ORIGINAL ARTICLE

Paralogue annotation identifies novel pathogenic variants in patients with Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia

Roddy Walsh,1 Nicholas S Peters,2 Stuart A Cook,2,3,4 James S Ware1,2

ABSTRACT

Background Distinguishing genetic variants that cause disease from variants that are rare but benign is one of the principal challenges in contemporary clinical genetics, particularly as variants are identified at a pace exceeding the capacity of researchers to characterise them functionally.

Methods We previously developed a novel method, called paralogue annotation, which accurately and specifically identifies disease-causing missense variants by transferring disease-causing annotations across families of related proteins. Here we refine our approach, and apply it to novel variants found in 2266 patients across two large cohorts with inherited sudden death syndromes, namely catecholaminergic polymorphic ventricular tachycardia (CPVT) or Brugada syndrome (BrS).

Results Over one third of the novel non-synonymous variants found in these studies, which would otherwise be reported in a clinical diagnostics setting as ‘variants of unknown significance’, are categorised by our method as likely disease causing (positive predictive value 98.7%). This identified more than 500 new disease loci for BrS and CPVT.

Conclusions Our methodology is widely transferable across all human disease genes, with an estimated 150 000 potentially informative annotations in more than 1800 genes. We have developed a web resource that allows researchers and clinicians to annotate variants found in individuals with inherited arrhythmias, comprising a referenced compendium of known missense variants in these genes together with a user-friendly implementation of our approach. This tool will facilitate the interpretation of many novel variants that might otherwise remain unclassified.

INTRODUCTION

Inherited arrhythmias such as long QT syndrome (LQTS), Brugada syndrome (BrS) and catecholaminergic polymorphic ventricular tachycardia (CPVT) are life-threatening diseases, caused predominantly by genetic variation in ion channels. In BrS, loss-of-function mutations in the SCN5A-encoded cardiac sodium channel (MIM 601144) have been shown to be responsible for 15–30% of cases1 2 with mutations in other minor genes accounting for a proportion of remaining cases. In CPVT, gain-of-function mutations in the cardiac ryanodine receptor encoded by RYR2 (MIM 604772) are responsible for at least 50% of cases.3

With on-going developments in DNA sequencing technology, it is expected that clinical genetic testing will become very widely available. However, it is increasingly appreciated that many healthy individuals carry rare variants in disease-associated genes (3–6% for genes associated with inherited arrhythmia syndromes),4 5 and that additional information is required to determine whether a novel variant identified during genetic testing is pathogenic. This is particularly the case for missense variants (single amino acid substitutions) caused by single nucleotide variants.

In order to determine whether a rare variant found in a patient is disease causing, sufficiently powered segregation analysis or functional biochemical characterisation are ideally performed. However, these are often impractical due to cost and time constraints, or a lack of phenotypically characterised family members for segregation studies. Several in silico algorithms, such as SIFT and Polyphen6 7 try to predict the effect of novel variants, based on the conservation and physicochemical properties of the variant amino acid, and variants in certain protein regions and domains have a high probability of pathogenicity.8 9

However, more evidence is needed to classify variants with sufficient confidence for clinical application.

We have recently proposed a new method to analyse novel missense variants and assess their likelihood of pathogenicity.10 This method, ‘paralogue annotation’, identifies functionally important residues that are intolerant of variation across families of evolutionarily related proteins (paralogues), using clinically ascertained human genotype-phenotype relationships. By aligning the protein sequences of members of these protein families, we can identify amino acids that are functionally equivalent across the different proteins. A variant that is known to be pathogenic in one member of a protein family can then be used to annotate the equivalent amino acid of other members of the family for which no clinical genetic information exists (figure 1). For example, if a missense variant in RYR1 alters protein function and causes malignant hyperthermia when expressed in skeletal muscle, then we hypothesise that a novel variant affecting the equivalent amino acid in RYR2, expressed in cardiac muscle, is likely to be disease causing in a patient with CPVT.

To cite: Walsh R, Peters NS, Cook SA, et al. J Med Genet: first published as 10.1136/jmedgenet-2013-101917 on 17 October 2013. Downloaded from http://jmg.bmj.com/ on December 20, 2023 by guest. Protected by copyright.
This approach was developed and experimentally validated by application to a large set of known variants in eight LQTS genes, and was found to have a positive predictive value (PPV) of 98.4% in these genes. Here we present novel refinements to increase accuracy, and apply this approach in two large cohorts of patients with BrS or CPVT to determine whether it provides additional useful information in a clinical diagnostic setting. We also report a web application that allows researchers and clinicians to use paralogue annotation to interrogate novel variants in arrhythmia genes.

MATERIALS AND METHODS

Identification of variants in paralogues of RYR2 and SCNSA

Paralogues of RYR2 and SCNSA, that is, evolutionarily related genes with homology in sequence and structure, were identified using the IUPHAR database on receptor nomenclature and through homology searches (Blastp searches of the ENST00000328968 (Ensembl), LRG_289t1/LRG_289p1 (Locus Reference Genomic), NM_198056/NP_932173 (Refseq), ENST00000366574/ENSP0000035533 (Ensembl), LRG_402t1/LRG_402p1 (Locus Reference Genomic) and SCN5A: ENST00000366574/ENSP0000035533 (Ensembl), LRG_289t1/LRG_289p1 (Locus Reference Genomic)).

RYR2 has two paralogues, RYR1 and RYR3, which were included in the sequence alignment. Although RYR3 is not implicated in human disease it was included to improve the alignment quality for the protein family. Variants in RYR1 cause malignant hyperthermia, central core disease, multi-minicore disease and congenital myopathy. There are 329 RYR1 missense variants in the Human Gene Mutation Database (HGMD) affecting 284 distinct amino acid residues.

For SCNSA, 19 paralogues were included: nine voltage-gated sodium channels, and 10 voltage-gated calcium channels that show strong sequence and structural homology with the sodium channels. Fourteen of these 19 paralogues are associated with Mendelian disease in humans, typically attributable to altered cellular electrophysiology in the tissue where the paralogue is expressed, such as epilepsy, myotonia, pain disorders, night blindness and hemiplegic migraine. SCNSA paralogues and their disease associations are shown in Table 1.

For each paralogue gene, variants reported as pathogenic were identified using the HGMD Professional V2012.3. Only disease-causing missense mutations, that is, single nucleotide variants causing a single non-terminal amino acid acid change, were considered.

Multiple sequence alignment and paralogue annotation

The protein sequences of RYR2 and SCNSA and their respective paralogues were aligned using the T-Coffee/M-Coffee alignment packages. These packages combine the output of a number of alignment algorithms into a single consensus alignment and provide a consensus score at each point in the alignment (0–9), which is a measure of the reliability of the alignment at each amino acid residue. Using these alignments, each paralogue protein residue with a disease-causing variant was mapped onto the equivalent amino acid residue in RYR2 and SCNSA.

To distinguish aligned amino acids that are truly functionally equivalent from alignment artefacts, an amino acid alignment quality classification was devised. Mappings were classified as high quality if the reference amino acid was conserved between paralogues, and the alignment consensus score was greater than 3. Medium quality mappings required conservation of the reference amino acid or a consensus score greater than 3 or more than one paralogue variant mapping to the same residue. Low quality mappings did not meet these criteria, and were excluded from analyses.

Sequencing and paralogue annotation in cohorts with CPVT and BrS

Sequencing of two large cohorts of unrelated patients with inherited arrhythmias has previously been reported: all exons of RYR2 were sequenced in 155 patients with CPVT and 200 healthy controls, and all exons of SCNSA were sequenced in 2111 patients with BrS and 1300 healthy controls. Paralogue annotation was applied to novel missense variants identified in each cohort. Variants were classified as ‘novel’ if they had not
previously been reported at the time of the original report, and have not subsequently been confirmed in an independent study. These would conventionally be reported clinically as ‘variants of unknown significance’, pending segregation and/or functional molecular characterisation.

**Mutation status of RYR2 and SCN5A residues**

Previous reports of benign and pathogenic variation in RYR2 and SCN5A were collated. HGMD identified putative disease-causing variants in both genes. Variants were also extracted from dbSNP V137 as part of the Ensembl 69 release. Population frequency from 1000 Genomes (phase 1) and other datasets in dbSNP were also used—any variant with more than one observation and a minor allele frequency (MAF) of greater than or equal to 0.01 was considered to be benign, as this is considered incompatible with the population frequencies of BrS and CPVT. In addition, several large published datasets of RYR2 and SCN5A pathogenic and benign variants were assessed.

Residues with one or more missense variants reported as disease causing were classified as pathogenic. Residues with reported benign missense variants or high MAF scores as described above were classified benign. Any residue with conflicting benign and pathogenic reports (for the same variant or different variants affecting the same residue), or dbSNP variants with no population frequency data or MAF of less than 0.01, were classified as uncertain. Residues for which no missense variants were known were classified as unannotated.

**Protein domains of RYR2 and SCN5A**

Mutation hotspots in RYR2 were defined according to Yano et al.9 residues 77–466 (N-terminal hotspot), 2246–2534 (central hotspot) and 3778–4959 (channel hotspot). The SCN5A domains were defined in accordance with the Uniprot entry Q14524 as used by Kapa et al.8 transmembrane regions are between residues 127–415, 712–939, 1201–1470 and 1524–1772.

**RESULTS**

Paralogue annotation is informative in assessing novel missense variants in clinical cohorts

Variants identified in two large cohorts of unrelated patients with inherited arrhythmias were analysed. RYR2 sequencing in 155 CPVT patients identified 63 distinct variants that were absent in 200 healthy control references. Of these, 59 were missense variants, and 31 were not reported in any other previous or subsequent cases. SCN5A sequencing 2111 BrS patients identified 293 distinct SCN5A variants that were absent in 1300 healthy controls. Of these, 193 were missense variants, and 122 were not reported in any other previous or subsequent cases. Paralogue annotation was applied to these 31 RYR2 and 122 SCN5A variants, which would otherwise be reported clinically as ‘variants of unknown significance’, to determine whether paralogue mutation annotation is informative and provides further evidence to reclassify some of these variants as ‘disease causing’ (figure 2).

Ten of the 31 novel RYR2 missense variants (32.3%) affected amino acid residues that were annotated with paralogue mutations, that is, variants in the equivalent residue in RYR1 have been shown to be disease causing for other human diseases (table 2). Five of these 10 residues have more than one reported RYR1 pathogenic mutation. Three of these variants occurred in the N-terminal hotspot (residues 77–466), five occurred in the channel region hotspot (residues 3778–4959) and two occurred outside the recognised RYR2 mutation hotspots—A549 V and H2168Q. In addition, five of the 29 non-novel RYR2 missense

| Paralogue | HGMD missense mutations | HGMD missense residues | SCN5A annotations | Major diseases |
|-----------|-------------------------|------------------------|------------------|---------------|
| SCN1A     | 410                     | 327                    | 393              | Epilepsy, Dravet syndrome, Hemiplegic migraine |
| SCN2A     | 26                      | 26                     | 24               | Epilepsy, Neonatal-infantile seizures |
| SCN3A     | 1                       | 1                      | 1                | Epilepsy |
| SCN4A     | 67                      | 53                     | 67               | Hyperkalaemic periodic paralysis, myotonia, paramyotoniacongenita, periodic paralysis |
| SCN7A     | 0                       | 0                      |                  |               |
| SCN8A     | 1                       | 1                      | 1                | Infantile epileptic encephalopathy |
| SCN9A     | 53                      | 50                     | 45               | Congenital indifference to pain, primary erythromelalgia, paroxysmal extreme pain disorder, Small fibre neuropathy |
| SCN9OA    | 0                       | 0                      |                  |               |
| SCN1A     | 0                       | 0                      |                  |               |
| SCN1A     | 0                       | 0                      |                  |               |
| CACNA1A   | 63                      | 61                     | 54               | Episodic ataxia 2, hemiplegic migraine |
| CACNA1B   | 0                       | 0                      |                  |               |
| CACNA1C   | 9                       | 9                      | 4                | BrS, LQTS |
| CACNA1D   | 1                       | 1                      | 0                |               |
| CACNA1E   | 1                       | 1                      | 0                |               |
| CACNA1F   | 28                      | 28                     | 27               | Night blindness |
| CACNA1G   | 2                       | 2                      | 0                | Juvenile myoclonic epilepsy |
| CACNA1H   | 25                      | 25                     | 13               | Epilepsy, autism spectrum disorder |
| CACNA1I   | 0                       | 0                      |                  |               |
| CACNA1S   | 13                      | 8                      | 13               | Hypokalaemic periodic paralysis, malignant hyperthermia |

Nineteen paralogues of SCN5A were used in this study, including voltage-gated sodium channels (SCN-A) and the homologous voltage-gated calcium channels. For each paralogue, the table shows the total number of distinct missense variants reported in HGMD, the number of distinct amino acid residues affected by these mutations and the number of mutations that were mapped to equivalent residues in SCN5A. Some of the most prominent diseases associated with these paralogues are highlighted; these are typically diseases attributable to abnormalities of membrane excitability in a range of tissues. BrS, Brugada syndrome; HGMD, Human Gene Mutation Database; LQTS, long QT syndrome.
various identified in the study also had equivalent paralogue mutations, adding further evidence to the likely pathogenicity of these variants (see supplementary table S1, available online only).

In SCN5A, 45 of the 122 novel missense variants (36.9%) were annotated with paralogue mutations (table 3). Fourteen of these 45 residues have more than one reported paralogue pathogenic mutation. Forty-one of these variants occurred in the transmembrane domains of SCN5A, three in the N-terminal region and one in the interlinker domains. In addition, 26 of the 70 non-novel SCN5A missense variants identified in the study also had equivalent paralogue mutations, adding further evidence to the likely pathogenicity of these variants (see supplementary table S2, available online only).

Three variants that are detected in healthy controls in these two studies are predicted to be pathogenic using our approach, and are considered false positives for the purpose of this study. RYR2-M2389L was identified in one of 200 controls. This variant is found in the central domain hotspot of RYR2 and is not identified in any of the 1000 Genomes phase 1 samples. SCN5A-D3596G was identified in one of 1300 controls. Located in interlinker domain I–II, this variant is also absent in the phase 1 1000 Genomes data. SCN5A-S216L has conflicting reports as to its pathogenicity—it has been linked to LQTS, atrial fibrillation and dilated cardiomyopathy,21–23 but has also been identified in healthy controls and in low frequencies in dbSNP populations. Its exact role in causing disease is therefore still unclear. Although we consider these variants as false positives here, in the absence of functional data to the contrary, at least some may be functionally deleterious in controls with non-penetrant disease or incomplete phenotype data. Importantly, controls did not undergo any provocation testing (eg, ajmaline challenge for BrS, and exercise test or adrenaline challenge for CPVT), which is used to make a diagnosis in some forms of these diseases and hence latent disease was not excluded.

Figure 2 Genetic variation in cohorts with inherited arrhythmia syndromes. RYR2 or SCN5A were sequenced in 2266 patients with an inherited arrhythmia syndrome, as shown. Three hundred and fifty-seven variants were identified in 511 patients, of which 153 were novel missense variants. In the absence of segregation or functional data, these would typically be reported as variants of unknown significance. However, paralogue annotation provided additional information for 65 (36%) variants, that would further inform a clinical genetic report. BrS, Brugada syndrome; CPVT, catecholaminergic polymorphic ventricular tachycardia.

Table 2 Paralogue annotation of novel variants identified in RYR2

| CDS variant | Protein | Region | Cases (n=155) | Exon | Paralogue | Paralogue variant | Paralogue disease | Consensus |
|-------------|---------|--------|--------------|------|-----------|------------------|------------------|-----------|
| c.527G>A    | R176Q   | N-terminal hotspot | 1 | 8 | RYR1 | R163L | Malignant hyperthermia | 9 |
| c.994C>T    | R332W   | N-terminal hotspot | 1 | 12 | RYR1 | R316L | Malignant hyperthermia | 8 |
| c.1069G>A   | R357S   | N-terminal hotspot | 1 | 13 | RYR1 | G341R | Malignant hyperthermia | 9 |
| c.1646C>T   | A549V   | Outside hotspots | 1 | 17 | RYR1 | A537T | Congenital myopathy | 9 |
| c.6504C>G   | H2168Q  | Outside hotspots | 2 | 42 | RYR1 | H2040Q | Multiminicore disease | 9 |
| c.7258A>T   | R2420W  | Central hotspot | 1 | 48 | RYR1 | R2454C | Malignant hyperthermia | 9 |
| c.11989A>G  | K3997E  | Channel hotspot | 1 | 90 | RYR1 | R401W | Malignant hyperthermia | 9 |
| c.14369G>A  | R4790Q  | Channel hotspot | 1 | 100 | RYR1 | R4861C | Central core disease | 9 |
| c.14414A>G  | K4805R  | Channel hotspot | 1 | 100 | RYR1 | R4861H | Central core disease | 9 |
| c.14465G>A  | R4822H  | Channel hotspot | 1 | 101 | RYR1 | R4893W | Central core disease | 9 |

Ten out of 31 novel missense variants identified in 155 unrelated CPVT patients were annotated. This provides strong evidence of pathogenicity for these variants. In addition, five of the 29 RYR2 missense variants previously reported to be pathogenic were annotated (see supplementary table S1, available online only).

RYR2 coordinates given with respect to transcripts NM_001035/NP_001026 (Refseq), ENST00000368574/ENSP00000355533 (Ensembl), LRG_402t1/LRG_402p1 (Locus Reference Genomics).

CDS, coding DNA sequence; CPVT, catecholaminergic polymorphic ventricular tachycardia.
Table 3  Paralogue annotation of novel variants identified in SCN5A

| CDS          | Protein             | Region          | Cases (n=2111) | Exon | Paralogue | Paralogue variant | Paralogue disease                      | Consensus |
|--------------|---------------------|-----------------|----------------|------|-----------|-------------------|----------------------------------------|-----------|
| c.278T>C     | p.F93S              | N-terminus      | 1              | 3    | SCN1A     | F90S              | Myoclonic epilepsy of infancy           | 5         |
| c.281T>G     | p.I95S              | N-terminus      | 1              | 3    | SCN1A     | I91T              | Myoclonic epilepsy of infancy           | 5         |
| c.362G>A     | p.A178G             | TM domain 1     | 1              | 5    | SCN1A     | A175V             | Dravet syndrome                        | 9         |
| c.4694A>T    | p.R367L             | TM domain 1     | 1              | 9    | SCN1A     | R377Q             | Generalised epilepsy with febrile seizures | 9         |
| c.1120T>G    | p.W374G             | Interdomain Linker I-II | 1 | 14 | SCN9A     | C699Y             | Childhood absence epilepsy              | 3         |
| c.1156G>A    | p.G386R             | TM domain 1     | 1              | 10   | SCN1A     | G396E             | Dravet syndrome                        | 8         |
| c.1187T>C    | p.W374G             | Interdomain Linker I-II | 1 | 14 | SCN9A     | C699Y             | Childhood absence epilepsy              | 3         |
| c.2047T>G    | p.C683G             | Interdomain Linker I-II | 1 | 14 | SCN9A     | C699Y             | Childhood absence epilepsy              | 3         |

Continued
Paralogue annotation has previously been validated in eight LQTS proteins. Sites predicted to be disease associated by paralogue annotation were highly enriched for known disease-causing variants, with very few false positives. Applying the approach to RYR2 and SCN5A variants again demonstrates enrichment for known disease-associated residues (table 4). Two hundred and seventy-five RYR2 residues were annotated with paralogue mutations: if distributed at random, eight would be expected to coincide with known pathogenic residues. We observe 35 known disease-associated residues: a 4.4-fold enrichment. Similarly, 113 disease-associated SCN5A residues are annotated by paralogue mutations, a 1.4-fold increase over the 81 annotations expected by random distribution. There is also a significant depletion of known benign residues in SCN5A (expected, six; observed, one). The positive predictive value (PPV) of paralogue annotation assessed in the current study is 98.7%, in keeping with previous findings in LQTS genes (98.4% PPV). By comparison, of 519 known missense variants in RYR2 and SCN5A predicted by SIFT to be deleterious, 500 are correct (variants with unambiguous evidence of pathogenicity) with 19 false positives (benign variants), yielding a PPV of 96.5%. Similarly, of 518 variants predicted to be damaging by Polyphen, 497 are correct with 21 false positives, with a PPV of 95.9%.

Importantly, the paralogue approach is informative for a large number of residues at which no variation has previously been reported, when variation detected in clinical testing would otherwise be reported as a ‘variant of unknown significance’. Paralogue annotation is informative for 238 unannotated RYR2 residues and 321 unannotated SCN5A residues, and was informative for more than one-third of novel variants identified in these large clinical cohorts. It is expected that any missense variants identified in these residues in CPVT or BrS patients are likely to be disease causing.

**Distribution of variation across protein domains**

Protein regions that are known to be enriched for pathogenic variation were also enriched in paralogue annotations (compared with expectation under a random distribution; table 5). One hundred and seventy-two of the 275 paralogue mutations mapped to RYR2 occur in the three mutations hotspots, an enrichment factor of 1.52. Similarly, 227 of the 441 paralogue mutations mapped to SCN5A occur in the four transmembrane domains, an enrichment factor of 1.57. As might be expected, disease-causing variants display similar domain distributions across the protein families: RYR1 has similar overlapping mutation hotspots to RYR2 (figure 3) and 75.8% of the missense mutations in the voltage-gated sodium and calcium channel paralogues of SCN5A occur in their transmembrane domains, despite these regions only accounting for 50.5% of the protein length.

**Reciprocal paralogue annotation: RYR1**

This technique is also applicable in a reciprocal fashion, that is, arrhythmia-causing mutations in RYR2 and SCN5A can be used to interpret novel variants found in any of the paralogues analysed in this study. As an example, we have mapped all the mutations in RYR2 that cause CPVT and other cardiac diseases onto RYR1, which can be used to assess novel missense variants found in patients with malignant hyperthermia or central core disease.

As expected, residues predicted to be intolerant of variation by paralogue annotation are indeed highly enriched for known disease-associated variants in RYR1 (4.4-fold enrichment; observed, 35; expected, eight). None of the four RYR1 residues that host missense variants with a MAF greater than 0.01 in 1000 Genomes (that are therefore presumably benign) were annotated, giving a notional PPV of 100%. In addition, 99 previously unannotated RYR1 residues are predicted to be intolerant of variation, of which 22 lie outside the traditional RYR1 mutation hotspots.

**Web-based paralogue annotation application**

We have developed a web-based tool to allow researchers and clinicians to identify paralogue annotations for novel variants to inform pathogenicity. This is available for RYR2 and SCN5A, as well as other LQTS-associated genes previously reported. Users input the coordinates of a novel variant (protein or complementary DNA coordinates) in a gene of interest, and any paralogue mutations associated with that residue are returned, together with fully referenced details of any known (pathogenic or benign) missense variants in the query gene. The quality of the mapping is crucial when using paralogue annotation to predict
pathogenicity. The web interface will assign each mapping a high, medium or low quality rating (defined using the alignment consensus, reference amino acids in the query and paralogue proteins and number of mappings as described in the Methods section above). In addition, the segment of the multiple sequence alignment around the mapping is displayed to allow users to assess the degree of homology visually between the proteins at that residue, coloured by the consensus between the different alignment algorithms used by the M-Coffee and T-Coffee packages. Finally, links to the publications describing the pathogenic paralogue mutations are available, allowing users to assess how well the mutation has been characterised in the paralogue-associated disease. The application is shown in figure 4.

DISCUSSION

In this study we have shown that over one-third (36%) of novel missense variants identified in large cohorts of patients with CPVT (RYR2) and BrS (SCN5A) can be annotated by paralogue mutations. This additional evidence of pathogenicity may allow us to reclassify as disease causing up to one-third of variants currently reported as ‘variants of unknown significance’, with a positive PPV exceeding 98%.10 There is a bias towards reporting pathogenic variants over benign variants in the literature, and although we have supplemented literature reports with data from 1000 Genomes and the Exome Sequencing Project, it is possible that these PPV values may be overestimated. However, as these data derive from clinically observed genotype–phenotype relationships within closely related protein families, rather than theoretical predictions based on the sequence and structure of much more distantly related proteins, we have confidence in this approach and these annotations.

In a clinical setting, the identification of a novel single amino acid substitution in a gene associated with inherited disease is generally insufficient to report the variant as disease causing. This is due to the relatively high prevalence of rare, benign variants in healthy controls in many genes. Missense variants have been shown to be present in 6% of controls in RYR24 and in 2.7% of controls in SCN5A.8 Segregation studies and functional biochemical characterisation are powerful, but resource intensive and are not always applicable. Computational predictions are attractive and widely applicable, but while informative they are not sufficiently accurate for robust clinical reporting. It therefore behoves us to make full use of available reports linking human variation to clinical phenotype.

When a variant in a disease gene has previously been observed, either in association with disease or in apparently healthy individuals, this provides invaluable additional information for clinical reporting. However, there is a huge amount of additional data linking genetic variation to clinical phenotype in paralogues of disease genes that is not presently used. By mapping functionally equivalent amino acid residues across protein families, and transferring genotype–phenotype annotations between related proteins, we can leverage this wealth of existing data to aid our interpretation of the significance of novel variants in a clinical setting.

The accuracy of this method depends on the quality of the protein sequence alignment (to ensure that aligned amino acids can be confidently regarded as functionally equivalent) and the

| Table 4 | Paralogue annotation accurately identifies disease-associated residues in RYR2 and SCN5A |
|---------|---------------------------------------------------------------|
| Protein | Pathogenic | Benign | Uncertain | Unannotated | Total |
| SCN5A | Published observations | 368 | 28 | 60 | 1560 | 2016 |
| Paralogue annotations: observed (expected) | 113 (81) | 1 (6) | 6 (13) | 321 (341) | 441 |
| RYR2 | Published observations | 139 | 20 | 71 | 4737 | 4967 |
| Paralogue annotations: observed (expected) | 35 (8) | 1 (1) | 1 (4) | 238 (262) | 275 |

Distribution of paralogue annotations across the amino acid residues of the SCN5A and RYR2 proteins. ‘Published observations’ shows the number of amino acid residues with known missense variants, classified as pathogenic, benign or uncertain, and the number of residues at which missense variation has not previously been observed (unannotated). ‘Uncertain’ refers to dbSNP variants without clinical information or residues with variants with conflicting reports as to pathogenicity. ‘Paralogue annotations observed’ shows the number of residues of each class that are annotated by variants in paralogues, and which would therefore be expected to be sites of pathogenic variation. ‘Paralogue annotations expected’ shows the number of residues in each class that would be expected to be annotated if paralogue annotation was random, with no predictive value. Variants annotated in this way are highly enriched for pathogenic variation in both genes (2 x 2 Fisher’s exact test p = 0.0009), with a positive predictive value (PPV) of 98.7%. 559 previously unannotated residues (321 in SCN5A, 238 in RYR2) are identified as putative disease-associated residues.

| Table 5 | Annotation across protein domains of RYR2 and SCN5A |
|---------|---------------------------------------------|
| Gene | Protein domains | Known missense variants | Paralogue mappings |
| | | Pathogenic | Benign | Uncertain | Actual | Enrichment |
| SCN5A | N-terminus | 27 | 2 | 6 | 14 | 0.51 |
| | Transmembrane | 277 | 8 | 26 | 355 | 1.57 |
| | Interlinker domains | 99 | 25 | 35 | 41 | 0.31 |
| | C-terminus | 40 | 5 | 16 | 31 | 0.58 |
| | Total | 443 | 40 | 83 | 441 | 10.00 |
| RYR2 | Hotspots | 134 | 10 | 17 | 157 | 1.52 |
| | Outside hotspots | 20 | 10 | 59 | 118 | 0.69 |
| | Total | 154 | 20 | 76 | 275 | |

Distribution of known variants and paralogue mutation annotations across protein domains of SCN5A and mutation hotspots of RYR2. There is significant enrichment of both known pathogenic mutations and paralogue mutation mappings in the protein regions recognised to be susceptible to pathogenic variation, that is, the trans-membrane domains of SCN5A and the three mutation hotspots of RYR2.
quality of the evidence relating genotype to phenotype for the paralogue variant. For the alignment, appropriate paralogues should be carefully and manually chosen, while we have described methods above to ensure that only mappings of sufficient quality are noted. In addition, our web application allows users to assess the quality of the alignment visually at the mapping residue. The web application also provides Pubmed links to allow researchers to assess the quality of the reports linking the paralogue variants to disease—a functionally characterised variant, or one that has been shown to segregate with disease, is more informative than a variant simply observed once in an affected individual. It is critical for researchers to consider this evidence carefully to avoid inferring pathogenicity from erroneous reports and also check that the direction of effect of the variant in the paralogue is compatible with the observed phenotype in the gene of interest.

The sensitivity of this approach depends on the availability of known, pathogenic mutations in the paralogues of genes of interest. The lack of paralogue annotation for a novel variant does not imply that the variant is non-pathogenic, simply that a disease-causing mutation in the equivalent paralogue residue has not yet been observed. As genetic testing becomes more widely available across the whole spectrum of inherited disease, for example, through the ongoing development of sequencing technologies.

Figure 3  Disease-causing variation in human ryanodine receptors, RYR1 and RYR2. An alignment of RYR1 and RYR2 reveals the structural similarity of the proteins and homologous clustering of pathogenic variation in these related proteins. The protein is represented in light grey, with reported ‘mutation hotspots’ marked in dark grey, and exon boundaries highlighted. The locations of missense variants previously reported to be pathogenic are shown with black lines above and below the protein graphic—longer lines indicate more than one pathogenic DNA variant affecting the same protein residue.

Figure 4  A web-based application makes paralogue annotation easily accessible for genes causing inherited arrhythmia syndromes. A web-based application is available at http://cardiodb.org/Paralogue_Annotation/. Users enter the position of a novel variant using complementary DNA or protein coordinates: in this example a substitution has been found in RYR2, affecting glycine residue 357. This residue maps to RYR1 residue 341, and two cDNA variants at that location (c.1021G>A and c.1021G>C) that each cause substitution of Arg for Gly at that position have been reported to cause malignant hyperthermia. Users should check the alignment quality—here the mapping quality is high: the surrounding region is highly homologous, the reference amino acid is the same in both proteins, and the alignment has a high consensus score. Pubmed links give access to the reports relating to the paralogue mutation(s), allowing users to assess the quality of evidence for pathogenicity. Here functional characterisation has been performed on this variant in the highlighted publication, adding confidence that the variant is disease causing and the residue is intolerant of variation in both RYR1 and its paralogue RYR2.
technologies, we expect the number of informative paralogue mutations to expand greatly, increasing the sensitivity of this method. Here, an additional 238 residues in RYR2 (a 171% increase compared with previous reports), and 323 residues in SCN5A (87% increase) are putatively annotated as intolerant of variation.

Paralogue annotation is designed to interpret individual variants in monogenic disease and does not currently help to clarify the genetics of complex multi-allelic disease. There have been recent reports linking common genetic variation with predisposition to BrS, and there is a debate about the role of genetic testing in this condition. However as BrS does segregate as a monogenic Mendelian trait in some pedigrees and most clinicians still feel that genetic testing of SCN5A is of value, paralogue annotation can help to interpret these novel variants.

Some regions of RYR2 and SCN5A are enriched for disease-associated variants, yet contain relatively few benign polymorphisms, so that variants in these regions have a higher-than-average previous probability of pathogenicity. While this is to some extent informative, the accuracy and sensitivity of a purely domain-based prediction are limited, as many pathogenic and benign variants do not follow this trend. Paralogue annotation provides a higher resolution annotation, down to the level of the individual amino acid. It is also not limited to any specific domains or protein regions and is less likely to mis-annotate a rare benign variant as pathogenic given the high PPV scores we have seen in this and previous studies.

In RYR2, 134 of the 154 known pathogenic missense mutations occur within the recognised mutation hotspots (figure 3). However, 10 known benign missense variants also occur in these regions, suggesting that novel variants found here cannot be definitively classified as disease causing. In our study we found that 57% of the paralogue annotations of RYR2 occur in the mutation hotspots. This is due to the fact that RYR1 contains equivalent, and largely overlapping, mutation hotspots (figure 3). Pathogenic mutations in RYR1 are, however, less concentrated in these hotspots than in RYR2, which allows us to annotate an additional 114 residues outside conventional RYR2 hotspots. It is likely that this clustering of mutations is at least partly caused by sequencing bias, that is, often only exons with known mutations are sequenced in genetic testing. The increasing use of next generation sequencing technologies allows laboratories to sequence all the exons of ryanodine receptors and is likely to expand the numbers of mutations found outside the hotspots of both RYR1 and RYR2 for their respective diseases—paralogue annotation will be able to utilise all of these findings to assess variants in either gene.

Pathogenic mutations are not as discreetly clustered in SCN5A as in RYR2. However, variation in the transmembrane regions is rare in apparently healthy individuals, and substitutions in these regions have been reported to have an 88% probability of pathogenicity when found in clinically affected individuals. Here, over 80% of disease-causing variants in SCN5A paralogues map to transmembrane residues, providing further evidence that these domains are functionally important and intolerant of variation. Most importantly, paralogue annotation also annotates some variants at residues outside these regions, whose significance would otherwise be unappreciated.

Paralogue annotation of variants is widely applicable: approximately half of all disease-associated genes have one or more paralogues with disease-causing variants. The technique can also be applied in a reciprocal manner, as illustrated here with RYR1 and RYR2, and our web-based application will be informative for researchers and clinicians interested in a range of diseases mediated by ion channels and membrane excitability, such as malignant hyperthermia and central core disease, and as well as the cardiovascular genetics community.

CONCLUSION

This study demonstrates an accurate and widely applicable approach to interpret novel missense variants, here applied to annotate variants in CPVT (RYR2) and BrS (SCN5A). This is informative for more than one-third of novel variants in these genes, and may provide sufficient evidence to report these variants as disease causing, rather than as ‘variants of unknown significance’, with a positive PPV of 98.7%. This approach is based on clinical genotypetype–phenotype relationships in humans, rather than computational prediction, giving further confidence in its application. While this method will not be applied in isolation, due to our incomplete inventory of pathogenic variation in paralogues, it provides an invaluable additional tool for clinicians and researchers. Our web-based application provides a user-friendly implementation of this technique for immediate application in the interpretation of novel variants, with an initial focus on inherited arrhythmia syndrome genes, together with an annotated compendium of previously reported missense variants in these genes.

Contributors RW and JSW devised the study. RW analysed the data and drafted the paper. JSW, SAC and NSP reviewed the data and revised the manuscript.

Funding This research was supported by the Academy of Medical Sciences, the Wellcome Trust, the British Heart Foundation (grant number SP/10/10/28431), Arthritis Research UK, Fondation Leducq and the NIHR Cardiovascular Biomedical Research Unit at Royal Brompton & Harefield NHS Foundation Trust and Imperial College London.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 3.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/3.0/

REFERENCES

1 Alings M, Wilde A. Brugada syndrome: clinical data and suggested pathophysiological mechanism. Circulation 1999;99:666–73.
2 Priori SG, Napolitano C, Gasparini M, Pappone C, Della Bella P, Giordano U, Bloise R, Giustetto C, De Nardis R, Grillo M, Ronchetti E, Faggiano G, Nastoli J. Natural history of Brugada syndrome: insights for risk stratification and management. Circulation 2002;105:1342–7.
3 Priori SG, Napolitano C, Memmi M, Colombi B, Drago F, Gasparini M, Desimone L, Coltoni F, Blowe R, Keegan R, Cruz Filho FE, Vignati G, Benatar A, DeLong A. Clinical and molecular characterization of patients with catecholaminergic polymorphic ventricular tachycardia. Circulation 2002;106:69–74.
4 Medeiros-Domingo A, Bhuiyan ZA, Tester DJ, Hofman N, Bikker H, van Tintelen JP, Mannens MM, Wilde AA, Ackerman MJ. The RYR2-encoded ryanodine receptor/calcium release channel in patients diagnosed previously with either catecholaminergic polymorphic ventricular tachycardia or genotype negative, exercise-induced long QT syndrome: a comprehensive open reading frame mutational analysis. J Am Coll Cardiol 2009;54:2065–74.
5 Kapplinger JD, Tester DJ, Alders M, Benito B, Berthet M, Brugada J, Brugada P, Fressart V, Guercielloff A, Harris-Kerr C, Kamakura S, Kyndt F, Koopmann TT, Miyamoto Y, Pfeiffer R, Pollevick GD, Protas V, Zuzanges S, Votta M, Toebin JA, Shimizu W, Schulze-Bahr E, Antzelevitch C, Salisbury BA, Guicheney P, Wilde AA, Brugada R, Schott JI, Ackerman MJ. An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. Heart Rhythm 2009;7:33–46.
6 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
7 Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
8 Kapa S, Tester DJ, Salisbury BA, Harris-Kerr C, Pungliya MS, Alders M, Wilde AA, Ackerman MJ. Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. Circulation 2009;120:1752–60.
Yano M, Yamamoto T, Ikeda Y, Matsuizaki M. Mechanisms of disease: ryanodine receptor defects in heart failure and fatal arrhythmia. Nat Clin Pract Cardiovasc Med 2006;3:43–52.

Ware JS, Walsh R, Cunningham F, Bimey E, Cook SA. Paralogous annotation of disease-causing variants in long QT syndrome genes. Hum Mutat 2012;33:1188–91.

Sharan JL, Mparthanga CP, Spedlin M, Germain P, Staels B, Ducquet C, Laudet V, Harmar AJ, NC-IUPHAR. IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. Nucl Acids Res 2011;39:DS34–8.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.

Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics 2011;39:1.13.1–1.13.20.

Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for multiple sequence alignments. J Mol Biol 2000;302:205–17.

Splawski I, Shen J, Timothy KW, Lehmann MH, Priori S, Robinson JL, Moss AJ, Schwartz PJ, Toybin JA, Vincent GM, Keating MT. Spectrum of mutations in long-QT syndrome genes. KVLQT1, HERG, SCN5A, KCNE1, and KCNE2. Circulation 2000;102:1178–85.

Tester DJ, Will ML, Haglund CM, Ackerman MJ. Compendium of cardiac channel mutations in 541 consecutive unrelated patients referred for long QT syndrome genetic testing. Heart Rhythm 2005;2:507–17.

Napolitano C, Priori SG, Schwartz PJ, Bloise R, Ronchetti E, Nastoli J, Bottelli G, Cerone M, Leonardi S. Genetic testing in the long QT syndrome: development and validation of an efficient approach to genotyping in clinical practice. JAMA 2005;294:2975–80.

Hedley PL, Jørgensen P, Schlamowitz S, Moolman-Smok J, Kanters JK, Corfield VA, Christiansen M. The genetic basis of Brugada syndrome: a mutation update. Hum Mutat 2009;30:1236–66.

Kaplinger JD, Tester DJ, Salisbury BA, Carr JL, Harris-Kerr C, Pollevick GD, Wilde AA, Ackerman MJ. Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. Heart Rhythm 2009;6:1297–303.

Arnestad M, Crotti L, Rognum TO, Insolera R, Pedrazzini M, Ferrandi C, Vege A, Wang DW, Rhodes TE, George AL, Schwartz PJ. Prevalence of long-QT syndrome gene variants in sudden infant death syndrome. Circulation 2007;115:361–7.

Olesen MS, Yuan L, Liang B, Holst AG, Nielsen N, Nielsen JB, Hedley PL, Christiansen M, Olesen SP, Haunsø S, Schmitt N, Jespersen T, Svendsen JH. High prevalence of long QT syndrome-associated SCN5A variants in patients with early-onset lone atrial fibrillation. Circ Cardiovasc Genet 2012;5:450–9.

Hershberger RE, Parks SB, Kushner JD, Li D, Ludwigsen S, Jakobs P, Nauman D, Burgess D, Partain J, Litt M. Coding sequence mutations identified in MYH7, TNNT2, SCNSA, CSPRS2, LBD3, and TAC3 from 313 patients with familial or idiopathic dilated cardiomyopathy. Clin Transl Sci 2008;1:21–6.

Bezzina CR, Barc J, Músawaya Y, Remme CA, Gourraud JB, Simonet F, Verkerk AO, Schwartz PJ, Crotti L, Dagradi F, Guichény P, Fressart V, Leenhart A, Antzelevitch C, Bartkowiak S, Schütze-Bahr E, Zumhagen S, Behr ER, Bastaenren R, Tfelt-Hansen J, Olesen MS, Kaab S, Beckmann BM, Weeke P, Watanabe H, Endo N, Minamino T, Horie M, Ohno S, Hasugawa M, Makita N, Nogami A, Shimizu W, Alba T, Freguell P, Balkau B, Lantier O, Torchio M, Wiese C, Weber D, Wolfsink R, Coronel R, Boukens BJ, Beüzéa S, Charpentier E, Chatel S, Despres A, Gros F, Kyndt F, Lecointe S, Lindenbaum P, Porterio V, Violette J, Gessler M, Tan HL, Roden DM, Christoffels VM, Le Marec H, Wilde AA, Probst V, Schott U, Dina C, Redon R. Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. Nat Genet 2013;45:1044–9.