# Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance

Bian Li, MSc; Jeffrey L. Mendenhall, MSc; Brett M. Kroncke, PhD; Keenan C. Taylor, PhD; Hui Huang, PhD; Derek K. Smith, DDS, PhD; Carlos G. Vanoye, PhD; Jeffrey D. Blume, PhD; Alfred L. George, Jr., MD; Charles R. Sanders, PhD; Jens Meiler, PhD

***Background***—An emerging standard-of-care for long-QT syndrome uses clinical genetic testing to identify genetic variants of the KCNQ1 potassium channel. However, interpreting results from genetic testing is confounded by the presence of variants of unknown significance for which there is inadequate evidence of pathogenicity.

***Methods and Results***—In this study, we curated from the literature a high-quality set of 107 functionally characterized KCNQ1 variants. Based on this data set, we completed a detailed quantitative analysis on the sequence conservation patterns of subdomains of KCNQ1 and the distribution of pathogenic variants therein. We found that conserved subdomains generally are critical for channel function and are enriched with dysfunctional variants. Using this experimentally validated data set, we trained a neural network, designated Q1VarPred, specifically for predicting the functional impact of KCNQ1 variants of unknown significance. The estimated predictive performance of Q1VarPred in terms of Matthew's correlation coefficient and area under the receiver operating characteristic curve were 0.581 and 0.884, respectively, superior to the performance of 8 previous methods tested in parallel. Q1VarPred is publicly available as a web server at http://meilerlab.org/q1varpred.

***Conclusions***—Although a plethora of tools are available for making pathogenicity predictions over a genome-wide scale, previous tools fail to perform in a robust manner when applied to KCNQ1. The contrasting and favorable results for Q1VarPred suggest a promising approach, where a machine-learning algorithm is tailored to a specific protein target and trained with a functionally validated data set to calibrate informatics tools. **(*Circ Cardiovasc Genet*. 2017;10:e001754. DOI: 10.1161/CIRCGENETICS.117.001754.)**

**Key Words:** genetic variation ■ KCNQ1 potassium channel ■ long QT syndrome ■ neural network models



Congenital long-QT syndrome (LQTS) is a heart rhythm disorder that affects ≈1 in 2500 births.[1] It predisposes children and young adults to a type of ventricular tachycardia (torsades de pointes) and sudden cardiac death.[2] LQTS is associated with pathogenic variants in several genes that lead to dysfunctional cardiac ion channels. Among the 16 known LQTS-associated genes, *KCNQ1* variants account for ≈30% to ≈35% of all LQTS cases. *KCNQ1* encodes the α-subunit of the voltage-gated $K^+$ channel KCNQ1 (also known as $K_V7.1$) that regulates the slow delayed rectifier current ($I_{Ks}$), a major driver of cardiac repolarization.[3] Loss of KCNQ1 function leads to diminished or dysfunctional $I_{Ks}$, impaired myocardial repolarization, and LQTS.[4]

**See Editorial by Giudicessi**
**See Clinical Perspective**

An emerging standard-of-care for LQTS uses clinical genetic testing to identify LQTS-associated variants.[4] Established genotype–phenotype relations should be factored into the assessment of the risk of sudden cardiac death and the selection of appropriate therapeutic interventions.[5] However, variants of unknown significance for which there is inadequate evidence to classify as being pathogenic are common findings.[6] This issue is further confounded by the presence of background genetic noise (the frequency of genetic variations of a particular gene in a healthy population) and variants with incomplete penetrance.[5–7] Variant interpretation is bound to present an increasingly daunting challenge in the era of next-generation sequencing.[7–9]

Ideally, except for certain well-established disease-causing variants, positive LQTS genetic testing results should be evaluated by physiologically relevant experimental functional assays, but experimental characterization remains labor-intensive and costly to scale.[9,10] Under such constraints, computational methods, which are usually machine learning based, represent a common predictive approach.[8,11,12] However, hardly any computational methods are sufficiently accurate for clinical use related to channelopathies or other genetic disorders.[13,14] Most existing computational methods have been

Received March 6, 2017; accepted August 24, 2017.

From the Department of Chemistry (B.L., J.L.M., J.M.), Center for Structural Biology (B.L., J.L.M., B.M.K., K.C.T., H.H., C.R.S., J.M.), Department of Biochemistry (B.M.K., H.H., C.R.S.), and Department of Biostatistics (D.K.S., J.D.B.), Vanderbilt University, Nashville, TN; Department of Medicine, Vanderbilt University Medical Center, Nashville, TN (B.M.K., C.R.S.); and Department of Pharmacology, Northwestern University Feinberg School of Medicine, Chicago, IL (C.G.V., A.L.G.).

**The Data Supplement is available at http://circgenetics.ahajournals.org/lookup/suppl/doi:10.1161/CIRCGENETICS.117.001754/-/DC1.**

Correspondence to Jens Meiler, PhD, Departments of Chemistry and Pharmacology, Center for Structural Biology, Vanderbilt University, 465 21st Ave S, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232–8725. E-mail jens.meiler@vanderbilt.edu

trained on data sets pulled from online databases that have not been subjected to rigorous functional validation.[12] These data sets may be significantly contaminated with erroneous annotations and thereby provide machine-learning algorithms with misleading information.[12,15] Furthermore, a potentially even more crucial issue is that current methods intermingle 2 related but separate questions: whether a given variant causes functional impact at the molecular level and, if so, whether that functional effect will be manifested at the organismal level. Making such distinctions is important when delivering predictions because dysfunction at molecular level does not necessarily equate to organismal deleteriousness.[7,8]

In this study, we sought to develop a protein-specific algorithm capable of accurately predicting functional consequences of KCNQ1 variants. We first curated a set of functionally validated KCNQ1 variants. We then trained a neural network-based, KCNQ1-specific genotype–channel function relationship predictor Q1VarPred. In contrast to genome-wide methods, whose performances have experienced data set contamination and heterogeneity and do not differentiate between functional impact and organismal deleteriousness when delivering predictions, Q1VarPred was trained on the functionally validated data set to predict molecular functional impact.

## Materials and Methods

### Data Set and Criteria for Annotating Functional Impact

KCNQ1 variants and their associated electrophysiological effects in the data set for this study were collected from the literature (Table I in the Data Supplement). We only considered data from experiments where the auxiliary subunit KCNE1 was also expressed. Each variant was annotated in terms of functional impact based on 2 experimental parameters (peak current relative to the wild type and change in voltage of half-maximal activation V½). Specifically, a variant was defined as normal if (1) 75% ≤ peak current ≤ 125%, and (2) there was ≤10 mV depolarization or hyperpolarization shift in V½. Mild loss of function was defined as (1) 25% < peak current < 75% or (2) 10 to 20 mV depolarization shift in V½. Severe loss of function was defined as (1) peak current < 25% or (2) >20 mV depolarization shift in V½. Severe gain of function was defined as (1) >150% peak current or (2) 120% to 150% peak current and >15 mV hyperpolarization shift in V½. Clinical classification (case variant versus control) of each variant was sourced from previous large-scale clinical studies[16,17] or electrophysiological studies that reported such information. Case variants were identified in patient cohort, whereas control variants were found in healthy cohort. In addition, in accordance to the recent American College of Medical Genetics and Genomic and the Association for Molecular Pathology standards and guidelines for the interpretation of sequence variants,[12] variants with a minor allele frequency of >1/2500 (LQTS prevalence) in the general population were removed. For training the binary classification model Q1VarPred, loss of function and gain of function variants were grouped together as dysfunctional, and a mild loss of function variant was either labeled as dysfunctional if its peak current was <50% or normal otherwise. The common variant G643S was classified as having normal function.[18]

### Neural Network Architecture and Training

The neural network in the present study was a fully connected 3-layer feed-forward network with a sigmoid transfer function. The input layer consists of 2 nodes, 1 for each predictive feature. The output layer consists of a single neuron that outputs a numeric prediction of the functional impact of a given variant on the scale of 0 to 1 with 1 being complete dysfunction. A hidden layer with 3 neurons was chosen considering the fact that the dropout technique[19] was adopted to prevent the neural network from overfitting, a phenomenon in which the learned model is excessively complex (eg, too many model parameters relative to the number of observations for training) and is poorly generalizable. However, we also tested hidden layers with up to 8 neurons, the results of which showed that the size of the hidden layer did not affect the performance of the neural network in a significant way (Table II in the Data Supplement). The neural network was trained on numeric encoding of variant functional labels (1 for complete dysfunction 0 for normal), with back-propagation of errors. The learning rate was set to 0.05 and momentum was set to 0.8. Weights were updated after each presentation of a variant to the network, and a constant weight decay of 0.02 was applied to reduce model flexibility.

### Predictive Features

We used 2 features to characterize an amino acid substitution, namely rate of evolution, which quantifies the conservation of the sequence position where the substitution has occurred, and position-specific scoring matrix–based perturbation, which measures the radicalness of the substitution itself. These 2 features were chosen, before the data set was inspected, based on the rationale that a conserved position may tolerate less radical substitutions while a variable position may not tolerate more radical substitutions as—for example—observed in a systematic mutation study of bacteriophage T4 lysozyme.[20] We confirmed that these features are critical among a limited number of features tested (Table III in the Data Supplement). Details on how these 2 features were computed can be found in Methods in the Data Supplement.

### Performance Metrics

The performance of the learned neural network model and other evaluated methods were quantified using the following metrics: true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), negative predictive value (NPV), accuracy, Matthew's correlation coefficient (MCC),[21] and area under the receiver operating characteristic curve (AUC). Note that the first 6 metrics can be computed only after all variants are classified at a specific threshold. Using the notation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), these metrics are defined as:

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \quad (1)$$

$$\mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{TN+FP}} \quad (2)$$

$$\mathrm{PPV}=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (3)$$

$$\mathrm{NPV}=\frac{\mathrm{TN}}{\mathrm{TN+FN}} \quad (4)$$

$$\mathrm{Accuracy}=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \quad (5)$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TN+FN})(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})}} \quad (6)$$

respectively. A TP is a dysfunctional variant classified as dysfunctional, and TN is a normal variant classified as normal. MCC measures the correlation between predicted and observed binary classifications, with a value between −1 and 1. An MCC of 1 means perfect classification, a value of 0 means no better than random classification, and −1 indicates a completely reversed classification. Because MCC is unaffected by class size, it is a particularly useful measure of classification quality when the 2 classes are of different sizes.[21] Computation of all performance metrics was accomplished using the ROCR package[22] implemented in the R programming environment.[23]

### Estimating Generalization Ability

The generalization ability of a learned model is defined as its performance in predicting new variants that are not used for training. A model with higher generalization ability is favored over ones with lower generalization ability. A common practice to estimate a model's generalization ability is through a procedure called $k$-fold cross-validation where the data set is randomly divided into $k$ equally-sized mutually exclusive subsets. The model is trained on $k-1$ subsets (collectively known as the training set), and its generalization ability is estimated on the remaining 1 subset (test set). Specifically, after the model is trained, a threshold is determined which maximizes the MCC on the training set, the same threshold is then used for computing the performance metrics on the test set. This process is repeated $k$ times each using a different one of the $k$ subsets as the test set and the remaining $k-1$ subsets as the training set. Every time a model is trained, its performance metrics are computed on the test set. In a $k$-fold cross-validation, the generalization ability is estimated as the average of performance metrics over $k$ test sets. Because the number of ways a data set can be split into $k$ subsets is enormous, it is desirable to repeat the random splitting $p$ times to reduce artifacts. In the current study, we chose $k=3$ and $p=200$, similar to a previous study.[24] The splitting was stratified such that the class proportions of the training set and the test set are as close to that of the whole data set as possible. To ensure the consistency of comparison, the performance metrics of all evaluated methods were estimated using the exact same data. This means that every time the data set was randomly split into 3 subsets, these subsets were used for calculating the performance metrics of all methods. The variability in performance metrics associated with random splitting of data set is presented in Table IV in the Data Supplement.

## Results

### Functional Studies Do Not Always Agree With Clinical Testing

We compiled a total of 107 functionally characterized KCNQ1 variants (Table I in the Data Supplement). Two important observations were made on this data set. First, not all case variants (variants identified in LQTS patient cohort, a total of 99 in our data set) are severely dysfunctional. Per our scheme of functional annotation (see Data Set and Criteria for Annotating Functional Impact), 6 of 99 case variants are functionally normal and 8 of 99 cause only mild loss of function. Interestingly, these 2 fractions roughly agree with the previous estimate that ≈10% case variants may be false positives.[16] However, a few variants identified in presumed healthy controls are severely dysfunctional (eg, V110I and A300T). A300T, which occurs within the pore-helix of the channel, was shown to cause a massive reduction of $I_{Ks}$ and hyperpolarization of the voltage of half-activation of the channel both with and without the presence of the wild-type subunit.[25] The V110I variant showed significant reduction in $I_{Ks}$ and depolarization of voltage of half-maximal activation when expressed in the absence of the wild-type subunit.[26] This analysis reinforces the argument that translating protein dysfunction at the molecular level to clinical manifestation and also attributing clinical manifestation to protein dysfunction both need to be performed with caution.[5]

### Position-Specific Rate of Evolution Reflects Functionally Critical Subdomains

The importance of a sequence site for protein structure or function can often be inferred from its conservation over evolution. We computed the position-specific rate of evolution for the entire sequence, as well as the mean rate of evolution for each of the 24 subdomains of KCNQ1 (Methods in the Data Supplement). A lower rate of evolution indicates higher conservation.

Overall, the N-terminal domain (NTD) and C-terminal domain (CTD) are generally less conserved than subdomains within the channel domain (CD), as shown in Figure 1. The average rates of evolution for the NTD and CTD are 3.2 and 2.5, respectively, whereas the average rate of evolution in the CD is 1.0. Within the CD, 6 subdomains have a mean rate of evolution below 1.0 (S4, S4-S5, S5, pore-helix, pore-loop, and S6). As expected, the pore-helix (residues 299–312) and pore-loop (residues 313–322) of the channel are the most conserved subdomains, with mean rates of evolution of only 0.38 and 0.41, respectively. This correlates with the critical role played by these components in achieving high ion selectivity for $K^+$ over $Na^+$ ions.[27] The S4 segment of the CD, which harbors basic residues for sensing and responding to changes in membrane potential,[28] has a mean rate of evolution of 0.61. The S4-S5 linker, which is thought to be responsible for transferring the conformational changes in the voltage sensor domain to the pore[29] and serve as binding sites for phosphatidylinositol-4,5-bisphosphate to modulate the deactivation rate of the channel,[30] has a mean rate of evolution of 0.92. The S2-S3 linker, proposed in a recent study to also bind phosphatidylinositol-4,5-bisphosphate,[31] is only moderately conserved. Interestingly, although most subdomains of the CD exhibit a low mean rate of evolution, 2 subdomains, namely the S1-S2 linker and the S5-Pore linker, show substantially higher mean rates of evolution (2.5 and 1.9, respectively) than the rest of the CD.

Because the CTD has been shown to have 4 helices designated A to D,[32] we computed the mean rate of evolution of each of these helices and their linkers to see if any of these subdomains are conserved. Our analysis shows that only helices A, B, and C have a mean rate of evolution <1.0, whereas the mean rate of evolution of helix D is substantially higher (1.9). This observation agrees with the functional role of helices A and B in binding calmodulin and the critical role of helix C in tetramerization of the intracellular C-terminal domain.[32,33] The juxtramembrane subdomain S6-A, with a mean rate of evolution of 0.88, as well as the B-C linker, considered extremely conserved according to its mean rate of evolution (0.24), have yet to be shown to play any particular functional role.

### Dysfunctional Variants Are Enriched in Selected Subdomains

Results from a recent study suggested that the probability of pathogenicity of a KCNQ1 variant depends in part on the topological location of the variant.[17] However, in the previous study, the protein was only divided into 3 topological domains namely NTD, CD, and CTD. We mapped all variants in our data set onto the curve of position-specific rates of evolution (Figure 1A). We observed that dysfunctional variants preferentially occur at positions with low rate of evolution, especially within a selected set of subdomains.

In fact, 95.7% (90 of 94) dysfunctional variants occur at positions where the rate of evolution is <2. In contrast, 61.5% (8 of 13) of normal variants occur at positions with rates of evolution >2. The 5 normal variants that occur at positions with a rate of evolution under 2 are Q147R, G179S, T391I,

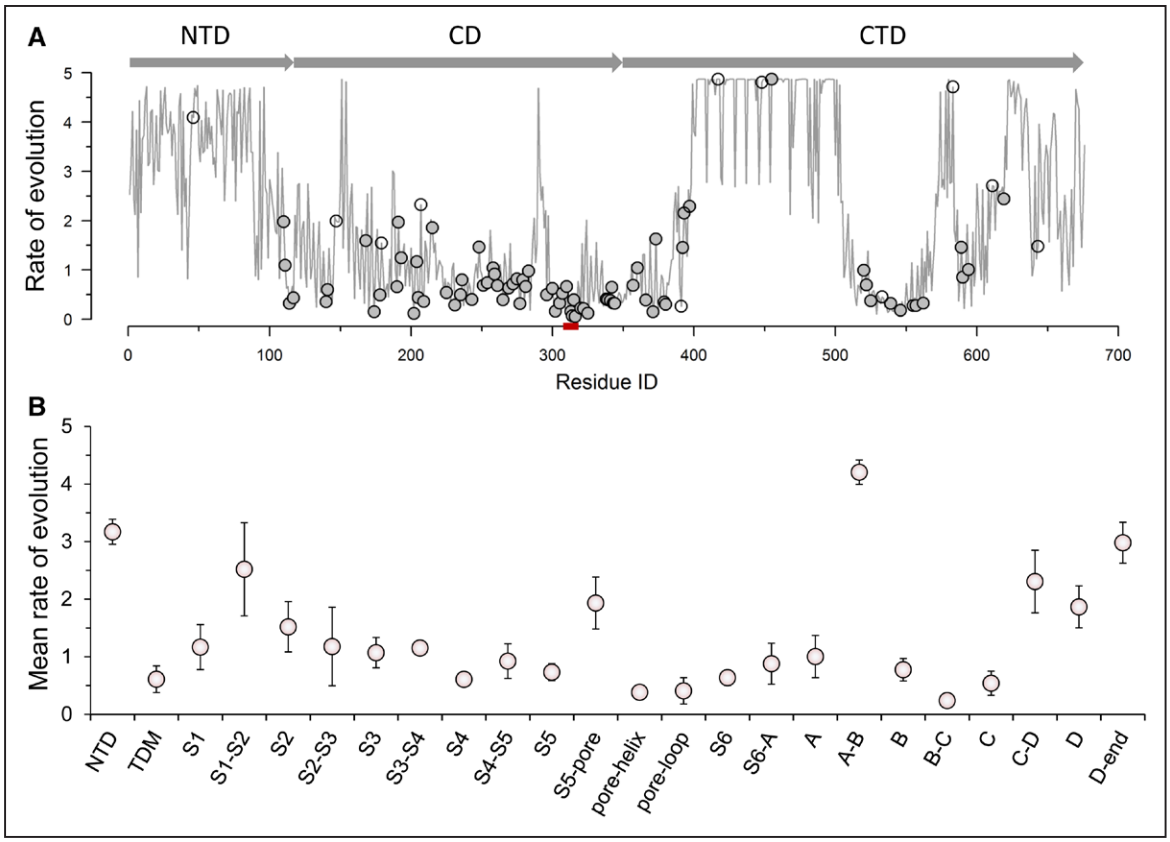


**Figure 1.** Analysis on the evolutionary variability of the KCNQ1 sequence. **A**, Position-specific rate of evolution. Shaded arrow bars on the top indicate the sequence range of N-terminal domain (NTD), channel domain (CD), and C-terminal domain (CTD), respectively. The small red bar on the horizontal axis highlights the selectivity filter TIGYG. Closed circles represent dysfunctional variants, and open circles represent normal variants. **B**, Mean rates of evolution for structurally distinct subdomains of NTD, CD, and CTD. Note that the trafficking determinant motif (TDM), which resides within the NTD, is singled out for its distinct functional role. Error bars indicate the 95% confidence intervals (under Student *t* distribution) for the mean rate of evolution.

R533W, and G643S. Interestingly, Q147R, G179S, T391I, and G643S are chemically conserved as judged by their Grantham distances[34]: Q→R (68), G→S (56), T→I (89). Nevertheless, this clear segregation of the functional impact of variants with respect to position-specific rate of evolution indicates that the rate of evolution of a sequence site preselected as one of the predictive features is indeed a strong predictor on whether variants occurring at the site will be dysfunctional or not.

In addition, we also computed the enrichment of dysfunctional variants for each subdomain to confirm that such variants are indeed localized within a selected set of subdomains (Methods and Table V in the Data Supplement). An enrichment of >1.0 indicates that the corresponding subdomain has higher than random chance of harboring dysfunctional variants. As shown in Figure 2, subdomains with higher than random chance for dysfunctional variants are S0, S2-S3 linker, S3, S4, S4-S5, S5, pore-helix, pore-loop, S6, S6-A, B-C, and C. In particular, S0, S3, S4-S5 linker, S5, pore-loop, and S6-A each have an enrichment ≥3. As discussed in the previous section, these subdomains are highly conserved.

## Q1VarPred: A KCNQ1-Specific Predictor

A schematic representation of the architecture of Q1VarPred is shown in Figure 3A. Figure 3B shows a visualization of the Q1VarPred model of the relationship between predictive features (rate of evolution and position-specific scoring matrix–based perturbation) and the prediction about functional impact (impact score 0, most likely normal; 1, most likely dysfunctional). The contour surface indicates that the impact score has a sharper dependence on the rate of evolution than it does on position-specific scoring matrix–based perturbation. In particular, variants at conserved positions (rate of evolution close to 0) are likely to be dysfunctional (impact score >0.5) even if the perturbation is small. An example of such variants is the dysfunctional V307L whose impact was predicted to be 0.68. The estimated rate of evolution of this position is 0.52, whereas the perturbation introduced by substituting Val for Leu at this position is considerably small (3.7). Similarly, variants at evolutionarily tolerated positions (eg, rate of evolution >3.0) tend to be normal even if the perturbation is large (eg, R583H). However, the impact score does rise along with increasing magnitude of perturbation, which is particularly important for predicting the impact at positions exhibiting intermediate rates of evolution.

## Comparing Q1VarPred With Other Methods

We used a procedure called repeated cross-validation[24] to estimate the generalization ability of Q1VarPred and other methods (see Estimating Generalization Ability). Seven commonly used genome-wide methods: PhD-SNP, Polyphen-2, PredictSNP, PROVEAN, SIFT, SNAP, and SNPs&GO and

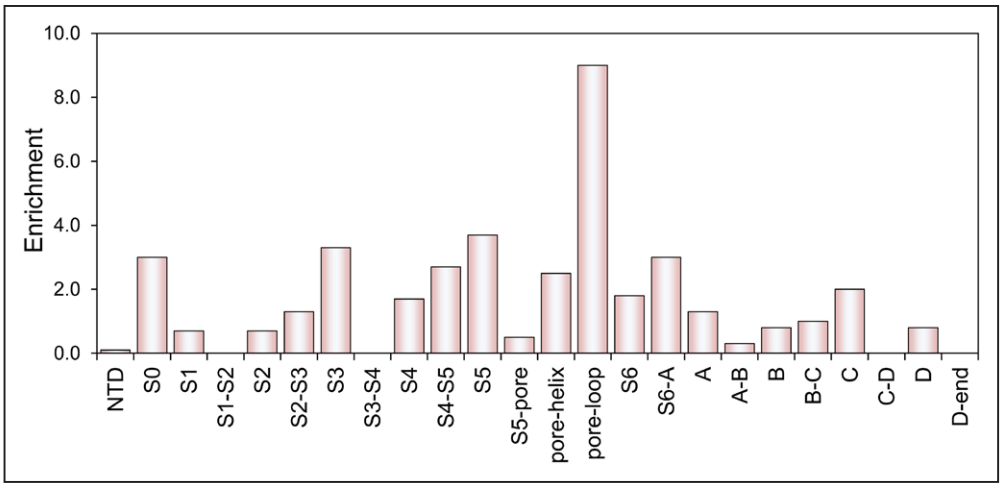


**Figure 2.** This bar graph plots subdomain-specific enrichment of dysfunctional variants, showing that dysfunctional variants are enriched in a selected set of subdomains (S0, S3, S4-S5, S5, pore-helix, pore-loop, S6-A, see Table V in the Data Supplement for the residue ranges these subdomains correspond to). One needs to keep in mind that because of the sparsity of functionally characterized variants, the estimates of enrichments are likely to be biased. NTD indicates N-terminal domain.



1 potassium channel-specific method called KvSNP were examined (Methods and Table VI in the Data Supplement). The Table shows that all performance metrics rank Q1VarPred the best, except for NPV and TPR. In general, AUC and MCC are considered the most robust metrics for evaluating classifiers. AUC is independent of user-chosen and therefore possibly biased thresholds. MCC has the advantage to consider all 4 numbers (TP, TN, FP, and FN) and provides a much more balanced evaluation than TPR or TNR individually.[35] In terms of AUC, Q1VarPred>PROVEAN>PhD-SNP>SNPs&GO>SIFT>KvSNP>PredictSNP>PolyPhen-2>SNAP. This is similar to the findings of Leong et al[36] except that PolyPhen-2 was shown to rank between PROVEAN and SNP&GO, and PhD-SNP and KvSNP were not evaluated in Leong et al.[36] Methods that perform better than Q1VarPred in TPR do so at a cost of a low TNR, that is, the threshold is chosen to minimize the loss of TP at the cost of predicting many FP. In some disease conditions, a high fraction of FP might be acceptable. However, in LQTS and related channelopathies, the cost of FP is as drastic as that of FN.[6] It is also worth noting that while KvSNP is gene specific, our evaluation shows that its performance is worse than most genome-wide methods on this data set. The primary cause of the inflation in KvSNP's claimed performance is probably its convolution of data set preparation and feature selection, where 85.5% of neutral variants were generated from variable sequence positions, and later several sequence conservation-based features were selected as predictive features.[37]

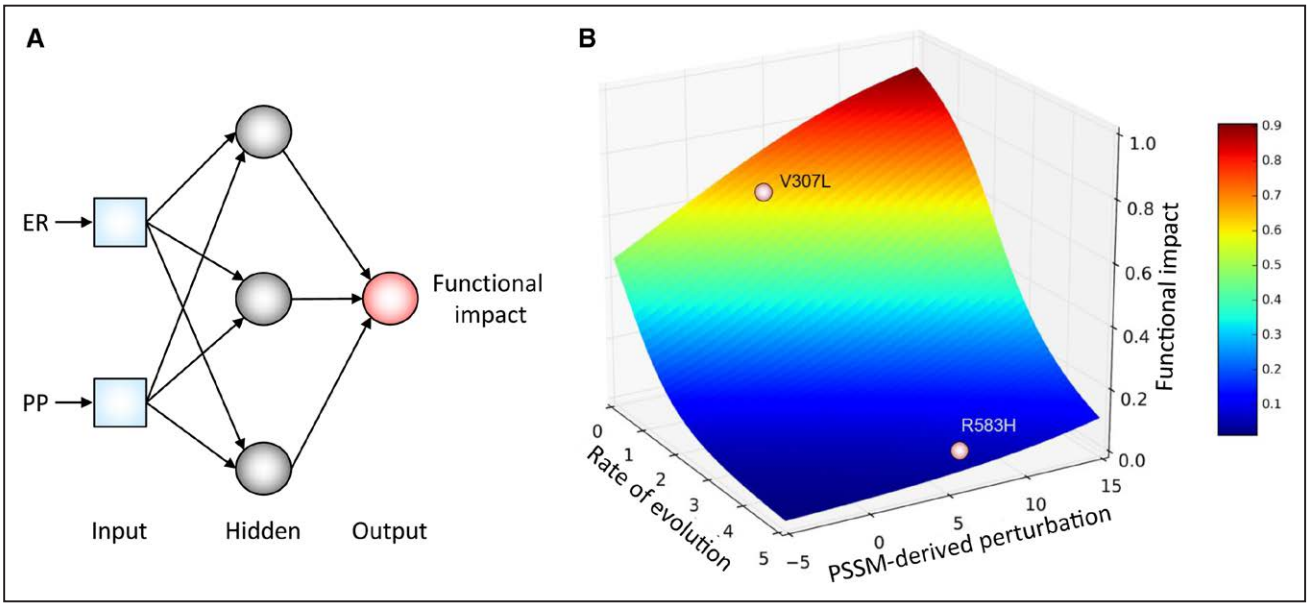


**Figure 3.** **A**, A schematic representation of the architecture of Q1VarPred. The input layer is composed of 2 predictive features: rate of evolution (ER) and perturbation derived from position-specific scoring matrix (PSSM; PP). The hidden layer has 3 neurons and the output layer has 1 neuron that computes the final predicted functional impact. **B**, A visualization of the Q1VarPred-mapped mathematical relationship between predictive features (rate of evolution and perturbation) and functional impact. The vertical axis is functional impact on the scale of 0 to 1 with 1 being complete dysfunction.

**Table. Comparison of Q1VarPred With Other Methods**

| Method | Mean Performance Metric | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AUC | MCC | PPV | NPV | Accuracy | TPR+TNR | TPR | TNR |
| Q1VarPred | 0.884 | 0.581 | 0.968 | 0.537 | 0.881 | 1.680 | 0.895 | 0.785 |
| KvSNP | 0.662 | 0.313 | 0.922 | 0.344 | 0.832 | 1.255 | 0.887 | 0.438 |
| PhD-SNP | 0.727 | 0.386 | 0.941 | 0.390 | 0.820 | 1.453 | 0.850 | 0.603 |
| PolyPhen-2 | 0.636 | 0.340 | 0.912 | 0.547 | 0.866 | 1.272 | 0.939 | 0.333 |
| PredictSNP | 0.652 | 0.355 | 0.918 | 0.459 | 0.850 | 1.303 | 0.912 | 0.391 |
| PROVEAN | 0.770 | 0.510 | 0.949 | 0.537 | 0.869 | 1.536 | 0.902 | 0.634 |
| SIFT | 0.680 | 0.360 | 0.927 | 0.503 | 0.861 | 1.364 | 0.921 | 0.443 |
| SNAP | 0.542 | 0.101 | 0.895 | 0.158 | 0.771 | 1.085 | 0.844 | 0.241 |
| SNPs&GO | 0.697 | 0.307 | 0.939 | 0.296 | 0.767 | 1.384 | 0.792 | 0.592 |

AUC indicates area under the receiver operating characteristic curve; MCC, Matthew's correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; TNR, true negative rate; and TPR, true positive rate.



## Discussion

### From Functional Impact to Clinical Disease Diagnosis

The goal of our study was to create a highly tailored computational method to predict functional impact. However, translating evidence on functional impact to clinical disease diagnosis is far from trivial. First, every computational method has a certain degree of accuracy and reliability, and those of genome-wide methods are particularly limited. In fact, this is one of the primary motivations of the present study. Second, variants that are dysfunctional at the molecular level may not have clinical manifestation. For example, the A300T variant, which was confirmed experimentally to be severely dysfunctional,[25] was later identified in a cohort considered to be clinically normal.[16] Such dysfunctional variants may have been rescued by compensating genetic variations. Third, trying to predict the clinical outcome without considering the mode of inheritance of LQTS may be problematic. The mode of inheritance is a key factor when determining the clinical relevance of a genotype for LQTS. For example, 4 variants in our data set (R231H, W305S, A525T, and R594Q) were functionally normal when expressed in combination with the wild-type channel but were severely dysfunctional in the absence of the wild type. W305S was identified in members of 2 consanguineous families with the recessive Jervell and Lange-Nielsen syndrome,[38] and A525T was suspected to cause the recessive form of Romano-Ward syndrome.[39] Moreover, a functionally normal variant may have compound genetic variations within the same gene or other genes that may obviate or, alternatively, contribute to the clinical phenotype.[40] In light of these considerations, Q1VarPred was intended for judicious use by researchers or clinicians in conjunction with complementary clinical and genetic evidence to assess the disease susceptibility caused by KCNQ1 variants.

### Unexpected Conserved Subdomains in the C-Terminal Domain

Figure 4 shows the topological distributions of position-specific rate of evolution and subdomain-specific enrichment of dysfunctional variants. In our analysis of the rate of evolution in the CTD, we found a few topological subdomains with conserved mean rate of evolution (Figure 1B), predicting important functional or structural roles. Two subdomains, the S6-A linker and the B-C linker, were shown to have a surprisingly low mean rate of evolution (0.88 and 0.24, respectively). While S6-A has an estimated enrichment of dysfunctional variants of 3.0, that of the B-C linker is unexpectedly low (1.0; Figure 2; Table V in the Data Supplement). The low enrichment of the B-C linker is likely biased because of the sparsity of functionally validated variants (eg, only 3 functionally validated variants are located in the B-C linker). In fact, another 6 variants (Table VII in the Data Supplement) found in this subdomain have been deposited in ClinVar.[41] However, they were not included into our data set because we were not able to find literature describing their functional validation. The enrichment of the B-C linker is likely to increase when larger data sets of functionally validated variants become available for estimating enrichments. More importantly, there seems to be a lack of study documenting the functional roles the S6-A linker and the B-C linker. Nevertheless, based on their low rate of evolution, we alert investigators about the potential high functional impact of variants found in these 2 subdomains.

### Machine-Learning Model

Ideally, a machine-learning algorithm should produce a learned model that is accurate at predicting new observations and, at the same time, simple enough to allow straightforward interpretation. In general, linear models are easier to interpret while nonlinear models are more powerful in cases where classes are not linearly separable. We chose a neural network, which generally is considered to be a nonlinear model, for the present study to leverage our extensive experience with neural networks and an established library for feature engineering and model building.[42–46] Admittedly, a logistic regression model performed only slightly worse (AUC=0.855), and a linear discriminant classifier performed comparably (AUC=0.870). However, given the complexity in the mechanisms behind KCNQ1 dysfunction, we expect that the true decision boundary between normal and dysfunctional variants is complex. When additional experimental data become available, the advantage of neural networks for prediction over linear models is likely to become more substantial.

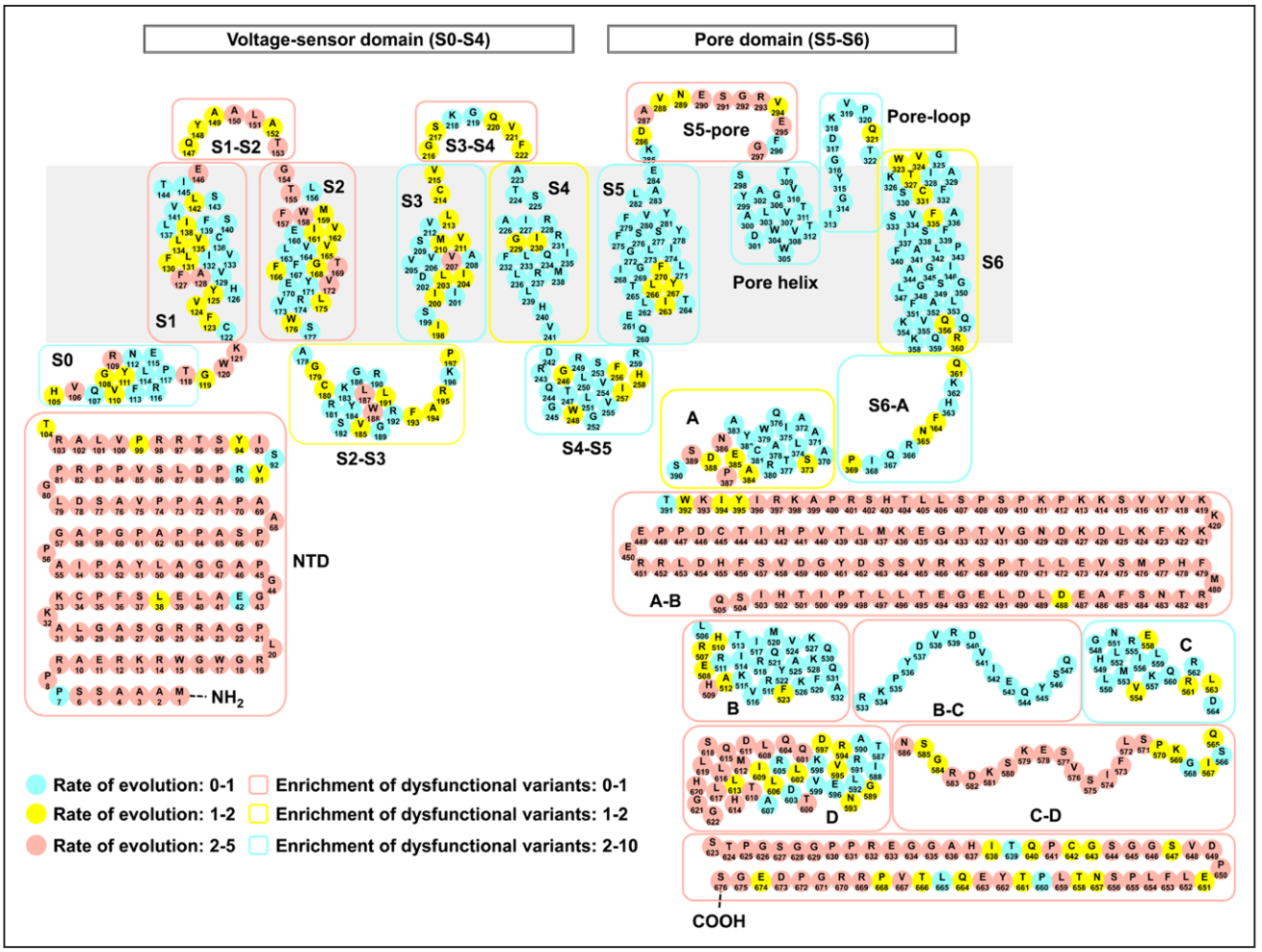


**Figure 4.** A global view of the topological distribution of rate of evolution and enrichment of dysfunctional variants. NTD indicates N-terminal domain.

## Factors Contributing to the Improved Performance of Q1VarPred

Q1VarPred offers improved overall performance in predicting functional impact of variants on a KCNQ1-specific basis compared with the other evaluated tools (Table). Although most tools allow for predictions for a wide range of proteins, the fact that each method applies a single threshold to classify variants on all proteins may be partially responsible for their weaker overall performance on KCNQ1 variants. In addition, recent work has shown that contemporary variant–phenotype and variant–stability prediction algorithms are substantially worse at predicting outcomes for membrane proteins, such as KCNQ1, than for water soluble proteins.[47]

The observed higher performance of Q1VarPred may also be attributed to better predictive features. Many methods use multiple sequence alignment-derived position-specific conservation scores as predictive feature, presumably based on the assumptions that the functional importance of a given position dictates how conserved this position is and, conversely, that the degree of conservation indicates the functional importance of this position. Although this latter assumption is often valid, position-specific conservation scores computed directly from multiple sequence alignment without considering the evolutionary history of the aligned protein family may be biased because of unevenly sampled sequence space. Numerous position-specific quantitative conservation scores have been proposed over the years,[48] and all evaluated methods except the meta-predictor PredictSNP use as position-specific conservation measures of some sort derived from multiple sequence alignment as predictive features. However, none of these methods consider the topology and branch lengths of phylogenetic trees as the method used in the current study does (Methods in the Data Supplement). Thus, these conservation measures may lead to less accurate estimations of rate of evolution.

The other predictive feature used in Q1VarPred is the perturbation derived in the context of a position-specific scoring matrix. This feature measures how much less likely it is for the variant to occur at a sequence position relative to the wild type. The higher the perturbation, the less likely for the variant to replace the wild-type residue at a specific position. Although the position-specific rate of evolution presumably is a strong predictor of functional impact, it only indicates how likely it is that the wild-type amino acid at this sequence position changes. It does not, however, tell how likely it is that the wild-type amino acid is changed to one particular amino acid type over the others. In other words, the perturbation adds

additional information by complementing position-specific rates of evolution with what the actual variants are.

### Limitations and Future Direction

The primary limitation of the current study is the size of the data set. Although a substantial amount of effort was spent by many laboratories to experimentally characterize the 107 variants treated in this study, the data set used in this study is still small, relative to that used to train other contemporary variant-effect predictors. As a result, we were limited from selecting a set of most relevant features in a systematic, algorithmic manner. Thus, it is very likely that we missed some very informative sequence-based features. When larger data sets become available, Q1VarPred can be retrained and new predictive features can be tested. In addition, our estimation of enrichment of dysfunctional variants for each subdomain is also likely to be biased because of this data sparsity. Even though the enrichment values correlate well with average rates of evolution and our analysis shows that functionally important subdomains tend to be more enriched with dysfunctional variants, there is currently not enough data available to demonstrate that such relationship for KCNQ1 is statistically significant.

Recent investigations into machine learning have shown that training neural networks on multiple traits/outcomes per training example can improve performance.[49,50] Specifically, the advantages of simultaneously training a neural network to predict multiple outcome variables (disease severity, electrophysiological parameters, etc.) may enable a more accurate prediction of phenotype traits as well. Previous work aimed at predicting secondary structure and membrane burial for residues has suggested that neural networks trained to predict multiple outcomes are particularly beneficial when the data set size is especially small.[43] This suggests that such neural networks may be particularly suitable to leverage the diverse experimental parameters available for LQTS variants and phenotypes.

The method developed in this study is modular in the sense that one possible future direction is to combine this method with other predictors—such as estimation of the impact of genetic variations on protein stability, to come up with predictions that are both more reliable and that also suggest mechanisms underlying variation-induced gain or loss of function.

## Acknowledgments

Data and software availability: The curated data set is included in the Data Supplement and designated as Table I in the Data Supplement. The data set for training Q1VarPred is provided as a spreadsheet in the Materials in the Data Supplement. Q1VarPred was developed under the framework of the Biochemical Library (available at http://www.meilerlab.org/bclcommons) and is made publicly available as a web server at http://meilerlab.org/q1varpred.

## Sources of Funding

This project was supported by National Institutes of Health (NIH) Grant R01 HL122010 and R01 GM080403. B. Li was also supported by American Heart Association Pre-doctoral Fellowship Award 16PRE27260211. Dr Kroncke was also supported by NIH Grant F32 GM113355. Dr Taylor was also supported by F32 GM11777 and T32 NS00749.

## Disclosures

None.

## References

1. Schwartz PJ, Stramba-Badiale M, Crotti L, Pedrazzini M, Besana A, Bosi G, et al. Prevalence of the congenital long-QT syndrome. *Circulation*. 2009;120:1761–1767. doi: 10.1161/CIRCULATIONAHA.109.863209.
2. Goldenberg I, Moss AJ. Long QT syndrome. *J Am Coll Cardiol*. 2008;51:2291–2300. doi: 10.1016/j.jacc.2008.02.068.
3. Barhanin J, Lesage F, Guillemare E, Fink M, Lazdunski M, Romey G. K(V)LQT1 and lsK (minK) proteins associate to form the I(Ks) cardiac potassium current. *Nature*. 1996;384:78–80. doi: 10.1038/384078a0.
4. Schwartz PJ, Ackerman MJ, George AL Jr, Wilde AA. Impact of genetics on the clinical management of channelopathies. *J Am Coll Cardiol*. 2013;62:169–180. doi: 10.1016/j.jacc.2013.04.044.
5. Giudicessi JR, Ackerman MJ. Genetic testing in heritable cardiac arrhythmia syndromes: differentiating pathogenic mutations from background genetic noise. *Curr Opin Cardiol*. 2013;28:63–71. doi: 10.1097/HCO.0b013e32835b0a41.
6. Ackerman MJ. Genetic purgatory and the cardiac channelopathies: exposing the variants of uncertain/unknown significance issue. *Heart Rhythm*. 2015;12:2325–2331. doi: 10.1016/j.hrthm.2015.07.002.
7. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Hum Mol Genet*. 2010;19:R125–R130. doi: 10.1093/hmg/ddq365.
8. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet*. 2011;12:628–640. doi: 10.1038/nrg3046.
9. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet*. 2013;14:415–426. doi: 10.1038/nrg3493.
10. Bhuiyan ZA. Silent mutation in long QT syndrome: pathogenicity prediction by computer simulation. *Heart Rhythm*. 2012;9:283–284. doi: 10.1016/j.hrthm.2011.10.012.
11. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet*. 2006;7:61–80. doi: 10.1146/annurev.genom.7.080505.115630.
12. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405–424. doi: 10.1038/gim.2015.30.
13. Ohanian M, Otway R, Fatkin D. Heuristic methods for finding pathogenic variants in gene coding sequences. *J Am Heart Assoc*. 2012;1:e002642. doi: 10.1161/JAHA.112.002642.
14. Tchernitchko D, Goossens M, Wajcman H. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. *Clin Chem*. 2004;50:1974–1978. doi: 10.1373/clinchem.2004.036053.
15. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: be mindful of your training data! *Bioinformatics*. 2007;23:664–672. doi: 10.1093/bioinformatics/btl649.
16. Kapa S, Tester DJ, Salisbury BA, Harris-Kerr C, Pungliya MS, Alders M, et al. Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation*. 2009;120:1752–1760. doi: 10.1161/CIRCULATIONAHA.109.863076.
17. Giudicessi JR, Kapplinger JD, Tester DJ, Alders M, Salisbury BA, Wilde AA, et al. Phylogenetic and physicochemical analyses enhance the classification of rare nonsynonymous single nucleotide variants in type 1 and 2 long-QT syndrome. *Circ Cardiovasc Genet*. 2012;5:519–528. doi: 10.1161/CIRCGENETICS.112.963785.
18. Modell SM, Lehmann MH. The long QT syndrome family of cardiac ion channelopathies: a HuGE review. *Genet Med*. 2006;8:143–155. doi: 10.109701.gim.0000204468.85308.86.
19. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15:1929–1958.
20. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. *J Mol Biol*. 1991;222:67–88.
21. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405:442–451.
22. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. 2005;21:3940–3941. doi: 10.1093/bioinformatics/bti623.

23. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2015.
24. Smith GC, Seaman SR, Wood AM, Royston P, White IR. Correcting for optimistic prediction in small data sets. *Am J Epidemiol*. 2014;180:318–324. doi: 10.1093/aje/kwu140.
25. Bianchi L, Priori SG, Napolitano C, Surewicz KA, Dennis AT, Memmi M, et al. Mechanisms of I(Ks) suppression in LQT1 mutants. *Am J Physiol Heart Circ Physiol*. 2000;279:H3003–H3011.
26. Cordeiro JM, Perez GJ, Schmitt N, Pfeiffer R, Nesterenko VV, Burashnikov E, et al. Overlapping LQT1 and LQT2 phenotype in a patient with long QT syndrome associated with loss-of-function variations in KCNQ1 and KCNH2. *Can J Physiol Pharmacol*. 2010;88:1181–1190. doi: 10.1139/Y10-094.
27. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, et al. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*. 1998;280:69–77.
28. Choe S. Potassium channel structures. *Nat Rev Neurosci*. 2002;3:115–121. doi: 10.1038/nrn727.
29. Labro AJ, Boulet IR, Choveau FS, Mayeur E, Bruyns T, Loussouarn G, et al. The S4-S5 linker of KCNQ1 channels forms a structural scaffold with the S6 segment controlling gate closure. *J Biol Chem*. 2011;286:717–725. doi: 10.1074/jbc.M110.146977.
30. Taylor KC, Sanders CR. Regulation of KCNQ/Kv7 family voltage-gated K(+) channels by lipids. *Biochim Biophys Acta*. 2017;1859:586–597. doi: 10.1016/j.bbamem.2016.10.023.
31. Chen L, Zhang Q, Qiu Y, Li Z, Chen Z, Jiang H, et al. Migration of PIP2 lipids on voltage-gated potassium channel surface influences channel deactivation. *Sci Rep*. 2015;5:15079. doi: 10.1038/srep15079.
32. Wiener R, Haitin Y, Shamgar L, Fernández-Alonso MC, Martos A, Chomsky-Hecht O, et al. The KCNQ1 (Kv7.1) COOH terminus, a multitiered scaffold for subunit assembly and protein interaction. *J Biol Chem*. 2008;283:5815–5830. doi: 10.1074/jbc.M707541200.
33. Sachyani D, Dvir M, Strulovich R, Tria G, Tobelaim W, Peretz A, et al. Structural basis of a Kv7.1 potassium channel gating module: studies of the intracellular c-terminal domain in complex with calmodulin. *Structure*. 2014;22:1582–1594. doi: 10.1016/j.str.2014.07.016.
34. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185:862–864.
35. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412–424.
36. Leong IU, Stuckey A, Lai D, Skinner JR, Love DR. Assessment of the predictive accuracy of five in silico prediction tools, alone or in combination, and two metaservers to classify long QT syndrome gene mutations. *BMC Med Genet*. 2015;16:34. doi: 10.1186/s12881-015-0176-z.
37. Stead LF, Wood IC, Westhead DR. KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics*. 2011;27:2181–2186. doi: 10.1093/bioinformatics/btr365.
38. Neyroud N, Denjoy I, Donger C, Gary F, Villain E, Leenhardt A, et al. Heterozygous mutation in the pore of potassium channel gene KvLQT1 causes an apparently normal phenotype in long QT syndrome. *Eur J Hum Genet*. 1998;6:129–133. doi: 10.1038/sj.ejhg.5200165.
39. Larsen LA, Fosdal I, Andersen PS, Kanters JK, Vuust J, Wettrell G, et al. Recessive Romano-Ward syndrome associated with compound heterozygosity for two mutations in the KVLQT1 gene. *Eur J Hum Genet*. 1999;7:724–728. doi: 10.1038/sj.ejhg.5200323.
40. Westenskow P, Splawski I, Timothy KW, Keating MT, Sanguinetti MC. Compound mutations: a common cause of severe long-QT syndrome. *Circulation*. 2004;109:1834–1841. doi: 10.1161/01.CIR.0000125524.34234.13.
41. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42:D980–D985. doi: 10.1093/nar/gkt1113.
42. Butkiewicz M, Lowe EW Jr, Mueller R, Mendenhall JL, Teixeira PL, Weaver CD, et al. Benchmarking ligand-based virtual high-throughput screening with the PubChem database. *Molecules*. 2013;18:735–756. doi: 10.3390/molecules18010735.
43. Leman JK, Mueller R, Karakas M, Woetzel N, Meiler J. Simultaneous prediction of protein secondary structure and transmembrane spans. *Proteins*. 2013;81:1127–1140. doi: 10.1002/prot.24258.
44. Mendenhall J, Meiler J. Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout. *J Comput Aided Mol Des*. 2016;30:177–189. doi: 10.1007/s10822-016-9895-2.
45. Li B, Mendenhall J, Nguyen ED, Weiner BE, Fischer AW, Meiler J. Accurate prediction of contact numbers for multi-spanning helical membrane proteins. *J Chem Inf Model*. 2016;56:423–434.
46. Li B, Mendenhall J, Nguyen ED, Weiner BE, Fischer AW, Meiler J. Improving prediction of helix–helix packing in membrane proteins using predicted contact numbers as restraints. *Proteins*. 2017;85:1212–1221.
47. Kroncke BM, Duran AM, Mendenhall JL, Meiler J, Blume JD, Sanders CR. Documentation of an imperative to improve methods for predicting membrane protein stability. *Biochemistry*. 2016;55:5002–5009. doi: 10.1021/acs.biochem.6b00537.
48. Valdar WS. Scoring residue conservation. *Proteins*. 2002;48:227–241. doi: 10.1002/prot.10146.
49. Qi Y, Oja M, Weston J, Noble WS. A unified multitask architecture for predicting local protein properties. *PLoS One*. 2012;7:e32235. doi: 10.1371/journal.pone.0032235.
50. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep*. 2015;5:11476. doi: 10.1038/srep11476.


## CLINICAL PERSPECTIVE

Congenital long-QT syndrome is a heart rhythm disorder that affects ≈1 in 2500 births. It predisposes children and young adults to a type of ventricular tachycardia (torsades de pointes) and sudden cardiac death. An emerging standard-of-care for long-QT syndrome uses clinical genetic testing to identify long-QT syndrome–associated variants in the KCNQ1 potassium channel. However, variants of unknown significance for which there is inadequate evidence to classify as being pathogenic are common findings. Although computational methods, mostly developed for genome-scale predictions, have been a common predictive approach to suggest genotype–phenotype relations for variants of unknown significance, hardly any is sufficiently accurate for clinical use related to channelopathies. This study presents a KCNQ1-specific genotype–channel function relationship predictor Q1VarPred, which was trained on a data set of KCNQ1 variants whose functional impact has been experimentally validated. Q1VarPred offers substantially improved overall performance in predicting functional impact of variants on a KCNQ1-specific basis compared with the other 8 methods evaluated in the study. It is publicly available as a web server at http://meilerlab.org/q1varpred to ease its access by researchers and clinicians. Along with developing this method, a detailed analysis on the conservation of the amino acid sequence of KCNQ1 showed that dysfunctional variants are enriched in a selected set of highly conserved subdomains. This finding together with the functional impact predicted by Q1VarPred may be considered as supplementary information to the interpretation of variants of known significance.

Circulation
Genomic and Precision Medicine

# Predicting the Functional Impact of KCNQ1 Variants of Unknown Significance

Bian Li, Jeffrey L. Mendenhall, Brett M. Kroncke, Keenan C. Taylor, Hui Huang, Derek K. Smith, Carlos G. Vanoye, Jeffrey D. Blume, Alfred L. George, Jr., Charles R. Sanders and Jens Meiler

*Circ Cardiovasc Genet.* 2017;10:
doi: 10.1161/CIRCGENETICS.117.001754
*Circulation: Cardiovascular Genetics* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1942-325X. Online ISSN: 1942-3268

# SUPPLEMENTAL MATERIAL

## Methods

### Computation of predictive features

Position-specific rate of evolution was estimated using the Rate4Site method.[1] While rates of evolution are commonly measured as the number of substitutions per sequence position per year,[2] it should be noted that the rate estimated by Rate4Site is relative to the average evolutionary rate across all positions and hence is unitless. The input multiple sequence alignment (MSA) of KCNQ1 homologs to Rate4Site was obtained by running HHblits against the Uniprot20 sequence database,[3] with minimum coverage of master sequence (KCNQ1 wild-type sequence) set to 25%, minimum sequence identity to master sequence set to 15%, maximum pairwise sequence identity set to 90%, and E-value cutoff for inclusion in result alignment set to 0.001. The total number of aligned sequences was limited to 300 as our testing showed that Rate4Site suffered from underflow problems when larger numbers of sequences were used.

For characterizing the severity of amino acid substitutions at a position, it is important to conduct the assessment in the context of MSA where the perturbation resulting from amino acid substitution can be quantified from the perspective of protein evolution. We derived this perturbation from the position-specific scoring matrix (PSSM, Figure S1) obtained by searching the NCBI non-redundant sequence database[4] with PSI-BLAST[5] for four iterations. The E-value inclusion threshold was set to 0.00001. For a protein of length $L$, a PSSM is a $L \times 20$ matrix containing log ratios of the estimated frequency of each of the 20 amino acids to occur at each position relative to the expected frequency of the wild-type amino acid in a random sequence. If $P_A$ is the probability for amino acid A to occupy a position and $P_A^0$ is its background probability, then the PSSM entry for A at this position equals $\lambda \ln \frac{P_A}{P_A^0}$, where $\lambda$ is a scaling factor built in PSI-

BLAST.[5] One might recognize that this formula resembles the equation for calculating Gibbs free energy change for a chemical reaction ($\Delta G = -RT \ln K$). Similar in spirit to free energy perturbation, we define the perturbation introduced by amino acid substitution from A to B in the context of MSA as: $\lambda\left(\ln\frac{P_A}{P_A^0} - \ln\frac{P_B}{P_B^0}\right)$. Intuitively, the more substantial the perturbation the less likely it is for a variation to occur without a functional or structural impact.

**Tested genome-wide tools**

Seven genome-wide prediction tools: PhD-SNP,[6] PolyPhen-2,[7] PredictSNP,[8] PROVEAN, SIFT,[9] SNAP[10], and SNPs&GO and a potassium channel-specific method KvSNP[11] were tested for their ability to predict functionality of KCNQ1 variants. PhD-SNP, PolyPhen-2, PredictSNP, SIFT, and SNAP were recently shown to have an overall Matthew's correlation coefficient (MCC) > 0.35 and an overall area under the receiver-operating characteristics curve (AUC) > 0.70 on a fully independent test set consisting of variants from multiple genes.[8] PROVEAN and SNPs&GO were shown to have high accuracy to classify LQTS gene variants. These selected tools differ in the machine learning algorithms with which they were trained and in the required input features. A summary of these tools is presented in Table S4.

**Calculation of enrichment of dysfunctional variants**

Based on a homology model of the homotetrameric transmembrane channel domain,[12] and a structural study of the C-terminal domain of KCNQ1,[13] we mapped the sequence of KCNQ1 into 24 topologically distinct regions and assigned each variant to the region within which it is located (Table S5). The enrichment of dysfunctional variants for a region is computed as the ratio of observed number of dysfunctional variants ($O_v$) to the number of dysfunctional variants that would otherwise be observed if each sequence position were equally likely to raise dysfunctional variants, denoted as $E_v$. $E_v$ can be easily obtained with

$$E_v = \frac{L_s}{L_p} \times N_v$$

where $L_s$ and $L_p$ are the length of the segment and the protein, respectively, and $N_v$ is the total number of dysfunctional variants in the data set.

**Supplemental Tables:**

**Table S1.** Functionally characterized KCNQ1 variants curated from the literature. (See excel file for supplemental table.)

**Table S2**. Performance of the neural network model with varied sizes of hidden layer. (See excel file for supplemental table.)

**Table S3.** Information gain of a set of tested predictive features. (See excel file for supplemental table.)

**Table S4.** Summary of the median and interquartile interval [Q1, Q3] of each performance metric. (See excel file for supplemental table.)

**Table S5**. Topological subdomains of KCNQ1 and the enrichment of dysfunctional variants within each region. (See excel file for supplemental table.)

**Table S6**. Summary of methods evaluated in this study. (See excel file for supplemental table.)

**Table S7.** Six other variants in the B-C linker deposited in the ClinVar database as of June 2017. (See excel file for supplemental table.)

**Supplemental Figure:**

20

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 198 | I | -0.1 | -2.9 | -3.3 | -3.6 | -2.2 | -2.7 | -3.1 | -3.2 | -2.0 | 2.5 | 1.4 | -2.9 | 1.1 | 4.3 | -3.1 | -2.2 | -1.7 | -0.9 | 2.6 | 1.2 |
| 199 | S | -0.8 | -2.2 | -3.5 | -4.2 | 8.9 | -3.5 | -4.0 | -3.7 | -4.0 | -1.8 | 1.7 | -3.5 | -1.6 | -3.0 | -3.9 | 1.1 | -1.9 | -3.8 | -3.4 | -1.2 |
| 200 | I | 0.0 | -1.9 | -2.1 | 0.4 | 2.9 | -1.6 | -1.7 | -2.5 | -2.4 | 0.8 | 0.4 | 0.1 | 1.8 | 1.3 | 1.4 | -1.5 | -1.4 | 4.4 | -0.9 | 0.5 |
| 201 | I | -1.8 | -4.1 | -4.3 | -4.3 | -2.6 | -3.7 | -3.0 | -3.2 | -4.4 | 5.6 | 1.9 | -3.8 | -0.2 | -1.7 | -2.5 | -2.4 | -1.0 | -4.0 | -2.8 | 2.3 |
| 202 | D | -2.6 | -3.1 | -1.0 | 5.8 | -4.3 | -2.0 | 1.3 | 2.4 | -2.8 | -3.6 | -4.2 | -2.4 | -3.8 | -4.4 | -3.2 | -2.0 | -2.6 | -4.7 | -4.1 | -0.8 |
| 203 | L | -1.6 | 0.1 | -2.8 | -3.4 | 1.3 | -2.1 | -2.8 | -0.9 | -2.9 | 2.3 | 2.6 | -0.5 | 2.5 | 0.7 | -3.3 | -0.4 | -1.1 | -2.7 | -1.7 | 0.7 |
| 204 | I | -1.6 | -3.3 | -3.1 | -3.7 | 1.2 | -2.9 | -3.4 | 0.4 | 1.3 | 3.6 | 1.3 | -3.1 | 1.6 | 0.7 | 0.6 | -2.5 | -1.3 | -2.6 | 0.8 | 1.0 |
| 205 | V | -1.2 | 0.8 | -3.3 | -3.8 | 2.8 | -2.6 | -3.1 | -1.7 | -2.8 | 1.5 | -0.6 | -2.3 | 0.7 | -1.6 | 0.7 | -2.2 | -0.1 | -2.9 | 1.8 | 3.9 |
| 206 | V | -1.5 | -0.5 | -2.6 | -3.2 | -2.3 | -0.3 | -2.5 | -0.2 | -2.4 | 2.7 | 0.9 | -2.1 | -0.2 | 1.8 | -3.1 | -0.8 | 0.3 | -2.2 | 1.1 | 2.1 |
| 207 | V | -0.7 | -2.5 | -0.4 | -2.6 | -0.2 | -2.2 | -2.5 | -1.0 | -2.7 | 2.5 | 1.5 | -2.3 | 1.0 | -0.5 | -2.5 | -0.7 | 0.1 | -2.7 | -1.7 | 3.0 |
| 208 | A | 3.3 | 0.2 | -1.7 | -2.1 | -1.7 | 0.9 | -1.3 | -1.4 | 0.7 | -1.5 | -1.0 | -1.1 | -1.3 | 0.7 | 0.9 | 0.5 | -0.6 | -2.4 | -1.4 | 0.1 |
| 209 | S | 0.5 | -2.1 | 1.2 | -1.6 | -2.3 | -1.5 | -1.7 | 1.9 | -1.7 | -3.2 | -3.2 | -1.6 | -2.6 | -2.5 | -2.1 | 4.5 | 0.2 | -2.8 | 1.5 | -2.7 |
| 210 | M | -1.1 | -2.8 | -2.8 | -3.1 | -1.9 | -2.5 | -2.8 | -3.1 | -2.8 | 3.4 | 1.6 | -2.6 | 1.0 | 2.6 | -2.8 | -0.4 | -0.8 | 0.6 | -0.7 | 2.3 |
| 211 | V | 0.1 | -1.5 | 0.7 | -1.8 | 2.5 | 0.7 | -1.4 | 0.1 | -1.4 | 0.1 | -1.6 | 0.2 | -1.3 | 0.9 | 1.2 | 0.3 | 0.5 | -2.0 | 1.3 | 0.1 |
| 212 | V | -0.6 | -1.5 | -3.0 | -3.3 | -2.1 | -2.5 | -2.8 | -1.5 | -3.3 | 2.5 | 1.0 | -2.5 | 1.3 | -1.8 | 0.6 | -2.1 | -0.2 | -3.2 | -2.3 | 4.0 |
| 213 | L | -2.2 | -2.6 | -3.4 | -3.6 | -2.5 | 0.3 | -2.8 | -4.0 | -3.3 | 3.2 | 3.5 | -0.8 | 0.1 | 0.0 | -3.3 | -2.7 | -0.5 | -3.0 | -2.1 | 1.7 |
| 214 | C | 2.7 | -2.8 | -2.3 | -2.9 | 4.6 | -2.3 | -2.5 | 0.4 | -2.5 | -0.9 | -0.9 | -2.3 | 0.9 | 1.1 | -2.6 | 1.3 | 0.4 | -2.6 | 0.6 | -0.4 |
| 215 | V | 0.9 | -2.3 | -0.7 | -2.9 | -0.4 | -0.9 | -2.3 | -2.0 | -2.6 | 0.8 | 0.2 | -0.5 | 1.7 | 2.8 | -0.2 | 0.3 | -1.3 | -0.2 | -0.9 | 1.9 |

L

**Figure S1.** An illustration of position-specific scoring matrix (PSSM). For a protein of length $L$, a PSSM is a $L \times 20$ matrix containing log ratios of the estimated frequency of each of the 20 amino acids to occur at each position relative to the expected frequency of the wild-type amino acid in a random sequence. If $P_A$ is the probability for amino acid A to occupy a position and $P_A^0$ is its background probability, then the PSSM entry for A at this position equals $\lambda \ln \frac{P_A}{P_A^0}$, where $\lambda$ is a scaling factor built in PSI-BLAST.[5]

**References:**


1. Pupko, T.; Bell, R. E.; Mayrose, I.; Glaser, F.; Ben-Tal, N., Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **2002,** *18 Suppl 1*, S71-7.

2. Lanfear, R.; Kokko, H.; Eyre-Walker, A., Population size and the rate of evolution. *Trends Ecol Evol* **2014,** *29* (1), 33-41.

3. Remmert, M.; Biegert, A.; Hauser, A.; Söding, J., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **2012,** *9* (2), 173-5.

4. Pruitt, K. D.; Tatusova, T.; Maglott, D. R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **2007,** *35* (Database issue), D61-5.

5. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **1997,** *25* (17), 3389-402.

6. Capriotti, E.; Calabrese, R.; Casadio, R., Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **2006,** *22* (22), 2729-34.

7. Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R., A method and server for predicting damaging missense mutations. *Nature methods* **2010,** *7* (4), 248-9.

8. Bendl, J.; Stourac, J.; Salanda, O.; Pavelka, A.; Wieben, E. D.; Zendulka, J.; Brezovsky, J.; Damborsky, J., PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS computational biology* **2014,** *10* (1), e1003440.

9. Ng, P. C.; Henikoff, S., Predicting deleterious amino acid substitutions. *Genome Res* **2001,** *11* (5), 863-74.

10. Bromberg, Y.; Rost, B., SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **2007,** *35* (11), 3823-35.

11. Stead, L. F.; Wood, I. C.; Westhead, D. R., KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics* **2011,** *27* (16), 2181-2186.

12. Smith, J. A.; Vanoye, C. G.; George, A. L., Jr.; Meiler, J.; Sanders, C. R., Structural models for the KCNQ1 voltage-gated potassium channel. *Biochemistry* **2007,** *46* (49), 14141-52.

13. Wiener, R.; Haitin, Y.; Shamgar, L.; Fernández-Alonso, M. C.; Martos, A.; Chomsky-Hecht, O.; Rivas, G.; Attali, B.; Hirsch, J. A., The KCNQ1 (Kv7.1) COOH terminus, a multitiered scaffold for subunit assembly and protein interaction. *J Biol Chem* **2008,** *283* (9), 5815-30.

14. Yang, T.; Chung, S. K.; Zhang, W.; Mullins, J. G.; McCulley, C. H.; Crawford, J.; MacCormick, J.; Eddy, C. A.; Shelling, A. N.; French, J. K.; Yang, P.; Skinner, J. R.; Roden, D. M.; Rees, M. I., Biophysical properties of 9 KCNQ1 mutations associated with long-QT syndrome. *Circ Arrhythm Electrophysiol* **2009,** *2* (4), 417-26.

15. Cordeiro, J. M.; Perez, G. J.; Schmitt, N.; Pfeiffer, R.; Nesterenko, V. V.; Burashnikov, E.; Veltmann, C.; Borggrefe, M.; Wolpert, C.; Schimpf, R.; Antzelevitch, C., Overlapping LQT1 and LQT2 phenotype in a patient with long QT syndrome associated with loss-of-function variations in KCNQ1 and KCNH2. *Can J Physiol Pharmacol* **2010,** *88* (12), 1181-90.

16. Dahimène, S.; Alcoléa, S.; Naud, P.; Jourdon, P.; Escande, D.; Brasseur, R.; Thomas, A.; Baró, I.; Mérot, J., The N-terminal juxtamembranous domain of KCNQ1 is critical for channel surface expression: implications in the Romano-Ward LQT1 syndrome. *Circ Res* **2006,** *99* (10), 1076-83.

17. Campbell, C. M.; Campbell, J. D.; Thompson, C. H.; Galimberti, E. S.; Darbar, D.; Vanoye, C. G.; George, A. L., Jr., Selective targeting of gain-of-function KCNQ1 mutations predisposing to atrial fibrillation. *Circ Arrhythm Electrophysiol* **2013,** *6* (5), 960-6.

18. Hong, K.; Piper, D. R.; Diaz-Valdecantos, A.; Brugada, J.; Oliva, A.; Burashnikov, E.; Santos-de-Soto, J.; Grueso-Montero, J.; Diaz-Enfante, E.; Brugada, P.; Sachse, F.; Sanguinetti, M. C.; Brugada, R., De novo KCNQ1 mutation responsible for atrial fibrillation and short QT syndrome in utero. *Cardiovasc Res* **2005,** *68* (3), 433-40.

19. Lundby, A.; Ravn, L. S.; Svendsen, J. H.; Olesen, S. P.; Schmitt, N., KCNQ1 mutation Q147R is associated with atrial fibrillation and prolonged QT interval. *Heart Rhythm* **2007,** *4* (12), 1532-41.

20. Westenskow, P.; Splawski, I.; Timothy, K. W.; Keating, M. T.; Sanguinetti, M. C., Compound mutations: a common cause of severe long-QT syndrome. *Circulation* **2004,** *109* (15), 1834-41.

21. Matavel, A.; Medei, E.; Lopes, C. M., PKA and PKC partially rescue long QT type 1 phenotype by restoring channel-PIP2 interactions. *Channels (Austin)* **2010,** *4* (1), 3-11.

22. Harmer, S. C.; Mohal, J. S.; Royal, A. A.; McKenna, W. J.; Lambiase, P. D.; Tinker, A., Cellular mechanisms underlying the increased disease severity seen for patients with long QT syndrome caused by compound mutations in KCNQ1. *Biochem J* **2014,** *462* (1), 133-42.

23. Chouabe, C.; Neyroud, N.; Richard, P.; Denjoy, I.; Hainque, B.; Romey, G.; Drici, M. D.; Guicheney, P.; Barhanin, J., Novel mutations in KvLQT1 that affect Iks activation through interactions with Isk. *Cardiovasc Res* **2000,** *45* (4), 971-80.

24. Pan, N.; Sun, J.; Lv, C.; Li, H.; Ding, J., A hydrophobicity-dependent motif responsible for surface expression of cardiac potassium channel. *Cell Signal* **2009,** *21* (2), 349-55.

25. Yamaguchi, M.; Shimizu, M.; Ino, H.; Terai, H.; Hayashi, K.; Mabuchi, H.; Hoshi, N.; Higashida, H., Clinical and electrophysiological characterization of a novel mutation (F193L) in the KCNQ1 gene associated with long QT syndrome. *Clin Sci (Lond)* **2003,** *104* (4), 377-82.

26. Eldstrom, J.; Xu, H.; Werry, D.; Kang, C.; Loewen, M. E.; Degenhardt, A.; Sanatani, S.; Tibbits, G. F.; Sanders, C.; Fedida, D., Mechanistic basis for LQT1 caused by S3 mutations in the KCNQ1 subunit of IKs. *J Gen Physiol* **2010,** *135* (5), 433-48.

27. Eldstrom, J.; Wang, Z.; Werry, D.; Wong, N.; Fedida, D., Microscopic mechanisms for long QT syndrome type 1 revealed by single-channel analysis of I(Ks) with S3 domain mutations in KCNQ1. *Heart Rhythm* **2015,** *12* (2), 386-94.

28. Das, S.; Makino, S.; Melman, Y. F.; Shea, M. A.; Goyal, S. B.; Rosenzweig, A.; Macrae, C. A.; Ellinor, P. T., Mutation in the S3 segment of KCNQ1 results in familial lone atrial fibrillation. *Heart Rhythm* **2009,** *6* (8), 1146-53.

29. Bianchi, L.; Priori, S. G.; Napolitano, C.; Surewicz, K. A.; Dennis, A. T.; Memmi, M.; Schwartz, P. J.; Brown, A. M., Mechanisms of I(Ks) suppression in LQT1 mutants. *Am J Physiol Heart Circ Physiol* **2000,** *279* (6), H3003-11.

30. Itoh, H.; Sakaguchi, T.; Ding, W. G.; Watanabe, E.; Watanabe, I.; Nishio, Y.; Makiyama, T.; Ohno, S.; Akao, M.; Higashi, Y.; Zenda, N.; Kubota, T.; Mori, C.; Okajima, K.; Haruna, T.; Miyamoto, A.; Kawamura, M.; Ishida, K.; Nagaoka, I.; Oka, Y.; Nakazawa, Y.; Yao, T.; Jo, H.; Sugimoto, Y.; Ashihara, T.; Hayashi, H.; Ito, M.; Imoto, K.; Matsuura, H.; Horie, M., Latent genetic backgrounds and molecular pathogenesis in drug-induced long-QT syndrome. *Circ Arrhythm Electrophysiol* **2009,** *2* (5), 511-23.

31. Bartos, D. C.; Giudicessi, J. R.; Tester, D. J.; Ackerman, M. J.; Ohno, S.; Horie, M.; Gollob, M. H.; Burgess, D. E.; Delisle, B. P., A KCNQ1 mutation contributes to the concealed type 1 long QT phenotype by limiting the Kv7.1 channel conformational changes associated with protein kinase A phosphorylation. *Heart Rhythm* **2014,** *11* (3), 459-68.

32. Steffensen, A. B.; Refaat, M. M.; David, J. P.; Mujezinovic, A.; Calloe, K.; Wojciak, J.; Nussbaum, R. L.; Scheinman, M. M.; Schmitt, N., High incidence of functional ion-channel abnormalities in a consecutive Long QT cohort with novel missense genetic variants of unknown significance. *Sci Rep* **2015,** *5* (doi), 10009.

33. Huang, L.; Bitner-Glindzicz, M.; Tranebjaerg, L.; Tinker, A., A spectrum of functional effects for disease causing mutations in the Jervell and Lange-Nielsen syndrome. *Cardiovasc Res* **2001,** *51* (4), 670-80.

34. Franqueza, L.; Lin, M.; Shen, J.; Splawski, I.; Keating, M. T.; Sanguinetti, M. C., Long QT syndrome-associated mutations in the S4-S5 linker of KvLQT1 potassium channels modify gating and interaction with minK subunits. *J Biol Chem* **1999,** *274* (30), 21063-70.

35. Deschenes, D.; Acharfi, S.; Pouliot, V.; Hegele, R.; Krahn, A.; Daleau, P.; Chahine, M., Biophysical characteristics of a new mutation on the KCNQ1 potassium channel (L251P) causing long QT syndrome. *Can J Physiol Pharmacol* **2003,** *81* (2), 129-34.

36. Wedekind, H.; Schwarz, M.; Hauenschild, S.; Djonlagic, H.; Haverkamp, W.; Breithardt, G.; Wulfing, T.; Pongs, O.; Isbrandt, D.; Schulze-Bahr, E., Effective long-term control of cardiac events with beta-blockers in a family with a common LQT1 mutation. *Clin Genet* **2004,** *65* (3), 233-41.

37. Labro, A. J.; Boulet, I. R.; Timmermans, J. P.; Ottschytsch, N.; Snyders, D. J., The rate-dependent biophysical properties of the LQT1 H258R mutant are counteracted by a dominant negative effect on channel trafficking. *J Mol Cell Cardiol* **2010,** *48* (6), 1096-104.

38. Kubota, T.; Shimizu, W.; Kamakura, S.; Horie, M., Hypokalemia-induced long QT syndrome with an underlying novel missense mutation in S4-S5 linker of KCNQ1. *J Cardiovasc Electrophysiol* **2000,** *11* (9), 1048-54.

39. Wu, Z. J.; Huang, Y.; Fu, Y. C.; Zhao, X. J.; Zhu, C.; Zhang, Y.; Xu, B.; Zhu, Q. L.; Li, Y., Characterization of a Chinese KCNQ1 mutation (R259H) that shortens repolarization and causes short QT syndrome 2. *J Geriatr Cardiol* **2015,** *12* (4), 394-401.

40. Chouabe, C.; Neyroud, N.; Guicheney, P.; Lazdunski, M.; Romey, G.; Barhanin, J., Properties of KvLQT1 K+ channel mutations in Romano-Ward and Jervell and Lange-Nielsen inherited cardiac arrhythmias. *The EMBO journal* **1997,** *16* (17), 5472-9.

41. Wu, J.; Naiki, N.; Ding, W. G.; Ohno, S.; Kato, K.; Zang, W. J.; Delisle, B. P.; Matsuura, H.; Horie, M., A molecular mechanism for adrenergic-induced long QT syndrome. *J Am Coll Cardiol* **2014,** *63* (8), 819-27.

42. Oka, Y.; Itoh, H.; Ding, W. G.; Shimizu, W.; Makiyama, T.; Ohno, S.; Nishio, Y.; Sakaguchi, T.; Miyamoto, A.; Kawamura, M.; Matsuura, H.; Horie, M., Atrioventricular block-induced Torsades de Pointes with clinical and molecular backgrounds similar to congenital long QT syndrome. *Circ J* **2010,** *74* (12), 2562-71.

43. Li, W.; Wang, Q. F.; Du, R.; Xu, Q. M.; Ke, Q. M.; Wang, B.; Chen, X. L.; Tian, L.; Zhang, S. Y.; Kang, C. L.; Guan, S. M.; Yang, J. G.; Song, Z. F., Congenital long QT syndrome caused by the F275S KCNQ1 mutation: mechanism of impaired channel function. *Biochem Biophys Res Commun* **2009,** *380* (1), 127-31.

44. Aidery, P.; Kisselbach, J.; Schweizer, P. A.; Becker, R.; Katus, H. A.; Thomas, D., Biophysical properties of mutant KCNQ1 S277L channels linked to hereditary long QT syndrome with phenotypic variability. *Biochimica et biophysica acta* **2011,** *1812* (4), 488-94.

45. Moreno, C.; Oliveras, A.; de la Cruz, A.; Bartolucci, C.; Munoz, C.; Salar, E.; Gimeno, J. R.; Severi, S.; Comes, N.; Felipe, A.; Gonzalez, T.; Lambiase, P.; Valenzuela, C., A new KCNQ1 mutation at the S5 segment that impairs its association with KCNE1 is responsible for short QT syndrome. *Cardiovasc Res* **2015,** *107* (4), 613-23.

46. Crotti, L.; Tester, D. J.; White, W. M.; Bartos, D. C.; Insolia, R.; Besana, A.; Kunic, J. D.; Will, M. L.; Velasco, E. J.; Bair, J. J.; Ghidoni, A.; Cetin, I.; Van Dyke, D. L.; Wick, M. J.; Brost, B.; Delisle, B. P.; Facchinetti, F.; George, A. L.; Schwartz, P. J.; Ackerman, M. J., Long QT syndrome-associated mutations in intrauterine fetal death. *JAMA* **2013,** *309* (14), 1473-82.

47. Bellocq, C.; van Ginneken, A. C.; Bezzina, C. R.; Alders, M.; Escande, D.; Mannens, M. M.; Baro, I.; Wilde, A. A., Mutation in the KCNQ1 gene leading to the short QT-interval syndrome. *Circulation* **2004,** *109* (20), 2394-7.

48. Ikrar, T.; Hanawa, H.; Watanabe, H.; Aizawa, Y.; Ramadan, M. M.; Chinushi, M.; Horie, M.; Aizawa, Y., Evaluation of channel function after alteration of amino acid residues at the pore center of KCNQ1 channel. *Biochem Biophys Res Commun* **2009,** *378* (3), 589-94.

49. Li, W.; Du, R.; Wang, Q. F.; Tian, L.; Yang, J. G.; Song, Z. F., The G314S KCNQ1 mutation exerts a dominant-negative effect on expression of KCNQ1 channels in oocytes. *Biochem Biophys Res Commun* **2009,** *383* (2), 206-9.

50. Thomas, D.; Khalil, M.; Alter, M.; Schweizer, P. A.; Karle, C. A.; Wimmer, A. B.; Licka, M.; Katus, H. A.; Koenen, M.; Ulmer, H. E.; Zehelein, J., Biophysical characterization of KCNQ1 P320 mutations linked to long QT syndrome 1. *J Mol Cell Cardiol* **2010,** *48* (1), 230-7.

51. Burgess, D. E.; Bartos, D. C.; Reloj, A. R.; Campbell, K. S.; Johnson, J. N.; Tester, D. J.; Ackerman, M. J.; Fressart, V.; Denjoy, I.; Guicheney, P.; Moss, A. J.; Ohno, S.; Horie, M.; Delisle, B. P., High-risk long QT syndrome mutations in the Kv7.1 (KCNQ1) pore disrupt the molecular basis for rapid K(+) permeation. *Biochemistry* **2012,** *51* (45), 9076-85.

52. Aidery, P.; Kisselbach, J.; Schweizer, P. A.; Becker, R.; Katus, H. A.; Thomas, D., Impaired ion channel function related to a common KCNQ1 mutation - implications for risk stratification in long QT syndrome 1. *Gene* **2012,** *511* (1), 26-33.

53. Hoosien, M.; Ahearn, M. E.; Myerburg, R. J.; Pham, T. V.; Miller, T. E.; Smets, M. J.; Baumbach-Reardon, L.; Young, M. L.; Farooq, A.; Bishopric, N. H., Dysfunctional potassium channel subunit interaction as a novel mechanism of long QT syndrome. *Heart Rhythm* **2013,** *10* (5), 728-37.

54. Heijman, J.; Spatjens, R. L.; Seyen, S. R.; Lentink, V.; Kuijpers, H. J.; Boulet, I. R.; de Windt, L. J.; David, M.; Volders, P. G., Dominant-negative control of cAMP-dependent IKs upregulation in human long-QT syndrome type 1. *Circ Res* **2012,** *110* (2), 211-9.

55. Zehelein, J.; Thomas, D.; Khalil, M.; Wimmer, A. B.; Koenen, M.; Licka, M.; Wu, K.; Kiehn, J.; Brockmeier, K.; Kreye, V. A.; Karle, C. A.; Katus, H. A.; Ulmer, H. E.; Schoels, W., Identification and characterisation of a novel KCNQ1 mutation in a family with Romano-Ward syndrome. *Biochimica et biophysica acta* **2004,** *1690* (3), 185-92.

56. Siebrands, C. C.; Binder, S.; Eckhoff, U.; Schmitt, N.; Friederich, P., Long QT 1 mutation KCNQ1A344V increases local anesthetic sensitivity of the slowly activating delayed rectifier potassium current. *Anesthesiology* **2006,** *105* (3), 511-20.

57. Boulet, I. R.; Raes, A. L.; Ottschytsch, N.; Snyders, D. J., Functional effects of a KCNQ1 mutation associated with the long QT syndrome. *Cardiovasc Res* **2006,** *70* (3), 466-74.

58. Shamgar, L.; Ma, L.; Schmitt, N.; Haitin, Y.; Peretz, A.; Wiener, R.; Hirsch, J.; Pongs, O.; Attali, B., Calmodulin is essential for cardiac IKS channel gating and assembly: impaired function in long-QT mutations. *Circ Res* **2006,** *98* (8), 1055-63.

59. Li, Y.; Gao, J.; Lu, Z.; McFarland, K.; Shi, J.; Bock, K.; Cohen, I. S.; Cui, J., Intracellular ATP binding is required to activate the slowly activating K+ channel I(Ks). *Proc Natl Acad Sci U S A* **2013,** *110* (47), 18922-7.

60. Xiong, Q.; Cao, Q.; Zhou, Q.; Xie, J.; Shen, Y.; Wan, R.; Yu, J.; Yan, S.; Marian, A. J.; Hong, K., Arrhythmogenic cardiomyopathy in a patient with a rare loss-of-function KCNQ1 mutation. *J Am Heart Assoc* **2015,** *4* (1), e001526.

61. Schmitt, N.; Calloe, K.; Nielsen, N. H.; Buschmann, M.; Speckmann, E. J.; Schulze-Bahr, E.; Schwarz, M., The novel C-terminal KCNQ1 mutation M520R alters protein trafficking. *Biochem Biophys Res Commun* **2007,** *358* (1), 304-10.

62. Dvir, M.; Strulovich, R.; Sachyani, D.; Ben-Tal Cohen, I.; Haitin, Y.; Dessauer, C.; Pongs, O.; Kass, R.; Hirsch, J. A.; Attali, B., Long QT mutations at the interface between KCNQ1 helix C and KCNE1 disrupt I(KS) regulation by PKA and PIP(2). *J Cell Sci* **2014,** *127* (Pt 18), 3943-55.

63. Aromolaran, A. S.; Subramanyam, P.; Chang, D. D.; Kobertz, W. R.; Colecraft, H. M., LQT1 mutations in KCNQ1 C-terminus assembly domain suppress IKs using different mechanisms. *Cardiovasc Res* **2014,** *104* (3), 501-11.

64. Spatjens, R. L.; Bebarova, M.; Seyen, S. R.; Lentink, V.; Jongbloed, R. J.; Arens, Y. H.; Heijman, J.; Volders, P. G., Long-QT mutation p.K557E-Kv7.1: dominant-negative suppression of IKs, but preserved cAMP-dependent up-regulation. *Cardiovasc Res* **2014,** *104* (1), 216-25.

65. Detta, N. Molecular Basis of Cardiac Arrhythmias: Genetics of Natural Variants and Electrophysiological Investigation of Mutant Proteins. University of Napoli Federico II, Napoli, 2010.

66. Piippo, K.; Swan, H.; Pasternack, M.; Chapman, H.; Paavonen, K.; Viitasalo, M.; Toivonen, L.; Kontula, K., A founder mutation of the potassium channel KCNQ1 in long QT syndrome: implications for estimation of disease prevalence and molecular diagnostics. *J Am Coll Cardiol* **2001,** *37* (2), 562-8.

67. Kinoshita, K.; Komatsu, T.; Nishide, K.; Hata, Y.; Hisajima, N.; Takahashi, H.; Kimoto, K.; Aonuma, K.; Tsushima, E.; Tabata, T.; Yoshida, T.; Mori, H.; Nishida, K.; Yamaguchi, Y.; Ichida, F.; Fukurotani, K.; Inoue, H.; Nishida, N., A590T mutation in KCNQ1 C-terminal helix D decreases IKs channel trafficking and function but not Yotiao interaction. *J Mol Cell Cardiol* **2014,** *72*, 273-80.

68. Yamaguchi, M.; Shimizu, M.; Ino, H.; Terai, H.; Hayashi, K.; Kaneda, T.; Mabuchi, H.; Sumita, R.; Oshima, T.; Hoshi, N.; Higashida, H., Compound heterozygosity for mutations Asp611-->Tyr in KCNQ1 and Asp609-->Gly in KCNH2 associated with severe long QT syndrome. *Clin Sci (Lond)* **2005,** *108* (2), 143-50.

69. Kubota, T.; Horie, M.; Takano, M.; Yoshida, H.; Takenaka, K.; Watanabe, E.; Tsuchiya, T.; Otani, H.; Sasayama, S., Evidence for a single nucleotide polymorphism in the KCNQ1 potassium channel that underlies susceptibility to life-threatening arrhythmias. *J Cardiovasc Electrophysiol* **2001,** *12* (11), 1223-9.

70. Li, B.; Mendenhall, J.; Nguyen, E. D.; Weiner, B. E.; Fischer, A. W.; Meiler, J., Accurate Prediction of Contact Numbers for Multi-Spanning Helical Membrane Proteins. *J Chem Inf Model* **2016,** *56* (2), 423-34.

71. Grantham, R., Amino acid difference formula to help explain protein evolution. *Science* **1974,** *185* (4154), 862-4.

72. Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R., A method and server for predicting damaging missense mutations. *Nature methods* **2010,** *7* (4), 248-9.

73. Bendl, J.; Stourac, J.; Salanda, O.; Pavelka, A.; Wieben, E. D.; Zendulka, J.; Brezovsky, J.; Damborsky, J., PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS computational biology* **2014,** *10* (1), e1003440.

74. Ng, P. C.; Henikoff, S., Predicting deleterious amino acid substitutions. *Genome research* **2001,** *11* (5), 863-74.

**Table S1**

| | Wild | Variant | Clinical | $I_{ks}$ ratio | $V_{1/2}$ (mV) | Activation τ ratio | Deactivation τ ratio | | Surface | Annotation | Label | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 46 | A | T | Case | 100% | 0 | 0.6 | | | | Normal | Normal | [14] |
| 110 | V | I | Control | 40% | 30 | | | | Normal | Severe LOF | Dysfunctional | [15] |
| 111 | Y | C | Case | 0% | | | | | Absent | Severe LOF | Dysfunctional | [16] |
| 114 | L | P | Case | 0% | | | | | Absent | Severe LOF | Dysfunctional | [16] |
| 117 | P | L | Case | 0% | | | | | Impaired | Severe LOF | Dysfunctional | [16] |
| 140 | S | G | Case | 150% | | | | | | Severe GOF | Dysfunctional | [17] |
| 141 | V | M | Case | 300% | 0 | | | | | Severe GOF | Dysfunctional | [18] |
| 147 | Q | R | Case | 60% | 0 | | | | | Mild LOF | Normal | [19] |
| 168 | G | R | Case | 5% | | | | | | Severe LOF | Dysfunctional | [20] |
| 174 | R | C | Case | 47% | 17 | 1 | | | | Mild LOF | Dysfunctional | [21] |
| 178 | A | T | Case | 41% | 45 | 1.68 | | 0.86 | Impaired | Severe LOF | Dysfunctional | [22] |
| 179 | G | S | Control | 54% | -12 | | | | | Mild LOF | Normal | [20] |
| 190 | R | Q | Case | 0% | | | | | | Severe LOF | Dysfunctional | [23] |
| 191 | L | P | Case | 22% | 0 | | | | Impaired | Severe LOF | Dysfunctional | [24] |
| 193 | F | L | Case | 80% | 0 | 1.83 | | | | Severe LOF | Dysfunctional | [25] |
| 202 | D | E | Case | 11% | 54.6 | 1 | | 0.33 | | Severe LOF | Dysfunctional | [26] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 202 | D | H | Case | 41% | 16.6 | 0.83 | 0.26 | Normal | Severe LOF | Dysfunctional | [26] |
| 202 | D | N | Case | 20% | 23.8 | 0.55 | 0.09 | Normal | Severe LOF | Dysfunctional | [26] |
| 204 | I | F | Case | 23% | 53.3 | 7.25 | 0.43 | Normal | Severe LOF | Dysfunctional | [26] |
| 204 | I | M | Case | 34% | 36.1 | 1.16 | 0.65 | Normal | Severe LOF | Dysfunctional | [26] |
| 204 | I | N | Case | | 32.9 | 2.47 | 0.7 | | Severe LOF | Dysfunctional | [26] |
| 205 | V | M | Case | 36% | 20 | 1.48 | 0.42 | | Severe LOF | Dysfunctional | [27] |
| 207 | V | M | Control | 93% | 7.1 | 1.4 | 1.2 | | Normal | Normal | [26] |
| 209 | S | F | Case | 35% | -48.7 | | | | Severe LOF | Dysfunctional | [26] |
| 209 | S | P | Case | 200% | -42.4 | | 5.7 | | Severe GOF | Dysfunctional | [28] |
| 215 | V | M | Case | 41% | 20.2 | | | | Severe LOF | Dysfunctional | [26] |
| 225 | S | L | Case | 10% | 11 | | | Normal | Severe LOF | Dysfunctional | [29] |
| 231 | R | C | Case | 5% | | | | | Severe LOF | Dysfunctional | [30] |
| 231 | R | H | Case | 15% | 40 | | | | Severe LOF | Dysfunctional | [30] |
| 235 | I | N | Case | 10% | | | | | Severe LOF | Dysfunctional | [31] |
| 236 | L | R | Case | 0% | 54 | | | Impaired | Severe LOF | Dysfunctional | [32] |
| 243 | R | C | Case | 12% | 67 | 1 | | | Severe LOF | Dysfunctional | [21] |
| 243 | R | H | Case | 13% | | | | Normal | Severe LOF | Dysfunctional | [33] |
| 248 | W | R | Case | 0% | | | | | Severe LOF | Dysfunctional | [34] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 251 | L | P | Case | 0% | | | | Normal | Severe LOF | Dysfunctional | [35] |
| 254 | V | M | Case | 7% | 41.5 | | | | Severe LOF | Dysfunctional | [36] |
| 258 | H | R | Case | 5% | -44 | 0.5 | 2.5 | Impaired | Severe LOF | Dysfunctional | [37] |
| 259 | R | C | Case | 30% | 10 | | | | Severe LOF | Dysfunctional | [38] |
| 259 | R | H | Case | 200% | 1 | | 1.7 | Normal | Severe GOF | Dysfunctional | [39] |
| 261 | E | D | Case | 9% | | | | | Severe LOF | Dysfunctional | [33] |
| 261 | E | K | Case | 5% | | | | | Severe LOF | Dysfunctional | [34] |
| 265 | T | I | Case | 100% | 8 | 2 | | | Severe LOF | Dysfunctional | [14] |
| 269 | G | D | Case | 0% | | | | | Severe LOF | Dysfunctional | [40] |
| 269 | G | S | Case | 15% | 70.7 | 1 | 0.4 | Impaired | Severe LOF | Dysfunctional | [41] |
| 272 | G | V | Case | 34% | 10 | | | | Severe LOF | Dysfunctional | [42] |
| 275 | F | S | Case | 34% | 27 | 1.5 | 2 | Impaired | Severe LOF | Dysfunctional | [43] |
| 277 | S | L | Case | 0% | -8.7 | | | | Severe LOF | Dysfunctional | [44] |
| 279 | F | I | Case | 150% | -25 | 0.42 | 1 | Normal | Severe GOF | Dysfunctional | [45] |
| 281 | Y | C | Case | 0% | | | | Normal | Severe LOF | Dysfunctional | [29] |
| 283 | A | T | Case | 20% | 9 | | | | Severe LOF | Dysfunctional | [46] |
| 296 | F | S | Case | 12% | -10 | | | | Severe LOF | Dysfunctional | [14] |
| 300 | A | T | Control | 15% | -19 | | | Normal | Severe LOF | Dysfunctional | [29] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 302 | A | V | Case | 5% | | | | | Severe LOF | Dysfunctional | [14] |
| 305 | W | S | Case | 0% | | | | | Severe LOF | Dysfunctional | [40] |
| 307 | V | L | Case | 130% | -18 | 0.52 | | | Severe GOF | Dysfunctional | [47] |
| 310 | V | I | Case | 0% | 60 | | | | Severe LOF | Dysfunctional | [20] |
| 313 | I | K | Case | 0% | | | | | Severe LOF | Dysfunctional | [48] |
| 314 | G | S | Case | 12% | | | | | Severe LOF | Dysfunctional | [49] |
| 315 | Y | C | Case | 0% | | | | Normal | Severe LOF | Dysfunctional | [29] |
| 315 | Y | S | Case | 0% | | | | | Severe LOF | Dysfunctional | [40] |
| 316 | G | E | Case | 18% | 0 | | | | Severe LOF | Dysfunctional | [14] |
| 320 | P | A | Case | 0% | | | | | Severe LOF | Dysfunctional | [50] |
| 320 | P | H | Case | 0% | | | | | Severe LOF | Dysfunctional | [50] |
| 322 | T | A | Case | 0% | | | | Impaired | Severe LOF | Dysfunctional | [51] |
| 322 | T | M | Case | 0% | | | | Impaired | Severe LOF | Dysfunctional | [51] |
| 325 | G | R | Case | 0% | | | | | Severe LOF | Dysfunctional | [52] |
| 338 | S | F | Case | 5% | 12 | | | Normal | Severe LOF | Dysfunctional | [53] |
| 339 | F | S | Case | 4% | 1 | | | Normal | Severe LOF | Dysfunctional | [53] |
| 341 | A | V | Case | 6% | 60 | 5.59 | 0.29 | Normal | Severe LOF | Dysfunctional | [54] |
| 342 | L | F | Case | 0% | | | | | Severe LOF | Dysfunctional | [40] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 343 | P | S | Case | 0% | | | | | Severe LOF | Dysfunctional | [55] |
| 344 | A | V | Case | 100% | 40 | | | | Severe LOF | Dysfunctional | [56] |
| 357 | Q | R | Case | 27% | 20 | 3 | 1 | Impaired | Severe LOF | Dysfunctional | [57] |
| 360 | R | G | Case | 20% | | | | | Severe LOF | Dysfunctional | [14] |
| 366 | R | P | Case | 0% | 24.1 | | | | Severe LOF | Dysfunctional | [58] |
| 366 | R | Q | Case | 22% | 29 | 1 | | | Severe LOF | Dysfunctional | [21] |
| 366 | R | W | Case | 30% | 39.2 | | | Impaired | Severe LOF | Dysfunctional | [58] |
| 371 | A | T | Case | 0% | 21.9 | | | | Severe LOF | Dysfunctional | [58] |
| 373 | S | P | Case | 5% | 37.9 | | | Impaired | Severe LOF | Dysfunctional | [58] |
| 379 | W | R | Case | 0% | | | | Impaired | Severe LOF | Dysfunctional | [32] |
| 380 | R | S | Case | 33% | 0 | | | Normal | Mild LOF | Dysfunctional | [59] |
| 391 | T | I | Case | 85% | 0 | | | | Normal | Normal | [20] |
| 392 | W | R | Case | 0% | 28.3 | | | | Severe LOF | Dysfunctional | [58] |
| 393 | K | M | Case | 33% | 0 | | | Normal | Mild LOF | Dysfunctional | [59] |
| 393 | K | N | Control | 100% | 13.3 | | | | Normal | Normal | [58] |
| 397 | R | Q | Control | 90% | 0 | | | Impaired | Normal | Normal | [60] |
| 397 | R | W | Control | 40% | 0 | 1 | 1 | Normal | Mild LOF | Dysfunctional | [59] |
| 417 | V | M | Case | 100% | 0 | 1 | | | Normal | Normal | [36] |
| 448 | P | R | Control | 120% | 0 | | | | Normal | Normal | [20] |
| 455 | H | Y | Case | 43% | 0 | 0.6 | | | Mild LOF | Dysfunctional | [14] |
| 520 | M | R | Case | 0% | | | | Absent | Severe LOF | Dysfunctional | [61] |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 522 | Y | S | Case | 10% | 7 | | | Impaired | Severe LOF | Dysfunctional | [32] |
| 525 | A | T | Case | 36% | 22 | 1.34 | 1.08 | Impaired | Severe LOF | Dysfunctional | [22] |
| 533 | R | W | Case | 72% | 13.9 | 1 | | | Normal | Normal | [23] |
| 539 | R | W | Case | 17% | 33.9 | 1 | 0.41 | | Severe LOF | Dysfunctional | [23] |
| 546 | S | L | Case | 25% | 50.7 | 1.3 | 0.81 | Normal | Severe LOF | Dysfunctional | [62] |
| 555 | R | C | Case | 25% | 60 | | | | Severe LOF | Dysfunctional | [40] |
| 555 | R | H | Case | 12% | 50 | 1.1 | 0.72 | Normal | Severe LOF | Dysfunctional | [63] |
| 557 | K | E | Case | 0% | | | | Normal | Severe LOF | Dysfunctional | [64] |
| 562 | R | M | Case | 43% | 43.3 | 1.55 | 1.07 | Normal | Severe LOF | Dysfunctional | [62] |
| 583 | R | H | Case | 100% | 0 | | | | Normal | Normal | [65] |
| 589 | G | D | Case | 15% | 33 | | | Impaired | Severe LOF | Dysfunctional | [66] |
| 590 | A | T | Case | 45% | 10 | | | Normal | Mild LOF | Dysfunctional | [67] |
| 594 | R | Q | Case | 5% | 60 | | | | Severe LOF | Dysfunctional | [20] |
| 611 | D | Y | Case | 100% | 0 | | | | Normal | Normal | [68] |
| 619 | L | M | Case | 1% | | | | Normal | Severe LOF | Dysfunctional | [63] |
| 643 | G | S | Control | 35% | 1.1 | 1 | 0.72 | | Mild LOF | Normal | [69] |

**Table S2**

| # hidden neurons | MCC | AUC |
|---|---|---|
| 1 | 0.568 | 0.882 |
| 2 | 0.567 | 0.881 |
| 3 | 0.572 | 0.884 |
| 4 | 0.562 | 0.883 |
| 5 | 0.581 | 0.881 |
| 6 | 0.584 | 0.886 |
| 7 | 0.581 | 0.885 |
| 8 | 0.559 | 0.88 |

**Table S3**

| Feature | Information gain | Threshold maximizes information gain |
|---|---|---|
| Rate of evolution | 0.22 | 1.46 |
| PSSM perturbation | 0.18 | 5.89 |
| Change in hydrophobicity | 0.035 | 0.01 |
| Predicted residue packing density [70] | 0.024 | 11.96 |
| Grantham score [71] | 0.02 | 103 |
| Change in charge | 0.018 | NA |
| Change in SASA* | 0.017 | 29.91 |

**SASA: solvent accessible surface area*

**Table S4**

| Method | Medians and [Q1, Q3] intervals of performance metrics | | | | | | |
|---|---|---|---|---|---|---|---|
| | AUC | MCC | PPV | NPV | Accuracy | TPR | TNR |
| Q1VarPred | 0.884 [0.876, 0.890] | 0.584 [0.560, 0.608] | 0.967 [0.966, 0.968] | 0.533 [0.502, 0.565] | 0.889 [0.871, 0.890] | 0.905 [0.885, 0.906] | 0.783 [0.767, 0.783] |
| KvSNP | 0.669 [0.577, 0.753] | 0.306 [0.213, 0.462] | 0.926 [0.900, 0.938] | 0.333 [0.250, 0.429] | 0.829 [0.800, 0.865] | 0.903 [0.839, 0.935] | 0.500 [0.250, 0.600] |
| PhD-SNP | 0.726 [0.653, 0.794] | 0.369 [0.293, 0.494] | 0.935 [0.913, 0.963] | 0.364 [0.273, 0.500] | 0.829 [0.771, 0.865] | 0.871 [0.774, 0.935] | 0.600 [0.500, 0.750] |
| PolyPhen-2 | 0.625 [0.593, 0.718] | 0.372 [0.298, 0.477] | 0.912 [0.899, 0.935] | 0.500 [0.333, 0.667] | 0.886 [0.857, 0.914] | 0.968 [0.935, 1.000] | 0.250 [0.250, 0.500] |
| PredictSNP | 0.653 [0.593, 0.718] | 0.306 [0.211, 0.470] | 0.912 [0.906, 0.936] | 0.500 [0.333, 0.600] | 0.865 [0.838, 0.892] | 0.935 [0.903, 0.968] | 0.400 [0.250, 0.500] |
| PROVEAN | 0.788 [0.722, 0.810] | 0.556 [0.468, 0.576] | 0.956 [0.938, 0.957] | 0.557 [0.500, 0.593] | 0.896 [0.880, 0.899] | 0.926 [0.925, 0.936] | 0.683 [0.579, 0.700] |
| SIFT | 0.684 [0.593, 0.786] | 0.435 [0.313, 0.532] | 0.926 [0.900, 0.962] | 0.500 [0.333, 0.600] | 0.865 [0.838, 0.886] | 0.935 [0.875, 0.969] | 0.500 [0.250, 0.750] |
| SNAP | 0.512 [0.484, 0.605] | 0.170 [0.089, 0.255] | 0.886 [0.875, 0.909] | 0.167 [0.000, 0.222] | 0.800 [0.714, 0.865] | 0.875 [0.750, 0.969] | 0.200 [0.000, 0.400] |
| SNPs&GO | 0.706 [0.638, 0.762] | 0.326 [0.232, 0.405] | 0.933 [0.920, 0.960] | 0.286 [0.250, 0.333] | 0.771 [0.730, 0.829] | 0.806 [0.742, 0.871] | 0.600 [0.500, 0.750] |

**Table S5**

| Subdomain | Range | Length | Observed number of variants | Expected number of variants | Enrichment |
|---|---|---|---|---|---|
| NTD | 1-110+118-121 | 114 | 1 | 16 | 0.1 |
| S0 | 111-117 | 7 | 3 | 1 | 3 |
| S1 | 122-146 | 25 | 2 | 3 | 0.7 |
| S1-S2 | 147-153 | 7 | 0 | 1 | 0 |
| S2 | 154-177 | 24 | 2 | 3 | 0.7 |
| S2-S3 | 178-197 | 20 | 4 | 3 | 1.3 |
| S3 | 198-215 | 18 | 10 | 3 | 3.3 |
| S3-S4 | 216-222 | 7 | 0 | 1 | 0 |
| S4 | 223-241 | 19 | 5 | 3 | 1.7 |
| S4-S5 | 242-259 | 18 | 8 | 3 | 2.7 |
| S5 | 260-284 | 25 | 11 | 3 | 3.7 |
| S5-pore | 285-298 | 14 | 1 | 2 | 0.5 |
| pore-helix | 299-312 | 14 | 5 | 2 | 2.5 |
| pore-loop | 313-322 | 10 | 9 | 1 | 9 |
| S6 | 323-360 | 38 | 9 | 5 | 1.8 |
| S6-A | 361-369 | 9 | 3 | 1 | 3 |
| A | 370-389 | 20 | 4 | 3 | 1.3 |
| A-B | 390-506 | 117 | 4 | 16 | 0.3 |
| B | 507-532 | 26 | 3 | 4 | 0.8 |
| B-C | 533-547 | 15 | 2 | 2 | 1 |
| C | 548-562 | 15 | 4 | 2 | 2 |
| C-D | 563-587 | 25 | 0 | 3 | 0 |
| D | 588-622 | 35 | 4 | 5 | 0.8 |
| D-end | 623-676 | 54 | 0 | 8 | 0 |

NTD: N-terminal domain

**Table S6**

| Tool | Algorithm | Link | Reference |
|---|---|---|---|
| KvSNP | Fast random forest | http://www.bioinformatics.leeds.ac.uk/KvDB/KvSNP.html | 6 |
| PhD-SNP | Support vector machine | http://snps.biofold.org/phd-snp/phd-snp.html | 6, 72 |
| PolyPhen-2 | Naïve Bayes classification | http://genetics.bwh.harvard.edu/pph2/bgi.shtml | 72 |
| PredictSNP | Metaserver | http://loschmidt.chemi.muni.cz/predictsnp1/ | 73 |
| PROVEAN | Sequence conservation | http://provean.jcvi.org/seq_submit.php | |
| SIFT | Sequence conservation | http://siftdna.org/www/SIFT_pid_subst_all_submit.html | 74 |
| SNAP | Neural networks | https://rostlab.org/services/snap2web/ | 10 |
| SNPs&GO | Support vector machine | http://snps.biofold.org/snps-and-go/snps-and-go.html | |

**Table S7**

| Variant | Clinical significance | Review status |
|---|---|---|
| R539Q | Uncertain significance | Criteria provided, single submitter |
| V541I | Uncertain significance | Criteria provided, multiple submitters, no conflicts |
| E543K | Not provided | No assertion provided |
| Q544L | Uncertain significance | Criteria provided, single submitter |
| S546L | Pathogenic/likely pathogenic, not provided | Criteria provided, multiple submitters, no conflicts |
| Q547R | Not provided | No assertion provided |

**Training Dataset**

| Residue ID | Wild Type | Variant | PSSM | Rate of Evolution | Label |
|---|---|---|---|---|---|
| 46 | A | T | 1.22 | 4.094 | 0 |
| 110 | V | I | 1.77 | 1.976 | 1 |
| 111 | Y | C | 12.38 | 1.094 | 1 |
| 114 | L | P | 8.91 | 0.3189 | 1 |
| 117 | P | L | 12.45 | 0.4344 | 1 |
| 140 | S | G | 7.58 | 0.3528 | 1 |
| 141 | V | M | 6.46 | 0.5978 | 1 |
| 147 | Q | R | 5.88 | 1.995 | 0 |
| 168 | G | R | 9.72 | 1.595 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 174 | R | C | 11.94 | 0.151 | 1 |
| 178 | A | T | 7.14 | 0.4896 | 1 |
| 179 | G | S | 8.23 | 1.544 | 0 |
| 190 | R | Q | 7.69 | 0.6565 | 1 |
| 191 | L | P | 8.57 | 1.967 | 1 |
| 193 | F | L | 8.26 | 1.243 | 1 |
| 202 | D | E | 7.31 | 0.1174 | 1 |
| 202 | D | H | 10.14 | 0.1174 | 1 |
| 202 | D | N | 7.56 | 0.1174 | 1 |
| 204 | I | F | 7.48 | 1.166 | 1 |
| 204 | I | M | 2.76 | 1.166 | 1 |
| 204 | I | N | 10.53 | 1.166 | 1 |
| 205 | V | M | 6.81 | 0.4373 | 1 |
| 207 | V | M | 4.48 | 2.322 | 0 |
| 209 | S | F | 10.14 | 0.3612 | 1 |
| 209 | S | P | 8.45 | 0.3612 | 1 |
| 215 | V | M | 4.18 | 1.855 | 1 |
| 225 | S | L | 10.22 | 0.5382 | 1 |
| 231 | R | C | 12.58 | 0.2855 | 1 |
| 231 | R | H | 9.19 | 0.2855 | 1 |
| 235 | I | N | 11.64 | 0.4978 | 1 |
| 236 | L | R | 9.01 | 0.7966 | 1 |
| 243 | R | C | 12.58 | 0.3992 | 1 |
| 243 | R | H | 9.19 | 0.3992 | 1 |
| 248 | W | R | 17.24 | 1.461 | 1 |
| 251 | L | P | 9.77 | 0.6864 | 1 |
| 254 | V | M | 6.81 | 0.738 | 1 |
| 258 | H | R | 12.12 | 1.042 | 1 |
| 259 | R | C | 9.64 | 0.914 | 1 |
| 259 | R | H | 7.68 | 0.914 | 1 |
| 261 | E | D | 6.62 | 0.6798 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 261 | E | K | 7.41 | 0.6798 | 1 |
| 265 | T | I | 7.52 | 0.3957 | 1 |
| 269 | G | D | 9.51 | 0.6323 | 1 |
| 269 | G | S | 8.41 | 0.6323 | 1 |
| 272 | G | V | 3.55 | 0.7165 | 1 |
| 275 | F | S | 11.68 | 0.8148 | 1 |
| 277 | S | L | 10.22 | 0.3139 | 1 |
| 279 | F | I | 7.32 | 0.7992 | 1 |
| 281 | Y | C | 13.2 | 0.6599 | 1 |
| 283 | A | T | 6.33 | 0.9759 | 1 |
| 296 | F | S | 12.32 | 0.4958 | 1 |
| 300 | A | T | 7.26 | 0.62 | 1 |
| 302 | A | V | 7.24 | 0.1591 | 1 |
| 305 | W | S | 17.32 | 0.3242 | 1 |
| 307 | V | L | 3.71 | 0.5225 | 1 |
| 310 | V | I | 3.08 | 0.6607 | 1 |
| 313 | I | K | 11.05 | 0.1534 | 1 |
| 314 | G | S | 8.41 | 0.06888 | 1 |
| 315 | Y | C | 13.2 | 0.3916 | 1 |
| 315 | Y | S | 12.42 | 0.3916 | 1 |
| 316 | G | E | 10.37 | 0.05287 | 1 |
| 320 | P | A | 10.86 | 0.2219 | 1 |
| 320 | P | H | 12.32 | 0.2219 | 1 |
| 322 | T | A | 8.32 | 0.2149 | 1 |
| 322 | T | M | 9 | 0.2149 | 1 |
| 325 | G | R | 10.58 | 0.1235 | 1 |
| 338 | S | F | 10.14 | 0.409 | 1 |
| 339 | F | S | 12.32 | 0.3956 | 1 |
| 341 | A | V | 7.42 | 0.3778 | 1 |
| 342 | L | F | 6.23 | 0.6439 | 1 |
| 343 | P | S | 10.86 | 0.3279 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 344 | A | V | 7.4 | 0.3204 | 1 |
| 357 | Q | R | 7.09 | 0.6863 | 1 |
| 360 | R | G | 11.41 | 1.042 | 1 |
| 366 | R | P | 9.15 | 0.3871 | 1 |
| 366 | R | Q | 6.16 | 0.3871 | 1 |
| 366 | R | W | 10.32 | 0.3871 | 1 |
| 371 | A | T | 7.26 | 0.1523 | 1 |
| 373 | S | P | 6.06 | 1.629 | 1 |
| 379 | W | R | 17.24 | 0.3489 | 1 |
| 380 | R | S | 9.75 | 0.3027 | 1 |
| 391 | T | I | 9.03 | 0.261 | 0 |
| 392 | W | R | 16.87 | 1.454 | 1 |
| 393 | K | M | 7.75 | 2.154 | 1 |
| 393 | K | N | 3.56 | 2.154 | 0 |
| 397 | R | Q | 4.41 | 2.289 | 0 |
| 397 | R | W | 8.23 | 2.289 | 1 |
| 417 | V | M | -4.46 | 4.87 | 0 |
| 448 | P | R | 2.7 | 4.808 | 0 |
| 455 | H | Y | 5.53 | 4.871 | 1 |
| 520 | M | R | 12.09 | 0.9933 | 1 |
| 522 | Y | S | 11.33 | 0.6934 | 1 |
| 525 | A | T | 6.71 | 0.3731 | 1 |
| 533 | R | W | 9.46 | 0.4498 | 0 |
| 539 | R | W | 10.11 | 0.3226 | 1 |
| 546 | S | L | 10.26 | 0.1776 | 1 |
| 555 | R | C | 12.73 | 0.2751 | 1 |
| 555 | R | H | 9.34 | 0.2751 | 1 |
| 557 | K | E | 7.53 | 0.2782 | 1 |
| 562 | R | M | 10.54 | 0.3266 | 1 |
| 583 | R | H | 5.69 | 4.713 | 0 |
| 589 | G | D | 8.38 | 1.456 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| 590 | A | T | 5.2 | 0.8497 | 1 |
| 594 | R | Q | 6.31 | 1.004 | 1 |
| 611 | D | Y | 4.14 | 2.709 | 0 |
| 619 | L | M | 0.83 | 2.44 | 1 |
| 643 | G | S | -1.1 | 1.479 | 0 |