eLife

Reviewed Preprint
v2 • November 18, 2025
Revised by authors

Reviewed Preprint
v1 • May 9, 2025

Biochemistry and Chemical Biology
Computational and Systems Biology

Identification and classification of ion-channels across the tree of life provide functional insights into understudied CALHM channels

Rahil Taujale, Sung Jin Park, Nathan Gravel, Saber Soleymani, Rayna Carter, Kennady Boyd, Sarah Keuning, Zheng Ruan, Wei Lü ✉, Natarajan Kannan ✉

Institute of Bioinformatics, University of Georgia, Athens, United States • Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States • Department of Molecular Biosciences, Northwestern University, Evanston, United States • Department of Biochemistry & Molecular Biology, Thomas Jefferson University, Philadelphia, United States • Department of Pharmacology, Northwestern University, Evanston, United States • Chemistry of Life Processes Institute, Northwestern University, Evanston, United States

eLife Assessment

In this manuscript Taujale et al describe an interdisciplinary approach to mine the human channelome and further discover orthologues across diverse organisms. Further, this work provides evidence that supports a role for conserved residues in CALHM channel gating. Overall this **important** work presents findings that can be helpful to the ion channel community, as well as to those interested in improved methods for mining sequence space for their protein of interest. However, further validation of the improvements their approach shows over previous approaches is needed, making this a **solid** contribution to the literature in this field.

https://doi.org/10.7554/eLife.106134.2.sa3

Abstract

The ion channel (IC) genes encoded in the human genome play fundamental roles in cellular functions and disease and are one of the largest classes of druggable proteins. However, limited knowledge of the diverse molecular and cellular functions carried out by ICs presents a major bottleneck in developing selective chemical probes for modulating their functions in disease states. The wealth of sequence data available on ICs from diverse organisms provides a valuable source of untapped information for illuminating the unique modes of channel regulation and functional specialization. However, the extensive diversification of IC sequences and the lack of a unified resource present a challenge in effectively using existing data for IC research. Here, we perform integrative mining of available sequence, structure, and functional data on 419 human ICs across disparate sources, including extensive literature mining by leveraging advances in large language models to annotate and curate the full complement of the "channelome". We employ a well-established orthology inference approach to identify and extend the IC orthologs across diverse organisms to above 48,000. We show that the depth of conservation and taxonomic representation of IC sequences can further be translated to functional similarities by clustering them into functionally relevant groups, which can be used for downstream functional prediction on understudied members. We demonstrate this by delineating co-conserved patterns characteristic of the understudied family of the Calcium Homeostasis Modulator (CALHM) family of ICs. Through mutational analysis of co-conserved residues altered in human diseases and electrophysiological studies, we show that these evolutionarily-constrained residues play an important role in channel gating functions. Thus, by providing new tools and resources for performing large comparative analyses on ICs, this study addresses the unique needs of the IC community and provides the groundwork for accelerating the functional characterization of dark channels for therapeutic intervention.

Introduction

Ion Channels (ICs) are membrane-bound proteins critical to many physiological processes in the human body, including regulating cell volume, neurotransmitter release, muscle contraction, and glandular secretion [1–4]. Abnormal channel functions have been causally associated with "channelopathies" such as Parkinson's disease, epilepsy, cardiac arrhythmia, cancer, and cystic fibrosis, to name but a few [5–8]. The development of selective chemical probes for channels in disease states is currently hindered by the limited knowledge of the diverse gating mechanisms, ion selectivity, and pathway associations displayed by the various channels encoded in the human genome [8, 9]. While IC sequences from diverse organisms encode critical information regarding underlying functions and effective mining of sequence data can provide important context for predicting and testing understudied IC functions, traditional bioinformatic approaches have had limited success due to the challenges in consistently defining the full IC complement [10, 11], and accurately aligning and mining large sequence datasets [12]. Online resources like the collaborative platform Channelpedia [13], the Guide to Pharmacology (GtoP) database [14], structure centric ChanFAD [15] and ChannelsDB [16], and several other efforts have previously catalogued and curated ICs and provide extensive information on a number of ion channel sequences, but do not include all the human encoded ICs, nor provide information on ICs from other organisms and taxa [17–20]. Most IC studies across taxa are largely confined to closely related families or limited taxonomic groups [21, 22].

The extensive diversification and widespread abundance of ICs across cell types and their similarities to other transmembrane protein families, such as transporters, have led to an inconsistent definition of the human IC complement. For example, previous literature has reported around 230 human ICs [17]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database has cataloged around 316 human ICs [23, 24], the GtoP database lists 278 ICs [14], while Pharos [25] and The Human Genome Organization (HUGO) [26] list around 344 and 330 proteins, respectively. Moreover, these resources classify the human ICs based on their function, which is not known for many. Furthermore, current classification rarely accounts for the pore-forming or auxiliary roles of ICs. Channel proteins often form large macromolecular complexes in association with other IC and auxiliary subunits. Separate studies have highlighted the myriad of roles that the non-conducting auxiliary subunits play in an ion channel complex that control channel trafficking, gating and pharmacology [27–29], making them suitable therapeutic targets for drugs to regulate ion channel activity. Unfortunately, many of the existing resources fail to identify, include, and annotate such auxiliary channels, leading to their exclusion from large scale analyses for clinical outcomes. Due to these inconsistencies, downstream analyses have been hindered and limited to certain families or groups of closely related channels. Consequently, our current knowledge of IC functions is skewed towards a subset of well-studied "light" channels, while a significant portion of the channelome remains understudied and is referred to as "dark" channels by the Illuminating Druggable Genome (IDG) consortium [30, 31]. Concerted efforts such as IDG, and the Structural Genomics Consortium [29] help address these gaps in knowledge by systematically identifying and prioritizing such understudied members. However, it is imperative to develop a functional and an evolutionary classification framework that provides tools for researchers to extend knowledge from light members to their dark counterparts enabling them to make meaningful biological inferences and shed light on their critical roles in function and disease states.

The current classification of channel genes is largely based on experimentally determined electrophysiological parameters such as the type of ions they transport (Na^+, K^+, Ca^{2+}, Cl^-) and the channel gating mechanism such as gating by voltage potential (voltage-gated) or ligand binding (ligand-gated), which has been quite helpful in placing ICs in a functional context [14]. One of the first comprehensive classification of ICs was done by Jegla et. al. [17], where they described 235 human ICs distributed in 24 families, along with their conservation across select metazoan species, and family specific phylogeny and evolutionary analyses. A more recent collection of ICs along with their classification can be obtained from the GtoP database [14], a resource dedicated to providing detailed overviews of over 1900 human drug targets, spanning six major pharmacological targets – ICs being one of them. GtoP classifies 278 human ICs into 3 major groups (Voltage-Gated Ion Channels (VGICs), Ligand-Gated Ion Channels (LGICs), and Other) and 31 families, and provides detailed overview of these channels in a friendly interface. However, it does not include a lot of now well-known ion channel families such as Calcium Homeostasis Modulator channels (CALHMs) or Bestrophins (Best). Additionally, the resource being a general drug target resource, is not IC-centric, and lacks evolutionary and residue level functional insights that might help hypotheses generation. The role or annotation for auxiliary channels, that play critical roles in function and regulation of ICs [27, 32] have not been addressed in these resources either.

On the other hand, cryo-electron microscopy (cryo-EM) studies and analysis of IC structures have shown that despite extensive variation in primary sequences, several channels adopt common structural folds and mechanisms of action [33]. For example, several ICs - Potassium (K+), Sodium (Na+), Calcium (Ca2+), Transient Receptor Potential (TRP), and Hyperpolarization-activated Cyclic Nucleotide-gated (HCN), grouped together as voltage-gated-like (VGL), are proposed to have evolved from a common ancestor because they share key sequence and structural features [34, 35]. Thus, quantitative comparisons of these commonly conserved regions at the sequence and structural level can provide new residue-level insights into IC function and a stronger basis for functional classification, much like in other large protein families such as kinases and glycosyltransferases [36–39]. This is imperative since selective targeting of channels in diseases and a mechanistic understanding of disease-associated mutations in channelopathies will require a residue-level understanding of function-determining features in both well-studied “light” and understudied “dark” channels. Thus, there is a critical need for an updated classification that places ICs in functionally, and evolutionarily related groups, enabling a deeper residue level analysis of light and dark channels alike.

To develop a comprehensive classification of ICs, we first mined literature and sequence data sources to collect information on ICs stored across various databases and in disparate data sources and formats to build a well-curated knowledgebase towards defining the human ion “channelome.” A comprehensive list encompassing established and putative IC sequences in the human proteome is provided as a unified table alongside other contextual data for functional annotations such as regulatory interactions, ion-selectivity, and pore-lining residues. Leveraging the identification of a pore region and pore-lining residues, we make a clear distinction between pore-containing versus auxiliary ICs. Based on this curation, we then perform a comprehensive classification of ICs into well-defined groups and families and identify the extent of understudied channels in each family, providing a premise for prioritizing studies in IC families with more understudied members.

We further performed a comprehensive evolutionary analysis to define and collect more than 48,000 well-defined IC orthologous relationships from diverse taxonomic groups spanning metazoans, fungi, protists, bacteria, and archaea. We demonstrate the application of these sequences in annotating understudied CALHMs at residue-level resolution using Bayesian statistical approaches [40] and in predicting and experimentally validating the structural and functional impact of disease-associated mutations. These studies support our working premise that integrative mining of available sequence, structure, and functional data on ICs from diverse organisms (well-studied and understudied) will provide an important context for defining sequence and structural features associated with understudied channel functions.

Results

Defining the IC complement of the human genome using informatics approaches

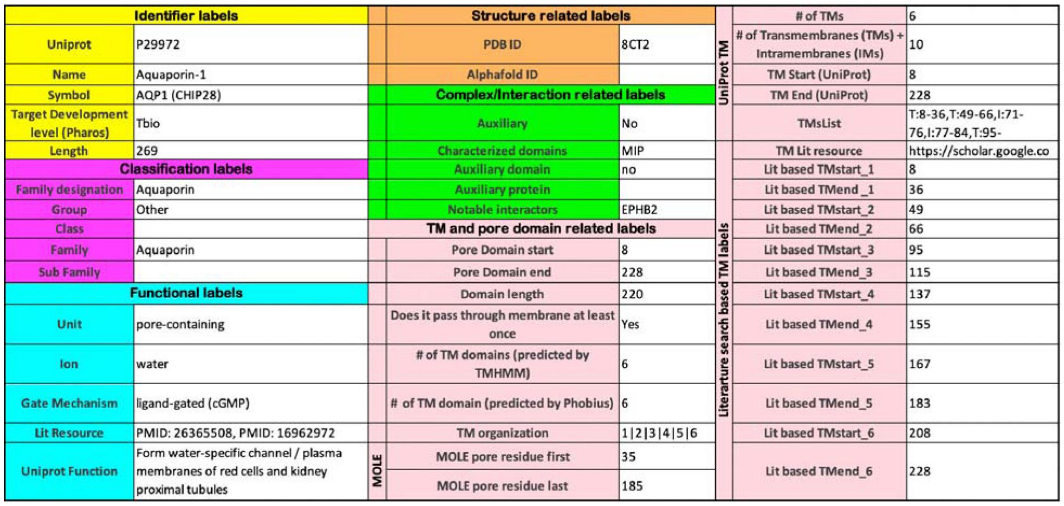
We first sought to collect and curate the full IC complement across the human proteome by systematically mining all the existing data sources, collating information from literature, and running various informatics tools (see methods) to identify related sequences based on overall similarity and pore-defining regions. In brief, we first collected all the sequences listed as ICs in the comprehensive UniProt database [41], as well as Pharos [25], KEGG [23, 24], GtoP [14], and HUGO Gene Nomenclature Committee (HGNC) [26] databases and subjected them to a series of annotations as described in **Table 1**. The complete annotation table for all human ICs is available in Supplementary Table 1. Primarily, we mined the literature for functional and structural annotations on these sequences. Broadly, we have included 6 annotation categories. 1) The identifier labels include the UniProt ID, name, and Target Development Level (TDL) designation from Pharos (as of 3 July 2025). The IDG consortium, through Pharos, assigns TDL, which can be Tclin, Tchem, Tbio or Tdark, based on the literature data available for proteins in terms of their potential as drug targets. Including this designation in the annotation is especially relevant for identifying and prioritizing understudied ICs for further investigation. 2) The classification labels describe the Group, Class, and Family the IC falls into. This information has been mined from literature including UniProt, Pharos, GtoP, and HGNC, and the field “Family designation” provides a consensus family label for the IC. 3) The functional labels have been manually mined from various sources, relying mostly on previous literature in the “Lit Resource” column. An important column in this section is the “Unit” column, which describes whether the IC is directly involved in forming the pore region, where it can be pore-forming (directly involved in forming an ion-conducting pore), 2-pore (contains 2 tandem pore-conducting regions), or auxiliary (does not have its own pore domain but interacts with other pore conducting IC subunits as part of a complex) (please see below and Methods for further definition of an auxiliary IC). We also provide the ions, gating mechanism, and UniProt functional description in this annotation category. To verify the ion and gating mechanism annotations, we mined the literature using Large Language Models (LLM) and Retrieval Augmented Generation (RAG) systems (Methods, Supplementary Figure 1). The RAG system was able to successfully verify about 40% of the ion and gating mechanism annotations, while for others, it was unable to find supporting evidence, either because no evidence was available in the existing literature or no relevant references were found. The annotations that were not verified by the RAG system are indicated with an asterisk in Supplementary Table 1. 4) The structure-related category includes the PDB ID of any experimentally generated crystal structure or an Alpha-Fold ID for a computationally predicted structure of the IC. 5) The Complex/Interaction label provides information on the types of complex the IC is involved in forming, along with the interactors. 6) The last section for transmembrane (TM) and pore domain-related labels has been curated extensively using various sources and predictors. The TM regions and pore containing functional domains are annotated based on literature references and prediction tools, as outlined in the Methods section. TM predictions by TMHMM [42] and Phobius [43] are provided alongside the TM annotations provided in

UniProt. First, we supplemented this information by adding TM information based on the literature review and the source. Then, we further analyzed any ICs with an experimentally determined structure using the MOLE software [44] to identify and annotate the pore region and pore-lining residues. The MOLE software starts by defining the membrane region of the protein, followed by the identification of cavities and computation of the pore boundaries to predict the pore lining residues. These predictions are used as additional evidence for defining the pore containing functional domain region and a pore containing IC sequence. Sequences without an identified pore region were classified as auxiliary, as defined by Gurnett et al. 1996 [45], to provide functional context for this classification. A snapshot of the list of annotations we collected through this process is shown in **Table 1**, using Aquaporin 1 as an example.

In the final phase, we conducted sequence similarity searches to detect sequences exhibiting significant homology with any curated sequences, subsequently subjecting these candidates to all annotation procedures to determine their eligibility as ICs. By combining the results from this annotation process, we were able to achieve the following: 1) compile a list of human ICs based on the presence of distinct transmembrane regions, a detectable pore, and evidence of ion conductivity, 2) distinguish between a pore containing IC vs an auxiliary IC, 3) identify novel putative IC sequences and 4) curate and classify the full IC complement in the human proteome. These results are summarized in **Figure 1** and Supplementary Table 1.

419 human ion channel (IC) sequences were curated using this approach, representing an increase of 75 sequences compared to the previous consensus of 344 ICs in humans [8[25]]. A comparison of the curated IC sequences in this study against the list of ICs in other resources – KEGG, GtoP, and Pharos, is provided in Supplementary Figure 2 and Supplementary Table 2. These sequences were classified into four major groups—VGICs, LGICs, Chloride Channels, and Others—defined previously [12], and 55 families. VGICs constitute the largest group, comprising 186 sequences distributed across 21 families, followed by LGICs with 82 sequences in 10 families. Of the 419 sequences, 62 could not be assigned to the four major groups and were categorized into outlier families. Among these, 28 are pore-containing ICs, with 19 of the 28 distributed across four families (CALHM, otopetrin (OTOP), transmembrane channel-like (TMC), and tweety homolog (TTYH)). These families, collectively referred to as “Unclassified” families, are labeled as such in **Figure 1**. The remaining 9 pore-containing sequences could not be assigned to a distinct family and were grouped together under a single “Unclassified” family. Based on the evaluation of the pore-containing regions, 343 out of 419 ICs were annotated as pore-containing ICs, while the remaining 76 were classified as auxiliary and fell into 17 different families. 23 of these auxiliary ICs are soluble with no detectable transmembrane domains. Any auxiliary ICs that did not belong to a distinct family were collectively classified into an “Auxiliary unclassified” family.

Next, we sought to use the curated set to define similarities across IC families. Traditional sequence-based bioinformatics approaches, such as pairwise sequence alignment (e.g., BLASTp) [46] or profile hidden Markov models (e.g., HMMER) [47], are limited in their ability to detect relationships among human ICs due to the extensive divergence in their primary sequences. These methods often fail to identify homologous ICs across different families, as the sequence similarity falls below detectable thresholds. Consequently, such approaches are inadequate for comprehensive classification or comparison of ion channel families in humans. Instead, we relied on representations (embeddings) derived from evolutionary scale protein language models [48] to capture sequence, structure, and evolutionary information and use them to generate pairwise sequence alignment. Specifically, for the 343 pore-containing ICs, we passed their pore-containing functional domains to a protein sequence embedding model called DEDAL [49] to perform a pairwise sequence alignment. Given that sequence similarity is largely restricted to the pore-containing functional domains of ICs, we computed protein embeddings using only these regions. As auxiliary ICs lack such domains, they were excluded from this part of the analysis. The resulting all-vs-all sequence similarity scores were used to generate uniform manifold approximation and projection (UMAP) embedding scores (Supplementary Figure 3). An average of



Identifier labels			**Structure related labels**			# of TMs	6
Uniprot	P29972		PDB ID	8CT2	UniProt TM	# of Transmembranes (TMs) + Intramembranes (IMs)	10
Name	Aquaporin-1		Alphafold ID			TM Start (UniProt)	8
Symbol	AQP1 (CHIP28)		**Complex/Interaction related labels**			TM End (UniProt)	228
Target Development level (Pharos)	Tbio		Auxiliary	No		TMsList	T:8-36,T:49-66,I:71-76,I:77-84,T:95-
Length	269		Characterized domains	MIP	Literarture search based TM labels	TM Lit resource	https://scholar.google.co
Classification labels			Auxiliary domain	no		Lit based TMstart_1	8
Family designation	Aquaporin		Auxiliary protein			Lit based TMend _1	36
Group	Other		Notable interactors	EPHB2		Lit based TMstart_2	49
Class			**TM and pore domain related labels**			Lit based TMend_2	66
Family	Aquaporin		Pore Domain start	8		Lit based TMstart_3	95
Sub Family			Pore Domain end	228		Lit based TMend_3	115
Functional labels			Domain length	220		Lit based TMstart_4	137
Unit	pore-containing		Does it pass through membrane at least once	Yes		Lit based TMend_4	155
Ion	water		# of TM domains (predicted by TMHMM)	6		Lit based TMstart_5	167
Gate Mechanism	ligand-gated (cGMP)		# of TM domain (predicted by Phobius)	6		Lit based TMend_5	183
Lit Resource	PMID: 26365508, PMID: 16962972		TM organization	1|2|3|4|5|6		Lit based TMstart_6	208
Uniprot Function	Form water-specific channel / plasma membranes of red cells and kidney proximal tubules	MOLE	MOLE pore residue first	35		Lit based TMend_6	228
			MOLE pore residue last	185			

Table 1

List of features annotated for the collected IC sequences.

The labels are colored by their annotation category. Labels for Aquaporin 1 are shown as examples of each annotation label.

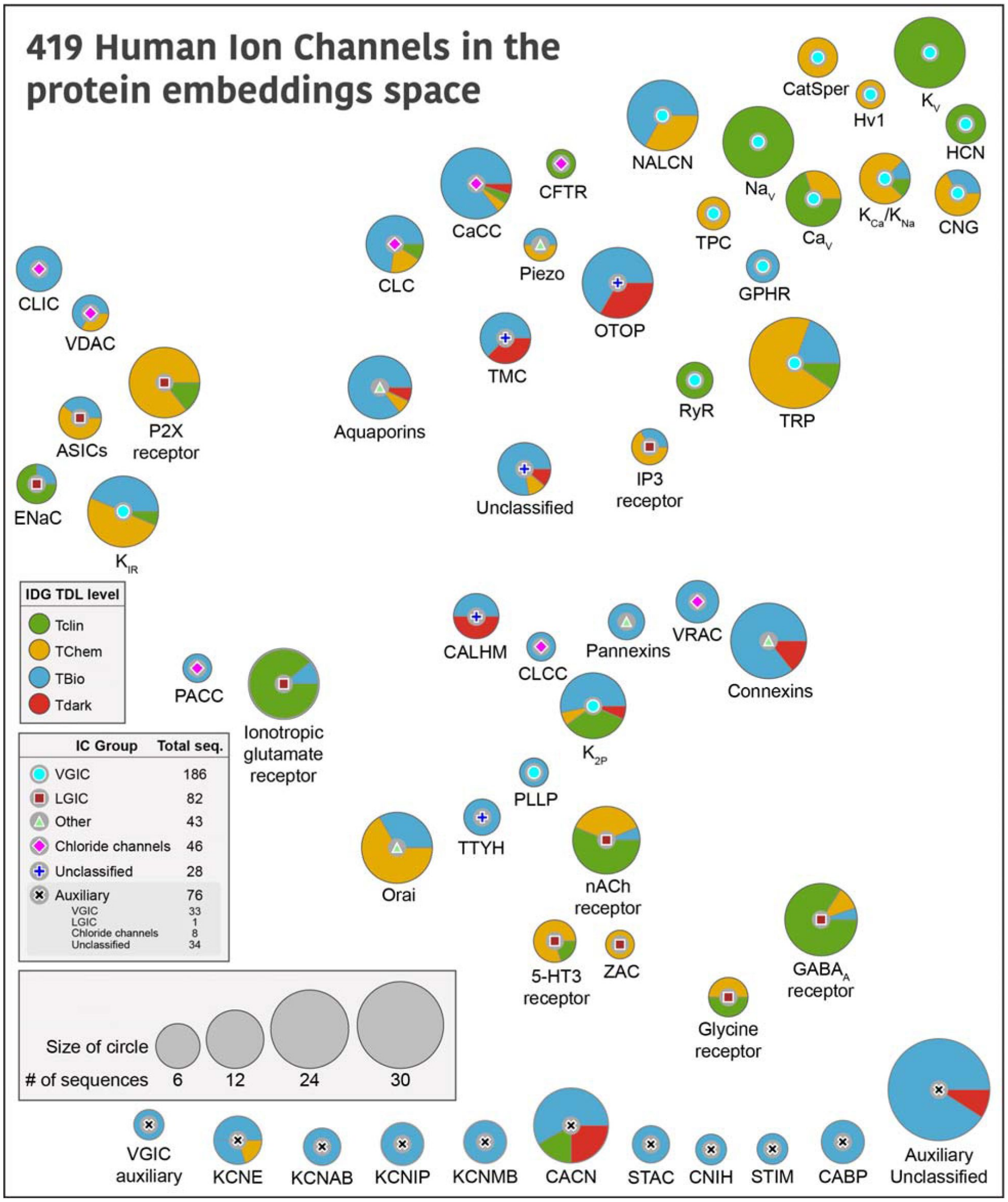


Figure 1

Distribution of human ICs across different families.

Each circle represents a human IC family, with the symbol at the center indicating its Group. The size of the circles is proportional to the number of sequences in that family, and the colored pies indicate the proportion of their TDL status as designated by IDG. The placement of the bubbles is based on the average co-ordinates of all the members within that family in a distribution of UMAP embeddings generated using protein embedding-based pairwise sequence alignments. An embedding-based sequence alignment approach was used to overcome the vast divergence of IC sequences with minimal sequence similarity. A family level abstraction was done to provide an intuitive view undeterred by the relationships of individual ICs across families. A detailed view of the full UMAP plot showing the placement of individual ICs is provided in Supplementary Figure 3 for reference. Auxiliary IC families at the bottom were not part of the embedding-based analysis due to the lack of a pore-containing domain, and thus placed arbitrarily at the bottom of the figure.

all IC sequences within a family was calculated to place the IC family bubbles in **Figure 1**. Thus, families placed close to each other in **Figure 1** indicate their similarities in sequence embeddings, which capture learned representations of sequence, structure, evolutionary conservation, and functional properties. Since the auxiliary IC families did not have UMAP embedding scores, they are arbitrarily placed at the bottom of the figure.

Most of the VGIC families group together on the top right except for inward rectifier potassium channels (K_{IR}), two-pore domain potassium channels (K_{2P}), and plasmolipin (PLLP), where K_{2P} and PLLP are more centrally dispersed, while K_{IR} falls closer to LGIC families acid-sensing ion channels (ASICs), epithelial sodium channel (ENAC), and P2X purinoceptors, all of which share similar transmembrane topology with only two transmembrane helices. The inositol 1,4,5-trisphosphate (IP3) and ionotropic glutamate receptors are also centrally located, while the Cys loop receptor LGIC families group separately from others towards the bottom right, indicating their shared functional and structural similarities. Interestingly, pannexins, volume-regulated anion channels (VRAC), connexins, chloride channel CLIC-like (CLCC), and CALHM families group closer together centrally, indicating shared features.

We also mapped TDL annotations defined by the IDG [30] in Pharos [25] to highlight the depth of understudied ICs within each group, and hopefully, spark additional studies of those members. Briefly, IDG describes Tclin as targets that have an approved drug; Tchem has known high potency small molecule binding targets; Tbio has experimentally supported Gene Ontology [50] annotations based on published literature; while Tdark is manually curated at the primary sequence level in UniProt, but do not meet any of the Tclin, Tchem or Tbio criteria [25]. There are 19 Tdark, 185 Tbio, 88 Tchem and 127 Tclin ICs spread across all families. The proportion of these levels for each family is shown as pie charts in **Figure 1**. CALHM family has the highest proportion of Tdark ICs, with OTOP, TMC, Connexins, calcium-activated chloride channels (CaCC), calcium voltage-gated channel (CACN), $K_{2P,}$ and the unclassified families accounting for the remaining Tdark ICs, emphasizing the need for additional research on these families. On the other hand, VGICs voltage-gated potassium channel (K_V), voltage-gated sodium channel (Na_V), voltage-gated calcium channel (Ca_V), ryanodine receptor (RyR), and LGICs Gamma-Aminobutyric acid type A ($GABA_A$) receptor, nicotinic acetylcholine (nACh) receptor, ENAC have the highest proportion of Tclin ICs, solidifying their status as the well-studied families.

Mining and cataloging IC orthologs across the tree of life

Next, we sought to extend this annotation and collect related ICs from other organisms across different taxonomic lineages from the tree of life [41] using the full complement of the annotated human IC sequences. To this end, we used a graph-based orthology inference approach [51] starting from the full length and the pore-containing functional domains of the human IC sequences to define orthologous relationships across more than 1,500 proteomes from the UniProt Proteomes database [41]. These selected proteomes dataset includes 353 Archaea, 696 Bacteria, and 547 Eukaryota organisms. This method has previously been successfully employed to define orthology across such a large collection of proteomes for protein kinases [51], and since it relies on both the full-length sequence and the annotated pore-containing domains, it can more accurately define true orthologs that are indeed evolutionarily related across organisms. Since this method relies on the pore containing domain annotations for a domain-based orthology inference, we only used the 343 pore-containing IC sequences for this analysis, and did not include the auxiliary ICs that lack a pore containing domain.

By combining domain and full-length searches, we could identify more than 48,000 IC orthologous relationships across diverse organisms. **Figure 2A** shows a heatmap depicting the phylogenetic profiling of human IC orthologs along the horizontal axis across the taxonomic lineages shown by the tree along the vertical axis. The color intensity indicates the percentage of orthologs detected for all the sequences in the IC family for a given taxonomic group. The highest number of orthologs were found for Aquaporins, including bacterial and archaeal orthologs, indicating that this family of ICs is the most conserved across lineages. Along with Aquaporins, the two paralogs of Golgi pH regulator ICs (GPHRs), both indicated as Tbio channels, were found to have the largest number of orthologs across metazoans. Along with GPHRs, several other ICs, notably members of the K_V family and $GABA_A$ receptors, display some of the most widespread orthologs across organisms. Some dark channels, such as members of the TMC (TMC7 and TMC3) and CALHM (CALHM5 and CALHM3) families, are well conserved across metazoans and vertebrates, respectively, yet very little is known about their functions. On the other hand, some Tdark channels, such as Potassium channel K member 7 (KCNK7) and TMC4 have a lineage-specific set of orthologs that extend only up to mammals. A full list of all the orthologs detected through this analysis is provided in Supplementary Table 3.

To use this evolutionary conservation for functional inference of dark ICs, we first used hierarchical clustering to group ICs with similar orthology profiles into related clusters, which led to the definition of 9 clusters, as shown in **Figure 2B**. Since the orthology profiles in prokaryotic lineages were very sparse, only the eukaryotic orthology profiles were retained for this analysis. Within eukaryotes, each cluster has a well-defined signature of orthology conservation. For example, ICs in cluster 2 have most of their orthologs only in mammals, whereas clusters 4, 5, and 6 have orthologs spanning other vertebrates. Similarly, clusters 7 and 8 have orthologs extending to other metazoans, while seven ICs in cluster 9, including the 2 Golgi pH regulators A and B and the sodium leak channel NALCN, have detectable orthologs in Fungi, Plants, Protists, and other eukaryotic lineages. To translate these patterns of orthology profiles to function, we performed a functional Gene Ontology (GO) term enrichment analysis for the human ICs in each cluster (**Figure 2C**). As expected, the most significant GO terms were related to ion channel function (colored green in Supplementary Table 4). However, beyond channel function, the enriched functional GO terms obtained for each cluster correlate well with the physiological functions present in the orthologous organismal groups. For example, clusters 4, 5, and 6 are conserved in vertebrates and have six dark IC sequences, including unclassified ICs such as CALHM 3 and 5, and connexins GJA10, GJD4, and GJE1, which had functional enrichment for regulation of heart contraction and heart rate, traits that could be closely related to a closed blood circulatory system with an endothelium, present in vertebrates [52]. On the other hand, cluster 2 with orthologs in mammals was conserved in mammalian-specific reproductive functions such as sperm capacitation.

Using the orthology information to identify evolutionary constraints in CALHMs

Once the orthologs have been identified across different organisms, they can be leveraged to find evolutionarily conserved signals that could point to functional similarities within related groups of sequences. To this end, we classified subsets of orthologous IC sequences into evolutionarily related clusters using a Bayesian Partitioning with Pattern Selection (BPPS) algorithm, which classifies sequences based on patterns of amino acid conservation and variation in a large multiple sequence alignment (see methods) [40]. For this analysis, we focused on the orthologs of CALHM family of IC sequences. The CALHM proteins constitute a family of large pore channels, forming oligomeric assemblies of different sizes [53]. It has 6 member sequences in humans (CALHM1-6), 3 labeled as dark channels, constituting one of the families with the highest prevalence of dark ICs. It has been well established that CALHM1 is activated by removing extracellular Ca^{2+} and membrane depolarization and that the heteromeric CALHM1 and CALHM3 channels are implicated in the ATP release during taste perception [54]. CALHM1 has also been shown to regulate cortical neuron excitability [55], locomotion and induces neurodegeneration in *C. elegans* [56]. On the other hand, CALHM6 has been shown to be important for immune system functions by facilitating induction of immune cells during infection [57]. In contrast, the activation stimuli and physiological roles of other family members remain largely unexplored. All CALHM family members share a common arrangement in the transmembrane domain, with each subunit consisting of four transmembrane segments (S1–4), forming a large cylindrical pore lined

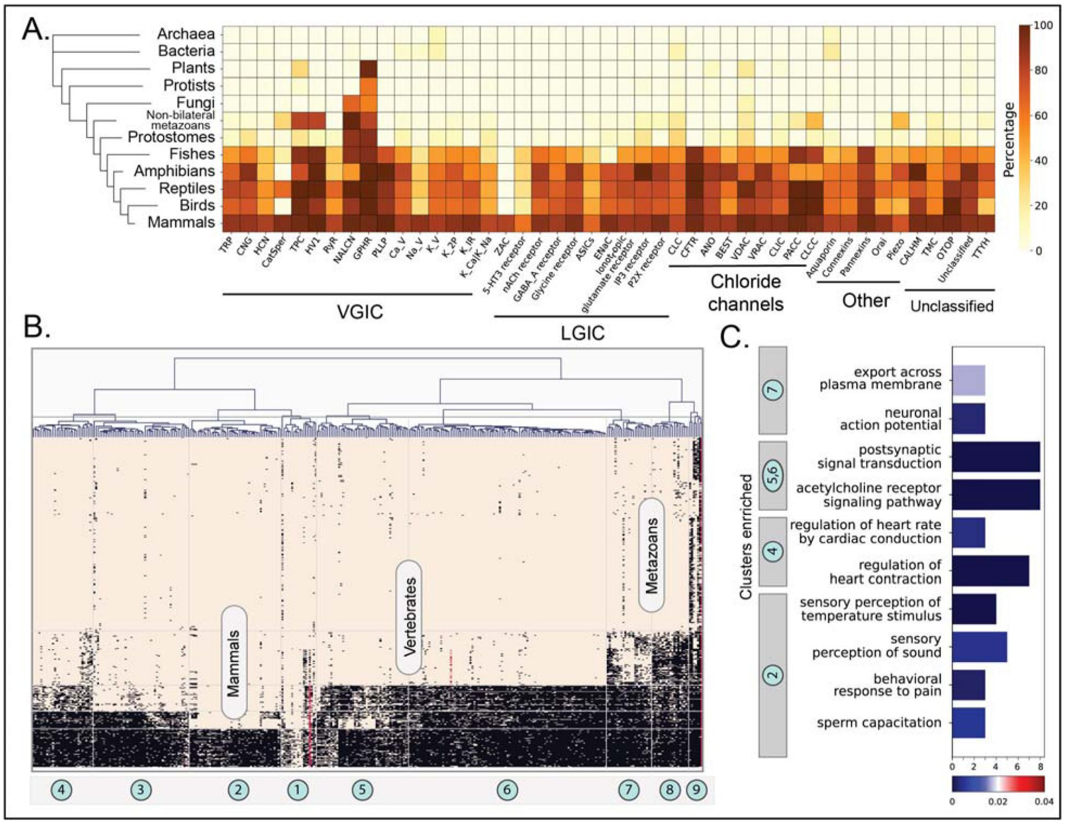


Figure 2

Orthology profiling of human ICs.

(A) Heatmap showing the percent of orthologs detected for each IC family within a given taxonomic lineage. The taxonomic groups are shown in the vertical axis with a tree on the left. Darker color represents a higher percentage of orthologs detected. Percentages were calculated as (total number of orthologs found for all ICs in a family)/(total number of organisms queried in the taxonomic lineage * number of sequences in the family) (B) Clustergram depicting the presence/absence of orthologous sequences of ICs across eukaryotic taxonomic lineages. ICs are clustered along the horizontal axis into 9 distinct clusters. Taxonomic groups are shown on the vertical axis. Each square in the heatmap is colored based on the orthology relationship found for a specific IC in a specific organism (black: one-to-one ortholog present, red: co-ortholog detected, brown: no orthology detected). (C) Results from the enrichment analysis performed on human ICs of each cluster. The x-axis shows the number of ICs in the cluster enriched for the GO term shown on the y-axis. The bars are colored based on their FDR values for the enriched term. For a full list of enriched terms, please refer to Supplementary Table 4.

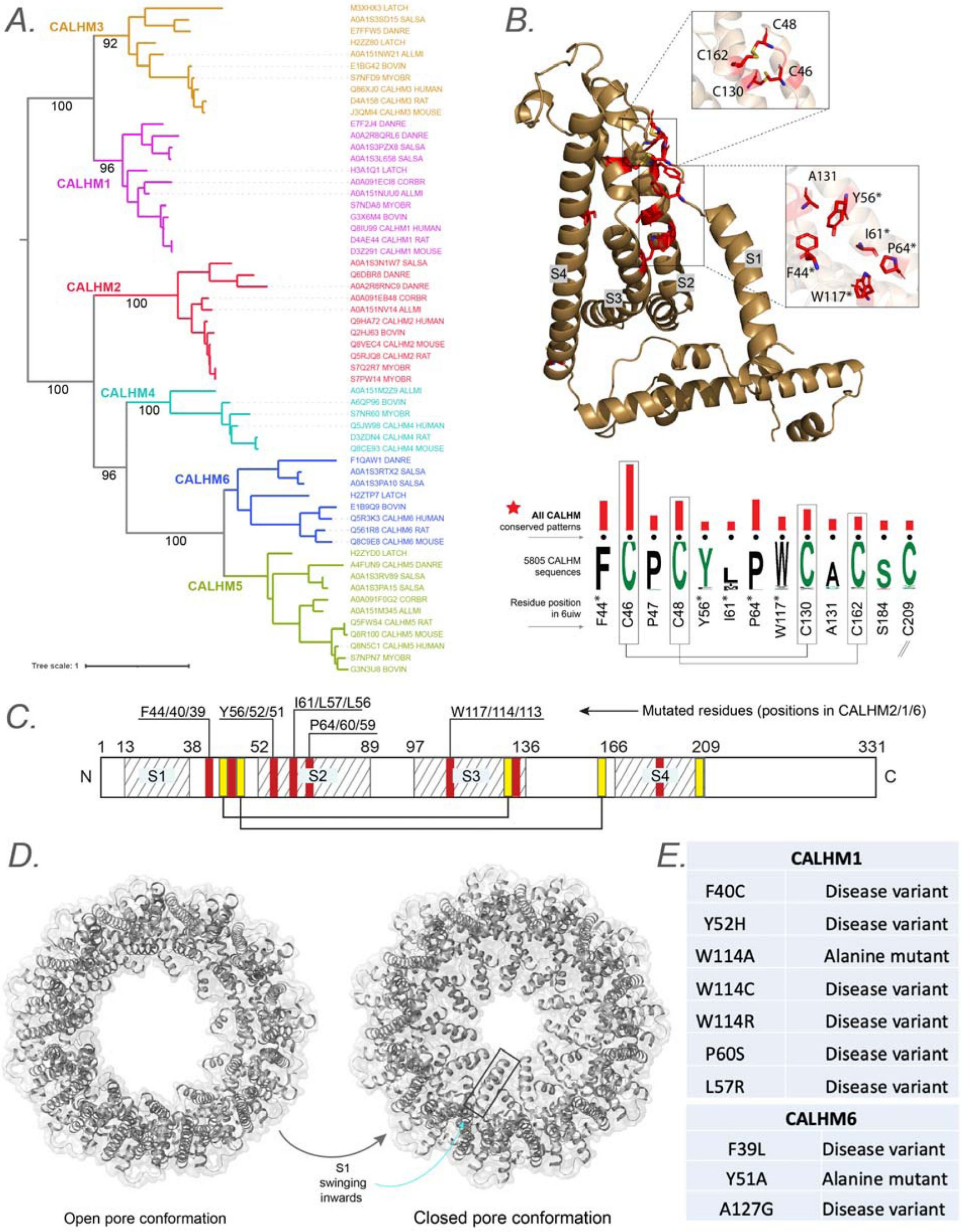
by the first transmembrane segment S1 along with a short N-terminal helix preceding S1. The exact gating mechanism of the CALHM channels is still unclear, partly since all their cryo-EM structures have been in an open state. However, previous studies have indicated that the N-terminal region, called the amino-terminal helix (NTH) plays a crucial role in modulating voltage dependence and stabilizing the closed channel state [58], while a more recent study has suggested that the voltage dependent gate is formed by the proximal regions of S1 [59]. Collectively, these studies indicate the role of the amino termini – the NTH and S1 regions, to play a crucial role in the gating mechanism and the determination of pore size [60] of CALHMs.

Thus, we performed a phylogenetic analysis to find evolutionarily conserved residues that might shed more light on these critical mechanisms that govern CALHM function. **Figure 3A** shows a phylogenetic tree depicting the evolutionary relationship across the 6 members of the CALHM family across diverse taxa. CALHM1 and 3 form a distinct clade from CALHM 2, 4, 5 and 6. We first analyzed 5805 CALHM homologs to identify pattern positions conserved across all these sequences with the hypothesis that such conserved positions could point to shared functional features across all CALHMs. The Bayesian analysis identified 13 aligned positions conserved across all 6 CALHM homologs, which we will refer to as CALHM shared patterns (**Figure 3B,C**). Most of these conserved positions were hydrophobic amino acid residues, and 5 conserved Cystine residues, 4 of which are involved in forming inter-molecular disulfide bridges (C46=C130, C48=C162). Residue position numbers reflect the numbering based on human CALHM2 (PDB id: 6uiv). Interestingly, five of the conserved positions (F44, Y56, I61, P64, W117) are located close to the amino termini in a functionally important linker region connecting S1 and S2 (S1-S2 linker). Specifically, this linker region could regulate the dynamic conformational changes of S1, where the S1 could adopt either a vertical conformation relative to the membrane plane, resulting in an enlarged pore size, or a lifted conformation, leading to a reduced pore size (**Figure 3D**) [60]. Therefore, mutations in this linker are expected to affect channel functions. Based on the positioning of these conserved residues, and previous studies that highlight the importance of this region in gating functions, we hypothesized that these conserved pattern positions play a role in the gating mechanism of CALHM. The conservation of these residues across orthologs of all 6 CALHM sequences further suggests that all CALHM paralogs could share this gating mechanism.

To test our hypothesis and determine the functional importance of the CALHM shared patterns, we sought to perform a series of mutational experiments to determine the functional implications of perturbations at these positions. To achieve this, we methodically determined target mutations for each position. We first scanned the Genome Aggregation Database (gnomAD) [61] to check for any prevalent variations at these conserved positions within the sampled population to use as our mutational targets. We found several disease variants at these positions that are listed in **Figure 3E** that were used to prioritize target mutations. For positions where a variation was not found, we tested their significance by performing alanine mutations, causing a deletion of the side chain at the β-carbon.

Targeted mutational and electrophysiological studies of CALHM conserved residues

To assess the functional roles of the predicted conserved residues in CALHM channels, we performed targeted mutational and electrophysiological analyses in two representative human CALHMs: the well-studied human CALHM1 and the relatively understudied human CALHM6. Building on the insights from previous structural and functional studies [59, 60], which implicated the NTH/S1 region in channel gating, we prioritized the constraints located in the S1-S2 linker and at the interface between S1 and the transmembrane domain (**Figure 3B,C**,E). These two regions were hypothesized to play distinct roles in gating: the S1-S2 linker as a flexible hinge and the S1-TMD interface as a stabilizing contact for S1 movement.



CALHM1	
F40C	Disease variant
Y52H	Disease variant
W114A	Alanine mutant
W114C	Disease variant
W114R	Disease variant
P60S	Disease variant
L57R	Disease variant
CALHM6	
F39L	Disease variant
Y51A	Alanine mutant
A127G	Disease variant

Figure 3

Evolutionary analysis of CALHM reveals conserved pattern positions.

(A) A phylogenetic tree depicting the evolutionary relationships across all 6 CALHM members with orthologs across different taxa. (B) The conserved pattern positions conserved across all 6 CALHM members are shown as a weblogo with the red bars indicating the significance (longer bars indicate higher significance) of conservation and are mapped into a representative structure of human CALHM2. The four transmembrane helices are labeled S1-S4. Conserved residues that are targeted for mutation are highlighted using a asterisk symbol in the labels. (C) Schematic representing the location of transmembrane regions and identified conserved pattern positions in a representative human CALHM2 sequence. The residues targeted for mutations are labelled with their corresponding positions for CALHM2, 1 and 6 shown in the labels. (D) Cartoon representation of CALHM2 structure (PDB ID: 6uiv and 6uiw) in open and closed conformation, respectively. (E) List of disease variants and mutations performed in the conserved pattern positions for functional studies.

Both total protein Western blotting and surface biotinylation assays showed that the mutants have either comparable or substantially higher expression levels than wild type, confirming that neither protein expression nor trafficking to the plasma membrane was impaired (Supplementary Figure 4). We then established electrophysiological measurements for wild-type CALHM1 and CALHM6 following protocols described in the methods. Our results were consistent with previous findings [55] (**Figure 4**). In addition to room temperature, which is commonly used for patch-clamp studies of CALHM channels, we also conducted measurements at physiological temperatures (37°C). We observed that CALHM1 activity was significantly higher at physiological temperature than at room temperature, as reported previously [62, 63], while the CALHM6 currents showed only a small increase (**Figure 4**). Notably, both CALHM1 and CALHM6 currents were inhibited by extracellular Gd^{3+} (**Figure 5C,I**), a commonly used inhibitor for the CALHM family, consistent with previous studies [55, 57, 59]. The robust currents at 37°C, particularly those of CALHM1, provide a solid basis to interpret the phenotypes of the mutants tested, as detailed below.

Mutations of a predicted conserved residue in the S1–S2 linker (F40 in CALHM1; F39 in CALHM6) either abolished or markedly reduced channel activity in both CALHM1 and CALHM6, presumably by impeding the conformational dynamics of S1 required for channel gating (**Figure 5P,Q**; Supplementary Figure 5). Similarly, mutation of a conserved tyrosine residue on S2 (Y52 and Y51 in CALHM1 and CALHM6, respectively), whose side chain directly contacts the S1-S2 linker, also resulted in strong phenotypic changes: Y52A in CALHM1 abolished channel activation (**Figure 5P,Q**; Supplementary Figure 5), while Y51A in CALHM6 converted the channel from voltage-independent to voltage-dependent gating, resulting in outward rectification (**Figure 5J-L**,P,Q; Supplementary Figure 5). These residues are likely key determinants controlling the conformational dynamics during gating.

Interestingly, mutations in other conserved residues, near – but not within – the S1–S2 linker (W114, L57, and P60 in CALHM1; W113 and A127 in CALHM6), also abolished or markedly reduced channel activity (**Figure 5P,Q**; Supplementary Figure 5). Among these, W114 in CALHM1 and W113 in CALHM6 directly contact S1. Thus, substituting these bulky hydrophobic residues with smaller (cysteine or alanine) or positively charged (arginine) residues is expected to alter the conformational dynamics of S1 and therefore impair channel gating. Notably, however, the CALHM6 W113A was an exception, retaining wild-type like currents at 37°C and exhibiting voltage dependence and inhibition by extracellular Gd^{3+} similar to wild-type CALHM6 (**Figure 5M-O**), suggesting that this mutation does not impair fundamental gating properties.

Finally, since most mutants showed either reduced or completely abolished activity, we included a positive control in the electrophysiological experiments: I109W on CALHM1, which has been previously documented to increase channel activity. As expected, we reproduced this gain-of-function phenotype (**Figure 5**). To further explore channel-specific differences, we also examined the corresponding mutant in CALHM6 (L108W). Interestingly, L108W decreased CALHM6 channel activity (**Figure 5**), possibly reflecting inherent differences between the two channels.

Discussion

Here, for the first time, we have computationally defined the full IC complement of the human genome – the “channelome”. We provide this comprehensive list and a rich annotation of functional, structural, and sequence features, mapping them to 4 widely accepted IC groups and further classifying them into 55 families. We also highlight unclassified outlier groups that contain most of the understudied “dark” and potentially novel unclassified IC sequences. As part of our curation, we also provide annotation for the pore-containing domain that is based on multiple sources, including an extensive literature review, further strengthened by the application of

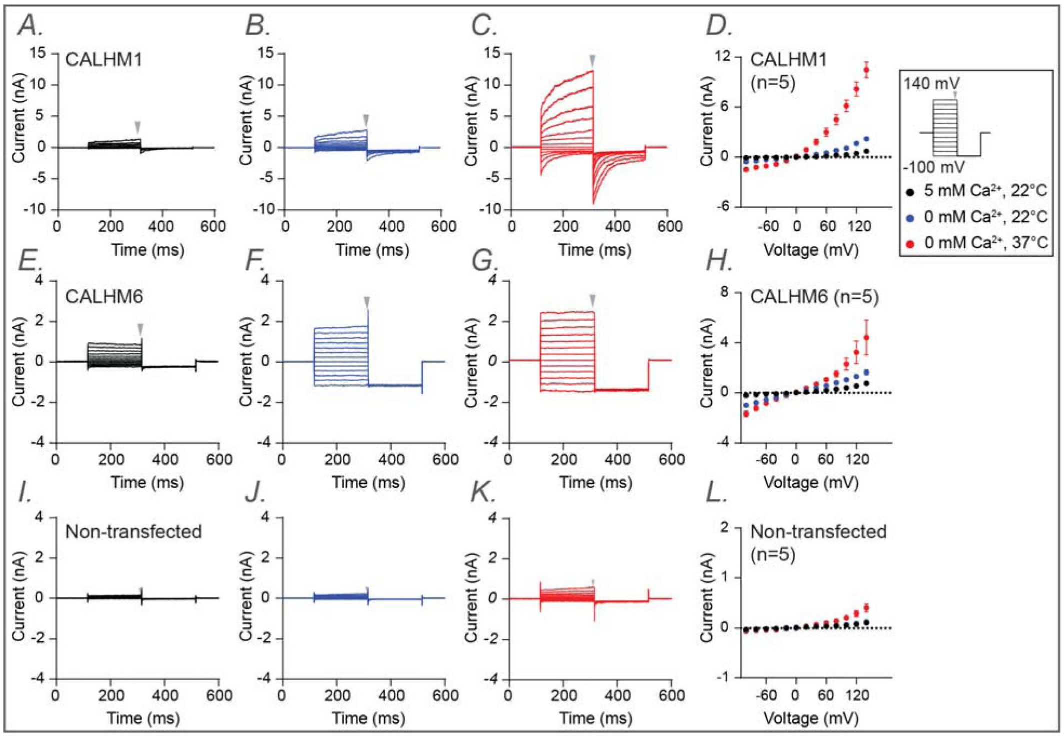


Figure 4.

Electrophysiological studies of human CALHM1 and CALHM6.

Whole-cell voltage-clamp recordings were performed in tsA cells overexpressing wild-type CALHM1 (A–D) and wild-type CALHM6 (E–H), as well as in non-transfected cells (I–L). Currents were measured under three conditions sequentially from the same cell: 5 mM Ca^{2+} at 22 °C (a, e, i; black), 0 mM Ca^{2+} at 22 °C (B,F,J; blue), and 0 mM Ca^{2+} at 37 °C (C,G,K; red). Voltage steps ranged from −100 mV to +140 mV, followed by a final tail pulse at −100 mV, with a holding potential of 0 mV (protocol illustrated in the box on the right). Current-voltage (I–V) relationships were plotted in D, H, and L using mean current amplitudes (averaged across independent cells) measured at the end of the voltage steps. The arrow indicates the time point at which current amplitudes were measured. Error bars represent SEM (D, n=5; H, n=5; L, n=5)

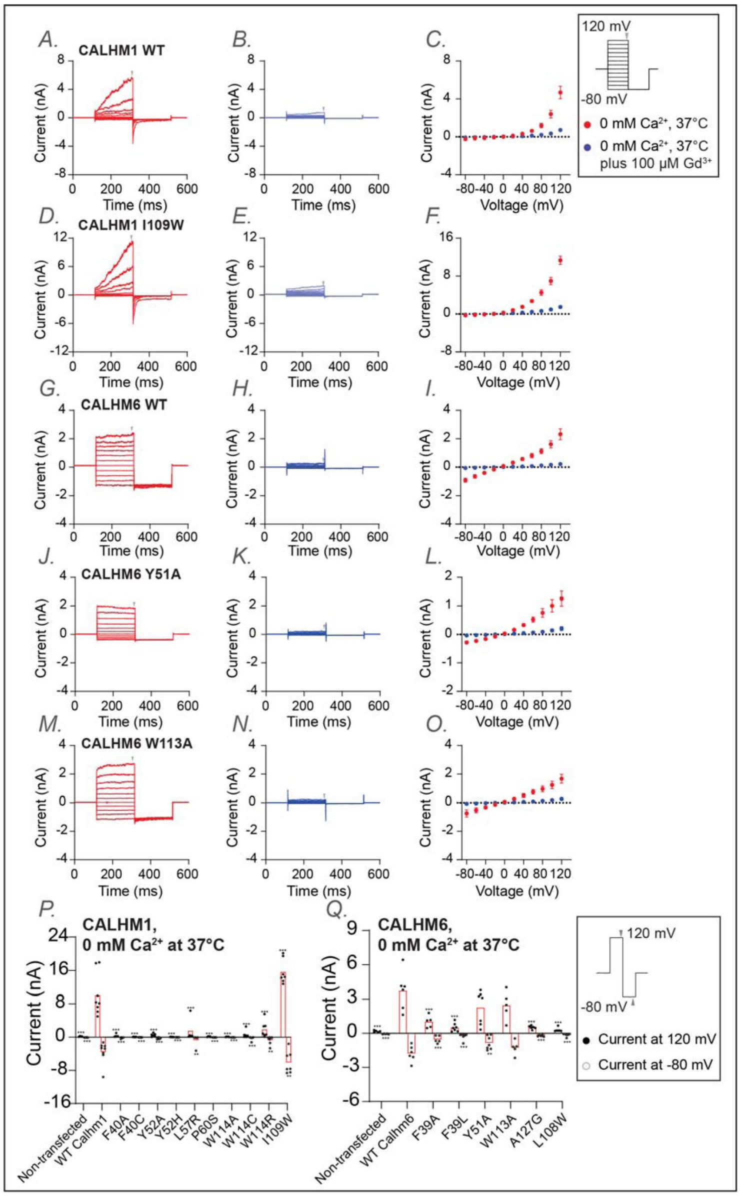


Figure 5

Functional characterization of CALHM1 and CALHM6 mutants at conserved residues at 37 °C.

Whole-cell voltage-clamp recordings were performed in tsA cells overexpressing wild-type CALHM1 (A-C), CALHM1 mutant (I109W) (D-F), wild-type CALHM6 (G-I), and CALHM6 mutants (Y51A (J-L)and W113A(M-O)). Currents were measured under two conditions: 0 mM Ca^{2}LJ at 37 °C (A,D,G,J,M; red) and 0 mM Ca^{2}LJ at 37°C plus 100 µM Gd^{3}LJ (B,E,H,K,N; blue), with both conditions recorded sequentially from the same cell. Voltage steps ranged from −80 mV to +120 mV, followed by a final tail pulse at −80 mV, with a holding potential of 0 mV (protocol shown in the box on the right). Current-voltage (I–V) relationships were plotted in C,F,I,L, and O using mean current amplitudes (averaged across independent cells) measured at the end of the voltage steps. The arrow indicates the time point at which the current was measured. The number of independent measurements (cells) were: n = 5 (C); n = 5 (F); n = 5 (I); n = 5 (L); n=5 (O). Error bars represent SEM. (P,Q) Current amplitudes obtained using a two-step voltage protocol (from +120 mV to −80 mV; protocol shown in the box on the right) are compared between wild-type CALHM1 and its mutants (P) and between wild-type CALHM6 and its mutants (Q). Each dot represents an independent measurement (cell), and bar represents the mean current amplitude across cells. The number of independent measurements (cells) for each bar in P and Q are shown from left to right: 5, 8, 5, 8, 6, 7, 5, 5, 5, 6, 7, 7 (P); 5, 6, 5, 7, 7, 5, 6, 6 (Q). Statistical analysis was performed using one-way ANOVA with Bonferroni’s post hoc test, comparing each mutant to wild type (*p < 0.05; **p < 0.01; ***p < 0.001).

literature mining using LLMs and structural analysis tools. We use the pore-containing domain to distinguish pore-containing from auxiliary ICs, providing additional functional context. Our annotations can serve as a reference for comparing ICs within or across families, designing experiments for functional studies, and identifying potential targets to illuminate the functional relevance of understudied ICs further.

The application of LLMs, specifically RAG systems, in our research demonstrates significant potential in automating literature mining and correcting certain inaccuracies. However, the system currently faces limitations related to diverse data presentation formats, complex channel subunit relationships, and variant gene nomenclature, which hinder its effectiveness. The high frequency of responses where evidence was not found underscores the need for enhanced entity recognition capabilities, particularly for accurately identifying gene and protein names across different nomenclatures. Additionally, the RAG system's ability to critically engage with data by identifying and amending presumed inaccuracies can be widely used to verify and correct annotation entries, though this also raises concerns about potential overcorrections or misinterpretations without adequate validation. By more data cleaning and standardizations, LLMs and RAG systems can become more reliable and integral tools for curating and expanding bioinformatics databases, ultimately enhancing the accuracy and comprehensiveness of our IC knowledgebase.

To the best of our knowledge, identifying true orthologs across the tree of life performed here is the first large-scale effort to consolidate ion channel orthologs across diverse organisms and provides a comprehensive view of the evolutionary conservation of individual ICs. Based on this analysis, we could pinpoint the Aquaporins family of ICs as the most ancient family of ICs with orthologous sequences present back to bacterial and archeal species. More importantly, the clustering of human ICs based on their depth of conservation placed them into functionally meaningful clusters with many enriched functions exclusive to their orthologous taxonomic group. The placement of understudied ICs in such groups allows for meaningful functional predictions, which can be further elucidated using experimental approaches.

The orthologous sequences identified and presented here can serve as a valuable resource for performing evolutionary and functional analysis using statistical and machine learning approaches. We demonstrate one such use by performing a Bayesian pattern-based classification on the CALHM subset of ICs to identify amino acid positions significantly conserved across all CALHM sequences. These features reside on a functionally important S1-S2 linker region that potentially governs their gating mechanism. Because these features were conserved and shared across all CALHM sequences, we also hypothesize that all CALHM sequences share this gating mechanism. By performing targeted mutations on these conserved residues, we show that a mutation in any of these evolutionarily conserved residues results in a dramatic loss of gating function, thus highlighting the functional importance of the identified pattern positions.

Among the conserved positions targeted for mutations, mutations of residues in the S1-S2 linker (F40C in CALHM1; F39L in CALHM6 and Y52A in CALHM1; Y51A in CALHM6) resulted in strong phenotypic changes, most likely due to their direct involvement in the conformational dynamics of S1 for channel gating. Mutations in W114 in CALHM1,which is not directly on the S1-S2 linker, also resulted in abolished activity, however mutations in an analogous position W113 in CALHM6 did not impair channel function, presenting an exception compared to other residues across CALHM1 and 6. Among other residues that are near – but not within – the S1-S2 linker, the positioning of P60 on the S2 helix of CALHM1 is particularly intriguing. In the cryo-EM structure, P60 resides approximately midway along S2, inducing a distortion or bending of the helix due to the rigid ring structure of proline. This bent configuration allows extensive contact between the extracellular portion of S2 and the S1–S2 linker, while the intracellular portion of S2 interacts extensively with S1. Furthermore, we hypothesize that this bent conformation of S2 contributes to the flexibility of the S1–S2 linker, as if S2 were straightened, it would consequently straighten the S1–S2 linker, potentially reducing its flexibility. Replacing the proline with glycine eliminates the structural constraint imposed by proline's cyclic side chain, allowing S2 to adopt a more regular, uninterrupted helical structure. Consequently, a straightened S2 may lead to reduced flexibility of the S1–S2 linker, ultimately impacting channel gating. Similarly, L57 in CALHM1 is positioned near P60 on the protrusion of the bend of S2, with its sidechain facing residues on the adjacent S3. We hypothesize that the larger sidechain of the L57R mutant would force S2 to straighten due to steric collision between the bulky arginine sidechain and residues on S2, again reducing the flexibility of the S1-S2 linker and alter channel gating. A127 in CALHM6 resides in a short alpha helix within the extracellular domain. While A127 does not directly interact with the S1–S2 linker, it is adjacent to the highly conserved disulfide bond (between C41 and C126), linking the helix containing A127 to the S1– S2 linker. Thus, it is conceivable that the A127G mutant may indirectly affect the conformation of the S1–S2 linker through this disulfide bond.

These findings complement prior studies and provide additional insights into CALHM gating mechanisms. Several previous studies examined conserved residues in CALHM channels and proposed roles in gating [56–59, 62, 64]. These studies were based on comparisons of a limited set of homologs from model organisms such as *C. elegans* or mouse, and identified residues in various parts of the protein that influence gating, providing important clues towards their functioning. Here, we have extended the analysis to encompass thousands of CALHM sequences collected from the entire tree of life, allowing us to identify residues conserved across all family members through evolution. Guided by a hypothesis from previous structural and functional studies [59, 60] which highlighted the NTH/S1 region as a key element in channel gating, we focused on evolutionarily conserved residues in the S1–S2 linker and at the interface of S1 with the rest of the TMD. We reasoned that if S1 movement is critical for gating, then these two structural elements—the S1–S2 linker, acting as a hinge, and the S1 interface with the TMD, serving as a stabilizing contact—would be key determinants of the conformational dynamics of S1.

Together, our experimental data indicate that the conserved residues at the S1–S2 linker and its immediate surroundings play an important role in CALHM channel gating, supporting a model in which S1 dynamics are central to the gating mechanism [59, 60]. Furthermore, the contrasting phenotypes of Y52A in CALHM1 versus Y51A in CALHM6, W114A in CALHM1 versus W113A of CALHM6, and I109W in CALHM1 versus L108W in CALHM6 highlight differences in gating properties specific to each family member. Overall, the complementary results from this study and the published literature highlight the complexity of CALHM gating and suggest that distinct conserved elements contribute to both shared and lineage-specific gating mechanisms in this unique family of large-pore channels. Importantly, this was demonstrated not only for the well-studied CALHM1 but also for the relatively understudied human CALHM6, providing valuable clues into shared gating features across light and dark CALHM sequence sets.

In addition to the CALHM shared patterns, we also present a second subset of conserved positions shared only within CALHM 2,4,5 and 6, paralogs that are evolutionarily closer compared to CALHM1 and 3 (Supplementary Figure 6). These subsets of conserved residues fall in the intracellular helical segment that is involved in oligomerization and formation of the pore complex governing CALHM function. Previous studies have indicated that CALHM2 adopts a unique undecameric oligomer different from the CALHM1 octamer [65]. Thus, with the presence of shared patterns within the subset of CALHM2 and the understudied CALHM4,5 and 6 in this intracellular helical region, we hypothesize that these conserved positions help maintain a similar mode of oligomerization between these evolutionarily related subsets of CALHMs.

Because many disease mutations map to these conserved positions, we believe the datasets and approaches offer a powerful avenue for elucidating the functional and clinical relevance of the understudied ICs.

Methods

Identification and annotation of human ICs

The annotation of human ICs was performed using a semi-automated pipeline. First, all the protein sequences that were labeled as ICs were collected from the UniProt [41], KEGG [23, 24], Pharos [25], GtoP [14], and HGNC [26] databases. The annotations described in **Table 1** were then collected from literature sources manually and compiled together. The UniProt specific labels and the complex formation information were extracted from the UniProt database. The sequences were run through TMHMM [42] and Phobius [43] to predict transmembrane and helical regions. The predictions were cross-referenced and confirmed against the CDD [66], Pfam [67], PrositePattern , PrositeProfiles [68] and Simple Modular Architecture Research Tool (SMART) [69] databases, where a reference was available. Finally, the prediction of the pore region and pore-lining residues was supplemented using the MOLE software [44]. The prediction of pore region and the information available in previous literature was combined to annotate the final set of auxiliary IC sequences. An auxiliary IC is defined as any sequence that itself does not have a pore domain, but has experimental evidence of being part of an IC complex. Thus, any IC where a pore domain could not be found was subjected to additional literature review to find evidence for their interactions with the pore containing ICs, and if the evidence was found, they were included as auxiliary ICs, otherwise they were removed from our curated IC list.

Retrieval-Augmented Generation (RAG) system for verifying ion-selectivity and gating-mechanism annotations

To systematically evaluate the ion-selectivity and gating-mechanism fields in the curated human IC data set, we built a RAG pipeline comprising three sequential stages: corpus construction and vectorization, query formulation and similarity search, and evidence synthesis with a LLM. All scripts and configuration files are archived in the project repository (*https://github.com/esbgkannan/ionchannels-final-pdf*).

Corpus construction and vectorization

PubMed identifiers (PMIDs) linked to each IC were obtained from our annotation pipeline. Full-text PDFs were downloaded, converted to plain text with the *pdfminer.six* Python library, and segmented into overlapping fragments of approximately 1000 tokens (50-token overlap) using the *CharacterTextSplitter* module of LangChain. Each fragment was embedded with the OpenAI *text-embedding-3-large* model (3072-dimensional vectors). Embeddings and fragment metadata (PMID, page number, fragment index) were stored in a local *Qdrant* vector database. Prior to embedding, boiler-plate headers, footers and reference lists were removed to retain only article body text.

Query formulation and similarity search

For every IC, two natural-language questions were generated automatically: 1. "Is there any evidence that <*ION*> is the ion selectivity of the <*IC-NAME*> ion channel?", and 2."Does this article provide evidence for <*GATING*> as the gating mechanism for the <*IC-NAME*> ion channel?".

Here, <*IC-NAME*> encompasses the UniProt primary name and all recognized synonyms (produced by extract_alternative_names.py script in *https://github.com/esbgkannan/ionchannels-final-pdf*). Each query was submitted to Qdrant to retrieve relevant fragments from the database. When PMIDs were available in the dataset, an initial PMID-filtered search was performed; filtered and unfiltered results were subsequently merged and fragments with the highest cosine similarity retained.

Evidence synthesis with an LLM

The retrieved fragments were inserted into a fixed prompt and supplied to GPT-4o (temperature = 0). The model was instructed to return a JSON object having three keys and values of answer (Found, or Not Found), confidence (a number between 0 to 1), and evidence (an explanation of LLM's decision for the answer). Outputs were parsed with LangChain's *JsonOutputParser*, combined with fragment metadata, and compiled to a single JSON file for further analysis. Predictions with confidence ≥ 0.80 were accepted automatically; lower-confidence or *Not Found* results triggered manual review. The structured records were merged with the annotation table, enabling automatic comparison with pre-existing entries and flagging of discrepancies for curator review.

Entries for which no supporting evidence could be located (model answer 'Not Found' confirmed on manual inspection) are marked with an asterisk in Supplementary Table 1. An illustrative exchange is provided in Supplementary Figure 1.

Defining a pore containing functional domain

We aimed to define the pore-containing functional domain so that it spanned the pore region and included all the TM domains present in an IC. This ensured that our domain definition included the pore region and any other functionally important TMs, while excluding any accessory domains on the flanking sequences. To ensure we did not miss any TM regions, four different sources of annotation information were combined to define the pore containing functional domain for the 343 ICs. 1) The TM predictions from TMHMM [42] and Phobius [43] was first used to identify the TM regions. 2) The UniProt-based TM annotations were matched with the predictions to verify the TM regions further. 3) The literature-based TM annotations were then used to verify the TM positions and organization, where available manually. 4) Where experimentally resolved crystal structure coordinates were available, the MOLE software was used to identify the pore lining residues.

Pairwise sequence alignment using sequence embeddings

The pore containing functional domains for the 343 pore containing ICs were passed to DEDAL [49] which was run using standard parameters and resulted in homology logit scores based on an all-vs-all pairwise sequence alignment using their sequence embeddings. These scores were used to generate a sequence similarity matrix passed to the umap function of the umap_learn package v0.5.7 [70] in python 3.9 to generate 2D UMAP embeddings. This was used to generate a 2D scatterplot that defined the placement of individual ICs shown in Supplementary Figure 3. Finally, an average position of ICs within a family was used to define the placement of that family in **Figure 1**.

Orthology detection and analysis

The KinOrtho pipeline [51] was used for defining the orthologs and co-orthologs of human ICs across the tree of life. KinOrtho employs a graph-based orthology inference approach using both the full-length and the pore domain regions for inferring orthologous relationships primarily relying on sequence similarity within these regions. The pipeline began with pairwise sequence similarity searches using both full-length and pore-containing domain sequences of human ICs against the target proteomes database. Since a pore domain needs to be defined for this analysis, we only used the 343 pore containing IC sequences to perform orthology detection.

They were used as query for running KinOrtho against a reference database of curated canonical proteomes from the UniProt Proteomes Release 2022_05 that consisted of 343 Archaea, 696 Bacteria and 548 Eukaryota proteomes. The initial sequence similarity search was conducted using BLASTp [46] with default parameters and an e-value cutoff of 1e-5 to retain high-confidence hits. The top hit from this forward search was then used as query in a reciprocal BLASTp search against the human proteome, applying a more stringent e-value threshold of 1e-200 to retain only the top and high confidence hits. Pairs of sequences were retained only if each was the top hit for the other in this reciprocal search. Such a reciprocal best-hit (RBH) strategy minimizes false positives by ensuring that the candidate orthologs are each other's most similar sequence in the respective proteomes, which is particularly important when distinguishing true orthologs from paralogs or other homologs that may share partial similarity but have diverged functionally [71].

All the retained hit pairs were then passed to cluster analysis using OrthoMCL [72] and filtering of relationships to keep only relationships within the same cluster. This process was conducted separately using both the full-length human IC sequences and their pore domain sequences. Only sequences that were identified as orthologs in both the full-length and domain-based analyses were retained as high-confidence orthologs. This ensured that both overall sequence similarity and conservation of functionally critical domains were satisfied.

Finally, the obtained ortholog sequence sets were subjected to a series of validation tests that provide more guardrails and ensure we avoid duplicate, fragments and extraneous hits ensuring that only sequences with high quality annotation were retained as true orthologous relationships. These validation tests followed the following steps:

Step 1: Check UniProt Entry Type: We first check the "entryType" flag in UniProt to check whether it is: "Reviewed" or "Unreviewed". "Reviewed" sequences are manually annotated and reviewed, thus are of the highest quality and safe to include. Any sequences with this status are passed. Sequences with "Unreviewed" status proceed to Step 2.

Step 2: Check for Protein Existence Evidence: For the "Unreviewed" sequences, check their "proteinExistence" flag. This flag can have one of five values: 1. Evidence at protein level 2. Evidence at Transcript level 3. Inferred from homology 4. Predicted 5. Uncertain. Flags 1,2, and 3 are dependable evidence of protein existence, thus are passed, whereas 4 and 5 are low confidence and subjected to further verification in Step 3.

Step 3: For the proteins with low confidence (4 or 5) protein existence values, we next check five different measures/flags to ensure its integrity:

1. Check their sequence version for unusual patterns. Multiple updates with high sequence versions or a very old version update indicate unstable or deprecated sequences, respectively.
2. Perform sequence level checks.
    1. Check for compositional bias: Sequences with compositional bias might be low complexity regions or repeats.
    2. Check for proportion of non-standard amino acids: Presence of large number of non-standard amino acids could also indicate poor protein annotation.
    3. Check sequence length: Unusually short (<30 aa) or long (>5000 aa) sequences could indicate fragments or fusion errors and misannotation.
3. Check for cross-references: If the protein completely lacks annotated domains, or additional cross referenced metadata, it might suggest poorly characterized or erroneous proteins. For any sequence that is subjected to Step 3, it should pass all the checks in this step to be included as an orthologous hit.

48694 unique orthologous relationships from 36846 sequences passed the orthology pipeline and validation checks and were included in the final list of orthologous sequences. The list of all the orthologus sequences that passed the validation check are provided in Supplementary Table 3.

To illustrate the extent of these orthologous sequences across different taxonomic lineages, they were used to create a presence/absence matrix of orthologs, with rows representing each human IC and columns representing the different organisms. To reduce individual granularity and get estimates at the family and lineage level for visualization, first, the orthologs were grouped by IC families and then, by their defined taxonomic lineages. For each group representing one IC family and one taxonomic lineage, a percentage value representing the proportion of detected orthologs was calculated using the following: (total number of orthologs found for all ICs in a family)/(total number of organisms queried in the taxonomic lineage * number of human IC sequences in the family). These percentages are depicted in the heatmap in **Figure 2A** where each cell represents an IC family and its proportion of orthologs in a given taxonomic lineage.

Next, the presence/absence matrix of individual ICs was used to perform an orthology profiling clustering and enrichment analysis. Since this analysis is human IC-centric, only the orthologs from eukaryotic lineages were selected. Hierarchical clustering was performed with the Ward method of clustering and Euclidean distance metric using the scipy package [73] in Python. The resulting dendrogram was used to define 9 clusters that group ICs with similar presence/absence patterns. Human ICs falling in these 9 clusters were then subjected to a functional enrichment analysis using the Gene Ontology resource GO enrichment tool [74]. GO terms with a corrected FDR <0.01 were retained as significantly enriched terms.

CALHM evolutionary analysis

BPPS was used to perform a pattern-based classification of the CALHM homologs. First, orthologs for all 6 human CALHMs were collected. This ortholog dataset was supplemented with more hits from the UniProt database using MAPGAPS, a multiply-aligned profile for global alignment of protein sequences [12]. Along with finding the best hits for the human CALHMs, MAPGAPS also aligns those hits to the template profile alignment to generate a large multiple-sequence alignment of the resulting 5805 CALHM. This large alignment was then subjected to BPPS, which performs a hierarchical classification of the sequence sets based on conserved pattern positions shared by subsets of sequences using a Bayesian statistical procedure [40]. This generates a hierarchical cluster where sequences within each cluster are defined by distinct conserved patterns.

Cell lines

HEK293T cells are purchased from Sigma-Aldrich (Catalogue Number: 96121229). Neuro2A cells are purchased from ATCC (Catalogue Number: CCL-131). The cells are authenticated and are mycoplasma contamination tested negative by the vendor.

CALHM plasmid construction and expression by transient transfection

Full-length human CALHM1 and CALHM6 in the pEGC Bacmam vector was used. The translated product contains the human CALHM1 or CALHM6 protein, a thrombin digestion site (LVPRGS), an enhanced GFP protein, and an 8x His tag. Primers for site-directed mutagenesis were designed using Snapgene and synthesized by Eurofins Genomics. The QuikChange mutagenesis protocol was used to generate all the mutants of the study. Sanger sequencing was performed to identify positive clones.

Adherent HEK293T (ECACC, Catalogue Number: 96121229) cells were grown in DMEM media supplemented with 10% fetal bovine serum. Transient transfection was conducted using lipofectamine-2000 by following the manufacturer's protocol. Specifically, the cells were cultured in 60 mm Petri dishes until 80% confluency. Transfection solution was made by mixing 500 ng of plasmid DNA, 4 uL of lipofectamine-2000 reagent, and 100 uL Opti-MEM media. After 20 min incubation at room temperature, the DNA-lipid complexes were added to the cell culture and incubated at 37°C. The next day, 10 mM sodium butyrate was added to the cells to boost protein expression. The cell culture was then grown at 30°C for another day before harvesting. The cell pellet was flash-frozen with liquid nitrogen and stored at -80°C. Each CALHM1 and CALHM6 mutant were transfected in triplicate as biological replicates.

Expression analysis of CALHM1, CALHM6, and their mutants

To analyze total expression levels of wild-type CALHM1, wild-type CALHM6, and their mutants, cells were lysed in TBS buffer (20 mM Tris, pH 8.0, 150 mM NaCl) supplemented with 10% lauryl maltose neopentyl glycol and cholesterol hemisuccinate detergents on ice. Lysates were solubilized at 4°C for 1 hour and clarified by centrifugation at 13,000 rpm for 5 min. The supernatant was mixed with 4× SDS loading buffer containing 5% 2-mercaptoethanol and resolved on a 4–20% gradient SDS-PAGE gel.

In-gel fluorescence imaging of the C-terminal GFP tag was performed immediately after electrophoresis using a Chemidoc system to visualize CALHM1 and CALHM6 proteins. Following imaging, proteins were transferred to a nitrocellulose membrane using semi-dry transfer buffer (48 mM Tris base, 39 mM glycine, 20% methanol). Membranes were blocked with TBST (20 mM Tris, pH 8.0, 150 mM NaCl, 0.1% Tween-80) containing 4% non-fat milk for 1 hour at room temperature. β-actin was detected as a loading control by incubating membranes with HRP-conjugated anti-β-actin antibody (Proteintech, cat. no. HRP-60008; 1:2000 dilution) for 1 hour at 4°C. After four washes with TBST (15 min each), chemiluminescence signals were developed using Pierce ECL substrate and imaged on a Chemidoc system. A brightfield image was overlaid to visualize protein marker positions.

Each mutant was analyzed in triplicate from independently transfected cell pellets. GFP signal intensities for CALHM1 and CALHM6 were quantified using ImageJ and normalized to β-actin chemiluminescence signals. Mean values and SEM were calculated for each mutant, and relative expression levels were compared to wild-type proteins using bar graphs generated in Microsoft Excel.

Surface expression analysis

Surface biotinylation of CALHM1, CALHM6, and their mutants was performed using the Pierce™ Cell Surface Biotinylation and Isolation Kit (Thermo Fisher Scientific) following the manufacturer's protocol. Cells were cultured and harvested 48 hours post-transfection as described above. Surface-isolated proteins were analyzed by SDS-PAGE, and in-gel fluorescence imaging of the C-terminal GFP tag was performed using a Chemidoc system.

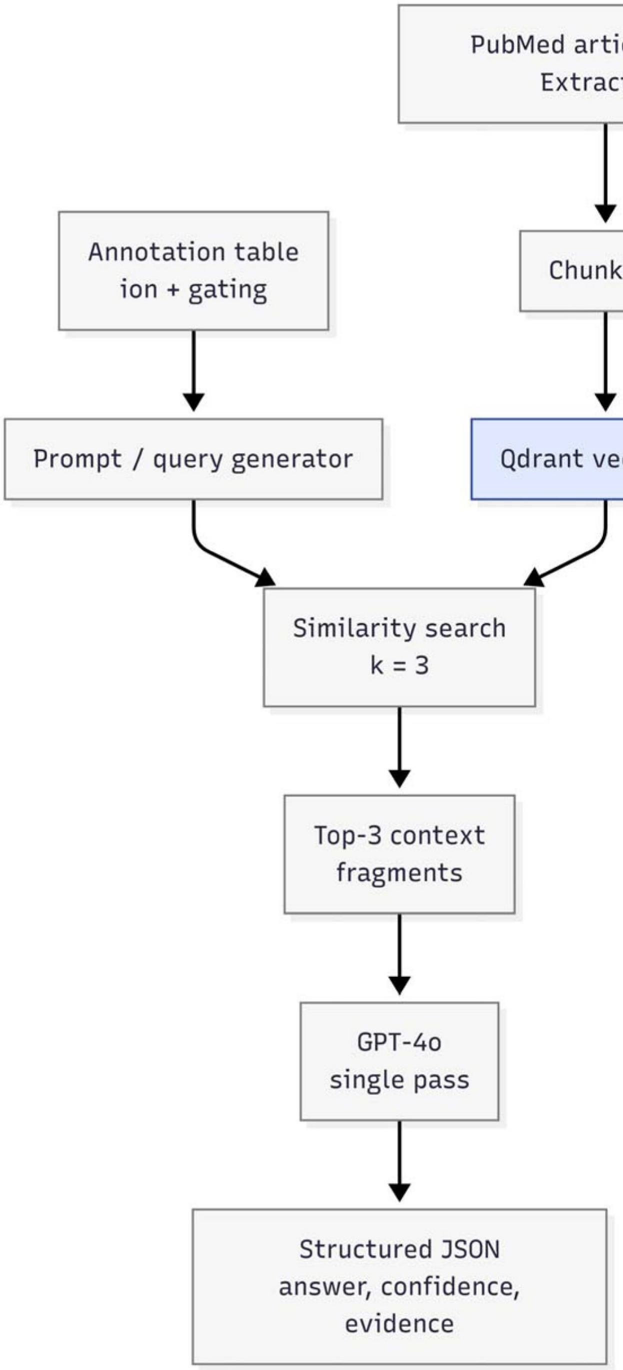
Electrophysiology

TsA201 cells expressing plasmids encoding N-terminal GFP tagged human CALHM1 or CALHM6 were used. After 1-day post-transfection with plasmid DNA (100 ng/mL) and Lipofectamine 2000 (Invitrogen, 11668019), the cells were trypsinized and replated onto poly-L-lysine-coated (Sigma P4707) glass coverslips After cell attachment, the coverslip was transferred to a recording chamber. Whole-cell patch-clamp recordings were performed at room temperature (21-23°C) or body temperature (36-38°C). Signals were amplified using a Multiclamp 700B amplifier and digitized using a Digidata 1550B A/D converter (Molecular Devices, Sunnyvale CA). The whole-cell current was measured on the cells with an access resistance of less than 10 MΩ after the whole-cell configuration was obtained. The amplifier circuitry compensated the whole-cell capacitance. The two-step pulse from 120 to -80 mV for 50 ms was continuously applied to the cell membrane every 5 sec to monitor the activation of the CALHM current. The step pulse from -100 to 140 mV (or –80 to 120 mV) for 200 msec with a holding potential of 0 mV was applied to plot the current-voltage relationship. Electrical signals were digitized at 10 kHz and filtered at 2 kHz. Recordings were analyzed using Clampfit 11.3 (Axon Instruments Inc), GraphPad Prism 10 (La Jolla, CA), and OriginPro 2024 (OriginLab, Northampton, MA). The standard bath solution contains (in mM): 150

NaCl, 5 KCl, 1 $MgCl_2$, 2 $CaCl_2$, 12 Mannitol, 10 HEPES, pH=7.4 with NaOH. For a whole-cell recording, the extracellular solution contains (in mM): 150 NaCl, 10 HEPES, 1 $MgCl_2$, 5 $CaCl_2$. To establish a zero Ca^{2+} condition, 5 mM $CaCl_2$ was replaced with 5 mM EGTA. For the zero Ca^{2+} condition with 100 µM Gd^{3+}, 5 mM $CaCl_2$ was omitted entirely without adding additional EGTA. The intracellular solution contains (in mM): 150 NaCl, 10 HEPES, 1 $MgCl_2$, 5 EGTA.

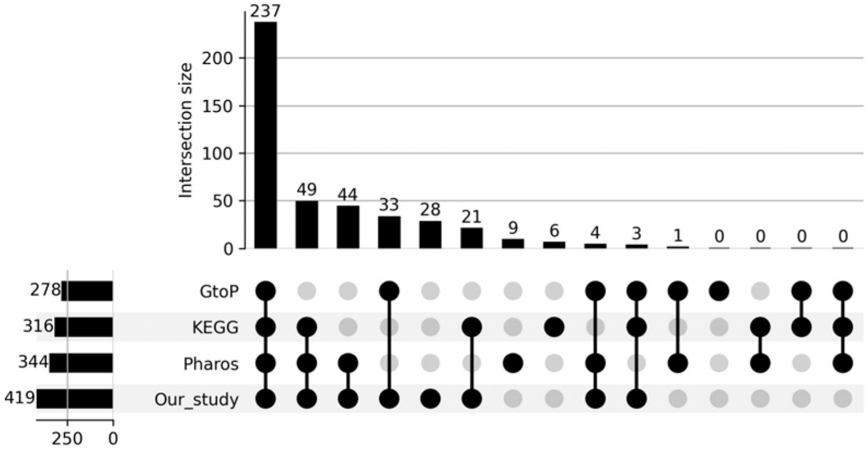
All data are expressed as mean±SEM. Multiple comparisons were performed by one-way or two-way ANOVA with Bonferroni's post hoc test. n indicates the number of cells. Significance was defined as: *P<0.05, **P<0.01, ***P<0.001. The absence of an asterisk indicates non-significance.

Supplementary Figures



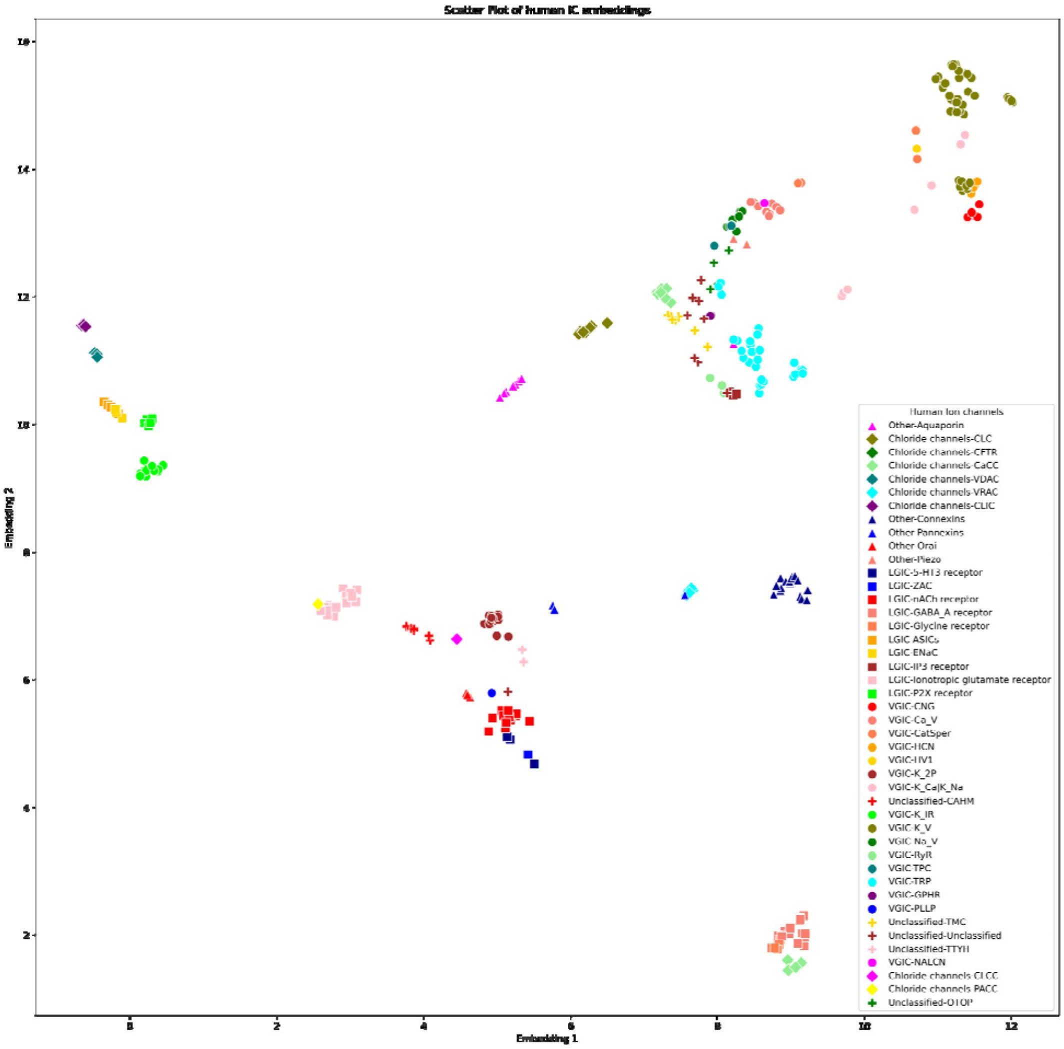
Supplementary Figure 1

RAG annotation pipeline used for validating the annotation of ion specificity and gating mechanism



Supplementary Figure 2

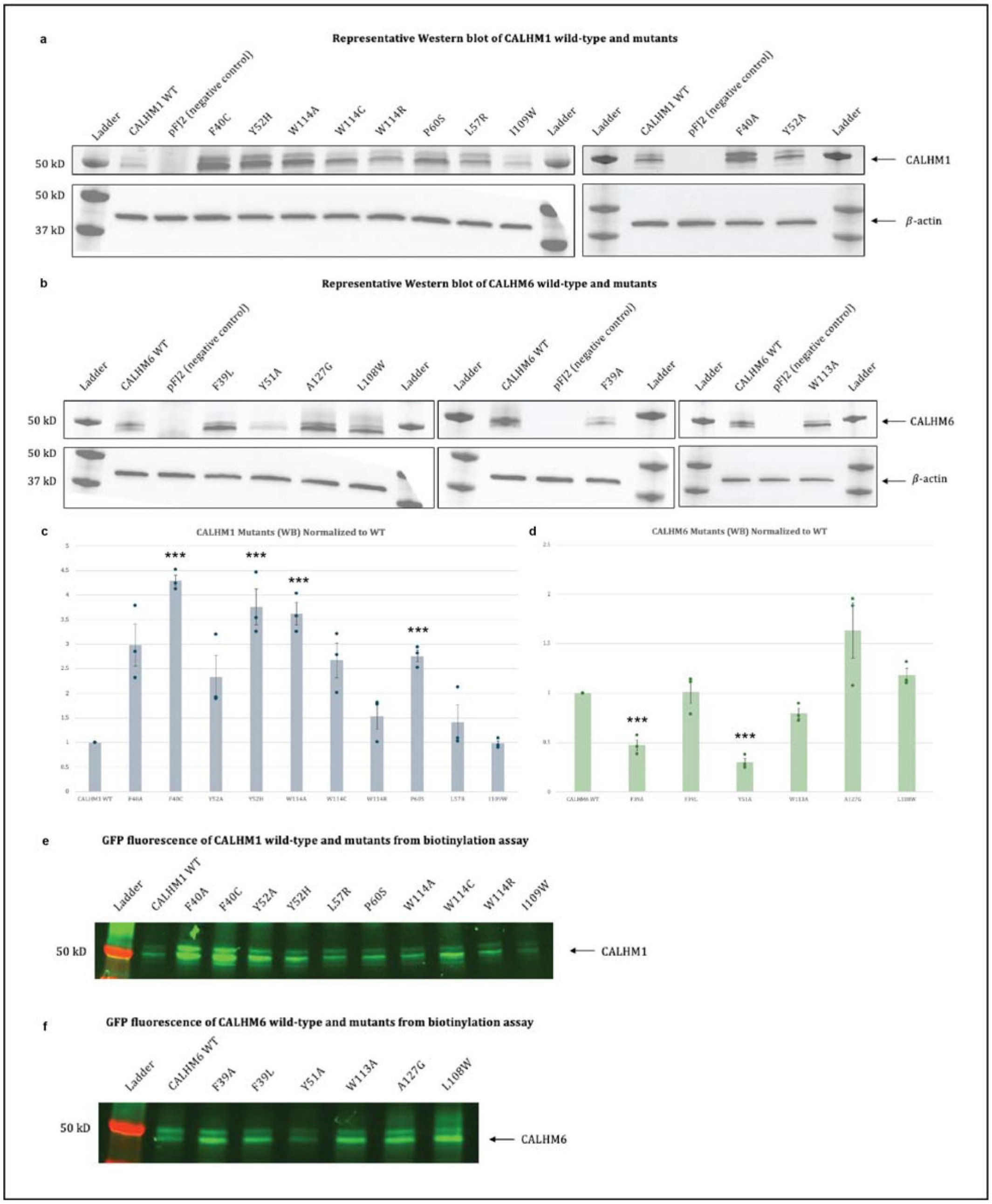
Upset plot showing the overlap of IC sequences based on their UniProt IDs across the current study, the KEGG database, GtoP and Pharos.



Supplementary Figure 3

UMAP embeddings plot of all human IC sequences.

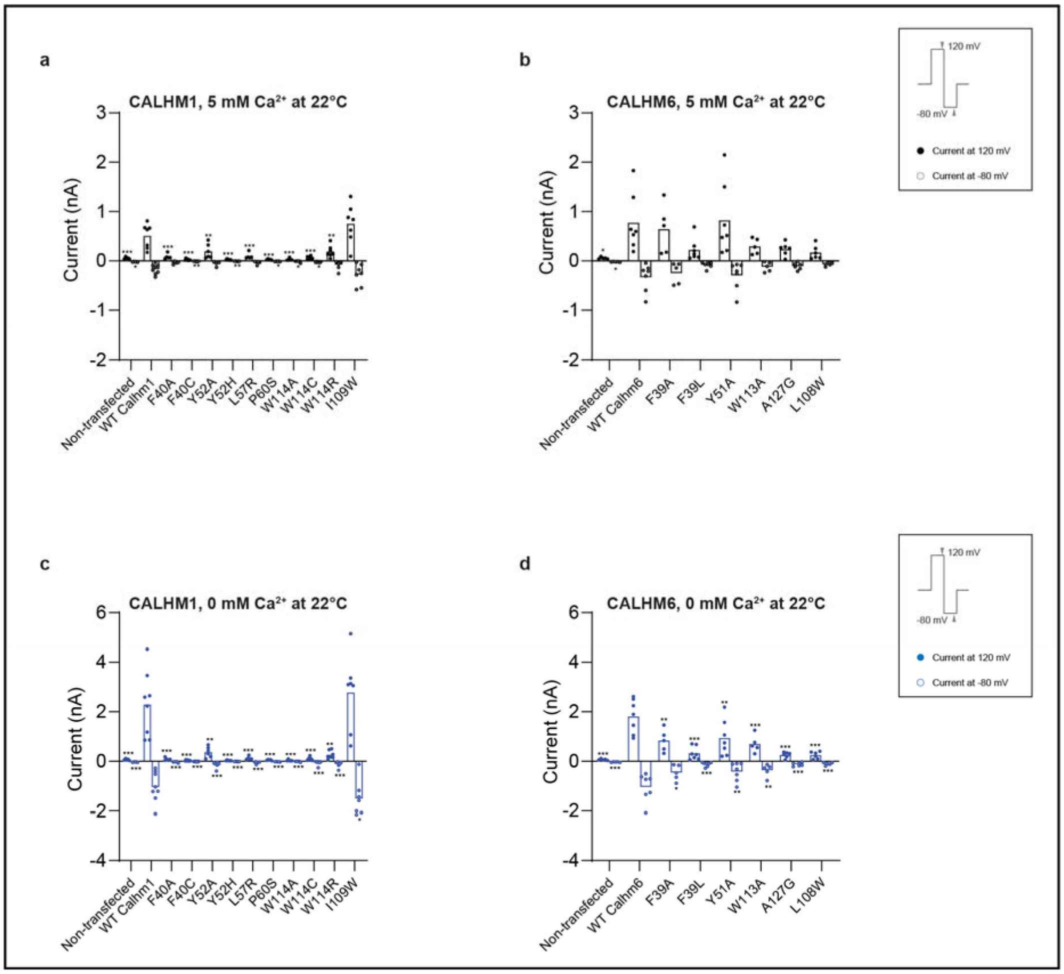
The UMAP embeddings were generated based on the pairwise similarity scores of the protein embedding alignment generated for all pairs of human ICs. Shapes of the markers indicate the different IC groups (circle: VGIC, square: LGIC, diamond: Chloride channel, triangle: Other, plus: Unclassified) and the colors indicate different IC families within each group as shown in the legend above.



Supplementary Figure 4.

Expression analysis of wild-type CALHM1, wild-type CALHM6, and their mutants.

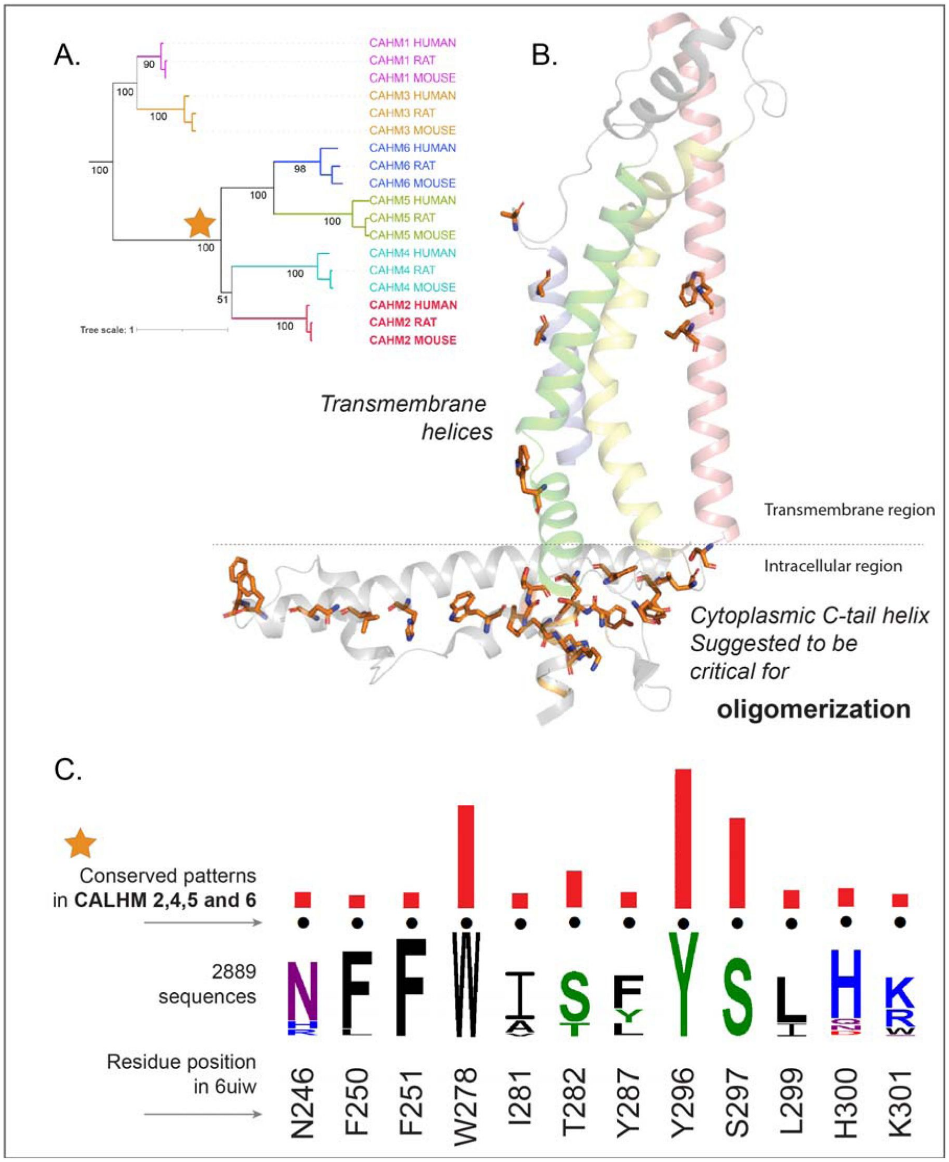
(A,B) Representative gels showing CALHM1 and its mutants (A), and CALHM6 and its mutants (B). CALHM1 and CALHM6 signals were detected using in-gel fluorescence of the C-terminal GFP tag, while β-actin was detected by Western blotting as a loading control. (C,D) Quantification of total protein expression levels of wild-type CALHM1 and its mutants (C), and wild-type CALHM6 and its mutants (D). Each dot represents an independent measurement (transfection) and error bars represent SEM (c, n = 3; d, n = 3). Statistical analysis was performed using one-way ANOVA with Bonferroni's post hoc test, comparing each mutant to wild type (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). (E,F) Surface biotinylation assays of wild-type CALHM1 and its mutants (E), and wild-type CALHM6 and its mutants (F), detected using the C-terminal GFP tag by in-gel fluorescence.



Supplementary Figure 5

Functional characterization of CALHM1 and CALHM6 mutants at conserved residues at 22 °C.

Current amplitudes obtained using a two-step voltage protocol (from +120 mV to −80 mV; protocol shown in the box on the right) are compared between wild-type CALHM1 and its mutants (A, B) and between wild-type CALHM6 and its mutants (C, D). The cells analyzed here are the same as those in **Figure 5P**, Q. Briefly, for each cell, currents were measured sequentially under three conditions: 5 mM Ca^{2+} at 22 °C, 0 mM Ca^{2+} at 22 °C, and 0 mM Ca^{2+} at 37 °C. The currents from the first two conditions are plotted here, while currents at 0 mM Ca^{2+} at 37 °C are shown in **Figure 5R**, S. Each dot represents an independent measurement (cell), and bar represents the mean current amplitude across cells. The number of independent measurements (cells) for each bar in a–d are shown from left to right: 5, 8, 5, 8, 6, 7, 5, 5, 5, 6, 7, 7 (A, B); 5, 6, 5, 7, 7, 6, 6, 5 (C, D). Statistical analysis was performed using one-way ANOVA with Bonferroni's post hoc test, comparing each mutant to wild type (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$).



Supplementary Figure 6

Conserved pattern positions identified within the clade for CALHM2,4,5 and 6.

A) Phylogenetic tree of CALHM sequences where the orange star indicates the clade for CALHM2,4,5 and 6. B) The identified pattern positions are mapped into a representative structure of human CALHM2 (PDB: 6uiw). C) Weblogo showing the conserved pattern positions. The red bar indicates the significance of conservation where a taller bar indicates higher significance.

Data availability

The human IC annotation table with all the curation information is made available with the manuscript as Supplementary Table 1. - The fasta sequences for the human ICs (both full-length sequences and the pore domain sequences) are available through Zenodo at doi 10.5281/zenodo.16232527. The full length sequences for all the identified orthologs are also available through the Zenodo doi link. - The code and results related to the RAG annotation pipeline are available at github: *https://github.com/esbgkannan/ionchannels-final-pdf*

Acknowledgements

We thank members of the Kannan Lab for feedback and rotation student William N Lantz for the manual evaluation of RAG LLM-based annotations. Maya Salcedo is acknowledged for onboarding the undergraduates involved in the project. Funding for NK from NIH Common Fund (IDG) U01CA271376 is acknowledged. ZR is supported by 5R00NS128258 from NINDS.

Additional information

Author Contributions

Conceptualization: RT, NK, WL. Methodology: RT, NK, SJP, SK, NG, SS, WL. Formal analysis and investigation: RT, SJP, SK, SS, RC, KB, NK. Data curation: RT, RC, KB, SS. Validation: RT, NG, RC, SS, WL, ZR. Writing – original draft: RT, SJP, SS, NK. Writing - review and editing: RT, SS, ZR, WL, NK. Supervision, project administration, and funding acquisition: NK, WL.

Funding

HHS | National Institutes of Health (NIH) (Common Fund, IDG, U01CA271376)

- Saber Soleymani
- Kennady Boyd
- Wei Lü
- Rahil Taujale
- Rayna Carter
- Nathan Gravel
- Natarajan Kannan

HHS | NIH | National Institute of Neurological Disorders and Stroke (NINDS) (5R00NS128258)

- Zheng Ruan

References

1. Como M, Koppala BR, Hasan MN, Han VL, Arora I, Sun D (2021) **Cell Volume Regulation in Immune Cell Function, Activation and Survival** *Cell Physiol Biochem* **55**:71–88 https://doi.org/10.33594/000000331 | PubMed | Google Scholar

2. Meir A, Ginsburg S, Butkevich A, Kachalsky SG, Kaiserman I, Ahdut R, et al. (1999) **Ion channels in presynaptic nerve terminals and control of transmitter release** *Physiol Rev* **79**:1019–88 https://doi.org/10.1152/physrev.1999.79.3.1019 | PubMed | Google Scholar

3. Thorneloe KS, Nelson MT (2005) **Ion channels in smooth muscle: regulators of intracellular calcium and contractility** *Can J Physiol Pharmacol* **83**:215–42 https://doi.org/10.1139/y05-016 | PubMed | Google Scholar

4. Rajan AS, Aguilar-Bryan L, Nelson DA, Yaney GC, Hsu WH, Kunze DL, Boyd AE (1990) **3rd. Ion channels and insulin secretion** *Diabetes Care* **13**:340–63 https://doi.org/10.2337/diacare.13.3.340 | PubMed | Google Scholar

5. Edelman A, Saussereau E. (2012) **[Cystic fibrosis and other channelopathies]** *Arch Pediatr* **19**:S13–6 https://doi.org/10.1016/s0929-693x(12)71101-6 | PubMed | Google Scholar

6. Kim JB. (2014) **Channelopathies** *Korean J Pediatr* **57**:1–18 https://doi.org/10.3345/kjp.2014.57.1.1 | PubMed | Google Scholar

7. Rivolta I, Binda A, Masi A, DiFrancesco JC. (2020) **Cardiac and neuronal HCN channelopathies** *Pflugers Arch* **472**:931–51 https://doi.org/10.1007/s00424-020-02384-3 | PubMed | Google Scholar

8. Bagal SK, Brown AD, Cox PJ, Omoto K, Owen RM, Pryde DC, et al. (2013) **Ion Channels as Therapeutic Targets: A Drug Discovery Perspective** *Journal of Medicinal Chemistry* **56**:593–624 https://doi.org/10.1021/jm3011433 | Google Scholar

9. Braun N, Sheikh ZP, Pless SA (2020) **The current chemical biology tool box for studying ion channels** *J Physiol* **598**:4455–71 https://doi.org/10.1113/jp276695 | PubMed | Google Scholar

10. Wickenden A, Priest B, Erdemli G (2012) **Ion Channel Drug Discovery: Challenges And Future Directions** *Future Medicinal Chemistry* **4**:661–79 https://doi.org/10.4155/fmc.12.4 | Google Scholar

11. Fodor AA, Aldrich RW (2006) **Statistical limits to the identification of ion channel domains by sequence similarity** *J Gen Physiol* **127**:755–66 https://doi.org/10.1085/jgp.200509419 | PubMed | Google Scholar

12. Neuwald AF (2009) **Rapid detection, classification and accurate alignment of up to a million or more related protein sequences** *Bioinformatics* **25**:1869–75 https://doi.org/10.1093/bioinformatics/btp342 | PubMed | Google Scholar

13. Ranjan R, Khazen G, Gambazzi L, Ramaswamy S, Hill SL, Schürmann F, Markram H (2011) **Channelpedia: an integrative and interactive database for ion channels** *Front Neuroinform* **5**:36 https://doi.org/10.3389/fninf.2011.00036 | PubMed | Google Scholar

14. Alexander SPH, Mathie AA, Peters JA, Veale EL, Striessnig J, Kelly E, et al. (2023) **The Concise Guide to PHARMACOLOGY 2023/24: Ion channels** *Br J Pharmacol* **180**:S145–s222 https://doi.org/10.1111/bph.16178 | PubMed | Google Scholar

15. Castro EV, Shepherd JW, Guggenheim RS, Sengvoravong M, Hall BC, Chappell MK, et al. (2022) **ChanFAD: A Functional Annotation Database for Ion Channels** *Frontiers in Bioinformatics* **2** https://doi.org/10.3389/fbinf.2022.835805 | Google Scholar

16. Špačková A, Vávra O, Raček T, Bazgier V, Sehnal D, Damborský J, et al. (2023) **ChannelsDB 2.0: a comprehensive database of protein tunnels and pores in AlphaFold era** *Nucleic Acids Research* **52**:D413–D8 https://doi.org/10.1093/nar/gkad1012 | Google Scholar

17. Jegla TJ, Zmasek CM, Batalov S, Nayak SK (2009) **Evolution of the human ion channel set** *Comb Chem High Throughput Screen* **12**:2–23 https://doi.org/10.2174/138620709787047957 | PubMed | Google Scholar

18. Jianzhao G, Zhen M, Zhaopeng Z, Hong W, Lukasz K (2019) **Prediction of Ion Channels and their Types from Protein Sequences: Comprehensive Review and Comparative Assessment** *Current Drug Targets* **20**:579–92 https://doi.org/10.2174/1389450119666181022153942 | Google Scholar

19. Li B, Gallin WJ (2004) **VKCDB: Voltage-gated potassium channel database** *BMC Bioinformatics* **5**:3 https://doi.org/10.1186/1471-2105-5-3 | Google Scholar

20. Moran Y, Barzilai MG, Liebeskind BJ, Zakon HH (2015) **Evolution of voltage-gated ion channels at the emergence of Metazoa** *Journal of Experimental Biology* **218**:515–25 https://doi.org/10.1242/jeb.110270 | Google Scholar

21. Lara A, Simonson BT, Ryan JF, Jegla T (2023) **Genome-Scale Analysis Reveals Extensive Diversification of Voltage-Gated K+ Channels in Stem Cnidarians** *Genome Biology and Evolution* **15** https://doi.org/10.1093/gbe/evad009 | Google Scholar

22. Uribe C, Nery MF, Zavala K, Mardones GA, Riadi G, Opazo JC (2024) **Evolution of ion channels in cetaceans: a natural experiment in the tree of life** *Scientific Reports* **14**:17024 https://doi.org/10.1038/s41598-024-66082-1 | Google Scholar

23. Kanehisa M, Goto S (2000) **KEGG: kyoto encyclopedia of genes and genomes** *Nucleic Acids Res* **28**:27–30 https://doi.org/10.1093/nar/28.1.27 | PubMed | Google Scholar

24. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M (2023) **KEGG for taxonomy-based analysis of pathways and genomes** *Nucleic Acids Res* **51**:D587–d92 https://doi.org/10.1093/nar/gkac963 | PubMed | Google Scholar

25. Kelleher KJ, Sheils TK, Mathias SL, Yang JJ, Metzger Vincent T, Siramshetty Vishal B, et al. (2022) **Pharos 2023: an integrated resource for the understudied human proteome** *Nucleic Acids Research* **51**:D1405–D16 https://doi.org/10.1093/nar/gkac1033 | Google Scholar

26. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, Bruford EA. (2023) **Genenames.org: the HGNC resources in 2023** *Nucleic Acids Res* **51**:D1003–d9 https://doi.org/10.1093/nar/gkac888 | PubMed | Google Scholar

27. Kato AS, Bredt DS (2007) **Pharmacological regulation of ion channels by auxiliary subunits** *Curr Opin Drug Discov Devel* **10**:565–72 PubMed | Google Scholar

28. Li Y, Um SY, McDonald TV (2006) **Voltage-gated potassium channels: regulation by accessory subunits** *Neuroscientist* **12**:199–210 https://doi.org/10.1177/1073858406287717 | PubMed | Google Scholar

29. Abbott GW (2022) **Kv Channel Ancillary Subunits: Where Do We Go from Here?** *Physiology (Bethesda)* **37**:0 https://doi.org/10.1152/physiol.00005.2022 | PubMed | Google Scholar

30. Sharma K, Nadler L (2021) **Identifying New Drug Targets by Illuminating the Druggable Genome** *The FASEB Journal* **35** https://doi.org/10.1096/fasebj.2021.35.S1.01799 | Google Scholar

31. Sheils T, Mathias SL, Siramshetty VB, Bocci G, Bologa CG, Yang JJ, et al. (2020) **How to Illuminate the Druggable Genome Using Pharos** *Curr Protoc Bioinformatics* **69**:e92 https://doi.org/10.1002/cpbi.92 | PubMed | Google Scholar

32. Vacher H, Trimmer JS (2011) **Diverse roles for auxiliary subunits in phosphorylation-dependent regulation of mammalian brain voltage-gated potassium channels** *Pflügers Archiv - European Journal of Physiology* **462**:631–43 https://doi.org/10.1007/s00424-011-1004-8 | Google Scholar

33. Chen GL, Li J, Zhang J, Zeng B (2023) **To Be or Not to Be an Ion Channel: Cryo-EM Structures Have a Say** *Cells* **12** https://doi.org/10.3390/cells12141870 | PubMed | Google Scholar

34. Yu FH, Yarov-Yarovoy V, Gutman GA, Catterall WA (2005) **Overview of molecular relationships in the voltage-gated ion channel superfamily** *Pharmacol Rev* **57**:387–95 https://doi.org/10.1124/pr.57.4.13 | PubMed | Google Scholar

35. Huffer KE, Aleksandrova AA, Jara-Oseguera A, Forrest LR, Swartz KJ (2020) **Global alignment and assessment of TRP channel transmembrane domain structures to explore functional mechanisms** *eLife* **9**:e58660 https://doi.org/10.7554/eLife.58660 | Google Scholar

36. Shrestha S, Byrne DP, Harris JA, Kannan N, Eyers PA (2020) **Cataloguing the dead: breathing new life into pseudokinase research** *Febs j* **287**:4150–69 https://doi.org/10.1111/febs.15246 | PubMed | Google Scholar

37. Picado A, Chaikuad A, Wells CI, Shrestha S, Zuercher WJ, Pickett JE, et al. (2020) **A Chemical Probe for Dark Kinase STK17B Derives Its Potency and High Selectivity through a Unique P-Loop Conformation** *J Med Chem* **63**:14626–46 https://doi.org/10.1021/acs.jmedchem.0c01174 | PubMed | Google Scholar

38. Taujale R, Venkat A, Huang L-C, Zhou Z, Yeung W, Rasheed KM, et al. (2020) **Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases** *eLife* **9**:e54532 https://doi.org/10.7554/eLife.54532 | Google Scholar

39. Taujale R, Zhou Z, Yeung W, Moremen KW, Li S, Kannan N (2021) **Mapping the glycosyltransferase fold landscape using interpretable deep learning** *Nature Communications* **12**:5656 https://doi.org/10.1038/s41467-021-25975-9 | Google Scholar

40. Neuwald AF (2014) **A Bayesian sampler for optimization of protein domain hierarchies** *J Comput Biol* **21**:269–86 https://doi.org/10.1089/cmb.2013.0099 | PubMed | Google Scholar

41. UniProt Consortium T. (2018) **UniProt: the universal protein knowledgebase** *Nucleic Acids Res* **46**:2699 https://doi.org/10.1093/nar/gky092 | PubMed | Google Scholar

42. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes** *J Mol Biol* **305**:567–80 https://doi.org/10.1006/jmbi.2000.4315 | PubMed | Google Scholar

43. Käll L, Krogh A, Sonnhammer EL (2007) **Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server** *Nucleic Acids Res* **35**:W429–32 https://doi.org/10.1093/nar/gkm256 | PubMed | Google Scholar

44. Sehnal D, Svobodová Vařeková R, Berka K, Pravda L, Navrátilová V, Banáš P, et al. (2013) **MOLE 2.0: advanced approach for analysis of biomacromolecular channels** *J Cheminform* **5**:39 https://doi.org/10.1186/1758-2946-5-39 | PubMed | Google Scholar

45. Gurnett CA, Campbell KP (1996) **Transmembrane auxiliary subunits of voltage-dependent ion channels** *J Biol Chem* **271**:27975–8 https://doi.org/10.1074/jbc.271.45.27975 | PubMed | Google Scholar

46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009) **BLAST+: architecture and applications** *BMC Bioinformatics* **10**:421 https://doi.org/10.1186/1471-2105-10-421 | Google Scholar

47. Eddy SR (1998) **Profile hidden Markov models** *Bioinformatics* **14**:755–63 https://doi.org/10.1093/bioinformatics/14.9.755 | PubMed | Google Scholar

48. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. (2023) **Evolutionary-scale prediction of atomic-level protein structure with a language model** *Science* **379**:1123–30 https://doi.org/10.1126/science.ade2574 | PubMed | Google Scholar

49. Llinares-López F, Berthet Q, Blondel M, Teboul O, Vert J-P (2023) **Deep embedding and alignment of protein sequences** *Nature Methods* **20**:104–11 https://doi.org/10.1038/s41592-022-01700-2 | Google Scholar

50. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium** *Nat Genet* **25**:25–9 https://doi.org/10.1038/75556 | PubMed | Google Scholar

51. Huang LC, Taujale R, Gravel N, Venkat A, Yeung W, Byrne DP, et al. (2021) **KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases** *BMC Bioinformatics* **22**:446 https://doi.org/10.1186/s12859-021-04358-3 | PubMed | Google Scholar

52. Monahan-Earley R, Dvorak AM, Aird WC (2013) **Evolutionary origins of the blood vascular system and endothelium** *J Thromb Haemost* **11**:46–66 https://doi.org/10.1111/jth.12253 | PubMed | Google Scholar

53. Ma Z, Tanis JE, Taruno A, Foskett JK (2016) **Calcium homeostasis modulator (CALHM) ion channels** *Pflugers Arch* **468**:395–403 https://doi.org/10.1007/s00424-015-1757-6 | PubMed | Google Scholar

54. Ma Z, Taruno A, Ohmoto M, Jyotaki M, Lim JC, Miyazaki H, et al. (2018) **CALHM3 Is Essential for Rapid Ion Channel-Mediated Purinergic Neurotransmission of GPCR-Mediated Tastes** *Neuron* **98**:547–61 https://doi.org/10.1016/j.neuron.2018.03.043 | PubMed | Google Scholar

55. Ma Z, Siebert AP, Cheung KH, Lee RJ, Johnson B, Cohen AS, et al. (2012) **Calcium homeostasis modulator 1 (CALHM1) is the pore-forming subunit of an ion channel that mediates**

extracellular Ca2+ regulation of neuronal excitability *Proc Natl Acad Sci U S A* **109**:E1963–71 https://doi.org/10.1073/pnas.1204023109 | PubMed | Google Scholar

56. Tanis JE, Ma Z, Krajacic P, He L, Foskett JK, Lamitina T (2013) **CLHM-1 is a functionally conserved and conditionally toxic Ca2+-permeable ion channel in Caenorhabditis elegans** *J Neurosci* **33**:12275–86 https://doi.org/10.1523/jneurosci.5919-12.2013 | PubMed | Google Scholar

57. Danielli S, Ma Z, Pantazi E, Kumar A, Demarco B, Fischer FA, et al. (2023) **The ion channel CALHM6 controls bacterial infection-induced cellular cross-talk at the immunological synapse** *Embo j* **42**:e111450 https://doi.org/10.15252/embj.2022111450 | PubMed | Google Scholar

58. Tanis JE, Ma Z, Foskett JK (2017) **The NH(2) terminus regulates voltage-dependent gating of CALHM ion channels** *Am J Physiol Cell Physiol* **313**:C173–c86 https://doi.org/10.1152/ajpcell.00318.2016 | PubMed | Google Scholar

59. Ma Z, Paudel U, Wang M, Foskett JK (2025) **A mechanism of CALHM1 ion channel gating** *Am J Physiol Cell Physiol* **328**:C1109–c24 https://doi.org/10.1152/ajpcell.00925.2024 | PubMed | Google Scholar

60. Choi W, Clemente N, Sun W, Du J, Lü W (2019) **The structures and gating mechanism of human calcium homeostasis modulator 2** *Nature* **576**:163–7 https://doi.org/10.1038/s41586-019-1781-3 | Google Scholar

61. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. (2022) **Variant interpretation using population databases: Lessons from gnomAD** *Human Mutation* **43**:1012–30 https://doi.org/10.1002/humu.24309 | Google Scholar

62. Kwon JW, Jeon YK, Kim J, Kim SJ, Kim SJ (2021) **Intramolecular Disulfide Bonds for Biogenesis of CALHM1 Ion Channel Are Dispensable for Voltage-Dependent Activation** *Mol Cells* **44**:758–69 https://doi.org/10.14348/molcells.2021.0131 | PubMed | Google Scholar

63. Jeon YK, Choi SW, Kwon JW, Woo J, Choi SW, Kim SJ, Kim SJ (2021) **Thermosensitivity of the voltage-dependent activation of calcium homeostasis modulator 1 (calhm1) ion channel** *Biochem Biophys Res Commun* **534**:590–6 https://doi.org/10.1016/j.bbrc.2020.11.035 | PubMed | Google Scholar

64. Syrjänen JL, Epstein M, Gómez R, Furukawa H (2023) **Structure of human CALHM1 reveals key locations for channel regulation and blockade by ruthenium red** *Nat Commun* **14**:3821 https://doi.org/10.1038/s41467-023-39388-3 | PubMed | Google Scholar

65. Syrjanen JL, Michalski K, Chou TH, Grant T, Rao S, Simorowski N, et al. (2020) **Structure and assembly of calcium homeostasis modulator proteins** *Nat Struct Mol Biol* **27**:150–9 https://doi.org/10.1038/s41594-019-0369-9 | PubMed | Google Scholar

66. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. (2020) **CDD/SPARCLE: the conserved domain database in 2020** *Nucleic Acids Res* **48**:D265–d8 https://doi.org/10.1093/nar/gkz991 | PubMed | Google Scholar

67. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. (2021) **Pfam: The protein families database in 2021** *Nucleic Acids Res* **49**:D412–d9 https://doi.org/10.1093/nar/gkaa913 | PubMed | Google Scholar

68. Henschel A, Winter C, Kim WK, Schroeder M (2007) **Using structural motif descriptors for sequence-based binding site prediction** *BMC Bioinformatics* **8**:S5 https://doi.org/10.1186/1471-2105-8-s4-s5 | PubMed | Google Scholar

69. Letunic I, Khedkar S, Bork P (2020) **SMART: recent updates, new developments and status in 2020** *Nucleic Acids Research* **49**:D458–D60 https://doi.org/10.1093/nar/gkaa937 | Google Scholar

70. McInnes L, Healy J, Saul N, Großberger L (2018) **Großberger LMaJHaNSaL. UMAP: Uniform Manifold Approximation and Projection** *Journal of Open Source Software* **3**:861 Google Scholar

71. Hernández-Salmerón JE, Moreno-Hagelsieb G (2020) **Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2** *BMC Genomics* **21**:741 https://doi.org/10.1186/s12864-020-07132-6 | Google Scholar

72. Li L, Stoeckert CJ, Roos DS (2003) **OrthoMCL: identification of ortholog groups for eukaryotic genomes** *Genome Res* **13**:2178–89 https://doi.org/10.1101/gr.1224503 | PubMed | Google Scholar

73. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. (2020) **SciPy 1.0: fundamental algorithms for scientific computing in Python** *Nature Methods* **17**:261–72 https://doi.org/10.1038/s41592-019-0686-2 | Google Scholar

74. The Gene Ontology Consortium (2021) **The Gene Ontology resource: enriching a GOld mine** *Nucleic Acids Res* **49**:D325–d34 https://doi.org/10.1093/nar/gkaa1113 | PubMed | Google Scholar

Author information

Rahil Taujale[†]

Institute of Bioinformatics, University of Georgia, Athens, United States, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States
ORCID iD: 0000-0003-1292-1619

[†]These authors contributed equally to the presented work.

Sung Jin Park[†]

Department of Molecular Biosciences, Northwestern University, Evanston, United States

[†]These authors contributed equally to the presented work.

Nathan Gravel

Institute of Bioinformatics, University of Georgia, Athens, United States

Saber Soleymani

Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States

Rayna Carter

Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States

Kennady Boyd

Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States

Sarah Keuning

Department of Molecular Biosciences, Northwestern University, Evanston, United States

Zheng Ruan

Department of Biochemistry & Molecular Biology, Thomas Jefferson University, Philadelphia, United States

Wei Lü

Department of Molecular Biosciences, Northwestern University, Evanston, United States, Department of Pharmacology, Northwestern University, Evanston, United States, Chemistry of Life Processes Institute, Northwestern University, Evanston, United States
ORCID iD: 0000-0002-3009-1025

For correspondence: wei.lu@northwestern.edu

Natarajan Kannan

Institute of Bioinformatics, University of Georgia, Athens, United States, Department of Biochemistry and Molecular Biology, University of Georgia, Athens, United States
ORCID iD: 0000-0002-2833-8375

For correspondence: nkannan@uga.edu

Editors

Reviewing Editor
Andres Jara-Oseguera
The University of Texas at Austin, Austin TX, United States of America

Senior Editor
Kenton Swartz
National Institute of Neurological Disorders and Stroke, Bethesda, United States of America

Reviewer #1 (Public review):

Summary:

In the manuscript "Identification and classification of ion-channels across the tree of life: Insights into understudied CALHM channels" Taujale et al describe an interdisciplinary approach to mine the human channelome and further discover orthologues across diverse organisms, culminating in delineating co-conserved patterns in an example ion channel: CALHM. Overall, this paper comes in two sections, one where 419 human ion channels and 48,000+ channels from diverse organisms are found through a multidisciplinary data mining approach, and a second where this data is used to find co-conserved sequences, whose functional significance is validated via experiments on CALHM1 and CALHM6. Overall, this is an intriguing data-first approach to better understand even understudied ion channels like CALHM6. However, more needs to be done to pull this story together into a single coherent narrative.

Strengths:

This manuscript takes advantage of modern-day LLM tools to better mine the literature for ion channel sequences in humans and other species with orthologous ion channel sequences. They explore the 'dark channome' of understudied ion channels to better reveal the information evolution has to tell us about our own proteins, and illustrate the information this provides access to in experimental studies in the final section of the paper. Finally, they provide a wealth of information in the supplementary tables (in the form of Excel spreadsheets and a dataset on Zenodo) for others to explore. Overall, this is a creative approach to a wide-reaching problem that can be applied to other families of proteins.

Weaknesses:

Overall, while a considerable amount of work has been done for this manuscript, the presentation, both in terms of writing and figures, still can use more work even after a first round of revisions. While they have improved their discussion to more clearly describe the need for a better-curated sequence database of ion channels, and how existing resources fall short, some aspects of this process and the motivation remain unclear, especially when it comes to the CALHM sequences.

Overall, this manuscript is a valuable contribution to the field, but requires a few main things to make it truly useful. Namely, how has this approach really improved their ability to identify conserved residues in CALHM over a less-involved approach? And better organization of the first results section of the paper, which is critical to the downstream understanding of the paper, as well as some cosmetic improvements.

https://doi.org/10.7554/eLife.106134.2.sa2

Reviewer #2 (Public review):

Summary:

In this paper, the authors defined the "channelome," consisting of 419 predicted human ion channels as well as 48,000 ion channel orthologs from other organisms. Using this information, the ion channels were clustered into groups, which can potentially be used to make predictions about understudied ion channels in the groups. The authors then focused on the CALHM ion channel family, mutating conserved residues and assessing channel function.

Strengths:

The curation of the channelome provides an excellent resource for researchers studying ion channels. Supplemental Table 1 is well organized with an abundance of useful information.

Comments on revisions:

The authors have thoroughly addressed my concerns and the manuscript is substantially improved. I have just a few suggestions regarding wording/clarification.

In Supplemental Figure 4, the Western blots (n=3) were quantitated, but the surface biotinylation was not. While I suppose that it is fine to just show one representative experiment for the biotinylation assay, the authors should indicate in the legend how many times this was done. It is essential to know whether these data in Supplemental Figure 4E, F are reproducible as they are absolutely critical for interpretation of all of the data in Figure 5.

https://doi.org/10.7554/eLife.106134.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Reviewing Editor Comments:***
>
> *(A) Revisions related to the first part, regarding data mining and curation:*
>
> *(1) One question that arises with the part of the manuscript that discusses the identification and classification of ion channels is whether these will be made available to the wider public. For the 419 human sequences, making a small database to share this result so that these sequences can be easily searched and downloaded would be desirable. There are a variety of acceptable formats for this: GitHub/figshare/zenodo/university website that allows a wider community to access their hard work. Providing such a resource would greatly expand the impact of this paper. The same question can be asked of the 48,000+ ion channels from diverse organisms.*

We thank the reviewer for providing this important feedback. While the long term plan is to provide access to these sequences and annotations through a knowledge base resource like Pharos, we agree with the comments that it would be beneficial to have these sequences made available with the manuscript as well. We have compiled 3 fasta files containing the following: 1) Full length sequences for the curated 419 ion channel sequences. 2) Pore containing domain sequences for the 343 pore domain containing human ion channel sequences. 3) All the identified orthologs for the human ion channels.

For each sequence in these files, we have extended the ID line to include the most pertinent annotation information to make it readily available. For example, the id>sp|P48995|TRPC1_HUMAN|TRP:VGIC--TRP-TRPC|pore-forming|dom:387-637 provides the classification, unit and domain bounds for the human TRPC1 in the fasta file itself.

These files have been uploaded to Zenodo and are available for download with doi 10.5281/zenodo.16232527. We have included this in the Data Availability statement of the manuscript as well.

> *(2) Regarding the 48,000+ sequences, what checks have been done to confirm that they all represent bona fide, full-length ion channel sequences? Uniprot contains a good deal of unreviewed sequences, especially from single-celled organisms. The process by which true orthologues were identified and extraneous hits discarded should be discussed in more detail, and all inclusion criteria should be described and justified, clearly illustrating that the risk of gene duplicates and fragments in this final set of ion channel orthologues has been avoided. Related to this, does this analysis include or exclude isoforms?*

We thank the reviewer for raising this important point. Our selection of curated proteomes and the KinOrtho pipeline for orthology detection returns, up to an extent, reliable orthologous sequence sets. In brief, our database sequences are retrieved from full proteomes that only include proteins that are part of an official proteome release. Thus, they are mapped from a reference genome to ensure species-specific relevance and avoid redundancy. The >1500 proteomes in this analysis were selected based on their wider use in other orthology detection pipelines like OMA and InParanoid. Our orthology detection pipeline, KinOrtho, performs a fulllength and a domain-based orthology detection which ensures that the orthologous relationships are being defined based on the pore-domain sequence similarity.

But we agree with the reviewer that this might leave room for extraneous, fragments or misannotated sequences to be included in our results. Taking this into careful consideration, we have expanded our sequence validation pipeline to include additional checks such as checking the uniport entry type, protein existence evidence and sequence level checks such as evaluating the compositional bias, non-standard codons and sequence lengths. These validation steps are now described in detail in the Methods section under orthology analysis (lines 768-808). All the originally listed orthologous sequences passed this validation pipeline and thus provide additional confidence that they are bona fide full length ion channel sequences.

We have also expanded this section (lines 758 – 766) to provide more details of the KinOrtho pipeline for orthology detection, which is a previously published method used for orthology detection in kinases by our lab.

Finally, our orthology analysis excludes isoforms and only spans the primary canonical sequences that are part of the UniProt Proteomes annotated sequence set. The isoforms that are generally available in UniProt Proteomes in a separate file named *_additional.fasta were not included in this analysis.

> *(3) The decision to show the families of ion channels in Figure 1 as pie charts within a UMAP embedding is intriguing but somewhat non-intuitive and difficult to understand. Illustrating these results with a standard tree-like visualization of the relationship of these channels to each other would be preferred.*

We appreciate the feedback provided by the reviewer, and understand that a standard tree-like visualization would be much easier to interpret and familiar than a bubble chart based on UMAP embeddings. However, we opted to use the bubble chart for the following reasons:

Low sequence similarity: the 419 human ICs share very minimal sequence similarity, falling in the twilight zone or lower (Dolittle, 1992; PMID:1339026). Thus, traditional multiple sequence alignment and phylogenetic reconstruction methods perform very poorly and generate unreliable or even misleading results. To explore the practicality of this option, we pursued performing a multiple sequence alignment of just 3 of the possibly related IC families as suggested by reviewer 2 (CALHM, Pannexins, and Connexins) using the state of the art structure based sequence alignment method Foldmason (doi: https://doi.org/10.1101/2024.08.01.606130). Even then, the sequence alignment and the resulting tree for just these 3 families were poor and unreliable, as illustrated in the attached Author response Image 2.

Protein embeddings based clustering: Novel LLM based approaches such as the protein language model embeddings offer ways to overcome these limitations by capturing sequence, structure, function and evolutionary properties in a high-dimensional space. Thus, we employed this model using DEDAL followed by UMAP for dimensionality reduction, which preserves biologically meaningful local and global relationships.

Abstraction at family level: In Figure 1, we aggregate individual channels into family bubbles with their positions representing the average UMAP coordinates of their members. This offers a balance between an intuitive view of how IC families are distributed in the embedding space and reflects potential functional and evolutionary proximities, while not being impeded by individual IC relationships across families.

We have revised the figure legend (lines 1221 – 1234) with additional description of the visualization and the process used to generate it, and the manuscript text (lines 248-270) provides the rationale behind the selection of this method.

> *(4) A strength of this paper is the visualization of 'dark' ion channels. However, throughout the paper, this could be emphasized more as the key advantage of this approach and how this or similar approaches could be used for other families of proteins. Specifically, in the initial statement describing 'light' vs 'dark channels', the importance of this distinction and the historical preference in science to study that which has already been studied can be discussed more, even including references to other studies that take this kind of approach. An example of a relevant reference here is to the Structural Genomics Consortium and its goals to achieve structures of proteins for which functions may not be well-characterized. Clarifying these motivations throughout the entire paper would strengthen it considerably.*

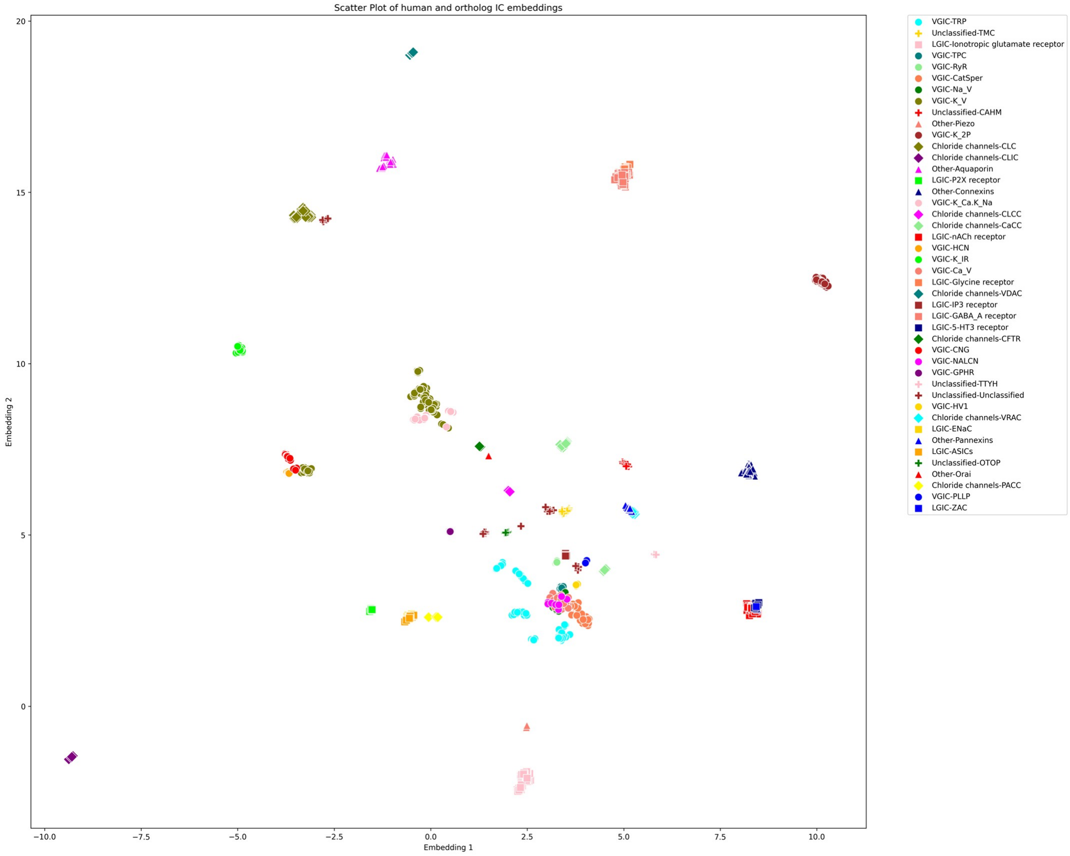
We thank the reviewer for this constructive comment and agree that highlighting the strength of visualizing “dark” channels and prioritizing them for future studies would strengthen the paper. As suggested, we have revised the text throughout the paper (lines 84-89, 176-180) to contextualize and emphasize this distinction. We have also added a reference for the Structural Genomics Consortium, which, along with resources like IDG, has provided significant resources for prioritizing understudied proteins.

> *(5) Since the authors have generated the UMAP visualization of the channome, it would be interesting to understand how the human vs orthologue gene sets compare in this space.*

We appreciate the reviewer’s input. It is an interesting idea to explore the UMAP embedding space for the human ICs along with their orthologs. The large number of orthologous sequences (>37,000) would certainly impose a computational challenge to generate embeddings-based pairwise alignments across all of them. Downstream dimensionality reduction from such a large set and the subsequent visualization would also suffer from accuracy and interpretability concerns. However, to follow up on the reviewer’s comments, we selected orthologous sequences from a subset of 12 model organisms spanning all taxa (such as mouse, zebrafish, fruit fly, C. elegans, A. thaliana, S. cerevisiae, E. coli, etc.).This increased the number of sequences for analysis to 1094 from 343, which is still manageable for UMAP. Using the exact same method, we generated the UMAP embeddings plot for this set as shown below.

Author response image 1.

UMAP embeddings of the human ICs alongside orthologs from 12 model organisms



As shown above, we observed that each orthologous set forms tight, well-defined clusters, preserving local relationships among closely related sequences. For example, a large number of VGICs cluster more closely together compared to Supplementary Figure 1 (with only the human ICs). However, families that were previously distant from others now appear to be even more scattered or pushed further away, indicating a loss of global structure. This pattern suggests that while local distances are well preserved, the global topology of the embedding space could be compromised. Moreover, we find that the placement of ICs with respect to other families is highly sensitive to the parameter choices (e.g., n_neighbors and min_dist), an issue which we did not encounter when using only the human IC sequences. The inclusion of a large number of orthologous sequences that are highly similar to a single human IC but dissimilar to others skews the embedding space, emphasizing local structure at the expense of global relationships.

Since UMAP and similar dimensionality reduction methods prioritize local over global structure, the resulting embeddings accurately reflect strong ortholog clustering but obscure broader interfamily relationships. Consequently, interpreting the spatial arrangement of human IC families with respect to one another becomes unreliable. We have made this plot available as part of this response, and anyone interested can access this in the response document.

> *(6) Figure 1 should say more clearly that this is an analysis of the human gene set and include more of the information in the text: 419 human ion channel sequences, 75 sequences previously unidentified, 4 major groups and 55 families, 62 outliers, etc. Clearer visualizations of these categories and numbers within the UMAP (and newly included tree) visualization would help guide the reader to better understand these results. Specifically, which are the 75 previously unidentified sequences?*

We thank the reviewer for the comments. To address this, we have revised Figure 1 and added more information, including a clear header that states that these are only human IC sets, numbers showing the total number of ICs, and the number of ICs in each group. We have further included new Supplementary Figure 2 and Supplementary Table 2, which show the overlap of IC sequences across the different resources. Supplementary Figure 2 is an upset plot that provides a snapshot of the overlap between curated human ICs in this study compared to KEGG, GtoP, and Pharos. Supplementary Table 2 provides more details on this overlap by listing, for each human IC, whether they are curated as an IC in the 3 IC annotation resources. We believe these additions should provide all the information, including the unidentified sequences we are adding to this resource.

> *(7) Overall, the manuscript needs to provide a clearer description of the need for a better-curated sequence database of ion channels, as well as how existing resources fall short.*

We thank the reviewer for pointing out this important gap in the description. As suggested, we have revised the text thoroughly in the Introduction section to address this comment. Specifically, we have added sections to describe existing resources at sequence and structure levels that currently provide details and/or classification of human ion channels. Then, we highlight the facts that these resources are missing some characterized pore-containing ICs, do not include any information on auxiliary channels, and lack a holistic evolutionary perspective, which raises the need for a better-curated database of ion channels. Please refer to lines 57-63, 73-79, and 95 – 119 for these changes and additions.

> *(8) Some of the analysis pipeline is unclear. Specifically, the RAG analysis seems critical, but it is unclear how this works - is it on top of the GPT framework and recursively inquires about the answer to prompts? Some example prompts would be useful to understand this.*

We thank the reviewer for highlighting this gap in explanation. We understand that the details provided in the Methods and Supplementary Figure 1 may not have sufficiently explained the pipeline, and are missing some important details. The RAG pipeline leverages vector-based retrieval integrated with OpenAI's GPT-4o model to systematically search literature and generate evidence-based answers. The process is as follows:

Literature sources (PubMed articles) relevant to the annotated ion channels were converted into vector representations stored in a Qdrant database.

Queries constructed from the annotated IC dataset were submitted to the vector database, retrieving contextually relevant literature segments.

Retrieved contexts served as inputs to the GPT-4o model, which produced structured JSON-formatted responses containing direct evidence regarding ion selectivity and gating mechanisms, along with associated confidence scores.

To clarify this further, we have rewritten the relevant subsection in lines 649 - 718. Now, this section provides a detailed description of the RAG pipeline. Also, we have improved Supplementary Figure 1 to provide a clearer description of the pipeline. We have also provided an example prompt template to illustrate the query. These additions clarify how the pipeline functions and demonstrate its practical utility for IC annotation.

> *(9) The existence of 76 auxiliary non-pore containing 'ion channel' genes in this analysis is a little confusing, as it seems a part of the pipeline is looking for pore-lining residues. Furthermore, how many of these are picked up in the larger orthologues search? Are these harder to perform checks on to ensure that they are indeed ion channel genes? A*

> *further discussion of the choice to include these auxiliary sequences would be relevant. This could just be further discussion of the literature that has decided to do this in the past.*

We thank the reviewer for this comment, and agree that further clarification of our selection and definition of auxiliary IC sequences would be helpful. As the reviewer has pointed out, one of the annotation pipeline steps is indeed looking for the pore-lining residues. Any sequences that do not have a pore-containing domain are then considered to be auxiliary, and we search for additional evidence of their binding with one of the annotated pore-containing ICs. If such evidence is not found in the literature, we remove them from our curated IC list.

In response to the above comment, we have revised the manuscript text to provide these details. In the Introduction section, we have added references to previous literature that have described auxiliary ICs and also pointed out that the existing ion channel resources do not account for such auxiliary channels (lines 73-79, 107-108,148-149). We have also expanded the Methods section to describe the selection and definition of auxiliary channels (lines 640-646).

With regards to the orthology analysis, since auxiliary channels do not have a pore domain, and our orthology pipeline requires a pore domain similarity search and hit, we did not include them in this part of the analysis. We have clarified the text in the Results section to ensure this is communicated properly throughout the manuscript (lines 212-215, 260-263).

> *(10) Why are only evolutionary relationships between rat, mouse, and human shown in Figure 3A? These species are all close on the evolutionary timeline.*

We thank the reviewer for this comment. Figure 3A currently provides a high-level evolutionary relationship across the 6 human CALHM members as a pretext for the pattern based Bayesian analysis. However, since this analysis is based on a wider set of orthologs that span taxa, we agree that a larger tree that includes more orthologs is warranted.

We have now revised Figure 3A to include an expanded tree that includes 83 orthologs from all 6 human CALHM members spanning 14 organisms from different taxa, ranging from mammals, fishes, birds, nematodes, and cnidarians. The overall structure of the tree is still consistent with 2 major clades as before, with CALHM 1 and 3 in the first clade and CALHM 2,4,5, and 6 in the second clade, with good branch support.

> *(B) Revisions related to the second part, regarding the analysis of CAHLM channel mutations:*
>
> *(1) It would strengthen the manuscript if it included additional discussion and references to show that previous methods to analyze conserved residues in CALHM were significantly lacking. What results would previous methods give, and why was this not enough? Were there just not enough identified CALHM orthologues to give strong signals in conservation analysis? Also, the amino acid conservation between CLHM-1 and CALHM1 is extremely low. Thus, there are other CALHM orthologs that give strong signals in conservation analysis. There are ~6 papers that perform in-depth analysis of the role of conserved residues in the gating of CALHM channels (human and C. elegans) that were not cited (Ma et al, Am J Physiol Cell Physiol, 2025; Syrjanen et al, Nat Commun, 2023; Danielli et al, EMBO J, 2023; Kwon et al, Mol Cells, 2021; Tanis et al, Am J Physiol Cell Physiol, 2017; Tanis et al, J Neurosci, 2013; Ma et al, PNAS, 2013) - these data needs to be discussed in the context of the present work.*

We thank the reviewer for the comment and agree that these are excellent studies that have advanced understanding of conserved residues in CALHM gating. While their analyses compared a limited set of sequences, focusing on residues conserved in specific CALHM homologs or species like C. elegans, our analysis encompasses thousands of sequences across the entire CALHM family, allowing us to identify residues conserved across all family members over evolution. We also coupled this sequence analysis with hypotheses derived from our published structural studies (Choi et al., Nature, 2019), which highlighted the NTH/S1 region as a critical element in channel gating. Based on this, we focused on evolutionarily conserved residues in the S1–S2 linker and at the interface of S1 with the rest of the TMD, reasoning that if S1 movement is essential for gating, these two structural elements (acting as a hinge and stabilizing interface, respectively) would be key determinants of the conformational dynamics of S1. These regions have been largely overlooked in previous studies. As a result, the residues highlighted in our study do not overlap with those previously reported but instead provide complementary insights into gating mechanisms in this unique channel family. Together, our study and the published literature suggest that many regions and residues in CALHM proteins are critical for gating: while some are conserved across the entire family evolutionarily, others appear conserved only within certain species or subfamilies.

To address the reviewer's comment, and to highlight the points mentioned above, we have added a brief discussion of these studies and the relevant citations in the revised manuscript (lines 378– 385, 563–576).

> *(2) Whereas the current-voltage relations for WT channels are clearly displayed, the data that is shown for the mutants does not allow for determining if their gating properties are indeed different than WT.*
>
> *First, the current amplitudes for the mutants were quantified at just one voltage, which makes it impossible to determine if their voltage-dependence was different than WT, which would be a strong indicator for an effect in gating. Current-voltage relations as done for the WT channels should be included for at least some key mutations, which should include additional relevant controls like the use of Gd3+ as an inhibitor to rule out the contribution of some endogenous currents.*

We thank the reviewer for this comment. To address this, we performed additional experiments using a multi-step pulse protocol to obtain current-voltage relations for WT CALHM1, CALHM1(I109W), WT CALHM6, and CALHM6(W113A). Our initial two-step protocol (−80 mV and +120 mV) covers both the physiological voltage range and the extended range commonly used in biophysical characterization of ion channels. Most mutants did not exhibit channel activation even within this broad range. We therefore focused on the three mutants that did show substantial activation to perform full I–V analysis as suggested. In all groups, currents activated at 37 °C were significantly inhibited by Gd^{3+}, consistent with published reports (Ma et al., AJP 2025; Danielli et al., EMBO J 2023; Syrjänen et al., Nat Commun 2023). Notably, for CALHM6(Y51A), while this mutation did not significantly alter current amplitudes at positive membrane potentials, it markedly reduced currents at negative potentials, rendering the channel outwardly rectifying and altering its voltage dependence. These new data are incorporated into Figure 5 (panels A–O) and discussed in the manuscript. Figure 5 now also shows current amplitudes at both +120 mV and −80 mV in 0 mM Ca^{2+} at 37 °C to facilitate direct comparison between WT and mutants. The previous data at 5 mM Ca^{2+} and 0 mM Ca^{2+} at 22 °C have been moved to Supplementary Figure 5 as requested.

> *Second, it is unclear whether the three experimental conditions (5 mM Ca^{2+}, and 0 Ca^{2+}, at 22 and 37C) were measured in the same cell in each experiment, or if they represent different experiments. This should be clarified. If measurements at each condition were done in the same experiment, direct comparison between the three conditions within each individual experiment could further help identify mutations with altered gating.*

We thank the reviewer for pointing this out and apologize for the confusion. All three conditions (5 mM Ca^{2+} at 22 °C, 0 mM Ca^{2+} at 22 °C, and 0 mM Ca^{2+} at 37 °C) were sequentially measured in the same cell within each experiment. The currents were then averaged across cells and plotted for each group.

> *Third, in line 334, the authors state that "expression levels of wild-type proteins and mutants are comparable." However, Western blots showing CALHM protein abundance (Supplementary Fig. 3) are not of acceptable quality; in the top blot, WT CALHM1 appears too dim, representative blots were not shown for all mutants, and individual data points should be included on the group data quantitation of the blots, together with a statistical test comparing mutants with the WT control.*

We thank the reviewer for the comment and agree that representative blots were not shown for all mutants. Supplementary Figure 4 (previously Supplementary Figure 3) has been updated to include representative blots for all mutants, individual data points in the quantification, and statistical tests comparing each mutant to the WT control.

> *A more serious concern is that the total protein quantitation is not very informative about the functional impact of mutations in ion channels, because mutations can severely impact channel localization in the plasma membrane without reducing the total protein that is translated. In mammalian cells, CALHM6 is localized to intracellular compartments and only translocates to the plasma membrane in response to an activating stimulus (Danielli et al, EMBO J, 2023). Thus, if CALHM6 is only intracellular, the protein amount would not change, but the measured current would. Abundant intracellular CALHM1 has also been observed in mammalian cells transfected with this protein (Dreses-Werringloer et al., Cell, 2008). Quantitation of surface-biotinylated channels would provide information on whether there are differences between the constructs in relation to surface expression rather than gating. An alternative approach to biotinylation would be to express GFP-tagged constructs in Xenopus oocytes and look for surface expression. This is what has been done in previous CALHM channel studies.*
>
> *Without evidence for the absence of defects in localization or clear alterations in gating properties, it is not possible to conclude whether mutant channels have altered activity. Does the analysis of sequences provide any testable hypotheses about substitutions with different side chains at the same position in the sequence?*

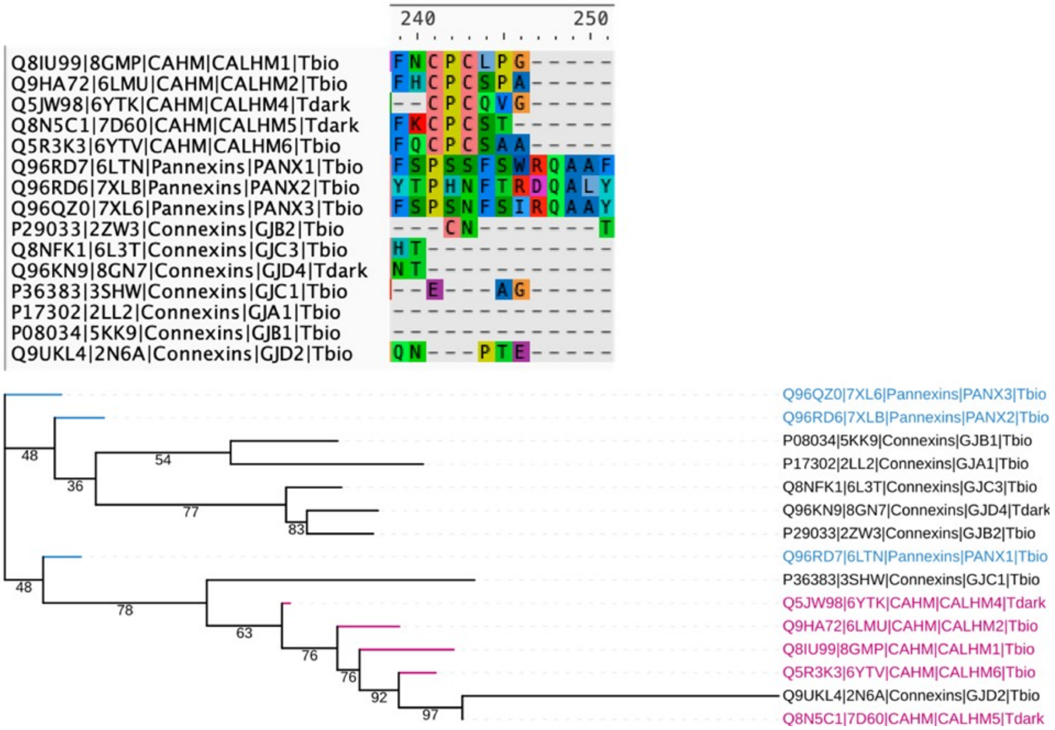
We thank the reviewer for this very important comment. We agree that total protein levels alone do not distinguish between intracellular retention and proper trafficking to the plasma membrane. To address this, we performed surface biotinylation assays for all WT and mutant CALHM1 and CALHM6 constructs to assess their plasma membrane localization. The results show that mutants have either comparable or substantially higher surface expression levels than WT, consistent with the Western blot data. Together, these findings support our original interpretation that the observed differences in electrophysiological currents are not due to trafficking defects but reflect functional effects. These new data are presented in Supplementary Figure 5.

> *(3) Line 303 - 13 aligned amino acids were conserved across all CALHM homologs - are these also aligned in related connexin and pannexin families? It is likely that cysteines and proline in TM2 are since CALHM channels overall share a lot of similarities with connexins and pannexins (Siebert et al, JBC, 2013). As in line 207, it would be expected that pannexins, connexins, and CALHM channel families would group together. Related to this, see Line 406 - in connexins, there is also a proline kink in TM2 that may play a role in mediating conformational changes between channel states (Ri et al, Biophysical Journal, 1999). This should be discussed.*

We thank the reviewer for the suggestion. We attempted a structure based sequence alignment of representative structures from all 3 families (CALHM, connexins and pannexins), but the resulting alignments are very poor and have a lot of gapped regions, making it very difficult to comment on the similarities mentioned in this comment. This is actually expected, as although CALHM, connexins, and pannexins are all considered “large-pore” channels, the TMD arrangement and conformation of CALHM are distinct from those of connexins and pannexins. Below, we have included a snapshot of the alignment at the conserved cysteine regions of the CALHM homologs, along with the resulting tree, which has very low support values and has difficulty placing the connexins properly, making it difficult to interpret.

Author response image 2.

Structure based sequence alignment and phylogenetic analysis of available crystal structures of members from the CALHM, Pannexin and Connexin families. Top: The resulting sequence alignment is very sparse and does not show conservation of residues in the TM regions. The CPC motif with conserved cysteines in CALHM family is shown. Bottom: Phylogenetic tree based on the alignment has low support values making it difficult to interpret.


240
250
Q8IU99|8GMP|CAHM|CALHM1|Tbio FNCPCLPG-----
Q9HA72|6LMU|CAHM|CALHM2|Tbio FHCPCSPA-----
Q5JW98|6YTK|CAHM|CALHM4|Tdark --CPCQVG-----
Q8N5C1|7D60|CAHM|CALHM5|Tdark FKCPCST------
Q5R3K3|6YTV|CAHM|CALHM6|Tbio FQCPCSAA-----
Q96RD7|6LTN|Pannexins|PANX1|Tbio FSPSSFSWRQAAF
Q96RD6|7XLB|Pannexins|PANX2|Tbio YTPHNFTRDQALY
Q96QZ0|7XL6|Pannexins|PANX3|Tbio FSPSNFSIRQAAY
P29033|2ZW3|Connexins|GJB2|Tbio ---CN-------T
Q8NFK1|6L3T|Connexins|GJC3|Tbio HT-----------
Q96KN9|8GN7|Connexins|GJD4|Tdark NT-----------
P36383|3SHW|Connexins|GJC1|Tbio --E---AG-----
P17302|2LL2|Connexins|GJA1|Tbio -------------
P08034|5KK9|Connexins|GJB1|Tbio -------------
Q9UKL4|2N6A|Connexins|GJD2|Tbio QN---PTE-----
Q96QZ0|7XL6|Pannexins|PANX3|Tbio
Q96RD6|7XLB|Pannexins|PANX2|Tbio
P08034|5KK9|Connexins|GJB1|Tbio
P17302|2LL2|Connexins|GJA1|Tbio
Q8NFK1|6L3T|Connexins|GJC3|Tbio
Q96KN9|8GN7|Connexins|GJD4|Tdark
P29033|2ZW3|Connexins|GJB2|Tbio
Q96RD7|6LTN|Pannexins|PANX1|Tbio
P36383|3SHW|Connexins|GJC1|Tbio
Q5JW98|6YTK|CAHM|CALHM4|Tdark
Q9HA72|6LMU|CAHM|CALHM2|Tbio
Q8IU99|8GMP|CAHM|CALHM1|Tbio
Q5R3K3|6YTV|CAHM|CALHM6|Tbio
Q9UKL4|2N6A|Connexins|GJD2|Tbio
Q8N5C1|7D60|CAHM|CALHM5|Tdark
48
54
36
77
83
48
78
63
76
76
92
97

> *(4) Line 36 - This work does not have experimental evidence to show that the selected evolutionarily conserved residues alter gating functions.*

Our electrophysiology data demonstrate that the selected evolutionarily conserved residues have a major impact on CALHM1 and CALHM6 gating. As shown in Figure 5, mutations at these residues produce two distinct phenotypes: (1) nonconductive channels, and (2) altered voltage dependence, resulting in outward rectification. Importantly, these functional changes occur despite normal total expression and surface trafficking, as confirmed by Western blotting and surface biotinylation (Supplementary Figure 4). These findings indicate that the affected residues are critical for the conformational dynamics underlying channel gating rather than for protein expression or localization.

> *(5) Line 296-297 - This could also be put in the context of what we already know about CALHM gating. While all cryo EM structures of CALHM channels are in the open state, we still do understand some things about gating mechanism (Tanis et al Am J Physiol Cell Physiol, Cell Physiol 2017; Ma et al Am J Physiol Cell Physiol, Cell Physiol 2025) with the NT modulating voltage dependence and stabilizing closed channel states and the voltage dependent gate being formed by proximal regions of TM1.*

Thank you for providing this suggestion. As suggested, we have revised the text to place our findings in the context of current knowledge about CALHM gating and have added the relevant citations (lines 370-373).

> *(6) Lines 314-315 - Just because residues are conserved does not mean that they play a role in channel gating. These residues could also be important for structure, ion selectivity, etc.*

We agree that evolutionary conservation alone does not imply a role in gating. However, our hypothesis derives from the positioning of these conserved residues, and previous studies that have indicated the importance of the NTH/S1 region for channel gating function. More importantly, our electrophysiology data indicate that these conserved residues specifically impact channel gating in CALHM1 and CALHM6. We have revised the text in lines 404-406 to clarify this further.

> *(7) Line 333 - while CALHM6 is less studied than CALHM1, there is knowledge of its function and gating properties. Should CALHM6 be considered a "dark" channel? The IDG development level in Pharos is Tbio. There have been multiple papers published on this channel (ex: Ebihara et al, J Exp Med, 2010; Kasamatsu et al, J Immunol 2014; Danielli et al, EMBO J, 2023).*

We thank the reviewer for noting this important discrepancy. We have updated the text and labels related to CALHM6 to reflect its status as Tbio in the manuscript.

> *(8) Please cite Jeon et al., (Biochem Biophys Res Commun, 2021), who have already shown temperature-dependence of CALHM1.*

Thank you for the comment. We have added the citation.

> *(9) It would be helpful to have a schematic showing amino acid residues, TM domains, highlighted residues mutated, etc.*

Thank you for the suggestion. We have revised the figure and added labels for the TM domains, and highlighted the mutated residues.

> ***Reviewer #1 (Recommendations for the authors):***
>
> *(1) Why in the title is 'ion-channels' hyphenated but in the text it is not?*

This has been changed.

> *(2) Line 78: 'Cryo-EM' is not defined before the acronym is used.*

This has been fixed.

> *(3) Typo in line 519: KinOrthto.*

This has been fixed.

> *(4) Capitalizing 'Tree of Life' is a bit strange in section 2 of the results and the Discussion.*

We have removed the capitalization as suggested.

> *(5) In Figure 3 and Supplementary Figure 4A, the gene names in the tree are CAHM and not CALHM - I assume this is an error.*

This has been made consistent to CALHM.

> *(6) Font sizes throughout all figures, with the exception of Figure 1, need to be more legible. The X-axis labels in Figure 2A are hard to read, for example (though I can see that there is also the CAHM/CALHM typo here...). A good rule of thumb is that they should be the same size as the manuscript text. Furthermore, the grey backgrounds of Figure 4 and Figure 5 are off-putting; just having a white background here should be sufficient.*

This has been addressed. We have increased the font size in all figures with these revisions. The styling for Figure 4 and 5 has also been made consistent with other figures.

> ***Reviewer #2 (Recommendations for the authors):***
>
> *(1) Line 36 - This work does not have experimental evidence to show that the selected evolutionarily conserved residues alter gating functions.*

Addressed in comment #4 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(2) Line 168 - should also be Supplemental Table 1.*

This has been addressed.

> *(3) Line 170 - 419 human ion channel sequences were identified and this was an increase of 75 sequences over previous number. Which 75 proteins are these?*

This is now shown in Supplementary Figure 2 and Supplementary Table 2. Supplementary Figure 2 shows an Upset plot with the number of sequences that overlap across databases and the novel sequences that we have added as part of this study. The 75 specifically refers to the sequences that were not included in Pharos, which was chosen to refer to this number since it has the highest number of ICs listed out of all the other resources. Further, Supplementary Table 2 now provides a list of individual ICs and whether they were present in each of the 3 databases compared.

> *(4) Line 289 - Ca2+ (not Ca); other similar mistakes throughout the manuscript*

These have been fixed.

> *(5) Line 291-292 - Please include more about functions for CALHM channels; ex. CALHM1 regulates cortical neuron excitability (Ma et al, PNAS 2012), CLHM-1 regulates locomotion and induces neurodegeneration in C. elegans (Tanis et al. Journal of Neuroscience 2013); see above for references on CALHM6 function.*

We have added the functions as suggested.

> *(6) Line 296-297 - This could also be put in the context of what we already know about CALHM gating. While all cryo EM structures of CALHM channels are in the open state, we still do understand some things about gating mechanism (Tanis et al Am J Physiol Cell Physiol, Cell Physiol 2017; Ma et al Am J Physiol Cell Physiol, Cell Physiol 2025) with the NT modulating voltage dependence and stabilizing closed channel states and the voltage dependent gate being formed by proximal regions of TM1.*

Addressed in comment #5 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(7) Lines 314-315 - Just because residues are conserved does not mean that they play a role in channel gating. These residues could also be important for structure, ion selectivity, etc.*

Addressed in comment #6 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(8) Line 333 - While CALHM6 is less studied than CALHM1, there is knowledge of its function and gating properties. Should CALHM6 be considered a "dark" channel? The IDG development level in Pharos is Tbio. There have been multiple papers published on this channel (ex: Ebihara et al, J Exp Med, 2010; Kasamatsu et al, J Immunol 2014; Danielli et al, EMBO J, 2023).*

Addressed in comment #7 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(9) Line 627 - Do you mean that 5 mM CaCl2 was replaced with 5 mM EGTA in 0 Ca2+ solution?*

This is correct.

> *(10) Why are only evolutionary relationships between rat, mouse, and human shown in Figure 3A? These species are all close on the evolutionary timeline.*

Addressed in comment #10 for Part A Revisions related to the first part, regarding data mining and curation above.

> *(11) Figure 5 - no need to show the currents at room temperature in the main text since there are robust currents at 37 degrees; this could go into the supplement. Also, please cite Jeon et al. (Biochem Biophys Res Commun, 2021), who have already shown temperature-dependence of CALHM1.*

Addressed in comment #8 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(12) It would be helpful to have a schematic showing amino acid residues, TM domains, highlighted residues mutated etc.*

Addressed in comment #9 for Part B Revisions related to the second part, regarding the analysis of CAHLM channel mutations above.

> *(13) Use of S1-S4 to refer to the transmembrane "segments" is not standard; rather, TM1-TM4 would generally be used to refer to transmembrane domains.*

We have used the S1–S4 helix notation to maintain consistency with the nomenclature employed in our previous study (Choi et al., Nature, 2019).

https://doi.org/10.7554/eLife.106134.2.sa0