Published in final edited form as:

J Proteome Res. 2014 January 3; 13(1): 107–113. doi:10.1021/pr400884c.

MOPED enables discoveries through consistently processed proteomics data

Roger Higdon*,

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab
Seattle Children's Hospital, Predictive Analytics

Elizabeth Stewart,

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab; Data
Enable Life Sciences Alliance (DELSA Global)

Larissa Stanberry,

Seattle Children's Hospital, Predictive Analytics; Seattle Children's Research Institute,
Bioinformatics and High-throughput Analysis Lab

Winston Haynes,

Stanford University, Biomedical Informatics

John Choiniere,

John; Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab

Elizabeth Montague,

Seattle Children's Hospital, Predictive Analytics; Seattle Children's Research Institute,
Bioinformatics and High-throughput Analysis Lab

Nathaniel Anderson,

Seattle Children's Hospital, Predictive Analytics; Seattle Children's Research Institute,
Bioinformatics and High-throughput Analysis Lab

Gregory Yandl,

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab
Seattle Children's Hospital, Predictive Analytics

Imre Janko,

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab

William Broomall,

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab

Simon Fishilevich,

Weizmann Institute of Science, Molecular Genetics

Doron Lancet,

Weizmann Institute of Science, Molecular Genetics

Natali Kolker, and

*Corresponding Author.

Seattle Children's Research Institute, Bioinformatics and High-throughput Analysis Lab

Eugene Kolker

Seattle Children's Hospital, Predictive Analytics
Seattle Children's Research Institute,
Bioinformatics and High-throughput Analysis Lab
Data Enable Life Sciences Alliance (DELSA
Global)
University of Washington, Biomedical Informatics and Medical Education

Abstract

The Model Organism Protein Expression Database (MOPED, <http://moped.proteinspire.org>), is an expanding proteomics resource to enable biological and biomedical discoveries. MOPED aggregates simple, standardized and consistently processed summaries of protein expression and metadata from proteomics (mass spectrometry) experiments from human and model organisms (mouse, worm and yeast). The latest version of MOPED adds new estimates of protein abundance and concentration, as well as relative (differential) expression data. MOPED provides a new updated query interface that allows users to explore information by organism, tissue, localization, condition, experiment, or keyword. MOPED supports the Human Proteome Project's efforts to generate chromosome and diseases specific proteomes by providing links from proteins to chromosome and disease information, as well as many complementary resources. MOPED supports a new omics metadata checklist in order to harmonize data integration, analysis and use. MOPED's development is driven by the user community, which spans 90 countries guiding future development that will transform MOPED into a multi-omics resource. MOPED encourages users to submit data in a simple format. They can use the metadata a checklist generate a data publication for this submission. As a result, MOPED will provide even greater insights into complex biological processes and systems and enable deeper and more comprehensive biological and biomedical discoveries.

Introduction

Researchers in the life sciences have, in recent years, produced truly enormous amounts of data with the volume increasing at an exponential rate [1–3]. Unfortunately, despite a desire and willingness to share these data amongst the community, sharing and integration efforts are hampered by a number of issues. These include varied and often incompatible formats, widely dispersed resources and repositories, and myriads of different tools used to process and analyze the data. Even more data are locked away in inaccessible hard drives or published without the appropriate metadata to make them useful. Obviously, this makes the data extremely difficult to use effectively. These circumstances hold true for high throughput proteomics despite the availability of many resources and public repositories [4–8]. In an effort to address these concerns in proteomics, the Model Organism Protein Expression Database (MOPED, <http://moped.proteinspire.org>) was introduced in 2011. It is an expanding proteomics resource that serves to aggregate, standardize, simplify, and make easily accessible mass spectrometry proteomics data and metadata for researchers [9]. MOPED provides protein expression data, meta-analysis capabilities and standardized analysis of raw data within the context of external protein and pathway databases. MOPED uses data that is consistently processed using statistical models, normalization methods and experimental standards developed in house and by the community at large. Users can query

this data with keywords or protein IDs and browse based on organism, tissue, localization and condition (Figure 1).

MOPED was inspired by the research community's feedback, gathered through a survey conducted by University of Washington business students working with the Kolker Lab in 2011. The majority of respondents asked for a complementary resource to already available data repositories. MOPED was developed in response to these stated needs. Its feature development is driven by users, whose engagement is facilitated through DELSA, the Data-Enabled Life Sciences Alliance (delsaglobal.org) [10,11]. DELSA has provided an avenue for ideas and feedback as the alliance of multidisciplinary experts focuses on translation of the data influx into tangible innovations and discoveries in life sciences.

Vigorous community engagement is necessary to capitalize on the exciting data opportunities available to the research community. The Human Proteome Project (HPP, hpp.org) is a project ideally suited to give feedback on and provide information for MOPED [12]. Its work to characterize all proteins originating from the 20,300 known protein coding genes in the human genome clearly connects with the protein expression information researchers can access through MOPED. Two main programs are currently being pursued by the HPP: the Chromosome Centric Human Proteome Project (C-HPP) and the Biology and Disease Driven Human Proteome Project (B/D-HPP) [13–15]. These projects demonstrate how useful different aspects of the same data are when integrated with different types of knowledge. In addition, MOPED complements the existing chromosome specific and chromosome spanning resources developed for the HPP [16–26]. MOPED achieves this through its summaries of large amounts of experimental data on proteins spanning all human chromosomes and by providing a resource for these projects to share their unique data.

Since the 2012 release, MOPED has added many new features and data including a completely redesigned query interface, new measures of protein concentration and abundance, relative expression data, new visualization tools, links to chromosome and disease information, and a meta-data checklist for omics data.

MOPED Resource

MOPED provides concise summaries of absolute expression incorporating newly implemented measures of concentration and abundance from experiments conducted using tissues and samples from human and three model organisms: mouse, worm and yeast. MOPED now provides summaries of relative expression experiments and metadata derived from a newly developed meta-data checklist. The summaries in MOPED are based on a standardized analysis of mass spectrometry proteomics data from public repositories and collaborators using our proteomics analysis pipeline, SPIRE (Systematic Protein Investigative Research Environment) [27]. SPIRE (proteinspire.org) integrates the best search tools and statistical models into a proteomics research pipeline, utilizing such open source search and data analysis methods as X!Tandem, OMSSA and IPM [28–30]. SPIRE also incorporates the experimental design and employs novel methods to identify proteins, produce accurate error estimations and normalize expression data [31–34]. Experimental standards are employed to aid in normalization and to validate the methods employed by

MOPED [35, 36]. MOPED supports querying, browsing, and data visualization across organisms, tissues, conditions, and pathways (Figure. 2). In 2012, the proteomics prototype MOPED had 17,000 unique users. For the first 7 months of 2013, this number grew to over 22,000 unique users from over 90 countries. Users can link to external resources, including Entrez, GeneCards, UniProt, KEGG, Reactome, Metacyc and Reactome [37–42] and have access to absolute and relative expression for 50,000 proteins based on 20 million high certainty spectra. MOPED is regularly updated and enhanced, and is steadily gaining recognition from the scientific community (e.g. [43, 44]).

Data Query Interface

At the heart of MOPED is a data query interface (Figure 1) allowing the user access to absolute and relative protein expression data. The query interface allows users to query both absolute and relative expression data by, protein identifiers or names, or by keyword search related to function, disease, pathways, and chromosome. Users may also perform an advanced search by organism, tissue or cell type, cellular localization, condition, or specific experiment using pull down menus. MOPED also enables search by a combination of key words. The interface also features new visualization tools such as the chord plot that breaks down proteins in experiments by organism, tissues, localization and condition (Figure 1). In addition, users can submit feature requests, report bugs, and obtain answers to their questions directly through the interface. Emails submitted through the website are sent directly to the MOPED team, and are generally answered within 24 hours. Surveys to solicit additional user feedback about MOPED are continuing through Survey Monkey®.

Absolute Expression

MOPED generates absolute expression measures and estimates of concentration for each protein in each experiment based on spectral counts (number of peptide spectrum matches per protein). It has been well documented that spectral counts are correlated with concentration [35, 36, 45, 46]. Approaches such as emPAI and APEX have been used to estimate absolute expression in shotgun proteomics experiments [47,48]. Since probabilities of identifying peptides are dependent on chemical properties of the peptide sequence, the APEX approach weights spectral counts using estimates of these probabilities in order to improve the accuracy of absolute expression measures. A number of tools have been developed to estimate identification probabilities based upon the peptide sequence and mass spectrometry data, including one we have developed ourselves [47, 49–51]. Our approach uses a simple logistic regression model based on five peptide properties [51]. A weight p for each protein is generated by summing all of the peptide probabilities from an in silico digest of the protein. The number of spectral counts is approximately proportional to the product of the protein concentration and p , therefore we estimate the fraction of total molecules in parts per million (PPM) by

$$PPM_i = \frac{SC_i / p_i}{\sum_j SC_j / p_j} \cdot 10^6 \quad (1)$$

where, p_i is the weight and SC_i is the number of spectral counts for protein i and the summation in the denominator is over all the proteins identified in the experiment. In addition to a PPM estimate, the proportion of protein concentration by weight (PCW) is calculated by incorporating molecular weights (MW) into equation 1.

$$PCW_i = \frac{\frac{SC_i \cdot MW_i}{p_i}}{\sum_j \frac{SC_j \cdot MW_j}{p_j}} \quad (2)$$

From the PCW estimate, we calculate concentration by weight (ng/ μ L) by multiplying the PCW estimate by an estimate of total protein concentration (ng/ μ L), which is dependent on organism and cell or tissue type and based upon published values [52]. For example, the PCW for tenascin (uniprot ID: P24821) in a plasma based study of trauma was 8.6×10^{-4} and a published estimate for the total protein concentration of plasma is 70 g/L, therefore the resulting concentration estimate for tenascin protein is 6.0×10^4 ng/ml [41–42]. Protein concentration by weight estimates can easily be converted to molarity estimates (nmol/ml) by dividing by the molecular weight of each protein. For each individual protein, MOPED provides a table of concentration estimates and shows a bar graph comparing the protein concentration across different experiments grouped by tissue, localization and condition (Figure 2). The data in MOPED also reflects the inherent heterogeneity of absolute expression estimates that characterizes shotgun proteomics.

One application of MOPED absolute expression data is to provide cross-tissue and cross-platform comparisons in GeneCards [37, 38]. For each of nearly 20,000 human protein coding genes, GeneCards V3.10 shows comparative expression levels for 23 tissues and fluids. This levels largely based on the MOPED data. GeneCards also shows tissue transcriptome diagrams, thus allowing researchers a unique way to compare protein and RNA gene-centric expression values from multiple sources.

Relative Expression

Many proteomics experiments are comparative and are designed to measure the relative expression of a large number of proteins. For such experiments, MOPED includes data on relative expression experiments that include both labeled [55–58] and unlabeled studies.

For unlabeled experiments, we provide analyzed spectral count data using our SPIRE pipeline with the data analysis approach based upon the Linear Models for Microarray Analysis (LIMMA) [59]. For labeled analyses, we use an ANOVA based approach that incorporates effects from different peptide sequences, different mass spectrometry experiments, and other features of the experimental design such as blocking or time points.

Relative expression data in MOPED are displayed in terms of pairwise comparisons of conditions using the expression ratio for each protein. When experiments are replicated, MOPED provides both p-values and false discovery rate (FDR) estimates. P-values and expression ratios are often better than FDR for comparing the same protein across different conditions as FDR estimates are dependent on expression of other proteins within the

experiment [60,61]. Relative expression data can be selected via the data query interface (Figure 1) and for individual proteins on the protein page (Figure 2).

Chromosome Specific and Disease Specific Proteomics

To support the efforts of the HUPO's C-HPP and B/D-HPP projects [13–15], MOPED maps all human proteins to their specific chromosome allowing for querying and analysis based upon a specific chromosome. Proteins are also matched to different human diseases through UniProt disease descriptions allowing data querying by disease keywords. This linking of proteins to chromosome and disease will facilitate a future link-out by MOPED to data from the C-HPP and B/D-HPP projects.

Meta-data Checklist and Data Submission

The ability to integrate complex data from many different sources is highly dependent on the availability of accurate metadata. To this end, we have developed a multiomics metadata checklist. The checklist will not only be used for collecting metadata for MOPED but also has been made available to the community at large through DELSA (www.delsaglobal.org) [63]. The metadata checklist is a harmonization strategy for omics data integration, analysis and use. In addition, the checklist can be used as a framework for a data publication that will allow those who produce and share data to be cited and receive a proper credit [64].

Users of MOPED are encouraged to submit data and this is facilitated by use of the metadata checklist and a simple format for data upload. An example of this format is given in table 1. Users can then generate a data publication for this submission based upon the checklist and be able to have their data cited in the future.

Future Directions

While proteomics data give great insight into cellular function and disease, they provide a limited view into the functioning of complex biological systems. Multi-omics data are essential to advancing the knowledge of biological systems, understanding cell function and regulatory systems, modeling the “normal” condition in an organism, or predicting ecological responses to environmental changes. A better approximation of the “normal” condition will enable new insights in biological variability, genotype to phenotype relationship and organism changes due to environmental interactions, time passage or damage, to name a few points of interest.

Surveys of life scientists, omics experts, bioinformaticians, and students identified two of the most critical resource needs of the community: simple integrated multi-omics summaries of expression data and the linking of this information to biological processes. At their recent biannual meeting in May 2013, DELSA experts reiterated the need of the life sciences community for a multi-omics resource capable of integrating expression data across omics, pathways, networks, and experiments.

With these needs in mind, the Kolker Lab will use the MOPED proteomics resource as a prototype to build the MultiOmics Profiling Expression Database. This resource will

integrate proteomics data with transcriptomics and metabolomics data through biological pathways and networks. A key to this integration will be multiomics pathway and network analysis approaches such as those we have previously developed and employed and innovative visualization tools to help comprehend the complex associations found in multiomics data [65,66].

As an integrated multiomics resource, MOPED will enable assembly of communitywide, data-driven biological discovery data. The resource will be transparent and actionable for users through intuitive interfaces and effective education. MOPED feature development will be driven by users, whose engagement will be facilitated through DELSA. Together, we will build understanding one step at a time.

Acknowledgments

We thank Maggie Lackey for her critical reading. Research reported in this study was supported by the National Science Foundation under the Division of Biological Infrastructure award 0969929, National Institute of Diabetes and Digestive and Kidney Metabolites 2013, xx 15 Diseases of the National Institutes of Health under awards U01DK089571 and U01DK-072473, Seattle Children's Research Institute award, The Robert B. McMillen Foundation award, and The Gordon and Betty Moore Foundation award to EK.

References

- Hey, T.; Tansley, S.; Tolle, K., editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research; 2009.
- Stein LD. The case for cloud computing in genome informatics. *Genome Biology*. 2010; 11:207. [PubMed: 20441614]
- Higdon R, Haynes W, Stanberry L, Stewart E, Yandl G, Howard C, Broomall W, Kolker N, Kolker E. Unraveling the complexities of life sciences data. *Big Data*. 2013; 1:42–50.
- Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2005; 33:D154–159. [PubMed: 15608167]
- Vizcaino JA, Côté RG, Csordas A, Dianas JA, Fabregat A, Foster JM, Griss J, Alpi E, Birim M, Contell J, O'Kelly G, Schoenegger A, Ovelheiro D, Pérez-Riverol Y, Reisinger F, Ríos D, Wang R, Hermjakob H. The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res*. 2013; 41:D1063–1069. [PubMed: 23203882]
- Desiere F. The PeptideAtlas project. *Nucleic Acids Research*. 2006; 34:D655–D658. [PubMed: 16381952]
- Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*. 1977; 112:535–542. [PubMed: 875032]
- Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res*. 2012; 40:D48–53. [PubMed: 22144687]
- Kolker E, Higdon R, Haynes W, Welch D, Broomall W, Lancet D, Stanberry L, Kolker N. MOPED: Model Organism Protein Expression Database. *Nucleic Acids Res*. 2012; 40:D1093–1099. [PubMed: 22139914]
- Kolker E, Stewart E, Özdemir V. DELSA Global for “Big Data” and the Bioeconomy: Catalyzing Collective Innovation. *Industrial Biotechnology*. 2012; 8:176–178.
- Ozdemir V, Rosenblatt DS, Warnich L, Srivastava S, Tadmouri GO, Aziz RK, Reddy PJ, Manamperi A, Dove ES, Joly Y, Zawati MH, Hizel C, Yazan Y, John L, Vaast E, Ptolemy AS, Faraj SA, Kolker E, Cotton RGH. Towards an Ecology of Collective Innovation: Human Variome Project (HVP), Rare Disease Consortium for Autosomal Loci (RaDiCAL) and Data-Enabled Life Sciences Alliance (DELSA). *Current pharmacogenomics and personalized medicine*. 2011; 9:243–251. [PubMed: 22523528]

12. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, Beretta L, Bergeron J, Borchers C, Corthals GL, Costello CE, Deutsch EW, Domon B, Hancock W, He F, Hochstrasser D, Marko-Varga G, Salekdeh GH, Sechi S, Snyder M, Srivastava S, Uhlen M, Hu CH, Yamamoto T, Paik YK, Omenn GS. The human proteome project: Current state and future direction. *Mol Cell Proteomics*. 2011
13. Paik Y-K, Jeong S-K, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee H-J, Na K, Choi E-Y, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon J-Y, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee E-Y, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat Biotechnol*. 2012; 30:221–223. [PubMed: 22398612]
14. Paik Y-K, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee H-J, Na K, Jeong S-K, He F, Binz P-A, Nishimura T, Keown P, Baker MS, Yoo JS, Garin J, Archakov A, Bergeron J, Salekdeh GH, Hancock WS. Standard guidelines for the chromosome-centric human proteome project. *J Proteome Res*. 2012; 11:2005–2013. [PubMed: 22443261]
15. Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, Qin J, Omenn GS. The biology/disease-driven human proteome project (B/D-HPP): enabling protein research for the life sciences community. *J Proteome Res*. 2013; 12:23–27. [PubMed: 23259511]
16. Fanayan S, Smith JT, Sethi MK, Cantor D, Goode R, Simpson RJ, Baker MS, Hancock WS, Nice E. Chromosome 7-Centric Analysis of Proteomics Data from a Panel of Human Colon Carcinoma Cell Lines. *J Proteome Res*. 2013; 12:89–96. [PubMed: 23244008]
17. Chen L-C, Liu M-Y, Hsiao Y-C, Choong W-K, Wu H-Y, Hsu W-L, Liao P-C, Sung T-Y, Tsai S-F, Yu J-S, Chen Y-J. Decoding the disease-associated proteins encoded in the human chromosome 4. *J Proteome Res*. 2013; 12:33–44. [PubMed: 23256888]
18. Wu S, Li N, Ma J, Shen H, Jiang D, Chang C, Zhang C, Li L, Zhang H, Jiang J, Xu Z, Ping L, Chen T, Zhang W, Zhang T, Xing X, Yi T, Li Y, Fan F, Li X, Zhong F, Wang Q, Zhang Y, Wen B, Yan G, Lin L, Yao J, Lin Z, Wu F, Xie L, Yu H, Liu M, Lu H, Mu H, Li D, Zhu W, Zhen B, Qian X, Qin J, Liu S, Yang P, Zhu Y, Xu P, He F. First Proteomic Exploration of Protein-Encoding Genes on Chromosome 1 in Human Liver Stomach, and Colon. *J Proteome Res*. 2013; 12:67–80. [PubMed: 23256928]
19. Jeong S-K, Lee H-J, Na K, Cho J-Y, Lee MJ, Kwon J-Y, Kim H, Park Y-M, Yoo JS, Hancock WS, Paik Y-K. GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes. *J Proteome Res*. 2013; 12:106–111. [PubMed: 23252913]
20. Yamamoto T, Nakayama K, Hirano H, Tomonaga T, Ishihama Y, Yamada T, Kondo T, Kadera Y, Sato Y, Araki N, Mamitsuka H, Goshima N. Integrated View of the Human Chromosome X-centric Proteome Project. *J Proteome Res*. 2013; 12:58–61. [PubMed: 23259409]
21. Gaudet P, Argoud-Puy G, Cusin I, Duek P, Evalet O, Gateau A, Gleizes A, Pereira M, Zahn-Zabal M, Zwahlen C, Bairoch A, Lane L. neXtProt: organizing protein knowledge in the context of human proteome projects. *J Proteome Res*. 2013; 12:293–298. [PubMed: 23205526]
22. Zhang Y, Yan G, Zhai L, Xu S, Shen H, Yao J, Wu F, Xie L, Tang H, Yu H, Liu M, Yang P, Xu P, Zhang C, Li L, Chang C, Li N, Wu S, Zhu Y, Wang Q, Wen B, Lin L, Wang Y, Zheng G, Zhou L, Lu H, Liu S, He F, Zhong F. Proteome atlas of human chromosome 8 and its multiple 8p deficiencies in tumorigenesis of the stomach, colon, and liver. *J Proteome Res*. 2013; 12:81–88. [PubMed: 23256868]
23. Segura V, Medina-Aunon JA, Guruceaga E, Gharbi SI, González-Tejedo C, Sánchez del Pino MM, Canals F, Fuentes M, Casal JI, Martínez-Bartolomé S, Elortza F, Mato JM, Arizmendi JM, Abian J, Oliveira E, Gil C, Vivanco F, Blanco F, Albar JP, Corrales FJ. Spanish Human Proteome Project: Dissection of Chromosome 16. *J Proteome Res*. 2013; 12:112–122. [PubMed: 23234512]
24. Goode RJA, Yu S, Kannan A, Christiansen JH, Beitz A, Hancock WS, Nice E, Smith AI. The Proteome Browser Web Portal. *J Proteome Res*. 2013; 12:172–178. [PubMed: 23215242]
25. Farrah T, Deutsch EW, Hoopmann MR, Hallows JL, Sun Z, Huang C-Y, Moritz RL. The State of the Human Proteome in 2012 as Viewed through PeptideAtlas. *J Proteome Res*. 2013; 12:162–171. [PubMed: 23215161]

26. Zhou H, Di Palma S, Preisinger C, Peng M, Polat AN, Heck AJR, Mohammed S. Toward a comprehensive characterization of a human cancer cell phosphoproteome. *J Proteome Res.* 2013; 12:260–271. [PubMed: 23186163]
27. Kolker E, Higdon R, Welch D, Bauman A, Stewart E, Haynes W, Broomall W, Kolker N. SPIRE: Systematic protein investigative research environment. *J Proteomics.* 2011; 75:122–126. [PubMed: 21609792]
28. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res.* 2004; 3:1234–1242. [PubMed: 15595733]
29. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. *J Proteome Res.* 2004; 3:958–964. [PubMed: 15473683]
30. Higdon R, Reiter L, Hather G, Haynes W, Kolker N, Stewart E, Bauman AT, Picotti P, Schmidt A, van Belle G, Aebersold R, Kolker E. IPM: An integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J Proteomics.* 2011; 75:116–121. [PubMed: 21718813]
31. Higdon R, Kolker E. A predictive model for identifying proteins by a single peptide match. *Bioinformatics.* 2006; 23:277–280. [PubMed: 17121779]
32. Hather G, Higdon R, Bauman A, von Haller PD, Kolker E. Estimating false discovery rates for peptide and protein identification using randomized databases. *PROTEOMICS.* 2010; 10:2369–2376. [PubMed: 20391536]
33. Higdon R, Hogan JM, Kolker N, van Belle G, Kolker E. Experiment-Specific Estimation of Peptide Identification Probabilities Using a Randomized Database. *OMICS: A Journal of Integrative Biology.* 2007; 11:351–366. [PubMed: 18092908]
34. Higdon R, Kolker N, Picone A, van Belle G, Kolker E. LIP index for peptide classification using MS/MS and SEQUEST search via logistic regression. *OMICS.* 2004; 8:357–369. [PubMed: 15703482]
35. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. *Trends Genet.* 1997; 13:163. [PubMed: 9097728]
36. Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I, Krug H, Perco P, Mayer B, Kolker E, Safran M, Lancet D. Insilico human genomics with GeneCards. *Hum Genomics.* 2011; 5:709–717. [PubMed: 22155609]
37. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 1999; 27:29–34. [PubMed: 9847135]
38. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* 2008; 36:D623–631. [PubMed: 17965431]
39. Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, Ladunga I, Ulitsky-Lazareva B, Muruganujan A, Rabkin S, Vandergriff JA, Doremioux O. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* 2003; 31:334–341. [PubMed: 12520017]
40. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 2011; 39:D691–D697. [PubMed: 21067998]
41. Zhao D, Wu J, Zhou Y, Gong W, Xiao J, Yu J. WikiCell: a unified resource platform for human transcriptomics research. *OMICS.* 2012; 16:357–362. [PubMed: 22702248]
42. Starkey JM, Tilton RG. Proteomics and systems biology for understanding diabetic nephropathy. *J Cardiovasc Transl Res.* 2012; 5:479–490. [PubMed: 22581264]
43. Kolker E, Hogan JM, Higdon R, Kolker N, Landorf E, Yakunin AF, Collart FR, van Belle G. Development of BIATECH-54 standard mixtures for assessment of protein identification and relative expression. *PROTEOMICS.* 2007; 7:3693–3698. [PubMed: 17890649]
44. Bauman A, Higdon R, Rapson S, Loieue B, Hogan J, Stacy R, Napuli A, Guo W, van Voorhis W, Roach J, Lu V, Landorf E, Stewart E, Kolker N, Collart F, Myler P, van Belle G, Kolker E. Design

- and Initial Characterization of the SC-200 Proteomics Standard Mixture. *OMICS: A Journal of Integrative Biology*. 2011; 15:73–82. [PubMed: 21250827]
45. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*. 2004; 76:4193–4201. [PubMed: 15253663]
 46. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotechnol*. 2006; 24:333–338. [PubMed: 16525410]
 47. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*. 2007; 25:117–124. [PubMed: 17187058]
 48. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics*. 2005; 4:1265–1272. [PubMed: 15958392]
 49. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol*. 2007; 25:125–131. [PubMed: 17195840]
 50. Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat Biotechnol*. 2009; 27:190–198. [PubMed: 19169245]
 51. Louie B, Higdon R, Kolker E. The necessity of adjusting tests of protein category enrichment in discovery proteomics. *Bioinformatics*. 2010; 26:3007–3011. [PubMed: 21068002]
 52. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res*. 2010; 38:D750–D753. [PubMed: 19854939]
 53. Liu T, Qian W-J, Gritsenko MA, Xiao W, Moldawer LL, Kaushal A, Monroe ME, Varnum SM, Moore RJ, Purvine SO, Maier RV, Davis RW, Tompkins RG, Camp DG 2nd, Smith RD. Inflammation and the Host Response to Injury Large Scale Collaborative Research Programm High dynamic range characterization of the trauma patient plasma proteome. *Mol Cell Proteomics*. 2006; 5:1899–1913. [PubMed: 16684767]
 54. Slyke DDV, Hiller A, Phillips RA, Hamilton PB, Dole VP, Archibald RM, Eder HA. The Estimation of Plasma Protein Concentration from Plasma Specific Gravity. *J Biol Chem*. 1950; 183:331–347.
 55. Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*. 2002; 1:376–386. [PubMed: 12118079]
 56. Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R, Kolker E. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun Mass Spectrom*. 2001; 15:1214–1221. [PubMed: 11445905]
 57. Wiese S, Reidegeld KA, Meyer HE, Warscheid B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics*. 2007; 7:340–350. [PubMed: 17177251]
 58. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, Neumann T, Johnstone R, Mohammed AKA, Hamon C. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem*. 2003; 75:1895–1904. [PubMed: 12713048]
 59. Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.; Carey, V.; Dudoit, S.; Irizarry, R.; Huber, W., editors. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer; New York, pages: 2005. p. 397–420.
 60. Holzman T, Kolker E. Statistical analysis of global gene expression data: some practical considerations. *Current Opinion in Biotechnology*. 2004; 15:52–57. [PubMed: 15102467]
 61. Higdon R, van Belle G, Kolker E. A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics*. 2008; 24:1225–1228. [PubMed: 18424815]
 62. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, Bahir I, Belinky F, Morrey CP, Safran M, Lancet D. MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)*. 2013; 2013 bat018.

63. Kolker E, Özdemir V, Martens L, Hancock W, Anderson G, Naderson, et al. Towards more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics*. 2013 In press.
64. Snyder MS, Mias GI, Stanberry L, Kolker E. Metadata checklist for the integrated personal omics study: proteomics and metabolomics experiments. *Big Data*. 2013 In press.
65. Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E. Differential expression analysis for pathways. *PLoS Comput Biol*. 2013; 9:e1002967. [PubMed: 23516350]
66. Stanberry L, Mias GI, Haynes W, Higdon R, Snyder M, Kolker E. Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites*. 2013

Figure 1. MOPED Absolute Expression Query Results page. A Query by Organism (Mouse), Tissue (Lung), and Localization (Mitochondrion). Results table not fully shown. Also, shown is a chord plot detailing the proteins in MOPED broken down by tissue, localization, condition and experiment.

Figure 2. Individual protein page displaying links to outside resources and pathways and providing a comparison of absolute and relative expression across different proteomics experiments. The page also presents visualization of expression grouped by tissues comparing different experiments and conditions.

Table 1

Example of table for data upload to MOPED. Table shows only the first 5 rows. Tissue is from the BRENDA tissue ontology (BTO) and localization is from the Gene Ontology (GO).

Uniprot ID	Spectra Counts	Unique Peptides	FDR	Sequence Coverage	Organism	Tissue	Localization	Condition	Experiment
A0AV96	1	1	0.034	1.52	Human	BTO:0002096	GO:0016020	Cystic Fibrosis	charro_CF
A0FGR8	8	4	0	5.54	Human	BTO:0002096	GO:0016020	Cystic Fibrosis	charro_CF
A0PJW6	2	1	0	5.45	Human	BTO:0002096	GO:0016020	Cystic Fibrosis	charro_CF
A1L020	3	1	0.014	3.08	Human	BTO:0002096	GO:0016020	Cystic Fibrosis	charro_CF
A1L0T0	15	7	0	14.4	Human	BTO:0002096	GO:0016020	Cystic Fibrosis	charro_CF