SOFTWARE SOLUTIONS FOR IMPROVED MASS SPECTROMETRY DATA DISSEMINATION

by

Dain R. Brademan

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2020

Date of final oral examination: 06/29/2020

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Chemistry and Biomolecular Chemistry
Lingjun Li, Professor, Chemistry
Lloyd M. Smith, Professor, Chemistry
Judith Simcox, Assistant Professor, Biochemistry

© Copyright by Dain R. Brademan 2020

All Rights Reserved

To those who taught, uplifted, and inspired me.

ACKNOWLEDGMENTS

It is often said that the personality of an individual is an amalgamation of the five people they spend the most time with. I would argue that number is too low. I have an immeasurable number of people in my life to thank for the professional and personal development I underwent during my time at UW-Madison. I wish I could name you all individually, but I don't have the space in this thesis for everybody who has ever influenced me, even in passing, throughout the years. With that being said, here are some highlights:

First, I want to give a big 'thank you' to my advisor, Professor Josh Coon. As a prospective student, I applied to UW-Madison with the aspiration to join your lab—I admired your accomplishments and your ability to foster community. As your student, you trusted me with the freedom to explore science while challenging me to find what I am passionate about. I would not have discovered my love for programming without this flexibility. It is through your mentorship, guidance, and supportive high-expectations, that I have grown leaps and bounds as a scientist. I look forward to continuing my research with you as a post-doc.

I also want to thank the past and present members past and present of the Coon Lab. Mike Westphall, you were always there to support me with a joke and advice. You are the best distinguished instrument innovator that we could have. You somehow fixed all the things I broke back when I did work inside the lab and not just at my desk. Alex Hebert, I am forever thankful for your mentorship and for teaching me how to get a cat out of a

wall. Now I have two cats in my wall. Evgenia Shishkova, I am grateful for your influential leadership as you seamlessly transitioned from a senior student to a staff member. Katherine Overmyer, you always had time to lend an ear and offer empathetic encouragement to help overcome the hurdles I encountered. Ian Miller, working with you on all the different computational projects was a treat. Brett Paulson, the smell of Chez Brett's daily pots of coffee brought me energy on those slow mornings. Emily Wilkerson, your guidance during my first few formative years in graduate school led me to persevere. Nick Riley, you were the worst. But seriously– despite your busy schedule, you always able to find time to answer questions, brainstorm ideas, or proofread a manuscript. I do not know how you did it. Nick Kwiecien, I am grateful for your mentorship when I began programming. I would not have accomplished what I did without your guidance (and old code—thanks!). Paul Hutchins, you are one of the reasons I joined the lab. I am eternally grateful I worked under you during rotations. Erin Weisenhorn, I always looked forward to lab board game night at your apartment. Anji Trujillo and Vanessa Linke, (I felt it was unfair to separate you two into different sentences) you both are glowing beacons of positivity and friendship. It was hard to be in low spirits when you both were always there with a quick word of encouragement and a smile, or a coffee, or tequila, or dinner, etc. Gary Wilson, I will miss our coding discussions over the barrier between our desks. Justin McKetney, you were the best bay buddy I could ask for. Also, thanks for eating all my excess candy. My waistline didn't need it. Jean Lodge, it was great to have somebody who understands the Duluth culture and accent. Also, seeing Teddy regularly was a great bonus. Trent Peters-Clarke

and Ben Anderson, I enjoyed our slow food days and fried cauliflower taco late nights. Laura, Yunyun, Yuchen and Lia, I know you will succeed in the rest of your years at UW and beyond.

I am thankful my amazing committee members: Dr. Lingjun Li, Dr. Lloyd Smith, and Dr. Judith Simcox. You were my 'outside eyes' and I do not undervalue the ways in which you have helped me grow. Your support, professional advice, and service on my committee through the past five years is much appreciated.

I also want to thank my undergraduate advisor, Professor Douglas Beussman. You gave me my first opportunity to conduct research during summer term at St. Olaf College, and that summer cemented my decision to attend graduate school. It was under your guidance where I initially learned about mass spectrometry while isotopically fingerprinting thread fibers for forensics. I am forever thankful that you also gave me my first opportunity to present my research at conferences—it pushed me out of my comfort zone and made me grow professionally.

On a few more personal notes,

I am appreciative of my adopted bargain cat, Seraphina. I did not choose you; you chose me. You made my apartment home, and your unsolicited contributions to this thesis kept the writing process entertaining. ;;;;;/////////////////////////

I cannot thank my parents, Eric and Cindy, enough for all that they have done throughout my life. You have always been there to encourage me to do new things, and you were there to offer support if that led to me biting off more than I could chew. You taught me to work

hard and inspired me to keep my nose to the grindstone. I would not have even considered graduate school without your unwavering support, and I would not have gotten where I am today without learning the art of stubborn persistence from you both.

Trey and Jace, thank you for always helping me unwind from a stressful week by our team video games, board games when we were together at holidays, and D&D sessions. I do not tell you often enough how much I appreciate having you two as my brothers. To my extended Brademan and Tuominen family and to all my friends from Esko– you always helped me to see the bigger picture in life beyond my graduate school bubble. While you may not have always understood my research, I am glad you asked me about it every time you saw me. They say it takes a village and I am glad that you are mine.

Finally, I want to express my gratitude and love to my fiancèe, Krista. You were my strength before I learned how to be strong myself. When I found myself overwhelmed, you were always there to remind me to take a breath, re-center myself, and get back to it. You pushed me to be the best version of myself, and I will be forever grateful for it. I cannot wait for the rest of our lives together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

m/z	Mass-to-charge ratio
m	Mass
z	Charge
inSeq	Instant sequencing algorithm
a.u.	Arbitrary unit
AC	Alternating current
ACN	Acetonitrile
AGC	Automatic gain control
AI-ETD	Activated-ion electrontransfer dissociation
AI-ETD+	Activated-ion electrontransfer dissociation supplemented by an IR pulse after reaction
AI-NETD	Activated-ion negative electron transfer dissociation
AngularJS	Angular Javascript, a Javascript framework
API	Application programming interface
ATP	Adenosine triphosphate
BCA	Bicinchoninic acid protein assay
BLAST	Basic local alignment search tool
C#	C Sharp, a programming language
C	Celcius
CAA	Chloroacetamide

CAD	Collision-activated dissociation
CE	Cholesteryl Ester
CEO	Calculated elution order
Cer	Ceramide
Chr	Chromosome
CI	Chemical ionization
CID	Collision-induced dissociation
CL	Cardiolipin
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
CSMSL	C# Mass Spectrometry Library
CSV	Comma separated value
cURL	PHP Client URL library
CV	Coefficient of variation
D3.js	Javascript plotting library
Da	Dalton, the atomic mass unit
DB	Database
DC	Direct current
DDA	Data-dependent acquisition
DG	Diglyceride
DIA	Data-independent acquisition

DNA	Deoxyribonucleic acid
DO	Diversity Outbred
DOMSSA	OMSSA search using HTCondor
DT	Decision tree
DTT	Dithiothreitol
ECD	Electron-capture dissociation
EI	Electrospray ionization
EO	Elution order
EOA	Elution order algorithm
ESI	Electrospray ionization
ETcaD	ETD with supplemental activation of all reaction products via CAD
ETD	Electron-transfer dissociation
EThcD	ETD with supplemental activation of all reaction products via HCD
E-value	Expectation value
FA	Formic acid
FAB	Fast-atom bombardment
FASTA	A format for storing protein sequences
FDR	False discovery rate
FT	Fourier transform
FT-ICR	Fourier transform ion cyclotron resonance

GB	Gigabyte
GC	Gas chromatography
GO	Gene Ontology
GUI	Graphical user interface
HCD	Higher-energy collisional dissociation
HPLC	High-performance liquid chromatography
HTCondor	High-throughput Condor
HTML	Hypertext markup language
Hz	Hertz, inverse seconds
ID	Identifier
IDA	Intelligent data acquisition
INC	Inclusion list
IP	Intellectual property
IPSA	Interactive Peptide Spectral Annotator
IT	Ion trap
ITCL	Ion trap control language
JSON	Javascript object notation
LC/MS	Liquid chromatography/mass spectrometry
LC	Liquid chromatography
LOD	Logarithm Of Odds

M	Molar
MALDI	Matrix-assisted laser desorption-ionization
Mbp	Megabase Pair
MGF	Mascot generic format
MGI	Mouse Genome Informatics consortium
min	Minute
MQ	MaxQuant
mRNA	Messenger RNA
MS^1	Survey mass analysis
MS^2	Tandem mass spectrometry
MS^n	Tandem mass spectrometry
MS/MS	Tandem mass spectrometry
MS	Mass spectrometry
MySQL	A relation database which uses SQL
NCE	Normalized collisional energy
NCQBCS	National Center for Quantitative Biology of Complex Systems
NETD	Negative electron transfer dissociation
nLC	Nanoflow liquid chromatography
Ome, Omics	Pertaining to a class or several classes of biomolecules
OMSSA	Open Mass Spectrometry Search Algorithm

PCA	Principal component analysis
PDO	PHP data objects
PE	Phosphatidylethanolamine
pH	Potential of hydrogen
PHP	PHP: hypertext preprocessor
PI	Phosphatidylinositol
ppm	parts per million, a measure of mass error
PRM	Parallel reaction monitoring
PSI	Proteomics Standards Initiative
PSM	Peptide-spectrum match
PTM	Post-translational modification
QA	Quality assurance
QC	Quality control
QM	QuantMode
QTL	Quantitative trait loci
RNA	Ribonucleic acid
RP	Reverse phase
RT	Retention time
S/N	Signal-to-noise ratio
s	Second

SCX	Strong-cation exchange
SDF	Site-determining fragment
SILAC	Stable isotope labeling by amino acids in cell culture
SIM	Single ion monitoring
SM	Sphingomyelin
SNP	Single nucleotide polymorphism
SQL	Structured query language
SRM	Selected reaction monitoring
SVG	Scalable vector graphics
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
TG	Triglyceride
Th	Thomson, the unit of the mass-to-charge ratio
TIC	Total ion current or chromatogram
TMT	Tandem mass tag
TRIS	Tris(hydroxymethyl)aminomethane
μg	Microgram
UHP	Ultra high pressure
μL	Microliter
μM	Micromolar

UPLC	Ultra performance liquid chromatography
UHPLC	Ultra-high performance liquid chromatography
UVPD	Ultraviolet photodissociation
x g	Times gravity
XIC	Extracted-ion chromatogram
XML	Extensible markup language

ABSTRACT

Mass spectrometry (MS) has proven itself as an indispensable technique for the deep characterization of multiple distinct biomolecule classes. The research described within this dissertation focuses on addressing the challenges in interpreting previously disparate data sets and the implementation of software-based solutions. **Chapter 1** provides an overview of the fundamentals of bioanalytical mass spectrometry and highlights the implementation of new online tools for the cloud-based dissemination of mass spectrometry data. **Chapter 2** describes the implementation of an online spectral annotation platform which enables the high-throughput annotation of both positive and negative mode peptide tandem mass spectra. A novel platform enabling the longitudinal tracking of proteomic instrument performance is described in **Chapter 3**. Using the diversity outbred (DO) mouse model, **Chapter 4** characterizes the relationships between genetics and lipids and leverages this information to identify previously unknown lipid species using an online data exploration tool. In **Chapter 5**, a reusable, web-based multi-omic data portal is developed and applied to a large cohort of gene deletion yeast strains. **Chapter 6** summarizes the improvements made with respect to the dissemination of mass spectrometry-based omics data by the resources described within this dissertation. Potential future developments are highlighted. Finally, the proposed research I will conduct as a post-doctoral researcher in the Coon and Simcox Labs is discussed in brief.

Chapter 1

BACKGROUND AND INTRODUCTION

Background

The complete characterization of all aspects of a biological system is an end goal for life science researchers. Unfortunately, this goal is much easier said than done. The complex regulatory relationships which exists within and between an organism's diverse classes of biomolecules present a constant challenge for researchers correlating different biological phenomena. **Figure 1.1** provides a simple graphical representation of increasing biomolecular complexity as one travels further from the genome. Deoxyribonucleic acid (DNA), or the genome, holds the blueprints for life inside a cell. These blueprints are encoded using two complementary chains of the 4 canonical nucleobases, adenine (A), guanine (G), thymine (T), and cytosine (C). The nucleobases within a single chain are connected by a phosphate-sugar backbone, while the complementary chains are held together through hydrogen bonding. The instructions for life are encoded in specific segments of DNA called genes. These genes are read by the enzyme RNA polymerase and are transcribed into messenger ribonucleic acid (mRNA). mRNAs are transported to the ribosome where they are translated into proteins. Specifically, sets of three consecutive mRNA nucleobases encode for a particular amino acid. The ribosome iterates over an mRNA molecule and uses transport RNAs (tRNA) to methodically construct a polypeptide. The process of transcribing DNA to mRNA, and translating mRNA to proteins is known as the Central Dogma of Molecular Biology.[1,2] Alternative splicing, or the ability of a single mRNA to produce several protein isoforms with distinct functions, is a regulatory method which increases protein

diversity.[3] Proteins serve as the main machinery of a cell. To name a few functions, proteins act as enzymatic catalysts to enable energetically unfavorable biochemical reactions, they transport materials across cell membranes, and they can regulate other proteins through the addition or removal of post-translational modifications (PTMs). PTMs, when bound, cause slight conformational shifts in proteins. These conformation changes impact protein function and can be used to regulate cellular metabolism. Metabolites and lipids are small molecules which participate with enzymatic proteins to drive the biochemical pathways within an organism. As these small molecules are heavily involved in cellular metabolism and the structure of cellular components, their regulation is extremely dynamic.

DNA, RNA, proteins, lipids, and metabolites comprise the vast majority of biomolecules within a cell and together comprise an organism's phenotype. As such, these five biomolecular classes must be fully characterized to best understand the complex regulatory relationships which drive biology. With the discovery of DNA polymerase I by Arthur Kornberg in 1957 and the invention of the polymerase chain reaction (PCR), DNA and expressed RNA could be isolated and replicated to high yields.[4,5] The field of nucleic acid sequencing exploded thanks to the implementation of high-throughput sequencing technologies, leading to the sequencing of the human and other genomes.[6–8] However, no such biomolecule replication technique exists for proteins, lipids, or metabolites. Characterization of these biomolecules lagged behind until the mainstream introduction of mass spectrometry, an analytical technique with the sensitivity necessary to detect and quantify proteins, lipids, and metabolites at biological levels reproducibly.

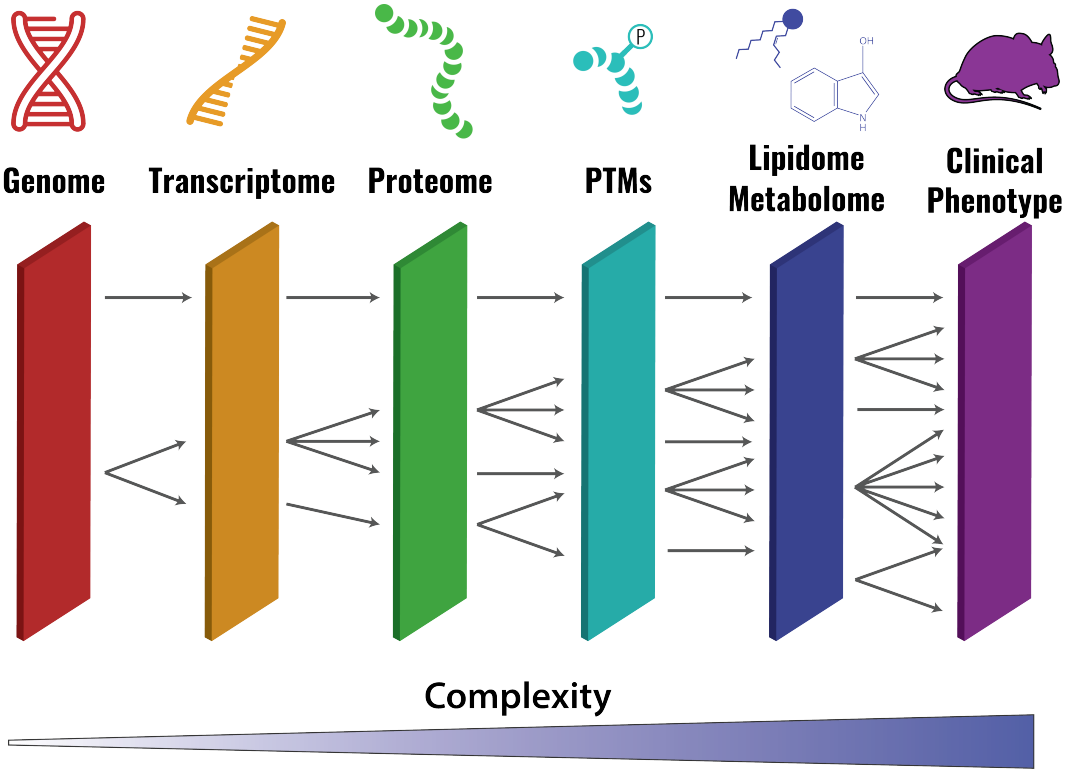


Figure 1.1: Omics Relationships. The DNA contained within the genome is transcribed to mRNA. These mRNAs are then translated into polypeptides or proteins. Proteins participate in a wide variety of cellular functions. Proteins can be post-translationally modified to change their chemical properties. Lipids and metabolites are essential components of cellular structure and metabolism. Metabolite and lipid abundances are often regulated via enzymatic proteins. All these components compose an organism's phenotype.

Bioanalytical Mass Spectrometry

A mass spectrometer operates on the simple principle that a charged particle in the gas phase will be deflected in the presence of an electric or magnetic field. The magnitude of the deflection force on the ion can be varied by modulating the field strength inside a mass spectrometer. It is important to note that a mass spectrometer does not measure mass directly. It instead measures the mass-to-charge ratio of an ion in the gas phase. J.J. Thomson was the first individual to exploit this phenomenon in his cathode-ray experiments.[9] Using his apparatus, he was able to generate evidence for the existence of the electron. Later, Francis Aston collected the first mass spectrum of several positively-charged ions derived from residual gas.[10,11] Arthur Dempster created the first magnetic sector mass spectrometer in 1918, beating Aston to the creation of the first 'modern' mass spectrometer, though Aston completed his platform the next year.[12] For the next several decades, development of mass spectrometer technologies stagnated as only atoms or small (in)-organic molecules could be induced into the gas phase without decomposition, limiting the purview of MS technology to other fields outside of physics for a time.

Ionization Early mass spectrometers almost unilaterally employed electron impact ionization, now known as electron ionization (EI).[13] EI is conducted by passing an analyte through a focused beam of electrons generated from a filament. The electron beam will knock an electron off the analyte, generating an ion. EI is categorized as a "hard" ionization

process, meaning EI typically induces fragmentation of the parent species. While these fragments are structurally informative, occasionally the parent ion will be missing from the spectra. Chemical ionization (CI) was developed as a "soft" alternative to EI.[14] A chemical reagent (commonly CH_4) is ionized by EI to form radical cations. This reagent cation is then mixed with the analyte, and a proton can be transferred on a cation-analyte collision. EI and CI still required the analyte to be in the gas phase before ionization. It wasn't until 1988 when Fenn and Tanaka concurrently released the new soft ionization techniques matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) that larger, more fragile biomolecules like proteins or lipids could be induced into the gas phase intact to undergo mass analysis.[15–17]

Mass Analysis and Detection As described above, early mass spectrometers used electronic or magnet sectors to conduct mass analysis of ions. The work described in this dissertation utilizes data collected on instruments which employ quadrupole, 2-dimensional quadrupole linear ion trap (2D-QLIT), and Orbitrap mass analyzers.[18–20] Quadrupole mass analyzers are constructed using 4 parallel rods arranged in a square configuration. Ions are shuttled from the inlet of the mass spectrometer to one end of the quadrupole.[18] A direct current (DC) offset potential is applied to two opposite rods, while an oscillating radio frequency (RF) potential is applied to the remaining two rods. By modulating the magnitudes of the DC offset potential and the RF frequency, different m/z values can be stabilized as described by the Mathieu equation.[21] Quadrupole mass analyzers can also be

run in 'pass-through' mode, where only the RF voltage is applied. This places all incoming ions in the region of stability. The 2D-QLIT employed on our mass spectrometers function similarly to a normal quadrupole with added ability to trap ions within the cell. To enable trapping, the central quadrupole operates in pass-through mode, while two end electrodes are added which induce large potential wells at both ends of the 2D-QLIT to prevent ions from exiting the trap from both ends. Once ions are trapped, they are confined radially by the central electrode's RF field. Trapped ions can be selectively destabilized by modulating the central quadrupole's DC and RF voltages to detect specific ions.

Ion detection is considered a distinct function from most mass analyzers. The simplest ion detector is a Faraday cup which is basically a shielded metal plate. Ions are neutralized on the detector, and charge is transferred to the plate during neutralization. When the Faraday cup is discharged, the measured electric potential is proportional to the number of charges neutralized.[22] Ions in our instrument's 2D-QLITs are detected using electron multipliers. When a singly-charged ion impacts the electron multiplier, several electrons are generated and they in turn impact the electron multiplier again. The signal gain electron multipliers provide enable the detection of lower abundance species. The Orbitrap mass analyzer is composed of an inner spindle-like electrode and an outer shell-like electrode. Ion packets are injected into the Orbitrap's electric field and oscillate back and forth in a frequency proportional to the ion m/z. Ions are not detected directly in the Orbitrap.[23] An image current is induced as ions oscillate around the center spindle. The image current takes the form of an interferogram. The interferogram is Fourier transformed to translate

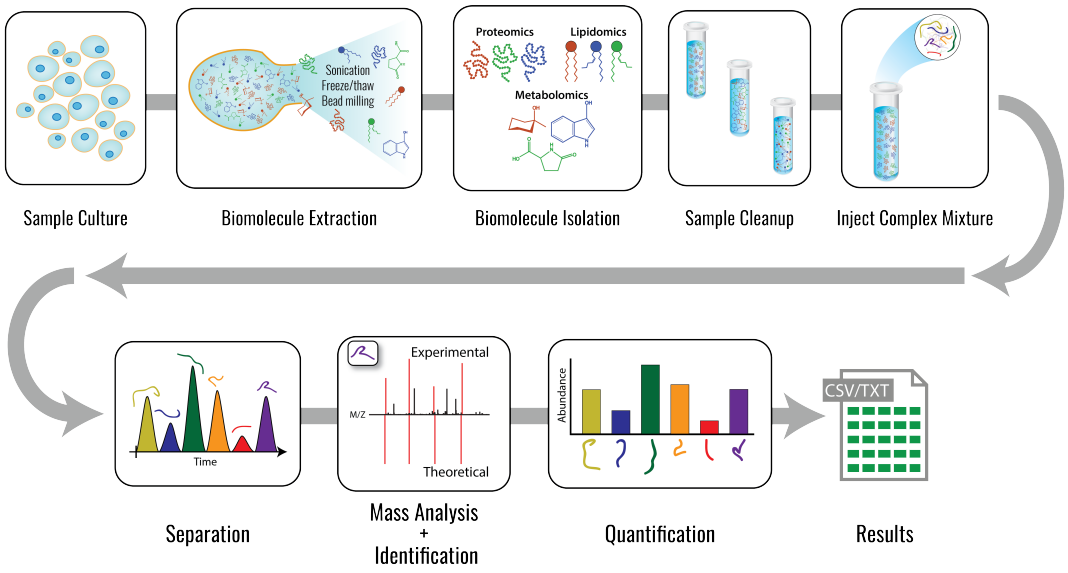


Figure 1.2: General MS Omics Workflow. Regardless of targeted biomolecule class, experimental preparation begins with procuring a biological sample. The sample undergoes lysis to break open cellular compartments and extract biomolecules of interest. Classes of biomolecules are often separated at this point in sample preparation, and the sample is prepared for mass analysis. Complex mixtures of biomolecules typically undergo chromatographic separation. As biomolecules elute, they are ionized, typically using some form of ESI or EI. Biomolecules are isolated in the mass spectrometer and fragmented. These fragment ions are scanned out and the resulting mass spectrum can be compared to theoretical or library spectra for identification. Finally, chromatographic peaks are quantified, linked to the tandem mass spectra identifications, and a list of profiled biomolecules is written out.

from the frequency domain to m/z. Additionally, the Orbitrap offers the highest resolving power of the three listed mass analyzers. Mass resolution is proportional to the square root of the number of oscillations on the spindle. This implies mass resolution in the Orbitrap is dependent on m/z since oscillation speed is also dependent on m/z.

Omics Sample Analysis The Coon lab has a high level of expertise in the collection and analysis of proteomic, lipidomic, and metabolomic data sets.[24–28] **Figure 1.2** provides an general overview for the sample preparation and data collection steps which are shared between our omics pipelines. Specific steps pertaining to individual biomolecular classes are described in the sections below. Sample preparation begins for all biomolecule classes with a lysis step to extract target biomolecules from the sample. Depending on the sample's durability, lysis can be accomplished through bead milling, probe sonication, or using freeze/thaw or boil/cool cycles. Once cellular compartments are disrupted, a single class of biomolecule is isolated from the complex lysate. Next sample cleanup is conducted to prepare the sample for chromatographic separations. The sample is then injected into the gas or liquid chromatography platform. As biomolecules elute from the chromatographic column, they are ionized and introduced to the mass spectrometer for analysis. Once data collection has finished, the raw spectral data is entered into data processing pipeline, where biomolecules are identified, quantified, and compiled into a list of biomolecule abundances.

Proteomics The majority of proteomic analyses conducted utilize a bottom-up approach.[29] In a bottom-up, or shotgun, proteomics workflow, intact proteins are isolated, denatured, disulfide bonds are reduced and alkylated, and are finally digested using a proteolytic enzyme. Digestion cleaves the protein into smaller, more MS-amenable peptides. Trypsin is a popular enzyme choice as it generate peptides with an average length of 8-9 amino acids.[30] Additionally, trypsin cleaves at the C-terminus of lysine and arginine, leaving a free amine to take on an additional charge during ionization. These peptides are then loaded onto a RP-UHPLC column and undergo chromatographic separation. As they elute, peptides are transition to the gas phase and are ionized using ESI, and the peptides enter the MS through the inlet capillary.

Spectral data can be collected using either a data-dependent acquisition (DDA) or data-independent acquisition (DIA) strategy.[31] DDA approaches first utilize a survey scan to detect what m/z features are eluting at the moment. Next, the instrument isolates a single m/z feature (ranked by intensity) and fragments the species to generate sequence informative ions. Many ion activation methods have been developed. They can be grouped into collision-based activation: Collision-induced dissociation (CID)[32], and higher-energy collisional dissociation (HCD)[33]; electron-based activation: Electron capture dissociation (ECD)[34], electron transfer dissociation (ETD)[35]; or photon-induced activation: Ultraviolet photodissociation (UVPD)[36], and infrared multiphoton photodissociation (IRMPD)[37]. DIA approaches divide the instrument's mass range into a number of equally spaced bins. Iteratively, eluting peptides contained in each mass bin are simultaneously isolated and

fragmented. This process repeats until the run has completed.

After data collection, tandem mass spectra are searched against a database containing theoretical peptides. These theoretical peptides are calculated from an *in-silico* digestion of a protein database.[38] SEQUEST was the first implementation of an algorithm to systematically correlate MS^2s to a protein database.[39] In brief, theoretical peptides which are close in mass to a collected MS^2 undergo *in-silico* fragmentation. The resulting theoretical spectrum is compared to the experimental spectrum, and a scoring metric is returned which is representative of how well the experimental and theoretical spectra matched. It is possible for incorrect sequences to be assigned to an MS^2 through random chance. The frequency of these incorrect assignments are controlled for by using a target-decoy strategy.[40] Target-decoy approaches inject known incorrect theoretical peptides into the protein database. These decoy peptides are used to estimate the relative frequency of incorrect sequence assignments in the final dataset. Relative peptide quantification can be conducted using label-free quantification. Label-free quantification is conducted by integrating the measured MS^1 elution profile of a peptide species. However, label-free methods cannot determine the absolute abundance of a peptide in a sample. Isobaric tagging strategies implemented before mass analysis such as TMT[41], iTRAQ[42], or SILAC[43] can enable absolute quantification. Finally, peptides can be reassembled into protein groups. Some peptides can match back to multiple proteins, so protein groups are formed to indicate this ambiguity.[44]

Lipidomics Lipids and metabolites can be isolated concurrently using a biphasic methyl-*tert*-butyl ether extraction as described by Matyash, *et al*.[45] Lipids are extracted into the top organic layer while polar metabolites are extracted into the lower aqueous layer. Proteins precipitate out of solution and can be discarded. After extraction, lipids can be dried down and reconstituted in mobile phase A for lipidomics LC/MS analysis. There are some similarities between proteomics and lipidomics LC/MS acquisitions. Lipids are amenable to both reverse phase chromatography and hydrophilic interaction liquid chromatography (HILIC).[46] HILIC separates lipids mainly by head group composition, while reverse phase separates lipids through stationary phase interactions with the fatty acid chains. Since both peptides and lipids separate well under reverse phase conditions, it may be possible to analyze both biomolecule classes in a single injection. The Coon lab employs a DDA approach with HCD activation to conduct mass analysis on eluting lipids. Many lipids ionize preferentially as either cations or anions. To account for this factor, mass analysis is conducted in polarity-switching mode. In polarity-switching mode, the instrument collects a survey scan and respective tandem mass spectra for lipid cations. The LC/MS then switches its electronics to negative mode to analyze anionic lipids for a single survey scan and its respective tandem mass spectra. This process repeats until the run completes.

Post-acquisition, tandem mass spectra are searched to assign putative identifications to sampled chromatographic features. Similar to peptides, lipids also fragment in a predictable manner. As such, fragmentation rules for individual lipid classes can be derived from several high-quality experimental spectra.[47] However, lipids do not have a comparable

'protein database' to generate theoretical lipid species. Instead, lipidomics search algorithms must use spectral libraries for spectral matching.[48] Spectral libraries contain annotated tandem mass spectra generated from high-quality lipids standards. Unfortunately, these libraries do not contain spectra for all existing lipid species. This leads to the presence of many unknown lipid-like chromatographic features in the final quantitative output. LipiDex is a recently released lipidomics software suite which is capable of learning fragmentation rules for new lipid classes if provided with several high-quality representative spectra, so a matching lipid standard is not necessary in all cases.[49] This functionality allows the *post-hoc* assignment of identifications to previous unknowns.

Metabolomics Metabolites can be analyzed using either gas chromatography (GC)/MS or LC/MS. For LC/MS analysis, metabolites are injected onto a front-end column and are analyzed in a similar fashion to proteomics DDA experiments, where a survey scan is collected and the most intense features are selected for tandem mass analysis. Gas chromatography separates metabolites by boiling point instead of hydrophobicity.[50] Metabolites elute as the temperature of the GC oven is ramped up during the chromatographic run. This has the advantage of allowing both polar and non-polar metabolites to be analyzed in the same run. During sample preparation for GC/MS analysis, polar functional groups are often derivatized with trimethylsiloxane to lower the boiling point of larger metabolites through the prevention of hydrogen bonding. Additionally, metabolite retention times are extremely reproducible in GC/MS analyses, enabling retention time indexing to filter

out poor identifications. Metabolites elute from the GC column in the gas phase, so they must be ionized using either EI or CI. Since EI is a hard ionization technique metabolites fragment upon ionization. Because of this only MS^1 scans are typically collected. After data acquisition is complete, coeluting metabolite fragments are grouped together in a similar fashion to DIA data processing methodology[31], and the high mass accuracy of the Orbitrap is used to filter poor coeluting chromatographic feature groups.[51]. Similar to lipids, metabolites also rely on spectral libraries to conduct spectral matching.[52]

Online MS Informatics

Today's mass spectrometrists conduct research in an increasingly collaborative environment. **Figure 1.3** visualizes the country-level contribution of international collaborative authors on published chemistry manuscripts from 2017-2018.[53] Addressing the need for technological platforms to share data with other researchers across the globe in an intuitive format is paramount, especially as not all collaborators are MS experts. Additionally, improvements to the depth and throughput of MS profiling technologies have led to increasing data storage needs. The PRIDE database is a centralized repository which allows MS researchers to upload high-quality experimental data for perpetual data storage and to facilitate public reuse of MS data. As of 2019, The PRIDE database accepts ~300 new datasets a month, and it permanently stores over 300 terabytes of experimental data.[54] While PRIDE is a rich data resource, researchers who are not experts in mass spectrometry data analysis may not have the experience necessary to properly leverage resources like PRIDE to their full potential.

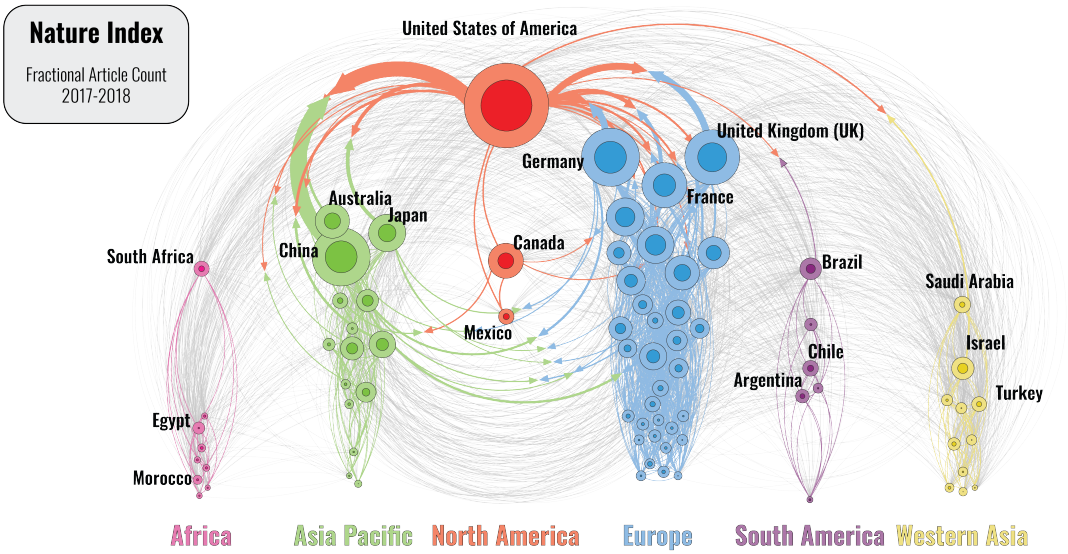


Figure 1.3: International Manuscript Collaborations. Chemistry manuscripts across the world often contain one or more authors from another nation than where the work originated. Each nested circle represents a country with the two areas being proportional to the number of authors from foreign nations divided by the total number of authors.

Scientific software applications hosted on 'The Cloud' have demonstrated great utility in disseminating study results.[25,55–57] Web applications benefit from centralization. The software only exists in one location which eases the burden of resource maintenance. Any researcher with access to a computer can navigate to these web applications using their web browser. Finally, the development of an online resource removes the burden of data interpretation and exploration from a non-expert user onto the resource's developer.

The work described in these successive thesis chapters describe the implementation and use cases of four web-based applications built for interpreting and sharing mass spectrometry omics data with the scientific community. **Chapter 2** describes a web-based platform which facilitates the post-acquisition annotation of peptide tandem mass spectra. This software supports processing spectra collected in both positive and negative modes, and additionally provides a method for users to extract fragment ion statistics from thousands of peptide spectral matches simultaneously. **Chapter 3** presents a quality control website internally utilized by the Coon laboratory to characterize proteomic instrument performance. Quality control metrics which are diagnostic of specific instrument issues are explored. Developmental features which enable instrument-agnostic performance tracking are then discussed in brief. **Chapter 4** introduces a preconfigured software package which enables the codeless generation of custom online data portals. These portals allow researchers to upload their multi-omic data for case-control style experiments. Uploaded data is processed and configured to display a set of interactive data visualizations. Additionally, these data portals are password protected and can be securely shared with collaborators. Finally,

Chapter 5 demonstrates the utility of a online visualization tool named LipidGenie which aggregates the quantitative lipidomics measurements from 384 diversity outbred mice. Additionally, LipidGenie enables the exploration of the connections between genetics and lipids, leading to the identification of new lipid biology.

References

[1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[2] F. H. Crick, "On protein synthesis.," *Symposia of the Society for Experimental Biology*, vol. 12, pp. 138–163, 1958.

[3] J. Darnell, "Implications of RNA-RNA splicing in evolution of eukaryotic cells," *Science*, vol. 202, no. 4374, pp. 1257–1260, 1978.

[4] I. R. Lehman, M. J. Bessman, E. S. Simms, and A. Kornberg, "Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli.," *The Journal of biological chemistry*, vol. 233, no. 1, pp. 163–170, 1958.

[5] K. B. Mullis, "The unusual origin of the polymerase chain reaction," *Scientific American*, vol. 262, no. 4, pp. 56–65, 1990.

[6] F. Sanger, A. R. Coulson, B. G. Barrell, A. J. Smith, and B. A. Roe, "Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing," *Journal of Molecular Biology*, vol. 143, no. 2, pp. 161–178, 1980.

[7] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. Fitzhugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. Levine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama,

M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, M. L. Hong, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. De La Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, and M. J. Morgan, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[8] J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Ko-

duru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[9] J. Thomson, "XL. Cathode Rays," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 44, no. 269, pp. 293–316, 1897.

[10] J. J. Thomson, "Rays of Positive Electricity and their Application to Chemical Analysis," *Nature*, vol. 92, no. 2307, pp. 549–550, 1914.

[11] F. Aston, "LXXIV. A positive ray spectrograph," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 38, no. 228, pp. 707–714, 1919.

[12] F. W. Aston and S. C. Lind, "Mass-spectra and isotopes. by f. w. aston.," *The Journal of Physical Chemistry*, vol. 47, no. 6, pp. 465–465, 1943.

[13] A. J. Dempster, "A new method of positive ray analysis," *Physical Review*, vol. 11, no. 4, pp. 316–325, 1918.

[14] M. S. Munson and F. H. Field, "Chemical Ionization Mass Spectrometry. I. General Introduction," *Journal of the American Chemical Society*, vol. 88, no. 12, pp. 2621–2630, 1966.

[15] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, vol. 246, no. 4926, pp. 64–71, 1989.

[16] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, T. Yoshida, and T. Matsuo, "Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 2, no. 8, pp. 151–153, 1988.

[17] M. Karas and F. Hillenkamp, "Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons," *Analytical Chemistry*, vol. 60, no. 20, pp. 2299–2301, 1988.

[18] P. E. Miller and M. B. Denton, "The quadrupole mass filter: Basic operating concepts," *Journal of Chemical Education*, vol. 63, no. 7, pp. 617–622, 1986.

[19] J. C. Schwartz, M. W. Senko, and J. E. Syka, "A two-dimensional quadrupole ion trap mass spectrometer," *Journal of the American Society for Mass Spectrometry*, vol. 13, no. 6, pp. 659–669, 2002.

[20] A. Makarov, "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis," *Analytical Chemistry*, vol. 72, no. 6, pp. 1156–1162, 2000.

[21] I. Kovacic, R. Rand, and S. M. Sah, "Mathieu's equation and its generalizations: Overview of stability charts and their features," *Applied Mechanics Reviews*, vol. 70, no. 2, 2018.

[22] K. L. Brown and G. W. Tautfest, "Faraday-cup monitors for high-energy electron beams," *Review of Scientific Instruments*, vol. 27, no. 9, pp. 696–702, 1956.

[23] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. G. Cooks, "The Orbitrap: A new mass spectrometer," *Journal of Mass Spectrometry*, vol. 40, no. 4, pp. 430–443, 2005.

[24] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, no. 5, pp. 701–714, 2015.

[25] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[26] E. M. Weisenhorn, T. J. Van 'T Erve, N. M. Riley, J. R. Hess, T. J. Raife, and J. J. Coon, "Multi-omics evidence for inheritance of energy pathways in red blood cells," *Molecular and Cellular Proteomics*, vol. 15, no. 12, pp. 3614–3623, 2016.

[27] E. A. Trujillo, A. S. Hebert, D. R. Brademan, and J. J. Coon, "Maximizing Tandem Mass Spectrometry Acquisition Rates for Shotgun Proteomics," *Analytical Chemistry*, vol. 91, no. 20, pp. 12625–12629, 2019.

[28] A. S. Hebert, C. Thöing, N. M. Riley, N. W. Kwiecien, E. Shiskova, R. Huguet, H. L. Cardasis, A. Kuehn, S. Eliuk, V. Zabrouskov, M. S. Westphall, G. C. McAlister, and J. J. Coon, "Improved Precursor Characterization for Data-Dependent Mass Spectrometry," *Analytical Chemistry*, vol. 90, no. 3, pp. 2333–2340, 2018.

[29] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates, "Protein analysis by

shotgun/bottom-up proteomics," *Chemical Reviews*, vol. 113, no. 4, pp. 2343–2394, 2013.

[30] D. L. Swaney, C. D. Wenger, and J. J. Coon, "Value of using multiple proteases for large-scale mass spectrometry-based proteomics," *Journal of Proteome Research*, vol. 9, no. 3, pp. 1323–1329, 2010.

[31] J. D. Chapman, D. R. Goodlett, and C. D. Masselon, "Multiplexed and data-independent tandem mass spectrometry for global proteome profiling," *Mass Spectrometry Reviews*, vol. 33, no. 6, pp. 452–470, 2014.

[32] J. Mitchell Wells and S. A. McLuckey, "Collision-induced dissociation (CID) of peptides and proteins," *Methods in Enzymology*, vol. 402, pp. 148–185, 2005.

[33] J. V. Olsen, B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, "Higher-energy C-trap dissociation for peptide modification analysis," *Nature Methods*, vol. 4, no. 9, pp. 709–712, 2007.

[34] R. A. Zubarev, N. L. Kelleher, and F. W. McLafferty, "Electron capture dissociation of multiply charged protein cations. A nonergodic process," *Journal of the American Chemical Society*, vol. 120, no. 13, pp. 3265–3266, 1998.

[35] J. E. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry," *Pro-*

ceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 26, pp. 9528–9533, 2004.

[36] J. P. Reilly, "Ultraviolet photofragmentation of biomolecular ions," *Mass Spectrometry Reviews*, vol. 28, no. 3, pp. 425–447, 2009.

[37] D. P. Little, J. P. Speir, M. W. Senko, P. B. O'Connor, and F. W. McLafferty, "Infrared Multiphoton Dissociation of Large Multiply Charged Ions for Biomolecule Sequencing," *Analytical Chemistry*, vol. 66, no. 18, pp. 2809–2815, 1994.

[38] R. Apweiler, M. J. Martin, C. O'Donovan, M. Magrane, Y. Alam-Faruque, R. Antunes, D. Barrell, B. Bely, M. Bingley, D. Binns, L. Bower, P. Browne, W. M. Chan, E. Dimmer, R. Eberhardt, F. Fazzini, A. Fedotov, R. Foulger, J. Garavelli, L. G. Castro, R. Huntley, J. Jacobsen, M. Kleen, K. Laiho, D. Legge, Q. Lin, W. Liu, J. Luo, S. Orchard, S. Patient, K. Pichler, D. Poggioli, N. Pontikos, M. Pruess, S. Rosanoff, T. Sawford, H. Sehra, E. Turner, M. Corbett, M. Donnelly, P. Van Rensburg, I. Xenarios, L. Bougueleret, A. Auchincloss, G. Argoud-Puy, K. Axelsen, A. Bairoch, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, L. Bollondi, E. Boutet, S. B. Quintaje, L. Breuza, A. Bridge, E. De Castro, E. Coudert, I. Cusin, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, S. Gehant, S. Ferro, E. Gasteiger, A. Gateau, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hulo, J. James, S. Jimenez, F. Jungo, T. Kappler, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, X. Martin, P. Masson, M. Moinat, A. Morgat, S. Paesano, I. Pedruzzi, S. Pilbout, S. Poux, M. Pozzato,

N. Redaschi, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, E. Stanley, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, W. C. Barker, C. Chen, Y. Chen, P. Dubey, H. Huang, R. Mazumder, P. McGarvey, D. A. Natale, T. G. Natarajan, J. Nchoutmboube, N. V. Roberts, B. E. Suzek, U. Ugochukwu, C. R. Vinayaka, Q. Wang, Y. Wang, L. S. Yeh, and J. Zhang, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Research*, vol. 39, no. SUPPL. 1, 2011.

[39] J. K. Eng, A. L. McCormack, and J. R. 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*, vol. 5, no. 11, pp. 976–989, 1994.

[40] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature Methods*, vol. 4, no. 3, pp. 207–214, 2007.

[41] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS," *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003.

[42] S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid, "Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research," *Proteomics*, vol. 7, no. 3, pp. 340–350, 2007.

[43] S. E. Ong and M. Mann, "A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC)," *Nature Protocols*, vol. 1, no. 6, pp. 2650–2660, 2007.

[44] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data: The protein inference problem," *Molecular and Cellular Proteomics*, vol. 4, no. 10, pp. 1419–1440, 2005.

[45] V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke, "Lipid extraction by methyl-terf-butyl ether for high-throughput lipidomics," in *Journal of Lipid Research*, vol. 49, pp. 1137–1146, American Society for Biochemistry and Molecular Biology, 2008.

[46] A. Anesi and G. Guella, "A fast liquid chromatography-mass Spectrometry methodology for membrane lipid profiling through hydrophilic interaction liquid chromatography," *Journal of Chromatography A*, vol. 1384, pp. 44–52, mar 2015.

[47] P. D. Hutchins, J. D. Russell, and J. J. Coon, "Mapping Lipid Fragmentation for Tailored Mass Spectral Libraries," *Journal of the American Society for Mass Spectrometry*, vol. 30, no. 4, pp. 659–668, 2019.

[48] T. Cajka and O. Fiehn, "LC–MS-based lipidomics and automated identification of lipids using the LipidBlast in-silico MS/MS library," *Methods in Molecular Biology*, vol. 1609, pp. 149–170, 2017.

[49] P. D. Hutchins, J. D. Russell, and J. J. Coon, "LipiDex: An integrated software package for High-Confidence lipid identification," *Cell Syst*, vol. 6, no. 5, pp. 621–625.e5, 2018.

[50] A. T. James and A. J. P. Martin, "GAS-LIQUID CHROMATOGRAPHY: A Technique for the Analysis and Identification of Volatile Materials," *British Medical Bulletin*, vol. 10, no. 3, pp. 170–176, 1954.

[51] N. W. Kwiecien, D. J. Bailey, M. J. Rush, J. S. Cole, A. Ulbrich, A. S. Hebert, M. S. Westphall, and J. J. Coon, "High-Resolution Filtering for Improved Small Molecule Identification via GC/MS," *Analytical Chemistry*, vol. 87, no. 16, pp. 8328–8335, 2015.

[52] "NIST Mass Spectral Library," 2012.

[53] Springer Nature, "Country Collaboration | Nature Index," 2019.

[54] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Perez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yılmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma, and J. A. Vizcaino, "The PRIDE database and related tools and resources in 2019: improving support for quantification data," *Nucleic Acids Research*, vol. 47, pp. D442–D450, 11 2018.

[55] D. Torre, A. Lachmann, and A. Ma'ayan, "BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud," *Cell Systems*, vol. 7, no. 5, pp. 556–561.e3, 2018.

[56] H. Marx, C. E. Minogue, D. Jayaraman, A. L. Richards, N. W. Kwiecien, A. F. Siahpirani, S. Rajasekar, J. Maeda, K. Garcia, A. R. Del Valle-Echevarria, J. D. Volkening, M. S. Westphall, S. Roy, M. R. Sussman, J. M. Ané, and J. J. Coon, "A proteomic atlas of the legume Medicago truncatula and its nitrogen-fixing endosymbiont Sinorhizobium meliloti," *Nature Biotechnology*, vol. 34, no. 11, pp. 1198–1205, 2016.

[57] M. T. Veling, A. G. Reidenbach, E. C. Freiberger, N. W. Kwiecien, P. D. Hutchins, M. J. Drahnak, A. Jochem, A. Ulbrich, M. J. Rush, J. D. Russell, J. J. Coon, and D. J. Pagliarini, "Multi-omic Mitoprotease Profiling Defines a Role for Oct1p in Coenzyme Q Production," *Molecular Cell*, vol. 68, no. 5, pp. 970–977.e11, 2017.

Chapter 2

INTERACTIVE PEPTIDE SPECTRAL ANNOTATOR: A VERSATILE WEB-BASED TOOL FOR PROTEOMIC APPLICATIONS

This chapter is adapted from a published manuscript and is reprinted with permission from:

Brademan DR, Riley NM, Kwiecien NW, Coon JJ. *Interactive Peptide Spectral Annotator: A Versatile Web-based Tool for Proteomic Applications.* Mol. Cell. Proteomics. **2019**, 18, S193–S201. DOI: 10.1074/mcp. TIR118.001209

Abstract

Here we present IPSA, an innovative web-based spectrum annotator that visualizes and characterizes peptide tandem mass spectra. A tool for the scientific community, IPSA can visualize peptides collected using a wide variety of experimental and instrumental configurations. Annotated spectra are customizable via a selection of interactive features and can be exported as editable scalable vector graphics to aid in the production of publication-quality figures. Single spectra can be analyzed through provided web forms, whereas data for multiple peptide spectral matches can be uploaded using the Proteomics Standards Initiative file formats mzTab, mzIdentML, and mzML. Alternatively, peptide identifications and spectral data can be provided using generic file formats. IPSA provides supports for annotating spectra collecting using negative-mode ionization and facilitates the characterization of experimental MS/MS performance through the optional export of fragment ion statistics from one to many peptide spectral matches. This resource is made freely accessible at `http://interactivepeptidespectralannotator.com`, whereas the source code and user guides are available at `https://github.com/coongroup/IPSA` for private hosting or custom implementations.

Introduction

Tandem mass spectrometry (MS/MS) is the centerpiece of modern proteome analysis. Advances in instrument design and acquisition software have enabled collection of well over 100,000 MS/MS scans in less than an hour of analysis.[1–9] Researchers have developed a wide variety of search algorithms and related computational tools to rapidly translate this large volume of experimental data to peptide spectral matches (PSMs), where peptide sequences are assigned to spectra to identify the proteins present in a sample.[10–16] An important component to this process is matching expected product ions to those observed in the experimental spectra. Annotation of spectra in this sense usually involves labeling observed m/z features with matched fragment ion designations (*e.g.* a/x-, b/y-, or c/z-type product ions) derived from the reported peptide sequence. Expert manual annotation is a valuable but greatly time-consuming process—unfeasible for the large volume of spectra generated in modern proteomic experiments.

Proteomic field guidelines have increasingly emphasized the importance of providing access to annotated MS/MS spectra for publication, which allows others to inspect reported PSMs and validate their assignment to a given sequence.[17–20] Many software tools have been created to aid researchers annotating individual PSMs contained in bulk datasets. Most such tools are downloadable and often integrated directly into data-analysis suites, although a handful have been developed as web browser-based platforms.[21–23] Lorikeet (`https://uwpr.github.io/Lorikeet`) is a well-established web-based spectral annotator

which has been integrated into several online mass spectrometry resources to visualize routine shotgun and cross-linked proteomics data.[24–27] However, Lorikeet does not render generated annotated spectra in scalable vector graphics (SVG) format, limiting the flexibility of exported visualizations with regards to figure creation. XiSPEC is a recently release tool which supports the annotation of shotgun and cross-linked proteomics data.[28] However, it does not currently support annotation of spectra collected in the negative mode.

Although powerful for the platforms for which they were designed, many of these tools are inseparable from their respective analytical pipelines; data visualization in MaxQuant is only available following processing with the integrated Andromeda search engine, for example. Their purview is therefore limited, and facile spectral annotation is restricted to only those search algorithms packaged in a pipeline with a developed annotator. This restriction poses a problem for numerous applications, especially for alternative peptide fragmentation methods such as ultraviolet photodissociation (UVPD), collisionally supplemented electron-transfer dissociation (EThcD), or activated-ion electron-transfer dissociation (AI-ETD).[27,29–31] Often these methods can be integrated into established analytical pipelines adopted by the field over the course of several years. But flexible annotation tools are largely unavailable in the beginning stages of method development—arguably when they are needed most. For example, Lorikeet bundles annotation calculations directly with its spectrum viewer. This requires in-depth knowledge of Lorikeet's architecture to add functionality for new technologies. However, separating the annotation process from the spectrum renderer is amenable toward a more stable platform for spectral annotation as

the components can be maintained and implemented independently.

Here we present the Interactive Peptide Spectra Annotator (IPSA) to provide a standalone web platform for annotation and interpretation of peptide tandem mass spectra independent of instrumental platform, identification pipeline, and peptide fragmentation technique. IPSA provides flexibility to annotate spectra containing any of the six common peptide fragment ion types. Importantly, it can export annotated data in a tabular format, which enables the rapid culmination of fragment ion statistics for individual or multiple peptide tandem mass spectra, a useful tool in a wide range of proteomic experiments. We have also built in compatibility with spectra collected in the negative mode, providing a much-needed resource for the continued development of negative-mode proteomic approaches. Further, IPSA offers a platform for the generation and exportation of figure-ready annotated spectra in an editable format. In all, IPSA expands spectral annotation capabilities to all types of shotgun proteomic data regardless of how data was collected or processed.

Methods

Software Development IPSA is composed of two major components: a client-facing interactive web visualizer and a server-side data processor which handles the data processing required for spectral annotation. Client-side visualization software was developed using AngularJS. The D3.js library is leveraged to generate interactive annotated spectra using SVG from annotated data returned from the server after analysis.[32]

Server-side software consists of a set of modular PHP scripts, which perform form

validation, data processing and annotation, file upload handling, and data export. A MySQL database is incorporated to securely cache parsed peptide identifications and spectral information extracted from uploaded data. MySQL integration facilitates data storage and retrieval when annotation requests are submitted to the server.

Example Data Sets Cell pellets of *Saccharomyces cerevisiae* (strain BY4742) containing approximately 1×10^8 cells were harvested from liquid culture by centrifugation ($3000 \times g$, 3 minutes, 4 °C). The supernatant was removed, and the cell pellet was resuspended in 8 M urea, 100 mM tris (pH 8.0). Methanol was added to 90% by volume and vortexed to lyse the cells and induce protein precipitation. The resulting solution was centrifuged ($14{,}000 \times g$, 3 min) to form a protein pellet. The supernatant was removed, and the pellet was resuspended in 8 M urea, 100 mM tris (pH 8.0), 10 mM tris(2-carboxyethyl)phosphine, and 40 mM chloroacetamide. The solution was then diluted to 1.5 M urea with 50 mM tris. Trypsin (Promega, Madison, WI) was added (1:50 enzyme/protein) and was allowed to digest overnight (22 °C). The resultant peptides were acidified ($\text{pH} < 2.0$) using 0.1% (v:v) trifluoroacetic acid (TFA) and were desalted using polymeric reverse phase Strata-X columns. Columns were equilibrated using one bed volume of 100% acetonitrile (ACN), then one bed volume of 0.1% TFA. Peptides were loaded onto the column and washed with two bed volumes of 0.1% TFA. Peptides were eluted by an addition of 500 µL 40% ACN, 0.1% TFA followed by an addition of 650 µL 70% ACN, 0.1% TFA and were then dried and resuspended in 0.2% formic acid. Peptide concentration was determined using a Pierce

quantitative colorimetric peptide assay (Thermo Fisher Scientific, Rockford, IL).

Low pH reverse-phase liquid chromatography was conducted using a Dionex UltiMate 3000 UPLC as described previously.[1,2] Eluting peptides were analyzed using a Q Exactive HF hybrid quadrupole Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) and were fragmented at HCD at 25% normalized collisional energy. Survey scans were taken at a resolution of 60,000 at 200 m/z, whereas tandem mass spectra were collected using a resolution of 15,000 at 200 m/z. The resulting tandem mass spectra were searched using the Coon OMSSA Proteomic Analysis Software Suite (v1.4.1).[33,34] A precursor mass tolerance of $\pm$150 ppm was used, whereas fragment ions were searched using a mass tolerance of $\pm$0.01 Da. A maximum of 3 missed tryptic cleavages were permitted. Carbamidomethylation of cysteine was set as a fixed modification, whereas oxidation of methionine was set as a variable modification. Data was searched against a canonical and isoform *Saccharomyces cerevisiae* database (UniProt, June 10, 2016) concatenated with the reverse protein sequence for decoy generation. A 1% FDR threshold was used at the peptide level, using both e-value and precursor mass accuracy to filter results.

Additional peptide identifications and spectral data were acquired from the previous work of Riley *et al.* to demonstrate IPSA's ability to process PSMs fragmented using alternative dissociation techniques. These include ETD; collisionally supplemented ETD (ETcaD and EThcD); AI-ETD; AI-NETD; and AI-ETD with supplemental infrared photon irradiation post-reaction (AI-ETD+).[35,36]

Results

Design of IPSA IPSA was developed as a versatile web-based spectral analysis tool capable of individual or *en masse* annotation of PSMs generated from experiments that produce any of the six common peptide fragment ion types (**Figure 2.1**). Single spectra can be annotated by entering peptide and spectral data into an intuitive web form, whereas multiple spectra can be uploaded directly to the website to be individually queried or batch processed. Single annotations are conducted using the metrics provided by the user through the web form and are returned client-side to generate an exportable, annotated spectrum. Exported spectra can easily be shared or integrated into figures. Because the individual interrogation of large numbers of PSMs can quickly become tedious, we added functionality to batch process all uploaded PSMs and export the annotations in a tabular format. This feature permits the rapid characterization of tens of thousands of tandem mass spectra.

Single Spectrum Annotation A single peptide spectrum can be annotated by providing the peptide's sequence, precursor charge, maximum allowed fragment charge, and spectral data to the user interface shown in ((**Figure 2.2A**. Expected fragmentation patterns and neutral losses can be selected to specify which theoretical peptide fragment ions are generated during data processing.[37,38] The mass tolerance for matching experimental features to theoretical fragment ions can be set in either ppm or Daltons. A relative intensity, raw intensity, or S/N (if supplied with spectral data) cutoff can be defined to ignore low-

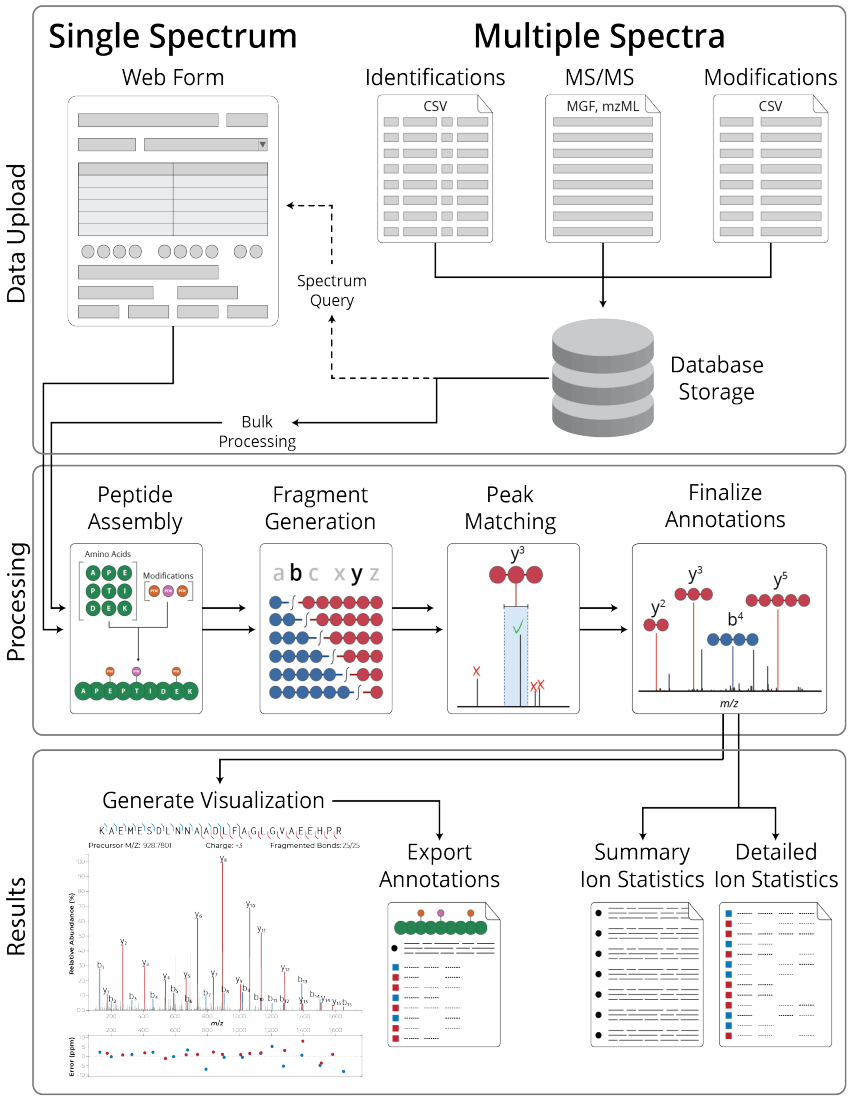


Figure 2.1: IPSA Software Flowchart. Single spectra can be annotated by entering peptide and spectral data into provided web forms. Files containing multiple peptide identifications and spectra can be uploaded either in PSI-supported or generic text-based formats to be individually annotated or to be bulk processed for ion statistic extraction. Theoretical peptides are assembled *in-silico*, fragmented, and matched to the experimental spectrum. The annotated experimental spectrum is then returned and visualized client-side. This visualization can be exported as an SVG image for figure generation or as a CSV file containing ion statistics for the single spectrum. Alternatively, ion statistics for all uploaded peptide spectral matches can be calculated and exported through bulk processing, returning two files containing summary and detailed metrics for each uploaded PSM.

abundance or insignificant features during matching. Visualization colors can additionally be customized.

A predefined list of common protein post-translational modifications (PTMs) can be queried and selected using a searchable dropdown below the fragmentation options. Available PTMs for a peptide are intelligently filtered to only show PTMs relevant to the entered peptide sequence. If a desired PTM is not included in the predefined modification list, new PTMs can be defined and are stored locally in the user's web browser. The user can provide a new modification name, target site, and mass shift to create a custom PTM option.

When the server receives an annotation request, data entered into the user interface is validated and sent for processing. The peptide sequence is parsed and assembled into an intact peptide *in-silico*. Theoretical peptide fragment ions are created from the intact peptide using the fragmentation schema selected by the user. Each fragment is matched to m/z peak within the specified mass tolerance. To address the case that multiple theoretical fragments are mapped to the same experimental feature, only the theoretical fragment that matches with the smallest mass error is reported. Once annotation mapping has been finalized, annotated spectral data is formatted into JSON and is returned to the client for visualization.

Immediately upon this return, IPSA generates the interactive annotated spectrum (**Figure 2.2B**). This visualization consists of three portions: a peptide sequence marked with detected fragment ion locations and summary statistics, an interactive annotated spectrum,

and an interactive scatterplot of the matched fragment-ion mass errors. The visualization supports many interactive features to help facilitate data interpretation. Both axes allow contextual zooming for deeper investigation of congested sections of annotated spectra, whereas tooltips provide exact values for any highlighted plotted experimental features. Highlighted fragments are mirrored in each section of the visualization to emphasize all aspects of the feature of interest. Additionally, annotation labels can be dragged to clearer locations to declutter busy regions. The generated visualization can be exported as an SVG file for figure creation as it appears on screen or in a tabular format at any time.

Bulk Data Upload If many spectra need to be rapidly interrogated, IPSA provides functionality to serially process multiple PSMs by directly uploading files containing peptide identifications and spectral data to the server. Identifications can be provided either in the Proteomics Standards Initiative file formats mzTab or mzIdentML, or in a generic CSV format.[18,19,39] Each row in the generic CSV lists a scan number, peptide sequence, precursor charge, and all PTM names and locations for each peptide identification. We chose this architecture for its simplicity; peptide identifications produced from a wide variety of search algorithms can easily be converted into this format. Spectral data can be uploaded as a Mascot Generic Format (MGF) or mzML file. Finally, a modifications file can be uploaded to link peptide modification names to their respective masses. We provide a set of example files on IPSA's file upload page to demonstrate how each of these files should be structured. MGF and mzML files can easily be generated from vendor or open file formats

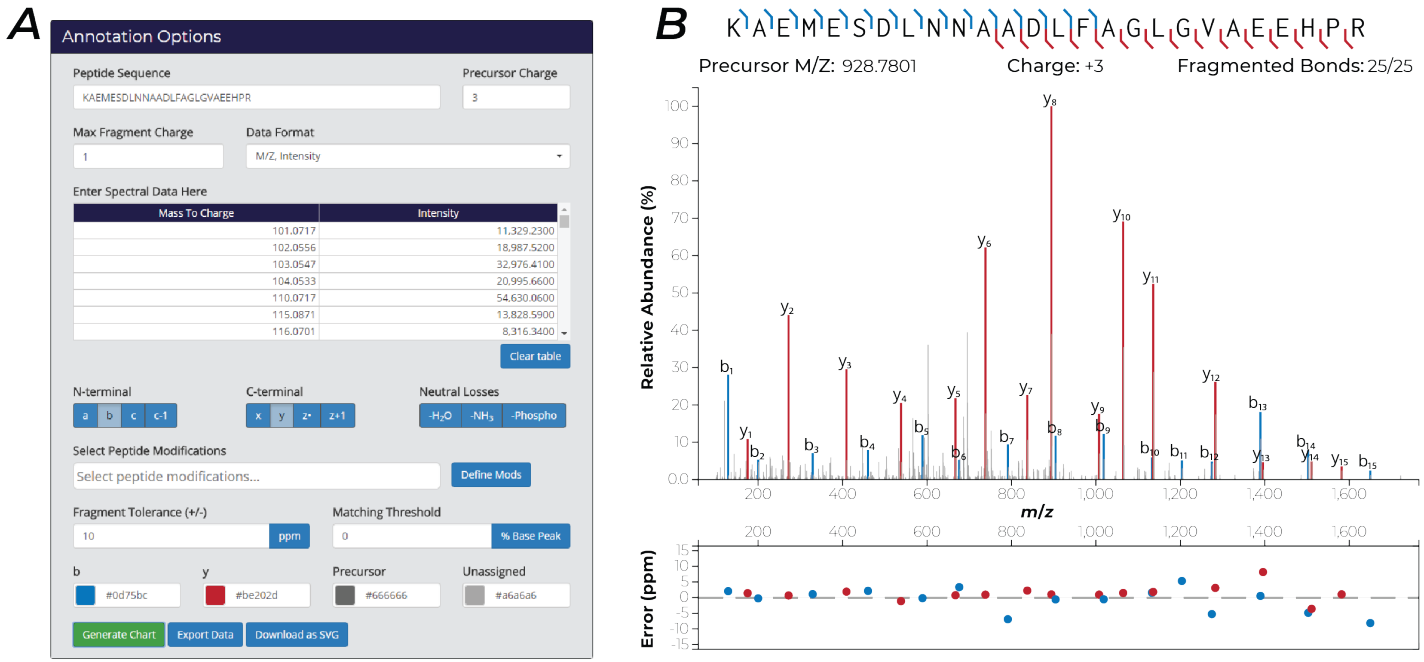


Figure 2.2: Spectrum Annotation. The peptide KAEMESDLNNAADLFAGLGVAEEHPR and its spectral information provide an example PSM. (**A**), Peptide information, fragment ion characteristics, tolerances, and chart colors can be easily set using the provided form. (**B**), The generated interactive visualization after server processing containing the peptide sequence marked with the locations of matched fragment ions, an annotated mass spectrum, and visualization of mass error in either parts-per-million or daltons for all matched fragment ions.

using conversion tools such as MSConvert.[12,40]

Data parsed from bulk identification and spectral data uploads are stored securely server-side in a MySQL database. On data upload, a unique identifier is assigned to the user's browser which is used to exclusively access the uploader's data. After data extraction, uploaded files are deleted to reduce server footprint. Only one data set can be stored at a time.[41]

Negative Mode Annotation Proteomic analyses are typically conducted using low-pH separations and positive-mode electrospray ionization to create peptide cations. This tendency leads to a systematic underrepresentation of acidic peptide species, which preferentially ionize as anions.[42–44] High-pH separations using negative-mode ionization can be used to better study these acidic species, but the complexity of tandem mass spectra generated using traditional collision-based activational methods has precluded the widespread adoption of this mode. This spectral complexity arises in part from a multitude of neutral losses originating from precursor and fragment ions.[45] Alternative fragmentation techniques such as UVPD or AI-NETD, producing a/x-, b/y-, c/z-*type* and $a\bullet/x$-type product ions respectively, have recently demonstrated their utility in producing informative tandem mass spectra from peptide anions[35,44]. However, many spectral annotators do not support these data types. IPSA is capable of annotating PSMs collected using negative-mode electrospray ionization. (**Figure 2.3**) demonstrates an IPSA-annotated spectrum of the triply deprotonated peptide LIPSDFILAAQSHNPIENK dissociated using AI-NETD.[35]

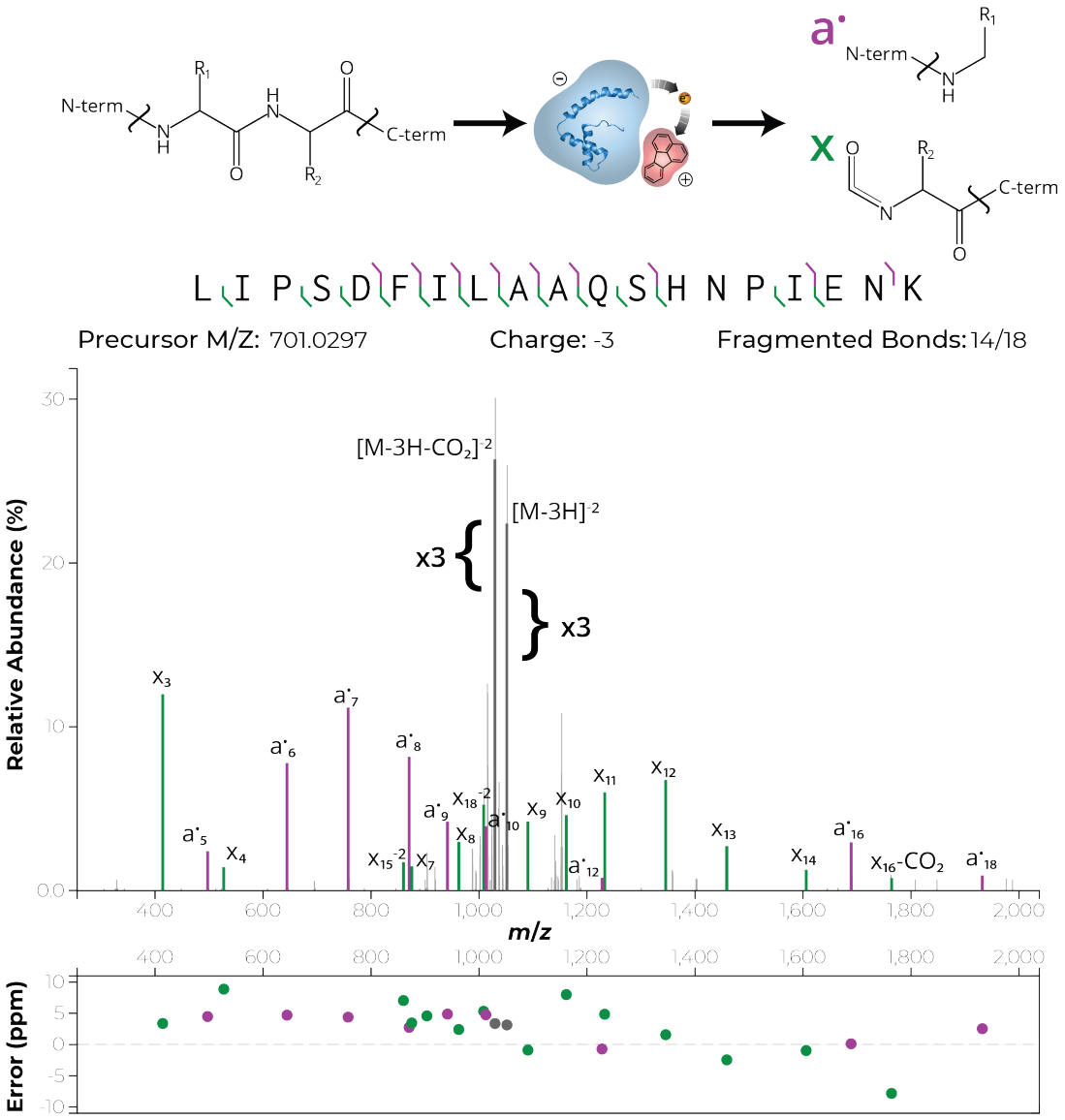


Figure 2.3: Negative-Mode Annotation. The annotated AI-NETD spectrum of the triply deprotonated peptide LIPSDFILAAQSHNPIENK generated using IPSA. The charge-reduced precursor was downscaled by a factor of 3. Unreacted precursor was cleaned from this spectrum.

Ion Statistics Obtaining fragment ion statistics in an automated fashion for an entire mass spectrometry experiment is no trivial task. Fragment ion statistics can be greatly informative during method optimization and can be used to monitor MS/MS performance by providing information on what ion types (and in what amounts) are being generated. Additional informative metrics include the sequence coverage of all detected peptide fragments, fragment ion mass errors, and the percent of the total ion current (TIC) that can be explained by annotated fragment ions.

IPSA provides a unique utility among web-based spectral annotators to compute and export all detected fragment ions for an uploaded experiment in a tabular format. The server extracts the fragment ion series, mass tolerances, and any intensity threshold from the provided user interface and serially processes every uploaded peptide identification. The annotation results are continuously written to a set of two downloadable CSVs. The first file contains summary statistics for the matched fragment ions for each uploaded PSM. This file reports the number of matched fragment ions, unique peptide bonds broken, and the percent of the total ion current explained by matched fragment ions. The second file contains detailed information concerning every detected fragment ion for all uploaded identifications; more specifically, the raw intensity, theoretical m/z, experimental m/z, mass error, percent of base peak, and percent of the total ion current explained is reported.

A series of experiments were previously described by Riley et al. to examine the efficacy of ETD, ETcaD, EThcD, AI-ETD, and AI-ETD+ fragmentation on a liquid chromatography timescale.[36] The authors found AI-ETD+ to be the optimal supplemental ETD fragmenta-

tion technique. Using the authors' reported peptide identifications and spectral data, we created a set of detailed comparisons similar to those made in the referenced manuscript using the ion statistics files directly exported from IPSA (**Figure 2.4**). No further programming was required to extract these data or make this figure, and all data manipulation post-export was performed in a spreadsheet using basic arithmetical functions.

In summary, IPSA is capable of both cleanly annotating peptide spectra collected using a wide variety of dissociation techniques in both positive and negative mode and of exporting the generated annotated spectra in the editable SVG format. Additionally, IPSA allows the bulk analysis of detected fragment ions for any number of uploaded spectra, permitting in turn the deep interrogation of data without requiring programming experience.

Discussion

Modern MS-based proteomics techniques are widely used to identify and characterize tens of thousands of peptides and proteins originating from a variety of biological samples. The annotation of the tandem mass spectra used to identify these species is an arduous task requiring extensive expertise. Our web-based and open-source peptide spectral annotator, IPSA, provides a resource for generating and investigating annotated spectra for peptide identifications to a wide research community. IPSA can generate customizable annotated peptide spectra using a clean and intuitive user interface, allowing researchers to export customizable, publication-ready annotated spectra as vector graphics to aid in figure creation. It can process MS/MS spectra from both anionic and cationic precursors,

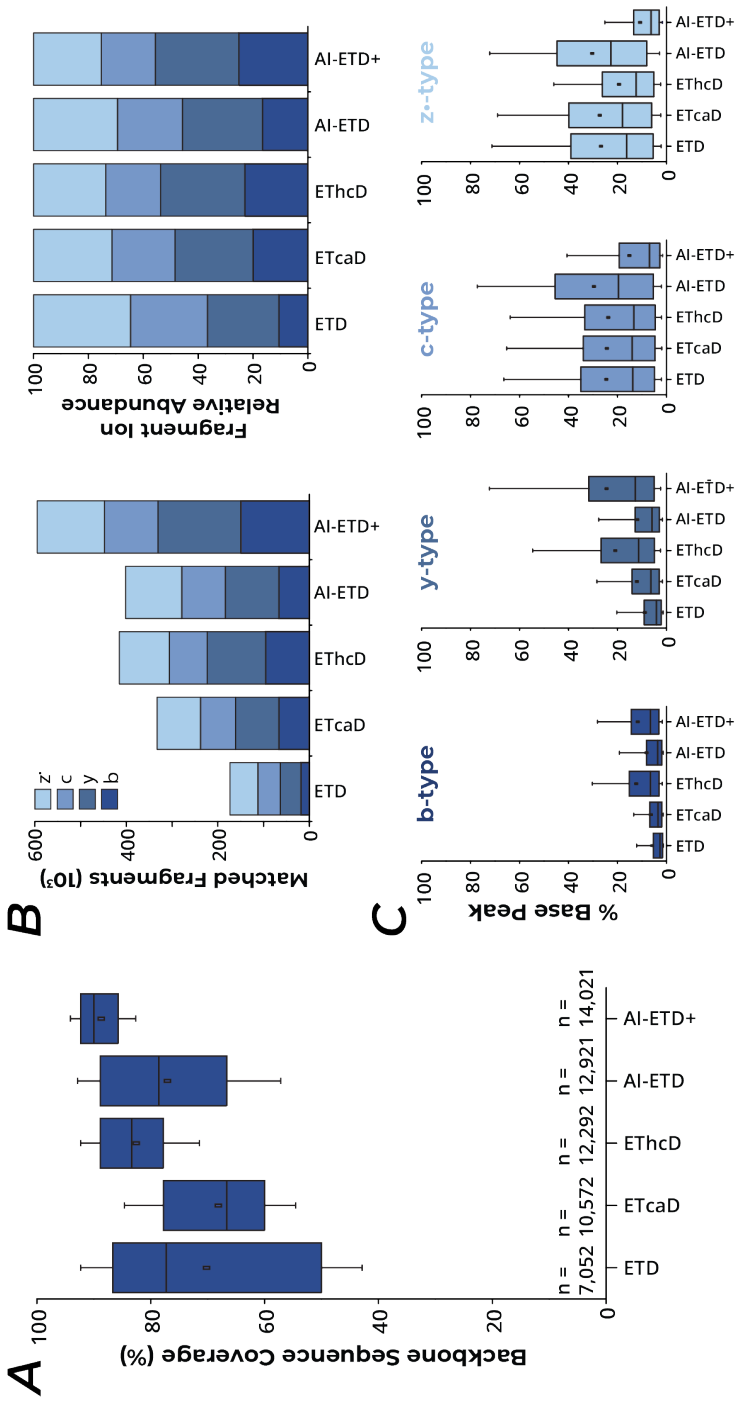


Figure 2.4: Fragment Ion Statistics. Exported fragment statistics facilitate comparisons between different electron transfer dissociation techniques. IPSA matching parameters were set to a ± 10 ppm mass error window and a 1% base peak intensity cutoff (**A**), AI-ETD+ provides superior peptide backbone fragmentation compared with ETD, ETcaD, EThcD, and AI-ETD. (**B**), AI-ETD+ generates higher counts of fragment ions over an experiment compared with other techniques. (**C**), Distributions of all matched fragment ions for each respective dissociative technique.

and it has built-in support to annotate fragment ions generated from a diverse assortment of dissociative techniques. Additionally, IPSA can extract fragment ion statistics from any number of peptide spectra and return results in a tabular format, giving researchers a deeper and more comprehensive view of their peptide analyses.

We chose to develop IPSA as an online platform to reach a wide audience of proteomics researchers: those with an Internet connection on a computer with a web browser. Web-based software also allowed us to use the flexibility of the well-established JavaScript visualization library D3.js while avoiding software compatibility issues and version control. Through IPSA, we aim to increase the approachability of spectral annotation to proteomics novices and experts alike.

The IPSA source code is freely available for inspection and download at `https://github.com/coongroup/IPSA` alongside additional guides regarding software usage. We recommend using an updated web browser to access IPSA at `http://interactivepeptidespectralannotator.com` as outdated browsers may not provide support for critical functions. IPSA can be easily installed on a private desktop or server using a prebuilt Docker[46] image and instructions at `https://hub.docker.com/r/dbrademan/ipsa`, or IPSA's project files can be manually configured to operate on private web servers with full functionality. Additionally, the JavaScript file used to render the interactive visualization, *IPSA.js*, is configured to be used as an AngularJS directive. This directive can be attached to custom annotation scripts in many website environments, allowing the use of our software beyond that of the platform we described here.

References

[1] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, no. 1, pp. 339–347, 2014.

[2] A. L. Richards, A. E. Merrill, and J. J. Coon, "Proteome sequencing goes deep," *Current Opinion in Chemical Biology*, vol. 24, pp. 11–17, 2015.

[3] M. W. Senko, P. M. Remes, J. D. Canterbury, R. Mathur, Q. Song, S. M. Eliuk, C. Mullen, L. Earley, M. Hardman, J. D. Blethrow, H. Bui, A. Specht, O. Lange, E. Denisov, A. Makarov, S. Horning, and V. Zabrouskov, "Novel parallelized quadrupole/linear ion trap/orbitrap tribrid mass spectrometer improving proteome coverage and peptide identification rates," *Analytical Chemistry*, vol. 85, no. 24, pp. 11710–11714, 2013.

[4] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, no. 5, pp. 701–714, 2015.

[5] A. S. Hebert, C. Thöing, N. M. Riley, N. W. Kwiecien, E. Shiskova, R. Huguet, H. L. Cardasis, A. Kuehn, S. Eliuk, V. Zabrouskov, M. S. Westphall, G. C. McAlister, and J. J. Coon, "Improved Precursor Characterization for Data-Dependent Mass Spectrometry," *Analytical Chemistry*, vol. 90, no. 3, pp. 2333–2340, 2018.

[6] R. A. Scheltema, J.-P. Hauschild, O. Lange, D. Hornburg, E. Denisov, E. Damoc, A. Kuehn, A. Makarov, and M. Mann, "The Q Exactive HF, a Benchtop Mass Spectrometer with a Pre-filter, High-performance Quadrupole and an Ultra-high-field Orbitrap Analyzer," *Molecular & Cellular Proteomics*, vol. 13, no. 12, pp. 3698–3708, 2014.

[7] C. D. Kelstrup, D. B. Bekker-Jensen, T. N. Arrey, A. Hogrebe, A. Harder, and J. V. Olsen, "Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics," *Journal of Proteome Research*, vol. 17, no. 1, pp. 727–738, 2018.

[8] D. B. Bekker-Jensen, C. D. Kelstrup, T. S. Batth, S. C. Larsen, C. Haldrup, J. B. Bramsen, K. D. Sørensen, S. Høyer, T. F. Ørntoft, C. L. Andersen, M. L. Nielsen, and J. V. Olsen, "An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes," *Cell Systems*, vol. 4, no. 6, pp. 587–599.e4, 2017.

[9] E. Shishkova, A. S. Hebert, and J. J. Coon, "Now, More Than Ever, Proteomics Needs Better Chromatography," *Cell Systems*, vol. 3, no. 4, pp. 321–324, 2016.

[10] J. K. Eng, A. L. Mccormack, and J. R. Yates, "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *American society for Mass Spectrometry*, vol. 5, no. 11, pp. 976–989, 1994.

[11] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nature Biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008.

[12] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[13] J. A. Taylor and R. S. Johnson, "Sequence database searches via de Novo peptide sequencing by tandem mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 11, no. 9, pp. 1067–1075, 1997.

[14] B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 17, no. 20, pp. 2337–2342, 2003.

[15] H. Chi, R. X. Sun, B. Yang, C. Q. Song, L. H. Wang, C. Liu, Y. Fu, Z. F. Yuan, H. P. Wang, S. M. He, and M. Q. Dong, "PNovo: De novo peptide sequencing and identification using HCD spectra," *Journal of Proteome Research*, vol. 9, no. 5, pp. 2713–2724, 2010.

[16] P. Sinitcyn, J. Daniel Rudolph, J. Cox, J. D. Rudolph, and J. Cox, "Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data," *Annual Review of Biomedical Data Science*, vol. 1, no. 1, pp. annurev–biodatasci–080917–013516, 2018.

[17] R. A. Bradshaw, A. L. Burlingame, S. Carr, and R. Aebersold, "Reporting Protein Identification Data: The next Generation of Guidelines," *Molecular & Cellular Proteomics*, vol. 5, no. 5, pp. 787–788, 2006.

[18] 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. Vizcaíno, M. Chambers, A. Pizarro, and D. Creasy, "The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results," *Molecular & Cellular Proteomics*, vol. 11, no. 7, p. M111.014381, 2012.

[19] S. L. Seymour, T. Farrah, P. A. Binz, R. J. Chalkley, J. S. Cottrell, B. C. Searle, D. L. Tabb, J. A. Vizcaíno, G. Prieto, J. Uszkoreit, M. Eisenacher, S. Martínez-Bartolomé, F. Ghali, and A. R. Jones, "A standardized framing for reporting protein identifications in mzIdentML 1.2," *Proteomics*, vol. 14, no. 21-22, pp. 2389–2399, 2014.

[20] A. Burlingame, S. A. Carr, R. A. Bradshaw, and R. J. Chalkley, "On Credibility, Clarity, and Compliance," *Molecular & Cellular Proteomics*, vol. 14, no. 7, pp. 1731–1733, 2015.

[21] P. R. Baker and R. J. Chalkley, "MS-Viewer: A web-based spectral viewer for proteomics results," *Molecular and Cellular Proteomics*, vol. 13, no. 5, pp. 1392–1396, 2014.

[22] M. Strohalm, M. Hassman, B. Košata, and M. Kodíček, "mMass data miner: An open source alternative for mass spectrometric data analysis," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 6, pp. 905–908, 2008.

[23] J. Colinge, A. Masselot, P. Carbonell, and R. D. Appel, "InSilicoSpectro: An open-source proteomics library," *Journal of Proteome Research*, vol. 5, no. 3, pp. 619–624, 2006.

[24] V. Sharma, J. K. Eng, M. J. MacCoss, and M. Riffle, "A mass spectrometry proteomics data management platform," *Molecular and Cellular Proteomics*, vol. 11, no. 9, pp. 824–831, 2012.

[25] M. Riffle, D. Jaschob, A. Zelter, and T. N. Davis, "ProXL (Protein Cross-Linking Database): A Platform for Analysis, Visualization, and Sharing of Protein Cross-Linking Mass Spectrometry Data.," *Journal of proteome research*, vol. 15, no. 8, pp. 2863–70, 2016.

[26] Y. Perez-Riverol, E. Alpi, R. Wang, H. Hermjakob, and J. A. Vizcaíno, "Making proteomics data accessible and reusable: Current state of proteomics databases and repositories," *PROTEOMICS*, vol. 15, no. 5-6, pp. 930–950, 2015.

[27] M. Riffle, G. E. Merrihew, D. Jaschob, V. Sharma, T. N. Davis, W. S. Noble, and M. J. MacCoss, "Visualization and Dissemination of Multidimensional Proteomics Data Comparing Protein Abundance during Caenorhabditis elegans Development," *Journal of the American Society for Mass Spectrometry*, vol. 26, no. 11, pp. 1827–1836, 2015.

[28] L. Kolbowski, C. Combe, and J. Rappsilber, "xiSPEC: web-based visualization, analysis and sharing of proteomics data," *Nucleic Acids Research*, 2018.

[29] T. Ly and R. R. Julian, "Ultraviolet photodissociation: developments towards applications for mass-spectrometry-based proteomics," *Angewandte Chemie - International Edition*, vol. 48, no. 39, pp. 7130–7137, 2009.

[30] Q. Yu, B. Wang, Z. Chen, G. Urabe, M. S. Glover, X. Shi, L. W. Guo, K. C. Kent, and L. Li, "Electron-Transfer/Higher-Energy Collision Dissociation (EThcD)-Enabled Intact Glycopeptide/Glycoproteome Characterization," *Journal of the American Society for Mass Spectrometry*, vol. 28, no. 9, pp. 1751–1764, 2017.

[31] A. R. Ledvina, N. A. Beauchene, G. C. McAlister, J. E. P. Syka, J. C. Schwartz, J. Griep-Raming, M. S. Westphall, and J. J. Coon, "Activated-ion electron transfer dissociation improves the ability of electron transfer dissociation to identify peptides in a complex mixture," *Analytical Chemistry*, vol. 82, no. 24, pp. 10068–10074, 2010.

[32] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[33] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–964, 2004.

[34] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "COMPASS: A suite of pre- and post-search proteomics software tools for OMSSA," *Proteomics*, vol. 11, no. 6, pp. 1064–1074, 2011.

[35] N. M. Riley, M. J. P. Rush, C. M. Rose, A. L. Richards, N. W. Kwiecien, D. J. Bailey, A. S. Hebert, M. S. Westphall, and J. J. Coon, "The Negative Mode Proteome with Ac-

tivated Ion Negative Electron Transfer Dissociation (AI-NETD)," *Molecular & Cellular Proteomics*, vol. 14, no. 10, pp. 2644–2660, 2015.

[36] N. M. Riley, M. S. Westphall, A. S. Hebert, and J. J. Coon, "Implementation of Activated Ion Electron Transfer Dissociation on a Quadrupole-Orbitrap-Linear Ion Trap Hybrid Mass Spectrometer," *Analytical Chemistry*, vol. 89, no. 12, pp. 6358–6366, 2017.

[37] P. Roepstorff and J. Fohlman, "Letter to the editors," *Biological Mass Spectrometry*, vol. 11, no. 11, pp. 601–601, 1984.

[38] R. S. Johnson, S. A. Martin, K. Biemann, J. T. Stults, and J. T. Watson, "Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine," *Analytical Chemistry*, vol. 59, no. 21, pp. 2621–2625, 1987.

[39] 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. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno, and H. Hermjakob, "The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience," *Molecular and Cellular Proteomics*, vol. 13, no. 10, pp. 2765–2775, 2014.

[40] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, 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*, vol. 10, no. 1, p. R110.000133, 2011.

[41] M. C. Chambers, B. MacLean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick, "A cross-platform toolkit for mass spectrometry and proteomics," *Nature Biotechnology*, vol. 30, no. 10, pp. 918–920, 2012.

[42] G. C. McAlister, J. D. Russell, N. G. Rumachik, A. S. Hebert, J. E. Syka, L. Y. Geer, M. S. Westphall, D. J. Pagliarini, and J. J. Coon, "Analysis of the acidic proteome with negative electron-transfer dissociation mass spectrometry," *Analytical Chemistry*, vol. 84, no. 6, pp. 2875–2882, 2012.

[43] N. G. Rumachik, G. C. McAlister, J. D. Russell, D. J. Bailey, C. D. Wenger, and J. J. Coon, "Characterizing peptide neutral losses induced by negative electron-transfer dissociation (NETD)," *Journal of the American Society for Mass Spectrometry*, vol. 23, no. 4, pp. 718–727, 2012.

[44] J. A. Madsen, H. Xu, M. R. Robinson, A. P. Horton, J. B. Shaw, D. K. Giles, T. S. Kaoud, K. N. Dalby, M. S. Trent, and J. S. Brodbelt, "High-throughput Database Search and Large-scale Negative Polarity Liquid Chromatography–Tandem Mass Spectrometry with Ultraviolet Photodissociation for Complex Proteomic Samples," *Molecular & Cellular Proteomics*, vol. 12, no. 9, pp. 2604–2614, 2013.

[45] N. P. Ewing and C. J. Cassady, "Dissociation of multiply charged negative ions for hirudin (54–65), fibrinopeptide b, and insulin a (oxidized)," *Journal of the American Society for Mass Spectrometry*, vol. 12, no. 1, pp. 105–116, 2001. PMID: 11142354.

[46] D. Merkel, "Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood," *Linux Journal*, vol. 2014, no. 239, pp. 2–7, 2014.

Chapter 3

THE NCQBCS CONTROLLER: SOFTWARE FOR TRACKING SYSTEM SUITABILITY OF LC/MS PLATFORMS

Portions of this chapter are part of a manuscript in preparation

Brademan DR, Shishkova E, Westphall MS, Coon JJ. *Software for Tracking System Suitability of LC/MS Platforms* **2020**

Abstract

Liquid chromatography-mass spectrometry (LC/MS) is the preferred analytical platform for the identification and quantitation of multiple classes of biomolecules contained in complex biological samples. The high-throughput nature of today's LC/MS experiments produces large volumes of data, even from studies containing only a handful of experimental conditions. The collection, analysis, and interpretation of the resulting data sets from experiments are costly in both time and resources. As such, quality assurance regarding collected data is paramount. Without the assurance of appropriate quality controls (QCs), the biological interpretation of experimental results may be insurmountably obfuscated by shifts in instrumental performance. Several QC software packages have been developed specifically for the LC/MS community, but as of yet none have been widely adopted as many laboratories prefer to design their own quality control pipeline. In this chapter we describe an internal software application used to longitudinally benchmark instrumental performance. This resource would be useful for the LC/MS community. This platform consists of a centralized website that supports drag-and-drop QC file upload and is currently capable of displaying identification-based metrics in an interactive environment. We also describe several key features which are under active development which would enable longitudinal tracking of any metric for any analytical platform.

Background and Introduction

Quality Assurance in LC/MS Experiments The ability of modern mass spectrometers to collect significant amounts of complex biological data makes mass spectrometry a desirable analytical technique for clinical researchers, especially regarding the determination of potential disease biomarkers,[1,2] uncovering novel insights to biological pathways,[3,4] or predicting the biomechanical impact of drugs in a particular biological system.[5] To ensure that the conclusions of these studies are founded in biological truth, steps must be taken to guarantee experimental reproducibility.[6] As analytical techniques can vary between laboratories, a "quality by design" approach to experimental development should be adopted to ensure a predetermined standard of data quality.[7] As such, each step in the construction of a biological study should be structured for scrupulous experimental design alongside the incorporation of quality control(s) where suitable.[8] Without the consideration of quality assurance, significant experimental discoveries can be stymied by unaccounted variables originating from sample preparation, analytical techniques, or instrumental drift.[9,10]

Quality controls should be integrated into experimental workflows to account for experimental variability. Different types of quality controls can be prepared and analyzed before, during, and after experimental samples depending on what aspects of variability regarding sample preparation and instrument performance need to be monitored. For large studies spanning many days, pooled quality controls are often prepared daily alongside each sample batch to account for a variety of factors which could induce batch effects.[11]

Normalization of the experimental samples to these pooled standards can aid in the reduction of variation resulting from slight differences in inter-day sample preparation and instrument performance.[12,13] Quality controls should be representative of the samples in an experiment. For example, discovery-based 'omics' studies may benefit from using complex cellular lysates as a quality control as the depth and reproducibility of biomolecule identifications are a good metric for system suitability. In a different case, a mixture of several digested standard proteins may be more appropriate for targeted assays where characterization quality for a handful of compounds is paramount.[14]

The metrics tracked from quality control can be primarily classified in two ways, either as intra-experimental or inter-experimental metrics. Intra-experimental metrics focus on the different properties contained internally to a single LC/MS run, while inter-experimental metrics compare summary results longitudinally.[8] QC metrics can further be classified as either identification-free, identification-based, or instrumental. Identification-free metrics are derived directly from collected spectral data without the computational overhead of spectral searching.[14] For example, survey mass spectra (MS^1) and tandem mass spectrum (MS^2) counts, injection times, and total ion current (TIC) metrics can be immediately extracted from instrument files once data acquisition has completed. Identification-based metrics can provide complimentary insights to the qualitative health of an analytical system but require computational resources to search. Metrics such as number of unique identified biomolecules, spectral similarity score distributions, and retention time distributions fall into the above category. Finally, instrument-level metrics pertain to direct readouts from

instrument components and can be used to quickly pinpoint component failures.

Current Quality Control Software Packages Many different quality control tools have been developed to measure the readiness of an LC/MS system. A seminal tool for proteomics QC was a software package named NIST MSQC, which collected 46 metrics per experiment.[15] Rudnick, *et al.* demonstrated how a comprehensive investigation of these metrics improved upon the determination of system suitability compared to manually tracking a select few. Support for MSQC has been dropped, but several open-source reimplementations of this program remain available.[16,17] QuaMeter, one of MSQC's reimplementations, is operated via a command-line application that extracts 42 identification-based and 45 identification-free metrics from raw experimental data and supports vendor file types.[16,18] QuaMeter solely generates identification-free inter-experimental summary metrics from raw input files. If identification-based metrics are desired, preprocessing of experimental data must be done using IDPicker, an algorithm that takes standard mzIdentML[19] or pepXML[20] files for peptide and protein filtering and identification.[21] Additionally, no inter-experimental metrics are collected through this software, nor is there any functionality for visualization or longitudinal tracking for trend analyses.

Other quality control softwares that permit longitudinal examination of various quality control metrics. SIMPATIQCO and iMonDB both use Structured Query Language (SQL) databases to store and retrieve current and historical quality control metrics.[22,23] SIMPATIQCO operates as a locally hosted application which automatically queries QC files using

a 'hot folder' and visualized via a web interface. Quality is determined statistically from previously uploaded controls to calculated acceptable ranges for the collected metrics. However, experience with web server administration is needed, as direct manipulation of SIMPATIQCO's database is required to configure instruments, making the installation process daunting for the novice user. IMonDB is unique compared to most other quality control programs as it solely tracks instrumental metrics. These instrument metrics, although not always directly helpful for troubleshooting common system issues, excel at indicating the rare case of mass spectrometer component failure. Other QC programs of note are Metriculator[17], OpenMS[24], and AutoQC.[25]

The ideal QC software would be capable of the rapid extraction of many identification-free, identification-based, and instrumental metrics which track and visualize instrument performance in a decentralized environment such as a web server. QCloud is a recently released QC tool which combines aspects of the above software into an automated cloud-computing QC platform.[26] A QCloud script detects collected QC samples, transfers the QC files to the cloud, and analyzes and tracks sets of performance predictive metrics for both targeted and discovery proteomics systems. However, QCloud presents very few customization options to the end user. The software supports only a predefined set of LC/MS instrument methods and samples types, limiting it's applicability to research labs which may employ alternative analytical methods.

The Coon laboratory has traditionally utilized a complex mixture of tryptic peptides to determine the system suitability our of dedicated proteomics platforms. Previously we

ran digested yeast extract (~4,000 expressed genes)[27], but improvements in separation technologies[28] and data acquisition algorithms[29] required increased sample complexity to properly benchmark periods of high-end performance. We have since transitioned to analyzing lysate derived from the readily available human cell line K562 (~10,500 expressed genes).[30] Instrument users intersperse these human controls with their own samples at a minimum frequency of 1 control a week. During periods of high user turnover, we found it challenging to manually aggregate performance metrics between multiple individuals. To streamline the lab-wide dissemination of instrument readiness and maintenance requirements, we developed an in-house web tool named the National Center for Quantitative Biology of Complex Systems (NCQBCS) Controller. This resource can be found at `http://www.ncqbcscontroller.com`. The NCQBCS Controller currently only tracks QC metrics pertaining to our proteomic platforms, but features under active development will soon enable omics-agnostic benchmarking for all our instrumentation. Herein we describe the current state of the NCQBCS Controller, and we discuss future developments which will enable usage of our resource by the scientific community.

Software Structure and Features

NCQBCS Controller Structure Overview The NCQBCS Controller structure can be broken down into three general components. The first is a client-facing front-end composed of HTML and Javascript files. These files are rendered in a user's browser and allow them to interact with the platform in a structured and intuitive manner. Software libraries of note

are the Twitter Bootstrap HTML templating library and the Angular Javascript framework which are used to procedural build the client interface. The second component is a relational database which serves as a secure centralized location to store data pertaining to users, instruments, and QC metrics. The third and final component consists of a set of modular server-side PHP scripts which are called upon by the front-end to process newly uploaded control data or retrieve historical performance metrics. These scripts are separated from the front end in order to provide a layer of abstraction from users. This abstraction improves the security and robustness of the Controller.

Uploading Quality Controls Before using the NCQBCS Controller, a new user must first create an account. On account creation, a user's credentials are immediately encrypted to ensure plain text passwords cannot be retrieved in the case the server is compromised. Once logged in, new QC files can be uploaded using the web page shown in **Figure 3.1**. First, the user must specify an instrument with which the uploaded QC results will be associated. If a particular instrument does not exist, it can easily be created using the 'Manage Instruments' section of the website (not shown). To create a new instrument, the user simply need to supply an instrument name, a text descriptor describing the instrument platform, and informative comments which could describe the LC/MS platform or specific parameters pertaining to the QCs which will be uploaded to this platform. All instrument-associated metadata can be updated at any time. The NCQBCS Controller currently has 3 proteomics instruments defined: two Orbitrap Fusion Lumos and one Orbitrap Eclipse platform(s).

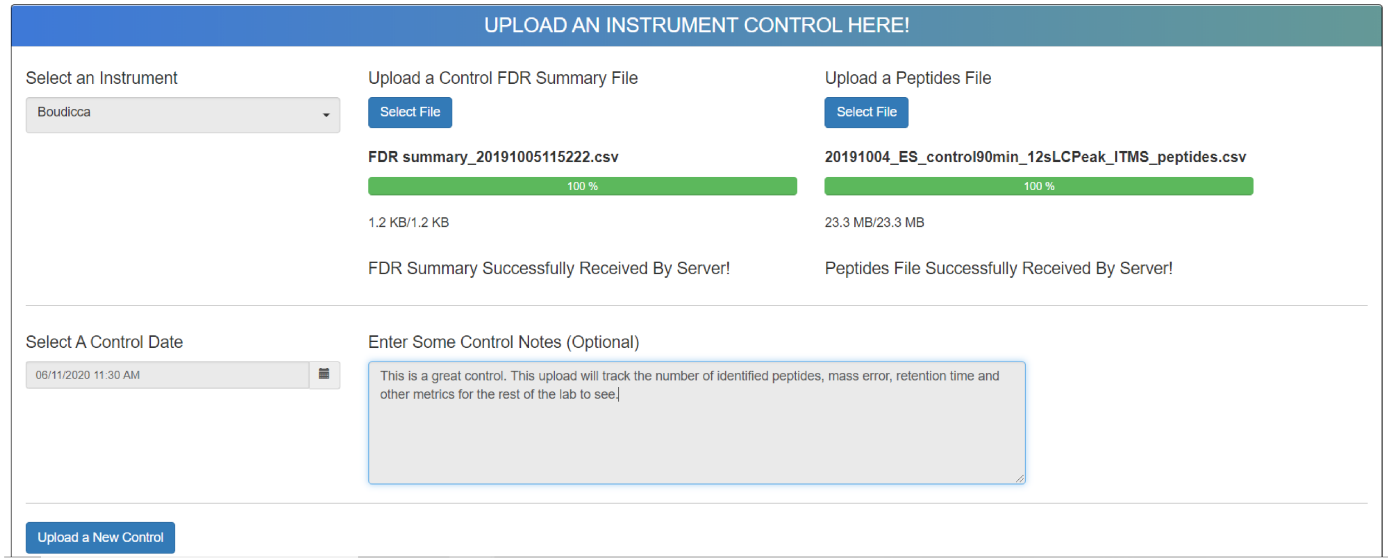


Figure 3.1: NCQBS Controller Data Upload. This control panel is used to upload new QC results to the NCQBCS controller. The user must select an instrument to upload the QC results to. They then can retrieve the searched QC FDRsummary and peptide files from COMPASS and upload them to the website. Once these files are uploaded and a control date is specified, the server will automatically extract relevant metrics for later visualization.

The Controller in its current implementation only accepts comma-separated value output files from the COMPASS software suite.[31] It does not support the processing of raw data. This requires users to have previously searched their QCs. Communication between individuals is needed to ensure QCs are analyzed using the same settings. We have found it useful to specify these parameters in the instrument notes. Once the FDRsummary.csv and peptides.csv files are uploaded, a QC run date along with optional detailed control notes are specified. The server then begins data processing. Occasionally QCs are erroneously uploaded to an incorrect instrument, or the wrong files are uploaded. In either case, QCs can be re-associated with the correct instrument or be deleted using the 'Manage Uploaded Data' section of the website (not shown).

Data Processing The NCQBCS Controller uses a back-end relational database (MySQL) which centrally stores all data contained within the Controller. The schema of this database is shown in **Figure 3.2A**. While a diverse set of database architectures exist (*e.g.* SQLite, MongoDB, Postgres), we have implemented MySQL as it supports multiple database connections, meaning multiple users can concurrently use the website. Additionally, the relational database aspect of MySQL strikes an appropriate balance between implementation difficulty and performance. Unique identifiers contained in each table are used to index common lookup operations. This improves data lookup performance which manifests as a more-responsive website.

Data processing begins with moving the uploaded QC files from the server's /tmp

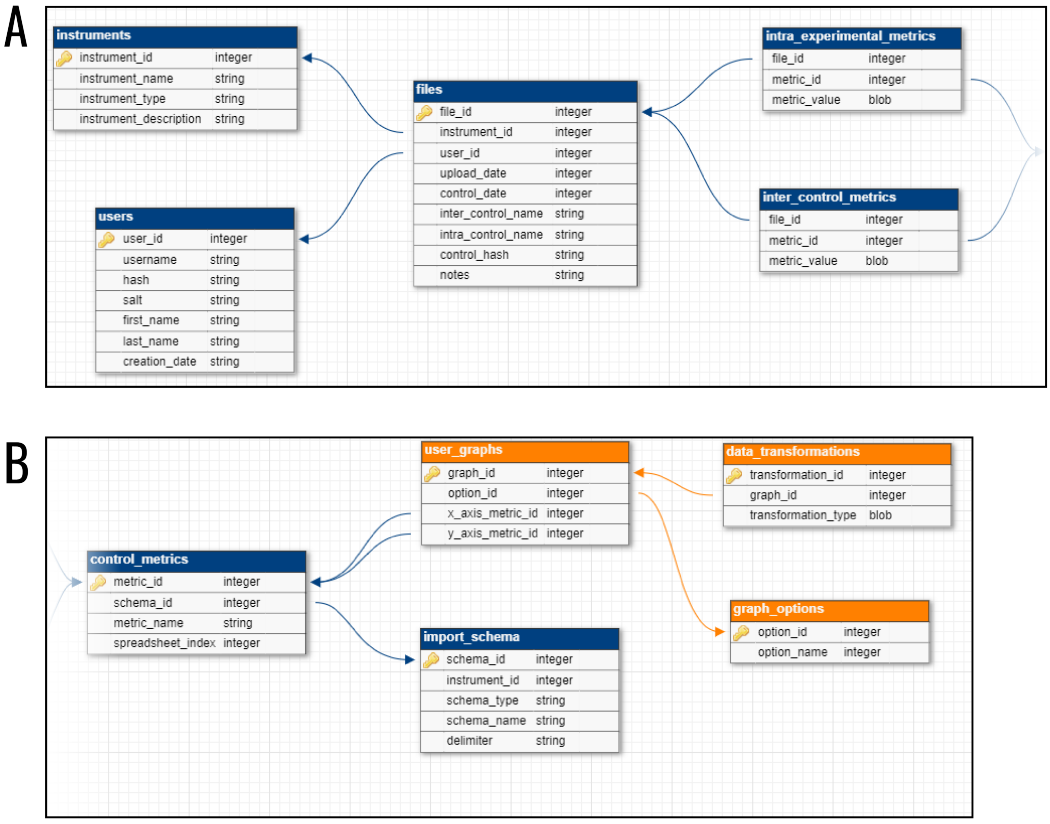


Figure 3.2: NCQBCS Current and Future Database Schema. Panel A shows the database structure of the proteomics-only version NCQBCS controller. **Panel B** contains table additions which will enable platform-agnostic QC tracking. The two metrics tables on the right of **panel A** join to the control_metrics table in **panel B**. Tables with blue headers are attached to implemented features as of this writing. Tables with orange headers will be implemented in the near future. These schema were made using DBDesigner.net

directory into a permanent location to preserve a collection of QC results which can be reprocessed if different metrics need to be extracted at a future date. To prevent duplicate control uploads, the contents of the newly uploaded file pair are cryptographically hashed to determine if these exact files had been uploaded previously. Next, the uploaded files are checked to see if they are in the correct format. If not, data processing is halted, and the user is alerted to recheck their uploaded files. The FDR Summary file is then scraped to extract the name of the QC raw file, the unique number of identified peptides, the number of peptide-spectral matches (PSMs), MS^2 counts, the average and maximum number of MS^2s per survey scan, and the average and maximum MS^2 inject times across the QC. These metrics will be stored in the inter-experimental-metrics table. The Peptides file is then scraped to extract spectrum numbers, peptide sequences with modifications, peptide charge state, peptide theoretical and experimental m/z, quadrupole isolation m/z, calculated mass error in ppm, and retention time. These metrics will be stored in the intra-control-metrics table. Once these values are successfully extracted from the uploaded CSVs, they are rapidly inserted into the MySQL database using MySQL transactions and the PDO extension. If all metrics are successfully entered, the results are immediately made available to view. Otherwise, the user is alerted that the uploaded failed and is given a reason why.

Instrument Performance Visualization Any user can navigate to the 'View Control History' page to explore current and historic instrument performance (**Figure 3.3**. When an instrument is selected using the droplist in the upper left, an asynchronous query is sent

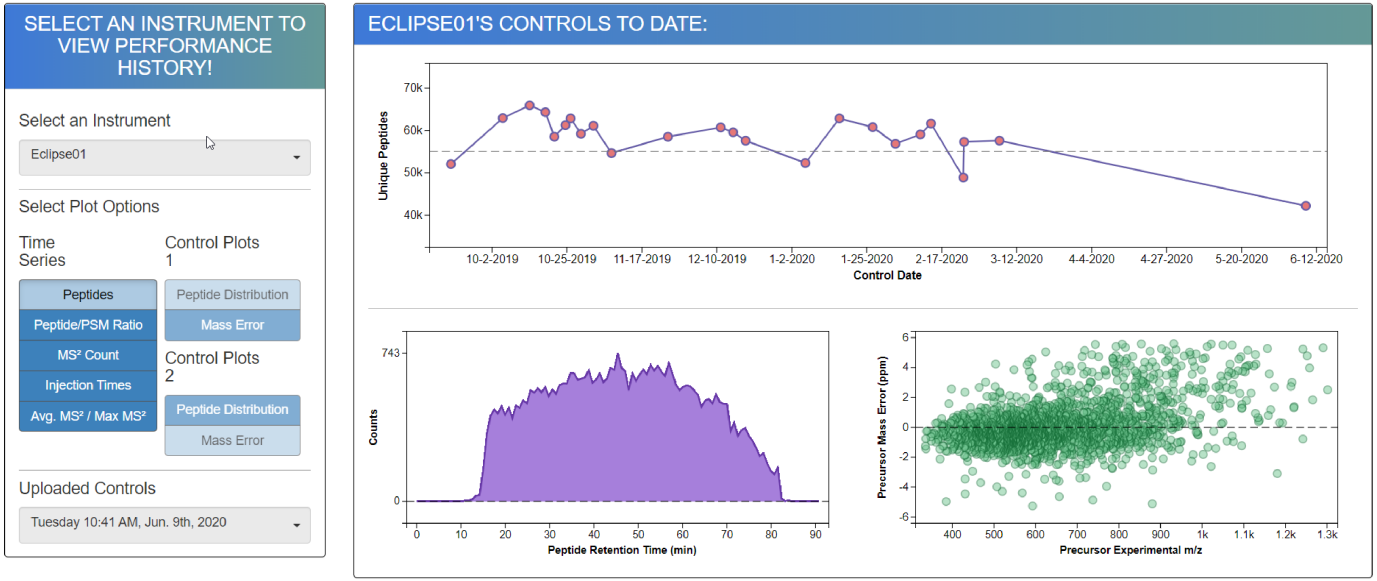


Figure 3.3: QC Data Visualization. This dashboard is the main data visualization page of the NCQBCS Controller. When a user selects an instrument, a longitudinal visualization of unique identified peptides per uploaded QC is rendered in the upper panel. Other inter-control metrics can be inspected using the control panel. The threshold for acceptable performance is visualized by the dashed line. The lower two panels visualize intra-control performance metrics. A peptide retention time distribution is charted in the lower left panel. The lower right panel enables an inspection of systematic peptide mass error trends.

to the server to retrieve the historic peptide counts for the selected instrument along with the peptide-retention time distribution of identified peptides (100 bins) and peptide mass errors of the most recently run control. As we routinely identify over 50,000 peptides per QC, peptides mass errors are subsampled by a factor of 20 to maintain website performance while still conferring mass calibration trends.

We find unique peptide counts to be a robust representative metric for the overall health of an LC/MS system. We mark peptide counts of greater than 55,000 peptides are representative of periods of high-performance. As a LC/MS mass spectrometer has many components, most if not all parts of the instrument must perform well to achieve these numbers. Additionally, not all instrument components have an easily obtainable metric which is indicative of performance. If the peptide counts drop, a user can use this 'smoking gun' to begin investigating other metrics which may be more informative for troubleshooting purposes. Other currently tracked inter-control metrics are shown in **Figure 3.4**. We also note longitudinal tracking of instrument performance can prevent what may be described as the gradual lowering of performance expectations of an instrument. For example, the most recent control collected in **Figure 3.3** was collected after the Orbitrap Eclipse was left idle for an extended period of time due to research restrictions from the Covid-19 pandemic. Without a record of historical performance, it's possible a lower performance threshold would be adopted as the new standard.

To generate these graphs, we use the D3.js visualization library.[32] The D3 library enables the construction of custom interactive visualizations in scalable vector graphic (SVG) format

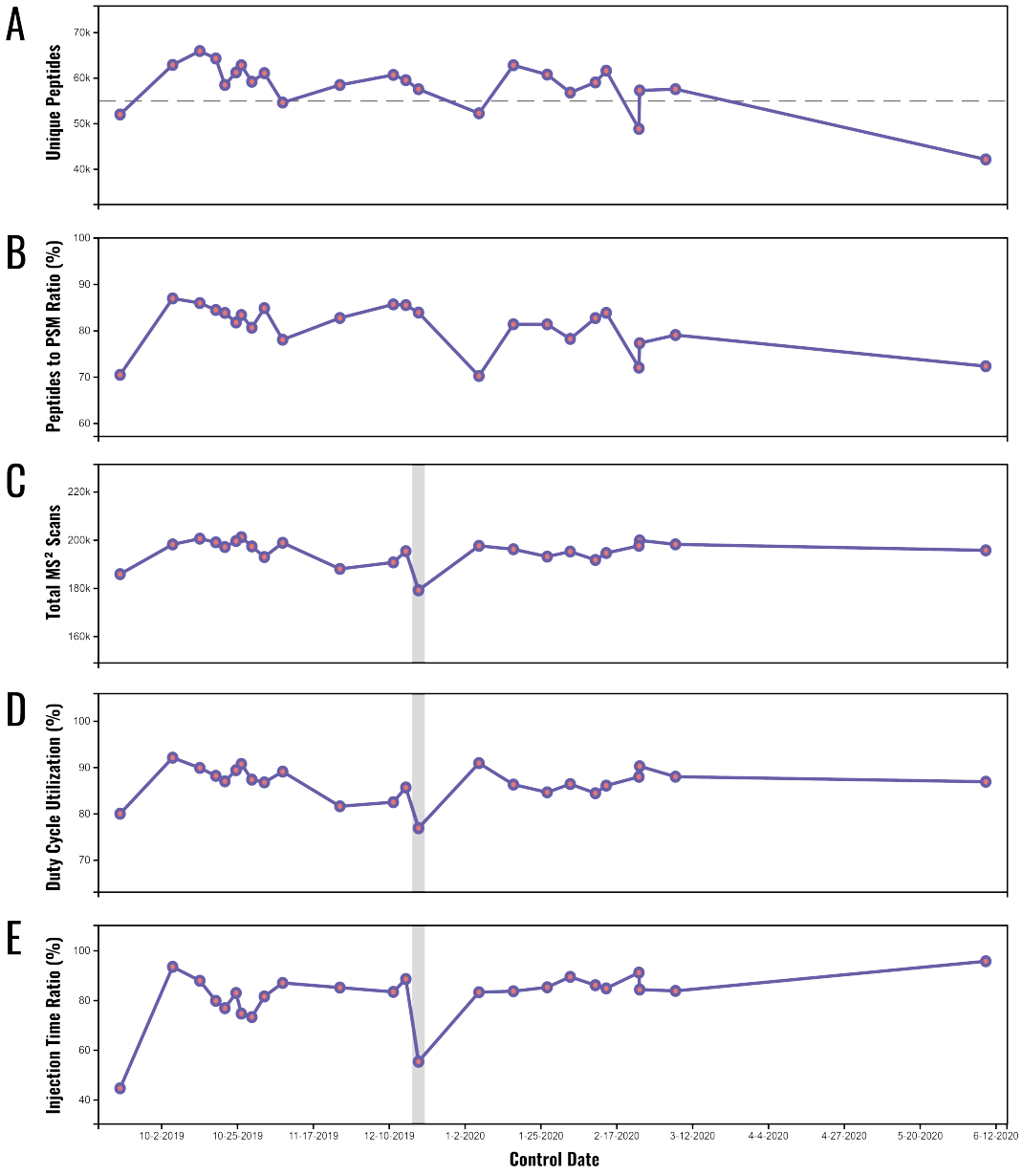


Figure 3.4: Inter-control Metrics Tracked by the NCQBCS Controller. All metrics shown in the figure are derived from QCs collected on an Orbitrap Eclipse. **A**, Unique peptides sequences can be a good indicator of overall performance. **B**, A low peptide to PSM ratio implies the peptides are being redundantly sampled. **C**, Low MS^2 scan counts could indicate problems with spray stability or issues with transferring ions through the instrument. **D**, Duty cycle utilization is a measure of sampling speed. Slower sampling could suggest low ion flux or poor chromatography. **E**, The injection time ratio is the average MS^2 inject time divided by the maximum allowed injection time. The gray region in panels **C**, **D**, and **E** highlights a problematic control referenced in **Figure3.6C**.

at the cost of challenging implementation. However, by combining AngularJS directives with D3, we can abstract shared D3 constructs from each visualization (e.g. a plot axis) to prevent unnecessary code repetition. SVG plots are ideal as they can be extracted from the Controller and be integrated directly into figure-ready visualizations.

Diagnosing Performance Issues As of this writing, the NCQBCS Controller has processed and stored data from 72 unique human controls that contain in total ~3,800,00 identified peptides spread between our three proteomic platforms. The metrics tracked by the NCQBCS Controller have already proven useful in troubleshooting instrumental issues. **Figure 3.5A** is a peptide/retention time distribution of a recently run control on an Orbitrap Fusion Lumos. The bimodal distribution of the beginning and end of the gradient indicate issues with longer, more hydrophobic peptides failing to elute from the column as normal. This could be caused by the incorrect delivery of mobile phase B due to a leaking pump head. These results could be explained in part due to a failing column. **Figure 3.5B** demonstrates a control where there are two distinct populations of peptides in the mass error graphic. Ideally, all peptides are tightly grouped at 0 ppm across m/z space. However, search algorithms often can account for systematic mass error as long as peptides are still tightly grouped. However, the two peptide populations shown here caused a reduction in identifications. Mass calibration alleviated this issue. **Figure 3.5C** shows periods in the peptide-retention time distribution where no peptides were detected. This observation is correlated with a marked drop in MS^2 scans for this control. During the examination of the QC rawfile, the

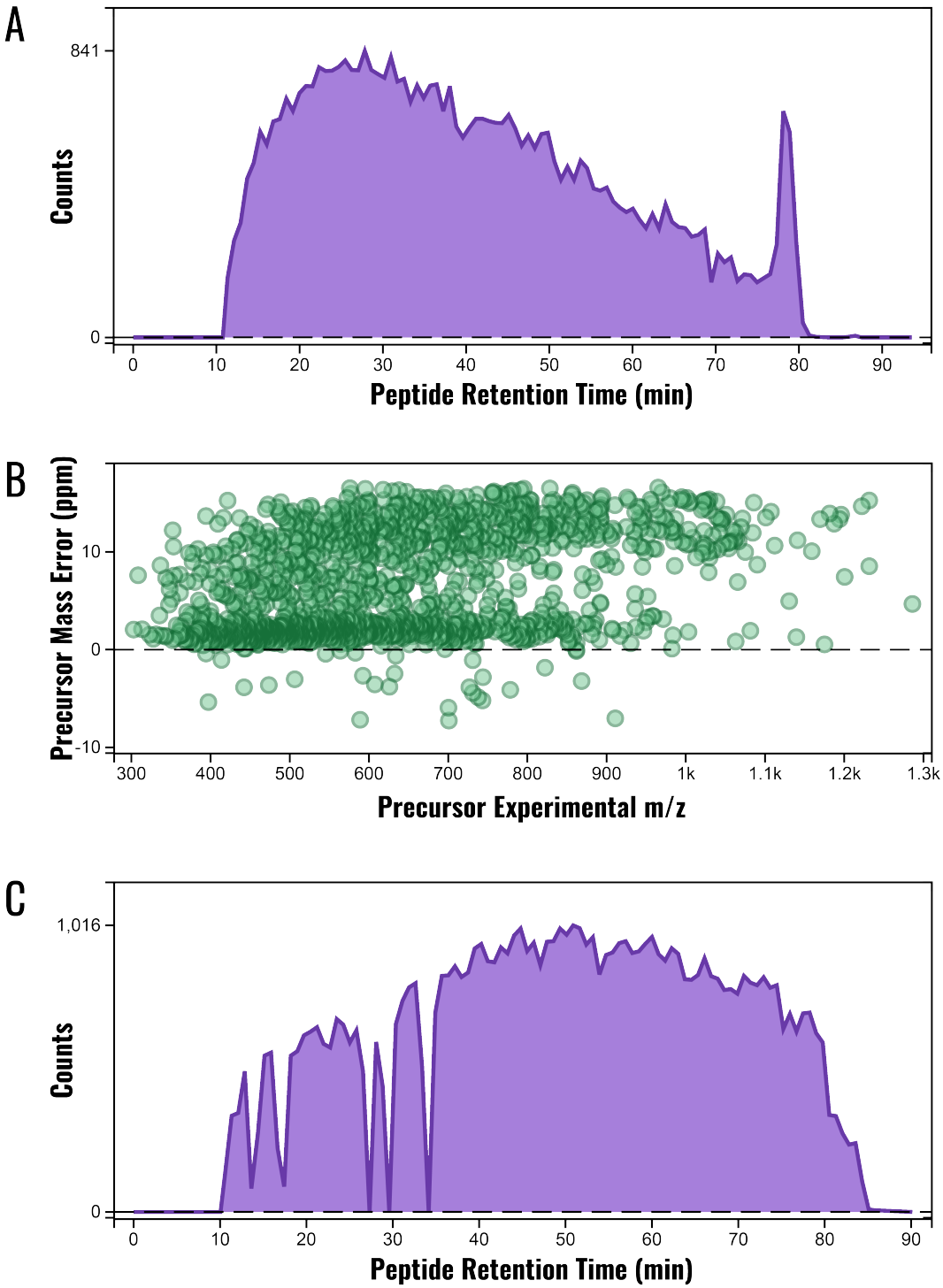


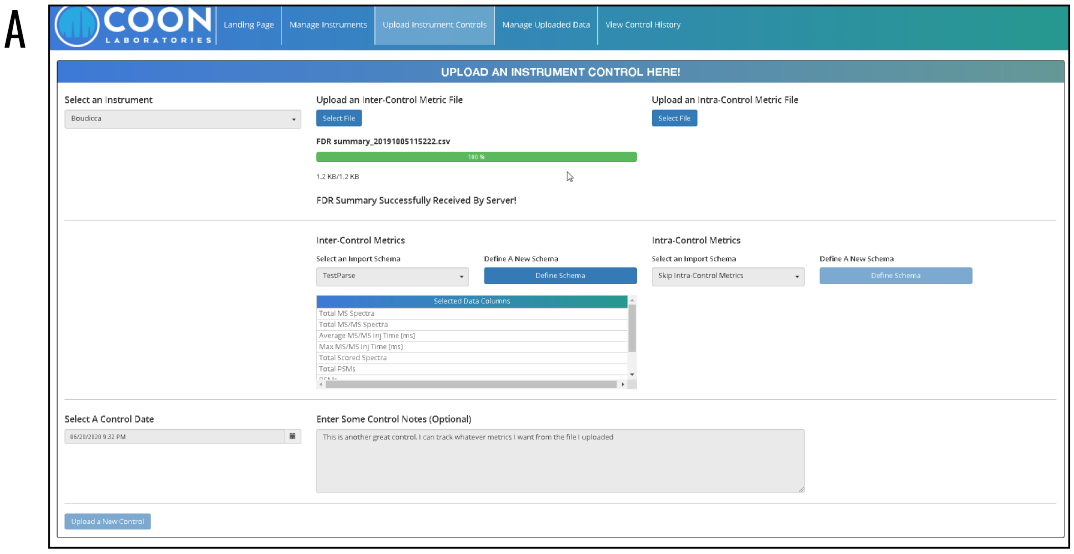
Figure 3.5: Troubleshooting using Intra-control Metrics. **A**, The peptide retention time distribution indicates issues with chromatography. This could be caused either by column degradation or a lack of mobile phase B near the end of the gradient, preventing the longer, more hydrophobic peptides from eluting before the column wash. **B**, Errors with mass calibration cause an irregular, divided population of peptides. **C**, At several points throughout the gradient, no peptides are identified. Charging inside the instrument prevented ions from properly navigating the MS components, causing complete signal drops.

ion signal would drop to zero at random intervals throughout the gradient, indicating a failure of the MS to shuttle ions correctly to the Orbitrap and ion traps for mass analysis and detection. On further inspection, it was determined this behavior was cause by charging. A thorough cleaning of the instrument's inlet capillary, s-lens, and quadrupole ameliorated this issue.

Discussion

While the NCQBCS Controller is a powerful tool for tracking instrument performance, it is custom-built for the Coon lab's proteomics pipeline. If widespread adoption of this tool in the MS community is to take place, we need to: 1) provide the platform in a format that can be quickly set up without requiring large amounts of configuration, and 2) implement the platform in a way that is agnostic to what QC samples are used and what metrics are tracked. In **Chapters 2** and **4**, we give examples of how a complete website can be embedded inside of a Docker container.[33] In brief, Docker is a tool which allows developers to preconfigure software in a virtual environment to guarantee software runs out of the gate regardless of a user's personal computer configuration. Using the **Chapters 2** and **4** implementations as references, we could embed the NCQBCS Controller inside a Docker container in a similar fashion.

To address the second point, we have begun developing a modified version of the NCQBCS Controller which is capable of extracting metrics from generic, tabular files delimited by commas, tabs, or semicolons. **Figure 3.6A** shows a fully functional version the



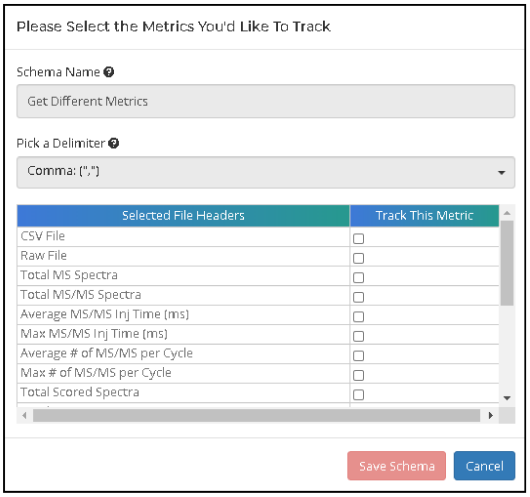


Figure 3.6: Developmental Features. Panel A shows a modified QC upload page. This page allows a user to hand-pick which metrics they would like to track. Import schema can be defined using the controls shown in **Panel B**. Once a new import schema is defined, it can used to parse custom QC metric spreadsheets.

Controller's newly modified file import page. A user is allowed to upload any text file to either of the inter-control (previously the FDR summary file) or intra-control (previously the peptides file) sections. If a file is uploaded, the user can define what metrics they want to track using the 'Import Schema Designer' popup (**Figure 3.6B**). They can save this schema and apply it to new QC files in the future. These metrics are then saved to the newly added tables shown in **Figure 3.3B**. When applying a previously defined import schema to an uploaded file, the first line of the file is parsed and presented to the user to confirm correct parsing behavior. To fully enable platform agnostic QC tracking, we plan on building a set of common QC visualization templates using D3 (*e.g.* density distributions, scatter plots, histograms, temporal metric tracking, ect.) and allow users to assign their tracked QC metrics to chart properties so they may design their own QC figures-of-merit. We have designed a similar resource for custom dataset exploration that is described in **Chapter 4**. We will also need to implement common data transformations for visualization purposes, such as log-transformation and regular expression parsing.

Some QC platforms support the automated processing of raw QCs, meaning a user simply has to supply the MS vendor file. The software will then automatically search the raw data and extract the necessary metrics without requiring user input. The NCQBCS Controller is the spiritual successor to a previously-used internal QC tool named Yeast Controller. While the NCQBCS Controller does not support automated processing, the Yeast Controller did. We have opted to not support processing of raw data for several reasons. Integrating automated processing usually involves constraining what data inputs are

supported, and the integration of additional workflows often comes with increasing levels of technical debt. We believe standardizing accepted inputs to tabular files as described in **Figures 3.3B** and **3.6** will increase applicability to most research groups. Additionally, the storage requirements for raw mass spectrometer files is rapidly increasing. A standard control from 3 years ago was ~1 GB in size. On newer instruments long runs, result in files as large as 3 GB. Accepting only post-processed metrics reduces the storage footprint of the website at the cost of the ability to reprocess the raw data internally. If automated processing is desired, an automated processing pipeline could be developed externally to the Controller. Upon completion, the automated pipeline could insert QC metrics into the MySQL database using a remote connection.

Supplemental Methods

Quality Control Preparation and Mass Analysis A pellet of K562 cells is suspended in 6 M guanidine to inhibit native protease activity, and the cells are lysed using probe sonication. The protein is precipitated from the resulting lysate by the addition of methanol to 90% v:v. The sample is centrifuged, the supernatant is removed, and the remaining protein pellet is suspended in 8 M urea. Initial protein concentration is determined using a Pierce colorimetric protein assay. Next, Tris (100 mM, pH 8), TCEP (10 mM), and CAA (40 mM) are added to the sample at and digested with LysC (1:100 enzyme:protein). The digested sample is then diluted to 1.5 M urea with 100 mM Tris (pH 8.0) and a second overnight digestion is performed using trypsin (1:50 enzyme:protein). The next day, trypsin activity is

halted through acidification using TFA. Desalting of the sample is performed using Sep-Pak SPE cartridge. Peptide concentration is determined using a Pierce colorimetric peptide assay. The desalted peptides are resuspended in 0.2% FA at a concentration of 2 μg/μL and are frozen for later use.

LC separations are performed using a Thermo Dionex Ultimate 3000 RSLC-nanoliquid chromatography instrument. An in-house fabricated column heater is set to 50 °C to reduce column pressure.[34] LC–MS/MS analysis is performed with 2 μg of peptides injected onto a reverse phase nano-UHP column. Separations are performed on a 30 cm, 75–360 μM (inner-outer) diameter, PicoFrit nanospray column (New Objective, Woburn, MA) that is packed with 130 Å pore size, 1.7 μm particle size BEH C18 (Waters, Milford, MA) as previously described.[28] The QC is loaded using a carrier fluid composed of 100% mobile phase A (0.2% FA). At 4 minutes, mobile phase B ((0.2% FA, 70% ACN) is increased to 10% to begin peptide elution. Mobile phase B is gradually increased to 55% over the next 70 minutes. Peptides undergo mass analysis as they elute during this main window. Finally, a 6 minute wash at 100% B and 10 minutes of re-equilibration time at 100% A is conducted for a total of 90 min LC–MS/MS analysis.

Instrument-specific settings vary depending on the specific Thermo Tribrid platform (Fusion Lumos or Eclipse), but generally, both instrumental methods utilize 240k resolving power MS^1 scans with an AGC of 10^6, followed by sampling the most intense peptide precursors for up to 1 second with a 20 second dynamic exclusion window. MS^2 analyses are performed using a 0.7 m/z isolation width using the quadrupole. The AGC target is

set to 3x10^4 with a maximum inject time of 18 ms. Fragmentation is conducted using a normalized HCD collision energy of 30%, and the resulting fragment ions are detected in the low-pressure ion trap using rapid or turbo scans.

Quality Control Data Analysis The COMPASS software suite is used in the Coon lab to analyze all QC files.[31] Briefly, spectral data is extracted from Thermo .raw files using DTA Converter. DTA Convert extracts all MS^2 spectra from the .raw file into a plain-text .dta format. These .dta files are then loaded into DOMSSA, an in-house C# reimplementation of the COMPASS OMSSA[35] search. DOMSSA leverages parallel computing on HTCondor[36,37] to search all QC spectra against a target-decoy protein database[38,39] derived from the UniProt canonical human proteome. The enzyme used to perform *in-silico* digestion is set to trypsin with up to three allowed missed cleavages. Cysteine carbamidomethylation is set as a static modification and methionine oxidation is set as a variable modification. Spectra were searched with a 25 ppm tolerance around the theoretical peptide monoisotopic m/z and a 0.3 Da tolerance on theoretical fragment ion m/z. After target and decoy PSMs are returned and concatenated, they are fed into FDR Optimizer, which filters results to a 1% peptide FDR (sorted on E-vaue) and a maximum allowed precursor mass error of 25 ppm.[31] The resulting FDRsummary and peptides.csv files can be uploaded to the NCQBCS controller.

References

[1] K. A. Overmyer, T. W. Rhoads, A. E. Merrill, Z. Ye, M. S. Westphall, A. Acharya, S. K. Shukla, and J. J. Coon, "Proteomics, lipidomics, metabolomics and 16S DNA sequencing of dental plaque from patients with diabetes and periodontal disease," *bioRxiv*, p. 2020.02.25.963967, 2020.

[2] B. Shen, X. Yi, Y. Sun, X. Bi, J. Du, C. Zhang, S. Quan, F. Zhang, R. Sun, L. Qian, W. Ge, W. Liu, S. Liang, H. Chen, Y. Zhang, J. Li, J. Xu, Z. He, B. Chen, J. Wang, H. Yan, Y. Zheng, D. Wang, J. Zhu, Z. Kong, Z. Kang, X. Liang, X. Ding, G. Ruan, N. Xiang, X. Cai, H. Gao, L. Li, S. Li, Q. Xiao, T. Lu, Y. Zhu, H. Liu, H. Chen, and T. Guo, "Proteomic and Metabolomic Characterization of COVID-19 Patient Sera," *Cell*, vol. 0, no. 0, 2020.

[3] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[4] M. T. Veling, A. G. Reidenbach, E. C. Freiberger, N. W. Kwiecien, P. D. Hutchins, M. J. Drahnak, A. Jochem, A. Ulbrich, M. J. Rush, J. D. Russell, J. J. Coon, and D. J.

Pagliarini, "Multi-omic Mitoprotease Profiling Defines a Role for Oct1p in Coenzyme Q Production," *Molecular Cell*, vol. 68, no. 5, pp. 970–977.e11, 2017.

[5] J. G. Meyer, S. Liu, I. J. Miller, J. J. Coon, and A. Gitter, "Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests," *Journal of Chemical Information and Modeling*, 2019.

[6] J. Hu, K. R. Coombes, J. S. Morris, and K. A. Baggerly, "The importance of experimental design in proteomic mass spectrometry experiments: Some cautionary tales," *Briefings in Functional Genomics and Proteomics*, vol. 3, no. 4, pp. 322–331, 2005.

[7] D. L. Tabb, "Quality assessment for clinical proteomics," *Clinical Biochemistry*, vol. 46, no. 6, pp. 411–420, 2013.

[8] W. Bittremieux, D. Valkenborg, L. Martens, and K. Laukens, "Computational quality control tools for mass spectrometry proteomics," *Proteomics*, vol. 17, no. 3-4, 2017.

[9] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, 2002.

[10] K. A. Baggerly, J. S. Morris, S. R. Edmonson, and K. R. Coombes, "Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer," *Journal of the National Cancer Institute*, vol. 97, no. 4, pp. 307–309, 2005.

[11] B. J. A. Mertens, *Transformation, Normalization, and Batch Effect in the Analysis of Mass Spectrometry Data for Omics Studies*, pp. 1–21. Cham: Springer International Publishing, 2017.

[12] D. A. Cairns, D. N. Perkins, A. J. Stanley, D. Thompson, J. H. Barrett, P. J. Selby, and R. E. Banks, "Integrated multi-level quality control for proteomic profiling studies using mass spectrometry," *BMC Bioinformatics*, vol. 9, no. 1, p. 519, 2008.

[13] D. Broadhurst, R. Goodacre, S. N. Reinke, J. Kuligowski, I. D. Wilson, M. R. Lewis, and W. B. Dunn, "Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies," *Metabolomics*, vol. 14, no. 6, pp. 1–17, 2018.

[14] W. Bittremieux, P. Meysman, L. Martens, D. Valkenborg, and K. Laukens, "Unsupervised Quality Assessment of Mass Spectrometry Proteomics Experiments by Multivariate Quality Control Metrics," *Journal of Proteome Research*, vol. 15, no. 4, pp. 1300–1307, 2016.

[15] P. A. Rudnick, K. R. Clauser, L. E. Kilpatrick, D. V. Tchekhovskoi, P. Neta, N. Blonder, D. D. Billheimer, R. K. Blackman, D. M. Bunk, H. L. Cardasis, A. J. L. Ham, J. D. Jaffe, C. R. Kinsinger, M. Mesri, T. A. Neubert, B. Schilling, D. L. Tabb, T. J. Tegeler, L. Vega-Montoto, A. M. Variyath, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. A. Carr, S. J. Fisher, B. W. Gibson, A. G. Paulovich, F. E. Regnier, H. Rodriguez, C. Spiegelman, P. Tempst, D. C. Liebler, and S. E. Stein, "Performance metrics for

liquid chromatography-tandem mass spectrometry systems in proteomics analyses," *Molecular and Cellular Proteomics*, vol. 9, no. 2, pp. 225–241, 2010.

[16] Z. Q. Ma, K. O. Polzin, S. Dasari, M. C. Chambers, B. Schilling, B. W. Gibson, B. Q. Tran, L. Vega-Montoto, D. C. Liebler, and D. L. Tabb, "QuaMeter: Multivendor performance metrics for LC-MS/MS proteomics instrumentation," *Analytical Chemistry*, vol. 84, no. 14, pp. 5845–5850, 2012.

[17] R. M. Taylor, J. Dance, R. J. Taylor, and J. T. Prince, "Metriculator: Quality assessment for mass spectrometry-based proteomics," *Bioinformatics*, vol. 29, no. 22, pp. 2948–2949, 2013.

[18] X. Wang, M. C. Chambers, L. J. Vega-Montoto, D. M. Bunk, S. E. Stein, and D. L. Tabb, "QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics," *Analytical Chemistry*, vol. 86, no. 5, pp. 2497–2509, 2014.

[19] 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. Vizcaíno, M. Chambers, A. Pizarro, and D. Creasy, "The mzIdentML data standard for mass spectrometry-based proteomics results," *Molecular and Cellular Proteomics*, vol. 11, no. 7, 2012.

[20] A. Keller, J. Eng, N. Zhang, X. Li, and R. Aebersold, "A uniform proteomics MS/MS

analysis platform utilizing open XML file formats," *Molecular Systems Biology*, vol. 1, no. 1, 2005.

[21] Z. Q. Ma, S. Dasari, M. C. Chambers, M. D. Litton, S. M. Sobecki, L. J. Zimmerman, P. J. Halvey, B. Schilling, P. M. Drake, B. W. Gibson, and D. L. Tabb, "IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering," *Journal of Proteome Research*, vol. 8, no. 8, pp. 3872–3881, 2009.

[22] W. Bittremieux, H. Willems, P. Kelchtermans, L. Martens, K. Laukens, and D. Valkenborg, "IMonDB: Mass spectrometry quality control through instrument monitoring," *Journal of Proteome Research*, vol. 14, no. 5, pp. 2360–2366, 2015.

[23] P. Pichler, M. Mazanek, F. Dusberger, L. Weilnböck, C. G. Huber, C. Stingl, T. M. Luider, W. L. Straube, T. Köcher, and K. Mechtler, "SIMPATIQCO: A server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on orbitrap instruments," *Journal of Proteome Research*, vol. 11, no. 11, pp. 5540–5547, 2012.

[24] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H. C. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher, "OpenMS: A flexible open-source software platform for mass spectrometry data analysis," *Nature Methods*, vol. 13, no. 9, pp. 741–748, 2016.

[25] M. S. Bereman, J. Beri, V. Sharma, C. Nathe, J. Eckels, B. MacLean, and M. J. MacCoss, "An Automated Pipeline to Monitor System Performance in Liquid Chromatography-Tandem Mass Spectrometry Proteomic Experiments," *Journal of Proteome Research*, vol. 15, no. 12, pp. 4763–4769, 2016.

[26] C. Chiva, R. Olivella, E. Borràs, G. Espadas, O. Pastor, A. Solé, and E. Sabidó, "QCloud: A cloud-based quality control system for mass spectrometry-based proteomics laboratories," *PLoS ONE*, vol. 13, jan 2018.

[27] B. Futcher, G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels, "A Sampling of the Yeast Proteome," *Molecular and Cellular Biology*, vol. 19, no. 11, pp. 7357–7368, 1999.

[28] E. Shishkova, A. S. Hebert, M. S. Westphall, and J. J. Coon, "Ultra-High Pressure (>30,000 psi) Packing of Capillary Columns Enhancing Depth of Shotgun Proteomic Analyses," *Analytical Chemistry*, vol. 90, no. 19, pp. 11503–11508, 2018.

[29] A. S. Hebert, C. Thöing, N. M. Riley, N. W. Kwiecien, E. Shiskova, R. Huguet, H. L. Cardasis, A. Kuehn, S. Eliuk, V. Zabrouskov, M. S. Westphall, G. C. McAlister, and J. J. Coon, "Improved Precursor Characterization for Data-Dependent Mass Spectrometry," *Analytical Chemistry*, vol. 90, no. 3, pp. 2333–2340, 2018.

[30] E. A. Ponomarenko, E. V. Poverennaya, E. V. Ilgisonis, M. A. Pyatnitskiy, A. T. Kopylov,

V. G. Zgoda, A. V. Lisitsa, and A. I. Archakov, "The Size of the Human Proteome: The Width and Depth," *International Journal of Analytical Chemistry*, vol. 2016, 2016.

[31] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "COMPASS: A suite of pre- and post-search proteomics software tools for OMSSA," *Proteomics*, vol. 11, no. 6, pp. 1064–1074, 2011.

[32] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[33] D. Merkel, "Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood," *Linux Journal*, vol. 2014, no. 239, pp. 2–7, 2014.

[34] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Molecular and Cellular Proteomics*, vol. 13, no. 1, pp. 339–347, 2014.

[35] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of Proteome Research*, vol. 3, no. 5, pp. 958–964, 2004.

[36] Livny, M., Basney, J., Raman, R., Tannenbaum, T., "Mechanisms for High Throughput Computing," tech. rep., Madison, WI, 1997.

[37] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: The Condor experience," *Concurrency Computation Practice and Experience*, vol. 17, no. 2-4, pp. 323–356, 2005.

[38] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature Methods*, vol. 4, no. 3, pp. 207–214, 2007.

[39] A. I. Nesvizhskii and R. Aebersold, "Interpretation of shotgun proteomic data: The protein inference problem," *Molecular and Cellular Proteomics*, vol. 4, no. 10, pp. 1419–1440, 2005.

Chapter 4

ARGONAUT: A WEB PLATFORM FOR COLLABORATIVE MULTI-OMIC DATA VISUALIZATION AND EXPLORATION

Portions of this chapter are part of a submitted manuscript:

Brademan DR, Miller IJ, Kwiecien NW, Pagliarini DJ, Westphall MS, Coon JJ, Shishkova E. *Argonaut: a Web Platform for Collaborative Multi-Omic Data Visualization and Exploration.* **2020**

Abstract

The astounding pace at which researchers can now generate large multi-omic datasets using increasingly mature mass spectrometry (MS) techniques presents a new challenge: "Big Data" dissemination, visualization, and exploration. In facilitating this process, web-based data portals accommodate the complexity of multi-omic experiments and the many experts involved. However, developing these tailored companion resources requires programming expertise and extensive knowledge of web server architecture – a substantial burden for many in the multi-omics community. Here we describe Argonaut: a simple, code-free, and user-friendly program for creating customizable, interactive data-hosting websites. The platform carries out real-time statistical analyses of the uploaded data, which it organizes into easily explorable and sharable projects. Collaborating researchers across the globe can view the results, visualized through a variety of common plots, and modify them as desired to streamline data interpretation. Increasing the pace, ease, and quality of access to multi-omic data in these ways, Argonaut ultimately aims to propel discovery of new biological insights. We showcase the capabilities of this tool using a published multi-omics dataset on yeast mitochondrial protease deletion collection, representing several hundred liquid-chromatography MS experiments. Individual Argonaut websites can be set up by using the Docker image freely provided at `https://hub.docker.com/r/coonlabs/Argonaut`, or by using the multi-portal management system provided at `https://github.com/coongroup/Argonaut`.

Introduction

Multi-omics is a powerful and versatile approach for probing biological systems. Encompassing many layers of biological information, multi-omic data can holistically describe a living system and its response to perturbations, as metabolites, lipids, and proteins co-function to orchestrate responses to various stimuli[1,2]. Recent advances in mass spectrometry (MS) profiling technologies have revealed this coordination by enabling simultaneous measurement of multiple molecular classes[3–9]. Specifically, improvements in experimental throughput of multi-omic analyses have opened the door to large-scale MS-based profiling studies, where the analysis of diverse biomolecules in many samples under dozens of different conditions is considered nearly routine[10–17].

The rapid creation of these large and complex datasets, however, presents a new challenge: quickly processing raw MS data into sets of quantified biomolecules and extracting rigorous biological insights from these results. To this end, tools for processing mass spectral data – primarily proteomic data, such as Perseus and MSstats[18,19] – have enabled a number of analyses and visualizations. Nonetheless, major challenges persist: (i) designed for use by MS experts, these tools require both a thorough understanding of statistics and knowledge of common nuances in MS data; (ii) because data processing is not fully streamlined, considerable hands-on and potentially taxing interaction with large datasets is required, and (iii) these tools' tabular outputs are not conducive for dissemination to and exploration by a non-expert user base. Making results accessible to a broader scientific community is

essential to realizing the full potential of biological mass spectrometry, particularly as MS technologies become increasingly application-driven and therefore collaborative[20,21].

Online data analysis and visualization tools have become increasingly popular in other areas of science as they stand to alleviate many of the issues associated with analysis and communication of large datasets[22–24]. These online tools also avoid issues commonly associated with software distribution, eliminating the need for version control by centralizing the software to a standardized web server environment. Functional web-based utilities thus provide an efficient means to share results with collaborators, minimize the challenges of data transfer between laboratories, and improve scientific discussion. In fact, to augment dissemination of study findings, many large-scale resource projects feature tailored companion websites that facilitate interactive data exploration[14,25–27].

Though ideal, such custom web-based interfaces are tedious and time-consuming to develop – even for a single research project. Construction of these tools requires programming experience and familiarity with web server architecture. Recently, Torre *et al*. presented BioJupies, a web-based utility that greatly augmented the analysis and distribution of transcriptomic data[28,29]. Other research groups have released web applications that facilitate online exploration and sharing of MS datasets[30–33]. The next generation of tools should be available to non-programmers, able to convert general multi-omics MS data into a cloud-friendly format, comprehensively interfaced with interactive visualizations, and sharable with collaborators for intuitive hands-on exploration.

To fulfill this need, we present a new platform called Argonaut. Our tool enables rapid

and codeless generation of MS data exploration portals, allowing users to create project-specific websites hosted on standalone web servers using the Docker environment[34]. We describe this process in detail below, demonstrating its use with a large multi-omic data set generated by a study on yeast mitochondrial protease deletion. Briefly, users upload their quantitative data (formatted in simple generic spreadsheets) directly to a browser. Argonaut then performs on-the-fly statistical analyses of that data and allows users to select several interactive visualizations, which are automatically embedded into the custom website. Once created, the data portals can be securely shared with researchers worldwide in just a few clicks.

Results

Creation of the Multi-omic Data Portal Argonaut Overview. The portal creation process can be completed through a series of intuitive steps (Figure 4.1). First, a new data portal is initialized using Docker. Then, the project owner can log into the data portal through their preferred web browser and begin to customize the newly created portal by providing a project title and description and uploading hierarchically organized quantitative data in accordance with their experimental design. The upload procedure allows for experimental or technical replicates to be easily grouped into separate branches under experimental treatment and an ome classifier. Argonaut utilizes an HTML upload page that accepts files containing quantitative data in a post-processed form, e.g., tabular sets of biomolecule abundances. While many pipelines use a variety of standardized file formats to store data[35,36],

Argonaut supports solely tabular, text-based spreadsheets; thanks to their simplicity and flexibility, many search algorithms and processing pipelines are capable of exporting results in this format. Following the initial data upload, users can select individual visualizations to add to their custom web portal from a menu of options, such as volcano plots and correlations, among many others. Based on these selections, Argonaut constructs a complete data exploration webpage with all associated functionality embedded. These custom web portals can then be shared with other researchers – at the discretion of the creator – via a tiered-permission sharing scheme.

Use with testbed dataset. To demonstrate how our tool creates interactive data portals for improved analysis and exploration, we acquired data from a multi-omics study investigating the biological functions of mitochondrial proteases and their substrates in *Saccharomyces cerevisaie*[37]. In this study, 19 single-gene deletion yeast strains and a control wild-type strain were analyzed in biological triplicate under two growth conditions for a total of 120 unique samples per ome. We reasoned that this rich dataset with validated biological insights could serve as a suitable and sufficiently challenging testbed for our tool. The publicly hosted data portal containing this dataset can be accessed at `https://coonlabdatadev.com` with the username "guest" and password "password".

The Veling study identified and quantified >3,000 biomolecules. Highlighting the compatibility of this dataset with our platform, we converted the abundance measurements of these biomolecules into three comma-separated value tables, each corresponding to one of the omes profiled in the study, i.e. the proteome, metabolome, and lipidome.

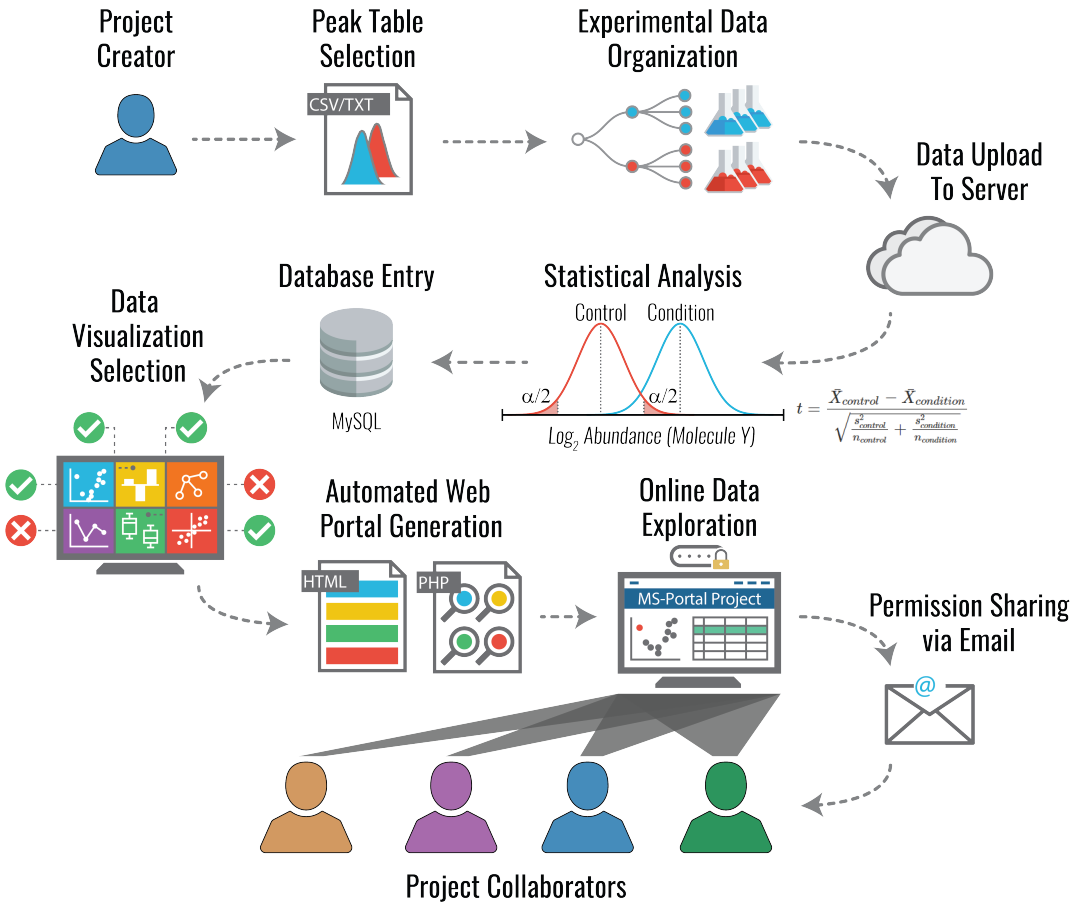


Figure 4.1: The Argonaut Workflow. Argonaut is designed as a portable platform to share multi-omics data in an online environment using customizable interactive visualizations. Processed quantitative measurements from case/control-style experiments are uploaded to the online platform in a variety of text-based formats. Uploaded data is then categorized according to the uploader's experimental design. Common data transformations, such as missing value imputation, filtering missing values, control normalization, or $\log_2$ transformations, can be conducted. Inferential statistics are used to determine the significance of molecular perturbations. Data portals can be customized in a variety of ways, allowing detailed project and data descriptions, selection of visualization options, and project management. Data portal access can also be shared directly with collaborators using a secure permission sharing scheme, allowing multiple laboratories to concurrently explore large datasets to rapidly generate biological insight.

The columns of the files included non-redundant biomolecule identifiers (e.g., UniProt, KEGG, HMDB), unique names of experimental conditions, including names of the strains, condition and replicate information, and optional sample metadata (alternative biomolecule names, FASTA headers, etc.), as illustrated in **Figure 4.2A**. A category of information in the columns must be specified during the data upload. The rows contained quantitative values of each biomolecule in the respective sample conditions. An in-depth description of file structure along with example quantitative files can be found in the supplemental materials and at Argonaut's GitHub.

Like in many other large-scale studies, samples in this study were processed in experimental batches with a designated batch control (the wild-type strain) and included biological replicates. To accommodate this common experimental design, Argonaut utilizes a tree-based hierarchy to organize replicates of experimental conditions in a batch-based format (**Figure 4.3B**). Within each branch, samples are grouped to determine their relationship to the rest of the samples within the uploaded dataset, including designation of replicate sets. Here, three replicates of the wild-type strain were denoted as batch controls, and average molecular abundances in the three replicates of each deletion strains were normalized to those in the selected control. (Note if the batch control is not specified, Argonaut automatically normalizes the condition measurements to the population mean).

During data processing, the significance of molecular perturbations between the control and the experimental conditions are calculated on-the-fly using a stochastic two-sided T test and stored in a database for later querying. Correction for multiple hypothesis testing

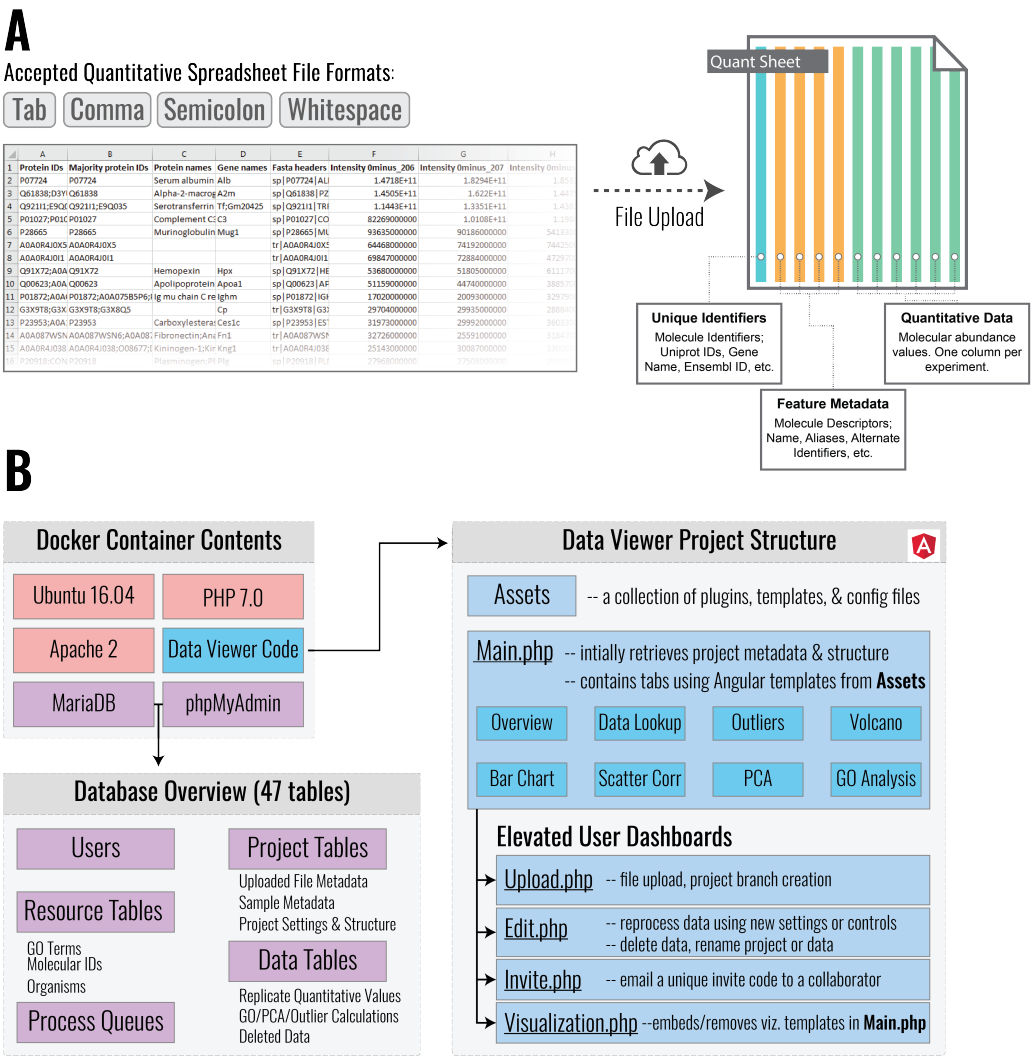


Figure 4.2: Data Upload and Docker Container Structure. (**A**) Currently, Argonaut supports the upload and processing of four formats of plain-text files to supply a data portal with new quantitative data: tab, comma, semicolon, and whitespace. Once uploaded, the user can select which columns contain a unique identifier, metadata, or quantitative values. In the figure, the user would select Protein IDs as a unique identifier; Majority Protein IDs, Protein Names, Gene Names, and Fasta Headers as feature metadata; and the remaining columns as quantitative data. (**B**) A brief graphical breakdown of the backend structure of an Argonaut project running in a Docker container. The project database contains 47 individual tables used to store data for five general purposes: user account information, static processing resources, job processing queues, project-specific metadata, and uploaded quantitative values. The Angular JavaScript framework drives the client-side application. Only elevated users can create or change projects, upload new data, and edit existing data in a data portal.

is available upon request using either the Bonferroni or Benjamini-Hochberg procedures. Once data are processed, the project creator can navigate to a list of predefined options to choose which visualizations and analyses are presented to the portal's users. If downstream Gene Ontology (GO) enrichment is desired and the uploaded file includes GO-compatible unique identifiers, GO enrichment analysis can be enabled (as it was here) by specifying the sample organism, the column containing the unique identifier, and the identifier type.

As a completed study, Veling *et al*. had already conducted data transformations and filtered quantitative values. Note, however, that users can enable these operations during upload. For example, raw quantitative values can be log2 transformed to facilitate fold-change visualizations; missing values can be imputed utilizing a modified left-censored imputation algorithm 4.3; or biomolecules can be filtered from downstream analysis if their abundance measurements are missing in a user-defined number of samples. (Refer to methods section and the GitHub repository for more details).

Interactive Data Examination The tree-based architecture of Argonaut is well suited to multi-omics datasets because it segregates data from a single project into distinct branches (e.g. proteomics, lipidomics, and metabolomics). When exploring the data, users can rapidly switch between branches to compare trends in abundance across samples and to integrate multi-omic data. On entering the data portal, users are presented with an overview, highlighting each branch (i.e. ome) of the project to briefly summarize the number of experimental conditions, replicates, quantified biomolecules, and the average

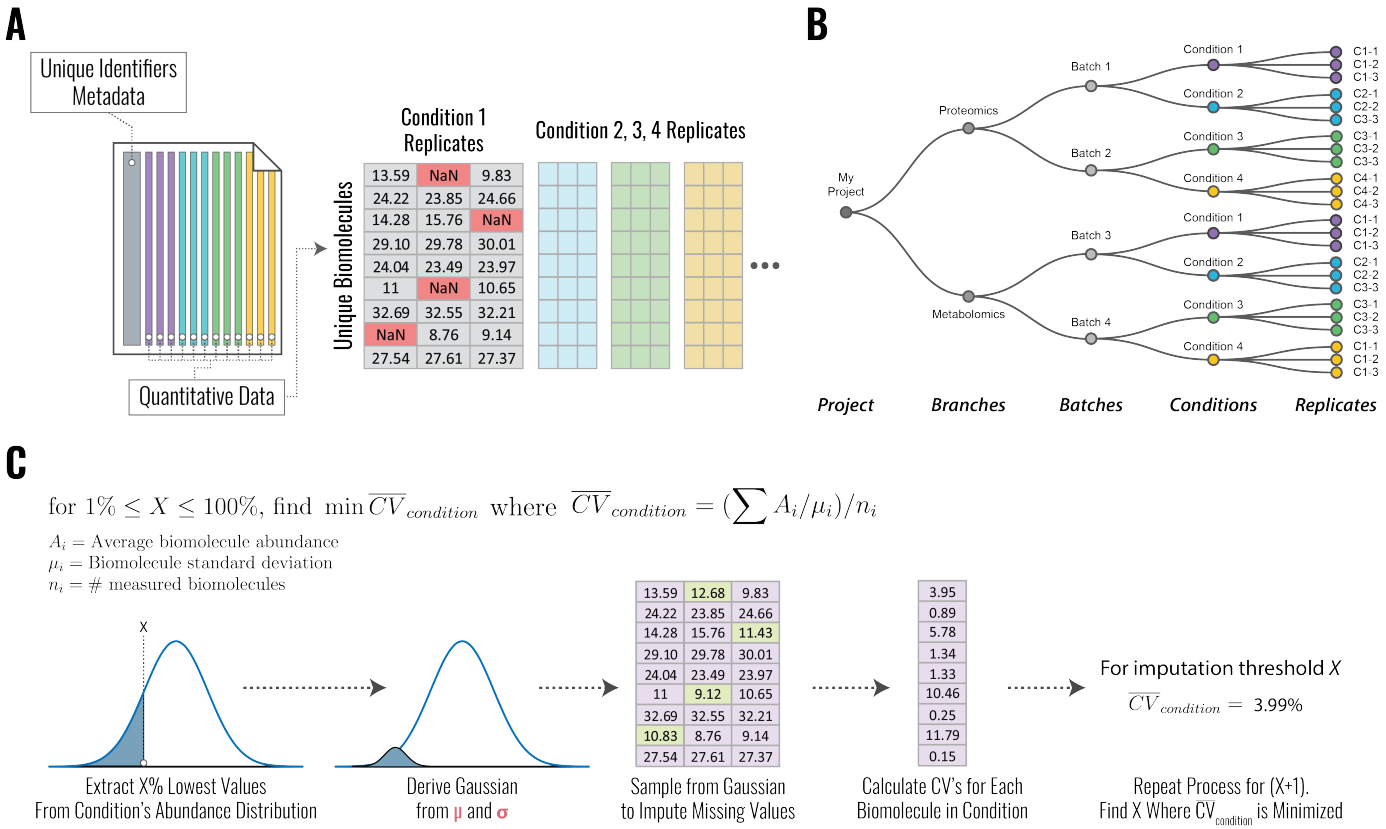


Figure 4.3: Data Portal Organization and Missing Value Imputation. Panel A demonstrates how after a new dataset has been uploaded to the portal by an elevated user, experimental replicates are grouped into their respective condition classifiers. However, some biomolecules may be missing quantitative values. **Panel B** presents a detailed representation of the tree-based hierarchy that Argonaut utilizes to classify experimental replicates into different categories. **Panel C** shows a visual representation of the modified left-censored imputation algorithm used by Argonaut. Imputation is done on each condition separately. The algorithm iteratively selects a percentage of the lowest biomolecular measurements to find a sampling threshold to minimize the average biomolecule coefficient of variation for the specified condition.

biomolecular coefficient of variation. From there, users can navigate to data visualization tabs to explore their dataset through six staple bioinformatic analyses (shown in **Figure 4.4**): volcano plots, principle component analysis (PCA), condition-condition correlation, bar charts of biomolecule abundances, gene ontology (GO) enrichment, and the outlier analysis. All visualizations are generated using the JavaScript library D3.js, which enables real-time customization and interactivity[38]. Significance thresholds can be modified by the user, and many plots support data point lookups by unique identifiers. Any visualization can be downloaded in scalable vector graphics (SVG) format, permitting easy integration into publication-quality figures, such as **Figure4.4**. Additionally, any uploaded or processed data used to generate the visualizations can be exported from the data portal. All interactive visualization options can be inspected on our publicly hosted portal. Further details explaining each visualization can be found on the Argonaut GitHub wiki.

Using our platform, we rapidly recapitulated and visualized several key findings of the Veling *et al*. study. For example, the volcano plot in **Figure 4.4A** demonstrates the upregulation of the iron sulfer assembly protein ISA1 in the absence of the PIM1 gene, revealing a novel relationship between the two proteins. The PCA (**4.4B**) clearly separates respiration-deficient and respiration-competent deletion strains. Biomolecule abundance correlation analysis (**4.4C**) shows a functional relationship between the closely related inner membrane proteases IMP1 and IMP2. **Figure 4.4D** visualizes the log2 fold-change abundance of 3-hexaprenyl-4-hydroxybenzoate (HHB) across all respiration-competent deletion strains and uncovers an increase in the Oct1 mitochondrial protease deletion

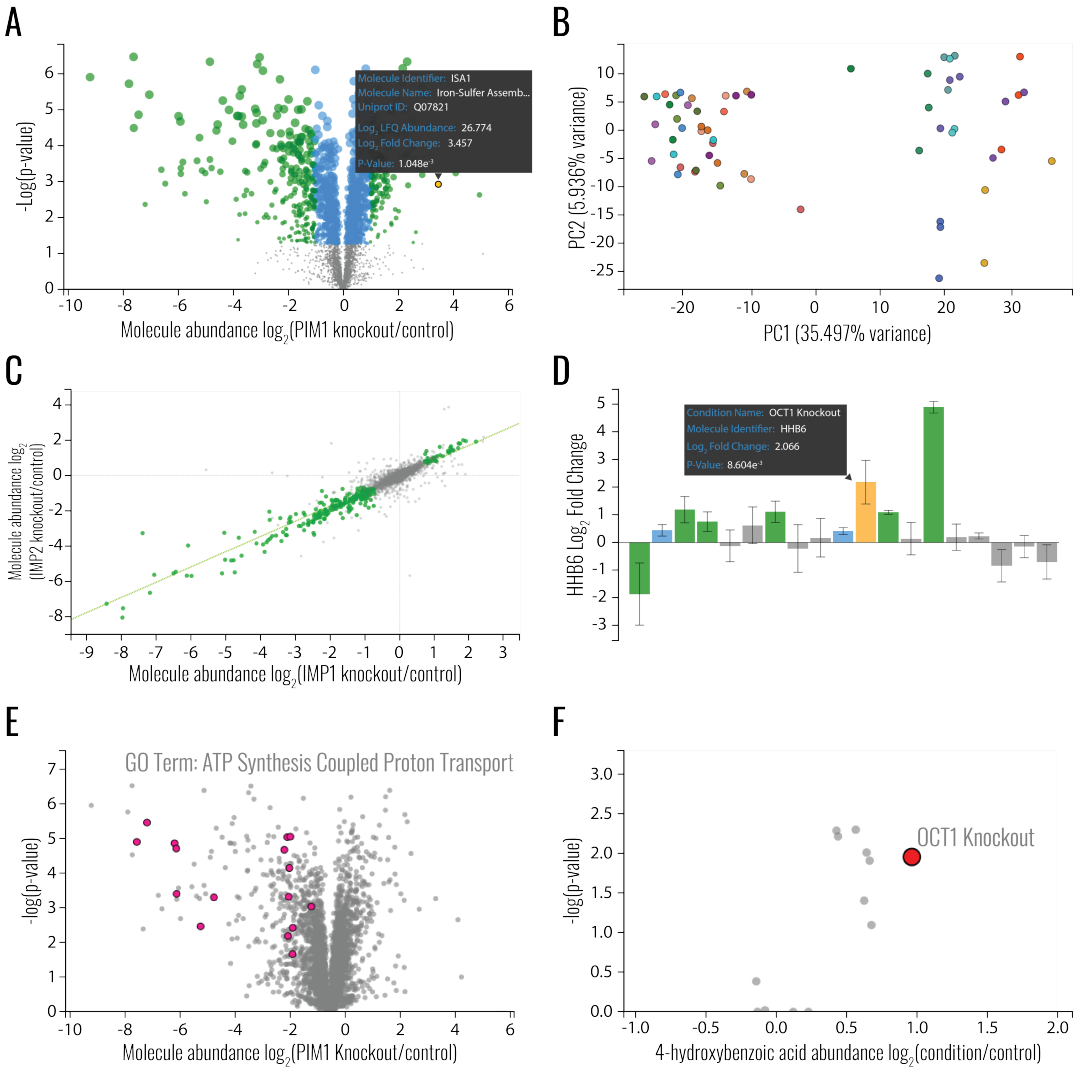


Figure 4.4: Visualization Options in Argonaut. The website generated using the multi-omic data from the Veling *et al*. study features a set of six analyses that are commonly used in omics experiments. All visualizations are fully interactive and generated on-demand using queries from the uploaded data. Significance and fold-change thresholds for data highlighting can be adjusted as desired. Visualizations and data can be exported from the portal as vector graphics, such as ones used to produce this figure, and text-based spreadsheets, respectively.

strain. GO enrichment using the term 'ATP synthesis coupled protein transport' (**4.4E**) recapitulates known relations between the mitochondrial protease PIM1 and ATP. An outlier analysis (**4.4F**) reveals the Oct1 mitochondrial protease deletion to be a significant outlier regarding 4-hydroxybenzoate abundance, a cytosolic precursor to coenzyme Q. These and potentially many other novel biological insights are readily accessible to all portal users. As the website navigation is intuitive, engaging with the hosted data requires almost zero prior guidance from the project creator, lowering barriers to entry into the world of omics data for non-experts.

Collaborative Data Exploration To facilitate the collaborative aspect of Argonaut, we have developed a three-tier accessibility scheme that allows for flexible utilization of a generated data portal's functionality. The first tier provides read-only access to a portal, allowing users with this permission level to view the created portal, interact with the visualizations, and download the processed data. The second tier upgrades the user accesses with additional edit permissions, which allow addition, removal, and editing of the uploaded data. The third permission tier corresponds to that of the project creator, allowing the users to invite collaborators, select visualizations or to delete the portal website entirely. Only project owners and tier three users have comprehensive discretion over access to the portal data. Upon initial data portal creation, the portal creator is the only user with permission to access the site, but with only a few clicks, they can send any collaborator an automatic invitation via email, which grants the new user access to the associated project with the

predefined permission level. Like with the Veling *et al*. dataset here, upon publication project-specific data portals can be made public through the creation of generic usernames and passwords with the tier one permission level. This approach enables online exploration of the study findings worldwide, while preserving the integrity of the underlying dataset.

Discussion

As MS researchers increasingly leverage "Big Data" offered by high-throughput studies to answer complex biological questions, multi-omics has become a particularly powerful approach – one that generates deep, multi-faceted descriptions of the biological system. However, tools to quickly interpret the results of multi-omics experiments across laboratories have not kept pace. Further, in the absence of tools that enhance data accessibility, the discovery potential of rapidly evolving MS techniques remains untapped for many researchers unfamiliar with systems biology data.

We thus developed Argonaut to provide the scientific community with a much-needed tool for the analysis of complex, data-rich resources. To the best our knowledge, there is no other online platform enabling users to compile their multi-omics data into a single resource and present it in the easy-to-explore manner offered by Argonaut. The codeless generation of web portals for data analysis, visualization, and sharing makes this tool uniquely accessible. The only processing steps required to take advantage of the platform are fold-change normalization and statistical testing. These requirements are compatible with the most generic batch-based experimental designs that contain biological or technical

replicates. For additional flexibility, the platform allows users to conduct different data filtering, missing value imputation, or data transformation operations external to Argonaut, if desired. Note that due to the considerable computational overhead, Argonaut does not currently support processing of raw MS data. We also elected to decouple Argonaut from any "searching" operations such as those offered by Trans-Proteomic Pipeline[39]. In doing so we aim to keep our platform lightweight, to increase performance for both data retrieval and visualization, and to widen the platform's utility for the broadest MS and omics community.

Although the analyses conducted using Argonaut are not exhaustive or exclusive to our platform, we believe the ability to securely share experimental omics data in a unified and intuitive format is transformative. By encouraging broad data sharing among the research community, Argonaut is directed to two goals. First, to leverage the expertise of individual researchers from different fields, it allows data portals to be hosted on a public server as companion resources for manuscripts using a few simple Docker commands. Second, to sharpen the significance of novel biological findings, our tool allows multiple portals from our platform to be hosted in a singular location that permits facile comparisons across multiple datasets. Indeed, Argonaut is an agnostic platform that can be used to host the transcriptomic, epigenetic, and phenotypic data that are often generated in the course of comprehensive multi-omic studies. While researchers with the relevant expertise may adapt Argonaut to serve specific projects, applications, or frameworks, for the broader scientific community it provides a stable platform for teams of researchers to concurrently

conduct in-depth analyses of their datasets and readily share their data in an intuitive, accessible format.

Supplemental Methods

The Argonaut Website The Argonaut platform is served using a Dockerized Linux, Apache, MySQL, and PHP (LAMP) web server. The client-side platform was built using the HTML templating framework Bootstrap (3.0) and the Angular JavaScript framework (1.3). Server-side scripts written in PHP (7.0) conduct database operations and relay data from the server to client. On container initialization, a blank data portal is assembled with a predefined administrative user account for data portal management. When an administrative user makes edits to a data portal's architecture (e.g. name, project description, or visualization options), the server utilizes the new data portal settings and embed the new settings together with required HTML templates (**Figure 4.2B**) to generate an updated data portal. Only administrative user accounts are permitted to upload data, edit project architecture, or invite new users. A running data portal can be accessed by any web browser capable of communicating with the Docker machine.

The Docker Image The Docker container consists of a base Ubuntu image (16.04) with an Apache web server (2.4.33), a MySQL relational database (1.6), and the server-side scripting language PHP (7.2). The *PHP Data Object* (*PDO*) extension is used for abstracted database accession and automatic query sanitization. The MySQL database contains forty-seven

tables which serve to rapidly store and retrieve user submitted data. To enable easier project database management, phpMyAdmin, a common web server administration platform, is installed to enable database management (phpMyAdmin, `https://www.phpmyadmin.net`). Details on how to access the Docker phpMyAdmin administrator account can be found in the Argonaut wiki.

Submitting Data to Argonaut Administrative users can submit text-based spreadsheets containing quantified biomolecules from an experiment using the *angular-file-upload* directive (`https://github.com/nervgh/angular-file-upload`). Argonaut is capable of parsing text-based quantitative spreadsheets delimited by tabs, commas, semicolons, or whitespace. When a new text file is uploaded, the file is temporarily saved and the column headers from the file are extracted. A selection of these column headers can be assigned as either unique identifiers, metadata, or quantitative values (**Figure 4.2A**). The user must also indicate which quantitative columns belong to the same condition (i.e. are experimental replicates). The uploaded data are then organized into a hierarchical structure to bin the uploaded experimental replicates into experimental conditions, experimental batches, and branches (**Figure 4.3B**). Branches can be used to separate data generated from the measurement of different biomolecule classes (i.e. different omes). Each uploaded file is denoted as by the keyword batch (i.e. batch of samples), and as such, missing value imputation, control normalization, and $\log_2$ transformation is conducted on each uploaded set of quantitative values independently. If standard molecular identifiers are included in

the uploaded file's metadata, specifying the sample organism, the column containing the standard identifier, and the type of standard identifier can enable optional downstream GO enrichment analysis. UniProt, ChEBI, and KEGG identifiers are currently supported for GO enrichment analysis, though support for other identifiers will be added in the future.

Once all settings are finalized, the data's tree-based hierarchy is presented to the user for review, and raw quantitative values are then uploaded to the database to begin data processing. Submitted files are queued for processing using the *PHP Client URL* library to enable the asynchronous processing of concurrent file uploads and appropriately meter computational resources. After data processing, the uploaded spreadsheet is preserved on the server to allow retrieval of the stored data at any time. Examples of Argonaut compatible files can be found in the supplemental materials, downloaded from the Argonaut GitHub, or downloaded extracted from the example portal provided at: `https://coonlabdatadev.com` using the username "guest" and the password "password".

Quantification and Statistical Analysis

Data Organization When a new set of experimental data is uploaded, the server begins a multi-stage process to group the user-provided quantitative measurements into conditions, conduct an optional $\log_2$ transform for raw quantitative values, optionally filter and/or impute missing values, and conduct significance testing and other analyses. Initially, the raw quantitative values are loaded into memory. Any non-numeric or negative entries found in the quantitative value columns are initially set to 0. Biomolecular identifiers

are checked for uniqueness. If duplicates in the unique identifier values are found, the duplicates are appended with an additional text qualifier. The experimental condition classifiers provided by the user are used to group experimental replicates into conditions using custom PHP objects. These grouped experimental replicates then undergo optional data filtration and missing value imputation.

Data Filtering and Missing Value Imputation Before statistical testing can be conducted, the dataset first must be considered complete, meaning there can be no missing values. The best strategy to account for missing quantitative values is an active area of debate regarding large-scale mass spectrometry profiling experiments[40,41]. To provide an one-size-fits-most solution to this issue, Argonaut provides functionality for the user to remove sparsely quantified biomolecules which are missing in a user-specified proportion of the experimental replicates and additionally offers a left-censored missing value imputation algorithm. Alternatively, data filtering and missing value imputation can be conducted externally to Argonaut if other data cleaning approaches are more appropriate for a particular dataset.

Missing values in mass spectrometry profiling experiments often arise from low-abundance molecular species below the limit of quantification[40,41]. Argonaut's missing value imputation algorithm is adapted from the imputation strategy implemented in Perseus[19,42], where a set of the smallest quantitative values are leveraged to impute missing data. Argonaut's imputation algorithm is visualized in **Figure 4.3A** and **Figure 4.3C**. For each condition, log2 transformed quantitative measurements are placed into an array and placed in ascending

order, generating a roughly normal distribution of quantitative values. An iterative loop is used to subset the smallest $x\%$ of existing quantitative values, $(1\% \leqslant x \leqslant 100\%)$, as demonstrated in**Figure 4.3C**. A gaussian distribution is drawn the mean and standard deviation of the subset data. This gaussian is then randomly sampled to populate all missing values for this condition, and the average biomolecule coefficient of variation (CV) within the condition is calculated for the cutoff x. The ideal cutoff x is selected by minimizing the average biomolecular CV. This calculation is then iteratively applied to all other conditions. If data filtration and missing imputation are not conducted and missing values remain, they will be excluded from further statistical analysis.

Normalization and Statistical Testing After data filtration and missing value imputation are completed, the mean $(\bar{\mathrm{X}})$ and standard deviation (s) of all remaining quantified biomolecules are calculated within each condition. Calculated means then undergo a linear control normalization to better scale raw abundances for visualizations. If a control condition was not specified for an uploaded batch, this calculation uses the mean abundance across all conditions for a biomolecule instead.

$$\bar{\mathrm{X}}_{\mathrm{normalized}} = \bar{\mathrm{X}}_{raw} - \bar{\mathrm{X}}_{\mathrm{control}} \tag{4.1}$$

These quantitative data then undergo statistical testing against the newly uploaded batch's control if one was specified. Otherwise the testing is conducted against the batch's log2 transformed average biomolecular abundance for each respective biomolecule. The test

statistic of differential biomolecule expressions are calculated using an unpaired two-tailed Student's T test, as shown below.

$$t = \frac{(\bar{X}_{condition} - \bar{X}_{control})}{\sqrt{\frac{s_{condition}}{n^2_{condition}} + \frac{s_{control}}{n^2_{control}}}} \tag{4.2}$$

The test statistic is converted to a p-value and stored in the MySQL database. Multiple hypothesis corrected p-values are then calculated using both the Bonferroni (Dunn, 1961) and Benjamini-Hochberg correction methods (Benjamini and Hochberg, 1995). These corrected p-values can be selected for use in Argonaut's interactive visualizations to enable users to be more stringent in what is labeled as statistically significant. Finally, outlier analysis and PCA are conducted as described previously by Stefely et al., 2016.

References

[1] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, no. 1, pp. 1–15, 2017.

[2] F. R. Pinu, D. J. Beale, A. M. Paten, K. Kouremenos, S. Swarup, H. J. Schirra, and D. Wishart, "Systems biology and multi-omics integration: Viewpoints from the metabolomics research community," *Metabolites*, vol. 9, no. 4, 2019.

[3] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent

Acquisition: A New Concept for Consistent and Accurate Proteome Analysis," *Molecular & Cellular Proteomics*, vol. 11, no. 6, p. O111.016717, 2012.

[4] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, no. 1, pp. 339–347, 2013.

[5] F. Meier, P. E. Geyer, S. Virreira Winter, J. Cox, and M. Mann, "BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes," *Nature Methods*, vol. 15, no. 6, pp. 440–448, 2018.

[6] K. A. Overmyer, T. W. Rhoads, A. E. Merrill, Z. Ye, M. S. Westphall, A. Acharya, S. K. Shukla, and J. J. Coon, "Proteomics, lipidomics, metabolomics and 16S DNA sequencing of dental plaque from patients with diabetes and periodontal disease," *bioRxiv*, p. 2020.02.25.963967, 2020.

[7] G. Rosenberger, C. C. Koh, T. Guo, H. L. Röst, P. Kouvonen, B. C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O. T. Schubert, P. Faridi, H. A. Ebhardt, M. Matondo, H. Lam, S. L. Bader, D. S. Campbell, E. W. Deutsch, R. L. Moritz, S. Tate, and R. Aebersold, "A repository of assays to quantify 10,000 human proteins by SWATH-MS," *Scientific Data*, vol. 1, no. 1, p. 140031, 2014.

[8] J. R. Wiśniewski, A. Zougman, N. Nagaraj, and M. Mann, "Universal sample preparation method for proteome analysis.," *Nature methods*, vol. 6, no. 5, pp. 359–62, 2009.

[9] Y. Zhang, J. M. Vera, D. Xie, J. Serate, E. Pohlmann, J. D. Russell, A. S. Hebert, J. J. Coon, T. K. Sato, and R. Landick, "Multiomic Fermentation Using Chemically Defined Synthetic Hydrolyzates Revealed Multiple Effects of Lignocellulose-Derived Inhibitors on Cell Physiology and Xylose Utilization in Zymomonas mobilis," *Frontiers in Microbiology*, vol. 10, p. 2596, 2019.

[10] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, no. 7608, pp. 500–505, 2016.

[11] C. P. Lapointe, J. A. Stefely, A. Jochem, P. D. Hutchins, G. M. Wilson, N. W. Kwiecien, J. J. Coon, M. Wickens, and D. J. Pagliarini, "Multi-omics Reveal Specific Targets of the RNA-Binding Protein Puf3p and Its Orchestration of Mitochondrial Biogenesis," *Cell Systems*, vol. 6, no. 1, pp. 125–135.e6, 2018.

[12] J. G. Meyer, S. Liu, I. J. Miller, J. J. Coon, and A. Gitter, "Learning Drug Functions from Chemical Structures with Convolutional Neural Networks and Random Forests," *Journal of Chemical Information and Modeling*, 2019.

[13] K. Overmyer, P. Muir, and J. Coon, "Discovery metabolomics and lipidomics of canine synovial fluid and serum," *Osteoarthritis and Cartilage*, vol. 26, p. S172, 2018.

[14] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer,

K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[15] D. Wang, B. Eraslan, T. Wieland, B. Hallström, T. Hopf, D. P. Zolg, J. Zecha, A. Asplund, L.-H. Li, C. Meng, M. Frejno, T. Schmidt, K. Schnatbaum, M. Wilhelm, F. Ponten, M. Uhlen, J. Gagneur, H. Hahne, and B. Kuster, "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues," *Molecular Systems Biology*, vol. 15, no. 2, p. e8503, 2019.

[16] E. M. Weisenhorn, T. J. Van 'T Erve, N. M. Riley, J. R. Hess, T. J. Raife, and J. J. Coon, "Multi-omics evidence for inheritance of energy pathways in red blood cells," *Molecular and Cellular Proteomics*, vol. 15, no. 12, pp. 3614–3623, 2016.

[17] E. G. Williams, Y. Wu, W. Wolski, J. Y. Kim, J. Lan, M. Hasan, C. Halter, P. Jha, D. Ryu, J. Auwerx, and R. Aebersold, "Quantifying and Localizing the Mitochondrial Proteome Across Five Tissues in A Mouse Population," *Molecular & Cellular Proteomics*, vol. 17, no. 9, pp. 1766–1777, 2018.

[18] M. Choi, C. Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek, "MSstats: An R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments," *Bioinformatics*, vol. 30, no. 17, pp. 2524–2526, 2014.

[19] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox,

"The Perseus computational platform for comprehensive analysis of (prote)omics data," *Nature Methods*, vol. 13, no. 9, pp. 731–740, 2016.

[20] M. Palmblad and N. J. van Eck, "Bibliometric Analyses Reveal Patterns of Collaboration between ASMS Members," *Journal of the American Society for Mass Spectrometry*, vol. 29, no. 3, pp. 447–454, 2018.

[21] S. Sidoli, K. Kulej, and B. A. Garcia, "Why proteomics is not the new genomics and the future of mass spectrometry in cell biology," *Journal of Cell Biology*, vol. 216, no. 1, pp. 21–24, 2017.

[22] G. Dennis, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: Database for Annotation, Visualization, and Integrated Discovery.," *Genome biology*, vol. 4, no. 5, p. P3, 2003.

[23] W. James Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[24] J. Severin, M. Lizio, J. Harshbarger, H. Kawaji, C. O. Daub, Y. Hayashizaki, N. Bertin, A. R. Forrest, A. Beckhouse, C. Wells, D. Vijayan, E. Mason, E. Wolvetang, K. Hitchens, K. A. Le Cao, L. Nielsen, Y. Hayashizaki, L. Fearnley, and T. Kenna, "Interactive visualization and analysis of large-scale sequencing datasets using ZENBU," *Nature Biotechnology*, vol. 32, no. 3, pp. 217–219, 2014.

[25] N. J. Krogan, S. Lippman, D. A. Agard, A. Ashworth, and T. Ideker, "The Cancer Cell Map Initiative: Defining the Hallmark Networks of Cancer," *Molecular Cell*, vol. 58, no. 4, pp. 690–698, 2015.

[26] D. K. Schweppe, E. L. Huttlin, J. W. Harper, and S. P. Gygi, "BioPlex Display: An Interactive Suite for Large-Scale AP-MS Protein-Protein Interaction Data," *Journal of Proteome Research*, vol. 17, no. 1, pp. 722–726, 2018.

[27] P. J. Thul and C. Lindskog, "The human protein atlas: A spatial map of the human proteome," *Protein Science*, vol. 27, no. 1, pp. 233–244, 2018.

[28] T. Kluyver, B. Ragan-kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter Notebooks—a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pp. 87–90, 2016.

[29] D. Torre, A. Lachmann, and A. Ma'ayan, "BioJupies: Automated Generation of Interactive Notebooks for RNA-Seq Data Analysis in the Cloud," *Cell Systems*, vol. 7, no. 5, pp. 556–561.e3, 2018.

[30] J. L. Norris, M. A. Farrow, D. B. Gutierrez, L. D. Palmer, N. Muszynski, S. D. Sherrod, J. C. Pino, J. L. Allen, J. M. Spraggins, A. L. Lubbock, A. Jordan, W. Burns, J. C. Poland, C. Romer, M. L. Manier, Y. W. Nei, B. M. Prentice, K. L. Rose, S. Hill, R. Van De

Plas, T. Tsui, N. M. Braman, M. R. Keller, S. A. Rutherford, N. Lobdell, C. F. Lopez, D. B. Lacy, J. A. McLean, J. P. Wikswo, E. P. Skaar, and R. M. Caprioli, "Integrated, High-Throughput, Multiomics Platform Enables Data-Driven Construction of Cellular Responses and Reveals Global Drug Mechanisms of Action," *Journal of Proteome Research*, vol. 16, no. 3, pp. 1364–1375, 2017.

[31] V. Sharma, J. Eckels, G. K. Taylor, N. J. Shulman, A. B. Stergachis, S. A. Joyner, P. Yan, J. R. Whiteaker, G. N. Halusa, B. Schilling, B. W. Gibson, C. M. Colangelo, A. G. Paulovich, S. A. Carr, J. D. Jaffe, M. J. Maccoss, and B. Maclean, "Panorama: A targeted proteomics knowledge base," *Journal of Proteome Research*, 2014.

[32] J. A. Vizcaíno, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P. A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H. J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, and H. Hermjakob, "ProteomeXchange provides globally coordinated proteomics data submission and dissemination," *Nature Biotechnology*, vol. 32, no. 3, pp. 223–226, 2014.

[33] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, "MetaboAnalyst: A web server for metabolomic data analysis and interpretation," *Nucleic Acids Research*, 2009.

[34] D. Merkel, "Docker : Lightweight Linux Containers for Consistent Development and Deployment Docker : a Little Background Under the Hood," *Linux Journal*, vol. 2014, no. 239, pp. 2–7, 2014.

[35] 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. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno, and H. Hermjakob, "The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience," *Molecular and Cellular Proteomics*, vol. 13, no. 10, pp. 2765–2775, 2014.

[36] 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. Vizcaíno, M. Chambers, A. Pizarro, and D. Creasy, "The mzIdentML Data Standard for Mass Spectrometry-Based Proteomics Results," *Molecular & Cellular Proteomics*, vol. 11, no. 7, p. M111.014381, 2012.

[37] M. T. Veling, A. G. Reidenbach, E. C. Freiberger, N. W. Kwiecien, P. D. Hutchins, M. J. Drahnak, A. Jochem, A. Ulbrich, M. J. Rush, J. D. Russell, J. J. Coon, and D. J. Pagliarini, "Multi-omic Mitoprotease Profiling Defines a Role for Oct1p in Coenzyme Q Production," *Molecular Cell*, vol. 68, no. 5, pp. 970–977.e11, 2017.

[38] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[39] E. W. Deutsch, L. Mendoza, D. Shteynberg, T. Farrah, H. Lam, N. Tasman, Z. Sun, E. Nilsson, B. Pratt, B. Prazen, J. K. Eng, D. B. Martin, A. I. Nesvizhskii, and R. Ae-

bersold, "A guided tour of the Trans-Proteomic Pipeline," *Proteomics*, vol. 10, no. 6, pp. 1150–1159, 2010.

[40] K. T. Do, S. Wahl, J. Raffler, S. Molnos, M. Laimighofer, J. Adamski, K. Suhre, K. Strauch, A. Peters, C. Gieger, C. Langenberg, I. D. Stewart, F. J. Theis, H. Grallert, G. Kastenmüller, and J. Krumsiek, "Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies," *Metabolomics*, vol. 14, no. 10, p. 128, 2018.

[41] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies," *Journal of Proteome Research*, vol. 15, no. 4, pp. 1116–1125, 2016.

[42] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B* (*Methodological*), vol. 57, no. 1, pp. 289–300, 1995.

Chapter 5

GENOME-GUIDED LIPID IDENTIFICATION ENABLED BY LIPIDGENIE

DRB designed and developed LipidGenie and analysed data for the quantitative trait locus clustering portions in the chapter.
Portions of this chapter are included in a submitted manuscript under peer review:

Linke V, Overmyer KA, Miller IJ, **Brademan DR**, Hutchins PD, Trujillo EA, Reddy T, Russell JD, Schueler KL, Stapleton DS, Rabaglia ME, Keller MP, Gatti DM, Keele G, Pham D, Broman KW, Churchill GA, Attie AD, Coon JJ. *Genome-guided lipid identification.* Nature Metabolism. **2020**.

Abstract

Lipids are metabolic actors in health and disease including diabetes and obesity. Mass spectrometry (MS)-based discovery lipidomics offers global and unbiased access to these crucial molecules. Current lipid identification strategies, which leverage mass-to-charge ratio, chromatographic retention time, and fragmentations pattern, are unable to identify the majority of detected lipid chromatographic features. We present genome-lipid association data as an orthogonal tool for lipid identification. Using high resolution LC-MS/MS, we analyzed liver and plasma derived from 384 Diversity Outbred (DO) mice and quantified 3,283 lipid-like features. These features were mapped to 5,622 lipid quantitative trait loci (QTL) and were compiled into a public web-resource, termed LipidGenie. Leveraging the genome-lipid associations embedded in this resource, we were able to derive identifications for an additional number of lipids, including gangliosides through their association with the protein *B4galnt1* as well as a novel group of sex-specific phosphatidylcholines. Finally, LipidGenie allows a user to query QTLs from either a mass feature or genetic locus perspective. Using this functionality, we uncover evidence suggesting the genes ABHD1 and ABHD2 possess acyl chain-specific functions.

Introduction

Beyond their roles in energy storage and membrane structure, lipids are central actors of myriad metabolic functions and molecular signaling.[1,2] As our understanding of these

diverse lipid functions grows, so too does our appreciation for the complexity of the lipidome of mammalian systems.[3] Mass spectrometry has emerged as the central tool to dissect and quantify lipid species.[4–6] Specifically, using liquid chromatography (LC) coupled with high resolution tandem mass spectrometry (MS/MS), over one thousand unique lipid features from a complex mixture can be quantified in under an hour.[7] From these features hundreds of individual lipids are routinely identified; however, the majority of the features remain unannotated.[8,9] The result is that more often than not, the majority of MS data are not leveraged.[8,10,11]

One strategy for lipid feature identification is to group compounds likely to be related. For example, members of a lipid class often (i) appear within a defined chromatographic retention time, (ii) occupy a characteristic mass range, and (iii) exhibit similar dissociation patterns when subjected to fragmentation.[12] Most efforts to improve lipid identification rates exploit one or all of these steps[3,13–15] - including our laboratories recent description of a software suite that constructs tailored MS/MS libraries for automated lipid spectral identification.[16,17] Others have sought to build on this information by adding external complementary data, such as measurement of collisional cross section,[10] or labile hydrogen counting, among others.[18] All of these methods show great promise but share in the common theme that they incorporate lipid chemical properties into their identification inference.

In the field of shotgun proteomics, genome sequence data are used to identify experimental peptide tandem mass spectra. Given the success of this field we wondered whether

genomic information could be leveraged similarly in the field of lipidomics. Unfortunately, it is not possible to directly predict an organism's theoretical lipid identities from genomic data in a similar fashion to how theoretical peptides are derived from known genes;[19] however, shared genetic regulation among lipids could provide key information to facilitate identification of uncommon lipid species. For example, a recent large-scale multi-omic study of a knock-out yeast library demonstrated dramatic regulation of the lipidome,[20] nuclear magnetic resonance (NMR)-base untargeted metabolomics identified disease-associated metabolites and genomic regions via quantitative trait loci (QTL) mapping,[21,22] and recent genome wide association studies (GWAS) were used to assist in small molecule identifications from both MS and NMR data.[23–26] We propose that a global genome-lipid association map would add a fourth (iv), orthogonal dimension of data to assist in lipid feature identification.

To construct a global genome-lipid association map we measured plasma and liver lipids from a mouse population using LC-MS/MS and performed quantitative trait loci (QTL) mapping.[28–30] We have selected the diversity outbred mice (DO), a multiparent population (MPP) derived from 8 highly diverse founder strains (**Figure 5.1A**.[30–32] A key advantage of the MPP is that we can identify the additive genetic effects contributed by each founder strain at a quantitative trait locus (QTL). Unlike standard bi-parental crosses where the founder haplotype effects are either increasing or decreasing, the haplotype effects in a MPP are complex and enable us to distinguish chance co-localization from pleiotropic effects. In addition, we can compare founder haplotype effects across different studies using

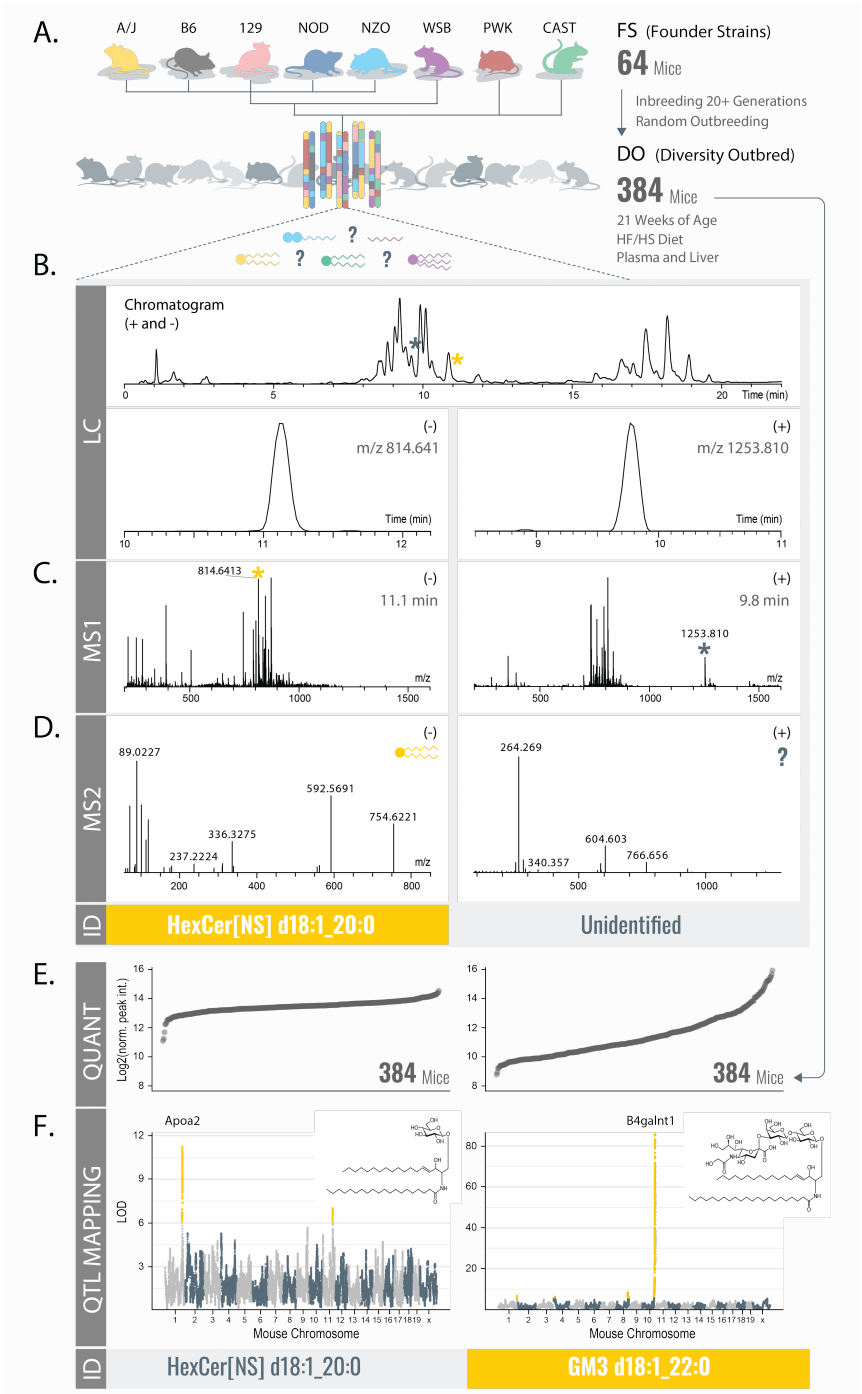


Figure 5.1: LC-MS/MS Lipidomics and QTL Mapping as Ways to Bolster Lipid Identifications. **a**, A modified MTBE lipid extraction[27] was performed on plasma and liver from 384 DO mice. **b-d**, Lipid extracts were analyzed by LC-MS/MS. Identifications were obtained through LipiDex[16] based on retention time window (**b**), exact mass (**c**), and tandem mass fragmentation (**d**). **e**, All lipidomic features (identified and unidentified) were mapped onto the mouse genome via QTL mapping, revealing genomic position and founder strain allele effect pattern as results for each QTL. This additional information enabled identification of otherwise unidentified features.

DO mice to identify shared effects on traits that were not directly measured in only one group of DO mice. Specifically, we were able to use liver gene expression from a previously published study to propose candidate genes for the lipid features in this study.[32] Further, each of the eight inbred DO founder mouse strains (129, A/J, B6, CAST, NOD, NZO, PWK, WSB) contributes to generate distinct allele effect patterns at each locus, thus providing an additional criterion for gene identification.[33] Finally, careful control of external sources of variation such as diet and environmental conditions, allows the extraordinary phenotypic diversity of the DO[34] to be directly attributed to genetic diversity. The DO population has already been used extensively to map clinical traits,[35] transcripts,[34] proteins,[36] gut microbiota and bile acids[37] providing a wealth of existing data to integrate with global genome-lipid associations.

Here we describe the first discovery lipidomics analysis on a cohort of DO mice. In doing so, we present QTL position as an independent piece of information to guide lipid identification and apply it to define unidentified mass spectral features. We demonstrate the utility of genome-lipid associations to assign identification or function in independent studies through a novel web-based resource - LipidGenie (`http://lipidgenie.com/`). LipidGenie compiles generated genome-lipid associations in an easy-to-access web application, enabling the scientific community to interrogate our newly-generated resource using their own expertise and data.

Results

QTL mapping connects lipids to their genetic regulators. To explore the hypothesis that global association of mass spectrometry data to genomic coordinates could assist lipid identification, we collected whole lipidome profiles of plasma and liver tissue from 64 founder strain mice (FS) and 384 DO mice using high resolution LC-MS/MS (**Figure 5.1a**). Altogether, we performed 894 LC-MS/MS discovery lipidomics experiments from which we extracted approximately 4,500,000 tandem mass spectra(**Figure 5.1d**). From the full scan mass spectrometry data (**Figure 5.1c**), we detected and quantified 19,636 molecular features; 12,429 in plasma and 7,207 in liver (**Figure 5.1b**). Next, we applied the LipiDex algorithm[16] to (1) match the tandem mass spectra to their respective features, (2) eliminate features derived from adduction, dimerization, in-source fragmentation, etc., and (3) to assign molecular identities when possible (**Figure 5.1d**. From the 3,283 distinct molecular features that remained, we identified 594 lipids (from 1,721 features) in plasma and 584 lipids (from 1,562 features) in liver (see Methods: Lipidomics Data Analysis for details). **Supplementary Figure S5.2a-b** and **Table 5.1** provides an overview of the identified lipids that span roughly 30 lipid subclasses from five of the major classes: fatty acyls, glycerolipids, phospholipids, sphingolipids, and sterol lipids.[38] For 70% of these identifications we find MS/MS evidence to detail fatty acid composition, otherwise we report sum composition.

Supplementary Figure S5.1a and **b** present a bird's-eye view of these plasma and liver lipidomes. Here each distinct molecular feature is plotted as a function of its m/z and

			Plasma				Liver			
Lipid Category	Lipid Class	Abbreviation(s)	Count	% of IDd	Molecular Level	% of Class	Count	% of IDd	Molecular Level	% of Class
Fatty Acyl			**29**	**4.6%**	**29**	**100.0%**	**37**	**5.9%**	**37**	**100.0%**
	Acyl Carnitine	AC	2	0.3%	2	100.0%	6	1.0%	6	100.0%
	Fatty Acid*	FA	27	4.3%	27	100.0%	31	5.0%	31	100.0%
Glycerolipid			**210**	**33.2%**	**165**	**78.6%**	**185**	**29.7%**	**111**	**60.0%**
	Diglyceride	DG, Alkenyl-DG	2	0.3%	2	100.0%	17	2.7%	15	88.2%
	Triglyceride	TG, Alkanyl-TG, Alkenyl-TG	208	32.9%	163	78.4%	168	27.0%	96	57.1%
Phospholipid			**287**	**45.3%**	**194**	**67.6%**	**303**	**48.6%**	**238**	**78.5%**
	Cardiolipin	CL	0	0.0%			9	1.4%	8	88.9%
	Lyso-Phosphocholine	Lyso-PC	27	4.3%	27	100.0%	18	2.9%	18	100.0%
	Lyso-Phosphoethanolamine	Lyso-PE	5	0.8%	5	100.0%	10	1.6%	10	100.0%
	Lyso-Phosphoinositol	Lyso-PI	1	0.2%	1	100.0%	3	0.5%	3	100.0%
	Phosphocholine	PC	129	20.4%	69	53.5%	88	14.1%	62	70.5%
	Phosphoethanolamine	PE, PE-NMe2	24	3.8%	22	91.7%	80	12.8%	57	71.3%
	Phosphoglycerol	PG	2	0.3%	2	100.0%	26	4.2%	26	100.0%
	Phosphoinositol	PI	23	3.6%	19	82.6%	23	3.7%	21	91.3%
	Plasmalogen	Plasmanyl-PC, Plasmenyl-PE, Plasmanyl-PE, Plasmenyl-PC	76	12.0%	49	64.5%	46	7.4%	33	71.7%
Sphingolipid			**102**	**16.1%**	**41**	**40.2%**	**96**	**15.4%**	**46**	**47.9%**
	Ceramide	Cer[NS], HexCer[NS], HexCer[AP], Cer[NP], Cer[AS], Cer[AP]	40	6.3%	34	85.0%	58	9.3%	39	67.2%
	Ganglioside*	GM2/GM3	13	2.1%	7	53.8%	8	1.3%	6	75.0%
	Sphingomyelin	SM	49	7.7%	0	0.0%	29	4.7%	0	0.0%
	Sphingosine	SP	0	0.0%			1	0.2%	1	100.0%
Sterol Lipid			**6**	**0.9%**	**6**	**100.0%**	**2**	**0.3%**	**2**	**100.0%**
	Cholesteryl Ester	CE	6	0.9%	6	100.0%	2	0.3%	2	100.0%
	* hand-identified		Count	% of Total	Molecular Level	% of IDd	Count	% of Total	Molecular Level	% of IDd
		Identified	634	36.8%	**435**	**68.6%**	623	39.9%	**434**	**69.7%**
		Unidentified	1087	63.2%			939	60.1%		
		Total	**1721**	**100.0%**			**1562**	**100.0%**		

Table 5.1: Breakdown of Lipid Identifications in Plasma and Liver Samples by Lipid Class. "Molecular Level" refers to lipids identified with individual fatty acid rather than as a sum composition.

chromatographic retention time. Identified lipids are colored by class; we note members of individual lipid classes group well, adding confidence to their identification. Triglycerides (TG),[39] for example, as hydrophobic lipids with three fatty acids can be found at high m/z and late chromatographic retention. From this perspective, we observe that the unidentified molecular features, frac23 of all detected species, are either clustered around identified lipid classes or class coverage and (2) reveal the presence of additional lipid classes. exist on m/z and retention islands. We conclude that these data can be further interrogated to (1) expand existing lipid class coverage and (2) reveal the presence of additional lipid classes.

Next we extracted quantitative information from all detected molecular features across all 384 animals, creating a molecular trait for each feature. **Figure 5.1e** displays the quantitative values of two such individual molecular traits from plasma; one identified as a phospholipid and one unidentified. Plasma HexCer[NS] d18:1_20:0 has a relative abundance dynamic range of 15-fold across all 384 animals. For comparison, we plot the abundance of a molecular feature with a mass of 1252.8028 Da. Here we see an even greater dynamic range of 75-fold, however, the feature was unidentified using our traditional data processing. Correlation to a candidate gene region ultimately led to the identification of the feature (**Figure 5.1f**). To correlate these MS-derived lipid quantitative phenotypes (*vide supra*) with genomic variation we performed quantitative trait locus (QTL) mapping using R/qtl2.[29]

Figure 5.2a displays a hierarchically clustered heatmap of these quantitative results for all measured molecular traits (1,721 and 1,562 for plasma and liver, respectively) across all

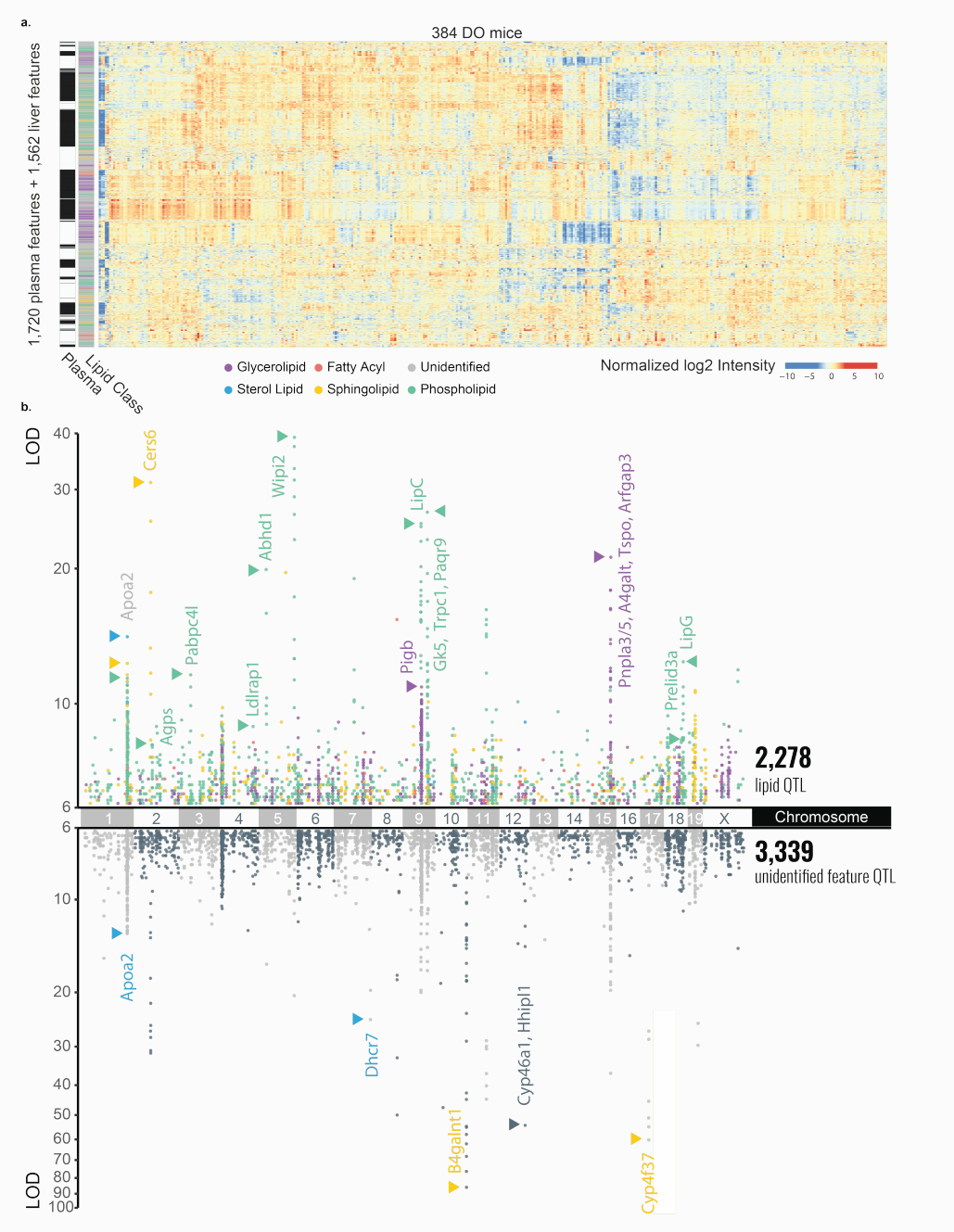


Figure 5.2: Large-Scale Lipid Quantitative Profiling and Subsequent QTL Mapping Reveals Hotspots of Associated Lipids. a, In plasma, we quantified 1,721 lipidomic features, 621 of which were identified, and in liver, we quantified 1,562 lipidomic features, 615 of which were identified. Hierarchical clustering of all 3,283 lipidomic features' intensities by the 384 DO mice resulted in distinct clustering by lipid class, notably across tissue type. **b**, When mapped onto the mouse genome, 1,405 plasma and 1,190 liver features showed at least one QTL with an LOD > 6 as displayed in a Manhattan plot (n = 3,353 + 2,269 = 5,622 total QTLs). A number of lipid hotspots are shared by identified lipids and unidentified features (e. g. at Apoa2), while others only appear among the unidentified features (*e.g.* at *B4galnt1*).

384 animals. Notably, we observe considerable clustering by lipid class, even across tissue type (y-axis). We detected 3,348 plasma lipid QTL for 1,405 of the 1,721 (81.6%) traits (logarithm of odds (LOD) score > 6). 1,351 of these were from identified lipids, while 1,997 were from unidentified features. Similarly, in liver, we detected 2,269 lipid QTL for 1,190 of the 1,562 (76.2%) traits, of which 927 were from identified lipids while 1,342 were from unidentified features. **Figure 5.2b** and **Supplementary Figure S5.2b** present the genetic correlations for this entire collection of significant QTL extracted in a Manhattan plot. We note that the unidentified molecular traits cluster among the various identified lipid classes, which provides further evidence that these features are of biological origin and amenable for further interrogation. Secondly, the unidentified features occupy additional distinct loci, implicating previously unidentified lipid classes.

QTL map recapitulates known APOA2 biology and informs cholesteryl ester lipid identifications. Several genetic loci are strongly associated with lipids and appear as hotspots - locations on the genome where multiple lipid QTL co-map (**Figure 5.1E**). To better explore these regions, we asked whether these co-mapping lipid QTL shared a common genetic relationship to segregating alleles at the locus. One advantage of the DO mice is that shared founder strain allele effect patterns can be indicative of a common genetic regulator.[30] Thus, we define a lipid QTL hotspot as multiple lipid QTL co-mapping ($\pm$ 2 Mbp) with a shared founder strain allele effect pattern. We identified a number of hotspots; To garner additional support for founder strain specific genetic effects on lipid abundance, we profiled plasma

and liver lipids for each of the founder strains (4 males, 4 females).

Figure 5.3a highlights a lipid QTL hotspot on chromosome 1:171 Mbp. Here, 255 lipid traits, all from plasma, co-localize with a shared allele effect pattern of upregulation associated with alleles derived from the founder strain 129 (**Supplementary Figure S5.3a**). The lipid with the highest LOD at this locus was a cholesteryl ester (CE 18:2), which was also elevated in founder strain 129 plasma **Supplementary Figure S5.3b**. At 171 Mbp on chromosome 1, strain 129 possesses a missense SNP in the *Apoa2* gene (rs8258226), resulting in a 61Ala >Val substitution in the protein apolipoprotein-II.[40] A prior DO study identified APOA2 protein and mRNA expression QTL in liver tissue but these displayed different allele effects than plasma lipid QTL, suggesting that the causal variants that modulate their respective levels differ **Supplementary Figure S5.3c**.[36] Notably, APOA2 protein is a major component of high density lipoprotein (HDL) particles in plasma, corroborated by human HDL traits mapping to APOA2 in GWAS,[41], and is considered a principal genetic regulator of plasma HDL levels in mice.[42–45] The other major components of HDL particles are phospholipids (35-50%) and cholesteryl esters (30-40%) (**Figure 5.3b**).[46] Consistent with this composition, seven sub-types of phospholipids and various cholesteryl esters (CE) map to the *Apoa2* locus (**Figure 5.3e**). Sphingolipids, a minor component of HDL particles, map in four different sub-classes to this *Apoa2* locus. We conclude that this hotspot illuminates the molecular composition of HDL particles in mice, while also linking an additional 130 unidentified lipid features to this locus.

To test if a shared QTL would enable identification of additional lipids, we plotted the

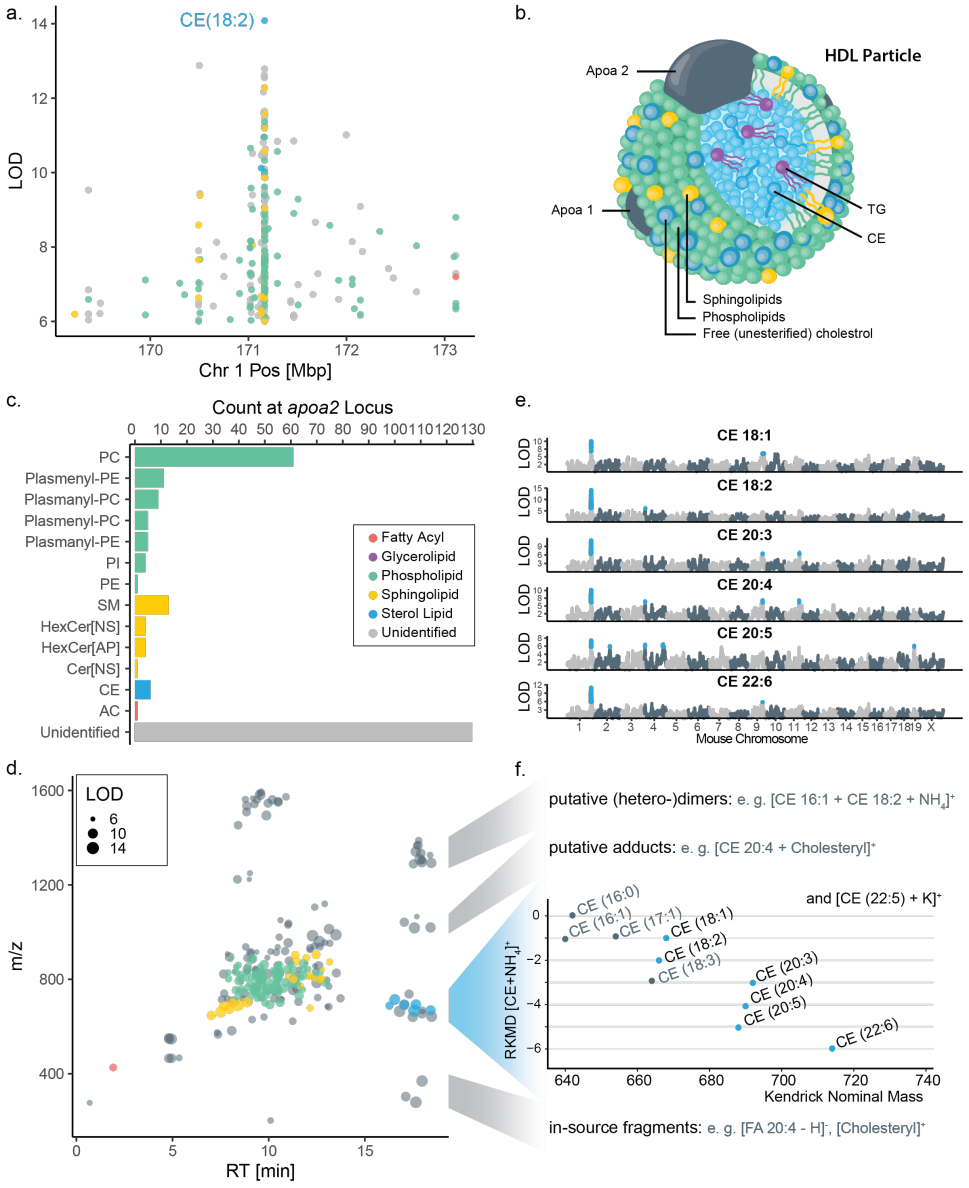


Figure 5.3: Co-mapping of Lipids at the Apoa2 Locus Facilitated Identification of Additional Cholesteryl Esters. a, One lipid hotspot on chromosome 1 at 171 Mbp is shared by 255 plasma lipid features co-mapping with a common 129 high allele effect(**Supplementary Figure S5.3a**). **b**, The candidate gene at this locus is Apoa2, which encodes for apolipoprotein II, which is carried on HDL cholesterol particles along with **c**, a variety of lipid classes, mostly phospho- and sphingolipids, which mapped to the locus. **d**, When plotting all 255 Apoa2-specific lipid features in the m/z-RT plane, a group of unidentified features sharing the RT region with CEs stood out. **e**, Notably, all six CEs show their primary QTL at this locus, as visible from their individual LOD plots. **f**, Subsets of the unidentified features could subsequently be identified as CE-related features, including heterodimers, cholesterol-adducts and in-source fragments.

255 lipid traits that map to the *Apoa2* locus as a function of chromatographic retention time, mass, and identification status (**Figure 5.3d**). A cluster of unidentified lipid features shared retention time with CEs, a class of lipids that are often devoid of informative fragments.[47] All CEs showed their major QTL at the *Apoa2* locus (**Figure 5.3e**) providing greater confidence in their identification. The shared genetic regulation further allowed us to predict a CE identity for the cluster of unidentified co-mapping features. Examination of their total masses[48] and tandem mass spectra supported the annotation of five additional CEs (**Figure 5.3f**), while another 18 lipid features' m/z and RT were consistent with technical artifacts of CEs: eleven heterodimers, four cholesterol adducts, and three in-source fragments.

QTL map provides an orthogonal tool for lipid identification - the case of polygenic gangliosides. On chromosome 10, at 127 Mbp we observed a significant lipid QTL hotspot. At this site, over twenty-five plasma and liver lipid features mapped with the highest overall significance (**Figure 5.4a**). These features also shared a common allele dependence; i.e., NOD-driven and split between NOD high vs. low effect (**Figure 5.4b**). None of these lipid features were identified following our conventional data analysis strategy, which leverages retention time, mass, and tandem mass spectra. The features were observed in two distinct clusters based on m/z and RT, suggesting they could derive from two distinct lipid classes (**Figure 5.4c**). Given that these unidentified features (1) appeared as two defined lipid classes and (2) were high scoring at a genetic locus with opposite allele effects, we reasoned that identification of the causal gene may enable their identification.

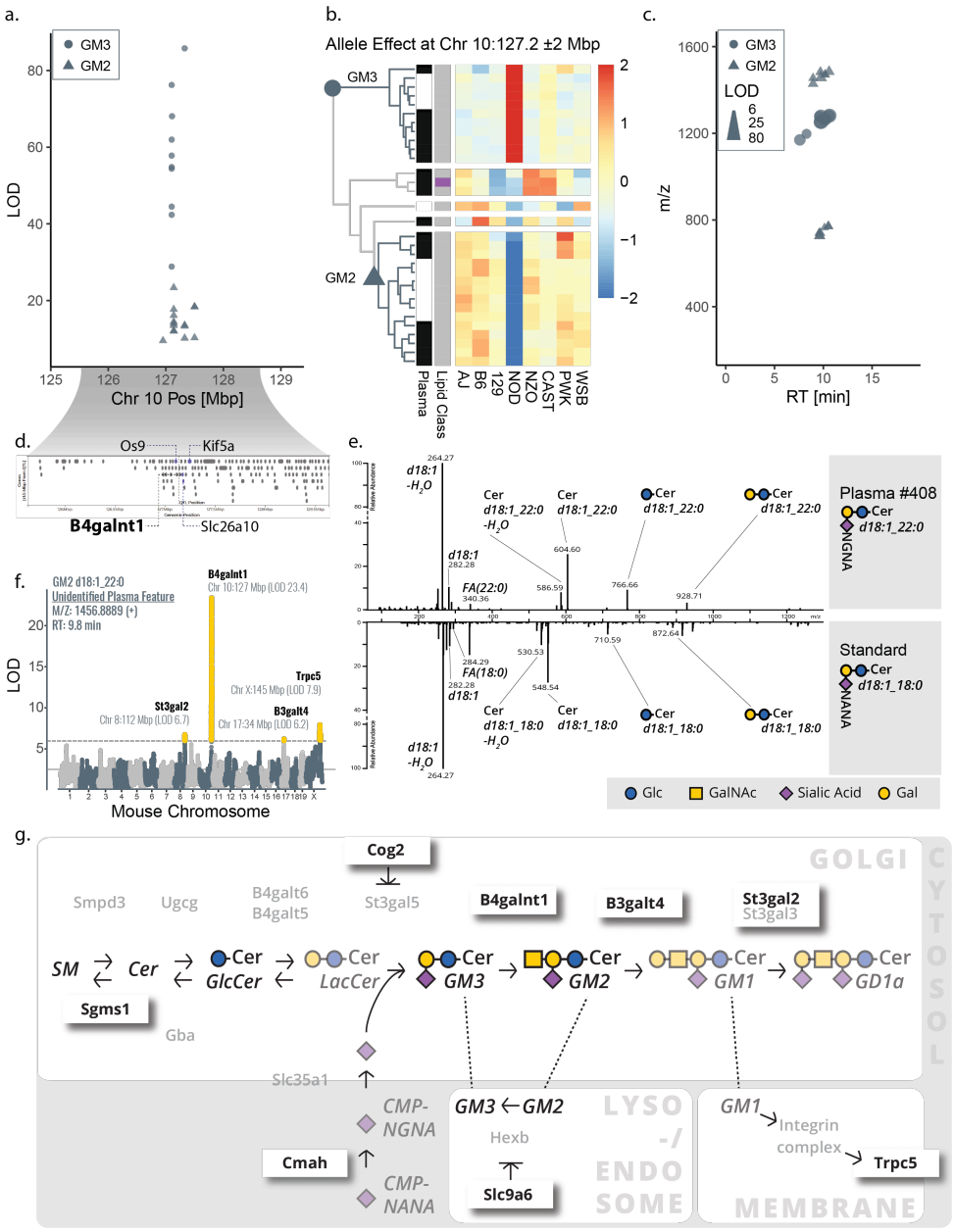


Figure 5.4: Lipid Features Mapping to B4galnt1 Lead to Identification of GM2 and GM3 Gangliosides. a, A hotspot of solely unidentified features with exceptionally strong correlation was composed of 11 liver and 15 plasma features mapping to chromosome 10:127 Mbp with **b**, a similar NOD-driver allele pattern (two main clusters from hierarchical clustering, row-scaled, Euclidean cutoff of h = 2.5). Two groups of lipid features (circles vs. triangles) emerged as distinct in strength of LOD (**a**), directionality of allele effect (**b**), and m/z space (**c**). **d**, The candidate gene B4galnt1 pointed us to the putative identifications of GM3 (circles) and GM2 (triangles) gangliosides, which were confirmed by **e**, spectral matching with a human GM3 standard. **f**, Secondary QTL for these gangliosides, as exemplary shown for GM2 d18 : $1_2$2 : 0, mapped to eight additional candidate genes within 4 Mbp of the 15 total ganglioside hotpots that were previously linked to ganglioside metabolism. **g**, The various candidate genes influencing GM3 and GM2 levels span well-known enzymes (*e.g.* B3galt4) but also include indirect affectors including Cog2 and Slc9a6.

The genetic effects of SNPs and other genomic variants can influence lipid abundance. For example, SNPs in coding regions can affect protein product function. In the extreme, a missense variant in proteins involved in lipid metabolism could very likely affect lipid abundance. SNPs in non-coding regions, such as promoters and enhancers, can alter gene expression. To identify candidate SNPs, we analyzed the SNPs associated with the lipids by identifying those with the founder strain SNP database at each QTL (R/qtl2; scan1snps()[29]) and subsequently causing missense, frameshift, stop lost/gained, incomplete terminal codons, in-frame deletions/insertions, altering 3′ or 5′ UTR sequences, splice acceptor/donor/region, predicated to cause nonsense-mediated decay, initiator codon or mature miRNA variants (according to the Sequence Ontology (SO) consortium)[49]. At the chromosome 10 hotspot we identified several candidate genes with potentially causal mutations (**Figure 5.4d**). We included in our analysis, but did not focus on genes with synonymous, stop retained, up-/downstream, intergenic, intron and non-coding transcript (exon) variants (which represent 97% of all SNPs in the database).

In cases of altered gene expression, we further narrowed down the list of candidates by directly assessing transcriptomics data. While we did not profile hepatic gene expression in the DO cohort used for lipid QTL analysis, we surveyed a recently published hepatic QTL data set to match allele effects of mRNA expression and protein QTL that are within the location of the candidate gene (cis-eQTL and pQTL, respectively).[36] We asked if any transcripts or proteins presented a similar NOD-driven allele effect at the lipid locus on chromosome 10. Of the protein coding genes within $\pm$2 Mbp of the lipid QTL, 55 showed a

cis-eQTL. However, the only cis-eQTL that was strongly and uniquely driven by NOD alleles was *B4galnt1* (**Supplementary Figure S5.4a**). Furthermore, 16 cis-pQTL were identified for genes within this region, including B4GALNT1. Similar to the cis-eQTL, the only pQTL that showed an NOD-driven allele effect pattern was for B4GALNT1 (**Supplementary Figure S5.4b**). Consequently, the 3' UTR variant in *B4galnt1* SNP rs13462597 was our strongest candidate as the genetic regulation of hepatic *B4galnt1* transcript and protein expression matches that of the unidentified lipids.

B4galnt1 encodes for β-1,4 N-acetylgalactosaminyltransferase 1, an enzyme that catalyzes the conversion of GM3 to GM2 gangliosides.[50] With this candidate gene in mind, we investigated whether the unidentified lipids could be classified as gangliosides. Their precursor m/z and tandem mass spectra were consistent with monosialic gangliosides, which we further confirmed by comparison with a GM3 ganglioside standard (**Figure 5.4e**). In total, we confidently identified 26 lipid features as six unique GM2 and seven unique GM3 species (Supplementary Table S4). Consistent with an NOD-driven effect, NOD mice have higher abundance of GM3 gangliosides in pancreas,[51] and we confirmed NOD had higher abundance of GM3 in plasma in an independent lipidomic analysis of founder strain mice (**Supplementary Figure S5.4c**).

By identifying the features mapping to chromosome 10:127 Mbp as gangliosides, we recognized that ganglioside abundances, like the levels of most lipid species, were polygenic, that is regulated by multiple loci (**Figure 5.4f**). From the 26 identified ganglioside features we gain a total of 62 QTL annotations, describing more than 15 unique loci (at least two gan-

glioside features with LOD > 6.0) on 10 chromosomes. Interestingly, these newly annotated ganglioside QTL mapped to candidate genes of the ganglioside pathway (*Sgms1*[52], *B3galt4*, *St3gal2*, *Cmah*[53]), even more distant regulators of ganglioside metabolism (*Slc9a6*[54], *Cog2*[55], *Trcp5*[56], *Cdh13*), and regions of the genome with yet undescribed ganglioside regulation (**Figure 5.4g**).

LipidGenie: A resource to identify candidate genetic regulators for lipid features. Up to this point, the interrogation of QTLs in this study had been conducted on a piecemeal basis via scripting with the R/QTL2[29] package. There existed no software tooled specifically for the facile examination of large-scale QTL mapping studies. To facilitate the exploration of this high-dimensional dataset and to make the described genome-lipid associations accessible to the scientific community we created a web-based resource: LipidGenie (`http://lipidgenie.com`). LipidGenie aggregregates the lipid identifications generated in this study using Lipidex,[17] known mouse genes as annotated by the Mouse Genome Informatics (MGI) consortium,[57], and QTL mapping outputs from the R/QTL2 package into a centralized web-based application.

QTLs can be queried from either a mass spectrometry-based lipidomics or genetic locus perspective 5.5. When LipidGenie's Lipid Viewer is accessed, all 3,295 lipid features profiled in this study are loaded into the web page. These features can be subset using lipid precursor m/z, tissue type, or lipid class as filtering properties (**Figure 5.5A**). Once a lipid feature is selected, all QTLs associated with this feature populate the QTL dropdown.

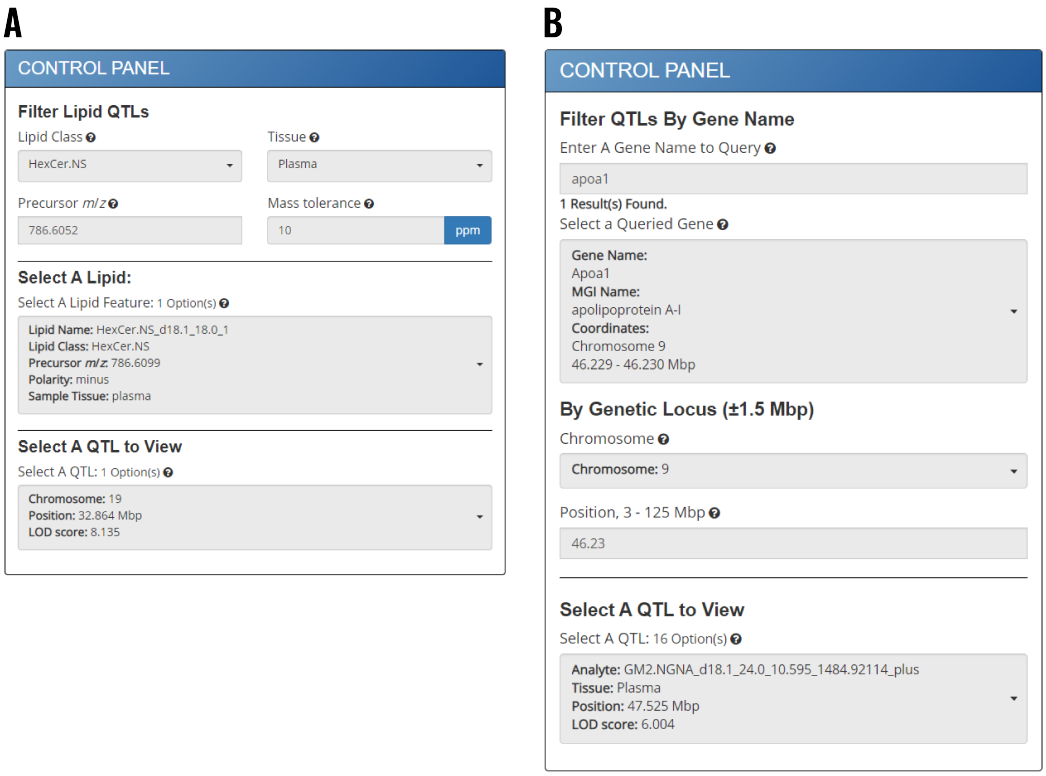


Figure 5.5: Control Panels for LipidGenie's Lipid and Genetic Locus QTL Viewers. Panel A shows the QTL query controls for the lipid-driven QTL viewer. A researcher can filter all the identified and unknown lipid features in this study by lipid class, sample type, and theoretical lipid m/z. Selecting a lipid species will return all QTLs associated with the lipid feature. These QTLs can then be individually inspected. **Panel B** shows the QTL query controls for the genetic locus QTL viewer. The name of a mouse or human gene of interest can be entered into the first textbox, populating the second dropdown with genes approximately matching the entered gene. Selecting a gene will automatically populate the genetic locus and chromosome fields. Alternatively, a chromosome and genetic locus can be manually entered. All QTLs within ± 1.5 Mbp of the specified locus will populate the final QTL dropdown. Selecting a QTL in both control panels will asynchronously query the data necessary to generate the visualizations shown in **Figure 5.6**.

Using the Lipid Viewer, researchers can quickly compare the LOD and Allele Effect plots of multiple lipid species. As described above, we were able to recover identifications for a set of co-mapping gangliosides (**Figure 5.4**). LipidGenie quickly uncovered that these previously unknown lipid features had similar LOD and Allele Effect plots on chromosome ten, suggesting they may be involved in a biological process encoded at the QTL locus.

LipidGenie's Gene Viewer (**Figure 5.5B**) enables QTL lookups through specifying a genetic locus (*i.e.* a chromosome and genetic location). Specific genes can be queried by entering a gene name into the first text box. To expand the purview of LipidGenie to human biology, we mapped homologous human genes to our dataset using shared MGI identifiers.[58] Up to 200 potential gene hits are returned using partial string matching on MGI-annotated mouse genes or human homologues if they exist. Selecting a gene will autopopulate the Genetic Locus fields chromosome and position. Alternatively, these coordinates can be manually specified if. QTLs within $\pm$1.5 Mbp of the provided locus are inserted into the QTL dropdown at the bottom of the control panel.

When a QTL is selected using either the lipid or genetic locus control panel, an asynchronous server query is executed to retrieve the respective QTL's LOD and Allele Effect values across the QTL's respective chromosome alongside all MGI genes near the QTL locus. The returned data is used to generate the visualizations shown in **Figure 5.6**. The textbfLOD plot represents the additive effects of DO mice genetics on the expression of a particular lipid species. The apex LOD score is chosen in this study as the QTL locus. Larger LOD scores are associated with strong correlations of particular alleles to feature

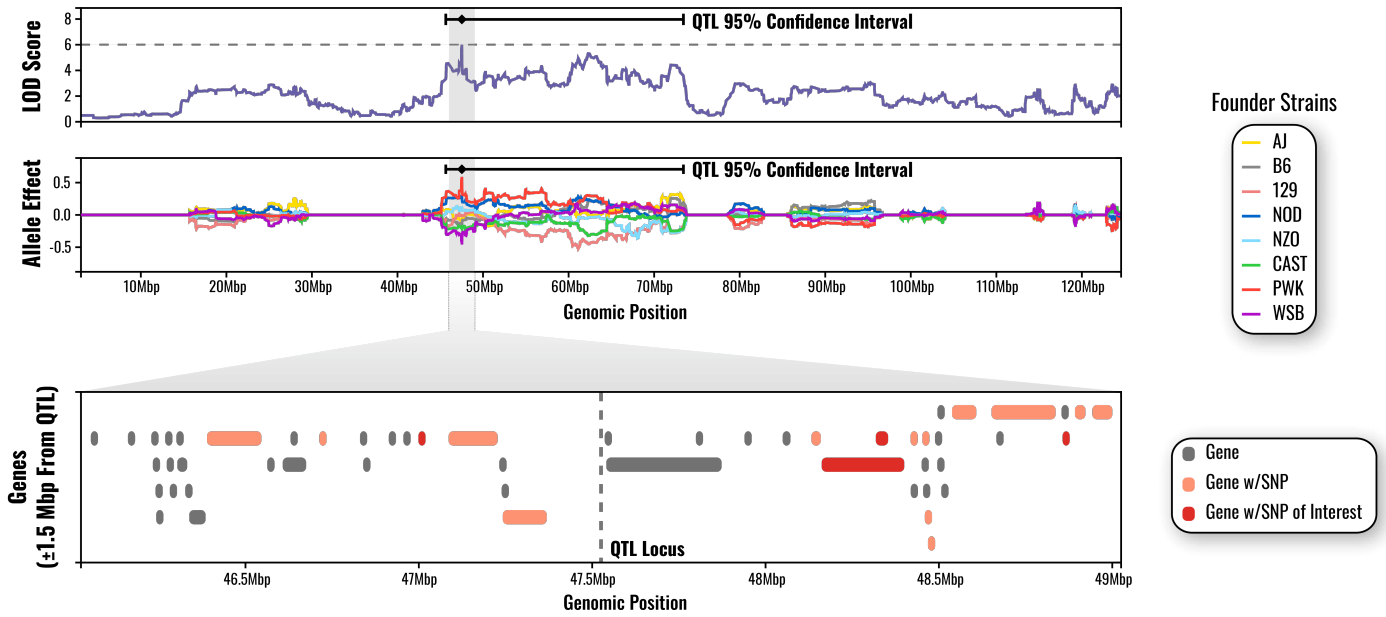


Figure 5.6: LipidGenie Visualizations's: LOD plots, Allele Effect Plots, and Candidate Genes. LOD Plot, The calculated LOD score is drawn across a chromosome for a selected QTL. The apex LOD score on a particular chromosome is chosen to be the QTL's locus. A 95% confidence interval for the QTL is indicated using shown the whisker plot. The QTL locus is marked using the diamond on the whisker plot. **Allele Effect Plots** highlight the contribution of the alleles of each Founder Strain to the observed LOD curve. The **Gene Plot** visualizes all known MGI genes within a ± 1.5 Mbp window of the QTL locus. Genes which contain at least one SNP are highlighted in orange. Genes which contain SNPs that have the potential to influence lipid abundance are highlighted in red.

abundance. The textbfallele effect plot highlights the contribution of all Founder Strain's alleles to the calculated QTL effects shown in the LOD plot. The biggest divergence between the allele effect curves co-occurs with the QTL locus. Finally, the **gene plot** visualizes all MGI-annoted genes within $\pm$1.5 Mbp of the apex LOD score. Genes that contain an annotated MGI SNP are highlighted in salmon, while genes which contain SNPs that have a likely phenotypic impact are highlighted in red. Finally, after a QTL is selected, the data above as well as a list of genes contained with the QTL 95% confidence interval can be downloaded in a tabular format if desired.

Validating LipidGenie using the DO mouse dataset To validate LipidGenie we explored sex-associated lipid features that were observed within the B6 founder strain. In this study we quantified 2,558 lipid features in B6 plasma and found 254 features that showed significantly different levels by sex (**Figure 5.7a**). As is common in LC-MS lipidomics, most of these sex-specific features were unidentified after the database search (n=197). Utilizing LipidGenie's m/z search parameter and a 10 ppm m/z window, we found significant genome-lipid associations for 127 of the sex-specific features, of which 79 were unidentified. Strikingly, a group of six unidentified lipids mapped to the same genetic locus on chromosome 6 at 91 Mbp (**Figure 5.7b**); all had similar allele effect patterns (**Figure S5.5a**) and were elevated in males (**Figure 5.7a** and **Figure S5.5b**). At the locus, a total of 12 out of 21 co-mapping features shared a lipid class-like behavior, i.e., clustered in m/z-RT space (**Figure 5.7c**). To further characterize these lipids, we collected additional tandem

mass spectra in both positive and negative mode (**Figure 5.7d-g**).

The spectra showed shared fragmentation patterns consistent with a phosphatidylcholine (PC) class identity. Strikingly, one fatty acid seemed to be either FA 22:6 (**Figure 5.7e**) or FA 16:0, but only MS3 spectra showed the presence of a second acyl chain expected for PCs (**Figure 5.7f-g**). These features also shared a m/z 522 fragment that matched the formula of lyso-PC 19:0, C26H53NO7P-. These fragmentation patterns suggest a third acyl chain, and this would be in accordance with the observed 4 min increase in RT and 300 Da increase in precursor m/z compared to typical PCs.

We next leveraged the LipidGenie associations to generate hypotheses about the nature of this lipid class. At chromosome 6 at 91 Mbp, we found SNPs with matching allele effects in several genes including *Txnrd3*, *Vmn1r*, *Uroc1*, *Aldh1l1*, *Slc41a3*, *Grip2*, and *Trh* (**Figure 5.7b**). One possible candidate on chromosome 6 is Vmn1r, encoding for vomeronasal receptors, the organs that sense pheromones. Not only could this gene explain the observed sex difference, it also points us to PC estolides as a potential class identity. Estolides are lipids containing fatty acid esters of hydroxy fatty acids (FAHFAs). Consistent with the observed 16:0 or 22:6 fragments in MS2 spectra of the unidentified lipids, 16:0 and 22:6 can be esterified to hydroxy fatty acids to form FAHFAs.[60] This hypothesis is further supported by accounts of FAHFAs as pheromones in spiders and TG estolides in mammalian scent glands.[61] The potential estolide identity is intriguing, but definitive identification will require follow-up studies. Further evidence is likely contained in the genetic associations. Similar to our earlier example with gangliosides, we observed co-mapping of these 12

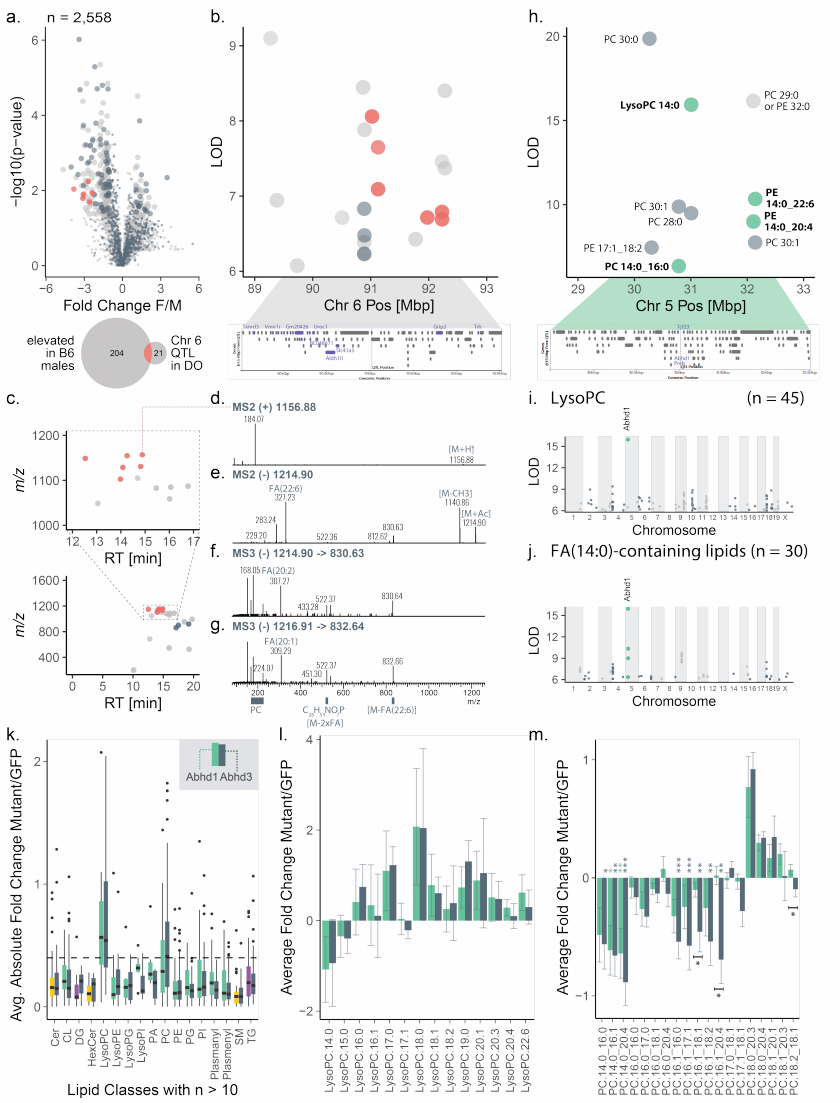


Figure 5.7: Web Resource LipidGenie Guides Exploration of Genome-lipid Connections. a, In a lipidomics experiment of B6 mouse plasma (n = 4 for each sex), we quantified 2,558 features. We found 254 features to be sex-specific (FC > 1.0, $p < 0.05$, non-paired, two-sided Student's t-test) of which $\frac{2}{3}$ could not initially be identified. Matching of their precursor m/z (±10 ppm) to our DO database provided genetic information for $\frac{1}{3}$ of the otherwise unidentified features. **b**, Highlighted in red is a group of six male-specific unidentified features that share a QTL on Chr 6:91 Mbp with a common A/J down allele effect (**Supplementary Figure S5.5a**). We identified several candidate genes within 4 Mbp that have matching SNP variants, including *Txnrd3*, *Vmn1r(43-45,49)*, *Aldh1l1*, *Slc41a3*, *Grip2*, and *Trh*. **c**, The six unidentified features further clustered in m/z-RT space, suggesting they were members of a lipid class. **d-g**, This grouping of features allowed us to acquire targeted fragmentation spectra (exemplary spectra for two of these species ([M+H]+ m/z 1156 and 1158) in positive (MS2) and negative (MS2 and MS3) mode). The exhibited signals were consistent with a lipid class built of a PC headgroup and three fatty acids. **h**, LipidGenie further informed a recent finding of LysoPC 14:0 mapping to *Abhd1*.[59] **j**, Importantly, we observed an enrichment of FA 14:0 containing lipids but not of LysoPCs (i), pointing to a putative function of ABHD1 similar to the phospholipase ABHD3. **k**, To examine this hypothesis, we overexpressed both *Abhd1* and *Abhd3* in Hepa1-6 cells and compared their lipidomes to a control overexpressing GFP (n = 12 for each, 4 biological replicates x 3 distinct technical replicates). We observed the largest absolute FC for LysoPC and PC lipids, as visible in boxplots of average absolute FC of each mutant over GFP by lipid class (of lipid classes with more than 10 identified members). The dashed line represents an arbitrary cut-off of a FC of 0.4, that both LysoPC and PC lipids surpassed. **l**, When plotting the average FC per LysoPC species, a FA dependency is observed. The max. negative FC for both *Abhd1* and *Abhd3* is observed for LysoPC 14:0, while the max.positive FC is observed for LysoPC 18:0. Note that each species contains the summed values of all isomers of the respective LysoPC. m, A similar pattern is observed for PC species identified on the FA level. All 14:0 containing PCs exhibit a negative FC for Abhd1 and Abhd3 mutants consistently, while 18:0 containing species are showing positive FC. Plotted are sum-normalized, log2-transformed FC. Error bars in **l** and **m** represent 95% confidence interval, significance indicated by * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$) of non-paired Student's t-test, equal variance, n = 12 for each.

lipids at other loci (e. g., Chr 10, at 84 Mbp, Chr 12, at 84 Mbp), thereby offering potential pathway information and highlighting the power of genome-lipid associations obtained with LipidGenie.

We next explored whether LipidGenie would also offer novel insights when querying for identified lipid features. Recently, Parker et al. found an association between lysoPC 14:0 and chromosome 5 at 31 Mbp using multi-omic QTL mapping of the hybrid mouse diversity panel[59]. From these data, they postulated that the protein encoded by candidate gene *Abhd1* (alpha beta hydrolase domain containing 1) regulates plasma levels of LysoPCs. Note, ABHD1 has no annotated function. LipidGenie's lipid search provides a direct means to test this putative functional annotation of ABHD1.

LipidGenie confirmed that plasma LysoPC 14:0 has a strong QTL at the abhd1 locus (**Figure 5.7h**), and further found the B6 and NZO high allele effect consistent with the 3' UTR variant rs29681817 (**Figure S5.5c**). This observation is further supported by an independent measure of the founder strain mice and a hepatic cis-eQTL in *Abhd1* with matching opposite allele effects (**Figure S5.5d-e**). To connect the function of ABHD1 protein to LysoPCs, we asked whether other LysoPCs (n = 45) mapped to this gene region. However, in contrast to Parker and co-workers' postulation, we did not find general mapping of LysoPCs to this locus (**Figure 5.7i**). Instead we found other lipids co-mapping on chromosome 5, at 31 Mbp, including PC 14:0_16:0, PE 14:0_20:4, PE 14:0_22:6, PC 28:0, PC 30:0, and PC 30:1 (**Figure 5.7h**). These fatty acid signatures suggest a myristic acid (14:0) specific association. Given the high degree of lipid structural resolution contained

within LipidGenie, we demonstrate that 14:0-containing lipids (n = 30) have an enriched hotspot at the *Abhd1* locus (**Figure 5.7j**). With these data we propose that ABHD1 is a phospholipase for myristic acid containing phospholipids; consistent with the function of a related and highly homologous gene, *abhd3*.[62,63] 14:0-containing phospholipids have also been mapped to ABHD3 in human GWAS.[64] To validate this hypothesis, we overexpressed ABHD1 and ABHD3 in Hepa1-6 cells (**Supplementary Figure S5.6a-b**) and measured their lipidome with respect to cells overexpressing GFP as control. Hierarchical clustering of the top 49 features showed two clusters, one with increased levels in the mutants over control and the other one decreased (**Supplementary Figure S5.6c**). We noticed a majority of identified lipids among the most significantly different features, and when plotting the average fold change by lipid class, LysoPC and PC phospholipids stood out (**Figure 5.7k**). Upon closer look, we could confirm the predicted fatty acid dependency for both LysoPC and PC lipids, particularly prominent in 14:0 containing phospholipids (**Figure 5.7l-m**). While ABHD1 and ABHD3 mutants exhibited largely similar lipidomic profiles, differences as in PC 16:1_20:4 that was only decreased in the *Abhd3* mutant, may also point to differential functions. The 14:0 specificity could be relevant to human health as plasma LysoPC 14:0 is a predictor of diabetes risk in humans. Finally, our proposed function might provide a clue to understanding why ABHD1 is associated with oxidative stress, a prominent hallmark of metabolic diseases.[62,65,66]

Having documented the diverse utility of LipidGenie for lipid queries, we lastly tested its use for gene-based queries. ABHD2, another member of the alpha beta hydrolase do-

main protein family, acts on arachidonylglycerol, among other substrates.[67,68] A LipidGenie query of Abhd2 does indeed provide evidence for this polyunsaturated fatty acid pathway specificity. Specifically, within ±2 Mbp of Abhd2, LipiGenie returned ten liver phospholipids. Eight of these lipids shared an allele effect pattern, and contained poly-unsaturated fatty acids - i.e., 18:2, 18:3, 20:3, 20:4, and 22:6 (**Figure S5.5f-g**). Further, ABHD2 showed matching opposite WSB and ST effects in both liver cis-eQTL and pQTL (**Figure S5.5h-i**).[36]

Discussion

Discovery lipidomics presently relies on measurement of various chemical properties for lipid identification. These properties are most often hydrophobicity, mass, and fragmentation pattern. Unfortunately, application of only these strategies to complex mammalian lipid mixtures results in many unidentified lipid features. Here we proposed that genome-lipid associations can facilitate lipid identification.

To test the power of genome-lipid associations for lipid identification, we performed QTL analysis for over 5,000 plasma and liver lipid QTL, of which over 60% stem from unidentified spectral features to construct a large scale mapping of QTL across the features. To our knowledge, this QTL map is the broadest in scope and depth of lipids analyzed and QTL identified; but more importantly, it is the first to map unidentified spectral features to genomic loci.[59,69,70] With these data, we first tested our hypothesis - that such associations could facilitate lipid identification - by analyzing one of many QTL hotspots; the Apoa2 locus. The identified lipids mapping to this locus belonged to 11 different classes and,

together with APOA2, constitute the known components of HDL particles. With this association, 23 unidentified lipid features could be classified as cholesteryl esters and related features.

To further test the concept, we selected a second hotspot containing only unidentified lipid features (10:127 Mbp). Genetic mapping to *B4galnt1* enabled their identification as GM3 and GM2 gangliosides. In fact, the identification allowed for a comprehensive investigation of their complex polygenic regulation. We identified a total of eight candidate genes that likely contribute different functions in the pathway, including three (*Slc9a6*, *Cog2*, *Trpc5*) that exert indirect effects on ganglioside biosynthetic enzymes.

Having confirmed the value of genome-lipid associations for lipid mass spectral data annotation, we built an interactive, query-able resource - LipidGenie. Using the lipid query function, we demonstrated LipidGenie's ability to facilitate lipid identification and in one instance revealed a potentially new sub-class of PC lipids (PC-estolides). Beyond assisting lipid identification, LipidGenie can provide evidence for gene function, and when queried for either lipid ID or gene ID, LipidGenie revealed acyl-chain specificity for ABHD1 and ABHD2, respectively. We confirmed the putative phospholipase function of ABHD1 in cells overexpressing the mouse protein while comparing to ABHD3.

We envision the genome-lipid associations contained within LipidGenie to be a valuable resource for researchers across multiple fields. We anticipate it will be immediately useful for directed analysis of key unidentified features in exploratory lipidomics analyses and lead to recovery of more data for biological studies. Simultaneously, we hope it will garner

excitement for potentially novel genetic regulation of lipid metabolism. While LipidGenie has specifically been built as a resource for this study, it's architecture is poised to be repurposed for the QTL mapping of other mass spectrometry-based omics measurements, such as proteomics and metabolomics. Finally, through integration with other large data resources, e.g., protein-protein interactions, pathway tools, tissue-specific QTL, etc., these genome-lipid associations will allow more global integration of lipid data into current knowledge bases. The integration with human loci will especially allow for cross-validation to inform human health and disease.[71,72]

Methods

Animal Husbandry and Sample Collection. All experiments involving mice were preapproved by an AAALAC-accredited Institutional Animal Care and Use Committee of the College of Agricultural Life Sciences (CALS) at the University of Wisconsin-Madison. The CALS Animal Care and Use Protocol number associated with the study is A005821, A.D. Attie, Principal Investigator. Equal numbers of male and female Diversity Outbred (DO) mice and the eight founder strains (C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), PWK/PhJ (PWK), WSB/EiJ (WSB), and CAST/EiJ (CAST)) were all obtained from the Jackson Labs and have been previously described.[34,35,73] Briefly, all mice were housed within the vivarium at the Biochemistry Department, University of Wisconsin-Madison, and maintained on a Western-style high-fat/high-sucrose (HF/HS) diet (44.6% kcal fat, 34% carbohydrate and 17.3% protein) from Envigo Teklad (TD.08811)

for 16 weeks. All mice were maintained in a temperature and humidity-controlled room on a 12 hr light/dark cycle (lights on at 6AM and off at 6PM), and provided water ad libitum. At 22 weeks of age, mice were sacrificed following a 4 hr fast. Plasma and liver were collected from each mouse, flash frozen in liquid nitrogen. One sample from each tissue per mouse was used for lipidomic analyses.

Mouse Genotyping and Haplotype Reconstruction. We collected tail biopsies for DNA extraction[30] at 4 to 6 weeks of age when animals arrived at the University of Wisconsin and were assigned to single-housed pens. We shipped DNA to Neogen (Lincoln, NE) for genotyping using the Mouse Universal Genotyping Array (GigaMUGA; 143,259 markers). Genotype calls were subject to quality control as described in Broman et al.[74] Genotypes were used to reconstruct the 8-founder haplotype mosaic of each DO mouse using the hidden Markov model in the R/qtl2 software package.[29,34] The haplotype-reconstruction uses information at each genetic markers and its neighbors to assign an eight-state haplotype probability that accounts for both heterozygosity and uncertainty in haplotype assignments.[28] We interpolated the founder haplotype probabilities onto an evenly spaced grid of 69,005 pseudo-markers for mapping analysis. Sample mix-ups (one pair of samples) were resolved using islet gene expression data as described in Keller et al. 2018.[34]

Plasmids and Cell Culture Expression. Mouse Abhd1 (CMV6 promoter, Myc-DDK-tagged, MR206471) and mouse Abhd3 (CMV6 promoter, Myc-DDK-tagged, MR206458) plasmids

were obtained from Origene. Manufacturer's sequencing primers were used to confirm plasmid insert. His-tagged CMV6-GFP plasmid was a gift from J. Simcox. All plasmids were transformed into E. coli (ThermoFisher Scientific, 18258012). Plasmids were maxiprepped according to manufacturer's instructions (Qiagen, 12362).

5x105 Hepa1-6 cells (ATCC® CRL-1830) were seeded in 6-well plates with DMEM (ThermoFisher Scientific, 12100061). After 16 hours, cells were reconditioned with fresh media for 2 hours. Cells were transfected in triplicate with Lipofectamine2000 (ThermoFisher Scientific, 11668019) according to manufacturer's instructions. Transfection efficiency was confirmed by visualizing GFP. After 24 hours, media was replaced. 48 hours after transfection, cells were washed in cold 1X PBS and scraped to be released from the plate. Released cells were pelleted by centrifugation and snap-frozen in liquid nitrogen. The frozen cell pellets were stored at -80 °C until lysis. Hepa1-6 cells were provided by J. Simcox.

For Western Blots, cell pellets were lysed in 2x SDS-PAGE loading buffer and boiled at 95C for 5 min. Samples were run on a 10% SDS-PAGE gel for 1.5 h at 120V, standard is Precision Plus Dual Color Protein Standards (Bio-Rad, 1610394). Samples were wet-transferred onto PVDF membrane (Bio-Rad, 1620177) for 1.5 h at 100 V. Following transfer, membrane was blocked in 5% milk in TBST for 1 h at room temperature. Membrane was incubated overnight at 4C with 1:2000 rabbit anti-MYC antibody (CST, 2278) in blocking buffer. Primary antibody was removed by washing 3X with 1X TBST. Membrane was incubated with 1:2000 goat anti-rabbit-HRP conjugated antibody (CST, 7074S) in blocking buffer. Samples were visualized with Clarity Western ECL Substrate (Bio-Rad, 1705060) on

a ThermoFisher iBright FL1500 Imaging System.

Lipidomics Sample Preparation. Plasma. 40 μL (30 μL for founder strains, FS) of plasma and 10 μL SPLASH Lipidomix internal standard mixture (Avanti Polar Lipids, Inc.) were aliquoted into a tube. Protein was precipitated by addition of 215 μL MeOH. Control samples comprised an aliquot of mixed male and female B6 plasma (Chow diet), extracted with each batch. After the mixture was vortexed for 10 s, 750 μL methyl tert-butyl ether (MTBE) were added as extraction solvent and the mixture was vortexed for 10 s and mixed on an orbital shaker for 6 min. Phase separation was induced by adding 187.5 μL of water followed by 20 s of vortexing. All steps were performed at 4 °C on ice. Finally, the mixture was centrifuged for 4 min at 14,000 x g at 4 °C and 150 μL of the lipophilic upper layer were transferred to glass vials and dried by vacuum centrifuge for 60 min. The dried extracts were re-suspended in 100 μL MeOH/Toluene (9:1, v/v).

Liver. 20 (±2) mg liver tissue, frozen in liquid nitrogen along with 20 μL SPLASH Lipidomix internal standard mixture were aliquoted into a tube with a metal bead and 1150 μL of MTBE/MeOH (10:3, v/v) were added for protein precipitation. Control samples for DO comprised aliquots of sample pooled from FS, extracted with each batch. All steps were performed at 4 °C on ice. The mixture was homogenized by bead beating for 4 min at 25 Hz and shaking on an orbital shaker for 6 min. After bead removal, 225 μL of water were added to each tube and the mixture was vortexed for 20 s. Finally, the mixture was centrifuged for 20 min at 13,000 x g at 4 °C after which 200 μL of the lipophilic upper layer were transferred

to glass vials and dried by vacuum centrifuge for 60 min. The dried lipophilic extracts were re-suspended in 100 µL MeOH/Toluene (9:1, v/v).

Hepa1-6 Cells. Hepa1-6 cells were scraped off of six well plates and transferred to 1.5 ml Eppendorf tubes. Cell pellets were kept frozen (less than -20 °C) until extraction. Cells were lysed and protein was precipitated by addition of 225 µL MeOH. 750 µL methyl tert-butyl ether (MTBE) were added as extraction solvent. The mixture was homogenized by vortexing for 10 s and shaking on an orbital shaker for 6 min. Phase separation was induced by adding 187.5 µL of water followed by 20 s of vortexing. All steps were performed at 4 °C on ice. Finally, the mixture was centrifuged for 8 min at 14,000 x g at 4 °C and 200 µL of the lipophilic upper layer were transferred to glass vials and dried by vacuum centrifuge for 60+ min. The dried extracts were re-suspended in 100 µL MeOH/Toluene (9:1, v/v).

LC-MS/MS. Sample analysis by LC-MS/MS, running data-dependent acquisition (DDA) with dynamic exclusion and polarity switching, was performed in randomized order on an Acquity CSH C18 column held at 50 °C (2.1 mm x 100 mm x 1.7 µm particle diameter; Waters) using an Ultimate 3000 RSLC Binary Pump (400 µL/min flow rate; Thermo Scientific) for plasma, while for the liver samples a Vanquish Binary Pump (400 µL/min flow rate; Thermo Scientific) was used. Mobile phase A consisted of 10 mM ammonium acetate in ACN/H2O (70:30, v/v) containing 250 µL/L acetic acid. Mobile phase B consisted of 10 mM ammonium acetate in IPA/ACN (90:10, v/v) with the same additives. Mobile phase B was initially held at 2% for 2 min and then increased to 30% over 3 min. Mobile phase B was further

increased to 50% over 1 min and 85% over 14 min and then raised to 95% over 1 min and held for 7 min. The column was re-equilibrated for 2 min before the next injection.

Plasma: Ten microliters of lipid extract were injected by an Ultimate 3000 RSLC autosampler (Thermo Scientific). The LC system was coupled to a Q Exactive Focus mass spectrometer by a HESI II heated ESI source kept at 300 °C (Thermo Scientific). The inlet capillary was kept at 300 °C, sheath gas was set to 25 units, auxiliary gas to 10 units, and the spray voltage was set to 5,000 V (+) and 4,000 V (-), respectively. The MS was operated in polarity switching mode acquiring positive and negative mode MS1 and MS2 spectra (Top2) during the same separation. MS acquisition parameters were 17,500 resolving power, 1 $\times 10^6$ automatic gain control (AGC) target for MS1 and 1 $\times 10^5$ AGC target for MS2 scans, 100-ms MS1 and 50-ms MS2 ion accumulation time, 200- to 1,600-Th MS1 and 200- to 2,000-Th MS2 scan range, 1-Th isolation width for fragmentation, stepped HCD collision energy (20, 30, 40 units), 1.0% under fill ratio, and 10 second dynamic exclusion.

Liver: One microliter of lipid extract was injected by a Vanquish Split Sampler HT autosampler (Thermo Scientific). The LC system was coupled to a Q Exactive HF mass spectrometer by a HESI II heated ESI source kept at 300 °C (Thermo Scientific). The inlet capillary was kept at 300 °C, sheath gas was set to 25 units, auxiliary gas to 10 units, and the spray voltage was set to 4,000 V (+) and 3,500 V (-), respectively. The MS was operated in polarity switching dd-MS2 mode acquiring positive and negative mode MS1 and MS2 spectra (Top2 for positive, Top3 for negative mode) during the same separation. MS acquisition parameters were 60,000 resolution and 3 $\times 10^6$ automatic gain control (AGC)

target for MS1 and 15,000 resolution and 5 $\times 10^5$ AGC target for MS2 scans, 100-ms MS1 and 35-ms MS2 ion accumulation time, 240 to 1,200-Th MS1 scan range for positive and to 1,600-Th for negative mode, and 200- to 2,000-Th MS2 scan range, 1.4-Th isolation width for fragmentation, stepped HCD collision energy (20, 25 units for positive, 20,30 units for negative mode), and 10 second dynamic exclusion.

Hepa1-6 Cells: Ten microliters of lipid extract were injected by a Vanquish Split Sampler HT autosampler (Thermo Scientific). The LC system was coupled to a Q Exactive HF mass spectrometer by a HESI II heated ESI source kept at 300 °C (Thermo Scientific). The inlet capillary was kept at 300 °C, sheath gas was set to 25 units, auxiliary gas to 10 units, and the spray voltage was set to 4,000 V (+) and 3,500 V (-), respectively. The MS was operated in polarity switching mode acquiring positive and negative mode MS1 and MS2 spectra (Top2) during the same separation. MS acquisition parameters were 30,000 resolving power, 1 × 106 automatic gain control (AGC) target for MS1 and 1 × 105 AGC target for MS2 scans, 100-ms MS1 and 50-ms MS2 ion accumulation time, 200- to 1,600-Th MS1 scan range, 1-Th isolation width for fragmentation, stepped HCD collision energy (20, 30, 40 units), 1.0% under fill ratio, and 10 second dynamic exclusion.

Lipidomics Data Analysis. The resulting LC-MS lipidomics raw files were converted to mgf files via MSConvertGUI (ProteoWizard, Dr. Parag Mallick, Stanford University)[75] and processed using Compound Discoverer 2.0 (Thermo Fisher Scientific) and an in-house developed open-source software suite, LipiDex[16]. All raw files were loaded into Compound

Discoverer with blanks marked as such to generate two result files using the following Workflow Processing Nodes: Input Files, Select Spectra, Align Retention Times, Detect Unknown Compounds, Group Unknown Compounds, Fill Gaps and Mark Background Compounds for the so called "Aligned" result and solely Input Files, Select Spectra, and Detect Unknown Compounds for an "Unaligned" Result. Under Select Spectra, the retention time limits were set between 0.4 and 21 min, MS order as well as unrecognized MS order replacements were set to MS1. Under Align Retention Times the mass tolerance was set to 10 ppm and the maximum shift according to the dataset to 0.5 min. Under Detect Unknown Compounds, the mass tolerance was also set to 10 ppm, with an S/N threshold of 3, and a minimum peak intensity of 5E5 (DO) or 1E5 (FS). Further, [M+H]+1 and [M-H]-1 were selected as ions and a maximum peak width of 0.75 min as well as a minimum number of scans per peak equaling 5 were set. Lastly, for Group Unknown Compounds as well as Fill Gaps, mass tolerance was set to 10 ppm and retention time tolerance to 0.2 minutes. For best compound selection rules #1 and #2 were set to unspecified, while MS1 was selected for preferred MS order and [M+H]+1 as the preferred ion. For everything else, the default settings were used. Resulting peak tables were exported as excel files in three levels of Compounds, Compound per File and Features (just Features for the "Unaligned") and later saved as csvs. In LipiDex' Spectrum Searcher "LipiDex_HCD_Acetate", "LipiDex_HCD_Plants", "LipiDex_Splash_ISTD_Acetate", "LipiDex_HCD_ULCFA", and "Ganglioside_20171205" were selected as libraries for the DO while "LipidBlast2_Reformatted_CoonLab", "LB_cleaned" and "Lipid_Spectral_Library_20170523" were selected for the FS. We further kept the de-

faults of 0.01-Th for MS1 and MS2 search tolerances, a maximum of 1 returned search result, and an MS2 low mass cutoff of 61-Th. Under the Peak Finder tab, Compound Discoverer was chosen as peak table type, and its "Aligned" and "Unaligned" results, as well as the MS/MS results from Spectrum Researcher uploaded. Features had to be identified in a minimum of 1 file (4 files for the FS), however, the average lipid ID was based on a much higher average of 344 features found in plasma and 310 features in the liver dataset. We kept the defaults of a minimum of 75% of lipid spectral purity, an MS2 search dot product of at least 500 and reverse dot product of at least 700, as well as a multiplier of 2.0 (3.0 for FS) for FWHM window, a maximum 15 ppm mass difference, adduct/dimer and in-source fragment filtering, and a maximum RT M.A.D Factor of 3.5. As post-processing all features that were only found in 1 file and had no ID were deleted, and artifactual duplicates deleted.

For the FS liver dataset, peak areas were normalized to the 15:0-18:1(d7)-PC internal standard by dividing each peak area by the internal standards' peak area of that sample and multiplying the result with the median of all internal standard peak areas. The quantification of the internal standard was obtained through TraceFinder 4.0 (Thermo Fisher Scientific). FS plasma results were normalized by dividing each peak area by the feature's average batch control and multiplying with the median feature's peak area over average batch controls. Reported is the log2 of all normalized values. Note that there is no data available for two CAST females as one animal died before sacrifice (CAST-4) and for another there was not enough plasma (CAST-3).

QTL Mapping. Prior to mapping analysis, the lipid metabolite data were adjusted for batch effects using the Combat algorithm[76] as implemented in the R/sva software package.[77] We reduced the 36-state founder probablities to an additive founder dosage, scaled to 1, thus implying an additive genetic model for the genome scans we performed for each lipid feature using the scan1() function in R/qtl2[29] with sex and DO breeding generation included as covariates. This model assumes a normal distribution after transformation and adjusts for the genome scan. We identified suggestive QTL at LOD > 6.0 and significant QTL with a 95% significance level threshold of LOD > 7.4, determined through permutation analysis.[78]

Data Analysis and Plotting. Data analysis was largely performed using R[79] in RStudio[80]. Data formatting was performed utilizing R/dplyr (0.8.3)[81], R/tidyr (1.0.0)[82] and R/reshape2 (1.4.3)[83] and visualisations were created using R/ggplot2 (3.2.1)[84], R/RColorBrewer (1.1-2)[85], and for exploratory analysis, R/plotly (4.9.0)[86]. Heatmaps were generated using R/pheatmap (1.0.12)[87] and manhattan plots were generated based on code accessible via the R graph gallery.[88] All boxplots were generated by ggplot2:geom_boxplot with the first and third quartiles (25th and 75th percentile) for lower and upper hinges, 1.5x interquartile range for the length of the whiskers, center line at median (50% quantile), and all data points, including outliers shown.

Allele effects for each QTL were generated using the scan1blup() function of R/qtl2.[29] SNP associations were performed using the scan1snps() function in R/qtl2_0.20[29] accessing

variants from the database cc_variants.sqlite (available here: `https://ndownloader.figshare.com/files/18533342`) and genes from mouse_genes_mgi.sqlite (available here: `https://ndownloader.figshare.com/files/17609252`) via R/RSQLite_2.1.2.[89]

To nominate candidate gene drivers at lipid-associated QTL, we integrated the lipid data collected in the present study, with hepatic gene expression data previously obtained from a separate cohort of DO mice.[36] We reasoned a locus that demonstrated a hepatic cis-eQTL and a lipid-associated QTL with a similar allele effect patterns is likely to be driving the two phenotypes. We focused on cis-eQTL, as these are expression traits responding to local genetic variation. We computed the Pearson's correlation between the allele effect patterns for all cis-eQTL at a locus to which one or more lipids co-mapped. We performed the same calculation for hepatic cis-pQTL identified in the previous study.[36] For example, at the Chr 10 locus, we identified >25 QTL of unknown lipids in plasma and liver, all of which showed a strong NOD-driven allele effect pattern. About half of these QTL showed NOD as the high allele and half showed NOD as the low allele. We first computed the average allele effect pattern for the NOD-high lipids and the NOD-low lipids. We then identified 55 cis-eQTL and 16 cis-pQTL that were within $\pm$2 Mbp of the lipid QTL at 127 Mbp on Chr 10, and calculated the correlation between their allele effect patterns and the NOD-high and NOD-low lipid QTL. One gene showed a very strong correlation; B4galnt1. The overall correlation between the allele effects of the lipid QTL and the cis-eQTL or cis-pQTL was very low (e.g., 0), suggesting that the vast majority expression traits are responding to genetic variants different than the lipid traits. However, the correlation between the lipid

traits and either the expression or protein level for B4galnt1 was >|0.97|. As B4galnt1 is a known gangliosidase, we then asked if the MS fragmentation pattern for the unknown lipids is consistent with gangliosides. It is worth noting that GM3 ganglioside standard (Cayman Chemicals, Ann Arbor, MI, Item No. 15587) contained N-acetyl-neuraminidate (NANA) - the only sialic acid made by humans. All gangliosides observed in the DO samples contain N-glycolyl-neuraminidate (NGNA), a major sialic acid in mice.[71,72] This powerful approach enabled us to combine the lipid data from one DO study with the gene expression and proteomic data of another DO study to nominate one candidate gene.

Development of LipidGenie. LipidGenie runs on a traditional Linux (CentOS 6), Apache2 (`https://httpd.apache.org/`), MySQL (`https://www.mysql.com/`), and PHP (7.0) web-server. Dynamic HTML documents using were constructed using Bootstrap templates (3.0, `https://getbootstrap.com/`). These HTML documents utilize the JavaScript Framework AngularJS (3.0) to facilitate two-way data binding between user input and server queries. The D3.js library is used to generate all visualizations.[90]. PHP scripts handle database operations and text file creation for QTL data downloads.

LipidGenie's MySQL database structure is shown in **Figure S5.7**. To populate the MySQL database, in-house C# scripts were used to scrape and link the following data from their respective sources: lipid identifications from Lipidex[17] search files, QTLs and LOD curves from the R/QTL2[29] `scan1()` function output, Allele Effect curves from the R/QTL2[29] `scan1blup()` function output, SNPs from the cc_variants.sqlite database and

scan1snps() function output, and genes from the mouse_genes_mgi.sqlite database. The remaining helper tables (shown in green) were built manually.

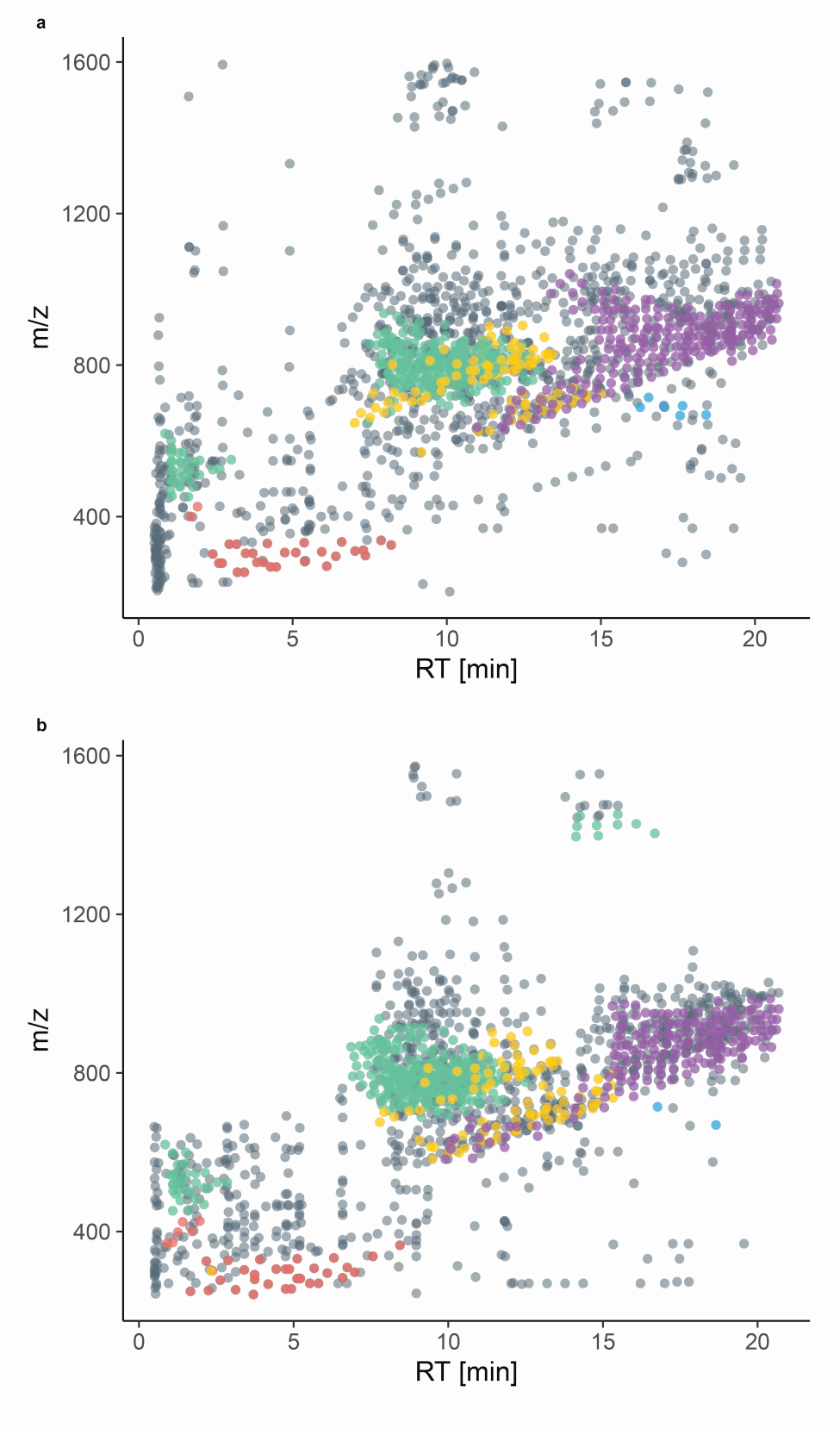
Data Availability. Genotypes and additional phenotype data associated with the DO mouse population have been deposited with Dryad (doi:10.5061/dryad.pj105; data files: Attie Islet eQTL data) (see Keller et al. 2018 for details).[34]

Mass spectrometry files can be found under ID 1610 at Chorus (http://chorusproject.org/). In addition, the lipidomics and QTL data reported here are available at https://uwmadison.box.com/s/2ahtpna8xlhs5j0esnto3zy95upca05j(password: DOLipids_guest01). Figures 1, 2, 3, 4, 5, 6, 7, 8 have associated raw data.

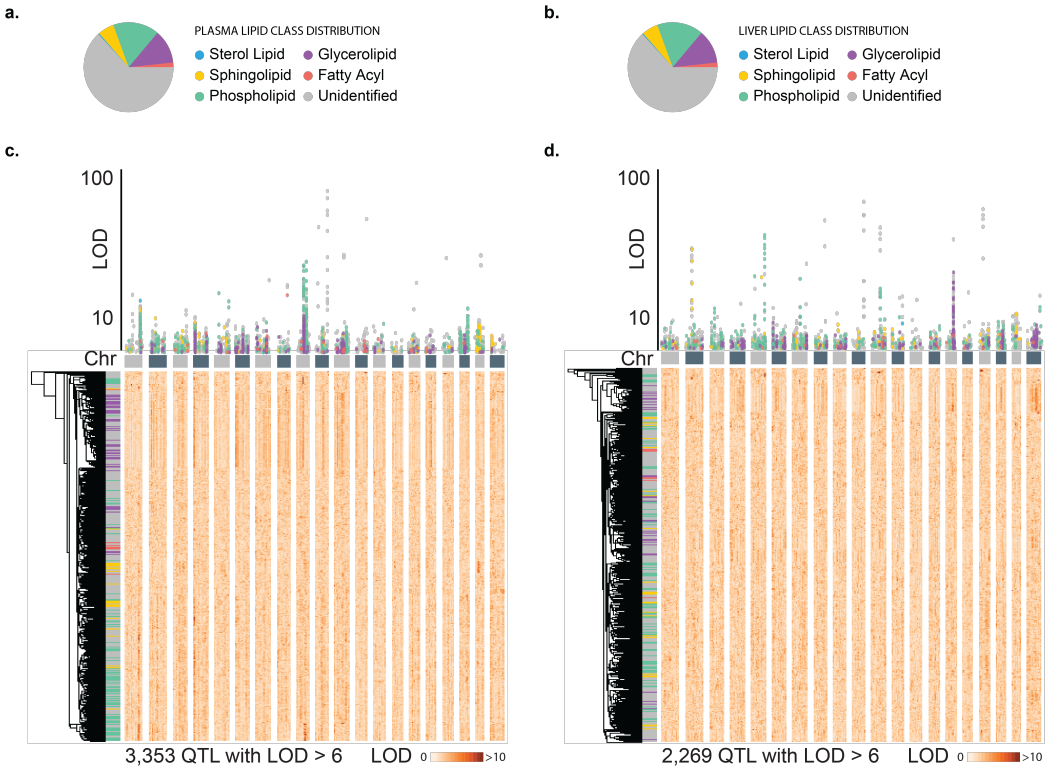
Code Availability. Code for data analysis and plotting is available at https://github.com/vanilink/DOLipids/. The genome-lipid associations are also accessible through an interactive web-based analysis tool that will allow users to replicate the analyses reported here (http://lipidgenie.com/). The source code for this resource can be found at https://github.com/coongroup/LipidGenie.

References

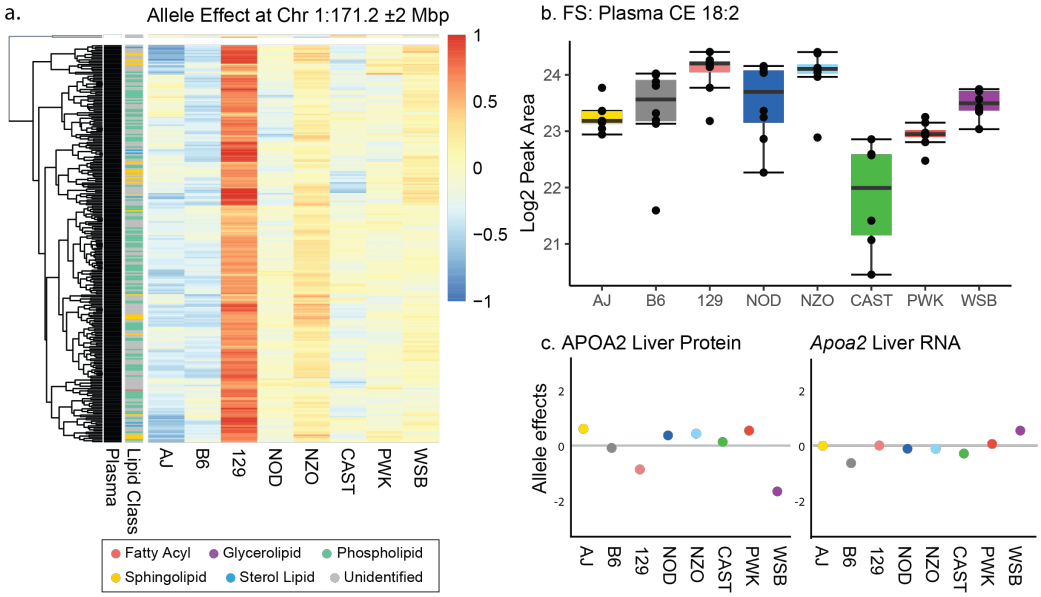
[1] X. Han, "Lipidomics for studying metabolism," *Nat. Rev. Endocrinol.*, vol. 12, no. 11, pp. 668–679, 2016.



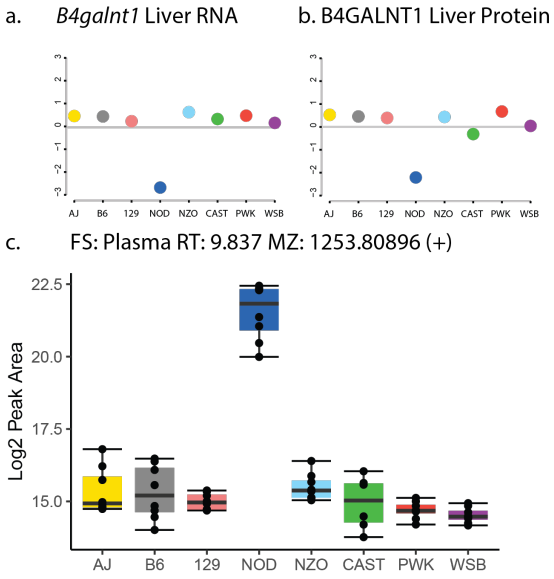
Supplementary Figure S5.1: Identified Lipids and Unidentified Features Occupy Characteristic Regions in the m/z vs. RT space. a, In plasma, we quantified 1,721 lipidomic features, 621 of which were identified, and **b**, In liver, we quantified 1,562 lipidomic features, 615 of which were identified.



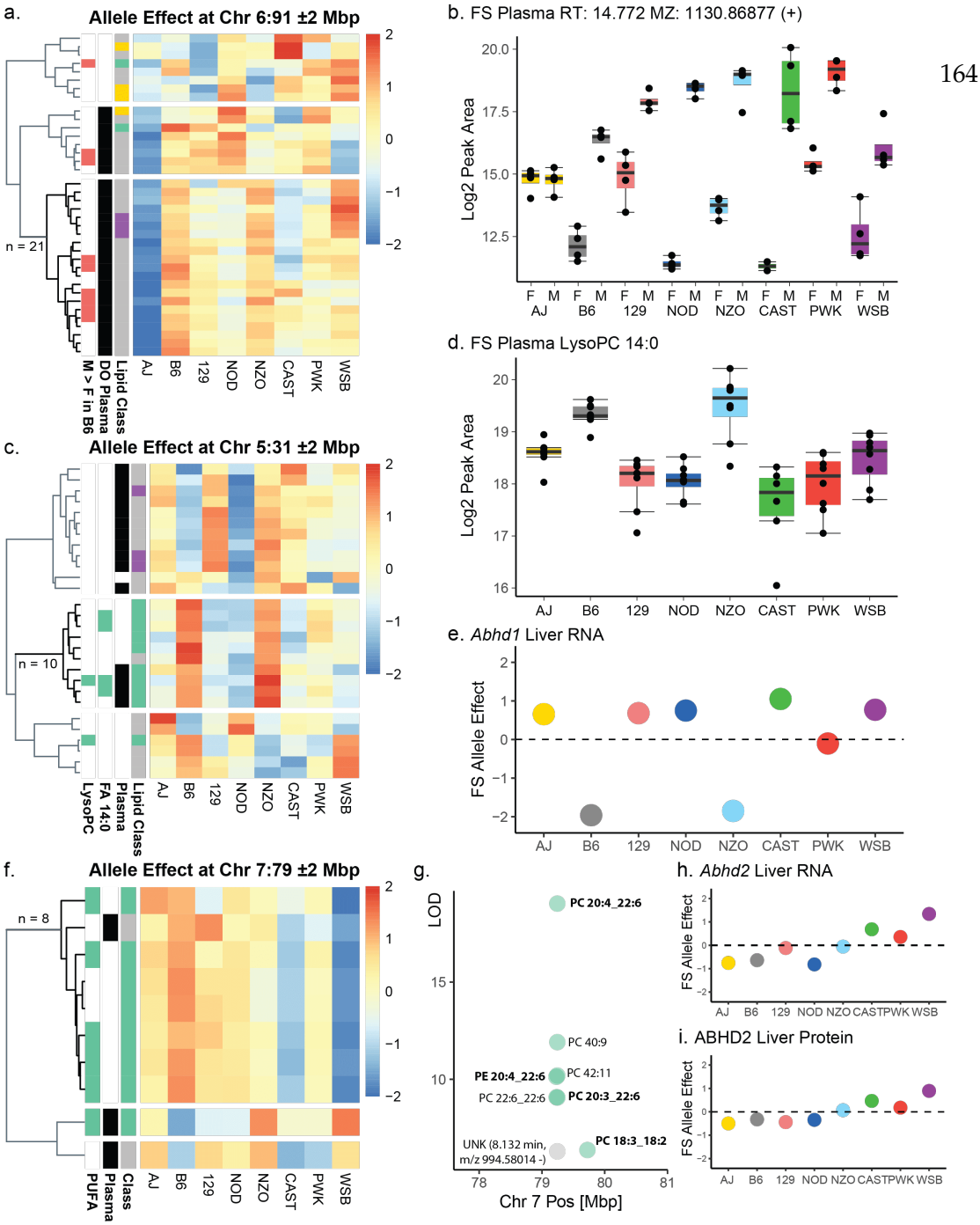
Supplementary Figure S5.2: Lipid Profiling and Subsequent QTL Mapping Reveals Clusters of Associated Lipids. **a**, Lipid class distribution of all 1,721 plasma and **b**, 1,562 liver lipidomic features. **c**, 1,405 plasma and **d**, 1,190 lipid features showed at least one QTL with an LOD > 6 as displayed in a Manhattan plot (n = 3,353 and 2,269 total QTLs, respectively). Hierarchical clustering of these features against the 69,005 markers on the mouse genome, resulted in clustering of lipid class based on hotspots at the genetic level.



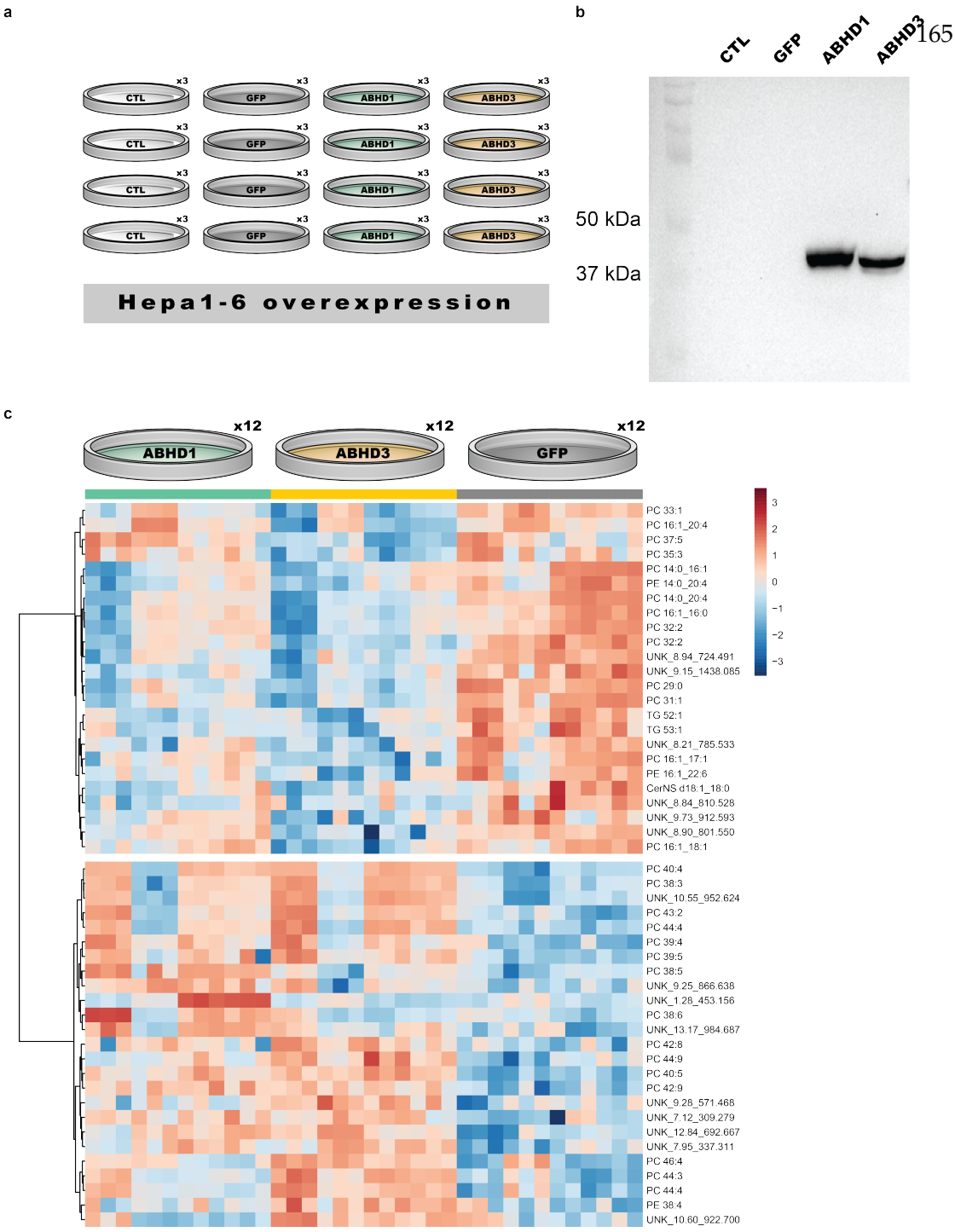
Supplementary Figure S5.3: *Apoa2* as the Candidate Gene at the Largest Lipid Hotspot. **a**, 255 plasma (black) features mapping to the APOA2 locus on chromosome 1 share an allele effect pattern with upregulation in the 129 allele, while 2 mapping liver features (white) do not share the pattern (based on hierarchical clustering on allele effects, with a euclidean distance cutoff of h = 1.5). **b**, The allele effect is exemplary replicated in an independent experiment of founder strain plasma CE(18:2) levels. **c**, The same pattern was not visible in previously reported [36] *Apoa2* liver protein and RNA allele effects.



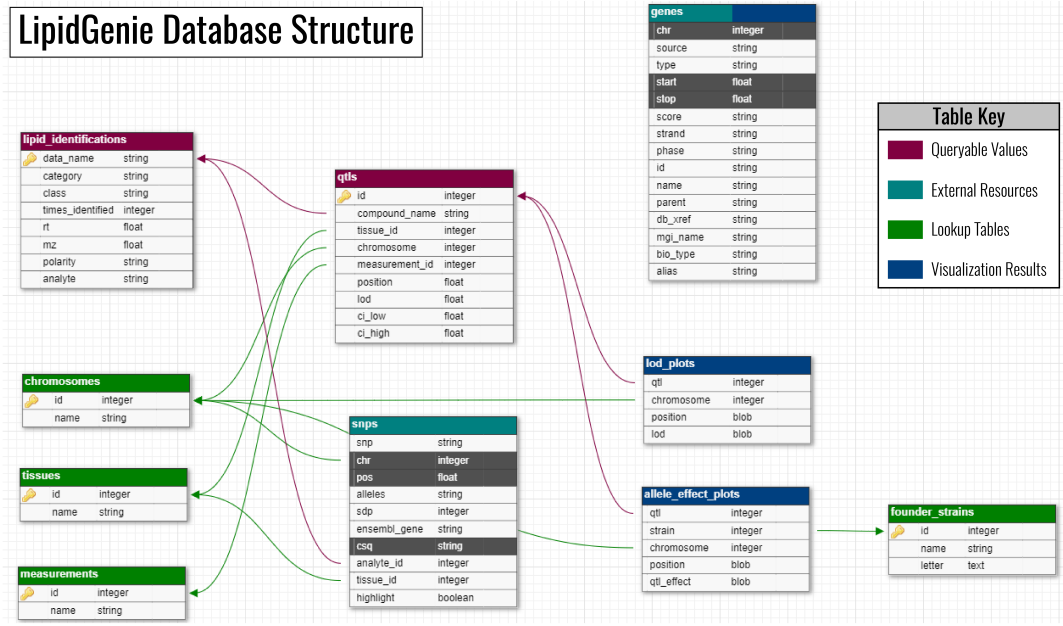
Supplementary Figure S5.4: *B4galnt1* as the Candidate Gene at theHotspot with the Largest LOD. a, The selection of *B4galnt1* as the candidate gene for the chromosome 10:127 Mbp locus was corroborated by NOD-specific allele effects in previously reported liver eQTL and **b**, pQTL.34 **c**, The allele effect patterns of the later as gangliosides identified features mapping to the B4galnt1 locus could further be validated in an independent experiment of founder strain mice (exemplar GM3 pattern).



Supplementary Figure S5.5: Allele Effects Characterize Genome-lipid Hotspots. a, Hierarchical clustering of allele effects at Chr 6:91 Mbp resulted in 21 features with matching A/J down effect (main cluster featuring the six B6 male specific features (red) after row-scaling and Ward clustering, cutoff at h=5). **b**, Consistently, the pattern of male » female was observed for each of the founder strains except for A/J mice as visible in the example boxplots for m/z 1130. **c**, Hierarchical clustering of allele effects at Chr 5:31 Mbp locus resulted in 10 features with matching B6 and NZO up effect (main cluster featuring LysoPC 14:0 (turquoise) after row-scaling and Ward clustering, cutoff at h=8). **d**, This pattern could be replicated in the founder strains, as shown for LysoPC 14:0, as well as **e**, in opposite directionality in a liver eQTL of a previously published dataset.[36] **f**, Hierarchical clustering of allele effects at Chr 7:79 Mbp locus resulted in 8 features with matching WSB down effect (main cluster featuring PUFA-containing phospholipids (turquoise) after row-scaling and Ward clustering, cutoff at h = 2.5). **g**, The mapping phospholipids contained polyunsaturated fatty acids such as 20:4 and 22:6. **h-i**, *Abhd2* liver RNA and protein allele effects of a previously published dataset[36] matched with an opposite WSB high effect.



Supplementary Figure S5.6: Cell Experiments Confirm Fatty Acid Specificity of ABHD1 and ABHD3. a, Experimental design of the validation experiment featuring three technical and four biological replicates of Hepa1-6 cells either untransfected (CTL), transfected with a His-tag GFP control (GFP), or transfected with MYC-tagged *Abhd1* or *Abhd3*. **b**, Western blot of Hepa1-6 overexpression of ABHD1 and ABHD3. Shown is an overlay of membrane and ECL blot for MYC-tag. **c**, Heatmap of top 49 features from discovery lipidomics experiment with $p < 0.05$ (ANOVA, Fisher's LSD post-hoc). Features were sum-normalized and log2-transformed. Hierarchical clustering (Ward clustering, Euclidean distance) shows two clusters with opposite fold changes distinguishing between ABHD1 and ABHD3 and the GFP control.



Supplementary Figure S5.7: LipidGenie Database Structure. The MySQL database behind LipidGenie contains ten distinct tables which are accessed when a QTL query is received. The directional arrows represent unique identifier connections between tables which are used to join multiple table segments together. The `lipid_identifications` and `qtls` tables are populated from LipiDex and QTL mapping outputs. LOD and Allele Effect plots were pregenerated for each QTL and stored in their respective tables. Properties which are shaded gray in the `snps` and `genes` tables are used to query proximal genes and SNPs for the gene visualization in **Figure 5.6**. This schema was made using DBDesigner.net

[2] L. Yang, M. Li, Y. Shan, S. Shen, Y. Bai, and H. Liu, "Recent advances in lipidomics for disease research," *J. Sep. Sci.*, vol. 39, no. 1, pp. 38–50, 2016.

[3] T. Kind, K.-H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, and O. Fiehn, "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nat. Methods*, vol. 10, no. 8, pp. 755–758, 2013.

[4] R. W. Gross and X. Han, "Lipidomics at the interface of structure and function in systems biology," *Chemistry & Biology*, vol. 18, no. 3, pp. 284–291, 2011.

[5] T. Cajka and O. Fiehn, "Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry," *Trends Analyt. Chem.*, vol. 61, pp. 192–206, 2014.

[6] R. Tabassum, J. T. Rämö, P. Ripatti, J. T. Koskela, M. Kurki, J. Karjalainen, P. Palta, S. Hassan, J. Nunez-Fontarnau, T. T. Kiiskinen, S. Söderlund, N. Matikainen, M. J. Gerl, M. A. Surma, C. Klose, N. O. Stitziel, H. Laivuori, A. S. Havulinna, S. K. Service, V. Salomaa, M. Pirinen, M. Jauhiainen, M. J. Daly, N. B. Freimer, A. Palotie, M. R. Taskinen, K. Simons, and S. Ripatti, "Genetic architecture of human plasma lipidome and its link to cardiovascular disease," *Nature Communications*, vol. 10, no. 1, pp. 1–14, 2019.

[7] R. Kiyonami, D. A. Peake, X. Liu, and Y. Huang, "Large-Scale lipid profiling of a human

serum lipidome using a High-Resolution, Accurate-Mass LC/MS/MS approach," in *LIPID MAPS Annual Meeting*, pp. 12–13, pdfs.semanticscholar.org, 2015.

[8] D. A. Slatter, M. Aldrovandi, A. O'Connor, S. M. Allen, C. J. Brasher, R. C. Murphy, S. Mecklemann, S. Ravi, V. Darley-Usmar, and V. B. O'Donnell, "Mapping the human platelet lipidome reveals cytosolic phospholipase A2 as a regulator of mitochondrial bioenergetics during activation," *Cell Metab.*, vol. 23, no. 5, pp. 930–944, 2016.

[9] K. Contrepois, S. Mahmoudi, B. K. Ubhi, K. Papsdorf, D. Hornburg, A. Brunet, and M. Snyder, "Cross-Platform comparison of untargeted and targeted lipidomics approaches on aging mouse plasma," *Sci. Rep.*, vol. 8, no. 1, p. 17747, 2018.

[10] I. Blaženović, T. Shen, S. S. Mehta, T. Kind, J. Ji, M. Piparo, F. Cacciola, L. Mondello, and O. Fiehn, "Increasing compound identification rates in untargeted lipidomics research with liquid chromatography drift Time-Ion mobility mass spectrometry," *Anal. Chem.*, vol. 90, no. 18, pp. 10758–10764, 2018.

[11] N. G. Mahieu and G. J. Patti, "Systems-Level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites," *Analytical Chemistry*, vol. 89, no. 19, pp. 10397–10406, 2017.

[12] I. Blaženović, T. Kind, J. Ji, and O. Fiehn, "Software tools and approaches for compound identification of LC-MS/MS data in metabolomics," *Metabolites*, vol. 8, no. 2, 2018.

[13] R. W. Gross, "The evolution of lipidomics through space and time," *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, vol. 1862, no. 8, pp. 731–739, 2017.

[14] J. P. Koelmel, N. M. Kroeger, C. Z. Ulmer, J. A. Bowden, R. E. Patterson, J. A. Cochran, C. W. W. Beecher, T. J. Garrett, and R. A. Yost, "LipidMatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data," *BMC Bioinformatics*, vol. 18, no. 1, p. 331, 2017.

[15] J. Hartler, A. Triebl, A. Ziegl, M. Trötzmüller, G. N. Rechberger, O. A. Zeleznik, K. A. Zierler, F. Torta, A. Cazenave-Gassiot, M. R. Wenk, A. Fauland, C. E. Wheelock, A. M. Armando, O. Quehenberger, Q. Zhang, M. J. O. Wakelam, G. Haemmerle, F. Spener, H. C. Köfeler, and G. G. Thallinger, "Deciphering lipid structures based on platform-independent decision rules," *Nature Methods*, vol. 14, no. 12, pp. 1171–1174, 2017.

[16] P. D. Hutchins, J. D. Russell, and J. J. Coon, "LipiDex: An integrated software package for High-Confidence lipid identification," *Cell Syst*, vol. 6, no. 5, pp. 621–625.e5, 2018.

[17] P. D. Hutchins, J. D. Russell, and J. J. Coon, "Mapping lipid fragmentation for tailored mass spectral libraries," *J. Am. Soc. Mass Spectrom.*, vol. 30, no. 4, pp. 659–668, 2019.

[18] Y. Kostyukevich, G. Vladimirov, E. Stekolschikova, D. Ivanov, A. Yablokov, A. Zherebker, S. Sosnin, A. Orlov, M. Fedorov, P. Khaitovich, and E. Nikolaev, "Hydrogen/Deu-

terium exchange aiding compound identification for LC-MS and MALDI imaging lipidomics," *Analytical Chemistry*, 2019.

[19] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek, and J. R. Yates, "Protein analysis by shotgun/bottom-up proteomics," *Chemical Reviews*, vol. 113, no. 4, pp. 2343–2394, 2013.

[20] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[21] M. E. Dumas, C. Domange, S. Calderari, A. R. Martínez, R. Ayala, S. P. Wilder, N. Suárez-Zamorano, S. C. Collins, R. H. Wallis, Q. Gu, Y. Wang, C. Hue, G. W. Otto, K. Argoud, V. Navratil, S. C. Mitchell, J. C. Lindon, E. Holmes, J. B. Cazier, J. K. Nicholson, and D. Gauguier, "Topological analysis of metabolic networks integrating co-segregating transcriptomes and metabolomes in type 2 diabetic rat congenic series," *Genome Medicine*, vol. 8, no. 1, p. 101, 2016.

[22] J. B. Cazier, P. J. Kaisaki, K. Argoud, B. J. Blaise, K. Veselkov, T. M. Ebbels, T. Tsang, Y. Wang, M. T. Bihoreau, S. C. Mitchell, E. C. Holmes, J. C. Lindon, J. Scott, J. K. Nicholson, M. E. Dumas, and D. Gauguier, "Untargeted metabolome quantitative

trait locus mapping associates variation in urine glycerate to mutant glycerate kinase," *Journal of Proteome Research*, vol. 11, no. 2, pp. 631–642, 2012.

[23] J. Krumsiek, K. Suhre, A. M. Evans, M. W. Mitchell, R. P. Mohney, M. V. Milburn, B. Wägele, W. Römisch-Margl, T. Illig, J. Adamski, C. Gieger, F. J. Theis, and G. Kastenmüller, "Mining the unknown: A systems approach to metabolite identification combining genetic and metabolic information," *PLoS Genet.*, vol. 8, no. 10, p. e1003005, 2012.

[24] S.-Y. Shin, E. B. Fauman, A.-K. Petersen, J. Krumsiek, R. Santos, J. Huang, M. Arnold, I. Erte, V. Forgetta, T.-P. Yang, K. Walter, C. Menni, L. Chen, L. Vasquez, A. M. Valdes, C. L. Hyde, V. Wang, D. Ziemek, P. Roberts, L. Xi, E. Grundberg, Multiple Tissue Human Expression Resource (MuTHER) Consortium, M. Waldenberger, J. B. Richards, R. P. Mohney, M. V. Milburn, S. L. John, J. Trimmer, F. J. Theis, J. P. Overington, K. Suhre, M. J. Brosnan, C. Gieger, G. Kastenmüller, T. D. Spector, and N. Soranzo, "An atlas of genetic influences on human blood metabolites," *Nat. Genet.*, vol. 46, no. 6, pp. 543–550, 2014.

[25] J. Raffler, W. Römisch-Margl, A.-K. Petersen, P. Pagel, F. Blöchl, C. Hengstenberg, T. Illig, C. Meisinger, K. Stark, H.-E. Wichmann, J. Adamski, C. Gieger, G. Kastenmüller, and K. Suhre, "Identification and MS-assisted interpretation of genetically influenced NMR signals in human plasma," *Genome Medicine*, vol. 5, no. 2, p. 13, 2013.

[26] R. Rueedi, R. Mallol, J. Raffler, D. Lamparter, N. Friedrich, P. Vollenweider, G. Waeber, G. Kastenmüller, Z. Kutalik, and S. Bergmann, "Metabomatching: Using genetic association to identify metabolites in proton NMR spectroscopy," *PLoS Computational Biology*, vol. 13, no. 12, 2017.

[27] V. Matyash, G. Liebisch, T. V. Kurzchalia, A. Shevchenko, and D. Schwudke, "Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics," *J. Lipid Res.*, vol. 49, no. 5, pp. 1137–1146, 2008.

[28] D. M. Gatti, K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, and Others, "Quantitative trait locus mapping methods for diversity outbred mice," *G3: Genes, Genomes, Genetics*, vol. 4, no. 9, pp. 1623–1633, 2014.

[29] K. W. Broman, D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, Ś. Sen, B. S. Yandell, and G. A. Churchill, "R/qtl2: Software for mapping quantitative trait loci with High-Dimensional data and multiparent populations," *Genetics*, vol. 211, no. 2, pp. 495–502, 2019.

[30] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill, "High-resolution genetic mapping using the mouse diversity outbred population," *Genetics*, vol. 190, no. 2, pp. 437–447, 2012.

[31] E. J. Chesler, D. M. Gatti, A. P. Morgan, M. Strobel, L. Trepanier, D. Oberbeck,

S. McWeeney, R. Hitzemann, M. Ferris, R. McMullan, A. Clayshultle, T. A. Bell, F. P. Manuel de Villena, and G. A. Churchill, "Diversity outbred mice at 21: Maintaining allelic variation in the face of selection," *G3*, vol. 6, no. 12, pp. 3893–3902, 2016.

[32] R. Mayer, A. Brero, J. von Hase, T. Schroeder, T. Cremer, and S. Dietzel, "Common themes and cell type specific variations of higher order chromatin arrangements in the mouse," *BMC Cell Biol.*, vol. 6, p. 44, 2005.

[33] D. L. Aylor, W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, R. S. Baric, M. T. Ferris, J. A. Frelinger, M. Heise, M. B. Frieman, L. E. Gralinski, T. A. Bell, J. D. Didion, K. Hua, D. L. Nehrenberg, C. L. Powell, J. Steigerwalt, Y. Xie, S. N. P. Kelada, F. S. Collins, I. V. Yang, D. A. Schwartz, L. A. Branstetter, E. J. Chesler, D. R. Miller, J. Spence, E. Y. Liu, L. McMillan, A. Sarkar, J. Wang, W. Wang, Q. Zhang, K. W. Broman, R. Korstanje, C. Durrant, R. Mott, F. A. Iraqi, D. Pomp, D. Threadgill, F. P.-M. de Villena, and G. A. Churchill, "Genetic analysis of complex traits in the emerging collaborative cross," *Genome Res.*, vol. 21, no. 8, pp. 1213–1222, 2011.

[34] M. P. Keller, D. M. Gatti, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, P. Simecek, M. Vincent, S. Allen, A. T. Broman, R. Bacher, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie, "Genetic drivers of pancreatic islet function," *Genetics*, vol. 209, no. 1, pp. 335–356, 2018.

[35] M. P. Keller, M. E. Rabaglia, K. L. Schueler, D. S. Stapleton, D. M. Gatti, M. Vincent, K. A. Mitok, Z. Wang, T. Ishimura, S. P. Simonett, C. H. Emfinger, R. Das, T. Beck,

C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie, "Gene loci associated with insulin secretion in islets from nondiabetic mice," *Journal of Clinical Investigation*, vol. 129, no. 10, pp. 4419–4432, 2019.

[36] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, no. 7608, pp. 500–505, 2016.

[37] J. H. Kemis, V. Linke, K. L. Barrett, F. J. Boehm, L. L. Traeger, M. P. Keller, M. E. Rabaglia, K. L. Schueler, D. S. Stapleton, D. M. Gatti, G. A. Churchill, D. Amador-Noguez, J. D. Russel, B. S. Yandell, K. W. Broman, J. J. Coon, A. D. Attie, and F. E. Rey, "Genetic determinants of gut microbiota composition and bile acid profiles in mice," *PLoS Genetics*, vol. 15, no. 8, p. e1008073, 2019.

[38] E. Fahy, S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. H. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. O. Wakelam, and E. A. Dennis, "Update of the LIPID MAPS comprehensive classification system for lipids," *J. Lipid Res.*, vol. 50, no. Supplement, pp. S9–S14, 2009.

[39] G. Liebisch, J. A. Vizcaíno, H. Köfeler, M. Trötzmüller, W. J. Griffiths, G. Schmitz, F. Spener, and M. J. O. Wakelam, "Shorthand notation for lipid structures derived from mass spectrometry," *Journal of Lipid Research*, vol. 54, no. 6, pp. 1523–1530, 2013.

[40] Z. Su, X. Wang, S.-W. Tsaih, A. Zhang, A. Cox, S. Sheehan, and B. Paigen, "Genetic basis

of HDL variation in 129/SvImJ and C57BL/6J mice: importance of testing candidate genes in targeted mutant mice," *J. Lipid Res.*, vol. 50, no. 1, pp. 116–125, 2009.

[41] J. Kettunen, A. Demirkan, P. Würtz, H. H. Draisma, T. Haller, R. Rawal, A. Vaarhorst, A. J. Kangas, L. P. Lyytikäinen, M. Pirinen, R. Pool, A. P. Sarin, P. Soininen, T. Tukiainen, Q. Wang, M. Tiainen, T. Tynkkynen, N. Amin, T. Zeller, M. Beekman, J. Deelen, K. W. Van Dijk, T. Esko, J. J. Hottenga, E. M. Van Leeuwen, T. Lehtimäki, E. Mihailov, R. J. Rose, A. J. De Craen, C. Gieger, M. Kähönen, M. Perola, S. Blankenberg, M. J. Savolainen, A. Verhoeven, J. Viikari, G. Willemsen, D. I. Boomsma, C. M. Van Duijn, J. Eriksson, A. Jula, M. R. Järvelin, J. Kaprio, A. Metspalu, O. Raitakari, V. Salomaa, P. Eline Slagboom, M. Waldenberger, S. Ripatti, and M. Ala-Korpela, "Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA," *Nature Communications*, vol. 7, no. 1, pp. 1–9, 2016.

[42] W. Zhang, R. Korstanje, J. Thaisz, F. Staedtler, N. Harttman, L. Xu, M. Feng, L. Yanas, H. Yang, W. Valdar, G. A. Churchill, and K. Dipetrillo, "Genome-wide association mapping of quantitative traits in outbred mice," *G3*, vol. 2, no. 2, pp. 167–174, 2012.

[43] N. Pamir, C. Pan, D. L. Plubell, P. M. Hutchins, C. Tang, J. Wimberger, A. Irwin, T. Q. d. A. Vallim, J. W. Heinecke, and A. J. Lusis, "Genetic control of the HDL proteome." 2018.

[44] X. Wang, R. Korstanje, D. Higgins, and B. Paigen, "Haplotype analysis in multiple crosses to identify a QTL gene," *Genome Res.*, vol. 14, no. 9, pp. 1767–1772, 2004.

[45] F. Blanco-Vaca, J. C. Escolà-Gil, J. M. Martín-Campos, and J. Julve, "Role of apoA-II in lipid metabolism and atherosclerosis: advances in the study of an enigmatic protein," *J. Lipid Res.*, vol. 42, no. 11, pp. 1727–1739, 2001.

[46] A. Kontush, M. Lhomme, and M. J. Chapman, "Unraveling the complexities of the HDL lipidome," *J. Lipid Res.*, vol. 54, no. 11, pp. 2950–2963, 2013.

[47] R. C. Murphy, T. J. Leiker, and R. M. Barkley, "Glycerolipid and cholesterol ester analyses in biological samples by mass spectrometry," *Biochim. Biophys. Acta*, vol. 1811, no. 11, p. 776, 2011.

[48] L. A. Lerno, Jr, J. B. German, and C. B. Lebrilla, "Method for the identification of lipid classes based on referenced kendrick mass analysis," *Anal. Chem.*, vol. 82, no. 10, pp. 4236–4245, 2010.

[49] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner, "The sequence ontology: a tool for the unification of genome annotations," *Genome Biol.*, vol. 6, no. 5, p. R44, 2005.

[50] Y. Nagata, S. Yamashiro, J. Yodoi, K. O. Lloyd, H. Shiku, and K. Furukawa, "Expression cloning of beta 1,4 n-acetylgalactosaminyltransferase cDNAs that determine the expression of GM2 and GD2 gangliosides," *J. Biol. Chem.*, vol. 267, no. 17, pp. 12082–12089, 1992.

[51] F. Dotta, L. B. Peterson, M. Previti, J. Metzger, C. Tiberti, E. Anastasi, P. Zoppitelli, L. S. Wicker, and U. Di Mario, "Pancreatic islet ganglioside expression in nonobese diabetic mice: comparison with C57BL/10 mice and changes after autoimmune beta-cell destruction," *Endocrinology*, vol. 130, no. 1, pp. 37–42, 1992.

[52] Z. Li, Y. Fan, J. Liu, Y. Li, C. Huan, H. H. Bui, M.-S. Kuo, T.-S. Park, G. Cao, and X.-C. Jiang, "Impact of sphingomyelin synthase 1 deficiency on sphingolipid metabolism and atherosclerosis in mice," *Arterioscler. Thromb. Vasc. Biol.*, vol. 32, no. 7, pp. 1577–1584, 2012.

[53] A. K. Bergfeld, R. Lawrence, S. L. Diaz, O. M. T. Pearce, D. Ghaderi, P. Gagneux, M. G. Leakey, and A. Varki, "-glycolyl groups of nonhuman chondroitin sulfates survive in ancient fossils," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 39, pp. E8155–E8164, 2017.

[54] P. Strømme, K. Dobrenis, R. V. Sillitoe, M. Gulinello, N. F. Ali, C. Davidson, M. C. Micsenyi, G. Stephney, L. Ellevog, A. Klungland, and S. U. Walkley, "X-linked angelman-like syndrome caused by slc9a6 knockout in mice exhibits evidence of endosomal–lysosomal dysfunction," *Brain*, vol. 134, no. 11, pp. 3369–3383, 2011.

[55] W. Spessott, A. Uliana, and H. J. F. Maccioni, "Defective GM3 synthesis in cog2 null mutant CHO cells associates to mislocalization of lactosylceramide sialyltransferase in the golgi complex," *Neurochem. Res.*, vol. 35, no. 12, pp. 2161–2167, 2010.

[56] R. W. Ledeen and G. Wu, "The multi-tasked life of GM1 ganglioside, a true factotum of nature," *Trends in Biochemical Sciences*, vol. 40, no. 7, pp. 407–418, 2015.

[57] C. J. Bult, J. A. Blake, C. L. Smith, J. A. Kadin, J. E. Richardson, A. Anagnostopoulos, R. Asabor, R. M. Baldarelli, J. S. Beal, S. M. Bello, O. Blodgett, N. E. Butler, K. R. Christie, L. E. Corbani, J. Creelman, M. E. Dolan, H. J. Drabkin, S. L. Giannatto, P. Hale, D. P. Hill, M. Law, A. Mendoza, M. McAndrews, D. Miers, H. Motenko, L. Ni, H. Onda, M. Perry, J. M. Recla, B. Richards-Smith, D. Sitnikov, M. Tomczuk, G. Tonorio, L. Wilming, and Y. Zhu, "Mouse Genome Database (MGD) 2019," *Nucleic Acids Research*, vol. 47, no. D1, pp. D801–D806, 2019.

[58] The Jackson Laboratory, "Human-Mouse Homologue Database."

[59] B. L. Parker, A. C. Calkin, M. M. Seldin, M. F. Keating, E. J. Tarling, P. Yang, S. C. Moody, Y. Liu, E. J. Zerenturk, E. J. Needham, M. L. Miller, B. L. Clifford, P. Morand, M. J. Watt, R. C. R. Meex, K.-Y. Peng, R. Lee, K. Jayawardana, C. Pan, N. A. Mellett, J. M. Weir, R. Lazarus, A. J. Lusis, P. J. Meikle, D. E. James, T. Q. de Aguiar Vallim, and B. G. Drew, "An integrative systems genetic analysis of mammalian lipid metabolism," *Nature*, 2019.

[60] M. M. Yore, I. Syed, P. M. Moraes-Vieira, T. Zhang, M. A. Herman, E. A. Homan, R. T. Patel, J. Lee, S. Chen, O. D. Peroni, A. S. Dhaneshwar, A. Hammarstedt, U. Smith, T. E. McGraw, A. Saghatelian, and B. B. Kahn, "Discovery of a class of endogenous

mammalian lipids with anti-diabetic and anti-inflammatory effects," *Cell*, vol. 159, no. 2, pp. 318–332, 2014.

[61] S. McLean, N. W. Davies, D. S. Nichols, and B. J. Mcleod, "Triacylglycerol estolides, a new class of mammalian lipids, in the paracloacal gland of the brushtail possum (trichosurus vulpecula)," *Lipids*, vol. 50, no. 6, pp. 591–604, 2015.

[62] C. C. Lord, G. Thomas, and J. M. Brown, "Mammalian alpha beta hydrolase domain (ABHD) proteins: Lipid metabolizing enzymes at the interface of cell signaling and energy metabolism," *Biochim. Biophys. Acta*, vol. 1831, no. 4, pp. 792–802, 2013.

[63] J. Z. Long, J. S. Cisar, D. Milliken, S. Niessen, C. Wang, S. A. Trauger, G. Siuzdak, and B. F. Cravatt, "Metabolomics annotates ABHD3 as a physiologic regulator of medium-chain phospholipids," *Nat. Chem. Biol.*, vol. 7, no. 11, pp. 763–765, 2011.

[64] H. H. Draisma, R. Pool, M. Kobl, R. Jansen, A. K. Petersen, A. A. Vaarhorst, I. Yet, T. Haller, A. Demirkan, T. Esko, G. Zhu, S. Böhringer, M. Beekman, J. B. Van Klinken, W. Römisch-Margl, C. Prehn, J. Adamski, A. J. De Craen, E. M. Van Leeuwen, N. Amin, H. Dharuri, H. J. Westra, L. Franke, E. J. De Geus, J. J. Hottenga, G. Willemsen, A. K. Henders, G. W. Montgomery, D. R. Nyholt, J. B. Whitfield, B. W. Penninx, T. D. Spector, A. Metspalu, P. Eline Slagboom, K. W. Van Dijk, P. A. 'T Hoen, K. Strauch, N. G. Martin, G. J. B. Van Ommen, T. Illig, J. T. Bell, M. Mangino, K. Suhre, M. I. McCarthy, C. Gieger, A. Isaacs, C. M. Van Duijn, and D. I. Boomsma, "Genome-wide association study

identifies novel genetic variants contributing to variation in blood metabolite levels," *Nature Communications*, vol. 6, 2015.

[65] C. Y. Ha, J. Y. Kim, J. K. Paik, O. Y. Kim, Y.-H. Paik, E. J. Lee, and J. H. Lee, "The association of specific metabolites of lipid metabolism with markers of oxidative stress, inflammation and arterial stiffness in men with newly diagnosed type 2 diabetes," *Clin. Endocrinol.*, vol. 76, no. 5, pp. 674–682, 2012.

[66] A. Demirkan, C. M. van Duijn, P. Ugocsai, A. Isaacs, P. P. Pramstaller, G. Liebisch, J. F. Wilson, Å. Johansson, I. Rudan, Y. S. Aulchenko, A. V. Kirichenko, A. C. J. W. Janssens, R. C. Jansen, C. Gnewuch, F. S. Domingues, C. Pattaro, S. H. Wild, I. Jonasson, O. Polasek, I. V. Zorkoltseva, A. Hofman, L. C. Karssen, M. Struchalin, J. Floyd, W. Igl, Z. Biloglav, L. Broer, A. Pfeufer, I. Pichler, S. Campbell, G. Zaboli, I. Kolcic, F. Rivadeneira, J. Huffman, N. D. Hastie, A. Uitterlinden, L. Franke, C. S. Franklin, V. Vitart, DIAGRAM Consortium, C. P. Nelson, M. Preuss, CARDIoGRAM Consortium, J. C. Bis, C. J. O'Donnell, N. Franceschini, CHARGE Consortium, J. C. M. Witteman, T. Axenovich, B. A. Oostra, T. Meitinger, A. A. Hicks, C. Hayward, A. F. Wright, U. Gyllensten, H. Campbell, G. Schmitz, and EUROSPAN consortium, "Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations," *PLoS Genet.*, vol. 8, no. 2, p. e1002490, 2012.

[67] M. R. Miller, N. Mannowetz, A. T. Iavarone, R. Safavi, E. O. Gracheva, J. F. Smith, R. Z. Hill, D. M. Bautista, Y. Kirichok, and P. V. Lishko, "Unconventional endocannabi-

noid signaling governs sperm activation via the sex hormone progesterone," *Science*, vol. 352, no. 6285, pp. 555–559, 2016.

[68] M. P. Baggelaar, M. Maccarrone, and M. van der Stelt, "2-arachidonoylglycerol: A signaling lipid with manifold actions in the brain," *Progress in Lipid Research*, vol. 71, pp. 1–17, 2018.

[69] P. Jha, M. T. McDevitt, R. Gupta, P. M. Quiros, E. G. Williams, K. Gariani, M. B. Sleiman, L. Diserens, A. Jochem, A. Ulbrich, J. J. Coon, J. Auwerx, and D. J. Pagliarini, "Systems analyses reveal physiological roles and genetic regulators of liver lipid species," *Cell Syst*, vol. 6, no. 6, pp. 722–733.e6, 2018.

[70] P. Jha, M. T. McDevitt, E. Halilbasic, E. G. Williams, P. M. Quiros, K. Gariani, M. B. Sleiman, R. Gupta, A. Ulbrich, A. Jochem, J. J. Coon, M. Trauner, D. J. Pagliarini, and J. Auwerx, "Genetic regulation of plasma lipid species and their association with metabolic phenotypes," *Cell Syst*, vol. 6, no. 6, pp. 709–721.e6, 2018.

[71] S. Kavaler, H. Morinaga, A. Jih, W. Fan, M. Hedlund, A. Varki, and J. J. Kim, "Pancreatic beta-cell failure in obese mice with human-like CMP-Neu5Ac hydroxylase deficiency," *FASEB J.*, vol. 25, no. 6, pp. 1887–1893, 2011.

[72] A. Salama, M. Mosser, X. Lévêque, A. Perota, J.-P. Judor, C. Danna, S. Pogu, A. Mouré, D. Jégou, N. Gaide, J. Abadie, O. Gauthier, J.-P. Concordet, S. Le Bas-Bernardet, D. Riochet, L. Le Berre, J. Hervouet, D. Minault, P. Weiss, J. Guicheux, S. Brouard,

S. Bosch, I. Lagutina, R. Duchi, G. Lazzari, E. Cozzi, G. Blancho, S. Conchon, C. Galli, J.-P. Soulillou, and J.-M. Bach, "Neu5Gc and α1-3 GAL xenoantigen knockout does not affect glycemia homeostasis and insulin secretion in pigs," *Diabetes*, vol. 66, no. 4, pp. 987–993, 2017.

[73] K. A. Mitok, E. C. Freiberger, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, N. W. Kwiecien, P. A. Malec, A. S. Hebert, A. T. Broman, R. T. Kennedy, M. P. Keller, J. J. Coon, and A. D. Attie, "Islet proteomics reveals genetic variation in dopamine production resulting in altered insulin secretion," *Journal of Biological Chemistry*, vol. 293, no. 16, pp. 5860–5877, 2018.

[74] K. W. Broman, D. M. Gatti, K. L. Svenson, Ś. Sen, and G. A. Churchill, "Cleaning genotype data from diversity outbred mice," *G3*, vol. 9, no. 5, pp. 1571–1579, 2019.

[75] R. Adusumilli and P. Mallick, "Data conversion with ProteoWizard msconvert," *Methods Mol. Biol.*, vol. 1550, pp. 339–368, 2017.

[76] W. E. Johnson, W. Evan Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[77] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[78] G. A. Churchill and R. W. Doerge, "Empirical threshold values for quantitative trait mapping," *Genetics*, vol. 138, no. 3, pp. 963–971, 1994.

[79] R Core Team, "R: A language and environment for statistical computing," 2019.

[80] RStudio Team, "RStudio: Integrated development environment for R," 2016.

[81] H. Wickham, R. François, L. Henry, and K. Müller, "dplyr: A grammar of data manipulation," 2019.

[82] H. Wickham and L. Henry, "tidyr: Tidy messy data," 2019.

[83] H. Wickham and Others, "Reshaping data with the reshape package," *J. Stat. Softw.*, vol. 21, no. 12, pp. 1–20, 2007.

[84] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2016.

[85] E. Neuwirth, "RColorBrewer: ColorBrewer palettes," 2014.

[86] C. Sievert, "plotly for R," 2018.

[87] R. Kolde, "pheatmap: Pretty heatmaps," 2019.

[88] Y. Holtz, "Manhattan plot in r: a review." `https://www.r-graph-gallery.com/101_Manhattan_plot.html`. Accessed: 2019-11-14.

[89] K. Müller, H. Wickham, D. A. James, and S. Falcon, "RSQLite: 'SQLite' interface for R," 2019.

[90] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

Chapter 6

CONCLUSION

DRB wrote this chapter.

Portions of this chapter are part of a manuscript in preparation:

Brademan DR, Serrano L, Linke V, Westphall MS, Simcox J, Coon JJ. *Implementation of a Real-Time Intelligent Data Acquisition Algorithm for Deeper Lipidome Characterization*. **2020**

Summary

Over the last several decades the analytical capabilities of mass spectrometry technology have improved dramatically. While sequencing single protein digests was once a daunting task, we now routinely characterize whole proteomes in a relatively short amount of time. This rapid increase in data quality and sample throughput has raised an equally strong demand regarding improvements to data analysis and dissemination of results. In this dissertation I have described and developed four separate software platforms which are aimed to address these demands. The first software platform described in **Chapter 2** provides a new resource to the mass spectrometry community to generate and explore sequence-annotated peptide tandem mass spectra. Since its release, this tool has been used extensively by the MS community and has processed tens of thousands of spectra. The second software platform described in **Chapter 3** concerns the development of a quality control web application is used internally by the Coon lab to track proteomics instrument performance metrics over time. We have found this tool to be critical for maintaining high instrument performance. The NCQBCS Controller quickly indicates when an instrument is under-performing, and tracked QC metrics The third software platform described in **Chapter 4** consists of a customizable data portal which allows users to upload and organize their own omics data. Data portals enable users to explore their data using a set of common case-control visualization options. Finally, the fourth software platform described in **Chapter 5** facilitates the interpretation of genome-lipid associations.

Future Work

Expanding the Purview of LipidGenie to Other Biomolecule Classes LipidGenie as described in **Chapter 4** enables the exploration of data resulting from solely QTL mapped lipids extracted from liver and plasma. During the data collection phase of the DO mouse collaboration, the Coon lab collected proteomics analyses on mouse islets, discovery metabolomics analyses on mouse liver, plasma, and cecum, discovery lipidomics analyses on mouse liver, plasma, and cecum, and targeted bile acid assays on mouse cecum and plasma. QTL mapping was conducted for all the above analyses, and all the resulting QTLs-biomolecule associations were inserted into LipidGenie's database. Specifically, these entries were placed into the 'qtls' table.

As of yet, SNP association, LOD plot generation, and allele effect plot generation have not been conducted in the format required for integration into LipidGenie's current architecture. These data are required to properly visualize QTL effects. A slight database refactor would also be required to support data from multiple omes. Once generated, these data could be inserted alongside the lipid dataset. Once these new data are inserted, I would be able to quickly adapt LipidGenie's client-side environment to enable exploration of QTLs generated from multiple omes.

Intelligent Lipidomics Data Acquisition Mass spectrometry has proven to be an excellent resource for the global characterization of multiple biomolecular classes: namely proteins

and their post-translational modifications, lipids, and metabolites. The abundances of these biomolecules are greatly reactive to genetic and environmental stressors, and perturbations of their abundances can be used to glean a deeper understanding of biological systems.[1] Proteomic methodology has been well developed. Global proteome characterization is possible with just over an hour of analysis time.[2] We can also identify and quantify over 1,000 and 200 unique lipid and metabolomic species in a single-shot analysis, respectfully.[3]

Lipids are mainly known for their role as major cell membrane components. Lipids also participate in signaling pathways, energy storage, metabolism, mitochondrial respiration, and as enzyme cofactors, and as such are very responsive to cell health.[4] Zhao, et al. have demonstrated specific lipid class dysregulation resulting from exposure to bisphenol F, a purportedly safe alternative to BPA.[5] There exist many other studies which also indicate lipid dysregulation resulting from diseases and exposures to foreign compounds.[4,6,7] However, there is much room for improvement considering that the lipidome is estimated to contain somewhere between 10,000-100,000 distinct lipid species, which is at least an order of magnitude larger than what we currently can detect reproducibly.[8]

There are several factors which drive the discrepancy between the number of theoretical lipids and how many we can identify experimentally. First, we currently fail to chromatographically separate lipid structural isomers. There is a great amount of heterogeneity between many lipid species with species sometimes only differing in the placement of a single site of unsaturation on a fatty acid chain. This issue can be ameliorated somewhat with improvements to separation technologies. Second, caveats regarding how lipids behave

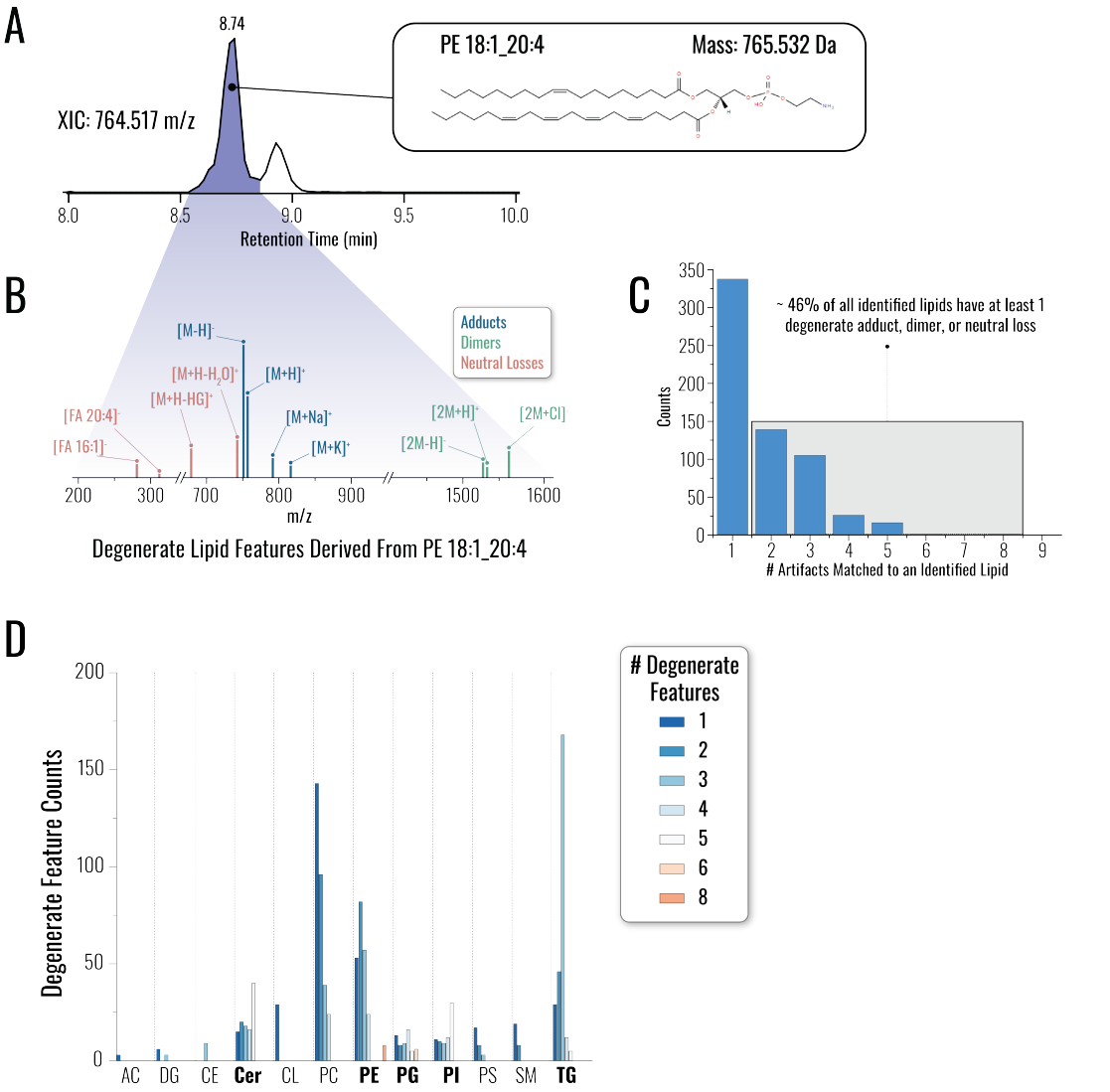


Figure 6.1: Degenerate Lipid Mass Features. **Panel A** shows an XIC of the $[M\text{-}H]^-$ peak of the lipid PE 18:1_20:4. **Panel B**, 11 total PE 18:1_20:4 lipid derivitives are mapped in a survey scan at 8.74 minutes. **Panel C**, of the ~700 identified lipid species in a single-shot experiment, 46% map to at least one degenerate artifact. **Panel D** shows a histogram of degenerate lipid features grouped by lipid class. Ceremides, phosphatidylethanolamines, phosphatidylglycerols, phosphatidylinositols, and triglycerides are particularly affected.

during liquid chromatography mass spectrometry analysis is not properly accounted for in the spectral acquisition software. Lipids, unlike peptides or metabolites, are particularly susceptible to adduction, dimerization, and in-source fragmentation. Similar to what is shown in **Figure 6.1B**, combinations of these artifacts can cause a single lipid species to appear as up to 15 distinct lipid features to the mass spectrometer, leading to a large sampling redundancy. This greatly limits the depth of our lipid analyses, as lowly-abundant lipid species are may not sampled in lieu of high-abundance, redundant artifacts. Missing these low abundance biomolecules can limit how deeply we mechanistically understand a biological system.

To overcome this limitation, we propose the development of a novel intelligent data acquisition strategy which can be loaded onboard our mass spectrometers to reduce the occurrences of redundant lipid sampling. Recent developments in proteomics have demonstrated the feasibility of identifying peptides in real time as they elute off the chromatographic column.[9] Depending on the results of each peptide search, an informed decision is be made real-time by the instrument (e.g. is this peptide sent off for additional analysis, or is it passed on in favor of the next species), permitting the creation of complex decision trees on-the-fly. A similar strategy can be implemented onboard our discovery lipidomic platforms to improve the efficiency of the instrumental duty cycle.

Specifically, using both new and previously collected experimental datasets collected either in-house or from public repositories, we will generate graph-based networks of lipid features via clustering on shared experimental properties like retention time, mass-to-

charge ratio, and isotopic distributions. We hypothesis features which cluster tightly are potentially non-informative artifacts originating from the same lipid species. We can then integrate these clusters into our cutting-edge lipid identification software suite, LipiDex.[3] LipiDex's algorithm will return a scored list of potential lipid structures. We will optimize LipiDex's search algorithm to not only conduct spectral matching on a millisecond time scale, but also return a list of potential lipid artifacts that are related to a just-sampled lipid species. Finally, we will inject these data into the mass spectrometer's data acquisition software. We will extract spectral data as it is collected and search each spectrum on-the-fly. If a high-scoring lipid identification is returned, we will inform the instrument to reject all other redundant lipid artifacts from further analysis, freeing the instrument to sample low-abundance lipid species that were previously ignored.

References

[1] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, no. 11, pp. 1191–1197, 2016.

[2] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The one hour yeast proteome," *Molecular and Cellular Proteomics*, vol. 13,

no. 1, pp. 339–347, 2014.

[3] P. D. Hutchins, J. D. Russell, and J. J. Coon, "LipiDex: An Integrated Software Package for High-Confidence Lipid Identification," *Cell Systems*, vol. 6, no. 5, pp. 621–625.e5, 2018.

[4] X. Han, "Lipidomics for studying metabolism," *Nature Reviews Endocrinology*, vol. 12, no. 11, pp. 668–679, 2016.

[5] C. Zhao, P. Xie, H. Wang, and Z. Cai, "Liquid chromatography-mass spectrometry-based metabolomics and lipidomics reveal toxicological mechanisms of bisphenol F in breast cancer xenografts," *Journal of Hazardous Materials*, vol. 358, pp. 503–507, 2018.

[6] H. Zhang, X. Shao, H. Zhao, X. Li, J. Wei, C. Yang, and Z. Cai, "Integration of Metabolomics and Lipidomics Reveals Metabolic Mechanisms of Triclosan-Induced Toxicity in Human Hepatocytes," *Environmental Science & Technology*, vol. 53, no. 9, pp. 5406–5415, 2019.

[7] I. Kania-Korwel, X. Wu, K. Wang, and H. J. Lehmler, "Identification of lipidomic markers of chronic 3,3′,4,4′,5-pentachlorobiphenyl (PCB 126) exposure in the male rat liver," *Toxicology*, vol. 390, pp. 124–134, 2017.

[8] G. Van Meer, "Cellular lipidomics," *EMBO Journal*, vol. 24, no. 18, pp. 3159–3165, 2005.

[9] D. K. Schweppe, J. K. Eng, D. Bailey, R. Rad, Q. Yu, J. Navarrete-Perea, E. L. Huttlin, B. K. Erickson, J. A. Paulo, and S. P. Gygi, "Full-featured, real-time database search-

ing platform enables fast and accurate multiplexed quantitative proteomics," *bioRxiv*, p. 668533, 2019.

COLOPHON

This document was typesetted with LaTeX 2ε using Overleaf. It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).