Improving Interpretation of Cardiac Phenotypes and Enhancing Discovery With Expanded Knowledge in the Gene Ontology

BACKGROUND: A systems biology approach to cardiac physiology requires a comprehensive representation of how coordinated processes operate in the heart, as well as the ability to interpret relevant transcriptomic and proteomic experiments. The Gene Ontology (GO) Consortium provides structured, controlled vocabularies of biological terms that can be used to summarize and analyze functional knowledge for gene products.

METHODS AND RESULTS: In this study, we created a computational resource to facilitate genetic studies of cardiac physiology by integrating literature curation with attention to an improved and expanded ontological representation of heart processes in the Gene Ontology. As a result, the Gene Ontology now contains terms that comprehensively describe the roles of proteins in cardiac muscle cell action potential, electrical coupling, and the transmission of the electrical impulse from the sinoatrial node to the ventricles. Evaluating the effectiveness of this approach to inform data analysis demonstrated that Gene Ontology annotations, analyzed within an expanded ontological context of heart processes, can help to identify candidate genes associated with arrhythmic disease risk loci.

CONCLUSIONS: We determined that a combination of curation and ontology development for heart-specific genes and processes supports the identification and downstream analysis of genes responsible for the spread of the cardiac action potential through the heart. Annotating these genes and processes in a structured format facilitates data analysis and supports effective retrieval of gene-centric information about cardiac defects.

Ruth C. Lovering, PhD
Paola Roncaglia, PhD
Douglas G. Howe, PhD
Stanley J.F. Laulederkind, PhD
Varsha K. Khodiyar, PhD
Tanya Z. Berardini, PhD
Susan Tweedie, PhD
Rebecca E. Foulger, PhD
David Osumi-Sutherland, PhD
Nancy H. Campbell, PhD
Rachael P. Huntley, PhD
Philippa J. Talmud, PhD
Judith A. Blake, PhD
Ross Breckenridge, MRCP, PhD
Paul R. Riley, PhD
Pier D. Lambiase, MRCP, PhD
Perry M. Elliott, MRCP, PhD
Lucie Clapp, PhD
Andrew Tinker, MRCP, PhD
David P. Hill, PhD

Correspondence to: Ruth C. Lovering, PhD, Functional Gene Annotation, Institute of Cardiovascular Genetics, University College London, Rayne Bldg, 5 University St, London WC1E 6JF, United Kingdom. E-mail r.lovering@ucl.ac.uk

Clinical Perspective

Identification of critical proteins and RNAs as potential drug targets requires computationally accessible descriptors, both in terms of their roles in biological pathways and the molecules they interact with to fulfill their actions. This information is also vital in choosing prognostic and diagnostic biomarkers, particularly for multifactorial diseases such as heart disease that might benefit from measurement of multiple biomarkers. Recent improvements in omic technologies have led to projects such as the 100 000 Genomes Project and genome-wide association studies, which are producing vast amounts of genetic data pertinent to human health. Furthermore, it is now possible to catalog which proteins or RNAs are present in normal or disease tissues, or individual cells. Understanding the molecular network or biological processes associated with a drug target can help predict off-target effects or the potential for drug repurposing because many gene products are active in multiple pathways. In addition, these networks can be used to predict the most efficacious molecules within the networks, through the identification of key positions at which the whole networks may be perturbed; these molecules are often associated with disease-causing mutations or identified as drug targets. One of the major resources used by omic researchers is the Gene Ontology. This article explains the considerable improvements made by the Gene Ontology Consortium in the bioinformatic description of cardiac physiology. These new descriptions are all based on published data and are now included in all major biological databases, thus available for use by the global scientific community to enhance the understanding of cardiovascular-relevant data.

Cardiac electrical conduction systems enable coordinated regulation of heart contraction in metazoans, ranging from fly to human. Disturbances of normal heart rhythm can occur de novo, but, more critically and commonly, they are an important feature of many cardiac diseases, and have substantial impacts on patient morbidity and mortality. To gain insight into the mechanisms of arrhythmia, high-throughput genome-scale methodologies (including genome-wide association studies [GWAS], transcriptomics, exon sequencing, and proteomics) are being used.[1–3] However, interpretation of these high-throughput experiments relying on descriptions of the cellular and physiological roles of gene products, and a computational approach to interrogation of cardiac gene function, is a bottleneck in these analyses. Our work aims to fill this gap by capturing information in the Gene Ontology (GO) resource in a structured way, thus integrating knowledge about genes, cells, tissues and organs. To support data interpretation, the GO Consortium (GOC) provides a freely available, structured, controlled vocabulary, the ontology,[4] that enables association of defined terms, describing cellular roles and locations, with a protein or RNA. This association process, called GO annotation, provides a computer-interpretable summary of the results of many independent experiments. GO terms describe Molecular Functions (molecular activities of a gene product), Biological Processes (the broader context in which a gene product acts), and Cellular Components (the subcellular location of a gene product). Some examples of GO terms relevant to cardiac research are, respectively, *voltage-gated sodium channel activity, Purkinje myocyte action potential*, and *Z disc*.

GO provides users with a summary of experimentally verified or predicted functions of genes, proteins, and RNAs.[3,5] Because of this high-impact information, GO data are incorporated into over 50 functional analysis tools and is routinely used to analyze large data sets.[6] The coordinated action of gene products involved in cardiac conduction at the cellular level results in proper heart functioning at the organ level and in healthy conditions at the whole organism level. Therefore, detailed and interoperable knowledge about gene products' roles is essential to better analyze heart failure phenotypes and ultimately to address potentially fatal conditions. The features listed above make GO an ideal resource to aid the interpretation of large-scale cardiac physiology investigations.

Before our expansion of the cardiac function domain, there were only 3 GO terms to describe how the cardiac cycle is coordinated: *cardiac conduction, membrane depolarization during atrial cardiac muscle cell action potential*, and *membrane repolarization during atrial cardiac muscle cell action potential*. In this article, we detail how GO editors and biocurators worked together with experimental researchers to ensure provision of accurate structured terminology[7] that describes features of gene products involved in cardiac physiology, and to generate GO annotations[8] that capture those roles based on published literature, thereby improving representation of this area of biology. In particular, we focused on processes involved in and regulating the coordinated contraction of the heart. We also show that our effort improves the investigation of heart-related GWAS and transcriptomic data sets.

METHODS

The data, analytic methods, and study materials have been made available to other researchers for purposes of reproducing the results or replicating the procedure. Protein, RNA, and macromolecular complex GO annotations are available from the UniProt GO annotation and GOC sites (https://www.ebi.ac.uk/GOA/downloads, http://www.geneontology.org/gene-associations/) and searchable in the AmiGO2[7] and QuickGO[9] browsers; GO

terms are downloadable from the GOC site (http://www.geneontology.org/ontology/) and searchable in AmiGO2 and QuickGO. The annotations have also been incorporated into many biological knowledge bases, including Ensembl,[10] UniProtKB,[11] and NCBI-Gene,[12] and exploited by numerous freely available functional analysis tools, including Database for Annotation, Visualization and Integrated Discovery,[13] g:profiler,[14] Protein Analysis Through Evolutionary Relationships (PANTHER),[15] and Visual Annotation Display (VLAD).[16] Full methods are available in the Data Supplement.

Generation of a Prioritized List of Cardiac Physiology-Relevant Gene Products

A list of 88 human proteins known to be required for normal cardiac function was compiled based on 8 literature reviews[17–24] (Table I in the Data Supplement). To keep the focus more specifically on ion channels, the following were not retained in the prioritized list of genes: ATPases, ATPase regulatory proteins, hormones, hormone receptors, and proteins required for normal heart development. However, angiotensinogen and 2 angiotensin I–converting enzymes were included in the prioritized protein list to ensure that the ontology was sufficiently developed to enable future projects to capture the role of gene products that regulate heart rate and contraction.[25] Orthologs of these 88 human proteins were identified in mouse, rat, zebrafish, and *Drosophila* using the Protein Analysis Through Evolutionary Relationships orthology prediction tool.[15] In addition, a list of 7 human microRNAs (miRNAs) identified as key players in modulating cardiac excitability at the time we started our curation effort was determined from 3 reviews[26–28]; literature available for these miRNAs was curated up to March 2016. Prioritized human gene products described above are listed in Table I in the Data Supplement, and for all of them, GO annotations supported by experimental data were included in the GO database according to established procedural guidelines[8,29,30] (details are given in Methods in the Data Supplement).

Identification of Candidate Genes Associated With Arrhythmia Risk Loci

We compiled a list of Mendelian inherited arrhythmia disorder genes associated with atrial fibrillation, long-QT syndrome, short-QT syndrome, catecholaminergic polymorphic ventricular tachycardia, and Brugada syndrome[1,31,32] and candidate genes associated with GWAS risk loci for these disorders[2,31] (literature search February 2016; Table II in the Data Supplement). Nine GO terms relevant to normal cardiac physiology (such as *cardiac conduction, muscle contraction,* and *response to oxygen levels*) were selected to investigate whether gene annotations to these normal processes could be correlated with arrhythmia candidate genes (Table III in the Data Supplement), and the associated annotations were downloaded on March 11, 2016 using the QuickGO browser[9] (http://www.ebi.ac.uk/QuickGO/; Methods in the Data Supplement).

GO Functional Analysis of Transcriptomic Data Sets

A ventricular cardiomyocyte (VCM) data set[3] was analyzed using the BinGO plugin[33] within the Cytoscape v3.3.0 tool,[34] applying

the recommended hypergeometric test and Benjamini and Hochberg False Discovery Rate correction, and a *P* value <0.05 (full details in Methods in the Data Supplement). To determine the impact of this focused project, we used gene association files from November 2011 (ie, the start of this project) and February 2016 (ie, the end of the project; files available from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/). For both analyses, the ontology version March 7, 2016 was included (available from http://geneontology.org/page/download-ontology).

RESULTS
Representation of Cardiac-Relevant Gene Products' Features Via Ontology Development

A coordinated working group of GO editors, GO biocurators, and field experts expanded the cardiac physiology domain by adding 197 new cardiac-relevant GO terms (Methods and Table IV in the Data Supplement). Of these, 87 refer to processes that contribute to cardiac conduction and thus describe the propagation of the action potential through the conduction system and cardiac chambers. An example of an ontology branch representing cardiac-relevant GO terms is shown in Figure.

The expansion of the cardiac-related processes in GO describes events that occur at 3 key biological levels: the single cell, the tissue (multicellular), and the organ. In addition to these biological levels, the ontology was expanded to cover generic processes (such as *regulation of potassium ion export*) or developmental processes (such as *cardiac pacemaker cell differentiation*) required for the annotation of cardiac-relevant gene products. These new ontology terms enable biocurators to capture the role of gene products in specific tissues of the heart (by applying GO terms such as *adrenergic receptor signaling pathway involved in cardiac muscle relaxation*), specific cell types (eg, *atrioventricular (AV) node cell to bundle of His cell communication*), and the whole organ (eg, *regulation of the force of heart contraction by cardiac conduction*). The addition of both generic terms, and terms that reference specific anatomical structures, allowed us to overcome challenges with respect to granularity of organism specificity. For example, the term *cell communication involved in cardiac conduction* can be used for any organism with a cardiac organ, and a child term like *AV node cell to bundle of His cell communication* can be used specifically for species that have the relevant anatomical structures, in this case AV node and bundle of His. By describing processes at the various levels, curators can choose appropriate terms with respect to the experiments and organisms they are curating. Where possible, terms were defined by creating necessary and sufficient statements using relationships between GO terms and terms from cross-referenced external ontologies, such as Uberon for anatomical structures, the Cell Ontol-
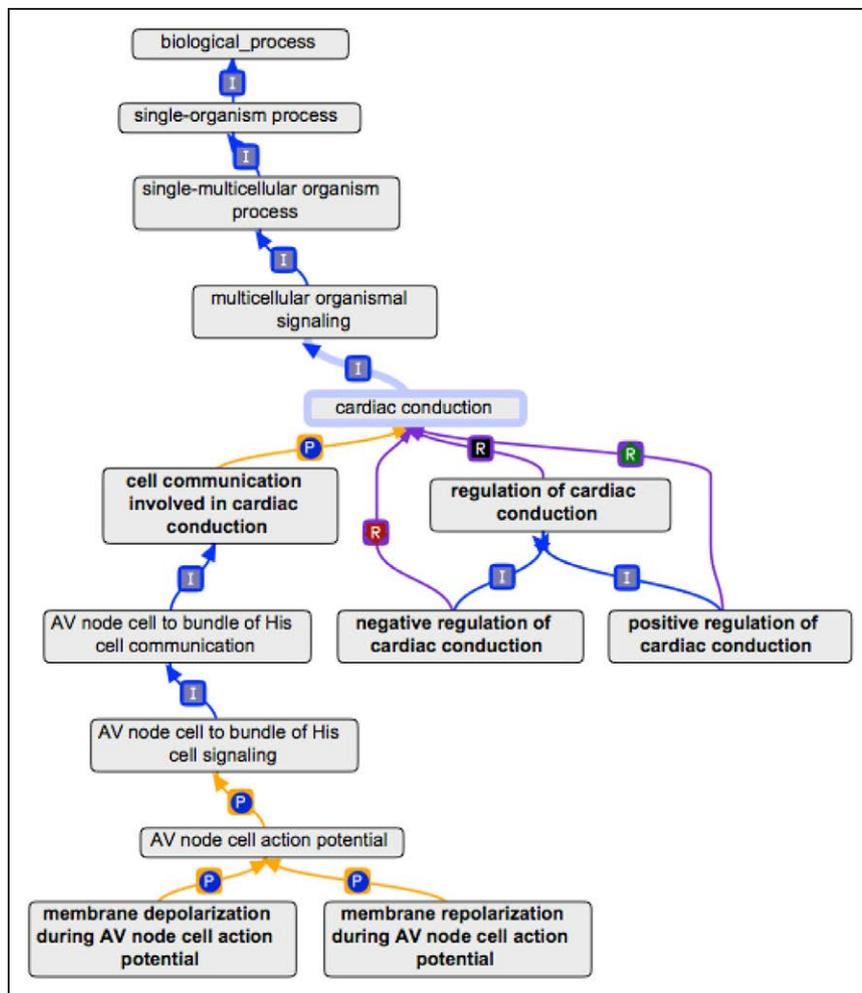
Figure. Part of the ontology describing AV node cell action potential and the regulation of ventricular cardiac muscle cell action potential. The entire model is not shown; this portion graphically illustrates how ontology terms relate to each other. Blue lines marked with I indicate the child term is a type of its parent, and orange lines marked with P indicate that child term is always a part of its parent. Purple lines marked with R indicate regulates, positively regulates, or negatively regulates relations.

ogy for cell types, and Chemicals of Biological Interest for chemicals (ChEBI), taking advantage of their species interoperability and thus making the terms accessible to computational reasoning.[35]

Cell-to-Cell Communication in the Heart

Cell communication is a key aspect of the coordinated activity of the heart. We created new terms, such as *cell communication involved in cardiac conduction*, to bring together all processes that mediate interactions between a cell and its surroundings and that also contribute to the process of cardiac conduction. This part of the ontology now includes terms to describe cell-to-cell impulse propagation by means of cardiac action potentials and electrical coupling (such as *AV node cell to bundle of His cell communication by electrical coupling*), as well as terms to describe the attachment of one cell to another or to the extracellular matrix (such as *bundle of His cell-Purkinje myocyte adhesion involved in cell communication*). In addition, many of the 22 new cardiac conduction-relevant Molecular Function terms describe channel or transporter activities that are necessary to propagate the electrical signals between specific cells and across the heart (eg, *gap junction channel activity involved in cardiac conduction electric coupling*).

The cardiac action potential branch of GO includes terms referring to a variety of cell types within the heart (Figure), such as *AV node cell action potential*, *membrane depolarization during AV node cell action potential* and *membrane repolarization during AV node cell action potential*, as well as terms that describe the regulation of the action potential (not shown). Similar terms also describe action potentials in SA node cells, bundle of His cells, and Purkinje myocytes (not shown).

Ensuring Capture of Robust Information About Genetic Contributions to Cardiac Physiology via GO Annotation

There is a considerable volume of literature describing the physiological characterization of the heart. This project brought together expertise from University College London,[36] Queen Mary University of London, University of Oxford, Mouse Genome Informatics,[37] the Rat Genome Database,[38] the Zebrafish Information Network,[39] and FlyBase.[40] It was unrealistic for the information from all

cardiac physiology research papers to be captured during this project; however, over 500 papers have been reviewed and used to annotate cardiac-relevant gene products with the expanded GO structure. A key aspect of gene product GO annotation is the recognition that a single gene product may have several functions, may participate in multiple processes, and can be found in a variety of subcellular locations. Consequently, fully describing how a gene functions typically results in >1 GO term associated with a single gene product[4] (Table 1).

The aim of creating multiple statements (annotations) about a single gene product is to comprehensively capture the knowledge about its biological role, as supported by published experimental data. For example, over 30 papers were curated to provide information about the human sodium channel SCN5A (Q14524). Individual experiments, such as immunostaining, expression of various sodium channel constructs in *Xenopus* oocytes or rat cardiomyocytes, combined with descriptions of cardiac disease associated with the protein, such as long-QT and Brugada syndromes, have provided evidence that this protein is found in the *Z disk*,[41] has *voltage-gated sodium channel activity involved in cardiac muscle cell action potential*,[42] and is involved in *membrane depolarization during Purkinje myocyte cell action potential*.[43] In contrast, the experimental data described in a single paper[44] supported 9 annotations

for the human protein RNF207 (Q6ZRF8). This article also resulted in 18 experimentally supported annotations to 5 other human proteins (DNAJA1, P31689; HSPA1A, P0DMV8; HSPA1B, P0DMV9; HSPA8, P11142; KCNH2, Q12809) and 1 zebrafish protein (Rnf207b, E9QHE3).

There is increasing evidence for the role of miRNAs in controlling cardiac physiology, particularly in the regulation of potassium and calcium ion channel function, where dysregulation can lead to aberrant action potential duration and cardiac conduction.[27,28] For this project, 7 miRNAs were prioritized for annotation (Methods and Table I in the Data Supplement), resulting in 27 annotations with relevance to cardiac function. For example, Li et al[45] demonstrated that miR-1 (URS00001DC04F_9606) and miR-133a (URS00004C9052_9606) could regulate the slow delayed rectifier potassium current (I_{Ks}) in human cells during simulated hyperglycemia. Furthermore, they showed that this was because of their ability to regulate the expression of 2 potassium channel proteins, KCNE1 (P15382) and KCNQ1 (P51787), which mediate I_{Ks}. These roles of miR-1 and miR-133a have been captured by the terms *negative regulation of membrane repolarization during cardiac muscle cell action potential* and *negative regulation of delayed rectifier potassium channel activity*, as well as *gene silencing by miRNA*, with KCNE1 included as the target of miR-1 and KCNQ1 as the target of miR-133a in annotation extensions (Methods in the Data Supplement).

Table 1. Selection of GO Annotations Associated With Human ACE2

GO Identifier	GO Term Name	Code	Reference	With
Biological process				
GO:0002005	Angiotensin catabolic process in blood	IC	PMID:10924499	GO:0004180
GO:0003081	Regulation of systemic arterial blood pressure by renin–angiotensin	IMP	PMID:18258853	…
GO:0019229	Regulation of vasoconstriction	IC	PMID:15380922	GO:0004180
GO:0046813	Receptor-mediated virion attachment to host cell	IDA	PMID:18343844	…
GO:0060452	Positive regulation of cardiac muscle contraction	IEA	GO_REF:0000019	Ensembl: ENSMUSP00000107890
GO:1903598	Positive regulation of gap junction assembly	IMP	PMID:12967627	…
GO:1903779	Regulation of cardiac conduction	IMP	PMID:12967627	…
Molecular function				
GO:0004175	Endopeptidase activity	IDA	PMID:15283675	…
GO:0004180	Carboxypeptidase activity	IDA	PMID:10969042	…
GO:0005515	Protein binding	IPI	PMID:21068237	UniProtKB:O15393
Cellular component				
GO:0005576	Extracellular region	IEA	GO_REF:0000037	UniProtKB-KW:KW-0964
GO:0070062	Extracellular exosome	IDA	PMID:19056867	…
GO:0016020	Membrane	IEA	GO_REF:0000002	InterPro:IPR001548

A selection of the 61 annotations associated with human ACE2 (Q9BYF1) illustrates the range of annotations associated with a single protein. Code indicates the evidence code used to support the annotation (www.geneontology.org/GO.evidence.shtml)[8]; Reference lists the source of the data supporting the annotation, this may be the curated article or information about the electronic annotation pipeline; and the With field provides additional information to support the annotation, such as the Gene Ontology (GO) identifier which supports an IC annotation or the Ensembl identifier from which the annotation is propagated.[8]

IC indicates Inferred by Curator; IMP, Inferred from Mutant Phenotype; IDA, Inferred from Direct Assay; IEA, Inferred from Electronic Annotation; IPI, Inferred from Physical Interaction.

The Application of Information Derived From Orthologs

Experimental data to support GO annotations are not always available for human proteins and especially not for miRNAs, but may be obtainable for model organism orthologs. Inferential assertions about the roles of human gene products can be achieved by mapping annotations from their orthologous gene products. The electronic pipeline Ensembl Compara ensures that experimentally supported annotations associated with those proteins that have a 1-to-1 ortholog across human, mouse, and rat are applied to all orthologs.[10] For example, the predicted 1-to-1 orthology between the mouse ACE2 (Q8R0I0) and human ACE2 (Q9BYF1) proteins as defined by Compara has enabled 3 experimentally supported mouse ACE2 annotations (including *positive regulation of cardiac muscle contraction*) to be associated with human ACE2. The GOC also provides an expert curation tool, called Phylogenetic Annotation and Inference Tool,[46] that enables biocurators to infer annotations across many species based on phylogenetic relationships and protein family membership. In addition, in some cases where knowledge from Compara and Phylogenetic Annotation and Inference Tool was not available, orthology was reviewed on a case-by-case basis and used to support transfer of annotations to the orthologous proteins[47] (see next section and Methods in the Data Supplement).

Comprehensive Capture of Information About Prioritized Gene Products Through Annotation

The manual annotation of the 88 heart-relevant human proteins that were prioritized in this effort (Table I in the Data Supplement) led to the submission of over 3100 annotations to the GO database. This represents a 4-fold increase in the number of manual annotations associated with these proteins. Furthermore, annotation of model organism experimental data has provided a further 2000 annotations to 82 of these proteins through the transfer of experimentally supported annotations to orthologous human proteins via the Ensembl Compara pipeline,[10] Phylogenetic Annotation and Inference Tool,[46] and other expert curation methods.[47] To supplement the orthology-based curation, we manually transferred 50 relevant biological process annotations from mouse or rat to human, using the Inferred From Sequence or Structural Similarity evidence code.[8] When annotations provided through other electronic pipelines[47] (such as mappings between UniProtKB Keywords and InterPro) are included, the 88 prioritized human proteins have a total of 7263 annotations (as of March 2017). The number of annotations per protein ranges from 20 associated with RNF207 (Q6ZRF8) to 272 associated with CAV1 (Q03135), with an average of 82 annotations per protein. Over 140 annotations

are associated with the 7 prioritized human miRNAs. Therefore, this curation effort allowed bioinformatic capture of knowledge of genes that were not represented before, and it enables more informative data analysis, as shown below.

Using GO to Interpret GWAS

GWAS have identified many risk loci associated with cardiac disorders.[1,31] In some cases, the impact of a variant on a protein-coding gene is relatively easy to identify because of predicted, and often experimentally verified, deleterious nonsynonymous substitutions. However, often risk variants identified in a genomic area fall within an intronic, intergenic, or regulatory region, making it difficult to identify the true causative variant(s) associated with the risk, and which protein-coding genes or functional RNA genes should be considered as candidates contributing to a disease.[2,31] The cost of investigating a candidate gene's role in a disease is considerable. Consequently, before experimental investigation, various in silico approaches are generally undertaken to try to narrow down the choice of which gene in a gene-rich region is involved in a disease. If nonsynonymous deleterious gene variants in a gene in the same region have already been associated with a similar phenotype, this is good supporting evidence for candidate gene choice. However, other approaches, such as those described by MacArthur et al,[48] may also be valuable. These include considering whether any of the potential candidate genes encode a protein that interacts with proteins previously implicated in the disease (either using in silico network analysis or through coimmunoprecipitation experiments), or if the gene is expressed in tissues relevant to the disease (eg, by using expression quantitative trait loci data[31]).

MacArthur et al[48] also describe selecting candidate genes based on a known function or role which is shared with other genes established as associated with the disease of interest or is consistent with their mutant phenotype. This approach can be easily undertaken using GO annotation data, and yet it is rarely included in the in silico investigations of GWAS results. After our focused annotation effort, the ability to use GO annotations to identify candidate genes associated with arrhythmia disorders was evaluated. Human proteins associated with 9 GO terms that are relevant to normal cardiac physiology were downloaded using QuickGO[9] (GO terms are shown in Table IV in the Data Supplement). The corresponding genes were then compared with Mendelian genes for arrhythmia disorders (atrial fibrillation, long-QT syndrome, short-QT syndrome, catecholaminergic polymorphic ventricular tachycardia, Brugada syndrome[1,31,32]) and with GWAS candidate variants associated with QT interval length and atrial fibrillation risk loci, described by Tucker et al[2] and Arking

et al.[31] Of 45 Mendelian arrhythmia-associated genes (described in 3 reviews to 2015[1,31,32]), 43 were annotated with ≥1 of the cardiac-relevant GO terms (Table II in the Data Supplement), with *SRL* and *GREM2* being the exceptions. The GO annotations associated with the 83 GWAS-identified candidate genes (associated with 48 risk loci) suggest that GO data can contribute to prioritization of candidate risk genes. Cardiac physiology-relevant GO terms suggest the prioritization of 36 of the 83 candidate genes. This selection approach compares favorably with that taken by Arking et al.[31] In that study, authors identified 35 common variant QT interval loci and used expression quantitative trait loci, protein interaction, and coimmunoprecipitation data to identify 70 candidate genes associated with these loci (Table II in the Data Supplement). Both the GO and the expression quantitative trait loci approaches suggest 27 QT interval candidate genes, whereas coimmunoprecipitation suggests 12 genes and in silico protein interaction networks 15 genes. For example, GO annotations hint that *RNF207*, at 1p36, is a good candidate for variations in QT interval length, with annotations to 6 of the key GO terms, and that potassium transport-associated *KCNAB2* is a good alternative candidate. In contrast, in silico protein interaction data support 3 other genes at this locus (*ACOT7*, *KCNAB2*, and *RPL22*; Table 2). Similarly, GO annotation information narrows the candidates at the 3p21 locus to 3 genes (*MYL3*, *SCAP*, and *SETD2*; Table 2), with *MYL3* having the strongest anno-

tation support, because of its role in muscle contraction and heart development; the analysis provided by Arking et al[31] suggested 5 candidate genes at this locus. Notably, at the 1q24 risk locus, GO annotations, expression quantitative trait loci, and coimmunoprecipitation data all suggest *ATP1B1* as the candidate gene (Table 2). This illustrates that GO annotations can provide additional data to either support the choice of candidate genes or suggest alternative candidates.

Functional Analysis of Atrial and Ventricular High-Throughput Data Sets

To provide a measure of the impact of our cardiac-focused annotation project on data interpretation, we compared the functional analysis of a specific heart transcriptomic data set using GO annotations available in November 2016 to an equivalent analysis using GO annotation data available in February 2011. A VCM transcriptomic data set[3] was used to investigate this, as we anticipated that a ventricular-specific transcriptome was likely to include the ion channels, and their regulators, required for cardiac repolarization and depolarization. Poon et al[3] profiled the transcriptome of human embryonic stem cells and compared this to the transcriptome of adult human VCMs, after filtering out the transcripts that were not significantly differentially expressed between the 2 cell types.[3] Poon et al[3] found that GO terms describing translation elongation,

Table 2. Using GO to Support the Identification of the Likely Causative Gene Associated With 3 QT Interval Risk Loci[31]

QT Interval Risk	rs846111				rs10919070		rs17784882					
Nearest Gene	*RNF207*				*ATP1B1*	*ELP6*						
GWAS Candidate Gene		*ACOT7*	*KCNAB2*	*RPL22*		*NME7*		*KLHL18*	*MYL3*	*PTPN23*	*SCAP*	*SETD2*
GO term name												
Cardiac conduction	*				*							
Regulation of membrane potential	*				*							
Actin filament-based process	*								*			
Potassium ion transmembrane transport	*		*		*							
Sodium ion transmembrane transport					*							
Calcium ion transmembrane transport					*							
Muscle contraction	*				*				*			
Heart development	*								*			*
Response to oxygen levels					*						*	
Candidate gene predicted by												
GO annotation data	*RNF207*		*KCNAB2*		*ATP1B1*				*MYL3*		*SCAP*	*SETD2*
eQTL or protein interaction data		*ACOT7*	*KCNAB2*	*RPL22*	*ATP1B1*	*NME7*		*KLHL18*	*MYL3*	*PTPN23*	*SCAP*	*SETD2*

Gene Ontology (GO) annotation data can be used to provide additional information about candidate genes, located close to the risk loci, and add to the information provided by expression quantitative trait loci data (eQTL) or protein interaction data (full list of arrhythmia associated genes in Table II in the Data Supplement). Association of these terms with the genes in close proximity to the QT interval risk loci suggests that these are candidate causative genes. At the bottom of the table, the candidate gene predicted by either GO annotation, eQTL, or protein interaction data is listed: GO annotation data (based on the annotations listed in the table); eQTL or protein interaction data (based on eQTL, in silico QT interval loci protein network analysis or coimmunoprecipitation with 1 of 5 LQTS Mendelian proteins[31]).

*The association of 1 of the 9 selected GO terms with a listed gene.

as well as muscle system, contraction, and energy generation (their wording), were enriched within the 200 most abundant transcripts in adult human VCMs. Our analysis using the 2011 and 2016 GO annotation data sets confirms findings of Poon et al,[3] with the majority of these proteins annotated to either *muscle contraction, mitochondrion organization, respiratory electron transport chain, gene expression,* or *developmental process* terms. Notably, our analysis also demonstrates that the terms *cardiac conduction, regulation of cardiac conduction,* and *regulation of the force of heart contraction* (among others) were significantly enriched in the 2016 analysis. Although VCMs are not involved in cardiac conduction, it is not unexpected that *cardiac conduction* and child terms will be enriched in a ventricular data set, because many of the same action potential-associated proteins are present in ventricular, atrial, and conduction tissues.[23] The identification of the cardiac-specific terms within this data set does, however, confirm that the GO terms and annotations created by this focused project are sufficient to enable the identification of the increased expression of genes involved in these processes. In 2011, many of the cardiac physiology GO terms did not exist, and consequently, these terms are (understandably) not enriched using the 2011 annotation data set (Table 3; Table V in the Data Supplement).

DISCUSSION

In this article, we describe work to improve the GO resource in representing relevant cardiac processes, and in capturing roles of gene products from published literature through GO annotation. The majority of publications describing heart or cardiomyocyte transcriptomic and proteomic data sets use GO annotations to investigate the pathways associated with this important organ.[3,5] We tested whether our recent cardiac-relevant annotations could be used to support a more informative interpretation of high-throughput studies. Our reanalysis of the top 200 differentially expressed transcripts in adult human VCMs[3] (compared with the human embryonic stem cells transcriptome; Table V in the Data Supplement) confirms the original findings of Poon et al[3] and, in addition, shows that the transcriptome of these cells is enriched for GO terms referring to cardiac physiology, including *cardiac conduction, regulation of cardiac conduction, regulation of cardiac muscle contraction by calcium ion signaling,* and *regulation of the force of heart contraction*. Notably, the enrichment of these terms is only possible because of the expansion of the GO resource in this area. These cardiac-relevant terms are identified despite more than half of these highly expressed proteins being associated with protein synthesis, mitochondrial respiratory systems, or having a structural cellular role. The significance of our work lies in the enhanced ability to identify cardiac-relevant genes within cardiac tissue; cardiac physiologists can now use the GO to test genomic profiles and confirm the presence of relevant genes. For example, researchers who are using in vitro cardiac cell differentiation systems can now test the profiles of gene expression in their cells to examine and confirm progression toward mature cardiomyocytes. This has potential application in the field of heart cell regeneration.

Furthermore, our analysis of QT interval length and atrial fibrillation risk loci, described by Tucker et al[2] and Arking et al,[31] demonstrates that our focused GO annotation of cardiac physiological processes provides an enhanced GO resource that can be used, in combination with other experimental and in silico approaches, to suggest candidate genes associated with cardiac dis-

Table 3. Comparison of Functional Analysis of Adult Ventricular Cardiomyocyte Data[3] Using Gene Ontology (Full List of Enriched Terms in Table V in the Data Supplement)

GO Identifier	GO Term Name	2011			2016		
		n	Corrected P Value	x	n	Corrected P Value	x
GO:0007005	Mitochondrion organization	#N/A	#N/A	#N/A	679	4.56E-12	24
GO:0022904	Respiratory electron transport chain	135	5.48E-34	29	157	1.67E-34	29
GO:0010467	Gene expression	2206	4.58E-11	45	5483	1.09E-13	73
GO:0032502	Developmental process	5680	4.09E-12	80	6093	1.82E-07	64
GO:0061337	Cardiac conduction	#N/A	#N/A	#N/A	126	1.71E-02	4
GO:1903779	Regulation of cardiac conduction	#N/A	#N/A	#N/A	68	1.57E-05	6
GO:0002026	Regulation of the force of heart contraction	#N/A	#N/A	#N/A	31	5.93E-06	5

A selection of Biological Process terms extracted from 2 analyses of the top 200 differentially expressed transcripts in hA-VCMs (adult human ventricular cardiomyocytes; compared with the human embryonic stem cells [hESCs] transcriptome[3]). Terms selected are either cardiac physiology relevant or align with terms identified by Poon et al.[3] Analyses were conducted using the BinGO plugin[33] within the Cytoscape v3.3.0 tool[34] using either the 2011 or 2016 Gene Ontology (GO) annotation data set. n indicates the number of protein IDs associated with the GO term, and x is the number of IDs in both the submitted list and associated with the GO term The inclusion of only the most abundant VCM transcripts probably accounts for the low number of genes in this data set associated with the cardiac terms. #N/A indicates term was not significantly enriched in the data set.

ease. In particular, our results suggest that *ATP1B1*, a Na⁺/K⁺-ATPase β-subunit, is the most likely candidate gene for further study at the 1q24 risk locus because it is associated with cardiac-relevant GO annotations. Another potential use of cardiac GO annotations is to extend protein network analysis. Overlaying proteins with relevant annotations (and hence known cardiac role) onto an existing network would highlight their interacting partners, and these could be explored as candidate genes.

With researchers turning to *Drosophila* and zebrafish as model systems for heart disease,[49,50] there is a requirement for GO to represent cardiac physiology using terms that are species independent.[6] The *Drosophila* pulsatile heart tube, with its anterior aorta, looks structurally very distinct from the zebrafish 2-chambered and the mammalian 4-chambered hearts. However, despite this, many of the new GO terms created during this project can be applied across all species because many orthologous genes have similar functions. The major limiting factor to the GO annotation of nonmammalian systems is the paucity of research describing the roles of individual gene products in the cardiac cycle in those species, as the majority of these data relate to mammals. Indeed, the curation effort described in this article focuses on improving analysis tools for human cardiovascular research.

Our results show that collaboration between members of relevant model organism databases and focused curation groups can improve and extend the capture of information about physiological processes in GO, in this particular study, cardiac processes. This effort was enhanced by the collaboration with experts in the cardiac research field who helped us identify key aspects of the process, key genes, and key publications on which to focus. As knowledge of the cardiac system continues to advance, there will be opportunities to capture additional cardiac information in GO. We have provided a structured framework for the addition of new knowledge to the resource, for example, to better describe information about the proteins involved in the mechanical aspect of the heart, and to fully capture roles of gene products and hormones in regulating cardiac physiology. Our work also shows that collaboration between groups developing ontologies and creating biological annotations and scientists and clinicians engaged in active research can lead to substantial improvements in the computational representation of biological knowledge. The GOC welcomes input from researchers about any aspect of our work, including changes to the ontology and suggestions of papers and gene products to annotate (goannotation@ucl.ac.uk, http://geneontology. org/page/contributing-go). A variety of GitHub repositories with issue trackers are in place for specific queries: for general inquiries about GO (https://github. com/geneontology/helpdesk), for specific questions about annotations or annotation-related topics (https:// github.com/geneontology/go-annotation/), and for specific questions or suggestions about the content and structure of the ontology (https://github.com/geneontology/go-ontology/). Such collaborations improve the value of the GO resource for the benefit of the entire cardiovascular research community and facilitate the interpretation of high-throughput data sets, toward the identification of dysregulated cardiac pathways, as well as variants and risk alleles, associated with cardiovascular diseases. This has important implications for cardiovascular physicians needing to interpret potentially pathological variants in their patients as a resource for the most up-to-date bioinformatic data to inform diagnosis and cascade screening in families.

DISCLOSURES

None.

AFFILIATIONS

From the Institute of Cardiovascular Science (R.C.L., V.K.K., R.E.F., N.H.C., R.P.H., P.J.T., P.D.L., P.M.E., L.C.) and Metabolism and Experimental Therapeutics, Division of Medicine (R.B.), University College London, United Kingdom; European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Hinxton, United Kingdom (P.R., D.O.-S.); Gene Ontology Consortium (P.R., T.Z.B., D.O.-S., J.A.B., D.P.H.); The Zebrafish Model Organism Database, University of Oregon,

Eugene (D.G.H.); Rat Genome Database, Human Molecular Genetics Center, Medical College of Wisconsin, Milwaukee (S.J.F.L.); Arabidopsis Information Resource, Phoenix Bioinformatics, Fremont, CA (T.Z.B.); FlyBase, University of Cambridge, United Kingdom (S.T.); Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME (J.A.B., D.P.H.); Oxbridge BHF Centre of Regenerative Medicine, Department of Physiology, Anatomy and Genetics, University of Oxford, United Kingdom (P.R.R.); and William Harvey Heart Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, United Kingdom (A.T.).

FOOTNOTES

REFERENCES

1. Spears DA, Gollob MH. Genetics of inherited primary arrhythmia disorders. *Appl Clin Genet*. 2015;8:215–233. doi: 10.2147/TACG.S55762.
2. Tucker NR, Clauss S, Ellinor PT. Common variation in atrial fibrillation: navigating the path from genetic association to mechanism. *Cardiovasc Res*. 2016;109:493–501. doi: 10.1093/cvr/cvv283.
3. Poon E, Yan B, Zhang S, Rushing S, Keung W, Ren L, et al. Transcriptome-guided functional analyses reveal novel biological properties and regulatory hierarchy of human embryonic stem cell-derived ventricular cardiomyocytes crucial for maturation. *PLoS One*. 2013;8:e77784. doi: 10.1371/journal.pone.0077784.
4. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25:25–29. doi: 10.1038/75556.
5. Bousette N, Kislinger T, Fong V, Isserlin R, Hewel JA, Emil A, et al. Large-scale characterization and analysis of the murine cardiac proteome. *J Proteome Res*. 2009;8:1887–1901.
6. Carnielli CM, Winck FV, Paes Leme AF. Functional annotation and biological interpretation of proteomics data. *Biochim Biophys Acta*. 2015;1854:46–54. doi: 10.1016/j.bbapap.2014.10.019.
7. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43:D1049–D1056. doi: 10.1093/nar/gku1179.
8. Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM. A guide to best practices for Gene Ontology (GO) manual annotation. *Database (Oxford)*. 2013;2013:bat054. doi: 10.1093/database/bat054.
9. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*. 2009;25:3045–3046. doi: 10.1093/bioinformatics/btp536.
10. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 2009;19:327–335. doi: 10.1101/gr.073585.107.
11. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43:D204–D212. doi: 10.1093/nar/gku989.
12. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. 2015;43(Database issue):D36–D42. doi: 10.1093/nar/gku1055.
13. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4:44–57. doi: 10.1038/nprot.2008.211.
14. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res*. 2016;44(W1):W83–W89. doi: 10.1093/nar/gkw199.
15. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res*. 2016;44(D1):D336–D342. doi: 10.1093/nar/gkv1194.
16. Richardson JE, Bult CJ. Visual annotation display (VLAD): a tool for finding functional themes in lists of genes. *Mamm Genome*. 2015;26:567–573. doi: 10.1007/s00335-015-9570-2.
17. De Giusti VC, Nolly MB, Yeves AM, Caldiz CI, Villa-Abrille MC, Chiappe de Cingolani GE, et al. Aldosterone stimulates the cardiac Na(+)/H(+) exchanger via transactivation of the epidermal growth factor receptor. *Hypertension*. 2011;58:912–919. doi: 10.1161/HYPERTENSIONAHA.111.176024.
18. Arnolds DE, Chu A, McNally EM, Nobrega MA, Moskowitz IP. The emerging genetic landscape underlying cardiac conduction system function. *Birth Defects Res A Clin Mol Teratol*. 2011;91:578–585. doi: 10.1002/bdra.20800.
19. Campuzano O, Brugada R. Genetics of familial atrial fibrillation. *Europace*. 2009;11:1267–1271. doi: 10.1093/europace/eup199.
20. Abriel H. Cardiac sodium channel Na(v)1.5 and interacting proteins: physiology and pathophysiology. *J Mol Cell Cardiol*. 2010;48:2–11. doi: 10.1016/j.yjmcc.2009.08.025.
21. Donoso P, Sanchez G, Bull R, Hidalgo C. Modulation of cardiac ryanodine receptor activity by ROS and RNS. *Front Biosci (Landmark Ed)*. 2011;16:553–567.
22. Dulhunty AF, Hewawasam R, Liu D, Casarotto MG, Board PG. Regulation of the cardiac muscle ryanodine receptor by glutathione transferases. *Drug Metab Rev*. 2011;43:236–252. doi: 10.3109/03602532.2010.549134.
23. Willis BC, Ponce-Balbuena D, Jalife J. Protein assemblies of sodium and inward rectifier potassium channels control cardiac excitability and arrhythmogenesis. *Am J Physiol Heart Circ Physiol*. 2015;308:H1463–H1473. doi: 10.1152/ajpheart.00176.2015.
24. Campuzano O, Allegue C, Fernandez A, Iglesias A, Brugada R. Determining the pathogenicity of genetic variants associated with cardiac channelopathies. *Sci Rep*. 2015;5:7953. doi: 10.1038/srep07953.
25. Fischer R, Dechend R, Gapelyuk A, Shagdarsuren E, Gruner K, Gruner A, et al. Angiotensin II-induced sudden arrhythmic death and electrical remodeling. *Am J Physiol Heart Circ Physiol*. 2007;293:H1242–H1253. doi: 10.1152/ajpheart.01400.2006.
26. Latronico MV, Condorelli G. MicroRNAs and cardiac conduction. *Curr Drug Targets*. 2010;11:907–912.
27. Kim GH. MicroRNA regulation of cardiac conduction and arrhythmias. *Transl Res*. 2013;161:381–392. doi: 10.1016/j.trsl.2012.12.004.
28. Poudel P, Xu Y, Cui Z, Sharma D, Tian B, Paudel S. Atrial fibrillation: recent advances in understanding the role of microRNAs in atrial remodeling with an electrophysiological overview. *Cardiology*. 2015;131:58–67. doi: 10.1159/000375403.
29. Huntley RP, Sitnikov D, Orlic-Milacic M, Balakrishnan R, D'Eustachio P, Gillespie ME, et al. Guidelines for the functional annotation of microRNAs using the Gene Ontology. *RNA*. 2016;22:667–676. doi: 10.1261/rna.055301.115.
30. Huntley RP, Harris MA, Alam-Faruque Y, Blake JA, Carbon S, Dietze H, et al. A method for increasing expressivity of Gene Ontology annotations using a compositional approach. *BMC Bioinformatics*. 2014;15:155. doi: 10.1186/1471-2105-15-155.
31. Arking DE, Pulit SL, Crotti L, van der Harst P, Munroe PB, Koopmann TT, et al; CARe Consortium; COGENT Consortium; DCCT/EDIC; eMERGE Consortium; HRGEN Consortium. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nat Genet*. 2014;46:826–836. doi: 10.1038/ng.3014.
32. Tucker NR, Ellinor PT. Emerging directions in the genetics of atrial fibrillation. *Circ Res*. 2014;114:1469–1482. doi: 10.1161/CIRCRESAHA.114.302225.
33. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*. 2005;21:3448–3449. doi: 10.1093/bioinformatics/bti551.
34. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 2011;27:431–432. doi: 10.1093/bioinformatics/btq675.
35. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, et al. Cross-product extensions of the Gene Ontology. *J Biomed Inform*. 2011;44:80–86. doi: 10.1016/j.jbi.2010.02.002.
36. Lovering RC, Dimmer E, Khodiyar VK, Barrell DG, Scambler P, Hubank M, et al. Cardiovascular GO annotation initiative year 1 report: why cardiovascular GO? *Proteomics*. 2008;8:1950–1953. doi: 10.1002/pmic.200800078.

37. Dowell KG, McAndrews-Hill MS, Hill DP, Drabkin HJ, Blake JA. Integrating text mining into the MGI biocuration workflow. *Database (Oxford)*. 2009;2009:bap019. doi: 10.1093/database/bap019.

38. Wang SJ, Laulederkind SJ, Hayman GT, Smith JR, Petri V, Lowry TF, et al. Analysis of disease-associated objects at the Rat Genome Database. *Database (Oxford)*. 2013;2013:bat046. doi: 10.1093/database/bat046.

39. Ruzicka L, Bradford YM, Frazer K, Howe DG, Paddock H, Ramachandran S, et al. ZFIN, The zebrafish model organism database: updates and new directions. *Genesis*. 2015;53:498–509. doi: 10.1002/dvg.22868.

40. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al.; FlyBase Consortium. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*. 2009;37(Database issue):D555–D559. doi: 10.1093/nar/gkn788.

41. Ziane R, Huang H, Moghadaszadeh B, Beggs AH, Levesque G, Chahine M. Cell membrane expression of cardiac sodium channel Na(v)1.5 is modulated by alpha-actinin-2 interaction. *Biochemistry*. 2010;49:166–178. doi: 10.1021/bi901086v.

42. Surber R, Hensellek S, Prochnau D, Werner GS, Benndorf K, Figulla HR, et al. Combination of cardiac conduction disease and long QT syndrome caused by mutation T1620K in the cardiac sodium channel. *Cardiovasc Res*. 2008;77:740–748. doi: 10.1093/cvr/cvm096.

43. Laurent G, Saal S, Amarouch MY, Béziau DM, Marsman RF, Faivre L, et al. Multifocal ectopic Purkinje-related premature contractions: a new SCN5A-related cardiac channelopathy. *J Am Coll Cardiol*. 2012;60:144–156. doi: 10.1016/j.jacc.2012.02.052.

44. Roder K, Werdich AA, Li W, Liu M, Kim TY, Organ-Darling LE, et al. RING finger protein RNF207, a novel regulator of cardiac excitation. *J Biol Chem*. 2014;289:33730–33740. doi: 10.1074/jbc.M114.592295.

45. Li Y, Yang CM, Xi Y, Wu G, Shelat H, Gao S, et al. MicroRNA-1/133 targeted dysfunction of potassium channels KCNE1 and KCNQ1 in human cardiac progenitor cells with simulated hyperglycemia. *Int J Cardiol*. 2013;167:1076–1078. doi: 10.1016/j.ijcard.2012.10.060.

46. Gaudet P, Livstone MS, Lewis SE, Thomas PD. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform*. 2011;12:449–462. doi: 10.1093/bib/bbr042.

47. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, et al. The UniProt-GO annotation database in 2011. *Nucleic Acids Res*. 2012;40(Database issue):D565–D570. doi: 10.1093/nar/gkr1048.

48. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508:469–476. doi: 10.1038/nature13127.

49. Cammarato A, Ahrens CH, Alayari NN, Qeli E, Rucker J, Reedy MC, et al. A mighty small heart: the cardiac proteome of adult Drosophila melanogaster. *PLoS One*. 2011;6:e18497. doi: 10.1371/journal.pone.0018497.

50. Asnani A, Peterson RT. The zebrafish as a tool to identify novel therapies for human cardiovascular disease. *Dis Model Mech*. 2014;7:763–767. doi: 10.1242/dmm.016170.