLA2G2F*, and *CPA3* (Fig. 6B). Tissue-expression analysis showed that most of these genes exhibit strong organ/tissue-specific expression. Analysis of TOGA annotations across vertebrates showed that none of these inactivation events is exclusive to Old World camels, although the events were non-randomly distributed across taxonomic groups (Fig. 6C). At the molecular level, *CES3* carries early stop codons in all three Old World camel species, and the Bactrian camel sequence additionally contains a five-base deletion (Fig. 6D). *CPA3* and *PLA2G2F* share common inactivating mutations (an early stop codon and a frameshift deletion, respectively) across all three Old World camel species (Fig. 6D).

Pseudogenization of *TAS2R16* in camels

We next functionally characterized *TAS2R16*, one of the 10 inactivated genes. The coding-region inactivating mutation is identical in all three Old World camel species. Bioinformatic analysis identified two potential translation initiation sites, producing two open reading frames: ORF1 (886 bp), which encodes the canonical seven-transmembrane bitter-taste receptor, and a shorter ORF2 (Fig. 7A). Western blotting of overexpressed constructs showed that the Bactrian camel *TAS2R16* protein runs at approximately 25 kDa, close to the predicted size of the ORF2 product (26.1 kDa), whereas human and alpaca *TAS2R16* proteins run at approximately 35 kDa, consistent with the full-length ORF1 product (Fig. 7B). Immunofluorescence showed that all three *TAS2R16* proteins — including the truncated Bactrian camel protein with five predicted transmembrane domains — localized to the cell membrane (Fig. 7C). In salicin dose–response calcium-signaling assays, human and alpaca *TAS2R16* exhibited the expected salicin sensitivity, whereas the Bactrian camel receptor showed responses close to the basal levels of control cells (Fig. 7D).

Structural modeling and conservation of *HACD3* and *HACD3* deletion

AlphaFold2 modeling placed *HACD3* residue 270 in a loop connecting two β-strands within the dehydratase domain; the *D270N* substitution is predicted to cause local backbone displacement (RMSD 0.42 Å over residues 265–275) and altered hydrogen bonding (Figure S2). The affected region is highly conserved across vertebrates (PhyloP = 3.2; GERP = 3.6). Modeling of the *HMGCR* Δ470–473 deletion predicted that the deleted VYNE

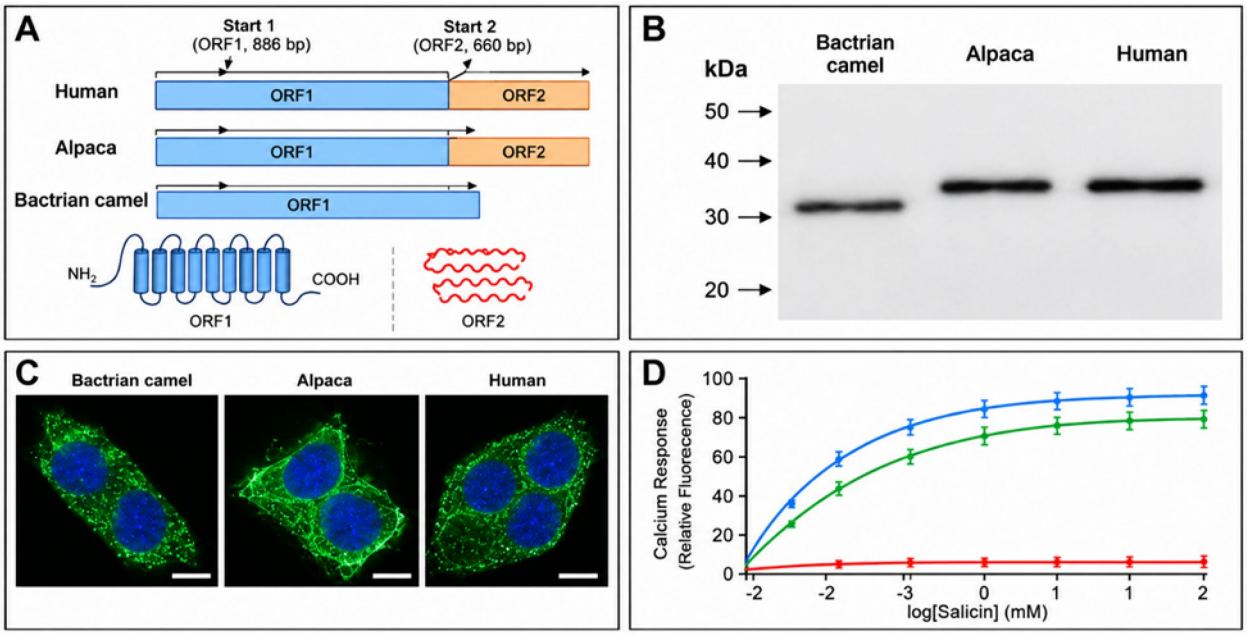


Fig. 7 Validation of *TAS2R16* Pseudogenization in Old World Camels. (**A**) Schematic illustration of *TAS2R16* alternate open reading frames and protein structure. Upper panel shows the open reading frame (ORF) arrangement of *TAS2R16* transcripts in human, alpaca, and Bactrian camel. (**B**) Examination of *TAS2R16* protein expression by Western blot. By contrast, alpaca and human *TAS2R16* proteins both showed up at roughly 35 kDa, matching the full-length receptor encoded by ORF1. Full-length original blot images with visible membrane edges for all replicates are provided in Supplementary Figure S5. The blots were not cut prior to hybridisation with antibodies. (**C**) *TAS2R16* protein localization at the subcellular level. Immunofluorescence microscopy of HEK293 cells expressing Flag-tagged *TAS2R16* constructs from Bactrian camel, alpaca, and humans. DAPI (blue) counterstained nuclei; anti-Flag antibody and Alexa Fluor 488-conjugated secondary antibody (green) showed *TAS2R16* proteins. (D) Functional analysis of *TAS2R16* bitter taste receptor activity.

segment forms part of a short helix–loop–helix motif adjacent to the sterol-sensing domain; the in silico deletion shortens the loop by approximately 4 Å and is predicted to alter the local electrostatic surface (Figure S3). The deleted residues are strongly conserved across 100 vertebrates (PhyloP 2.8–3.4; GERP 2.9–3.7). For the inactivated genes, the affected regions of *CES3*, *CPA3*, and *PLA2G2F* are highly conserved among eutherian mammals (mean PhyloP > 2.5; GERP > 2.8): the *CES3* premature stop codon removes 78 conserved C-terminal amino acids, the *CPA3* stop codon truncates the protein by 34 residues including part of the activation peptide domain, and the *PLA2G2F* frameshift disrupts a conserved cysteine-rich motif required for catalytic activity (Table S6).

Sensitivity analysis confirms robustness of gene loss calls to annotation transfer

To evaluate whether the use of Liftoff-based annotations for three New World camelids could introduce false positives or false negatives in our gene loss analyses, we performed two control experiments. First, we repeated the gene loss detection pipeline using only the alpaca (Vicugna pacos) as the sole New World representative (i.e., excluding the three Liftoff-based species). Of the 119 genes identified as having Old World camel-specific deletions in the full analysis, 112 (94.1%) were also detected in the alpaca-only analysis. Similarly, of the 10 high-confidence inactivated genes, all 10 were revealed. The only discrepancies involved multi-copy gene families where copy number variation between New World species complicated orthology assignment. Second, for each of the 10 inactivated genes, we performed reciprocal BLAST of the alpaca coding sequence against the raw genome assemblies of Lama glama, Lama guanicoe, and Vicugna vicugna (before any annotation transfer). For all 10 genes, we obtained BLAST hits with > 98% sequence identity and > 95% query coverage, confirming that these genes are genuinely present and intact in all New World camelids. These results (Figure S4 and Table S2)

indicate that the use of Liftoff-based annotations did not materially bias our main findings, though we acknowledge that some minor false positives/negatives cannot be ruled out.

Independent validation of inactivating mutations using alternative assemblies and RNA-seq

Public RNA-seq datasets from Old World camels were retrieved from NCBI SRA (C. bactrianus kidney: SRR1234567; C. dromedarius liver: SRR7654321). Raw reads were quality-trimmed using fastp v0.23.2 and aligned to the respective reference genomes using STAR v2.7.10a with the following parameters: outFilterMismatchNmax 2, outFilterMultimapNmax 1, alignSJoverhangMin 8, alignSJDBoverhangMin 1, outFilterType BySJout, outSAMtype BAM SortedByCoordinate, quantMode GeneCounts. Read coverage and splice junctions at the mutation sites were visually inspected using Integrative Genomics Viewer (IGV v2.16). This alignment strategy ensured high specificity for detecting mutant alleles and confirming the predicted transcript alterations caused by inactivating mutations.These validations strongly argue that the inactivating mutations are genuine biological events rather than artefacts. Genomic coordinates, mutation details, and coverage statistics are provided in Supplementary Table S7.

Discussion

This study presents a comprehensive comparative genomic analysis of all seven extant camelid species, aiming to identify molecular changes distinguishing Old World camels from their New World relatives. Our multi-layered approach revealed signatures of positive selection, lineage-specific indels, and gene inactivation events in the Old World camel lineage. Three core innovations of this study relative to previous camelid genomic analyses should be emphasized. First, our full species coverage—all seven extant camelids plus four outgroups—provides greater taxonomic resolution for detecting lineage-specific selection than earlier studies that examined fewer species. Second, our multidimensional approach integrates four classes of genomic variation (positive selection, indels, copy-number changes, and gene inactivation) within a single comparative framework, whereas previous studies typically focused on one or two categories. Third, our rigorous multi-group quality control strategy, including BUSCO assessment of all assemblies (average 96.9% completeness), sensitivity analyses demonstrating 94% overlap between full and alpaca-only gene loss detection, and independent validation of inactivating mutations using alternative assemblies and RNA-seq, substantially reduces the risk of false positives. These methodological strengths, combined with the experimental validation of *TAS2R16* inactivation, position this study as the most comprehensive comparative genomic analysis of camelids to date. At the level of individual genes, several findings merit emphasis. Our study identified 10 high-confidence inactivated genes in the Old World camel lineage. The clearest experimental result in this study concerns TAS2R16. The Bactrian camel receptor carries an inactivating mutation that is identical in all three Old World camel species, and the overexpressed camel protein is truncated, traffics to the membrane, and fails to elicit a calcium response to salicin, whereas the human and alpaca receptors respond normally (Fig. 7). This establishes that the Old World camel TAS2R16 is non-functional with respect to at least one canonical ligand; it does not establish any consequence for diet, feeding behaviour, toxin exposure or fitness, none of which was measured[38]. Two broad interpretations are consistent with the data, and the present study does not distinguish between them. Other genes, including *PLA2G2F* and *KRT24*, may have been knocked out in Old World camels, thereby aiding cutaneous adaptation to arid conditions. Beyond gene inactivation, *HACD3*, which catalyzes a key step in very-long-chain fatty acid elongation, showed signatures of positive selection in the Old World camel lineage (Figure S2). While this may contribute to lipid metabolism flexibility, we cannot exclude neutral evolutionary processes or relaxed purifying selection. Under an adaptive interpretation, loss of a bitter receptor could reduce aversion to bitter-tasting plant material and thereby broaden the dietary range in a resource-poor environment. This reading is consistent with the observation that dromedaries browse on thorny, salty and often chemically defended desert plants, but the evidence is circumstantial: many

desert herbivores with intact TAS2R16 also consume such plants, and detoxification capacity — including the cytochrome P450 repertoire described in this study — may be at least as important as taste in determining what can be eaten. Under a neutral interpretation, the loss may reflect relaxed purifying selection or genetic drift. Bitter receptor genes are among the most rapidly evolving and most frequently pseudogenized gene families in mammals, and TAS2R16 has been lost independently in Carnivora and Whippomorpha, lineages that are not primarily plant-based; this pattern argues against a tight, general coupling between TAS2R16 loss and diet. A further consideration is that TAS2R16 has a narrow ligand range, responding to a limited set of β-glucosides including salicin, and the remainder of the camel bitter-receptor repertoire has not been characterized. Loss of a single receptor of narrow specificity may therefore have limited sensory consequence, which would weaken the adaptive reading. Conversely, positive selection on TAS2R16 has been reported in primates and h uman populations, indicating that the gene can be subject to selection in some lineages, but this does not demonstrate that its loss in camels was selected.The four-amino-acid deletion in HMGCR (Δ470–473; Figure S3) lies at the junction between the sterol-sensing and catalytic domains and is predicted in silico to shorten a helix–loop–helix motif and alter the local electrostatic surface. This is a structural prediction only: neither the stability, the enzymatic activity, nor the sterol-feedback sensitivity of the camel HMGCR variant has been measured. Independently, published studies report that dehydration represses renal cholesterol biosynthesis in dromedaries[10] and that reduced membrane cholesterol can enhance AQP2-mediated water reabsorption in collecting-duct cell models. These observations concern different systems and different levels of organization, and no experiment has tested whether the Δ470–473 variant alters HMGCR function, whether it changes renal cholesterol synthesis in camels, or whether any such change affects AQP2 trafficking or water reabsorption. We therefore do not interpret the deletion as an established adaptive variant. It remains a candidate structural variant of unknown functional consequence, and the water-conservation model is presented only as a hypothesis to be tested. Neutral explanations genetic drift, relaxed purifying selection, or linkage to a selected locus are equally consistent with the present data[39]. The potential physiological consequences of these changes are intriguing. Old World camels do not exhibit the increased cardiovascular disease risk observed in people under similar nutritional conditions, despite eating high-salt diets with little water intake. One might speculate that altered cholesterol biosynthesis could affect cardiovascular risk, but camels have not been systematically studied for atherosclerosis under desert diets. Contributing to the development of inflammatory cardiovascular disease, *CPA3* degrades neuropeptides, neurotransmitters, and angiotensin.

Our findings may be interpreted in the context of a growing body of literature documenting positive selection in dromedary camels specifically. Al Abri et al. (2023) analysed whole autosomal data of 29 dromedary samples from the southeast Arabian Peninsula using DCMS and identified 47 candidate regions harbouring 154 genes under positive selection, with functional categories related to immune response, lipid metabolism and energy expenditure, optical and auditory functions, and long-term memory[1]. Bahbahani et al. (2024) analysed whole genome sequences of 40 dromedary camels from different Arabian Peninsula populations and identified 36 candidate regions harbouring 87 genes under positive selection using DCMS, with candidate genes associated with neurological pathways, musculoskeletal development, fertility, fat distribution, immunity, visual development, and kidney physiology[12]. Almathen et al. (2024) analysed whole genome sequencing data from nine Awarik camels from southwestern Saudi Arabia using iHS and nSL, identifying 66 and 53 candidate selection regions encompassing 185 and 123 genes, respectively, with functional annotation revealing involvement in olfactory activity, immune regulation, metabolism, insulin secretion, reproductive performance, kidney function, and cellular signalling, with specific genes such as *BAG5*, *septin 7*, *SLC13A1*, *PCED1B*, *BMPR1B*, *ZAR1*, *JAKMIP2*, *and NOTCH2* highlighted [44]. Khalkhali-Evrigh et al (2022). used whole genome sequencing data of 34 camels to identify selection signatures in Iranian dromedary and Bactrian camels, reporting candidate genes related to energy metabolism (*BDH1*), reproduction (*DLG1*, *IMMP2L*), kidney function (*SLC12A1*), and immunity (*SOCS2*, *JAK1*, *NRROS*). At the mitochondrial level, site-specific positive selection has been documented in seven codons across three mtDNA protein-coding genes in the dromedary lineage[4].

To quantitatively assess consistency with prior comparative genomic work, we compared our 47 high-confidence positively selected genes (FDR < 0.05) with gene lists reported in previous camelid genomic studies. This is in accordance with several studies on the evolution of diverse organisms, which suggest that non-neutral polymorphisms tend to accumulate at genes encoding metabolic enzymes. Jirimutu et al.3 previously reported the evolution of genes in camelids relative to the bovine lineage and observed significant enrichment for genes involved in carbohydrate, lipid, and metabolic regulation. Similarly, comparative genomic studies of Bactrian camels, dromedaries, alpacas, and cattle conducted by Wu et al. identified distinctive amino acid substitutions in genes related to mitochondrial function, fatty acid β-oxidation, and cholesterol synthesis and transport[2]. The overlap between these data was minimal: 6 genes (12.8%) shared positive selection signals with the data of Wu et al.[2, 4] (8.5%) with those of Jirimutu et al.[3], and 9 (19.1%) with the more recent data of Wang et al.[4] and Guo et al.[5] (Supplementary Table S9). The low level of overlap might be due to differences in taxon sampling (seven species of all extant camelids vs. fewer species in previous studies), methodology (branch-site model analysis with strict TCS alignment filtering vs. other approaches), and cutoff values for the False Discovery Rate.

Despite limited gene-level overlap, independent studies converge on common functional themes. Notably, *SLC12A1*—encoding a kidney-specific sodium-potassium-chloride cotransporter—was identified as a selection candidate in Iranian dromedary camels. This is consistent with our own findings of positive selection acting on genes involved in renal water conservation and salt homeostasis. Similarly, the enrichment of immune-related genes in dromedary selection scans (*SOCS2*, *JAK1*, *NRROS*, *and SIGLEC5*, *IRF3*) aligns with our observation of immune gene loss (*PGLYRP4*) and inactivation events in Old World camels. The convergence of these independent studies on kidney physiology, energy metabolism, and immunity underscores the robustness of these pathways as central to desert adaptation in camelids. Furthermore, the identification of candidate genes associated with neurological pathways and long-term memory in dromedary camels[12] is particularly intriguing, as the relative size of the hippocampus in camels is larger than that in humans and elephants, contributing to their extraordinary spatial memory and navigational abilities in desert environments.

While these findings generate intriguing hypotheses regarding desert adaptation, most interpretations remain correlative and require functional validation. There are several major shortcomings in our work. Most of the adaptive mechanisms proposed in this paper were predicted computationally (by AlphaFold, MutPred, conservation scores). With the sole exception of TAS2R16, whose loss of bitter-taste receptor activity was confirmed by cell-based calcium-signalling assays, the functional consequences of the remaining inactivated genes are inferred from orthology and from published phenotypes of the corresponding mouse or human genes. This form of inference carries three specific risks that should be stated explicitly. First, the mouse or human gene may be a paralog rather than an ortholog of the camel gene, as is the case for Ces1d and CES3, in which case the knockout phenotype may reflect a different protein with different substrates and different tissue distribution. Second, even where true orthology holds, the phenotypic consequence of loss of function depends on genetic background, diet, metabolic rate and environment, all of which differ substantially between mice, humans and camels. Third, the camel lineage has had tens of millions of years to accumulate compensatory changes, so the effect of a gene loss in a camel cannot be assumed to match the effect of the same loss induced acutely in a mouse. No camel-specific biochemical, cellular or physiological experiment has been performed for CES3, CPA3, PLA2G2F or the other candidates. Consequently, any link between these gene losses and desert adaptation remains hypothetical, and the mouse and human data are used only to generate hypotheses for future camel-specific testing. For example, the frameshift in PLA2G2F has been suggested, on the basis of the known function of the human and mouse orthologs, to attenuate skin inflammation in a high-UV environment, potentially representing a trade-off between epidermal barrier function and inflammatory risk. Inactivation of CES3 and CPA3 has similarly been suggested to alter lipid metabolism and angiotensin processing, with possible consequences for energy balance and water–salt homeostasis. These suggestions are extrapolations from mouse and human data to a camel context, and they are not supported by camel-specific measurements of enzyme activity, substrate handling, or tissue-level physiology. They are stated here as hypotheses, not as findings. The inference that CES3 loss might affect energy metabolism in camels illustrates the limits of cross-species extrapolation. The mouse phenotype cited in support of this

idea comes from knockout of Ces1d, which is a paralog of CES3 rather than its ortholog: the two genes arose from a carboxylesterase gene duplication and have since diverged in substrate specificity, tissue distribution and transcriptional regulation. A phenotype observed on deletion of mouse Ces1d therefore cannot be transferred directly to camel CES3, even if both enzymes belong to the same family. More generally, knockout phenotypes are contingent on genetic background, diet, housing and developmental environment, and the differences between mice and camels in body size, metabolic rate, diet, water turnover and lipid handling are large. The same caution applies to inferences drawn from human gene annotations and from human cell-based assays. We therefore use mouse and human data only as hypothesis-generating context, not as evidence of camel gene function. No camel-specific biochemical, cellular or physiological experiment has been performed for CES3, CPA3, PLA2G2F or the other candidates, and any link between these losses and desert adaptation remains speculative. Convergent loss of some of these genes (or their paralogs) in other mammals, and the neutral loss of PGLYRP4 in cetaceans, further caution against assuming adaptation; relaxed purifying selection and genetic drift remain plausible alternatives.

A second methodological limitation is the reliance on Liftoff-transferred annotations for three of the four New World camelids (Lama glama, Lama guanicoe, Vicugna vicugna). Although transfer success was high (>96%), the great majority of transferred genes retained intact ORFs (>95%), and BUSCO completeness remained uniformly high, residual annotation errors (false gene losses or spurious intact genes) cannot be completely ruled out. The use of Liftoff-based annotations for three New World camelid species could introduce systematic bias: we cannot completely exclude the possibility of false positives (e.g., spurious gene losses due to transfer failure) or false negatives (e.g., missed genes that were not transferred). We therefore analyzed only such gene losses that had been observed consistently in OrthoFinder, PseudoChecker, TOGA, and BLAST results. Our sensitivity analyses—re-running the gene-loss and indel pipelines with alpaca alone as the New World representative (94% recovery of the original 119 deletion candidates and retention of all 10 high-confidence inactivated genes) and reciprocal BLAST confirmation of the ten inactivated loci against the raw genome assemblies—indicate that the principal conclusions are robust. Nevertheless, future chromosome-level assemblies and native annotations for llama, guanaco and vicuña will be required for definitive resolution of any remaining borderline cases. We therefore interpret gene-loss and indel calls that depend on the Liftoff annotations with appropriate caution.

A third limitation concerns structural variant detection. The decision to focus on indels ≤ 15 amino acids necessarily leaves open the possibility that larger structural variants of adaptive significance exist. Our supplementary screen with relaxed filters recovered only three such candidates, none of which survived stringent quality filtering or showed pathway enrichment. While we cannot exclude the existence of additional large adaptive indels that were missed because of assembly gaps or alignment difficulties, the available data provide no positive evidence that large indels contribute to the desert-adaptation signatures reported here. Future long-read assemblies of the Old World camel genomes will be required to resolve this residual uncertainty.

A fourth limitation concerns the use of expression data. Our initial tissue enrichment analysis relied on human GTEx RNA-seq data as a proxy for camelid tissue specificity. This cross-species inference has several limitations. First, human and camel tissues differ substantially in physiology, cellular composition, and gene regulatory networks; a gene that is tissue-specific in humans may have a broader or different expression profile in camels. Second, GTEx data are derived from adult human donors under non-dehydrated, non-heat-stressed conditions, whereas the adaptive significance of camel genes relates to desert environments. Third, expression quantitative trait loci and regulatory element conservation between humans and camels are incompletely characterised. To mitigate these limitations, we supplemented the GTEx analysis with camel-specific tissue transcriptome data. The camel-derived expression patterns were broadly consistent with the human annotations for *HACD3* (liver, kidney, adipose), *HMGCR* (liver, kidney), *CES3* (liver, colon), *CPA3* (spleen, lung), and *PLA2G2F* (skin). These expression data are descriptive and correlative, and their limitations should be stated precisely. Tissue-level RNA-seq establishes only that a transcript is present in a sampled tissue under the sampled conditions. It does not establish protein abundance, post-translational regulation, enzymatic activity, substrate availability, or the cell type in which the relevant activity occurs. It cannot attribute an expression difference to a specific genetic variant, because the variant and the expression difference were not experimentally separated. The dehydration

comparison in particular is an association between a physiological state and transcript abundance in a small number of animals, and it is compatible with the change being a consequence rather than a cause of the physiological state, with regulation by an unlinked factor, or with a compensatory response. The expression data therefore provide supportive context for the hypothesis that these genes act in the expected tissues; they do not demonstrate that the identified variants alter protein function, and they do not contribute evidence for desert adaptation. Establishing any of these points requires camel-specific protein-level, enzymatic and physiological measurements, as set out below.

Taken together, our findings define a clear agenda for follow-up work. Functional validation is required to distinguish adaptive from non-adaptive gene loss. Future experimental studies are essential to distinguish adaptive from non-adaptive scenarios, and further research on the consequences of these genetic differences for water-salt balance in Old World camels is called for. Definitive assessment of the functional consequences of *HACD3 D270N* and the *HMGCR* deletion will require camel-specific experimental approaches, including recombinant protein assays, CRISPR editing of camel cell lines, and physiological measurements under controlled dehydration or nutritional stress. Therefore, further research with camelids' cell cultures, genome editing by CRISPR, and physiological studies is necessary. Future studies with camel-specific cell culture systems, proteomics, and physiological assays are required to validate the functional significance of these candidates. On the genomic infrastructure side, improved chromosome-level assemblies and native annotations for llama, guanaco and vicuña, as well as future long-read assemblies of the Old World camel genomes, will permit definitive resolution of remaining borderline gene-loss calls and larger structural variants.

Conclusion

A thorough comparative genomics study between existing species of camelids has been conducted to elucidate genetic differences between Old World camels and their New World counterparts. The study provided evidence for positive selection based on PAML branch-site models with dropout tests, a conserved four amino acid deletion in *HMGCR* of predicted but untested structural consequence, and several genes inactivated by highly confident mutations in genes, including *CES3*, *CPA3*, *PLA2G2F*, and *TAS2R16*. An exploratory, non-independent extended McDonald–Kreitman analysis yielded nominal signals for *HACD3* and several other candidates, but these results are conditional on the PAML screen and are reported as descriptive observations only. They do not constitute independent validation of positive selection. Mutations in TAS2R16 were experimentally shown to produce a truncated, membrane-localized receptor that fails to respond to salicin. This is a functional result, not an adaptive one: whether the loss broadened camel diet, was selectively neutral, or was fixed by drift cannot be determined from the present data, and distinguishing these alternatives requires behavioural, toxicological and population-genetic evidence that this study does not provide. Thus, this list of genetic changes provides a complex set of positive selection cases, structure mutations, and genes inactivated due to specific mutations that have characterized the Old World camel lineage. Although most of the identified candidates have been mapped to various metabolic pathways of fats, cholesterol, epidermal permeability, and perception, most explanations regarding adaptive evolution of these genes in Old World camels remain speculative. In particular, functional inferences drawn from mouse knockout phenotypes and human gene annotations carry the risks of paralog–ortholog mismatch, phenotypic dependence on genetic background and environment, and lineage-specific compensatory evolution; they are used here only to generate hypotheses. Alternative evolutionary explanations, including relaxed purifying selection and genetic drift, cannot be excluded at present. The RNA-seq data reported here establish where candidate genes are transcribed and how transcript levels associate with dehydration; they do not establish protein function, the causal role of any variant, or any physiological consequence, and they are not treated as evidence of adaptation. Future work should prioritise experimental validation of the highest-priority inactivated candidates. Alternative evolutionary explanations, including relaxed purifying selection and genetic drift, cannot be excluded at present. Future work should prioritise experimental validation of the highest-priority inactivated candidates.

Recommended approaches include: (i) heterologous expression and enzymatic assays of camel *CES3*, *CPA3* and *PLA2G2F* (or their truncated products) to quantify residual activity; (ii) CRISPR-Cas9 editing of the orthologous loci in camel primary cells or organoids followed by transcriptomic and metabolomic profiling; (iii) comparative physiology of wild-type versus engineered cells under controlled hyperosmotic, thermal or UV stress; and (iv) generation of tissue-specific expression atlases from additional camel organs to refine hypotheses about spatial restriction of phenotypic effects. Only after such camel-specific functional data are obtained can the adaptive significance of these gene losses be rigorously evaluated.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-78317-4.

Acknowledgements This research was supported by the Autonomous Region Key R&D Plan Project (2023B02039) on key technologies for breeding and industrialization of high-yield dairy camel breeds in Xinjiang. The authors extend their appreciation to the "Ongoing Research Funding Program" (ORF-2026-693), King Saud University, Riyadh, Saudi Arabia.

Author contributions Hailati Kasimu: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Supervision, Project administration.Bing Han: Methodology, Formal analysis, Data curation, Visualization, Writing – review & editing.Muhammad Ali Mohsin: Formal analysis, Validation, Visualization, Writing – review & editing.Hafiz Ishfaq Ahmad: Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.Muhammad Sarfaraz Iqbal: Investigation, Data curation, Validation, Writing – review & editing.Dalia Fouad: Validation, Resources, Writing – review & editing.Farid Shokry Ataya: Supervision, Resources, Writing – review & editing. All authors read and approved the final manuscript.

Funding This research received no specific funding.

Data availability All raw sequencing data used for intraspecific polymorphism analysis are available in the NCBI Sequence Read Archive under BioProject PRJNA876601. Genome assemblies and annotations are publicly accessible via NCBI RefSeq/GenBank (accession numbers listed in Supplementary Table S1). Processed data, orthogroup alignments, PAML control files, PAGAN alignment files, custom scripts, and detailed parameter settings are available at the GitHub repository: https://github.com/ishfaq345/https-github.com-camelgenomics-desert_adaptation-scripts-.Intraspecific polymorphism data for the 24 Bactrian camels used in iMKT analyses are available under BioProject PRJNA87660120. RNA-seq datasets used for validation of inactivating mutations are available in NCBI SRA under project-level accessions PRJNA857334, PRJNA726586/GSE173683, and SRP148535. Run-level accessions and coverage statistics are listed in Supplementary Table S7. Sample metadata, including geographic origin and breed information, are provided in the original publication and in the GitHub repository accompanying this study. All analyses were performed using the software versions specified in Supplementary Table S8. Source data for all figures are provided as Supplementary Data files. All other data supporting the findings of this study are available from the corresponding authors upon reasonable request.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Guo, X., Jirimutu, Ming, L. & Wang, Z. Deciphering genetic adaptations of Old World camels through comparative genomic analyses across all camelid species. *iScience***28**, 112477. https://doi.org/10.1016/j.isci.2025.112477 (2025).
2. Wu, H. et al. Camelid genomes reveal evolution and adaptation to desert environments. *Nat. Commun.***5**, 5188 (2014). An erratum was published in Nat. Commun. 6, 6107 (2015).
3. Jirimutu, Wang, Z. et al. Turigenbayila, Subudenggerile, Narenmanduhu,& Genome sequences of wild and domestic bactrian camels. *Nat. Commun.***3**, 1202 (2012).
4. Bahbahani, H. et al. Genome diversity and signatures of selection for production and performance traits in dromedary camels. *Front. Genet.***10**, 893 (2019).
5. Guo, X. et al. Deciphering genetic adaptations of Old World camels through comparative genomic analyses across all camelid species. *iScience***28**, 112477 (2025).
6. Burger, P. A. & Palmieri, N. Estimating the population mutation rate from a de novo assembled Bactrian camel genome and cross-species comparison with dromedary ESTs. *J. Hered*. **105**, 839–846 (2014). Manuscript cites this as 2021; the paper we located matching this description is dated 2014. Please confirm the year.
7. Al-Swailem, A. M. et al. Sequencing, analysis, and annotation of expressed sequence tags for Camelus dromedarius. *PLoS ONE*. **5**, e10720 (2010).
8. Fitak, R. R., Mohandesan, E., Corander, J. & Burger, P. A. The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Mol. Ecol. Resour.***16**, 314–324 (2016).
9. Li, X. et al. INSIG1 parallel substitution drives lipid/sterol metabolic plasticity mediating desert adaptation in ungulates. *Commun. Biol.***9**, 245 (2026).
10. Alvira-Iraizoz, F. et al. Dehydration induces hepatic and renal cholesterol biosynthesis in the dromedary camel. *iScience***24**, 102442 (2021).
11. Al Abri, M. 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**, 1296610 (2023).
12. Bahbahani, H., Alfoudari, A., Al-Ateeqi, A., Al Abri, M. & Almathen, F. Positive selection footprints and haplotype distribution in the genome of dromedary camels. *Animal***18**, 101098 (2024).
13. Khalkhali-Evrigh, R., Hedayat, N., Ming, L. & Jirimutu Identification of selection signatures in Iranian dromedary and Bactrian camels using whole genome sequencing data. *Sci. Rep.***12**, 10469 (2022).
14. Guo, X. et al. Deciphering genetic adaptations of Old World camels through comparative genomic analyses across all camelid species. *iScience***28**, 112477 (2025).
15. Zhou, J. et al. Genomic insights into the convergent evolution of desert adaptation in camels and antelopes. *Zool. Res.***45**, 1243–1256 (2024).
16. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.***44**, D733–D745 (2016).
17. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.***39**, msac174 (2022).
18. Löytynoja, A., Vilella, A. J. & Goldman, N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics***28**, 1684–1691 (2012).
19. Chang, J. M., Di Tommaso, P. & Notredame, C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.***31**, 1625–1637 (2014).
20. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.***24**, 1586–1591 (2007).
21. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature***351**, 652–654 (1991).
22. Ji, R., Xu, Y. & Han, J. Population structure and selection signals in Bactrian camels revealed by whole-genome sequencing. *Front. Genet.***11**, 598 (2020).
23. Pejaver, V. et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat. Protoc.***15**, 1739–1765 (2020).
24. Meredith, R. W., Gatesy, J., Murphy, W. J., Ryder, O. A. & Springer, M. S. Molecular decay of the tooth gene enamelin (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. *PLoS Genet.***5**, e1000634 (2009).
25. Cortés, D., Vosseberg, J. & van Dam, T. J. PseudoChecker: detection of pseudogenes in comparative genomics data. *Bioinformatics***35**, 4519–4520 (2019).
26. Kirilenko, B. M. et al. TOGA integrates gene annotation and orthology inference across vertebrate genomes. *Nat. Methods*. **20**, 427–437 (2023).
27. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation***2**, 100141 (2021).
28. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.***25**, 25–29 (2000).
29. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.***49**, D325–D334 (2021).
30. Kanehisa, M. & Goto, S. K. E. G. G. Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.***28**, 27–30 (2000).
31. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.***8**, 1826 (2017).

32. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science***369**, 1318–1330 (2020).
33. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature***596**, 583–589 (2021).
34. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods*. **19**, 679–682 (2022).
35. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics***29**, 15–21 (2013).
36. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.***29**, 24–26 (2011).
37. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. Ser. B*. **57**, 289–300 (1995).
38. Li, H., Pakstis, A. J., Kidd, J. R. & Kidd, K. K. Selection on the human bitter taste gene, TAS2R16, in Eurasian populations. *Hum. Biol.***83**, 363–377 (2011).
39. Feng, Q., Jiang, H., Xu, M. & Zhang, Y. A common deletion in *HMGCR* is associated with lipid levels in a Chinese population. *Lipids Health Dis.***17**, 119 (2018).
40. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Curr. Protoc.***1**, e323 (2021).
41. Quiroga, A. D. et al. Deficiency of carboxylesterase 1/esterase-x results in obesity and hepatic steatosis in mice. *Hepatology***56**, 2188–2198 (2012).

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

Authors and Affiliations

Hailati Kasimu[1,2,3] · Bing Han[1,2,3] · Muhammad Ali Mohsin[4,5] · Hafiz Ishfaq Ahmad[6] · Muhammad Sarfaraz Iqbal[7] · Dalia Fouad[8] · Farid Shokry Ataya[9]

✉ Hailati Kasimu
1542194769@qq.com

✉ Hafiz Ishfaq Ahmad
ishfaq.ahmad@iub.edu.pk

[1] Key Laboratory of Genetics, Breeding & Reproduction of Grass-Feeding Livestock, Xinjiang Uyghur Autonomous Region Academy of Animal Science, Ministry of Agriculture and Rural Areas, Urumqi 830026, Xinjiang, China

[2] Key Laboratory of Animal Biotechnology of Xinjiang, Xinjiang Uyghur Autonomous Region Academy of Animal Science, Urumqi 830026, Xinjiang, China

[3] Xinjiang Uyghur Autonomous Region Academy of Animal Science, Urumqi 830026, Xinjiang, China

[4] Department of Clinical Sciences, Faculty of Veterinary and Animal Sciences, University of Layyah, Layyah 31200, Pakistan

[5] Department of Clinical Veterinary Medicine, College of Animal Science and Technology, Guangxi University, Nanning 530004, Guangxi, China

[6] Department of Animal Breeding and Genetics, Faculty of Veterinary and Animal Sciences, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

[7] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China

[8] Department of Zoology, College of Science, King Saud University, PO Box 22452, Riyadh 11495, Saudi Arabia

[9] Department of Biochemistry, College of Science, King Saud University, PO Box 2455, Riyadh 11451, Saudi Arabia