Cagliani *et al. Biology Direct* (2026) 21:32
https://doi.org/10.1186/s13062-026-00741-3

Biology Direct

RESEARCH Open Access

Urea transporter evolution: deep conservation, recent adaptation in mammals, and maintenance of the Jka/Jkb polymorphism in the past ~110,000 years

Rachele Cagliani[1*†], Uberto Pozzoli[1†], Diego Forni[1], Alessandra Mozzi[1] and Manuela Sironi[2,3]

Abstract

Urea is a central metabolite and many organisms encode membrane integral urea transporters (UTs). Here, we combined motif homology searches, phylogenetics, and molecular evolution with paleogenomics to investigate the evolutionary dynamics of UTs from prokaryotes to human populations. We discovered previously unknown combinations of UT domains with other functional modules in metazoa, protists, and bacteria, suggesting a wider range of functions and regulatory mechanisms for UTs than previously understood. The early origin of UT domains allowed the identification of specific residues that have remained conserved across billions of years. Contestually, we found evidence that UT-B was a target of positive selection in mammals. In particular, selection targeted UT-B in bats, primates and rodents, possibly as the result of a pressure exerted by blood pathogens. In human populations, a single variant (Asp280Asn) in the UT-B protein is responsible for the common Kidd blood group antigens. Here we provide direct evidence for the temporal stability of the Asp280Asn polymorphism over ~110,000 years in Europe and Asia. While we previously proposed this variant to be under balancing selection, this study is the first to use ancient DNA to track its allele frequency through deep time. This novel application of paleogenomics confirms that the polymorphism was maintained at stable frequencies over time, in different European sub-regions, and across continents. Our data show how paleogenomics can provide information on the selective processes in humans, not only limited to directional selection, but also to balancing selection.

Keywords Urea transporters, Positive selection, Kidd blood group, SLC14A1 Asp280Asn variant, Balancing selection

[†]Rachele Cagliani and Uberto Pozzoli contributed equally to this work.

*Correspondence:
Rachele Cagliani
rachele.cagliani@lanostrafamiglia.it
[1]Scientific Institute IRCCS E. MEDEA, Computational Biology Unit, 23842 Bosisio Parini, Italy
[2]School of Medicine and Surgery, University of Milano-Bicocca, 20900 Monza, Italy
[3]Fondazione IRCCS San Gerardo dei Tintori, 20900 Monza, Italy

Introduction

Urea is the main end product of protein metabolism in animals and, for many bacteria, it is an important nitrogen source. Urea also functions as an essential osmolyte for many aquatic vertebrates and a cryoprotectant in some hibernating species [1]. To facilitate its transport across membranes, many organisms encode urea transporters (UTs). UTs are membrane integral proteins that mediate the rapid transmembrane permeation of urea. The majority of UTs share a common structure, with 10 transmembrane domains that allow urea transport with a channel-like mechanism [1]. Both in metazoa and in bacteria, UTs are homomultimers (dimers or trimers) with a pore in each subunit, and the overall three dimensional structure is conserved [2].

In vertebrates, where the urea transport system has been intensely investigated, distinct species encode a different repertoire of UTs, which are likely to derive from duplications of an ancestral gene [1]. The most notable expansion occurred in fishes, whereas mammals possess only two UTs, usually referred to as UT-A (encoded by the human *SLC14A2* gene) and UT-B (encoded by *SLC14A1*). The gene product of *SLC14A1* contains a single UT domain and is expressed in the kidneys and bladder, as well as in a number of other tissues including erythrocytes, heart, colon, and the brain [1]. In contrast, *SLC14A2* encodes two UT domains in tandem, and produces a variety of isoforms via alternative splicing (UT-A1 to -A6) [3]. The expression of UT-A isoforms is narrower compared to UT-B and mainly limited to ureogenic tissues. In line with this differential expression, mice that are genetically deficient for UT-A1/A3 exhibit strong urine-concentrating defects, whereas the deficit is much less evident in animals lacking UT-B [3, 4]. In contrast, UT-B null mice were shown to present extrarenal phenotypes such as exacerbated heart problems with aging, depression-like behavior, and earlier sexual maturation in males [5]. In humans, individuals genetically deficient for UT-B reach a frequency higher than 1% in some areas of Polynesia [6]. This is known because the protein product of *SLC14A1* is responsible for one of the human blood group systems, namely the Kidd blood group. A single variant (Asp280Asn, rs1058396) in the UT-B protein is responsible for the common *Jka/Jkb* antigens, whereas different alleles (the most common one in Polynesia being a splice site mutation) account for Kidd-null phenotypes [6]. Jk(a–b–) individuals are healthy and suffer from minor urine concentration defects, but are at risk of hemolytic reaction after transfusion and hemolytic disease of the newborn [6].

Because of the central relevance of urea metabolism across the domains of life and given the extremely diverse metabolic and homeostatic requirements among organisms, we aimed to perform a comprehensive evolutionary analysis of UT genes across different timeframes. We thus combined domain searches, phylogenetics, molecular evolution analysis, and ancient human DNA data to resolve the evolutionary dynamics of these transporters across time scales.

Materials and methods

Identification of the urea transporter domain

The profile of the urea transporter domain (PFAM PF03253) was retrieved from the InterPro website (https://www.ebi.ac.uk/interpro/) and searched against the UniProt database (https://www.uniprot.org/, last accessed August 5, 2025) using hmmsearch (version 2.39.0) [7]. In order to reduce false positives, we have applied stringent filters: we selected hits having an E-value < 1.10^{-5}, in proteins longer than 150 amino acids and in which the identified domain was longer than 100 amino acids. Finally, using Uniprot annotation, we removed all proteins annotated as "fragment". In the event that the same gene appeared more than once (e.g.: isoforms), the longest one was selected.

Proteins were screened using hmmscan [7] against the Pfam-A database [8] to identify all possible domains. The analysis was performed with default parameters, and results with an E-value lower than 1×10^{-5} were considered significant.

UT domains were aligned using MAFFT v7.475 with default parameters [9] and the resulting alignment was used to build a maximum likelihood phylogenetic tree with the IQ-TREE software (version 3.0.1) [10]. ModelFinder [11] identified as the best substitution model the Pfam-derived empirical substitution matrix with observed amino acid frequencies, and a four-category discrete gamma distribution to account for variation among-sites. Finally, branch support was evaluated using 1,000 ultrafast bootstraps [12]. The final tree was visualized using FigTree v1.4.4 (https://tree.bio.ed.ac.uk/software/figtree/).

Tree representatives were aligned using MAFFT, and the identity of each amino acid position was estimated. Archaea domains were merged and considered as a single unit.

Protein three dimensional structures were modeled using the AlphaFold server [13] with default parameters. The models with the highest predicted local distance difference test scores were selected for further analyses and visualized using Pymol [14]. Amino acid conservation of the UT domain was plotted on the predicted structure of the trimeric form of human UT-B using Pymol.

Molecular evolution analyses

Coding sequences of *SLC14A1* and *SLC14A2* were retrieved from the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.

gov, last accessed September 23, 2025). In particular, we retrieved sequence information for species belonging to the five most populous orders: Primates, Rodentia, Chiroptera, Carnivora, and Artiodactyla. Sequences having low sequence coverage were excluded. A list of species is reported in Table S1.

The RevTrans 2.0 utility was used to generate multiple sequence alignments (MSAs) using MAFFT v6.240 as an aligner [15]. Phylogenetic trees were reconstructed using the phyML program (version 3.1) with a General Time Reversible (GTR) model plus gamma-distributed rates and 4 substitution rate categories with a fixed proportion of invariable sites [16]. Each resulting alignment was manually inspected and was analyzed for the presence of recombination signals using GARD (Genetic Algorithm Recombination Detection) [17]. GARD is a Genetic Algorithm implemented in the HYPHY suite (version 2.2.4), which uses phylogenetic incongruence among segments in the alignment to detect the best-fit number and location of recombination breakpoints.

To detect positive selection, the codon-based *codeml* program implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood, v4.9) suite was applied [18]. Using F3 × 4 codon frequencies model (codon frequencies estimated from the nucleotide frequencies in the data at each codon site) [18, 19], a model (M8, positive selection model) that allows a class of sites to evolve with dN/dS > 1 was compared to two models (M7 and M8a, neutral models) that do not allow dN/dS > 1. To assess statistical significance, twice the difference of the likelihood (ΔlnL) for the models (M8a vs. M8 and M7 vs. M8) was compared to a $\chi 2$ distribution (1 or 2 degrees of freedom for M8a vs. M8 and M7 vs. M8 comparisons, respectively). To be conservative and to obtain robust results, we called a gene as a target of positive selection only if it was detected by both comparisons.

In order to identify specific sites subject to positive selection, we applied four different methods: (1) the Bayes Empirical Bayes (BEB) analysis (with a cutoff of 0.90), which calculates the posterior probability that each codon is from the positive selection site class (under M8 model) [20]; (2) Fast Unbiased Bayesian AppRoximation (FUBAR) (with a cutoff of 0.90), an approximate hierarchical Bayesian method that generates an unconstrained distribution of selection parameters to estimate the posterior probability of positive diversifying selection at each site in a given alignment [21]; (3) Mixed Effects Model of Evolution (MEME) (with a p-value cutoff < 0.1), which allows the distribution of dN/dS to vary from site to site and from branch to branch at a site [22]; (4) Fixed Effects Likelihood (FEL) (with a p-value cutoff < 0.1), a maximum-likelihood (ML) approach to infer dN/dS on a per-site basis, assuming that the selection pressure for each site is constant along the entire phylogeny [23]. Again, to be conservative and to limit false positives, only sites detected using at least two methods were considered as positive selection targets.

GARD, FEL, FUBAR, and MEME analyses were run locally through the HyPhy suite [24].

Primate polymorphism data

Polymorphis data for chimpanzee (*Pan troglodytes*), rhesus macaque (*Macaca mulatta*), crab-eating macaque (*Macaca fascicularis*), olive baboon (*Papio anubis*), and marmoset (*Callithrix jacchus*) were derived from the Genome variation Map [25] (https://ngdc.cncb.ac.cn/gvm/home). Only missense variants with a frequency higher than 0.10 were included.

Ancient DNA data analysis

To analyze the frequency of the rs1058396 derived allele over time, we used a compilation of ancient DNA genome data made available from the Allen Ancient DNA Resource (AADR, V54.1.p1, Dataverse 8.0, March 6 2023) [26]. Because ancient DNA sequences often have a very low coverage and a relatively high genotyping error rate, data are pseudohaploidized by sampling one allele per variant site. Individuals were assigned to continents and countries based on geographic coordinates or information on the country of sampling. The boundary between Europe and Asia was set at a longitude of 45°. Due to the sample size, only data from Asia and Europe were analyzed. Individuals from the Middle East (Israel, Lebanon, Jordan, and Syria) were excluded for these analyses.

To compare the average derived allele frequency (DAF) and standard deviation (SD) of the study variant with those of other SNPs and to obtain the quantile values, we retrieved all variants that are polymorphic in the European (number of variants = 1,094,500) and Asian (number of variants = 1,094,523) samples for which the derived allele could be inferred from comparison with chimpanzee positions (as available in the AADR).

For geographic representation, European regions were defined following the EuroVoc classification of sub-regions (Central and Eastern Europe, Southern Europe, Northern Europe, and Western Europe) (https://op.europa.eu/en/web/eu-vocabularies/concept-scheme/-/resource?uri=http://eurovoc.europa.eu/100277). Allele frequencies were calculated over time intervals and sub-regions.

Results

Evolutionary conservation and diversification of the UT module in the three domains of life

Because urea transportation is thought to have originated early in evolution [1], we searched for the presence of UTs in all three domains of life, meaning Eukaryotes, Archaea, and Bacteria. Using the UT Pfam profile we scanned the UniProt database and, after filtering,

we found more than 1330 proteins showing at least one UT domain (Table S2). As expected [1], multiple species exhibited gene duplications (Table S2): bony fish represented the most prominent group, showing both the largest number of species and the highest frequency of duplication events. Given the large number of UT module-containing proteins we obtained, we set out to explore their diversity in terms of domain architecture. We thus searched for domain information in these proteins. In vertebrates, no additional domains were identified in UTs, with the exclusion of a few fish proteins that display galactosyltransferase or immunoglobulin modules (Table S3). Conversely, in invertebrates, unicellular eukaryotes, and bacteria, UT domains were found to occur with other modules, the most common being ankyrin repeats (rotifers, arthropods, and protists), EF-hand domains (protists and bacteria), and M23 peptidase domains (bacteria) (Table S3). Most of these domains are located C-terminal to the UT (see Fig. 1A for representative structures). Thus, despite its overall conservation across the tree of life, the UT domain may contribute diverse functions or undergo specific regulation in some eukaryotic and prokaryotic species.

To better understand the evolutionary and phylogenetic relationships among UTs from different taxa, we next extracted the UT domain(s) from each protein sequence and constructed a phylogenetic tree (Fig. 1B). Results showed a clear separation between vertebrates and all other groups, whereas most non-vertebrate species did not clearly cluster based on taxonomy (Fig. 1B). The evolution and separation of fish domains (UT-C and UT-D) from all other vertebrate species was also clearly evident (Fig. 1B). Another noteworthy aspect is the positioning of Archaea domains. We found 9 Archaea genes carrying at least 2 UT modules, six having 3 of them. A closer inspection revealed that these domains are much shorter compared to the Bacteria/Eukaryotic ones, probably due to the splitting of the ancestral sequence.

Despite the lack of clear phylogenetic separation among taxa, we sought to evaluate the sequence conservation of the UT modules, with the aim to identify highly conserved positions or regions. For doing that, we selected a few representatives from the tree and we generated a multiple sequence alignment (Fig. S1). Due to their being split, Archaea domains were considered as a single unit. Despite the huge phylogenetic distances among the taxa, results showed a good general conservation, with 24 amino acids (8%) being conserved in more than 80% of the sequences. To evaluate the three-dimensional position of the conserved sites, we generated a model of the human UT-B domain (Fig. 1C). We found that most of the conserved residues are buried in the structure and define the channel-like pore responsible for urea transport.

Overall, these results indicate that the evolution of UTs has been dynamic, with frequent gene gains and losses, and acquisition of different domain architectures. Nonetheless, at the sequence level, a remarkable conservation of relevant amino acids is evident.

Positive selection drove the evolution of UT-B gene in specific mammalian orders

We next aimed to analyze the evolutionary history of UT genes in a more recent timeframe, namely during the evolution of mammals. We thus retrieved sequence information of the coding sequences of *SLC14A1* and *SLC14A2* orthologs from public databases, and we retained the five more populous orders: Primates, Carnivora, Chiroptera, Rodentia, and Artiodactyla (Table S1). To test whether positive selection has been driving the evolution of these genes, we applied likelihood ratio tests (LRTs) implemented in the PAML suite. Before running the LRTs, in order to avoid false inferences, we tested for recombination, which was not detected. LRT results indicated significant evidence of positive selection in Primates, Chiroptera, and Rodentia for *SLC14A1*, whereas no selection signature was found for *SLC14A2* (Table S4, Fig. 2A).

To gain further insight into the evolution of *SLC14A1* in mammals, we identified positively selected sites in the distinct orders where selection was detected (see methods). Several positively selected sites were identified in primates and bats, whereas only one was found in rodents (Table S4, Fig. 2A). Interestingly, some sites showed independent evidence of selection in more than one order (position 332 in rodents and humans, positions 204 and 283 in primates and bats), and additional sites were very close to each other, suggesting recurrent selective pressures acting on specific sites or regions. In general, positively selected sites were more abundant in the extracellular loops or transmembrane domains compared to the cytoplasmic region, with the exclusion of a cluster of three sites (Cys25, Gln26, and Gly27) in the intrinsically disordered intracellular N-terminal tail (Fig. 2A). One of these (Cys25) was previously shown to contribute to membrane localization [27]. Interestingly, some of the positively selected sites directly flanked positions that have instead been conserved for billions of years, as determined by our analysis across the domains of life, highlighting the different selective forces acting on this protein (Fig. 2A). Mapping of the selection signals onto the 3D structure of UT-B confirmed that several of them are located in the flexible extracellular loops, one of which carries the Asp280Asn that specifies the Kidd blood group antigens (Fig. 2A and B).

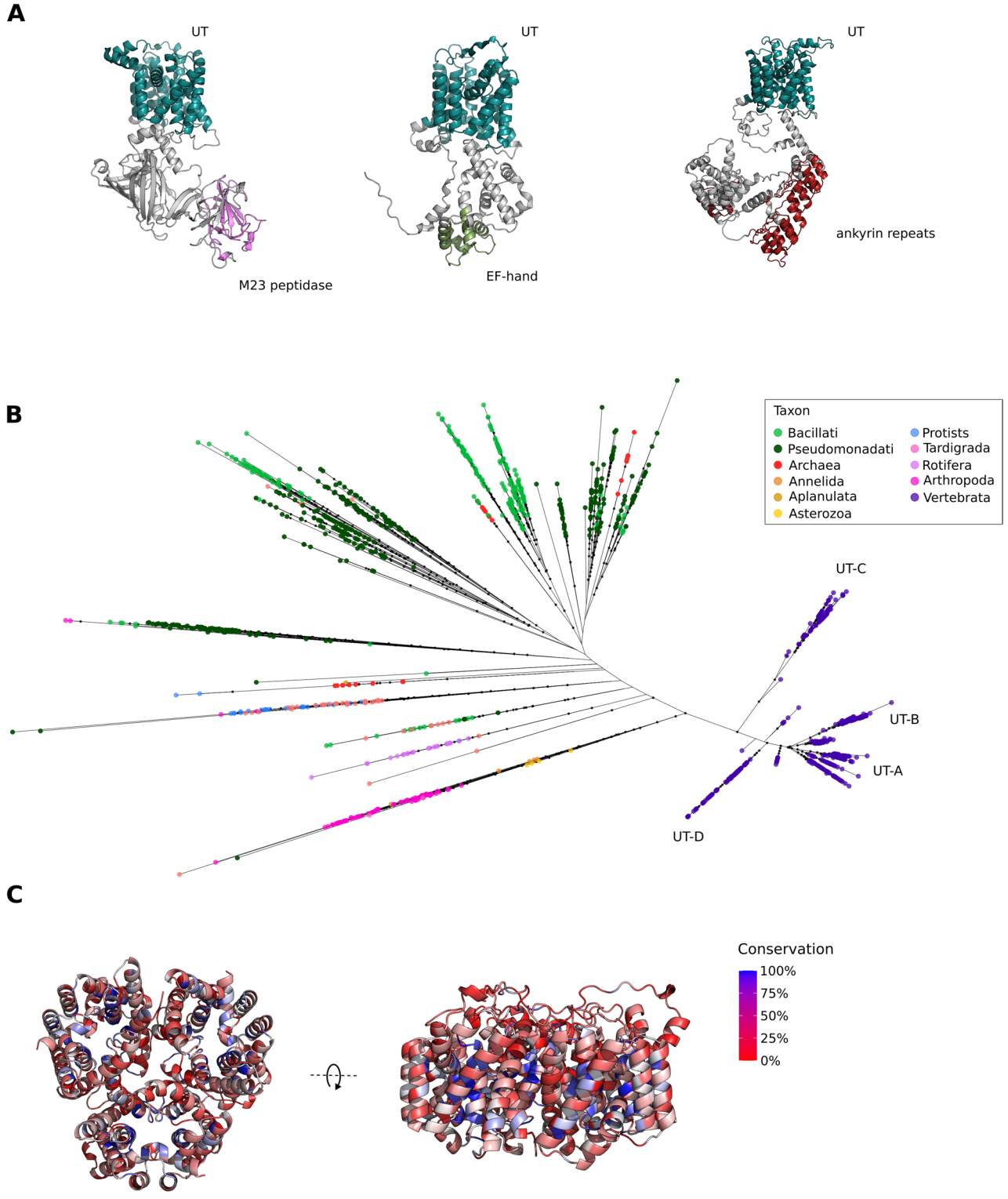

Fig. 1 Urea transporter conservation in the three domains of life (A) Structural model prediction of proteins that carry functional domains associated with the UT domain (deep teal); three representative proteins are shown (peptidase M23, UniProt ID: A0A399T3G1; EF-hand, UniProt ID:A0A261KJD6; ankyrin repeat, UniProt ID: A0A814UYH3). (B) Phylogenetic tree for the UT domain in Eukaryotes, Bacteria, and Archaea. Species were coloured as shown in the inset legend to better highlight the conservation and distribution of the domain. The vertebrate repertoire of UTs is also shown. Tree nodes with a bootstrap support higher than 80 are indicated with a black dot. (C) Structural model of the human UT-B domain trimer colored based on amino acid conservation in different taxa

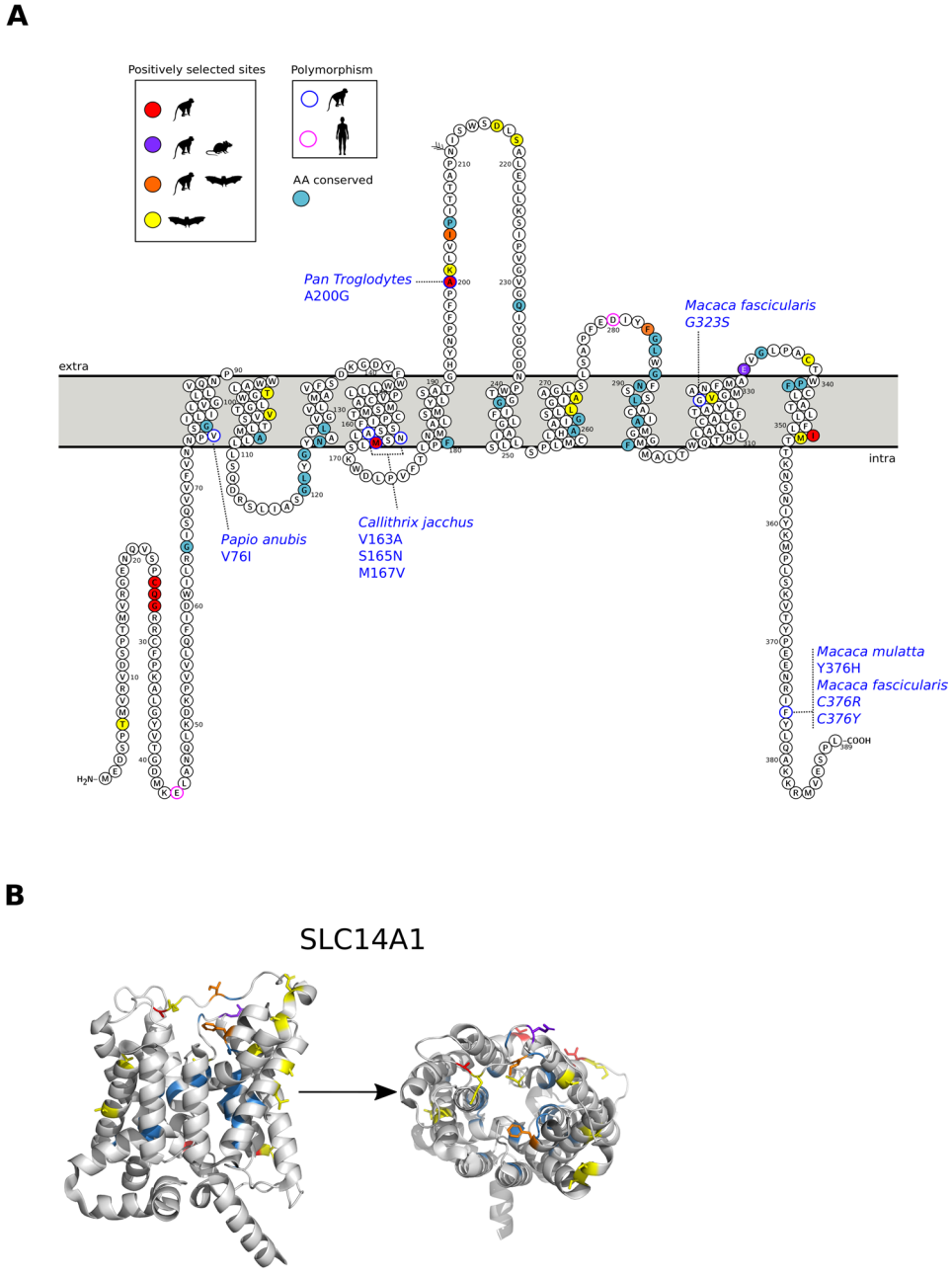

Fig. 2 Positive selection in mammalian orders. **(A)** Positively selected sites in different mammalian orders are highlighted on the SLC14A1 (UT-B) protein topology (see inset legend). Amino acids conserved across the three domains of life are in light blue. Polymorphic variants in human and non-human primates are circled in magenta and in blue, respectively. The secondary structure of human UT-B1 was generated using Protter (http://wlab.ethz.ch/protter/) based on Chi et al. data [2]. **(B)** Crystal structure of human SLC14A1 (UT-B, PDB ID: 8BLP). Positively selected sites are shown as sticks and highlighted as in panel (A); conserved amino acids are in blue

Ongoing selection at the *SLC14A1* gene in primate populations

Because of the abundant signatures of fast evolution in the primate *SLC14A1* gene, and due to the previously suggested action of balancing selection acting on the human Asp280Asn variant (see below) [28], we sought to determine whether polymorphic nonsynonymous substitutions are maintained in non-human primate populations. We thus interrogated the Genome variation Map database and retrieved nonsynonymous variants with a frequency higher than 0.10 in chimpanzee, macaque (rhesus and crab-eating), baboon, and marmoset populations. We detected common variants in all of them (Table S5). In the case of chimpanzees and marmoset two of the polymorphic positions (Ala200Gly and Met167Val, respectively) correspond to sites that were

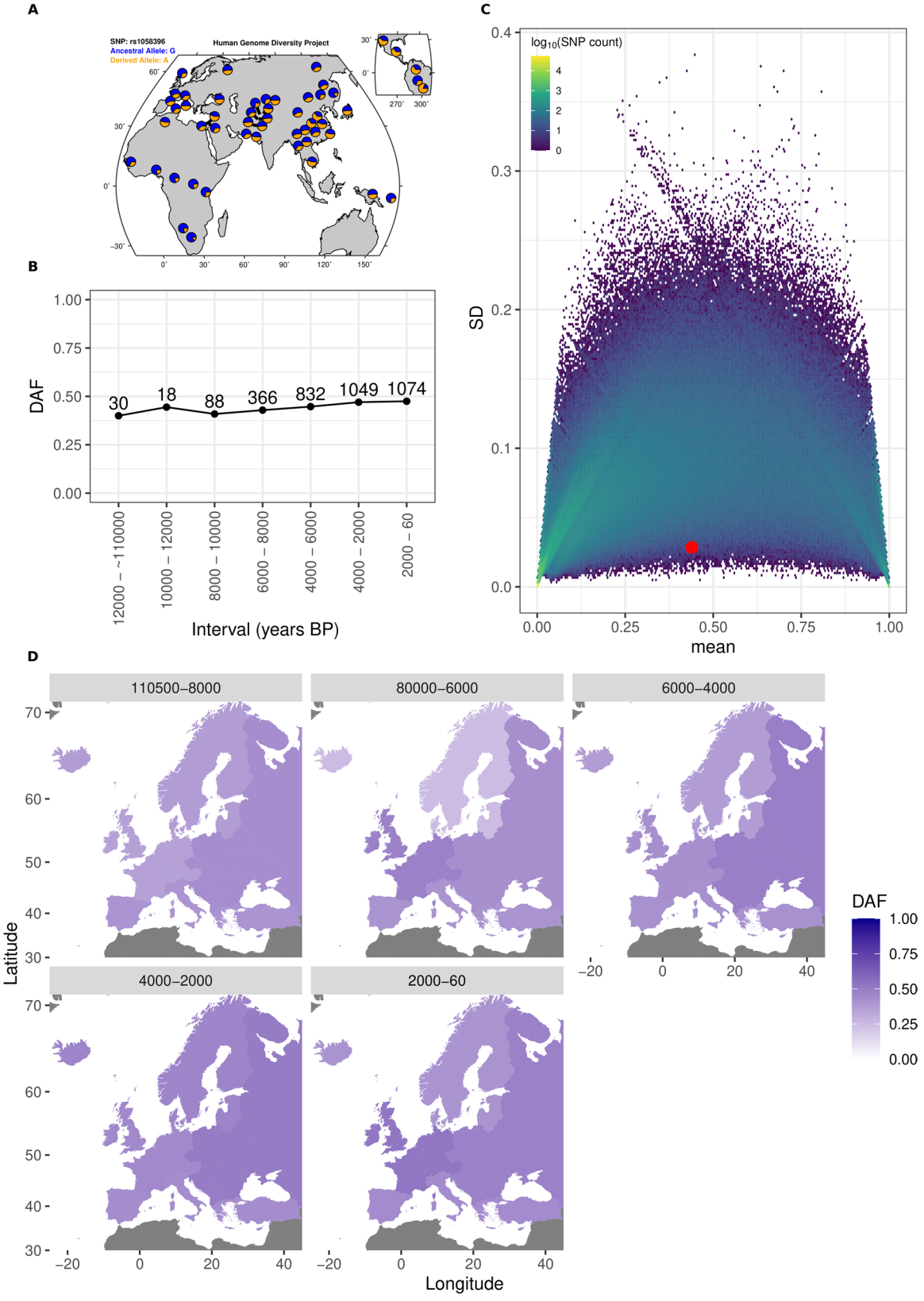

Fig. 3 (See legend on next page.)

(See figure on previous page.)

Fig. 3 Temporal dynamics of the Asp280Asn variant in Europe. **(A)** Present-day allele frequency of the rs1058396 variant in 54 human populations of the Human Genome Diversity Project [60] **(B)** Frequency of the derived allele (280Asn) over seven time intervals covering a time transect from the Paleolithic to the near past (contemporary samples are not included). The number of individuals is reported above each data point. **(C)** Density plot of the mean DAF and SD for ~ 1 million variants. For visualization purposes, the data space was divided into hexagonal bins that are color-coded according to SNP count. The rs1058396 SNP is represented as a red dot. **(D)** Frequency of the derived allele over different time periods in four European sub-regions: Central and Eastern Europe, Southern Europe, Northern Europe, and Western Europe

targeted by positive selection during primate evolution, suggesting ongoing selection (Fig. 2A). Interestingly, in rhesus macaques and crab-eating macaques, the same position (Phe376 in the human protein) is polymorphic and two distinct amino acids are found in the latter, due to changes in the first or second codon position. This suggests that the 376 residue, located in the intracellular C-terminal loop, is subject to diversifying selection in macaque populations.

Ancient DNA data support the maintenance of the SLC14A1 Asp280Asn variant as a balanced polymorphism in Europe

As introduced above, the SLC14A1 Asp280Asn variant (rs1058396) determines the *Jka/Jkb* antigens of the Kidd blood group system. Residue 280 is located on an extracellular loop close to a site that was positively selected in bats and primates (Fig. 2A). In contemporary human populations, the derived allele (280Asn) is very common in Europe and Asia (~ 50% frequency or higher), while its frequency is around 25% in Africa (Fig. 3). We have previously shown that the variant has been maintained in human populations for at least two million years as a result of long-standing balancing selection [28]. The latter refers to different evolutionary processes that maintain polymorphisms in a population. Thus, balancing selection may result from heterozygote advantage, frequency-dependent selection, variable environmental conditions across time or space, or other forces [29]. However, balancing selection is often transient in time, raising the possibility that the signatures we detected were the results of selective events that occurred in a distant past. Also, the increased genetic diversity that is a hallmark of balancing selection can in principle also be caused by population structure [29]. It has thus been suggested that temporally sampled data are ideally suited to detect or confirm evidence of balancing selection, as the frequency of alleles evolving under this selective regime is expected to be relatively stable over time [30]. Nonetheless, this expectation has never been tested using real-world human genetic data.

To explore the dynamics of the rs1058396 variant through time, we took advantage of the fact that it is covered in genotype data from a large collection of ancient DNA samples. Most of such samples derive from individuals who lived in Europe and we thus initially focused on this continent [26]. From 6,480 ancient European genomes in the Allen Ancient DNA Repository [26], 3,457 genomes had genotype calls at the position of interest. In the time frame ranging from ~ 110,000 to 12,000 years before present (BP), data are sparse, with only 30 individuals having information for the rs1058396 variant. Starting from 12,000 years BP, information is more abundant. Thus, to analyze the temporal dynamics of rs1058396, we divided the genomes based on their estimated dating into 7 time periods (2,000 years each, except for the oldest, ranging from ~ 110,000 to 12,000 years BP), allowing analysis of allele frequencies over a time transect from the Paleolithic to present days. Results indicated that the frequency of the derived 280Asn allele (rs1058396-A) has remained remarkably stable across time periods in the past ~ 110,000 years. To compare the trajectory of the rs1058396 variant with those of other SNPs genotyped in the dataset, we calculated the mean and standard deviation (SD) of the derived allele frequency (DAF) across the seven time periods. The same calculation was performed for more than 1 million variant positions where the derived allele could be inferred based on the chimpanzee allele status (see methods). Results showed a tendency for variants with intermediate average DAF to have higher standard deviations than variants at the extremes of the DAF spectrum. Despite its intermediate DAF (mean over time periods = 0.439), the SD of rs1058396 was very low (SD = 0.028). In order to determine whether, given its average DAF, the SD of the variant is indeed lower than expected by chance, we applied an empirical approach. Specifically, we binned all variants in 1,000 average DAF intervals and we calculated the SDs in each bin. We then calculated the quantile corresponding to the SD of rs1058396, which resulted equal to 0.0086. This result clearly indicates that the DAF of the variant under study has remained unexpectedly stable over the time transect.

We next asked whether the frequency of the Asp280Asn variant has also remained stable over different geographic locations in Europe. We thus divided the genomic data based on time periods and European subregions (see methods). Calculation of the average subregion-wise DAFs indicated limited differences, with the DAF always being within the 0.25–0.53 range. This analysis should be regarded as qualitative, as the sample size is limited and different in the distinct subregion/time groups. Overall, these results are consistent with balancing selection having maintained the rs1058396 at

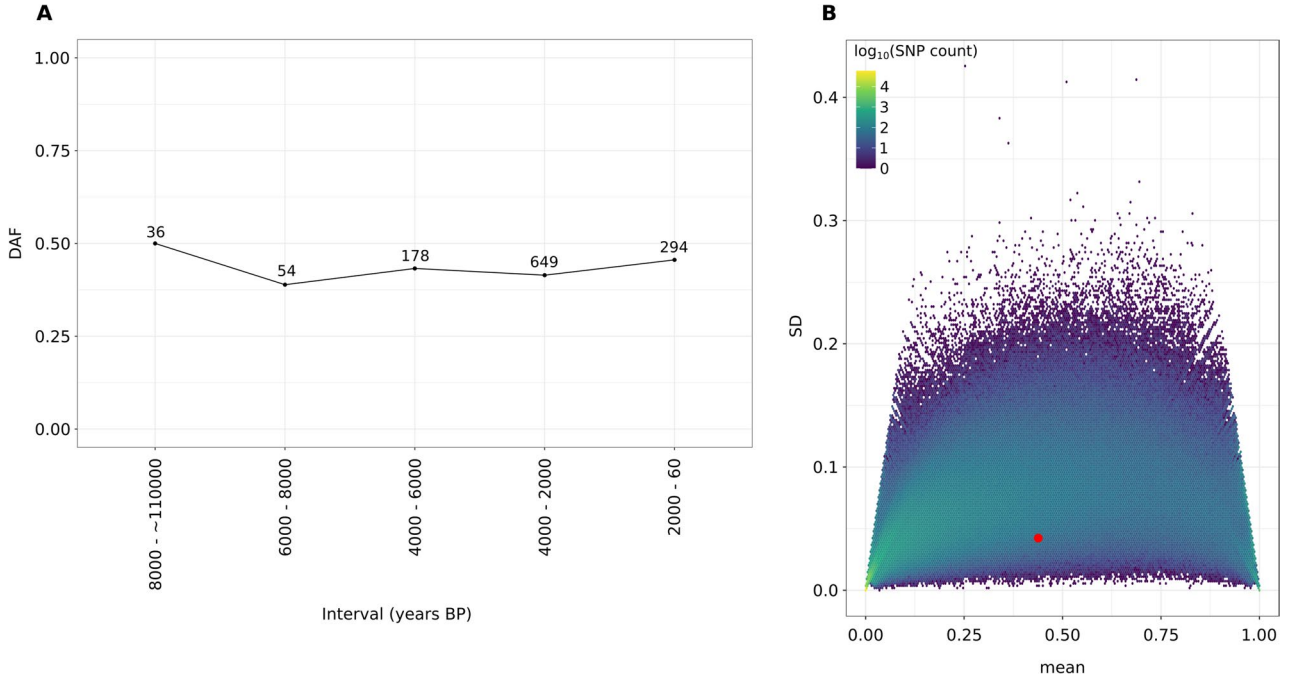

Fig. 4 Temporal dynamics of the Asp280Asn variant in Asia. **(A)** Frequency of the derived allele (280Asn) over five time intervals covering a time transect from the Paleolithic to the near past (contemporary samples are not included). The number of individuals is reported above each data point. **(B)** Density plot of the mean DAF and SD for ~1 million variants. Visualization is as described in the legend of Fig. 3. The rs1058396 SNP is represented as a red dot

intermediate frequency across Europe from the Paleolithic to the present.

Evidence of balancing selection acting on rs1058396 in Asia

We next sought to determine whether the Asp280Asn variant has also evolved under a balancing selection regime in Asia, the second continent where ancient DNA data are most aboundant. In the Allen Ancient DNA Repository, Asian 1,211 genomes had genotype calls for rs1058396. In the time period ranging from ~110,000 to 8,000 years BP, only 36 genotypes were available, thus these time intervals were merged. Calculation of the average DAF over the 5 time periods showed a general stability, with only a slight initial increase (Fig. 4).

As in the case of the European data, we analyzed the across-interval average frequency and SD of the variant together with those of ~1 million polymorphic SNPs in samples from this continent with available ancestral state information. We obtained similar results as in the European samples, with variants with intermediate average DAF tending to have higher SDs (Fig. 4). The SD of rs1058396 was in the low range given its DAF and binning into DAF classes indicated that it corresponds to the 0.087 quantile.

In summary, both in Europe and in Asia, the 280Asn allele has remained significantly more stable in frequency than the majority of variants in its DAF class. The observation of the same pattern in two independent sets of data samples from different continents strongly supports the view that the stability of the Asp280Asn variant is not due to drift but rather to its having been a target of balancing selection in the past 110,000 years.

Discussion

Our study unveils key aspects of the molecular evolution of UTs at different time scales, highlighting deep conservation across the three domains of life, but also positive selection in mammals, as well as balancing selection of the Kidd antigen variant in human populations along a time transect from the Paleolithic to the present.

Because a comprehensive analysis of the representation and domain architecture of UT proteins was missing, we systematically searched for UT modules across the three domains of life. We found a wide diversity of UT domains in prokaryotic proteins and some protist UT domains clustered with bacterial ones, possibly suggesting recent horizontal transfer events. A notable finding of the UT module search across the tree of life was the identification of diverse domain architectures. These were particularly common in invertebrates, protists, and bacteria and most likely resulted from domain shuffling, which is a major driver of protein evolution [31]. For instance, all UT-containing proteins in rotifers also present ankyrin domains, and EF-hand modules are common in stramenopiles (protists). Ankyrin repeats were shown to modulate the activity of metazoan ion channels of the transient receptor potential family [32–34], whereas

EF-hand motifs can regulate the activation/inactivation of some calcium channels [35, 36]. These observations suggest that, in some species, UTs may be regulated in their function by Ca++ or other stimuli.

The early origin of UT domains also allowed the identification of specific residues that have remained conserved across billions of years of evolution. At the same time, though, we found evidence that UT-B has undergone rapid evolution at specific sites as the result of positive selection in mammals. Some such sites impinge on the same region where highly conserved residues are located, highlighting the co-occurrence of different selective forces at nearby sites. We found a considerable number of sites targeted by selection in primates and chiroptera. It is clearly difficult to speculate on the possible selective forces underlying the signatures we detected in these mammalian orders. Because UT-B is expressed on the erythrocyte membrane, one possibility is that selection operated as a response to protozoan blood parasites (orders Haemosporida and Piroplasmida). Bats and primates were shown to host a wide diversity of haemosporidians [37–43], whereas piroplasms are widely distributed in mammals, including humans, with the seroprevalence for *Babesia* spp. being around 2% in the general population in Europe [44–46]. In fact, piroplasms and haemosporidians were previously suggested to have represented a strong selective pressure in mammals [47]. Unfortunately, little is known about the molecular mechanisms that these parasites use to invade blood cells, but adhesion to surface molecules such as UT-B, may have a role. Likewise, plasma membrane proteins or alterations in the intracellular concentration of urea may impact the ability of bacteria such as those in the *Bartonella* and *Anaplasma* genera to establish intraerythrocytic bacteraemia. In this respect, it is worth noting that bats contributed to the radiation of mammal-associated *Bartonella* species and display a long-standing association with these pathogens [48]. Experimental investigations will be necessary to determine whether SLC14A1 functions as a receptor for blood pathogens and if the positively selected sites impact its functionality in terms of intracellular urea concentration or erythrocyte membrane structure.

In this work, we also provide direct evidence for the temporal stability of Jka/Jkb polymorphism over approximately 110,000 years in Europe and Asia. While we previously showed that this variant was a target of balancing selection [28], we here use ancient DNA data to track its allele frequency through deep time. This novel application of paleogenomics confirms that the polymorphism was maintained at stable intermediate frequencies and that this is unlikely due to chance alone. Specifically, we compared the allele frequency trajectory to those of more than 1 million variants from the same time-binned intervals, which provided the background empirical distribution. Whereas this approach does not adjust for temporal shifts in ancestry composition or population replacements, which were common in Eurasia [49], it relies on the concept that demographic changes affect the whole genome and that a large number of variants can provide a neutral expectation. Moreover, the stability of rs1058396 frequency over time, European sub-regions, and across continents is compelling. In this respect, we wish to mention that the boundary between Europe and Asia was set at 45° longitude, an arbitrary divide. Whereas this is common practice to compare allele trajectories between continents (e.g [50, 51]), using a fixed geographic split does not account for the complex patterns of genetic and cultural overlap across Eurasia. This represents a limitation of our study, although we consider that the use of empirical distributions may alleviate potential biases. Another limitation of our approach, which is inherent to paleogenomic analyses, is the use of pseudohaploid data, which clearly hamper analyses based on heterozygosity and thus affects interpretations of balancing selection. In terms of limitations, we should also add that, with only 30 European individuals and the merged early Asian intervals, the oldest time period estimates may carry large uncertainty. However, the later Holocene bins have larger sample sizes and stability across these intervals is reproducible.

Whereas these results are clearly consistent with a role for balancing selection in maintaining the Asp280Asn variant at intermediate frequency, the underlying selective pressure remains unknown. The amino acid change occurs in an extracellular loop that also carries a site that was positively selected independently in bats and primates, as well as some highly conserved residues. No human phenotype has been associated with Jka or Jkb carriers and even Jk-null individuals only suffer minor urine concentration defects [6]. However variants in *SLC14A1* have been associated with urinary bladder cancer, reticulocyte count, mean corpuscular hemoglobin concentration, and other hematological traits [52–59]. Whereas cancer is an unlikely driver of the selection signals we detected, due to its usual onset after reproductive age, changes in erythrocyte-related traits might confer differential susceptibility to blood pathogens, including the ones we mentioned above. The testing of these hypotheses will need epidemiological and/or experimental analysis. Despite the missing underlying causative explanation, our data show how paleogenomics can provide information on the selective processes in humans, not only limited to positive, directional selection, but also to other selective regimes, including balancing selection.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13062-026-00741-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
Not applicable.

Author contributions
Conceptualization: M.S., R.C., U.P.; Methodology: M.S., R.C., U.P., D.F.; Investigation: M.S., R.C, U.P., D.F., A.M.; Formal analysis: M.S., R.C., U.P., D.F., A.M.; Visualization: R.C., U.P., D.F., A.M.; Funding acquisition: R.C.; Supervision: M.S.; Writing—original draft: M.S., R.C., U.P., D.F.; Writing—review and editing: M.S., R.C., U.P., D.F., A.M.

Funding
This work was supported by the Italian Ministry of Health ("Ricerca Corrente" to RC).

Data availability
The data underlying this article are available in the article and in its online supplementary material. Multiple sequence alignments and trees are available in Zenodo, at https://doi.org/10.5281/zenodo.17521170.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 January 2026 / Accepted: 6 February 2026
Published online: 11 February 2026

References

1. LeMoine CMR, Walsh PJ. Evolution of urea transporters in vertebrates: adaptation to urea's multiple roles and metabolic sources. In: Podrabsky JE, Stillman JH, Tomanek L, editors. J Exp Biol. 2015;218:1936–45. https://doi.org/10.1242/jeb.114223.
2. Chi G, Dietz L, Tang H, Snee M, Scacioc A, Wang D, et al. Structural characterization of human Urea transporters UT-A and UT-B and their Inhibition. Sci Adv. 2023;9:eadg8229. https://doi.org/10.1126/sciadv.adg8229.
3. Shayakul C, Clémençon B, Hediger MA. The Urea transporter family (SLC14): Physiological, pathological and structural aspects. Mol Aspects Med. 2013;34:313–22. https://doi.org/10.1016/j.mam.2012.12.003.
4. Geng X, Zhang S, He J, Ma A, Li Y, Li M, et al. The Urea transporter UT-A1 plays a predominant role in a Urea-dependent urine-concentrating mechanism. J Biol Chem. 2020;295:9893–900. https://doi.org/10.1074/jbc.RA120.013628.
5. Yang B, Li X, Guo L, Meng Y, Dong Z, Zhao X. Extrarenal phenotypes of the UT-B knockout mouse. In: Yang B, Sands JM, editors. Urea transporters [Internet]. Dordrecht: Springer Netherlands; 2014. pp. 153–64. [cited 2025 Oct 29]. https://doi.org/10.1007/978-94-017-9343-8_10.
6. Lawicki S, Covin RB, Powers AA. The Kidd (JK) blood group system. Transfus Med Rev. 2017;31:165–72. https://doi.org/10.1016/j.tmrv.2016.10.003.
7. Eddy SR, Accelerated Profile HMM, Searches. Pearson WR, editors. PLoS Comput Biol. 2011;7:e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
8. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9. https://doi.org/10.1093/nar/gkaa913.
9. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80. https://doi.org/10.1093/molbev/mst010.
10. Wong T, Ly-Trong N, Ren H, Baños H, Roger A, Susko E, et al. IQ-TREE 3: phylogenomic inference software using complex evolutionary models. Life Sci. 2025 [cited 2025 Oct 30]. https://doi.org/10.32942/X2P62N.
11. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9. https://doi.org/10.1038/nmeth.4285.
12. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35:518–22. https://doi.org/10.1093/molbev/msx281.
13. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature. 2024;630:493–500. https://doi.org/10.1038/s41586-024-07487-w.
14. Schrödinger LLC. The PyMOL molecular graphics system, version 2.0 Schrödinger. Google Scholar There is no corresponding record for this reference; 2017.
15. Wernersson R. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31:3537–9. https://doi.org/10.1093/nar/gkg609.
16. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods in molecular biology. (Clifton NJ). 2009;537:113–37. https://doi.org/10.1007/978-1-59745-251-9_6.
17. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination detection. Bioinformatics. 2006;22:3096–8. https://doi.org/10.1093/bioinformatics/btl474.
18. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91. https://doi.org/10.1093/molbev/msm088.
19. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosciences: CABIOS. 1997;13:555–6.
20. Anisimova M, Bielawski JP, Yang Z. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 2002;19:950–8. https://doi.org/10.1093/oxfordjournals.molbev.a004152.
21. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30:1196–205. https://doi.org/10.1093/molbev/mst030.
22. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8:e1002764. https://doi.org/10.1371/journal.pgen.1002764.
23. Kosakovsky Pond SL, Frost SDW. Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005;22:1208–22. https://doi.org/10.1093/molbev/msi105.
24. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9. https://doi.org/10.1093/bioinformatics/bti079.
25. Li C, Tian D, Tang B, Liu X, Teng X, Zhao W, et al. Genome variation map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res. 2021;49:D1186–91. https://doi.org/10.1093/nar/gkaa1005.
26. Mallick S, Micco A, Mah M, Ringbauer H, Lazaridis I, Olalde I, et al. The Allen ancient DNA resource (AADR) a curated compendium of ancient human genomes. Sci Data. 2024;11:182. https://doi.org/10.1038/s41597-024-03031-7.
27. Lucien N, Sidoux-Walter F, Roudier N, Ripoche P, Huet M, Trinh-Trang-Tan M-M, et al. Antigenic and functional properties of the human red blood cell Urea transporter hUT-B1. J Biol Chem. 2002;277:34101–8. https://doi.org/10.1074/jbc.M205073200.
28. Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, et al. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res. 2009;19:199–212. https://doi.org/10.1101/gr.082768.108.
29. Fijarczyk A, Babik W. Detecting balancing selection in genomes: limits and prospects. Mol Ecol. 2015;24:3529–45. https://doi.org/10.1111/mec.13226.
30. Dehasque M, Ávila-Arcos MC, Díez-del-Molino D, Fumagalli M, Guschanski K, Lorenzen ED, et al. Inference of natural selection from ancient DNA. Evol Lett. 2020;4:94–108. https://doi.org/10.1002/evl3.165.
31. Chothia C, Gough J. Genomic and structural aspects of protein evolution. Biochem J. 2009;419:15–28. https://doi.org/10.1042/BJ20090122.
32. Zhang W, Cheng LE, Kittelmann M, Li J, Petkovic M, Cheng T, et al. Ankyrin repeats convey force to gate the NOMPC mechanotransduction channel. Cell. 2015;162:1391–403. https://doi.org/10.1016/j.cell.2015.08.024.

33. Howard J, Bechstedt S, Hypothesis. A helix of Ankyrin repeats of the NOMPC-TRP ion channel is the gating spring of mechanoreceptors. Curr Biol. 2004;14:R224–6. https://doi.org/10.1016/j.cub.2004.02.050.
34. Hori S, Tateyama M, Shirai T, Kubo Y, Saitoh O. Two single-point mutations in Ankyrin repeat one drastically change the threshold temperature of TRPV1. Nat Commun. 2023;14:2415. https://doi.org/10.1038/s41467-023-38051-1.
35. Petri ET, Ćelić A, Kennedy SD, Ehrlich BE, Boggon TJ, Hodsdon ME. Structure of the EF-hand domain of polycystin-2 suggests a mechanism for Ca^{2+} -dependent regulation of polycystin-2 channel activity. Proc Natl Acad Sci USA. 2010;107:9176–81. https://doi.org/10.1073/pnas.0912295107.
36. Peterson BZ, Lee JS, Mulle JG, Wang Y, De Leon M, Yue DT. Critical determinants of Ca2+-Dependent inactivation within an EF-Hand motif of L-Type Ca2 + Channels. Biophys J. 2000;78:1906–20. https://doi.org/10.1016/S0006-3495(00)76739-7.
37. Ayouba A, Mouacha F, Learn GH, Mpoudi-Ngole E, Rayner JC, Sharp PM, et al. Ubiquitous hepatocystis infections, but no evidence of plasmodium falciparum-like malaria parasites in wild greater spot-nosed monkeys (Cercopithecus nictitans). Int J Parasitol. 2012;42:709–13. https://doi.org/10.1016/j.ijpara.2012.05.004.
38. Thurber MI, Ghai RR, Hyeroba D, Weny G, Tumukunde A, Chapman CA, et al. Co-infection and cross-species transmission of divergent hepatocystis lineages in a wild African primate community. Int J Parasitol. 2013;43:613–9. https://doi.org/10.1016/j.ijpara.2013.03.002.
39. Schaer J, Perkins SL, Decher J, Leendertz FH, Fahr J, Weber N, et al. High diversity of West African Bat malaria parasites and a tight link with rodent plasmodium taxa. Proc Natl Acad Sci USA. 2013;110:17415–9. https://doi.org/10.1073/pnas.1311016110.
40. Lutz HL, Patterson BD, Kerbis Peterhans JC, Stanley WT, Webala PW, Gnoske TP, et al. Diverse sampling of East African haemosporidians reveals chiropteran origin of malaria parasites in primates and rodents. Mol Phylogenet Evol. 2016;99:7–15. https://doi.org/10.1016/j.ympev.2016.03.004.
41. Schaer J, Perkins SL, Ejotre I, Vodzak ME, Matuschewski K, Reeder DM. Epauletted fruit bats display exceptionally high infections with a hepatocystis species complex in South Sudan. Sci Rep. 2017;7:6928. https://doi.org/10.1038/s41598-017-07093-z.
42. Boundenga L, Ngoubangoye B, Mombo IM, Tsoubmou TA, Renaud F, Rougeron V, et al. Extensive diversity of malaria parasites Circulating in central African bats and monkeys. Ecol Evol. 2018;8:10578–86. https://doi.org/10.1002/ece3.4539.
43. Haffener PE, Hopson HD, Herbert-Mainero A, Ramirez A, Leffler EM. Phylogenetics and genomic variation of Hepatocystis isolated from shotgun sequencing of wild primate hosts. In: Sharp PM, editor. PLoS Pathog. 2025;21:e1013240. https://doi.org/10.1371/journal.ppat.1013240.
44. Jalovecka M, Sojka D, Ascencio M, Schnittger L. Babesia life Cycle – When phylogeny Meets biology. Trends Parasitol. 2019;35:356–68. https://doi.org/10.1016/j.pt.2019.01.007.
45. Hunfeld K-P, Lambert A, Kampen H, Albert S, Epe C, Brade V, et al. Seroprevalence of *Babesia* infections in humans exposed to ticks in Midwestern Germany. J Clin Microbiol. 2002;40:2431–6. https://doi.org/10.1128/JCM.40.7.2431-2436.2002.
46. Svensson J, Hunfeld K-P, Persson KEM. High Seroprevalence of Babesia antibodies among borrelia burgdorferi-infected humans in Sweden. Ticks Tick-borne Dis. 2019;10:186–90. https://doi.org/10.1016/j.ttbdis.2018.10.007.
47. Ebel ER, Telis N, Venkataram S, Petrov DA, Enard D. High rate of adaptation of mammalian proteins that interact with plasmodium and related parasites. PLoS Genet. 2017;13:e1007023. https://doi.org/10.1371/journal.pgen.1007023.
48. McKee CD, Bai Y, Webb CT, Kosoy MY. Bats are key hosts in the radiation of mammal-associated Bartonella bacteria. Infect Genet Evol. 2021;89:104719. https://doi.org/10.1016/j.meegid.2021.104719.
49. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. Nature. 2017;541:302–10. https://doi.org/10.1038/nature21347.
50. Nordgren J, Ågren R, Hu DZ, Neijd M, Childebayeva A, Prüfer K, et al. Natural selection of a virus-protective FUT2 variant following the transition to agriculture. In: Ávila-Arcos MC, editor. Mol Biol Evol. 2025;42:msaf243. https://doi.org/10.1093/molbev/msaf243.
51. Segurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, et al. Why and when was lactase persistence selected for? Insights from central Asian herders and ancient DNA. PLoS Biol. 2020;18:e3000742. https://doi.org/10.1371/journal.pbio.3000742.
52. Verma A, Huffman JE, Rodriguez A, Conery M, Liu M, Ho Y-L, et al. Diversity and scale: genetic architecture of 2068 traits in the VA million veteran program. Science. 2024;385:eadj1182. https://doi.org/10.1126/science.adj1182.
53. Garcia-Closas M, Ye Y, Rothman N, Figueroa JD, Malats N, Dinney CP, et al. A genome-wide association study of bladder cancer identifies a new susceptibility locus within SLC14A1, a Urea transporter gene on chromosome 18q12.3. Hum Mol Genet. 2011;20:4282–9. https://doi.org/10.1093/hmg/ddr342.
54. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The polygenic and Monogenic basis of blood traits and diseases. Cell. 2020;182:1214–e123111. https://doi.org/10.1016/j.cell.2020.08.008.
55. Loya H, Kalantzis G, Cooper F, Palamara PF. A scalable variational inference approach for increased mixed-model association power. Nat Genet. 2025;57:461–8. https://doi.org/10.1038/s41588-024-02044-7.
56. Lee C-J, Chen T-H, Lim AMW, Chang C-C, Sie J-J, Chen P-L, et al. Phenome-wide analysis of Taiwan biobank reveals novel glycemia-related loci and genetic risks for diabetes. Commun Biol. 2022;5:1175. https://doi.org/10.1038/s42003-022-04168-0.
57. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50:390–400. https://doi.org/10.1038/s41588-018-0047-6.
58. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021;53:1415–24. https://doi.org/10.1038/s41588-021-00931-x.
59. Jee YH, Wang Y, Jung KJ, Lee J-Y, Kimm H, Duan R, et al. Genome-wide association studies in a large Korean cohort identify quantitative trait loci for 36 traits and illuminate their genetic architectures. Nat Commun. 2025;16:4935. https://doi.org/10.1038/s41467-025-59950-5.
60. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367:eaay5012. https://doi.org/10.1126/science.aay5012.

Publisher's note