The structural effects of mutations can aid in differential phenotype prediction of beta-myosin heavy chain (Myosin-7) missense variants

Nouf S. Al-Numair, Luis Lopes, Petros Syrris,
Lorenzo Monserrat, Perry Elliott and Andrew C.R. Martin

Supplementary Figures and Tables

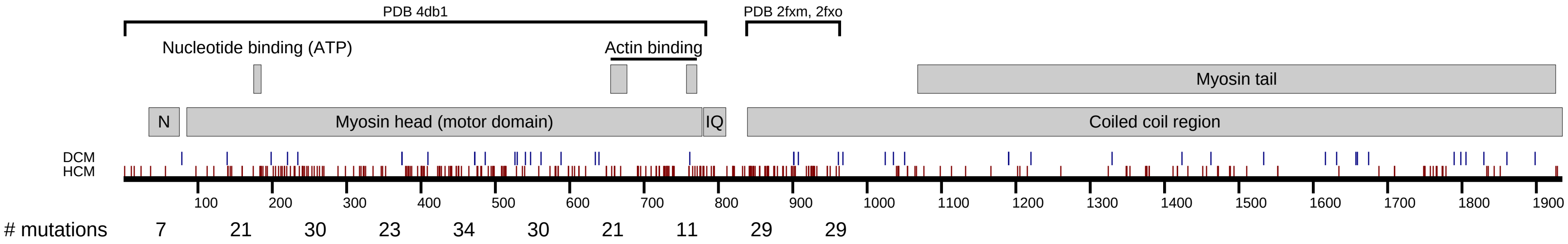

Figure S1: Annotated regions of the Myosin-7 sequence. Regions for which structures are known are indicated (top line), together with the number of known mutations from Table S2 in each 100 amino acids of the sequence (bottom line). Annotated domains from secondary databanks are also indicated: **N (Myosin N-terminal domain)** Pfam annotation, residues 34–75; **Myosin head (motor domain)** Pfam and InterPro annotation, residues 85–778; **IQ motif** UniProtKB/SwissProt and InterPro annotation, residues 781–810, SMART annotation, residues 780–802; **Coiled coil region** UniProtKB/SwissProt annotation, residues 839–1935, SMART annotation, residues 841–1927; **Nucleotide binding (ATP) region** UniProtKB/SwissProt annotation, residues 178–185; **Actin-binding region** UniProtKB/SwissProt annotation, residues 655–677; **Actin-binding region** UniProtKB/SwissProt annotation, residues 757–771; **Myosin tail** Pfam and InterPro annotation, residues 1068–1926.

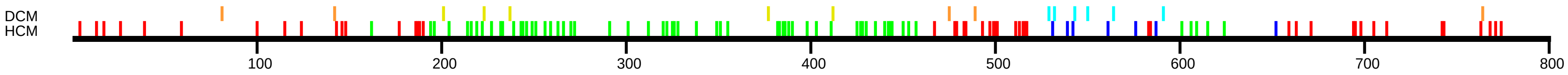

Figure S2: HCM and DCM mutations mapped to structural clusters as shown in Figure 2. For the three clusters, HCM mutations are shown in 1: red, 2: green and 3: blue, while DCM mutations are shown in 1: orange, 2: yellow and 3: cyan.

Analysis	Features	Type
Binding	Is the residue involved in binding (defined by presence of specific contacts with another protein chain or ligand)?	Boolean
Interface	Is the residue in an interface (defined by change in solvent accessibility between complexed and uncomplexed forms)?	Boolean
SProtFT	Is the residue annotated with a functionally relevant SwissProt feature?	Boolean
	Which of 12 SwissProt features appear? (ACT_SITE, BINDING, CA_BIND, DNA_BIND, NP_BIND, METAL, MOD_RES, CARBOHYD, MOTIF, LIPID, DISULFID, CROSSLNK)?	12 x Boolean
RelAccess	Relative solvent accessibility of the residue	Percentage
ImPACT	ImPACT conservation score for the residue if it is found to be significantly conserved	Real
HBond	If the native residue was involved in a hydrogen bond, the difference in hydrogen bonding pseudo-energy	Real
SurfacePhobic	The difference in hydrophobicity if the residue is on the surface and the hydrophobicity has increased	Real
CorePhilic	The difference in hydrophobicity if the residue is buried and the hydrophobicity has decreased	Real
BuriedCharge	The difference in charge if the residue is buried	Integer
SSGeom	Was the native residue involved in a disulphide bond?	Boolean
Void	The difference in size of the largest void	Real
	The sizes of the 10 largest voids in the native protein	10 x Real
	The sizes of the 10 largest voids in the mutant protein	10 x Real
Clash	The sum of the van der Waals and torsional energy for the minimum perturbation protocol modelled sidechain replacement	Real
Glycine	If the native residue was a glycine, the Ramachandran pseudo-energy difference of the mutation	Real
Proline	If the mutant residue was a proline, the Ramachandran pseudo-energy difference of the mutation	Real
CisPro	Was the native residue a cis-proline?	Boolean

Table S1: The 47 features used in SAAPpred machine learning derived from the 14 structural analyses in SAAPdap.

Disease (Phenotype)	Unique[†] mutations	Mutations mapped to PDB
HCM	290	190
DCM	46	21
RCM	1	0
LVNC	17	9
LVNC/ASD	1	1
DCM/Endocardial Fibroelastosis	1	1
DCM/LVNC	3	1
HCM/LVNC	1	0
HCM/DCM/LVNC	2	0
HCM/DCM	3	0
HCM/RCM/DCM	1	0
HCM/Myopathy central core	1	1
Laing distal myopathy	1	0
Distal myopathy	3	0
Ebstein	5	3
Cardiomyopathy and distal myopathy	2	1
Myosin storage myopathy	3	0
Hyaline body myopathy	1	0
No recorded phenotype	13	10
Total	395	238

Table S2: Numbers of *MYH7* mutations for each phenotype. Abbreviations: PDB, Protein DataBank; DCM, Dilated Cardiomyopathy; HCM, Hypertrophic Cardiomyopathy; RCM, Restrictive Cardiomyopathy; LVNC, Left Ventricular Non-compaction; ASD, Atrial Septal Defect. The mutations for which there was no recorded phenotype were excluded from structural analysis, meaning that only 228 mutations which mapped to PDB structures could be analysed. For the novel differential phenotype predictor, only the 211 unique HCM and DCM mutations that mapped to PDB structures were used. [†]Unique mutations represents the number of non-redundant mutations at the protein level. Multiple observations of the same mutation (because the DNA level mutation is different or because of redundancy between different data sources) have been removed.

PDB	Resolution	Used in paper[†]	Release date	Start residue	End residue	Notes
4db1	2.6Å	Y	25.01.12	2	777	
4p7h	3.2Å	X	21.05.14	1	787	Chimera
4pa0	2.25Å	X	08.07.15	1	787	Chimera
1ik2	Model	N	01.05.01	1	841	Model
2fxm	2.7Å	Y	21.11.06	838	963	
2fxo	2.5Å	Y	21.11.06	838	963	
3dtp	20.0Å	N	07.10.08	842	963	Chimera
4xa1	3.2Å	X	01.07.15	1173	1238	Chimera
4xa3	2.55Å	X	01.07.15	1361	1425	Chimera
5cj1	2.1Å	X	02.12.15	1526	1571	Chimera
4xa4	2.33Å	X	01.07.15	1551	1609	Chimera
5chx	2.3Å	X	02.12.15	1590	1657	Chimera
5cj0	2.3Å	X	02.12.15	1631	1692	Chimera
5cj4	3.1Å	X	02.12.15	1562	1622	Chimera
4xa6	3.42Å	X	01.07.15	1777	1855	Chimera

Table S3: Structures of regions of the MYOSIN-7 protein (UniProtKB/SwissProt accession code P12883) available in the Protein Databank (PDB). PDB files may be accessed at `http://www.pdb.org/` or viewed using PDBSum (`http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/`). Note that PDB file 2fxo contains a mutation Glu924Lys. [†]X = Extra files, all of which are chimeric structures, added to the PDB since our dataset was built.

Feature	χ^2
Impact (conservation)	19.9
Glycine	11.9
Binding	4.9
CisPro	0.9
Clash	0.89
Buried Charge	0.625
Voids	0.199
Surface Phobic	0.15
Core Philic	0.09
Proline	0.03
Interface	N/A[†]
Disulphide	N/A
Hbonds	N/A

Table S4: Chi-squared tests were performed using each of the features to judge their ability to discriminate between HCM and DCM in order to inform feature selection. In each case where a parameter is a real number (rather than boolean, see Table S1), a threshold was used as described by Al-Numair*et al.* (2013) to classify a local structural effect as being present or not. For each feature, a 2x2 contingency table was constructed (Effect: present/not-present *vs.* Phenotype: HCM/DCM). [†]N/A: χ^2 tests were not performed where the same result (effect either present or not-present) was seen for all the mutations analyzed. Relative accessibility was not included as a parameter in the SAAPdap work and therefore no threshold was available and χ^2 values could not be calculated.

	Total		With Structure	
Domain	HCM	DCM	HCM	DCM
ATP binding	1	0	1	0
Actin binding 1	5	0	5	0
Actin binding 2	4	1	4	1
Coiled coil region	110	26	48	3
IQ domain	8	0	1	0
Myosin N-terminus	2	0	2	0
Myosin head (motor domain)	157	19	133	17
Myosin tail	51	18	0	0

Table S5: Numbers of HCM and DCM mutations seen in each of the annotated domains. Mutations occurring in the ATP binding region and the two actin binding regions are also counted as being in the Myosin head (motor domain).

		Pathogenicity		Phenotype	
Mutation	Phenotype	Prediction	Confidence	Prediction	Confidence
R869C	Undefined	PD	0.74	DCM	0.194
L908V	HCM+MCC	PD	0.44	HCM	0.482
E903K	Undefined	PD	0.80	HCM	0.754
Y501H	Undefined	PD	0.53	DCM	0.410
D955N	LVNC	PD	0.60	HCM	0.481
G584R	Undefined	PD	0.34	HCM	0.034
L390P	Ebstein	PD	0.47	HCM	0.341
R422H	EF	PD	0.26	HCM	0.324
I909M	Undefined	PD	0.70	HCM	0.649
R403L	Undefined	SNP	0.12	HCM	0.004

Table S6: Pathogenicity and differential phenotype predictions for 10 randomly chosen ‘other’ mutations some collected after the main dataset. MCC: myopathy central core; LVNC: Left Ventricular Non-compaction; EF: Endocardial Fibroelastosis.