Published in final edited form as:
Circ Res. 2013 October 12; 113(9): 1043–1053. doi:10.1161/CIRCRESAHA.113.301151.

Integration of Cardiac Proteome Biology and Medicine by a Specialized Knowledgebase

Nobel C. Zong[1,2], **Haomin Li**[1,2,3], **Hua Li**[2,4], **Maggie P.Y. Lam**[1,2], **Rafael C. Jimenez**[1,5], **Christina S. Kim**[1,2], **Ning Deng**[1,3], **Allen K. Kim**[1,2], **Jeong Ho Choi**[1,2], **Ivette Zelaya**[1,2], **David Liem**[1,2], **David Meyer**[2], **Jacob Odeberg**[1,6], **Caiyun Fang**[7], **Hao-jie Lu**[7], **Tao Xu**[1,8], **James Weiss**[1,2], **Huilong Duan**[1,3], **Mathias Uhlen**[1,6], **John R. Yates III**[1,8], **Rolf Apweiler**[1,5], **Junbo Ge**[4], **Henning Hermjakob**[1,5], and **Peipei Ping**[1,2,3,4]

[1]NHLBI Proteomics Center at UCLA/NHLBI Proteomics Program, UCLA School of Medicine, Los Angeles, California 90095, USA

[2]Departments of Physiology and Medicine/CVRL, UCLA School of Medicine, Los Angeles, California 90095, USA

[3]College of Biomedical Engineering and Instrumentation, Zhejiang University, the Key Laboratory for Biomedical Engineering of Ministry of Education, Hangzhou 310027, China

[4]Shanghai Institute of Cardiovascular Diseases, Zhongshan Hospital, Fudan University, Shanghai, 200032, China

[5]EMBL Outstation European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[6]Science for Life Laboratory, the Royal Institute of Technology (KTH), Stockholm, SE-171 65, Sweden

[7]Department of Chemistry, Fudan University, Shanghai 200433, China

[8]Department of Chemical Physiology, the Scripps Research Institute (TSRI), La Jolla, California 92037, USA

Abstract

Rationale—Omics sciences enable a systems-level perspective in characterizing cardiovascular biology. Integration of diverse proteomics data via a computational strategy will catalyze the

Address correspondence to: Dr. Nobel C. Zong, UCLA School of Medicine, 675 Charles E. Young Drive South, MRL Building Suite 1-609, Los Angeles, CA 90095, Tel: 310-267-5624, Fax: 310-267-5623, czong@mednet.ucla.edu, Dr. PeiPei Ping, UCLA School of Medicine, 675 Charles E. Young Drive South, MRL Building Suite 1-609, Los Angeles, CA 90095, Tel: 310-267-5624, Fax: 310-267-5623, pping@mednet.ucla.edu.
N.C.Z., Haomin L., and Hua L. contributed equally to this study.

DISCLOSURES
None.

assembly of contextualized knowledge, foster discoveries through multidisciplinary investigations, and minimize unnecessary redundancy in research efforts.

Objective—The goal of this project is to develop a consolidated cardiac proteome knowledgebase with novel bioinformatics pipeline and web portals, thereby serving as a new resource to advance cardiovascular biology and medicine.

Methods and Results—We created Cardiac Organellar Protein Atlas Knowledgebase (COPaKB), a centralized platform of high quality cardiac proteomic data, bioinformatics tools and relevant cardiovascular phenotypes. Currently, COPaKB features eight organellar modules, comprising 4,203 LC-MS/MS experiments from human, mouse, drosophila and *C. elegans* as well as expression images of 10,924 proteins in human myocardium. In addition, the Java-coded bioinformatics tools provided by COPaKB enable cardiovascular investigators in all disciplines to retrieve and analyze pertinent organellar protein properties of interest.

Conclusions—COPaKB (www.HeartProteome.org) provides an innovative and interactive resource, which connects research interests with the new biological discoveries in protein sciences. With an array of intuitive tools in this unified web server, non-proteomics investigators can conveniently collaborate with proteomics specialists to dissect the molecular signatures of cardiovascular phenotypes.

Keywords

Cardiovascular Proteomics; COPaKB; Spectral Library; Omics Science; knowledge translation; bioinformatics; organelle; proteomics; mitochondria

INTRODUCTION

Recent studies on cardiovascular biology have been transformed by growing applications of Omics technologies[1–5]. For example, large-scale proteomic investigations have discovered the protein anatomy and dynamics of individual cardiac organelles as well as new principles of cardiac regulations in forms of widespread post-translational modifications. Numerous other large-scale datasets are now being generated that enable investigators to ask more complex biological questions.

Efficient bioinformatics resources are vital for connecting these increasingly voluminous and diversified data to experts of various disciplines to formulate new biological insights[6, 7]. However, at this moment, the data are often distributed in forms that are not readily accessible, and as a result, they are often of limited value to researchers outside a particular area of expertise. Furthermore, efficient utilization of these datasets has been impeded by inconsistent annotation guidelines. Addressing these challenges therefore requires new computational tools and bioinformatics infrastructures to integrate, analyze and visualize multi-disciplinary datasets[8].

Here we present the Cardiac Organellar Protein Atlas Knowledgebase (COPaKB), a specialized resource for the cardiovascular community with three distinct components. Firstly, it comprises comprehensive spectral libraries of individual cardiac organelles and a search engine for investigators to quickly identify proteins from supplied datasets with high

coverage. Secondly, it contains a curated database and a set of bioinformatics tools to integrate the identified proteins with relevant biomedical attributes (e.g. genetic mutations, disease phenotypes) and orthogonal biomolecular properties (e.g. protein expression imaging, gene transcription activity) (Fig. 1). Lastly, COPaKB provides a unified web portal with a robust web service infrastructure to allow proteomic data to be efficiently analyzed, distributed, and queried. The Wiki component of the web portal, in particular, facilitates interactions and collaborations among investigators, supporting a knowledge-building process in cardiovascular biology and medicine.

A primary benefit of this unified knowledgebase is that it contextualizes and distributes cardiac proteome data within a consistent set of standards. The datasets from multiple studies of different investigators can be combined through COPaKB as the shared reference for effective comparative analyses. Moreover, COPaKB Client is created to enable high-speed analysis of large datasets on a proteome scale. Overall, COPaKB encapsulates carefully curated data, new informatics schema and an effective web portal in a complete package. We anticipate this platform will broaden the utility of proteomic data for the entire cardiovascular community and help bridge discovery-driven and hypothesis-driven studies.

METHODS

Construction of a modular knowledgebase for cardiovascular proteome biology

COPaKB contains the following components: a relational database supporting multiple modules of proteome knowledge, a Wiki interface promoting user input, and a computational toolbox facilitating data analyses (Fig. 1). All components are regularly updated and maintained.

Structure and organization of COPa Knowledgebase

COPa knowledgebase contains a repertoire of protein properties based on their subcellular compartments. The underlying relational database is configured using spectral libraries as a backbone structure. The selection of representative spectra has been based on the cross-correlation score (Xcorr) assigned by ProLuCID[8]. An Oracle database has been used to manage orthogonal proteomic datasets in the COPaKB. Known associations between protein function and cardiac diseases were retrieved from OMIM web service[9] and from peer-reviewed publications via PubMed using keyword combinations of protein name, gene symbol and heart diseases.

Protein expression profiles probed with specific antibodies were integrated from the Human Protein Atlas[10]. The differential expression of gene transcript was integrated from Gene Expression Atlas[11]. Gene Ontology (GO) annotations were obtained using UniProt API[12] and QuickGO services[13]. The relationships among different GO terms were delineated using Ontology Lookup Service[14] by EMBL-EBI. The schema of this relational database is readily expandable to accommodate additional forms of knowledge (Fig. 1). Each component of the COPaKB are constantly updated and maintained. The release history is outlined in the COPaKB website below (Please visit http://www.heartproteome.org/copa/ReleaseHistory.aspx).

COPaKB computational toolbox

We have implemented the COPaKB web server on a DELL Precision T7500 workstation. Details on server configuration are documented in the *Online Supplement*.

We developed the COPaKB Client software to enable web-based data transfer via the Simple Object Access Protocol (SOAP). Details on the coding and application of this program are documented in the Online Supplement.

The analysis of proteomic data files from COPaKB users has been supported by a spectral library search engine that we previously developed[15].

We processed mass spectral data files to create spectral libraries for COPaKB. Thus far, ten modules have been configured upon multiple replicate analyses (biological and technical), eight of which are organellar modules. They include human heart mitochondria (29 replicates), human heart proteasomes (20 replicates), murine heart mitochondria (34 replicates), murine heart proteasomes (22 replicates), murine heart nucleus (30 replicates), murine heart cytosol (9 replicates), drosophila mitochondria (18 replicates), and *C. elegans* mitochondria (9 replicates). Two modules are total tissue lysates, human heart lysate (20 replicates) and mouse heart lysate (1 replicate), Details on data source and data processing are documented in the *Online Supplement*.

We created a publicly accessible website to interface COPaKB with the scientific community. The specific procedures of implementing both the website and its Wiki web portal (software development for each component) are described in the *Online Supplement*.

Data source and tissue collection

As of May 31st 2013, COPaKB hosts ten modules. The modules of murine heart mitochondria, murine heart proteasome, murine heart cytosol, drosophila mitochondria, human heart mitochondria and human heart proteasome were created using the data collected at UCLA; the modules of murine heart nuclei, murine heart total lysates, *C. elegans* mitochondria, and human heart total lysates were curated from public resources.

Regarding data created at UCLA, all procedures involving mice (ICR strain) were performed in accordance with the Animal Research Committee (ARC) guidelines at UCLA and the Guide for the Care and Use of Laboratory Animals, published by the National Institutes of Health. Drosophila mitochondria were extracted from the Oregon-R-C strain. The Experimental procedures involving human samples were approved by the UCLA Human Subjects Protection Committee (HSPC) and the UCLA Institutional Review Boards (IRBs). The phenotypes of all samples are documented online at COPaKB website. Additional information can be found in the Online Supplement.

Demonstration of the COPaKB-assisted proteomics workflow

A test dataset was downloaded from the Peptide Atlas Repository[16](Project ID: PAe000353)[17], containing a total of 111 raw proteomic data files. In this dataset, murine heart mitochondrial proteins were extracted from the female ICR mouse strain and analyzed

on a LCQ Deca XP mass spectrometer. This dataset was processed by the COPaKB-directed workflow to benchmark its performance against that of the SEQUEST-assisted workflow.

The analytical efficiency and robustness of the COPaKB Client-directed workflow were evaluated using spectral files in mzML format by scientists at six different test centers globally. Each center conducted three replicate tests and reported the fastest rate.

The utility of COPaKB in integrating discoveries from multiple analyses was examined using three sets of LTQ-Orbitrap-collected data on murine heart mitochondria, with the mass resolutions of MS1 scan set at 60,000, 15,000, and 7,500, respectively. Each set contains 21 LC-MS/MS experiments.

RESULTS

To create COPaKB (www.HeartProteome.org), we compiled a large collection of annotated protein mass spectral datasets on human, mouse, drosophila and *C. elegans* samples. We integrated these data in a modular structure that parallel the organization of subcellular organelles inside the cardiac cell. The following examples demonstrate the utility of COPaKB for efficiently conducting comparative analyses of multiple datasets.

Assembly of cardiac spectral datasets

Mass spectral datasets of proteins were organized into modular fashions based on their subcellular locations. Each module included multiple replicate analyses, which encapsulate the dynamic range of protein expression. For the human mitochondria module, a total of 6 biological replicates were integrated (Fig. 2A); for the human proteasome module, a total of 5 biological replicates were integrated (Fig. 2B). Altogether, the human mitochondria mass spectral library module was built upon 856 LC-MS/MS experiments, and the human proteasome module incorporated 160 LC-MS/MS experiments.

A total of 41,758 non-redundant mass spectra representing 1,398 proteins and 28,031 peptides were compiled for the human mitochondria module. A total of 5,668 mass spectra representing 283 proteins and 3,482 peptides were assembled for the human proteasome module; these data include proteasome subunits and their associated proteins. A total of 59,020 mass spectra representing 1,619 proteins and 38,421 peptides were organized into the murine mitochondria module. A total of 9,442 mass spectra representing 151 proteins and 6,409 peptides were collected for the murine proteasome module (Table 1).

The data coverage of knowledgebase modules was evaluated by analyzing the cumulative number of protein entries as more replicates were incorporated. For the four example modules (human heart mitochondria, human heart proteasome, murine heart mitochondria, and murine heart proteasome) presented in Figure 2, the coverage of these modules reached plateaus when sufficient number of replicates were included.

Data integration in COPa Knowledgebase

COPaKB organizes complex cardiac proteome knowledgebase using a relational database (Fig. 1). The core structure of this database is built upon an integrated framework using

mass spectrometry datasets, which capture multi-dimensional molecular features ranging from peptides to proteins. This unique hierarchy structure of the database renders seamless incorporation of diverse properties.

Protein expression image datasets were synchronized with those curated by the Human Protein Atlas (http://www.proteinatlas.org)[10] and compiled into a HPA-reference table of COPaKB. Immunofluorescence and immunohistochemistry images of a total of 10,924 human proteins were included (Table 1). Immunofluorescence images enlist protein expression profiles with subcellular resolution; immunohistochemistry images assist visualization of the protein expression profiles in different cardiac cell types.

Changes in protein properties associated with cardiovascular pathogenesis were documented by performing a systematic literature search on peer-reviewed sources; a total of 413 non-redundant perturbations were found. In parallel, biomedical data from OMIM[18] (Online Mendelian Inheritance in Man) and Gene Expression Atlas[19] were incorporated into COPaKB to present the relevance of individual proteins to heart diseases.

Application of COPa Knowledgebase to support new biology in cardiovascular studies

The main utilities and functional outputs of COPaKB are summarized in Table 2. Queries can take the formats of (i) a protein identifier of the interest as “Protein Identifier”; (ii) a particular amino acid sequence of the interest as “Amino Acid Sequence”; (iii) any mass spectrometry data files from users as “MS Data File(s)”; or (iv) identifiers of existing analyses by any investigator team(s) as “Analysis of Multiple Datasets”. Specifically, a protein identifier can be a name of a protein; the glycogen synthase kinase-3 alpha is shown in Table 2 as an example. Moreover, the protein identifier can also be its gene symbol, for example “GSK3α”; or the protein identifier can be its UniProt ID, for example, “P49840”. When a query is made, COPaKB will generate reports regarding the relevance of this protein in cardiac phenotypes, information in the literature on its mRNA expression, its interacting protein partners, its immunohistochemistry images in the myocardium, and its immunocytofluorescence images in human cells. In addition, COPaKB will also report a list of peptides identified with corresponding mass spectra data about this protein; all peer-reviewed publications on this protein as documented by iHOP. Furthermore, COPaKB welcomes user-inputs to further annotate this protein in the format of a Wiki page. An example output of this query by COPaKB is presented in Table 2 as “www.heartproteome.org/copa/proteinInfo.aspx?qType=protein%20ID&qValue= P49840”. In similar fashions, Table 2 details examples highlighting query formats made as “Amino Acid Sequence”, “MS Data File(s)”, or “Analysis of Multiple Datasets”.

Furthermore, to demonstrate the utility of COPaKB-facilitated analyses, we acquired a test dataset of murine mitochondria proteins from the Peptide Atlas Repository (PAe000353)[17]. This test dataset was reprocessed using the murine mitochondria module of COPaKB as a reference. In this analysis, we identified a total of 261 proteins with a statistical confidence of 95%. In contrast, when this test dataset was analyzed using the SEQUEST, a commonly used mass-spectra search engine, we identified 183 proteins at the same confidence level (Table SII). Specifically, 78.7% of proteins were commonly identified by the two approaches (Fig. 4A), whereas COPaKB identified an additional 64.5% proteins (i.e., 117

proteins) that were not covered by the SEQUEST search (Fig. 4B). This added protein coverage significantly expanded the search outcome; the additional 117 proteins include mitochondrial proteins (e.g., cytochrome c oxidase 7A1 of the electron transport chain complex IV).

Moreover, an automatic query to the knowledgebase for protein identification was integrated with functional annotations. Among the 117 proteins uniquely identified by COPaKB, 92 proteins (79%) had a Gene Ontology annotation of the mitochondrion as their primary subcellular location (Fig. 4A). 74 out of the 92 mitochondrial proteins (80%) were involved in metabolism, 3 were involved in transport, and 4 were involved in apoptosis. Furthermore, 199 protein expression images (among the 261 identified proteins) were available in HPA and were automatically retrieved (Fig. 4C). According to peer-reviewed publications, 49 of the 261 proteins were involved in the processes of cardiac pathogenesis.

Performance efficiency of COPaKB workflow

A reliable delivery of Omics-scale information requires robust web portals. The traditional web-based data transfer protocol is limited by its ability to effectively transfer data within a defined time frame (e.g. 60 minutes). To overcome this challenge, a COPaKB Client program was engineered to implement a SOAP-assisted workflow (Fig. 3). Its performance efficiency was benchmarked against that of the traditional HTTP-assisted workflow by participating investigators from six test centers (Table 3 & Table SI). In all these tests, large-scale data files were reliably and consistently delivered by the COPaKB Client. In a load test, the COPaKB server was able to process 50 simultaneous search requests without compromising its reliability.

The analytical efficiency of the COPaKB-supported search engine was also benchmarked against that of the SEQUEST. Analysis using COPaKB requires an extra time to upload the data onto the knowledgebase server. Despite the required extra time, the COPaKB-supported workflow completed the analyses faster than that performed by the SEQUEST-supported workflow. This test was repeated on a PC and on a computing cluster of moderate size (7 nodes) (Table 4). This enhanced efficiency of COPaKB workflow is accomplished by a condensed search space within the mass spectral library, a simplified spectral matching algorithm, a strategy of selecting only MS2 spectra for web-based protein identification, and the integration of the SOAP protocol.

Platform to promote new discoveries through collaborative effort

Discoveries from multiple studies were integrated in real time via the COPaKB server (Table 2). COPaKB receives the identifier of each analysis task as input and returns a list of proteins that are identified in each of these analyses. Three test datasets of murine mitochondrial proteins were collected using an LTQ-Orbitrap mass spectrometer with different settings of MS1 resolution; they were independently analyzed via COPaKB (Figure 5A). These results were then combined using the murine mitochondria module of COPaKB as a reference. Comparative analysis on these datasets was subsequently conducted (Fig. 5B).

COPaKB relies on inputs from the scientific community. Integration of mass spectrometry-based datasets from various sources necessitates standard operation protocols. COPaKB achieves consistency following the principles established by the Minimum Information about a Proteomics Experiment (MIAPE)[20]. This guideline describes experimental procedures with sufficient details using pre-defined and controlled vocabularies. The complexity of the controlled vocabulary, however, often serves as a double-edged sword; the redundant terminology allows the same experiments to be described in different terms. Using a specialized subset of the data submission utility[21], COPaKB conducts standardized collection and propagation of proteomics data (Fig. SIII).

Along with the effort to streamline the knowledgebase with consistent vocabularies, a text-based Wiki component has been created to facilitate communication among investigators. The Wiki component is coupled to each peptide, protein, and spectrum entry of the knowledgebase, welcoming users to add or edit the content relevant to the subject. The open nature of the Wiki component fosters a worldwide collaborative effort (Fig. SIV).

DISCUSSION

Our goal is to create a protein knowledgebase to support the long-term advancement of cardiovascular biology in an informatics-driven era. As the continued growth of any bioinformatics platform hinges upon community support, we designed COPaKB from the grounds up with community participation in mind.

Systems integration and software engineering for COPaKB

COPaKB assembles its individual modules based on protein localizations. Each module consists of orthogonal datasets curated from either a public resource or a large cohort of experiments. The public resources also include UniProt[12], Human Protein Atlas[22], OMIM[9] and Expression Atlas[19]. The annotated datasets are integrated using a relational database schema, which offers the investigators easy access to an array of protein properties.

COPaKB operates via an innovative workflow. First, it has the design features to support multidimensional data integration. Specifically, utilizing mass spectrometry-based proteomics data has the benefit that information from different disciplines can be synthesized. Second, it carries a new mechanism of conducting data query. Many online proteomics resources exist, including repertoires of 2D-PAGE images[23], mass spectra[10], and protein expression images[16, 21]. Retrieving these data from multiple databases remains laborious and requires repetitive mining efforts. COPaKB overcomes these challenges with two unique functional features. One is to connect specific interests of individual investigators with vast information hosted in the knowledgebase; another is to support large file transfer in an efficient manner. Both features are supported by computational strategies created for COPaKB. These strategies include a spectral library and a search engine that automatically decodes raw spectral data[24–26]; subsequently, combining them with orthogonal properties of proteins and genes. This computational toolbox provides a new mechanism of utilizing biomedical knowledgebase where functional annotations of each protein in the sample are presented in a cohesive context. Accordingly, this built-in

bioinformatics pipeline in COPaKB supports investigators to complement their targeted investigation with discovery-based or hypothesis-driven approaches.

Finally, COPaKB Client addresses the technical limitations of transferring large data files on a proteome scale. We have engineered this program to utilize a web-based SOAP portal to enable robust connections with the COPaKB server, allowing reliable access to contextualized properties on proteins of interest.

Building COPaKB via a community-driven paradigm

Comprehensive understanding of cardiovascular biology on a proteome scale is a long-term goal that often exceeds the capacity of individual investigators. COPaKB alleviates this limitation by implementing the bioinformatics infrastructure necessary to facilitate effective collaboration. In particular, COPaKB allows real-time integration of analyses on multiple datasets from numerous investigators in parallel. This strategy surmounts technical challenges in long-distance collaborations; namely, geographic boundaries and platform discrepancies. Additionally, the Wiki component in COPaKB supports investigators to communicate and contribute information on individual proteins in a variety of formats (e.g., images or text).

Investigator participation is essential to the future growth of a community-driven knowledgebase. Currently the capacity of COPaKB modules is primarily contributed by datasets publicly available. However, the current coverage of COPaKB on several modules is at their infant stage. Particularly, this situation applies to the module of mouse heart total lysate. As the ProteomeXchange consortium[27] led by EBI has accelerated the supply of raw proteomic data, COPaKB is expected to grow and will cover additional modules on organelles and cells from cardiovascular relevant model systems. The content of each module will expand with the increasingly available public data.

Despite the rapid development in proteomics science, the gap in translating the new tools to biological applications has widened. This is largely due to limited access to high-end instrumentation and difficulties to manage large data analyses. The user-friendly workflow of the COPaKB helps non-proteomics investigators to directly benefit from the technological advancement and new datasets. Investigators are no longer restricted amid lacking a direct access to high-end mass spectrometry. For example, COPaKB documents the parameters (e.g. molecular mass and charge state) that are associated with peptides; the institutional proteomic core can then adjust instrument settings accordingly to improve the detection of their selected proteins. In this scenario, the cost and time in using the proteomic technologies are optimized. Taken together, COPaKB provides a number of effective workflows to guide cardiovascular investigators from proteomics data to systematic interpretation of biomedical properties. We are continuing to develop innovative workflows to aid better understanding of protein functions in cardiovascular diseases.

In conclusion, COPaKB is a novel computational platform with its unique bioinformatics pipelines and web portals, engaging a community effort to build a knowledgebase. As proteome biology has become increasingly integrated into cardiovascular medicine, we envision a growing importance of this new resource.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

SOURCES OF FUNDING

This work was supported, in part, by NHLBI Proteomics Center Award HHSN268201000035C, NIH R01 HL063901, an endowment from Theodore C. Laubisch to Dr. Peipei Ping, R37HL063901 diversity supplement award to Mr. David Meyer and R37HHSN268201000035C diversity supplement award to Ms. Ivette Zelaya.

Nonstandard Abbreviations and Acronyms

API	Application Programming Interface
COPaKB	Cardiac Organellar Protein Atlas Knowledgebase
HPA	Human Protein Atlas
HTTP	Hyper Text Transfer Protocol
iHOP	Information Hyperlinked over Proteins
MIAPE	Minimum Information about a Proteomics Experiment
PRIDE	Proteomics Identifications Database
SOAP	Simple Object Access Protocol
UniProt	Universal Protein Resource

REFERENCES

1. Pazdrak K, Young TW, Straub C, Stafford S, Kurosky A. Priming of eosinophils by gm-csf is mediated by protein kinase cbetaii-phosphorylated l-plastin. J Immunol. 2011; 186:6485–6496. [PubMed: 21525390]
2. Wang SB, Foster DB, Rucker J, O'Rourke B, Kass DA, Van Eyk JE. Redox regulation of mitochondrial atp synthase: Implications for cardiac resynchronization therapy. Circ Res. 2011; 109:750–757. [PubMed: 21817160]
3. Zhang J, Guy MJ, Norman HS, Chen YC, Xu Q, Dong X, Guner H, Wang S, Kohmoto T, Young KH, Moss RL, Ge Y. Top-down quantitative proteomics identified phosphorylation of cardiac troponin i as a candidate biomarker for chronic heart failure. J Proteome Res. 2011; 10:4054–4065. [PubMed: 21751783]
4. Theberge R, Infusini G, Tong W, McComb ME, Costello CE. Top-down analysis of small plasma proteins using an ltq-orbitrap. Potential for mass spectrometry-based clinical assays for transthyretin and hemoglobin. Int J Mass Spectrom. 2011; 300:130–142. [PubMed: 21607198]
5. Lam MP, Lau E, Scruggs SB, Wang D, Kim TY, Liem DA, Zhang J, Ryan CM, Faull KF, Ping P. Site-specific quantitative analysis of cardiac mitochondrial protein phosphorylation. J Proteomics. 2013; 81:15–23. [PubMed: 23022582]
6. Mazumder R, Natale DA, Julio JA, Yeh LS, Wu CH. Community annotation in biology. Biology direct. 2010; 5:12. [PubMed: 20167071]
7. Chen C, McGarvey PB, Huang H, Wu CH. Protein bioinformatics infrastructure for the integration and analysis of multiple high-throughput “omics” data. Advances in bioinformatics. 2010:423589. [PubMed: 20369061]
8. Xu T, Wong CC, Kashina A, Yates JR 3rd. Identification of n-terminally arginylated proteins and peptides by mass spectrometry. Nat Protoc. 2009; 4:325–332. [PubMed: 19229197]

9. McKusick VA. Mendelian inheritance in man and its online version, omim. American journal of human genetics. 2007; 80:588–604. [PubMed: 17357067]
10. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L, Ponten F. Towards a knowledge-based human protein atlas. Nat Biotechnol. 2010; 28:1248–1250. [PubMed: 21139605]
11. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N, Kurnosov P, Malone J, Melnichuk O, Petryszak R, Pultsin N, Rustici G, Tikhonov A, Travillian RS, Williams E, Zorin A, Parkinson H, Brazma A. Gene expression atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res. 2012; 40:D1077–D1081. [PubMed: 22064864]
12. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. The universal protein resource (uniprot): An expanding universe of protein information. Nucleic Acids Res. 2006; 34:D187–D191. [PubMed: 16381842]
13. 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. [PubMed: 19744993]
14. Cote R, Reisinger F, Martens L, Barsnes H, Vizcaino JA, Hermjakob H. The ontology lookup service: Bigger and better. Nucleic Acids Res. 2010; 38:W155–W160. [PubMed: 20460452]
15. Li H, Zong NC, Liang X, Kim A, Choi JH, Deng N, Zelaya I, Lam M, Duan H, Ping P. A novel spectral library workflow to enhance protein identifications. J Proteomics. 2013
16. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The peptideatlas project. Nucleic Acids Res. 2006; 34:D655–D658. [PubMed: 16381952]
17. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A. Global survey of organ and organelle protein expression in mouse: Combined proteomic and transcriptomic profiling. Cell. 2006; 125:173–186. [PubMed: 16615898]
18. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005; 33:D514–D517. [PubMed: 15608251]
19. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A. Gene expression atlas at the european bioinformatics institute. Nucleic Acids Res. 2010; 38:D690–D698. [PubMed: 19906730]
20. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr, Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW, Dunn MJ, Heck AJ, Leitner A, Macht M, Mann M, Martens L, Neubert TA, Patterson SD, Ping P, Seymour SL, Souda P, Tsugita A, Vandekerckhove J, Vondriska TM, Whitelegge JP, Wilkins MR, Xenarios I, Yates JR 3rd, Hermjakob H. The minimum information about a proteomics experiment (miape). Nat Biotechnol. 2007; 25:887–893. [PubMed: 17687369]
21. Jones P, Cote RG, Cho SY, Klie S, Martens L, Quinn AF, Thorneycroft D, Hermjakob H. Pride: New developments and new datasets. Nucleic Acids Res. 2008; 36:D878–D883. [PubMed: 18033805]
22. Ponten F, Jirstrom K, Uhlen M. The human protein atlas--a tool for pathology. J Pathol. 2008; 216:387–393. [PubMed: 18853439]
23. Hoogland C, Mostaguir K, Appel RD, Lisacek F. The world-2dpage constellation to promote and publish gel-based proteomics data through the expasy server. J Proteomics. 2008; 71:245–248. [PubMed: 18617148]
24. Yates JR 3rd, Morgan SF, Gatlin CL, Griffin PR, Eng JK. Method to compare collision-induced dissociation spectra of peptides: Potential for library searching and subtractive analysis. Anal Chem. 1998; 70:3557–3565. [PubMed: 9737207]
25. Lam H, Deutsch EW, Eddes JS, Eng JK, King N, Stein SE, Aebersold R. Development and validation of a spectral library searching method for peptide identification from ms/ms. Proteomics. 2007; 7:655–667. [PubMed: 17295354]

26. Li H, Zong N, Liang X, Kim A, Choi J, Deng N, Zelaya I, Lam M, Duan H, Ping P. A novel spectral library workflow to enhance protein identifications. J Proteomics. 2013 JProt-D-12-00514 (in Press).

27. Orchard S, Albar JP, Deutsch EW, Eisenacher M, Binz PA, Martinez-Bartolome S, Vizcaino JA, Hermjakob H. From proteomics data representation to public data flow: A report on the hupo-psi workshop september 2011, geneva, switzerland. Proteomics. 2012; 12:351–355. [PubMed: 22290802]

28. Zhai P, Sadoshima J. Overcoming an energy crisis?: An adaptive role of glycogen synthase kinase-3 inhibition in ischemia/reperfusion. Circ Res. 2008; 103:910–913. [PubMed: 18948628]

29. Shiojima I, Walsh K. Role of akt signaling in vascular homeostasis and angiogenesis. Circ Res. 2002; 90:1243–1250. [PubMed: 12089061]

Novelty and Significance

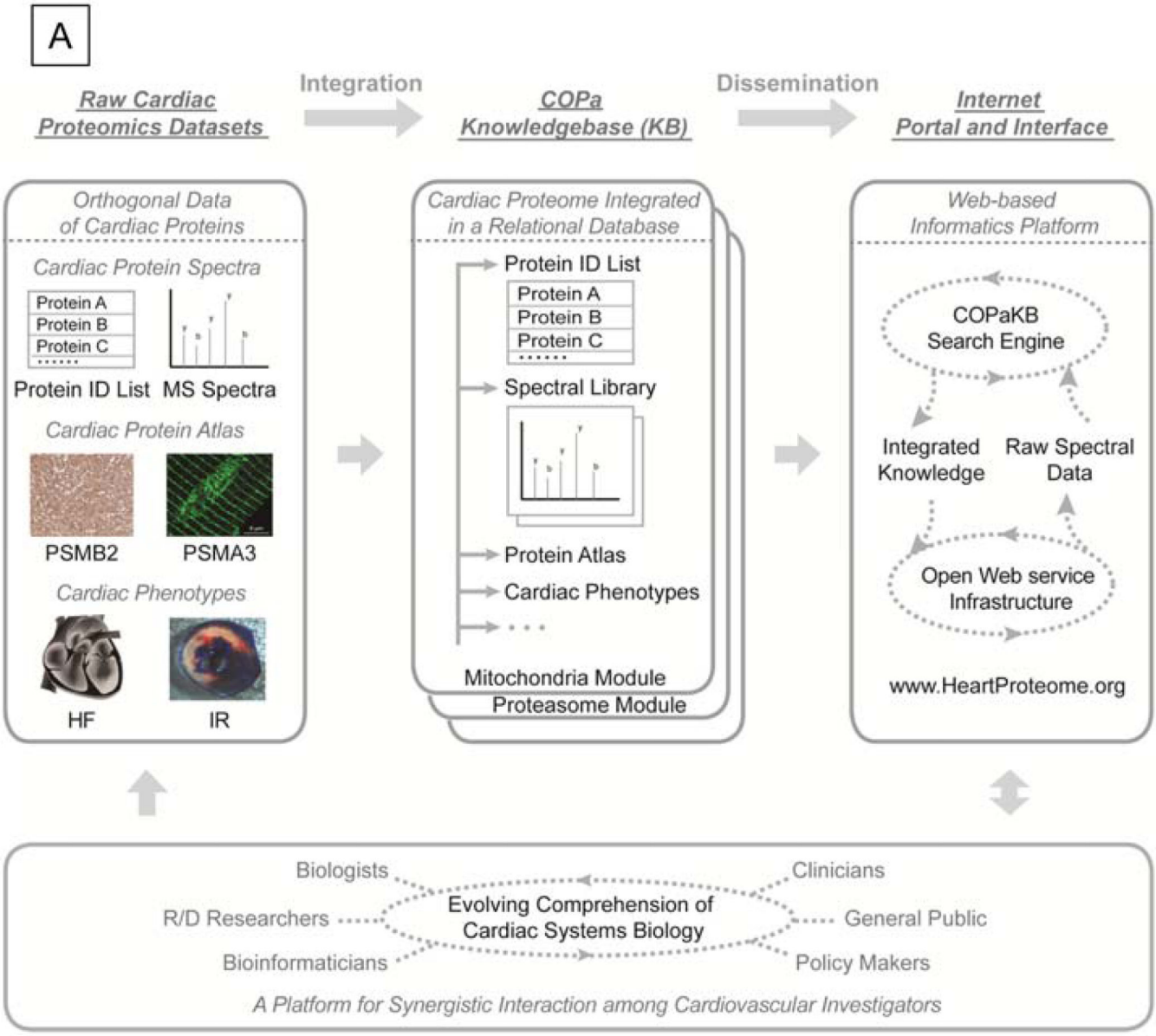
What Is Known?

- Proteomics techniques allow for large-scale analysis of global protein expression, but are not commonly accessible due to the need for specialized computational and informatics expertise.

What New Information Does This Article Contribute?

- Cardiac Organellar Protein Atlas Knowledgebase (COPaKB) is a new resource consolidating relevant protein datasets from multiple scientific disciplines and linking protein molecular properties to their functional phenotypes.
- COPaKB features a novel algorithm supporting user-directed protein pathway studies in their specified biological context of interests.
- COPaKB could serve as a centralized web portal, allowing remote data management, and as an online platform for collaboration among investigators.

Proteomics investigations have received increasing attention in cardiovascular research, but several obstacles remain to effective translation and utilization of proteomic data. These challenges include fragmented data structure, inconsistent data annotations, and, often, investigator inaccessibility to relevant technology platforms. COPaKB was created as a unique resource to facilitate better understanding of proteomic datasets: This platform is a curated relational database of protein molecular and biomedical phenotype properties, interfaced to a website for public data retrieval. It allows any investigator to process raw proteomic datasets without the need of accessing high-end instrumentation, and it returns a consistently annotated report of protein properties. The platform also offers a wide range of informatics tools for investigators to analyze different studies in parallel and to conduct meta-analyses.



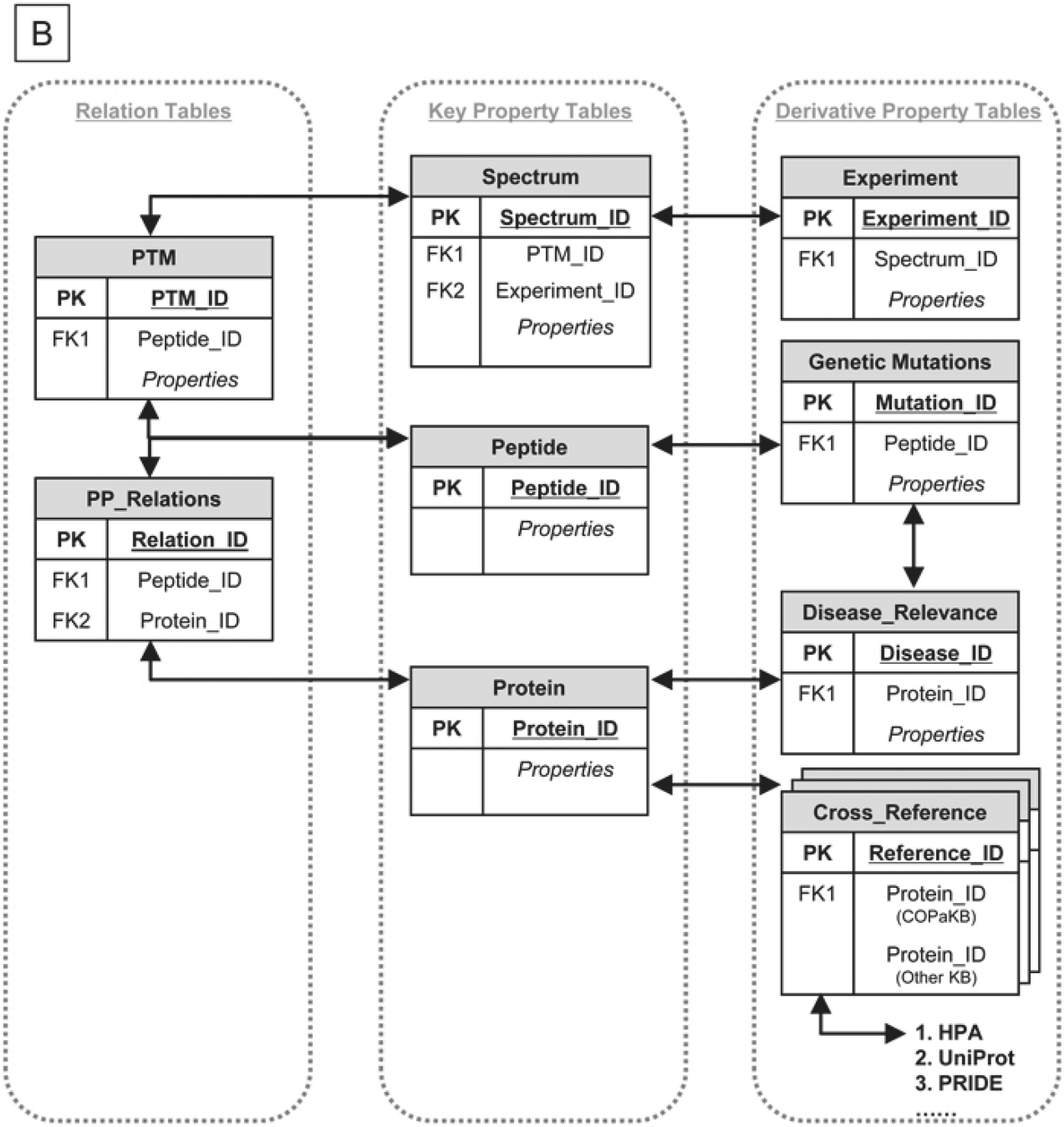


Figure 1. The Schema of the Cardiac Organellar Protein Atlas Knowledgebase (COPaKB)

A. Differential datasets of the cardiac proteome served as the basis for implementing COPa Knowledgebase (KB), including protein mass spectra, protein expression images (e.g. immunohistochemical image of PSMB2 expression, immunofluorescence image of PSMA3 expression), and cardiac phenotypes (e.g. heart failure, HF, ischemia reperfusion injury, IR). Using a protein identification list as a common index, these datasets were integrated into COPaKB in a relational database. In a modular structure, proteins were organized according to their organellar origins. Communication between COPaKB and cardiovascular

investigators was mediated by a dedicated internet portal and interface featuring an open web service infrastructure and a search engine. This infrastructure enables the delivery of integrated knowledge on the cardiac proteome in response to an input of raw spectral dataset. Collectively, COPaKB acts as a platform for synergistic interaction among cardiovascular investigators. **B**. The relational database uses relation tables (e.g. PTM table, PP relations table) to connect key property tables (e.g. spectrum table, protein table) of the cardiac proteome. Derivative property tables (e.g. disease relevance table) are connected to key property tables to archive diverse attributes of the cardiac proteome. Primary keys (PK) and foreign keys (FK) were used to establish correlations among these tables. (HPA stands for Human Protein Atlas; UniProt stands for Universal Protein Resource; PRIDE stands for PRoteomics IDEntifications database).

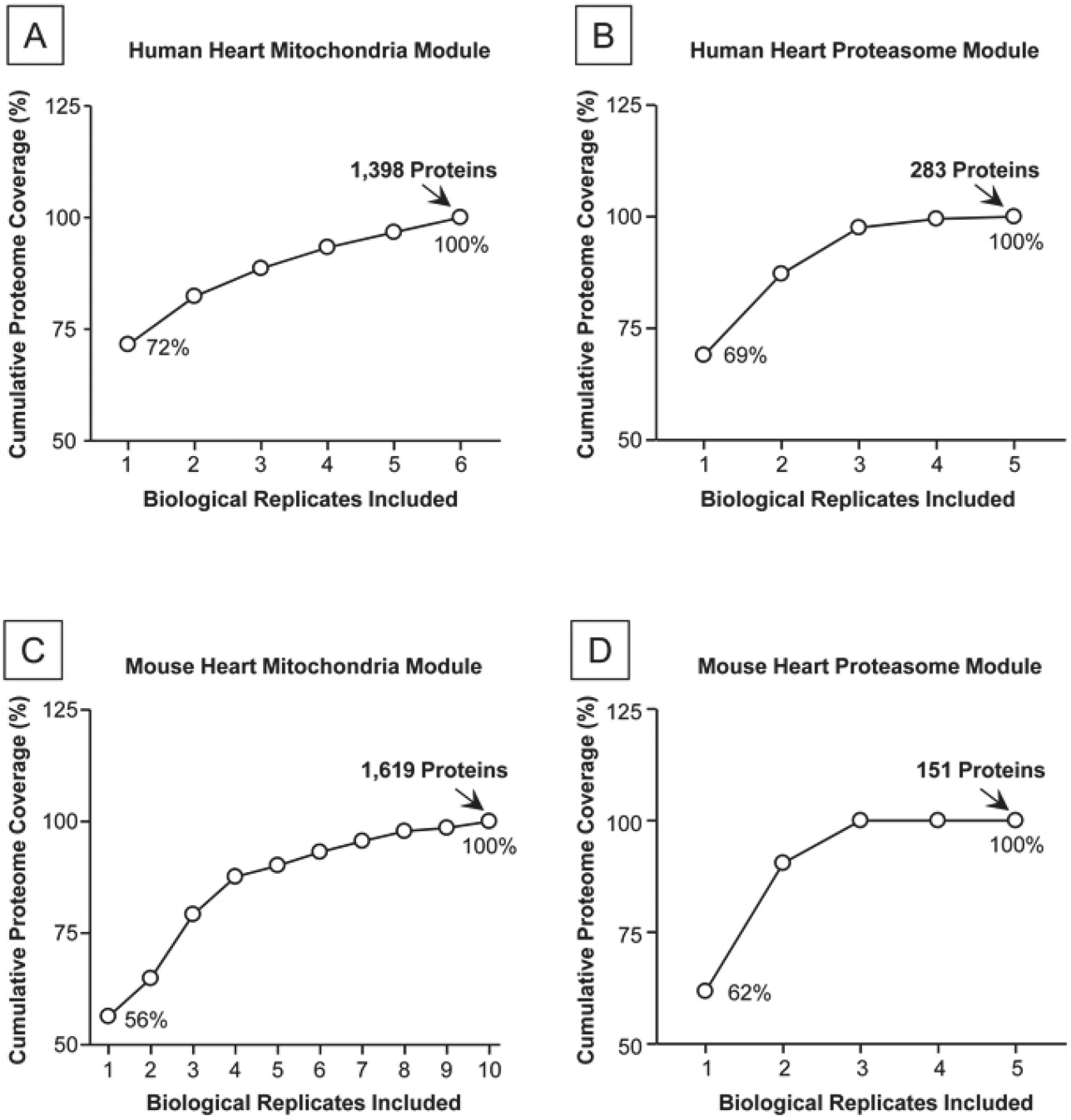


Figure 2. Coverage of the Four Organellar Modules in the COPa Knowledgebase

A. Six biological replicates of human mitochondria preparations were analyzed to construct this spectral library module. A total of 1,398 proteins were identified and compiled; the cumulative proteome coverage reached the limits of the LC-MS/MS platform. **B**. Five biological replicates were analyzed for the human proteasome module to reach a plateau in protein identification with 283 proteins. **C**. Ten biological replicates were analyzed for the mouse mitochondria module to reach a plateau of 1,619 proteins. **D**. Five biological

replicates were analyzed for the mouse proteasome module to reach a plateau of 151 proteins.

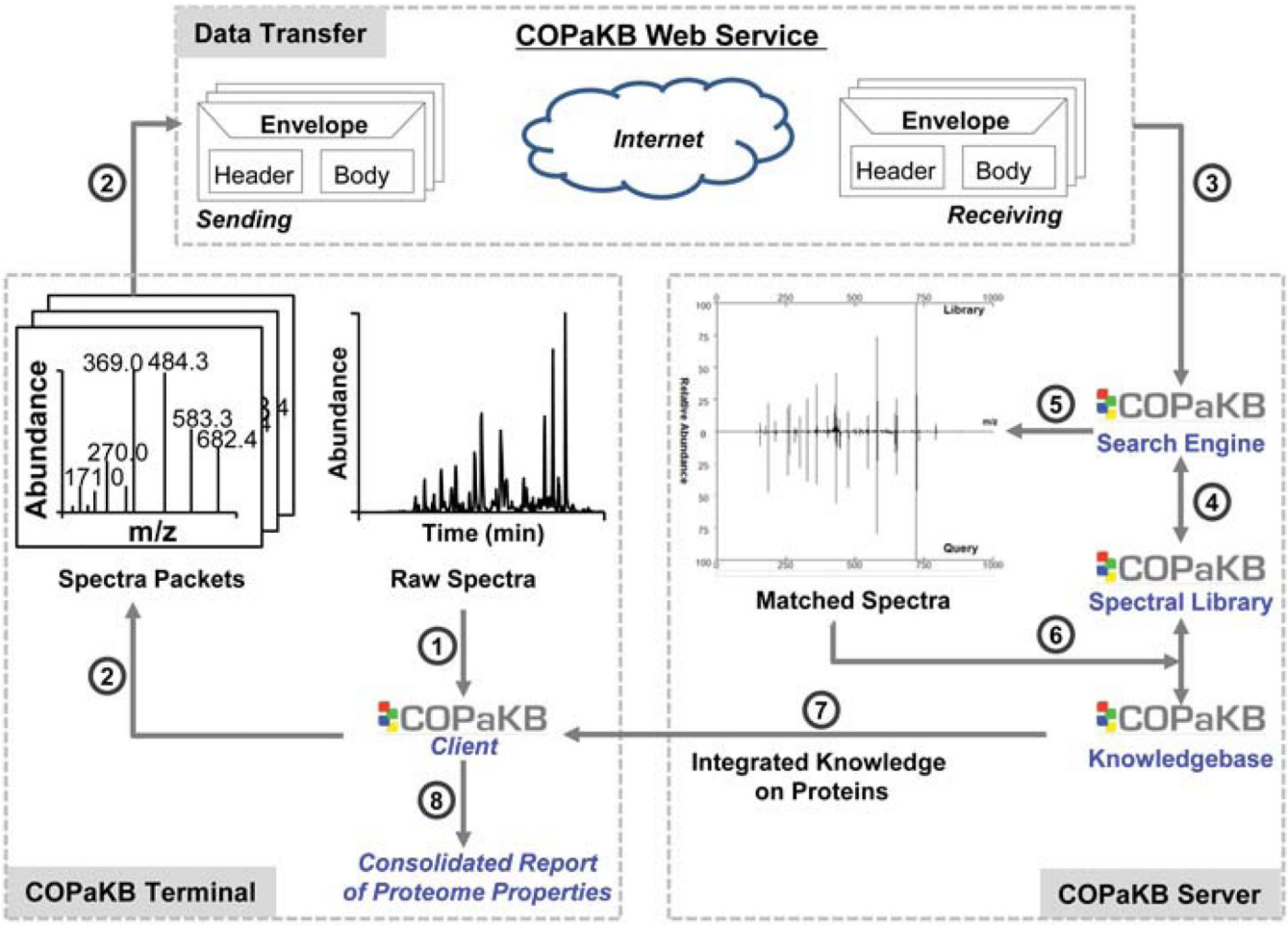
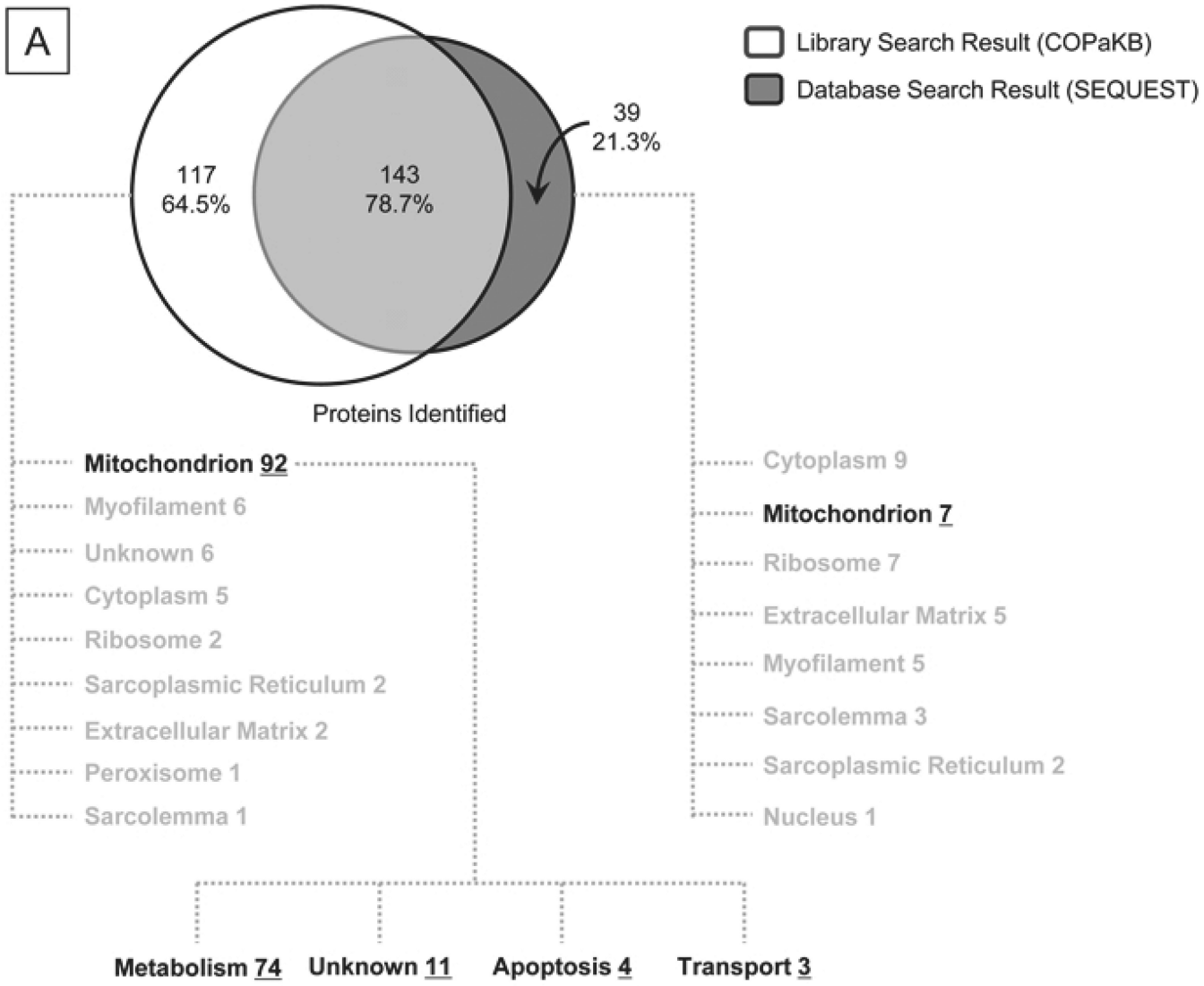
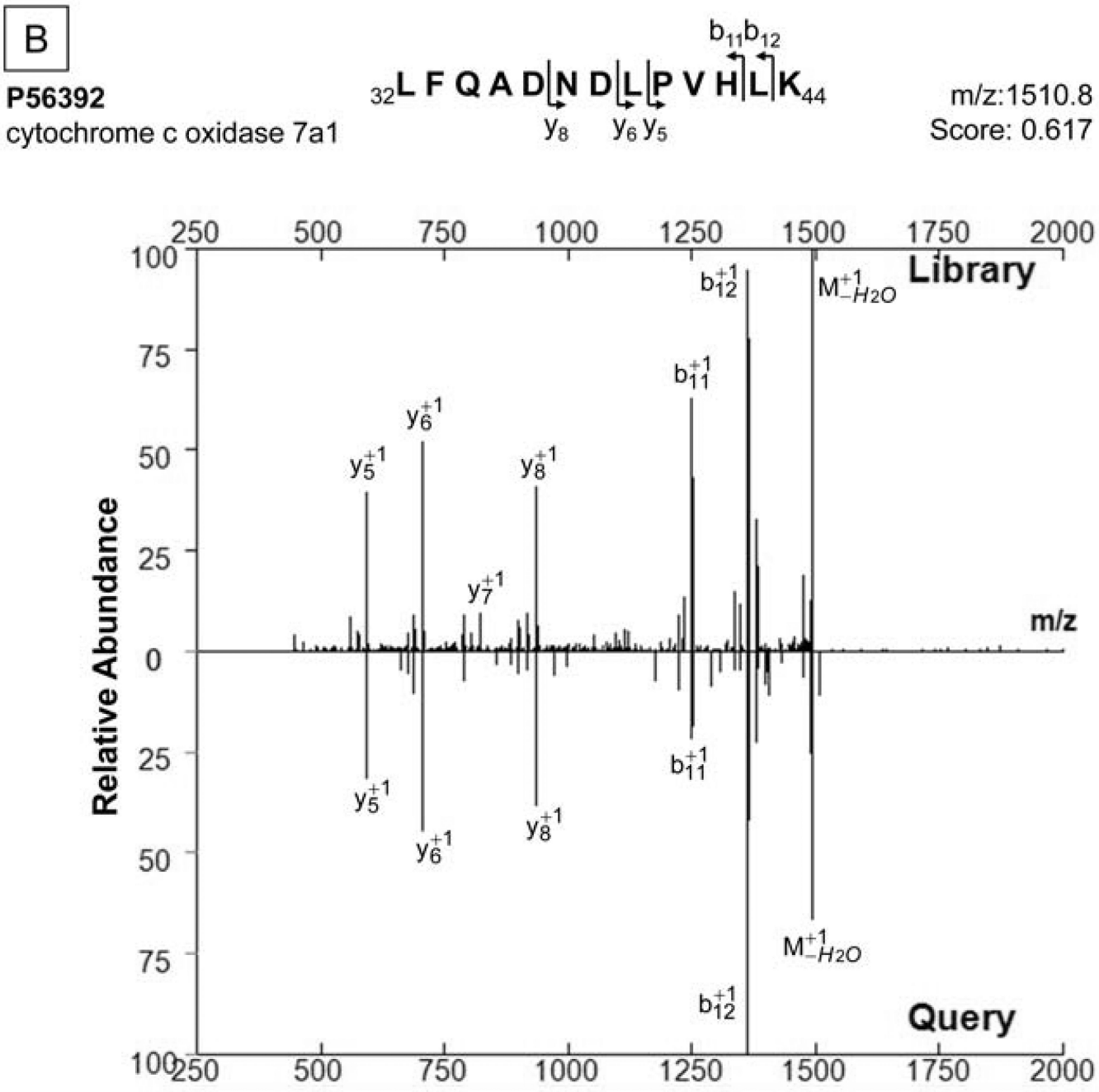


Figure 3. Large-scale Spectral Analysis over the Web

COPaKB Client orchestrates segmentation (1) and submission (2) of mass spectral data through the Internet to the COPaKB server. Data packets (3) received by the COPaKB server were analyzed using the spectral library as a reference (4). For matched spectra (5), the properties of their corresponding proteins are retrieved from COPaKB automatically (6), which are returned to COPaKB Client (7). By the end of the analysis, COPaKB Client presents a consolidated report (8) outlining the proteome properties encoded in the raw spectral files.





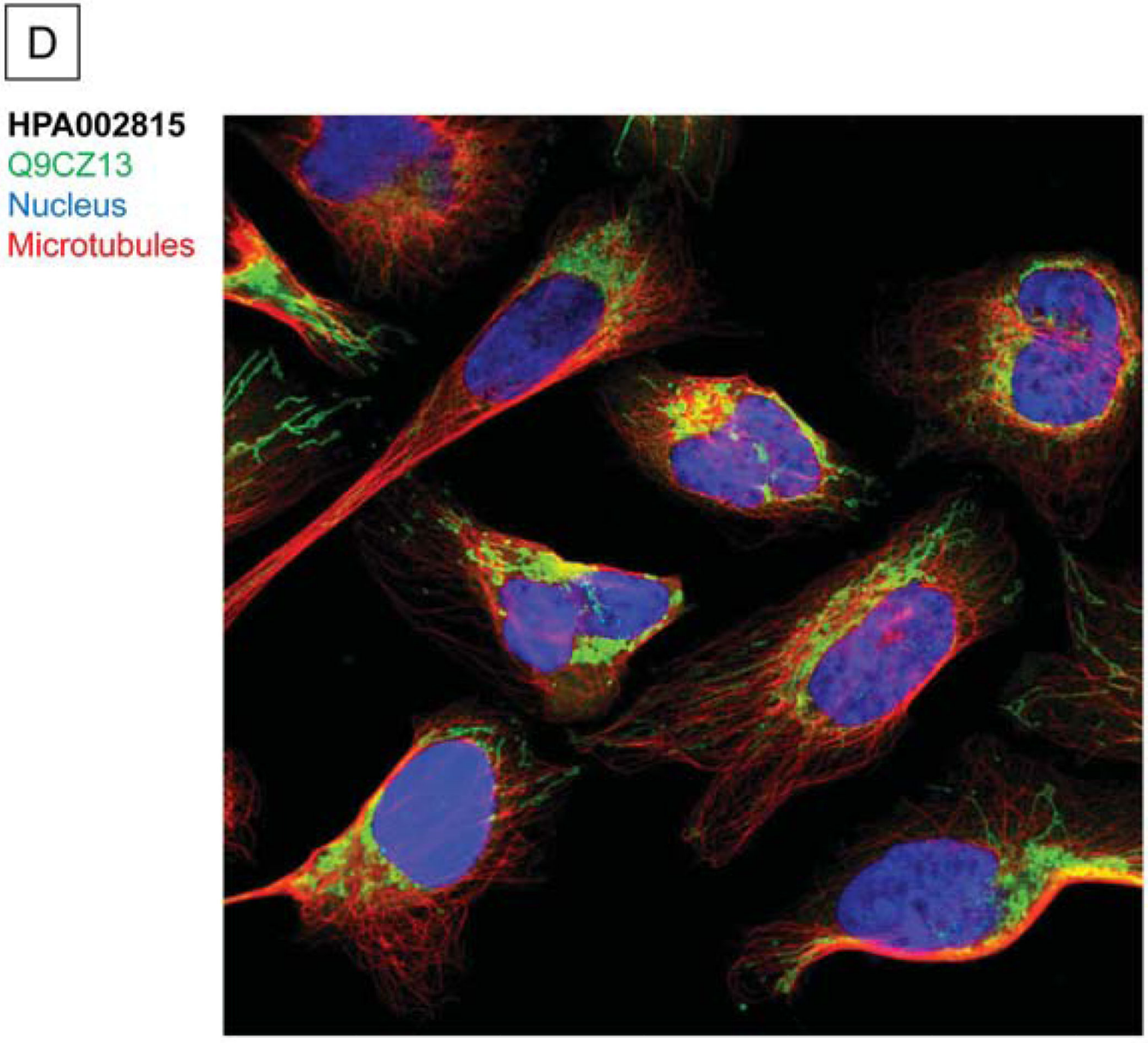


Figure 4. Sensitive Protein Identification via the COPaKB-mediated Web service

The test dataset of the murine mitochondrial proteome containing 111 raw data files was downloaded from the Peptide Atlas Repository[16](PAe000353)[17]. **A**. Compared to a database search workflow (SEQUEST), the COPaKB web service covered 144 shared proteins (78.7%) and an additional 117 identifications (63.9%). The 117 proteins uniquely identified by the COPaKB workflow were categorized according to their Gene Ontology annotations. Among them, 92 proteins had a subcellular location annotation of the mitochondrion with functional implications in metabolism (74), apoptosis (4), transport (3) and unknown (11). Among the 39 proteins identified uniquely with the sequence database search engine, only 7 proteins had a subcellular location annotation of the mitochondrion. **B**. Mass spectrum corresponding to peptide LFQADNDLPVHLK was identified as belonging to cytochrome c

oxidase 7a1 of the mitochondrial electron transport chain complex IV, which was identified by the COPaKB web service. **C**. The expression profile of cytochrome b-c1 complex subunit 1 (Q9CZ13) in human myocardium was probed by its specific antibody (ID: HPA002815). This image was automatically retrieved by the COPaKB web service from the Human Protein Atlas after its identification. **D**. Immunofluorescence image of this protein with the same antibody provided organellar resolution of protein expression. With the reference of organellar markers, Q9CZ13 was probed to express in mitochondria.

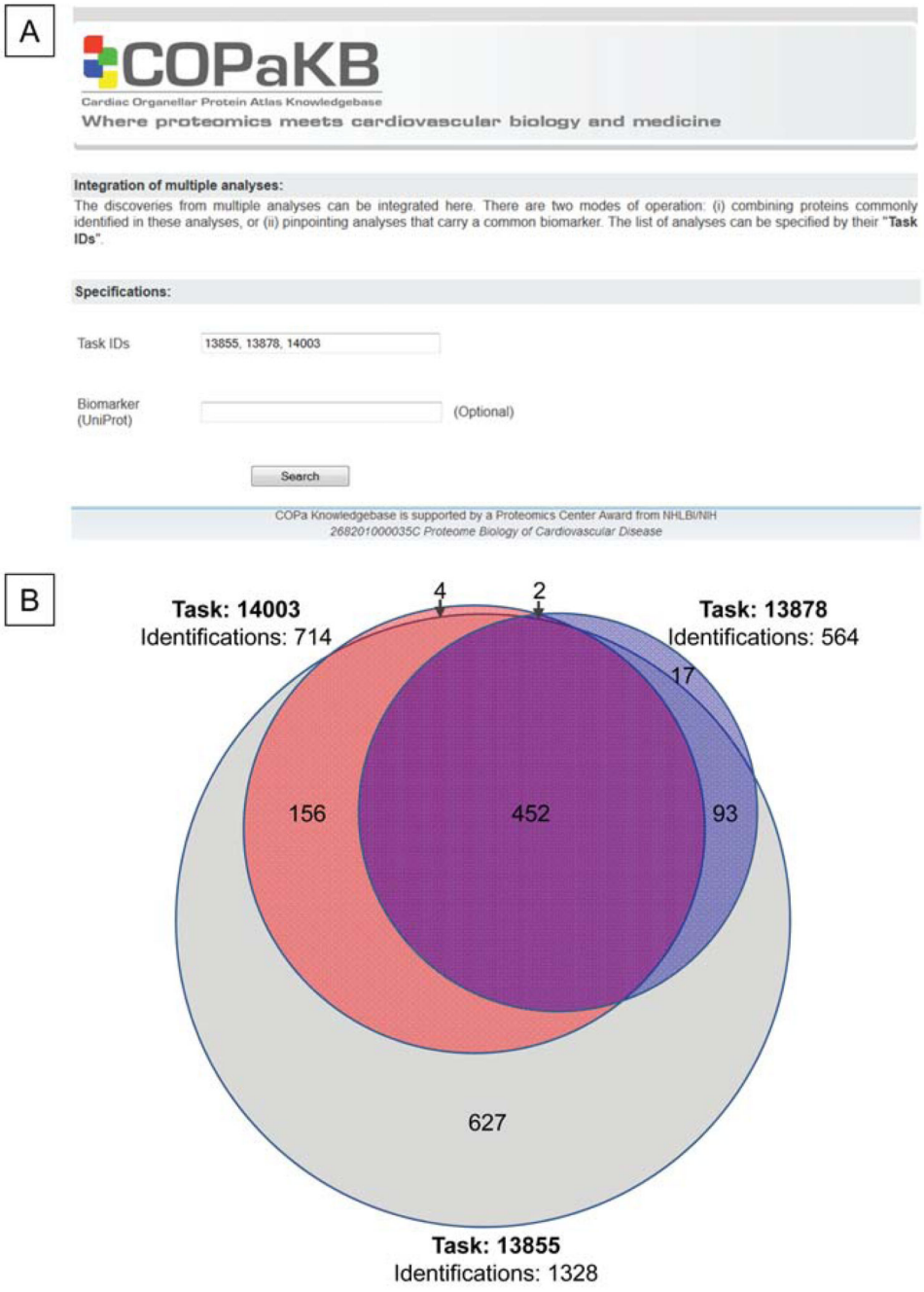


Figure 5. Integration of Discoveries from Multiple Analyses

A. Via COPaKB, discoveries from multiple proteomic investigations from discrete research group can be aligned by specifying the task ID of each analysis. Alternatively, the expression of selected biomarkers in these studies can be probed. **B**. There were 452 shared protein identities among the tasks 13855, 13878 and 14003. Meanwhile, 627 proteins were detected only in task 13855, 4 proteins only in 14003 and 17 proteins only in 13878.

Table 1

Cardiovascular Proteome Biology Integrated in a Modular Knowledgebase

Modules*	Proteins†	Peptides‡	Spectra§	Images//	Diseases#
Human Mitochondria	1,398	28,031	41,758	822	142
Human Proteasomes	283	3,482	5,668	146	53
Mouse Mitochondria	1,619	38,421	59,020		189
Mouse Proteasomes	151	6,409	9,409		29
Mouse Nuclei	1,048	6,918	9,115		
Mouse Cytosol	2,558	13,983	19,141		
Human Lysate	21,834	43,086	61,745	9,956	
Mouse Lysate	4,247	49,068	69,156		
Drosophila Mitochondria	1,015	13,770	27,185		
C. elegans Mitochondria	1,117	18,291	27,493		

*The organellar modules currently available in COPaKB.

†The number of non-redundant proteins entries in each module.

‡The number of non-redundant peptides entries in each module.

§The number of non-redundant spectra entries in each module.

//The number of proteins with an expression profile available in the Human Protein Atlas (HPA).

#The number of proteins with references to cardiovascular diseases in publications.

Table 2

Utilities and Functional Outputs of COPaKB

Queries*	Protein Identifier	Amino Acid Sequence	MS Data File(s)	Analysis of Multiple Datasets
Query Format†	Protein Name, its Gene Symbol, or its UniProt ID	Full or partial amino acid sequence of a protein, a peptide, or a specific domain	Directory or URL address(es) of existing MS data files on user protein samples	Existing analyses by any investigator team and/or by different groups
Entry Point‡	www.heartproteome.org/copa/KeywordSearch.aspx?type=1	www.heartproteome.org/copa/KeywordSearch.aspx?type=2	www.heartproteome.org/copa/MSRUNSearch.aspx or www.heartproteome.org/copa/COPaKBClient.aspx	www.heartproteome.org/epiSearch.aspx
Example Inputs§	Example 1: Protein Name, *Glycogen synthase kinase-3 alpha*[28]; its Gene Symbol, *GSK3α*; or its UniProt ID, *P49840*. Example 2: Protein Name, *RAC-alpha serine/threonine-protein kinase*[29]; its Gene Symbol, *AKT1*; or its UniProt ID, *P31749*.	Example 1: TTTLAFKFR Example 2: GGVTTFVALYDYESRTETD LSFKKGERLQIVNNTEGDW WLAHSLSTGQTGYIPSNYV APSDS	Example 1: www.heartproteome.org/sample/human_proteasome_sample.mzML Example 2: https://dl.dropboxusercontent.com/u/65582675/1Orbi_Mito_B12_A.mzML	Example 1: 14455, 14457, 14458 Example 2: See Figure 5
Reports by COPaKB//	• Reported cardiac phenotypes • mRNA expression profiles • Interacting partners • Tissue immunohisochemistry images • Immunocytofluorescence images • A list of peptides identified w/MS data • Peer-reviewed reports provided by iHOP • A Wiki page for further annotations	• PTM sites identified# • Its protein origin when querying a peptide or a domain. • Expression in organelles • A Wiki page for further annotations	• Protein expression profiles • Interacting partners • PTM sites identified w/MS data • Protein abundance information • MS spectra of identified proteins • A unique identifier for each user sample analysis	• Proteins commonly expressed in multiple samples • Interacting partners • Protein expression comparison • Protein abundance comparison
Example Outputs**	www.heartproteome.org/copa/proteinInfo.aspx?qType=protein%20ID&qValue= P49840	www.heartproteome.org/copa/PeptideInfo.aspx?qType=Peptide%20Sequence&qValue=TTTLAFKFR	www.heartproteome.org/copa/Report.aspx?TaskID=14455	www.heartproteome.org/copa/epiTest.aspx?Tasks=14455,%2014457,%2014458&Markers=

*Users can query COPaKB using multiple specific inputs.

†User-directed queries allow several formats.

‡Users can query COPaKB by entering these query requests at a number of sites available in COPaKB.

§Users can submit inputs in formats as shown here in the example.

//Users receive a report of these information from COPaKB in response to their queries.

#Analysis of post-translational modification sites includes phosphorylation, acetylation, ubiquitination, oxidation and carbamidomethylation.

**Users can view examples of outputs as shown via the links listed here.

Table 3

Performance of the COPaKB Web Service across the Globe

Test Dataset*	Centers†	SOAP (mm:ss)‡	HTTP (mm:ss)§	Terminal Bandwidth//
Human Mitochondria (LTQ-Orbitrap) 12,197 scans 1.34 GB (mzML file)	Los Angeles, CA	1:28	3:56	100 Mbps
	San Diego, CA	2:53	21:12	1 Gbps
	Woods Hole, MA	5:56	29:12	130 Mbps
	Ann Arbor, MI	2:45	59:22	144 Mbps
	Cambridge, UK	8:27	46:42	1 Gbps
	Shanghai, CN	18:25	53:48	100 Mbps
Mouse Proteasomes (LTQ-Orbitrap) 17,746 scans 1.88 GB (mzML file)	Los Angeles, CA	1:17	4:05	100 Mbps
	San Diego, CA	5:45	30:24	1 Gbps
	Woods Hole, MA	26:45	45:46	130 Mbps
	Ann Arbor, MI	4:18	>60	144 Mbps
	Cambridge, UK	14:01	57:21	1 Gbps
	Shanghai, CN	13:01	>60	100 Mbps
Human Mitochondria (LTQ-XL) 20,677 scans 292 MB (mzML file)	Los Angeles, CA	5:35	6:05	100 Mbps
	San Diego, CA	14:37	7:45	1 Gbps
	Woods Hole, MA	22:21	10:43	130 Mbps
	Ann Arbor, MI	11:15	16:08	144 Mbps
	Cambridge, UK	28:31	12:48	1 Gbps
	Shanghai, CN	23:10	12:22	100 Mbps
Human Proteasomes (LTQ-XL) 20,736 scans 249 MB (mzML file)	Los Angeles, CA	2:56	3:00	100 Mbps
	San Diego, CA	9:09	5:11	1 Gbps
	Woods Hole, MA	13:05	7:02	130 Mbps
	Ann Arbor, MI	6:52	13:01	144 Mbps
	Cambridge, UK	28:21	10:40	1 Gbps
	Shanghai, CN	12:00	8:53	100 Mbps

*The source and size of each test file.

†The location of the six test center.

‡The time needed for completing an analysis via a SOAP portal.

§The time needed for completing an analysis via a HTTP portal.

//The internet bandwidth available at each test center.

Table 4

Comparison of Analytical Efficiency of COPaKB Client and SEQUEST

Test Dataset*	SEQUEST (min)†				COPaKB Client (min)‡	
	PC	Head Node	Cluster 7 Nodes	Cluster 13 Nodes	PC	Head Node
Human Mitochondria (LTQ-Orbitrap)	21:40	16:39	1:14	0:45	1:35	0:57
Mouse Proteasomes (LTQ-Orbitrap)	21:49	19:32	1:15	0:44	2:24	1:15
Human Mitochondria (LTQ-XL)	286:00	277:00	12:54	0:29	5:05	2:13
Human Proteasomes (LTQ-XL)	263:00	259:00	11:57	0:23	2:39	2:09

*The source of each test file.

†The time needed for completing an analysis by SEQUEST (BioWork).

‡The time needed for completing an analysis by COPaKB Client.