Continuous Bayesian Variant Interpretation Accounts for Incomplete Penetrance among Mendelian Cardiac Channelopathies

Short Title: Continuous Bayesian Penetrance Probabilities

Matthew J. O'Neill[1], Luca Sala[2], Isabelle Denjoy[3], Yuko Wada[4], Krystian Kozek[1], Lia Crotti[2], Federica Dagradi[2], Maria-Christina Kotta[2], Carla Spazzolini[2], Antoine Leenhardt[3], Joe-Elie Salem[3], Asami Kashiwa[5], Seiko Ohno[6], Ran Tao[7], Dan M. Roden[8], Minoru Horie[9], Fabrice Extramiana[3], Peter J. Schwartz[2], and Brett M. Kroncke[4]*

1. Vanderbilt University School of Medicine, Medical Scientist Training Program, Vanderbilt University, Nashville, TN
2. IRCCS, Istituto Auxologico Italiano, Center for Cardiac Arrhythmias of Genetic Origin and Laboratory of Cardiovascular Genetics, Milano, Italy
3. Department of Cardiovascular Medicine, Hôpital Bichat, APHP, Université de Paris Cité, Paris, France
4. Vanderbilt Center for Arrhythmia Research and Therapeutics (VanCART), Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
5. Department of Cardiovascular Medicine, Kyoto University Graduate School of Medicine Kyoto, Japan
6. Department of Bioscience and Genetics, National Cerebral and Cardiovascular Center, Osaka, Japan

7. Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN
8. Vanderbilt Center for Arrhythmia Research and Therapeutics (VanCART), Departments of Medicine, Pharmacology, and Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
9. Department of Cardiovascular Medicine, Shiga University of Medical Science, Shiga, Japan

*Correspondence to Brett M. Kroncke:
Office 1215E Light Hall/Medical Research Building IV
2215 Garland Avenue
Nashville, TN, 37232
brett.m.kroncke.1@vumc.org

Abstract

Background: The congenital Long QT Syndrome (LQTS) and Brugada Syndrome (BrS) are Mendelian autosomal dominant diseases which frequently precipitate fatal cardiac arrhythmias. Incomplete penetrance is a barrier to clinical management of heterozygotes harboring variants in the major implicated disease genes *KCNQ1, KCNH2*, and *SCN5A*. We apply and evaluate a Bayesian penetrance estimation strategy that accounts for this phenomenon and evaluate penetrance distributions and rationalize their structural underpinnings across four genotype-phenotype pairs.

Methods: We generated Bayesian penetrance estimation models for *KCNQ1*-LQT1 and *SCN5A*-LQT3 using variant-specific features and clinical data from the literature, international arrhythmia genetic centers, and population controls. We analyzed the distribution of posterior penetrance estimates across four-genotype phenotype relationships and compared continuous estimates to ClinVar annotations. Posterior estimates were mapped onto protein structure.

Results: Bayesian models of *KCNQ1*-LQT1 and *SCN5A*-LQT3 are well-calibrated to clinical observations. Variant-informed penetrance estimates of *KCNQ1*-LQT1 and *SCN5A*-LQT3 are empirically equivalent to 10 and 5 heterozygote clinical phenotypes, respectively. Posterior penetrance estimates were bimodal for *KCNQ1*-LQT1 and *KCNH2*-LQT2, with a higher fraction of missense variants with high penetrance among *KCNQ1* variants. *SCN5A*-LQT3 and *SCN5A*-BrS had comparatively more variants with predicted low penetrance. There was a wide distribution of variant penetrance estimates among similar categories of ClinVar annotations. Structural mapping revealed heterogeneity among ‘hot spot’ regions and featured high penetrance estimates for *KCNQ1* variants in contact with calmodulin and the S6 domain.

62 **Conclusions:** Bayesian penetrance estimates provide a continuous framework for variant
63 interpretation, provide higher resolution within 'hot spot' domains, and facilitate prospective
64 clinical management of variant heterozygotes.
65

Introduction

Incomplete penetrance is a hallmark of the inconsistency of genetic determinism, where heterozygotes harboring the same variant may manifest a continuum of the disease phenotype, or none at all[1-3]. The nuance of incomplete penetrance is frequently unaccounted for in contemporary genetic methods development; probabilistic approaches may more accurately account for variable genetic contributions to disease presentation. The American College of Medical Genetics and Genomics (ACMG) has established criteria for adjudicating the 'pathogenicity' of variants, drawing to varying extents on population frequency data, segregation within families, *in silico* methods, and functional data, among other sources[4]. However, this categorical approach is unable to account for incomplete penetrance, especially among ClinVar-annotated Pathogenic or Likely Pathogenic variants[5].

Clinical management of patients affected by inherited arrhythmic syndromes is a salient example of this challenge. The Long QT Syndromes Type 1-3 (LQT1-3 [MIM: 192500, 613688, 603830], respectively) and Brugada Syndrome (BrS [MIM 601144]) are frequently responsible for Sudden Cardiac Death[6]. The most rigorously established genotype-phenotype relationships are loss-of-function variants in *KCNQ1* encoding K_V7.1 (LQT1), *KCNH2* encoding K_V11.1 (LQT2), and *SCN5A* encoding Na_V1.5 (BrS) and gain-of-function variants in *SCN5A* (LQT3)[7,8]. These membrane proteins are responsible for orchestrating the cardiac action potential through cardiomyocyte depolarization and repolarization. All four of these genotype-phenotype pairs are characterized by a significant degree of incomplete penetrance[9], which may in part be accounted for by environmental risk factors as well as polygenic liability[10-12]. The ability to accurately predict variant penetrance is especially important, since medical or interventional therapies are well established to lower the risk of fatal ventricular arrhythmias[6]; however, overly aggressive management may conversely cause harm[13,14]. New methods for variant interpretation that capture the granularity of clinical penetrance in a probabilistic manner will help facilitate the clinical management of variant heterozygotes.

Herein, we elaborate a framework where variant-specific features inform a continuous, quantitative probability of disease manifestation (penetrance) for missense variants – in contrast to the probability of 'pathogenic' or 'benign' classification[15,16]. These prior estimates are calibrated to clinically observed penetrance from the spectrum of all available variant-specific feature data. Within a Bayesian framework, we apply clinical evidence from phenotyped individuals from the literature, arrhythmia genetics clinical cohorts, and putative population controls to these calibrated prior estimates of disease probability to yield posterior penetrance estimates (Figure 1). This approach recasts nuanced variant features, previously only interpretable by specialists, e.g., structural and channel biophysical properties, to a wider audience as an equivalent number of clinically phenotyped heterozygotes. These estimates can therefore inform clinical management in the absence of prior clinical observations of a variant, or aid in interpretation of variants when few heterozygotes have been clinically phenotyped.

Implementing two instances of this approach, we develop penetrance models for *KCNQ1*-LQT1 and *SCN5A*-LQT3, incorporate data from *KCNH2*-LQT2[17] and *SCN5A*-BrS1[18], and investigate penetrance across these four genotype-phenotype relationships. We demonstrate how penetrance varies among these genotype-phenotype relationships and correlate penetrance with structural features of each gene product.

Materials and Methods

All data and code used to generate and evaluate the models are available on GitHub. Results are also curated in the *KCNQ1* and *SCN5A* variant browsers (see Web Resources). Detailed methods are available in the Supplemental Methods.

Results

Evaluation and Performance of KCNQ1 *Variant Penetrance Model.* We implemented a Bayesian penetrance model using all available data for the *KCNQ1*-LQT1 genotype-phenotype pair (Figure 1)[17,18]. We curated >2,000 manuscripts from the literature, and in combination with phenotyped heterozygotes from international arrhythmia clinics and putative heterozygote controls from gnomAD[19], found 9,969 *KCNQ1* heterozygotes harboring 630 unique missense and in-frame insertion/deletion variants. Plotting the observed penetrance of each variant by residue recapitulated known associations with 'hot spot' regions of the protein; plotting our structural penetrance density metric[20] (residue weighted by penetrance of adjacent residues in 3D environment) by residue resulted in more continuous estimates (Supplemental Figure 1).

We evaluated the functional, *in silico*, and structural features against the estimated LQT1 penetrance (the observed penetrance modestly biased towards the mean, called the empirical posterior, see Materials and Methods) using Spearman-rank order coefficients, weighted towards variants with many clinical observations (Figure 2 and Supplemental Table 1). We found many channel biophysical properties were statistically significant (nominal $p < 0.05$). For example, reductions in peak current decrease repolarization and, as the primary molecular correlate of LQT1, predict higher LQT1 penetrance (Figure 2; Spearman correlation coefficient -0.55 homozygous, -0.54 heterozygous). CardioBoost[21] had the highest correlation coefficient (+0.64) of the *in silico* predictors, followed by REVEL[22] (+0.64) and Provean[23] (-0.61). The structural LQT1-dist covariate had a Spearman correlation coefficient of +0.54. Integrating the most relevant features using our previously described method[17], we derived a LQT1 Bayesian prior with a coefficient of +0.71 (10-fold cross validated Spearman correlation coefficient +0.65 – see Materials and Methods). We additionally evaluated model covariates using Brier scores[24], a probabilistic forecasting metric where 0 is most favorable, and 1 is least favorable (Table 1). We

observe Brier scores for the LQT1-dist metric of 0.062 while the *in silico* covariates CardioBoost (0.19), REVEL (0.17), and PolyPhen2 (0.22) performed less well. The Brier score for the Bayesian prior was 0.028 (10-fold cross validated Brier score 0.033) and 0.011 for the Bayesian posterior.

A central advantage of our Bayesian penetrance estimate is the quantification of uncertainty (Supplemental Figure 2 – see Materials and Methods)[17]. This is expressed as a hypothetical number of heterozygotes, e.g., 7 of 10 clinically phenotyped individuals with a phenotype has a greater uncertainty than 700 of 1,000 clinically phenotyped heterozygotes with a phenotype, though the estimated penetrance is the same. To arrive at this number, we empirically test a tuning parameter, equivalent to the number of hypothetically phenotyped heterozygotes, such that 95% of the true observations (per variant empirical posterior) are contained within the EM posterior's 95% credible interval across the distribution from 0 to 1.0 using boot strap sampling (Supplemental Figure 2 – see outcomes with several tuning parameters tested). From this analysis, for *KCNQ1*-LQT1, we suggest the Bayesian prior is empirically equivalent to the information contained in approximately 10 clinically phenotyped individuals.

In Figure 3, we provide a representative example of the prior and posterior penetrance estimates obtained by our approach imposed on the latter part of the S6 transmembrane segment into the C-terminus (residues 350-400). The forest plot shows prior (green, blue, and red) and posterior (black) estimates and their corresponding 95% credible intervals for variants harbored by at least five heterozygotes. The 95% credible interval of the Bayesian prior estimate is narrowed after the application of the likelihood (heterozygote status). For example, for the variant p.Ala370Val, there is a moderate Bayesian prior probability (0.28). After applying the clinical likelihood (1 affected heterozygote of 45 observed heterozygotes), the Bayesian posterior is decreased (.067) along with the respective 95% credible interval. This method

provides a best probability estimate for these variants, which intrinsically accounts for incomplete penetrance, even among high-penetrance variants.

LQT3 Penetrance. We performed a similar analysis for the *SCN5A*-LQT3 genotype-phenotype pair. In contrast to LQT1, the LQT3 phenotype arises from a specific loss of channel inactivation, which may arise through late current, window current, or altered kinetics of inactivation[25]. We previously curated a dataset to evaluate this relationship empirically[18], which includes a total of 86,118 *SCN5A* heterozygotes with 1,243 manifesting the LQT3 phenotype. Analogous to LQT1, we updated prior penetrance estimates based on functional, structural, and *in silico* features with this clinical evidence. The posterior estimates were empirically shown to be equivalent to the information contained in 5 clinical observations of heterozygotes using the same framework described above (Supplemental Figure 3).

Bayesian Estimates Reflects Gene-specific Penetrance Among Channelopathies. To assess differences in penetrance among the four channelopathies (LQT1-3 and BrS), we compared the Bayesian posterior penetrance estimate distributions for each gene-disease pair (Figure 4). Observed penetrance (strictly frequentist) tends towards 0 or 1 owing to most unique variants being observed in a single or few phenotyped heterozygotes (Figure 4A). The Bayesian posterior penetrance estimates are distributed over a continuous bimodal distribution (Figure 4B). We observe a cluster of variants with low penetrance estimates, followed by a second cluster with higher estimates which may loosely map onto the categorical ACMG framework. We interpret these estimates as indicating ‘benign’ at low estimated penetrance, followed by a continuous gradation of “risk allele’ to ‘pathogenic low penetrance’ to ‘pathogenic’ variants. Formalizing these estimates as continuous disease-presentation probabilities explicitly accounts for incomplete penetrance along this continuum. *SCN5A* variants associated with BrS and LQT3 do not display the same extent of bimodal distribution as those in *KCNQ1* and feature

comparatively more variants predicted to have very low penetrance (Figure 4C and 4D). This finding partially reflects the composition of *SCN5A,* at approximately four times the size of *KCNQ1* with large linkers (400 residues) between membrane spanning segments largely uninvolved in channel function and correspondingly low estimated penetrance. Finally, *KCNH2* variants also have a bimodal distribution like those for *KCNQ1* variants, but with a smaller fraction predicted at high penetrance (Figure 4E). These observations also reflect the relative fraction of functionally relevant residues within the gene. These differences are quantified in Figure 4F. These distributions indicate half of identified *KCNQ1* variants have a penetrance estimate of at least 0.41, suggesting a randomly chosen *KCNQ1* variant has a 50% chance of producing a LQT1 penetrance greater than 41%. The probability distributions skew comparably lower for *KCNH2*-LQT2 and *SCN5A*-LQT3 where half of observed variants have penetrance estimates lower than 0.157 and 0.063, respectively.

Distribution of Posterior Estimates within Individual ClinVar Annotations. We next investigated how our Bayesian estimates and observed penetrance compared with ClinVar categorical designations. In Figure 5A, we present posterior estimates for Benign/Likely Benign (B/LB) and Pathogenic/Likely Pathogenic (P/LP) variants for *KCNQ1*. Plotting the observed penetrance vs category reveals that many variants are incompletely penetrant (Figure 5A). We observe that the Bayesian estimates for B/LB cluster close to 0, although we show remarkable heterogeneity for estimates of the P/LP variants (ranging from 0.17-0.90). We repeated this analysis for variants within *KCNH2* (Figures 5C and 5D). We again observe distributions of Bayesian penetrance estimates (0.093-0.95) as well as a high degree of incomplete penetrance among P/LP variants. Some P/LP variants have been observed in relatively few individuals without a phenotype, and therefore have an observed penetrance of 0 (Figure 5A/C). Further investigation revealed that several of these variants have a P/LP ClinVar annotation which includes private

218 communications about heterozygote clinical phenotypes (such as *KCNH2* p.Gly657Ser), which
219 were therefore unavailable to incorporate into our model.
220
221 *Structural Implications of Variable Penetrance.* We mapped posterior penetrance estimates onto
222 Cryo-EM structures (PDBIDs 5VMS, 5VA1, and 5X0M for KCNQ1, KCNH2, and SCN5A
223 [homology modeled to include the c-terminal EF hand domain], respectively) of each protein and
224 interrogated global and 'hot spot' penetrance heterogeneity. As shown in Figure 6A, there is
225 substantial heterogeneity in penetrance across traditionally labelled 'hot spot' regions in K_V7.1,
226 including the pore domain, voltage sensing domains, and subunit assembly domains (Figure 6A
227 and 6B)[26]. Whereas several variants in these 'hot spot' domains are observed to have high
228 posterior estimates, as indicated by the color and thickness of the putty rendering, there remain
229 several residues with comparatively low penetrance within the same domain. Interestingly, we
230 observe enriched penetrance across the known binding site of calmodulin (Figure 6C)[27]. For
231 example, surface residues on two helices in contact with calmodulin both have high penetrance
232 estimates, including p.Trp379 (3 observed variants, 4 of 5 heterozygotes affected) and
233 p.Met520Arg (4 of 4 heterozygotes affected). Structural comparisons revealed that the K_V7.1
234 S5-S6 domain have higher penetrance estimates than those of K_V11.1. We observe that the
235 high pathogenicity of variants in the S5 helix continues into the S6 helix of K_V7.1 (p.Ala344Val,
236 17 of 20 heterozygotes affected; p.Gly345Arg, 2 of 2 heterozygotes affected; p.Leu353Pro, 8/8
237 heterozygotes affected), whereas those for K_v11.1 rapidly taper (residue p.Met645 – 4 unique
238 variants – 9/10 heterozygotes affected; p.Ser660Leu, 2/7 heterozygotes affected). This finding
239 is consistent with a recent report implicating variants in the Kv7.1 S6 domain as particularly
240 pathogenic, causing a significant decrease in event-free survival[28]. This may be structurally
241 rationalized by the greater sensitivity of flexible motifs within the Kv7.1 S6 helix including the
242 'GSG' motif[29] (p.Ser349Trp, 2 of 2 heterozygotes affected, and p.Gly350, 3 variants, 9 of 9
243 heterozygotes affected).

We next mapped Bayesian posteriors for BrS and LQT3 onto the structure of *SCN5A* (Figure 7A and 7B). The BrS posterior estimates were highest in the pore region, especially near the selectivity filter, with some additionally high penetrance variants in the voltage sensing domains (VSDs) and the C-terminal EF-hand domain; however, even in the highly conserved pore helices, there were several BrS low-penetrance residues. These results contrast with penetrance estimates for LQT3, where high penetrance residues were concentrated in and near the inactivation gate on the intracellular side of the channel. While some domains are associated with both phenotypes (such as the C-terminal EF-hand domain), other regions of $Na_V1.5$ are exclusively linked to BrS or LQT3 (VSD and pore for BrS and inactivation gate for LQT3). For example, of the 164 heterozygotes for the variant p.Glu1784Lys, we observed 31 with BrS and 114 with LQT3 (posterior estimates 0.19 and 0.69, respectively). This variant is in the C-terminus of the S6 helix, and recent structural studies have implicated its role in fast inactivation[30]. In contrast, the in-frame deletion p.Gln1507-1509del in the domain III-domain IV linker (inactivation-gate) was LQT3 specific (85 of 87 heterozygotes LQT3 affected, 0 BrS). Our highest BrS posterior estimate was for the variant p.Arg878Cys (32 of 36 heterozygotes affected) in the domain II pore.

Discussion

Bayesian Penetrance Estimates for Mendelian Arrhythmias. We applied a probabilistic model to generate prior and posterior penetrance estimates for the *KCNQ1*-LQT1 and *SCN5A*-LQT3 genotype-phenotype pairs that explicitly account for incomplete penetrance in the model framework. Functional, *in silico*, and structural features of these channels were significantly associated with LQT1 and LQT3 penetrance. We found that these models carried information equivalent to 10 and 5 heterozygote clinical observations, respectively. The *KCNQ1*-LQT1 model had a Bayesian prior 10-fold cross-validated Brier scores of 0.032, indicating excellent probabilistic forecasting. We then analyzed the distributions of penetrance estimates for four genotype-phenotype pairs and found *KCNQ1*-LQT1 had the largest fraction of variants with high estimated penetrance (Figure 5E). These continuous estimates were then mapped onto each protein structure, which revealed a structural basis of these findings as well as significant heterogeneity in traditionally annotated 'hot spot' domains. We curated these data and host them in a convenient format for research and prospective clinical management.

Variable Penetrance Among Genotype-Phenotype Pairs. Incomplete penetrance is observed even among highly deleterious variants. For example, the *KCNQ1* variant p.Ala341Val has sharply decreased event-free survival compared to other *KCNQ1* missense variants[28], but still provokes the phenotype incompletely (301 of 375 heterozygotes manifest LQT1). Extended to all variants observed in at least one heterozygote, Figure 4 shows penetrance estimates for relatively few variants exhibit >90% penetrance (quantified in Figure 4E) and a bimodal distribution for LQT1 and LQT2 penetrance, a loosely bimodal distribution for BrS, and a taper for LQT3. The fraction of significantly high penetrance variants is consistent with the fraction of LQTS cases attributed to variants in these genes: roughly 50% of LQTS cases are attributable

to variants in *KCNQ1*, 40% in *KCNH2* and 10% in *SCN5A*[31]. This trend inversely correlates with gene size: (K_V7.1 676 amino acids, K_V11.1 1159 amino acids, and Na_V1.5 2016 amino acids). For BrS and LQT3, more variants can confer loss-of-function for BrS, compared to the more limited set of possible variants that could induce a 'gain-of-function' phenotype associated with LQT3. This difference may explain the relative decrease in precision of the *SCN5A*-LQT3 estimates reflected in the lower number of phenotype observations compared to *KCNQ1*-LQT1 (5 vs 10). The predictive features are comparatively weaker due to the bias to estimate loss-of-function rather than gain-of-function as a single score. Compared to LQT1 and LQT2, BrS (also associated with loss-of-function) penetrance distribution has a lower fraction of variants with penetrance >0.5, likely due to contributions to the phenotype from other genes or polygenic modifiers (Figure 5E). This is consistent with recent suggestions of the oligogenic nature of BrS as compared to more strictly Mendelian LQT1-3 with polygenic modifiers[32]. The complication of both arrhythmia-associated gain-of-function and loss-of-function phenotypes for missense variants in SCN5A likely contributes to the worse correlation between SCN5A predictive features and LQT3 penetrance leading to lower hypothetical heterozygote count for SCN5A-LQT3.

Continuous Bayesian Penetrance Provides Higher Resolution Variant Interpretation. The ACMG variant classification scheme is a clinically useful categorical framework to interpret variants[4]. While this approach enables convenient approximations for clinical decision making, variants within each of these categories may nevertheless exert divergent risk due to incomplete penetrance. The *KCNQ1* variant p.Val254Met has not been observed in gnomAD and is associated with a high frequency of sudden cardiac death and syncope in Japanese heterozygotes[33]. From variant-specific features alone, we obtained a penetrance estimate of 0.65 and a posterior LQT1 penetrance estimate of 0.84. In contrast, the disease-associated

312 variant p.Gly589Asp has a higher allele-frequency (4.96e-5 in gnomAD) than expected for a
313 highly pathogenic variant (2.5e-5 rare allele cutoff[34]), despite being annotated as Pathogenic in
314 ClinVar[5]. Accordingly, our model predicted a prior penetrance estimate of 0.37, and a posterior
315 estimate of 0.23 after incorporating clinical data (8 of 40 affected heterozygotes). Such a variant
316 clearly carries risk above the baseline population rate, but not as high as that conferred by
317 p.Val254Met. A strength of our approach is therefore the ability to capture this granular level of
318 detail, whereas classification methods lose this nuance. Our continuous estimates suggest that
319 certain variants may be considered 'modifiers' or exert subtle impact, consistent with the
320 terminology of "LQT-lite" proposed by Ackerman and colleagues[35-37]. This idea is additionally
321 consistent with recent work showing variable environment interactions in the presence of
322 common *SCN5A* variants, where the variant may act as a 'modifier' of penetrance[38].

323

324 The divergence of our approach, leveraging information contained within the variant-specific
325 features, can also be appreciated when contrasted with ClinVar categorical descriptions (Figure
326 5). ClinVar annotated P/LP variants display a range of predicted penetrance estimates (0.167-
327 0.903 *KCNQ1* and 0.0930-0.945 *KCNH2*); the effect is also evident from non-Bayesian
328 penetrance observations (Figure 5A and 5C). Furthermore, our findings in Figure 4A show that
329 observed penetrance is not sufficient for forecasting penetrance, that most unique variants are
330 observed in very few people leading to too many LQTS penetrance estimates of 0 or 1. Variant-
331 specific features modify these crude estimates from low numbers of phenotyped variants to shift
332 estimates away from the extremes to more intuitive and biologically reasonable estimates
333 (Figure 4B). Collectively, these observations support the notion that our continuous Bayesian
334 estimates offer a higher resolution alternative to categorical classification.

335

Penetrance Heterogeneity within Structured Functional Domains. Previous works have implicated 'hot spot' regions within the three protein products based on collapsing of variants in cases and controls; however, many of these studies preceded recent large-scale population genome sequencing efforts[26,39]. Our combination of genetic and structural data shows variable penetrance throughout many traditional 'hot spot' domains (Figure 6 and 7). For example, not all residues in the *SCN5A* domain IV VSD or pore are associated with high BrS penetrance (Figure 6A). These differences are further rationalized between the *SCN5A*-BrS and *SCN5A*-LQT3 pairs, where different electrophysiological parameters correspond to discrete structural features. Even for higher penetrance variants within 'hot spots', many provoke the phenotype incompletely. In *KCNH2,* the ClinVar Pathogenic variant p.Val822Met is located in the pore domain 'hot spot' but affects only 27 of 42 heterozygotes with a Bayesian penetrance estimate of 0.65. We also observe heterogeneity between homologous S6 segments in K_V7.1 and K_V11.1, with a much higher LQTS penetrance throughout the helix in the former. These results highlight important structural differences in these proteins that contribute to clinical penetrance, despite similar channel characteristics (both being voltage-gated potassium channels). Altogether, these data suggest using variant location in 'hot spot' domains in ACMG variant classifications may be premature[40]. For instance, recent work from our lab performed a Deep Mutational Scan of a 'hot spot' exon in KCNH2, and found experimentally that only 42% of all possible amino acid substitutions substantially decreased trafficking[41]. We view the current approach as providing higher resolution interpretation in these domains – this work provides more information than categorical classification, or yes/no 'hot spot' annotations.

Mapping penetrance to structure also provides unique insights into the underlying channel biology. For example, K_V7.1 residues 364-383 and 510-533 form the A and B helices, respectively, with several high-penetrance sites, including non-buried residues which interact

with calmodulin (Figure 6C). Calmodulin is global calcium sensor within the cardiomyocyte and is important for regulating excitation-contraction coupling[42]. The functional implications of Calmodulin binding $K_V7.1$ are speculative, however, a recent study implicated calmodulin as a state-dependent switch to modulate channel gating[43]. Our study suggests a high LQT1 penetrance when these interactions are abrogated by genetic variation, likely being mediated through this loss of gating modulation. These results provide a complementary set of data to implicate loss of the Calmodulin-$K_V7.1$ signaling axis as an etiology for LQT1.

Comparison with Other Approaches. Recent studies in the cardiac space have leveraged machine learning to facilitate variant reclassification[15,16]. Approaches such as GENESIS[16] and MLb-LDLr[15] make progress towards this goal but rely on ClinVar data as a critical input. The approach we take relies on classic regression strategies and expectation-maximization to integrate data, and instead draws on all available variant-specific features to output a *continuous* penetrance estimate calibrated to curated, granular clinical data. Rather than focusing on classification, other interesting approaches have instead focused primarily on functional impact, such as prediction of functional properties in the presence of missense variants[44,45]. The use of granular clinical data, each heterozygote evaluated as above or below a clinically meaningful threshold represents an advance by which we can assert not only a penetrance estimate, but also the uncertainty of the estimate – an approach well-tailored to incompletely penetrant BrS and LQTS variants.

Precision Medicine Applications. An increasingly important goal of precision medicine is to enable 'genetics first approaches', wherein sequencing of unascertained clinical cohorts may prioritize individuals for tailored follow-up[46,47]; however, prioritizing variant heterozygotes for follow up remains difficult. Additionally, sequencing of large population cohorts has revealed an

386 unexpectedly large frequency of missense variants in genes associated with Mendelian
387 disease[19,48,49]. It is unlikely there will ever be sufficient discovered heterozygotes for clinical
388 phenotyping to adjudicate the clinical consequence of all possible single nucleotide variants in
389 Mendelian disease genes[50]. It is therefore imperative to maximally leverage all variant-specific
390 features to help facilitate an understanding of clinical consequence[51]. As many phenotypes
391 demonstrate variable penetrance, our method provides an initial gestalt from variant-specific
392 features which help tailor prospective precision medicine in concert with other clinical
393 observations. To calibrate the strength of these features, we provide an empirical equivalent of
394 heterozygote observations, an interpretable form of these data. Based on this progress, we
395 envision a future in which genetic testing forms a pre-test probability of disease which is
396 updated by clinical phenotyping. This schematic would represent an inversion of the current use
397 of genetics in medical practice, in which a genetic test most commonly modifies a clinical pre-
398 test probability.

399

400 *Limitations*. One limitation of our study is the ascertainment bias of highly penetrant variants in
401 clinical practice. Many asymptomatic heterozygotes will not be clinically phenotyped or
402 adequately sampled in current population sequencing studies. This bias is additionally reflected
403 in the availability of functional data, as most studies have interrogated variants with a high
404 suspicion of pathogenicity; however, this bias is increasingly tempered by functional evaluation
405 of less unascertained variants such as those in a recent study of arrhythmia genes in the
406 eMERGE network[52]. We are currently unable to satisfactorily account for variant combinations
407 such as compound heterozygotes[53]. Data availability remains a limitation for model construction,
408 as several ClinVar annotations were based off 'personal communications' which resulted in low
409 or zero observed penetrance among P/LP variants in select cases.

410

Conclusion

Incomplete penetrance demands nuance to variant interpretation for the realization of genome-first precision medicine approaches[46]. We make progress towards this goal by elaborating a probabilistic approach that considers each variant as having a continuous penetrance probability and provides higher resolution in 'hot spot' domains to facilitate prospective clinical management. This approach provides more information than classification and explicitly treats incomplete penetrance in model construction. We render variant-specific features as equivalents to clinically-phenotyped heterozygotes to democratize these data to a broader audience without electrophysiology and structural biology expertise.

Description of Supplemental Data

We have included all referenced supplemental figures and tables (Supplemental Figures 1-3 and Supplemental Table 1), as well as a detailed description of the penetrance estimation protocol in the Supplemental Methods section.

Declaration of Interests

The authors declare no competing interests.

Ethics Declaration

Institutional Review Board (#191563) of Vanderbilt University Medical Center granted a waiver for this work based on 45 CFR 46.104 (d) category (4) for Exempt Review.

Acknowledgements

This research was funded by American Heart Association 907581 (MJO), a Marie Skłodowska-Curie Individual Fellowship (H2020-MSCA-IF-2017; Grant Agreement No. 795209, LS), a Fondazione CARIPLO, “Biomedical Research Conducted by Young Researchers” (Grant No. 2019-1691, LS), NIH R01HL160863 (BMR), American Heart Association 848898, Leducq Foundation for Cardiovascular Research grant 18CVD05 (LC, JES, PJS, BMR), and partially supported by the Italian Ministry of Health (LS, LC, FD, MCK, CS, PJS). Also, we are grateful to Drs. Silvia Castelletti and Sabrina Bedogni for their help in the initial data collection.

Web Resources

OMIM, http://www.omim.org/

RCSB Protein Data Bank, http://www.rcsb.org/pdb/home/home.do/

Variant Browser, https://variantbrowser.org/

Data and Code Availability

All code used to generate the results are available at https://github.com/kroncke-lab.

Author Contributions

Conceptualization: MJO, BMK; Data curation: MJO, LS, ID, YW, KK, LC, FD, MCK, CS, AL, JES, AK, SO, MH, FE, PJS, BMK; Formal Analysis: MJO, LS, ID, YW, KK, RT, DMR, MH, FE, PSJ, BMK; Supervision: BMK; Visualization: MJO, BMK; Writing-original draft: MJO, BMK; Writing-review & editing: MJO, LS, ID, YW, KK, LC, MCK, CS, AL, JES, AK, SO, RT, DMR, MH, FE, PJS, BMK.

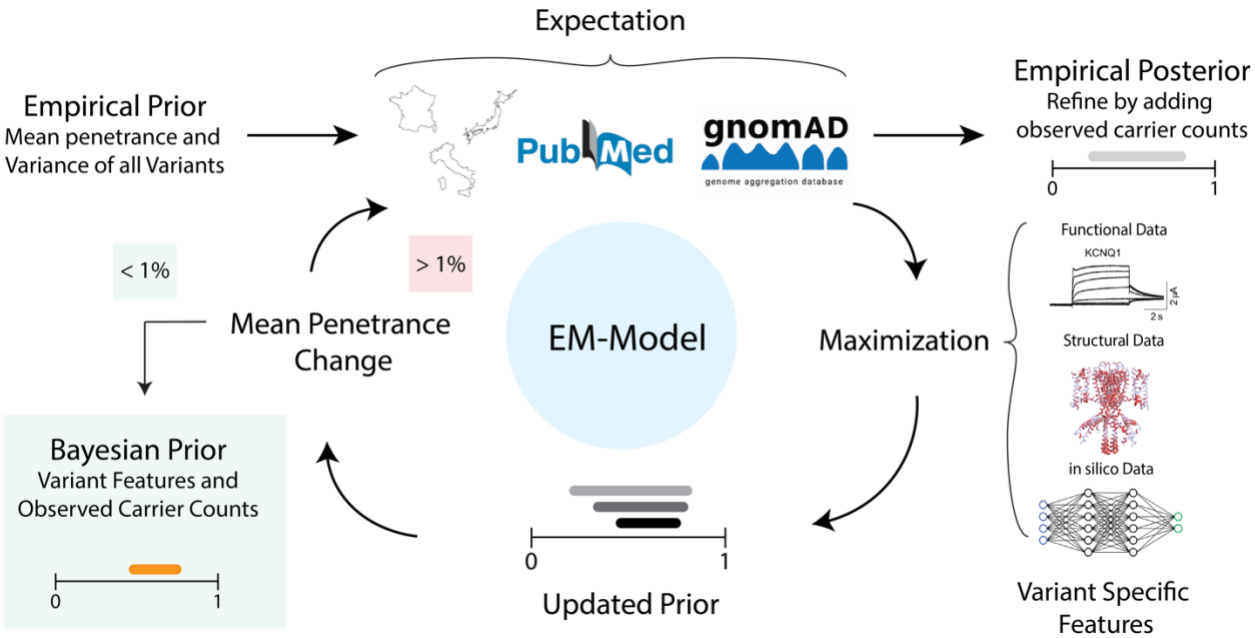


Figure 1. Schematic of Bayesian approach

An empirical prior is derived from the average penetrance and variance of all variants. We add variant specific counts to derive an empirical posterior for each variant. Within the context of an Expectation-Maximization model, we update the empirical posterior with variant specific features. This provides an updated prior, which is further iterated with clinical data and variant specific features until the penetrance does not continue to change. This outputs a Bayesian prior used in future analyses.

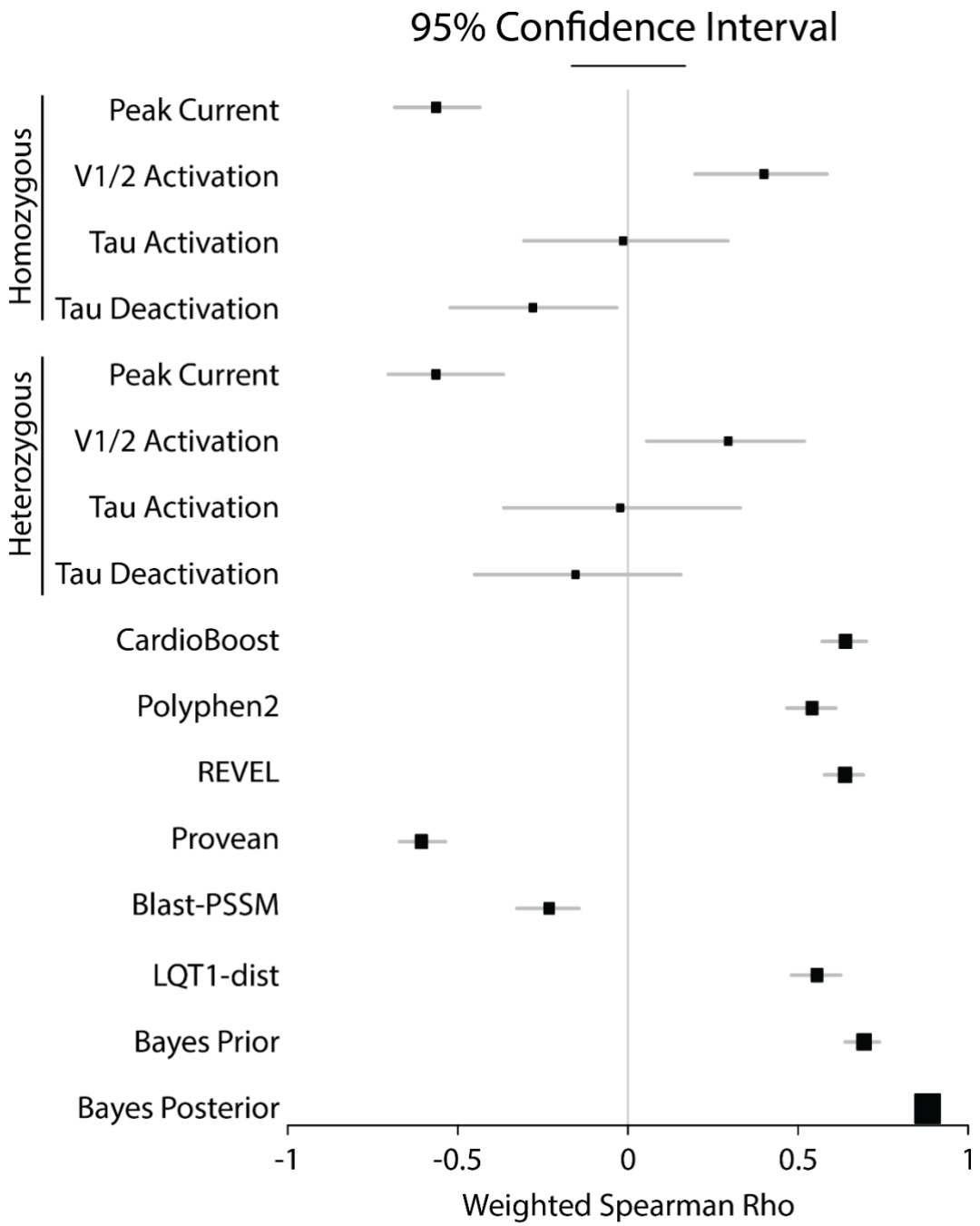


Figure 2. LQT1 Bayesian Model Covariate Evaluation

Weighted Spearman Rho correlations and 95% confidence intervals for each covariate against LQT1 empirical penetrance.

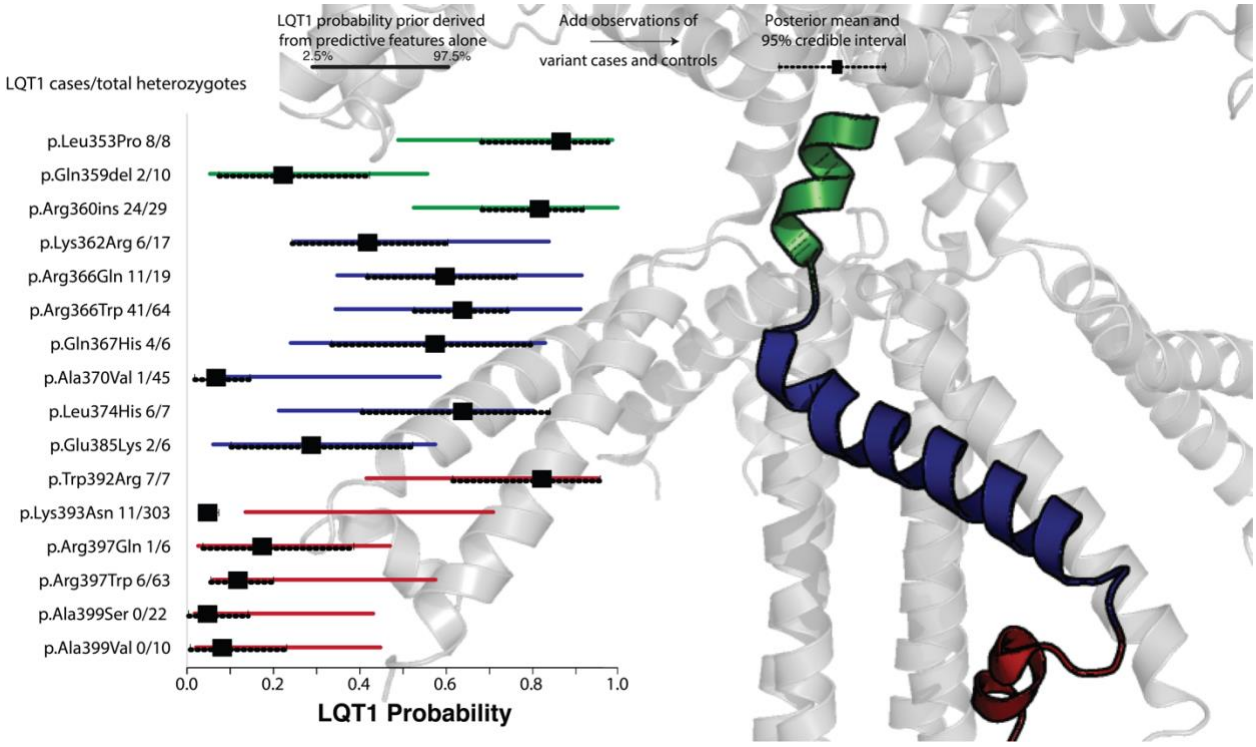
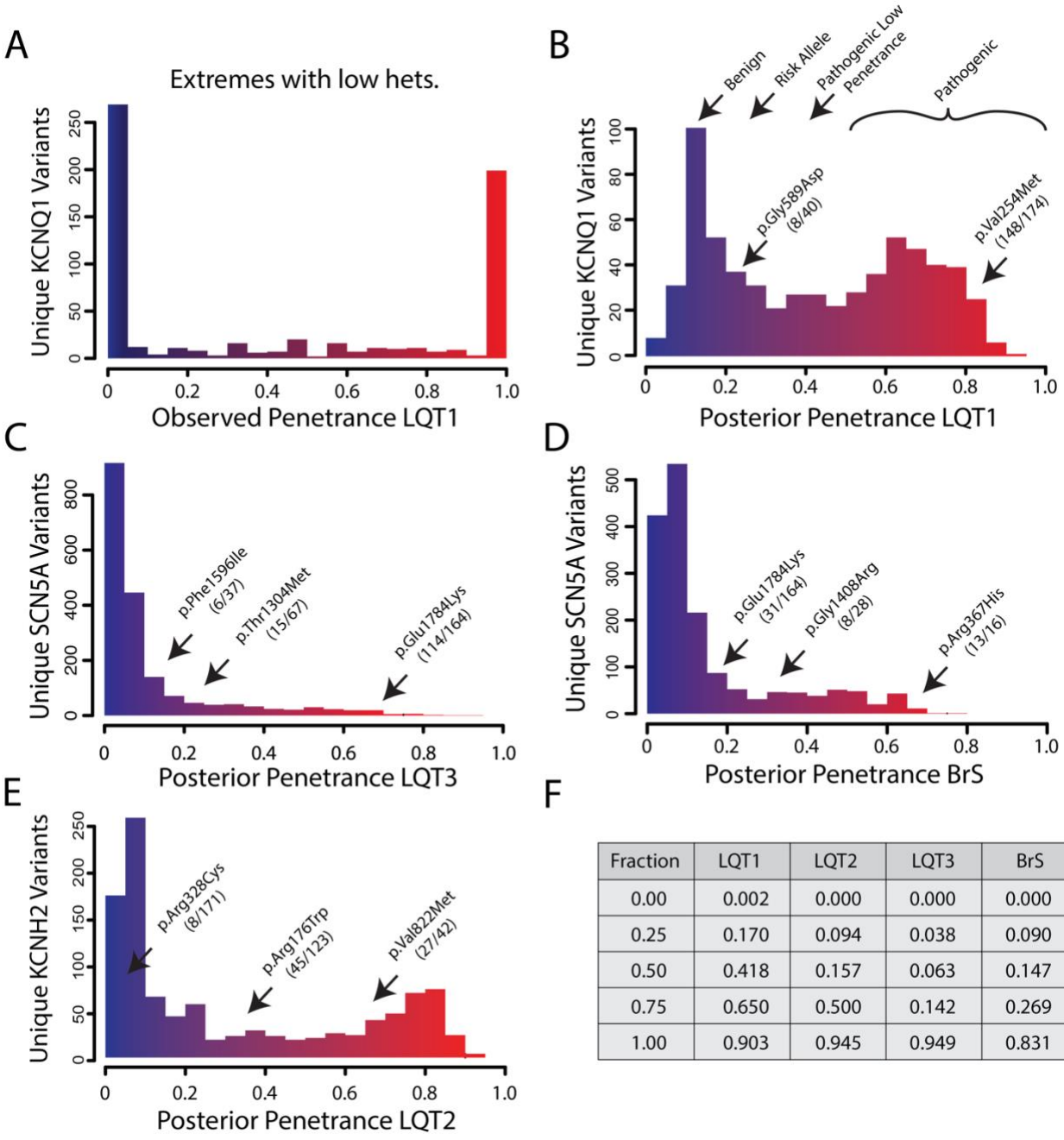


Figure 3. LQT1 Penetrance Estimate Forest Plots and Structure

Depiction of prior and posterior penetrance estimates for residues 350-400 (S6 transmembrane segment into the C-terminus) with >5 carriers. The prior penetrance estimates and 95% credible interval are depicted in gold, with the posterior and updated 95% credible intervals in black. Total number of observed heterozygotes are given. 95% credible interval changes proportionally with the number of clinical observations.

476



Fraction	LQT1	LQT2	LQT3	BrS
0.00	0.002	0.000	0.000	0.000
0.25	0.170	0.094	0.038	0.090
0.50	0.418	0.157	0.063	0.147
0.75	0.650	0.500	0.142	0.269
1.00	0.903	0.945	0.949	0.831

477

478 **Figure 4. Histograms of Variant Penetrance in 4 Mendelian Channelopathies**

479 A) Observed penetrance frequencies for *KCNQ1*-LQT1 variants. Note bimodal distribution of

480 penetrance frequencies in the observed penetrance.

481 B) Frequencies of Bayesian posterior penetrance estimates for *KCNQ1*-LQT1. Labels indicate

482 clinical 'gestalts' for different ranges of penetrance fractions to map ACMG classifications.

C) Bayesian posterior frequencies for *SCN5A*-LQT3. This distribution has the lowest penetrance estimates with few very highly penetrant variants.

D) Bayesian posterior frequencies for *SCN5A*-BrS. Covariates predict a more left-skewed distribution with a mild bimodal distribution. The highest decile of observations has less penetrance than those of LQT1 and LQT2.

E) Bayesian posterior frequencies for *KCNH2*-LQT2. Bimodal distribution of penetrance frequencies.

F) Quantile table of distributions of posterior penetrance fractions for each genotype-phenotype pair.

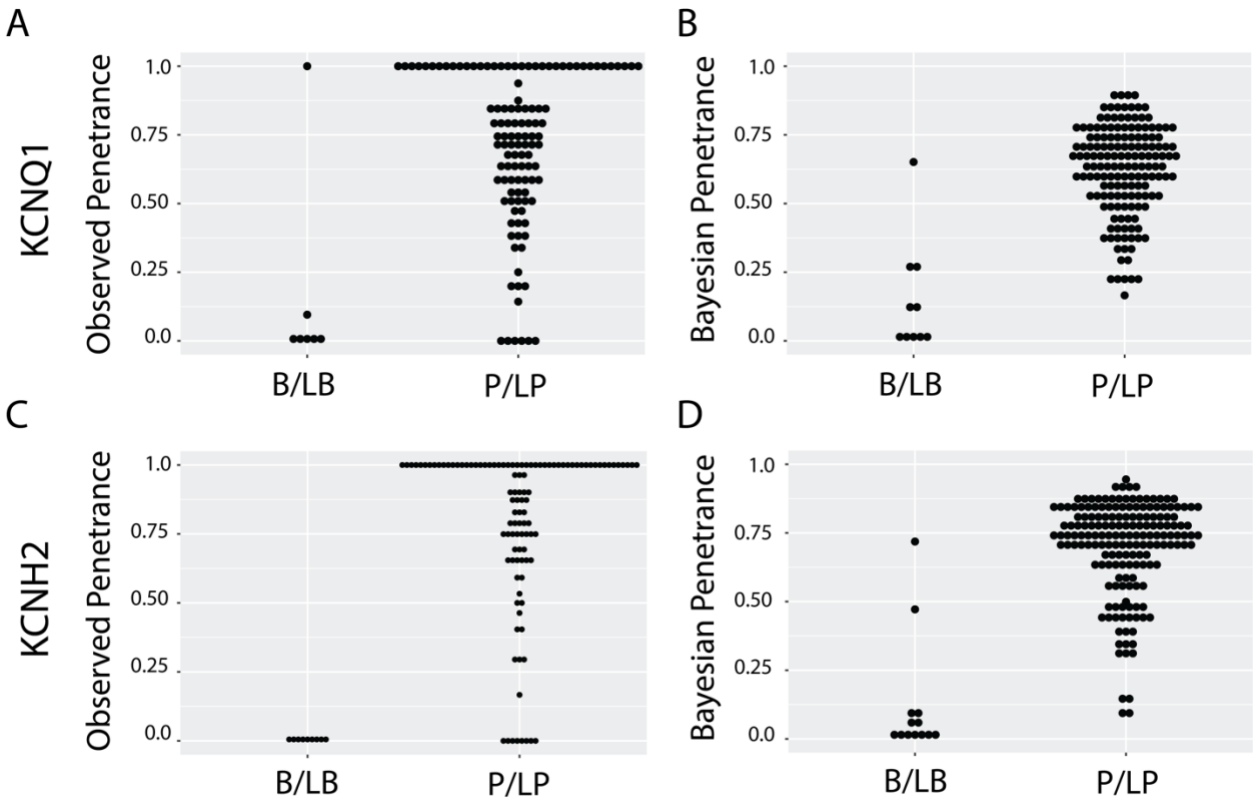


Figure 5. Bayesian and Observed Penetrance Associated with ClinVar Annotations.

A) Plotting the observed penetrance by ClinVar category shows that many *KCNQ1* P/LP variants are incompletely penetrant.

B) Most ClinVar Benign and Likely Benign *KCNQ1* variants have low Bayesian penetrance estimates. Notably, Pathogenic and Likely Pathogenic variants have a wide distribution of Bayesian penetrance.

C) P/LP *KCNH2* variants have a wide distribution of Bayesian penetrance scores in our model.

D) Several ClinVar P/LP variants within *KCNH2* are incompletely penetrant.

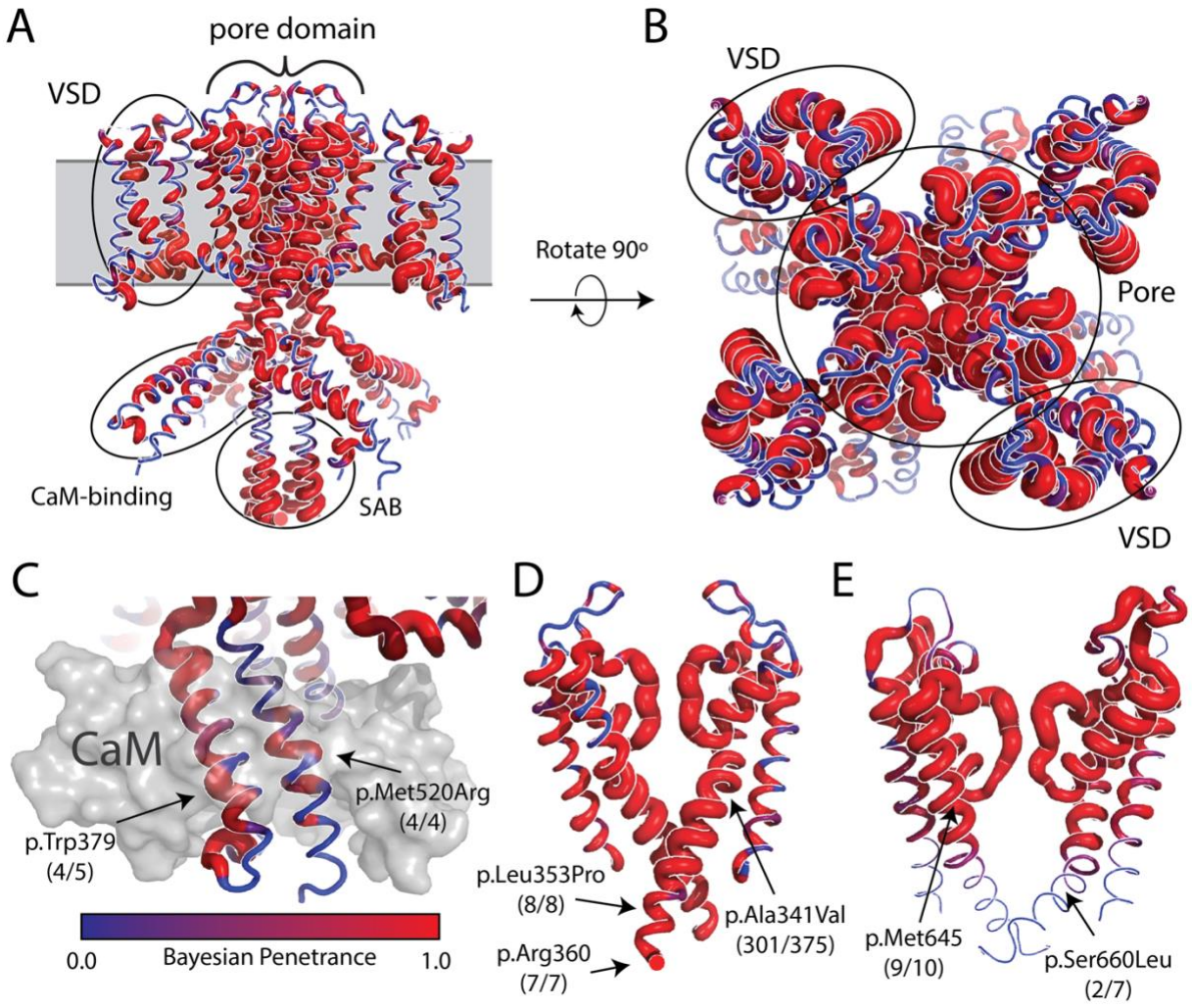


Figure 6. Structural Basis of *KCNQ1*-LQT1 Penetrance and 'Hot-Spot' Domain Heterogeneity

A) Posterior penetrance estimates imposed upon the tetrameric experimental $K_V7.1$ structure. The subassembly domain (SAB) and pore domain have a high degree of penetrance. Calmodulin (CaM) binding domains and Voltage Sensing Domains (VSD) also exhibit high, although heterogenous, penetrance.

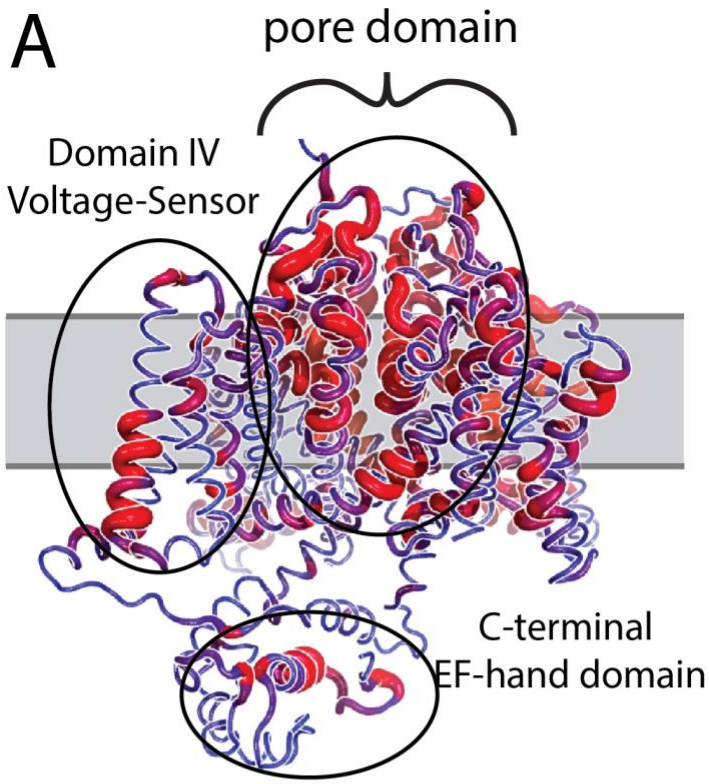
B) View of channel embedded in membrane from extracellular perspective. We observe very high penetrance in the channel pore and along VSD.

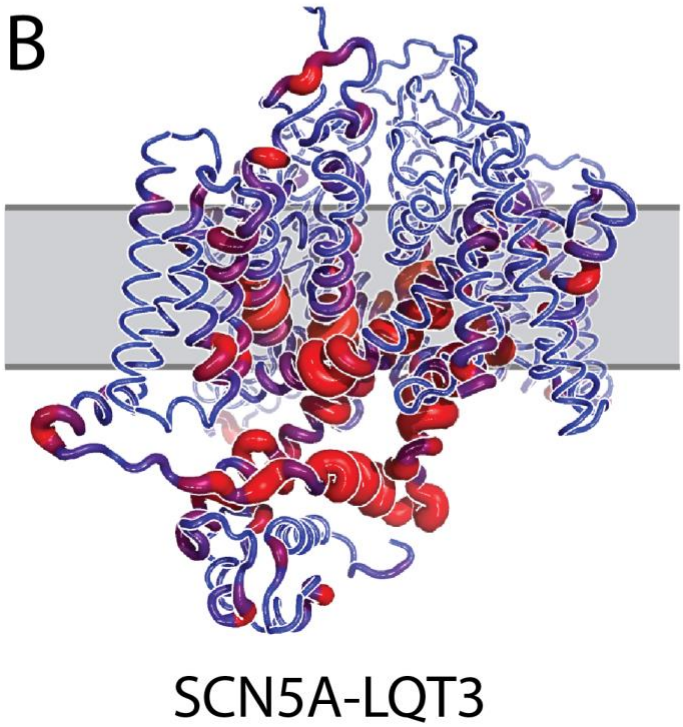
C) Close up of calmodulin binding domain with experimentally observed Calmodulin binding. Residues with high contact show enlarged penetrance estimates.

D) $K_V7.1$ S5-S6 helices with nearly uniform high penetrance estimates. Example variants provided.

E) $K_V11.1$ S5-S6 helices. Penetrance tapers along the S6 transmembrane domain compared to that observed for $K_V7.1$.

519





520

Figure 7. Structural heterogeneity of BrS and LQT3 penetrance

A) Bayesian posterior estimates for *SCN5A*-BrS. High degree of penetrance along the channel pore and voltage sensing domain (VSD).

B) Bayesian posterior estimates for *SCN5A*-LQT3. Fewer residues show high penetrance for LQT3 compared to BrS, consistent with the molecular defects precipitating each disease. Some overlap is observed in the C-terminal EF Hand domain, whereas variants in the pore are highly tolerated compared to BrS.

529 **Table 1.** Brier Scores for KCNQ1-LQT1 Model.

Covariate	Brier Score
CardioBoost	0.19
Polyphen2	0.22
REVEL (unadjusted)	0.17
LQT1-Dist	0.062
Bayesian Prior	0.028
Bayesian Posterior	0.010

530

References

1 Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am J Hum Genet* **104**, 275-286 (2019). https://doi.org:10.1016/j.ajhg.2018.12.015

2 McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210-217 (2010). https://doi.org:10.1016/j.cell.2010.03.032

3 Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* **132**, 1077-1130 (2013). https://doi.org:10.1007/s00439-013-1331-2

4 Richards, S. *et al.* 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* **17**, 405-424 (2015). https://doi.org:10.1038/gim.2015.30

5 Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980-985 (2014). https://doi.org:10.1093/nar/gkt1113

6 Schwartz, P. J. *et al.* Inherited cardiac arrhythmias. *Nat Rev Dis Primers* **6**, 58 (2020). https://doi.org:10.1038/s41572-020-0188-7

7 Hosseini, S. M. *et al.* Reappraisal of Reported Genes for Sudden Arrhythmic Death: Evidence-Based Evaluation of Gene Validity for Brugada Syndrome. *Circulation* **138**, 1195-1205 (2018). https://doi.org:10.1161/circulationaha.118.035070

8 Adler, A. *et al.* An International, Multicentered, Evidence-Based Reappraisal of Genes Reported to Cause Congenital Long QT Syndrome. *Circulation* **141**, 418-428 (2020). https://doi.org:10.1161/CIRCULATIONAHA.119.043132

9 Priori, S. G., Napolitano, C. & Schwartz, P. J. Low penetrance in the long-QT syndrome: clinical impact. *Circulation* **99**, 529-533 (1999). https://doi.org:10.1161/01.cir.99.4.529

10 Lahrouchi, N. *et al.* Transethnic Genome-Wide Association Study Provides Insights in the Genetic Architecture and Heritability of Long QT Syndrome. *Circulation* **142**, 324-338 (2020). https://doi.org:10.1161/circulationaha.120.045956

11 Wijeyeratne, Y. D. *et al.* SCN5A Mutation Type and a Genetic Risk Score Associate Variably With Brugada Syndrome Phenotype in SCN5A Families. *Circ Genom Precis Med* **13**, e002911 (2020). https://doi.org:10.1161/circgen.120.002911

12 Nauffal, V. *et al.* Monogenic and Polygenic Contributions to QTc Prolongation in the Population. *medRxiv*, 2021.2006.2018.21258578 (2021). https://doi.org:10.1101/2021.06.18.21258578

13 van Rees, J. B. *et al.* Inappropriate implantable cardioverter-defibrillator shocks: incidence, predictors, and impact on mortality. *J Am Coll Cardiol* **57**, 556-562 (2011). https://doi.org:10.1016/j.jacc.2010.06.059

14 Schwartz, P. J. *et al.* Who are the long-QT syndrome patients who receive an implantable cardioverter-defibrillator and what happens to them?: data from the European Long-QT Syndrome Implantable Cardioverter-Defibrillator (LQTS ICD) Registry. *Circulation* **122**, 1272-1282 (2010). https://doi.org:10.1161/circulationaha.110.950147

15 Larrea-Sebal, A. *et al.* MLb-LDLr: A Machine Learning Model for Predicting the Pathogenicity of LDLr Missense Variants. *JACC Basic Transl Sci* **6**, 815-827 (2021). https://doi.org:10.1016/j.jacbts.2021.08.009

16 Draelos, R. L. *et al.* GENESIS: Gene-Specific Machine Learning Models for Variants of Uncertain Significance Found in Catecholaminergic Polymorphic Ventricular Tachycardia and Long QT Syndrome-Associated Genes. *Circ Arrhythm Electrophysiol*, 101161circep121010326 (2022). https://doi.org:10.1161/circep.121.010326

17 Kozek, K. *et al.* Estimating the Post-Test Probability of Long QT Syndrome Diagnosis for Rare KCNH2 Variants. *Circ Genom Precis Med* (2021). https://doi.org:10.1161/circgen.120.003289

18 Kroncke, B. M. *et al.* A Bayesian method to estimate variant-induced disease penetrance. *PLoS Genet* **16**, e1008862 (2020). https://doi.org:10.1371/journal.pgen.1008862

19 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org:10.1038/s41586-020-2308-7

20 Kroncke, B. M. *et al.* Protein structure aids predicting functional perturbation of missense variants in SCN5A and KCNQ1. *Comput Struct Biotechnol J* **17**, 206-214 (2019). https://doi.org:10.1016/j.csbj.2019.01.008

21 Zhang, X. *et al.* Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genet Med* **23**, 69-79 (2021). https://doi.org:10.1038/s41436-020-00972-3

22 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016). https://doi.org:10.1016/j.ajhg.2016.08.016

23 Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745-2747 (2015). https://doi.org:10.1093/bioinformatics/btv195

24 BRIER, G. W. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review* **78**, 1-3 (1950). https://doi.org:10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

25 Wilde, A. A. M. & Amin, A. S. Clinical Spectrum of SCN5A Mutations: Long QT Syndrome, Brugada Syndrome, and Cardiomyopathy. *JACC Clin Electrophysiol* **4**, 569-579 (2018). https://doi.org:10.1016/j.jacep.2018.03.006

26 Kapa, S. *et al.* Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation* **120**, 1752-1760 (2009). https://doi.org:10.1161/CIRCULATIONAHA.109.863076

27 Sun, J. & MacKinnon, R. Cryo-EM Structure of a KCNQ1/CaM Complex Reveals Insights into Congenital Long QT Syndrome. *Cell* **169**, 1042-1050.e1049 (2017). https://doi.org:10.1016/j.cell.2017.05.019

28 Schwartz, P. J. *et al.* Mutation location and IKs regulation in the arrhythmic risk of long QT syndrome type 1: the importance of the KCNQ1 S6 region. *Eur Heart J* **42**, 4743-4755 (2021). https://doi.org:10.1093/eurheartj/ehab582

29 Sun, J. & MacKinnon, R. Structural Basis of Human KCNQ1 Modulation and Gating. *Cell* **180**, 340-347.e349 (2020). https://doi.org:10.1016/j.cell.2019.12.003

30 Li, Z. *et al.* Structure of human Na(v)1.5 reveals the fast inactivation-related segments as a mutational hotspot for the long QT syndrome. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org:10.1073/pnas.2100069118

31 Napolitano, C. *et al.* Genetic testing in the long QT syndrome: development and validation of an efficient approach to genotyping in clinical practice. *Jama* **294**, 2975-2980 (2005). https://doi.org:10.1001/jama.294.23.2975

32 Barc, J. *et al.* Genome-wide association analyses identify new Brugada syndrome risk loci and highlight a new mechanism of sodium channel regulation in disease susceptibility. *Nat Genet* **54**, 232-239 (2022). https://doi.org:10.1038/s41588-021-01007-6

33 Yagi, N. *et al.* A challenge for mutation specific risk stratification in long QT syndrome type 1. *J Cardiol* **72**, 56-65 (2018). https://doi.org:10.1016/j.jjcc.2017.12.011

34 Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* **19**, 1151-1158 (2017). https://doi.org:10.1038/gim.2017.26

35 Lane, C. M. *et al.* Long QT syndrome type 5-Lite: Defining the clinical phenotype associated with the potentially proarrhythmic p.Asp85Asn-KCNE1 common genetic variant. *Heart Rhythm* **15**, 1223-1230 (2018). https://doi.org:10.1016/j.hrthm.2018.03.038

36 Giudicessi, J. R., Roden, D. M., Wilde, A. A. M. & Ackerman, M. J. Classification and Reporting of Potentially Proarrhythmic Common Genetic Variation in Long QT Syndrome Genetic Testing. *Circulation* **137**, 619-630 (2018). https://doi.org:10.1161/circulationaha.117.030142

37 Schwartz, P. J., Crotti, L. & George, A. L., Jr. Modifier genes for sudden cardiac death. *Eur Heart J* **39**, 3925-3931 (2018). https://doi.org:10.1093/eurheartj/ehy502

38 Wada, Y. *et al.* Common Ancestry-specific Ion Channel Variants Predispose to Drug-induced Arrhythmias. *Circulation* (2022). https://doi.org:10.1161/circulationaha.121.054883

39 Kapplinger, J. D. *et al.* Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. *Heart Rhythm* **6**, 1297-1303 (2009). https://doi.org:10.1016/j.hrthm.2009.05.021

40 Walsh, R. *et al.* Enhancing rare variant interpretation in inherited arrhythmias through quantitative analysis of consortium disease cohorts and population controls. *Genet Med* **23**, 47-58 (2021). https://doi.org:10.1038/s41436-020-00946-5

41 Ng, C. A. *et al.* A massively parallel assay accurately discriminates between functionally normal and abnormal variants in a hotspot domain of KCNH2. *Am J Hum Genet* (2022). https://doi.org:10.1016/j.ajhg.2022.05.003

42 Crotti, L. *et al.* Calmodulin mutations and life-threatening cardiac arrhythmias: insights from the International Calmodulinopathy Registry. *Eur Heart J* **40**, 2964-2975 (2019). https://doi.org:10.1093/eurheartj/ehz311

43 Kang, P. W. *et al.* Calmodulin acts as a state-dependent switch to control a cardiac potassium channel opening. *Sci Adv* **6** (2020). https://doi.org:10.1126/sciadv.abd6798

44 Clerx, M., Heijman, J., Collins, P. & Volders, P. G. A. Predicting changes to I(Na) from missense mutations in human SCN5A. *Sci Rep* **8**, 12797 (2018). https://doi.org:10.1038/s41598-018-30577-5

45 Heyne, H. O. *et al.* Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci Transl Med* **12** (2020). https://doi.org:10.1126/scitranslmed.aay6848

46 Parikh, V. N. Promise and Peril of Population Genomics for the Development of Genome-First Approaches in Mendelian Cardiovascular Disease. *Circ Genom Precis Med* **14**, e002964 (2021). https://doi.org:10.1161/circgen.120.002964

47 Gnecchi, M., Sala, L. & Schwartz, P. J. Precision Medicine and cardiac channelopathies: when dreams meet reality. *Eur Heart J* **42**, 1661-1675 (2021). https://doi.org:10.1093/eurheartj/ehab007

48 Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628-634 (2021). https://doi.org:10.1038/s41586-021-04103-z

49 Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021). https://doi.org:10.1038/s41586-021-03205-y

50 Weile, J. & Roth, F. P. Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum Genet* **137**, 665-678 (2018). https://doi.org:10.1007/s00439-018-1916-x

51 Hille, B. *Ion channels of excitable membranes*. 3rd edn, (Sinauer, 2001).

52 Glazer, A. M. *et al.* Arrhythmia Variant Associations and Reclassifications in the eMERGE-III Sequencing Study. *Circulation* (2021). https://doi.org:10.1161/circulationaha.121.055562

53 Renaux, A. *et al.* ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res* **47**, W93-w98 (2019). https://doi.org:10.1093/nar/gkz437