



ELSEVIER

Contents lists available at ScienceDirect

## Data in Brief

journal homepage: [www.elsevier.com/locate/dib](http://www.elsevier.com/locate/dib)



Data article

# Prediction and visualization data for the interpretation of sarcomeric and non-sarcomeric DNA variants found in patients with hypertrophic cardiomyopathy



Irene Bottillo<sup>a,\*</sup>, Daniela D'Angelantonio<sup>a</sup>, Viviana Caputo<sup>b</sup>, Alessandro Paiardini<sup>c</sup>, Martina Lipari<sup>a</sup>, Carmelilia De Bernardo<sup>a</sup>, Silvia Majore<sup>a</sup>, Marco Castori<sup>a</sup>, Elisabetta Zachara<sup>d</sup>, Federica Re<sup>d</sup>, Paola Grammatico<sup>a</sup>

<sup>a</sup> Medical Genetics, Department of Molecular Medicine, Sapienza University, San Camillo-Forlanini Hospital, Rome, Italy

<sup>b</sup> Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy

<sup>c</sup> Department of Biochemical Sciences, Sapienza University of Rome, Rome, Italy

<sup>d</sup> Cardiomyopathies Unit, Division of Cardiology and Cardiac Arrhythmias, San Camillo-Forlanini Hospital, Rome, Italy

## ARTICLE INFO

### Article history:

Received 8 December 2015

Received in revised form

16 February 2016

Accepted 1 March 2016

Available online 10 March 2016

## ABSTRACT

Genomic technologies are redefining the understanding of genotype–phenotype relationships and over the past decade, many bioinformatics algorithms have been developed to predict functional consequences of single nucleotide variants. This article presents the data from a comprehensive computational workflow adopted to assess the biomedical impact of the DNA variants resulting from the experimental study “Molecular analysis of sarcomeric and non-sarcomeric genes in patients with hypertrophic cardiomyopathy” (Bottillo et al., 2016) [1]. Several different independently methods were employed to predict the functional consequences of alleles that result in amino acid substitutions, to study the effect of some DNA variants over the splicing process and

DOI of original article: <http://dx.doi.org/10.1016/j.gene.2015.11.048>

\* Correspondence to: Medical Genetics, Department of Molecular Medicine, Sapienza University, San Camillo-Forlanini Hospital Circonvallazione Gianicolense, 87 - 00152 Rome, Italy. Tel.: +39 06 58704622; fax: +39 06 5870 4657.

E-mail address: [i.bottillo@gmail.com](mailto:i.bottillo@gmail.com) (I. Bottillo).

<http://dx.doi.org/10.1016/j.dib.2016.03.004>

2352-3409/© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

to investigate the impact of a sequence variant with respect to the evolutionary conservation.

© 2016 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## Specifications Table

Subject area	Biology
More specific sub- ject area	In silico predictions of DNA variants
Type of data	Tables, figures
How data was acquired	Prediction tools: SIFT, Polyphen HDIV, Polyphen HVAR, Provean, LRT, Mutation Taster, Mutation Assessor, FATHMM, RadialSVM, LR, CADD, HSF, GERP++, PhyloP placental, PhyloP vertebrate, SiPhyMolecular Modeling
Data format	Processed, filtered and analyzed
Experimental factors	Genomic DNA from peripheral blood was tested by next generation sequencing on Ion Torrent PGM (ThermoFisher, Carlsbad, CA, USA) with a custom cardiomyopathy panel
Experimental features	The identified rare (Minor Allele Frequency $\leq 0,01$ ) non-synonymous DNA changes were subjected to different <i>in silico</i> predictions
Data source location	Rome, Italy
Data accessibility	These data are with this article

## Value of the data

- These data delineate a prompt informatic pipeline for the prioritization of the most likely pathogenic DNA variants in a clinical context.
- These data are supportive for the researchers to evaluate the prevalence of sarcomeric and non-sarcomeric gene variants in hypertrophic cardiomyopathy.
- The described computational strategy is helpful to researchers for the rapid interpretation of Variants of Unknown Significance (VUS) implicated in rare, common and complex diseases.

### 1. Data

Here we report the *in silico* predictions data of the non-synonymous changes found in 41 HCM patients and in 3 HCM-related cases [1] (Table 1).

### 2. Experimental design, materials and methods

#### 2.1. Analysis of the nucleotides' evolutionary conservation

Nucleotide-specific estimates of evolutionary constraint were explored by (i) GERP++ (Genomic Evolutionary Rate Profiling); (ii) PhyloP placental; (iii) PhyloP vertebrate and (iv) SiPhy.

#### 2.2. Analysis of the splicing variants

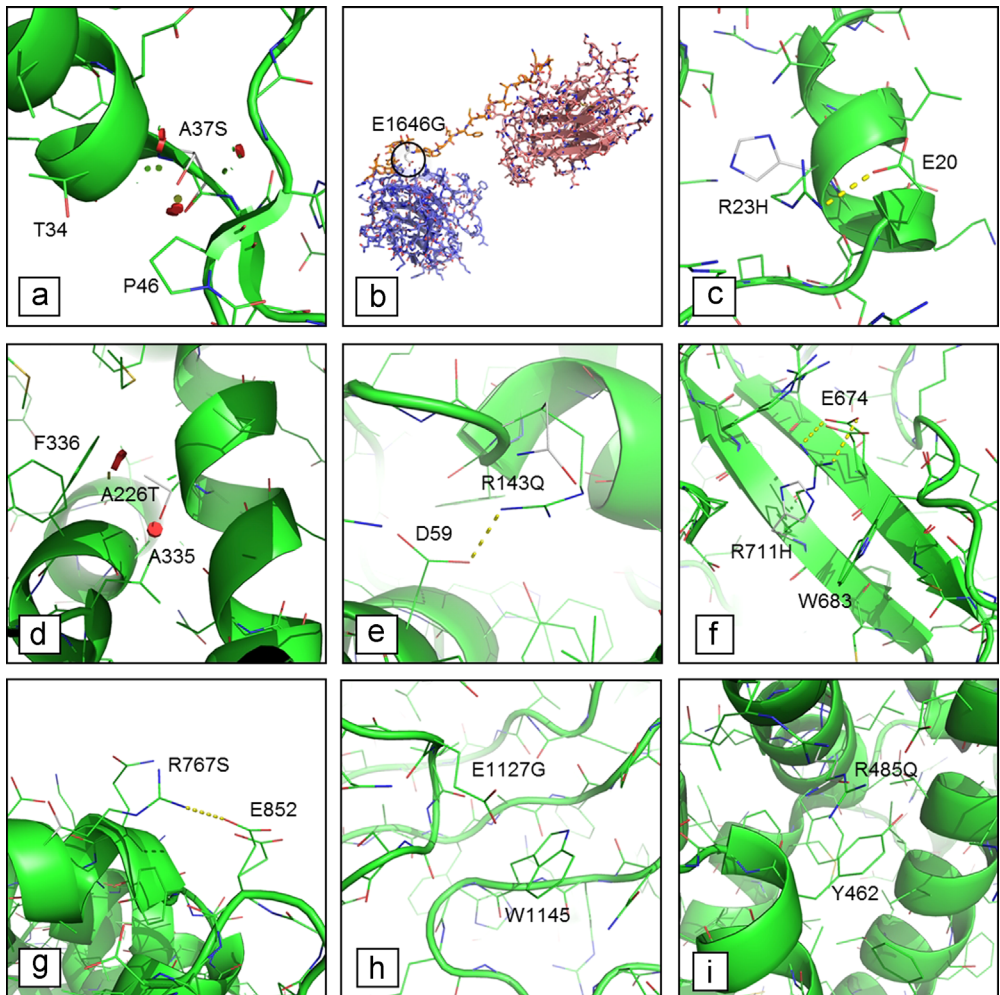
The analysis of intronic variants leading to splicing defects was tested by Human Splicing Finder (HSF) 3.0.



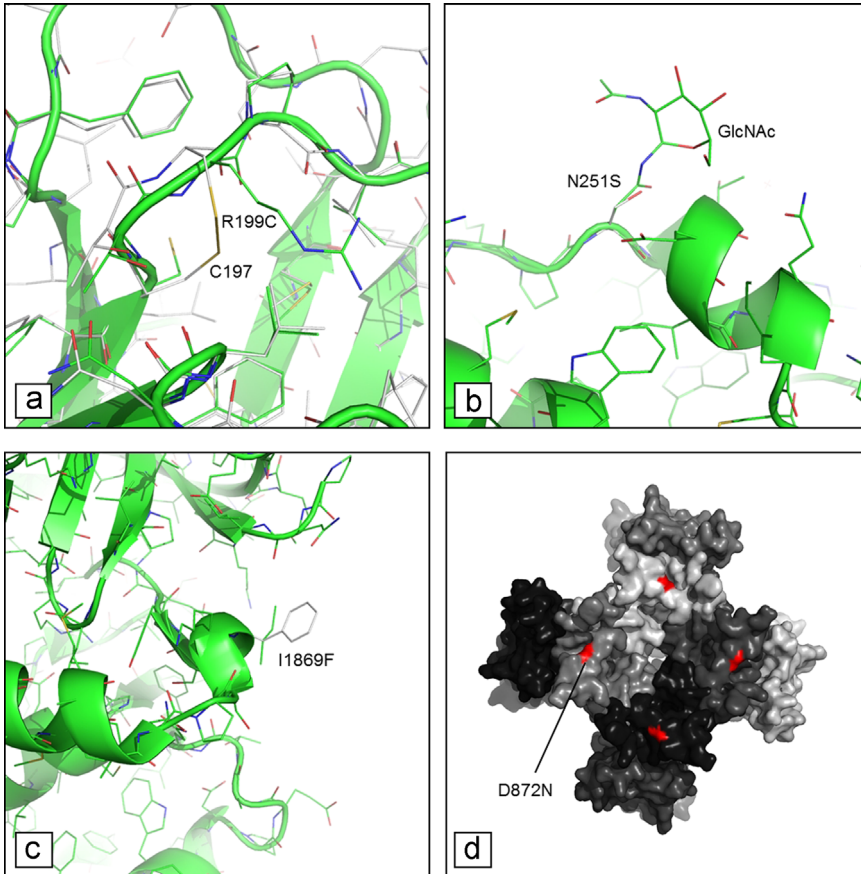


### 2.3. Analysis of the missense variants

The effect of missense changes on the structure and function of a human protein was predicted by: (i) SIFT (Sorting Intolerant From Tolerant), (ii) PolyPhen-2 (Polymorphism Phenotyping v2) HDIV, that identifies human damaging mutations by assuming differences between human proteins and their closely related mammalian homologs as non-damaging; (iii) PolyPhen-2 HVAR, that identifies human disease-causing mutations by assuming common human nsSNPs as non-damaging; (iv) Provean (Protein Variation Effect Analyzer); (v) LRT (Likelihood Ratio Test) that identifies conserved amino acid positions and deleterious mutations using a comparative genomics data set of multiple vertebrate species; (vi) Mutation Taster; (vii) Mutation Assessor; (viii) FATHMM (Functional Analysis through Hidden Markov Models); (ix) RadialSVM (Radial Support Vector Machine); (x) LRT (Logistic Regression Test); (xi) CADD v1.3 (Combined Annotation–Dependent Depletion), a method for objectively integrating many diverse annotations into a single measure (C score) for each variant; and (xii) molecular modeling.



**Fig. 1.** Structural comparison of wild-type and mutant forms for (a) FLH2 A37S; (b) LAMA4 E1646G; (c) MYH6 R23H; (d) MYH7 A226T; (e) MYH7 R143Q; (f) MYOM1 R711H; (g) PKP2 R767S; (h) RYR2 E1127G; (i) RYR2 R485Q. The mutation is indicated in white. The predicted structural effects of mutations are: (a, d) steric hindrance (red circles); (b) local misfolding of linker domain (orange); (c, e, f, g) loss of important inter-residues contacts; (h) loss of a  $\pi$ -anion interaction; (i) loss of a  $\pi$ -cation interaction.



**Fig. 2.** Effects of nsSNVs for: (a) the cadherin domain of DSC2. The mutant R199C in the cadherin domain of DSC2 is predicted to introduce a disulfide bond with the near Cys197 residue ( $C\alpha$ - $C\alpha$  distance  $\sim 6$  Å), and possibly to result in local misfolding of the cadherin domain; (b) the melibiase domain of GLA. Mutant N215S of the melibiase domain of GLA results in the loss of a glycosylated site probably affecting the protein structure and/or function; (c) the FGF13 interaction domain of SCN5. Mutation I869F localizes on a solvent-exposed hydrophobic path of the domain of interaction with fibroblast growth factor 13 (FGF13). The I869F mutation could affect the recognition of the FGF13 protein; (d) the Na-Channel of SCN5. The mutant D872N results in the loss of a negative charge that is approximately located at the Na-channel domain of SCN5, probably affecting cations conductance of the channel. The approximate position of the negatively charged Asp872 residue is shown in red, in each of the four protein subunits forming the channel.

Regarding the molecular modeling, protein structure were experimentally determined by X-ray crystallography, or were inferred by homology modeling means (i.e., availability of a structural template with percentage of identity  $> 20\%$ ). Protein models were built using the homology modeling approach implemented in modeller-9 package [2]. PSI-BLAST was used to find suitable structural templates for each sequence to model [3]. The sequences of each protein target to model and its structural template were then aligned by using the program CLUSTALW [4] and manually manipulated to optimize the matching of several characteristics, including the observed and predicted secondary structural elements, the hydrophobic regions in the three-dimensional structures, the structurally and functionally conserved residues, and *indel* regions in the structures. Then, ten different models were built for each target protein and evaluated using several criteria. The model displaying the lowest objective function [5], which measures the extent of violation of constraints from the structural templates, was taken as the representative model. Superimposition and root-mean-square deviation (RMSD) calculation of  $C\alpha$  traces of the 10 models were performed to detect the most variable and therefore less reliable modeled regions. These invariably corresponded to loop elements.

Procheck [6] was used to monitor the stereochemical quality of the representative models, whereas ProsaII [7] was used to measure the overall protein quality in packing and solvent exposure. Mutations on protein structures was carried out using the “Mutate model” script implemented in modeller-9 package [2]. The script takes as input a given three-dimensional structure of a protein (experimentally determined or predicted), and mutates a single residue. The residue sidechain's position is then optimized by energy minimization and refined by molecular dynamics simulations. Prediction of protein stability upon mutation was carried out using the DUET server [8]. Sequence identity between the modeled domain and its closest template ranged from 23% (Laminin G-like domain of LAMA4), to nearly 95% (N-terminal globular head domain of VCL). However, in spite of the low value of sequence identity measured in some cases, all of the models resulted in a good overall quality (Prosa Z-score < -2.00), except for CALR3 and SCN5. Given the short length of the predicted PB035848 domain of CALR3 (residues 294–347) and its sequence identity with its template (61%), the measured Prosa Z-score (-1.93) nonetheless indicated a model of quality comparable to a Nuclear Magnetic Resonance (NMR) structure [7] ( Figs. 1 and 2).

## Acknowledgments

This work funded by the Department of Molecular Medicine, Sapienza University of Rome. This work was also partially funded by the Department of Biochemical Sciences Sapienza University of Rome (prot. C26A149EC4).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.dib.2016.03.004>.

## References

- [1] I. Bottillo, D. D'Angelantonio, V. Caputo, A. Paiardini, M. Lipari, C. De Bernardo, D. Giannarelli, A. Pizzuti, S. Majore, M. Castori, E. Zachara, F. Re, P. Grammatico, Molecular analysis of sarcomeric and non-sarcomeric genes in patients with hypertrophic cardiomyopathy, *Gene* 577 (2) (2016) 227–235.
- [2] N. Eswar, B. Webb, M.A. Marti-Renom, M.S. Madhusudhan, D. Eramian, M.Y. Shen, U. Pieper, A. Sali, Comparative protein structure modeling using modeller, in: Andreas D. Baxevanis, et al., (Eds.), *Current Protocols in Bioinformatics*, 2006, Chapter 5 Unit 5 6.
- [3] I. Friedberg, T. Kaplan, H. Margalit, Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments, *protein science: a publication of the protein, Society* 9 (11) (2000) 2278–2284.
- [4] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic acids Res.* 22 (22) (1994) 4673–4680.
- [5] D.F. Burke, C.M. Deane, H.A. Nagarajaram, N. Campillo, M. Martin-Martinez, J. Mendes, F. Molina, J. Perry, B.V. Reddy, C. M. Soares, R.E. Steward, M. Williams, M.A. Carrondo, T.L. Blundell, K. Mizuguchi, An iterative structure-assisted approach to sequence alignment and comparative modeling, *Proteins (Suppl 3)* (1999) S55–S60.
- [6] R.A. Laskowski, J.A. Rullmann, M.W. MacArthur, R. Kaptein, J.M. Thornton, AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, *J. Biomol. NMR* 8 (4) (1996) 477–486.
- [7] M.J. Sippl, Recognition of errors in three-dimensional structures of proteins, *Proteins* 17 (4) (1993) 355–362.
- [8] D.E. Pires, D.B. Ascher, T.L. Blundell, DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach, *Nucleic Acids Res.* 42 (2014) W314–W319.