UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Spectral Library Construction and Matching of MS/MS Spectra at Repository Scales

Permalink

https://escholarship.org/uc/item/08g414s7

Author

Wang, Mingxun

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Spectral Library Construction and Matching of MS/MS Spectra at Repository Scales

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Computer Science

by

Mingxun Wang

Committee in charge:

Professor Nuno Bandeira, Chair
Professor Vineet Bafna
Professor Pieter C Dorrestein
Professor Elizabeth Komives
Professor Pavel Pevzner

2017

Copyright

Mingxun Wang, 2017

All rights reserved.

The Dissertation of Mingxun Wang is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

DEDICATION

To my friends and family. Their support and guidance have made me into the man I am today.

EPIGRAPH

It seems like an end, my schooling is through
with nothing to learn, and nothing to do.
Rather now's just the start, my real education
to learn of the world, through the lens of vocation.
And armed with my knowledge, and a tad ability
I'll help make my mark, with dutiful responsibility.
So looking to the future, I'm thankful of my past
To friends and to family, let's keep having a blast!

Mingxun Wang

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables xi

Acknowledgements xii

Vita xiv

Abstract of the Dissertation xviii

Chapter 1 Introduction 1

Chapter 2 Spectral Library Generating Function - Significance for Spectral Matching 5
2.1 Introduction 5
2.2 Methods 7
2.3 Datasets 12
2.4 Results and Discussion 14
2.4.1 SLGF distribution Evaluation 14
2.4.2 Peptide Spectral Library Search 14
2.5 Conclusion 22

Chapter 3 Community Analysis and Knowledge Aggregation for Natural Products 24
3.1 Introduction 24
3.2 GNPS Spectral Libraries 26
3.3 Dereplication using GNPS 29
3.4 Molecular Networking 32
3.5 Living Data by Continuous Analysis 34
3.6 Molecular Explorer 35
3.7 Conclusion 36

Chapter 4 The Mass spectrometry Interactive Virtual Environment (MassIVE) for sharing and community reanalysis of proteomics data 39
4.1 Introduction 39
4.2 Data Sharing 41

4.2.1 The Mass Spectrometry Virtual Interactive Environment 41
4.2.2 Complete Datasets and Provenance 43
4.3 Live Data 44
4.3.1 Dataset Owner Updates/Upgrades 44
4.3.2 Dataset Re-analyses of Datasets 45
4.3.3 Broad Re-analysis of Partial Datasets 46
4.3.4 MassIVE Search 47
4.3.5 In Depth Re-analysis 47
4.3.6 In Depth Re-analysis 47
4.3.7 Dataset Updates - Broadcasting 48
4.4 Data Community 50
4.4.1 Reaching Consensus - Results Comparison 51
4.5 Discussion 52
4.6 Methods 53
4.6.1 Automated Conversions 53
4.6.2 Automated Searches 53
4.6.3 Results Comparison 53
4.6.4 Results Estimation of FDR 54

Chapter 5 Automated Knowledge Aggregation for Repository Scale Peptide Spectral Libraries 55
5.1 Introduction 55
5.2 Aggregation/Search of Human Proteomics Data 57
5.3 Construction and Spectrum Quality of Peptide Spectral Library 58
5.4 Library Provenance 60
5.5 Proteome Coverage 60
5.6 Targeted Proteomics Design 61
5.7 Detecting Unobserved Proteins 62
5.8 Knowledgebase Redistribution 63
5.9 Discussion 64
5.10 Methods 65
5.10.1 Data Origin and Preparation 65
5.10.2 Single Pass Search 65
5.10.3 Dynamic Search Space Adjustment Search 65
5.10.4 Synthetic Searching 66
5.10.5 Spectra Extraction/Filtering 67
5.10.6 Spectral Library Precursor Selection for Inclusion 67
5.10.7 Online Spectral Library Augmentation 68
5.10.8 Filtering out Ambiguous Library Spectra 68
5.10.9 Novel Protein Calling 68
5.10.10 Gold Standard Library Construction 69

Appendix A SLGF Dynamic Programming Details 74

Appendix B GNPS Methods 77
B.0.1 Spectral Library Searching 77
B.0.2 Variable Dereplication 77
B.0.3 Molecular Network Construction 78
B.0.4 GNPS Collections - Sample Preparation 78
B.0.5 GNPS Collections - LC MS/MS Acquisition 79
B.0.6 GNPS Collections - Library Creation 79
B.0.7 GNPS-Community Contributed Spectral Library Processing and Control 80

Appendix C GNPS Supplementary Information 81
C.0.1 Note 1: Browsing Available Libraries 81
C.0.2 Note 2: GNPS Library Addition Procedure (Collections and Community) 81
C.0.3 Note 3: Library Search False Discovery Rate Estimation 82
C.0.4 Note 4: GNPS Dataset Creation and Access 84
C.0.5 Note 5: Public Data Molecular Networks 85
C.0.6 Note 6: Continuous Identification Procedure 85
C.0.7 Note 7: Molecular Explorer Creation and Presentation 86

Appendix D MassIVE-KB Supplementary Information 95
D.0.1 Online Library Creation 95
D.0.2 Database p-value Calculation 95
D.0.3 Dynamic Search Space Adjustment for Proteome Limited Data . 96
D.0.4 Database Identification of Synthetic Peptide Pools from ProteomeTools 97
D.0.5 Global Error Rate Controls for Spectral Library 97
D.0.6 Decoy Inclusion in Spectral Library 98
D.0.7 Library Representative Selection 98
D.0.8 Spectral Library Search 99
D.0.9 Proteins in Dataset 99

Bibliography 104

LIST OF FIGURES

Figure 2.1. Distributions of variation in ion intensity 9

Figure 2.2. Spectral Library Generating Function calculation of Spectrum-Spectrum Match Significance 13

Figure 2.3. Empirical vs calculated (SLGF) distributions 15

Figure 2.4. Distribution of total ion current 16

Figure 2.5. Peptide spectral library search sensitivity and specificity comparison 18

Figure 2.6. Gains of identification by Tremolo over SpectraST 19

Figure 3.1. Overview of GNPS 27

Figure 3.2. GNPS spectral libraries 30

Figure 3.3. Molecular Network Creation and Visualization 33

Figure 3.4. Living data in GNPS by crowdsourcing molecular annotations . . . 36

Figure 4.1. Overview of MassIVE ecosystem 42

Figure 4.2. MassIVE Reanalysis 49

Figure 4.3. MassIVE Community 52

Figure 5.1. Overview of MassIVE-KB 70

Figure 5.2. MassIVE-KB Coverage of Human Proteome 71

Figure 5.3. Growth of Proteome Coverage 72

Figure 5.4. Proteotypic Peptides in Proteome 73

Figure A.1. SLGF calculation example 75

Figure A.2. SLGF Yeast ROC Curve 76

Figure C.1. Spectral Library Search Match Assessment 89

Figure C.2. Replicate Count Histogram within Reported Matches 90

Figure C.3. Chemical Space Visualization with Molecular Networking 90

Figure C.4. Example Continuous Identification Digest Email 91
Figure C.5. GNPS Collaboration Initiation 92
Figure C.6. Molecular explorer highlights 93
Figure C.7. Dataset fold increase in identifications by continuous identification 94
Figure D.1. Technical Search Overview 100
Figure D.2. Search Space Score Adjustment 101
Figure D.3. Size Comparison to NIST 101
Figure D.4. MassIVE-KB Library Quality 102
Figure D.5. Selection Methods for Representative 103
Figure D.6. MassIVE-KB Library Identification Rate 103

LIST OF TABLES

Table 2.1. Tremolo Missed Spectra Table 20

Table 5.1. Top Contributing PIs to MassIVE-KB 69

Table C.1. GNPS Libraries Summary 87

Table C.2. Library Quality Levels Description 88

Table C.3. Number of Annotation Revisions 88

Table C.4. Rating Description 89

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Nuno Bandeira for his support as the chair of my committee.

Chapter 1, in part is a reprint of the material as it appears in ”Spectral Library Searching of Tandem Mass Spectra” as part of the UCSD Computer Science and Engineering Research Exam. Wang, M (2013). The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is a reprint of the material as it appears in Spectral library generating function for assessing spectrum-spectrum match significance in Journal of Proteome Research 2013. Wang, M, Bandeira, N (2013). Spectral library generating function for assessing spectrum-spectrum match significance. J. Proteome Res., 12, 9:3944-51. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part, of the material as it appears in Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking in Nature Biotechnology. Wang, M, Carver, JJ, Phelan, VV, Sanchez, LM, Garg, N, Peng, Y, Nguyen, DD, Watrous, J, Kapono, CA, Luzzatto-Knaan, T, Porto, C, Bouslimani, A, Melnik, AV, Meehan, MJ, Liu, WT, Crsemann, M, Boudreau, PD, Esquenazi, E, Sandoval-Caldern, M, Kersten, RD, Pace, LA, Quinn, RA, Duncan, KR, Hsu, CC, Floros, DJ, Gavilan, RG, Kleigrewe, K, Northen, T, Dutton, RJ, Parrot, D, Carlson, EE, Aigle, B, Michelsen, CF, Jelsbak, L, Sohlenkamp, C, Pevzner, P, Edlund, A, McLean, J, Piel, J, Murphy, BT, Gerwick, L, Liaw, CC, Yang, YL, Humpf, HU, Maansson, M, Keyzers, RA, Sims, AC, Johnson, AR, Sidebottom, AM, Sedio, BE, Klitgaard, A, Larson, CB, Boya P, CA, Torres-Mendoza, D, Gonzalez, DJ, Silva, DB, Marques, LM, Demarque, DP, Pociute, E, O’Neill, EC, Briand, E, Helfrich, EJ, Granatosky, EA, Glukhov, E, Ryffel, F, Houson, H, Mohimani, H, Kharbush, JJ, Zeng, Y, Vorholt, JA, Kurita, KL, Charusanti, P, McPhail, KL, Nielsen, KF, Vuong, L, Elfeki, M, Traxler, MF, Engene, N, Koyama,

N, Vining, OB, Baric, R, Silva, RR, Mascuch, SJ, Tomasi, S, Jenkins, S, Macherla, V, Hoffman, T, Agarwal, V, Williams, PG, Dai, J, Neupane, R, Gurr, J, Rodriguez, AM, Lamsa, A, Zhang, C, Dorrestein, K, Duggan, BM, Almaliti, J, Allard, PM, Phapale, P, Nothias, LF, Alexandrov, T, Litaudon, M, Wolfender, JL, Kyle, JE, Metz, TO, Peryea, T, Nguyen, DT, VanLeer, D, Shinn, P, Jadhav, A, Mller, R, Waters, KM, Shi, W, Liu, X, Zhang, L, Knight, R, Jensen, PR, Palsson, B, Pogliano, K, Linington, RG, Gutirrez, M, Lopes, NP, Gerwick, WH, Moore, BS, Dorrestein, PC, Bandeira, N (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol., 34, 8:828-37. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part is currently being prepared for submission for publication of the material. The Mass spectrometry Interactive Virtual Environment (MassIVE) for sharing and community reanalysis of proteomics data. Mingxun Wang, Jeremy Carver, Nuno Bandeira. The dissertation author was the primary investigator and author of this material.

Chapter 5, in part is currently being prepared for submission for publication of the material. Assembling the Community Scale Discoverable Human Proteome. Mingxun Wang, Jian Wang, Jeremy Carver, Nuno Bandeira. The dissertation author was the primary investigator and author of this material.

VITA

2009	Bachelor of Science in Computer Engineering, University of Illinois, Urbana Champaign
2013	Master of Science in Computer Science, University of California, San Diego
2017	Doctor of Philosophy in Computer Science, University of California, San Diego

PUBLICATIONS

Mingxun Wang, Jeremy Carver, Nuno Bandeira, "The Mass spectrometry Interactive Virtual Environment (MassIVE) for sharing and community reanalysis of proteomics data," (in preparation)

Mingxun Wang, Jian Wang, Jeremy Carver, Nuno Bandeira, "Assembling the Community Scale Discoverable Human Proteome," (in preparation)

Neha Garg, **Mingxun Wang**, Embriette Hyde, Ricardo R. da Silva, Alexey V. Melnik, Ivan Protsyuk, Amina Bouslimani, Yan Wei Lim, Richard Wong, Greg Humphrey, Gail Ackermann, Timothy Spivey, Sharon S. Brouha, Nuno Bandeira, Grace Y. Lin, Forest Rohwer, Douglas J. Conrad, Theodore Alexandrov, Rob Knight, Pieter C. Dorrestein, "Three dimensional cartography of microbiome and metabolome data onto radiological images of the human lung," (in revisions)

Kerstin Scheubert, Franziska Hufsky, Daniel Petras, **Mingxun Wang**, Louis-Felix Nothias, Kai Duehrkop, Nuno Bandeira, Pieter Dorrestein, Sebastian Boecker, "Significance estimation for large scale untargeted metabolomics annotations," (in revisions)

Tal Luzzatto Knaan, Neha Garg, **Mingxun Wang**, Evgenia Glukhov, Yao Peng, Gail Ackermann, Amnon Amir, Brendan M Duggan, Sergey Ryazanov, Lena Gerwick, Rob Knight, Theodore Alexandrov, Nuno Bandeira, William H Gerwick, Pieter C Dorrestein, "Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae," eLife 6 (2017): e24214.

Yasset Perez-Riverol, Mingze Bai, Felipe da Veiga Leprevost, Silvano Squizzato, Young Mi Park, Kenneth Haug, Adam J Carroll, Dylan Spalding, Justin Paschall, **Mingxun Wang**, Noemi del-Toro, Tobias Ternent, Peng Zhang, Nicola Buso, Nuno Bandeira, Eric W Deutsch, David S Campbell, Ronald C Beavis, Reza M Salek, Ugis Sarkans,

Robert Petryszak, Maria Keays, Ariana Barbera, Rafael C Jimnez, Alexey I Nesvizhskii, Susanna-Assunta Sansone, Christoph Steinbeck, Rodrigo Lopez, Juan Antonio Vizcano, Peipei Ping, Henning Hermjakob, "Omics Discovery Index Discovering and Linking Public Omics Datasets," Nature biotechnology 35.5 (2017): 406-409.

Amina Bouslimania, Alexey V Melnik, Zhenjiang Xu, Amnon Amir, Ricardo R da Silva, **Mingxun Wang**, Nuno Bandeira, Theodore Alexandrov, Rob Knight, and Pieter C. Dorrestein, "Lifestyle chemistries from phones for individual profiling," Proceedings of the National Academy of Sciences 113.48 (2016): E7645-E7654.

Gert Wohlgemuth1, Sajjan S Mehta, Ramon F Mejia, Steffen Neumann, Diego Pedrosa, Tom Pluskal, Emma L Schymanski, Egon L Willighagen, Michael Wilson, David S Wishart, Masanori Arita, Pieter C. Dorrestein, Nuno Bandeira, **Mingxun Wang**, Tobias Schulze, Reza M Salek, Christoph Steinbeck, Venkata Chandrasekhar Nainala, Robert Mistrik, Takaaki Nishioka, Oliver Fiehn, "SPLASH, A hashed identifier for mass spectra," Nature biotechnology 34.11 (2016): 1099-1101.

Eric W Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, Yasset Perez-Riverol, Tobias Ternent, David S Campbell, Manuel Bernal-Llinares, Shujiro Okuda, Shin Kawano, Robert L Moritz, Jeremy J Carver, **Mingxun Wang**, Yasushi Ishihama, Nuno Bandeira, Henning Hermjakob, Juan Antonio Vizcano, "The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition," Nucleic acids research 45.D1 (2017): D1100-D1106.

Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Caldern, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jrn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric JN Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins, Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams,

Jingqui Dai, Ram Neupane, Joshua Gurr, Andrs MC Rodrguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Mller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutirrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, Nuno Bandeira, "Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking," Nature biotechnology 34.8 (2016): 828-837.

Waqas Nasir, Alejandro Gomez Toledo, Fredrik Noborn, Jonas Nilsson, **Mingxun Wang**, Nuno Bandeira, Goran Larson, "SweetNET: A bioinformatics workflow for glycopeptide MS/MS spectral analysis," Journal of Proteome Research 15.8 (2016): 2826-2840.

Amina Bouslimani, Carla Porto, Christopher M Rath, **Mingxun Wang**, Yurong Guo, Antonio Gonzalez, Donna Berg-Lyon, Gail Ackermann, Gitte Julie Moeller Christensen, Teruaki Nakatsuji, Lingjuan Zhang, Andrew W Borkowski, Michael J Meehan, Kathleen Dorrestein, Richard L Gallo, Nuno Bandeira, Rob Knight, Theodore Alexandrov, Pieter C Dorrestein, Molecular cartography of the human skin surface in 3D," Proceedings of the National Academy of Sciences 112.17 (2015): E2120-E2129.

Katherine R Duncan, Max Crsemann, Anna Lechner, Anindita Sarkar, Jie Li, Nadine Ziemert, **Mingxun Wang**, Nuno Bandeira, Bradley S Moore, Pieter C Dorrestein, Paul R Jensen, Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species," Chemistry & biology 22.4 (2015): 460-471.

Hosein Mohimani, Roland D Kersten, Wei-Ting Liu, **Mingxun Wang**, Samuel O Purvine, Si Wu, Heather M Brewer, Ljiljana Pasa-Tolic, Nuno Bandeira, Bradley S Moore, Pavel A Pevzner, Pieter C Dorrestein, Automated genome mining of ribosomal peptide natural products," ACS chemical biology 9.7 (2014): 1545-1551.

Mingxun Wang, Nuno Bandeira, Spectral library generating function for assessing spectrum-spectrum match significance," Journal of proteome research 12.9 (2013): 3944-3951.

FIELDS OF STUDY

Major Field: Computer Science

Studies in Bioinformatics and Computational Mass Spectrometry

Professors Nuno Bandeira, Vineet Bafna, and Pavel A. Pevzner

ABSTRACT OF THE DISSERTATION

Spectral Library Construction and Matching of MS/MS Spectra at Repository Scales

by

Mingxun Wang

Doctor of Philosophy in Computer Science

University of California, San Diego, 2017

Professor Nuno Bandeira, Chair

The characterization of proteins, peptides, metabolites, and natural products are crucial to the understanding biological processes, discovering biomarkers, and uncovering new therapeutic molecules. Tandem mass spectrometry (MS/MS) has proven to be a high throughput and sensitive tool to assay these molecules, whereby the fragmentation observed in the MS/MS spectra functions as a reproducible signature for each molecule. Thus, any acquisition of a molecule's MS/MS fragmentation can be aggregated into a reusable collection of observed and annotated MS/MS spectra known as a spectral library.

Due to the reproducibility of a molecule's MS/MS spectrum, spectral libraries

have gained traction as a resource for the sensitive identification of newly acquired MS/MS spectra. Thus, the utility of spectral libraries rests on the reliability of MS/MS similarity metrics as well as the quality and size of the libraries themselves.

In this dissertation we highlight the computational methods that were developed to enable the creation of spectral libraries for proteomics, metabolomics, and natural products discovery. These methods include the aggregation and analysis of the entire community's mass spectrometry data along with online computational resources that crowd-source the annotation and curation of specialized spectral libraries. Further, by leveraging repository scale mass spectrometry data, we have developed methods to assign statistical significance to spectral similarity metrics in order to enable the automated identification of MS/MS data by matching to spectral libraries.

Chapter 1

Introduction

Mass spectrometry has proven to be a valuable tool in understanding the role of biomolecules at a molecular level and in guiding drug discovery. Metabolomics and proteomics [2], the study of metabolites and proteins respectively, have extensively leveraged mass spectrometry analysis to help elucidate the role of metabolites and proteins in biological processes. Additionally, proteins and metabolites along with natural products show promise as biomarkers as well as a source of antibiotics and other therapeutics [74]. Due to the complexity of biology, many of the discoveries in proteomics and metabolomics have only been made possible by computational methods. While once it was possible for a diligent scientist to manually examine all mass spectrometry data, today, high-throughput instruments are capable of generating hundreds to thousands of mass spectra in the matter of hours, making manual interpretation impractical [69].

The interpretation of tandem mass spectrometry (MS/MS) data has been a key focus in computational mass spectrometry research. Though the computational community has invested heavily in the past several decades on the interpretation of MS/MS spectra, the success of computational approaches to unambiguously assign an MS/MS spectrum to a putative structure varies vastly depending on the chemical domain.

In proteomics, amino acids are the building blocks of proteins and are coded for by three DNA base pairs. Thus, given a genome of an organism we cab compute a proposed

proteome, i.e. a list of proteins translated from the genome [5]. The most common method to computationally identify proteins is comparing tandem mass spectrometry data to these genome derived proteome databases, known as database search [26, 49]. A major limitation of database search is that the software must guess at what the tandem mass spectrum for a particular peptide will be. Additionally, the space of possible structures is not clearly defined simply by the genome derived proteome. Post translation modifications [65] and single amino acid variants (SAAVs) [113] drastically augment this space. The vast majority of this chemical space does not correspond to molecules in a biological sample, but presents opportunities for false identifications that also incur a large computational cost to explore. While there exist database search methods that attempt to skirt around these issues [73, 115], an alternative approach is to rely on spectral libraries for identifications. Since spectral libraries are collections of reference spectra of known compounds that have been previously observed, they avoid the need to predict fragmentation and also accurately capture the empirical chemical space (including modifications and SAAVs) observed in data, rather than a combinatorially growing space.

In contrast to proteomics, the metabolomics and natural products fields have not been blessed with computational techniques quite as successful. Though there exist several tools for identification of various non-peptidic molecules [24, 85, 3], the lack of sensitivity, absence of false discovery rate controls, and high computational cost limit their utility in large scale untargeted experiments. Again, spectral libraries present an alternative approach which capitalizes on previous discovery efforts. In this space, where computational tools lack maturity, the case for spectral libraries is even more compelling. To validate the identification of a molecule requires long and laborious chemical assays such as Nuclear Magnetic Resonance (NMR), Crystallography, or chemical synthesis. Though, in many instances this process is unavoidable (e.g. novel compound discovery), the process is necessary only once. By depositing MS/MS spectra of validated compounds

in spectral libraries, the community benefits in never repeating these costly validation procedures.

Spectral libraries provide a way to capture the mass spectrometry knowledge of a community. This knowledge can be compared to newly acquired MS/MS spectra to identify them in a sensitive manner [58, 114]. Further, spectral libraries can also inform experimental design, e.g. constructing targeted single reaction (SRM) monitoring experiments [54]. Thus the success of spectral libraries as a community resource critically hinges two factors. First, there must exist appropriate measure of similarity between library and query MS/MS spectra. While the field has generally converged upon using cosine based similarity metrics [88], the statistical meaning of these scores is not estimated. Chapter 2 presents a method to assign significance to spectrum matching. Second, the spectral libraries must comprehensively cover all biological molecules of interest. Currently available libraries in metabolomics/natural products discovery [91, 39, 106] have fallen short in being comprehensive and have been unable to identify a sizeable fraction of the community's data. Chapter 3 highlights an approach and platform that democratizes and crowd-sources the creation/curation of spectral libraries for metabolomics/natural products discovery, resulting in a drastic increase in spectral library search utility. In proteomics, the completeness of spectral libraries is highly dependent on the availability and diversity of community data as well as the ability to re-analyze these data in a high throughput manner. Chapter 4 presents a mass spectrometry data repository, the Mass Spectrometry Interactive Virtual Environment (MassIVE), and how this repository is enabling the community to share and re-analyze mass spectrometry data in meaningful ways. Finally, chapter 5 builds on these repository efforts and presents a method to take advantage of the entire proteomics communitiy's data and create a reusable peptide spectral library at repository scales.

Chapter 1, in part is a reprint of the material as it appears in "Spectral Library

Searching of Tandem Mass Spectra" as part of the UCSD Computer Science and Engineering Research Exam. Wang, M (2013). The dissertation author was the primary investigator and author of this paper.

Chapter 2

Spectral Library Generating Function - Significance for Spectral Matching

2.1 Introduction

High throughput identification of peptides and proteins in complex samples is enabled by tandem mass (MS/MS) spectrometry generation of hundreds of thousands to millions of spectra, from which many thousands of proteins are typically identified by matching the resulting MS/MS spectra against genome-derived databases of known protein sequences [2]. In difference from such database search algorithms [26, 16, 50], spectral library search approaches [111, 92, 14, 31, 58, 98, 112, 17] identify experimental MS/MS spectra by matching against collections of previously identified reference spectra (spectral libraries) and have been consistently found to identify more spectra than database search whenever the corresponding peptides have reference spectra in the library. But despite this demonstrated superior sensitivity, the development of methods to determine the statistical significance of Spectrum-Spectrum Matches (SSMs) in peptide spectral library searches is still in its early stages.

The most common approach to controlling the False Discovery Rate (FDR) in both database search [25] and spectral library search [57] is the Target-Decoy approach where one extends the database/library of true peptides with a complement of sequences/spectra

from ‘random’ peptides and uses matches to the latter to estimate the number of false matches to true sequences/spectra. But while these FDR approaches continue to be very valuable in correcting for multiple hypothesis testing in large-scale experiments, they provide little to no insight on the statistical significance of individual SSMs or Peptide Spectrum Matches (PSMs). In addition, it has been shown [49, 50, 35] that rigorous modeling of random PSM scores allows one to determine accurate p-values for true PSMs and thus substantially improve the performance of database search tools. In this MS-GF [49] approach, dynamic programming is used to exhaustively determine the distribution of PSM scores for all possible peptides matched to a given spectrum and then this distribution is used to determine the probability (p-value) of observing a random PSM score at least as high as the score of an observed PSM derived from the database of known peptide sequences. Unfortunately this approach does not have a direct analog in the realm of spectral library searches - while it is straightforward to traverse the space of all possible random peptide sequences (as in MS-GF), it remains unclear how to generate and/or traverse a space of ‘random’ spectra that would be representative of false matches to a true spectral library. First, truly random spectra[1] are easy to generate and could be traversed in a manner similar to MS-GF but such spectra would be mostly very different from the spectra that tend to be generated by mass spectrometry instruments and thus p-values obtained using this background distribution of random spectra would not accurately reflect the probability of false matches for experimental MS/MS spectra. Second, the approach used for the generation of decoy spectra [57] in FDR calculations continues to work well in practice for the generation of small collections of ‘semi-random’ peptide spectra but it is not sufficient to explore the space of *all* ‘random’ spectra because it only considers limited changes to peak masses, allows for no variations in peak intensities and is completely peptide-specific in that it is based on sequence permutations (and thus not

[1] For example, all spectra of Euclidean norm 1.0 at a pre-determined fixed resolution for peak intensities.

applicable to spectra from other types of molecules). Third, SSM scoring and p-value approaches have been proposed based on statistical models of random SSMs but these assume uniform distributions of peak masses and either ignore (e.g., hypergeometric models [112, 17]) or make limited use (e.g., peak ranks in Kendall-Tau statistic [17]) of MS/MS peak intensities. As a result, even though these approaches use a probabilistic model and calculate p-values, the underlying assumptions and their results on real MS/MS data suggest that these don't represent the statistics of SSMs well enough to increase the overall number identified SSMs (more details in Results).

2.2 Methods

A spectrum is defined as a set of (mass, intensity) pairs called peaks which are assigned into uniformly sized mass bins (e.g. 1 Th bins [31]). After transformation [2] a spectrum becomes a vector S with n bins, where each bin S_i contains the summed intensity of all peaks with masses in that bin; all subsequent references to "spectrum" refer to the respective spectrum's vector. Given a library spectrum L and a query spectrum S, the projection spectrum $Proj(S,L)$ is defined as:

$$Proj(S,L) = \{S_i : (L_i > 0), 0\ otherwise\}$$

All library spectra L are normalized to euclidean norm $||L|| = 1.0$, as are all projected spectra:

$$S^L = NormProj(S,L) = \left\{ \frac{Proj(S,L)_i}{||Proj(S,L)||} \right\},\ where\ ||Proj(S,L)|| = \sqrt{\sum_i Proj(S,L)_i^2}$$

[2] We note that even though spectrum binning is used here for ease of explanation of our approach, the actual implementation uses peak lists to improve performance as well as to provide the ability to use per-peak m/z tolerances.

The most common spectral similarity function used for spectral matching is cosine (also known as normalized dot product), defined as follows for vectors of Euclidian norm 1.0:

$$cos(L, S^L) = \sum_i L_i \times S_i^L$$

We define a replicate spectrum R (relative to a library spectrum L) to be a spectrum of the same molecule as L and acquired under the same or similar experimental conditions (i.e., charge state, instrument, collision energy, abundance, etc.). Due to stochastic factors in mass spectrometry fragmentation and instrument measurement error [95], some level of variation is expected between the intensities of peaks in R relative to the intensities of peaks in L. We model this variability with a log ratio of ion intensities, $log(R_i^L/L_i)$ where $R^L = NormProj(R, L)$; this ratio is calculated across all bins in R and L where L_i is not zero.

For all library spectra with replicate spectra in our training datasets, all observed log ratios were collected into an ion variation histogram. We use this histogram (scaled to total area under the curve 1.0) as the empirical probability mass function, $RatioFreq(r)$, of variation in ion intensities for all $L_i \neq 0$ (Figure 2.2A). In difference from varying intensities, the special cases of ion deletion (i.e., $R_i^L = 0$) are modeled separately with $DelFreq = \frac{\#Deletions}{\sum_{L \in Library} \#Peaks_L \times \#Replicates_L}$, where $\#Deletions$ is the total number of peak deletions in all replicate spectra, $\#Replicates_L$ is the number of replicate spectra in our training set of a Library spectrum L, and $\#Peaks_L$ is the number of peaks in a given library spectrum L. Combining our model of ion variance, $RatioFreq(r)$, and our model for ion deletion events $DelFreq$, the probability $Prob(R_i|L_i)$ of observing a replicate ion intensity R_i given a library ion intensity L_i is:

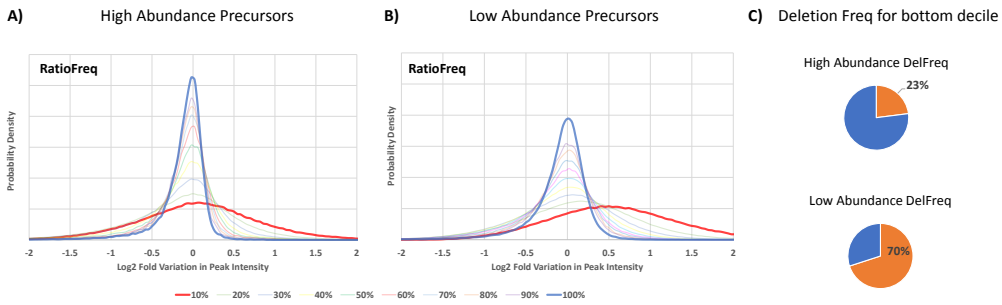


Figure 2.1. Distributions of variation in ion intensity for (A) high and (B) low TIC precursors. As illustrated by the differences between (A) and (B), the ion variation distributions for low TIC precursors tend to be wider than those of the high TIC precursors. In both cases it is shown in bold blue, the distribution of variation in ion intensities for the 10% most intense peaks in library spectra and in bold red, the distribution of variation in ion intensities for the 10% least intense peaks in library spectra, with other deciles shown in between. Note that the width of the distributions for the top decile distribution is markedly narrower than the bottom decile distribution suggesting the need to model the variation of ion intensity differently depending upon a peak's intensity in the library spectrum. It should also be noted that the lowest decile distribution shown in (B) in red is not centered at 0 log fold variation due to the significantly higher deletion percentage (C) of the lowest decile library peaks in low TIC precursors. The deletion of these peaks caused all other peaks in the spectrum to increase in normalized intensity, for the entire spectrum is normalized to Euclidean Norm 1, thus causing a shift in the specific ion variation distribution.

$$Prob(R_i|L_i) = \begin{cases} DelFreq & \text{if } R_i = 0 \\ (1 - DelFreq) \times RatioFreq(\log_2(\frac{R_i}{L_i})) & \text{if } R_i \neq 0 \end{cases}$$

Our goal is to calculate the distribution of cosines scores over all possible replicate spectra within instrument variability of a given library spectrum. To compute the generating function for each library spectrum, we use $RatioFreq$ and $DelFreq$. We consider every possible replicate spectrum R (Figure 2.2B) by exploring all possible intensity variations of every peak L_i to R_i and calculate their aggregate probability and

cosine similarity $cos(R^L, L)$.

A three dimensional dynamic programming table, $LibDP(i,c,p)$, is used to explore all possible ion variations, where i is the spectrum vector index, c is the cosine score, and p is the intermediate euclidean norm value. The value in each cell $LibDP(i,c,p)$ is the probability of a replicate spectrum obtaining cosine c, having p euclidean norm, and up to and including ion index i. The recurrence for $LibDP(i,c,p)$ is thus defined as follows for $i = 1..n$:

$$LibDP(i,c,p) = \sum_{y=0 \rightarrow p} LibDP(i-1, c-\sqrt{y} \times L_i, p-y) \times Prob(\sqrt{y}|L_i)$$

At ($i = 0$), before considering any ions from the library spectrum L, every cell $LibDP(0,>0,>0) = 0$, and when no intensity is used and the cosine score is zero, it is $LibDP(0,0,0) = 1$. (See Supplementary Figure 2.2 for an illustration of the recursion).

Since each step i only depends on the values from step $i-1$, one only needs to use two two-dimensional matrices of constant size to calculate all $LibDP$ values. The size of each 2D matrix is $n_c \times n_p$, where n_c is the number of cosine bins and n_p is the number of intensity bins, each set accordingly to the desired granularity. The time complexity of computing the entire dynamic programming table is $O(n_c \times n_p^2 \times n)$. The final SLGF distribution of cosines between a library spectrum and its replicates is extracted from $LibDP(n,*,1)$ and normalized to sum to 1. It is necessary to normalize at the end because of probability mass in discarded replicate spectra of Euclidean norm $\neq 1$.

The probability that a replicate spectrum R and corresponding library spectrum L exhibit a cosine less than a threshold is expressed as the following p-value:

$$Prob_L(cos(L,R^L) < T)$$

where T can be set according to observed cosines between query and library spectra to determine the probability of a query spectrum S being a replicate spectrum of L.

We developed a spectral library search tool called Tremolo in order to assess SLGF's utility in library search. Since SLGF assesses the quality of a single SSM but does not correct for multiple hypothesis testing when searching many spectra in a dataset, in Tremolo, FDR is estimated by the TDA [57]. In brief, decoy spectral libraries were generated using the peptide shuffle and reposition method to obtain a set of decoy spectra. For all Target and Decoy library spectra, SLGF distributions were calculated and used in the subsequent scoring function.

The scoring function of an SSM in the spectral library search between a library spectrum L identified as $Peptide(L)$ and a query S is:

$$SSM_{score} = SLGF_e = Prob_L(cos(L,R^L) < cos(L,S^L)) \times ExpInt(S, Peptide(L))$$

where we define $S^{e|Peptide(L)}$ to be the subset of peaks from S which are annotated by peptide of L and

$$ExpInt(S, Peptide(L)) = \frac{\sum_i (S_i^{e|Peptide(L)})}{\sum_i (S_i)}$$

The score $SLGF_e$ is also considered in addition to $SLGF = Prob_L(cos(L,R^L) < cos(L,S^L))$ because it is a closer comparison to SpectraST [58], which does not consider co-eluting peptides. In cases of co-eluting peptides, $SLGF$ is able to consider these spectra because it only uses peaks at m/z values $L_i \neq 0$. Yet since SpectraST penalizes for co-elution, $SLGF_e$ is similarity penalized by the explained intensity term, and thus $SLGF_e$ is used as the scoring function in Tremolo. While $SLGF$ may be useful towards identification of co-eluting peptides, additional considerations are required to correctly address co-eluting peptide identification [98] (e.g., addressing multiple molecules per

spectrum and FDR on mixture identifications).

2.3 Datasets

The *Training dataset* was composed of 236 CPTAC [79, 93] Study 6 Orbitrap files (2,766,504 spectra) and was used to train the distributions of variation in ion intensities. All spectra were searched with SpectraST v4.0 with a 2 Th m/z tolerance against the NIST Yeast Ion Trap peptide library (May 2011 build). The decoy spectral library was created using SpectraST's own decoy generation feature and the resulting SSMs were filtered to 1% FDR, yielding 396,526 identified spectra from 18,440 unique precursors. Replicate spectra in this filtered dataset were matched with their respective library spectra in the library and ion variance distributions were calculated from these replicates and library spectra.

Yeast and Human Ion Trap peptide spectral libraries, containing 78,825 and 310,688 spectra repectively, were acquired from NIST (May 2011 Build) and were used in evaluating spectral library search performance. The shuffle and reposition method proposed by Lam et al [57] was used to create the decoy spectral libraries for use in Tremolo's, SpectraST's, and Pepitome's search. SpectraST, Pepitome, and Tremolo searched the NIST library using a 2 Th precursor tolerance; Tremolo and Pepitome used a 0.5 Th tolerance to annotate MS/MS peaks.

The *Yeast Test dataset* from CPTAC was composed of an arbitrarily selected file from CPTAC Study 6 that was *not* included in the Training dataset and contained 9,809 MS/MS spectra used for evaluating SLGF's search of the NIST Yeast Ion Trap peptide library. A second test dataset, *Hela S3 Test* dataset, was also used to evaluate spectral search performance in addition to the CPTAC *Yeast Test dataset* described above. This dataset [84] was generated from human HeLa S3 cells, by tryptic digestion, and MS acquisition was performed by an LTQ Orbitrap Classic yielding high accuracy MS1 and

low mass accuracy MS2 spectra. *Hela S3 Test* contained 11,723 MS/MS spectra and was searched against the NIST Human Ion Trap peptide library.

For Tremolo peptide spectra library search we find that it is best to perform library preprocessing. Peaks in library spectra were annotated with the respective peptide sequence considering *b*,*y*,*b*++,*y*++ ions [82], their respective single ^{13}C isotope peaks ($+1$ Da mass shift), single H_2O losses (-18 Da mass shift), single NH_3 (-17 Da mass shift) losses and *a* ions; all non-annotated peaks and precursor peaks were removed. Additionally, all peak intensities were transformed by square root in both library and query spectra.

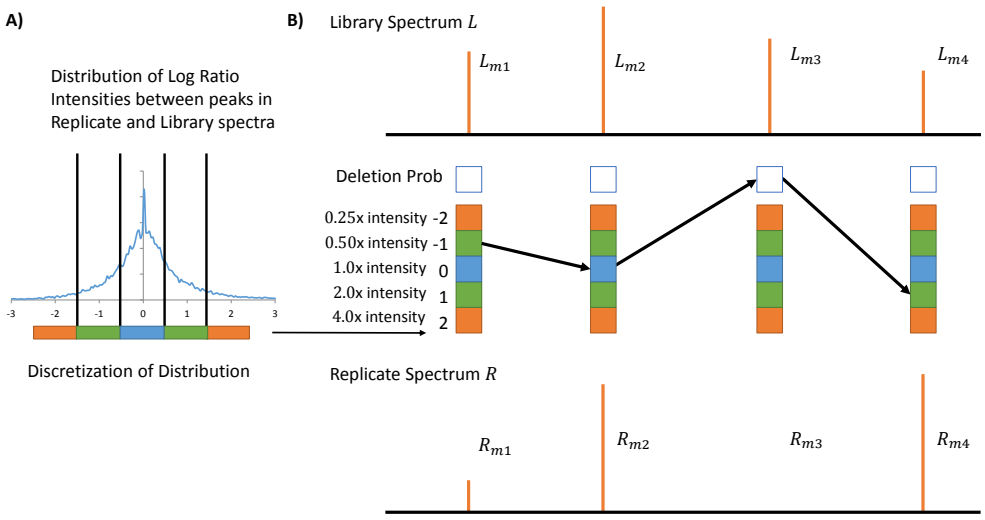


Figure 2.2. Spectral Library Generating Function (SLGF) calculation of Spectrum-Spectrum Match (SSM) significance by modeling instrument variation in peak intensity measurements in replicate spectra of the same compound. In (A) the empirical distribution of variations in intensity measurements is assessed and discretized. (B) Every possible replicate spectrum *R* is represented by a path through each library peak's possible intensity variations, thus representing every possible combination of variations of peaks in library spectrum *L*. Note that some replicate spectra (paths) are invalid, i.e. replicate spectra that do not have Euclidean norm 1; intermediate paths with norm < 1.0 are allowed during the calculation but paths resulting in euclidean norm $\neq 1.0$ must be discarded when calculating the distribution of cosine scores.

2.4 Results and Discussion

2.4.1 SLGF distribution Evaluation

In calculating the SLGF distributions for each peptide library spectrum, it was observed that it was more accurate to have 10 different distributions of variation in ion intensity based on the relative intensity of an ion peak in the library spectrum (instead of a single distribution for all ion peaks). Figure 2.1 illustrates the significant differences in the distributions of log-ratio ion variations for the top 10% most intense peaks and bottom 10% least intense ion peaks (other deciles also shown). Additionally, the ion variance is also substantially different between replicate spectra from high ($\geq 12,000$ ions) and low total ion current ($< 12,000$ ions) precursors. The total ion current (TIC) of a spectrum is calculated as the total ion current of that MS/MS spectrum. As expected, low intensity replicates exhibit more variation in ion intensities (i.e., wider ion variation distributions) than high intensity replicates. Using these two separate models for high and low TIC precursors, two SLGF distributions are pre-calculated for each library spectrum. Low and high TIC query spectra are then partitioned and searched separately, with SLGF p-values calculated from the corresponding SLGF distribution.

The SLGF distributions were visually assessed by comparing against empirical score distributions using replicates from the Training dataset, and it was found that SLGF distributions approximated the empirical distributions (See Figure 2.3), but further work is necessary to enable a more accurate p-value calculation. Thus, we use SLGF p-value as a score and evaluate its performance in the context of spectral library search.

2.4.2 Peptide Spectral Library Search

To systematically assess the performance of Tremolo as a search tool, we compared it against SpectraST and Pepitome at fixed FDR. SpectraST's performance exceeded

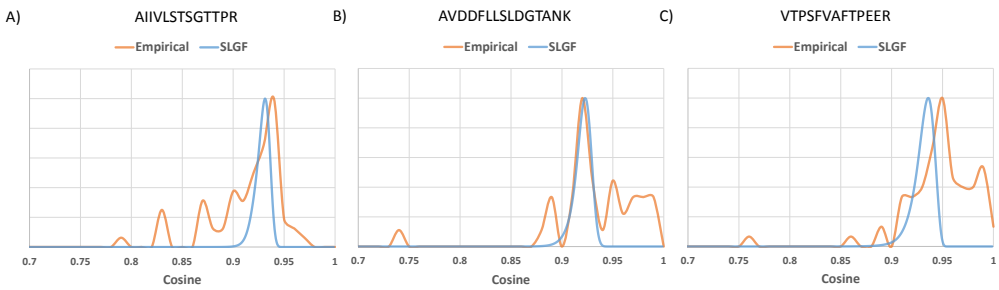


Figure 2.3. Empirical vs calculated (SLGF) distributions of cosines between library and replicate spectra. Empirical cosine distributions are histograms of cosine scores between a library spectrum and replicate experimental spectra ($cos(L, R^L)$ for all available R^L). Examples were selected from NIST Yeast library spectra with sufficient replicates from the *Training dataset* to derive an empirical distribution. The probability mass in the empirical distributions to the left of the theoretical SLGF distribution is mostly caused by cases of co-elution, leading to lower cosines. The discrepancies of higher cosines result from SLGF using an average model of ion variation derived from data acquired in many experiments and laboratories. As such, our average model of ion intensity variation for certain peptides has variance higher than that of the best calibrated instruments, thus causing SLGF distributions to expect lower average cosines than some empirical distributions.

that of Pepitome, and thus we focus our detailed analysis on the comparison with SpectraST results. Comparing the sensitivity of Tremolo to that of SpectraST on the *Yeast Test dataset* we find that at 1% spectrum-level FDR (as determined by TDA), Tremolo was able to identify 4,373 spectra versus 3,884 spectra by SpectraST (12.5% more, see Figure 2.5).

The gain in Tremolo sensitivity can be explained by analyzing the different components in SpectraST's SSM scores. SpectraST's score is $SSM_{spectrastscore} = 0.6D + 0.4DeltaD - b$, where D is the cosine score between library and query. $DeltaD = \frac{D_1 - D_2}{D_1}$ where D_1 and D_2 are the top and second cosine scores respectively from a set of library spectra to a query spectrum. It is argued that the larger this $DeltaD$ term, the more the top candidate stands out from the alternatives, thus implying a greater chance the top candidate is correct. b is the penalty applied to the score for *DotBias* scores that are

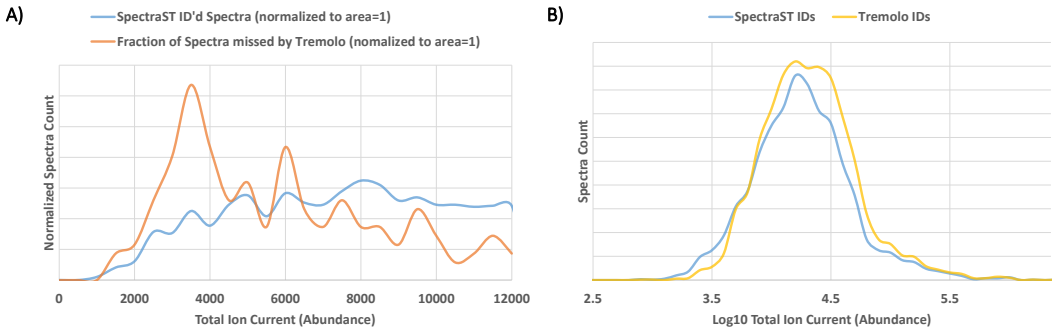


Figure 2.4. Distribution of total ion current for unidentified low TIC spectra and spectra identified at 1% FDR on the Training dataset. A) Low TIC spectra that were not identified by Tremolo but were identified by SpectraST (in orange) on the Test dataset at 1% FDR. In blue is the distribution of TICs for SpectraST identifications in the Training dataset and subsequently used to train SLGF. (B) Distribution of peptide TICs for IDs at 1% FDR on the Test dataset, with Tremolo shown in yellow and SpectraST in blue. Tremolo is found to be more sensitive in regions of high TIC ($>$ 12000 ions) but loses some sensitivity in the very low TIC region. The loss of these identifications is caused in part by the low TIC ion variation model being trained on spectra mostly in the 5000-12000 ions range and thus not optimally modeling variation in spectra in the $<$ 5000 ions range.

not preferable. *DotBias* is defined as $DotBias = \frac{\sqrt{\sum_{i=1 \to n} L_i^2 * (S_i^L)^2}}{D}$ and intuitively can be thought of a measure of how much a cosine score is dominated by a few peaks. A score of $DotBias = 1.0$ signifies one peak dominates the score and a score of $DotBias \approx 0.0$, the cosine contribution is evenly distributed over all peaks. Thus, high *DotBias* scores possibly imply dubious matches as there are only a few peaks leading to a high cosine score. Low *DotBias* scores also are not preferable as this means many equal intensity peaks are matching, which most likely would imply noise. SpectraST's penalty b is tuned to cause larger penalties for larger and very small *DotBias* values. In figure 2.6A it is shown the *DeltaD* versus the *DotBias* of all IDs at 1% FDR identified by SpectraST over the test dataset (orange dots). As expected SpectraST IDs tend to avoid high *DotBias* as well as exceedingly low *DeltaD*. In difference from these, Tremolo-only IDs are shown in gray dots. These tended towards lower *DeltaD* and higher *DotBias* and thus were missed by SpectraST, because of either low *DeltaD* (lacking the ability to distinguish

between the top two SSMs) or because of high *DotBias* (cosine score dominated by only a few peaks). Tremolo, however, is able to identify these low *DeltaD* spectra because each possible library match to the query spectrum has a different expected score distribution, and even though from an absolute cosine perspective there is little difference from the top and second hit, once the cosine p-value is calculated for each respective library spectrum, then the scores separate substantially. It is clearly show in figure 2.6B that even though the *DeltaD* for these spectra that SpectraST failed to identify was very low (x-axis), the Tremolo's delta score ($SLGF_{delta} = \frac{SLGF_1 - SLGF_2}{SLGF_1}$ where $SLGF_1$ and $SLGF_2$ are the top and second $SLGF_e$ scores respectively) is considerably higher because of the separation obtained from SLGF p-values.

The identification of spectra with high *DotBias* is also enabled by the calculated cosine distribution. For spectra whose intensities are dominated by very few peaks, cosines alone are not enough to distinguish between good SSMs and bad SSMs. Since these spectra are dominated by few peaks, the less intense peaks become especially informative in how their slight cosine changes (because of matching or not matching these small peaks) distinguish good and bad SSMs. SLGF's distributions are able to capture these slight changes in cosine (i.e. higher SLGF distributions with lower variance) to correctly identify spectra dominated by few peaks whereas SpectraST penalizes *all* spectra that are dominated by a few peaks. In general we observe that the effect of *DeltaD* is captured by SLGF's determination of the appropriate mean cosine per library spectrum and the effect of *DotBias* is captured by the variance of the SLGF distributions.

An additional source of IDs that Tremolo was able to recover were spectra that SpectraST did not consider in its search: spectra that have "negligible" intensity above 500 m/z. These spectra generally came from shorter (6-8mer) charge 2 precursors and moderate length (8-12mer) charge 3 precursors. While these spectra may be easier to match to decoys with SpectraST's scoring scheme, Tremolo is again able to use its

calculated distributions to identify 198 spectra that SpectraST discarded.

The 671 spectra that were identified by SpectraST and not by Tremolo are shown in Figure 2.6A as blue dots and exhibit no clear bias for or against *DeltaD* or *DotBias*. Upon closer examination we found that many of the spectra from low TIC precursors that Tremolo scored poorly seemed to exhibit relatively high cosine scores (~ 0.85) yet the SLGF distribution would expect a score significantly higher. However, these cases are mostly skewed towards very low TIC (< 5000 ions) and our training set of spectra for the low TIC ion variance models skewed toward higher TIC (Figure 2.4A), thus suggesting that a larger training set may be required to train ion variance models for precursors of TIC < 5000 ions.

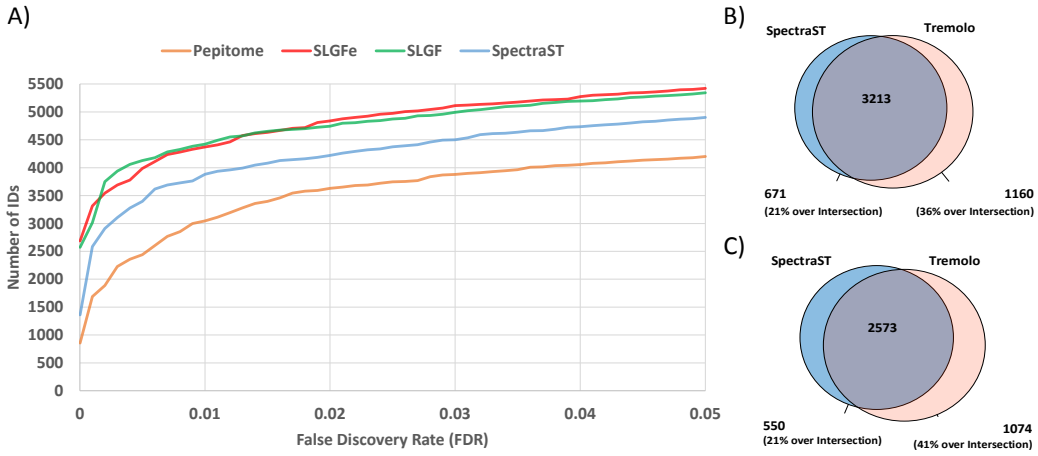


Figure 2.5. Peptide spectral library search sensitivity and specificity comparison between Tremolo, SpectraST, and Pepitome. In (A) the performance of the scoring function *SLGF* is shown to be comparable to that of $SLGF_e$, where one explicitly attempts to not consider mixture spectra (see text for details). This contrasts to the performance of SpectraST in blue, and Pepitome in orange. (B) Number of spectra identified exclusively by SpectraST, Tremolo, and by both tools at 1% spectrum level FDR. On this Test dataset, $SLGF_e$ was also found to be more sensitive than SpectraST across the whole range of FDR thresholds ($> 12\%$ increase in IDs at 1% FDR). (C) Number of peptide IDs at 1% FDR by SpectraST and Tremolo. Tremolo identified 16% more peptides than SpectraST.

Missed identifications by Tremolo on spectra from high TIC precursors separated

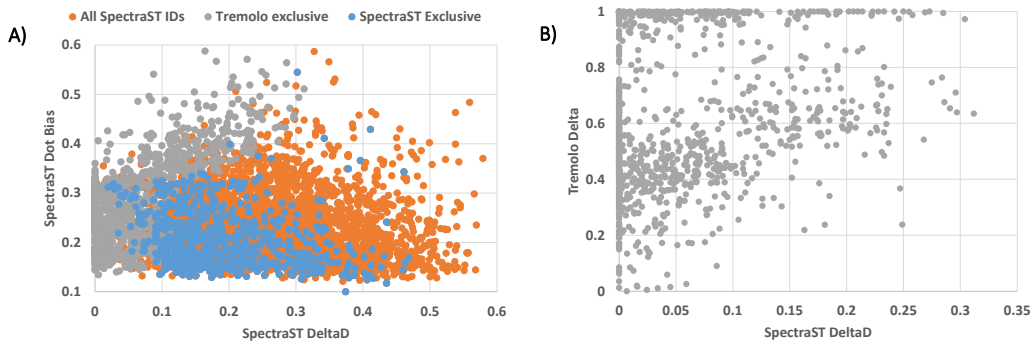


Figure 2.6. Gains of identification by Tremolo over SpectraST through analyzing SpectraST's score's *DotBias* and *DeltaD*. Each dot in (A) represents an identified spectra plotted with *DeltaD* versus *DotBias* as calculated by SpectraST. In orange are all the IDs by SpectraST; in blue are the identifications by SpectraST at 1% FDR that were missed by Tremolo's search at 1% FDR; in gray are the identifications by Tremolo at 1% FDR that were missed by SpectraST at 1% FDR. For this third category of spectra, there is a clear bias towards high *DotBias* and low *DeltaD*. It is shown in (B) that while SpectraST was unable to obtain a large *DeltaD* for these spectra, Tremolo's delta score was high and, since this delta score is not used anywhere in Tremolo scores, it thus reinforces the assertion that these Tremolo identifications are correct. Tremolo's exclusive identifications show that there are classes of spectra that remained unidentified (low *DeltaD* and high *DotBias*) that Tremolo is now able to identify. Note the change of scale for the DeltaD axis (x-axis) in B as opposed to A; since there were no spectra with SpectraST $DeltaD > 0.35$ we opted to omit those regions in the figure.

into several categories as shown in Table 2.1. Many examples of deamidation (a post translational modification that increases the mass of amino acids N or Q by 1 Da) were seen throughout our analysis for SpectraST is unable to distinguish between the two variants of the peptide because of the peak smoothing in SpectraST's spectrum preprocessing. However, we were able to distinguish these cases and *correctly* not identify them because deamidated versions of spectra were not present in the library. While identifying modified peptides from unmodified spectra is a worthy goal [6], we argue that such searches should *explicitly* identify query spectra as modified variants of library spectra instead of just reporting them as the same unmodified peptide identification.

Additionally, 63 spectra contained a high number (> 2) of ^{13}C isotope atoms. In

Table 2.1. SpectraST identifications missed by Tremolo at 1% FDR. There were a total of 334 high TIC and 337 low TIC spectra that were missed by Tremolo at 1% FDR. While the numbers are comparable, the proportion of low TIC IDs missed was much higher as there were only $\sim$ 1400 low TIC spectra identified by Tremolo at 1% FDR. This higher percentage of missed low TIC spectra can be attributed to suboptimal ion variation models for very low TIC spectra as described in the text and in Figure 2.4.

Reason	Number of spectra
Low TIC Deamidation or 1 ^{13}C	105
Low TIC precursor m/z with > 2 ^{13}C	36
Low TIC Other	193
High TIC Deamidation or 1 ^{13}C	60
High TIC precursor m/z with > 2 ^{13}C	63
High TIC Other	214
Total	671

these cases ^{13}C replaced the more common ^{12}C in the peptide, causing an increase in precursor mass because of the additional neutrons. The presence of these ^{13}C also affects the prefix and suffix ions in the MS/MS spectra as they skew a portion of the intensity of the b,y,b++,y++, etc. ions into peaks of 1 Da higher mass. This distorts the shape of the spectrum and exaggerates the variance in ion intensities of the query spectrum beyond what is expected by the SLGF distributions.

Of the remaining 214 other spectra that were not identified by Tremolo at 1% FDR, we manually examined a representative subset of these cases and determined that $\sim$ 30% of spectra contained a mixture of two or more peptides. Another $\sim$ 23% were matched to library spectra of questionable quality, exhibiting low signal to noise ratio and a high proportion of un-annotated peaks in the reference spectrum. While we have a low TIC ion variation model that accounts for low TIC *query* spectra matched to high quality library spectra, we could not account for lower quality library spectra since this information is not readily available. Several of these cases of lower quality library spectra were replaced in the subsequent release of the NIST yeast IT library indicating that NIST revisions also concluded that these library spectra were of lower quality. $\sim$ 28% of the

spectra were matched to high quality library spectra and exhibited high SSM cosine scores (~ 0.85) but SLGF distributions were too strict (e.g., expected mean cosine was too high), which may indicate that our average model of variation of ion intensities across all library spectra may not be the most appropriate for specific library peptides resulting in less reproducible spectra.

Despite marked gains over Pepitome [17] and SpectraST [58], our results suggest that 3 levels of precursor intensity models (< 5000 ions, $5000 - 12000$ ions, > 12000 ions) may be better suited to model peak intensity variations across the range of precursor TICs in our sample and could thus further improve Tremolo's performance (Figure 2.4A). In addition, while our models take into consideration the precursor TIC for query spectra, it would also be informative to know the precursor TIC of library spectra since fragmentation patterns in these are also very dependent on precursor TIC. Further studies will be able to determine the effect of both of these factors through the use of larger training and reference datasets.

In addition to comparing search performance on the *Yeast test dataset*, the *Hela S3 Test* dataset was searched against the NIST Human Ion Trap spectral library using both Tremolo and SpectraST. Filtering to 1% FDR, Tremolo was able to identify 7,332 MS/MS spectra compared to SpectraST's 6,723 MS/MS spectra, yielding a 9% increase in IDs. In this search, Tremolo further imposed a threshold of a minimum explained intensity of 0.55. Note that since we calculated the SLGF theoretical distributions for the NIST human library using the ion variance models trained on the CPTAC *training dataset*, the amount of ion variation seemed to exceed what was estimated in the *Yeast training dataset*. SLGF theoretical distributions were shifted slightly lower to adjust for this higher variability in ion intensities, resulting in an additional gain in IDs, bringing the total gain to 10.5% more. This indicates that some amount of retraining may be beneficial when using different spectral libraries and experimental conditions.

Even though we did not explicitly aim to identify mixture spectra (and did not evaluate it), we note that the proposed SLGF approach is based on matching reference library spectra to *subsets* of peaks in query spectra (i.e., normalized projections) and thus appears to be well suited to determining *containment* of compounds in mixture spectra. While a detailed assessment of SLGF's performance on mixture spectra would require a more comprehensive evaluation [98], our preliminary results illustrated in Figure 2.5A show that *SLGF*'s performance was essentially indistinguishable from that of $SLGF_e$ and thus strongly suggests that SLGF should be suitable for identification of peptides in mixture spectra.

Despite SLGF's utility in spectral library search, SLGF theoretical score distributions are currently unable to capture the multi-modality in empirical score distributions that are shown in figure 2.3 because SLGF is limited to modeling only one source of variation, i.e. instrument variability in intensity measurements. Other sources of variation, such as co-elution and alternate fragmentation pathways, would also need to be considered in order to obtain more accurate SLGF theoretical score distributions. Another limitation of SLGF p-values is that these are not usable to directly estimate false discovery rates in the absence of TDA. Unlike MS-GF's p-values for the distribution of *false* matches, SLGF models the distribution of *true* matches. Thus, SLGF p-values could possibly enable the calculation of false negative rates and thus allow one to estimate sensitivity (proportional to area under the curve for true matches) but not accuracy (proportional to area under the curve for false matches).

2.5 Conclusion

Having been repeatedly found [58, 112, 56, 17] that spectral library searching performs consistently better than database search of the same peptide identification search space in high-throughput proteomics, there is now renewed interest in establishing

statistical methods to further assess the quality of Spectrum-Spectrum Matches (SSMs) and increase the total number of reported SSM-based identifications. Here we propose a new Spectral Library Generating Function (SLGF) approach to assessing the significance of SSMs, show how to rigorously calculate SLGF distributions for any spectrum from any type of molecule and demonstrate that SLGF-based peptide spectral library searching identifies significantly more spectra than state-of-the-art alternative search tools. In difference from database search (and other fields) where statistical significance is estimated by calculating the p-value of observing a high match score when matching a random sequence, we circumvent the open problem of defining realistic 'random' MS/MS spectra by instead calculating the p-value of observing a low match score when matching a true (replicate) spectrum to a known reference spectrum. To achieve this goal, we explicitly model instrument variation in measurement of MS/MS peak intensities and show how these can be used to derive theoretical distributions of SSM cosines between replicate and reference library spectra.

Chapter 2, in part, is a reprint of the material as it appears in Spectral library generating function for assessing spectrum-spectrum match significance in Journal of Proteome Research 2013. Wang, M, Bandeira, N (2013). Spectral library generating function for assessing spectrum-spectrum match significance. J. Proteome Res., 12, 9:3944-51. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Community Analysis and Knowledge Aggregation for Natural Products

3.1 Introduction

Natural products (NPs) from marine and terrestrial environments, including their inhabiting microorganisms, plants, animals, and humans, are routinely analyzed using mass spectrometry. However a single mass spectrometry experiment can collect thousands of MS/MS spectra in minutes [10] and individual projects can acquire millions of spectra. These datasets are too large for manual analysis. Further, comprehensive software and proper computational infrastructure are not readily available and only low-throughput sharing of either raw or annotated spectra is feasible, even among members of the same lab. The potentially useful information in MS/MS datasets can thus remain buried in papers, laboratory notebooks, and private databases, hindering retrieval, mining, and sharing of data and knowledge. Although there are several NP databases Dictionary of Natural Products [1], AntiBase [55] and MarinLit [66] that assist in dereplication (identification of known compounds), these resources are not freely available and do not process mass spectrometry data. Conversely, mass spectrometry databases including Massbank [39], Metlin [91], mzCloud, and ReSpect [109] host MS/MS spectra but limit data analyses to several individual spectra or a few LC-MS files. While Metlin and

mzCloud provide a spectrum search function, unfortunately, their libraries are not freely available.

Global genomics and proteomics research has been facilitated by the development of integral resources such as the National Center for Biotechnology Information (NCBI) and UniProt KnowledgeBase (UniProtKB), which provide robust platforms for data sharing and knowledge dissemination [7, 63]. Recognizing the need for an analogous community platform to effectively share and analyze natural products MS data, we present the Global Natural Products Social Molecular Networking (GNPS, available at gnps.ucsd.edu). GNPS is a data-driven platform for the storage, analysis, and knowledge dissemination of MS/MS spectra that enables community sharing of raw spectra, continuous annotation of deposited data, and collaborative curation of reference spectra (referred to as spectral libraries) and experimental data (organized as datasets).

GNPS provides the ability to analyze a dataset and to compare it to all publicly available data. By building on the computational infrastructure of the University of California San Diego (UCSD) Center for Computational Mass Spectrometry (CCMS), GNPS provides public dataset deposition/retrieval through the Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository. The GNPS analysis infrastructure further enables online dereplication[91, 59, 42, 60], automated molecular networking analysis[72, 102, 75, 89, 97, 104, 27, 110], and crowdsourced MS/MS spectrum curation. Each dataset added to the GNPS repository is automatically reanalyzed in the next monthly cycle of continuous identification (see Living Data by Continuous Analysis below). Each of these tens of millions of spectra in GNPS datasets is matched to reference spectral libraries to annotate molecules and to discover putative analogs (Fig. 3.1a). From January 2014 to November 2015, GNPS has grown to serve 9,267 users from 100 countries (Fig. 3.1b), with 42,486 analysis sessions that have processed more than 93 million spectra as molecular networks from a quarter million LC-MS runs. Searches

against a combined catalog of over 221,000 MS/MS reference library spectra from 18,163 compounds (Supplementary Table C.1) are possible, and GNPS has matched almost one hundred million MS/MS spectra in all public and private search jobs using an estimated 84,000 compute hours.

3.2 GNPS Spectral Libraries

GNPS spectral libraries enable dereplication, variable dereplication (approximate matches to spectra of related molecules), and identification of spectra in molecular networks. GNPS has collected available MS/MS spectral libraries relevant to NPs (which also include other metabolites and molecules), including MassBank[39], ReSpect[109] and NIST[76] (Fig. 3.2a, and Supplementary Table C.1). Altogether, these third party libraries total 212,230 MS/MS spectra representing 12,694 unique compounds (Fig. 3.2b). While this combined collection of reference spectra, provides a starting point for dereplication, only 1.01% of all spectra public GNPS datasets has been matched to this collection, indicating insufficient chemical space coverage. Although the NP community is working to populate this missing chemical space, there is no way to report discoveries of chemistries in an easily verifiable and reusable format.

To begin to address this pressing need, GNPS houses both newly-acquired reference spectra (GNPS-Collections) as well as a crowd-sourced library of community-contributed reference spectra (GNPS-Community). GNPS-Collections includes NPs and pharmacologically active compounds totaling 6,629 MS/MS spectra of 4,243 compounds (Fig 3.2b, Supplementary Table C.1, Supplementary Note C.0.1.C.0.2). The GNPS-Community library has grown to include 2,224 MS/MS spectra of 1,325 compounds from 55 worldwide contributors. While the total number of MS/MS spectra in GNPS libraries is only 4% of the MS/MS spectra collected in third party libraries, GNPS libraries contribute matches of MS/MS spectra at a scale disproportionate to their size (Fig. 3.2c). The

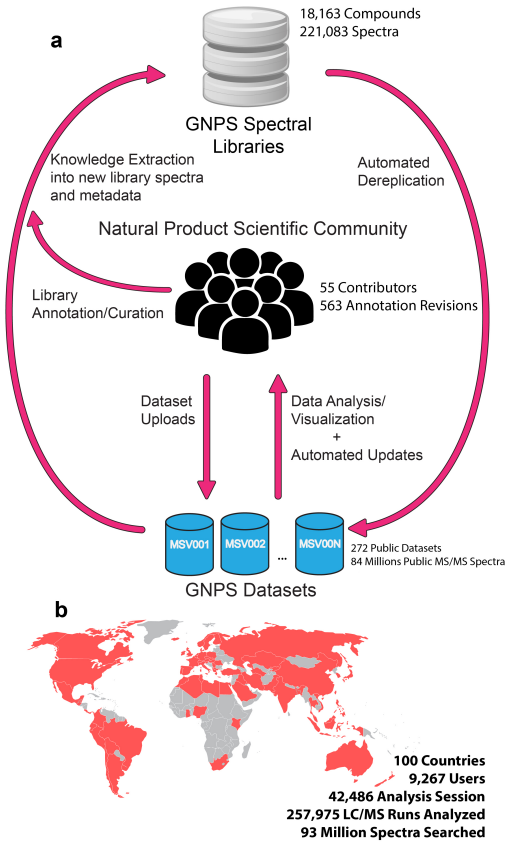


Figure 3.1. Overview of GNPS. (a) Representation of interactions between the natural product community, GNPS spectral libraries, and GNPS datasets. At present 221,083 MS/MS spectra from 18,163 unique compounds are used for the search in the GNPS. These include both 3rd party libraries such as MassBank, ReSpect, and NIST, as well as spectral libraries created for GNPS (GNPS-Collections) and spectra from the natural product community (GNPS-Community). GNPS spectral libraries grow through user contributions of new identifications of MS/MS spectra. To date, 55 community members have contributed 8,853 MS/MS spectra from 5,568 unique compounds (30.5% of the unique compounds available). In addition, on-going curation efforts have already yielded 563 annotation updates for library spectra. The utility of these libraries is to dereplicate compounds (recognition previously characterized and studied known compounds), in both public and private data. This dereplication process is performed on all public datasets and results are automatically reported, thus enabling users to query all datasets/organisms/conditions. Automatic reanalysis of all public data creates a virtuous cycle in which contributions to libraries can be matched to all public data. Combined with molecular networking (Fig. 3.3), this automatic reanalysis empowers community members to identify analogs that can then be added to GNPS spectral libraries. (b) The GNPS platform has grown to serve a global user base of 9,200+ users from 100 countries.

GNPS libraries account for 29% of unique compound matches and 59% of the MS/MS matches in public (88% of public+private) data. This indicates that the GNPS libraries contain compounds that are complementary to the chemical space represented in other libraries (Fig. 3.2c,d). Moreover, in contrast to third party libraries, spectra submitted to GNPS-Community libraries are immediately searchable by the whole community, such that submissions seamlessly transfer knowledge between laboratories (Fig. 3.1a) in a process that is akin to the addition of genome annotations to GenBank[7].

In order to create a robust library, it is important for submissions to be peer-reviewed and, if necessary, annotations corrected or updated as appropriate. Reference spectra submitted to the GNPS-Community library are categorized by the estimated reliability of the proposed submissions. Gold reference spectra must be derived from structurally characterized synthetic or purified compounds and can only be submitted by approved users. Approval is given to contributors who have undergone training. Training is initiated by contacting the corresponding authors or CCMS administrators. Silver reference spectra need to be supported by an associated publication, while Bronze reference spectra are all remaining putative annotations (Supplementary Table C.2). This type of division of spectra is reminiscent of RefSeq/TPA/GenBank[7, 81] (genomics) and Swiss-Prot/TrEMBL/UniProt[4, 5] (proteomics), allowing for varying tradeoffs between comprehensiveness and reliability of annotations defined as Gold, Silver, and Bronze (Fig. 3.2e).

To enable refinements or corrections of annotations, GNPS allows for community driven, iterative re-annotation of reference MS/MS spectra in a wiki-like fashion, to progressively improve the library and converge towards consensus annotation of all MS/MS spectra of interest. This is a process similar to the iterative annotation of the human genome (e.g., see series of papers on NCBI GenBank[7]). To date, 563 annotation revisions have been made in GNPS (Supplementary Table C.3), most of which added

metadata to library spectra or refined compound names. The history of each annotation is retained so that users can discuss the proper annotation and address disagreements via comment threads.

3.3 Dereplication using GNPS

High throughput dereplication of NP MS/MS data is implemented in GNPS by querying newly acquired MS/MS spectra against all the accumulated reference spectra in GNPS spectral libraries (Fig. 3.2a). To date, more than 93 million MS/MS spectra from various instruments (including Orbitrap, Ion Trap, qTof, and FT-ICR) have been searched at GNPS, yielding putative dereplication matches of 7.7 million spectra to 15,477 compounds. In the second stage of dereplication, GNPS goes beyond re-identification by utilizing variable dereplication, which is a modification-tolerant spectral library search that is mediated by a spectral alignment algorithm. Variable dereplication enables the detection of significant matches to either putative analogs of known compounds (e.g., differing by one modification or substitution of a chemical group) or compounds belonging to the same general class of molecules (Fig. 3.3b).Variable dereplication is not available through any other computational platform. For example, GNPS variable dereplication has detected compounds with different levels of glycosylation on various substrates. As MS/MS fragmentation preferentially results in peaks from glycan fragments, it is possible to detect sets of compounds with related glycans even when the substrates to which the glycans are attached are themselves unrelated[45]. To date, 3,891 putative analogs have been identified in public data using GNPS variable dereplication (Supplementary Table C.4). These 3,891 putative analogs include several unique molecules that could be user-curated and added to GNPS reference libraries (see Molecular Explorer below on accessing and annotating putative analogs).

To assess the reliability of the MS/MS matches found by GNPS dereplication,

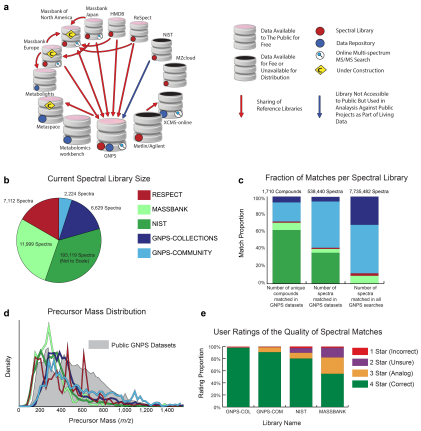


Figure 3.2. GNPS spectral libraries (a) The computational resources of the metabolomics and natural products community fall into two main categories: i) Reference collections (red dots) of MS/MS libraries and ii) Data Repositories (blue dots). Reference collection resources are contributors and aggregators of reference MS/MS spectra, some of which also include data analysis tools, e.g. online multi-spectrum MS/MS search (magnifying glass icon). Several resources have aggregated MS/MS spectra from various reference collections (red and blue arrows). GNPS has imported all freely available reference collections (>221,000 MS/MS spectra) and makes them available for online analyses. GNPS and several other resources provide both reference MS/MS spectra and data in an open and free manner to the public (pink caps). (b) Comparison of spectral library sizes of available libraries (MassBank, ReSpect, and NIST) and GNPS libraries; GNPS-Collections includes newly acquired spectra from synthetic or purified compounds and GNPS-Community includes all community-contributed spectra. (c) Searching all public GNPS datasets revealed that Massbank/ReSpect/NIST libraries matched to 1,217 unique compounds, with GNPS libraries increasing unique compound matches by 41% (corresponding to 29% of total unique matches) with an accompanying 4% increase in spectral library size. Overall, GNPS libraries increase the total number of spectra matched in public datasets by 144% (59% of total public MS/MS matches) and spectra matches across all GNPS public and private data by 767% (88% of all MS/MS matches). (d) The distribution of precursor masses in all GNPS public datasets (gray) and compared to the precursor mass distributions of Massbank, ReSpect, NIST, and GNPS libraries. Though GNPS libraries have a combined size that is smaller than MassBank/ReSpect/NIST, GNPS libraries have a higher proportion of molecules in the higher m/z range (e) The quality of spectrum matches obtained by searching against the available spectral libraries is assessed by user ratings (1 to 4 stars see Supplementary Table C.4) of continuous identification results. User ratings of 2.5+ stars for 98%+ of GNPS library matches compares favorably with the 90% mark for NIST matches.

GNPS users can rate the quality of matches returned by automated GNPS reanalysis (see below). These ratings are 4 star (correct), 3 star (likely correct, e.g. could also be isomers with similar fragmentation patterns), 2 star (unable to confirm the annotation due to limited information) and 1 star (incorrect) (Supplementary Table C.4). So far, of the 3,608 matches that have been rated, 139 (3.9%) matches were given 1 or 2 stars (insufficient information (2.9%) or incorrect (1%)) by user ratings. These percentages are consistent with the false discovery rates estimated using spectral library searches of benchmark LC-MS datasets with compound standards (Appendix C.0.3, Fig. C.1, C.2). Furthermore, these 3,608 match ratings were associated with 2,041 library spectra, therefore the average rating of a library spectrum can offer insight into the reliability of its reference annotation, not unlike Yelp ratings for restaurants. Incorrect matches can arise through either spurious high-scoring matches to library spectra or incorrect annotations for library spectra. Of the 2,041 library spectra with match ratings, 72 (3.5%) spectra had average ratings below 2.5 stars. These percentage ratings were further broken down by spectral library (Fig. 3.2e). We found that for GNPS-Collection and GNPS-Community libraries, only 29 out of 1746 (1.7%) of the rated library spectra had average ratings below 2.5 stars. These ratings demonstrate that the perceived reliability of GNPS spectral libraries compares favorably with established community resources such as NIST and Massbank, in which 10.5% and 20.1% of the ratings were below 2.5 stars respectively, and provides confidence that the community curation process is robust and that third party libraries integrate well with GNPS. The main advantages of searching using GNPS are the option to run simple or variable dereplication against all publicly accessible reference spectra, and that community-rated matches can be used to improve the quality of the reference libraries and matching algorithms. These dereplication capabilities are not possible with existing published resources.

3.4 Molecular Networking

Molecular networks are visual displays of the chemical space present in mass spectrometry experiments. GNPS can be used for molecular networking [72, 102, 75, 89, 97, 104, 27, 110, 36, 70], a spectral correlation and visualization approach that can detect sets of spectra from related molecules (so-called spectral networks[6]) even when the spectra themselves are not matched to any known compounds (Fig. 3.3a). Spectral alignment[102, 36] detects similar spectra from structurally related molecules, assuming these molecules fragment in similar ways reflected in their MS/MS patterns (Fig. 3.3b), analogous to the detection of related protein or nucleotide sequences by sequence alignment. GNPS is currently the only public infrastructure that enables molecular networking. The visualization of molecular networks in GNPS represents each spectrum as a node, and spectrum-to-spectrum alignments as edges (connections) between nodes. Nodes can be supplemented with metadata, including dereplication matches or information that is provided by the user, such as abundance, origin of product, biochemical activity or hydrophobicity which can be reflected in a nodes size or color. It is possible to visualize the map of related molecules as a molecular network[110, 105, 87, 47, 71] (Fig. C.3) both online at GNPS (Fig. 3.3c) or exported for analysis in Cytoscape[87]. Molecular networking analyses of 272 public datasets (Fig. 3.4a) from a diverse range of samples reveals that on average 35.2% of all unidentified nodes are significantly matched to other spectra of related molecules within a cosine score of 0.8 (increasing to 44.7% of all nodes in more exploratory networks with a cosine score of 0.65). This indicates that a large fraction of all unidentified spectra are identifiable if their or their neighboring nodes reference spectra were available in the reference spectral libraries.

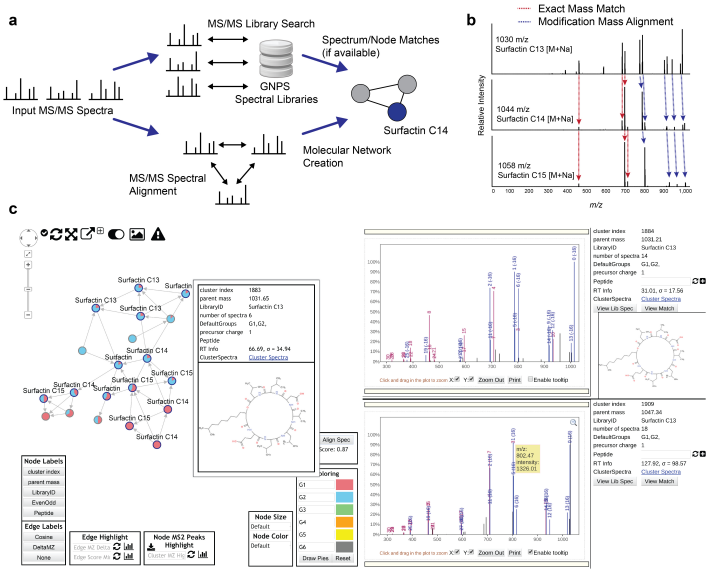


Figure 3.3. Molecular Network Creation and Visualization. (a) Molecular networks are constructed from the alignment of MS/MS spectra to one another. Edges connecting nodes (MS/MS spectra) are defined by a modified cosine scoring scheme determines the similarity of two MS/MS spectra with scores ranging from 0 (totally dissimilar) to 1 (completely identical). MS/MS spectra are also searched against GNPS Spectral Libraries, seeding putative nodes matches in the molecular networks. Networks are visualized online in-browser or exported for third party visualization software such as Cytoscape[87]. (b) An example alignment between three MS/MS spectra of compounds with structural modifications that are captured by modification tolerant spectral matching utilized in variable dereplication and molecular networking. (c) In-browser molecular network visualization enables users to interactively explore molecular networks without requiring any external software. To date, more than 11,000 molecular networks have been analyzed using this feature. Within this interface, (i) users are able to define cohorts of input data and correspondingly, nodes within the network are represented as pie charts to visualize spectral count differences for each molecule across cohorts. (ii) Node labels indicate matches made to GNPS spectral libraries, with additional information displayed with mouseovers. These matches provide users a starting point to annotate unidentified MS/MS spectra within the network. (iii) To facilitate identification of unknowns, users can display MS/MS spectra in the right panels by clicking on the nodes in the network, giving direct interactive access to the underlying MS/MS peak data. Furthermore, alignments between spectra are visualized between spectra in the top right and bottom right panels in order to gain insight as to what underlying characteristics of the molecule could elicit fragmentation perturbations.

3.5 Living Data by Continuous Analysis

Funding agencies and publishers have called for raw scientific data, including mass spectrometry data, and analysis methods to be made publicly available where possible. Consistent with this aim, GNPS datasets usually comprise the full set of mass spectrometry files produced during a NP research project or the full set of spectra analyzed for a peer-reviewed publication (Chapter C.0.4). While it is potentially advantageous to the community for all data to be made public, GNPS user data can remain private until users explicitly choose to make it public (private data is also analyzable and privately sharable, with >93 million spectra in >250,000 private LC/MS runs already searched using GNPS). GNPS has the largest collection of publicly accessible natural product and metabolomics MS/MS datasets and is the only infrastructure where public data sets can be reanalyzed together and compared with each other. To date, GNPS has made 272 public GNPS datasets openly available which are comprised of more than 30,000 mass spectrometry runs with approximately 84 million MS/MS spectra. In common with other public repositories[37, 68], GNPS datasets can be downloaded. However, data availability on its own does not serve to enable data reuse. GNPS is unique among MS repositories by enabling continuous identification: the periodic and automated re-analysis of all public datasets (Appendix C.0.5, C.0.6). This continuous re-analysis, which incorporates molecular networking and dereplication tools, implements a virtuous cycle as illustrated in Figure 3.1a. Because GNPS spectral libraries are constantly growing due to community contributions and continued generation of reference spectra, the number of matches made by successive re-analyses of public datasets has already grown and is expected to continue to grow over time (Fig. 3.4b). GNPS users are periodically updated with alerts of new search results.

For example, a Streptomyces roseosporeus project (MSV000078577) was de-

posited April 8, 2014. At first, only 7 MS/MS spectra were matched. However as of July 14, 2015 36 spectral matches have been made to GNPS libraries. Overall, the total number of compounds matched to GNPS datasets increased more than tenfold, while the number of matched MS/MS spectra in GNPS datasets increased more than twenty-fold in 2015 (Fig. 3.4b). GNPS users can also subscribe to specific datasets of interest, rather like following people on Twitter. When new matches are made, changed, or revoked, all subscribers are notified of new information by an email summarizing changes in identification. From April 2014 to July 2015, 45 updates were initiated by CCMS and automatically sent to subscribers (Fig. C.4). Update emails have led to substantially more views per dataset, compared to non-GNPS datasets (192 proteomics datasets) deposited in MassIVE. Continuous identification not only keeps a single dataset alive, it can create connections between datasets and users over time. Similarities between datasets could form the basis of a data-mediated social network of users with potentially related research interests despite seemingly disparate research fields, rather like the People You May Know feature on LinkedIn. On average each GNPS user already has 5 suggested collaborators (Fig. C.5).

3.6 Molecular Explorer

Molecular Explorer is a feature that can only be implemented on living data repositories and thus exists only in GNPS. Molecular Explorer allows users to find all datasets and putative analogs that have ever been observed for a given molecule of interest. We anticipate that this feature could guide the discovery of previously unknown analogs of existing antibiotics. Public NP data contains more than one hundred unidentified putative analogs of antibiotics such as valinomycin, actinomycin, etamycin, hormaomycin, stendomycin, daptomycin, erythromycin, napsamycin, clindamycin, arylomycin, and rifamycin, highlighting a clear potential to generate leads to discover structurally related

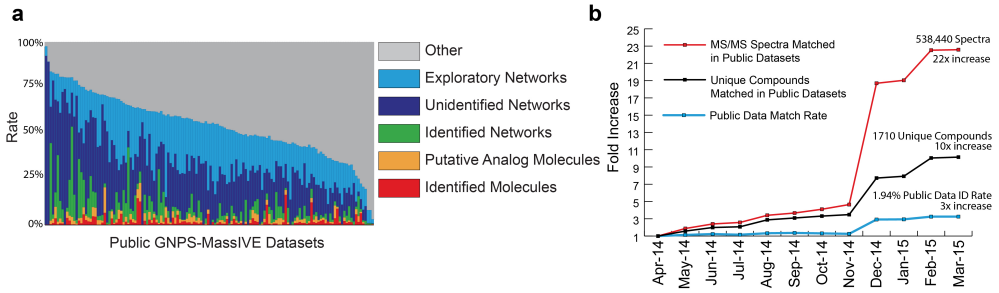


Figure 3.4. Living data in GNPS by crowdsourcing molecular annotations. (a) A global snapshot of the state of MS/MS matching of public natural product datasets available in GNPS using molecular networking and library search tools. Identified molecules (1.9% of the data) are MS/MS spectrum matches to library spectra with a cosine greater than 0.7. Putative Analog Molecules (another 1.9% of the data) are MS/MS spectra that are not identified by library search but rather are immediate neighbors of identified MS/MS spectra in molecular networks. Identified Networks (9.9% of the data) are connected components within a molecular network that have at least one spectrum match to library spectra. Unidentified Networks (25.2% of the data) are molecular networks where none of the spectra match to library spectra; these networks potentially represent compound classes that have not yet been characterized. Exploratory Networks (an additional 20.1% of the data) are unidentified connected components in molecular networks with more relaxed parameters. Thus, 55.3% of the MS/MS spectra at least have one related MS/MS spectrum in spectral networks, with 44.7% having none. In this 44.7% of the data, each MS/MS spectrum has been observed in two separate instances and should not constitute noise. Altogether, this analysis indicates that most of the chemical space captured by mass spectrometry remains unexplored. (b) In the past year, there has been significant growth in the GNPS spectral libraries, driving growth in the match rates of all public data. The number of unique compounds matched in the public data has increased 10x; the number of total spectra matched has increased 22x; and the average match rate has increased 3x. It is expected that identification rates will continue to grow with further contributions from the community to the GNPS-Community spectral library.

antibiotics though the application of GNPS (Fig. C.6, and Appendix C.0.7).

3.7 Conclusion

GNPS provides a community-led knowledge space in which NP data can be shared, analyzed and annotated by researchers worldwide. It enables a cycle of annotation, in which users curate data, continuous dereplication for product identification, and houses

a knowledge base of reference spectral libraries and public datasets.

The transformation of deposited spectra into living data that is enabled by the GNPS platform could mediate connections between researchers and has the potential to transform data networks into social networks. Of 1,272 compound identifications obtained by continuous identification with the GNPS-Community library, 1,063 (83.6%) were made using reference spectra that were not uploaded by the submitter. In other words, the vast majority of identifications were enabled by other community members. This reuse of knowledge and data is analogous to other community-wide curation efforts including Wikipedia and crowd-sourced dictionaries. Since their initial deposition, 59% of datasets have an increased number of identifications, with the average dataset more than doubling the number of identifications since submission (Fig. C.7). GNPS enables facile sharing of individual analyses and uses molecular networks to reveal connections between datasets from different laboratories and biological sources that would otherwise remain disconnected. To date, 3,145 analysis jobs have included files shared between GNPS users, encompassing 548 unique pairs of individuals collaborations. GNPS recasts public datasets as conversation starters in a data-mediated social network.

As previously shown in genomics[7] and protein structure analysis[8], the models of global collaboration and social cooperation that are present in GNPS could empower scientific communities to collectively translate big data into shared, reusable knowledge and profoundly influence the way we explore molecules using mass spectrometry.

Chapter 3, in part, of the material as it appears in Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking in Nature Biotechnology. Wang, M, Carver, JJ, Phelan, VV, Sanchez, LM, Garg, N, Peng, Y, Nguyen, DD, Watrous, J, Kapono, CA, Luzzatto-Knaan, T, Porto, C, Bouslimani, A, Melnik, AV, Meehan, MJ, Liu, WT, Crsemann, M, Boudreau, PD, Esquenazi, E, Sandoval-Caldern, M, Kersten, RD, Pace, LA, Quinn, RA, Duncan, KR, Hsu, CC, Floros,

DJ, Gavilan, RG, Kleigrewe, K, Northen, T, Dutton, RJ, Parrot, D, Carlson, EE, Aigle, B, Michelsen, CF, Jelsbak, L, Sohlenkamp, C, Pevzner, P, Edlund, A, McLean, J, Piel, J, Murphy, BT, Gerwick, L, Liaw, CC, Yang, YL, Humpf, HU, Maansson, M, Keyzers, RA, Sims, AC, Johnson, AR, Sidebottom, AM, Sedio, BE, Klitgaard, A, Larson, CB, Boya P, CA, Torres-Mendoza, D, Gonzalez, DJ, Silva, DB, Marques, LM, Demarque, DP, Pociute, E, O'Neill, EC, Briand, E, Helfrich, EJ, Granatosky, EA, Glukhov, E, Ryffel, F, Houson, H, Mohimani, H, Kharbush, JJ, Zeng, Y, Vorholt, JA, Kurita, KL, Charusanti, P, McPhail, KL, Nielsen, KF, Vuong, L, Elfeki, M, Traxler, MF, Engene, N, Koyama, N, Vining, OB, Baric, R, Silva, RR, Mascuch, SJ, Tomasi, S, Jenkins, S, Macherla, V, Hoffman, T, Agarwal, V, Williams, PG, Dai, J, Neupane, R, Gurr, J, Rodriguez, AM, Lamsa, A, Zhang, C, Dorrestein, K, Duggan, BM, Almaliti, J, Allard, PM, Phapale, P, Nothias, LF, Alexandrov, T, Litaudon, M, Wolfender, JL, Kyle, JE, Metz, TO, Peryea, T, Nguyen, DT, VanLeer, D, Shinn, P, Jadhav, A, Mller, R, Waters, KM, Shi, W, Liu, X, Zhang, L, Knight, R, Jensen, PR, Palsson, B, Pogliano, K, Linington, RG, Gutirrez, M, Lopes, NP, Gerwick, WH, Moore, BS, Dorrestein, PC, Bandeira, N (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol., 34, 8:828-37. The dissertation author was the primary investigator and author of this paper.

Chapter 4

The Mass spectrometry Interactive Virtual Environment (MassIVE) for sharing and community reanalysis of proteomics data

4.1 Introduction

Public data repositories for mass spectrometry based proteomics have grown immensely in the last decade with the proteomexchange consortium[21] coordinating the requirements of its member repositories, specifically the PRoteomics IDEntifications (PRIDE)[96], JPOSTrepo[77], and this resource the Mass Spectrometry Virtual Interactive Environment (MassIVE) at the Center for Computational Mass Spectrometry (CCMS) at the University of California San Diego. Even though the total amount of mass spectrometry data available at these public repositories has swelled to 106 TB for human data (229TB total), only 9TB (10%) included identifications and even less (4TB) have controlled false discovery rates (FDR). This severe shortage of identifications limits the utility and reusability of data for the broader biological sciences community.

Resources such as the Global Proteome Machine Database (GPMDB)[14] and PeptideAtlas[19] aimed to remedy problem through standard reprocessing pipelines

and present a distilled version of the identifications. However, these approaches fall short as the search workflows and results are not presented to the users, thus missing provenance of the identifications. This obfuscation precludes the community from making a critical assessment of the claims made by these resources (GPMDB and PeptideAtlas) on individual datasets.

MassIVE picks up where both primary data repositories and reprocessing resources left off. First MassIVE enables data deposition with visualization of all results as well as automated re-analysis of public deposited data. Second, MassIVE aims to give a voice to the community and democratize the re-analysis of these data.

Currently, all analyses presented in proteomics data repositories come from only two sources: the original dataset submitter or a reprocessing resource (GPMDB and PeptideAtlas). Even though MassIVE has endeavored to reanalyze a large swath of the public proteomics data with standard pipelines, having only repositories performing broad fixed analyses limits the progress that can be made in understanding a specific dataset. For example, the human draft proteomes[48, 103] were published in 2014 whose claims were critically assessed in subsequent publications[28]. These follow-up re-analyses do not share a life cycle with the original datasets and scientists in the future reusing these key datasets will miss out all the community conversation and accumulated knowledge. MassIVE supports the inclusion of community sourced re-analyses with each dataset, thus facilitating the accumulation of knowledge that is a prerequisite for creating a community consensus(Fig. 4.1).

While several laboratories possess the infrastructure and software pipelines to re-process terabytes of mass spectrometry data, much of the proteomics community does not – limiting the re-utilization of much of the public data available. MassIVE aims to empower more of the community to re-analyze existing data to help both the community reach a consensus about the data as well as further their own work. Backed

by a cloud scale computational infrastructure at the CCMS, MassIVE provides a set of analysis workflows to interrogate mass spectrometry data. Today, by offering database search (MSGF+[51]), proteogenomic search (ENOSI[107]), spectral library search (MSPLIT[99]), blind modification search (MODa [73]), SWATH library search (MSPLIT-DIA[100]), and molecular networks analysis (Maestro), MassIVE begins to give the community analysis flexibility. Additionally, these analysis workflows are seamlessly integrated with dataset re-analyses, i.e. the community can attach any search results to the associated dataset. These features thus recasts public datasets not as static entities, but rather dynamic and evolving focal points of knowledge aggregation.

The MassIVE repository infrastructure presented here highlights an evolution in the role of a data repository. Cued by the success of Global Mass Spectrometry Social Molecular Networkings[101] approaches for the metabolomic and natural products space, we envision MassIVE will create a more collaborative data centric community fostering more critical analyses of all the public data.

4.2 Data Sharing

4.2.1 The Mass Spectrometry Virtual Interactive Environment

The MassIVE mass spectrometry data repository provides support for the proteomics community to deposit their raw/processed data and results. Submissions fall into two categories, partial and complete. While both types of submissions include spectra, complete submissions require that spectra be available in open formats (e.g. mgf, mzXML[80], mzML[67]) along with accompanying identification results (e.g. mzID[44], mzTab[34]) that can support claims made in a publication.

To facilitate complete dataset deposition, file format conversions to open formats are provided for both mass spectrometry (to mzML) as well as identification files

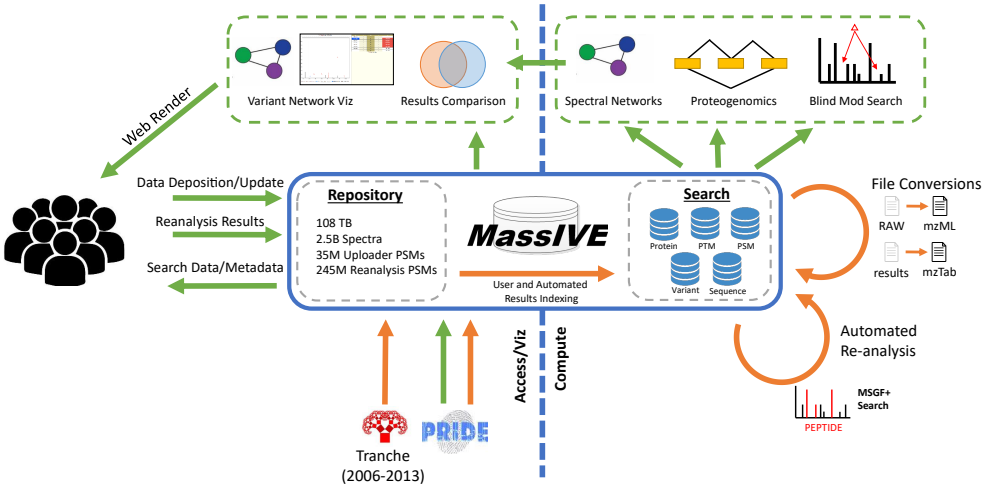


Figure 4.1. Overview of MassIVE ecosystem. Datasets are imported from Tranche, PRIDE, and deposited by the proteomics community. Each dataset in MassIVE is automatically processed to open data formats (mzML). Users can do in depth analysis on one or a combination of datasets with advance computation workflows (e.g. Database search, blind modification search, proteogenomics analysis, spectral networks analysis, etc.). These analyses can be further visualized and compared to existing analyses for the respective dataset at the spectrum, peptide, and protein levels. MassIVE additionally automatically analyzes each dataset with a baseline database search (MSGF+) and attaches each as a reanalysis for subsequent interrogation and search. All results attached to datasets are imported into the MassIVE Index, enabling the community to query on the 280M PSMs already indexed.

(to mzTab). Currently, mass spectrometry conversions support Thermo RAW, SciEx IDA, and SciEx SWATH data. Additionally, we support the conversion of standard tab separated files (tsv) and mzIdentML identification files into mzTab.

Private MassIVE datasets are given password-protected access to the data for reviewers, allowing them to visualize, download, and reanalyze the data.

As of May 20, 2017, MassIVE serves 7,806 datasets, containing 108TB of mass spectrometry data. These 7,806 datasets includes 6,324 Tranche recoveries[90], 639 Proteomexchange[21], 510 GNPS[101], and 333 other datasets.

4.2.2 Complete Datasets and Provenance

The Proteomexchange consortium[21] aims make data public and reusable by encouraging the publishing of datasets that include identifications (peptide and protein), as complete datasets. As a proteomexchange consortium member, MassIVE supports these goals by offering both complete and partial (lacking identifications) submissions. To facilitate user submissions of complete datasets, MassIVEs web user interface allows users to select data, enter metadata, and automatically reconcile filename mappings in results files. After deposition, complete datasets present online pages for all PSMs, peptides, and protein identifications – including visualizations of peptide spectrum matches. In difference from all other online resources, for each PSM, a full provenance record of each identification links back to a spectrum file and search task (including parameters if analyzed at CCMS).

Of all 1,412 Proteomexchange human datasets, only 330 (23%) are complete. However, this accounts for only 13% of the total data size. Furthermore, many of these complete datasets represent their identifications in formats that breaks the connection back to the raw data and presents issues for external resources to parse the results (e.g. prideXML). From 330, only 175 datasets (12% datasets and 9% of data volume) provide results in an open format connected back to the original peak files. However, the vast majority (136) of datasets contained identifications where the PSM level FDR was either high ($<$1%) or could not be calculated. Out of a total of 40M reported PSMs in complete datasets, only 15M were in search results that were reported under 1% PSM level FDR (representing 4TB of data) . Of these 15M, 10M (66%) came from the CPTAC colorectal set of datasets. With only 4TB of data containing error controlled identifications (of 106TB of human data), there is a need to begin providing identifications for the public proteomics data with a tightly controlled FDR much more broadly. The MassIVE

repository addresses this need by enabling the addition of identifications from 1) the dataset submitter, 2) automated reanalyses of public proteomics data, and 3) reanalyses generated by the community.

4.3 Live Data

4.3.1 Dataset Owner Updates/Upgrades

At the point of initial deposition, datasets are almost always incomplete (e.g. files and/or identifications are missing). To provide a path to completeness, MassIVE enabling dataset owners to augment their datasets after deposition. For example, the MSPLIT-DIA dataset[100] (MSV000079200) was originally deposited at MassIVE on July 27, 2015, but after publication the authors amended the dataset on July 28, 2016 with identifications supporting the publication, upgrading the dataset to complete.

As a repository, the need arises to update and enrich key datasets for the community, for example, the human draft proteome from the Pandey laboratory[48] (PXD000561 - MSV000079514) was initially deposited as a partial dataset at a ProteomeXchange repository (PRIDE). As part of MassIVEs import process, we have restructured the data and identifications in three major ways. 1. Organized data from a unstructured flat file layout into one that more accurately reflects the metadata and experimental parameters, i.e. organized by tissue, instrument, and experiment. 2. The datasets proprietary raw format were converted into an open format to facilitate its reuse and utility for the community. 3. We upgraded this dataset to a complete submission by painstakingly converting pepXML to mzTab and reverse engineering the filenames present in the identification files yielding over 10M PSMs as originally reported.

Dataset updates can also be useful even for complete datasets designed for sharing. For example, the CPTAC proteomic study of human colon and rectal cancer[113] contains

95 patient samples. This dataset was initially deposited as ten separate datasets and we reorganized them into a single dataset logically organized by patient (both the data and identifications) and further included 30 CPTAC normal colon epithelium samples as controls. This reorganized and aggregated dataset, complete with original identifications from the publication vastly improves navigability and facilitates the validation of the results (MSV000079852).

By taking all these steps to reorganize, enrich, and upgrade key datasets for the community, we have made it possible to explore the original claims made in their publication and enable the community to reanalyze/reuse these data/results in their own research.

4.3.2 Dataset Re-analyses of Datasets

Despite MassIVEs efforts to enrich key datasets for the community, the vast majority of public proteomics data lacks identifications (>90%). A second source of identifications is from the community. While each dataset can have an official set of identifications deposited/updated by the dataset submitter, we give the opportunity to the community to have a voice about what can be found in datasets, presented in a public and consistent manner. That is, MassIVE allows the community to attach their own reanalysis of public data – containing a description and references to open format spectrum files of the dataset that provides provenance and visualization of identifications.

While it is clear that there might arise conflicting interpretations of the same MS/MS data, the publishing of a re-analysis of a dataset is one step in teasing these interpretations into the open. Without a common platform to present and interrogate conflicting claims, reaching consensus is impossible. By giving such a platform for the community to express their insights and a method to compare differences in claims (see results comparison), we provide the tools necessary even to begin to reach a data

supported consensus.

4.3.3 Broad Re-analysis of Partial Datasets

We aimed to automatically reanalyze the public proteomics data in a error controlled and consistent manner. We acknowledge that other computational resources such as PeptideAtlas and GPMDB also perform reprocessing of publicly available proteomics data, but the identification results are not made available for each dataset, thus preventing a growth of knowledge of a given dataset. While PeptideAtlas has attempted in the past to perform this, only five such reprocessings exist and the results are only available by FTP access in zipped files rather than browseable online. In difference, MassIVE reanalysis results are available and visualizable online as well as available in open identification formats (mzTab).

First, all spectrum files for datasets are converted to open formats amenable to MS/MS bioinformatic analysis (centroided mzML). To date, this process has processed 47 TBs (1.06B MS/MS spectra) of human proteomics data and will continue on a periodic basis. Second, all human datasets were searched against the UniProt human reference database (May 23, 2016) with MSGF+ at 1% PSM level FDR . Results were attached to the appropriate dataset as a reanalysis. These reanalyses created error controlled identifications for 34 TB (770% increase) from 226 datasets (480% increase) with 245M PSMs (1600% increase). These reanalyses span the several major instrument vendors, e.g. Thermo QExactive (PXD003903 - MSV000080004), Thermo Velos/Elite (PXD000561 - MSV000079514), Sciex TripleTOF 5600 IDA (PXD000953 - MSV000079593), Sciex TripleTOF Sciex TripleTOF 5600 SWATH-MS (PXD001064 - MSV000079937), Bruker qTOF Impact II (PXD001592 - MSV000079935), etc, and show a significant increase in PSMs for each of these instrument types (Fig. 4.2a).

This increase in annotation across such a breadth of data provides a baseline

set of identifications for each dataset to allow the community to have a conversation. With a standardized analysis pipeline, comparisons of identifications and claims between datasets is now possible. While these baseline reanalyses do not aim to provide definitive evidence of existence of proteins, they do allow for the generation of hypothesis for where certain peptides and proteins might occur. For example, there are 19,314 proteins that putatively occur in greater than ten datasets and based on the metadata of each of these datasets, hypothesis can generated and further validated. Finally, these standardized re-analyses of a wide variety of datasets facilitates comparisons.

4.3.4 MassIVE Search

To explore the 280M PSMs across hundreds of datasets, we introduce MassIVE Search that allows querying across all error controlled ($<$1% PSM FDR) dataset identifications and re-analyses. Across the entire MassIVE repository, 280M PSMs, 10M variants, 5.9M sequences, and 260 modifications from 12,764 individual results were imported. All PSMs provide full provenance back to the original dataset and/or reanalysis and enables visualization for the identifications made. We envision this type of query can allow users to find where a variant, sequence, modification, or protein occurs. For example, the proteogenomic SNP event sequence: SALFAQINQGESITHALK can be found in 562 searches from 29 datasets (including both cancer and normal CPTAC data).

4.3.5 In Depth Re-analysis

4.3.6 In Depth Re-analysis

While the baseline set of MSGF+ database search identifications provides a starting point for the knowledge of a dataset, we also empower the community to expand this knowledge by customizing their own reanalysis utilizing MassIVEs online analysis workflows. As a starting point, users can reproduce and tweak the MSGF+

baseline searches and customize the search parameters, e.g. proteome database, mass tolerance values, post-translational modifications, etc. Further, officially supported analyses workflows include ENOSI (proteogenomic search), MODa (blind modification search), M-SPLIT (mixture spectra library search), and Maestro (spectral networks). Applied to a CPTAC Colorectal cancer dataset[113] (MSV000079852) revealed 260 modification 211 modification masses being observed more than 50 times, 19,728 proteogenomic events, and 386K spectra as mixtures. Analyzing a single patient's data with Maestro (TCGA-AA-3518-01A-11) uncovered a high diversity of variants of peptides in the sample. 181 peptides have five or more variants, with a maximum of 35 variants per peptide (Fig. 4.2b).

While all these workflows individually allow for detailed interrogation, further validation by corroborating identifications is difficult. For example, in order to validate the novel proteogenomic events, e.g. a proteogenomic nonsynonymous single nucleotide polymorphism event on the peptide SALFAQINQGESITHALK is corroborated by SSL-16FAQINQGESITHALK in MODa, to another search, e.g. MODa, can be a laborious process. With so many different analysis tools looking at the same data, where the majority of the easy identifications are identical, a comparison tool is necessary to both at a glance and in depth explore what new identifications and biology can be gleaned from these differences.

4.3.7 Dataset Updates - Broadcasting

Further, any updates/reanalyses to datasets are also reported to proteomexchange enabling the whole proteomics community to learn of new knowledge being deposited. Re-analyses of a dataset are announced to the community via a proteomexchange re-processed accession (RPXD). Additionally, datasets imported into MassIVE from other proteomexchange consortium members (e.g. PRIDE) which are updated/reanalyzed at

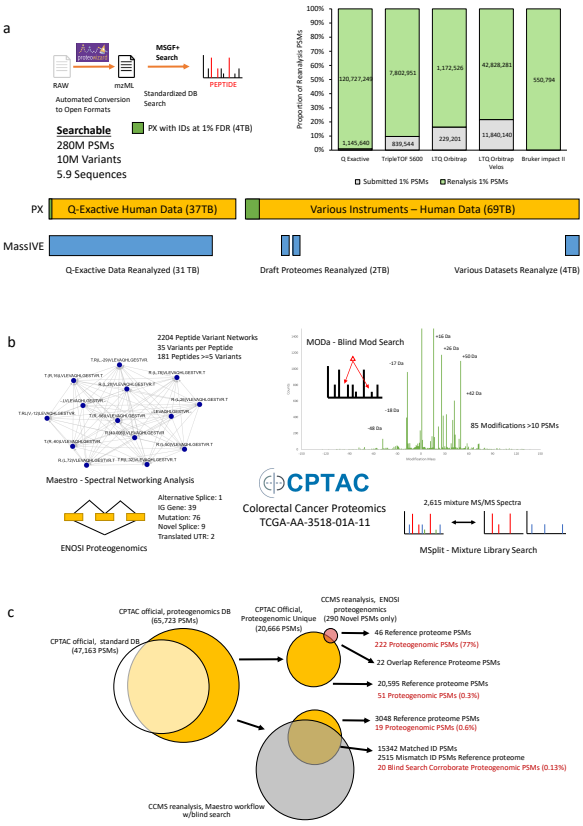


Figure 4.2. MassIVE Reanalysis (a) Even though human proteomics data has swelled to 106TB, scarcely 4TB contain reusable identifications with proper error controls. MassIVE has taken the steps to automatically convert from proprietary vendor formats to standardized open formats (mzML) and follow up with a standardized analysis pipeline. The total volume of data with identifications has expanded 16x along with a growth of 15M PSMs to 280M PSMs. (b) Powerful computational workflows from the CCMS (e.g. proteogenomics search, blind search, mixture library search, and spectral networks) enable deep analysis of a specific dataset. When applied to a single patients data from the CPTAC (TCGA-AA-3518-01A-11) project, 129 proteogenomic events, 260 blind modifications, 2,635 mixture spectra, and thousands of peptide variants. C) MassIVEs Results comparison visualizes and pinpoints unique identifications and disagreements between search results. For example, in comparing the CPTAC official standard DB results to the CPTAC official proteogenomic results, revealed 20,666 unique PSMs from the proteogenomic searches. However, comparing these 20,666 PSMs to ENOSI proteogenomic and Maestro results revealed that that only a handful could be proteogenomic events for the majority of these PSMs were already found to be from the UniProt Reference Proteome. Finally, 20 of these proteogenomics PSMs (denoted as SNVs) could be corroborated by concordant identifications in Maestros blind search results.

MassIVE will announce to the community with new identifications and discoveries.

4.4 Data Community

While MassIVE has endeavored to create a baseline set of identifications for all datasets, in many cases the default search tools and parameters can short of being the most sensitive tool. For example, iPRG 2012[12] challenged the community to identify a complex mixture of modified peptides. Identification results from 24 laboratories examining identical MS/MS spectra from a Sciex 5600 qTof yielded drastically different sensitivities (Fig. 4.3a). Further, the agreement of of a particular submissions results to the consensus identification varied from 60% to 99% (Fig. 4.3a) and the consistency of a consensus identification varied from 50% to 100% (Fig. 4.3b). Due to this great variability in results from different computational analysis pipelines, the community must be made aware of this fact. However, these 24 results were hidden in cryptically organized spreadsheets, making validation and comparison impossible. They were painstakingly converted to open formats and added as a re-analysis for the iPRG 2012 dataset (MSV000078492), giving transparency to the evaluation process.

To reiterate the need for the community to contribute their own reanalyses, a SILAC datasets[43] (MSV000080791) PSMs identified increased by 42% and precursors identified increased by 46% when searched with the sample specific SILAC modifications compared to the standard MSGF+ broad analysis workflow (Fig. 4.3c). Further, for a phosophoproteomic study[78] (MSV000080773), including appropriate phosphorylation modifications increased spectrum and precursors identification rates by 55% and 60% respectively (Fig. 4.3c).

4.4.1 Reaching Consensus - Results Comparison

With the huge variety of methods to analyze data and the multiple identification results for public datasets, e.g. CPTAC Colorectral cancer dataset (MSV000079852), a detailed comparison of identification results is necessary to determine what is new, conflicting, and consistent. We address this need with an online results comparison tool that reveals differences at the PSM, peptide, and protein levels. This can be applied intra-dataset and inter-tool to determine gains in sensitivity and corroborate extraordinary claims. Here, identification disagreements at the spectrum level can be uncovered in order to assess error in identifications or determine corroborating evidence. For example, in comparing proteogenomics analysis (ENOSI) and blind modification search for a CPTAC patient, it can be quickly ascertained that 275 PSMs from novel proteogenomic events break down into 64 matching PSMs, 168 mismatching PSMs (same spectrum, different identifications), and 89 unique PSMs to proteogenomics.

We applied our online results comparison CPTAC official proteogenomic results for patient TCGA-AA-3518-01A-11. In comparing to the standard database search, Maestro reanalysis, and ENOSI reanalysis, revealed that out of a total of 65,723 PSMs in the published results, only 39 remaining PSMs could be novel proteogenomic events (Fig. 4.2c).

Further, overlap and uniqueness of peptides and proteins (can be remapped to a common proteome) are determined. Inter-dataset comparisons reveal uniqueness at the peptide and protein levels of each result set. This can reveal biological and experimental variation by highlighting expression differences between different conditions, tissues, and organisms.

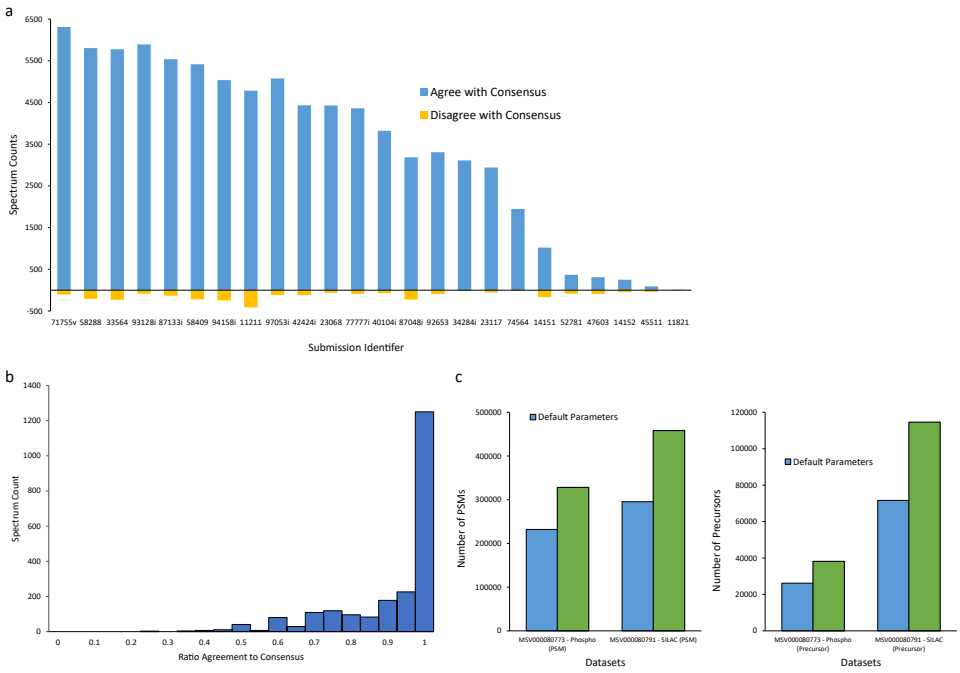


Figure 4.3. MassIVE Community (a) 2012 iPRG community identifications. The wide variance in number of identifications and deviations from the consensus identification illustrates the need for the community to be involved in providing identifications for data. (b) While the majority of consensus identifications have high agreement from community identifications, a large tail exists and these conflicts will need to resolved and explored with MassIVEs Results Comparison. (c) Understanding a datasets metadata can have a large influence on the identification rate of a dataset. By correctly choosing a database searchs modifications (e.g. SILAC or Phosphorylation) PSMs increased by up to 55% and precursors increased by up to 60%.

4.5 Discussion

We present the mass spectrometry data repository MassIVE. Data repositories have long been static institutions that only serve data. However, we envision that data should take on a life of its own after deposition that continues to bring value to the community in a more active way. By enabling automated reanalysis, user-initiated reanalyses, and whose results can be shared and indexed, MassIVE transforms the communitys past data into a reusable resource. The community can take advantage of

these reusable data and results, with the powerful analysis/comparison/search tools at MassIVE to corroborate, validate, and guide future discoveries.

4.6 Methods

4.6.1 Automated Conversions

All MassIVE datasets provide open format (mzML) of all proprietary vendor formats. Files are converted automatically by the MassIVE system on a periodic basis through msconvert in Proteowizard. These converted spectrum files are presented in a special collection for each dataset called ccms_peak.

4.6.2 Automated Searches

Public datasets are automatically re-analyzed at CCMS with a standard MSGF+ search workflow. Spectra are searched against Uniprot Human Reference (May 23rd, 2016) and are filtered to a 1% PSM level FDR. Common modifications (Oxidation, Pyro-Glu, Deamidation, N-term Acetylation) are included in the search and for high resolution instruments a precursor tolerance of 10ppm is used.

4.6.3 Results Comparison

Two sets of results (A and B) are compared at the spectrum, peptide, and protein level. At the spectrum level, results are categorized as matching, mismatching, unique A, unique B. PSMs that identify the same spectrum (filename and spectrum identifier match) with identical annotations are considered a match. When the annotations are not identical it is a mismatch. Unique PSMs are to spectra that were not identified by the other search. Peptides of a search are determined as the set of unique precursors present in the PSMs. The set of unique A, unique B, and intersection are presented. Proteins are what the search results reported in the resulting mzTab file. Similarly, unique and

overlapping proteins are reported. Additionally, peptides may be remapped to a common proteome in order to compare proteins in the same name space. Unique and overlapping proteins again are reported.

4.6.4 Results Estimation of FDR

The FDR of a search result file is calculated in one of the following ways:

1. If Q-value column present in the identifications file, maximum q-value of PSMs reported. Maximum q-value of passThreshold PSMs if input were mzIdentML
2. If decoys are present in results set, total number of decoys divided by the total number of targets (only including passThreshold PSMS if input were mzIdentML)
3. User entry as part of dataset submission or re-analysis attachement

If none of these conditions apply to an mzTab file, the FDR is undefined.

Chapter 4, in part is currently being prepared for submission for publication of the material. The Mass spectrometry Interactive Virtual Environment (MassIVE) for sharing and community reanalysis of proteomics data. Mingxun Wang, Jeremy Carver, Nuno Bandeira. The dissertation author was the primary investigator and author of this material.

Chapter 5

Automated Knowledge Aggregation for Repository Scale Peptide Spectral Libraries

5.1 Introduction

To map and understand the human proteome is essential in characterizing biological processes, improving health, and fighting disease. Despite several large scale proteomics efforts [48, 103, 83] to create draft maps of the human proteome, it remains impractical for a single study to be fully comprehensive. The enormous diversity of tissues, cell sub-types, and experimental conditions requires the entire communitys involvement. By leveraging the breadth and depth of the whole fields investigations, can we begin to build associations of peptide/protein observations to metadata charting the expression profile of these biomolecules. Further, any comprehensive map of the human proteome must include detailed provenance for all observations and should be delivered as a verifiable reusable collection (i.e. a spectral library) of reference spectra to drive the detection in both shotgun and targeted proteomics experiments, enabling biological discoveries.

These goals have motivated the construction of large scale spectral libraries (PeptideAtlas[30], NIST, GPMDB[15]), but unfortunately existing resources were not

designed to scale automatically with the growing availability of public human data. Thus, the requirement of manual intervention to update these resources has resulted in outdated collections that lack detailed provenance records linking library entries to the raw data and searches used to construct the libraries. The absence of provenance records and algorithmic details is especially relevant as the field evolves to require more stringent criteria to control false discovery rates at multiple levels for proteomics discovery.

We propose a new approach and MassIVE-KB resource for the reusable aggregation of community-scale detection of peptides and proteins observations. Our approach is composed of multiple steps designed to scale its applicability to all public human proteomics data (Fig. 5.1a). First, our approach is designed to independently search new data as it becomes available using only open source algorithms and search procedures. Second, we build on advanced statistical controls for peptide identifications to automatically adjust the effective search space for each dataset, thereby capturing the statistical advantage of simpler samples while retaining stringency over the whole range of sample complexity. Third, we impose additional statistical controls for the community-scale aggregation of peptide identifications, thereby avoiding the uncontrolled accumulation of false discoveries. Fourth, our online approach guarantees that incremental extension of the library yields the same results as full-scale reconstruction, thereby eliminating the need for time consuming reconstruction when integrating newly available data. Fifth, all identifications in the library are supported by detailed provenance records linking each identification to not only the source raw data but also the details of the search procedure used to derive the results (Fig. 5.1b). Sixth, MassIVE-KB is readily reusable with existing search tools, including both traditional data dependent (MSPLIT[99]) and more recent data-independent (MSPLIT-DIA[100]) acquisition modes as well as other popular tools such as Skyline[62], SpectraST[58], and ProHits[61]. Lastly, a publicly accessible pipeline that enables the community to create their own spectral libraries.

With human HCD proteomics data growing to over 37 TB in public data repositories (Q-Exactive is the largest fraction of human data with over twice as much data as the next instruments in ProteomeXchange[21]), we demonstrate MassIVE-KB by have reanalyzing this data and distilling it into community-scale reusable HCD spectral library resource. The reusable collection included 2.1M precursors from 1,088,645 unique sequence, expanding over previous state of the art HCD libraries by delivering 3x more precursors and 2x more sequences (Fig. 5.2a). Furthermore, MassIVE-KB expands the set of observed protein products by providing evidence for 723 proteins previously lacking evidence of protein translation, as well as reporting the expression of 16,801 proteins across 27,404 mass spectrometry experiments (out of 19,610 total proteins covered in the current release of MassIVE-KB).

5.2 Aggregation/Search of Human Proteomics Data

The MassIVE-KB human HCD spectral library is built from 227 human datasets available at the Center for Computational Mass Spectrometry's (CCMS) data repository, the Mass spectrometry Interactive Virtual Environment (MassIVE). Together these datasets contain 658M MS/MS spectra from 27,404 LC/MS runs totalling 31.4TB - approximately 85% of all Human HCD Q-Exactive data in ProteomeXchange[21] - grouped into three broad experimental categories. 1. Whole proteome analysis, 2. Limited Protein Purification (e.g. Affinity Purification Mass Spectrometry) 3. Synthetic peptide pools[116] (Fig. 5.1b). This gradient of sample complexity is modeled in the search search strategies used to analyze these data (Fig. D.1). We aimed to maximize identification sensitivity by ensuring that the confidence of a spectrum identification is based not only upon the peptide spectrum match score, but also in relation to the search space of a given spectrum (as dictated by the experimental procedures). By searching these data with the open source database search tool MSGF+[52], the proper adjustment

of the search space improved the database p-value scores up to three orders of magnitude (Fig. D.2a).

False discovery rates (FDR) in the MassIVE-KB are initially controlled at the level of the individual dataset searches (1% PSM-level FDR) but also need to be controlled at the aggregate level in the final spectral library to avoid an otherwise unacceptable accumulation of errors into a library-scale 28% precursor-level FDR. MassIVE-KB requirements for global precursor error controls are in difference to current library resources, e.g. PeptideAtlas[30] applies PSM level FDR controls at a per dataset level and protein FDR globally, but no explicit controls were described to assess the precursor FDR used for library creation. Similarly, GPMDB[15] relies on an expectation value for each peptide but there is no explicit FDR calculation showing whether the threshold needs to be adjusted. NISTs libraries[23] also do not provide guarantees for library level false discovery rates at the PSM, precursor, peptide, and protein level.

Searching this HCD human data generated 161,510,925 PSMs. Synthetic peptide pool PSMs were filtered to 0% empirical FDR in each individual search and at most 1% PSM level FDR in other searches. In aggregating all search results, 1% local FDR by length at the precursor level and 1% FDR at the protein level [86]. From the 161M candidate PSMs, spectrum quality (see Filtering Methods), precursor FDR, and protein FDR filters led to 2,176,235 precursors at an estimated 0.1065% global precursor FDR (Fig. 5.2 a).

5.3 Construction and Spectrum Quality of Peptide Spectral Library

For each precursor, the representative library spectrum was chosen from the top 100 scoring PSMs (capped at 20 PSMs per dataset to prevent over-dominance of one dataset over others) - 30,160,134 M spectra (universally 0% FDR in the original

search estimated by target decoy). The representative library spectrum was selected to be the most similar spectrum to all others from the same precursor (Fig. D.5 and Fig. D.4b). Of the representatives selected, 1,340,043 came from general proteomics data, 647,797 came from affinity purification, and 188,395 came from synthetics peptide pools, demonstrating the complementarity of these data sources.

To combat potentially false identification of MS/MS spectra to insufficiently distinguishable annotations via database search, all library spectra were re-searched with MSGF+[52] against the reference proteome. PSMs for each spectrum were retained that exceeded the 1% PSM FDR threshold in the spectrums original search. Library spectra with more than one distinct annotation are considered ambiguously identified and are removed from the library (2.5% of precursors), reducing the library to 2,122,890 precursors (Fig. D.4c) called the MassIVE-KB.

To estimate the quality of representative spectrum selection, we created a gold standard library spectrum from stable isotope labeled amino acids in cell culture (SILAC) samples[33]. 2,622,224 MS/MS spectra from 30 HCD LC/MS [18] runs were searched and filtered with strict spectral quality requirements and consistency of fragmentation of concordantly identified co-eluting light and heavy spectral pairs ($<$1% psm level FDR)[108]. The resulting library of light peptides (containing 29,574 precursors) was compared against NIST's HCD spectral library and CCMSs HCD library. The two libraries demonstrated nearly identical similarities to each other and to the expected similarity distribution between reference library spectra (Fig. D.4a), highlighting that the MassIVE-KB reference spectrum quality is on par with the state of the art libraries.

The MassIVE-KB libraries also aim to be augmented in an online basis as new public human data becomes available. The library addition procedure is order agnostic and is deterministic based upon the input data. The current version of MassIVE-KB has been augmented ten times, growing from 609K to 2.1M precursors over the past several

months as the total data in MassIVE-KB swelled from 2TB to over 30TB.

5.4 Library Provenance

As spectral libraries represent the extent a proteome has been observed by mass spectrometry, the provenance of the spectra in the library are valuable to the community for both validation and exploration. Though existing spectral library resources make attempts at preserving provenance information, all seem to fall short to linking back to the spectra, PSMs, and searches that were used to create the spectral library.

MassIVE-KB attempts to close this gap and provide user access to a full provenance record (Fig. 5.1a). For the entire spectral library, all LC/MS run, search jobs / parameters, and all intermediate library construction jobs are included. Further, representative spectra, spectra considered for representative selection (top 100 candidate spectra per precursor), and the full set of PSMs in library construction are available and point back directly to the specific LC/MS run, dataset, PI (Table 5.1), and original search job/parameters (Fig. 5.1b). This transparency in the library creation process empowers the community to validate search parameters and check the knowledge in the spectral libraries against the underlying data. Finally, it provides a springboard for users to customize their own reanalysis of the data and create their own customized libraries from existing libraries and existing/new searches.

5.5 Proteome Coverage

The MassIVE-KB spectral library currently covers 19,610 (97.4%) proteins out of the 20,129 proteins reported in SwissProt (Jan 2, 2017) at a 1% protein FDR. Without the inclusion of synthetic peptides, MassIVE-KB covers 16,801 proteins (83.4%) contributed from a wide variety of dataset (Fig. 5.3). This represents a gain over the NIST HCD peptide spectral librarys (July 20, 2016) 13,261 proteins (no Protein FDR) of 48% and

27% respectively. At the precursor level, MassIVE KBs 2.1M precursors represent an increase of 322% increase over NISTs 497K (Fig. 5.2a). This gain corresponds to a similar 235% gain in unique sequences that increased amino acid coverage of the proteome by 106% from 2.99M amino acids (26.4% of proteome) to 6.17M amino acids (54.5%), corresponding to covering 191,037 exons (71%) in the human genome. These gains in coverage represent a corresponding increase in important sites for protein function. One such site, glycosylation, serves an important roles in cellular signalling, especially in immune cells. NISTs peptides currently only cover 856 glycosylation sites with CCMS expanding 547% to 5,540 glycosylation sites out of a total of 17,042 reported sites by UniProtKB[11]. The majority contribution of these new sites come from the Bioplex dataset[40], Lung Adenomcarcenoma (PXD002612)[29], human draft proteomes (PXD000865, PXD000561)[48, 103], and ProteomeTools synthetic data (PXD004732)[116].

While there do exist some sequences that are unique to NISTs libraries (46,230 sequences), the majority of these sequences are short ($<$10 amino acids) and represent a potential gain of 129,261 amino acids (1%) in the proteome (Fig. D.3). Of the remaining 46% of the proteome not covered by MassIVE-KB, 1,197,093 amino acids occur in extremely long tryptic peptides ($>$40 aa) and are estimated to be uncoverable by trypsin based bottom up proteomics. Thus, without synthetic peptides, we have observed 5,335,526 (6,167,998 amino acids with synthetics) amino acids, with a remaining 4,869,110 (4,042,347 with synthetics) amino acids remaining in coverable regions (Fig. 5.2d).

5.6 Targeted Proteomics Design

In targeted proteomics experiments, e.g. Single Reaction Monitoring (SRM), the careful selection of proteotypic peptides is essential to the proper quantification of

targeted proteins. Previous efforts by PeptideAtlas[53, 64] empirically predicted the most proteotypic peptides for each protein. Unfortunately, the breadth of PeptideAtlas only enabled evaluation of 9,946 proteins. Building upon the provenance of all searches in creating the MassIVE-KB library, we have expanded this set to nearly 17K proteins.

In evaluating the proteotypic-ness of precursors, we observed a wide variance in the consistency ranging from 0% to 100%. Due to this variance, the number of precursors per protein that are necessary to observe the protein at least 90% (Fig. 5.4a) of the time varied from 1 to greater than 20; however, in the majority of cases less than 5 precursors were necessary (Fig. 5.4b). These measures of consistency on a per protein basis provide the groundwork to design more sensitive SRM experiments.

Further, library derived sets of proteotypic peptides can probe protein-protein interactions. 24K protein-protein interactions were reported by Bioplex 1.0[41], and were filtered down to 11,838 interactions between primary isoforms in swissprot. 9,983 interactions were observed in the data and utilizing the covering set of proteotypic peptides, these interactions can be observed on average greater than 90% of the time (Fig. 5.4c). 915 interactions were not observed because of the absence of the protein in the library. A remaining 924 reported interactions were not observed with co-occurring proteins within the same sample in our analysis.

5.7 Detecting Unobserved Proteins

The NeXtProt[32] resource aims to be a knowledgebase that catalogs the extent of the communitys knowledge of human proteins. As such, NeXtProt (Jan 2, 2017) claims that 16,641 proteins from SwissProt[9] (20,129 total) have been previously observed with protein evidence (PE1). Of these, 15,666 proteins were observed in the MassIVE-KB without synthetics at a 1% protein FDR (94% PE1 proteins). The remaining 6% of proteins anecdotally observed to be from labeled quantification (iTRAQ) experiments

and datasets (P98073), NeXtProt peptide evidence was not provided (P61578), or only observed in synthetic experiments (Q69383). We envision as more HCD data is generated and included in MassIVE-KB library, these remaining proteins will become part of the library.

NeXeProt claimed the remaining 3,488 proteins (PE2-PE5) to lack proteomics evidence. 723 of these 3,488 proteins were observed in the MassIVE-KB library originating from non-synthetic data with at least two non-overlapping sequences that could not map to another protein as a single amino acid variant (SAAV). The majority (649 proteins) were previously seen with transcriptomic evidence (PE2) with the remaining 32 inferred from homology (PE3), 12 uncertain (PE4), and 30 predicted (PE5) (Fig. 5.2c). 531 newly observed proteins contained peptides that could be validated against a synthetic peptide library generated from ProteomeTools[116] (Fig. 5.2b). The common precursors demonstrated very high spectral similarity, with minimum cosine scores of 0.7, and a median of 0.93 (Fig. 5.2c). 308 proteins contained at least two non-overlapping precursors, meeting US HUPOs Extraordinary Detection Claims[22] (Fig. 5.2b). Additionally, 302 of these 723 proteins were observed to have to have peptides derived from AP-MS Bioplex data. 267 of these proteins were observed in samples matching the exact same bait protein, reinforcing that these previously unobserved proteins peptides are indeed correct. Finally, the protein FDR of all proteins with at least two non-overlapping sequences not mapping to an SAAV was 0.013% estimated by TDA.

5.8 Knowledgebase Redistribution

The MassIVE-KB HCD spectral library is made to be redistributed and reused in several ways (Fig. 5.1b). 1. Downloaded as an annotated MGF for use with ProHits[61] and for further computational reanalyses. 2. Downloaded as an sptxt library file for use with SpectraST[58] in the Trans Proteomics Pipeline[20], and Skyline[62]. 3. Used in

CCMSs computational workflows engine for online library searching of data. These tools include M-SPLIT[99] (Data Dependent Acquisition Spectral Library Search), MSPLIT-DIA[100] (Data Independent Acquisition Peptide ID), and Maestro (Spectral Networks Analysis). Searching a HEK293 dataset[13] containing 1.1M MS/MS spectra with M-SPLIT showed MassIVE-KB identified 64% more precursors than NISTs HCD library (Fig. D.6).

5.9 Discussion

We have presented a platform that aggregates the community data into a reusable knowledge base - MassIVE-KB, in the form of a spectral library. This error controlled distillation of the communitys HCD data enables the measuring of progress to completion of proteomes. As new studies and data is deposited in ProteomeXchange consortium mass spectrometry data repositories (e.g. MassIVE, PRIDE, JPost), they will be automatically and appropriately searched and appended to the existing spectral library in an online and error controlled manner.

While the MassIVE-KB spectral library is a general purpose resource for the proteomics community, we envision the communitys need for bespoke spectral library for more targeted and niche purposes. We provide the exact workflows to search and create the MassIVE-KB spectral libraries to the community. Leveraging the provenance of the MassIVE-KB spectral libraries, the community can use the existing searches and library creation workflows as a starting point to create their own distilled and reusable library resources.

5.10 Methods

5.10.1 Data Origin and Preparation

All HCD data was either deposited at MassIVE or PRIDE[96] (and subsequently imported). The Bioplex dataset was acquired directly from the Bioplex project. All thermo raw files were converted using proteowizards msconvert[46] to centroided 32-bit mzXML or mzML[67] files and made available as part of the respective dataset on MassIVE. The initial creation of the CCMS library included 227 datasets, totalling over 30TB of data. The dataset list is in SI Table Input Datasets and all datasets are publically available on MassIVE.

5.10.2 Single Pass Search

The spectra are searched by MSGF+. The UniProt human reference proteome database (May 23, 2016) used contains $\sim 70K$ proteins. Allowed modifications are oxidation, n-term acetylation, n-term Carbamylation, Pyro-glu, and deamidation. Carbamidomethylation is a fixed modification on cysteine. Allowed 1 C_{13} and 1 non-enzymatic termini and a 10ppm parent mass tolerance. Each individual search is filtered to 1% PSM level FDR.

5.10.3 Dynamic Search Space Adjustment Search

In order to improve the sensitivity of searches and take advantage of the experimental metadata, we created a dynamic search space adjustment search. Since MSGF+ reports a database p-value[35] that takes into account both the quality of the peptide-spectrum-match and the search space, using a more appropriate search space will improve database p-value of peptide spectrum matches.

While the search space of synthetic peptide pooled runs are defined prior to a

search, for the class of datasets where there is a limited set of proteins (by purification or separation), that search space of proteins is discovered and utilized to boost certain PSM database p-value scores.

In this first phase, spectra are searched by MSGF+ UniProt proteome database (May 23, 2016) containing 70K proteins. Allowed modifications are oxidation, n-term acetylation, n-term Carbamylation, Pyro-glu, and deamidation. Carbamidomethylation is a fixed modification on cysteine. Allowed 1 C13 and 1 non-enzymatic termini and a 10ppm parent mass tolerance. Search results are filtered to 1% spectrum level FDR. A subset of the proteins from the database are selected such that each protein has at least two distinct sequences from identification set. Proteins with two or more peptides are collected. Input spectra are re-scored against this reduced protein set to update database p-values for spectra identified to these proteins in both stages. All other PSMs will have their database p-value scores unchanged. Results are then filtered to 1% PSM level FDR with database p-value as the primary score ranking.

5.10.4 Synthetic Searching

Synthetic spectra from ProteomeTools[116] were run in 365 pools of 1,000 peptides. Each pool of 1,000 peptides was acquired multiple times. Each LC/MS run per pool was searched separately by searched by MSGF+ (except the DDA sighting run). 10 random LC/MS runs from ProteomeTools were searched together as a background. These LC/MS runs were searched against the per pool specific database of peptides. All PSMs were merged and sorted by descending database p-value. A decoy PSM was defined as

1. A target or decoy PSM matched to an LC/MS run from the background files
2. A decoy PSM matched to an LC/MS run from the desired pool

PSM results were filtered to 0% FDR.

5.10.5 Spectra Extraction/Filtering

All filtered PSMs per search are considered eligible to enter the spectral library. For each PSM, the corresponding MS/MS spectrum is extracted from the appropriate file and match statistics are calculated. Each spectrums peaks are normalized to a euclidean norm of 1.0. All peaks are annotated by the PSM peptide sequence with the ions ”b”, ”b-iso”, ”y”, ”y-iso”, ”b-H2O”, ”b-NH3”, ”y-H2O”, ”y-NH3”, and ”a” with charges up to and including the precursor charge with a mass tolerance of 0.1 Da. Any spectrum with an explained intensity of less than 0.4 were discarded. Further, any MS/MS spectrum that deviated from the theoretical mass by more than 50ppm were filtered out, including any C13 isotopes.

5.10.6 Spectral Library Precursor Selection for Inclusion

For each precursor, the top 100 scoring spectra are retained (with a maximum of 20 per dataset) for future consideration. Each precursor is scored with the best database p-value of its PSMs and these precursors are filtered to 1% local precursor FDR by length. Multiple precursors that differ only by site localization of modifications were removed with only the highest scoring precursor retained. Further precursors are then mapped to the SwissProt Human Proteome (Jan 2, 2017). Protein level FDR is performed with only non-shared precursors. Each protein is scored as the sum of the -log10 of each precursors database p-value that maps to it. Proteins are filtered to 1% Protein FDR. All precursors (including shared sequences) that map to at least this filtered protein set are retained, others are dropped.

For each peptide length we determined the worst database p-value score at the FDR threshold to define a PSM cutoff threshold. All PSMs below this threshold were dropped. Of the remaining spectra for each precursor, the spectrum most similar (by cosine) to all others in this set is selected as the representative library spectrum.

5.10.7 Online Spectral Library Augmentation

New datasets are converted to open formats, searched with a standard database search workflow, and spectra are extracted. All new annotated spectra are merged together with the existing top 100 PSMs per precursor and again filtered to 1% precursor level local FDR with a new database p-value score cutoff for each length. The spectra remaining in contention are treated similarly to initial library creation. This procedure ensures order independence of dataset addition.

5.10.8 Filtering out Ambiguous Library Spectra

Library spectra are searched against the UniProt human reference proteome database (May 23, 2016) by MSGF+ with identical settings as the original database search. The top 20 identification candidates are returned for each library spectrum. For each library spectrum, we retain identification candidate if it exceeds the 1% PSM FDR cutoff score in the library spectrums original search. Identification candidates that differ from the highest scoring identification candidates by site localization of PTMs are removed. Library spectra with multiple identification candidates are considered to be ambiguous and are removed from the library.

5.10.9 Novel Protein Calling

Precursors greater than length 8 are mapped to the SwissProt (Jan 2, 2017) Human Proteome and filtered to 1% Protein FDR via the picked Protein FDR method[86]. Further, these precursors are mapped to proteins, allowing for a single amino acid variant (SAAV). Peptides mappings to multiple proteins are removed. For each protein, the number of non-overlapping peptides is counted. Only proteins with at least two non-overlapping peptides and fall under 1% Protein FDR are candidates for novel protein observation. Each proteins NeXtProt PE level is identified, and PE1 proteins are not reported.

Table 5.1. Top Contributing PIs to MassIVE-KB

PI	Top 100 Candidate Spectra	Representative Spectra
Steven Gygi	3,556,504 (12%)	660,801 (30%)
Bernhard Kuster	4,570,424 (15%)	252,229 (12%)
Matthias Mann	4,875,236 (16%)	224,867 (10%)
Akhilesh Pandey	2,356,169 (8%)	172,966 (8%)
Christopher Gerner	1,156,998 (4%)	66,701 (3%)

5.10.10 Gold Standard Library Construction

SILAC datasets were searched MSGF+ and variable SILAC modifications of +8,+10 on Arginine and Lysine. Results were filtered to 1% PSM level FDR. Next we determined concordant light and heavy pairs of MS/MS spectra. They had to fulfill the following criteria:

1. Light and heavy MS/MS spectra had to exhibit exact same peptide annotation aside from SILAC modification
2. Aligned cosine score of greater than 0.6
3. Explained Intensity greater than 0.5
4. Retention time within one minute

Light version of validated light/heavy pairs are considered for the SILAC library. The highest database p-value scoring spectrum for each precursor is selected to be the representative in the library.

Chapter 5, in part is currently being prepared for submission for publication of the material. Assembling the Community Scale Discoverable Human Proteome. Mingxun Wang, Jian Wang, Jeremy Carver, Nuno Bandeira. The dissertation author was the primary investigator and author of this material.

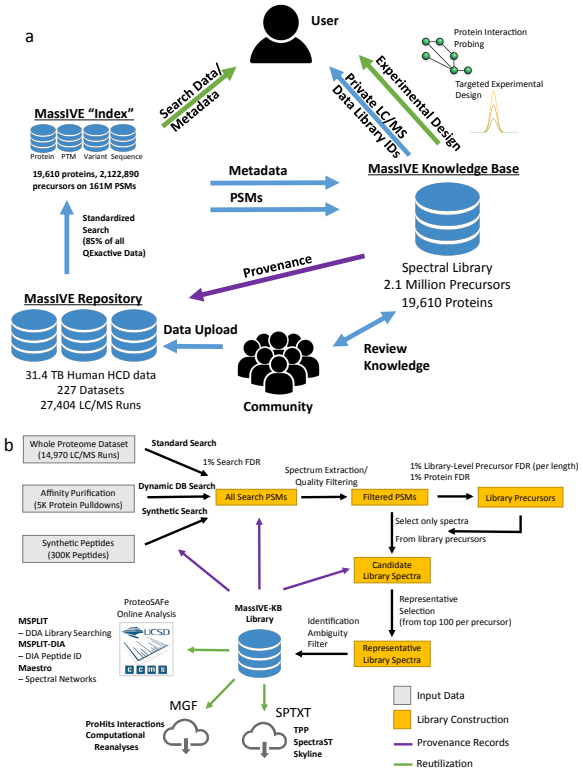


Figure 5.1. Overview of MassIVE-KB (a) Representation of the interactions between the proteomics community, MassIVE datasets, MassIVE Index, and MassIVE knowledge base. This eco-system empowers a user by guiding experimental design with a spectral library (e.g. designing an SRM experiment), making identifications in their own LC-MS/MS data by spectral library search, and validate extraordinary results against reference library spectra and all PSMs found in all public data via MassIVE Index. The MassIVE-KB spectral library draws from over 30 TBs of data with 658M MS/MS spectra yielding 161M PSMs resulting in a spectral library of 2.1M precursors. New HCD human datasets are added to the MassIVE-KB in an online fashion upon deposition in ProteomeXchange consortium repositories. Further, the community can review the knowledge in MassIVE-KB in order to scrutinize the full provenance record and request changes with supporting information. (b) Overview of library creation procedure. Beginning with different search strategies for different experimental procedures. PSMs and aggregated and filtered for spectrum quality. Precursor level FDR is applied on the remaining spectra that filter candidate spectra for inclusion in the library. Representative spectra are chosen and ambiguously identified spectra are removed. The MassIVE-KB spectral library is made available to browse, search new data, and download in both mgf and sptxt formats.

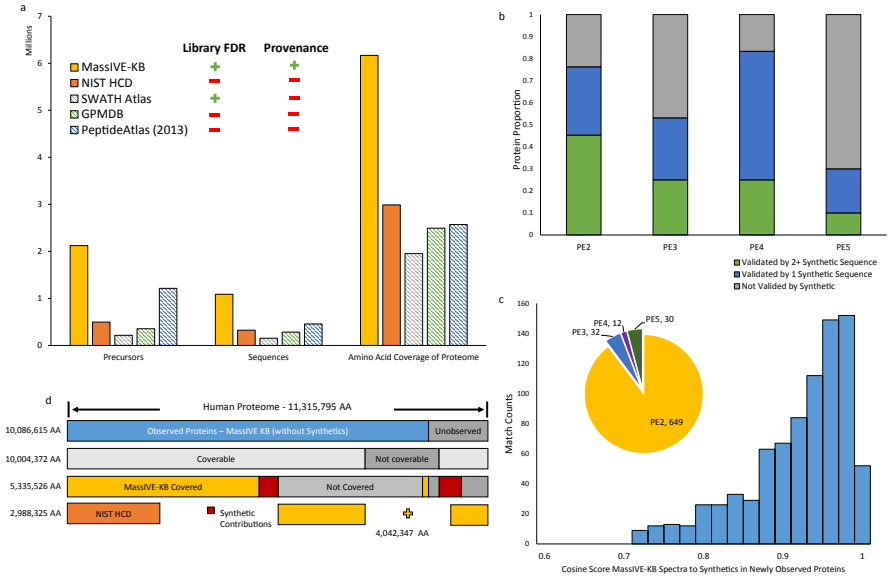


Figure 5.2. MassIVE-KB Coverage of Human Proteome (a) The MassIVE-KB HCD spectral library expands over NISTs HCD library by 323% for precursors, 235% for sequences, and 106% for proteome coverage. Further, in comparison from other available spectral library resources, MassIVE-KB controls library level FDR at the precursor and protein levels along with provenance information for each library spectrum. (b) 723 of 3,489 proteins not observed with proteomics evidence from NeXtProt were detected with at least two non-overlapping sequences in the MassIVE-KB. 531 proteins could be confirmed with synthetic peptides with 308 proteins matching 2+ synthetic sequences (green) and 223 proteins matched with 1 synthetic sequence (blue). (c) The library spectra calling these unobserved proteins confirmed by synthetic MS/MS spectra exhibited high spectral similarity with a minimum score of 0.7 and a median of 0.93. (d) Even with the marked gain in coverage by the MassIVE-KB of the human proteome, there remains 4M amino acids that are coverable (<40 amino acid tryptic sequences). This represents 3.4M amino acids remaining in observed proteins in MassIVE-KB (without synthetics) as well as 625K amino acids in unobserved proteins.

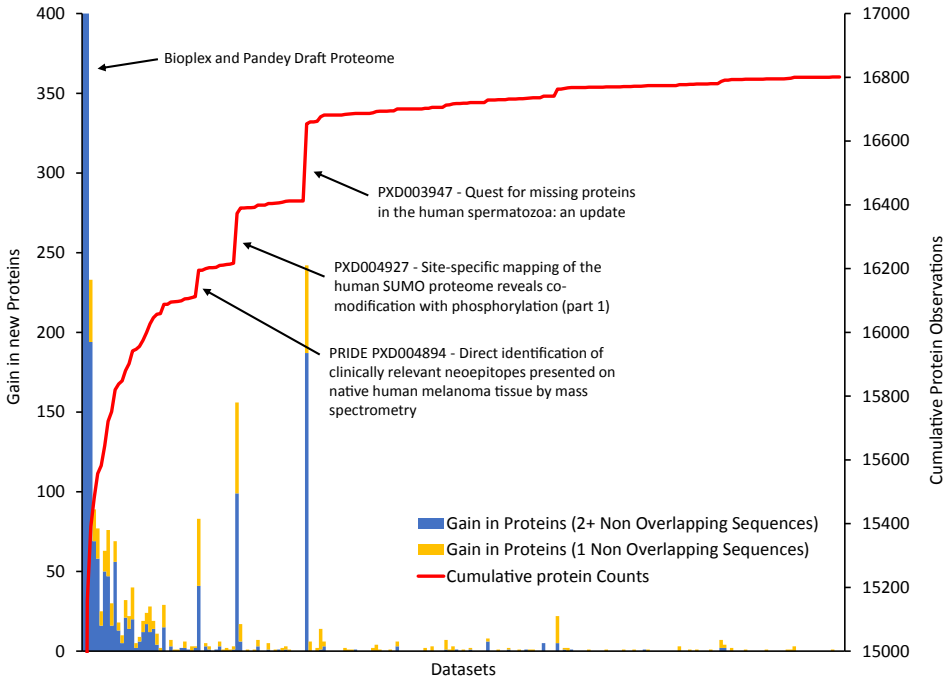


Figure 5.3. Growth of Proteome Coverage. Various datasets in the community contribute their own unique picture of the human proteome. Several foundational datasets , e.g. Bioplex and draft proteomes, provide vast comprehensive coverage of commonly observed proteins. However, several smaller datasets contributed unique proteins, e.g. PXD004927 SUMO proteome[38] and PXD003947 spermatozoa[94] proteins. All combined, the communities dataset (excluding synthetics) observe nearly 17K proteins of the human proteome.

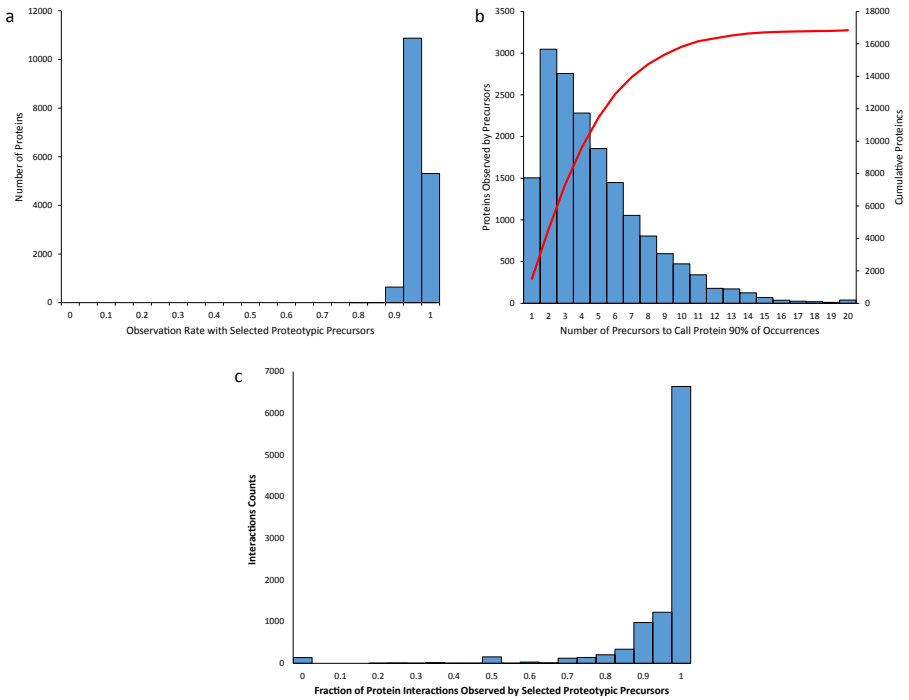


Figure 5.4. Proteotypic Peptides in Proteome (a) The percentage observability for proteins given a selected proteotypic peptide set. In the vast majority of cases (98%), proteins could be observed in >90% of searches with a selected set of proteotypic peptides. (b) The number of peptides that must be selected such that at least one peptide is present 90% of the time when a protein is called. For the majority of cases, less than four precursors is required to observe a protein in 90% of its occurrences. While there does exist a tail, upon closer investigation, a subset of datasets proteotypic peptides differ drastically from the norm. This discrepancy can be partially attributed to experimental procedures for peptide enrichment (e.g. phosphorylation and SUMOylation). (c) The protein-protein interactions as reported by the Bioplex project can also be detected with a subset of the peptides in the identifications. By using the top peptides that can call a protein >90% of the time, >88% of observed interactions can be called >90% of the time with this proteotypic peptide set.

Appendix A

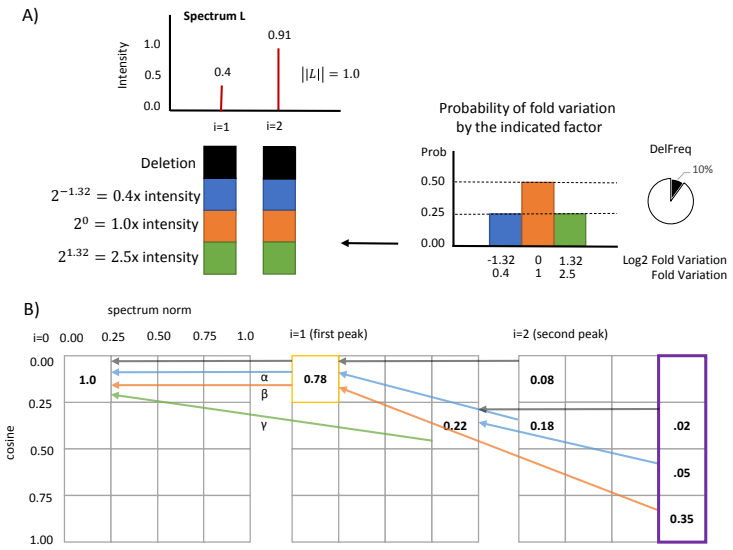
SLGF Dynamic Programming Details

$$LibDP(i,c,p) = \sum_{y=0 \rightarrow p} LibDP(i-1, c-\sqrt{y} \times L_i, p-y) \times Prob(\sqrt{y}|L_i)$$

In the equation for the LibDP recurrence, $c - \sqrt{y} * L_i$ is used because we are iterating over all possible increments in spectrum norm. For each increment y (third term in the LibDP recurrence, $p - y$), this translates into a $\sqrt{y}$ intensity for the replicate spectrum peak. Thus if we consider $\sqrt{y}$ intensity in a replicate spectrum peak, such a peak will contribute $\sqrt{y} \times L_i$ to the accumulated cosine c between the replicate and the library spectrum L, which explains the second term in the LibDP recurrence: $(c - \sqrt{y} \times L_i)$.

CPTAC Test Dataset Search with Sigma UPS1 search

In the initial analysis, the search library did not include Sigma48UPS1 spectra. To determine whether this affected the search results, the search was re-performed with the combined NIST Yeast and Sigma libraries. Supplementary Figure A.2 includes ROC curves for this search of the Test dataset against this combined library. At 1% Peptide FDR, Tremolo had 3902 IDs, and SpectraST had 3315, representing a 17% gain. The results show the relative difference between Tremolo and SpectraST to be comparable to



C) (α) if replicate peak intensity $\Delta cos = (0.16) \times 0.4 = 0.06$
is $0.4 \times L_1 = 0.16$: $\Delta norm = (0.16)^2 = 0.03$
(β) if replicate peak intensity $\Delta cos = (0.4) \times 0.4 = 0.16$
is $1.0 \times L_1 = 0.40$: $\Delta norm = (0.4)^2 = 0.16$
(γ) if replicate peak intensity $\Delta cos = (1.0) \times 0.4 = 0.4$
is $2.5 \times L_1 = 1.00$: $\Delta norm = (1.0)^2 = 1.0$

D) $(1.0 \times 0.1) + (1.0 \times 0.25 \times 0.9) + (1.0 \times 0.5 \times 0.9) = 0.775$

Figure A.1. SLGF calculation example using the LibDP dynamic programming recursion. In (A) a very simple library spectrum with only two peaks is considered (with intensity 0.4 and 0.91). Each of these peaks has identical deletion frequency and intensity variation distributions that are discretized to 0.4x, 1.0x, and 2.5x intensity, with 0.25, 0.5, and 0.25 probability respectively. In (B) the dynamic programming recurrence is illustrated with the probability of occurrence shown in each cell (0.0 if empty). The arrows represent the propagation of probability and the color of each arrow represents which portion of the intensity variation distribution yielded the probability update. The last column of the table for the last peak (i=2 in this case) corresponds to the distribution of cosine scores at Euclidean norm 1.0 (shown in purple). This last column is re-normalized to total probability 1.0 and used as the SLGF theoretical score distribution for the library spectrum. The calculation of the $\Delta Cosine$ and $\Delta Norm$ is shown in (C), which explains the differing indices for the dynamic programming for the dynamic programming recurrence between i=1 and i=0. Additionally, the calculation of the probability of the top left cell (shown in yellow) for the $i = 1$ peak is shown in (D). Note that while it is shown that the LibDP matrix has $n_p = 4$ and $n_c = 4$, the actual implementation uses $n_p = 400$ and $n_c = 800$.

that of the search without the Sigma library.

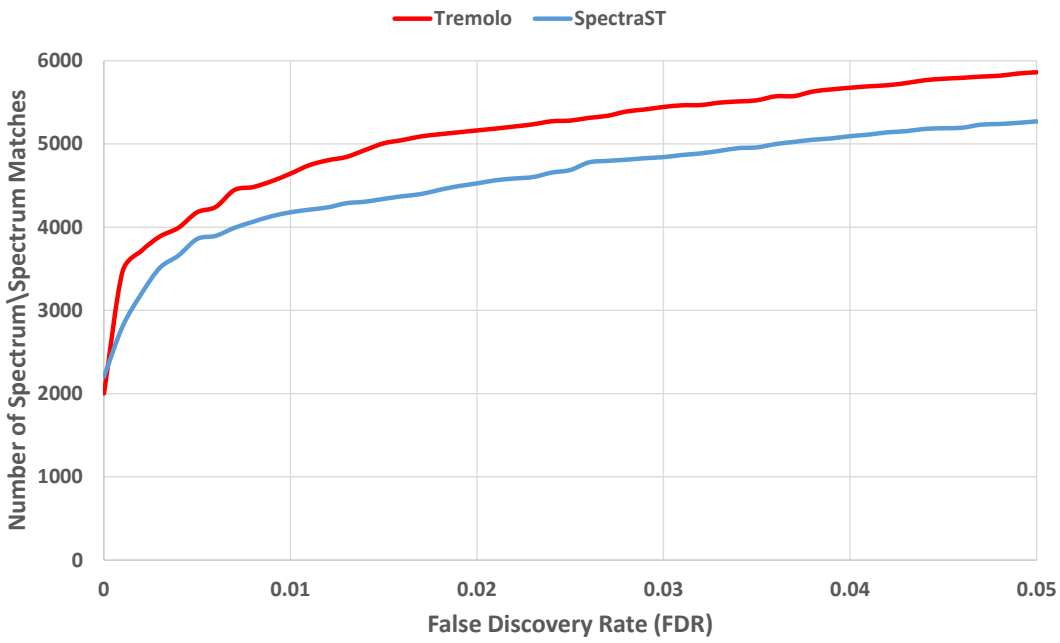


Figure A.2. SLGF Yeast ROC Curve. Peptide spectral library search sensitivity by FDR comparison between Tremolo and SpectraST on the CPTAC Test dataset with the combined NIST libraries for Sigma and Yeast. At 1% spectrum level FDR, the gain in IDs for Tremolo is approximately 11%.

Appendix B

GNPS Methods

B.0.1 Spectral Library Searching

Input MS/MS spectra (i.e., query spectra) are considered matched to library spectra if they meet the following criteria: same precursor charge state, precursor m/z is within a user defined Thompson tolerance, share a minimum number of matched peaks, and exceed a user-defined minimum spectral match score. Exact spectral matches between library and query spectra are scored with a normalized dot product5759. The matching of peaks between two spectra is formulated as a maximum bipartite matching problem15 where peaks from the library and query spectra are represented as nodes with edges connecting library and query peaks. Edges connect peaks that are within a user defined fragment mass tolerance. The bipartite match of library to query peaks that maximizes the normalized dot product is selected. The highest scoring library match for each query spectrum is reported. Estimated false discovery rates of the exact spectral library search are shown in Supplementary Note C.0.3. Source code can be found at the CCMS github page.

B.0.2 Variable Dereplication

Variable dereplication utilizes a modification tolerant spectral library search. Similar to exact spectral matches, except additional edges are added to the bipartite

matching between library and query peaks which differ by a δ (as determined by their precursor mass difference δ) +/- the user defined fragment mass tolerance.

B.0.3 Molecular Network Construction

Molecular networks can be constructed from any collection of MS/MS spectra. First, all MS/MS spectra are clustered with MSCluster60 such that MS/MS spectra found to be identical are merged into a consensus spectrum. Consensus spectra are then matched against each other using the modification tolerant spectral matching scheme15. All spectrum-to-spectrum matches that exceed a user defined minimum match score are retained. MS/MS spectra are then represented as nodes in a graph and significant matches between spectra are represented as edges. Further, edges in the graph are only retained if the two nodes, A and B, connected by a given edge satisfy the following properties: i) B must be in the top K highest scoring neighbors of A and ii) A must be in the top K highest scoring neighbors of B. All other edges are removed. Source code can be found at the CCMS github page.

B.0.4 GNPS Collections - Sample Preparation

The NIH Prestwick Phytochemical Library, NIH Natural Product Library, and NIH Small Molecule Pharmacologically Active Library compounds were received as stock solutions of pure compounds (10 mM in DMSO). They were reformatted by 1 L of each compound into 89 L of methanol into 96 well plates with 11 distinct compounds in each well. They were further diluted 100-fold for a final 1 M concentration. The NIH Clinical Collections and FDA Library part 2 were received as stock solutions of pure compounds (10 mM in DMSO). They were diluted to final concentration of 1 M in 50:50 methanol:water and formatted onto 96 well plates with 10 compounds per well.

B.0.5 GNPS Collections - LC MS/MS Acquisition

LC-MS/MS acquisition for all in house generated libraries was performed using a Bruker Daltonics Maxis qTOF mass spectrometer equipped with a standard electrospray ionization source (ESI). The mass spectrometer was tuned by infusion of Tuning Mix ESTOF (Agilent Technologies) at a 3 L/min flow rate. For accurate mass measurements, lock mass internal calibration used a wick saturated with hexakis phosphazene ions (Synquest Laboratories, m/z 922.0098) located within the source. Samples were introduced by a Thermo Scientific UltraMate 3000 Dionex UPLC using a 20 L injection volume. A Phenomenex Kinetex 2.6 μm C18 column (2.1 mm 50 mm) was used. Compounds from NIH Prestwick Phytochemical Library, NIH Natural Product Library, and NIH Small Molecule Pharmacologically Active Library were separated using a seven minute linear wateracetonitrile gradient (from 98:2 to 2:98 water:acetonitrile) containing 0.1% formic acid. Compounds from NIH Clinical Collections and FDA Library part 2 Library employed a step gradient for chromatographic separation [5% solvent B (2:98 water:acetonitrile) containing 0.1% formic acid for 1.5 min, a step gradient of 5% B-50% B in 0.5 min, held at 50% B for 2 min, a second step of 50% B-100% B in 6 min, held at 100% B for 0.5 min, 100%-5 % B in 0.5 min and kept at 5% B for 0.5 min]. The flow rate was 0.5 mL/min. The mass spectrometer was operated in data dependent positive ion mode; automatically switching between full scan MS and MS/MS acquisitions. Full scan MS spectra (m/z 50 - 1500) were acquired in the TOF and the top ten most intense ions in a particular scan were fragmented using collision induced dissociation (CID) utilizing stepping.

B.0.6 GNPS Collections - Library Creation

All raw data were centroided and converted to 32-bit uncompressed mzXML file using Bruker Data Analysis. A script was developed to select all possible MS/MS spectra

in each LC-MS/MS run that could correspond to a compound present in the sample. For each compound, we calculated the theoretical mass M from its chemical composition and searched for the M+H, M+2H, M+K, and M+Na adducts. Putative identifications included all MS/MS spectra whose precursor m/z had a ppm error $<$50 compared to the theoretical mass of each possible precursor m/z; all tandem MS/MS spectra with an MS1 precursor intensity of $<$1E4 were ignored. All candidate identifications were manually inspected and the most abundant representative spectrum for each compound was added to the corresponding library at the gold or bronze level based upon an expert evaluation of the spectrum quality. The best MS/MS spectrum per compound as added to the GNPS-Collections library without filtering or alteration from the mzXML files.

B.0.7 GNPS-Community Contributed Spectral Library Processing and Control

User contributed library spectra are not filtered or altered in any way from the user submission. MS/MS spectra are extracted from the submitted data and are made available in the GNPS libraries. The list and description of metadata fields can be found in GNPS online documentation. To preserve provenance information, the full input file is also retained and made available for download for each library spectrum. Different levels of reference spectra submissions are enforced with access restrictions on a per user basis. The description of each of the quality levels: Gold, Silver and Bronze and be found in Supplementary Table C.2. While any MS/MS spectrum can be Bronze quality level in the GNPS libraries, Silver contributions require peer-reviewed publication of the MS/MS spectra, and Gold contributions require MS/MS spectra to be of synthetics or purified compounds with complete structural characterization.

Appendix C

GNPS Supplementary Information

C.0.1 Note 1: Browsing Available Libraries

All available libraries at GNPS, including freely available 3rd party libraries can be browsed and or downloaded here. Users are able to see continually updated snapshots of each of the respective libraries described in Supplementary Table C.1. User instructions, available at gnps.ucsd.edu under documentation, for the Library Spectrum View are be found under the appropriate sections.

C.0.2 Note 2: GNPS Library Addition Procedure (Collections and Community)

To facilitate library growth, GNPS enables users to contribute to the GNPS-Community spectral library from both private and public datasets. Reference spectra can be uploaded as a set with a batch upload process, one spectrum at a time, or even directly from the output of an analysis tool at GNPS with a single click Add To Library button. The metadata fields that are required for user submissions and descriptions can be found at gnps.ucsd.edu under documentation under the Library Contribution section. All provenance information is automatically maintained so that proper credit can be given in follow-up analysis. Even with Gold, Silver, and Bronze quality levels in place, it is possible that i) some submissions will be tentative (especially in the Bronze category), ii)

some spectra will not yet have the highest possible signal (e.g., low abundance NPs), or iii) metadata might be incomplete at the time of submission (e.g. tentative structures). Therefore a wiki-style Update Annotation workflow is available, providing a path for annotations of library spectra to become more complete. This revision approach enables annotation updates for example from a hypothetical sugar containing natural product to a much more specific Erythromycin when new information becomes available and can also be used to fix incorrect or imprecise annotations. Also, missing metadata, such as complete structural information, can be added to further complete annotation of library spectra. To date 376 annotations been updated. Further, to reach a community consensus on certain annotations, a data centric exchange may need to occur. GNPS promotes these conversations by allowing comments on annotations with additional supporting information attachments.

C.0.3 Note 3: Library Search False Discovery Rate Estimation

With library search being an integral part of molecular networking, dereplication, and continuous identification, it is important to assess the false discovery rates of the library search available at GNPS. The massive datasets MSV000078708 and MSV000078710 are the full LC-MS/MS runs for to create the NIH Natural Product and NIH Small Molecule Pharmacological Libraries. They total 1,327,215 MS/MS query spectra. The dynamic exclusion settings were to fragment a precursor at most three times unless the MS1 peak intensity increased 2x since the last MS2 acquisition. This ensured that there are multiple MS2 spectra per precursor. Further details can be found in Methods. These datasets were searched against the NIH Natural Product and NIH Small Molecule Pharmacological Libraries. The matches returned were categorized into the following: identity match, consistent correct match, inconsistent correct match, and incorrect match. Identity matches were matches between the library spectrum and the identical spectrum

in the data that was extracted to create the library spectrum. Consistent Correct matches are between library spectra and query spectra where the query spectrum occurred in the same LC-MS/MS run as the library spectrum. Inconsistent Correct matches are between library spectra and query spectra such that the ppm error of the match is <50ppm and are within 200 scan numbers of each other, but the library and query did not come from the same LC-MS/MS file. These could have resulted from contamination and impurities in compound library. All other matches are considered incorrect. The 50ppm tolerance and 200 scan thresholds were chosen based on the empirical match data. We took all library spectra and matched them to all MS/MS query spectra with a 2 Da precursor tolerance. The top scoring match for each query spectrum was reported if the match score exceeded 0.7. The distribution of all possible ppm precursor m/z errors are shown in Supplementary Figure 3.1a. For all of these matches that were reported, the ppm precursor m/z error histogram is shown in Supplementary Figure 3.1b. With the possible ppm errors of reported matches ranging beyond 10,000, the reported matches having almost all PPMs below 50 is in indication the scoring scheme is discriminating between true and false matches. Further, scan number deltas for all possible matches are shown in Supplementary Figure 3.1c, which again shows a high variance in the possible deltas that could occur. However, of the reported matches, the scan number delta converges significantly as shown in Supplementary Figure 3.1d with nearly all of the identifications occurring under a scan delta of 200 (5% of the chromatographic time). The spectral library search at GNPS returned 8,355 matches, of which 2,428 were Identity matches, 4,557 were Consistent Correct matches. 1,138 were Inconsistent Correct matches, and 232 were incorrect matches. We calculated the false discovery rate to be 3.9% as the number of incorrect matches divided by the total number of correct identifications with the identity matches removed. Further, to show the sensitivity of the search, the library included a total of 2,716 MS/MS spectra. Of which, 2,428 library spectra matched to

the identical spectra used to create the library. Thus, 288 of the library spectra were not matched because they did not meet the minimum number of peaks required to be considered identifiable (6 peaks). Of the 2,428 library spectra that were deemed identifiable, 2,045 library spectra had at least one replicate identification (Supplementary Fig. 2). This allows us to estimate the sensitivity of the search at 84% by dividing the total number of library spectra that had at least one correct non-identity match (2,045 spectra) by the number of library spectra that had an identity match (2,428 spectra).

C.0.4 Note 4: GNPS Dataset Creation and Access

Entire project data sets, irrespective of size and number of data files, can be uploaded, managed, analyzed, and shared with the entire community. Each of these public dataset is given a unique identifier that can be referenced and linked in journal articles. A variety of open file formats (mzXML, mzML, and MGF) as well as several proprietary file formats can be uploaded to the system. Public datasets will be made available in their original submission format and in an open format that is compatible with all GNPS tools shortly after dataset publication. This process usually takes 48 hours and availability status is shown on the dataset page under the heading Analyze Data. Since GNPS datasets will be public and compatible with GNPS tools, public datasets can be re-analyzed online. While metadata such as PI and username are required, metadata such as protocols and similar details can be provided but are not mandatory (though can be updated in the future). For further documentation regarding dataset creation, dataset browsing (including downloading data, online re-analysis of data, and viewing continuous identification results), see the appropriate section in the GNPS online documentation (available at gnps.ucsd.edu under documentation).

C.0.5 Note 5: Public Data Molecular Networks

All public NP datasets were clustered1, producing consensus spectra as described in Methods Molecular Network Construction. For each dataset, the Identified Molecules is the number of consensus spectra matched to a library spectrum by continuous identification. The Putative Analog Molecules is the number of consensus spectra that had an edge in the molecular network to a library matched consensus spectrum. The Identified Networks is the number of consensus spectra that were a part of connected components (node count >1) in the molecular network that had at least one consensus spectrum match to a library spectrum. The Unidentified Networks is number of consensus spectra that were a part of connected components (node count >1) in the molecular network that did not have any matches to library spectra. The Exploratory Networks is the number of consensus spectra that were a part of connected components (node count >1) in the exploratory molecular network (Fig. 3.4a). The overall public data identification rate was Identified Molecules summed across all datasets divided the total number of consensus spectra across all datasets (Fig. 3.4b).

C.0.6 Note 6: Continuous Identification Procedure

Each single public GNPS dataset is run through the standard Molecular Networking workflow. The clustered set of spectra are then searched against the GNPS-Collections, GNPS-Community, MassBank, ReSpect, and NIST libraries with the molecular library search workflow. The most current search results are then compared with the previous continuous identifications on a per dataset basis. Any changes in matches (New, Different, or Deleted) are reported. Matches for previously unidentified spectra are defined as New. Changes in matches of spectra due to annotation updates or a better library match are defined as Different. Spectra that previously had a match but currently do not due to removal of library spectra or updates to library search scoring functions are

defined as Deleted. These results are reported in two different ways: i) all are shown on the GNPS Dataset page under the Continuous Identification section ii) users subscribing to each dataset are emailed a summary of changes in identifications (Each user receives only one email with all summaries to avoid spamming). An email is only when a change or new entry is obtained for a dataset to which a user is subscribed to.

C.0.7 Note 7: Molecular Explorer Creation and Presentation

All Continuous Identification matches are aggregated and grouped by matched compound name. The GNPS Molecular Explorer presents all matched library compounds and their respective occurrences in the public data. Thus, it is possible to easily ask the question in which public datasets contain a certain molecule. Further, since GNPS also constructs molecular networks for public datasets, the molecular explorer also presents occurrences of putative analogs of library compounds. Supplementary Figure C.6 highlight the amount of diversity of analogs captured by the molecular explorer as well as the fraction of this diversity that is already captured by reference spectra in GNPS spectral libraries. For further instructions, see the Molecular Explorer section GNPS online documentation.

Table C.1. GNPS Libraries Summary. Current spectral libraries available at GNPS, including compound libraries run exclusively for GNPS, community contributed spectra, and 3rd party libraries. Note: Original numbers created for the manuscript are inconsistent with this table as updated numbers were generated 3/21/2016.

Library Name	#Spectra	#Compounds	Data Generators
GNPS-Community	2,224	1,325	Community
GNPS-Collections (FDA Approved Libraryfrom Selleck Chem Pt 1)	2,389	297	Sirenas MD
GNPS-Collections (FDA Approved Libraryfrom Selleck Chem Pt 2)	656	535	Dorrestein Lab
GNPS-Collections (NIH Clinical Collection 1)	377	329	Dorrestein Lab
GNPS-Collections (NIH Clinical Collection 2)	195	164	Dorrestein Lab
GNPS-Collections (Natural Products in NIH Small Molecule Repository)	1,268	1,255	Dorrestein Lab
GNPS-Collections (Pharma Active Compounds in the NIH Small Mol Rep	1,460	1,398	Dorrestein Lab
GNPS-Collections (Prestwick Phytochemical Library)	143	140	Dorrestein Lab
GNPS-Collections (Faulkner Legacy Library)	127	125	Sirenas MD
MassBank ESI MS/MS Library (Feb 2016)	24,545	5,992	Various
ReSpect MS/MS Library (July 2014)	7,112	1,986	Various
HMDB (Feb 2015)	2,235	747	Various
NIST 2014 ESI MS/MS Library	193,119	8,351	NIST
Total	235,850	22,644	

Table C.2. Library Quality Levels Description. The general public is granted access to add to the Bronze quality levels. However, special permissions are required to have access to add to Silver and Gold level MS/MS spectra. The oversight of these permissions falls under the administrators of GNPS, and currently is managed by Pieter Dorrestein.

Quality Level	Description	Spectra Count
Gold	Synthetic or purified, Complete structural characterization with NMR, crystallography or other standard methods as defined in the publication guidelines for Journal of Natural Products. Requires administrative approval.	4308
Silver	Isolated or lysate/crude, Published data showing presence of molecule in the sample. Requires administrative approval. Recommended attendance of GNPS curator workshop.	446
Bronze	Any other putative, complete, or partial annotation	4085

Table C.3. Number of Annotation Revisions. This table summarizes the total number of spectra with a given number of annotation revisions. There are a total of 563 annotation revisions for 429 GNPS library spectra.

Number of Annotation Revisions	Number of Library Spectra
1	351
2	46
3	19
4	7
5	4
6	0
7	1
8	1

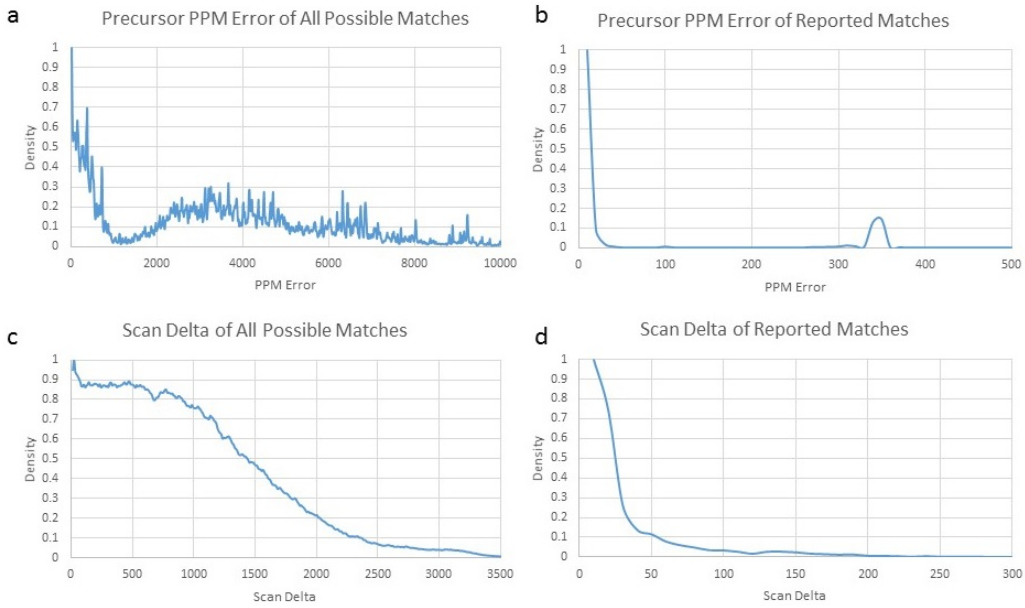


Figure C.1. Spectral Library Search Match Assessment. (a) The PPM error of all possible matches from the NIH Natural Product Library and NIH Pharmacologically Active Library to the query datasets MSV000078708 and MSV000078710. The spread of possible PPM errors for all possible matches illustrates the large possible space of erroneous matches as measured by PPM error. (b) The PPM errors of all reported matches. There is a slight bump around 350PPM as all LC-MS/MS runs from one of the plates was miscalibrated. The total number of non-identity matches that within 50 PPM error is 5,520 matches out of a 5,927 matches. (c) The scan deltas of all possible matches between the query spectra and library spectra. The large spread of possible scan number deltas illustrates large possible space of erroneous matches as measured by scan number deltas. (d) A histogram of scan number delta between query and library reported matches by library search. 5,789 of all 5,927 matches reported fell within 200 scans of the library spectrum.

Table C.4. Rating Description. Description of each of the ratings for matches made in continuous identification.

Rating	Description
1 Star	Incorrect Identification
2 Star	Not enough information, fragmentation is not sufficient to tell
3 Star	Putative Analog Identification, possibly not exactly correct isomer
4 Star	Correct Identification

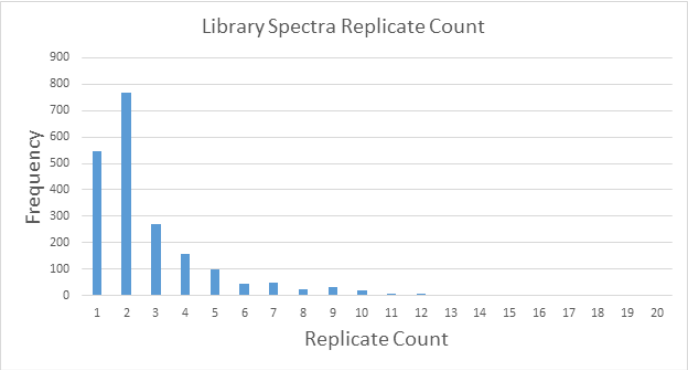


Figure C.2. Replicate Count Histogram within Reported Matches. A histogram of the number of replicates for each library spectrum. These do not include identity matches or incorrect matches.

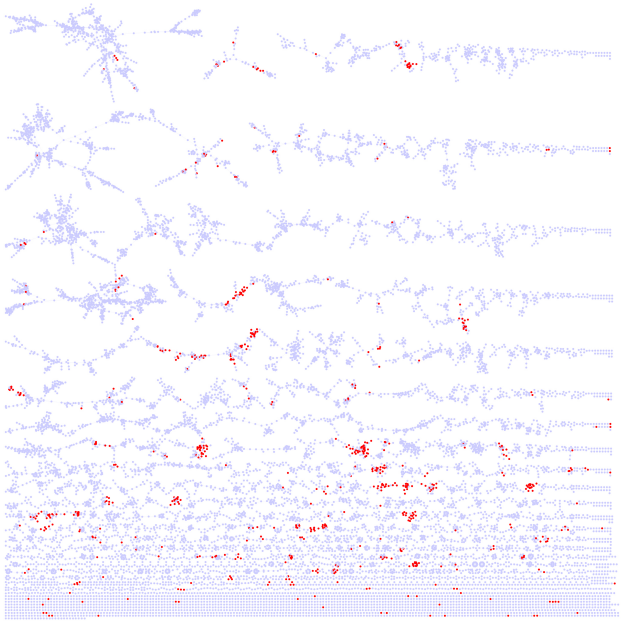
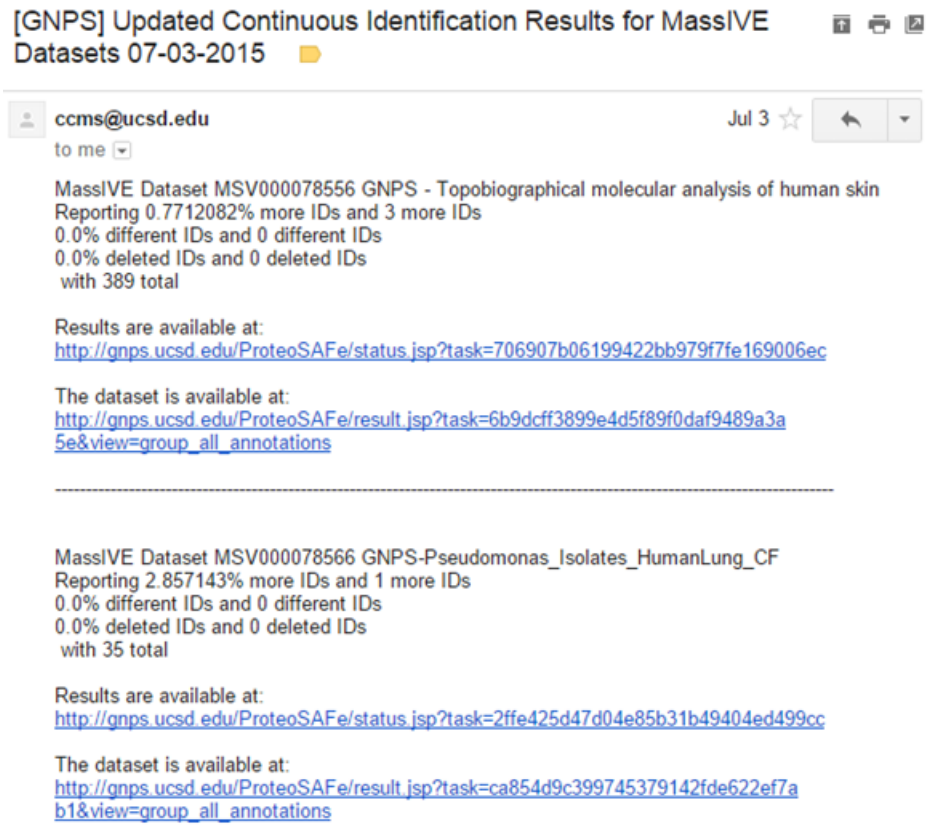


Figure C.3. Chemical Space Visualization with Molecular Networking. MS/MS data from a human skin metabolomics datasets (Bouslimani et al. 2015, MassIVE dataset MSV000078556) was analyzed with molecular networking at GNPS. Consensus MS/MS spectra matched to a GNPS library spectrum are highlighted as red nodes, unmatched in grey.



[GNPS] Updated Continuous Identification Results for MassIVE Datasets 07-03-2015

ccms@ucsd.edu Jul 3

to me

MassIVE Dataset MSV000078556 GNPS - Topobiographical molecular analysis of human skin
Reporting 0.7712082% more IDs and 3 more IDs
0.0% different IDs and 0 different IDs
0.0% deleted IDs and 0 deleted IDs
with 389 total

Results are available at:
http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=706907b06199422bb979f7fe169006ec

The dataset is available at:
http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=6b9dcff3899e4d5f89f0daf9489a3a5e&view=group_all_annotations

---

MassIVE Dataset MSV000078566 GNPS-Pseudomonas_Isolates_HumanLung_CF
Reporting 2.857143% more IDs and 1 more IDs
0.0% different IDs and 0 different IDs
0.0% deleted IDs and 0 deleted IDs
with 35 total

Results are available at:
http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ffe425d47d04e85b31b49404ed499cc

The dataset is available at:
http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=ca854d9c399745379142fde622ef7ab1&view=group_all_annotations

Figure C.4. Example Continuous Identification Digest Email. Digest email reporting new identifications per dataset. For each dataset, the spectral match changes are reported as well as total match count, links to results, and links to each dataset.

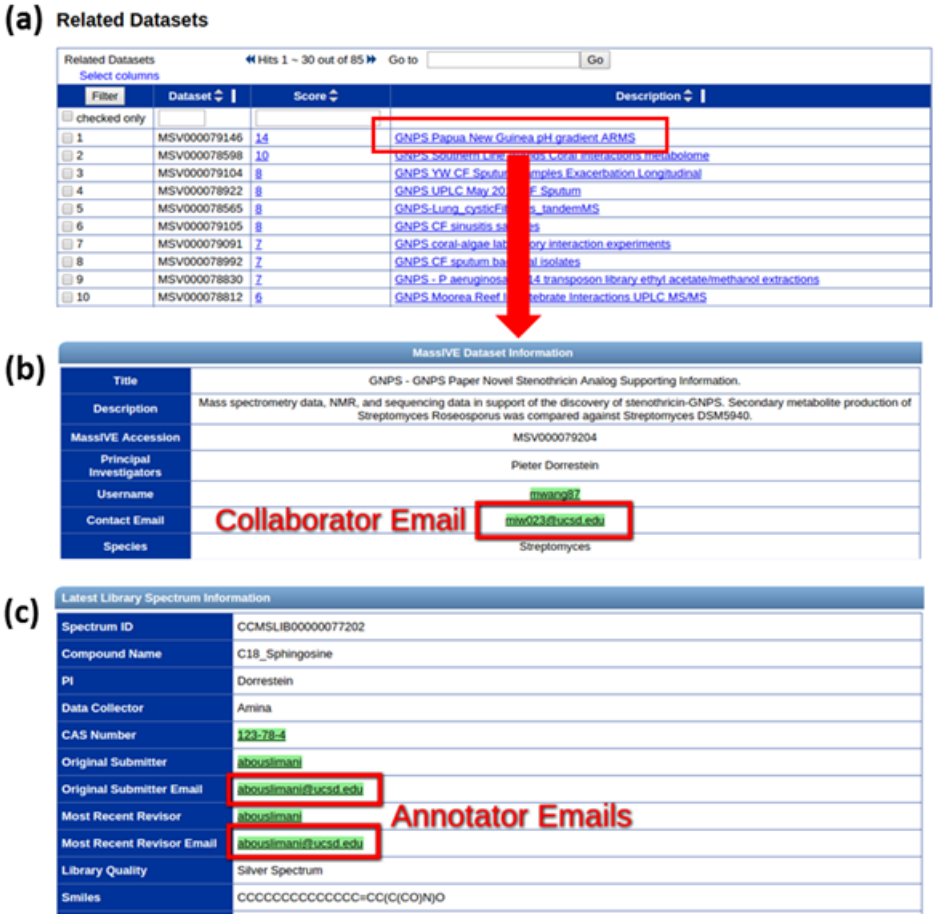


Figure C.5. GNPS Collaboration Initiation. (a) For each public dataset, GNPS provides a list of related datasets. This allows users to find related datasets to their own in order to spark collaboration. In selecting a related dataset, users reach a dataset page (b). From here more detailed metadata including an email link are available to initiate contact. (c) Further, identifications made via continuous identification link to spectrum library pages and enables users to initiate contact with the annotators of these library spectra.

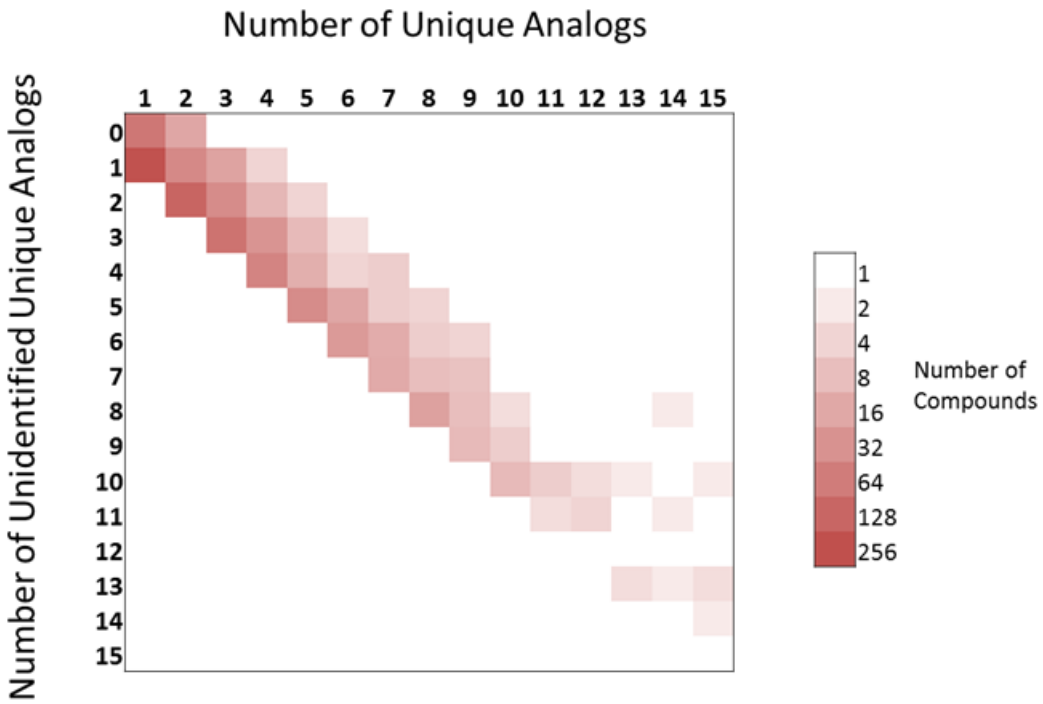


Figure C.6. Molecular explorer highlights the number and locations of related MS/MS spectra of known compounds, indicating they are related (defined as analogs). Analogs are direct neighbors of matched compounds in molecular networks. For each library compound we tallied the number of unique mass analogs. Here we illustrate exactly how much diversity of known compounds is found in all public GNPS datasets and how much of that diversity is actually captured in the GNPS spectral libraries. For each library spectrum matched to public data all unique mass analogs were counted (X axis) and all unidentified analogs were counted (Y axis); the relationship between these is shown as a 2D heatmap. Since most of the density of the plot is near the diagonal, it indicates that the majority of the putative analogs of known compounds in the data remained unidentified.

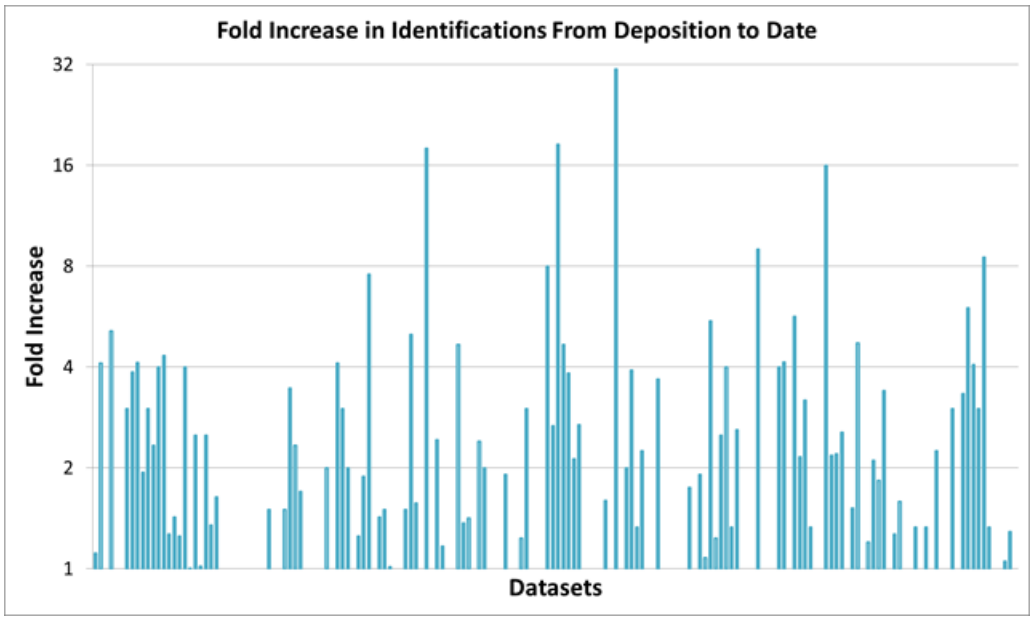


Figure C.7. Dataset fold increase in identifications by continuous identification. The fold increase of identifications on a per dataset basis from the time of deposition to date. 59% of these datasets increased their identifications, averaging a 143% increase in identifications since deposition.

Appendix D

MassIVE-KB Supplementary Information

D.0.1 Online Library Creation

The CCMS spectral library can be augmented in an incremental manner. When new datasets become available, they are automatically searched, PSMs extracted, and augmented to the existing library. The resulting spectral library from a series of augmentations is order agnostic, thus resulting in identical libraries given any permutation of augmentations. An augmentation of new spectra can only result in the loss of precursors from the library with the addition of new decoy spectra. The lost target precursors will always be from the lowest scoring precursors. This stability is in contrast to SpectraSTs construction process, where precursors can be lost if new additions become too similar to existing library entries regardless of the confidence of the identifications.

D.0.2 Database p-value Calculation

The primary score for all PSMs is the database p-value[35]. The database p-value captures the probability of at least one high scoring false match to a given spectrum in a database size N. Lower database p-values indicate low error.

The equation to calculate the database p-value is as follows:

$$1-(1-P(S|T))^N$$

$P(S|T)$ represents the spectral probability[49], i.e. the probability of a random peptide match to spectrum S with a score higher than T. This value captures the how good of a match a given peptide matches a spectrum S, with smaller values being a better explanation. $(1-P(S|T))$ represents the probability of a random match with low score. N represents the size of the database, i.e. the total number of candidates that are eligible to match against the spectrum S. $1-(1-P(S|T))^N$ is the probability that all random database matches are of low score. Then the final expression $1-(1-P(S|T))^N$ is the probability that at least one random database match to spectrum S exceeds score T. Since lower database p-values indicate higher confidence, we can aim to lower the database p-value in several ways: decreasing spectral probability $P(S|T)))$ and decreasing database size (N). To decrease spectral probability, by acquiring higher quality spectra or finding a peptide that better explains the data accomplishes this goal. To decrease the database size in samples where the true set of peptides and proteins that occur is unknown, we discover the appropriate database (Fig. D.1a, D.2a). When the database is known, e.g. synthetic peptide pools, we specifically constrain the database size and calculate the database p-value appropriately (Fig. D.1b).

D.0.3 Dynamic Search Space Adjustment for Proteome Limited Data

The dynamic adjustment of search space for these affinity purification mass spectrometry files on average filtering the search space from $\sim 70K$ proteins in the Uniprot Human Reference database down to approximately 1,500 proteins (Fig. D.2b). PSMs that matching to these proteins will have their scores boosted because of the reduced search space while all other PSMs scores will remain unchanged. Because of

the requirement that two distinct sequences much match to a protein for the protein to be included in the second pass search set, a single PSM will not boost its own score, rather a second distinct PSM of a different sequence is required.

By appropriately adjusting the search space for the samples from Bioplex, the MassIVE-KB spectral library increased by 172K precursors (8% of the MassIVE-KB library and 35% of Bioplex's unique precursors).

D.0.4 Database Identification of Synthetic Peptide Pools from ProteomeTools

Each LC/MS run of the 365 peptide pools from ProteomeTools is searched separately against the appropriate peptide database. Due to scheduled acquisition of MS/MS spectra, there existed very few background MS/MS that would not match to the database, decreasing the probability for any decoy spectra. We included 10 random LC/MS runs from different peptide pools, i.e. including completely different peptides, as background spectra for decoys.

Decoy PSMs are PSMs from any LC/MS run that matched to a reverse decoy database and PSMs from the background LC/MS runs that matched to the target database. Because of this expanded definition of decoy PSMs that can come from true target sequences, we cannot include these decoy PSMs in the library. Therefore, we have conservatively filtered to 0% empirical PSM FDR.

D.0.5 Global Error Rate Controls for Spectral Library

To control accumulation of errors in repository wide aggregation of identification results, we globally controlled error rates at the precursor and protein levels. Per search, we controlled FDR at 1% PSM level, however, if we naively aggregated all of the PSMs to form a library, the precursor FDR would grow to 28%. We control this precursor FDR by filtering to 1% local precursor FDR by length. Each precursors score is the best score

of any PSM from that precursor. All precursors are grouped by length and then sorted by decreasing score (-log(database p-value)). Local FDR is determined for each precursor as the FDR of a sliding window of 500 adjacent precursors.

Following that, we calculated a picked protein FDR and retained only precusros that mapped (including shared) to proteins falling below 1% FDR. A proteins score is the summation of all uniquely mapping precursors -log(database p-value).

While we did choose to filter to 1% PSM level FDR per search before aggregation, all candidate spectra that could be representatives spectra in the MassIVE-KB spectral library uniformly had a per search FDR of 0%. This indicates that we could have chosen a 0% PSM level FDR per search and would not change the resulting library. However, not including decoys would not enable us to determine a library FDR.

D.0.6 Decoy Inclusion in Spectral Library

The source of errors in spectral library search identifications arise from erroneous spectral matches and incorrect identifications of library spectra. The former is captured by spectral decoys[57] in library search, but the latter is ignored to exist. Database search decoys are carried throughout the library search process to this type of error. Thus, this introduces another class of decoys for library search to consider. We have implemented this awareness into an MSPLIT workflow that is available to use online at the CCMS. While the MassIVE-KB library contained only 1,910 (0.09% of library) decoy precursors, these decoys constituted 6% of the decoy matches (Fig. D.6), an over-representation of over 60x.

D.0.7 Library Representative Selection

In choosing the most similar spectrum to all other spectra in general also chose the highest scoring spectra, but not necessary all of the time (Fig. Spectral Quality b). In

nearly all cases however, the top 10 scoring spectrum was chosen. By choosing the most similar spectrum, an improvement in spectral similarity is observed when compared to the gold standard SILAC spectral library (Fig. D.2). Further, the representative selection criteria yielded MS/MS spectra that exhibited similarity to the SILAC library at least as good as the NIST HCD spectral library (Fig. D.4a).

D.0.8 Spectral Library Search

The MassIVE-KB spectral library is distributed both as an MGF and sptxt format for use in various tools. In its MGF format, it is directly compatible with the MSPLIT spectral library search workflow available at CCMS. A HEK293 dataset[13] containing 1.1M MS/MS spectra was searched with both the MassIVE-KB library and NISTs HCD library. Though this dataset was used in creating the MassIVE-KB spectral library, any from this dataset was removed from the MassIVE-KB for these search purposes. MassIVE-KB increased precursors (at 1% precursor FDR) identified from 77,472 to 127,419 (Fig. D.6), a 64% increase.

D.0.9 Proteins in Dataset

There are 1,270 proteins observed that included less than two non-overlapping sequences that did not match to a single SAAV in the proteome. 394 of these proteins occur in the Bioplex dataset. 112 of these proteins were bait proteins and 283 did contain two or more unique peptides without considering SAAVs in the proteome. These lines of evidence give credence that these protein identifications are correct. However, the remaining 95 proteins did only have one unique peptide and were not included in baits and potentially could be false identifications. However, out of the 12K proteins in the Bioplex dataset, if these 95 proteins were indeed false discoveries, it would represent a 0.7% FDR, in-line with the global 1% Protein FDR that was applied during library

creation.

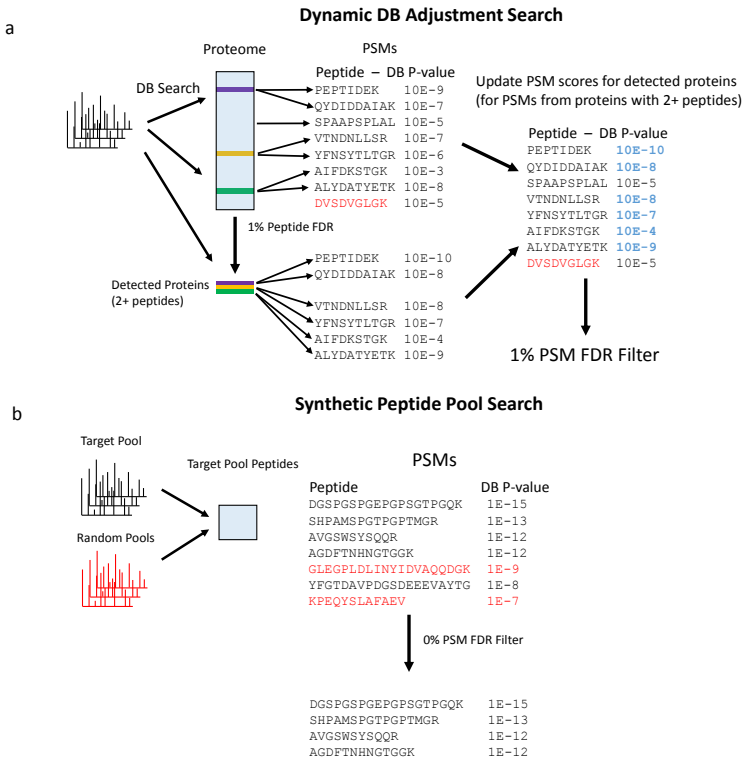


Figure D.1. Technical Search Overview. (a) Illustration of search strategy for protein constrained dataset (e.g. affinity purification). Spectra are first searched against the full proteome and PSMs are filtered to 1% PSM level FDR. Proteins with two or more peptides are collected. Input spectra are re-scored against this reduced protein set to update database p-values for spectra identified to these proteins in both stages. Results are filtered to 1% PSM level FDR to retain enough decoys for estimation of library-level FDR. (b) Synthetic peptide data is searched against the appropriate peptide database along with a random synthetic peptide run. All identifications from the target peptide pool data is considered targets and identifications to the random peptide pool data are considered decoys. All PSMs are ranked by their database p-value scores and all PSMs with a q-value greater than 0 are dropped.

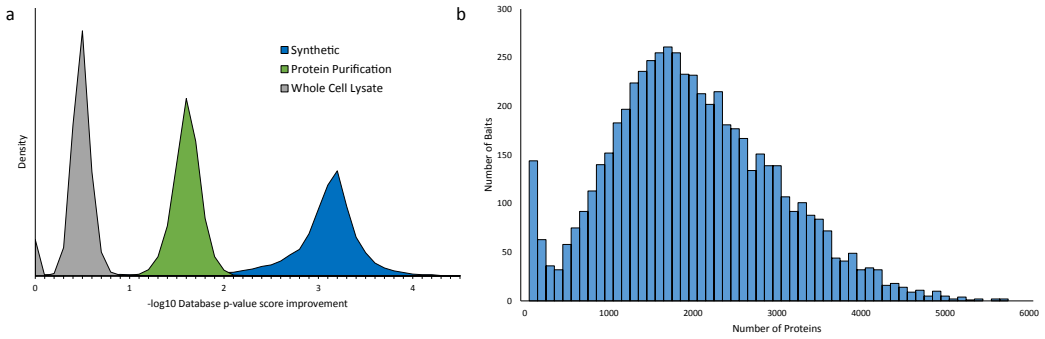


Figure D.2. Search Space Score Adjustment. (a) The appropriate search space for a mass spectrometry dataset can vary widely based on experimental procedures. These search spaces range from nearly full proteome for cell lysates down to a handful of peptides for synthesized peptide pools. By appropriately determining this search space for cell lysate, affinity purification, and synthetic peptide pools, database p-value scores improved by ~ 0.3, ~ 1.6, and ~ 3.2 orders of magnitude respectively. (b) Each bait in the Bioplex AP-MS dataset was searched separately. The distribution of proteins used in refined search space is centered around approximately 1,500 proteins.

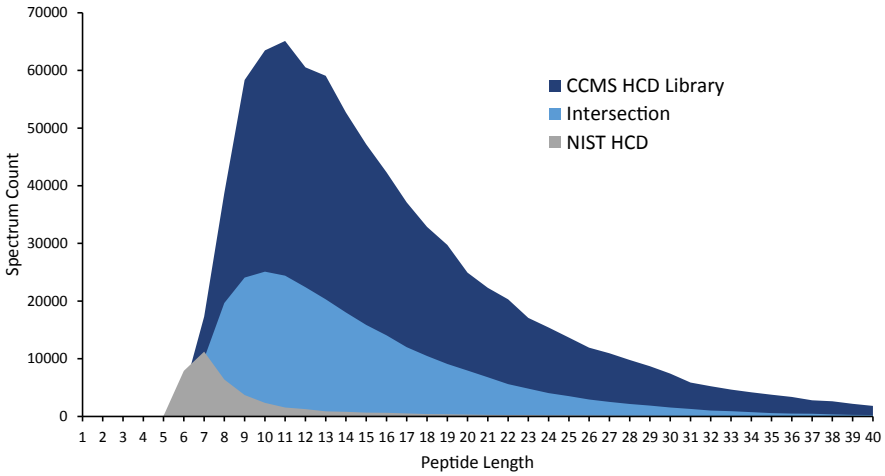


Figure D.3. Size Comparison to NIST. The shared and unique sequences in CCMS vs NISTs libraries. While NIST contains 50K unique sequences not contained in the MassIVE-KB libraries, the majority are short peptides (<9 amino acids). In contrast, the gains in sequences in MassIVE-KB library over NIST are predominantly sequences length that are most commonly observable and identifiable in mass spectrometry.

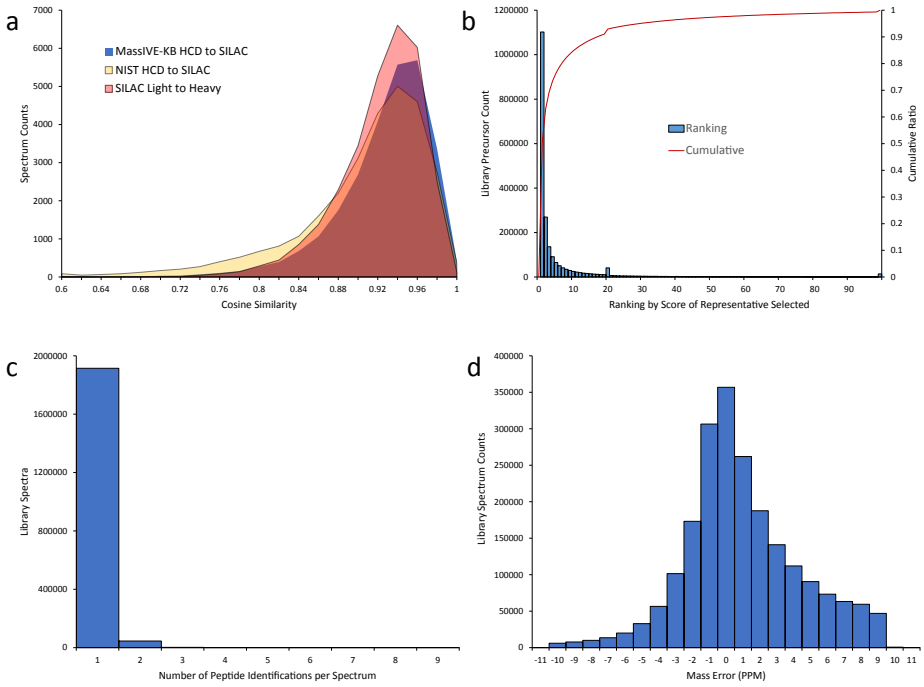


Figure D.4. MassIVE-KB Library Quality. (a) Spectral similarity of annotated ions in NIST and MassIVE-KB spectral library compared to SILAC gold standard library. Shown in red is the distribution of cosine scores between light and heavy SILAC MS/MS spectra in the SILAC dataset, representing the expected best similarity that can be expected from replicates. (b) The chosen representative library spectrum score rank of all considered spectra. In greater than 90% of cases, the representative spectrum was in the top 20 spectra. (c) The number of peptide identifications for each library spectrum. Only 2.5% of library spectra were called ambiguous (two or more identifications) and were removed from the MassIVE-KB spectral library. (d) PPM error of library spectra.

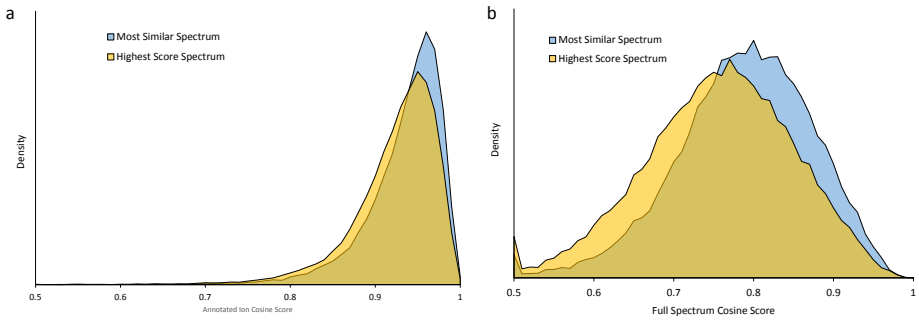


Figure D.5. Selection Methods for Representative. In comparison to the SILAC gold standard library, the the most similar representative spectra to all other replicates exhibits a better cosine similarity to the gold standard library compared with the highest scoring replicate spectrum.

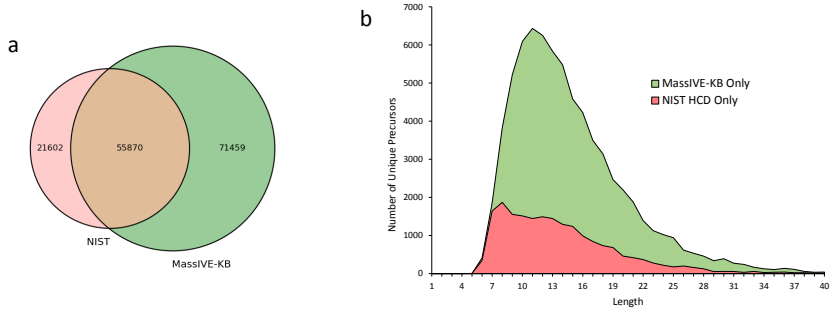


Figure D.6. Massive-KB Library Identification Rate. (a) The MassIVE-KB HCD spectral library when searching a HEK293 Q-Exactive dataset (PXD001468) increases the number of precursors identified by 64% (1% Precursor FDR) over NISTs HCD library. MassIVE-KBs identified 127,419 precursors and NISTs HCD library identified 77,029 precursors. (b) The 21,602 precursors uniquely identified by the NIST HCD library trended toward shorter peptides that were missing from the MassIVE-KB library.

Bibliography

[1] Dictionary natural products. 2013.

[2] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, 2003.

[3] Felicity Allen, Allison Pon, Michael Wilson, Russ Greiner, and David Wishart. Cfm-id: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic acids research*, 42(W1):W94–W99, 2014.

[4] Amos Bairoch and Rolf Apweiler. The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic acids research*, 28(1):45–48, 2000.

[5] Amos Bairoch, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, Hongzhan Huang, Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale, Claire O'Donovan, Nicole Redaschi, and Lai Su L. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33(DATABASE ISS.), 2005.

[6] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein Identification via Spectral Networks Analysis. *Proc Natl Acad Sci U S A*, 104:6140–6145, 2007.

[7] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2013.

[8] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[9] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, and Michel Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, 2003.

[10] Amina Bouslimani, Laura M Sanchez, Neha Garg, and Pieter C Dorrestein. Mass spectrometry of natural products: current, emerging and future technologies. *Natural product reports*, 31(6):718–729, 2014.

[11] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, Parit Bansal, Alan J Bridge, Sylvain Poux, Lydie Bougueleret, and Ioannis Xenarios. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: how to use the entry view. *Plant Bioinformatics: Methods and Protocols*, pages 23–54, 2016.

[12] R. J. Chalkley, N. Bandeira, M. C. Chambers, K. R. Clauser, J. S. Cottrell, E. W. Deutsch, E. A. Kapp, H. H. N. Lam, W. H. McDonald, T. A. Neubert, and R.-X. Sun. Proteome Informatics Research Group (iPRG)_2012: A Study on Detecting Modified Peptides in a Complex Mixture. *Molecular & Cellular Proteomics*, 13(1):360–371, 2014.

[13] Joel M Chick, Deepak Kolippakkam, David P Nusinow, Bo Zhai, Ramin Rad, Edward L Huttlin, and Steven P Gygi. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology*, 33(7):743–749, 2015.

[14] R Craig, J C Cortens, D Fenyo, and R C Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J Proteome Res*, 5:1843–1849, 2006.

[15] Robertson Craig, John P Cortens, and Ronald C Beavis. The use of proteotypic peptide libraries for protein identification. *Rapid communications in mass spectrometry*, 19(13):1844–1850, 2005.

[16] DM. Creasy and JS. Cottrell. Error tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics*, 2:1426–1434, 2002.

[17] S. Dasari, M. C. Chambers, M. A. Martinez, K. L. Carpenter, A. J. Ham, L. J. Vega-Montoto, and D. L. Tabb. Pepitome: evaluating improved spectral library search for identification complementarity and quality assessment. *J. Proteome Res.*, 11:1686–1695, 2012.

[18] Sally J Deeb, Rochelle CJ D'Souza, Jürgen Cox, Marc Schmidt-Supprian, and Matthias Mann. Super-silac allows classification of diffuse large b-cell lymphoma subtypes by their protein expression profiles. *Molecular & Cellular Proteomics*, 11(5):77–89, 2012.

[19] F Desiere, E W Deutsch, N L King, A I Nesvizhskii, P Mallick, J Eng, S Chen, J Eddes, S N Loevenich, and R Aebersold. The PeptideAtlas project. *Nucleic Acids Res*, 34:655–658, 2006.

[20] E W Deutsch, L Mendoza, D Shteynberg, T Farrah, H Lam, N Tasman, Z Sun, E Nilsson, B Pratt, and B Prazen. A guided tour of the Trans Proteomic Pipeline. *Proteomics*, 10(6):1150–1159, 2010.

[21] Eric W. Deutsch, Attila Csordas, Zhi Sun, Andrew Jarnuczak, Yasset Perez-Riverol, Tobias Ternent, David S. Campbell, Manuel Bernal-Llinares, Shujiro Okuda, Shin Kawano, Robert L. Moritz, Jeremy J. Carver, Mingxun Wang, Yasushi Ishihama, Nuno Bandeira, Henning Hermjakob, and Juan Antonio Vizcano. The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, 45(D1):D1100–D1106, 2017.

[22] Eric W. Deutsch, Christopher M. Overall, Jennifer E. Van Eyk, Mark S. Baker, Young Ki Paik, Susan T. Weintraub, Lydie Lane, Lennart Martens, Yves Vandenbrouck, Ulrike Kusebauch, William S. Hancock, Henning Hermjakob, Ruedi Aebersold, Robert L. Moritz, and Gilbert S. Omenn. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1, 2016.

[23] Qian Dong, Xinjian Yan, Lisa E Kilpatrick, Yuxue Liang, Yuri A Mirokhin, Jeri S Roth, Paul A Rudnick, and Stephen E Stein. Tandem mass spectral libraries of peptides in digests of individual proteins: Human serum albumin (hsa). *Molecular & Cellular Proteomics*, 13(9):2435–2449, 2014.

[24] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015.

[25] J E Elias and S P Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, 4:207–214, 2007.

[26] JK. Eng, McCormack AL., and JR. Yates. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *Journal Of The American Society For Mass Spectrometry*, 5:976–989, 1994.

[27] Philipp Engel, Maria I Vizcaino, and Jason M Crawford. Gut symbionts from distinct hosts exhibit genotoxic activity via divergent colibactin biosynthesis pathways. *Applied and environmental microbiology*, 81(4):1502–1512, 2015.

[28] Iakes Ezkurdia, Jesús Vázquez, Alfonso Valencia, and Michael Liam Tress. Analysing the First Drafts of the Human Proteome. *Journal of proteome research*, pages 3854–3855, 2014.

[29] Johannes F Fahrmann, Dmitry Grapov, Brett S Phinney, Carol Stroble, Brian C DeFelice, William Rom, David R Gandara, Yanhong Zhang, Oliver Fiehn, Harvey Pass, and Suzanne Miyamoto. Proteomic profiling of lung adenocarcinoma

indicates heightened dna repair, antioxidant mechanisms and identifies lasp1 as a potential negative predictor of survival. *Clinical proteomics*, 13(1):31, 2016.

[30] Terry Farrah, Eric W Deutsch, Michael R Hoopmann, Janice L Hallows, Zhi Sun, Chung-Ying Huang, and Robert L Moritz. The state of the human proteome in 2012 as viewed through peptideatlas. *Journal of proteome research*, 12(1):162–171, 2012.

[31] B E Frewen, G E Merrihew, C C Wu, W S Noble, and M J MacCoss. Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries. *Anal Chem*, 78:5678–5684, 2006.

[32] Pascale Gaudet, Pierre André Michel, Monique Zahn-Zabal, Isabelle Cusin, Paula D. Duek, Olivier Evalet, Alain Gateau, Anne Gleizes, Mario Pereira, Daniel Teixeira, Ying Zhang, Lydie Lane, and Amos Bairoch. The neXtProt knowledgebase on human proteins: Current status. *Nucleic Acids Research*, 43(D1):D764–D770, 2015.

[33] Tamar Geiger, Juergen Cox, Pawel Ostasiewicz, Jacek R Wisniewski, and Matthias Mann. Super-silac mix for quantitative proteomics of human tumor tissue. *Nature methods*, 7(5):383–385, 2010.

[34] J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q.-W. Xu, N. del Toro, Y. Perez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaino, and H. Hermjakob. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Molecular & Cellular Proteomics*, 13(10):2765–2775, 2014.

[35] Nitin Gupta, Nuno Bandeira, Uri Keich, and Pavel A Pevzner. Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry*, 22(7):1111–1120, 2011.

[36] A Guthals, J D Watrous, P C Dorrestein, and N Bandeira. The spectral networks paradigm in high throughput mass spectrometry. *Mol Biosyst*, (in press), 2012.

[37] Kenneth Haug, Reza M. Salek, Pablo Conesa, Janna Hastings, Paula De Matos, Mark Rijnbeek, Tejasvi Mahendraker, Mark Williams, Steffen Neumann, Philippe Rocca-Serra, Eamonn Maguire, Alejandra Gonzlez-Beltrn, Susanna Assunta Sansone, Julian L. Griffin, and Christoph Steinbeck. MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1), 2013.

[38] Ivo A Hendriks, David Lyon, Clifford Young, Lars J Jensen, Alfred CO Vertegaal, and Michael L Nielsen. Site-specific mapping of the human sumo proteome

reveals co-modification with phosphorylation. *Nature structural & molecular biology*, 2017.

[39] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, 45(7):703–714, 2010.

[40] Edward L. Huttlin, Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 2017.

[41] Edward L. Huttlin, Lily Ting, Raphael J. Bruckner, Fana Gebreab, Melanie P. Gygi, John Szpyt, Stanley Tam, Gabriela Zarraga, Greg Colby, Kurt Baltier, Rui Dong, Virginia Guarani, Laura Pontano Vaites, Alban Ordureau, Ramin Rad, Brian K. Erickson, Martin Wühr, Joel Chick, Bo Zhai, Deepak Kolippakkam, Julian Mintseris, Robert A. Obar, Tim Harris, Spyros Artavanis-Tsakonas, Mathew E. Sowa, Pietro De Camilli, Joao A. Paulo, J. Wade Harper, and Steven P. Gygi. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2):425–440, 2015.

[42] Tatsuya Ito and Miyako Masubuchi. Dereplication of microbial extracts and related analytical technologies. *The Journal of antibiotics*, 67(5):353–360, 2014.

[43] Yanxialei Jiang, Jeeyoung Lee, Jung Hoon Lee, Joon Won Lee, Ji Hyeon Kim, Won Hoon Choi, Young Dong Yoo, Hyunjoo Cha-Molstad, Bo Yeon Kim, Yong Tae Kwon, Sue Ah Noh, Kwang Pyo Kim, and Min Jae Lee. The arginylation branch of the N-end rule pathway positively regulates cellular autophagic flux and clearance of proteotoxic proteins. *Autophagy*, 12(11):2197–2212, 2016.

[44] A. R. Jones, M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. J. Hubbard, J. N. Selley, B. C. Searle, J. Shofstahl, S. L. Seymour, R. Julian, P.-A. Binz, E. W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J. A. Vizcaino, M. Chambers, A. Pizarro, and D. Creasy. The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results. *Molecular & Cellular Proteomics*, 11(7):M111.014381, 2012.

[45] Roland D Kersten, Nadine Ziemert, David J Gonzalez, Brendan M Duggan, Victor Nizet, Pieter C Dorrestein, and Bradley S Moore. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proceedings of the National Academy of Sciences*, 110(47):E4407–E4416, 2013.

[46] Darren Kessner, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. Proteowizard: open source software for rapid proteomics tools development. *Bioinformatics*, 24(21):2534–2536, 2008.

[47] Sara Kildgaard, Maria Mansson, Ina Dosen, Andreas Klitgaard, Jens C Frisvad, Thomas O Larsen, and Kristian F Nielsen. Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by uhplc-dad-qtofms and a ms/hrms library. *Marine drugs*, 12(6):3681–3705, 2014.

[48] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, Dhanashree S. Kelkar, Ruth Isserlin, Shobhit Jain, Joji K. Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A. Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N. Selvan, Arun H. Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J. Sathe, Sandip Chavan, Keshava K. Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D. Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R. Murthy, Nazia Syed, Renu Goel, Aafaque A. Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G. Shaw, Donald Freed, Muhammad S. Zahari, Kanchan K. Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J. Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T. Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D. Leach, Charles G. Drake, Marc K. Halushka, T. S. Keshava Prasad, Ralph H. Hruban, Candace L. Kerr, Gary D. Bader, Christine A. Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome. *Nature*, 509(7502):575–581, 2014.

[49] S Kim, N Gupta, and P A Pevzner. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*, 7:3354–3363, 2008.

[50] S Kim, N Mischerikow, N Bandeira, J D Navarro, L Wich, S Mohammed, A J Heck, and P A Pevzner. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol Cell Proteomics*, 9:2840–2852, 2010.

[51] S. Kim and P. A. Pevzner. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*, 5:5277, Oct 2014.

[52] Sangtae Kim and Pavel A Pevzner. Universal database search tool for proteomics. *Nature communications*, 5:5277, 2014.

[53] Ulrike Kusebauch, David S. Campbell, Eric W. Deutsch, Caroline S. Chu, Douglas A. Spicer, Mi Youn Brusniak, Joseph Slagel, Zhi Sun, Jeffrey Stevens, Barbara Grimes, David Shteynberg, Michael R. Hoopmann, Peter Blattmann, Alexander V. Ratushny, Oliver Rinner, Paola Picotti, Christine Carapito, Chung Ying Huang, Meghan Kapousouz, Henry Lam, Tommy Tran, Emek Demir, John D. Aitchison, Chris Sander, Leroy Hood, Ruedi Aebersold, and Robert L. Moritz. Human SRMAtlas: A Resource of Targeted Assays to Quantify the Complete Human Proteome. *Cell*, 166(3):766–778, 2016.

[54] Ulrike Kusebauch, Eric W Deutsch, David S Campbell, Zhi Sun, Terry Farrah, and Robert L Moritz. Using peptideatlas, srmatlas, and passel: comprehensive resources for discovery and targeted proteomics. *Current Protocols in Bioinformatics*, pages 13–25, 2014.

[55] Hartmut Laatsch. *Antibase 2012: The natural compound identifier*. Wiley-Vch, 2012.

[56] H Lam and R Aebersold. Spectral library searching for peptide identification via tandem ms. *Methods Mol Biol*, 604:95–103, 2010.

[57] H Lam, E W Deutsch, and R Aebersold. Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics. *J Proteome Res*, 9:605–610, 2010.

[58] H Lam, E W Deutsch, J S Eddes, J K Eng, N King, S E Stein, and R Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7:655–667, 2007.

[59] Gerhard Lang, Nor Ainy Mayhudin, Maya I Mitova, Lin Sun, Sonia van der Sar, John W Blunt, Anthony LJ Cole, Gill Ellis, Hartmut Laatsch, and Murray HG Munro. Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. *Journal of Natural Products*, 71(9):1595–1599, 2008.

[60] James L Little, Antony J Williams, Alexey Pshenichnov, and Valery Tkachenko. Identification of known unknowns utilizing accurate mass data and chemspider. *Journal of The American Society for Mass Spectrometry*, 23(1):179–185, 2012.

[61] Guomin Liu, Jianping Zhang, Brett Larsen, Chris Stark, Ashton Breitkreutz, Zhen-Yuan Lin, Bobby-Joe Breitkreutz, Yongmei Ding, Karen Colwill, Adrian

Pasculescu, Tony Pawson, Jeffrey L Wrana, Alexey I Nesvizhskii, Brian Raught, Mike Tyers, and Anne-Claude Gingras. ProHits: integrated software for mass spectrometrybased interaction proteomics. *Nature Biotechnology*, 28(10):1015–1017, 2010.

[62] Brendan MacLean, Daniela M Tomazela, Nicholas Shulman, Matthew Chambers, Gregory L Finney, Barbara Frewen, Randall Kern, David L Tabb, Daniel C Liebler, and Michael J MacCoss. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7):966–968, 2010.

[63] Michele Magrane and Uni Prot Consortium. UniProt Knowledgebase: A hub of integrated protein data. *Database*, 2011, 2011.

[64] Parag Mallick, Markus Schirle, Sharon S Chen, Mark R Flory, Hookeun Lee, Daniel Martin, Jeffrey Ranish, Brian Raught, Robert Schmitt, Thilo Werner, Bernhard Kuster, and Ruedi Aebersold. Computational prediction of proteotypic peptides for quantitative proteomics. *Nature Biotechnology*, 25(1):125–131, 2007.

[65] Matthias Mann and Ole N Jensen. Proteomic analysis of post-translational modifications. *Nature biotechnology*, 21(3):255–261, 2003.

[66] Version September MarinLit. A marine literature database produced and maintained by the department of chemistry. *University of Canterbury, New Zealand*, 2003.

[67] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz, and E. W. Deutsch. mzML–a Community Standard for Mass Spectrometry Data. *Molecular & Cellular Proteomics*, 10(1):R110.000133–R110.000133, 2011.

[68] Lennart Martens, Henning Hermjakob, Philip Jones, Marcin Adamski, Chris Taylor, David States, Kris Gevaert, Joël Vandekerckhove, and Rolf Apweiler. Pride: the proteomics identifications database. *Proteomics*, 5(13):3537–3545, 2005.

[69] David P Martinsen. Survey of computer aided methods for mass spectral interpretation. *Applied Spectroscopy*, 35(3):255–266, 1981.

[70] Samantha J Mascuch, Wilna J Moree, Cheng-Chih Hsu, Gregory G Turner, Tina L Cheng, David S Blehert, A Marm Kilpatrick, Winifred F Frick, Michael J Meehan, Pieter C Dorrestein, and Lena Gerwick. Direct detection of fungal siderophores on bats with white-nose syndrome via fluorescence microscopy-guided ambient ionization mass spectrometry. *PloS one*, 10(3):e0119668, jan 2015.

[71] Fumio Matsuda, Masami Y Hirai, Eriko Sasaki, Kenji Akiyama, Keiko Yonekura-Sakakibara, Nicholas J Provart, Tetsuya Sakurai, Yukihisa Shimada, and Kazuki Saito. Atmetexpress development: a phytochemical atlas of arabidopsis development. *Plant physiology*, 152(2):566–578, 2010.

[72] Wilna J Moree, Vanessa V Phelan, Cheng-Hsuan Wu, Nuno Bandeira, Dale S Cornett, Brendan M Duggan, and Pieter C Dorrestein. Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. *Proceedings of the National Academy of Sciences*, 109(34):13811–13816, 2012.

[73] Seungjin Na, Nuno Bandeira, and Eunok Paek. Fast multi-blind modification search through tandem mass spectrometry. *Molecular & Cellular Proteomics*, 11(4):M111–010199, 2012.

[74] David J Newman and Gordon M Cragg. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *Journal of natural products*, 75(3):311–335, 2012.

[75] Don Duy Nguyen, Cheng-Hsuan Wu, Wilna J Moree, Anne Lamsa, Marnix H Medema, Xiling Zhao, Ronnie G Gavilan, Marystella Aparicio, Librada Atencio, Chanaye Jackson, Javier Ballesteros, Joel Sanchez, Jeramie D Watrous, Vanessa V Phelan, Corine van de Wiel, Roland D Kersten, Samina Mehnaz, René De Mot, Elizabeth a Shank, Pep Charusanti, Harish Nagarajan, Brendan M Duggan, Bradley S Moore, Nuno Bandeira, Bernhard Ø Palsson, Kit Pogliano, Marcelino Gutiérrez, and Pieter C Dorrestein. MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America*, 110(28):E2611–20, 2013.

[76] NIST. The national institute of standards and technology. nist standard reference database 1a. 2014.

[77] Shujiro Okuda, Yu Watanabe, Yuki Moriya, Shin Kawano, Tadashi Yamamoto, Masaki Matsumoto, Tomoyo Takami, Daiki Kobayashi, Norie Araki, Akiyasu C. Yoshizawa, Tsuyoshi Tabata, Naoyuki Sugiyama, Susumu Goto, and Yasushi Ishihama. JPOSTrepo: An international standard data repository for proteomes. *Nucleic Acids Research*, 45(D1):D1107–D1111, 2017.

[78] Nerea Osinalde, Jone Mitxelena, Virginia Sánchez-Quiles, Vyacheslav Akimov, Kerman Aloria, Jesus M. Arizmendi, Ana M. Zubiaga, Blagoy Blagoev, and Irina Kratchmarova. Nuclear Phosphoproteomic Screen Uncovers ACLY as Mediator of IL-2-induced Proliferation of CD4 ⁺ T lymphocytes. *Molecular & Cellular Proteomics*, 15(6):2076–2092, 2016.

[79] A. G. Paulovich, D. Billheimer, A. J. Ham, L. Vega-Montoto, P. A. Rudnick, D. L. Tabb, P. Wang, R. K. Blackman, D. M. Bunk, H. L. Cardasis, K. R. Clauser, C. R.

Kinsinger, B. Schilling, T. J. Tegeler, A. M. Variyath, M. Wang, J. R. Whiteaker, L. J. Zimmerman, D. Fenyo, S. A. Carr, S. J. Fisher, B. W. Gibson, M. Mesri, T. A. Neubert, F. E. Regnier, H. Rodriguez, C. Spiegelman, S. E. Stein, P. Tempst, and D. C. Liebler. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell Proteomics*, 9(2):242–254, Feb 2010.

[80] Patrick G A Pedrioli, Jimmy K Eng, Robert Hubley, Mathijs Vogelzang, Eric W Deutsch, Brian Raught, Brian Pratt, Erik Nilsson, Ruth H Angeletti, Rolf Apweiler, Kei Cheung, Catherine E Costello, Henning Hermjakob, Sequin Huang, Randall K Julian, Eugene Kapp, Mark E McComb, Stephen G Oliver, Gilbert Omenn, Norman W Paton, Richard Simpson, Richard Smith, Chris F Taylor, Weimin Zhu, and Ruedi Aebersold. A common open representation of mass spectrometry data and its application to proteomics research. *Nature Biotechnology*, 22(11):1459–1466, 2004.

[81] Kim D Pruitt, Tatiana Tatusova, Garth R Brown, and Donna R Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic acids research*, 40(D1):D130–D135, 2012.

[82] P. Roepstorff and J. Fohlman. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, 11(11):601, Nov 1984.

[83] George Rosenberger, Ching Chiek Koh, Tiannan Guo, Hannes L. Röst, Petri Kouvonen, Ben C. Collins, Moritz Heusel, Yansheng Liu, Etienne Caron, Anton Vichalkovski, Marco Faini, Olga T. Schubert, Pouya Faridi, H. Alexander Ebhardt, Mariette Matondo, Henry Lam, Samuel L. Bader, David S. Campbell, Eric W. Deutsch, Robert L. Moritz, Stephen Tate, and Ruedi Aebersold. A repository of assays to quantify 10,000 human proteins by SWATH-MS. *Scientific Data*, 1:140031, 2014.

[84] E. L. Rudomin, S. A. Carr, and J. D. Jaffe. Directed sample interrogation utilizing an accurate mass exclusion-based data-dependent acquisition strategy (AMEx). *J. Proteome Res.*, 8(6):3154–3160, Jun 2009.

[85] Christoph Ruttkies, Emma L Schymanski, Sebastian Wolf, Juliane Hollender, and Steffen Neumann. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, 8(1):3, 2016.

[86] Mikhail M Savitski, Mathias Wilhelm, Hannes Hahne, Bernhard Kuster, and Marcus Bantscheff. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Molecular & Cellular Proteomics*, 14(9):2394–2404, 2015.

[87] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[88] Wenguang Shao and Henry Lam. Tandem mass spectral libraries of peptides and their roles in proteomics research. *Mass spectrometry reviews*, 2016.

[89] Ashley M Sidebottom, Andrew R Johnson, Jonathan A Karty, Darci J Trader, and Erin E Carlson. Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from streptomyces coelicolor m145. *ACS chemical biology*, 8(9):2009–2016, 2013.

[90] Bryan E Smith, James A Hill, Mark A Gjukich, and Philip C Andrews. Tranche Distributed Repository and ProteomeCommons.org. In *Methods in molecular biology (Clifton, N.J.)*, volume 696, pages 123–145. 2011.

[91] Colin A Smith, Grace O'Maille, Elizabeth J Want, Chuan Qin, Sunia A Trauger, Theodore R Brandon, Darlene E Custodio, Ruben Abagyan, and Gary Siuzdak. Metlin: a metabolite mass spectral database. *Therapeutic drug monitoring*, 27(6):747–751, 2005.

[92] SE Stein. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10:770 – 781, 1999.

[93] D. L. Tabb, L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A. J. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. A. Carr, K. R. Clauser, J. D. Jaffe, K. A. Kowalski, T. A. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.*, 9(2):761–776, Feb 2010.

[94] Yves Vandenbrouck, Lydie Lane, Christine Carapito, Paula Duek, Karine Rondel, Christophe Bruley, Charlotte Macron, Anne Gonzalez De Peredo, Yohann Couté, Karima Chaoui, Emmanuelle Com, Alain Gateau, Anne Marie Hesse, Marlene Marcellin, Loren Méar, Emmanuelle Mouton-Barbosa, Thibault Robin, Odile Burlet-Schiltz, Sarah Cianferani, Myriam Ferro, Thomas Fréour, Cecilia Lindskog, Jérôme Garin, and Charles Pineau. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *Journal of Proteome Research*, 15(11):3998–4019, 2016.

[95] JD. Venable and JR. Yates. Impact of ion trap tandem mass spectra variability on the identification of peptides. *Anal Chem*, 76:2928–2937, 2004.

[96] Juan Antonio Vizcaíno, Richard G. Côté, Attila Csordas, José A. Dianes, Antonio Fabregat, Joseph M. Foster, Johannes Griss, Emanuele Alpi, Melih Birim, Javier Contell, Gavin O'Kelly, Andreas Schoenegger, David Ovelleiro, Yasset Pérez-Riverol, Florian Reisinger, Daniel Ríos, Rui Wang, and Henning Hermjakob. The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013. *Nucleic Acids Research*, 41(D1), 2013.

[97] Maria I Vizcaino, Philipp Engel, Eric Trautman, and Jason M Crawford. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *Journal of the American Chemical Society*, 136(26):9244–9247, 2014.

[98] J Wang, J Pérez-Santiago, J E Katz, P Mallick, and N Bandeira. Peptide identification from mixture tandem mass spectra. *Mol Cell Proteomics*, 9:1476–1485, 2010.

[99] Jian Wang, Josué Pérez-Santiago, Jonathan E Katz, Parag Mallick, and Nuno Bandeira. Peptide identification from mixture tandem mass spectra. *Molecular & Cellular Proteomics*, 9(7):1476–1485, 2010.

[100] Jian Wang, Monika Tucholska, James DR Knight, Jean-Philippe Lambert, Stephen Tate, Brett Larsen, Anne-Claude Gingras, and Nuno Bandeira. Msplit-dia: sensitive peptide identification for data-independent acquisition. *Nature methods*, 12(12):1106–1108, 2015.

[101] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Wei-Ting Liu, Max Crüsemann, Paul D Boudreau, Eduardo Esquenazi, Mario Sandoval-Calderón, Roland D Kersten, Laura A Pace, Robert A Quinn, Katherine R Duncan, Cheng-Chih Hsu, Dimitrios J Floros, Ronnie G Gavilan, Karin Kleigrewe, Trent Northen, Rachel J Dutton, Delphine Parrot, Erin E Carlson, Bertrand Aigle, Charlotte F Michelsen, Lars Jelsbak, Christian Sohlenkamp, Pavel Pevzner, Anna Edlund, Jeffrey McLean, Jörn Piel, Brian T Murphy, Lena Gerwick, Chih-Chuang Liaw, Yu-Liang Yang, Hans-Ulrich Humpf, Maria Maansson, Robert A Keyzers, Amy C Sims, Andrew R Johnson, Ashley M Sidebottom, Brian E Sedio, Andreas Klitgaard, Charles B Larson, Cristopher A Boya P, Daniel Torres-Mendoza, David J Gonzalez, Denise B Silva, Lucas M Marques, Daniel P Demarque, Egle Pociute, Ellis C O'Neill, Enora Briand, Eric J N Helfrich, Eve A Granatosky, Evgenia Glukhov, Florian Ryffel, Hailey Houson, Hosein Mohimani, Jenan J Kharbush, Yi Zeng, Julia A Vorholt, Kenji L Kurita, Pep Charusanti, Kerry L McPhail, Kristian Fog Nielsen, Lisa Vuong, Maryam Elfeki, Matthew F Traxler, Niclas Engene, Nobuhiro Koyama, Oliver B Vining, Ralph Baric, Ricardo R Silva, Samantha J Mascuch, Sophie Tomasi, Stefan Jenkins,

Venkat Macherla, Thomas Hoffman, Vinayak Agarwal, Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr, Andrés M C Rodríguez, Anne Lamsa, Chen Zhang, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, Pierre-Marie Allard, Prasad Phapale, Louis-Felix Nothias, Theodore Alexandrov, Marc Litaudon, Jean-Luc Wolfender, Jennifer E Kyle, Thomas O Metz, Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav, Rolf Müller, Katrina M Waters, Wenyuan Shi, Xueting Liu, Lixin Zhang, Rob Knight, Paul R Jensen, Bernhard Ø Palsson, Kit Pogliano, Roger G Linington, Marcelino Gutiérrez, Norberto P Lopes, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, and Nuno Bandeira. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*, 34(8):828–837, 2016.

[102] J. Watrous, P. J. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. Kersten, M. Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira, and P. C. Dorrestein. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A*, 109:E1743–1752, 2012.

[103] Mathias Wilhelm, Judith Schlegl, Hannes Hahne, Amin Moghaddas Gholami, Marcus Lieberenz, Mikhail M. Savitski, Emanuel Ziegler, Lars Butzmann, Siegfried Gessulat, Harald Marx, Toby Mathieson, Simone Lemeer, Karsten Schnatbaum, Ulf Reimer, Holger Wenschuh, Martin Mollenhauer, Julia Slotta-Huspenina, Joos-Hendrik Boese, Marcus Bantscheff, Anja Gerstmair, Franz Faerber, and Bernhard Kuster. Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587, 2014.

[104] Micheal C Wilson, Tetsushi Mori, Christian Rückert, Agustinus R Uria, Maximilian J Helf, Kentaro Takada, Christine Gernert, Ursula a E Steffens, Nina Heycke, Susanne Schmitt, Christian Rinke, Eric J N Helfrich, Alexander O Brachmann, Cristian Gurgui, Toshiyuki Wakimoto, Matthias Kracht, Max Crüsemann, Ute Hentschel, Ikuro Abe, Shigeki Matsunaga, Jörn Kalinowski, Haruko Takeyama, and Jörn Piel. An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, 506(7486):58–62, 2014.

[105] Jacob R Winnikoff, Evgenia Glukhov, Jeramie Watrous, Pieter C Dorrestein, and William H Gerwick. Quantitative molecular networking to profile marine cyanobacterial metabolomes. *The Journal of antibiotics*, 67(1):105–112, 2014.

[106] David S. Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D. Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alim M.

Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: The human metabolome database. *Nucleic Acids Research*, 35, 2007.

[107] Sunghee Woo, Seong Won Cha, Stefano Bonissone, Seungjin Na, David L. Tabb, Pavel A. Pevzner, and Vineet Bafna. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. *Journal of Proteome Research*, 14(9):3555–3567, 2015.

[108] Mingguo Xu and Liang Li. Validation of peptide ms/ms spectra using metabolic isotope labeling for spectral matching-based shotgun proteome analysis. *Journal of proteome research*, 10(8):3632–3641, 2011.

[109] Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, Akiyama K, Sakurai T, Matsuda F, Aoki T, Hirai MY, and Saito K. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. aug 2012.

[110] Jane Y. Yang, Laura M. Sanchez, Christopher M. Rath, Xueting Liu, Paul D. Boudreau, Nicole Bruns, Evgenia Glukhov, Anne Wodtke, Rafael De Felicio, Amanda Fenner, Weng Ruh Wong, Roger G. Linington, Lixin Zhang, Hosana M. Debonsi, William H. Gerwick, and Pieter C. Dorrestein. Molecular networking as a dereplication strategy. *Journal of Natural Products*, 76(9):1686–1699, 2013.

[111] J R Yates, S F Morgan, C L Gatlin, P R Griffin, and J K Eng. Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis. *Anal Chem*, 70:3557–3565, 1998.

[112] C Y Yen, S Houel, N G Ahn, and W M Old. Spectrum-to-spectrum searching using a proteome-wide spectral library. *Mol Cell Proteomics*, 10:10.1074/mcp.M111.007666, 2011.

[113] Bing Zhang, Jing Wang, Xiaojing Wang, Jing Zhu, Qi Liu, Zhiao Shi, Matthew C. Chambers, Lisa J. Zimmerman, Kent F. Shaddox, Sangtae Kim, Sherri R. Davies, Sean Wang, Pei Wang, Christopher R. Kinsinger, Robert C. Rivers, Henry Rodriguez, R. Reid Townsend, Matthew J. C. Ellis, Steven A. Carr, David L. Tabb, Robert J. Coffey, Robbert J. C. Slebos, Daniel C. Liebler, Steven A. Carr, Michael A. Gillette, Karl R. Klauser, Eric Kuhn, D. R. Mani, Philipp Mertins, Karen A. Ketchum, Amanda G. Paulovich, Jeffrey R. Whiteaker, Nathan J. Edwards, Peter B. McGarvey, Subha Madhavan, Pei Wang, Daniel Chan, Akhilesh Pandey, Ie-Ming Shih, Hui Zhang, Zhen Zhang, Heng Zhu, Gordon A. Whiteley, Steven J. Skates, Forest M. White, Douglas A. Levine, Emily S. Boja, Christopher R. Kinsinger, Tara Hiltke, Mehdi Mesri, Robert C. Rivers, Henry Rodriguez, Kenna M. Shaw, Stephen E. Stein, David Fenyo, Tao Liu, Jason E. McDermott, Samuel H. Payne, Karin D. Rodland, Richard D. Smith, Paul Rudnick, Michael

Snyder, Yingming Zhao, Xian Chen, David F. Ransohoff, Andrew N. Hoofnagle, Daniel C. Liebler, Melinda E. Sanders, Zhiao Shi, Robbert J. C. Slebos, David L. Tabb, Bing Zhang, Lisa J. Zimmerman, Yue Wang, Sherri R. Davies, Li Ding, Matthew J. C. Ellis, and R. Reid Townsend. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–387, 2014.

[114] Xin Zhang, Yunzi Li, Wenguang Shao, and Henry Lam. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics*, 11(6):1075–1085, 2011.

[115] Zhongqi Zhang. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry*, 76(14):3908–3922, 2004.

[116] Daniel P Zolg, Mathias Wilhelm, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, Bernard Delanghe, Derek J Bailey, Siegfried Gessulat, Hans-Christian Ehrlich, Maximilian Weininger, Peng Yu, Judith Schlegl, Karl Kramer, Tobias Schmidt, Ulrike Kusebauch, Eric W Deutsch, Ruedi Aebersold, Robert L Moritz, Holger Wenschuh, Thomas Moehring, Stephan Aiche, Andreas Huhmer, Ulf Reimer, and Bernhard Kuster. Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods*, 2017.