Group project 1: Exploring variants in MYBPC family of genes in the context of complex genetic diseases

MYBPC paralogs play a major role in striated muscle contraction. Increasing evidence suggests that genetic alterations in MYBPC paralogs are directly linked to myopathies.

A systematic analysis can be carried out to determine nucleotide pattern, codon, and amino acid changes in existing genetic mutations among these three proteins.

Task
The VCF (File VCF_Group_Project_1.vcf) contains coding variants in each paralog obtained from gnomAD.

Ensembl
Use Ensembl VEP to compare the distribution of genetic variants in MYBPC paralogs:

1. Across the three genes, which gene has the highest prevalence of likely deleterious coding variants (missense, frameshift and truncation variants)
2. Are there any paralog-specific alterations in amino acids
3. Are there any domains that are highly prone to variants, which could, therefore, be a potential therapeutic target in disease conditions.

G2P
1. Are there any diseases associated with any of the MYBPC genes that you can identify in the Gene2Phenotype resource?
2. Looking at the information held for MYBPC1, would you expect someone with a genetic change in both copies of the gene which resulted in no gene product being made to have the same condition as someone with only one changed gene copy, where a protein is produced with only one amino acid is changed?
3. Looking at the information held for MYBPC3, what is the mechanism of disease and what variant types have been observed to cause it? Does the information available suggest that everyone with a potentially deleterious variant will develop the condition? Is there strong evidence for the disease association?

Uniprot/ProtVar

What are the UniProt IDs that correspond to the genetic variants we have been exploring?

Q. For MYBPC3 Q14896 what genetic variants have been highlighted in the UniProt record as likely significant for '*Cardiomyopathy, familial hypertrophic, 4 (CMH4)*'?

Hint: look at the "variants and disease" section.

Destabilisation of the protein by a variant is one potential functional mechanism. You may notice that some of the variants have been shown to be destabilising while for others there is no such data.

Click on P>S for position 161 to go to ProtVar and then on the functional annotation button  to investigate whether this variant is predicted to be destabilising.

Look at 237 Y>S in ProtVar as well and look at the stability. You can investigate the position of these variants in structure by using the "structural annotations" button: 

Returning to UniProt, at the end of the "variants and disease" section (or in a tab at the top of the page) you can open the variants viewer. This displays all the variants for the protein both curated and from large scale studies. There are filters to the left:

Q. Highlight all those variants that have been curated by UniProt and labelled as pathogenic. Click on one of the variant changes eg and investigate whether all of the variant effect predictors (in the "functional annotations" section) agree with the curators. Comment on the use of VEPs without annotations or expert knowledge.

PDB

PDB task 1:

Jupyter / Colab Notebook for this task: 01_structures_availability (https://github.com/PDBeurope/pdbe-notebooks/blob/main/variants_embl-ebi_july2025/01_structures_availability.ipynb)

- Fill-in information on predicted and experimentally-determined structures availability. *HINT: Use UniProt ID to help find structures.*

Table 1: General Structure Availability	
No of predicted & exp-determined structures	
No of predicted structures	
No of exp-determined structures	
No of exp-determined structures from PDB	
No of exp-determined structures with resolution better than 3 Ångstrom	
PDB id(s) with best resolution	
PDB id(s) with most coverage	
PDB id(s) without any mutations	

PDB task 2:
Jupyter / Colab Notebook for this task:01_structures_availability (https://github.com/PDBeurope/pdbe-notebooks/blob/main/variants_embl-ebi_july2025/01_structures_availability.ipynb)

The following structures may help give insight into the genetic variant:

1. Experimentally-determined structure with an amino acid change at the position
2. Experimentally-determined structure with information at the variant position and no other mutations or variants
3. Experimentally-determined structure with information at the variant position but with additional mutations or variants
4. Predicted structure with information at the variant position

The usefulness of the structures for insight into genetic variants is as follows:
(1) is best, BUT if (1) does not exist, should consider (2), and if (2) does not exist should consider (3) or (4).

List 4-10 single amino acid changes from previous analysis (included 2 examples)	222 (Tyr/His) for MYBPC1, 5 (Glu/Arg) for MYBPC3,

- Identify the corresponding data from predicted and experimentally-determined structures for a specific set of variants.
  HINT: Use UniProt numbering for variant position.

Table 2: Find PDB ids / AFDB ids with coverage for understanding the variant		
Variant position	Best structure type available	PDB or AFDB id(s)
222	(4) Predicted	AF-Q00872-F1
5	(2) Exp-det (no mut)	8g4l, 6cxi, 6cxj, 2k1m

PDB task 3:
Jupyter / Colab Notebook for this task:02_structural_visuallisation (https://github.com/PDBeurope/pdbe-notebooks/blob/main/variants_embl-ebi_july2025/02_structures_visuallisation.ipynb)

Visualising is a key part of understanding protein structures. We will use some of the structures previously identified for visualisation. Mutations that are clustered together spatially in a structure may be impacting in a similar manner. This visualisation will be useful for PDB task 5.

- Using the tool developed by PDBe and RSCB and partners called MolStar we can view where the variants are in a structure.

PDB task 4:

- Use the Ligand subpage of a PDBe-KB page to assess whether a ligand binding site is applicable for the variant or not.This analysis will be useful for PDB task 5.
  HINT1: Use UniProt ID to find the appropriate PDBe-KB page.
  HINT2: If no experimental structures are available, skip this task.

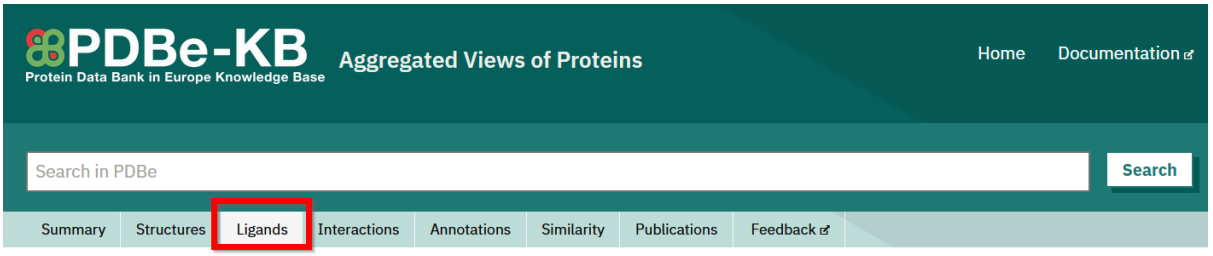

Table 3: Find PDB ids / AFDB ids if ligand is involved		
Variant position	Near ligand site? (CCD ID)	PDB ids
222	N/A	
5	N/A	

PDB task 5:
Jupyter / Colab Notebook for this task: 03_structural_assessment (https://github.com/PDBeurope/pdbe-notebooks/blob/main/variants_embl-ebi_july2025/03_structural_assessment.ipynb)

- If possible, propose an interpretation of why the genetic variant may impact on the protein's function and thus cause disease
  HINT: Use the tools from PDB task 3 & 4 to aid answering structural assessment questions.

Table 4: Propose why variant might impact on protein function	
Variant	**Variant interpretation (hypothesis based on structure)**
222 (Ala/Thr)	A charged amino acid has replaced an aromatic / hydrophobic amino acid in a hydrophobic pocket in the middle of a protein domain. This likely destabilizes the protein fold for this protein domain.
5 (Glu/Arg)	Unclear. May need further analysis or experimentation.

Open Targets

Based on the gene with a higher proportion of likely deleterious coding variants from the Ensemble exercise above:

- Go to the target page in the Open Target Platform, and list the gene symbol and Ensembl IDs for all human paralogs
- You should have a list of 11 targets that we now want to prioritise in terms of their tractability and doability properties for treating cardiomyopathies. To do so, go to the cardiomyopathy association page and use the facets to filter and pin our 11 targets by clicking on the gene symbol. This will put them together at the top of the page.
    - Which gene presents the highest association score? What is this value? Which source of data supports the association with cardiomyopathy?
    - Go to the Target prioritisation page to assess qualities that distinguish target doability in a drug development context. A key aspect is if any of those targets are already in the clinic, is that the case?
    - Paying special attention to target specificity, which might be an indicator of target safety, which target would you prioritise for further development?