Title

Cardiovascular proteomics in the era of big data: experimental and computational advances.

Authors

Lam, Maggie
Lau, Edward
Ng, Dominic
et al.

Publication Date

2016

DOI

10.1186/s12014-016-9124-y

Clinical Proteomics

REVIEW

Open Access

Cardiovascular proteomics in the era of big data: experimental and computational advances

Maggie P. Y. Lam[1*], Edward Lau[1], Dominic C. M. Ng[1], Ding Wang[1] and Peipei Ping[1,2,3]

Abstract

Proteomics plays an increasingly important role in our quest to understand cardiovascular biology. Fueled by analytical and computational advances in the past decade, proteomics applications can now go beyond merely inventorying protein species, and address sophisticated questions on cardiac physiology. The advent of massive mass spectrometry datasets has in turn led to increasing intersection between proteomics and big data science. Here we review new frontiers in technological developments and their applications to cardiovascular medicine. The impact of big data science on cardiovascular proteomics investigations and translation to medicine is highlighted.

Keywords: Cardiovascular medicine, Clinical proteomics, Shotgun proteomics, Mass spectrometry

Background

The heart is in many ways an exceptional organ. Proteins at the sarcolemma, sarcomere, mitochondrion, and other cardiac organelles must orchestrate vital functions seamlessly on a beat-by-beat basis, while dynamically adjusting energetic and contractile outputs to environmental cues within seconds. Heart diseases including cardiac hypertrophy and failure are characterized by complex remodeling of various protein signaling networks and subcellular components, which often involve a multitude of collaborating proteins. Therefore, understanding how multiple protein species interact to carry out higher physiological phenotypes and regulation has been an important objective of cardiovascular research. The power of proteomics to simultaneously provide information on the panoply of expressed proteins has made it uniquely suitable for resolving complex signaling conundrums and revealing disease mechanisms in the heart.

Advances in genome sequencing are often celebrated to have outpaced even the vaunted Moore's law of computing power [1]. Lesser known but equally impressive is the parallel surge in the capacity of mass spectrometry-based proteomics last decade. To wit, when the first draft of the human genome was published in 2001, a state-of-the-art two-dimensional electrophoresis technique could identify ~200 proteins in 3 days. Fast-forward to today, a modern mass spectrometer can generate over a million spectra per day and quantify ~4000 proteins in the course of 1 h [2]. This quantum leap is attributable to parallel advances in three areas: (1) analytical chemistry in sample processing and liquid chromatography tandem mass spectrometry (LC–MS/MS) instrumentation; (2) bioinformatics and computational tools in high-throughput data processing and analysis; and (3) completeness and accuracy of sequence and annotation databases. Riding on growing experimental capacity, there have been continued improvements to the experimental coverage of proteome analysis, the interpretability of data, the reliability of results, and the diversity of protein parameters that may be interrogated. New applications and experimental designs not possible a few years ago are now being exploited to explore new regulatory modalities in cardiac physiology.

Cardiovascular proteomics has grown rapidly in the intervening period, with >400 studies now being published yearly (Fig. 1) [3]. To put into context, we recall two landmark reviews of cardiovascular proteomics, in

*Correspondence: magelpy@ucla.edu
[1] NIH BD2K Center of Excellence at UCLA; Department of Physiology, University of California at Los Angeles, 675 Charles E. Young Drive, Los Angeles, CA 90095, USA
Full list of author information is available at the end of the article

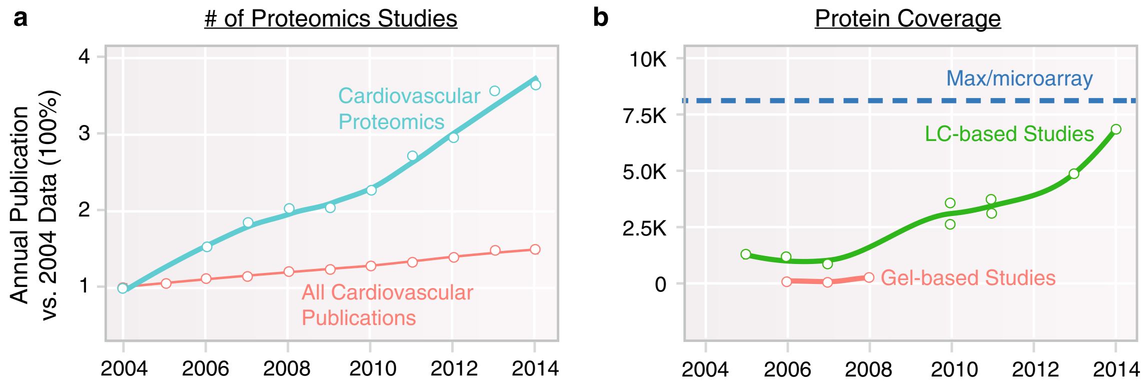

Fig. 1 Trends in cardiovascular proteomics. Both (**a**) the volume of proteomics studies, and (**b**) the size of proteomics dataset have skyrocketed in the last decade. **a** The number of cardiovascular proteomics studies has increased approximately 400 % from 2004 to 2014, far outpacing the natural growth of the cardiovascular field, indicating increasingly common adoption of the technologies. **b** The protein coverage of proteomics experiments in the same time period has experienced considerable growth also, quantified as the numbers of identifiable cardiac proteins in an experiment. The maximum number of cardiac proteins (*dashed lines*) is based on estimated significantly expressed loci in the mouse heart and does not take into account proteoforms such as resulting from alternative splicing. This increase is driven by parallel advances in hardware instrumentation and computational technology. Coinciding with the notion of "complete proteomics", proteomics studies can now interrogate more proteins of interest such as chromatin remodeling factors and transcription factors that express at low copy numbers. Effective means to analyze big proteomics dataset are becoming a new frontier of growth in cardiovascular proteomics

2001 [4] and 2006 [5], which noted that although many enabling technologies were emerging, cardiovascular proteomics remained a field 'on the threshold' of future applications. Fast forward to the present and it is clear that proteomics has had a transformative impact on cardiovascular sciences, as recounted in recent review articles. We attempt to complement these reviews here with a concise overview on the lockstep improvements in the analytical (separation sciences and mass spectrometry) and computational (data science and algorithms) advances of the past 5 years that enabled landmark studies, as well as ongoing developments driving the next stage of applications.

Experimental and analytical advances

Improvements in analytical methods

An early hurdle that bedeviled cardiovascular proteomics was the limitation in the sensitivity and dynamic range of protein detection, which skewed results towards few high-abundance proteins (e.g., contractile proteins) and masked low-abundance species. This is due to the complexity of proteomes. The mammalian heart is known to express at least ~8000 genes at significant levels [6], and at least 8325 human proteins have been referenced in the ~1.4 million cardiac-related publications on PubMed [7]. Each human gene can encode multiple proteoforms, e.g., due to the average ~4 alternative splicing isoforms per human gene plus many more post-transcriptional and post-translational editing processes, resulting in at least ~10^6 proteolytic peptides per sample. With the addition of post-translational modifications (PTMs)—e.g., the four histone proteins alone have identified PTMs on at least 105 different residues in myriad combinations [8]—the total proteome complexity is likely orders of magnitude more complex still.

In the past decade, great strides have been made to improve proteome coverage, from how protein samples are extracted to mass spectrometry instrumentation. To perform proteomics analysis, it follows that the proteins must be effectively extracted and released from biological samples. This is typically achieved via mechanical homogenization or chemical surfactants. Protein solubilization techniques in early proteomics protocols were at times ineffectual in extracting hydrophobic or membrane proteins, which often aggregated out of the sample and led to their non-detection. Development in this area in recent years have led to commercially available, mass spectrometry compatible surfactants [9], size exclusion filter-mediated buffer exchange techniques [10], as well as empirically refined experimental protocols that are optimized for the analysis of various cardiac subproteomes [11]; all of which serve to expand the portion of the proteome that is open to mass spectrometry exploration. The incompleteness of proteolytic digestion was once found to be a limiting factor of the maximal peptide coverage of the experiment and contribute to batch-to-batch variations. The use of optimized proteolysis protocols including double lys-C/trypsin proteolysis is gaining traction [1].

Advances in separation sciences have had a particularly tremendous impact on reducing the complexity of peptides prior to mass spectrometry signal acquisition. High-performance shotgun proteomics using mass spectrometry has supplanted two-dimensional (2D) electrophoresis to become the de facto standard for large-scale analysis of cardiac proteins (see general workflow of shotgun proteomics in Fig. 2). Whereas the now-dethroned 2D electrophoresis was limited to detecting a few hundred proteins, contemporary LC–MS experiments can resolve peptides from >10,000 proteins to allow their identification and quantification. Since 2001, separation science has led in a relentless pursuit to increase protein coverage [12, 13], with the success of strong cation exchange-reversed phase based MudPIT approaches followed successively by other 2D-LC approaches including reversed-phase-reversed-phase separation [14] as well as very-long separation columns with high peak capacity, nano-scale microfluidic devices driven by ultra-high pressure LC systems [15] and capillary electrophoresis separation (see [16, 17] for reviews on separation science developments). By separating the peptide samples into smaller subset based on their chemistry, a simpler mixture of peptides is introduced into the mass spectrometer in any given time, which decreases ion competition and increases sensitivity.

State-of-the-art instruments including hybrid Orbitraps and time-of-flight instruments achieve high performance by virtue of their high scan speed (allowing more peptides to be analyzed in the same analysis), sensitivity (allowing minute amounts of samples to be analyzed), and mass resolution (increasing power to differentiate similar peptide species). Recent proteome profiling experiments of the mammalian heart using the latest and greatest LC–MS combinations routinely achieve 5000 or more proteins identified in an experiment (Fig. 1b). In a recent survey we quantified the relative abundance of 8064 proteins in the mouse heart, covering more than 10 major organelles and 201 major cellular pathways [18]. As little as micrograms of proteins are sufficient for shotgun

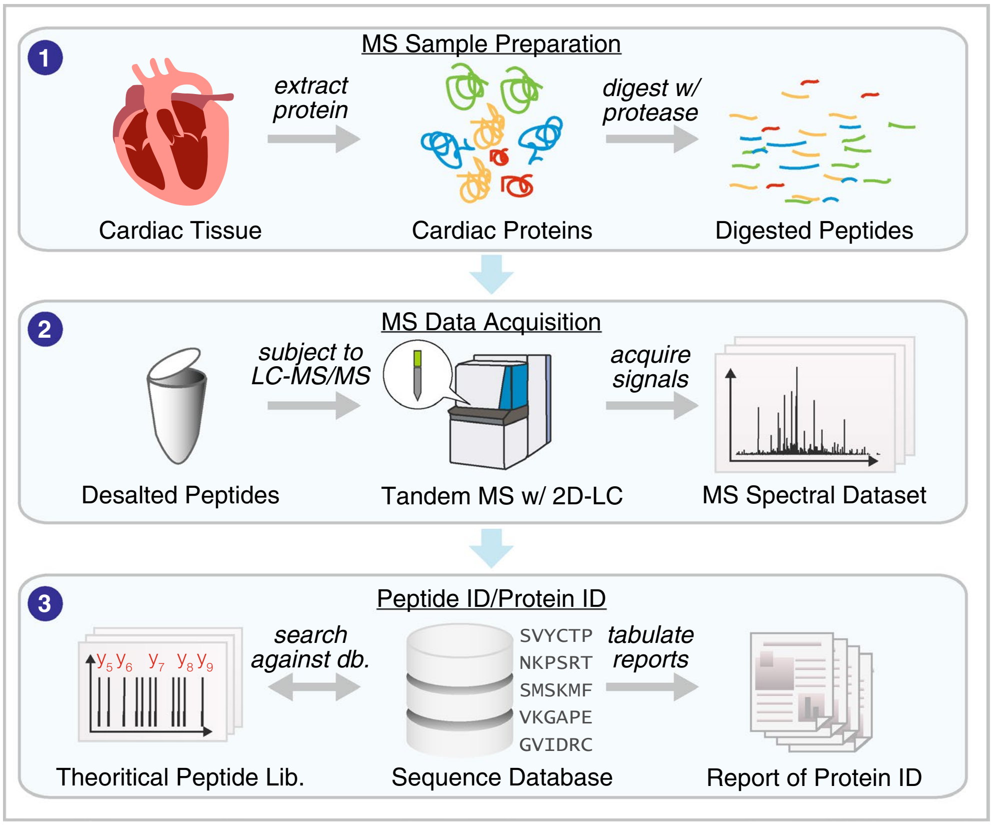

Fig. 2 Analytical and computational overview in protein identification. *1* Cardiac samples are processed to extract the proteomes or subproteomes of interest, which may then be proteolyzed to obtain peptide digests. *2* The resulting peptides are desalted and subjected to LC–MS/MS analysis to acquire MS1 and MS2 spectra. *3* The peptide sequences that are present in the MS dataset can be identified using a database search approach, which uses a sequence database (e.g., UniProt) to generate theoretical peptide sequence and predict their fragmentation patterns in silico, then automatically find the best-match theoretical spectra to the experimental spectra for protein identification. Alternatively, the proteins can be identified using a spectral library search. The resulting protein datasets can be further analyzed to extract other biomedically meaningful information (see Fig. 4)

proteomics analysis. This amount may come from milligrams or less of cardiac biopsies, or ~10^5 cultured cardiac cells, opening the door of proteome analysis to more experimental and clinical designs where sample amounts may be limiting.

Taken together, these advances have helped solve a principal challenge to proteomics applications, namely how to successfully detect the maximal number of peptides inside the overwhelmingly complex mixture that is the cardiac proteome. Although each technological development is incremental, over time they accrued into a qualitative transformation on the power and utility of proteomics, when proteins of biomedical interest gradually became measurable and discoverable in large-scale experiments. Catalogs of so-called "complete proteomes" (i.e., here narrowly defined as the detection of one protein product of every expressed locus in the genome) have now been described for many human tissues and organs, including the heart [19, 20]. Therefore, although earlier cardiovascular proteomics studies were best equipped to discover changes in structural, contractile, or metabolic housekeeping proteins, contemporary studies can now easily interrogate regulatory proteins including membrane receptors, kinases, ubiquitin ligases, and chromatin remodeling factors, whereas the analyses of yet scarcer species such as transcription factors and cytokines are now on the cusp of routine applications.

Early applications were also plagued by the variability of quantification results, which limited power to discover significant changes between disease model and control samples. A major source of variability in proteomics quantification originated from the variable detectability of peptides with different amino acid compositions. Two equimolar peptide sequences can show rather different intensities in MS signals. Accurate prediction of peptide intensity based on sequence information remains an unsolved issue in computational proteomics due to the large number of combinatorial variables that contribute to signal behaviors. Several methods have been developed to normalize peptide intensity and achieve accurate quantification. Targeted MS methods, such as Multiple Reaction Monitoring (MRM) [21], allow users to program the mass spectrometer to scan for only targeted peptide ions for quantification. An advantage of targeted MS is the gain in reproducibility and sensitivity, which can avail the detection of low-abundance proteins at their native concentration. Targeted assays have been successfully developed such that very low amount of proteins in the sample can be quantified, as in the case of troponin I [22]. Isotope labeling methods including SILAC and Super-SILAC [23, 24] can also reduce variability in relative quantification by ensuring peptides from multiple samples are compared in identical experimental conditions, but require additional labeling steps.

With advances in data acquisition methods, non-targeted label-free techniques can also reliably deduce accurate protein intensity from shotgun experiments directly through bioinformatics analysis. Label-free quantification is analogous to deducing transcript abundance from read counts in next-generation sequencing. Existing approaches largely fall into two categories (Fig. 3). Spectral counting exploits the bias of shotgun proteomics towards abundant proteins, and calculates protein quantity from the stochastic sampling frequency of peptide ions, i.e., the higher the protein abundance, the more of its MS spectra are likely to be identified. A major advantage of spectral counting is that it quantifies directly from the identification output and thus is compatible with most workflows. Spectral counting algorithms tally the number of redundant spectra for each identifiable peptide, then sum the numbers of spectra for all peptides assigned to a protein. On the other hand, ion intensity approaches integrate the intensity mass-specific ion signals over time in the chromatographic space. This utilizes a feature detection step in data analysis to read raw MS files and integrate the corresponding areas-under-curve of each peptide ion over time. Both labeled and label-free methods provide a useful guide to differential protein expression, and can now be used to discover candidate disease protein that can then be validated by further studies.

Improvements in software tools

The massive amount of MS data generated in proteomics experiments requires computational aid for effective data processing and analysis. A growing number of open-access computational tools concerning all steps of proteomics data analysis are now freely available to users, a subset of which are listed in Table 1.

A major computational task in shotgun proteomics is to efficiently interpret the mass and intensity information within mass spectral data to identify proteins. The computational task can be formulated thus: given a particular tandem mass spectrum, identify the peptide sequences most likely to have given rise to the set of observed parent molecular mass and fragment ion patterns in a reasonable time frame. A general solution to this problem is "database search", which involves generating theoretical spectra based on in silico fragmentation of peptide sequences contained in a protein database, and then systematically comparing the experimental MS spectra against the theoretical spectra to find the best peptide-spectrum matches. The SEQUEST algorithm was the first proposed to solve this peptide spectrum matching

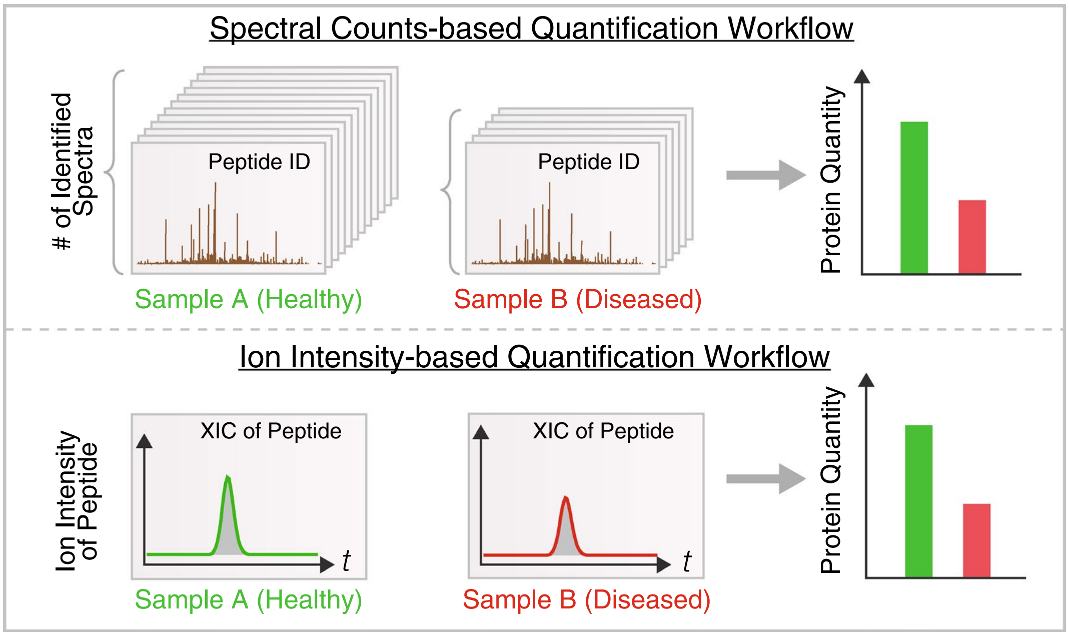

Fig. 3 Common label-free quantification in proteomics studies. Two common label-free quantification approaches in use are based on spectral count (*top*) and ion intensity (*bottom*). *Top*: spectral counting methods leverage the fact that in stochastic shotgun profiling, the frequency of a protein being sampled by the instrument scales with its relative abundance in the sample. The numbers of spectra matched to an identical protein in healthy (*green*) versus diseased (*red*) samples can therefore be compared if appropriate normalization and bioinformatics workflows are implemented. *Bottom*: ion intensity methods integrate the total signal intensity of peptide ion signals in the mass spectrometer to infer protein quantity. Software is now available to automatically identify and quantify peptide ion signals from mass spectra

problem in 1994 [25] and its variants (e.g., Comet, ProLuCID [26]) remain among the most widely utilized algorithms to-date for peptide identification. SEQUEST-style algorithms score peptide-spectrum matches in two steps, with the first step calculating a rough preliminary score which empirically restricts the number of sequences being analyzed, and the second step deriving a cross-correlation score to select the best peptide-spectrum match among the candidates. Recent descendants of the SEQUEST algorithms have focused on optimizing its searching speed as well as improving the statistical rigor of candidate sequence scoring, with some programs reporting ~30 % more peptides/proteins identifiable from identical MS datasets and better definition of true-/false-positive identifications [26–29]. Other search engines also exist that are commonly in use, including X!Tandem, which calculates the dot product between experimental and theoretical spectra, then derives the expectation value of the score being achieved in a random sequence match; MaxQuant/Andromeda, which considers fragment ion intensities and utilizes a probabilistic model for fragment observations [30], MS-GF+ [31], and others. Methods have also been developed to combine the unique strengths and biases of multiple search engines to improve total protein identifications [32].

Means to distinguish true and false positives are critical to all large-scale approaches. The “two-peptide rule” was once commonly adopted to decrease false positives at the protein level by requiring each protein to be identified by at least two independent peptides. However, this rather conservative rule could inflate false negatives, as some short or protease-incompatible proteins may only produce maximally one identifiable peptide. More recent conventions involve foregoing the two-peptide rule and instead estimating the false discovery rate (FDR) of identification through statistical models, often with the aid of decoy databases. The use of decoy databases/sequences (reversed or scrambled peptide sequences), allows a quick estimation of the number of false positive proteins, by assuming identical distribution in protein identification scores for false positive hits and the decoy hits. A maximum acceptable FDR can then be specified (conventionally 1–5 %) to determine which protein identifications are accepted in the final result. To explicitly reveal the posterior probability of any particular identification being correct (also called the local FDR), a mixture model has been used that assumes that the peptide identification result is a mixture of correct and incorrect peptides with two distinct Poisson distributions of identification scores [33]. Auxiliary determinants including the presence of other

Table 1 Selected open access software tools in proteomics

Open access tools	Language/framework	License	Publication	Website
Database search engine (untargeted proteomics)				
Comet*,†	C++	Apache 2.0	[93]	[94]
MS-GF+*,†	Java	Custom/Academic	[31]	[95]
MSAmanda†	C#/Mono	Custom/Academic	[96]	[97]
ProLuCID*,†	Java	Custom/Academic	[26]	[98]
X!Tandem*,†	C++	OSI Artistic	[99]	[100]
Targeted proteomics and/or data-independent acquisition				
Skyline*	C#	Apache 2.0	[101]	[102]
OpenSWATH*,†	C++	BSD 3-Clause	[103]	[104]
Protein inference and/or search post-processing				
Percolator*,†	C++	Apache 2.0	[34]	[105]
ProteinProphet*,†	C++	GNU LGPLv2	[106]	[107]
ProteinInferencer†	Java	Custom/Academic	[35]	[98]
Protein quantification				
MaxQuant	.NET	Custom/Academic	[108]	[109]
Census†	Java	Custom/Academic	[110]	[98]
PLGEM*,†	R	GNU GPLv2	[111]	[112]
QPROT*,†	C	GNU GPLv3	[37]	[113]
Pipelines and toolkits				
Perseus	.NET	Custom/Academic	[114]	[115]
Crux*	C++	Apache 2.0	[116]	[105]
OpenMS*	C++	BSD 3-Clause	[117]	[118]
TPP*	C++	GNU LGPLv2	[119]	[120]
Data access and reuse				
PeptideShaker*	Java	Apache 2.0	[121]	[122]
PRIDE inspector*	Java	Apache 2.0	[123]	[124]

Proteomics software tools that provide open access to users. Many of these tools are also open source which potentially allows users to participate in the continual development of the tools

* Available open-source source code repository at the time of writing

† Platform-independent (Windows, Linux, Mac)

identified peptides from the same proteins can also be applied to infer overall likelihood of protein assignment [33]. Machine learning algorithms (e.g., Percolator) have been demonstrated to build classifiers that automatically distribute peptide spectrum matches into true and false positives [34]. New inference approaches have also been demonstrated that consider peptide and protein level information together to improve the confidence of identification [35].

With the increase in data size and multiplexity (number of sample compared) in proteomics experiments, statistical approaches to analyze data have also evolved to tackle high-dimensionality data. Whereas early studies utilized mostly confirmatory statistics, modern proteomics datasets typically contain thousands of features (e.g., protein expression) over a handful of observations, hence simply testing whether each protein is significantly altered across experimental conditions can result in under-analysis and failure to distinguish latent structures across multiple dimensions, e.g., whether there exists a subproteome of co-regulated proteins across multiple treatment categories. To gain biological insights, quantitative proteomics datasets are now routinely mined using statistical learning strategies that comprise feature selection (e.g., penalized regression methods), dimensionality reduction (e.g., principal component analysis), and both supervised and unsupervised learning (e.g., support vector machine and hierarchical models) to discern significant protein signatures, disease-implicated pathways, or interconnected co-expression networks (Fig. 4).

Improvements to computational methods that allow more robust results from label-free quantification are an area of active research. For example, recent works (e.g., QProt) have attempted to resolve the respective quantities of multiple proteins that share common peptide sequences in spectral counting, either using weighted average methods or more statistically motivated models [36, 37]. In ion intensity approaches, chromatographic

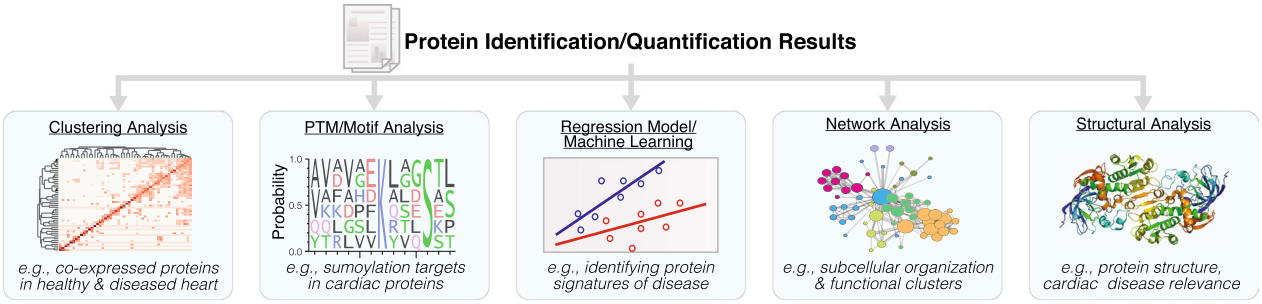
Protein Identification/Quantification Results
Clustering Analysis
e.g., co-expressed proteins in healthy & diseased heart
PTM/Motif Analysis
Probability
1.0
0.5
0.0
e.g., sumoylation targets in cardiac proteins
Regression Model/ Machine Learning
e.g., identifying protein signatures of disease
Network Analysis
e.g., subcellular organization & functional clusters
Structural Analysis
e.g., protein structure, cardiac disease relevance

Fig. 4 Proteomics data mining and functional annotations. Common computation approaches to extract information from massive proteomics datasets include (1) unsupervised cluster analysis, class discovery and visualization; (2) motif analysis and annotation term enrichment; (3) statistical learning methods for disease signature extraction; (4) network analysis; and (5) annotations with other functional information including protein motifs and cardiac disease relevance

features that correspond to peptide signals over mass- and retention time-dimensions are identified using image analysis or signal processing algorithms. Because LC gradients are seldom perfectly reproducible, nonlinear distortions in retention time may occur. To ensure identical ions are compared between experiments, automatic chromatographic alignment and clustering methods are used. Some software can identify small chromatographic features based on accurate mass/retention time alone, such that some peptides may be quantified even in experiments where they were not explicitly identified. These processes tend to become computationally expensive for large experimental files [38], thus faster solutions are continuously developed.

With the proliferation of inter-compatible tools, an ongoing trend is to daisy-chain individual tools into user-friendly pipelines that provide complete solutions to a set of related data analysis problems. An ideal proteomics pipeline may combine identification, quantification, and validation tools in a modular organization accessible from a single location. Computation may be performed on the cloud to avoid the need to repeatedly copy, transfer, and store large files. Researchers can carry out computational tasks remotely from the browser on any computer system, obviating the need for redundant infrastructure investments. Currently, the Trans-Proteomics Pipeline [39] and the Integrated Proteomics Pipeline [40] are two example “end-to-end” pipelines that connect raw MS proteomics data to analysis output, whereas comprehensive, open-access pipelines have also been demonstrated in other omics fields, including Galaxy for genomics/transcriptomics [41] and MetaboAnalyst for metabolomics [42]. In parallel, tools are also being federated into interoperable networks through open frameworks. A modular and open-source software development paradigm, where individual software functionalities can interoperate via common interfaces and standards, helps ensure that new software can dovetail with existing ones with ease, and that software development may continue following inactivity from the original research team. Examples of such frameworks include the GalaxyP proteomics extension [43], and the proteomics packages within the R/BioConductor framework [44].

Improvements in annotation resources

Not unlike other omics approaches, the success of proteomics experiments relies heavily on having adequate and up-to-date resources to analyze large-scale data. To wit, for protein database search to succeed, it follows that the protein being identified must first be documented on a sequence database. Fortunately, there have been tremendous advances on the completeness (number of true positive sequences recorded) and precision (removal of redundancy or artifacts) of sequence databases such as UniProt and RefSeq. Some databases are manually curated to contain precise information, whilst others strive to be more comprehensive; but many now list precise “complete proteomes” for commonly studied laboratory species. Databases for human and popular laboratory model organisms have seen particular progress in the completeness of annotation in recent years, such that proteomics studies can now be performed similarly well in mice, drosophila, rats, and other organisms to interrogate cardiac physiology. On the horizon, one can foresee an influx of sequence information on protein polymorphism and alternative splicing isoforms. Although current databases primarily originate from genomic translation or cDNA library of specific cell types or genetic backgrounds, “proteogenomics” efforts are underway to translate additional sequences for proteomics studies, which will expand the scope and precision of protein identification. For non-model organisms, alternative search methods have been devised, such as against a custom database generated by manual six-frame translation of genomic sequences.

Data interpretability problems arise when the complex results comprising changes of many proteins could not be easily digested and summarized in terms that are relatable and of value to biomedical and clinical researchers. The biological significance of the implicated protein targets may be interpreted and connected to the growing corpus of biomedical knowledge through functional annotations. Commonly used annotation resources include Gene Ontology for biological functions [45], Reactome or KEGG for curated biochemical and signaling pathways [46], PFAM for protein motifs and homology [47], PhosphoSitePlus for PTMs [48], and so on.

To map molecular data to curated annotations, a class discovery approach is commonly utilized, which looks for annotated properties that are preferentially shared amongst a subset of proteins with interesting quantitative features over the proteome-wide background, e.g., through sequence motif analysis or annotation term enrichment. The principle behind such analyses is to infer biologically relevant processes based on significant overlaps between data features and the data annotation classes. Computational and statistical approaches are used to determine whether particular annotations are over-represented in a particular subset results than would be expected by chance. This allows both bias-free discovery and specific questions to be asked of a dataset, e.g., whether the down-regulated proteins in a heart failure patient are preferentially involved in fatty acid metabolism, etc. Enrichment analyses can be easily carried out using online tools that perform binomial or hypergeometric tests on the over-representation of functional annotations, including NCBI DAVID, WebGestalt, and Pantherdb [49–51]. For users conversant in statistical programming and data analysis environment, open-access packages dedicated to proteomics data operation have been made available in languages such as R and Python, including the RforProteomics package [52] for MS data visualization on the R/Bioconductor repository, and the Python Pyteomics library [53] for parsing and processing MS data. These packages allow users to connect upstream MS analysis to downstream functional annotation services that are commonly employed.

We note that many proteomics functional analysis strategies were originally developed for microarray datasets. Although the analytical goals between proteomics and microarray experiments often overlap, i.e., identify functional associations from numerical molecular expression data, several methodological differences merit considerations. The stochastic nature of shotgun proteomics can lead to missing values and high variability in the data for low-abundance proteins. To address this, several proteomics-focused enrichment analysis tools have recently been developed to address specific features of proteomics datasets, e.g., using weighted sampling, which should account for relative abundance and variability of observations and further improve quantification performance. Monte Carlo sampling approaches have also been used to counter the bias of annotations on high abundance proteins when calculating enrichment significance [54].

As in the case for protein identification, the success of functional analysis is contingent upon the completeness and accuracy of annotations in knowledgebases. Commonly used knowledgebases such as Uniprot and Reactome have steadily improved in size and richness of functional information. Nevertheless, it sometimes remains the case that some classes of annotations are more complete, either because they are easily computable from sequence information (e.g., do these phosphoproteins with increased phosphorylation in heart failure share putative kinase domains?) or can be derived from popular experimental designs (e.g., whether a particular cardiac protein localizes to the mitochondrion?), and as a result are more likely to turn up in functional enrichment analyses. On the other hand, annotations on higher-level pathophysiology and tissue-specific regulations are more challenging, because a majority of such information is hidden in unstructured free text in the literature. With >2 million cardiovascular related articles alone on PubMed [55], however, the volume of literature data being published dwarfs the capacity of human expert biocurations to translate them. Hence in recent years many approaches are being pursued to improve biocuration, including crowdsourcing initiatives to leverage public contributors through web-based micro-tasks, as well as text-mining algorithms that comb through the literature and automatically convert free texts into computable formats. Domain-specific knowledgebases (including organelle specific databases [56], and cardiovascular disease specific knowledgebases [57] have also been developed to provide richer cardiovascular contexts in data interpretation.

Examples and frontiers in cardiovascular applications

The aforementioned analytical and computational advances have enabled novel and noteworthy applications in cardiovascular proteomics. An exciting trend is to venture beyond simply inventorying which proteins are present in the heart or the blood, and into quantifying their dynamic and spatiotemporal properties. Protein complexity necessitates that many parameters are needed to sufficiently describe the overall proteome in a particular physiological state. New methodologies continually arise that enable new insights into protein–protein interactions [58], protein homeostasis [59], and spatial distributions [60]. Many molecular parameters are now known to be directly involved in disease pathogenesis thanks to proteomics studies, including the PTMs of proteins, their spatiotemporal distributions, and interacting partners.

Quantifying diverse post-translational modifications

With increased experimental power to detect peptides, rare and hard-to-detect peptides are increasingly analyzable, including many modified by PTMs. Because translational modifications are attached to proteins following synthesis, their chemical identity, position, and fractional quantity cannot be easily predicted from transcripts, necessitating proteomics studies. PTMs have been the focus of proteomics studies for over a decade and these efforts have increasingly begun to bear fruit in various biomedical investigations. In a recent notable example, Lee et al. [61] used a global approach was used to discern the roles of phosphodiesterase (PDE) in cyclic guanosine monophosphate (cGMP) degradation in cardiac signaling. Although once assumed to be a common pathway acting through a single secondary messenger, the subcomponents are modulated by two different enzymes PDE5A and PDE9A at different cellular locations. A global, high-throughput phosphoproteomics profiling approach was used to differentiate the downstream signaling targets of the two pathways, allowing their precise regulations to be elucidated and subclassified. Therapeutic decision and precision medicine may be informed by targeting PDE9A versus PDE5A with different pharmacological compounds.

Evidence suggests that the current investigations into PTMs have barely scratched the surface of their complexity. Over 380,000 PTM events are documented on the PhosphoSitePlus database [48], including acetylation, di-methylation, mono-methylation, O-GlcNAcylation, phosphorylation, sumoylation and ubiquitiylation, etc. on a variety of proteins. But many additional, unknown modifications likely lurk in acquired spectra that await identification, which are the subject of ongoing developments such as using cascade search, open search, or spectral clustering approaches [62, 63]. In addition to classical studies of phosphorylation and ubiquitination, improved methods to isolate and identify PTMs have fueled investigations into many different kinds of modifications including glycosylation, acetylation, sumoylation, and oxidative modifications that are now known to play critical and indispensable roles in the regulations of core aspects of cardiac physiology. Publications from multiple groups have led to increasing appreciation of the fine regulations of oxidative cysteine modifications (including disulfide bridge, S-nitrosylation, and S-glutathionylation) in cardiac redox regulation [64]. Examples include the discovery of S-nitrosylation at TRIM72 in regulating ischemic injury [65], and the unexpected “moonlighting” of GAPDH in the mitochondria to confer S-nitrosylation in the heart [66]. An increasing number of other examined modifications are likewise now implicated in important cardiac processes, at a rate that far exceeds that which would be attainable in traditional single-target, hypothesis-driven investigations. These include for instance the role of acetylation in the context of mitochondrial metabolism [67, 68]; sumoylation in the context of heart failure [69]; O-GlcNAcylation in the context of diabetic hearts [70], and lysine succinylation in the context of ischemic injury [71]. These studies represent a broadening of our observable universe, and are driven by both advances in specific purification or labeling strategies, as well as a general increase in MS instrumentation and data analysis maturity that together propel the experimental scope, scale and reproducibility past the threshold of informativeness.

Tracing protein spatiotemporal distributions

The function and functionality of a protein are modulated to a great extent by the spatial milieu in which it is situated, which in turn dictates the substrates and interacting partners with which it comes across. Ongoing studies into the spatial distributions of cardiac proteins are inventorying the protein compositions in different cardiac organelles, with particular successes in cardiac mitochondria and nucleus proteomes, as elucidated via targeted enrichment of specific organelles. More recent studies in other organs are suggesting the possibility that proteins are actively and constantly translocalizing between organelles in response to cues, which can be measured by combining differential centrifugation, isotope labeling, and machine learning techniques. Recent advances in data analysis have allowed such differential centrifugation techniques to be used for pan-organellar mapping. Instead of enriching only for a pure sample of a particular organelle type, a centrifugation gradient here is matched to a supervised classification algorithm to classify proteins based on their sedimentation behaviors. The average intracellular position of many proteins can therefore be discovered by their grouping with known organellar markers. These approaches can be readily adopted to understand dynamic protein translocalization from one organelle to another, using new analytical frameworks that can quantify protein translocation in differential centrifugation experiments [60, 72]. Through proteomics studies, it is also demonstrated that many proteins important in the heart can have multiple localizational isoforms that carry out different functions [73].

Secondly, the synthesis rates of proteins have also proven important to tracing the progression of cardiac hypertrophy preceding heart failure. Interests in classical physiology to understand skeletal and cardiac muscle mass gain have propelled technological and software developments aimed at understanding protein turnover, which can be applied to other fields. There have been particular developments in isotope labeling and kinetic modeling methods, which have elucidated the half-life of proteins in many various cellular compartments in the heart [74–77]. Developments in data analysis methods are particular important in this area, as although stable isotope administration and mass spectrometry approaches have been in use for decades, MS data measuring the rate of isotope incorporation cannot be efficiently analyzed on a very large scale without specialized software that can deconvolute isotope patterns and fit massive datasets to kinetic models [78, 79]. Following these advances, in vivo studies are revealing a previously unknown regulatory layer and architecture of the proteome in which functionally associated proteins share more synchronous turnover rates. During disease development, protein pathways have been found to deviate from physiological baseline via elevated protein replacement but not any apparent change in steady-state abundance, a result consistent with increased protein synthesis counterpoised by increased degradation [79]. Hence the measurement of half-life is also pursued to identify candidate disease proteins and from which to infer significantly dysregulated biological processes during pathogenesis.

Mapping protein–protein interaction networks

There have been marked improvements in experimental protocols in affinity purification, as well as statistical and data science methods to filter out false positives in pull-down experiments. Chiang et al. recently elucidated the interactome of the protein phosphtase 1 catalytic subunit (PP1c), identifying 78 interacting partners in human heart. The proteomics results found increased binding to PDE5A in paroxysmal atrial fibrillation patients to impair proteins involved in electrical and calcium remodeling, a result that has implications in the understanding and treatment of atrial fibrillation [80]. Waldron et al. identified the TBX5 interactome in the developing heart to discover its interactions with the repressor complex NuRD, elucidating the mechanisms by which TBX5 mutations can influence cardiac development and confer congenital heart diseases. The accretion of public-domain protein–protein interactome data are also serving as a permanent resource that benefits other investigators outside the proteomics field, and in one but many recent examples provided important context to systems genetics experimental data to evidence the involvement of an interacting cilia protein network in congenital heart diseases [81]. More recently, the CoPIT method extends the scope of comparison to degrees of interactions among samples across cell states with more rigorous statistics, and is particularly notable in its suitability for quantifying differential interactomes of membrane proteins in human diseases [47]. Potential protein–protein interactions can now also be predicted in silico and de novo using machine learning algorithms that take in experimental data and auxiliary information [82].

At the same time, there is renewed interest to perform crosslinker studies on a large scale, which in addition to identifying protein–protein interaction partners, can provide information on the topology and protein domains involved in the interactions. Again we note that the development of new proteomics methodologies now necessitates hand-in-hand advances of novel data science solutions almost without exception. An example is the application of chemical cross-linkers in proteomics, which allows the linking of proximal proteins to quantify the degrees and likelihood of protein–protein interactions in their native cellular environment. Cross-linking proteomics experiments are however infeasible without specialized search engines that can consider the combinatorics of crosslinked peptide sequences, and identify interacting proteins whilst controlling for the FDR that result from the quadratic increase of search space [83, 84].

Outlooks: emergence of proteomics big data

Quantitative shotgun proteomics has developed into a remarkably powerful technology that enables sophisticated questions on cellular physiology to be asked. The total volume of proteomics data generated per year now ranges in the petabytes. This is paralleled by an increasing number of available proteomics datasets in the public domain that can be reused and reanalyzed, with as many as 100 new datasets being made available per month on the proteomics data repository PRIDE [85]. Hence joining next-generation genomics, proteomics has become a veritable source of biomedical "big data". As our capacity for data generation surges, opportunities for breakthroughs will increasingly come from not how much more data we can generate, but how well we can make sense of the results. As a corollary, the need for proteomics big data solutions is poised to skyrocket in the coming few years, where new resources, tools, and ways of doing science are needed to rethink how best to harness datasets and discern deeper meanings. The production of biological knowledge will involve tools and solutions devised in the field of data science, including those concerning data management, multivariate analysis, statistical learning, predictive modeling, software engineering, and crowdsourcing. Several current limitations and possible future frontiers, out of many, are discussed below:

Despite impressive gains, improvement of protein identification will likely continue to be an area of active research. It is estimated that up to 75–85 % of mass spectra generated in a proteomics experiment can remain unidentified by current data analysis workflows [62, 86], thus leaving room for continuous growth through better bioinformatics in the near future. Currently the unidentified "junk" spectra are mostly siloed or discarded, thus they constitute a major untapped source of biomedical big data. More inclusive search criteria (e.g., considering non-tryptic cleavage) can enhance identification, but there also exists a substantial portion of spectra that represent bona fide peptides not amenable to existing methods. These include peptides too short to score well in searching algorithms (≤5 amino acids), and peptides that are absent from protein sequence databases, e.g., variant peptides from polymorphisms or unknown splice isoforms. The advent of massive publicly available datasets has opened new avenues to tackle this problem [62, 63]. For instance, the millions of unidentified spectra that are uploaded to the PRIDE proteomics data repository may be systematically sorted and clustered, then analyzed via more exhaustive search protocols to identify what peptides are commonly present but unidentified across datasets and experiments.

Secondly, advances in protein quantification techniques, via both experimental and computational developments, will likely continue unabated. Many quantification techniques do not take into account the peptides that may become post-translationally modified or otherwise lost in a biological state. Currently, decreases in label-free measured quantity may be confounded by differences in protein modifications, digestion, or ionization, or matrix effects from different samples. For instance, the acquired spectral counts may be inflated by the existence of shared peptides among multiple (documented or undocumented) protein forms [87] as well as the sampling saturation for high-abundance peptides [88]. Statistical approaches pioneered in transcriptomics may be useful which can take into account the many-to-many mappings between proteins and peptides and to reconstruct proteoforms from individual peptide signals.

Lastly, the identification of unknown or unspecified PTMs will likely see continued progress. Because the multiplicity of the possible modification types on a peptide can shift peptide fragment masses combinatorially, they can greatly inflate the number of possible matches. Efforts are underway to develop custom sequence databases and devise new algorithms to extract information from existing cardiovascular proteomics datasets that is currently "hiding in plain sight". For instance, algorithms can be used to predict peptide fragment intensity to improve peptide identification [89]. To make PTM search computationally tractable, multi-pass or cascade search approaches have been implemented that restrict the possibility of modified peptides to only within proteins that were preliminarily identified in the initial search. To improve peptide identification, "spectral libraries" including library modules for the cardiovascular system have been constructed that contain previously identified spectra, against which new experimental spectra can be directly matched for identification [57, 90, 91]. Because spectral libraries contain only a small subset of all theoretical sequences, and contain precise peptide fragment intensity in addition to ion masses, library search can lead to faster and more accurate identification.

To summarize, we recall an apt analogy provided by Loscalzo to compare the understanding of cardiac proteome with that of building a house and the genome with that of its floor plan [92]. Genomics kick-started the era of high-throughput omics investigations, but building a house requires more than just the blueprint; the complexity of protein regulation and pathway functions is better approached with proteomics. Technological advances in the last decade have gained tremendous power to discover finer minutiae of the house of the cardiac proteome. Success in the next 5 years will likely come from interfacing big proteomics data and computational approaches to distill regulatory principles and to support diagnostic/prognostic process from seemingly overwhelming information. The notion that existing data contain additional latent information that may be extracted to answer future questions is a fundamental tenet of big data science. Extrapolating from current developments, one can envision sophisticated discovery-driven studies in cardiac biomedicine, where original research projects may be initiated by data scientists using publicly-accessible proteomics datasets to ask new and unanticipated questions. A vibrant data science culture that promotes interactions between data generators and informaticians will facilitate the design and validation of computational methods and promote continued development in proteomics.

Abbreviations

MS: mass spectrometry; LC: liquid chromatography; PTM: post-translational modification; MRM: multiple reaction monitoring; PDE: phosphodiesterase; cGMP: cyclic guanosine monophosphate; FDR: false discovery rates; 2D: two-dimensional; SAF: spectral-abundance-factor.

Authors' contributions

All authors contributed to the writing and conceptualization of the manuscript. All authors read and approved the final manuscript.

Author details

[1] NIH BD2K Center of Excellence at UCLA; Department of Physiology, University of California at Los Angeles, 675 Charles E. Young Drive, Los Angeles, CA 90095, USA. [2] Department of Medicine, University of California at Los Angeles, 675 Charles E. Young Drive, Los Angeles, CA 90095, USA. [3] Department of Bioinformatics, University of California at Los Angeles, 675 Charles E. Young Drive, Los Angeles, CA 90095, USA.

Acknowledgements

None.

Competing interests

The authors declare no financial and non-financial competing interests.

Funding

This work was supported by National Institutes of Health Grants: NIH U54GM114833, NIH R01HL129723 to Dr. Peipei Ping and NIH K99HL127302 to Dr. Maggie Lam.

Received: 3 April 2016 Accepted: 24 August 2016
Published online: 05 December 2016

References

1. Glatter T, Ludwig C, Ahrne E, Aebersold R, Heck AJ, Schmidt A. Large-scale quantitative assessment of different in-solution protein digestion protocols reveals superior cleavage efficiency of tandem Lys-C/trypsin proteolysis over trypsin digestion. J Proteome Res. 2012;11(11):5145–56.
2. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ. The one hour yeast proteome. Mol Cell Proteomics. 2014;13(1):339–47.
3. Lindsey ML, Mayr M, Gomes AV, Delles C, Arrell DK, Murphy AM, Lange RA, Costello CE, Jin Y-F, Laskowitz DT, et al. Transformative impact of proteomics on cardiovascular health and disease: a scientific statement from the American Heart Association. Circulation. 2015;132(9):852–72.
4. Van Eyk JE. Overview: the maturing of proteomics in cardiovascular research. Circ Res. 2011;108(4):490–8.
5. McGregor E, Dunn MJ. Proteomics of the heart: unraveling disease. Circ Res. 2006;98(3):309–21.
6. Weiss JN, Karma A, MacLellan WR, Deng M, Rau CD, Rees CM, Wang J, Wisniewski N, Eskin E, Horvath S, et al. "Good enough solutions" and the genetics of complex diseases. Circ Res. 2012;111(4):493–504.
7. Lam MP, Venkatraman V, Cao Q, Wang D, Dincer TU, Lau E, Su AI, Xing Y, Ge J, Ping P, et al. Prioritizing proteomics assay development for clinical translation. J Am Coll Cardiol. 2015;66(2):202–4.
8. Draizen EJ, Shaytan AK, Marino-Ramirez L, Talbert PB, Landsman D, Panchenko AR. HistoneDB 2.0: a histone database with variants–an integrated resource to explore histones and their variants. Database. 2016;2016. pii: baw014. doi:10.1093/database/baw014.
9. Waas M, Bhattacharya S, Chuppa S, Wu X, Jensen DR, Omasits U, Wollscheid B, Volkman BF, Noon KR, Gundry RL. Combine and conquer: surfactants, solvents, and chaotropes for robust mass spectrometry based analyses of membrane proteins. Anal Chem. 2014;86(3):1551–9.
10. Fonslow BR, Stein BD, Webb KJ, Xu T, Choi J, Park SK, Yates JR. Digestion and depletion of abundant proteins improves proteomic coverage. Nat Methods. 2013;10(1):54–6.
11. Lotz C, Lin AJ, Black CM, Zhang J, Lau E, Deng N, Wang Y, Zong NC, Choi JH, Xu T, et al. Characterization, design, and function of the mitochondrial proteome: from organs to organisms. J Proteome Res. 2014;13(2):433–46.
12. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001;19:242–7.
13. Webb KJ, Xu T, Park SK, Yates JR III. Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast. J Proteome Res. 2013;12(5):2177–84.
14. Lau E, Lam MP, Siu SO, Kong RP, Chan WL, Zhou Z, Huang J, Lo C, Chu IK. Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications. Mol BioSyst. 2011;7:1399–408.
15. Shen Y, Jacobs JM, Camp DG II, Fang R, Moore RJ, Smith RD, Xiao W, Davis RW, Tompkins RG. Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome. Anal Chem. 2004;76(4):1134–44.
16. Puangpila C, Mayadunne E, El Rassi Z. Liquid phase based separation systems for depletion, prefractionation, and enrichment of proteins in biological fluids and matrices for in-depth proteomics analysis—an update covering the period 2011–2014. Electrophoresis. 2015;36(1):238–52.
17. Huang J, Wang F, Ye M, Zou H. Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications. J Chromatogr A. 2014;1372C:1–17.
18. Lau E, Cao Q, Ng DC, Bleakley BJ, Dincer TU, Bot BM, Wang D, Liem DA, Lam MP, Ge J, et al. A large dataset of protein dynamics in the mammalian heart proteome. Sci Data. 2016;3:160015.
19. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–81.
20. Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–7.
21. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012;9(6):555–66.
22. Zhang P, Kirk JA, Ji W, dos Remedios CG, Kass DA, Van Eyk JE, Murphy AM. Multiple reaction monitoring to identify site-specific troponin I phosphorylated residues in the failing human heart. Circulation. 2012;126(15):1828–37.
23. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative proteomics of human tumor tissue. Nat Methods. 2010;7(5):383–5.
24. Geiger T, Wisniewski JR, Cox J, Zanivan S, Kruger M, Ishihama Y, Mann M. Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. Nat Protoc. 2011;6(2):147–57.
25. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 1994;5(11):976–89.
26. Xu T, Park SK, Venable JD, Wohlschlegel JA, Diedrich JK, Cociorva D, Lu B, Liao L, Hewel J, Han X, et al. ProLuCID: an improved SEQUEST-like algorithm with enhanced sensitivity and specificity. J Proteomics. 2015;129:16–24.
27. Yates JR. Pivotal role of computers and software in mass spectrometry—SEQUEST and 20 years of tandem MS database searching. J Am Soc Mass Spectrom. 2015;26(11):1804–13.
28. Tabb DL. The SEQUEST family tree. J Am Soc Mass Spectrom. 2015;26(11):1814–9.
29. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013;13(1):22–4.
30. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011;10(4):1794–805.
31. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014;5:5277.
32. Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM. MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. J Proteome Res. 2011;10(7):2949–58.
33. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics. 2010;73(11):2092–123.
34. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4(11):923–5.
35. Zhang Y, Xu T, Shan B, Hart J, Aslanian A, Han X, Zong N, Li H, Choi H, Wang D, et al. ProteinInferencer: Confident protein identification and multiple experiment comparison for large scale proteomics projects. J Proteomics. 2015;129:25–32.

36. Gerster S, Kwon T, Ludwig C, Matondo M, Vogel C, Marcotte EM, Aebersold R, Bühlmann P. Statistical approach to protein quantification. Mol Cell Proteomics. 2014;13(2):666–77.
37. Choi H, Kim S, Fermin D, Tsou CC, Nesvizhskii AI. QPROT: statistical method for testing differential expression using protein-level intensity data in label-free quantitative proteomics. J Proteomics. 2015;129:121–6.
38. Sandin M, Teleman J, Malmström J, Levander F. Data processing methods and quality control strategies for label-free LC–MS protein quantification. Biochim Biophys Acta. 2014;1844:29–41.
39. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. Trans-proteomic pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. Proteomics Clin Appl. 2015;9(7–8):745–54.
40. Lavallee-Adam M, Park SK, Martinez-Bartolome S, He L, Yates JR III. From raw data to biological discoveries: a computational analysis pipeline for mass spectrometry-based proteomics. J Am Soc Mass Spectrom. 2015;26(11):1820–6.
41. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86.
42. Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0—making metabolomics more meaningful. Nucl Acids Res. 2015;43(W1):W251–7.
43. Sheynkman GM, Johnson JE, Jagtap PD, Shortreed MR, Onsongo G, Frey BL, Griffin TJ, Smith LM. Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genom. 2014;15:703.
44. Gatto L, Breckels LM, Naake T, Gibb S. Visualization of proteomics data using R and bioconductor. Proteomics. 2015;15(8):1375–89.
45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25(1):25–9.
46. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucl Acids Res. 2016;44(D1):D457–62.
47. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The Pfam protein families database: towards a more sustainable future. Nucl Acids Res. 2016;44(D1):D279–85.
48. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics. 2004;4(6):1551–61.
49. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucl Acids Res. 2013;41:D377–86.
50. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:44–57.
51. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucl Acids Res. 2013;41:W77–83.
52. Gatto L, Christoforou A. Using R and bioconductor for proteomics data analysis. Biochim Biophys Acta. 2014;1844:42–51.
53. Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. J Am Soc Mass Spectrom. 2013;24(2):301–4.
54. Lavallee-Adam M, Rauniyar N, McClatchy DB, Yates JR III. PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. J Proteome Res. 2014;13(12):5496–509.
55. Lau E, Watson KE, Ping P. Connecting the dots: from big data to healthy heart. Circulation. 2016;134(5):362–4.
56. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong SE, Walford GA, Sugiana C, Boneh A, Chen WK, et al. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008;134(1):112–23.
57. Zong NC, Li H, Li H, Lam MPY, Jimenez RC, Kim CS, Deng N, Kim AK, Choi JH, Zelaya I, et al. Integration of cardiac proteome biology and medicine by a specialized knowledgebase. Circ Res. 2013;113(9):1043–53.
58. Pankow S, Bamberger C, Calzolari D, Martínez-Bartolomé S, Lavallée-Adam M, Balch WE, Yates JR. ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis. Nature. 2015;528(7583):510–6.
59. Kim TY, Wang D, Kim AK, Lau E, Lin AJ, Liem DA, Zhang J, Zong NC, Lam MP, Ping P. Metabolic labeling reveals proteome dynamics of mouse mitochondria. Mol Cell Proteomics. 2012;11:1586–94.
60. Christoforou A, Mulvey CM, Breckels LM, Geladaki A, Hurrell T, Hayward PC, Naake T, Gatto L, Viner R, Arias AM, et al. A draft map of the mouse pluripotent stem cell spatial proteome. Nat Commun. 2016;7:9992.
61. Lee DI, Zhu G, Sasaki T, Cho G-S, Hamdani N, Holewinski R, Jo S-H, Danner T, Zhang M, Rainer PP, et al. Phosphodiesterase 9A controls nitric-oxide-independent cGMP and hypertrophic heart disease. Nature. 2015;519:472–6.
62. Griss J, Perez-Riverol Y, Lewis S, Tabb DL, Dianes JA, Del-Toro N, Rurik M, Walzer MW, Kohlbacher O, Hermjakob H, et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. Nat Methods. 2016;13(8):651–6.
63. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, Gygi SP. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol. 2015;33(7):743–9.
64. Sverdlov AL, Elezaby A, Behring JB, Bachschmid MM, Luptak I, Tu VH, Siwik DA, Miller EJ, Liesa M, Shirihai OS, et al. High fat, high sucrose diet causes cardiac mitochondrial dysfunction due in part to oxidative post-translational modification of mitochondrial complex II. J Mol Cell Cardiol. 2015;78:165–73.
65. Kohr MJ, Evangelista AM, Ferlito M, Steenbergen C, Murphy E. S-nitrosylation of TRIM72 at cysteine 144 is critical for protection against oxidation-induced protein degradation and cell death. J Mol Cell Cardiol. 2014;69:67–74.
66. Kohr MJ, Murphy E, Steenbergen C. Glyceraldehyde-3-phosphate dehydrogenase acts as a mitochondrial trans-S-nitrosylase in the heart. PLoS One. 2014;9(10):e111448.
67. Lam MPY, Lau E, Liem DA, Ping P. Cyclophilin D and acetylation: a new link in cardiac signaling. Circ Res. 2013;113(12):1268–9.
68. Nguyen TTM, Wong R, Menazza S, Sun J, Chen Y, Wang G, Gucek M, Steenbergen C, Sack MN, Murphy E. Cyclophilin D modulates mitochondrial acetylome. Circ Res. 2013;113(12):1308–19.
69. Kho C, Lee A, Jeong D, Oh JG, Chaanine AH, Kizana E, Park WJ, Hajjar RJ. SUMO1-dependent modulation of SERCA2a in heart failure. Nature. 2011;477(7366):601–5.
70. Ramirez-Correa GA, Ma J, Slawson C, Zeidan Q, Lugo-Fagundo NS, Xu M, Shen X, Gao WD, Caceres V, Chakir K, et al. Removal of abnormal myofilament *O*-GlcNAcylation restores Ca^{2+} sensitivity in diabetic cardiac muscle. Diabetes. 2015;64(10):3573–87.
71. Boylston JA, Sun J, Chen Y, Gucek M, Sack MN, Murphy E. Characterization of the cardiac succinylome and its role in ischemia-reperfusion injury. J Mol Cell Cardiol. 2015;88:73–81.
72. Gatto L, Breckels LM, Burger T, Nightingale DJH, Groen AJ, Campbell C, Nikolovski N, Mulvey CM, Christoforou A, Ferro M, et al. A foundation for reliable spatial proteomics data analysis. Mol Cell Proteomics. 2014;13(8):1937–52.
73. Monaghan RM, Whitmarsh AJ. Mitochondrial proteins moonlighting in the nucleus. Trends Biochem Sci. 2015;40(12):728–35.
74. Dai D-F, Karunadharma PP, Chiao YA, Basisty N, Crispin D, Hsieh EJ, Chen T, Gu H, Djukovic D, Raftery D, et al. Altered proteome turnover and remodeling by short-term caloric restriction or rapamycin rejuvenate the aging heart. Aging Cell. 2014;13(3):529–39.
75. Kim T-Y, Wang D, Kim AK, Lau E, Lin AJ, Liem DA, Zhang J, Zong NC, Lam MPY, Ping P. Metabolic labeling reveals proteome dynamics of mouse mitochondria. Mol Cell Proteomics. 2012;11(12):1586–94.
76. Shekar KC, Li L, Dabkowski ER, Xu W, Ribeiro RF, Hecker PA, Recchia FA, Sadygov RG, Willard B, Kasumov T, et al. Cardiac mitochondrial proteome dynamics with heavy water reveals stable rate of mitochondrial protein synthesis in heart failure despite decline in mitochondrial oxidative capacity. J Mol Cell Cardiol. 2014;75:88–97.
77. XaCY Chan, Black CM, Lin AJ, Ping P, Lau E. Mitochondrial protein turnover: methods to measure turnover rates on a large scale. J Mol Cell Cardiol. 2015;78:54–61.
78. Hsieh EJ, Shulman NJ, Dai D-F, Vincow ES, Karunadharma PP, Pallanck L, Rabinovitch PS, MacCoss MJ. Topograph, a software platform for precursor enrichment corrected global protein turnover measurements. Mol Cell Proteomics. 2012;11(11):1468–74.

79. Lam MPY, Wang D, Lau E, Liem DA, Kim AK, Ng DCM, Liang X, Bleakley BJ, Liu C, Tabaraki JD, et al. Protein kinetic signatures of the remodeling heart following isoproterenol stimulation. J Clin Investig. 2014;124(4):1734–44.
80. Chiang DY, Lebesgue N, Beavers DL, Alsina KM, Damen JMA, Voigt N, Dobrev D, Wehrens XHT, Scholten A. Alterations in the interactome of serine/threonine protein phosphatase type-1 in atrial fibrillation patients. J Am Coll Cardiol. 2015;65(2):163–73.
81. Li Y, Klena NT, Gabriel GC, Liu X, Kim AJ, Lemke K, Chen Y, Chatterjee B, Devine W, Damerla RR, et al. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature. 2015;521(7553):520–4.
82. Mellacheruvu D, Wright Z, Couzens AL, Lambert JP, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY, et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat Methods. 2013;10:730–6.
83. Rinner O, Seebacher J, Walzthoeni T, Mueller LN, Beck M, Schmidt A, Mueller M, Aebersold R. Identification of cross-linked peptides from large sequence databases. Nat Methods. 2008;5(4):315–8.
84. Wang J, Anania VG, Knott J, Rush J, Lill JR, Bourne PE, Bandeira N. Combinatorial approach for large-scale identification of linked peptides from tandem mass spectrometry spectra. Mol Cell Proteomics. 2014;13(4):1128–36.
85. ProteomeXchange. http://www.proteomexchange.org. Accessed 10 Aug 2016.
86. Frank AM, Monroe ME, Shah AR, Carver JJ, Bandeira N, Moore RJ, Anderson GA, Smith RD, Pevzner PA. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat Methods. 2011;8(7):587–91.
87. Lundgren DH, Hwang SI, Wu L, Han DK. Role of spectral counting in quantitative proteomics. Expert Rev Proteomics. 2010;7(1):39–53.
88. Arike L, Peil L. Spectral counting label-free proteomics. Methods Mol Biol. 2014;1156:213–22.
89. Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics. 2008;24(13):i348–56.
90. Li H, Zong NC, Liang X, Kim AK, Choi JH, Deng N, Zelaya I, Lam M, Duan H, Ping P. A novel spectral library workflow to enhance protein identifications. J Proteomics. 2013;81:173–84.
91. Hu Y, Lam H. Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications. J Proteome Res. 2013;12(12):5971–7.
92. Loscalzo J. Proteomics in cardiovascular biology and medicine. Circulation. 2003;108(4):380–3.
93. Eng JK, Hoopmann MR, Jahan TA, Egertson JD, Noble WS, MacCoss MJ. A deeper look into Comet-implementation and features. J Am Soc Mass Spectrom. 2015;26(11):1865–74.
94. Comet. http://comet-ms.sourceforge.net. Accessed 10 Aug 2016.
95. Software Tools Developed at CCMS. http://proteomics.ucsd.edu/software-tools/. Accessed 10 Aug 2016.
96. Dorfer V, Pichler P, Stranzl T, Stadlmann J, Taus T, Winkler S, Mechtler K. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. J Proteome Res. 2014;13(8):3679–84.
97. Mass spectrometry, Protein Chemistry Facility. http://ms.imp.ac.at/?goto=msamanda. Accessed 10 Aug 2016.
98. Yates Laboratory. http://fields.scripps.edu. Accessed 10 Aug 2016.
99. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004;20(9):1466–7.
100. X! TANDEM Spectrum Modeler. http://www.thegpm.org/tandem/. Accessed 10 Aug 2016.
101. Egertson JD, MacLean B, Johnson R, Xuan Y, MacCoss MJ. Multiplexed peptide analysis using data-independent acquisition and Skyline. Nat Protoc. 2015;10(6):887–903.
102. Skyline. http://skyline.gs.washington.edu. Accessed 10 Aug 2016.
103. Rost HL, Rosenberger G, Navarro P, Gillet L, Miladinovic SM, Schubert OT, Wolski W, Collins BC, Malmstrom J, Malmstrom L, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nat Biotechnol. 2014;32(3):219–23.
104. OpenSWATH. http://openswath.org. Accessed 10 Aug 2016.
105. CRUX. http://cruxtoolkit.sourceforge.net. Accessed 10 Aug 2016.
106. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem. 2003;75:4646–58.
107. ProteinProphet. http://proteinprophet.sourceforge.net. Accessed 10 Aug 2016.
108. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72.
109. MaxQuant. http://maxquant.org. Accessed 10 Aug 2016.
110. Park SKR, Aslanian A, McClatchy DB, Han X, Shah H, Singh M, Rauniyar N, Moresco JJ, Pinto AFM, Diedrich JK, et al. Census 2: isobaric labeling data analysis. Bioinformatics. 2014;30(15):2208–9.
111. Pavelka N, Pelizzola M, Vizzardelli C, Capozzoli M, Splendiani A, Granucci F, Ricciardi-Castagnoli P. A power law global error model for the identification of differentially expressed genes in microarray data. BMC Bioinform. 2004;5:203.
112. Bioconductor. http://bioconductor.org/packages/plgem/. Accessed 10 Aug 2016.
113. Alexey Newvizhskii Lab. http://nesvilab.org/software.html. Accessed 10 Aug 2016.
114. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. The perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016. doi:10.1038/nmeth.3901.
115. Documentation from the Cox Group. http://coxdocs.org. Accessed 10 Aug 2016.
116. McIlwain S, Tamura K, Kertesz-Farkas A, Grant CE, Diament B, Frewen B, Howbert JJ, Hoopmann MR, Kall L, Eng JK, et al. Crux: rapid open source protein tandem mass spectrometry analysis. J Proteome Res. 2014;13(10):4488–91.
117. Sturm M, Bertsch A, Gropl C, Hildebrandt A, Hussong R, Lange E, Pfeifer N, Schulz-Trieglaff O, Zerck A, Reinert K, et al. OpenMS—an open-source software framework for mass spectrometry. BMC Bioinform. 2008;9:163.
118. OpenMS. http://www.openms.de/. Accessed 10 Aug 2016.
119. Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N, Sun Z, Nilsson E, Pratt B, Prazen B, et al. A guided tour of the trans-proteomic pipeline. Proteomics. 2010;10:1150–9.
120. Seattle Proteome Center (SPC)—Proteomics Tools. http://tools.proteomecenter.org. Accessed 10 Aug 2016.
121. Vaudel M, Burkhart JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, Martens L, Barsnes H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol. 2015;33(1):22–4.
122. Compomics. http://compomics.github.io. Accessed 10 Aug 2016.
123. Perez-Riverol Y, Xu QW, Wang R, Uszkoreit J, Griss J, Sanchez A, Reisinger F, Csordas A, Ternent T, Del-Toro N, et al. PRIDE inspector toolsuite: moving toward a universal visualization tool for proteomics data standard formats and quality assessment of ProteomeXchange datasets. Mol Cell Proteomics. 2016;15(1):305–17.
124. PRIDE. http://www.ebi.ac.uk/pride/archive/. Accessed 10 Aug 2016.