Title

Identification and classification of ion-channels across the tree of life: Insights into understudied CALHM channels

Authors and affiliations

Rahil Taujale[1,4], Sung Jin Park[2], Nathan Gravel[1], Saber Soleymani[4], Rayna Carter[4], Kennady Boyd[4], Sarah Keuning[2], Zheng Ruan[3], Wei Lü[2], Natarajan Kannan[1,4*]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA 30606

[2]Van Andel Institute, 333 Bostwick Ave., N.E., Grand Rapids, MI 49503

[3]Department of Biochemistry & Molecular Biology, Thomas Jefferson University, Philadelphia, PA 19107

[4]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30606

* Corresponding author: nkannan@uga.edu

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 evolutionary-constrained residues significantly alter 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.

Keywords: ion channel classification, LLM, RAG, human ion channel orthology, CALHM electrophysiology

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]. 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 [13]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database has cataloged around 315 human ICs [14, 15], while Pharos [16] and The Human Genome Organization (HUGO) [17] 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. 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 [18, 19].

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, 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 [20]. However, 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.

The structure space of the channelome is being rapidly filled in by crystal structures. Cryo-EM studies and analysis of these structures have shown that despite extensive variation in primary sequences, several channels adopt common structural folds and mechanisms of action [21]. For example, several ICs (K, Na, Ca, TRP, and 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 [22, 23]. Thus, quantitative comparisons of these commonly conserved regions at the sequence and structural level can provide new residue-level insights into IC function, much like in other large protein families such as kinases and glycosyltransferases [24-27]. Additionally, a comprehensive evolutionary analysis that includes orthologous sequences across different organisms from the Tree of Life can add valuable insights into the evolutionary constraints that govern the conservation of critical functional residues. Currently, such studies for ICs are limited to a select few closely related organisms and do not span the full IC complement [28, 29]. Thus, this lack of a unified resource that systematically identifies and classifies ICs across diverse organisms presents a major challenge in using evolutionary features for function prediction.

To address these challenges, 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 source alongside other contextual data for functional annotations such as regulatory interactions, ion-selectivity, and pore-lining residues. 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 orthologs from diverse taxonomic groups spanning metazoans, fungi, protists, bacteria, and archaea. We demonstrate the application of these sequences in annotating understudied Calcium Homeostasis Modulator channels (CALHMs) at residue-level resolution using Bayesian statistical approaches [30] 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 [31], as well as Pharos [16], KEGG [14, 15], Guide to Pharmacology (GtoP) [32], and HUGO Gene Nomenclature Committee (HGNC) [17] 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 (light/dark) designation from Pharos. 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). We also provide the ions, gating mechanism, and UniProt functional description in this annotation category. To further 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. 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 TM and pore domain-related labels has been curated extensively using various sources and predictors. We start by annotating the TM regions and pores containing functional domains using the literature and prediction tools, as described in the Methods. TM predictions by TMHMM [33] and Phobius [34] are provided alongside the TM annotations provided in UniProt. First, we supplement this information by adding TM information based on the literature review and the source. Then, we further analyze any ICs with an experimentally generated structure using the MOLE software [35] 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. To provide functional context to this classification, sequences, where a pore region was not identified, were classified as auxiliary following the definition provided by Gurnett et al. 1996 [36]. A snapshot of the list of annotations we collected through this process is shown in Table 1, using Aquaporin 1 as an example.

As a final step, we performed sequence similarity searches to identify sequences that share significant similarity to any of the curated sequences and subjected them to all the annotation steps to check if any of them qualify as an IC. 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.

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]. These sequences were classified into four major groups—Voltage-Gated Ion Channels (VGICs), Ligand-Gated Ion Channels (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, OTOP, TMC, and 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 bioinformatics approaches such as sequence alignment or HMM models could not account for the vast divergence in primary sequence across IC families and thus not applicable for ICs. Instead, we relied on representations (embeddings) derived from evolutionary scale protein language models [37] 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 [38] to perform a pairwise sequence alignment. The resulting all-vs-all sequence similarity scores were used to generate UMAP embedding scores (Supplementary Figure 2). An

average of 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 ICs did not have a defined domain, they were not used to generate the alignment. Hence, auxiliary IC families are placed at the bottom of the figure. Most of the VGIC families group together on the top right except for K_{IR}, $K_{2P,}$ and PLLP, where K_{2P} and PLLP are more centrally dispersed, while K_{IR} falls closer to LGIC families ASICs, ENAC, and P2X receptor, all of which share similar transmembrane topology with only two transmembrane helices. The 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, VRAC, Connexins, CLCC, and CALHM families group closer together centrally, indicating shared features.

We also mapped Target Development Level annotations defined by the IDG [18] in Pharos [16] to highlight the depth of understudied ICs within each group. 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 [39] 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 [16]. 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, CaCC, CACN, $K_{2P,}$ and the unclassified families accounting for the remaining Tdark ICs, highlighting the need for more studies on these families. On the other hand, VGICs K_V, Na_V, Ca_V, RyR, and LGICs $GABA_A$ receptor, 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 [31]using the full complement of the annotated human IC sequences. To this end, we used a graph-based orthology inference approach [40] 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 [31]. 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 [40], 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 conducting IC sequences for this analysis.

Using this method, we could identify more than 48,000 IC orthologs 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 Gamma-Aminobutyric acid type A ($GABA_A$) 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 Transmembrane channel-like protein 4 (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 2.

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 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 3). 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 [41]. 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) [30]. 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 [42]. 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 and membrane depolarization and that the heteromeric CALHM1 and CALHM3 channels are implicated in the ATP release during taste perception [43]. 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 by the first transmembrane segment S1 along with a short N-terminal helix preceding S1. While the exact gating mechanism of the CALHM channels is still unclear, it is hypothesized that the conformational state of S1 plays a crucial role in determining the pore size.

Figure 3A shows a phylogenetic tree depicting the evolutionary relationship across the 6 members of the CALHM family. 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, most other conserved positions (F44, Y56, I61, P64, W117) are located around 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) [44]. Therefore, mutations in this linker are expected to affect channel functions, and the conserved pattern positions identified in this analysis are hypothesized to 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 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) [45] 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.

Experimental studies to probe the functional implications of the evolutionary constraints

To assess the functional roles of the predicted conserved residues in CALHM channels, we performed several informed mutational analyses of these residues in representative human CALHMs and subjected them to patch clamp electrophysiology studies. We selected the well-studied human CALHM1 and the understudied human CALHM6 for our analyses. The expression levels of wild-type proteins and mutants are comparable, as judged by western blot and in-gel fluorescence images (Supplementary Figure 3). We established electrophysiological measurements for wild-type CALHM1 and CALHM6 following protocols described in the methods. Our results were consistent with previous findings [46] (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 CAHLM1 activity was significantly higher at physiological temperature than at room temperature, while the CALHM6 currents showed only a small increase (Figure 4). 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.

Two predicted conserved residues are in the S1–S2 linker (F40 and Y52 in CALHM1; F39 and Y51 in CALHM6). As expected, mutations in these residues 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 5). Interestingly, mutations in other conserved residues, not within the S1–S2 linker but near it (W114, L57, and P60 in CALHM1; W113 and A127 in CALHM6), also resulted in abolished or markedly reduced channel

activity (Figure 5). Among these residues, W114 (of CALHM1) and W113 (of CALHM6) have direct interaction with S1, thus changing its bulky hydrophobic side chain into a small side chain (cysteine or alanine) or positively charged residue (arginine) may significantly alter the conformational dynamics of S1 and therefore impair channel gating. Our experimental data indicate that the conserved pattern residues at the S1–S2 linker and its immediate surroundings are important for the gating mechanism across CALHMs.

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 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. This was shown not just on the well-studied CALHM1, but also on the understudied human CALHM6, thus providing valuable clues into the shared function of light and dark CALHM sequence sets.

Among the conserved residue set, the positioning of P60 on the S2 helix of CALHM1 is 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 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. Because 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. L57 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, thereby reducing the flexibility of the S1-S2 linker. A127 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. Since all mutants showed either reduced or completely abolished activity, we included a positive

control in the patch clamp experiments: I109W on CALHM1, which has been previously documented to increase channel activity and were able to reproduce this phenotype (Figure 5). We also examined the corresponding mutant in CALHM6, L108W (I109 in CALHM1 aligns with L108 in CALHM6), to further explore channel-specific differences. Interestingly, L108W decreased CALHM6 channel activity (Figure 5), possibly reflecting inherent differences between the two channels.

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 4). 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 [47]. 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 [31], KEGG [14, 15], Pharos [16], GtoP [20], and HGNC [17] 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 [33] and Phobius [34] to predict transmembrane and helical regions. The predictions were cross-referenced and confirmed against the CDD [48], Pfam [49], PrositePattern , PrositeProfiles [50] and SMART [51] databases, where a reference was available. Finally, the prediction of the pore region and pore-lining residues was supplemented using the MOLE software [35].

RAG system to verify ion specificity and gating mechanism annotations

To enhance the curation and annotation of human ion channels, we developed a Retrieval-Augmented Generation (RAG) system that leverages large language models (LLMs) to verify existing annotations. This method combines vector databases for efficient information retrieval with the natural language understanding capabilities of LLMs, allowing for systematic verification of IC annotations. The system can identify discrepancies or confirm existing information based on the latest available literature. The modular design of this approach facilitates future improvements and adaptations to other annotation verification tasks in bioinformatics.

We began the data collection by downloading relevant PubMed articles to each annotation, storing the PDF files locally, and converting them to text. After filtering out irrelevant content, the extracted text was split into manageable chunks. To create the vector database, we utilized OpenAI's text-embedding-ada-002 model to generate embeddings for each text chunk. Each processed and embedded chunk was then added to the vector database collection, with metadata including the source PubMed ID attached to each vector.

We designed the query system to construct two queries for each IC: one for verifying ion selectivity and another for confirming the gating mechanism. These questions were dynamically generated based on the manually curated annotation table. The RAG system architecture

performed similarity searches in the vector database using the query text, conducting filtered (using specific PubMed IDs) and unfiltered searches to ensure comprehensive retrieval. The top three most relevant chunks from the search results were combined to form the context for the LLM. A custom prompt template was designed to guide the LLM in providing evidence-based answers, including instructions for JSON-formatted output and emphasizing the importance of evidence-based responses.

For LLM integration, we passed queries to OpenAI's GPT-4 model. The LLM's responses were parsed and structured into a JSON format, including an answer (*Found* or *Not Found*), a confidence value between 0 and 1, and supporting evidence from the retrieved context. This structured output allows easy integration with existing annotation data, enabling systematic comparison and updating of ion channel information. Where the LLM suggested alternative ion selectivity or gating mechanism, such instances were manually verified and corrected. Finally, the ion selectivity and gating mechanism annotations not supported by LLM were labeled in our annotation table (Supplementary Table 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 and Phobius 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 [38] 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 [52] 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 2. 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 KinOrthto pipeline [40] was used for defining the orthologs and co-orthologs. In brief, KinOrtho uses 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 a homology search using both full-length and pore-containing domain sequences of human ICs against the 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 proteomes from the UniProt Proteomes Release 2022_05 that consisted of 343 Archaea, 696 Bacteria and 548 Eukaryota proteomes. After running reciprocal BLAST searches between the query and the hits, orthology detection was performed, followed by cluster analysis using OrthtoMCL and filtering of relationships to keep

only relationships within the same cluster. Finally, the results obtained from both the full-length and pore domains were retained as true orthologous relationships to remove extraneous hits.

48694 unique orthologous relationships were defined and then used to create a presence/absence matrix of orthologs, with rows representing each human IC and columns representing the different organisms. Firstly, the orthologs were grouped by IC families and defined taxonomic lineages. A percentage value representing the proportion of detected orthologs was calculated for each IC family and lineage. This 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 sequences in the family).

Only eukaryotic orthologs were retained for the orthology profiling clustering and enrichment analysis. Further, only organisms with an ortholog for at least one human IC were included. This was used to perform hierarchical clustering with the Ward method of clustering and Euclidean distance metric using the scipy package [53] in Python. To group ICs with similar presence/absence patterns, we broadly defined 9 clusters based on the dendrogram. Human ICs falling in these 9 clusters were then subjected to a functional enrichment analysis using the Gene Ontology resource GO enrichment tool. 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 [30]. This generates a hierarchical cluster where sequences within each cluster are defined by distinct conserved patterns.

CALHM construct design, transfection, and western blot analysis

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.

For the western blot experiment, the cell pellet was lysed in TBS buffer (20 mM Tris pH 8.0, 150 mM NaCl) with 10% lauryl maltose neopentyl glycol and cholesterol hemisuccinate detergent on ice. The lysate was then solubilized at 4°C for 1 hour before being clarified by centrifugation at 13,000 rpm for 5 min. The soluble portions were mixed with 4X SDS loading buffer supplemented with 5% 2-mercaptoethanol. The samples were resolved in precast gradient SDS gel (4-20%). A Chemidoc instrument was used to directly image the in-gel fluorescence signal after electrophoresis to obtain the GFP signal from the CALHM1 and CALHM6 proteins. Subsequently, the protein in the gel was transferred to the nitrocellulose membrane using the semi-dry transfer buffer (48 mM Tris base, 39 mM glycine, 20% methanol). The membrane was blocked in the TBST buffer (20 mM Tris pH 8.0, 150 mM NaCl, 0.1% Tween 80) with 4% non-fat milk for 1 hour at room temperature. Afterward the 1-step beta-actin antibody (1:2000 dilution) was incubated with the membrane for 1 hour at 4°. After incubation, the membrane was washed with TBST buffer 4 times, 15 mins each. The western blot signal was detected using ECL Pierce substrate and imaged using a Chemidoc instrument. A brightfield image was overlaid with the illuminance signal to visualize the position of protein markers in the membrane. The anti-beta-actin antibody was obtained from Proteintech (Catalogue number: HRP-60008). Each mutant was run in a western blot in triplicate from the transfected cell pellets as biological replicates, with normalization by beta-actin signal as a housekeeping gene.

For quantification of the western blot data, ImageJ was used to make measurements of the signal from the GFP fluorescence for CALHM1 and CALHM6 and to normalize the data with the chemiluminescence signal from the beta-actin blot. Measurements were taken through ImageJ of each lane of protein signal, then divided by the beta-actin signal as a loading control to normalize the data. The signal for each mutation was averaged, and the standard estimated mean was calculated before comparing it to the signal from the wild-type CALHM1 or CALHM6

protein. All mutants for each protein were compared in Microsoft Excel using a bar graph with the standard estimated mean as the error bar values.

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 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 $CaCl_2$. 5 mM $CaCl_2$ was replaced with 5 mM EGTA. The intracellular solution contains (in mM): 150 NaCl, 10 HEPES, 1 $MgCl_2$, 5 EGTA.

All data are expressed as mean±SE. 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.

Acknowledgments

We thank members of the Kannan Lab for feedback and rotation student William N Lantz for the manual evaluation of RAG LLM-based annotations. Funding for NK from NIH Common Fund (IDG) U01CA271376 is acknowledged.

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.

Competing Interests Statement

The authors declare no competing interests.

Reference:

1. Como M, Koppala BR, Hasan MN, Han VL, Arora I, Sun D. Cell Volume Regulation in Immune Cell Function, Activation and Survival. Cell Physiol Biochem. 2021;55(S1):71-88. doi: 10.33594/000000331. PubMed PMID: 33611867.

655 2. Meir A, Ginsburg S, Butkevich A, Kachalsky SG, Kaiserman I, Ahdut R, et al. Ion channels
656 in presynaptic nerve terminals and control of transmitter release. Physiol Rev. 1999;79(3):1019-
657 88. doi: 10.1152/physrev.1999.79.3.1019. PubMed PMID: 10390521.
658 3. Thorneloe KS, Nelson MT. Ion channels in smooth muscle: regulators of intracellular
659 calcium and contractility. Can J Physiol Pharmacol. 2005;83(3):215-42. doi: 10.1139/y05-016.
660 PubMed PMID: 15870837.
661 4. Rajan AS, Aguilar-Bryan L, Nelson DA, Yaney GC, Hsu WH, Kunze DL, Boyd AE, 3rd. Ion
662 channels and insulin secretion. Diabetes Care. 1990;13(3):340-63. doi:
663 10.2337/diacare.13.3.340. PubMed PMID: 1689632.
664 5. Edelman A, Saussereau E. [Cystic fibrosis and other channelopathies]. Arch Pediatr.
665 2012;19 Suppl 1:S13-6. doi: 10.1016/s0929-693x(12)71101-6. PubMed PMID: 22682482.
666 6. Kim JB. Channelopathies. Korean J Pediatr. 2014;57(1):1-18. Epub 20140131. doi:
667 10.3345/kjp.2014.57.1.1. PubMed PMID: 24578711; PubMed Central PMCID:
668 PMCPMC3935107.
669 7. Rivolta I, Binda A, Masi A, DiFrancesco JC. Cardiac and neuronal HCN channelopathies.
670 Pflugers Arch. 2020;472(7):931-51. Epub 20200518. doi: 10.1007/s00424-020-02384-3. PubMed
671 PMID: 32424620.
672 8. Bagal SK, Brown AD, Cox PJ, Omoto K, Owen RM, Pryde DC, et al. Ion Channels as
673 Therapeutic Targets: A Drug Discovery Perspective. Journal of Medicinal Chemistry.
674 2013;56(3):593-624. doi: 10.1021/jm3011433.

675 9. Braun N, Sheikh ZP, Pless SA. The current chemical biology tool box for studying ion
676 channels. J Physiol. 2020;598(20):4455-71. Epub 20200817. doi: 10.1113/jp276695. PubMed
677 PMID: 32715480.
678 10. Wickenden A, Priest B, Erdemli G. Ion Channel Drug Discovery: Challenges And Future
679 Directions. Future Medicinal Chemistry. 2012;4(5):661-79. doi: 10.4155/fmc.12.4.
680 11. Fodor AA, Aldrich RW. Statistical limits to the identification of ion channel domains by
681 sequence similarity. J Gen Physiol. 2006;127(6):755-66. doi: 10.1085/jgp.200509419. PubMed
682 PMID: 16735758; PubMed Central PMCID: PMCPMC2151544.
683 12. Neuwald AF. Rapid detection, classification and accurate alignment of up to a million or
684 more related protein sequences. Bioinformatics. 2009;25(15):1869-75. Epub 20090608. doi:
685 10.1093/bioinformatics/btp342. PubMed PMID: 19505947; PubMed Central PMCID:
686 PMCPMC2732367.
687 13. Jegla TJ, Zmasek CM, Batalov S, Nayak SK. Evolution of the human ion channel set. Comb
688 Chem High Throughput Screen. 2009;12(1):2-23. doi: 10.2174/138620709787047957. PubMed
689 PMID: 19149488.
690 14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids
691 Res. 2000;28(1):27-30. doi: 10.1093/nar/28.1.27. PubMed PMID: 10592173; PubMed Central
692 PMCID: PMCPMC102409.
693 15. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for
694 taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587-d92.
695 doi: 10.1093/nar/gkac963. PubMed PMID: 36300620; PubMed Central PMCID:
696 PMCPMC9825424.

16. Kelleher KJ, Sheils TK, Mathias SL, Yang JJ, Metzger Vincent T, Siramshetty Vishal B, et al. Pharos 2023: an integrated resource for the understudied human proteome. Nucleic Acids Research. 2022;51(D1):D1405-D16. doi: 10.1093/nar/gkac1033.

17. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, Bruford EA. Genenames.org: the HGNC resources in 2023. Nucleic Acids Res. 2023;51(D1):D1003-d9. doi: 10.1093/nar/gkac888. PubMed PMID: 36243972; PubMed Central PMCID: PMCPMC9825485.

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

19. Sheils T, Mathias SL, Siramshetty VB, Bocci G, Bologa CG, Yang JJ, et al. How to Illuminate the Druggable Genome Using Pharos. Curr Protoc Bioinformatics. 2020;69(1):e92. doi: 10.1002/cpbi.92. PubMed PMID: 31898878; PubMed Central PMCID: PMCPMC7818358.

20. Alexander SPH, Mathie A, Peters JA, Veale EL, Striessnig J, Kelly E, et al. THE CONCISE GUIDE TO PHARMACOLOGY 2019/20: Ion channels. Br J Pharmacol. 2019;176 Suppl 1(Suppl 1):S142-s228. doi: 10.1111/bph.14749. PubMed PMID: 31710715; PubMed Central PMCID: PMCPMC6844578.

21. Chen GL, Li J, Zhang J, Zeng B. To Be or Not to Be an Ion Channel: Cryo-EM Structures Have a Say. Cells. 2023;12(14). Epub 20230717. doi: 10.3390/cells12141870. PubMed PMID: 37508534; PubMed Central PMCID: PMCPMC10378246.

22. Yu FH, Yarov-Yarovoy V, Gutman GA, Catterall WA. Overview of molecular relationships in the voltage-gated ion channel superfamily. Pharmacol Rev. 2005;57(4):387-95. doi: 10.1124/pr.57.4.13. PubMed PMID: 16382097.

23. Huffer KE, Aleksandrova AA, Jara-Oseguera A, Forrest LR, Swartz KJ. Global alignment and assessment of TRP channel transmembrane domain structures to explore functional mechanisms. eLife. 2020;9:e58660. doi: 10.7554/eLife.58660.

24. Shrestha S, Byrne DP, Harris JA, Kannan N, Eyers PA. Cataloguing the dead: breathing new life into pseudokinase research. Febs j. 2020;287(19):4150-69. Epub 20200310. doi: 10.1111/febs.15246. PubMed PMID: 32053275; PubMed Central PMCID: PMCPMC7586955.

25. Picado A, Chaikuad A, Wells CI, Shrestha S, Zuercher WJ, Pickett JE, et al. A Chemical Probe for Dark Kinase STK17B Derives Its Potency and High Selectivity through a Unique P-Loop Conformation. J Med Chem. 2020;63(23):14626-46. Epub 20201120. doi: 10.1021/acs.jmedchem.0c01174. PubMed PMID: 33215924; PubMed Central PMCID: PMCPMC7816213.

26. Taujale R, Venkat A, Huang L-C, Zhou Z, Yeung W, Rasheed KM, et al. Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. eLife. 2020;9:e54532. doi: 10.7554/eLife.54532.

27. Taujale R, Zhou Z, Yeung W, Moremen KW, Li S, Kannan N. Mapping the glycosyltransferase fold landscape using interpretable deep learning. Nature Communications. 2021;12(1):5656. doi: 10.1038/s41467-021-25975-9.

28. Lara A, Simonson BT, Ryan JF, Jegla T. Genome-Scale Analysis Reveals Extensive Diversification of Voltage-Gated K+ Channels in Stem Cnidarians. Genome Biology and Evolution. 2023;15(3). doi: 10.1093/gbe/evad009.

739 29. Uribe C, Nery MF, Zavala K, Mardones GA, Riadi G, Opazo JC. Evolution of ion channels in
740 cetaceans: a natural experiment in the tree of life. Scientific Reports. 2024;14(1):17024. doi:
741 10.1038/s41598-024-66082-1.
742 30. Neuwald AF. A Bayesian sampler for optimization of protein domain hierarchies. J
743 Comput Biol. 2014;21(3):269-86. Epub 20140204. doi: 10.1089/cmb.2013.0099. PubMed PMID:
744 24494927; PubMed Central PMCID: PMCPMC3948484.
745 31. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res.
746 2018;46(5):2699. doi: 10.1093/nar/gky092. PubMed PMID: 29425356; PubMed Central PMCID:
747 PMCPMC5861450.
748 32. Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, et al.
749 The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19,
750 malaria and antibacterials. Nucleic Acids Res. 2022;50(D1):D1282-d94. doi:
751 10.1093/nar/gkab1010. PubMed PMID: 34718737; PubMed Central PMCID: PMCPMC8689838.
752 33. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein
753 topology with a hidden Markov model: application to complete genomes. J Mol Biol.
754 2001;305(3):567-80. doi: 10.1006/jmbi.2000.4315. PubMed PMID: 11152613.
755 34. Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and
756 signal peptide prediction--the Phobius web server. Nucleic Acids Res. 2007;35(Web Server
757 issue):W429-32. Epub 20070505. doi: 10.1093/nar/gkm256. PubMed PMID: 17483518; PubMed
758 Central PMCID: PMCPMC1933244.
759 35. Sehnal D, Svobodová Vařeková R, Berka K, Pravda L, Navrátilová V, Banáš P, et al. MOLE
760 2.0: advanced approach for analysis of biomacromolecular channels. J Cheminform.

761 2013;5(1):39. Epub 20130816. doi: 10.1186/1758-2946-5-39. PubMed PMID: 23953065;
762 PubMed Central PMCID: PMCPMC3765717.
763 36. Gurnett CA, Campbell KP. Transmembrane auxiliary subunits of voltage-dependent ion
764 channels. J Biol Chem. 1996;271(45):27975-8. doi: 10.1074/jbc.271.45.27975. PubMed PMID:
765 8910401.
766 37. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-
767 level protein structure with a language model. Science. 2023;379(6637):1123-30. Epub
768 20230316. doi: 10.1126/science.ade2574. PubMed PMID: 36927031.
769 38. Llinares-López F, Berthet Q, Blondel M, Teboul O, Vert J-P. Deep embedding and
770 alignment of protein sequences. Nature Methods. 2023;20(1):104-11. doi: 10.1038/s41592-
771 022-01700-2.
772 39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology:
773 tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.
774 doi: 10.1038/75556. PubMed PMID: 10802651; PubMed Central PMCID: PMCPMC3037419.
775 40. Huang LC, Taujale R, Gravel N, Venkat A, Yeung W, Byrne DP, et al. KinOrtho: a method
776 for mapping human kinase orthologs across the tree of life and illuminating understudied
777 kinases. BMC Bioinformatics. 2021;22(1):446. Epub 20210918. doi: 10.1186/s12859-021-04358-
778 3. PubMed PMID: 34537014; PubMed Central PMCID: PMCPMC8449880.
779 41. Monahan-Earley R, Dvorak AM, Aird WC. Evolutionary origins of the blood vascular
780 system and endothelium. J Thromb Haemost. 2013;11 Suppl 1(Suppl 1):46-66. doi:
781 10.1111/jth.12253. PubMed PMID: 23809110; PubMed Central PMCID: PMCPMC5378490.

42. Ma Z, Tanis JE, Taruno A, Foskett JK. Calcium homeostasis modulator (CALHM) ion channels. Pflugers Arch. 2016;468(3):395-403. Epub 20151125. doi: 10.1007/s00424-015-1757-6. PubMed PMID: 26603282; PubMed Central PMCID: PMCPMC5308871.

43. Ma Z, Taruno A, Ohmoto M, Jyotaki M, Lim JC, Miyazaki H, et al. CALHM3 Is Essential for Rapid Ion Channel-Mediated Purinergic Neurotransmission of GPCR-Mediated Tastes. Neuron. 2018;98(3):547-61.e10. Epub 20180419. doi: 10.1016/j.neuron.2018.03.043. PubMed PMID: 29681531; PubMed Central PMCID: PMCPMC5934295.

44. Choi W, Clemente N, Sun W, Du J, Lü W. The structures and gating mechanism of human calcium homeostasis modulator 2. Nature. 2019;576(7785):163-7. doi: 10.1038/s41586-019-1781-3.

45. Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: Lessons from gnomAD. Human Mutation. 2022;43(8):1012-30. doi: https://doi.org/10.1002/humu.24309.

46. Ma Z, Siebert AP, Cheung KH, Lee RJ, Johnson B, Cohen AS, et al. 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. 2012;109(28):E1963-71. Epub 20120618. doi: 10.1073/pnas.1204023109. PubMed PMID: 22711817; PubMed Central PMCID: PMCPMC3396471.

47. Syrjanen JL, Michalski K, Chou TH, Grant T, Rao S, Simorowski N, et al. Structure and assembly of calcium homeostasis modulator proteins. Nat Struct Mol Biol. 2020;27(2):150-9. Epub 20200127. doi: 10.1038/s41594-019-0369-9. PubMed PMID: 31988524; PubMed Central PMCID: PMCPMC7015811.

48. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265-d8. doi: 10.1093/nar/gkz991. PubMed PMID: 31777944; PubMed Central PMCID: PMCPMC6943070.

49. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412-d9. doi: 10.1093/nar/gkaa913. PubMed PMID: 33125078; PubMed Central PMCID: PMCPMC7779014.

50. Henschel A, Winter C, Kim WK, Schroeder M. Using structural motif descriptors for sequence-based binding site prediction. BMC Bioinformatics. 2007;8 Suppl 4(Suppl 4):S5. Epub 20070522. doi: 10.1186/1471-2105-8-s4-s5. PubMed PMID: 17570148; PubMed Central PMCID: PMCPMC1892084.

51. Letunic I, Khedkar S, Bork P. SMART: recent updates, new developments and status in 2020. Nucleic Acids Research. 2020;49(D1):D458-D60. doi: 10.1093/nar/gkaa937.

52. Großberger LMaJHaNSaL. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29):861.

53. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17(3):261-72. doi: 10.1038/s41592-019-0686-2.

Figures and Tables

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

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.

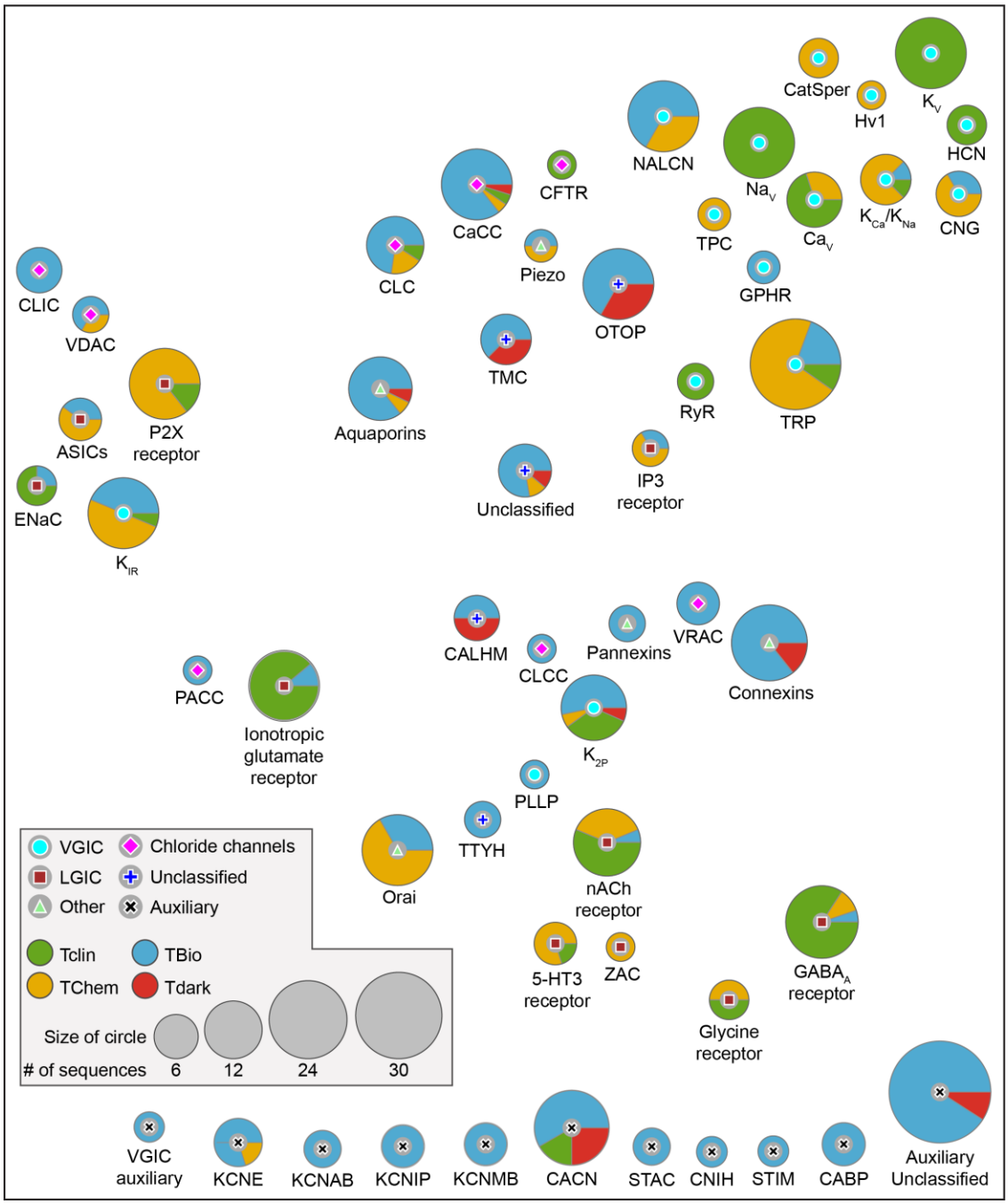

Figure 1: Distribution of ICs across different families. Each circle represents an IC family, with the symbol at the center indicating its Group. The size of the circles is proportional to the number of sequences in that family, and the colored pies indicate the proportion of their studied status as designated by IDG. Families are placed based on their distribution of UMAP embeddings generated using protein embedding-based pairwise sequence alignments (Supplementary Figure 2).

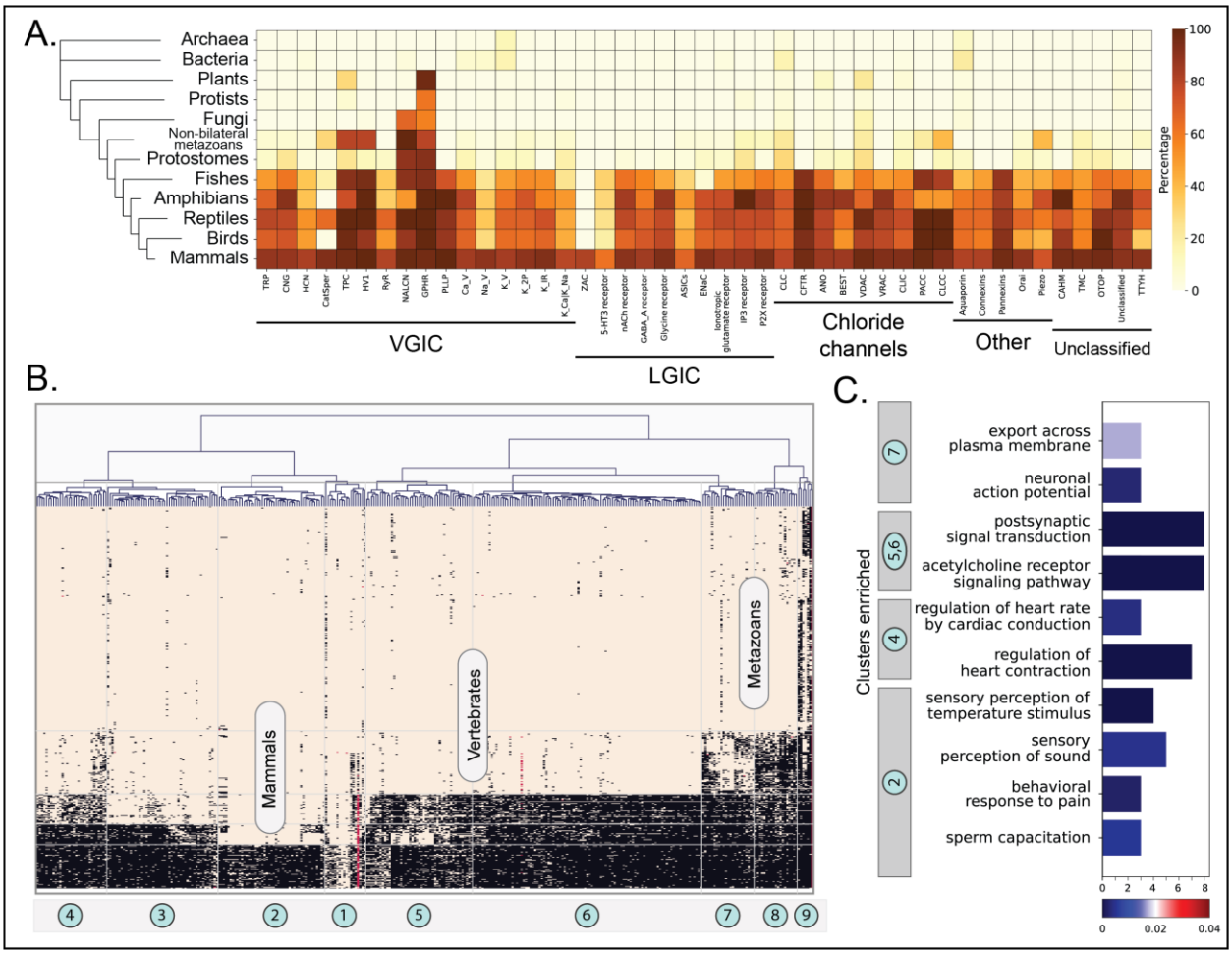

Figure 2: Orthology profiling of human ICs. (A) Heatmap showing the percent of orthologs detected for each IC family within a given taxonomic lineage. The taxonomic groups are shown in the vertical axis with a tree on the left. Darker color represents a higher percentage of orthologs detected. Percentages 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 3.

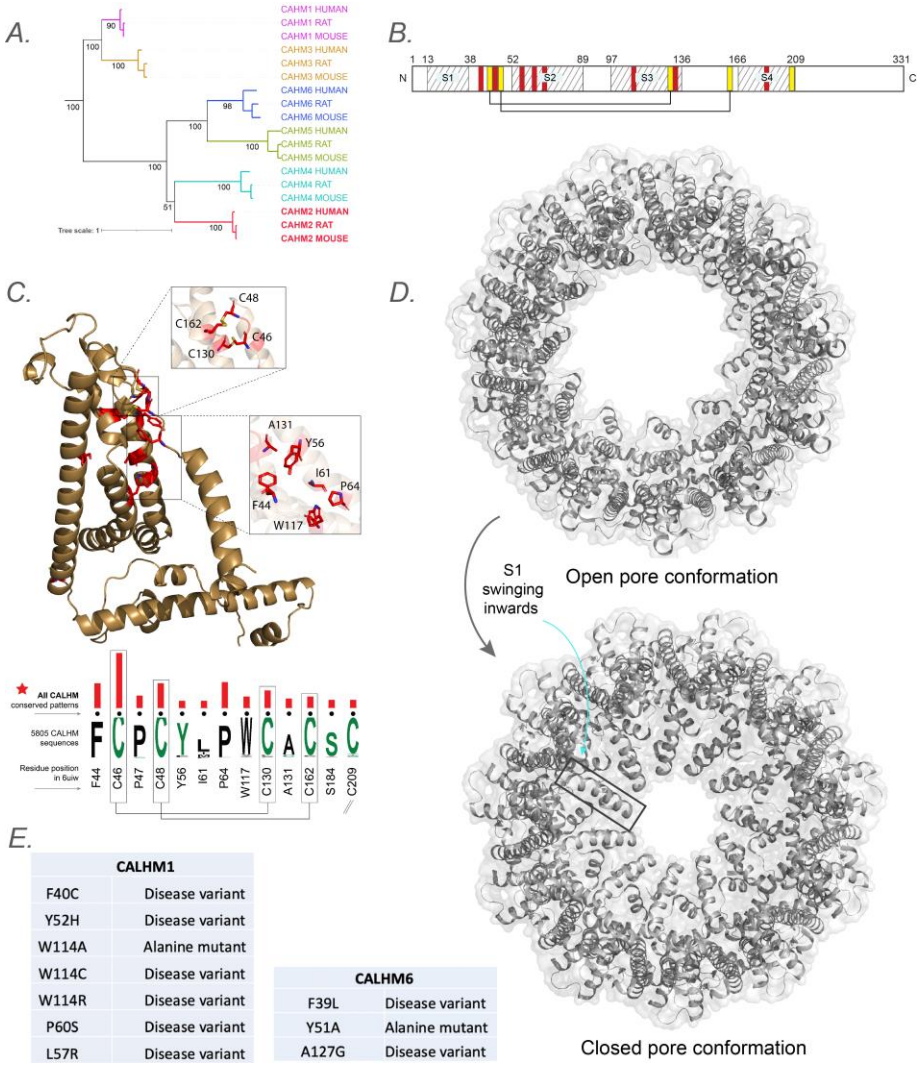

Figure 3: Evolutionary analysis of CALHM reveals conserved pattern positions. A) A phylogenetic tree depicting the evolutionary relationships across all 6 CALHM paralogs with representative sequences from humans, mouse and rat. B) Schematic representing the location of transmembrane regions and identified conserved pattern positions in a representative human CALHM2 sequence. C) The conserved pattern positions conserved across all 6 CALHM paralogs are shown as a weblogo with the red bars indicating the significance (longer bars indicate higher significance) and are mapped into a representative structure of human CALHM2. 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.

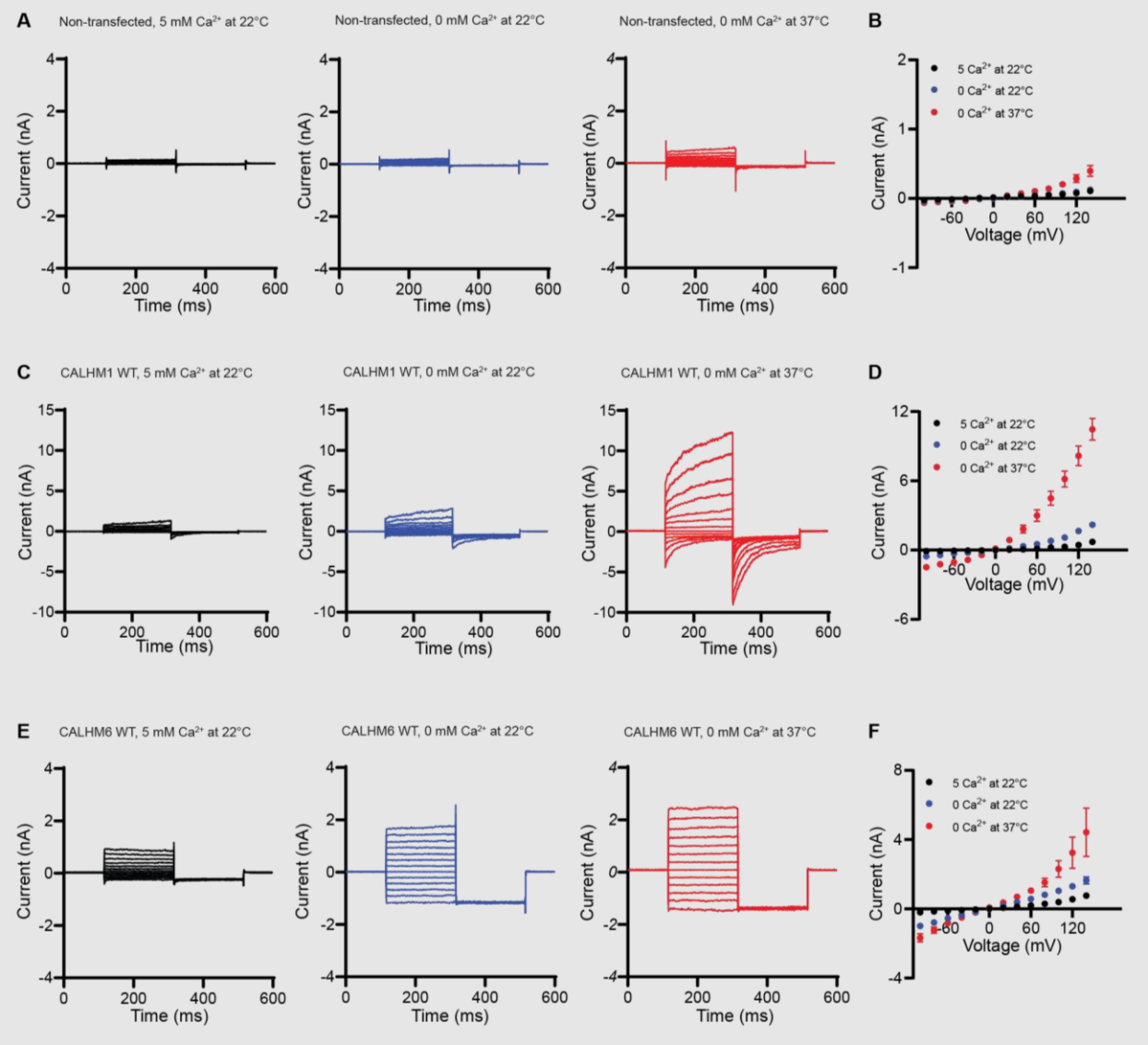

Figure 4: Electrophysiological studies of human CALHM1 and CALHM6. A,C,E. Representative current traces in the presence of 5 mM Ca2+ in the bath solution at 22°C (left), 0 mM Ca^{2+} at 22°C (middle), and 0 mM Ca^{2+} at 37°C (right) from non-transfected tsA201 cells (A), or tsA201 cells overexpressing wild-type CALHM1 (C), and wild-type CALHM6 (E), recorded in the whole-cell patch-clamp configuration. Voltage clamps were applied from −100 mV to 140 mV (200 ms each step) with a final tail pulse at −100 mV (200 ms), with a holding potential of 0 mV. B,D,F. The mean current amplitudes at the end of each pulse step of the experiments in (A,C,E) were plotted as a function of clamp voltage. Error bars represent SEM (CALHM1: n=5, CALHM6: n=3).

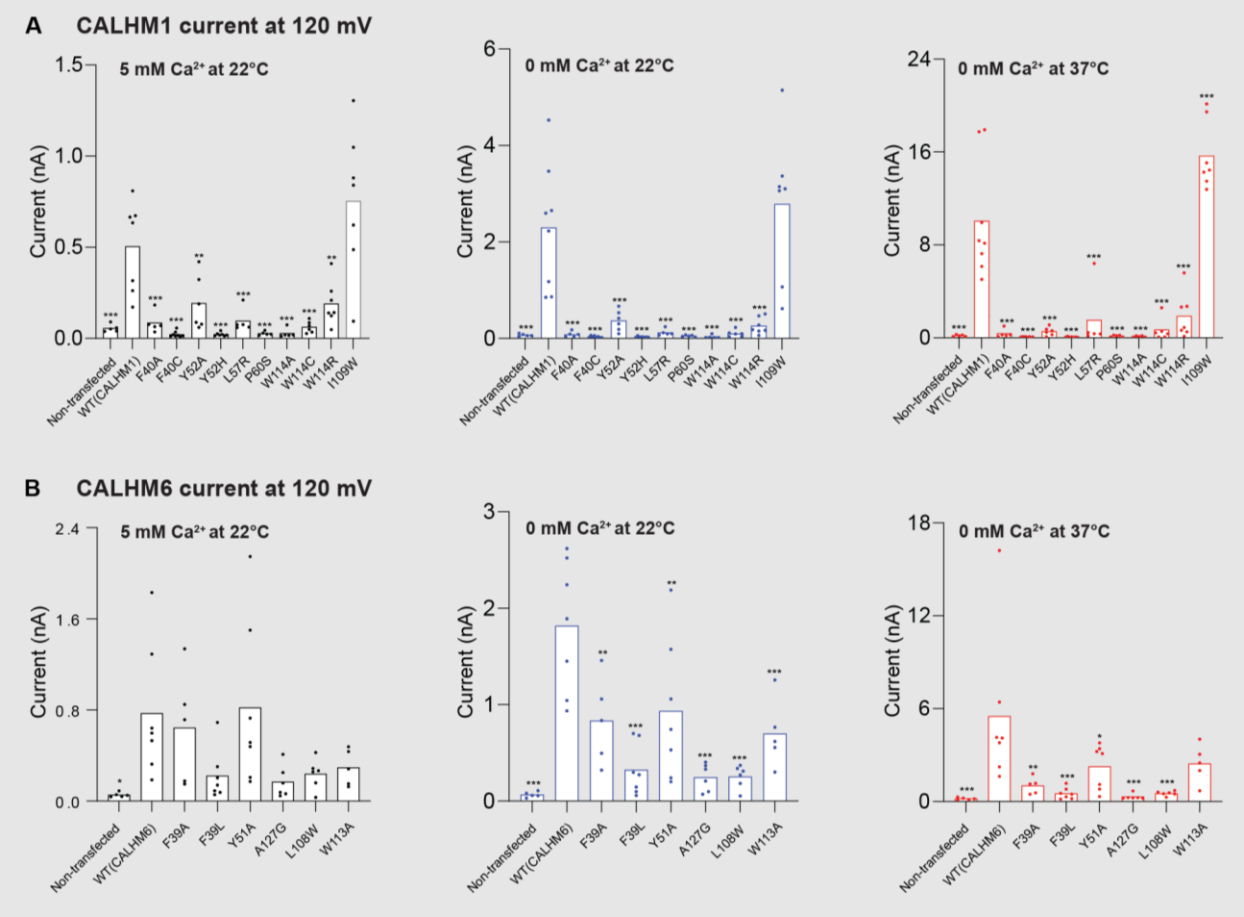

Figure 5: Current amplitudes of wild-type CALHM1, wild-type CALHM6, and mutants at conserved pattern positions in human CALHM1 and CALHM6. Currents were recorded at 120 mV in the presence of 5 mM Ca^{2+} in the bath solution at 22°C, 0 mM Ca^{2+} at 22°C, and 0 mM Ca^{2+} at 37°C. Bars represent the mean. One-way ANOVA with Bonferroni's post hoc test was used for statistical analysis (*p<0.05. **p<0.01, ***p<0.001).