Analysis of Single Nucleotide Polymorphisms in human Voltage-Gated Ion Channels

Katerina C. Nastou, Michail A. Batskinis, Zoi I. Litou, Stavros J. Hamodrakas and
Vassiliki A. Iconomidou*

Section of Cell Biology and Biophysics, Department of Biology, School of Sciences,
National and Kapodistrian University of Athens, Panepistimiopolis, Athens 15701, Greece

*Correspondence to: veconom@biol.uoa.gr

To whom correspondence should be addressed

Assist. Prof. Vassiliki A. Iconomidou

Section of Cell Biology and Biophysics, Department of Biology,
National and Kapodistrian University of Athens, Panepistimiopolis,
Athens 15701, Greece

Phone: +30 210 727 4871

Fax: +30 210 7274254

e-mail: veconom@biol.uoa.gr

<http://biophysics.biol.uoa.gr>

Abstract

Voltage-Gated Ion Channels (VGICs) are one of the largest groups of transmembrane proteins. Due to their major role in the generation and propagation of electrical signals, VGICs are considered important from a medical viewpoint and their dysfunction is often associated with a group of diseases known as “Channelopathies”. We identified disease associated mutations and polymorphisms in these proteins through mapping missense Single Nucleotide Polymorphisms (SNPs) from the UniProt and ClinVar databases on their amino acid sequence, taking into consideration their special topological and functional characteristics. Statistical analysis revealed that disease associated SNPs are mostly found in the Voltage Sensor Domain – and especially at its fourth transmembrane segment (S4) – and in the Pore Loop. Both these regions are extremely important for the activation and ion conductivity of VGICs. Moreover, amongst the most frequently observed mutations are those of arginine to glutamine, to histidine or to cysteine, which can probably be attributed to the extremely important role of arginine residues in the regulation of membrane potential in these proteins. We suggest that topological information in combination with genetic variation data can contribute towards a better evaluation of the effect of currently unclassified mutations in VGICs. It is hoped that potential associations with certain disease phenotypes will be revealed in the future, with the use of similar approaches.

Keywords

Voltage-Gated Ion Channel (VGIC), Single Nucleotide Polymorphism (SNP), Channelopathies, Pathogenic mutations, Polymorphisms

Abbreviations

VGICs: Voltage-Gated Ion Channels, LGICs: Ligand-Gated Ion Channels, SNPs: Single Nucleotide Polymorphisms, Ca_VS: Voltage-Gated Calcium Channels, Na_VS: Voltage-Gated Sodium Channels, K_VS: Voltage-Gated Potassium Channels, PL: Pore Loop, VSD: Voltage Sensor Domain, ILs: Intracellular Loops, ELs: Extracellular Loops, ANOVA: Analysis of Variance, GWAS: Genome-Wide Association Study

1. Introduction

Ion channels are transmembrane proteins that enable ion flow via lipid membranes [1]. These proteins govern the electrical properties of membranes of excitable cells and thus are crucial for many physiological functions, such as neuronal signaling, muscle contraction and hormone secretion [2]. Ion channels are mainly classified into Voltage-Gated Ion Channels (VGICs) and Ligand-Gated Ion Channels (LGICs) depending on their activation mechanism [3, 4]. VGICs form a superfamily characterized by the ability to respond rapidly to alterations in membrane potential (as their nomenclature “voltage-gated” suggests), resulting in the selective conductance of ions [5]. The most prominent, voltage-sensitive members of this superfamily can be categorized physiologically into 3 families based on the types of ions that they transport; Voltage-Gated Calcium Channels (Ca_{VS}), Voltage-Gated Sodium Channels (Na_{VS}) and Voltage-Gated Potassium Channels (K_{VS}) [6].

VGICs that belong to these families share some common characteristics. Specifically, they have α -subunits consisting of four domains (DI-DIV). Na_{VS} and Ca_{VS} consist of one polypeptide chain with four domains [7-9], while K_{VS} consist of four polypeptide chains, which assemble to form a functional tetrameric channel [10], which explains the extreme diversity of K_{VS} in comparison to the other two families [11, 12]. Each one of these domains shares a common “topological profile” with six transmembrane segments (S1-S6), four of which form the Voltage Sensor Domain (VSD) and a Pore Loop (PL) [7-9]. The PL is located between transmembrane segments S5 and S6 and includes the selectivity filter of the channel [13-16]. The VSD is a modular four transmembrane helix bundle (S1-S4) [17], whose conformational changes upon alterations in membrane potential are responsible for the opening of the channel pore [18-20]. Apart from these distinguished members of the superfamily of VGICs, other families have been described more recently. However, members of these families either do not follow the “topological profile” described above [21-25] or are non-selective, voltage-insensitive and their activation is affected by neurotransmitters and biochemical mediators [24, 26-29].

Each family of VGICs serves a different functional role in the cell. K_{VS} mostly control neuronal excitability [30] and participate in solute transport [31], while Na_{VS} contribute to the generation of neuronal action potential [32, 33]. Ca_{VS} , on the other hand, are involved in a wide range of physiological processes ranging from muscle contraction and neurotransmitter release [34] to gene expression [35]. Due to their versatility, defects in their expression, regulation or function has severe consequences that lead to the emergence of a plethora of

diseases [36]. This explains the fact that ion channels currently represent the second largest target group of proteins for drugs in the market, coming immediately after G-protein-coupled receptors [37, 38]. Moreover, advances on compound screening techniques have also designated ion channels as promising targets for the future development of drugs [39].

Progress in the field of human molecular genetics allowed the association between changes in genes coding for VGICs and a group of heterogeneous diseases known as ‘Channelopathies’ [40]. Channelopathies include a broad range of clinical conditions – from epilepsy [41] to cardiac dysfunctions [42] – which arise from gain-of-function or loss-of-function mutations in ion channels [43]. The majority of these disorders are genetically diverse. Same phenotypes arise from different mutations and the same mutation can present with different effects in individuals with distinctive genetic backgrounds [44]. Genetic testing, based on common genetic markers for some of the most common Channelopathies, is commercially available [45]. Some of the most important markers identified for these diseases include Single Nucleotide Polymorphisms (SNPs) [46]. SNPs are inherited single nucleotide substitutions among individuals of a species and are defined as such, when they occur at a rate greater than 1% in a population [47]. Normal substitutions are responsible for genetic variation among the human population. On the contrary, when SNPs result in mutations on regulatory or protein-coding regions of a gene, they can lead to adverse effects and even diseases, such as Channelopathies [46].

The aim of this study was to examine the distribution of missense SNPs and the range of amino acid substitutions in relation to the topology of human VGICs. Our goal was to elucidate the extent to which missense SNPs can be associated with predisposition towards the emergence of a disease. Moreover, we wanted to examine whether the special topological features of VGICs, in combination with genetic data, can aid towards the better characterization of the pathogenic potential of SNPs that are currently unclassified.

2. Materials and Methods

2.1. Datasets

2.1.1. VGICs dataset

The dataset used in this analysis contains all VGICs that are activated by voltage (voltage-dependent). Their α -subunit forms a functional tetramer, where each monomer comprises of six transmembrane segments and a pore forming loop, responsible for ion transport. From all families of VGICs included in IUPHAR [48], only three meet the aforementioned criteria and are analyzed in this work, namely Voltage-gated Potassium Channels (K_{VS}) [49], Voltage-gated Calcium Channels (Ca_{VS}) [50, 51] and Voltage-gated Sodium Channels (Na_{VS}) [32, 52]. Specifically, all channels used in this study, contain four domains with six transmembrane segments (S1-S6) – where four of them (S1-S4) form the Voltage Sensor Domain (VSD) – and a Pore-loop (PL). Other families characterized as VGICs in IUPHAR are either voltage-independent channels or have a totally different topology than the one described above (see Introduction). The dataset of Ion Channels from IUPHAR and the VGICs entries used in this analysis are shown in Table S1.

2.1.2. Missense SNPs dataset

Human genetic variation data (“Polymorphisms” and “Pathogenic” mutations) were gathered from UniProt database [53] (release date: 10-10-2018) and from ClinVar database [54] (release date: 22-10-2018). A manual mapping between RefSeq Nucleotide Codes from ClinVar to UniProt ACs was conducted (Table S2), in order to gather data from ClinVar only for the isoform that represents the canonical sequence of the VGIC in UniProt. This step was necessary, since position information in UniProt applies only to the “canonical” isoform and incorrect mapping between these two databases leads to numerous mistakes.

The final dataset used in this study emerges from the union of data collected from UniProt and ClinVar. For a more accurate analysis, a common reference code for clinical characterization was used, considering the different terminologies these databases use. Thus, missense SNPs characterized as ‘Disease’ (UniProt), ‘Pathogenic’ or ‘Likely Pathogenic’ (ClinVar) will be referred as “Pathogenic” in the final dataset, while those referred to as ‘Polymorphism’ (UniProt), ‘Benign’, ‘Likely benign’ or ‘Risk factor’ (ClinVar) will be referred simply as “Polymorphism”. Finally, SNPs characterized as ‘Unclassified’ (UniProt), ‘not provided’, ‘not reported for simple variant’, ‘Conflicting interpretations of pathogenicity’, ‘no interpretations for the single variant’ or ‘Uncertain significance’ (ClinVar) will be added to the “Unclassified” dataset. Duplicate entries were merged in one

entry in the final dataset, since there were cases where the two datasets overlapped. In cases with conflicting clinical significance for the same entry in the two databases, the clinical significance from ClinVar was adopted (Table S3).

2.2. VGICs' missense SNPs

2.2.1. Mapping of SNPs on biologically significant regions and categorization based on biophysical properties

The determination of three dimensional structures for VGICs, and especially human VGICs, has proven a rather difficult task for the scientific community in the past two decades [55]. The case is similar for the majority of transmembrane proteins [56, 57] and several computational tools have been developed through the years for the prediction of transmembrane segments [58], to provide researchers with a view of the topology of these proteins. Protein databases, like UniProt, have incorporated these data in their membrane protein entries. In this work, data regarding the transmembrane, extracellular and intracellular regions for all VGICs were extracted from UniProt and were transformed, in order to follow the known “topological profile” of VGICs. Specifically, regions were grouped as follows: N-terminal, C-terminal, Intracellular Loops (ILs), Extracellular Loops (ELs), Pore-Loop (PL), Voltage Sensor Domain (VSD) and S5-S6 Transmembrane Regions (S5-S6 TM).

Using this “topological profile”, all “Polymorphisms”, “Pathogenic” and “Unclassified” SNPs were mapped on the sequences of VGICs. Both the exact amino acid substitutions and substitutions grouped in terms of biophysical properties were counted, for the conduction of additional statistical analyses. Asparagine (N), Glutamine (Q), Serine (S), Threonine (T), Tyrosine (Y), Cysteine (C), Methionine (M), Tryptophan (W) and Histidine (H) were considered polar residues, Alanine (A), Glycine (G), Isoleucine (I), Leucine (L), Phenylalanine (F), Proline (P) and Valine (V) were designated as non-polar and Arginine (R), Aspartate (D), Glutamate (E) and Lysine (K) as charged residues [59].

2.2.2. Statistical analysis

Due to the complexity of our data, a simple statistical analysis (e.g. t-test) was not sufficient to examine the parameters we wanted. For this reason, more complex statistical methods were used. Data were subjected to Analysis of Variance (ANOVA [60]) and logistic regression in order to gain an estimate of the correlation between pathological SNPs and their appearance within certain topological regions. More specifically, ANOVA was utilized to

get an initial estimate of the variance of appearance of “Polymorphisms” and “Pathogenic” SNPs on transmembrane and non-transmembrane segments. However, ANOVA could not provide sufficient results when the special topological features of VGICs were taken into consideration, and thus, due to the increase of dimensions in our feature set, logistic regression [61] was applied. Regression analysis can aid towards the understanding of how the typical value of a dependent variable changes, when one of the independent variables is adjusted and others are held fixed. During the logistic regression analysis, our focus was on four different topological regions of VGICs. The VSD and PL were examined as the most biologically significant regions of VGICs and the extracellular and intracellular loops due to their role in the interactions between different VGIC domains. Moreover, logistic regression analysis was performed to test the association between the biophysical attributes of the amino acid changes and the pathogenicity status (“Polymorphism” or “Pathogenic”) of missense SNPs.

In order to estimate the statistical significance of all amino acid changes of “Polymorphisms” and “Pathogenic” SNPs in different topological sites, data were subjected to random sampling with replacement (Bootstrap [62]). Bootstrap is a statistical method that allows us to obtain a theoretical estimate regarding the distribution and the standard deviation of data for a given experimentally obtained dataset (amino acid changes in our case). The application of this method on biological data allows an increase in available data in a random and unbiased manner [63]. This analysis was performed on four different starting populations, specifically the samples of a) all pathological SNPs, b) all polymorphisms, c) all pathological SNPs on VSD and d) all pathological SNPs on PL. The method was applied 1000 times for each sample and the average and standard deviation values were determined for all substitutions. The significance of our results was assessed with the calculation of Confidence Intervals (CI) for each SNP in all samples.

Figure 1 shows a workflow of the entire process of data collection and analysis, presented in the previous sections.

Figure 1. The workflow of data collection and analysis. A dataset of human VGICs was collected from IUPHAR and data regarding missense SNPs of VGICs were gathered from UniProt and ClinVar. These data were manipulated with scripts written in Perl, and all SNPs were mapped on human VGICs in relation to the topological information collected from UniProt. When available, data was isolated from ClinVar regarding the association of specific SNPs with Channelopathies. Statistical analysis was carried out to draw conclusions regarding the significance of our results.

3. RESULTS AND DISCUSSION

3.1. VGICs' missense SNPs rates

The dataset of 5311 unique missense SNPs located on 51 out of the 59 known human VGICs was gathered from both UniProt and ClinVar. From those SNPs, 2035 are located on K_{VS}, 741 on Ca_{VS} and 2535 on Na_{VS} (Table S3). Concerning their status, 1765 “Pathogenic SNPs” are found on 33 VGICs, 457 “Polymorphisms” on 42 VGICs and the majority of SNPs are “Unclassified” and account for 3089 SNPs on 41 VGICs (Table S4). The rate of “Polymorphisms” per residue is 5.5×10^{-3} , while the rate of “Pathogenic” SNPs per residue is 3.8×10^{-2} . This suggests an enrichment of “Pathogenic” SNPs in VGICs and a reduced tolerance to mutations for this protein group, since the appearance of a random SNP in a VGIC usually leads to disease. All SNPs were mapped on VGICs' sequences and the total numbers per topological region are shown in Table 1. Results in more detail are presented in Table S5. The raw counts of SNPs indicate the VSD, Pore Loop, Intracellular Loops and the C-terminal as the main regions where SNPs appear.

Table 1. Total count of SNPs per topological region. The number of SNPs found in each topological region of VGICs.

<i>Topological region</i>	<i>Pathogenic SNPs</i>	<i>Polymorphisms</i>	<i>Unclassified</i>	<i>Total</i>
<i>N-terminal</i>	101	48	384	533
<i>VSD</i>	378	31	412	821
<i>S5</i>	139	5	104	248
<i>Pore Loop</i>	373	49	397	819
<i>S6</i>	178	12	131	321
<i>Intracellular Loops</i>	305	142	677	1124
<i>Extracellular Loops</i>	49	24	125	198
<i>C-terminal</i>	242	146	859	1247
<i>Total</i>	1765	457	3089	5311

However, since these regions differ vastly in their length, all data had to be subjected to normalization, based on the length of each topological region. The relative frequencies for each mutation were calculated and are shown in Figure 2. At this point, “Unclassified” SNPs were excluded from any future analysis, since a pathogenicity status could not be assigned for these SNPs. Highest frequencies for both “Polymorphisms” and “Pathogenic” mutations are found in the VSD and ILs, both regions of great biological significance for VGICs. These results were further statistically examined.

Figure 2. Normalized frequencies of SNPs per topological domain. Green bars represent the frequencies of “pathogenic” SNPs, blue bars show the frequencies of “polymorphisms” and yellow bars the total frequency of “polymorphisms” and “pathogenic” mutations. The VSD and ILs are designated as the most prominent regions for the appearance of SNPs.

3.2. Comparison of the distribution of pathogenic mutations’ frequency within topological regions of biological significance

3.2.1. Examining the relation between pathogenicity and transmembrane regions

At first, ANOVA analysis was performed in order to examine the connection between an SNP’s pathogenicity status and its appearance in transmembrane regions. Our null hypothesis was that these parameters are not associated. Results from this analysis showed that the means of observations for “Polymorphisms” and “Pathogenic” mutations differ significantly (Table 2, $P\text{-value} < 0.05$). However, no difference between transmembrane and non-transmembrane regions was observed and also pathogenicity and transmembrane regions were not found to be related (Table 2).

Table 2. Results from two-factor with replication ANOVA. The difference in the distribution of “Polymorphisms” and “Pathogenic” mutations in transmembrane and non-transmembrane regions is examined. The *Sample* difference suggests that the means of observations between “Polymorphisms” and “Pathogenic” mutations varies ($F > F_{crit}$, $P\text{-value} < 0.05$). The *Columns* difference examines the difference of means between transmembrane and non-transmembrane regions. In this case the null hypothesis cannot be rejected ($F < F_{crit}$, $P\text{-value} > 0.05$), meaning that no difference between SNPs in these two regions could be observed. Finally, the

Interaction difference of means examines the interaction between these two factors (pathogenicity status and appearance in transmembrane or non-transmembrane region). The null hypothesis that there is no interaction between these two factors could neither be rejected ($F < F_{crit}$, $P\text{-value} > 0.05$). Statistically significant results are shown in **bold**. *SS*: Sum of Squares due to the source, *df*: degrees of freedom in the source, *MS*: Mean of Squares due to the source, *F*: F-test statistic, *P-value*: P-value associated with the F-test, *F crit*: F critical (from the F distribution).

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	79853.04	1	79853.04	13.53	0.0003	3.88
<i>Columns</i>	1.06	1	1.06	0.00018	0.98	3.88
<i>Interaction</i>	210.4	1	210.4	0.035	0.85	3.88
<i>Within</i>	1368690	232	5899.5			
<i>Total</i>	1448755	235				

3.2.2. Examining the relation between pathogenicity and regions of biological significance

Even though ANOVA can provide a first estimate for the distribution of “Polymorphisms” and “Pathogenic” SNPs on VGICs (Table S6), the method is not sufficient to draw detailed conclusions, regarding the connection between biologically significant regions and pathogenicity. For this reason, logistic regression analysis was carried out, to examine the relationship between the “pathogenicity status” of SNPs and their appearance in specific topological regions. At this point, it should be noted that normalization is not a part of the procedure, since the effect of each independent variable has to be assessed and a normalization process would manipulate the data in such a way that would create bias to the analysis.

Results from the logistic regression analysis of the relationship of pathogenicity status and topological regions are shown in Table 3. *Extracellular and Intracellular Loops* appear as not statistically significant based on our model. As for the extremely statistically significant variables, the *VSD* has the lowest *P-value*, followed by *Other* regions and the *Pore Loop*. This suggests a strong association of these topological regions with the probability of an SNP to be linked with a specific pathogenicity status. The positive coefficient of the estimate in the *VSD* and the *Pore Loop* suggests that all other variables being equal, an SNP in these regions is more likely to be pathogenic. On the contrary, the negative coefficient for *Other* regions indicates that an SNP in these regions is probably not “Pathogenic”.

Table 3. Results from logistic regression for pathogenicity status in association with topological regions of biological significance. The distribution of “Polymorphisms” and “Pathogenic” mutations in different topological regions of biological significance is examined. Three regions are statistically significant based on our model. SNPs that are present in the *VSD* and *PL* are more likely to be “Pathogenic” (positive estimate

coefficient), while SNPs in *Other* regions of VGICs are probably “Polymorphisms”. Statistically significant results are shown in **bold**.

<i>Condition-Region</i>	<i>Estimate</i>	<i>P-value</i>	<i>Significance level</i>
VSD	0.60	0.018	95%-99%
PL	0.30	0.047	95%-99%
Extracellular Loops (ELs)	-0.55	0.082	90%-95%
Intracellular Loops (ILs)	-0.12	0.082	90%-95%
Other	-0.16	0.032	95%-99%

These results indicate that the VSD and PL are important for VGICs’ functionality. Specific networks of residues that interact with ions in these proteins are extremely important for ion flux regulation and thus, disturbances in these regions often lead to disease [64-69]. More specifically, it has been shown that mutations in the VSD region and especially in the positively charged S4 helix, can cause the pores of these channels to be permanently permeant to ions, at either depolarized or hyperpolarized voltages, where they are normally closed [70].

3.2.3. Examining the interaction between pathogenicity status and the biophysical attributes of mutations

In a different approach, SNPs were grouped based on the biophysical attributes of each amino acid change. The relative frequencies of the entire SNP dataset, as well as, those of “Pathogenic” mutations only, were counted (Table 4). These frequencies demonstrate that the majority of mutations, for all SNPs and for those that are disease associated specifically, are caused from non-polar to non-polar, non-polar to polar and charged to polar changes.

Table 4. Relative frequencies of Total and “Pathogenic” mutations in VGICs, grouped based on the biophysical attribute of the residue change. The majority of mutations are from non-polar to non-polar residues, followed by non-polar to polar and charged to polar. The residue changes with the highest frequency for each column are shown in **bold**, those with the second highest with underline, while those with the third highest are shown in *italics*.

% residue changes	<i>Total SNPs (%)</i>				<i>Pathogenic SNPs(%)</i>			
	<i>VGICs</i>	<i>Na_Vs</i>	<i>Ca_Vs</i>	<i>K_Vs</i>	<i>VGICs</i>	<i>Na_Vs</i>	<i>Ca_Vs</i>	<i>K_Vs</i>
polar to polar	11.25	11.34	10.10	11.55	11.22	11.63	8.47	11.08
polar to non-polar	10.17	11.25	7.32	9.56	10.82	11.43	6.78	10.61
polar to charged	7.16	9.05	4.88	5.05	7.71	9.54	6.85	5.20
non-polar to polar	<u>16.07</u>	<u>14.64</u>	20.91	<u>16.47</u>	<i>14.94</i>	<u>14.90</u>	16.42	<i>15.12</i>
non-polar to non-polar	21.29	22.25	<i>16.72</i>	21.51	21.50	22.01	<u>17.46</u>	22.01
non-polar to charged	8.78	7.87	9.41	9.96	9.92	8.68	<i>16.75</i>	10.87

charged to polar	<i>14.81</i>	<i>13.20</i>	<u>19.86</u>	<i>15.41</i>	<u>15.03</u>	<i>13.57</i>	20.47	<u>16.71</u>
charged to non-polar	5.45	5.33	4.18	6.11	6.01	5.90	3.59	6.78
charged to charged	5.04	5.08	6.62	4.38	4.86	5.00	6.35	4.51

In addition, logistic regression analysis was conducted, to model the relationship of the pathogenicity status and biophysical attributes of the mutations. Results from this analysis are presented in Table 5. In this model, only one case exhibited statistical significance: changes from non-polar to charged residues. The estimate has a positive coefficient (Table 5), suggesting that these changes have a higher probability to lead to the appearance of “Pathogenic” SNPs on VGICs. Despite being on the top three cases with the highest frequencies in our sample, non-polar to non-polar, non-polar to polar and charged to polar mutations were not designated as statistically significant based on our model. This means that even though these mutations appear often, they do not present a higher potential to lead to pathogenicity.

Table 5. Results from logistic regression for pathogenicity status in association with biophysical attributes of amino acid changes. The relationship between the pathogenicity status and the biophysical attributes of SNPs is analyzed. Changes from non-polar to charged residues are statistically significant and have a higher probability to lead to disease (positive estimate coefficient). Statistically significant results are shown in **bold**.

<i>Condition</i>	<i>Estimate</i>	<i>p-value</i>	<i>Significance level</i>
polar to polar	-0.47	0.16	<90%
polar to non-polar	0.10	0.71	<90%
polar to charged	0.34	0.41	<90%
non-polar to polar	-0.15	0.39	<90%
non-polar to non-polar	0.00	0.99	<90%
non-polar to charged	0.86	0.01	95%-99%
charged to polar	0.15	0.43	<90%
charged to non-polar	-0.07	0.83	<90%
charged to charged	-0.59	0.16	<90%

3.2.4. Analyzing the distribution of amino acid substitutions for “Polymorphisms” and “Pathogenic” mutations in VGICs

Apart from the analysis of relative frequencies of SNPs based on their biophysical nature, we conducted a more detailed analysis, for each amino acid change. Due to the large size of the tables (20x20) we constructed contour charts to visualize the frequencies of all changes in a concise manner (Figure 3 and Table S7). The contour chart in Figure 3 shows that “Pathogenic” mutations with the higher frequency are those from Glycine (G) to Arginine (R), Leucine (L) to Proline (P), Arginine (R) to Cysteine (C), Arginine (R) to Histidine (H),

Arginine (R) to Glutamine (Q), Alanine (A) to Valine (V) and Glutamate (E) to Lysine (K) (Table S7B). From those, E to K, R to H and R to Q mutations are more frequent in the VSD (Table S7C), while R to C and A to V appear mainly in the Pore Loop (Table S7D). All aforementioned mutations can be easily explained based on the standard genetic code, since changes in at least two codons for each substitution are able to induce all these changes.

Figure 3. Relative percentage frequencies of “Pathogenic” mutations in the entire length of VGICs. “Pathogenic” mutations with the higher frequency are those in Yellow and Dark Green Peaks in the Contour Chart and specifically G to R, L to P, R to C, R to H, R to Q, A to V and E to K.

In the case of “Polymorphisms”, three mutations emerge as the most frequent, A to T, P to L and R to H (Table S7A).

In an effort to examine the statistical significance of the aforementioned changes, we conducted a random sampling with replacement “Bootstrap” analysis. All observed “Polymorphisms” and “Pathogenic” mutations were counted and were compared to a dataset of mutations produced through 1000 random samplings. This allowed us to increase the

available SNP data in order to assess the statistical significance of each change. As shown in Table S8, the most prominent “Pathogenic” mutations, that were identified previously, are all statistically significant.

3.3. Focusing on statistically important mutations

Based on our analysis the most prominent disease associated mutations are those of G to R, L to P, A to V, E to K and R to C, H or Q. From those R mutations located on the VSD appear extremely significant.

Previous studies have shown that G to R and L to P mutations in human transmembrane proteins, lead to increased disease susceptibility [71]. VGICs are not an exception, since all such mutations, recorded in transmembrane regions, result to pathogenicity (Table S9A). Most ‘Pathogenic’ G to R mutations are located on the Pore Loops of VGICs. Almost half of them are found on K_{Vs} (Table S9D). This can perhaps be attributed to the fact that K⁺ selectivity filters contain a signature sequence (TVGYG) [72] with two glycine residues within each loop, and thus a disruption in this sequence can lead to adverse effects for the protein’s functionality. Moreover, when present in transmembrane regions, G to R mutations provide an additional charge in these segments, which leads to the disturbance of the hydrophobic core of these proteins. Even though L and P are both non-polar residues, these mutations are also harmful when located in transmembrane α -helices, since prolines’ side chains provide extra charge due to the exposure of hydrogen bridge acceptors.

In a similar manner, mutations from E to K alter the charge of the mutated residue from negative to positive. This leads to the disruption of the interactions with neighboring residues, explaining the link of such mutations to pathogenicity. Moreover, selectivity filters in the PLs of both Na_{Vs} (DEKA) [5] and Ca_{Vs} (EEEE/EEDD) [14, 73] contain E residues and thus such mutations can alter the activation or inactivation mechanism of these channels and the subsequent ion flow. Additionally, Na_{Vs} [74] alanine mutations and K_{Vs} [75] A to V mutations in residues adjacent to pore loops have been shown to alter the properties of their selectivity filters, often leading to atypical gating, suggesting an important role of these non-charged residues, as well.

Mutations in VSD have repeatedly been recorded to cause many Channelopathies. A detailed catalogue and analysis of diseases caused by SNPs on VGICs is presented in Section

3.4. Biophysical studies have been carried out to reveal the mechanism underlying these clinical conditions [64, 65, 76, 77]. It has been shown that mutations in the VSD, and specifically in the S4 segment, lead to the emergence of different pathological permeation pathways, known as gating pores [64, 65, 69]. Gating pores have been recorded when the coherence of the hydrophobic septum of VGICs is corrupted in general. These structures activate rapidly, but do not have a mechanism dedicated to their inactivation [70]. This disturbance in ion flow has been associated with many pure electrical disorders, such as Brugada and long QT syndrome [78]. More recently their association with cardiac disorders has also been reported [79]. Mutations in R residues appear to be the key behind the mechanism underlying these disorders [76, 80]. Specifically, S4 arginine neutralizing mutations, such as R to C [81-83], R to H [64] and R to Q [84, 85], have shown different impacts on gating pore currents. Since R residues control the movement of S4 segments, which in turn serve as the cornerstone of voltage sensing, this effect explains the connection of these mutations with diseases such as periodic paralyses [85, 86]. Consequently, gating pores caused by mutations in the VSD could constitute the pathological mechanism of many understudied Channelopathies. Interestingly, these 3 mutations have been designated as the most important arginine mutations in our dataset as well. When present on the VSD they always lead to pathogenicity, stressing the importance of almost another fifty (50) R to C, H or Q SNPs of uncertain significance in the same topological region (Table S9A). The data regarding these mutations is growing and in combination with topological information, potential pathogenic effects can be associated with known or newly identified unclassified SNPs.

3.4. Association with Channelopathies

SNPs associated with Channelopathies are found on 33 out of the 59 human VGICs. Some SNPs have more than one association with different diseases. In total 1964 associations were recorded for 1490 unique SNPs and a total of 132 different diseases were identified, excluding general disease categories (Table S10). These diseases can be classified based on the affected human organ system [43]. The majority affects the Nervous System and is mostly associated with Navs (Table 6). Over 65% of SNPs are associated with two major categories of diseases; various types of epilepsies (~35%) and long QT syndromes (~30%) (Table S10A). Furthermore, SNPs of VGICs belonging to different families present a

phenotypic overlap, since they lead to the same disorder. For these reasons we decided to look further into each VGIC family and its properties.

Na_v s physiological function is the initiation and propagation of action potentials in nerves and muscle fibers [87]. The importance of their inactivation, due to specific mutations, in the development of Channelopathies and particularly neurological disorders, has been thoroughly studied and well established [88, 89], which explains the plethora of data recorded for this type of SNPs in our work. However, it should also be mentioned that this result could be attributed to publication bias, since disorders associated with Na_v s are those studied mostly by the scientific community [90]. On the other hand, K_v s are implicated in the establishment of membrane potential and the repolarization after an action potential [87]. A large number of SNPs on K_v s associated with Heart Muscle Diseases has been recorded via our analysis, in accordance with the effects of the disturbance of their physiological function. Finally, Ca_v s are responsible for the generation of action potential in the heart and smooth muscles, muscle contraction, neurotransmitter release and intracellular signaling [91]. The biological processes in which they participate, explain why Hormonal Diseases were linked only with a few SNPs on Ca_v s (Table 6). The small number of representatives for this family though, does not undermine their importance, but rather stresses the need for their future extensive study.

Table 6. Total count of references to Channelopathies. References from different SNPs are grouped based on the type of disease they are associated with and the VGICs in which they are found. A total of 1490 ‘Pathogenic’ SNPs have 1964 references for 132 Channelopathies. Four major organ systems are affected in Channelopathies; the neuronal, the muscular, the hormonal and the heart muscle system. In the category ‘Other’ SNPs that reference general disease terms (e.g. CACNA-related conditions) and not specific diseases are included. The majority of SNPs is located on Na_v s and is associated either with Neuronal or Heart Muscle Diseases. ND: Neuronal Disease, MD: Muscular Disease, HMD: Heart Muscle Disease, HD: Hormonal Disease.

<i>VGICs (Channels with disease associations)</i>	<i>ND</i>	<i>MD</i>	<i>HMD</i>	<i>HD</i>	<i>Other</i>	<i>Total</i>
<i>Ca_vs (9/10)</i>	101	6	11	5	3	126
<i>K_vs (15/40)</i>	286	6	525	0	15	832
<i>Na_vs (9/9)</i>	667	54	270	0	15	1006
<i>Total</i>	1054	66	806	5	33	1964

It is obvious that the importance of Channelopathies in the scope of neurological and muscular disorders is gathering growing attention. Recognition of their clinical significance will aid in the profiling of these diseases and the elucidation of their heterogeneity. Moreover,

the investigation of rare single-gene pathologies will hopefully provide the appropriate foundation for the delineation of the mechanisms governing more common polygenic disorders, like migraines [88].

4. Conclusions

The cascade of data from genome sequencing [92] and Genome Wide Association Studies (GWAS) [93, 94] has enabled the identification of novel associations between SNPs and specific disease phenotypes [95]. However, concerns have been raised in the past about many SNPs discovered through GWAS. They may have auxiliary roles in the development of a disease phenotype and thus turn out to be unsuitable candidates for the improvement of health care through genetic testing [96, 97]. For this reason, efforts are being made so that current analytical protocols can shift into more integrated approaches that account for the genomic context of genotype-phenotype relationships [98]. Bioinformatics could take a more active role in the address of this issue.

In this work, we focused on VGICs, a protein family with great pharmacological significance. We have proved that ‘structural’ protein information, like the topological profile of VGICs can contribute to the evaluation of the implications of genetic variations in this protein group. Our results showed that mutations that alter the charge and structural properties of amino acid chains are extremely important in the development of pathogenicity. Furthermore, two regions have been designated as the most important topological regions for disease implication; the VSD and the PL. More specifically, arginine (R) mutations in the S4 helix of the VSD appear to be the most significant disease associated mutations, based on our analysis and in accordance with literature data. We hope that the integration of topological information – when available – in the biostatistics protocols, used to catalog SNPs associated with specific disease phenotypes, will aid in the identification of clinically relevant mutations and in the determination of the pathogenic status of the numerous available unclassified SNPs.

Acknowledgements

The authors thank the National and Kapodistrian University of Athens for support.

5. References

- [1] M.S. Sansom, I.H. Shrivastava, J.N. Bright, J. Tate, C.E. Capener, P.C. Biggin, Potassium channels: structures, models, simulations, *Biochim Biophys Acta*, 1565 (2002) 294-307.
- [2] B. Hille, *Ion channels of excitable membranes*, 3rd ed., Sinauer, Sunderland, Mass., 2001.
- [3] M.J. Ackerman, D.E. Clapham, Ion channels--basic science and clinical disease, *N Engl J Med*, 336 (1997) 1575-1586.
- [4] F. Miceli, M.V. Soldovieri, P. Ambrosino, M. De Maria, L. Manocchio, A. Medoro, M. Taglialatela, Molecular pathophysiology and pharmacology of the voltage-sensing module of neuronal ion channels, *Front Cell Neurosci*, 9 (2015) 259.
- [5] Y. Moran, M.G. Barzilai, B.J. Liebeskind, H.H. Zakon, Evolution of voltage-gated ion channels at the emergence of Metazoa, *J Exp Biol*, 218 (2015) 515-525.
- [6] S. Molinarolo, S. Lee, L. Leisle, J.D. Lueck, D. Granata, V. Carnevale, C.A. Ahern, Cross-kingdom auxiliary subunit modulation of a voltage-gated sodium channel, *J Biol Chem*, 293 (2018) 4981-4992.
- [7] M. Noda, T. Ikeda, H. Suzuki, H. Takeshima, T. Takahashi, M. Kuno, S. Numa, Expression of functional sodium channels from cloned cDNA, *Nature*, 322 (1986) 826-828.
- [8] M. Noda, S. Shimizu, T. Tanabe, T. Takai, T. Kayano, T. Ikeda, H. Takahashi, H. Nakayama, Y. Kanaoka, N. Minamino, et al., Primary structure of *Electrophorus electricus* sodium channel deduced from cDNA sequence, *Nature*, 312 (1984) 121-127.
- [9] T. Tanabe, H. Takeshima, A. Mikami, V. Flockerzi, H. Takahashi, K. Kangawa, M. Kojima, H. Matsuo, T. Hirose, S. Numa, Primary structure of the receptor for calcium channel blockers from skeletal muscle, *Nature*, 328 (1987) 313-318.
- [10] S.B. Long, E.B. Campbell, R. Mackinnon, Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel, *Science*, 309 (2005) 897-903.
- [11] D.A. Doyle, J. Morais Cabral, R.A. Pfuetzner, A. Kuo, J.M. Gulbis, S.L. Cohen, B.T. Chait, R. MacKinnon, The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity, *Science*, 280 (1998) 69-77.
- [12] E. Bocksteins, D.J. Snyders, Electrically silent Kv subunits: their molecular and functional characteristics, *Physiology (Bethesda)*, 27 (2012) 73-84.
- [13] A.J. Yool, T.L. Schwarz, Alteration of ionic selectivity of a K⁺ channel by mutation of the H5 region, *Nature*, 349 (1991) 700-704.
- [14] T. Schlieff, R. Schonherr, K. Imoto, S.H. Heinemann, Pore properties of rat brain II sodium channels mutated in the selectivity filter domain, *Eur Biophys J*, 25 (1996) 75-91.
- [15] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, R. MacKinnon, X-ray structure of a voltage-dependent K⁺ channel, *Nature*, 423 (2003) 33-41.
- [16] J. Payandeh, T. Scheuer, N. Zheng, W.A. Catterall, The crystal structure of a voltage-gated sodium channel, *Nature*, 475 (2011) 353-358.
- [17] Q. Kuang, P. Purhonen, H. Hebert, Structure of potassium channels, *Cell Mol Life Sci*, 72 (2015) 3677-3693.
- [18] W. Stuhmer, F. Conti, H. Suzuki, X.D. Wang, M. Noda, N. Yahagi, H. Kubo, S. Numa, Structural parts involved in activation and inactivation of the sodium channel, *Nature*, 339 (1989) 597-603.
- [19] D.M. Papazian, T.L. Schwarz, B.L. Tempel, Y.N. Jan, L.Y. Jan, Cloning of genomic and complementary DNA from Shaker, a putative potassium channel gene from *Drosophila*, *Science*, 237 (1987) 749-753.
- [20] N. Yang, A.L. George, Jr., R. Horn, Molecular basis of charge movement in voltage-gated sodium channels, *Neuron*, 16 (1996) 113-122.

- [21] X.D. Tang, R. Xu, M.F. Reynolds, M.L. Garcia, S.H. Heinemann, T. Hoshi, Haem can bind to and inhibit mammalian calcium-dependent Slo1 BK channels, *Nature*, 425 (2003) 531-535.
- [22] S.J. Pitt, B. Reilly-O'Donnell, R. Sitsapesan, Exploring the biophysical evidence that mammalian two-pore channels are NAADP-activated calcium-permeable channels, *J Physiol*, 594 (2016) 4171-4179.
- [23] H. Hibino, A. Inanobe, K. Furutani, S. Murakami, I. Findlay, Y. Kurachi, Inwardly rectifying potassium channels: their structure, function, and physiological roles, *Physiol Rev*, 90 (2010) 291-366.
- [24] A.P. Braun, Two-pore domain potassium channels: variation on a structural theme, *Channels (Austin)*, 6 (2012) 139-140.
- [25] T.E. DeCoursey, Voltage-gated proton channels: molecular biology, physiology, and pathophysiology of the H(V) family, *Physiol Rev*, 93 (2013) 599-652.
- [26] L.A. Swayne, A. Mezghrani, A. Varrault, J. Chemin, G. Bertrand, S. Dalle, E. Bourinet, P. Lory, R.J. Miller, J. Nargeot, A. Monteil, The NALCN ion channel is activated by M3 muscarinic receptors in a pancreatic beta-cell line, *EMBO Rep*, 10 (2009) 873-880.
- [27] U.B. Kaupp, R. Seifert, Cyclic nucleotide-gated ion channels, *Physiol Rev*, 82 (2002) 769-824.
- [28] F. Van Petegem, Ryanodine receptors: structure and function, *J Biol Chem*, 287 (2012) 31624-31632.
- [29] B. Nilius, G. Owsianik, The transient receptor potential family of ion channels, *Genome Biol*, 12 (2011) 218.
- [30] J. Johnston, I.D. Forsythe, C. Kopp-Scheinpflug, Going native: voltage-gated potassium channels controlling neuronal excitability, *J Physiol*, 588 (2010) 3187-3200.
- [31] D.L. Neverisky, G.W. Abbott, Ion channel-transporter interactions, *Crit Rev Biochem Mol Biol*, 51 (2015) 257-267.
- [32] W.A. Catterall, A.L. Goldin, S.G. Waxman, International Union of Pharmacology. XLVII. Nomenclature and structure-function relationships of voltage-gated sodium channels, *Pharmacol Rev*, 57 (2005) 397-409.
- [33] R.H. Lee, C.J. Heckman, Essential role of a fast persistent inward current in action potential initiation and control of rhythmic firing, *J Neurophysiol*, 85 (2001) 472-475.
- [34] G.W. Zamponi, J. Striessnig, A. Koschak, A.C. Dolphin, The Physiology, Pathology, and Pharmacology of Voltage-Gated Calcium Channels and Their Future Therapeutic Potential, *Pharmacol Rev*, 67 (2015) 821-870.
- [35] L. Lu, P. Sirish, Z. Zhang, R.L. Woltz, N. Li, V. Timofeyev, A.A. Knowlton, X.D. Zhang, E.N. Yamoah, N. Chiamvimonvat, Regulation of gene transcription by voltage-gated L-type calcium channel, Cav1.3, *J Biol Chem*, 290 (2015) 4663-4676.
- [36] T.J. Jentsch, C.A. Hubner, J.C. Fuhrmann, Ion channels: function unravelled by dysfunction, *Nat Cell Biol*, 6 (2004) 1039-1047.
- [37] Z. Jiang, Y. Zhou, Using bioinformatics for drug target identification from the genome, *Am J Pharmacogenomics*, 5 (2005) 387-396.
- [38] J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there?, *Nat Rev Drug Discov*, 5 (2006) 993-996.
- [39] J. Dunlop, M. Bowlby, R. Peri, D. Vasilyev, R. Arias, High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology, *Nat Rev Drug Discov*, 7 (2008) 358-368.
- [40] O. Campuzano, P. Beltran-Alvarez, A. Iglesias, F. Scornik, G. Perez, R. Brugada, Genetics and cardiac channelopathies, *Genet Med*, 12 (2010) 260-267.
- [41] H. Lerche, K. Jurkat-Rott, F. Lehmann-Horn, Ion channels and epilepsy, *Am J Med Genet*, 106 (2001) 146-159.

- [42] M.J. Ackerman, Cardiac channelopathies: it's in the genes, *Nat Med*, 10 (2004) 463-464.
- [43] G. Bernard, M.I. Shevell, Channelopathies: A Review, *Pediatr Neurol*, 38 (2008) 73-85.
- [44] F.M. Ashcroft, From molecule to malady, *Nature*, 440 (2006) 440-447.
- [45] R. Felix, Channelopathies: ion channel defects linked to heritable clinical disorders, *J Med Genet*, 37 (2000) 729-740.
- [46] J. Zheng, M.C. Trudeau, *Handbook of ion channels*.
- [47] M.W. Stehpen T. Sherry, Karl Sirotkin, dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation, *Genome Research*, 9 (1999) 677-679.
- [48] S.D. Harding, J.L. Sharman, E. Faccenda, C. Southan, A.J. Pawson, S. Ireland, A.J.G. Gray, L. Bruce, S.P.H. Alexander, S. Anderton, C. Bryant, A.P. Davenport, C. Doerig, D. Fabbro, F. Levi-Schaffer, M. Spedding, J.A. Davies, I. Nc, The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY, *Nucleic Acids Res*, 46 (2018) D1091-D1106.
- [49] G.A. Gutman, K.G. Chandy, J.P. Adelman, J. Aiyar, D.A. Bayliss, D.E. Clapham, M. Covarriubias, G.V. Desir, K. Furuichi, B. Ganetzky, M.L. Garcia, S. Grissmer, L.Y. Jan, A. Karschin, D. Kim, S. Kuperschmidt, Y. Kurachi, M. Lazdunski, F. Lesage, H.A. Lester, D. McKinnon, C.G. Nichols, I. O'Kelly, J. Robbins, G.A. Robertson, B. Rudy, M. Sanguinetti, S. Seino, W. Stuehmer, M.M. Tamkun, C.A. Vandenberg, A. Wei, H. Wulff, R.S. Wymore, P. International Union of, International Union of Pharmacology. XLI. Compendium of voltage-gated ion channels: potassium channels, *Pharmacol Rev*, 55 (2003) 583-586.
- [50] W.A. Catterall, E. Perez-Reyes, T.P. Snutch, J. Striessnig, International Union of Pharmacology. XLVIII. Nomenclature and structure-function relationships of voltage-gated calcium channels, *Pharmacol Rev*, 57 (2005) 411-425.
- [51] E.A. Ertel, K.P. Campbell, M.M. Harpold, F. Hofmann, Y. Mori, E. Perez-Reyes, A. Schwartz, T.P. Snutch, T. Tanabe, L. Birnbaumer, R.W. Tsien, W.A. Catterall, Nomenclature of voltage-gated calcium channels, *Neuron*, 25 (2000) 533-535.
- [52] A.L. Goldin, R.L. Barchi, J.H. Caldwell, F. Hofmann, J.R. Howe, J.C. Hunter, R.G. Kallen, G. Mandel, M.H. Meisler, Y.B. Netter, M. Noda, M.M. Tamkun, S.G. Waxman, J.N. Wood, W.A. Catterall, Nomenclature of voltage-gated sodium channels, *Neuron*, 28 (2000) 365-368.
- [53] UniProt_Consortium, UniProt: the universal protein knowledgebase, *Nucleic Acids Res*, 46 (2018) 2699.
- [54] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: public archive of interpretations of clinically relevant variants, *Nucleic Acids Res*, 44 (2016) D862-868.
- [55] W. Martinez-Ortiz, T.J. Cardozo, An Improved Method for Modeling Voltage-Gated Ion Channels at Atomic Accuracy Applied to Human Cav Channels, *Cell Rep*, 23 (2018) 1399-1408.
- [56] S. Newstead, S. Ferrandon, S. Iwata, Rationalizing alpha-helical membrane protein crystallization, *Protein Sci*, 17 (2008) 466-472.
- [57] P. Fromme, J.C. Spence, Femtosecond nanocrystallography using X-ray lasers for membrane protein structure determination, *Curr Opin Struct Biol*, 21 (2011) 509-516.
- [58] K.D. Tsirigos, A. Hennerdal, L. Kall, A. Elofsson, A guideline to proteome-wide alpha-helical membrane protein topology predictions, *Proteomics*, 12 (2012) 2282-2294.
- [59] C.-I. Brañndén, J. Tooze, Introduction to protein structure, 2nd ed., Garland Pub., New York, 1999.

- [60] H.Y. Kim, Statistical notes for clinical researchers: Two-way analysis of variance (ANOVA)-exploring possible interaction between factors, *Restor Dent Endod*, 39 (2014) 143-147.
- [61] P. McCullagh, J.A. Nelder, Generalized linear models, 2nd ed., Chapman and Hall, London ; New York, 1989.
- [62] B. Efron, R. Tibshirani, An introduction to the bootstrap, Chapman & Hall, New York, 1993.
- [63] A.R. Henderson, The bootstrap: a technique for data-driven statistics. Using computer-intensive analyses to explore experimental data, *Clin Chim Acta*, 359 (2005) 1-26.
- [64] P. Gosselin-Badaroudine, D.I. Keller, H. Huang, V. Pouliot, A. Chatelier, S. Osswald, M. Brink, M. Chahine, A proton leak current through the cardiac sodium channel is linked to mixed arrhythmia and the dilated cardiomyopathy phenotype, *PLoS One*, 7 (2012) e38331.
- [65] P. Gosselin-Badaroudine, A. Moreau, M. Chahine, Nav 1.5 mutations linked to dilated cardiomyopathy phenotypes: Is the gating pore current the missing link?, *Channels (Austin)*, 8 (2014) 90-94.
- [66] E. Matthews, R. Labrum, M.G. Sweeney, R. Sud, A. Haworth, P.F. Chinnery, G. Meola, S. Schorge, D.M. Kullmann, M.B. Davis, M.G. Hanna, Voltage sensor charge loss accounts for most cases of hypokalemic periodic paralysis, *Neurology*, 72 (2009) 1544-1547.
- [67] S.N. Misra, K.M. Kahlig, A.L. George, Jr., Impaired Nav1.2 function and reduced cell surface expression in benign familial neonatal-infantile seizures, *Epilepsia*, 49 (2008) 1535-1545.
- [68] L. Volkers, K.M. Kahlig, N.E. Verbeek, J.H. Das, M.J. van Kempen, H. Stroink, P. Augustijn, O. van Nieuwenhuizen, D. Lindhout, A.L. George, Jr., B.P. Koeleman, M.B. Rook, Nav 1.1 dysfunction in genetic epilepsy with febrile seizures-plus or Dravet syndrome, *Eur J Neurosci*, 34 (2011) 1268-1275.
- [69] T.V. Wuttke, K. Jurkat-Rott, W. Paulus, M. Garncarek, F. Lehmann-Horn, H. Lerche, Peripheral nerve hyperexcitability due to dominant-negative KCNQ2 mutations, *Neurology*, 69 (2007) 2045-2053.
- [70] A. Moreau, P. Gosselin-Badaroudine, M. Chahine, Molecular biology and biophysical properties of ion channel gating pores, *Q Rev Biophys*, 47 (2014) 364-388.
- [71] J. Molnar, G. Szakacs, G.E. Tusnady, Characterization of Disease-Associated Mutations in Human Transmembrane Proteins, *PLoS One*, 11 (2016) e0151760.
- [72] L. Heginbotham, Z. Lu, T. Abramson, R. MacKinnon, Mutations in the K⁺ channel signature sequence, *Biophys J*, 66 (1994) 1061-1067.
- [73] F. Neumaier, M. Dibue-Adjei, J. Hescheler, T. Schneider, Voltage-gated calcium channels: Determinants of channel function and modulation by inorganic cations, *Prog Neurobiol*, 129 (2015) 1-36.
- [74] S.H. Heinemann, H. Terlau, W. Stuhmer, K. Imoto, S. Numa, Calcium channel characteristics conferred on the sodium channel by single mutations, *Nature*, 356 (1992) 441-443.
- [75] M.V. Soldovieri, M.R. Cilio, F. Miceli, G. Bellini, E. Miraglia del Giudice, P. Castaldo, C.C. Hernandez, M.S. Shapiro, A. Pascotto, L. Annunziato, M. Taglialatela, Atypical gating of M-type potassium channels conferred by mutations in uncharged residues in the S4 region of KCNQ2 causing benign familial neonatal convulsions, *J Neurosci*, 27 (2007) 4919-4928.
- [76] S. Sokolov, T. Scheuer, W.A. Catterall, Gating pore current in an inherited ion channelopathy, *Nature*, 446 (2007) 76-78.
- [77] F. Tombola, M.M. Pathak, E.Y. Isacoff, Voltage-sensing arginines in a potassium channel permeate and occlude cation-selective pores, *Neuron*, 45 (2005) 379-388.

- [78] Y. Ruan, N. Liu, S.G. Priori, Sodium channel mutations and arrhythmias, *Nat Rev Cardiol*, 6 (2009) 337-348.
- [79] A. Moreau, P. Gosselin-Badaroudine, L. Delemotte, M.L. Klein, M. Chahine, Gating pore currents are defects in common with two Nav1.5 mutations in patients with mixed arrhythmias and dilated cardiomyopathy, *J Gen Physiol*, 145 (2015) 93-106.
- [80] C. Fan, F. Lehmann-Horn, M.A. Weber, M. Bednarz, J.R. Groome, M.K. Jonsson, K. Jurkat-Rott, Transient compartment-like syndrome and normokalaemic periodic paralysis due to a Ca(v)1.1 mutation, *Brain*, 136 (2013) 3775-3786.
- [81] M. Chahine, S. Pilote, V. Pouliot, H. Takami, C. Sato, Role of arginine residues on the S4 segment of the *Bacillus halodurans* Na⁺ channel in voltage-sensing, *J Membr Biol*, 201 (2004) 9-24.
- [82] J. Blanchet, M. Chahine, Accessibility of four arginine residues on the S4 segment of the *Bacillus halodurans* sodium channel, *J Membr Biol*, 215 (2007) 169-180.
- [83] M.H. Wang, S.P. Yusaf, D.J. Elliott, D. Wray, A. Sivaprasadarao, Effect of cysteine substitutions on the topology of the S4 segment of the Shaker potassium channel: implications for molecular models of gating, *J Physiol*, 521 Pt 2 (1999) 315-326.
- [84] F. Miceli, M.V. Soldovieri, C.C. Hernandez, M.S. Shapiro, L. Annunziato, M. Taglialatela, Gating consequences of charge neutralization of arginine residues in the S4 segment of K(v)7.2, an epilepsy-linked K⁺ channel subunit, *Biophys J*, 95 (2008) 2254-2264.
- [85] S. Sokolov, T. Scheuer, W.A. Catterall, Depolarization-activated gating pore current conducted by mutant sodium channels in potassium-sensitive normokalemic periodic paralysis, *Proc Natl Acad Sci U S A*, 105 (2008) 19980-19985.
- [86] A. Moreau, P. Gosselin-Badaroudine, M. Chahine, Biophysics, pathophysiology, and pharmacology of ion channel gating pores, *Front Pharmacol*, 5 (2014) 53.
- [87] H.F. Lodish, *Molecular cell biology*, 6th ed., W.H. Freeman, New York, 2008.
- [88] R.S. Kass, The channelopathies: novel insights into molecular and genetic mechanisms of human disease, *J Clin Invest*, 115 (2005) 1986-1989.
- [89] M.H. Meisler, J.A. Kearney, Sodium channel mutations in epilepsy and other neurological disorders, *J Clin Invest*, 115 (2005) 2010-2017.
- [90] M. de Lera Ruiz, R.L. Kraus, Voltage-Gated Sodium Channels: Structure, Function, Pharmacology, and Clinical Indications, *J Med Chem*, 58 (2015) 7093-7118.
- [91] W.A. Catterall, Voltage-gated calcium channels, *Cold Spring Harb Perspect Biol*, 3 (2011) a003947.
- [92] C. International Human Genome Sequencing, Finishing the euchromatic sequence of the human genome, *Nature*, 431 (2004) 931-945.
- [93] A.D. Johnson, C.J. O'Donnell, An open access database of genome-wide association results, *BMC Med Genet*, 10 (2009) 6.
- [94] J.N. Hirschhorn, M.J. Daly, Genome-wide association studies for common diseases and complex traits, *Nat Rev Genet*, 6 (2005) 95-108.
- [95] W.Y. Wang, B.J. Barratt, D.G. Clayton, J.A. Todd, Genome-wide association studies: theoretical and practical concerns, *Nat Rev Genet*, 6 (2005) 109-118.
- [96] J. Jakobsdottir, M.B. Gorin, Y.P. Conley, R.E. Ferrell, D.E. Weeks, Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers, *PLoS Genet*, 5 (2009) e1000337.
- [97] J.H. Moore, From genotypes to genomotypes: putting the genome back in genome-wide association studies, *Eur J Hum Genet*, 17 (2009) 1205-1206.
- [98] J.H. Moore, F.W. Asselbergs, S.M. Williams, Bioinformatics challenges for genome-wide association studies, *Bioinformatics*, 26 (2010) 445-455.

Supporting Information

Table S1: The dataset of Voltage Gated Ion Channels from IUPHAR. Green Color in Column AC indicates the Voltage dependent ion channels incorporated in the final dataset (namely KVs, CaVs and NaVs). The column "Comment" explains briefly the reason why a channel or an entire family of channels (orange cells) is excluded from our analysis (for more details please refer to the main manuscript).

Table S2: Mapping between RefSeq Nucleotide Codes and UniProt ACs for the canonical isoforms of VGICs. This mapping was used to gather data from ClinVar regarding the SNPs of VGICs in the GRCh38 version of the proteome.

Table S3 (A): The full dataset of missense SNPs located on KVs (2035 SNPs), CaVs (741 SNPs) and NaVs (2535 SNPs). The excel sheet named "Full" contains a dataset of unique SNPs from both databases (5311 SNPs). **(B): The dataset of missense SNPs** located on KVs (669 SNPs), CaVs (188 SNPs) and NaVs (1128 SNPs) as isolated from UniProt. The excel sheet "UniProt" contains all SNPs isolated from UniProtKB (1985 SNPs). **(C): The dataset of missense SNPs** located on KVs (1964 SNPs), CaVs (593 SNPs) and NaVs (1732 SNPs) as isolated from ClinVar. This excel sheet named "ClinVar" contains all SNPs isolated from ClinVar (4289 SNPs).

Table S4: The dataset of Voltage Gated Ion Channels with SNPs. In the sheet "Pathogenic SNPs" 33 VGICs with 1765 SNPs are shown. In the sheet "Polymorphisms" 42 VGICs with 457 SNPs are shown and in the sheet "Unclassified" 41 VGICs with 3089 SNPs are shown.

Table S5: The full dataset of missense SNPs located on KVs (2035 SNPs), CaVs (741 SNPs) and NaVs (2535 SNPs). The topological domain in which each SNP is located is given. Topological domains were characterized based on the topological profile of VGICs and are the following: C-terminal, IL1-11: Intracellular Loop 1-11, EL1-8: Extracellular Loop 1-8, S1-S6: Transmembrane Segment 1-6 (for KVs), S(1-6)D(I-IV): Transmembrane Segment 1-6 from Domain I-IV (for CaVs and NaVs), PL: Pore loop (for KVs), PL1-4: Pore loop 1-4 (for CaVs and NaVs), N-terminal.

Table S6: Results from two-factor with replication ANOVA. The difference in the distribution of "Polymorphisms" and "Pathogenic" mutations in regions of biological significance (VSD, PL, ILs, ELs, Other) is examined. The Sample difference suggests that the means of observations between "Polymorphisms" and "Pathogenic" mutations varies

($F > F_{crit}$, $P\text{-value} < 0.05$). The Columns difference examines the difference of means between the different regions of biological significance, and suggests that there is no difference observed between the appearance of SNPs in these regions ($F < F_{crit}$, $P\text{-value} > 0.05$). Moreover, the Interaction difference of means between these two factors (pathogenicity status and appearance in biologically significant regions) could not be rejected ($F < F_{crit}$, $P\text{-value} > 0.05$), and suggests that these factors are not associated. Statistically significant results are shown in bold. SS: Sum of Squares due to the source, df: degrees of freedom in the source, MS: Mean of Squares due to the source, F: F-test statistic, P-value: P-value associated with the F-test, F crit: F critical (from the F distribution).

Table S7: Relative frequencies of amino acid substitutions. The relative frequencies for “Polymorphisms” and “Pathogenic” Mutations for the entire VGICs (Sheets: *Polymorphisms_Total* and *Pathogenic_Total*) are given, along with the relative frequencies for regions with biological significance (*Pathogenic_VSD* and *Pathogenic_PL*).

Table S8: Bootstrap analysis. The significance of the different kind of substitutions. Tables represent the observed (OBS) values, the average values of random sampling of 1000 times (AVG), the standard deviation of the random sampling (SD), the upper (CIHI) and lower (CILO) limits of the 95% Confidence Intervals (CI) and the statistical significance testing for each observation (TEST). The result is considered statistically significant if the observed value lies within the upper and lower limits of the 95% confidence intervals. Amino acids before substitutions are shown in rows and amino acids after substitutions in columns. Four tests were performed. One for all “Polymorphisms” (*Polymorphisms_Total*), one for all “Pathogenic” mutations (*Pathogenic_Total*), one for “Pathogenic” mutations in the VSD (*Pathogenic_VSD*) and one for “Pathogenic” mutations in the Pore Loop (*Pathogenic_PL*).

Table S9: Counts of the 7 most prominent mutations identified in VGICs and Navs, Cavs and Kvs, grouped based on the topological regions they are located.

Table S10: (A) SNPs with disease associations. In total 1964 associations were recorded for 1490 unique SNPs. Some SNPs have more than one association with different diseases. ND: Neuronal Disease, MD: Muscular Disease, HMD: Heart Muscle Disease, HD: Hormonal Disease, **(B) List of Channelopathies VGICs' SNPs are associated with.** A total of 132 different diseases were identified (excluding general disease categories), **(C) List of Common Channelopathies between different VGIC families.**

Collection
of human
VGICs

Collection
of
missense
SNPs

Topological information

Mapping of all
missense SNPs
on human
VGICs and
disease
association

Disease association

Statistical
Analysis

CTGATGGTATGGGGCCAAGAGATA
AGGTACGGGTGTCTACACTTAGAC
AGGGCTGGGGAAGAAGTCAAGGCG
CATGGTGGCATCTGACTCTGAGGA
CAGGTTGGTATCAAGGTTACAAGA
GCACTGACTCTCTCTGGCTATTGG

CTGATGGTATGGGGCCAAGAGATA
AGGTACGGGTGTCTACACTTAGAC
AGGGCTGGGGAAGAAGTCAAGGCG
CATGGTGGCATCTGACTCTGAGGA
CAGGTTGGTATCAAGGTTACAAGA
GCACTGACTCTCTCTGGCTATTGG

Amino Acid After

A C D E F G H I K L M N P Q R S T V W Y

Amino Acid Before

0.00-1.00% 1.00-2.00% 2.00-3.00% 3.00-4.00%