

A readthrough-permissive stop context and C-terminal constraint distinguish AQP4ex from its orthodox aquaporin paralogues

Satyam Sangeet

satyam.sangeet@sydney.edu.au

University of Sydney <https://orcid.org/0000-0002-8274-6775>

Research Article

Keywords: AQP4ex, Comparative Genomics, Evolution

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11323608/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

AQP4ex is an isoform of aquaporin-4 (AQP4) produced by readthrough of the UGA stop codon that carries a C-terminal extension encoded in the 3' untranslated region (UTR). It is required to anchor AQP4 at perivascular astrocyte endfeet. Binding assays have mapped the interaction of the human peptide with α -syntrophin, but whether that surface is evolutionarily conserved and whether any of the remaining positions are constrained, is unknown. Here we use evolutionary constraint as a residue level map across AQP4 orthologues from 759 vertebrate species, 697 of them with a 3'-UTR record. The readthrough-permissive UGA-C context was present in 99.3% of these 697 species. Codon models identified purifying selection at 16 of 28 extension positions (12 after Bonferroni correction), concentrated in a C-terminal segment ending in KETDLL, which is retained in 72.8% of species. Synonymous rates were suppressed across the extension, indicating constraint on the nucleotide sequence in addition to the encoded peptide. None of five orthodox aquaporin paralogues showed this combination. AQP5 retains the UGA-C context in 58.6% of species, but shows no purifying selection on downstream frame, suggesting that a permissive context alone does not maintain a constrained extension. These findings identify the C terminus of AQP4ex as an evolutionarily conserved candidate anchoring interface.

Introduction

Aquaporin-4 (AQP4) is the main water channel of the central nervous system (CNS) and is highly expressed in astrocytes at the blood–brain barrier¹. In the brain, AQP4 is concentrated in astrocytic endfeet surrounding cerebral blood vessels^{2,3} where it is anchored within a specialised perivascular membrane domain and contributes to water exchange at the neurovascular interface^{4,5}. This perivascular polarisation distinguishes AQP4 from the more diffuse distribution reported for several other aquaporins in the CNS^{6,7}. Despite the continued presence of AQP4 protein and preservation of its intrinsic water permeability⁸, the loss of AQP4 polarisation has been associated with impaired glymphatic transport and reduced clearance of interstitial solutes^{9–12}, including amyloid- β . Understanding the molecular determinants of AQP4 localisation is therefore important for defining how AQP4 contributes to brain water homeostasis and clearance.

AQP4 is expressed as four principal isoforms generated by alternative translation initiation and stop-codon readthrough¹³. The canonical M1 and M23 isoforms differ in their N-terminal sequence and assemble into tetramers¹⁴. M23-rich tetramers, together with M1/M23 hetero-tetramers, can further organise into supramolecular orthogonal arrays of particles (OAPs) in the plasma membrane^{15–17}. In addition, translation of AQP4 mRNA can continue beyond the canonical termination codon, producing the C-terminally extended isoforms M1ex and M23ex, collectively referred to as AQP4ex^{18–20}. These isoforms carry an additional 29 residues. Among these additional residues, there is an amino acid incorporated at the decoded stop codon, followed by 28 residues encoded by the sequence downstream of it. The extended isoforms are expressed in the mammalian CNS²¹ and are enriched at astrocytic perivascular endfeet¹⁸, where the C-terminal extension contributes to the anchoring of AQP4 assemblies through interactions involving α -syntrophin within the dystrophin-associated protein complex¹⁸.

The functional importance of AQP4ex was investigated using a CRISPR/Cas9-targeted stop-codon modification that selectively prevents translational readthrough²⁰. In this model, conversion of the endogenous UGA termination codon to UAA, together with the introduction of additional downstream stop codons, eliminated AQP4ex production without removing the AQP4 transcript or the canonical AQP4 open reading frame. Loss of AQP4ex did not reduce total AQP4 expression or astrocytic osmotic water permeability, but it abolished the characteristic enrichment of AQP4 at astrocyte endfeet²⁰. The study also showed that the loss of AQP4ex reduced the size of AQP4-containing OAPs. It also reduced perivascular expression of α -syntrophin, a protein associated with AQP4 anchoring. These findings support a primary role for AQP4ex in the perivascular localisation and anchoring of AQP4 rather than in bulk astrocytic osmotic water permeability²⁰. But the requirement for AQP4ex does not establish how functional information is distributed across the extension. Experiments that eliminate translation of the entire extension cannot distinguish whether its activity depends on most of the peptide sequence, a limited number of residues, a short-conserved motif or a sequence-independent property such as peptide length, flexibility or spacing.

Comparative genomic analyses have provided evidence that some translational readthrough products are maintained by natural selection. Several readthrough regions remain under purifying selection, although generally at lower levels than canonical coding regions^{22,23}, supporting the view that these regions can encode biologically relevant peptides. Comparative analyses of AQP4ex have likewise identified conserved sequence features within the C-terminal extension. A previous analysis of AQP4ex sequences from a limited number of species identified an RXXS phosphorylation motif, including a serine residue that was subsequently investigated as a phosphorylation site in human AQP4ex²⁴. However, that analysis was based on limited taxonomic sampling and did not provide a residue-level assessment of evolutionary constraint across the extension. Because the extension is encoded within the 3' untranslated region (3'-UTR), its conservation could also reflect selection on nucleotide features, rather than on the encoded peptide itself. It therefore remains unclear whether evolutionary constraint extends across the complete peptide or is concentrated within specific regions or motifs, whether it acts on the peptide or on the underlying nucleotide sequence, and whether it is shared by analogous downstream reading frames of other aquaporin paralogues.

In this study, we investigate the evolutionary constraint acting on the AQP4ex readthrough extension using a comparative evolutionary analysis. We analyse AQP4 orthologues from 759 vertebrate species, 697 of them with a 3'-UTR record, to reconstruct the evolutionary history of the readthrough-permissive stop context. We apply site-wise codon models to estimate the distribution of selective pressure across the peptide against a gene-wide synonymous-rate reference. To determine whether the evolutionary constraint pattern is a broader feature of this protein family, we compared AQP4 with the orthodox aquaporins AQP1, AQP2, AQP5, AQP6 and AQP8.

Methods

Sequence data

Coding sequences (CDS) and 3'-UTRs for vertebrate orthologues of AQP4 and of five paralogous orthodox aquaporins (AQP1, AQP2, AQP5, AQP6 and AQP8) were retrieved from NCBI Orthologs²⁵ (accessed 27 July 2026). Human sequences were used as reference records for transcript selection and protein-length composition. For AQP4 this yielded 1,751 CDS and 1,621 3'-UTR records. Each 3'-UTR record begins at the nucleotide immediately following the annotated stop codon of the transcript with the same accession.

Sequence curation and species set

A CDS was retained if its length was a multiple of three, it began with ATG, it ended in a stop codon and it contained no internal stop codon. For AQP4, 1,738 of 1,751 records met these criteria (12 contained an internal stop codon and 1 lacked an initiator ATG). Where a species had more than one record, a single representative was selected by preferring curated reference sequence (RefSeq) records (NM_) over computationally predicted models (XM_) and then the record whose protein length was closest to that of the human reference (AQP4, 323 residues, NM_001650.7; AQP1, 269; AQP2, 271; AQP5, 265; AQP6, 282; AQP8, 261). Representatives whose protein length differed from the human reference by more than 25% were excluded as probable mis-annotations (none for AQP4; 0–19 per paralogue).

Six nested species sets were defined for every gene (Table 1), together with a seventh set, T, used only for the ancestral-state analysis of AQP4. Because the analyses differ in what they require of a record (stop codon context requires only the stop codon and the 4 nucleotides downstream of it, extension length requires a reading frame that terminates within the record and selection analysis requires an extension long enough to align), each analysis was performed on the largest set that satisfies its requirement. Set A1 comprised all species with a CDS that met these criteria (AQP4, $n = 759$). Set A2 comprised species in A1 with a 3'-UTR record ($n = 697$). Set A3 comprised species in A2 whose downstream reading frame terminated at an in-frame stop codon within the 3'-UTR record ($n = 687$). The ten excluded records had 3'-UTRs of 5–84 nucleotides, and translation ran to the end of the record without reaching a stop codon. Set A4 comprised species in A3 with two records removed whose reading frame exceeded 35 codons, lacked the C-terminal motif shared by all other long frames and encoded sequence unrelated to that of other species, consistent with a frameshift in the annotation (*Festucalex cinctus* and *Syngnathoides biaculeatus*; $n = 685$). Because this criterion depends on the AQP4 terminal motif, it was not applied to the paralogues, for which A4 is identical to A3. Of the 759 AQP4 records in A1, 11 (1.4%) were curated RefSeq entries.

Two smaller sets were derived from A4 and a third from A2. Set G kept one species per genus (AQP4, $n = 530$). If the genus had several species, we kept the human record where present and otherwise the species with the longest protein, with ties resolved alphabetically. Set S kept the species in G whose extension was at least 10 codons long (AQP4, $n = 524$). Set S was used for alignment, tree inference and selection analysis. Set T kept one species per genus from A2, restricted to species whose 3'-UTR record contained at least the four nucleotides after the stop codon, and was selected by the same rule as G (AQP4, $n = 540$). T was drawn from A2 rather than A4 because scoring the stop codon context does not require a complete downstream reading frame. Set T was used only for the ancestral-state analysis.

Species were assigned to clades from their NCBI Taxonomy lineage, obtained from a local copy of the taxonomy database (downloaded 30 July 2026) with ete3 v3.1.3²⁶. Clades were defined by lineage membership as Cyclostomata, Chondrichthyes, Non-Teleost Actinopterygii, Teleostei (divided into Ostariophysi, Acanthomorphata and other teleosts), Sarcopterygian fishes, Amphibia, Aves, Non-Avian Sauropsida, Monotremata, Marsupialia and Eutheria.

Downstream reading frame and stop-codon context

The 3'-UTR of each species was translated in frame with the CDS, beginning at its first nucleotide, up to the first in-frame stop codon. Extension length is reported as the number of codons preceding that stop codon. The canonical stop codon, which is decoded during readthrough, but whose decoded identity was not determined here, was excluded from all analyses. The human extension therefore comprises 28 residues, numbered 325–352 after the 323-residue channel/core protein and the decoded stop codon at position 324. The stop codon and the four downstream nucleotides (+ 4 to + 7) were tabulated for all species in A2. Contexts were classified hierarchically as UGA-C, UGA-CUA[G/U] and UGA-CUAG^{19,27}. Extension length and terminal sequence composition were computed on A4.

Alignment and phylogenetic inference

Channel-domain protein sequences were aligned with MAFFT v7.526 using the L-INS-i algorithm²⁸. Codon alignments were produced by projecting the protein alignment onto the corresponding nucleotide sequences, so that all indels comprise whole codons and the reading frame is preserved. Channel columns at which the human sequence was gapped were removed, so that channel codon *N* corresponds to human residue *N* (323 codons for AQP4). Extension protein sequences were aligned separately with MAFFT L-INS-i and projected onto codons in the same way. Extension columns were not masked by the human sequence. For AQP4, extension alignments were also generated with MUSCLE v3.8.1551²⁹ and with the phylogeny-aware aligner PRANK v.170427³⁰, using the gene tree as the guide tree. For each gene, a maximum-likelihood tree was inferred from the channel codon alignment of the selection-analysis set with IQ-TREE v2.0.7³¹ under a fixed GTR + F+R4 model, with 1,000 ultrafast bootstrap replicates³² and random seed. For every analysis of a taxon subset, the tree was pruned to exactly the taxa in the corresponding alignment. Branch-support values were not used in downstream analyses.

Site-wise selection analysis

Site-wise synonymous (α) and non-synonymous (β) substitution rates were estimated with the fixed-effects likelihood (FEL) method v2.6³³ in HyPhy v2.5.101³⁴. All branches were included and site-to-site synonymous rate variation was enabled. No multiple-hit model or parametric resampling was used. Branch lengths and nucleotide substitution parameters were estimated once under a global MG94×REV codon model³⁵ and held fixed. α and β were then estimated independently at each codon and $\beta = \alpha$ was tested by a likelihood-ratio test with one degree of freedom. Two hypotheses make distinguishable predictions for the extension.

Selection on the encoded peptide should suppress β while leaving α comparable to that of the surrounding coding sequence, whereas selection on the nucleotide sequence should suppress both rates. To measure α against a gene-wide reference, all site-level results were obtained from a joint alignment that combined the channel and extension codon alignments (AQP4: 323 + 59 codons, 524 species). Because a joint fit yields a single global dN/dS, module-specific global dN/dS values were obtained from separate channel-only and extension-only fits using the same taxa and tree. These values are labelled as such wherever they appear. The channel domain analysed in the joint fit served as a positive control for the sensitivity of the analysis in each gene. A codon position was tested only if (i) at least 50% of sequences had a non-gap codon in that column (ii) $\alpha \leq 10$ and (iii) $\alpha > 0.01$. The occupancy criterion excludes insertion columns represented by one or a few sequences, which carry no information on site-wise rates. The criterion on α excludes columns at the parameter boundary because ω (ratio of β/α) is undefined and the likelihood-ratio test is not valid. p values were corrected for multiple testing by the Benjamini–Hochberg procedure within each gene and module. A site was classified as under purifying selection if $q \leq 0.20$ and $\beta < \alpha$, and as under diversifying selection if $q \leq 0.20$ and $\beta > \alpha$. The threshold of $q \leq 0.20$ follows the reporting threshold used for Contrast-FEL³⁶. Bonferroni-corrected significance ($p \leq 0.05$) is reported alongside. For each extension column, the most frequent amino acid and its frequency among non-gap sequences were computed from the primary alignment. Synonymous rates of tested extension and channel sites from the primary joint fit were compared with a two-sided Mann–Whitney U test and with a permutation test of the difference in medians (10,000 permutations).

Robustness analyses

The primary analysis (MAFFT²⁸, 524 species) was repeated with the MUSCLE²⁹ and PRANK³⁰ extension alignments. It was also repeated after removing species with identical extension nucleotide sequences. To assess the effect of synonymous saturation at deep divergences, the analysis was repeated within Gnathostomata (521 species), Euteleostomi (505), Tetrapoda (383) and Mammalia (172) (Table 2). Each stratum was analysed with a joint fit for site-level inference and with an extension-only fit for the module-specific dN/dS. All robustness analyses used the tree pruned from the primary gene tree and the same filtering and correction procedures.

Ancestral state of the UGA-CUAG context

A time-calibrated species tree was obtained from TimeTree 5³⁷ for the 540 species of set T. TimeTree placed 524 of the 540 species. Tip names were matched to the submitted species through NCBI taxonomy, which reconciles the synonyms used by TimeTree. Thirty-nine species that TimeTree returned as several subspecies were reduced to a single tip. In each case, the subspecies formed a clade, so the species-level tree was unchanged. All 524 tips were scored for the presence or absence of UGA-CUAG. The root of this tree separates Cyclostomata from Gnathostomata. Ancestral states were reconstructed by unit-cost parsimony³⁸. The minimum number of state changes was computed separately with the root fixed as present and as absent. When costs were equal, the root was reported as equivocal. For each root state with the minimum parsimony cost, we traced one reconstruction that placed state changes as far from the root

as possible and one that placed them as close to the root as possible. State changes are reported as events on branches, together with the number of descendant tips and the lowest NCBI taxon that contains them. Two-state continuous-time Markov (Mk) models with equal rates and with unequal gain and loss rates were also fitted by maximum likelihood, using branch lengths in millions of years and Nelder–Mead optimisation. The marginal probability of each state at the root is reported under flat and stationary root priors.

Comparison with paralogous aquaporins

AQP1, AQP2, AQP5, AQP6 and AQP8 were analysed with the same procedures and parameter values as AQP4, using MAFFT alignments only. Gene-specific inputs were limited to the length of the human reference protein. AQP5, AQP6 and AQP8, which carry the UGA-C context in a subset of species, formed the comparison set. AQP1 and AQP2, in which UGA-C was absent from all sampled species, were analysed as controls for downstream frames that are not expected to be translated, providing an empirical reference for site-wise calls in untranslated 3'-UTR sequence. Because the paralogue selection sets consist mainly of tetrapods, the paralogues were also compared with the AQP4 Tetrapoda and Mammalia analyses, which have comparable phylogenetic depth. Proportions of tested downstream-frame columns under purifying selection were compared between genes with two-sided Fisher's exact tests. For AQP5, the site-wise analysis was repeated on the 124 species of set S that carry the UGA-C context.

Results

The readthrough-permissive stop context is conserved across vertebrates

Fig. 1a shows the translational readthrough mechanism of AQP4ex formation and Table 1 list out the species set. The ribosome reads through UGA-C context forming the 28-amino acid long extended tail. Of the 697 species with a 3'-UTR record (A2), 694 (99.57%) terminated the AQP4 coding sequence with UGA (Fig. 1b) and 693 (99.43%) carried cytosine at position +4. These were almost entirely the same species where 692 (99.28%) had both, giving the readthrough-permissive UGA-C context (Fig. 1b). Five species lacked UGA-C (Fig. S1). Three carried a stop codon other than UGA (UAA in *Alligator mississippiensis* and *Rattus rattus*; UAG in *Gymnodraco acuticeps*), four carried a nucleotide other than cytosine at +4 (*A. mississippiensis*, *R. rattus*, *Camelus ferus* and *Hippocampus zosterae*), and *A. mississippiensis* and *R. rattus* differed at both positions.

Conservation extended beyond position + 4. The UGA-CUA[G/U] context occurred in 665 of 697 species (95.41%) (Fig. 1b), comprising UGA-CUAG in 566 (81.21%) and UGA-CUAU in 99 (14.20%). UGA-CUAG is the context associated with efficient readthrough of AQP4 in mammalian cells¹⁹. No other context occurred in more than 1% of species (UGA-CGAG, 7 species, 1.00%, Fig. S2). UGA-CUAG had a distinct phylogenetic distribution (Fig. 1c). It was absent from all four lamprey records (Cy), which carried UGA-CUAC (three species) or UGA-CUCC (one species). It was present in 16 of 17 cartilaginous fishes (Ch), all 6 non-teleost actinopterygians (NTA), all 12 teleosts (OT) outside Ostariophysi and Acanthomorphata, 36 of 47 ostariophysans (Os, 76.6%), 32 of 33 amphibians (Am, 97.0%), 235 of 236 sauropsids (Av + NAS, 99.6%) and

227 of 240 mammals (Mo + Ma+Eu, 94.6%) (Fig. 1c). Two groups departed from this pattern. Acanthomorph fishes (Ac) carried UGA-CUAU in 85 of 100 species and UGA-CUAG in one, and all 11 marsupials (Ma) carried UGA-CUAU (Fig. 1c).

To determine whether the phylogenetic distribution of UGA-CUAG (Fig. 1c) reflects a few ancient events or repeated independent changes and whether this context was present in the vertebrate ancestor, we reconstructed its ancestral state. On a time-calibrated species tree of 524 species (Fig. 1d), the distribution of UGA-CUAG required nine state changes by parsimony. Eight changes were common to all most-parsimonious reconstructions which were losses in acanthomorph fishes (77 species, Ac(77)), marsupials (Me(9)), a clade of siluriform catfishes (Si(6)), catostomid suckers (Ca(2)) and four single species (*A. mississippiensis* (AM), *Latimeria chalumnae* (LC), *R. rattus* (RR) and *Rhinoderma darwinii* (RD)). The ninth change was either a gain on the stem of jawed vertebrates (Gn) or a loss on the lamprey stem (PM). Both reconstructions required nine steps, so parsimony did not resolve the state at the root of the vertebrate tree. Two-state Mk models assigned a probability of 0.82 to presence at the root under equal rates and 0.81–1.00 under unequal rates, depending on the root prior. With a single outgroup lineage, these values are determined by the model and prior rather than by outgroup evidence. Thus, UGA-CUAG was present in the common ancestor of jawed vertebrates and was subsequently lost at least 8 times. The core UGA-C context was present in all four lamprey records and in 99.3% of jawed vertebrates (Fig. 1c).

Table 1
Species sets used in each analysis.

Set	Definition	AQP1	AQP2	AQP4	AQP5	AQP6	AQP8
A1	One CDS per species meeting the sequence criteria	539	416	759	383	276	517
A2	A1 with 3'-UTR record, stop codon context	521	279	697	298	209	378
A3	A2 with in-frame stop codon inside the 3'-UTR record	475	218	687	285	188	328
A4	A3 without frameshifted records; used for length and terminal sequence	475	218	685	285	188	328
G	One species per genus from A4	362	173	530	225	136	245
S	Genus representatives with a reading frame ≥ 10 codons; used for alignment, phylogeny and selection	334	141	524	203	122	179
T	One species per genus from A2 with a resolvable stop codon context, placed on the TimeTree species tree; used for ancestral state reconstruction	-	-	524	-	-	-

For paralogues, A4 is identical to A3 because the frameshift criterion depends on the AQP4 terminal motif. T is defined for AQP4 only. 540 genus representatives were submitted to TimeTree and 524 were placed.

Length and terminal sequence of the downstream reading frame

Among the 685 species in A4, the reading frame downstream of the canonical stop codon had a median length of 28 codons (Fig. 2a). Lengths were concentrated at 28 codons (490 species, 71.53%) and 11 codons (88 species, 12.85%). The remaining 107 species (15.62%) were distributed across 16 other lengths, most frequently 24 (36 species, 5.26%), 27 (19, 2.77%) and 26 codons (14, 2.04%) (Fig. S3). In total, 529 reading frames (77.23%) were 25–35 codons long, and two (*Alosa alosa* and *Alosa sapidissima* (43 codons)), exceeded 35 codons. All 88 11-codon frames occurred in acanthomorph fishes (Fig. 2a, Ac).

Sequence conservation was concentrated at the C-terminus (Fig. 2b, 2c). The terminal hexapeptide KETDLL occurred in 499 of 685 species (72.85%, Fig. 2d), and the single-residue variants KETDLI and KETDIL in a further 54 (7.88%). 564 species (82.34%) terminated in D-[L/I]-[L/I]. The 11-codon frames terminated in ETNL in 85 of 88 species (Fig. 2c), most commonly as EMETNL (64 of 88; 9.34% of A4, Fig. 2d), and none terminated in KETDLL. The N-terminus was more variable (Fig. 2d). The most frequent pentapeptide, LARST, occurred in 229 species (33.43%), followed by LEDST (84, 12.26%), LSRAL (69, 10.07%, the N-terminus of most 11-codon frames), LEDGT (48, 7.01%) and the human sequence LEDRT (28, 4.09%). The alignment of the representative species shows the conservation clustering at the N- and C-terminal ends (Fig. 2e, 2f). For the 11-codon frames, among the 88 acanthomorph, C-terminal ETNL was the most common one with the N-terminal being highly variable (Fig. S4).

Purifying selection is concentrated at the C-terminus of the extension

We estimated site-wise synonymous (α) and non-synonymous (β) rates in a joint alignment of the channel domain (323 codons) and the extension (59 alignment columns) for 524 species (Fig. 3a). Of the 59 extension columns, 31 were insertions occupied by at most four sequences and were not tested. The 28 tested columns corresponded exactly to human residues 325–352. Sixteen of the 28 codons were under purifying selection ($q \leq 0.20$) and none were under diversifying selection (Fig. 3a). Twelve remained significant after Bonferroni correction ($p \leq 0.05/28$, D327, E330, Q333, D334 and P345–L352, Fig. 3a, filled blue square). At $q \leq 0.20$, the least significant residues were S335, L343 ($q = 0.093$) and S340 ($q = 0.166$). In the same fit, 308 of 316 tested channel codons (97.5%) were under purifying selection and one (codon 149, $q = 0.050$) was under diversifying selection (Fig. S5a). The joint model had a global dN/dS of 0.150 (Fig. S5b). Module-specific fits gave 0.134 for the channel and 0.430 for the extension (Fig. S5c).

Constraint was strongest at the C-terminus (Fig. 3b). At K347–L352, β ranged from 0.013 to 0.107 and site-wise ω from 0.016 to 0.177, whereas α (0.61–1.20) was close to the gene-wide mean of 1 (Fig. 3a). At P345–K347, ω was 0.18–0.28. At the N-terminus, E330 combined low β (0.155; $\omega = 0.276$) with conservation of glutamate in 92.4% of sequences. At D327, Q333 and D334, significance reflected elevated synonymous rates ($\alpha = 1.39$ –1.66) as well as reduced β (0.38–0.66). β at D334 (0.662) exceeded that at S335 (0.464), S340 (0.525) and L343 (0.376).

At several positions, the human residue was not the most frequent state (Fig. 3c). The dominant residue was arginine at position 327 (55.3% of non-gap sequences; aspartate 25.6%), glutamate at 333 (48.8%; glutamine 25.2%), glutamine at 334 (38.2%; aspartate 15.8%) and proline at 335 (69.6%; serine 15.6%). Position 340 was occupied by serine (45.9%) or threonine (42.6%). R332 was conserved in 83.1% of sequences but was

not under detectable selection ($\alpha = 0.33$, $\beta = 0.22$, $q = 0.39$). Its low synonymous rate limits the power to detect $\beta < \alpha$. In the C-terminal segment, the dominant residue matched the human sequence at every position (K347 (85.7%), E348 (99.2%), T349 (99.2%), D350 (84.1%; asparagine 13.4%), L351 (94.1%) and L352 (90.6%; isoleucine 8.7%), blue circles).

Synonymous rates were lower in the extension than in the channel domain (median $\alpha = 0.707$ versus 1.009 across tested sites, two-sided Mann–Whitney $p = 1.3 \times 10^{-4}$, permutation test on the difference in medians $p = 0.0012$, Fig. 3d). The lowest synonymous rate in the extension occurred at codon 325 ($\alpha = 0.029$, $\beta = 0.061$, Fig. 3a), which comprises positions + 4 to + 6 of the stop-codon context and encoded leucine in 98.1% of species. Both rates were suppressed at this codon, and it was not called under selection ($q = 0.52$).

To check the robustness of the result, we repeated the above analysis with different alignment software. The site-level result was insensitive to alignment method and sequence redundancy. The MUSCLE²⁹ and PRANK³⁰ alignments each yielded 28 tested codons and 16 sites under purifying selection where 13 of them were Bonferroni-significant (Fig. S6, S7, Table 2). Removing 153 species with duplicate extension sequences (371 remaining) yielded 16 sites where 12 were Bonferroni-significant (Table 2). Fifteen sites were recovered in all four analyses (D327, E330, Q333, D334, S335 and L343–L352) and the 12 Bonferroni-significant sites of the primary analysis were Bonferroni-significant in all four. S340 was not recovered with the PRANK³⁰ alignment (Fig. S6).

Restricting the analysis to Gnathostomata (521 species), Euteleostomi (505), Tetrapoda (383) and Mammalia (172) yielded 16, 16, 17 and 17 sites under purifying selection, respectively (Fig. 3e, Table 2). Fourteen sites (D327, E330, Q333, D334 and L343–L352) were recovered in all eight analyses. R332 was recovered only in tetrapods and mammals, in which arginine occupied 94.8% and 99.4% of sequences, and was Bonferroni-significant in mammals ($\beta = 0.031$). S335 was not recovered in mammals, in which proline predominated at this position (66.3%; serine 26.7%, Fig. 3c). The extension-only dN/dS were 0.434 (95% CI 0.389–0.483) in gnathostomes, 0.433 (95% CI 0.387–0.484) in euteleostomes, 0.392 (95% CI 0.341–0.449) in tetrapods and 0.335 (95% CI 0.269–0.413) in mammals (Table 2, Fig. S8b). Codon 328 in tetrapods ($q = 0.148$), was classified as under diversifying selection, but its α (0.10) was close to the parameter boundary (Fig. 3e).

Table 2
Sensitivity of site-level inference to alignment, redundancy and phylogenetic depth.

Analysis	Species	Codons tested	Purifying ($q \leq 0.20$)	Bonferroni	Extension dN/dS
MAFFT (primary)	524	28	16	12	0.430
MUSCLE	524	28	16	13	—
PRANK	524	28	16	13	—
MAFFT, duplicates removed	371	28	16	12	—
Gnathostomata	521	28	16	12	0.434
Euteleostomi	505	28	16	12	0.433
Tetrapoda	383	27	17	12	0.392
Mammalia	172	25	17	8	0.335

All analyses are joint channel-plus-extension fits; extension dN/dS are from the corresponding extension-only fit.

The combination of features is specific to AQP4 among orthodox aquaporins

To determine whether these features are specific to AQP4, we applied the same procedures to AQP1, AQP2, AQP5, AQP6 and AQP8 (Table 3). The UGA-C context was absent from AQP1 (0 of 519 species) and AQP2 (0 of 275 species) and present in 14.1% of AQP6, 23.0% of AQP8 and 58.6% of AQP5 species, compared with 99.3% for AQP4 (Fig. 4a). Among the stop codons of AQP1, UGA occurred in 7 of 519 species (Fig. 4b), so that the AQP1 downstream frame is the most stringent negative control. In AQP2, UGA occurred in 161 of 275 species (Fig. 4b), but none of these carried cytosine at + 4 position. The + 4 position was guanine in 93.8%, giving the less permissive UGA-G context (Fig. 4b). Because UGA-C was absent from every AQP1 and AQP2 species (Fig. 4a), their downstream frames are not expected to be translated, and these two genes were treated as controls, with AQP5, AQP6 and AQP8 forming the comparison set. Downstream reading frames occurred in all six genes, with median lengths from 11 (AQP5) to 34 codons (AQP2) (Fig. S9).

The downstream frames of AQP5, AQP6 and AQP8 had module-specific dN/dS of 0.670–0.859, compared with 0.430 for AQP4, and those of the control genes AQP1 and AQP2 had 0.896 and 0.725 (Fig. 4c, Table 3). For AQP1 and AQP2, these values provide a baseline for untranslated downstream sequence under the same model rather than a measure of constraint on an encoded peptide. The corresponding channel domains were strongly constrained (dN/dS 0.06–0.20, 78.8–95.5% of tested codons under purifying selection, Fig. 4c). After the occupancy criterion, 11–23 extension columns were testable in AQP5, AQP6 and AQP8, of which 1–4 (5.9–27.3%) were under purifying selection, compared with 16 of 28 (57.1%) for AQP4 (Fig. 4d, Table 3). In the control genes, 7 of 21 (AQP1) and 8 of 30 (AQP2) columns were classified as under purifying selection (Fig. 4d). The proportion of tested columns under purifying selection was higher in AQP4 than in the two control genes combined (16 of 28 versus 15 of 51; Fisher's exact $p = 0.029$, Fig. S8a), although not significantly higher than in AQP1 alone ($p = 0.15$, Fig. S8a). Because the paralogue selection sets consist

mainly of tetrapods, AQP4 was also compared at matched depth. In tetrapods and mammals, 17 of 27 and 17 of 25 AQP4 extension columns were under purifying selection, with extension-only dN/dS of 0.392 and 0.335 (Fig. 3e, Fig. S8b). The mammalian proportion exceeded that of the control genes (Fisher's exact $p = 0.003$, Fig. S8a). The downstream frames also differed in their synonymous rates. In all the five paralogues the median site-wise synonymous rate of the downstream frame equalled or exceeded that of the corresponding channel domain, whereas in AQP4 it was lower (0.71 vs 1.01, Fig. 4d). Suppression of the synonymous rate in the downstream frame is therefore specific to AQP4. Also, the most frequent C-terminal hexapeptide occurred in 4.6% for DTITYG in AQP8 to 16.8% for GQPLSP in AQP5, compared with 72.8% for KETDLL in AQP4 (Fig. 4e, Table 3).

Table 3
Downstream reading frames of AQP4 and five paralogous aquaporins.

Gene	Species (S)	UGA-C (% of A2)	Median frame (codons)	Most frequent C-terminal hexapeptide (% of A4)	Extension dN/dS	Channel dN/dS	Extension codons tested	Purifying ($q \leq 0.20$)
AQP4	524	99.3	28	KETDLL (72.8)	0.430	0.134	28	16
AQP1	334	0.0	26	SVIDSL (9.7)	0.896	0.083	21	7
AQP2	141	0.0	34	PPLPQV (6.9)	0.725	0.065	30	8
AQP5	203	58.6	11	GQPLSP (16.8)	0.855	0.077	11	3
AQP6	122	14.1	24	LLRGAL (10.1)	0.859	0.198	17	1
AQP8	179	23.0	22	DTITYG (4.6)	0.670	0.198	23	4
<i>dN/dS values are from module specific fits. Columns use different species sets: species (S) use selection set, UGA-C use percentage of A2 species with a resolvable stop codon, median frame and hexapeptide use A4</i>								

Across all tested sites of the downstream frame α ranged from 0.22 to 4.42 in the paralogues compared with 0.03–1.65 in AQP4 (Table 4, Fig. 5a-f). In AQP5, restricting the analysis to the 124 species in S that carry UGA-C identified no site under purifying selection (Fig. S10). By comparison, AQP4 retained 17 sites under purifying selection in the 172-species mammalian analysis (Fig. 3e), so the absence of calls in AQP5 is unlikely to reflect sample size alone.

Table 4
Synonymous and non-synonymous rates across the residues of downstream reading frames

Aquaporin	Synonymous (α) Range	Non-synonymous (β) Rate Range
AQP1	0.22–2.89	0.37–5.10
AQP2	1.01–4.01	0.07–2.89
AQP4	0.03–1.65	0.01–1.05
AQP5	0.77–2.03	0.46–1.36
AQP6	0.89–4.42	0.57–4.72
AQP8	0.76–3.62	0.10–2.07

Discussion

This study presents a comprehensive evolutionary analysis of the AQP4ex translational readthrough extension across vertebrates. The key findings are: (i) the readthrough-permissive UGA-C context is present in 99.3% of vertebrate species, including lampreys, and the extended UGA-CUAG context in 81.2%. But whether UGA-CUAG was gained in jawed vertebrates or lost in lampreys could not be resolved (ii) the downstream reading frame shows a bimodal length distribution with peaks at 28 codons and at 11 codons, the latter restricted to acanthomorph fishes (iii) 16 of the 28 positions of the extension are under purifying selection with the strongest and most robust constraint in the C-terminal segment P345–L352 and weaker constraint at N-terminal and central positions and (iv) the combination of a near-universal readthrough-permissive context, a conserved terminal motif and strong C-terminal constraint distinguishes AQP4 from AQP5, AQP6 and AQP8, which retain the UGA-C context only in a subset of species.

The near-universal conservation of the UGA-C readthrough-permissive context (99.3% of vertebrate species, Fig. 1b), together with purifying selection acting on the downstream reading frame, supports the view that the capacity for AQP4 translational readthrough has been maintained during vertebrate evolution. This is consistent with experimental studies showing that the sequence immediately downstream of UGA strongly influences readthrough efficiency and that the UGA-CUAG context found in AQP4 (Fig. 1b) is compatible with efficient readthrough in mammalian cells¹⁹. The presence of UGA-C in all four lamprey records (Fig. 1c) suggests that a readthrough-permissive context was already present before the divergence of cyclostomes and jawed vertebrates. But the history of the extended UGA-CUAG context is less certain. A gain on the gnathostome stem and a loss on the lamprey stem were equally parsimonious (Fig. 1d), and with lampreys as the only outgroup the direction of this change cannot be determined from the present data. The principal alternative, UGA-CUAU, differs by a single transversion at position + 7 and predominates in acanthomorph fishes (85 of 100 species) and marsupials (11 of 11) (Fig. 1c). Because the nucleotide immediately downstream of the UGA-CUA core can influence readthrough efficiency²⁷, these lineage-specific substitutions may alter the level of AQP4ex production, although this possibility requires experimental validation. The eight losses of UGA-CUAG (Fig. 1d) further show that the extended readthrough context has not remained invariant throughout vertebrate evolution. Together, these patterns indicate that the AQP4ex

readthrough context is not a recent lineage-specific innovation but a long-standing feature of vertebrate AQP4 that has been retained across most vertebrate lineages. Whether readthrough occurs at comparable levels outside mammals has not been tested experimentally.

The distribution of evolutionary constraint within AQP4ex suggests distinct functional roles for its C-terminal and N-terminal regions. The C-terminal segment shows the strongest constraint. All eight positions from P345 to L352 were under purifying selection in every analysis (Fig. 3a), including the tetrapod and mammal analyses (Fig. 3e), which contain none of the 69 acanthomorph 11-codon frames in S. The C-terminal signal therefore does not depend on aligning the short frames against the 28-codon frame and remained significant after Bonferroni correction with every alignment (Fig. S6). The site-wise ω at K347–L352 was 0.016–0.177 (Fig. 3a), and the human residue was the dominant residue at each position of K347–L352 (84.1–99.2% of sequences, Fig. 3c). This pattern is consistent with a functionally important C-terminal motif and resembles readthrough extensions in which the new C-terminus carries a functional terminal signal^{39,40}. The N-terminal region shows a different pattern. D327, E330, Q333 and D334 were recovered in all eight analyses (Fig. 3e), but only E330 combined low β with conservation of a single residue (glutamate, 92.4%, Fig. 3a). At positions 327, 333 and 334 the human residue is a minority state, the dominant residues being arginine, glutamate and glutamine, respectively (Fig. 3c). Also, the significance reflects elevated synonymous rates as well as reduced non-synonymous rates (Fig. 3a). Substitutions at 333 and 334 are largely among chemically related residues (E, Q and D), consistent with constraint on broader physicochemical properties such as charge, local structure or flexibility⁴¹. On the other hand, position 327 alternates between arginine and aspartate (Fig. 3c). The central segment of the extension is not unconstrained. S335, S340, L343 and P344 were under purifying selection, although with weaker support (Fig. 3a). This segment is also variable in length. 12.8% of species carry an 11-codon form, restricted to acanthomorph fishes (Fig. 2a), that lacks the central sequence of the 28-codon form and terminates in ETNL rather than KETDLL (Fig. 2e, Fig. S4). The ETNL terminus of the 11-codon acanthomorph form matches the canonical class I α -syntrophin consensus -E-S/T-X- Φ ⁴², placing the threonine at position -2 rather than at -3 as in KETDLL. Whether this short form binds to α -syntrophin has not been shown. This observation is consistent with a linker-like role in which the intervening region tolerates length variation⁴³, although experimental testing is required to determine whether its length controls the spatial relationship between the two terminal regions. Synonymous rates in the extension were lower than in the channel domain (median α :0.707 vs 1.009, Fig. 3d), and were lowest at codon 325, which forms part of the stop-codon context (Fig. 3a). This suggests that the constraint on the extension does not act on the encoded peptide alone. Nucleotide-level constraint, most plausibly on the sequence that specifies readthrough, operates alongside it. This suppression was not seen in any of the five paralogues whose downstream frames had synonymous rates equal to or above those of their channel domain (Fig. 4d), which argues against the generic difference in mutation rate between coding sequence and 3'-UTR. But differences in local mutation rate between coding sequence and 3'-UTR cannot be excluded.

This framework is consistent with previous experimental findings that AQP4ex promotes the association of AQP4 with syntrophin-containing dystrophin-associated complexes¹⁸ and is required for the normal enrichment of AQP4 at astrocytic perivascular endfeet²⁰. The strong conservation of the C-terminal segment (Fig. 2f, Fig. 3a, 3b) is therefore consistent with its contributing to a functionally important binding or anchoring surface for α -syntrophin or other components of the dystrophin-associated protein complex.

Single-molecule binding measurements with peptides of the human extension have since shown that AQP4ex binds the PDZ domain of α -syntrophin through its C-terminal residues rather than an internal motif⁴⁴. Also, alanine substitution of each of the six terminal residues reduces binding between extension and α -syntrophin, with substitution of the threonine at position 349 (T349, in our numbering) abolishing or nearly abolishing it⁴⁴. The residues identified experimentally coincide with the segment under the strongest constraint in our analysis, and T349, which is the most important residue for binding, is the most constrained site of the extension and, together with E348, the most conserved ($\omega = 0.016$; threonine in 99.2% of species, serine in 0.6% and arginine in 0.2%; Fig. 3a, 3c). Jacobo et al.⁴⁴ noted that the terminal leucine extending the class I-like motif has not been described in other PDZ ligands. Our data show that this residue is not specific to human AQP4ex but is retained as leucine or isoleucine in 99% of vertebrate sequences (Fig. 3c). The N-terminal segment contains the RXXS motif (R332-X-X-S335) of human AQP4ex (Fig. 3c), and previous mutagenesis and comparative analyses support a possible phosphorylation-dependent regulatory role for this region²⁴. Phosphorylation of AQP4ex has been proposed to influence water-channel activity and may also affect AQP4ex localisation or supramolecular assembly^{18,24}, but the responsible kinase and the direct molecular mechanism remain unresolved. Our data indicate that this motif is not conserved across vertebrates. Serine occupies position 335 in only 15.6% of species (proline, 69.6%) and in 26.7% of mammals, whereas R332 is conserved in 83.1% of species and was under detectable purifying selection only in tetrapods and mammals (Fig. 3c, 3e). Any regulatory role of S335 phosphorylation is therefore likely to be restricted to a subset of mammals. By contrast, position 340 carries serine or threonine in 88.5% of species and was under purifying selection in seven of eight analyses (Fig. 3a, 3c, 3e). This raises the possibility of a phosphorylation site that is conserved more broadly than S335. However, this residue has not been previously characterised.

The comparison with five paralogous orthodox water channels suggests that the evolutionary signature of AQP4ex is not a general feature of the aquaporin family. None of the paralogues displayed the combination of features observed in AQP4ex. AQP1 and AQP2 lack the UGA-C context in all sampled species, whereas AQP6 and AQP8 retain it in only 14.1% and 23.0% of species, respectively (Fig. 4a). This indicates that UGA-C readthrough-permissive context is not conserved in these paralogues. AQP5 presents an intermediate case. The UGA-C context was retained in 58.6% of sampled species (Fig. 4a), suggesting that the local sequence context could permit readthrough in a substantial fraction of AQP5 orthologues. However, its downstream reading frame shows no evidence of strong selective constraint. Its dN/dS was 0.855 (Fig. 4c), its median length was 11 codons (Fig. S9), and it has no widely shared terminal motif (the most frequent C-terminal hexapeptide occurred in 16.8% of species, Fig. 4e). Also, no site was under purifying selection when the analysis was restricted to the 124 species in S that carry UGA-C (Fig. S10). The same analysis applied to the AQP5 channel domain gave a fitted dN/dS of 0.08 (Fig. 4c), confirming that the procedure detects strong constraint in this gene where it is present. This indicates that a readthrough-permissive context is not sufficient to maintain a functional extension. Consistent with this, AQP4 was the only gene in which the synonymous rate of the downstream frame fell below that of its channel domain (Fig. 4d). In all five paralogues it was equal or higher. The functional maintenance appears to require additional selection acting on the encoded peptide or on linked RNA-level properties.

The AQP4 specificity of the pattern suggests (Fig. 4a-e) that AQP4ex has a unique functional requirement among orthodox aquaporins. Although AQP1, AQP2, AQP5, AQP6 and AQP8 are targeted to different cellular and membrane compartments^{6,7}, none is characterised by the same highly polarised enrichment in astrocytic perivascular endfeet as AQP4. AQP2 undergoes regulated trafficking between intracellular storage vesicles and the apical plasma membrane⁴⁵⁻⁴⁷, AQP6 is predominantly intracellular⁴⁸, and AQP8 can occur in both plasma membrane and intracellular compartments^{49,50}. Whereas AQP4 is concentrated at astrocytic endfeet, where its polarised localisation contributes to perivascular water exchange and glymphatic transport⁹⁻¹². This specialised distribution depends on anchoring through the dystrophin-associated protein complex, including α -syntrophin⁵¹⁻⁵³. The evolutionary maintenance of AQP4ex, particularly its strongly constrained C-terminal segment (Fig. 3a, 3b), is therefore consistent with a requirement for a protein-interaction or anchoring region that helps localise AQP4 to this specialised membrane domain. In AQP1 and AQP2 the downstream frame carries no readthrough-permissive context in any sampled species (Fig. 4a) and is therefore unlikely to be translated at all. These frames therefore serve as controls for untranslated sequence, and residues getting flagged as purifying (Fig. 4d, 5a) indicate that counts of constrained sites alone can overstate peptide-level constraint in 3'-UTR-encoded frames. AQP5 is the informative case here. The readthrough context is present in most species (Fig. 4a), yet the frame shows no constraint on the encoded peptide (Fig. 4c), which suggests that a permissive context is not by itself sufficient to maintain an extension.

Several aspects of the study warrant consideration. First, 98.6% of AQP4 records (748 of 759) derive from computationally predicted models (XM_ prefixes), reflecting the current state of vertebrate genome annotation. But the principal findings of a bimodal length distribution (Fig. 2a), motif conservation (Fig. 2f) and purifying selection (Fig. 3a) are unlikely to be explained entirely by random annotation errors, because they form a structured and internally coherent pattern across taxa. However, systematic errors cannot be excluded. Second, the inference that UGA-C predates jawed vertebrates rests on four computationally predicted lamprey records (Fig. 1c, 1d), and lampreys are the only outgroup available. The evolutionary timing is therefore conditional on annotation accuracy. Third, the present study identifies evolutionary constraints but cannot establish their molecular mechanisms. The strong conservation of the C-terminal segment is consistent with a protein-protein interaction or anchoring region. Attributing the C-terminal constraint to α -syntrophin binding rests on in-vitro data from human peptides⁴⁴. The site-wise rates cannot separate this from other constraints on the same residues. Whether the motif binds α -syntrophin in other species outside mammals, including the ETNL-terminated acanthomorph form, remains untested. Forth, the 524 species are not independent observations. FEL accounts for shared ancestry by fitting rates along the gene tree, so the site-wise inferences are conditional on that tree and on the extension alignment, which we varied across three aligners (Fig. S6). Finally, the functional roles of the N-terminal segment and of position 340 remain predictions that should be tested using targeted mutagenesis, phosphorylation assays, interaction studies and subcellular localisation experiments.

Declarations

Resource Availability

Data and Code Availability

All the data for performing the analysis was collected from NCBI Orthologs. The analysis code for the study, the codon alignments, gene trees, TimeTree species tree, FEL output, and AlphaFold 3 predicted structures are available at https://github.com/psychedelic2007/AQP4ex_analysis

Declaration of Interests

The author declares no competing interests.

Acknowledgements

I would like to thank A/Prof. Svetlana Postnova (University of Sydney) for her guidance on logic and consistency of methodology of the paper.

References

1. Papadopoulos MC, Verkman AS (2013) Aquaporin water channels in the nervous system. *Nat Rev Neurosci* 14:265–277. 10.1038/nrn3468
2. Frigeri A et al (1995) Localization of MIWC and GLIP water channel homologs in neuromuscular, epithelial and glandular tissues. *J Cell Sci* 108(9):2993–3002. 10.1242/jcs.108.9.2993
3. Nielsen S et al (1997) Specialized membrane domains for water transport in glial cells: high-resolution immunogold cytochemistry of aquaporin-4 in rat brain. *J neuroscience: official J Soc Neurosci* 17:171–180. 10.1523/jneurosci.17-01-00171.1997
4. Nagelhus EA, Ottersen OP (2013) Physiological roles of aquaporin-4 in brain. *Physiol Rev* 93:1543–1562. 10.1152/physrev.00011.2013
5. Rash JE, Yasumura T, Hudson CS, Agre P, Nielsen S (1998) Direct immunogold labeling of aquaporin-4 in square arrays of astrocyte and ependymocyte plasma membranes in rat brain and spinal cord. *Proc Natl Acad Sci USA* 95:11981–11986. 10.1073/pnas.95.20.11981
6. Albertini R, Bianchi R (2010) Aquaporins and glia. *Curr Neuropharmacol* 8:84–91. 10.2174/157015910791233178
7. Badaut J, Lasbennes F, Magistretti PJ, Regli L (2002) Aquaporins in brain: distribution, physiology, and pathophysiology. *J Cereb blood flow metabolism: official J Int Soc Cereb Blood Flow Metabolism* 22:367–378. 10.1097/00004647-200204000-00001
8. Simon M et al (2022) Loss of perivascular aquaporin-4 localization impairs glymphatic exchange and promotes amyloid β plaque formation in mice. *Alzheimers Res Ther* 14. 10.1186/s13195-022-00999-5
9. Bojarskaite L et al (2024) Role of aquaporin-4 polarization in extracellular solute clearance. *Fluids barriers CNS* 21:28. 10.1186/s12987-024-00527-7
10. Eidsvaag VA, Enger R, Hansson HA, Eide PK, Nagelhus EA (2017) Human and mouse cortical astrocytes differ in aquaporin-4 polarization toward microvessels. *Glia* 65:964–973. 10.1002/glia.23138

11. Yang J et al (2011) Loss of astrocyte polarization in the tg-ArcSwe mouse model of Alzheimer's disease. *J Alzheimer's disease: JAD* 27:711–722. 10.3233/jad-2011-110725
12. Zhu DD et al (2022) AQP4-A25Q Point Mutation in Mice Depolymerizes Orthogonal Arrays of Particles and Decreases Polarized Expression of AQP4 Protein in Astrocytic Endfeet at the Blood-Brain Barrier. *J neuroscience: official J Soc Neurosci* 42:8169–8183. 10.1523/jneurosci.0401-22.2022
13. Rossi A, Pisani F, Nicchia GP, Svelto M, Frigeri A (2010) Evidences for a leaky scanning mechanism for the synthesis of the shorter M23 protein isoform of aquaporin-4: implication in orthogonal array formation and neuromyelitis optica antibody interaction. *J Biol Chem* 285:4562–4569. 10.1074/jbc.M109.069245
14. Furman CS et al (2003) Aquaporin-4 square array assembly: opposing actions of M1 and M23 isoforms. *Proc Natl Acad Sci USA* 100:13609–13614. 10.1073/pnas.2235843100
15. Jin BJ, Rossi A, Verkman AS (2011) Model of aquaporin-4 supramolecular assembly in orthogonal arrays based on heterotetrameric association of M1-M23 isoforms. *Biophys J* 100:2936–2945. 10.1016/j.bpj.2011.05.012
16. Nicchia GP et al (2008) Expression of multiple AQP4 pools in the plasma membrane and their association with the dystrophin complex. *J Neurochem* 105:2156–2165. 10.1111/j.1471-4159.2008.05302.x
17. Nicchia GP et al (2010) Higher order structure of aquaporin-4. *Neuroscience* 168:903–914. 10.1016/j.neuroscience.2010.02.008
18. De Bellis M et al (2017) Translational readthrough generates new astrocyte AQP4 isoforms that modulate supramolecular clustering, glial endfeet localization, and water transport. *Glia* 65:790–803. 10.1002/glia.23126
19. Loughran G et al (2014) Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic Acids Res* 42:8928–8938. 10.1093/nar/gku608
20. Palazzo C et al (2019) AQP4ex is crucial for the anchoring of AQP4 at the astrocyte end-feet and for neuromyelitis optica antibody binding. *Acta Neuropathol Commun* 7. 10.1186/s40478-019-0707-5
21. Valente O et al (2022) Alteration of the translational readthrough isoform AQP4ex induces redistribution and downregulation of AQP4 in human glioblastoma. *Cell Mol Life Sci* 79:140. 10.1007/s00018-021-04123-y
22. Jungreis I et al (2016) Evolutionary Dynamics of Abundant Stop Codon Readthrough. *Mol Biol Evol* 33:3108–3132. 10.1093/molbev/msw189
23. Stark A et al (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450:219–232. 10.1038/nature06340
24. Pati R et al (2022) The Readthrough Isoform AQP4ex Is Constitutively Phosphorylated in the Perivascular Astrocyte Endfeet of Human Brain. *Biomolecules* 12. 10.3390/biom12050633
25. O'Leary NA et al (2024) Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci data* 11:732. 10.1038/s41597-024-03571-y
26. Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* 33:1635–1638. 10.1093/molbev/msw046

27. Stiebler AC et al (2014) Ribosomal readthrough at a short UGA stop codon context triggers dual localization of metabolic enzymes in Fungi and animals. *PLoS Genet* 10:e1004685. 10.1371/journal.pgen.1004685
28. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780. 10.1093/molbev/mst010
29. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797. 10.1093/nar/gkh340
30. Löytynoja A (2021) Phylogeny-Aware Alignment with PRANK and PAGAN. *Methods in molecular biology*. (Clifton N J) 2231:17–37. 10.1007/978-1-0716-1036-7_2
31. Minh BQ et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37:1530–1534. 10.1093/molbev/msaa015
32. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35:518–522. 10.1093/molbev/msx281
33. Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 22:1208–1222. 10.1093/molbev/msi105
34. Kosakovsky Pond SL et al (2020) HyPhy 2.5-A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol Biol Evol* 37:295–299. 10.1093/molbev/msz197
35. Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* 11:715–724. 10.1093/oxfordjournals.molbev.a040152
36. Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S (2021) Contrast-FEL-A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches. *Mol Biol Evol* 38:1184–1198. 10.1093/molbev/msaa263
37. Kumar S et al (2022) TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol Biol Evol* 39. 10.1093/molbev/msac174
38. Sankoff D (1975) Minimal Mutation Tree of Sequences. *SIAM J Appl Math* 28
39. Hofhuis J et al (2016) The functional readthrough extension of malate dehydrogenase reveals a modification of the genetic code. *Open biology* 6. 10.1098/rsob.160246
40. Schueren F, Thoms S (2016) Functional Translational Readthrough: A Systems Biology Perspective. *PLoS Genet* 12:e1006196. 10.1371/journal.pgen.1006196
41. Juritz E, Palopoli N, Fornasari MS, Fernandez-Alberti S, Parisi G (2013) Protein conformational diversity modulates sequence divergence. *Mol Biol Evol* 30:79–87. 10.1093/molbev/mss080
42. Liu X, Fuentes EJ (2019) Emerging Themes in PDZ Domain Signaling: Structure, Function, and Inhibition. *Int Rev cell Mol biology* 343:129–218. 10.1016/bs.ircmb.2018.05.013
43. Fahmi M, Ito M (2019) Evolutionary Approach of Intrinsically Disordered CIP/KIP Proteins. *Sci Rep* 9:1575. 10.1038/s41598-018-37917-5
44. Jacobo EP et al (2026) AQP4 Extended Region Binds to α -Syntrophin with a Novel Binding Motif. *J Phys Chem B*

45. Agarwal SK, Gupta A, Aquaporins (2008) The renal water channels. *Indian J Nephrol* 18:95–100. 10.4103/0971-4065.43687
46. Brown D, Katsura T, Gustafson CE (1998) Cellular mechanisms of aquaporin trafficking. *Am J Physiology-Renal Physiol* 275
47. Noda Y, Sasaki S (2005) Trafficking mechanism of water channel aquaporin-2. *Biol Cell* 97:885–892. 10.1042/bc20040120
48. Yasui M, Kwon TH, Knepper MA, Nielsen S, Agre P (1999) Aquaporin-6: An intracellular vesicle water channel protein in renal epithelia. *Proc Natl Acad Sci USA* 96:5808–5813. 10.1073/pnas.96.10.5808
49. Burghardt B et al (2003) Distribution of aquaporin water channels AQP1 and AQP5 in the ductal system of the human pancreas. *Gut* 52:1008–1016. 10.1136/gut.52.7.1008
50. Elkjaer ML et al (2001) Immunolocalization of aquaporin-8 in rat kidney, gastrointestinal tract, testis, and airways. *Am J Physiol Renal Physiol* 281:F1047–1057. 10.1152/ajprenal.0158.2001
51. Costea D et al (2025) The Glymphatic-Venous Axis in Brain Clearance Failure: Aquaporin-4 Dysfunction, Biomarker Imaging, and Precision Therapeutic Frontiers. *Int J Mol Sci* 26. 10.3390/ijms262110546
52. Ding J, Zhao C, Hao X, Jiao H (2025) Glymphatic and meningeal lymphatic dysfunction in Alzheimer's disease: Mechanisms and therapeutic perspectives. *Alzheimer's Dement J Alzheimer's Assoc* 21:e70709. 10.1002/alz.70709
53. Neely JD et al (2001) Syntrophin-dependent expression and localization of Aquaporin-4 water channel protein. *Proc Natl Acad Sci USA* 98:14108–14113. 10.1073/pnas.241508198

Figures

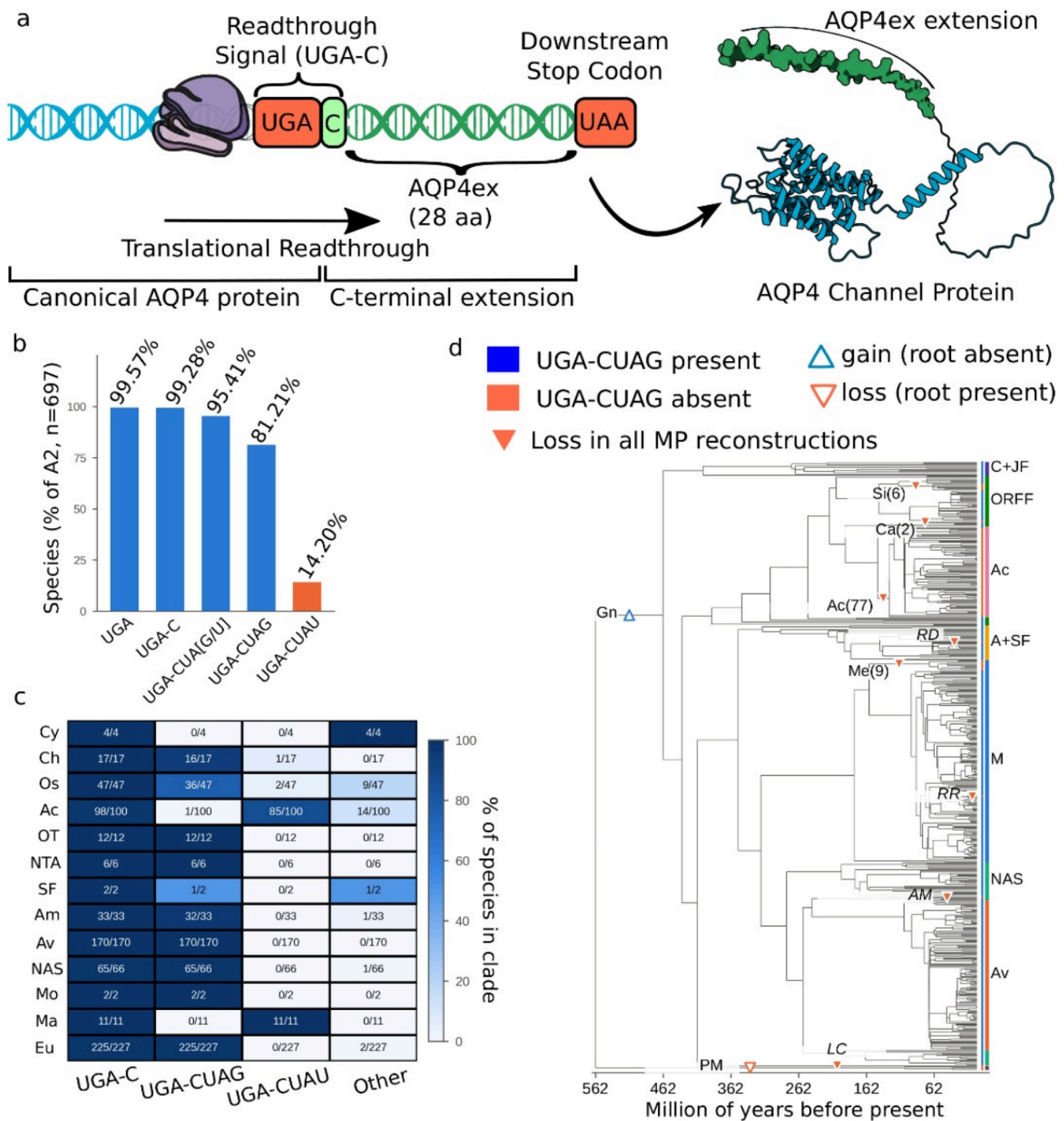


Figure 1

Readthrough permissive context across vertebrates for AQP4. (a) Readthrough mechanism showing the process of AQP4 extension (AQP4ex) formation with the core AQP4 channel protein (blue) and the extended tail (green) (b) Different AQP4 codon context showing the percentage appearance in the species set A2 (c) Distribution of UGA-C, UGA-CUAG, UGA-CUAU and other codon context across different phylogenetic class (Cy: Cyclostomata, Ch: Chondrichthyes, Os: Ostariophysi, Ac: Acanthomorphata, OT: Other Teleosts, NTA:

Non-Teleost Actinopterygians, SF: Sarcopterygian Fishes, Am: Amphibia, Av: Aves, NAS: Non-Avian Sauropsids, Mo: Monotremata, Ma: Marsupialia, Eu: Eutheria) (d) Time calibrated species phylogenetic tree showing the historical distribution of UGA-CUAG stop context in AQP4. The tree contains the 54 species of tree set T (one species per genus from A2) placed by TimeTree. The axis shows millions of years before present, with the root at about 563 million years. Squares beside the tip marks whether UGA-CUAG is present (blue) or absent (orange). Filled orange triangle are eight losses shared by every parsimonious reconstruction (Ac: Acanthomorphata, Si: Siluroidei, Ca: Catostomidae, RD: *Rhinoderma darwinii*, Me: Metatheria, RR: *Rattus rattus*, AM: *Alligator mississippiensis*, LC: *Latimeria chalumnae*, PM: Petromyzontidae, Gn: Gnathostomata). The hollow triangle shows the ninth change which the parsimony cannot resolve. It corresponds to either a gain on the jawed vertebrate stem (blue hollow triangle) or a loss on the lamprey stem (orange hollow triangle). The coloured strip on the right gives major clades by NCBI Taxonomy (C+JF: cartilaginous and jawless fishes, ORFF: other ray-finned fishes, Ac: Acanthomorphata, A+SF: amphibians and sarcopterygian fishes, M: Mammalia, NAS: non-avian sauropsids, Av: Aves).

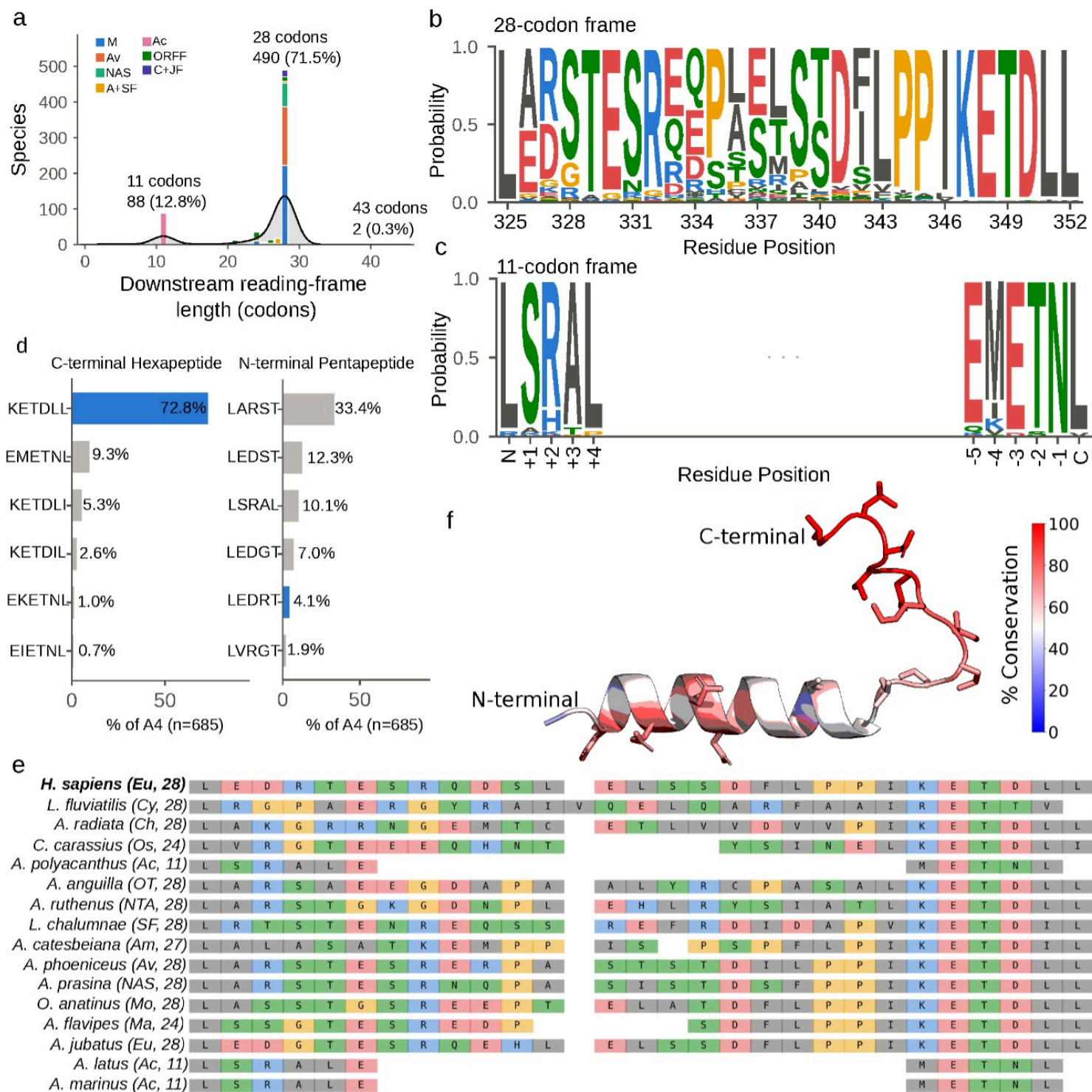


Figure 2

Length and terminal sequence of the downstream reading frame. (a) Length distribution of the AQP4 extension. Bars are stacked by clade group. Lengths cluster at 28 codons (490 species, 71.5%) and 11 codons (88 species, 12.8%; all acanthomorph fishes) (b) Probability logos for the extension sequence numbered by human AQP4ex position (325–352). (c) Probability logos of the 11-codon frames showing the first five residues (N, +1 to +4) and the last six (–5 to –1, C). The dots mark a break in the display. The 11-codon logo sits under the 28-codon logo for layout only, and this placement does not imply positional homology. Residues are coloured by physicochemical class: acidic (D, E) red; basic (K, R, H) blue;

hydrophobic (A, V, L, I, M, F, W, C) grey; polar (S, T, N, Q, Y) green; P and G yellow. (d) The six most frequent C-terminal hexapeptides and the five most frequent N-terminal pentapeptides. The human sequences are in blue (KETDLL and LEDRT) (e) MAFFT protein alignment of the extension sequence in representative species. Shown are *Homo sapiens* (bold), one representative per clade and two more 11-codon acanthomorph fishes (*A. lactus* and *A. marinus*). Labels give the species, its clade (Eu: Eutheria, Cy: Cyclostomata, Ch: Chondrichthyes, Os: Ostariophysi, Ac: Acanthomorphata, OT: Other teleosts, NTA: Non-Teleosts Actinopterygians, SF: Sarcopterygian Fishes, Am: Amphibians, Av: Aves, NAS: Non-Avian Sauropsids, Mo: Monotremata, Ma: Marsupialia) and its frame length in codons. Columns that are gaps in all of these rows are removed. Residue are coloured as in Fig. 2c. In the 11-codon frames, the N-terminal LSRAL and C-terminal (M/I)ETNL align with the start and end of the 28-codon frames, and the central segment is gapped. (f) AlphaFold3 predicted structure of the human AQP4 extension sequence coloured based on conservation (blue: less conserved, white: moderate, red: highly conserved)

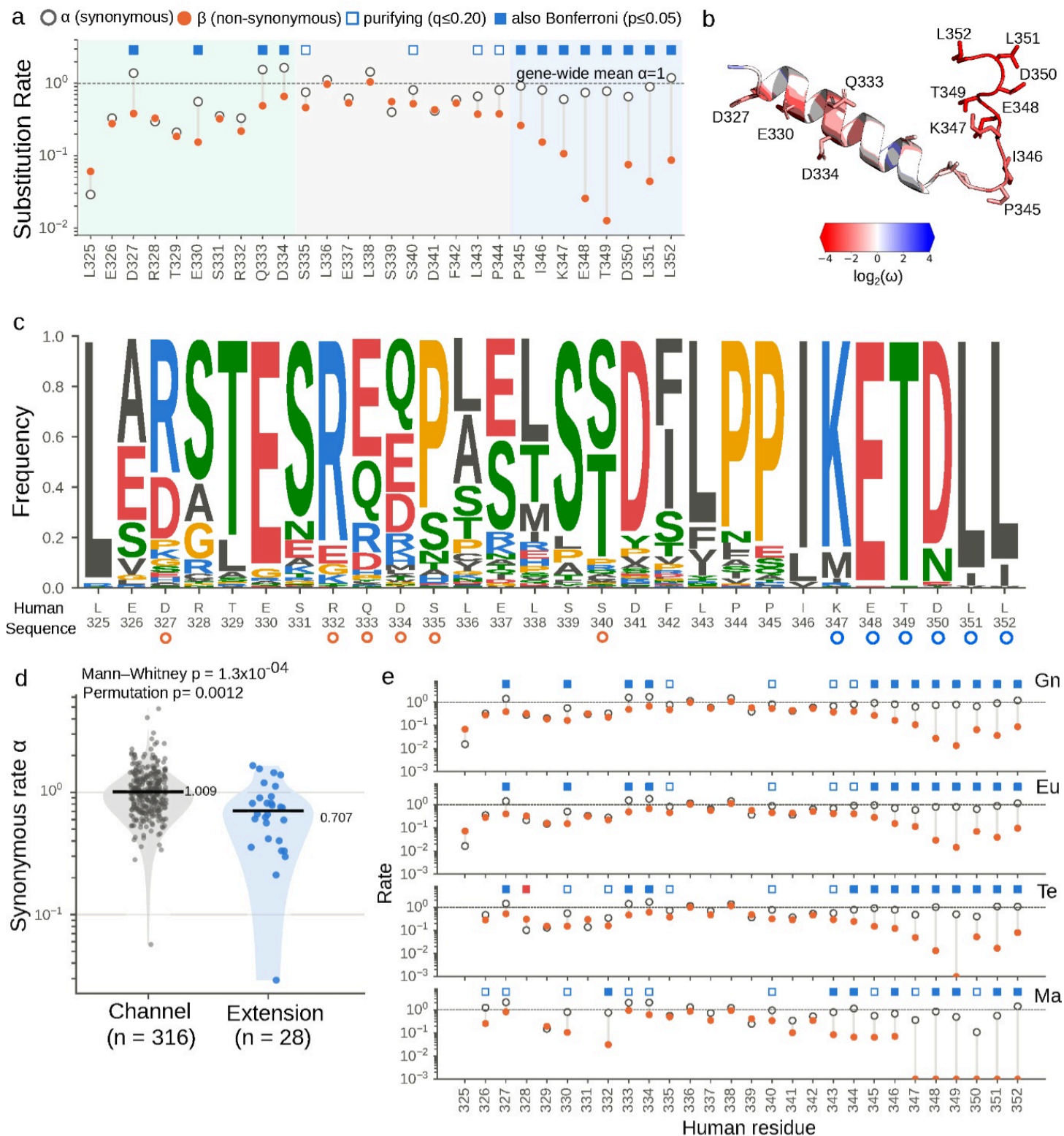


Figure 3

Site-wise selection on AQP4 extension. (a) Site-wise synonymous (α , open circles) and non-synonymous (β , filled orange circles) substitution rates from FEL fit to joint alignment of the channel domain and the extension corresponding to human residues L325-L352. Vertical grey lines join α and β at each site. The dashed line marks the gene wide mean synonymous rate ($\alpha=1$). Squares above the plot mark sites under purifying selection (open blue: $q \leq 0.20$, filled blue: $q \leq 0.20$ and Bonferroni significant). Shaded bands

distinguish N-terminal (L325-D334), Middle region (S335-P344) and C-terminal (P345-L352) (b) AlphaFold 3 predicted extension of AQP4 coloured by site wise $\log_2(\omega)$ from the fit in (a). Red indicates low ω (strong purifying selection) and blue indicated high ω (weak purifying selection). (c) Sequence logo of the 28 residue AQP4 extension. Letter height is the frequency of each residue at that position. The human residue and its position are provided below each column. Blue circles mark the six C-terminal positions at which the human residue is also the most frequent residue. Orange circles mark the positions where the human residues are not the most frequent. (d) Distribution of site-wise synonymous rate (α) for the 316 tested channel codons and the 28 tested extension codons from the same joint fit. Points are individual sites, the violin is the kernel density, and the horizontal bar is the median (channel: 1.009, extension; 0.707). Two-sided Mann-Whitney $p=1.3 \times 10^{-4}$, permutation test on difference in medians $p=0.0012$. (e) Site-wise synonymous (α) and non-synonymous (β) rate performed on nested taxonomic subset (Gn: Gnathostomata, Eu: Euteleostomi, Te: Tetrapoda, Ma: Mammalia). Symbols as in (a). the filled red square at codon 328 in the Tetrapoda panel marks the one site classified as under diversifying selection.

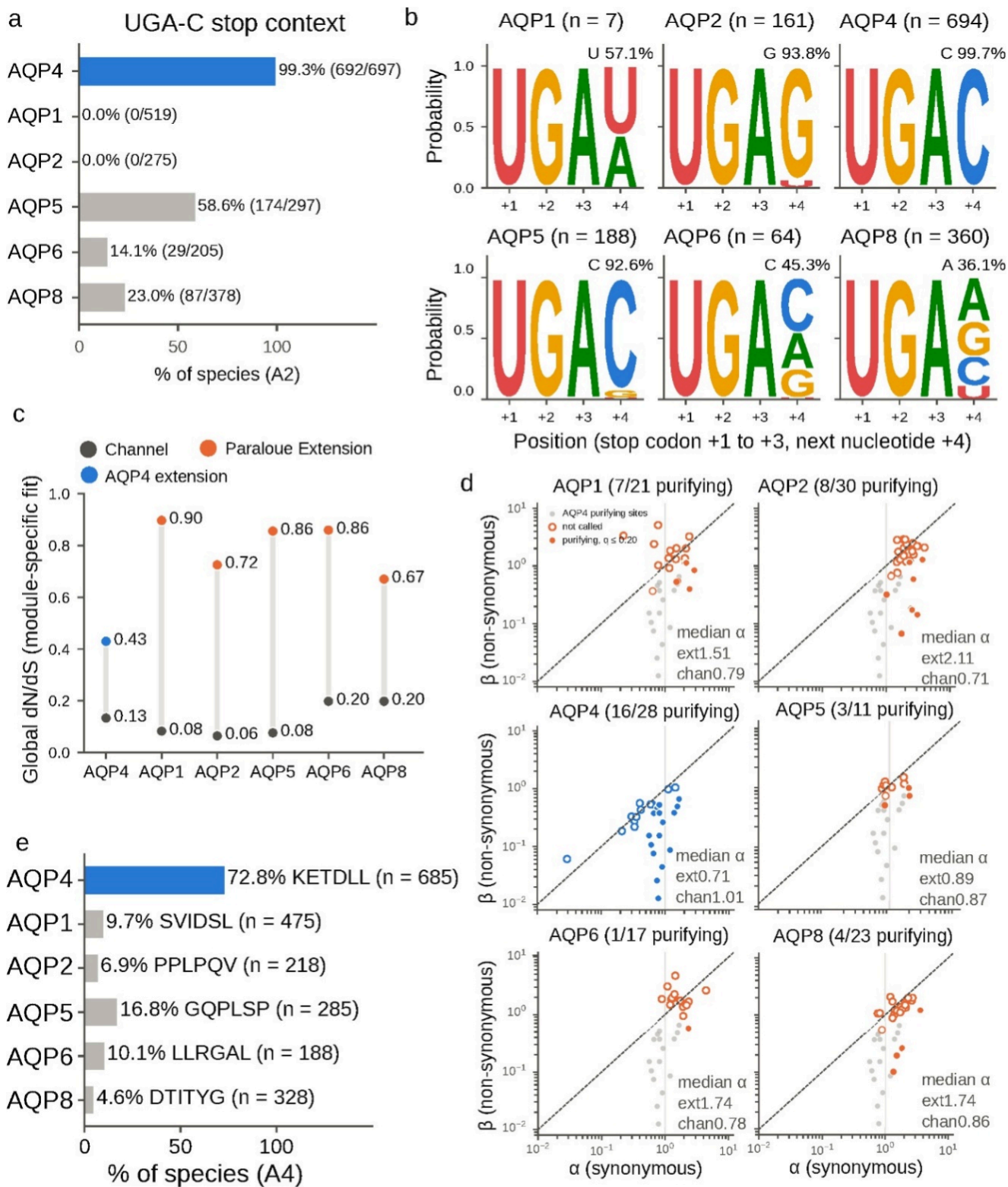


Figure 4

Four features distinguish AQP4 from five orthodox aquaporin paralogues (a) Percentage of species whose coding sequence ends in readthrough context UGA-C (b) Composition of stop codon (+1 to +3) and the following nucleotide (+4) shown as sequence logos. Each logo uses only the species carrying the UGA stop. (c) Global dN/dS from separate fits to the channel domain and to the downstream reading frame of each gene (d) Site-wise synonymous (α) against non-synonymous (β) rate at each tested codon of the

downstream frame. Filled symbols are sites under purifying selection ($q \leq 0.20$) and open symbols are sites not called. AQP4 purifying sites are represented in grey in each paralogue panel. The dashed line is $\beta = \alpha$. Panel titles give purifying over tested codons, and inner numbers give the median α of the downstream frame (ext) and of the channel (chan) domain (e) Percentage of species carrying the most frequent C-terminal hexapeptide of the downstream frame.

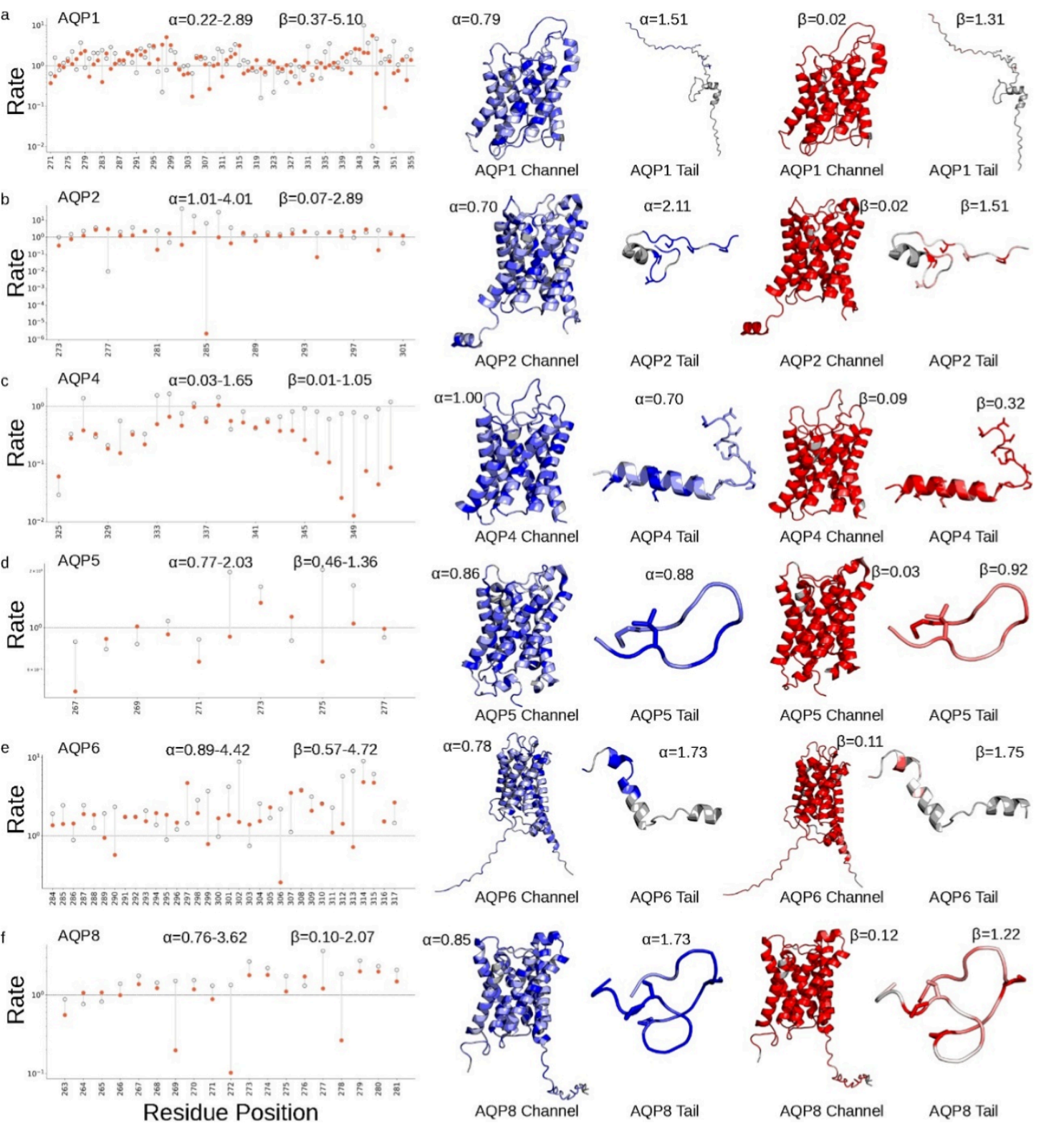


Figure 5

Site wise synonymous and non-synonymous rates in the channel domain and downstream frame of aquaporins. Rows show (a) AQP1 (b) AQP2 (c) AQP4 (d) AQP5 (e) AQP6 and (f) AQP8. Left panels show site-wise synonymous (α , open circles) and non-synonymous (β , filled orange circles) rates for each tested positions of the downstream frame. Vertical line joins the pair at each site. Right panel shows the AlphaFold 3 predicted structure of channel domain and the extension coloured by site-wise rates (blue: α and red: β)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryfiguresandtables.docx](#)