Proteomic Subprofiles Reveal the Organization of Biological Processes Across Cardiac Diseases

Juan Francisco Pérez López[1], Sergio Muñoz-Romero[1], Francisco-Javier Gimeno-Blanes[2], Fernando Corrales[3], Maria Sabater-Molina[4], Carmen Muñoz-Esparza[5], Juan Pedro Hernández del Rincón[6], Sergio Cánovas[5], Juan-Ramon Gimeno-Blanes[5], Jose Luis Rojo-Alvarez[1], Laura Martinez-Mateu[1]

[1]Universidad Rey Juan Carlos, Fuenlabrada, Spain
[2]Universidad Miguel Hernández, Elche, Spain
[3]Centro Nacional de Biotecnología (CSIC), Madrid, Spain
[4]Facultad de Medicina Universidad de Murcia, Murcia, Spain
[5]Hospital Clínico Universitario Virgen de la Arrixaca, Murcia, Spain
[6]Instituto de Medicina Legal y Ciencias Forenses, Murcia, Spain

Abstract

Protein abundance profiles often display similar alterations across cardiac diseases, limiting diagnostic discrimination, while patients with the same diagnosis show heterogeneous and poorly understood proteomic profiles. This study aimed to identify coordinated protein subprofiles driving patient variability in cardiac disease.

Cardiac samples (150 control and arrhythmogenic, ischemic, hypertrophic and dilated cardiomyopathy individuals) were analyzed using TMTpro quantitative proteomics, followed by functional annotation, enrichment analysis, hierarchical clustering, dimension reduction, and biomarker scoring, all integrated into a visualization dashboard.

Three protein clusters were identified: cardiac muscle contraction and sarcomere organization showed the strongest functional enrichment, followed by respiration and metabolic processes, and muscle contraction, cell differentiation and cardiac development ($p \approx 7.3\times10^{-27}$–$9.8\times10^{-8}$, 4.7×10^{-11}–2.4×10^{-4}, and 1.1×10^{-5}–3.7×10^{-3}, resp.). Clustering by disease or subcellular localization did not generate clear groups. Overall, coordinated protein patterns provide deeper insight into molecular variability across cardiac diseases than individual abundances.

1. Introduction

Cardiomyopathies comprise a heterogeneous group of myocardial disorders characterized by ventricular dysfunction, arrhythmias, heart failure, and increased risk of sudden cardiac death. Despite advances in imaging and genetic testing, the molecular mechanisms underlying different cardiomyopathy subtypes remain incompletely understood [1,2]. Proteomic analysis provides a functional view of disease biology by capturing alterations in protein expression that reflect the combined effects of genetic, metabolic, and environmental factors [3]. Mass spectrometry-based proteomics enables the unbiased identification of disease-associated proteins and pathways, offering valuable insights into cardiac pathophysiology [4]. However, the high dimensionality and complexity of proteomic datasets demand robust computational methods to uncover biologically meaningful patterns [5]. Moreover, analytical outputs must be readily interpretable by clinicians to facilitate clinical translation.

In this study, we characterized and compared the proteomic landscapes of control and hypertrophic (HCM), dilated (DCM), arrhythmogenic (ACM), and ischemic (ICM) cardiomyopathy individuals using quantitative proteomic data, aiming to identify coordinated protein signatures and facilitate their exploration through an interactive visualization dashboard.

2. Materials

This study used a quantitative proteomics dataset containing ~2900 proteins generated by the University of Murcia in 2024, after collecting 150 human cardiac tissue samples from controls and different cardiomyopathy phenotypes (HCM, DCM, ACM and ICM) at the Hospital Clínico Universitario Virgen de la Arrixaca (Murcia, Spain). All samples were obtained in accordance with applicable regulations and the Declaration of Helsinki.

Protein quantification was performed using TMTpro 16-plex isobaric labelling coupled with mass spectrometry. Samples were distributed across 10 independent TMT experiments, each including 15 biological samples and a pooled internal standard (IS). The IS, composed of 25

representative samples from the different disease groups, was included in every experiment to enable batch normalization and cross-experiment quantitative comparisons. For downstream analyses, scaled protein abundances normalized using the IS were selected.

3. Methods

The workflow followed in this study is depicted in Figure 1. First, proteins with 30% or more missing values across samples were filtered, and the remaining missing values were imputed with a low abundance value [5]. Then, protein identifiers were cross-referenced with the Human Protein Atlas (HPA) to identify proteins with documented expressions in cardiac tissue. As a result, the dataset was reduced to 158 cardiac-associated proteins, which constituted the input for all subsequent analyses.

Second, a differential expression analysis (DEA) was performed to identify proteins whose abundance significantly differs between different conditions. On the one hand, the Kruskal-Wallis test was applied independently to each of the 158 proteins across the five experimental groups to evaluate the potential differences among all disease groups simultaneously [6]. On the other hand, pairwise comparisons were performed using the Mann-Whitney U test to compare each disease group individually against the control group [7]. The Benjamini-Hochberg false discovery rate (FDR) procedure was applied in both cases to control the number of false positives. Proteins with an adjusted FDR < 0.05 were considered statistically significant [8]. In the second case, the magnitude of differential expression was quantified using the log2 fold-change (log2FC), calculated as $log_2FC_i = \bar{X}_{i,Didease}\bar{X}_{i,Control}$, where $\bar{X}$ represents the mean log2-transformed protein abundance for protein i in each experimental group. Then each protein was assigned to a regulation category according to both statistical significance and effect size. Proteins with log2FC > 0.5 and FDR < 0.05 were classified as upregulated, whereas proteins with log2FC < −0.5 and FDR < 0.05 were classified as downregulated. All remaining proteins were classified as not significant.

Third, proteins were grouped based on similar expression profiles across samples using agglomerative hierarchical clustering. Correlation distance was selected as the similarity metric to capture common expression patterns while reducing the influence of differences in overall abundance [9]. Cluster assignments were retained and used in all subsequent analyses. Functional annotations were retrieved from the UniProt database using accession numbers. Information on subcellular localization, biological processes, and molecular functions was extracted for each protein [10]. When multiple annotations were available, only the first curated term was retained. Proteins lacking annotation were labeled as *Unknown* for the corresponding category. Functional enrichment analysis was performed with GSEApy using the Enrichr platform [11,12]. Protein clusters were analyzed against the Gene Ontology (GO) Biological Process and Kyoto Encyclopedia of Genes and Genomes (KEGG) Human gene set libraries [13,14]. Enriched terms with adjusted p-value < 0.05 were retained, and top GO terms were used for cluster characterization.

Fourth, three complementary dimensionality reduction (DR) techniques were used: principal component analysis (PCA), which preserves global linear variance; t-distributed stochastic neighbor embedding (t-SNE), which captures local neighborhood relationships; and uniform

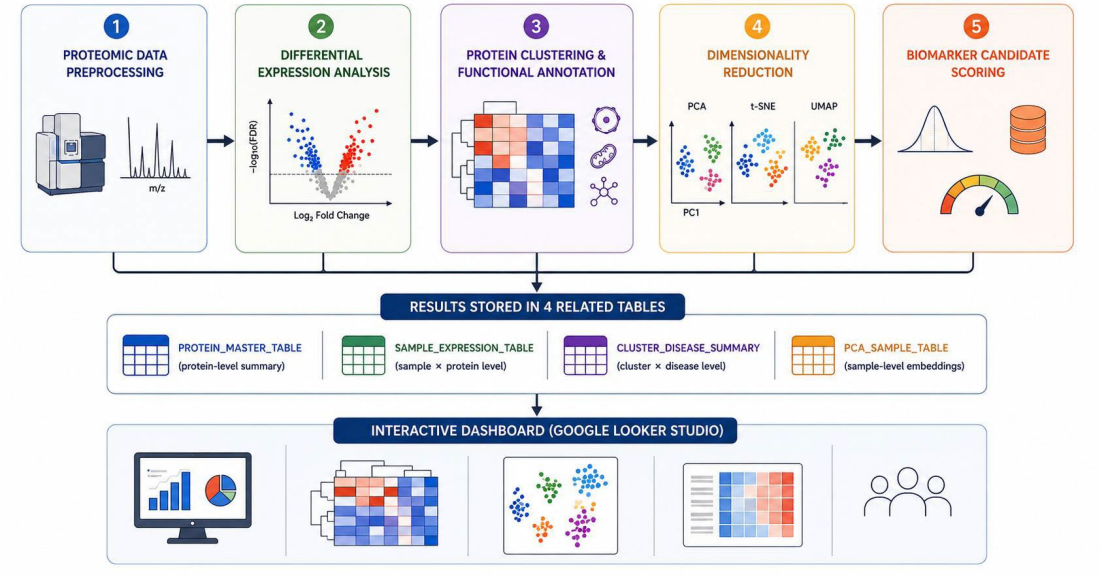

Figure 1. Workflow for cardiac proteomic data analysis and interactive visualization.

manifold approximation and projection (UMAP), which preserves both local and global data structure [15].

Fifth, a composite biomarker score (BS) was developed by integrating statistical evidence from DEA with external disease-association evidence from the Open Targets (OT) Platform [16], assigning equal weight to both components. Proteins lacking disease-association evidence were assigned a score of zero.

All data preprocessing, statistical analyses, machine learning model development, and result visualization were performed in Python using pandas, NumPy, SciPy, scikit-learn, Matplotlib, Plotly, among others.

The final stage of the methodological pipeline involved consolidating all previous outputs into an organized, interoperable data structure (four relational tables exported as CSV files) designed for seamless integration into an interactive web dashboard built in Google Data Studio.

4. Results and discussion

The preprocessing stage yielded a dataset comprised of 158 cardiac-associated proteins. Results of the DEA are summarized in Table 1 and Figure 2. First, the Kruskal-Wallis test followed by the Benjamini-Hochberg FDR correction provided 32 statistically significant proteins (FDRp < 0.05). Table 1 reports 5 top proteins.

Table 1. Five most significant proteins identified by Kruskal-Wallis test after Benjamini-Hochberg correction.

Gene	Accession	K-Wp	FDRp	Group
SMPX	Q9UHP9	0.000036	0.0032	DCM
CHCHD10	Q8WYQ3	0.000041	0.0032	ICM
CTNNA3	Q9UI47	0.000130	0.0051	DCM
CDH2	P19022	0.000115	0.0051	ICM
NCAM1	P13591	0.000206	0.0058	ICM

Second, the Mann–Whitney U test revealed heterogeneous differential expression across diseases. DCM showed the strongest response (27 upregulated and 5 downregulated proteins, Figure 2). The most altered proteins included key structural and cytoskeletal components (FLNC, ACTC1, DES), calcium-handling regulators (CAMK2B, RYR2), and intercalated disc proteins (CDH2, PKP2). HCM displayed a moderate but significant response (12 upregulated and 2 downregulated proteins, Figure 2). The strongest alteration was observed in TRDN, together with proteins involved in extracellular matrix organization (BGN), cytoskeletal architecture (SMPX), and cell adhesion (NCAM1). In ARVD and ICM no proteins met the log2FC and significance thresholds.

The extensive proteomic remodeling observed in DCM aligns with previous studies [3], whereas no proteins in ICM met significance thresholds, likely due to the unbalanced sample distribution across disease groups, which may have reduced the statistical power of the DEA.

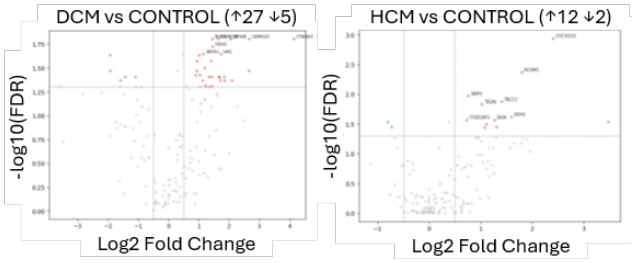

Figure 2. Volcano plot of differential protein expression in DCM/HCM (left/right) compared with control group. Upregulated/downregulated proteins in red/blue.

Hierarchical clustering revealed three protein clusters with distinct expression patterns (Figure 3), closely matching the organization reported in human cardiac proteome studies [3] and indicating that this approach recovered biologically meaningful networks.

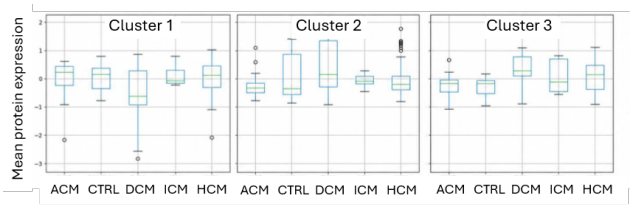

Figure 3. Protein expression by cluster and disease.

Cluster 1 (55 proteins) was strongly enriched for respiration and metabolic processes ($p \approx 4.7\times10^{-11}$ – 2.4×10^{-4}). It included both mitochondria-encoded proteins (e.g., MT-ND1, MT-ND3, MT-ND4, MT-ND5, MT-ND6, MT-CYB, MT-CO2, MT-ATP6) and nuclear-encoded components (e.g., NDUFA3, NDUFA8, COX5A, COX7A1), together with enzymes from the TCA cycle and fatty acid oxidation. Expression was lower in DCM than in other groups. Cluster 2 (33 proteins) showed the strongest functional enrichment dominated by cardiac muscle contraction and sarcomere organization ($p \approx 7.3\times10^{-27}$ – 9.8×10^{-8}). It contained core sarcomeric proteins (myosins, troponins, titin, α-cardiac actin) and Z-disc/cytoskeletal linkers. DCM displayed the highest median expression, while other groups remained closer to controls. Finally, cluster 3 (70 proteins) was the most diverse, enriched for regulation of muscle contraction, cell differentiation and cardiac development ($p \approx 1.1\times10^{-5} – 3.7\times10^{-3}$). It included prior differentially expressed proteins (CAMK2B/D, RYR2, TRDN), heat-shock proteins (CRYAB, HSPB2/3/7), and cytoskeletal remodelers (FLNC, SORBS2, CTNNA3). DCM and HCM showed higher expression. These results match those from the DEA.

The DR embeddings are shown in Figure 4. PCA separated the three protein clusters clearly, t-SNE produced a similar structure with sharper local separation, and UMAP yielded the most distinct boundaries, which

agrees with the literature [15]. The consistent separation reflects a robust biological structure rather than a methodological artifact. Instead, clustering by disease or subcellular localization did not yield well-defined groups.

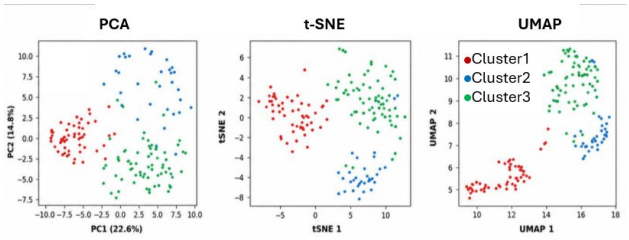

Figure 4. Visualization of the cardiac proteomic dataset using PCA, t-SNE, and UMAP.

Among the 32 significant proteins, 22 (68.8%) had documented cardiac disease associations in the OT, while 10 proteins lacked prior cardiac annotation. Consequently, proteins supported by both statistical significance and external evidence received higher BS: CHCHD10 (0.918) and CDH2 (0.914) were top candidates, both highly expressed in ICM. For DCM, the highest-ranked proteins were CRYAB (0.878), BVES (0.813), CTNNA3 (0.809), FLNC (0.766), and ACTC1 (0.746). In HCM, TRDN (0.756) ranked highest, and SMPX stood out despite lacking external disease associations.

All analytical outputs were integrated into a three-page interactive dashboard in Google Data Studio, linked to four data tables via Google Sheets. The dashboard was designed for clinician-friendly interpretation, providing immediate access to relevant information and contributing to better patient outcomes [17]. QR code linking to the interactive dashboard is shown in Figure 5.

Figure 5. QR code linking to the interactive dashboard.

5. Conclusions

Our results suggest that coordinated protein patterns offer deeper insight into molecular variability across cardiac diseases than individual abundances. Moreover, the BS prioritized proteins supported by both differential expression and external evidence, while still noting statistically significant proteins lacking prior annotation.

Acknowledgments

This work was partially supported by grant PID2022-140553OA-C42 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU.

References

[1] H. El Hadi et al., "Hypertrophic, Dilated, and Arrhythmogenic Cardiomyopathy: Where Are We?, " *Biomedicines*, vol. 11, no. 2, Feb. 2023, Art. no. 524.

[2] M. Bourfiss et al., "Prevalence and Disease Expression of Pathogenic and Likely Pathogenic Variants Associated With Inherited Cardiomyopathies in the General Population" *Circ Genom Precis Med*., vol. 15, no. 6, pp. 530-541, Dec. 2022.

[3] O. A. Karpov et al., "Proteomics of the heart," *Physiol Rev.*, vol. 104, no. 3, pp. 931-982, Feb. 2024.

[4] L. Anderson, "Candidate-based proteomics in the search for biomarkers of cardiovascular disease," *J Physiol*., vol. 563, no. 1, pp. 23-60, Feb. 2005.

[5] B. M. Webb-Robertson et al., "Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics," *J. Proteome Res*., vol. 14, no. 5, pp. 1993–2001, May 2015.

[6] W. H. Kruskal, and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, Dec. 1952.

[7] H. B. Mann, and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[8] Y. Benjamini, and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

[9] B. Meunier et al., "Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining," *J. Proteome Res.*, vol. 6, no. 1, pp. 358–366, Jan. 2007.

[10] S. Ahmad et al., "The UniProt website API: facilitating programmatic access to protein knowledge," *Nucleic Acids Res.*, vol. 53, no. W1, pp. W547–W553, Jul. 2025.

[11] Z. Fang et al., "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, vol. 39, no. 1, Jan. 2023, Art. No. 757.

[12] E. Y. Chen et al., "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, Dec. 2013, Art. No. 128.

[13] M. Ashburner et al., "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.

[14] H. Ogata et al., "Computation with the KEGG pathway database," *Biosystems*, vol. 47, pp. 119–128, 1998.

[15] E. Becht et al., "Dimensionality reduction for visualizing single-cell data using UMAP," *Nat Biotechnol*., vol. 37, no. 1, pp. 38–44, Jan. 2019.

[16] D. Ochoa et al., "The next-generation Open Targets Platform: reimagined, redesigned, rebuilt," *Nucleic Acids Res.*, vol. 51, Database issue, pp. D1353–D1359, Jan. 2023.

[17] D. Dowding et al., "Dashboards for improving patient care: Review of the literature," *Int J Med Inform.*, vol. 84, no. 2, pp. 87–100, Feb. 2015.

Address for correspondence:

Laura Martinez-Mateu.
Universidad Rey Juan Carlos, Departamental-III, D205, Camino del Molino 5, 28942 Fuenlabrada, Madrid, España.
laura.martinez.mateu@urjc.es