COMPUTATIONAL APPROACHES FOR IMPROVED IDENTIFICATION, QUANTITATION, AND INTERPRETATION OF MASS SPECTROMETRY-BASED "OMICS" DATA

by

Nicholas W. Kwiecien

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

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 11/18/16

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Chemistry
David J. Pagliarini, Associate Professor, Biochemistry
Jean-Michel Ané, Professor, Agronomy
Lloyd M. Smith, Professor, Chemistry
Anthony Gitter, Assistant Professor, Biostatistics and Medical Informatics

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Joshua J. Coon, for his mentorship and guidance throughout my graduate career at UW. I joined his lab having never written a single line of code which did not deter him in the slightest from encouraging me to learn how to program. At every turn he has provided access to all of the resources and tools I needed to be successful in my research efforts. I am extremely grateful for the trust he put in me to pursue ideas and deliver on research projects which I found meaningful and interesting. If nothing else, I have really enjoyed the work I did in his lab over the past four years, and will always look back fondly on this time in my life.

Throughout all of my time in graduate school, Professor Coon has always filled his lab with people who are remarkably creative, thoughtful, and hard-working. The work I did while at UW would not have been possible without the continued support and encouragement of members of the Coon Lab, past and present. I would like to thank Mike Westphall for his continued guidance and emotional support over these past four years. He was always willing to riff on new ideas and to help work through research challenges. I am tremendously thankful to Derek Bailey who was patient enough to help teach me the basics of programming, and the finer points of working with mass spectrometry data. Derek set an incredibly high

standard for what I should attempt to achieve in my work which I have always strived to meet.

I am very grateful to Alex Hebert who helped me to demonstrate my abilities to the lab early on by recruiting me to help him with data analysis. Throughout the years he has been, and continues to be, a wealth of knowledge on all things mass spec, and generally a pretty good guy. I was fortunate to sit next to, and be friends with Alicia Richards who similarly understood mass spec inside and out. She answered countless questions for me and was an awesome person to bounce ideas around with.

I would also like to thank Craig Wenger, Graeme McAllister, Amelia Peterson, Chris Rose, Anna Merrill, Catie Minogue, Greg Potts, Arne Ulbrich, Katie Overmeyer, Harald Marx, Evgenia Shishkova, Elyse Freiberger, Matt Rush, Emily Wilkerson, Paul Hutchins, Erin Weisenhorn, Dain Brademan, Anji Trujillo, Gary Wilson, Vanessa Linke, Kyle Connors, Rebecka Manis, Alan Higbee, and Jason Russell for all of their support, and for making the Coon Lab such a great place to work over the past four years.

Outside of our lab I was fortunate to work with some wonderful collaborators over the years. I would especially like to thank Jon Stefely who I worked with for many months on Y3K. Jon works as hard as any person I have ever met and he set

an incredible pace for that project. Working closely with Jon gave me no choice but to push myself harder and to try and bring more to the table. I learned an incredible amount and developed many new skills during Y3K, and I am very grateful to have been able to work alongside him throughout. Additionally, I would like to thank the numerous members of the Pagliarini lab who provided invaluable suggestions and assistance in testing the Y3K web portal. I am also very thankful to Trey Sato, Mark Keller, and Andrew Reidenbach for their assistance in developing and testing online data analysis tools.

I am thankful to my committee members Dr. Dave Pagliarini, Dr. Lloyd Smith, Dr. Jean-Michel Ané, and Dr. Anthony Gitter for their continued support over the years and for their service on my committee.

I am very thankful to my undergraduate research supervisors Dr. David M. Hercules and Dr. John A. McLean at Vanderbilt University. It was under their guidance that I first was exposed to mass spectrometry and its utility in proteomic and metabolomic applications. Together, they supported me throughout my undergraduate research career and convinced me to continue to explore mass spectrometry in graduate school. Looking back now, that was excellent advice which I am glad I took.

I would like to thank Tim Rhoads for his friendship, support, and sage advice

over the past four years. Tim's experience with the challenges and triumphs of graduate school was a tremendous benefit and helped me maintain perspective on every situation. It was a privilege to live and work with Tim who always was willing and eager to talk through new ideas and provide input whenever I was struggling. Tim went out of his way on multiple occasions to help me in my academic and personal life which I am extremely grateful for.

I was fortunate enough to meet Nick Riley on our visit weekend to UW where we immediately became friends. I cannot imagine a better person to have gone through this entire process with and owe so much to him. There are very few people I have met in life who have as much heart or who work as hard as Nick does. Every day over the last four years I have always challenged myself to achieve more in my research, because I knew Nick was going to do the same. I truly believe that I would not have achieved nearly as much in my graduate career if it was not for him, and I cannot possibly thank him enough for his friendship.

I would like to thank my parents Jim and Cindy for all of their love and support ver the past 27 years. I am unbelievably lucky to have such a great support system in my life. They have always stood behind me 100% and supported me unconditionally in whatever I chose to pursue—even if it was almost guaranteed that I would lose interest within a week. Both of my parents have been role models for me my entire

life, and with both having careers in science it really is no wonder that I ended up on the same path. I love you both very much, thank you for everything.

Finally, I would like to thank my girlfriend Jen Peotter for everything she has done for me. More times than I care to recall over the past few years I have found myself stressed, exhausted, and sometimes completely overwhelmed. No matter what, Jen took everything in stride. She was always there to support me, especially when I needed it most, and even if I didn't always deserve it. She has helped me to achieve more than I thought I could in graduate school and I couldn't be happier to have spent the past few years together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

2DE	Two-dimensional gel electrophoresis
4-HB	4-Hydroxybenzoic acid
4-HBz	4-Hydroxybenzaldehyde
AC	Alternating current
ACN	Acetonitrile
AGC	Automatic gain control
ALDH	Aldehyde dehydrogenase
ATP	Adenosine triphosphate
C#	C-Sharp, a Programming language
CHI	Chalcone isomerases
CI	Chemical ionization
CID	Collision-induced dissociation
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
CoQ	Coenzyme Q
CPU	Central processing unit
CSV	Comma-separated values
CV	Coefficient of variation
D3	Data-driven documents, a JavaScript code library

Da	Dalton, the atomic mass unit
DC	Direct current
DDA	Data dependent acquisition
DDA	Deoxyribonucleic acid
ECD	Electron capture dissociation
EI	Electron ionization
ETD	Electron transfer dissocation
ETF	Electron transfer flavoprotein
FAB	Fast atom bombardment
FALDH	Fatty aldehyde dehydrogenase
FDR	False discovery rate
FPR	False positive rate
FTICR	Fourier transform ion cyclotron resonance
FWHM	Full width at half maximum
GC	Gas chromatography
Δ*gene*	Yeast knockout strain
GO	Gene ontology
h	hour
HCD	Higher-energy C-trap dissocation

HRF	High-resolution filtering
HTCondor	High throughput condor
Hz	Hertz, inverse seconds
IRMPD	Infrared multiphoton dissociation
iTRAQ	Isobaric tags for relative and absolute quantification
JSON	JavaScript object notation
LC	Liquid chromatography
LFQ	Label free quantitation
m	mass
m/z	mass-to-charge ratio
M+H	Protonated molecular ion
MALDI	Matrix-assisted laser desorption/ionization
MCNA	Molecule covariance network analysis
MCN	Molecule covariance network
min	minute
MS	Mass spectrometry
MS^1	Survey mass analysis
MS^2	Tandem mass analysis
MSTFA	N-Methyl-N-(trimethylsilyl) trifluoroacetamide

MXP	Mitochondrial uncharacterized protein
NIST	National Institute of Standards and Technology
OD	Optical density
OMSSA	Open mass spectrometry search algorithm
ORF	Open reading frame
OxPhos	Oxidative phosphorylation
pABA	para-aminobenzoic acid
pABA$^-$	para-aminobenzoic acid depleted media
PC	Personal computer
PCA	Principal components analysis
PCR	Polymerase chain reaction
PHP	PHP, a programming language
PPAB	Aminated analog of 3-polyprenyl-4-hydroxybenzoate
PPHB	3-polyprenyl-4-hydroxybenzoate
ppm	Parts per million
PSM	Peptide–spectral match
Q	Quadrupole
QC	Quality control
QIT	Quadrupole ion trap

RC	Respiration competent
RD	Respiration deficient
RDR	Respiration deficiency response
RF	Radio frequency
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RP	Reverse phase
RPM	Rotations per minute
RT	Retention time
s	second
s.d.	Standard deviation
S/N	Signal-to-Noise
SILAC	Stable isotopic labeling of amino acids in cell culture
SQL	Structured query language
Th	Thomson, a unit of mass-to-charge
TIC	Total ion current
TMCS	Trimethylsilyl chloride
TMS	Trimethylsilyl
TMT	Tandem mass tags

TOF	Time-of-flight
TPR	True positive rate
UVPD	Ultraviolet photodissociation
w/v	weight-to-volume
WT	Wild-type
Y3K	Yeast three thousand
z	charge

ABSTRACT

The research described in this dissertation presents novel computational algorithms and strategies for (1) improving the assignment of molecular identities to analytes profiled by high-resolution gas chromatography-mass spectrometry (GC/MS), (2) performing relative quantitation of large sets of metabolites across expansive sets of mass spectrometry data files, (3) disseminating processed mass spectrometry data and *post hoc* statistical results in web-based platforms, and (4) monitoring mass spectrometer performance via a web-based data processing and analysis tool. An overview of the aforementioned computational strategies and developed software tools is presented in **Chapter 1**. A novel algorithm for leveraging accurate mass—afforded by high-resolution GC/MS systems—to discriminate between putative identifications assigned to profiled small molecules is described in **Chapter 2**. In **Chapter 3**, an algorithm and accompanying software suite designed to enable untargeted quantitation of small molecules across expansive sets of raw GC/MS data files is described. In **Chapter 4**, these algorithms are employed as part of a larger study wherein 174 single gene deletion strains of yeast were comprehensively profiled at the proteomic, metabolomic, and lipidomic levels. These multi-omic data were then integrated through various analysis planes in order to define functions of uncharacterized mitochondrial proteins. **Chapter 5** details numerous web-based

data visualization utilities developed for various projects designed to enable researchers to more rapidly interrogate MS data sets at depth. In **Chapter 6**, the development of a web-based mass spectrometry data deposition, processing, and visualization tool for automated quality control analysis is described.

Chapter 1

INTRODUCTION AND BACKGROUND

Mass Spectrometry

Mass spectrometry (MS) is a premier analytical technique for qualitative and quantitative molecular profiling. This technology has had widespread utility in numerous biological applications as it uniquely enables comprehensive protein, metabolite, and lipid analysis on a grand scale. Mass spectrometers are instruments used to generate these global molecular profiles. The primary function of a MS instrument is to ionize and measure the mass-to-charge (m/z) ratio of individual molecules. Instrument readouts are produced in the form of mass spectra, which inform both the chemical identity and relative abundance of analyzed entities. These MS data can provide researchers with an up-close look at the molecular composition of their system of interest—a perspective which few other analytical technologies can provide. Here, we will describe the fundamental processes which a mass spectrometer performs to generate these data—namely, ionization, mass analysis, and detection.

Ionization. Ionization is the process wherein charge is imparted to analyte molecules usually existing in either the liquid or gas phase. The choice of ionization technique is largely application dependent, and is often dictated by how a sample is delivered to the MS instrument. For the analysis of complex mixtures (i.e., complete proteomes, metabolomes, lipidomes, etc.) it is often advantageous to employ a

front-end chromatographic separation using either a gas or liquid chromatograph (GC/LC, respectively). GC is a useful technique for separating complex mixtures of volatile small molecules, such as metabolites. This mode of chromatography interfaces well with MS instrumentation and can be used in conjunction with electron ionization (EI) and chemical ionization (CI) techniques.

EI was one of the oldest ionization techniques to be used for MS analysis, dating back to the early 1900's[1,2]. In GC, a column is interfaced with an EI source, and eluting analyte molecules are bombarded with a beam of high-energy electrons. This process induces molecular fragmentation and formation of radical cations, which are then subjected to mass analysis. It is noteworthy, that these fragmentation processes are highly reproducible. Individual molecules give rise to characteristic sets of fragments across repeat experiments. Given that EI causes extensive fragmentation, this technique is considered a "hard" ionization process. Often it is desirable to analyze intact precursor molecules, which can be achieved using by alternative "soft" ionization techniques.

CI is a "soft" mode of ionization which is similarly amenable to GC/MS applications[3]. Here, a neutral reagent gas—often methane—is introduced to the MS at a concentration much higher than that of analyte molecules (10^3-10^4x). The reagent gas is ionized via interactions with high-energy electrons, and secondary reactions

cause formation of protonated species. Analyte molecules are then introduced to the MS, and are subsequently ionized via interactions with charged reagent gas molecules. The high concentration of reagent gas effectively shields analyte molecules from competing EI processes, making CI the preferred mode of ionization. CI processes provide a benefit over EI in that they often yield an intact pseudomolecular ion ($[M+H]^+$), albeit with less complementary fragmentation information. Both EI and CI techniques are useful for the analysis of small molecules, but lose efficacy for analytes >1000 Da.

For front-end LC separations, electrospray ionization (ESI) is the most popular and commonly employed ionization technique[4]. Here, a high voltage is applied to the tip of an LC column. This causes the eluent from the column to aerosolize and disperse in the form of highly charged droplets. These droplets are introduced into a vacuum region of a mass spectrometer where they gradually desorb until only charged molecules are left. The ionized molecules are then available for mass analysis inside the instrument. ESI is amenable for ionization of molecules of all sizes—from small molecule metabolites (<500 Da), to large macro protein complexes (800 kDa+)[5]—which has contributed to its extensive use in MS labs around the world.

Matrix-assisted laser desorption/ionization (MALDI)[6] and fast atom bombard-

ment (FAB)[7] are alternative ionization techniques, designed to be used in applications where samples exist in the solid phase. In MALDI, samples are mixed with a chemical matrix and deposited onto a metal plate, which is then injected into a MS under vacuum. A pulsed laser irradiates the surface of the plate causing desorption and ionization of molecules. This technique has been highly useful for MS imaging applications aiming to elucidate spatial molecular compositions of biological tissues. FAB is similar to MALDI with respect to sample preparation and delivery to the MS. However, in FAB molecular desorption and ionization is induced by firing a high energy beam of inert gas atoms (typically argon or xenon) at the sample surface.

Mass Analysis. Following ionization, mass analysis is the process wherein a MS separates molecules on the basis of *m/z*. Various mass analyzers are employed in commercial MS systems today, each of which affords different advantages with regards to resolution, mass accuracy, mass range, and speed of acquisition. The simplest mass analyzer is the time-of-flight (TOF)[8], which separates molecules longitudinally. Here, packets of ions are injected into and accelerated through an electric field of uniform strength, which imparts an equivalent amount of kinetic energy to each molecule. Each molecule's velocity is a function of its *m/z* ratio, which results in lighter molecules—of the same charge— travelling faster through

the field. By the same effect, molecules of equivalent mass, but higher charge, will also traverse the field more quickly. TOF analyzers are among the fastest scanning of all mass analyzers, and can afford relatively high resolving powers with sufficiently long flight tubes.

Quadrupole (Q)[9] mass analyzers consist of four parallel metal rods, arranged in a square, which form an open channel for ions to pass through. A direct current (DC) offset voltage is applied to two opposing rods, and an alternating current (AC) offset voltage is applied to the remaining two rods. Modulation of the voltages applied to the rod pairs allows selective transmission of ions and will selectively discriminate towards molecules having a specific *m/z* ratio. All other molecule's trajectories will become unstable and they will be ejected from the flight path. Quadrupole ion trap (QIT)[10] analyzers operate on similar principles. Here, electrodes are arranged to create a cell such that the electric field created by each can be used to trap ions. By modulating the potential across each electrode, ions of specific *m/z* can be selectively ejected and sent to a detector. Both Q and QIT analyzers are sensitive and fast-scanning, but do not match other analyzers in terms of resolving power.

Fourier transform ion cyclotron resonance (FTICR)[11] and Orbitrap[12] offer the highest resolution of all analyzers. Here, ions are injected tangentially into a magnetic or electrostatic field and excited to their cyclotron radius. Within an Orbitrap

mass analyzer, packets of ions having the same m/z will oscillate axially in phase. These measurements generate an image current—composed of sine waves from each discrete packet of ions–on the outer electrodes which can be converted into a mass spectrum by performing a Fourier transform. Alternative FTICR instrumentation operates on similar principles. Resolving power is a function of acquisition time in FTICR analyzers and increases in transient acquisition time lead to improved resolution.

Granted that analyzers afford different benefits in terms of resolving power, mass accuracy, mass range, and rate of acquisition, many instruments incorporate multiple analyzers in sequence, and are referred to as 'hybrids.' The advantage here, is that ions can be analyzed in different ways during the course of a single experiment. Tandem mass analysis (MS^2) is a technique wherein precursor molecules are isolated and fragmented, and those fragments are subsequently analyzed. Tandem mass analysis is useful for identification purposes as molecular fragmentation patterns serve as a characteristic signature of chemical identity.

Detection. Following ionization and mass analysis, a detector operates to register individual ions as electrical signals which can be converted into mass spectra. As previously mentioned, FTICR analyzers perform both mass analysis and detection simultaneously. Ion oscillations within the FTICR cell induce charges on internal

electrodes which are registered in an interferogram. Longer acquisition times increase both resolution and signal-to-noise (S/N) of individual peaks in the Fourier-transformed mass spectrum. For TOF, Q, and QIT mass analyzers, which separate molecules on the basis of *m/z*, a separate device must be employed to register ion signals. Faraday cups and electron multipliers are frequently used in mass spectrometers for this purpose. Faraday cups are simple detectors wherein ions strike a dynode surface causing an emission of electrons which induce a current that can be measured and recorded. No signal amplification is employed here which makes detection of low abundance ions challenging. Electron multipliers operate on similar principles, but these devices position multiple dynode surfaces in series to achieve amplification of signal. Ions strike the first dynode surface causing emittance of electrons directed towards a secondary dynode. These electrons strike the secondary surface causing emittance of an increased number of electrons directed towards a third dynode. This cascading process continues, and more electrons are released with each successive step. Eventually the electrons emitted from the final dynode are registered as a current and stored. Typically, electron multipliers can achieve an amplification of signal on the order of ~10^6.

MS-Based "Omic" Profiling

"Omics" is a neologism used to refer to the biological fields of study devoted to comprehensively characterizing a specific class of biomolecules in a system. For instance, genomics—the study of an organism's genetic material—focuses on determining which DNA molecules are present in a particular sample, and at what abundances. By comparing genomic maps between samples exposed to different treatments, researchers can identify deviations in molecular profiles. These characteristic deviations can then be leveraged to drive hypothesis generation and biochemical discovery. The central dogma of biology states that genetic information is coded into DNA, which is subsequently transcribed into RNA, and finally translated into functional protein units. Proteins carry out myriad biochemical reactions in the cell that yield smaller metabolic byproduct molecules called metabolites. Lipids are yet another class of essential biomolecules, and are synthesized for the purpose of signaling, storing energy, and maintaining cell membranes. Proteins, metabolites, and lipids are all amenable to analysis by MS, which has rapidly established itself as the leading analytical tool for proteomic, metabolomic, and lipidomic study.

Proteomics. MS-based proteomic analyses are typically carried out using a bottom-up approach. In a traditional discovery-oriented, bottom-up experiment, proteins

are first extracted from a sample of interest and then digested into smaller peptide subunits via proteolytic enzymes. The digestion of proteins into smaller pieces improves chromatographic separations—as compared to analysis of intact proteins—and simplifies downstream data processing. These complex mixtures of peptide species are loaded onto a front-end LC column, chromatographically separated, and sprayed into a MS. Throughout this LC gradient, peptides are analyzed in MS^1 survey scans and precursors are selected for subsequent MS^2 analysis in a data dependent (DDA) fashion. DDA strategies employ algorithms which select precursors on the basis of abundance and observed isotope pattern—a signature of peptidic species. Peptides selected for MS^2 analysis are isolated, fragmented—typically via a collisional (collision induced dissociation [CID][13], higher-energy c-trap dissociation [HCD][14]), photodissociative (infrared multiphoton photodissociation [IRMPD][15], ultraviolet photodissociation [UVPD][16]), or electron-based (electron transfer dissociation [ETD][17], electron capture dissociation [ECD][18]) dissociation technique, and then analyzed. Following MS^2 analysis, all selected precursors are placed on a timed exclusion list to avoid resampling of the same peptides. This procedure enables the MS to efficiently allocate scan time such that the largest possible pool of distinct precursor molecules can be analyzed per experiment.

Following data acquisition, acquired MS^2 spectra are compared against an *in sil-*

ico digest of a protein sequence database to assign identifications. Given the repeat polymeric structure of peptides, fragmentation patterns are easy to predict computationally. This *in silico* fragmentation information, coupled with a measurement of precursor mass, can be leveraged to assign an amino acid sequence to analyzed peptides[19]. False discovery rate (FDR) can automatically be controlled here by conducting searches against a concatenated target-decoy (forward and reverse-sense sequence) database[20]. Each identified peptide can be quantified by summing the total abundance of MS^1 peaks measured at that particular *m/z*; this quantitation approach is referred to as 'label-free'. Alternative quantitative approaches which make use of isotopically heavy amino acid labels (stable isotope labeling with amino acids in cell culture [SILAC][21], neutron encoded mass signatures [NeuCode][22]), or isobaric mass tags (tandem mass tags [TMT][23], isobaric tags for relative and absolute quantitation [iTRAQ][24]) can also be used for peptide quantitation, although these typically require modified data acquisition routines. In all cases, quantified peptides are aggregated to form consensus protein abundances, which can then be used for comparative proteomic analyses.

Metabolomics and Lipidomics. Although metabolites and lipids are chemically distinct classes of molecules, the approaches used for global profiling—analysis of as many chemical entities as possible—of these species are fundamentally similar.

In these paradigms, a front-end chromatograph—either GC or LC for metabolomics, and typically LC for lipidomics—is used to resolve complex mixtures of extracted metabolites and lipids. Eluting species are then sprayed into a MS instrument for subsequent analysis. For LC/MS experiments, a DDA-like approach is often employed wherein all molecules are measured in an MS^1 survey scan, and abundant precursors are selected for MS^2 analysis. In GC/MS applications, EI is the most commonly employed ionization technique. Given that EI induces molecular fragmentation upon ionization, typically only MS^1 survey scans are acquired during these experiments. Following data acquisition, spectral deconvolution algorithms can be employed to extract spectra containing fragments derived exclusively from a singular precursor. In both cases, fragmentation spectra can be used for identification purposes.

Unlike peptides, metabolites do not have repeat polymeric structures which readily lend themselves to *in silico* fragment generation. For metabolomic analyses, experimentally-derived fragmentation spectra are often compared against libraries of reference spectra for identification purposes. The majority of publicly available metabolite reference libraries contain spectra from the analysis of pure reference standards (NIST[25], Wiley[26]). Lipid fragmentation patterns are slightly more predictable, and extensive databases of theoretical fragmentation spectra have recently

been published[27]. For both metabolomic and lipidomic analyses, quantitative information can be obtained by extracting MS^1 chromatographic features, as is done in the previously described 'label-free' approach.

Challenges in MS-Based "Omic" Profiling

Given the depth and breadth of molecular coverage it can provide, mass spectrometry has positioned itself as the analytical tool of choice for "omic" profiling studies of small effector biomolecules. Recent advances in MS have enabled unprecedented throughput and have placed data acquisition speeds on a timescale commensurate with orthogonal genomic and transcriptomic profiling technologies. Despite these rapid advances in MS profiling capabilities, a number of challenges to routine profiling persist—particularly with regards to computational analysis. Assignment of confident identifications to profiled metabolite species has proven difficult, with many studies reporting identification of only a fraction of all profiled species. Streamlined metabolite quantitation software tools remain underdeveloped, particularly for GC/MS applications. The generation of increasingly large MS data sets creates a burden to data exploration and interpretation, and functional analysis tools are similarly underdeveloped. Finally, unified tools for monitoring MS instrument performance and assisting in troubleshooting are limited, and the

larger community lacks a standard solution to this widespread problem. The work described in this dissertation has been directed to address each of these issues.

Identification of Metabolites. Metabolomics in general suffers from a relatively low rate of compound identification. In a typical discovery experiment, an average of only ~30% of monitored species are assigned a confident identification[28]. This is greatly reduced from a typical proteomics experiment where we note (empirically) that ~50% of analyzed peptides or more are routinely identified. Metabolite identifications are assigned based on comparisons to spectral libraries (NIST[25], Wiley[26]) which, while expansive, are incomplete. Often, the absence of an appropriate reference spectrum in a library precludes identification altogether. Furthermore, we note that most reference libraries consist of spectra acquired exclusively on unit resolution mass spectrometers. The advent of high-resolution Orbitrap and FTICR mass spectrometry systems provides metabolomic analysts with a new dimension of information in accurate mass. Researchers were quick to leverage this in LC/MS applications wherein molecular assignments are based in part upon intact precursor mass. This increase in mass accuracy affords a substantial reduction in putative precursors considered, which generally translates to a higher overall identification rate. Conversely, there was not an obvious way to integrate this information into traditional GC/MS workflows, where identifications are assigned based on

comparisons to unit resolution reference spectra.

We note that the mass accuracy afforded by high-resolution GC/MS systems—such as the GC-Orbitrap[29–31]—yields measurements which are precise enough to be annotated with an exact chemical formula. We expanded upon this concept and developed an approach for leveraging accurate mass to discriminate against false matches, when used in conjunction with traditional spectral matching. In our approach—called High-Resolution Filtering (HRF)[32]—we extract fragmentation spectra from high-resolution GC-Orbitrap raw data files, and submit these spectra for matching against unit resolution reference libraries. Each submitted spectrum is returned with a set of putative identifications, complete with associated chemical formulas. We then generate all non-repeating combinations of atoms from these formulas and attempt to match these sub-formulas (theoretical fragments) to our spectrum using exact mass. The proportion of measured signal that can be annotated with a sub-formula is then reported as a metric of plausibility of the assigned identification. This approach is desirable as it capitalizes on the expansive reference libraries currently in existence. Furthermore, it exploits the richness of acquired GC-Orbitrap data to discriminate between true and false matches with higher fidelity. We also note that while the current implementation of HRF is built around spectral matching, this metric can be used independently to test the plausi-

bility of formula-spectrum matches in a non-biased manner. The HRF algorithm is described in detail in **Chapter 2**.

Untargeted Metabolite Profiling. Considering all MS-based "omic" profiling technologies, quantitation routines are most well-developed for proteomic analyses. Researchers can utilize a variety of techniques from label-free, to metabolic or chemical tagging, to acquire quantitative protein information. Furthermore, a panel of software tools is available to assist in extracting abundance measurements[33–35]. In most of these quantitation packages, identifications are required in order for quantitative information to be reported. This paradigm is challenging for metabolomic applications where often many biologically-relevant features go unidentified. Independent of identifications, profiling conserved sets of metabolites across experiments can be useful for elucidation of phenotypic similarities. We note that software solutions which enable untargeted metabolomic quantitation remain under-developed, notably for high-resolution GC/MS applications.

EI-GC/MS analysis is challenging in general as many signals arise from singular chemical entities. Additionally, most sample preparation procedures utilize solvents and a derivatization reagent—to increase molecular volatility—which adds unwanted background chemicals to already complex mixtures[36]. Quantitative profiling of phenotypes here requires aggregation of signals derived from singular

precursors, in addition to selection against those signals arising from background chemical noise. These operations collectively enable the isolation of a representative set of biologically-meaningful signals that can be quantified and compared across samples. Ideal software tools would perform all necessary feature extraction and background subtraction steps to identify this set of "targets," in addition to quantitation and normalization of features detected across MS data files. Furthermore, it is desirable that these results be packaged in a format which lends itself to rapid data visualization in order to expedite downstream data analysis. In **Chapter 3** we describe work on a suite of software tools designed to perform all of the steps required for untargeted GC/MS metabolomic profiling.

Interpretation of Large MS data sets. Over the past twenty years, mass spectrometry has experienced massive technological advances which are reflected in "omic" profiling studies of the day. In 1996, the Mann group sequenced ~150 yeast proteins using a 2D gel electrophoretic separation (2DE), followed by joint MALDI and LC/MS analysis[37]. In 2001, Yates and colleagues identified 1483 yeast proteins in 68 hours of LC/MS analysis[38]. In 2006, the Mann group again pushed the needle, and identified 2003 yeast proteins in only 48 hours[39]. In 2013, Zubarev and colleagues reported detection of ~5,000 proteins in human cell lines in only four hours of MS analysis time[40]. One year later in 2014, the Coon research group

published a groundbreaking study where the complete yeast proteome (~4000 proteins) was sequenced in just over one hour of MS analysis time[41]. Collectively, these recent studies have signaled a paradigm shift for proteomics. Now, comprehensive profiling of proteomes is possible in hours, not days, which opens the door to systems-level studies where the analysis of hundreds, or even thousands, of samples is considered routine.

The rapid expansion of MS data sets has been met with enthusiasm by the larger biochemical research community. However, the increased flux of data presents new challenges in dissemination, analysis, and interpretation of processed results. Tools for visualization, exploration, and sharing of large-scale MS data sets—particularly multi-omic data sets—remain underdeveloped. Web-based data exploration solutions have become popular in other areas of science and have helped to alleviate the burden of manual data analysis and file sharing. The UCSC Genome Browser is an online tool, (hosted by the University of California—Santa Cruz) which provides visualization of data from numerous large-scale genomic profiling studies[42]. This utility also supports functionality for users to upload their own data, compare against other data sets, and share results with collaborators. Tools such as the UCSC Genome Browser are ideal as they enable non-programmers to perform systems-level computational analyses without having to write any code. Furthermore, these

tools can be accessed from any web browser without the need for additional data or software downloads.

Web-based data exploration solutions are likely the wave of the future for the reasons mentioned above. However, the appearance of online analysis tools for MS-based "omic" applications in the literature remains limited. Construction of web portals requires computer programming and web development expertise—which many researchers lack—that creates a barrier to construction of these utilities. We contend that development of software tools which can convert MS-based "omics" data sets into interactive web portals are both critical and timely. In **Chapter 5** we discuss the creation and publication of two interactive web resources for exploration of large-scale "omics" data sets developed in-house[43,44]. Additionally, we describe the development of a prototype web-based platform, which facilitates the codeless generation of interactive web portals using MS peak tables as inputs.

Quality Control Monitoring Solutions. One of the largest practical challenges in MS-based "omic" profiling is the routine collection of high quality data. Mass spectrometers are complex and sensitive instruments comprised of numerous mechanical components. All of these integral pieces must perform specific functions, in concert, in order to acquire measurements. The loss of function in any singular component can diminish MS performance and dramatically influence data quality.

As such, it is critical that MS performance be routinely monitored so that lapses in expected operation can be corrected for. Within the larger proteomic, metabolomic, and lipidomic communities, there is not a widely accepted quality control (QC) analysis routine. Rather, individual labs often develop their own QC procedures, which are used to inform instrument maintenance needs.

In our own lab we employ a QC protocol for monitoring performance of LC/MS systems—dedicated to proteomic analysis—where tryptic digests of whole cell yeast lysates are analyzed on a weekly basis. From these experiments, the number of uniquely identified peptides is reported as a metric of performance. Granted that sampling of as many peptide species as possible is one of the overarching goals of proteomic analysis, this simple metric is a useful proxy for overall system performance. However, during times of subpar instrument operation, this metric does little to inform the root cause of instrument issues.

Optimal QC procedures—both data acquisition and analysis—are routinely executed in exactly the same manner in an effort to diminish variation. Static procedures, such as our own QC, often lend themselves to automation. We recognized that by developing completely automated computer scripts, we could rapidly extract unique peptide counts from each QC data file and reduce the workload for QC data analysts. Furthermore, automated data analysis afforded the opportunity

to extract additional orthogonal metrics of performance, which could be reported and stored in a historical record. These added metrics can also be used during troubleshooting to help localize the source of performance issues. In **Chapter 6** we describe the development of a web-based quality control analysis and monitoring tool—The Yeast Controller—which supports automated QC data upload, processing, and visualization.

References

[1] a. Dempster, "A new Method of Positive Ray Analysis," 1918.

[2] W. Bleakney, "A new method of positive ray analysis and its application to the measurement of ionization potentials in mercury vapor," *Physical Review*, vol. 34, pp. 157–160, 1929.

[3] M. S. B. Munson and F. H. Field, "Chemical Ionization Mass Spectrometry. I. General Introduction," *Journal of the American Chemical Society*, vol. 88, pp. 2621–2630, 1966.

[4] 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, pp. 64–71, 1989.

[5] R. J. Rose, E. Damoc, E. Denisov, A. Makarov, and A. J. R. Heck, "High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies.," *Nature methods*, vol. 9, pp. 1084–6, 2012.

[6] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, "Matrix-assisted ultraviolet laser desorption of non-volatile compounds," *International Journal of Mass Spectrometry and Ion Processes*, vol. 78, pp. 53–68, 1987.

[7] M. Barber, R. S. Bordoli, R. D. Sedgwick, and A. N. Tyler, "Fast atom bombardment of solids (F.A.B.): a new ion source for mass spectrometry," *Journal of the Chemical Society, Chemical Communications*, p. 325, 1981.

[8] W. E. Stephens, "A Pulsed Mass Spectrometer with Time Dispersion," *Proceedings of the American Physical Society*, vol. 69, no. 11,12, p. 691, 1946.

[9] W. Paul and H. Steinwedel, "Ein neues Massenspektrometer ohne Magnetfeld," 1953.

[10] G. Stafford, P. Kelley, J. Syka, W. Reynolds, and J. Todd, "Recent improvements in and analytical applications of advanced ion trap technology," *International Journal of Mass Spectrometry and Ion Processes*, vol. 60, pp. 85–98, 1984.

[11] M. Comisarow and A. Marshall, "Fourier transform ion cyclotron resonance spectroscopy," *Chem. Phys. Lett.*, vol. 25, pp. 282–283, 1974.

[12] A. Makarov, "Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis," *Analytical Chemistry*, vol. 72, pp. 1156–1162, 2000.

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

[14] 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, pp. 709–12, 2007.

[15] 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, pp. 2809–2815, 1994.

[16] W. D. Bowers, S. S. Delbert, R. L. Hunter, and R. T. McIver, "Fragmentation of oligopeptide ions using ultraviolet laser radiation and Fourier transform mass spectrometry," *Journal of the American Chemical Society*, vol. 106, pp. 7288–7289, Nov. 1984.

[17] J. E. P. Syka, J. J. Coon, M. J. Schroeder, J. Shabanowitz, and D. F. Hunt, "Peptide and protein sequence analysis by electron transfer dissociation mass spectrome-

try.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 9528–33, 2004.

[18] R. Zubarev, N. L. Kelleher, and F. W. McLafferty, "Electron capture dissociation of multiply charged protein cations. A …," *J. Am. Chem. Soc*, vol. 120, pp. 3265–3266, 1998.

[19] 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, pp. 976–989, 1994.

[20] 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, pp. 207–214, 2007.

[21] S.-E. Ong, "Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics," *Molecular & Cellular Proteomics*, vol. 1, pp. 376–386, 2002.

[22] A. S. Hebert, A. E. Merrill, D. J. Bailey, A. J. Still, M. S. Westphall, E. R. Strieter, D. J. Pagliarini, and J. J. Coon, "Neutron-encoded mass signatures for multiplexed proteome quantification.," *Nature methods*, vol. 10, pp. 332–4, 2013.

[23] 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, pp. 1895–1904, 2003.

[24] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin, "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.," *Molecular & cellular proteomics : MCP*, vol. 3, pp. 1154–69, Dec. 2004.

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

[26] "Wiley Registry of Mass Spectral Data," 2010.

[27] T. Kind, K. H. Liu, Y. Lee do, B. DeFelice, J. K. Meissen, and O. Fiehn, "LipidBlast in silico tandem mass spectrometry database for lipid identification," *Nat Methods*, vol. 10, pp. 755–758, 2013.

[28] O. Fiehn, J. Kopka, R. N. Trethewey, and L. Willmitzer, "Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry.," *Analytical chemistry*, vol. 72, pp. 3573–3580, 2000.

[29] A. C. Peterson, G. C. McAlister, S. T. Quarmby, J. Griep-Raming, and J. J. Coon, "Development and characterization of a GC-enabled QLT-orbitrap for high-resolution and high-mass accuracy GC/MS," *Analytical Chemistry*, vol. 82, pp. 8618–8628, 2010.

[30] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, 2014.

[31] A. Peterson and A. Balloon, "Development of a GC/Quadrupole-Orbitrap mass spectrometer, part II: new approaches for discovery metabolomics," *Analytical …*, vol. 86, pp. 10044–51, Oct. 2014.

[32] N. W. Kwiecien, D. J. Bailey, M. J. P. 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, pp. 8328–35, Aug. 2015.

[33] C. D. Wenger, D. H. Phanstiel, M. V. Lee, D. J. Bailey, and J. J. Coon, "COM-

PASS: A suite of pre- and post-search proteomics software tools for OMSSA," *Proteomics*, vol. 11, pp. 1064–1074, 2011.

[34] 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, pp. 1367–72, 2008.

[35] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss, "Skyline: An open source document editor for creating and analyzing targeted proteomics experiments," *Bioinformatics*, vol. 26, pp. 966–968, 2010.

[36] D. R. Knapp, "Handbook of Analytical Derivatization Reactions," *John Wiley Sons New York*, p. 741, 1979.

[37] A. Shevchenko, O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, and M. Mann, "Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, pp. 14440–5, 1996.

[38] M. P. Washburn, D. Wolters, and J. R. Yates, "Large-scale analysis of the yeast

proteome by multidimensional protein identification technology.," *Nature biotechnology*, vol. 19, pp. 242–7, 2001.

[39] L. M. F. de Godoy, J. V. Olsen, G. A. De Souza, G. Li, P. Mortensen, and M. Mann, "Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system.," *Genome biology*, vol. 7, p. R50, 2006.

[40] M. Pirmoradian, H. Budamgunta, K. Chingin, B. Zhang, J. Astorga-Wells, and R. a. Zubarev, "Rapid and deep human proteome analysis by single-dimension shotgun proteomics.," *Molecular & cellular proteomics : MCP*, vol. 12, pp. 3330–8, 2013.

[41] 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, pp. 339–347, 2014.

[42] W. Kent, C. Sugnet, and T. Furey, "The human genome browser at UCSC," *Genome …*, pp. 996–1006, 2002.

[43] 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*, pp. 1–11, Sept. 2016.

[44] H. Marx, C. E. Minogue, D. Jayaraman, A. L. Richards, N. W. Kwiecien, A. F. Sihapirani, 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*, Oct. 2016.

Chapter 2

HIGH-RESOLUTION FILTERING FOR IMPROVED SMALL MOLECULE IDENTIFICATION VIA GC/MS

This chapter has been published:

Kwiecien NW, Bailey DJ, Rush MJP, Cole JS, Ulbrich A, Hebert AS, Westphall MS, Coon JJ. *High-Resolution Filtering for Improved Small Molecule Identification via GC/MS.* Analytical Chemistry. **2015**, *87*, 8328-8335.

Abstract

Gas chromatography-mass spectrometry (GC/MS) has long been considered one of the premier analytical tools for small molecule analysis. Recently, a number of GC/MS systems equipped with high-resolution mass analyzers have been introduced. These systems provide analysts with a new dimension of information—accurate mass measurement to the third or fourth decimal place; however, existing data processing tools do not capitalize on this information. Beyond that, GC/MS spectral reference libraries, which have been curated over the last several decades, contain almost exclusively unit resolution MS spectra making integration of accurate mass data dubious. Here we present an informatic approach, called High-Resolution Filtering (HRF), which bridges this gap. During HRF, high-resolution mass spectra are assigned putative identifications through traditional spectral matching at unit resolution. Once candidate identities have been assigned, all unique combinations of atoms from these candidate precursors are generated and matched to *m/z* peaks using narrow mass tolerances. The total amount of measured signal that is annotated is used as a metric of plausibility for the presumed identification. Here we demonstrate that the HRF approach is both feasible and highly specific towards correct identifications.

Introduction

Gas chromatography-mass spectrometry (GC/MS) is a premier analytical tool for small molecule analysis[1–3]. Highly reproducible chromatographic separations combined with conserved molecular fragmentation lend this technique to both targeted and discovery assays, and has become particularly useful in the area of metabolite profiling[4,5]. Since the metabolome is closest to phenotype, metabolic profiling has great potential to propel biomedical research and is quickly emerging as a field of interest for both systems biologists and clinical researchers[6,7]. The ability to rapidly and comprehensively monitor metabolites will doubtless facilitate basic research into disease pathogenesis and also provide new opportunities for disease diagnosis. Moreover, metabolite screens are highly desirable in the clinical setting as they often rank among the least invasive biological assays. As an emergent field there is critical need for the development of advanced tools and technologies to enable deeper small molecule profiling in shorter time spans.

In traditional discovery experiments, volatile analytes are separated by GC and ionized using electron ionization (EI) prior to mass analysis. EI is a "hard" ionization technique and causes molecules to fragment in characteristic patterns. Spectra containing fragments from individual analytes, which may or may not include an intact molecular ion, are extracted and then compared to databases of

unit-resolution reference spectra[8]. Matches with sufficiently high spectral similarity are often presumed to be correct identifications. Identifying all of the observed spectral features resulting from a GC/MS experiment is a formidable challenge[9,10], so often the majority of features often remain unidentified. For those compounds where putative identifications have been assigned, subsequent validation often necessitates analysis of a pure reference standard. This process is laborious, especially when considering that for many spectral features there exist a large number of putative identifications. As such, any auxiliary information which can be used to discriminate between candidate precursors is highly valuable[11].

Unit resolution GC/MS instruments have been, and continue to be, the most widespread and commonly used mass spectrometers in the world. Given that, the largest publically available reference libraries are comprised of spectra exclusively acquired on these systems[12,13]. In the last few years, however, several GC/MS systems possessing mass analyzers capable of high-resolution and accurate mass measurement have become commercially available—i.e., time-of-flight and, most recently, Orbitrap. Despite these exciting technological advances and their potential impact on metabolomic research, data analysis tools have remained largely unchanged[14–17]. We conclude that, if coupled with novel informatic capability, this new generation of GC/MS systems offers considerable opportunity to drive

small molecule discovery. This nascent promise is reminiscent of the revolution that occurred in LC/MS-based proteomics following the introduction of high-resolution/accurate mass measurement. In this case, existing peptide-spectral matching algorithms were easily adapted to achieve a concomitant reduction in search space while affording increased precursor/product ion matching specificity; unfortunately, leveraging the specificity enabled by accurate mass GC/MS data with existing small molecule algorithms is not straightforward. The major EI reference databases comprise unit resolution spectra, precluding the ability to directly compare measured exact masses against their reference counterparts. An alternative route is to generate theoretical EI spectra *in silico*, though this has proven to be exceptionally challenging[18–20]. Of course, another approach is to generate new accurate mass libraries which would ostensibly allow for increased discrimination against spurious matches as fragments which are nominally the same but not equivalent within a narrow mass tolerance would no longer be matched. This increased specificity in spectral matching would hopefully make it easier to identify correct matches. Generating new reference databases is an admirable goal but one that will take years, if not decades, to achieve given that current spectral libraries have been compiled over the past fifty years from hundreds of thousands of individual analyses.

Here we describe a new approach to harness the existing unit resolution EI mass spectral databases while simultaneously exploiting the accurate mass measurement capabilities of high-resolution GC/MS systems. In this method accurate mass GC/MS data is searched via spectral matching to existing unit resolution EI spectral libraries as normal. Next, EI-MS top scoring putative identifications are tested for plausibility based on comparison of the experimentally measured accurate mass fragments to combinatorially generated theoretical fragments constrained by the atomic composition of the assigned precursor. This method avoids the pitfalls of theoretical EI spectral prediction by simply generating and testing all possible combinations of atoms, as theoretical fragments, in a precursor. We demonstrate that although this method makes minimal approximations it remains highly specific toward correct precursor identifications. By enabling discrimination between candidate molecular precursors on the basis of both measured fragmentation profiles and accurate mass, this method effectively bridges the current technology gap between high-resolution spectral acquisition and unit resolution mass spectral libraries.

Experimental Section

Materials and Reagents Unless otherwise specified all standard reference materials were purchased from Sigma-Aldrich (St. Louis, MO) with the exception

of the 37 pesticide reference standards analyzed which were contained in the *Organonitrogen Pesticide Mix #1 – EPA Method 525.2* and purchased from Restek (Bellefonte, PA). Methanol, ethyl acetate, acetone, hexane, dichloromethane, and isopropyl alcohol reagents were also purchased from Sigma-Aldrich. The N-methyl-N-trimethylsilytrifluoroacetamide with 1% trimethylchlorosilane derivatization reagent (MSTFA + 1% TMCS) was purchased from Pierce Biotechnology (Rockford, IL). Compressed gases (methane, helium, and nitrogen) were ultrahigh purity grade and purchased from Airgas (Madison, WI). 200 mg Clean Screen® Extraction Columns were purchased from United Chemical Technologies (Bristol, PA).

Sample Preparation and GC/MS Acquisition Stock solutions of the reported standards were prepared individually at a concentration of 1 mg/mL in appropriate solvents. Standards were processed in batches containing ~5-10 individual analytes. The EPA 525.2 pesticide mixture was diluted from 500 μg/mL to a working concentration of 3 ng/μL in acetone prior to mass spectral analysis. For the drug-spiked urine experiments, stock solutions of all drugs were first prepared at 1 mg/mL in methanol. These stock solutions were combined and diluted (again in methanol) to appropriate concentrations. For each gradient data point, 100 μL of the drug mixture was added to raw urine prior to extraction using the 200 mg Clean Screen extraction columns. Acidic and basic drug/metabolite fractions were extracted

according to manufacturer protocols[21]. Yeast metabolites were extracted by first washing cultured cells with buffered saline and submerging into a precooled 1.5 mL plastic tube containing 2:2:1 acetonitrile/methanol/H_2O mixture. For all materials (not including the pesticide mixture) 25 μL aliquots were resuspended in 25 μL of pyridine and vortexed. 25 μL of N-methyl-N-[trimethylsilyl]trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) was added and samples were incubated at 60° C for 30 minutes. All samples were analyzed using a GC/MS instrument comprising a Trace 1310 GC coupled to a Q Exactive Orbitrap mass spectrometer. For the yeast metabolite extracts a linear temperature gradient ranging from 50 °C to 320 °C was employed spanning a total runtime of 30 minutes. Analytes were injected using a 1:10 split at a temperature of 275 °C and ionized using electron ionization (EI). The mass spectrometer was operated in full scan mode using a resolution of 30,000 $(m/\Delta m)$ relative to 200 m/z. Instrumental parameters and specifications for all other experiments are provided in the Supporting Information. All MS experiments utilized Automatic Gain Control (AGC)[22–24] and all data was acquired in profile mode.

GC/MS Data Processing All GC/MS data processing was done using in-house algorithms designed to facilitate spectral deconvolution, spectral matching against a unit resolution reference database, and high-resolution filtering. The details of

each algorithmic component are described at length in the Supporting Information. Briefly, following mass spectral acquisition deconvolved spectra were extracted from raw data files. A pseudo-unit resolution copy of each spectrum was made by combining the intensities of peaks falling in the same nominal mass range, setting the measured *m/z* to the nearest integer value, and normalizing peak intensities relative to the base peak (set to 999). All 212,961 unit resolution reference spectra in the NIST 12 MS/EI Library were exported to a .JDX file through the NIST MS Search 2.0 program and converted to a format suitable for matching against acquired Q Exactive CG spectra. Extracted spectra were submitted for database searching and spectral similarity was measured using the following dot product equation:

$$100 \times \frac{\sum (m/z[\text{Intensity}_{\text{experimental}} \times \text{Intensity}_{\text{reference}}]^{0.5})^2}{\sum (\text{Intensity}_{\text{experimental}} \times m/z) \sum (\text{Intensity}_{\text{reference}} \times m/z)}$$

Following candidate identification retrieval the high-resolution filtering algorithm was employed by first generating all unique atomic combinations from a given precursor using the most abundant isotope of each considered atom. Starting with the smallest measured *m/z* peak, peaks were matched to theoretical fragments using a narrow ± 10 ppm tolerance centered around the *m/z* value. To account for isotopic clusters a variant of each matched theoretical fragment was created containing substituted heavy isotopes was placed back on the list of all candidate theoretical fragments. This process was repeated until every measured peak in a

given spectrum had been considered. The total amount of measured signal that could be annotated as calculated by:

$$\frac{\sum(\text{Intensity} \times m/z)_{\text{annotated}}}{\sum(\text{Intensity} \times m/z)_{\text{observed}}}$$

was returned in the form of an HRF score.

Results and Discussion

The HRF method is founded on one central tenet – all *m/z* peaks in a pure EI spectrum are derived from a single molecular precursor and, therefore, contain a subset of the atoms from the molecular precursor. For example, the EI mass spectrum of 3-methyl-3-hexanol (C7H16O) exhibits prominent features at *m/z* values 73, 87, and 101[25]. Expert annotation of this spectrum revealed the chemical identity of these fragments as C_4H_9O, $C_5H_{11}O$, and $C_6H_{13}O$, respectively[25]. Note each of these formulas is a subset of the molecular precursor, supporting our guiding supposition. Without regard for chemical structure feasibility, there are 271 unique atomic combinations of the parent molecule $C_7H_{16}O$. First consider the *m/z* peak at 73, only three of these combinations have this nominal mass – C_6H, C_5H_{13}, and C_4H_9O; however, only one (C_4H_9O) has an exact mass within ± 10 ppm of the correct annotation. Such is also the case for the other expertly annotated fragments. Given

that we can now routinely measure all *m/z* peaks in an EI mass spectrum with low ppm mass accuracy, we implement this annotation strategy on a large scale.

Figure 2.1 presents an outline of the HRF workflow. The process consists of three main steps: deconvolution, spectral matching, and high-resolution filtering. Spectral deconvolution is a standard part of processing GC/MS data; however, accurate mass is highly advantageous as it reduces, or eliminates, interference between nominally isobaric fragments. Extracted spectral features are subsequently grouped based on corresponding elution apex and a spectrum containing only fragments arising from a singular precursor is derived from each group **(Figure 2.1a-c)**. Next, by rounding accurate mass *m/z* peaks to the nearest integer value, a pseudo-unit resolution copy of each spectrum is created and then submitted for spectral matching against a unit resolution reference database. The intent is to retrieve candidate identifications based on spectral similarity. These steps represent a traditional workflow for spectral assignment in a discovery-based GC/MS experiment. In the HRF method, this workflow is further augmented to leverage accurate mass for discrimination between putative identifications.

The HRF method attempts to annotate every measured *m/z* peak in an EI mass spectrum using some combination of atoms from a putative precursor's chemical formula. The amount of total ion current that can be successfully annotated can be

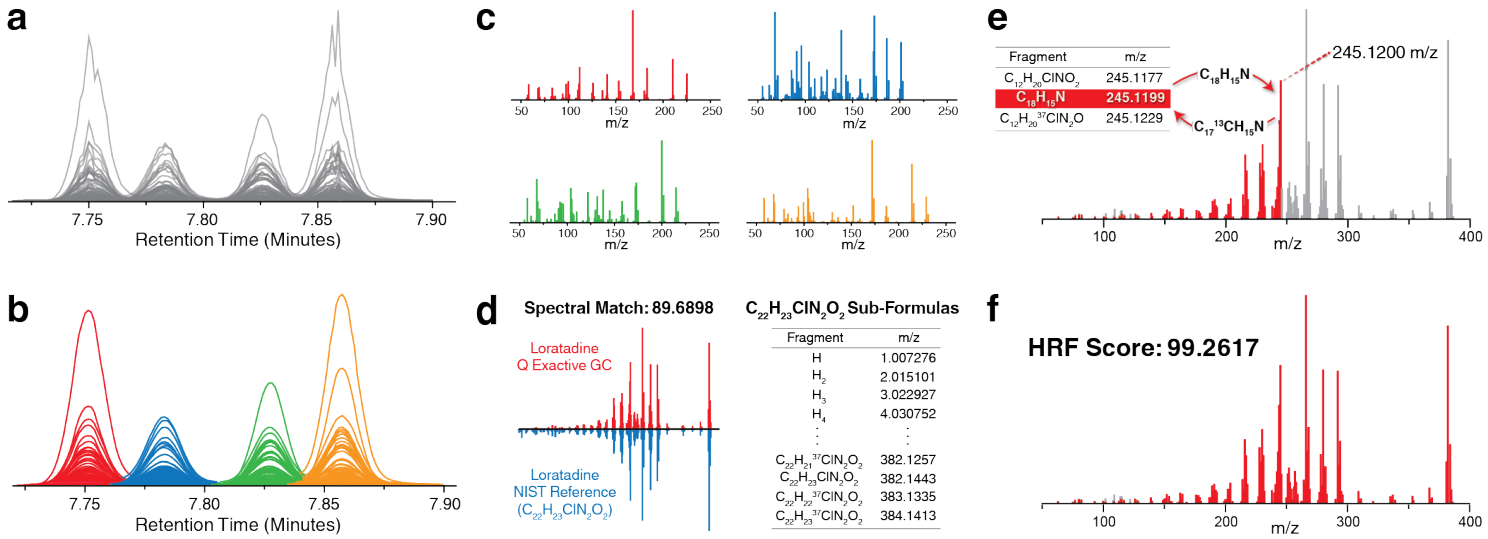


Figure 2.1: High-resolution filtering workflow with spectral matching. **(a)** Peaks observed across consecutive scans are condensed into data features. **(b)** Features are smoothed and grouped based on elution apex. All features within a group are assumed to arise from a singular precursor. **(c)** Individual spectra are derived from feature groups (using average *m/z* and apex intensity) and can then be submitted for spectral matching. **(d)** A strong spectral match of an experimentally-derived spectrum of loratadine against the corresponding NIST reference spectrum. All sub-formulas from $C_{22}H_{23}ClN_2O_2$ are generated and sorted by exact formula mass less an electron. **(e)** Sub-formulas are matched to peaks in ascending order based on *m/z*. For each matched theoretical fragment a variant containing appropriate heavy isotopes is created and placed into the list of sub-formulas in sorted-order. **(f)** For the high-resolution spectrum of loratadine 99.2617% of the measured ion current can be annotated with a sub-formula of $C_{22}H_{23}ClN_2O_2$.

used as metric of confidence in that putative identification. **Figure 2.1d-f** illustrates the HRF strategy using an EI mass spectrum of loratadine, a popular over-the-counter antihistamine, collected using a Q Exactive GC mass spectrometer. A unit resolution database search, returns a reference spectrum of loratadine as a strong candidate match. To evaluate the quality of this putative identification we next employ the HRF strategy. With the chemical formula of loratadine ($C_{22}H_{23}ClN_2O_2$) all non-repeating combinations of atoms (i.e., sub-formulas) are generated and ordered by ascending exact mass less an electron **(Figure 2.1d)**. Note that the theoretical fragment search space is restricted by the atomic composition of loratadine. Starting with the smallest measured *m/z* peak, sub-formulas are matched based on exact mass. To accommodate isotopic clusters present in spectra, a variant containing an appropriate number of heavy isotopes is created for each matched theoretical fragment and placed back onto the list of sub-formulas. For example, once the highlighted *m/z* peak at 245.1200 is matched to $C_{18}H_{15}N$ (theoretical *m/z*: 245.1199) a formula containing a substituted ^{13}C isotope ($C_{17}{}^{13}CH_{15}N$) is added to the list of candidate sub-formulas **(Figure 2.1e)**. This strategy of on-the-fly theoretical isotopic fragment generation enables annotation of non-monoisotopic fragments without unduly increasing sub-formula search space. Once every *m/z* peak in the spectrum has been considered the total percentage of measured ion current that

has been annotated is returned in the form of a HRF score. In the example case of loratadine we find that 99.2617% of all measured ion current can be annotated using a sub-formula of its true parent precursor **(Figure 2.1f)**. Here we demonstrate that the HRF method is viable, enables discrimination between putative identifications, is highly robust even in times of diminished signal-to-noise, and is uniquely enabled with high-resolution GC/MS. Finally, we establish that this approach stands to greatly improve how unknowns are identified in discovery-based analyses.

Reference Standard Analysis To ensure broad utility we benchmarked performance of the HRF algorithm as applied to spectra collected from a diverse array of small molecules. For this work, a data set of high-resolution Q Exactive GC spectra collected from 105 pure reference standards covering many classes of small molecules including metabolites, pesticides, pharmaceuticals, drugs of abuse, among others, was constructed. Following GC/MS analysis of all reference standards, individual spectra were extracted from raw data files using the described deconvolution algorithm. Each extracted spectrum was compared against its corresponding NIST reference spectrum and a weighted dot product score was calculated to measure spectral similarity. For these 105 spectra, a median spectral match score of 81.889, minimum of 42.599, and standard deviation of 9.587 was achieved. Following spectral matching, each spectrum was then subjected to our

HRF approach using the chemical formula of the true parent molecule. Considering all spectra in the data set, we report a median HRF score of 99.700, minimum of 93.497, and standard deviation of 1.022 **(Figure 2.2a)**.

From these data we conclude that performance of the HRF method is well-conserved across many different classes of small molecules. Next we wondered whether similar results could be obtained from other chemical formulas in the reference library. To test specificity, 60,560 HRF scores – all from unique formulas residing in the NIST database – were calculated for each of the 105 spectra. **Figure 2.2b** presents the results of this experiment for the spectrum of trimethylsilyl-derivatized beta-sitosterol ($C_{32}H_{58}OSi$). Note the true parent chemical composition is the smallest formula that can produce the maximal HRF score. We were curious as to the scores generated by subset formulas (some but not all of the atoms contained within the precursor formula) as well as superset formulas (all of the atoms contained by the precursor and then some) which are also highlighted. The annotated subsets lack the proper combination of atoms to achieve the same score. Not surprisingly, all supersets of $C_{32}H_{58}OSi$ produce similarly high scores. This is expected as all subformulas from the true parent will also be included in the subformula sets generated by these superset precursors. We note that in some cases very large formulas which are not true supersets but share a large percentage of

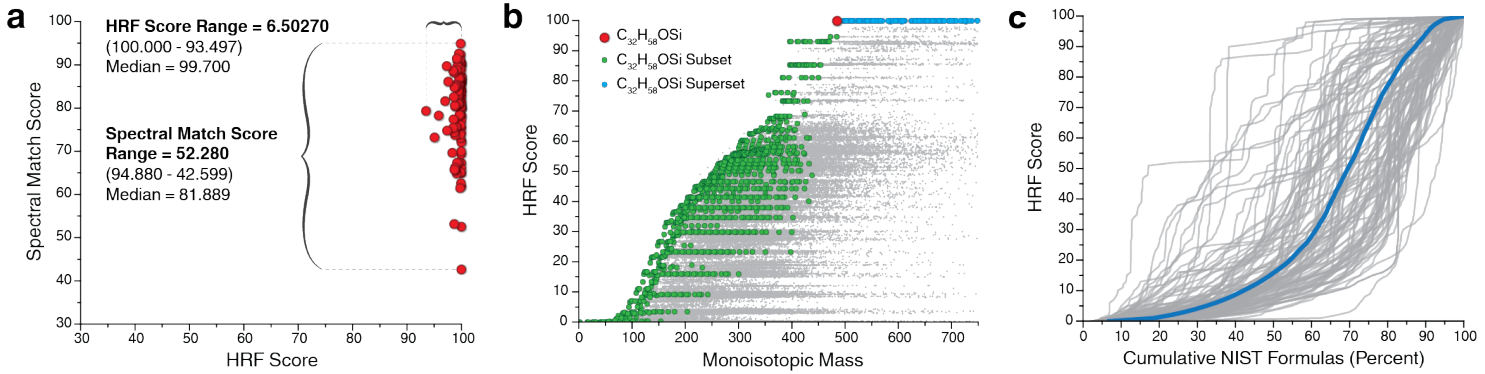


Figure 2.2: High-resolution filtering results. **(a)** Spectral match and HRF scores are shown for the 105 spectra in the dataset.**(b)** HRF scores for a spectrum of beta-sitosterol (TMS) using 60,560 different formulas are shown. The true parent ($C_{32}H_{58}OSi$) is shown in red. Sub- and supersets of $C_{32}H_{58}OSi$ are shown in green and blue respectively. **(c)** Cumulative distributions of the 60,560 HRF scores calculated for all 105 spectra are shown in gray. A representative distribution generated by combining all results is shown in blue.

atoms with the correct parent can also produce high scores.

For a global view of the method's specificity we plot the cumulative distributions of HRF scores to all 105 spectra along with the average distribution of all cumulative distributions **(Figure 2.2c)**. Note that this analysis provides a worst-case scenario given all 60,560 formulas considered have an equal chance of being selected as a putative parent for an acquired spectrum. In most cases this is not the case, as either spectral matching or a priori information held by the analyst allows discrimination against the majority of these candidates. Still, these data reveal that on average ~86.9% of considered formulas will return a HRF score $\geqslant 90$ and that only 3.560% of candidate formulas will produce a score greater than or equal to the median calculated HRF score (99.700). We also note that specificity is dependent on the complexity of the analyte in question, for example, increases in elemental complexity and atom count will often result in spectra which a smaller number of precursors can successfully annotate.

Urine Drug Testing Most analytical applications demand the identification of low level analytes, often present within complex matrices. In these situations spectral quality is eroded – manifested by the loss of key diagnostic fragments with diminished signal and increased chemical noise – limiting the ability to correctly assign identifications through traditional spectral matching. To test the benefits of

HRF in such situations we next analyzed a panel of drugs at varying concentration in a biological matrix. GC/MS is the ideal platform to test for drugs of abuse, pharmaceuticals, sports dopants, and their metabolites in human urine. These assays are highly desirable in the clinical setting as they are minimally invasive.

As a proof-of-concept, twelve drugs (amobarbital, Benadryl, caffeine, cotinine, glutethimide, ketamine, loratadine, methadone, methaqualone, nicotine, primidone, and scopoloamine) were spiked into human urine at eight concentations (10 ng/μL to ~78 pg/μL) and extracted prior to GC/MS analysis **(Figure 2.3a)**. Chromatographic resolution was insufficient to separate Benadryl and ketamine, and high native levels of caffeine diminished the ability to analyze the compound through a range of concentrations. As such, further analysis was not carried out and here we report results for nine of the twelve drugs.

The analysis of compounds in a complex background matrix, such as urine, presents two considerable challenges – extracting high-quality spectra and assigning confident identifications—with the latter being highly dependent on the former. Ideally, extracted spectra should retain all fragment *m/z* peaks stemming from the eluting precursor while eliminating all other chemical background, which can be of higher abundance. Deconvolution is the core technique for spectral extraction and this process, as we report here, is considerably improved by use of the FT-MS

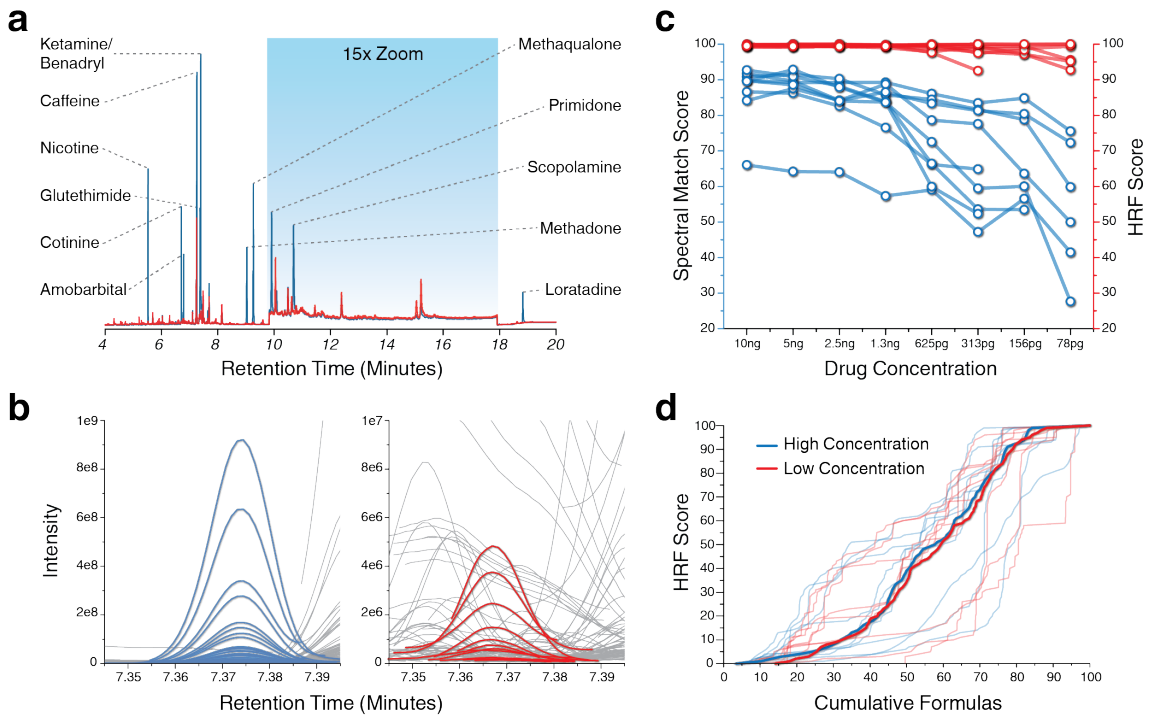


Figure 2.3: Analysis of drugs spiked into human urine at variable concentration. **(a)** GC/MS TIC chromatograms from the most concentrated (blue) and least concentrated (red) spiked samples are shown. **(b)** Deconvolved feature groups for the drug Glutethimide at high (blue) and low (red) concentrations. Background features are shown in gray. **(c)** Spectral match and HRF scores for each drug analyzed at all concentrations where analyte abundance was sufficient to produce a spectrum. **(d)** Two spectra were isolated for each drug (one at the most concentrated point, the other at the least) and 55,290 HRF scores were calculated using unique formulas (0-500 Da) in the NIST database. Cumulative HRF results are shown for each drug using a spectrum acquired at high and low concentration (blue and red, respectively). A combined distribution is also shown for each population of drug spectra.

systems. Accurate mass measurement largely eliminates interferences between nominally isobaric fragments and allows extraction of chromatographic profiles using narrow mass tolerances (~$\pm$ 10 ppm). Furthermore, the rapid scan rate (> 18 Hz) provides sufficient temporal resolution enabling more precise detection of chromatographic apex. Note that spectral deconvolution assumes that all peaks are derived from a singular precursor and if two compounds completely overlap the resulting EI-MS spectrum will be chimeric which may impede spectral identification. **Figure 2.3b** highlights spectral deconvolution of glutethimide at high and low concentrations. Note the numerous co-eluting interferants in the low concentration chromatogram that are easily distinguished. We conclude that spectral deconvolution is a key parameter for successful downstream identification and is improved by collection of spectra with high-resolution and accurate mass.

Extraction of high-quality spectra from raw data files is only the first step in assigning confident identifications. Mapping these spectra to structure is then commonly done by spectral matching against a library, which is most effective when experimental spectra very closely resemble those contained in the library. The specificity of this approach, however, is reduced as analyte abundance decreases and diagnostic fragments fall below the limit of detection. We surmised that the HRF approach could provide an orthogonal metric, allowing greater discrimination

between putative identifications. To test this hypothesis we applied the HRF approach to analyze the standard drug compounds spiked into the urine matrix across a wide range of concentrations. We required that all spectra contain at least 10 *m/z* peaks, eliminating 5 of 72 data points. In these instances the compound in question was at a sufficiently reduced concentration such that the extracted spectrum was either non-existent or of too low quality for any further processing. Extracted spectra were then compared to their corresponding NIST reference spectrum to generate both spectral match and HRF scores for each **(Figure 2.3c)**. As expected, the spectral match score decreases with diminishing analyte abundance, primarily due to the loss of low abundance peaks at decreased concentrations. HRF performance, however, is remarkably consistent, independent of analyte concentration, and remains high (> 92) for all observed spectra. From these data we draw two primary conclusions: First, FT-MS mass analyzers provide robust mass accuracy measurements, even for signals occurring at low S/N[26]; and second, unlike the conventional spectral matching strategy, the HRF scoring metric is conserved across a wide range of analyte concentrations.

While the experiment described above demonstrates strong HRF scoring performance, we wondered whether the method would maintain the ability to discriminate between candidate precursors, when provided with lower quality spectra.

To determine if the HRF scoring method had diminished specificity for spectra containing a reduced number of diagnostic *m/z* peaks, i.e., those collected at lower abundance, we calculated HRF scores from 55,290 unique formulas in the NIST spectral library (0-500 Da) using two EI-MS spectra for each drug analyzed (one from the most concentrated data point, the other from the least). These high and low concentration spectra present a striking spectral quality difference as the low abundance spectra contain only about 25% of the *m/z* peaks found in the higher quality analog (23 v. 96, on average). **Figure 2.3d** presents the cumulative distributions of these calculated HRF scores for either high (blue) or low (red) concentration spectra. The average distribution for each set of spectra is also displayed and no difference is readily observed. It is apparent that, whether analyzing low or high quality spectra, HRF specificity is maintained. The fundamental driving force for this indifference to spectral quality, as compared to traditional spectral matching, is the discriminatory power of mass accuracy which is retained even within low-quality spectra. Based on these data we surmise that the HRF strategy is less dependent upon input spectral quality—a characteristic that will propel the emergent area of small molecule discovery and profiling applications.

Application to Discovery Metabolomics High chromatographic resolution, excellent sensitivity, and conserved fragmentation of molecular precursors render

GC/MS a fitting method for discovery-based metabolic profiling. In recent years there has been a marked decrease in the time required to comprehensively sequence genomes, transcriptomes, and proteomes. These increases in throughput have largely come as a result of coincident improvements in instrumentation and informatics enabling faster sequencing than ever before. Discovery metabolomics has lagged behind these other "omics" technologies due in large part to the difficulty in assigning confident identifications to analyzed compounds. We assert that by coupling the recent advances in high-resolution GC/MS instrumentation with new data processing schemes, the depth and speed at which metabolomes can be fully characterized can be greatly increased. One approach to realizing this potential is to utilize the HRF approach as a data reduction strategy for eliminating spurious hits, and retaining only those which are chemically plausible.

To characterize the utility of the HRF approach for metabolomic applications the algorithm was applied to a discovery analysis of a yeast metabolite extract. Here a TMS-derivatized yeast metabolome and solvent blank were analyzed on a Q Exactive GC system in tandem. Following data acquisition individual spectra were extracted from both raw files using the described in-house deconvolution algorithm. Spectral deconvolution yielded 19,367 spectral features which were placed into 554 feature groups—each group containing fragments which are assumed to stem from

a singular precursor. Deconvolution results were manually validated and additional curation was employed where necessary. EI-MS spectra that were common to both the yeast extract and solvent blank were eliminated from consideration. In total, 232 EI-MS spectra (all containing no fewer than 10 *m/z* peaks) were considered for this analysis, post background subtraction. These spectra were then searched against the NIST database (~213,000 compounds) at unit resolution. The 20 highest scoring spectral matches were returned and HRF scores were then calculated for each – generating 4,640 HRF scores in total. **Figure 2.4a** displays the distributions for both scores. The orthogonality between these two metrics is readily apparent. While the majority of spectral match scores cluster around 30-40 with a skew towards higher scores—again, this distribution represents the 20 best hits to each spectrum, many of which were derived from lowly abundant precursors—the HRF distribution is bimodal with large clusters at both extremes. These clusters (greater than 90 and less than 10) comprise 60.69% of all returned results.

In the analysis of reference standards we observed no instances where a correct identification yielded spectral match or HRF scores lower than 20 or 90, respectively. To visualize these data we present a heat map **(Figure 2.4b)** displaying each EI-MS spectrum as a row with the calculated HRF score for each of the 20 putative spectral matching assignments as the columns. This plot reveals that top scoring

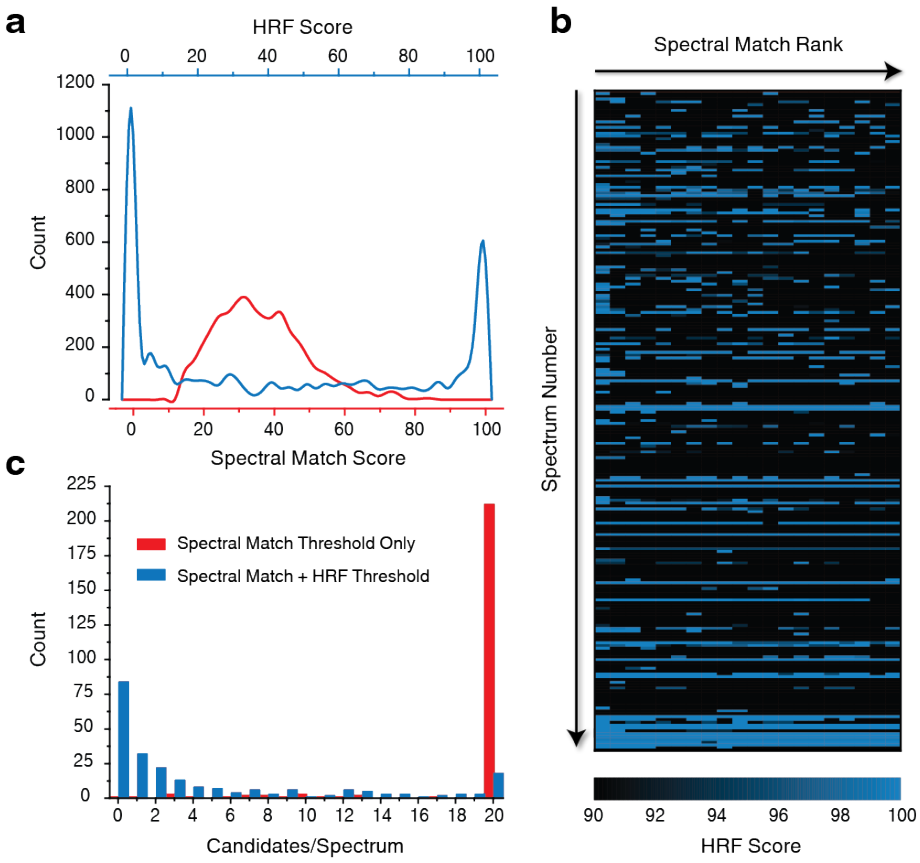


Figure 2.4: Discovery yeast metabolomic analysis. (a) Distributions of the top 20 spectral match/corresponding HRF scores to 232 spectra extracted from a yeast metabolomics experiment. **(b)** HRF scores corresponding to the 20 best spectral matches (left to right) for all 232 spectra (top to bottom) are shown in the blue heat map. The intensity of each pixel reflects the percentage of total ion current that can be annotated with an exact chemical formula. **(c)** Viable candidates/spectrum when applying spectral match and HRF score thresholds.

spectral matches are not always consistent with the chemical formula information gleaned by the HRF calculation. We find that 76.00% of returned identifications are eliminated after applying an HRF threshold (90) including 58.62% of all number one spectral match hits. We also note many instances in which lower spectral match scores to a given spectrum yield higher HRF scores suggesting that joint consideration of both metrics is advantageous. To determine the value of the HRF method to eliminate from consideration incorrect putative assignments we plotted the number of candidate identifications per spectrum before and after application of HRF scoring **(Figure 2.4c)**. Imposing a spectral match score cutoff of 20, eliminates only 5.28% (245) of hits, leaving the analyst to sort through the remaining 4,240 candidates. Application of the HRF score threshold in addition to the spectral match score threshold, however, allows dismissal of the majority of the putative candidate identifications – 79.78% (3,720). In fact, the HRF method allows the analyst to reduce the number of viable candidate structures with confidence; for example, 65.09% of spectra retain three or fewer valid candidates. While analysts will still find it useful to confirm candidate identifications by sampling pure reference standards, the tremendous reduction in candidate identifications will expedite the process of small molecule identification and provide a means to accelerate the pace of metabolomic discovery.

Conclusions

Small molecule analysis and discovery remains at the core of many fields—e.g., toxicology, sports doping, environmental analysis, food safety, clinical research, etc.—and is emerging as a key technology in the expanding area of metabolomics. GC/MS is a robust and mature method for profiling small molecules, but has recently undergone a transformation with the introduction of state-of-art mass analyzer capabilities that deliver routine high-resolution and accurate mass measurement. The new type of GC/MS data created by these modern systems has transformative potential—realizing this promise, however, requires new and innovative data processing approaches.

Here we describe a simple and straightforward method, HRF, which leverages accurate mass to both improve spectral deconvolution and increase confidence in small molecule identifications. The HRF approach can be used in conjunction with traditional spectral matching and effectively extends the utility of currently available unit-resolution reference libraries. Moreover, information provided by this approach is orthogonal to traditional spectral matching. In the future we predict this method will be of high value for the analysis of novel compounds, where a suitable reference spectrum is unavailable. In this application users would simply provide suspected chemical formulae and/or structures and utilize the HRF scoring

method to test candidate plausibility. No such technology currently exists. We note the HRF approach facilitates rapid annotation of EI-MS spectra, has potential for LC-MS/MS applications, and may prove useful for automated false-discovery rate calculations. In summary, by enabling discrimination between candidate molecular precursors on the basis of both measured fragmentation profiles and accurate mass, the HRF method capitalizes on new high-resolution GC/MS instrumentation and the large, existing unit resolution EI-MS spectral libraries.

Extended Methods

Urine Drug Analysis The following GC gradient was used: 2.5 min isothermal at 60 °C, ramp to 210 °C at 40 °C/min, ramp to 267 °C at 5 °C/min, ramp to 310 °C at 40 °C/min, then 6.2 min isothermal at 310 °C. The MS transfer line and source temperatures were held at 280 °C and 200 °C, respectively. The mass range from 50-500 m/z was mass analyzed using a resolution of 30,000 $(m/\Delta m)$, relative to 200 m/z. The AGC target was set to 10e6, and electron ionization (70 eV) was used. Lock mass calibration was employed during acquisition of these data. An unanticipated error occurred in calculation of the necessary mass correction, and many scans acquired during these experiments resulted in extreme mass errors (~25ppm). Large distortions in mass accuracy largely inhibit the described HRF

approach. As such, during data processing each spectrum was restored to its native-state by removing the applied mass correction as reported in each scan header. Subsequent analyses did not employ this lock-mass correction and mass accuracy was unaffected.

Preparation of a *Saccharomyces cerevisiae* metabolite extract Saccharomyces cerevisiae was grown on media containing dextrose and glycerol. $1x10^8$ cells were isolated by rapid vacuum filtration with a nylon filter membrane, washed with phosphate buffered saline, and submerged into a precooled 1.5 mL plastic tube containing a 2:2:1 acetonitrile/methanol/H_2O mixture.

Pesticide Analysis The mixture containing 37 EPA 525.2 pesticides was diluted from 500 μg/mL to a working concentration of 3 ng/μL in acetone. A 1 μL aliquot was injected using a 1:10 split at a temperature of 275 °C and separated at 1.2 mL/min He. The following GC oven gradient was used: isothermal at 100 °C for 1 min, 8 °C/min to 320 °C, and isothermal at 320 °C for 3 min. Transfer line and source temperatures were maintained at 275 °C and 225 °C, respectively. In each MS scan, the range from 50-650 *m/z* was analyzed using a resolution of 17,500 *(m/Δm)*, relative to 200 *m/z*. Maximum injection times of 100 ms were allowed at an AGC target of 1e6. Electron ionization (EI) at 70 eV was used.

Additional Reference Standard Analysis Stock solutions for all other reported standards were prepared individually at a concentration of 1 mg/mL in appropriate solvents. Mixtures containing ~5-10 reference standards were prepared by combining 20 μL aliquots of each standard using no specific organizational scheme. These mixtures were dried down under nitrogen, resuspended in 100 μL of the MSTFA + 1% TMCS derivatization reagent, capped, vortexed, and heated at 60 °C for 15 minutes. 100 μL of ethyl acetate was then added to each mixture before being transferred to an autosampler vial. The same GC oven gradient and MS parameters as described in Urine Drug Analysis were also used here.

Spectral Deconvolution Following data collection raw EI-MS spectral data was deconvolved into 'features' and then grouped into individual spectra containing only product ions stemming from a singular parent. This step was critical as the inclusion of extraneous fragment ions in a spectrum can diminish the ability of the algorithm to annotate all observed peaks with exact chemical formulas constrained by the atom set of the parent. Every peak in the raw data file was considered. Peaks observed in at least five consecutive scans having *m/z* values within ±10 ppm of their averaged *m/z* were grouped together as a data feature. Note that mass accuracy is a function of and S/N, and ppm tolerance a function of *m/z*. The 10 ppm tolerance was empirically observed to yield complete chromatographic profiles which were

free of interference from neighboring peaks. Peaks were added successively to these groups and the average *m/z* value was recalculated after each addition. Following aggregation of peaks into features, smoothed intensity profiles were created for each. Spurious features arising from noise were eliminated from consideration by requiring that each feature exhibit a "peak-like" shape. All features were required to rise to an apex having at least twice the intensity of the first and last peaks included. Any features arising from fragments common to closely eluting precursors were split into separate features at significant local minima. Features reaching an elution apex at approximately the same time were grouped together. Features were first sorted based on apex intensity. Starting with the most intense fragment a discrete time window around the apex was created. All features having an apex within this window were then grouped together. The width of this window was set to include all peaks having an intensity $\geqslant$ 96% of the apex peak's intensity as a default. More conservative criteria was used for the extraction of spectra in the urine drug spike-in and discovery metabolomics experiments given the complex background. Here the time window was set to include peaks having an intensity $\geqslant$ 99% of the apex. Following feature grouping, a new spectrum was created for each group and populated with peaks representing each feature in the group. Peak *m/z* and intensity values were set equal to the intensity-weighted *m/z* average of all peaks in

the corresponding feature and the intensity at the apex, respectively.

Small Molecule Identification via Spectral Matching Compound identifications for the small molecules analyzed were assigned by comparing deconvolved high-resolution spectra against unit-resolution reference spectra present in the NIST 12 MS/EI Library. All 212,961 unit-resolution reference spectra in the library were exported to a .JDX file through the NIST MS Search 2.0 program and converted to a format suitable for matching against acquired Q Exactive GC spectra. A pseudo-unit resolution copy of each high-resolution spectrum was created by combining the intensities of peaks falling within the same nominal mass range. The nominal mass value was reported as peak *m/z* and all intensity values were normalized relative to the spectrum's base peak (set to 999). To calculate spectral similarity between experimental and reference spectra a weighted dot product calculation was used. First, all peaks in a spectrum were scaled using the following normalization factors reported in the literature which were determined to provide optimal spectral matching results[27]:

$$m/z_{\text{normalized}} = m/z_{\text{measured}} \times 1.3$$

$$\text{Intensity}_{\text{normalized}} = \text{Intensity}_{\text{measured}}{}^{0.53}$$

These normalization factors redistribute the weight placed on any given spectral

peak in two ways: First, by scaling *m/z* by a factor of 1.3x, more massive peaks (which are inherently more diagnostic for spectral matching) are given greater weight. Second, by scaling intensity by a factor of x0.53 more intense peaks are given relatively less weight. This is done to ensure that no single peak can disproportionately influence spectral matches. The described normalizations were applied to all reference spectra as well. The following dot product equation was used to measure spectral similarity:

$$100 \times \frac{\sum (m/z[\text{Intensity}_{\text{experimental}} \times \text{Intensity}_{\text{reference}}]^{0.5})^2}{\sum (\text{Intensity}_{\text{experimental}} \times m/z) \sum (\text{Intensity}_{\text{reference}} \times m/z)}$$

Although simplistic, this approach was more than adequate for retrieving candidate compounds having similar fragmentation patterns to experimentally derived spectra. To increase search space as much as possible all reference spectra were matched against each unit resolution copy of a Q Exactive GC spectrum in the 'discovery metabolomics analysis'. All compounds reported yielded a confident spectral match with a reference spectrum in the NIST database.

High-Resolution Filtering: Theoretical Fragment Generation A set of theoretical fragments for each candidate compound was produced by generating all unique combinations of atoms from the set contained in the parent chemical formula which can be calculated by:

$$x = \sum_{i}^{n} (i_a + 1)$$

where x is the number of theoretical fragments stemming from a given chemical formula, n is the number of unique elements in the formula, and i_a represents the atom count of that element within the formula. The most abundant isotope for each atom was used with the exception of bromine and chlorine. ^{79}Br and ^{81}Br have natural isotopic abundances of 0.5069 and 0.4931, respectively. Similarly, ^{35}Cl and ^{37}Cl have natural abundances of 0.7576 and 0.2424. For each theoretical fragment containing either a bromine or chlorine an additional variant was generated where a heavier isotope was exchanged for its lighter counterpart. This process was repeated in a combinatorial manner for those theoretical fragments containing multiple Br and/or Cl atoms. Generation of additional isotopic theoretical fragments for those candidates containing atoms in the set ^{12}C, ^{32}S, ^{28}Si was done on a case-by-case basis during the theoretical fragment/peak matching process.

High-Resolution Filtering: Theoretical Fragment/Peak Matching It is assumed that all fragment peaks in an EI-MS spectrum are radical cations. Accordingly, the mass of an electron was subtracted from the monoisotopic mass of each fragment

in the set of candidates. Starting with the least massive peak in the Q Exactive GC spectrum, theoretical fragments falling within a $\pm$ 10 ppm tolerance centered around the peak's measured *m/z* were found. This tolerance was empirically determined to be the optimal allowed mass tolerance as it enabled annotation of low S/N fragments where mass accuracy is diminished while maintaining discrimination against spurious chemical formulas. If no fragments were present within this range, the algorithm moved to the next most massive peak and repeated the process. If a single fragment was found within this range, isotopic variants containing substituted ^{13}C, ^{33}S, ^{34}S, ^{29}Si, or ^{30}Si atoms were generated where appropriate and added to the list of candidate fragments. If multiple fragments were found within the allowed tolerance each fragment was independently evaluated to determine how many additional peaks/signal could be matched. The theoretical fragment resulting in the largest amount of additional matched signal was assumed to be correct and substituted isotopic theoretical fragments were added to the list of candidate theoretical fragments. All peaks which had matching theoretical fragments were stored. After all peaks were considered the total ion current that was matched to a theoretical fragment as calculated by:

$$\frac{\sum(\text{Intensity} \times m/z)_{\text{annotated}}}{\sum(\text{Intensity} \times m/z)_{\text{observed}}}$$

was returned. This scoring calculation was deemed appropriate as it gives ad-

ditional weight to larger ions which are inherently more diagnostic of a given precursor than less massive ions. Conceptually, there are fewer molecules in existence which can theoretically produce a fragment at 300 *m/z* than there are which can produce a fragment at 200 *m/z*.

References

[1] P. Westerhoff, Y. Yoon, S. Snyder, and E. Wert, "Fate of endocrine-disruptor, pharmaceutical, and personal care product chemicals during simulated drinking water treatment processes," *Environmental Science and Technology*, vol. 39, pp. 6649–6663, 2005.

[2] E. Tareke, P. Rydberg, P. Karlsson, S. Eriksson, and M. Törnqvist, "Analysis of acrylamide, a carcinogen formed in heated foodstuffs," *Journal of Agricultural and Food Chemistry*, vol. 50, pp. 4998–5006, 2002.

[3] H. Kataoka, H. Lord, and J. Pawliszyn, "Applications of solid-phase microextraction in food analysis," *Journal of chromatography A*, vol. 880, pp. 35–62, June 2000.

[4] C. Yang, A. C. Park, N. A. Davis, J. D. Russell, B. Kim, D. D. Brand, M. J. Lawrence, Y. Ge, M. S. Westphall, J. J. Coon, and D. S. Greenspan, "Compre-

hensive mass spectrometric mapping of the hydroxylated amino acid residues of the α1(V) collagen chain," *Journal of Biological Chemistry*, vol. 287, pp. 40598–40610, 2012.

[5] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. N. Trethewey, and L. Willmitzer, "Metabolite profiling for plant functional genomics.," *Nature biotechnology*, vol. 18, pp. 1157–61, 2000.

[6] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: Acquiring and understanding global metabolite data," 2004.

[7] J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, and D. B. Kell, "High-throughput classification of yeast mutants for functional genomics using metabolic footprinting.," *Nature biotechnology*, vol. 21, pp. 692–6, 2003.

[8] S. E. Stein, "An Integrated Method for Spectrum Extraction," vol. 0305, no. 99, 1999.

[9] O. Fiehn, "Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry," *TrAC - Trends in Analytical Chemistry*, vol. 27, pp. 261–269, 2008.

[10] O. Fiehn, J. Kopka, R. N. Trethewey, and L. Willmitzer, "Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry.," *Analytical chemistry*, vol. 72, pp. 3573–3580, 2000.

[11] T. Pluskal, T. Uehara, and M. Yanagida, "Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching.," *Analytical chemistry*, vol. 84, pp. 4396–403, May 2012.

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

[13] "Wiley Registry of Mass Spectral Data," 2010.

[14] S. Lewis and C. Kenyon, "High resolution gas chromatographic/real-time high resolution mass spectrometric identification of organic acids in human urine," *Analytical chemistry*, vol. 51, pp. 1275–1285, July 1979.

[15] A. Peterson and A. Balloon, "Development of a GC/Quadrupole-Orbitrap mass spectrometer, part II: new approaches for discovery metabolomics," *Analytical …*, vol. 86, pp. 10044–51, Oct. 2014.

[16] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J.

Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, 2014.

[17] A. C. Peterson, G. C. McAlister, S. T. Quarmby, J. Griep-Raming, and J. J. Coon, "Development and characterization of a GC-enabled QLT-orbitrap for high-resolution and high-mass accuracy GC/MS," *Analytical Chemistry*, vol. 82, pp. 8618–8628, 2010.

[18] S. Wolf, S. Schmidt, M. Müller-Hannemann, and S. Neumann, "In silico fragmentation for computer assisted identification of metabolite mass spectra.," *BMC Bioinformatics*, vol. 11, p. 148, 2010.

[19] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, and D. F. Grant, "Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra," *Analytical Chemistry*, vol. 80, pp. 5574–5582, 2008.

[20] A. Kerber and R. Laue, "MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation," *Adv Mass …*, vol. 5, no. 2, pp. 5–6, 2001.

[21] "Solid Phase Extraction Applications Manual," 2008.

[22] A. Michalski, E. Damoc, J.-P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning, "Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer.," *Molecular & cellular proteomics : MCP*, vol. 10, p. M111.011015, 2011.

[23] J. V. Olsen, J. Schwartz, J. Griep-Raming, M. L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E. R. Wouters, M. Senko, A. Makarov, M. Mann, and S. Horning, "A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed.," *Molecular & Cellular Proteomics : MCP*, vol. 8, pp. 2759–2769, 2009.

[24] T. P. Second, J. D. Blethrow, J. C. Schwartz, G. E. Merrihew, M. J. MacCoss, D. L. Swaney, J. D. Russell, J. J. Coon, and V. Zabrouskov, "Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures," *Anal.Chem.*, vol. 81, pp. 7757–7765, 2009.

[25] E. White V, "Fred W. McLafferty. Interpretation of mass spectra, third edition. University science books, Mill valley, California, 1980. pp. xvii + 303," *Biological Mass Spectrometry*, vol. 9, no. 6, pp. iii–iv, 1982.

[26] C. Wenger, G. McAlister, Q. Xia, and J. Coon, "Sub-part-per-million precursor

and product mass accuracy for high-throughput proteomics on an electron transfer dissociation-enabled orbitrap mass," *Molecular & Cellular …*, vol. 9, pp. 754–63, May 2010.

[27] S. Kim, I. Koo, X. Wei, and X. Zhang, "A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry," *Bioinformatics*, vol. 28, pp. 1158–1163, 2012.

Chapter 3

A SOFTWARE SUITE FOR THE ANALYSIS OF HIGH-RESOLUTION GC/MS METABOLOMIC DATA

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

Kwiecien NW, Rush MJP, Westphall MS, Coon JJ. *A Software Suite for the Analysis of High-Resolution GC/MS Metabolomic Data.* **2016**.

Introduction

Within the field of metabolomics, liquid chromatography-mass spectrometry (LC/MS) has emerged as the technology of choice for large-scale profiling studies. High-resolution LC/MS systems are prominent in many research facilities, and software tools for quantitative data processing are well developed[1,2]. Gas chromatography-mass spectrometry (GC/MS) is similarly a powerful tool for small molecule and metabolite analysis. GC/MS affords highly reproducible chromatographic separations and molecular fragmentation patterns, which greatly facilitate the comparison of metabolomic profiles across large sample sets. However, given the lack of high-resolution instrumentation and associated data processing software, GC/MS has been slow to gain traction in the metabolomics community. Recent developments in GC/MS technology stand to increase the viability and appeal for metabolite profiling applications. Thermo Fisher Scientific (Austin, TX) has recently commercialized a new high-resolution GC-Orbitrap mass spectrometer, which affords unparalleled resolution and mass accuracy[3–5]. The resolving powers achievable with the GC-Orbitrap enable detection and quantitation of more compounds per sample, given that nominally isobaric ions can be readily distinguished. Furthermore, the high mass accuracy provided allows annotation of *m/z* peaks with specific chemical formulae (as described in **Chapter 2**), a feature which can be leveraged to improve

assignment of putative chemical identities[6].

Although this new platform can be used to generate rich metabolomic data, processing software for extracting quantitative information remains underdeveloped. In fact, to the best of our knowledge, no software packages exist for performing untargeted metabolite quantitation from high-resolution GC/MS data. To propel high-resolution GC/MS forward as a functional tool for metabolite profiling, robust software solutions are needed. Ideal software tools would perform all necessary operations to identify and quantify a conserved set of metabolites across a large set of raw MS data files. Furthermore, these tools should present results to users in a manner that facilitates rapid data exploration and comparative analysis. To achieve these goals, a number of considerations must be made. First, developed packages must be able to extract and aggregate chemical features arising from singular chemical species. This is critical as traditional EI-GC/MS experiments generate numerous molecular fragments during the ionization process, which results in multiple features being analyzed from each precursor. Second, it is necessary to perform a conservative background subtraction to identify biologically relevant metabolites in an unbiased manner, and eliminate chemical noise from downstream quantitation. Finally, it is optimal that developed software solutions perform *post hoc* statistical analyses which will expedite data analysis. To facilitate meaningful

statistical testing, software tools must account for the hierarchical organization of experimental data.

Here, we report on the development of a comprehensive high-resolution GC/MS quantitation pipeline, to be used with GC-Orbitrap data. This pipeline was designed with all of the previously mentioned considerations in mind, and is intended to be run on personal computers (PCs). Our tools perform all operations required to extract quantitative metabolomic profiles from raw MS data files in an untargeted fashion. We incorporate user-provided data about experimental organization to perform automated statistical analyses. Additionally, we provide quantitative results to users in a format which can be immediately visualized through a provided data viewer. Each component of the developed pipeline is described in detail below. In addition, we highlight results from various biological studies where our pipeline was employed, successfully, for metabolite quantitation.

Pipeline Overview

Our developed suite consists of five standalone software utilities—designed to be used sequentially—which perform all steps required to quantify and compare a conserved set metabolites across an undetermined number of raw GC-Orbitrap MS data files. For the purposes of comparative analysis, it is desirable to profile the

largest set of metabolites possible, in any given experiment. Metabolomics suffers, in general, from a relatively low rate of compound identification[7]—an issue which is addressed extensively in **Chapter 2**. Even without an assignment of compound identity, examination of characteristic molecular abundance changes can afford valuable insight into phenotypic similarities between analyzed samples. With this in mind, we have designed our package to automatically select biologically-relevant chromatographic features (i.e. *m/z* peaks) which we attempt to locate and quantify across all user-provided MS data files. This untargeted approach negates the need for assignment of molecular identities prior to quantitation, which greatly expands the number of metabolites that can be monitored in a given MS study.

The five software tools included in this pipeline are: Deconvolution Engine, Deconvolution Studio, Experiment Builder, GC-Quant, and GC-Viewer. Deconvolution Engine extracts all chromatographic features from user provided GC-Orbitrap MS data files, and then exports that information to separate data files. Deconvolution Studio performs a multi-dimensional background subtraction between a 'blank' and an 'analyte' (read, representative sample) file, and automatically curates a list of chromatographic features to be used for quantitation. Experiment Builder is a tool for spreadsheet editing, wherein a user can define the hierarchical organization of MS data files in their experiment. GC-Quant performs all quantitation

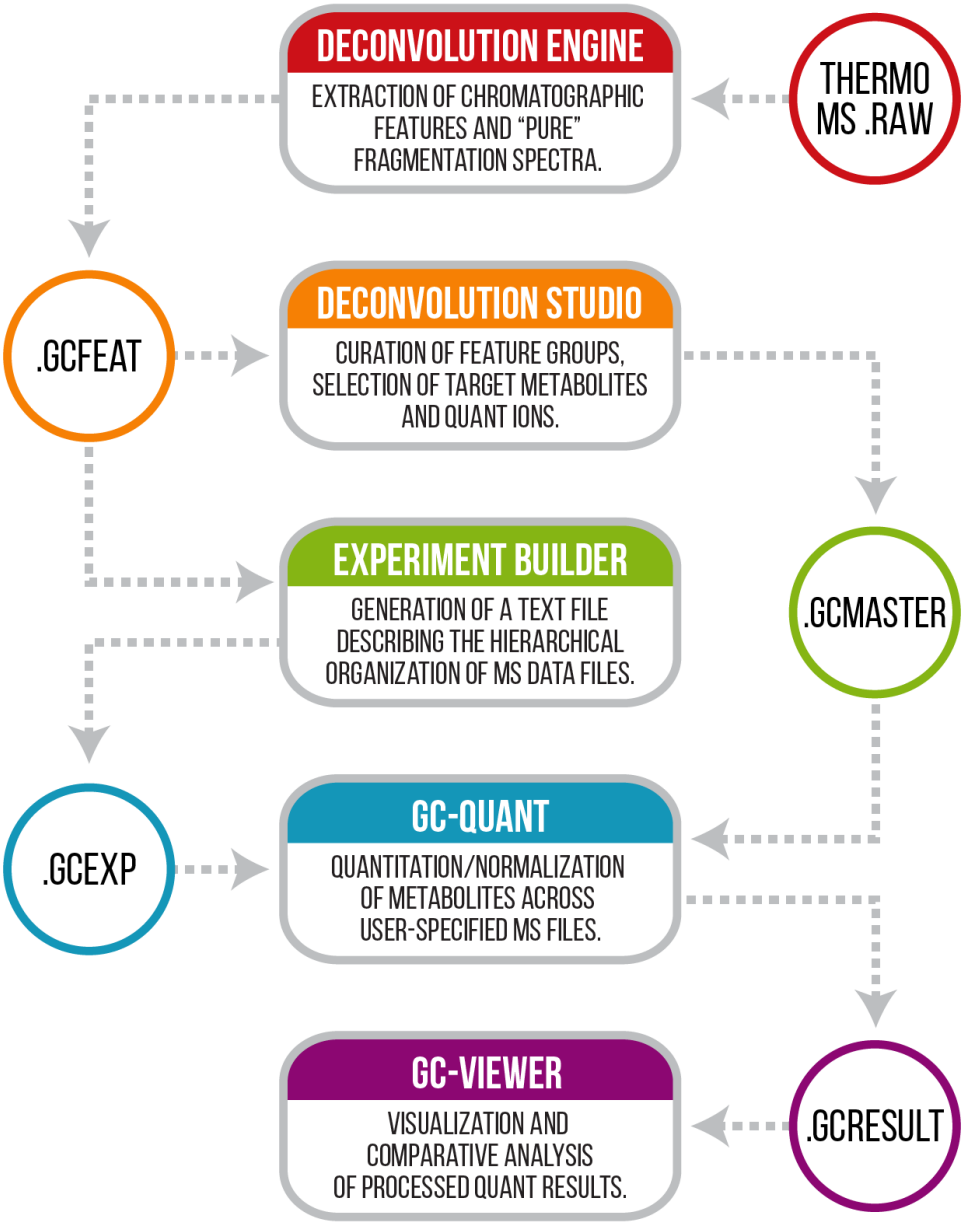


Figure 3.1: High-resolution GC/MS metabolite quantitation analysis workflow. Chromatographic features and "pure" fragmentation spectra are extracted from raw Thermo MS data files in Deconvolution Engine. Individual metabolites—and corresponding quant ions—are selected for quantitation in a semi-automated fashion in Deconvolution Studio. Users can define the hierarchical organization of their experimental data within experiment builder. GC-Quant performs quantitation across multiple files using output files from the three previoulsy mentioned programs. Results from quantitative analyses can be explored in GC-Viewer.

and normalization procedures using the previously generated "target-list," as well as deconvolved MS data files. Finally, GC-Viewer is a functional GUI wherein the results generated by GC-Quant can be automatically visualized and explored. We elaborate on the design and functionality of each of these tools below.

Deconvolution Engine

Design. Deconvolution Engine is the first software tool in our pipeline. It serves to extract chromatographic features from raw GC-Orbitrap MS data files, aggregate these features into consensus fragmentation spectra, and then export all results to a separate file. This tool was built using C# .NET and we provide a simple, but functional, GUI which only requires user input of MS data files and a file output directory **(Figure 3.2)**. By clicking 'Start,' the processes described above are carried to completion for each user-provided MS data file. A progress bar updates in real-time as the underlying algorithms are executed. On average, each raw file—assuming a 30-minute gradient—can be processed in approximately two minutes. Here, we describe in detail the underlying processes which Deconvolution Engine carries out.

Function. In a traditional GC/MS experiment, EI is used to impart charge to analyte molecules following a front-end chromatographic separation. EI is generally

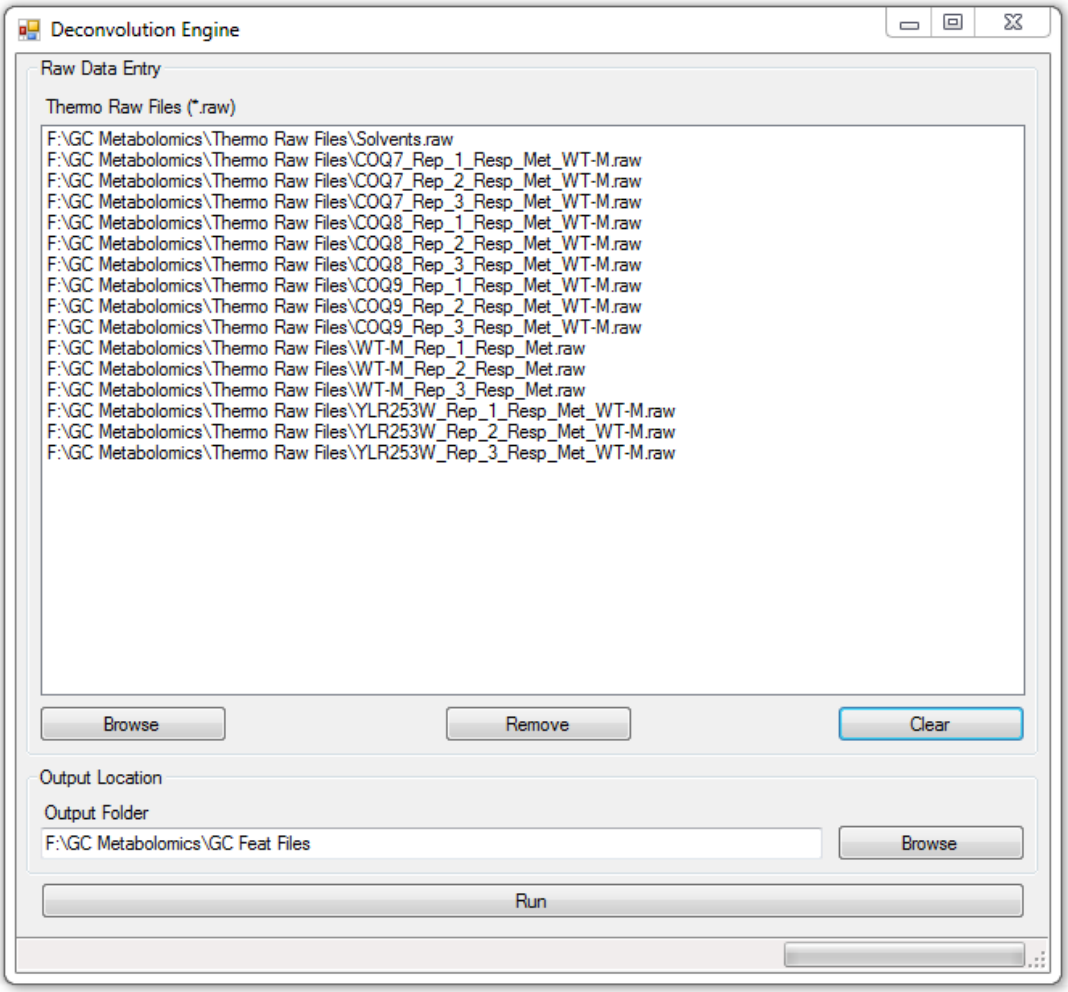


Figure 3.2: Deconvolution Engine. The GUI program used for extracting chromatographic features from raw GC-Orbitrap MS data files. Users can drag and drop data files into the 'Thermo Raw Files (*.raw)' pane and specify an output directory. One corresponding .gcfeat file is created for each user provided MS file.

considered a "hard" ionization technique as it induces molecular fragmentation during the ionization process[8,9]. It is noteworthy that this ionization process is highly reproducible, and yields a characteristic set of fragments from individual analytes. Following ionization, these charged fragments are subjected to mass analysis in the Orbitrap. The exclusive analysis of molecular fragments differentiates GC/MS-based metabolomics from related LC/MS studies. In the case of the latter, intact parent precursors are measured prior to selection for MS^2 analysis, yielding both an intact parent mass and a fragmentation spectrum.

Given that analytes elute continuously throughout the course of a GC gradient, the MS is constantly sampling pools of fragments. This results in acquisition of chimeric spectra comprised of fragment peaks derived from multiple precursors. Since putative identifications are assigned by matching fragmentation spectra against reference spectra (acquired from pure standards), resolving fragments from individual compounds is critical. Furthermore, quantitation is facilitated by use of at least one ion from a specific compound. In this regard, it is useful to identify and group together all fragments which are derived from a singular precursor. Deconvolution Engine was designed specifically to perform this procedure.

To begin, all chromatographic features—a collection of individual *m/z* peaks observed across consecutive scans—need to be extracted from a raw data file. This

process is done with no *a priori* knowledge of what fragments have been monitored. We start with the first acquired MS data scan. For each observed *m/z* peak, a new `Feature` object is created. This object serves as a container for all peaks having the same *m/z* observed across consecutive scans. The associated *m/z* peak is added to the `Feature` object and it's retention time (RT) is noted. Each of these objects is initially flagged as '*active*' asserting that we anticipate that a peak of similar *m/z* exists in the following scan. From here, we move to the second MS scan in the raw data file. For each active `Feature` we attempt to find a corresponding peak which has an *m/z* that falls within a ± 10 ppm tolerance of the measured *m/z*. If peak is found, it is added to the `Feature` along with its corresponding RT. If no peak is found, the `Feature` is flagged as '*inactive*', asserting that we do not expect to find additional *m/z* peaks derived from this particular fragment in subsequent scans. For all *m/z* peaks measured in the second scan which do not have an *m/z* matching that of an '*active*' `Feature`, a new `Feature` object is created, and again, it is flagged as '*active*.' This process is performed iteratively and to completion such that all scans and all *m/z* peaks are considered. Only `Features` which contain five or more consecutive *m/z* peaks are stored for further analysis.

Once the *m/z* peak aggregation process has completed, we are provided with a set of complete `Features`. A consideration which must be made is that not all

measured *m/z* peaks are derived from biologically-relevant species—some measured signal stems from background noise. It is desirable to omit `Features` which arise from background noise from further analysis. We can distinguish analyte signals from noise signals based on chromatographic peak shape. We anticipate that analyte signals will exist as a unimodal distribution wherein they rise to a local maxima and then fall. We note that signals from noise will be somewhat random, and instead expect fluctuations around a central mean, without a characteristic peak shape.

To make chromatographic peak shapes more obvious we apply an 11-pt boxcar average smoothing filter to all complete `Features`. From here we utilize a peak-splitting algorithm which separates multi-modal `Features` into unimodal `Features`, by identifying local minima and maxima. Essentially, for each chromatographic `Feature`, a first derivative is calculated and all patterns wherein a zero-crossing point is padded by positive and negative data points (left and right, respectively) before returning to a zero baseline are stored. Each instance of this pattern is saved as a new `Feature` object. Previously acquired `Feature` objects which do not meet this pattern are subsequently omitted.

At this point in execution, we have extracted a set of chromatographic `Features`, all of which meet our aforementioned "signal" criteria. Each of these `Features`

is assumed to represent a molecular fragment, derived from a singular parent precursor. For the purpose of obtaining a "pure" fragmentation spectrum, and identifying candidate ions for quantitation, it is necessary that features stemming from individual parents are grouped as such. Fragments from singular parents will co-elute, and they will reach an elution apex at the same time. We have developed and employ an algorithm which groups together fragments based on similarities in elution profiles.

Briefly, starting from all valid chromatographic `Features`, we group `Feature` objects into smaller sets based on similar retention times. All `Features` are ranked in descending order based on apex intensity. Starting with the most abundant `Feature`, we calculate a time window centered around its apex where measured intensities are $\geqslant$95% of that `Feature`'s apex intensity. We then iterate over all other valid `Features` in the set and group together those which have an apex within the calculated time range. These `Features` are then placed into a new `Feature Group` object—an object containing all features assumed to stem from a singular parent—and removed from further consideration. This process is carried out iteratively until all `Features` have been placed into a single `Feature Group`. `Feature Groups` containing fewer than five features are discarded from further consideration. Finally, we extract a consensus spectrum from each of these `Feature Groups` by creating a

peak for each Feature, with corresponding *m/z* and apex intensity.

Following aggregation of `Features` into `Feature Groups`, we export all results to a SQLite database file (the extension .gcfeat is used) to be used downstream in our pipeline. This database consists of three individual tables. First, is a total ion current (TIC) chromatogram table which contains individual entries for each data point in the TIC chromatogram (RT and TIC). Second, is a table containing entries for each valid chromatographic `Feature` including specific columns for *m/z*, apex RT, apex intensity, a string representation of the smoothed feature (RT_1:$Intensity_1$;RT_2:$Intensity_2$;…RT_n:$Intensity_n$), and a unique numerical identifier. Finally, a table describing all `Feature Groups` is included with columns containing apex RT, a string of all included feature IDs, and a string representation of all spectral peaks (m/z_1:$Intensity_1$;m/z_2:$Intensity_2$;…m/z_n:$Intensity_n$).

Deconvolution Studio

Design. Deconvolution Studio is the second software tool in our pipeline and serves to identify biologically relevant target species via comparison of a representative 'analyte' file against a 'blank' file. This target list is saved to a separate file and can be modified at the a user's discretion. We provide users with the ability to adjust the constituent features in all feature groups, select/deselect target feature

groups (read, molecules) for downstream quantitation, and to intelligently choose ions to use for quantitation. This tool was built using C# .NET and we provide a functional GUI which supports a number of tabs and embedded graph panes to facilitate the processes above **(Figure 3.3)**. This tool is designed such that individual files can be opened and edited *ad infinitum* and supports change save functionality. Here we will describe some of the key functions which this software tool supports.

GC Master Creation. As mentioned, one of the critical functions which Deconvolution Studio supports is creation of a "target" list (.gcmaster file) to be used for downstream quantitation. Upon initial launch, a user will select both 'analyte' and 'background' .gcfeat files—files containing deconvolved `Features` and `Feature Groups`, produced by Deconvolution Engine. After clicking 'Create GC Master File,' `Feature Groups` from both files are loaded into memory for subsequent comparison. The expectation is that all biologically relevant species will be present exclusively in the 'analyte' .gcfeat file. We note that the inclusion of a TMS-derivatization step in GC/MS sample preparation procedures[10] is common, and adds many background chemicals which give rise to spurious signals that should be omitted from further analysis. For each feature group in the 'analyte' file, the most abundant `Feature` is selected, and searched for in the corresponding 'background' file using a $\pm$10 ppm *m/z*, $\pm$0.05 minute RT, and 2x apex intensity

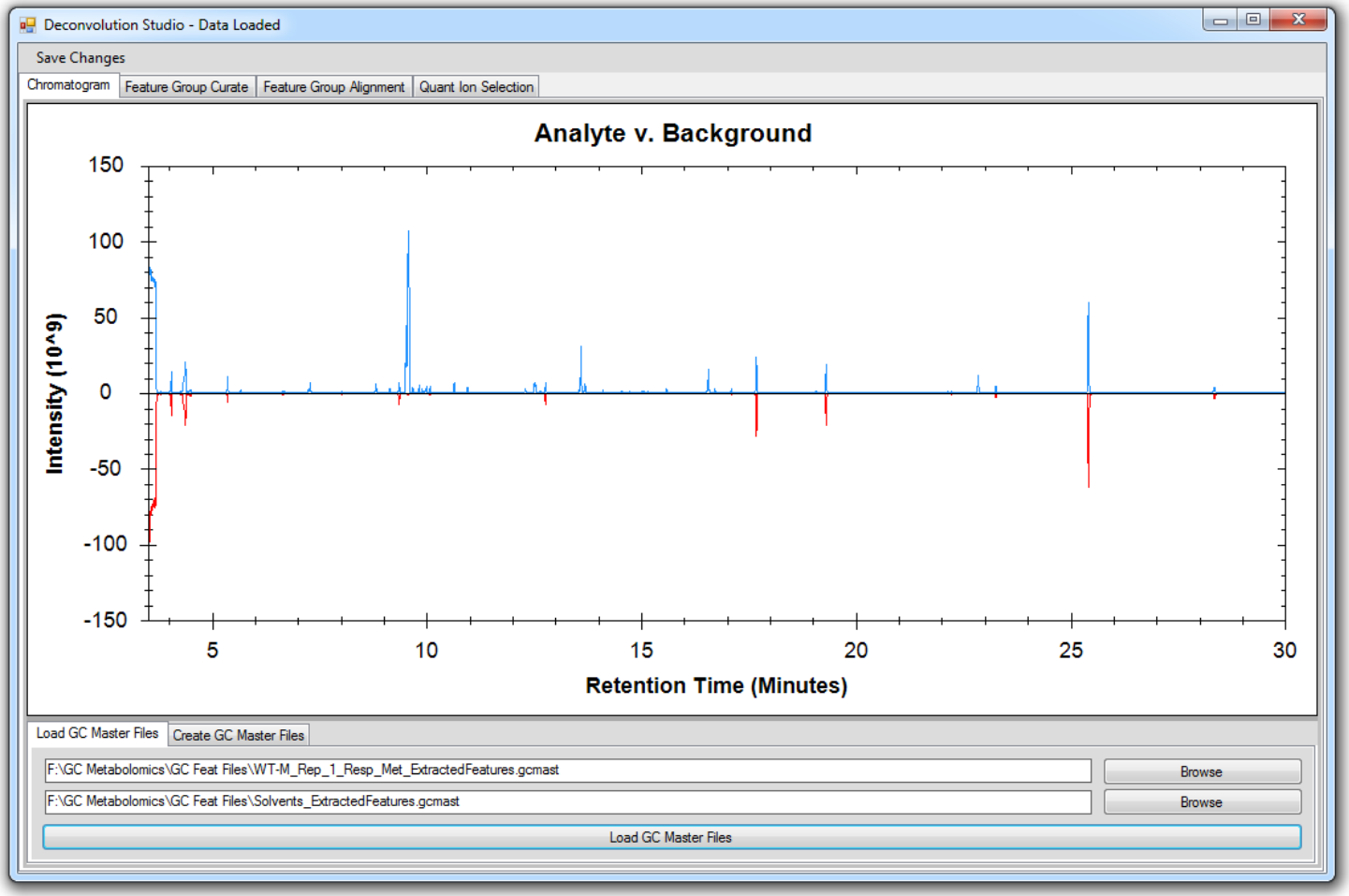


Figure 3.3: Deconvolution Studio. The GUI program used to create and edit .gcmaster files, used for downstream quantitation. Creation of 'analyte' and 'background' .gcmaster files can be done under the 'Create GC Master Files' input tab. These files can be further edited by loading them into the program. Upon load, opposing TIC chromatograms from the 'analyte' (blue) and 'noise' (red) files are displayed here.

tolerance. If no matching record is found, it is assumed that the selected `Feature Group` is biologically relevant, and it is listed as a downstream quantitation target. Additionally, a single quant ion is automatically selected. By default, the most abundant ion—which is not a known TMS fragment—is chosen, although this can be updated later at the user's discretion. All of this data is saved to the .gcmaster file, which can then be loaded back into Deconvolution Studio and edited. The following processes described occur after a .gcmaster file has been created and re-loaded into the tool.

Feature Group Curation. The construction of `Feature Groups` is handled in a completely automated fashion by Deconvolution Engine, using a set of predetermined parameters. While the employed algorithms are highly performant, we recognize that they can still fail. In these situations it is desirable to provide users with added control to correct improperly grouped `Features`. Under the 'Feature Group Curate' tab **(Figure 3.4)**, all `Feature Groups` are displayed in a list with both apex RT and peak count indicated. Selection of any `Feature Group` from this list will display related data in the associated graph panes. In the left pane, all `Features` surrounding the group's apex RT are shown. `Features` belonging to the selected `Feature Group` are shown in red, and all others are shown in gray. In the right pane, a mass spectrum containing peaks corresponding to all constituent `Features`

is displayed.

We support functionality such that `Features` can be added to and removed from the selected `Feature Group`. All `Features` can be selected from the left pane by double-clicking a curve of interest. Upon select, the chosen curve will change color and line thickness. If the `Feature` belongs to the current `Feature Group` it can be removed by clicking the 'Exclude' button. Alternatively, if it is not currently a member, it can be added by clicking 'Include.' In either case, after the composition of the `Feature Group` is changed, the right-hand spectrum is updated to reflect the current state. Additional functionality is supported to allow a user to lookup a `Feature` with characteristic *m/z* and RT, and then create an entirely new `Feature Group`. This option is useful for the case of lowly abundant metabolites which may not have been incorporated into a `Feature Group` containing the minimum number of required features. Once changes have been made here, a user can click the 'Save Changes' button and all edits will be stored in the .gcmaster file.

Feature Group Target Selection. One of the most notable advantages of our pipeline is that it supports untargeted quantitation, which negates the need for identifications to enable extraction of molecular abundances. In any untargeted, or discovery-based, quantitation routine it is necessary to designate a set of species which you will attempt to quantify across all samples. Our pipeline is designed around the idea

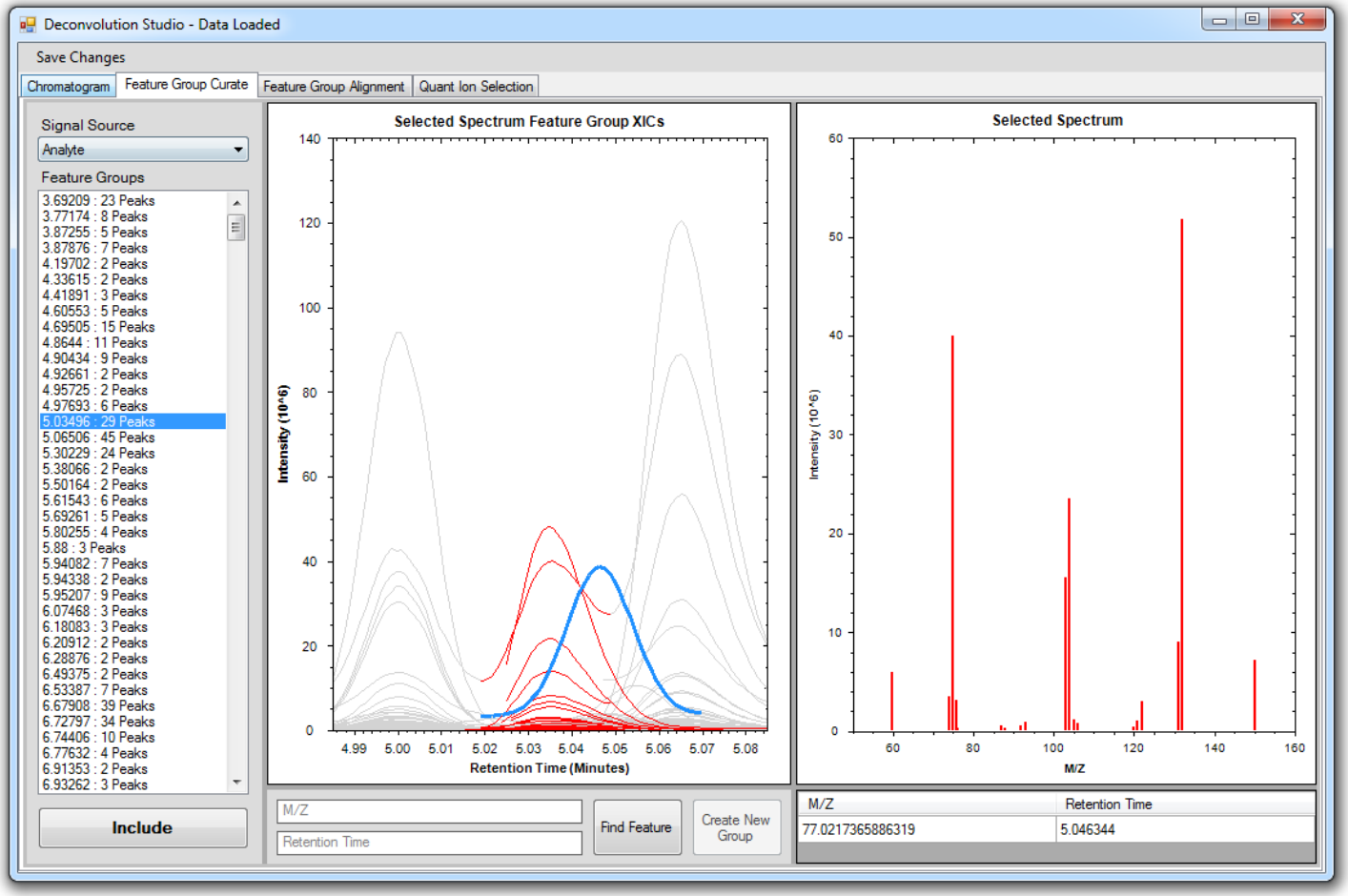


Figure 3.4: Deconvolution Studio–Feature group curation. `Feature Groups` can be manually modified under this tab. All `Feature Groups` from the 'analyte' .gcmaster file are displayed in the list along the left-hand side of the GUI. Selected `Feature Groups`, and closely eluting chromatographic `Features`, are displayed (left pane) upon click. The composition of a selected `Feature Group` can be edited in this tab. By double clicking any `Feature` currently contained in the `Feature Group` or otherwise, it will become selected—highlighted in blue as shown here—and can be added or removed. The `Feature Group`'s consensus mass spectrum (right pane) is automatically updated following `Feature` addition or removal.

that a control will be used for normalization purposes—which we acknowledge will not always be the case. This assumption greatly simplifies the quantitation process. Using this approach, we can determine which `Features` to quantify based on a single experiment, rather than multiple experiments. This streamlined `Feature` selection process opens to the door to manual result-checking and curation which we provide in Deconvolution Studio.

All `Features` to be quantified are initially selected during the GC master creation process—as described above. By navigating to the 'Feature Group Alignment' tab **(Figure 3.5)**, users can observe all `Feature Groups`—from the provided analyte file—displayed in a list. This list contains associated column labeled 'Included' which reflects whether that `Feature Group` is to be quantified across all MS data files. This tab provides four individual graph panes, each of which provides the user with a unique data view that serves to inform whether a `Feature Group` ought to be targeted for downstream quantitation.

Briefly, the views shown include a look at all `Features` from the analyte file over a narrow time range (upper-left), and all `Features` eluting over the same time range in the background file (lower-left). Individual `Feature Groups` can be selected in these two panes by double-clicking the group of interest. In the lower-right hand pane a reflection of all chromatographic `Features` across both analyte

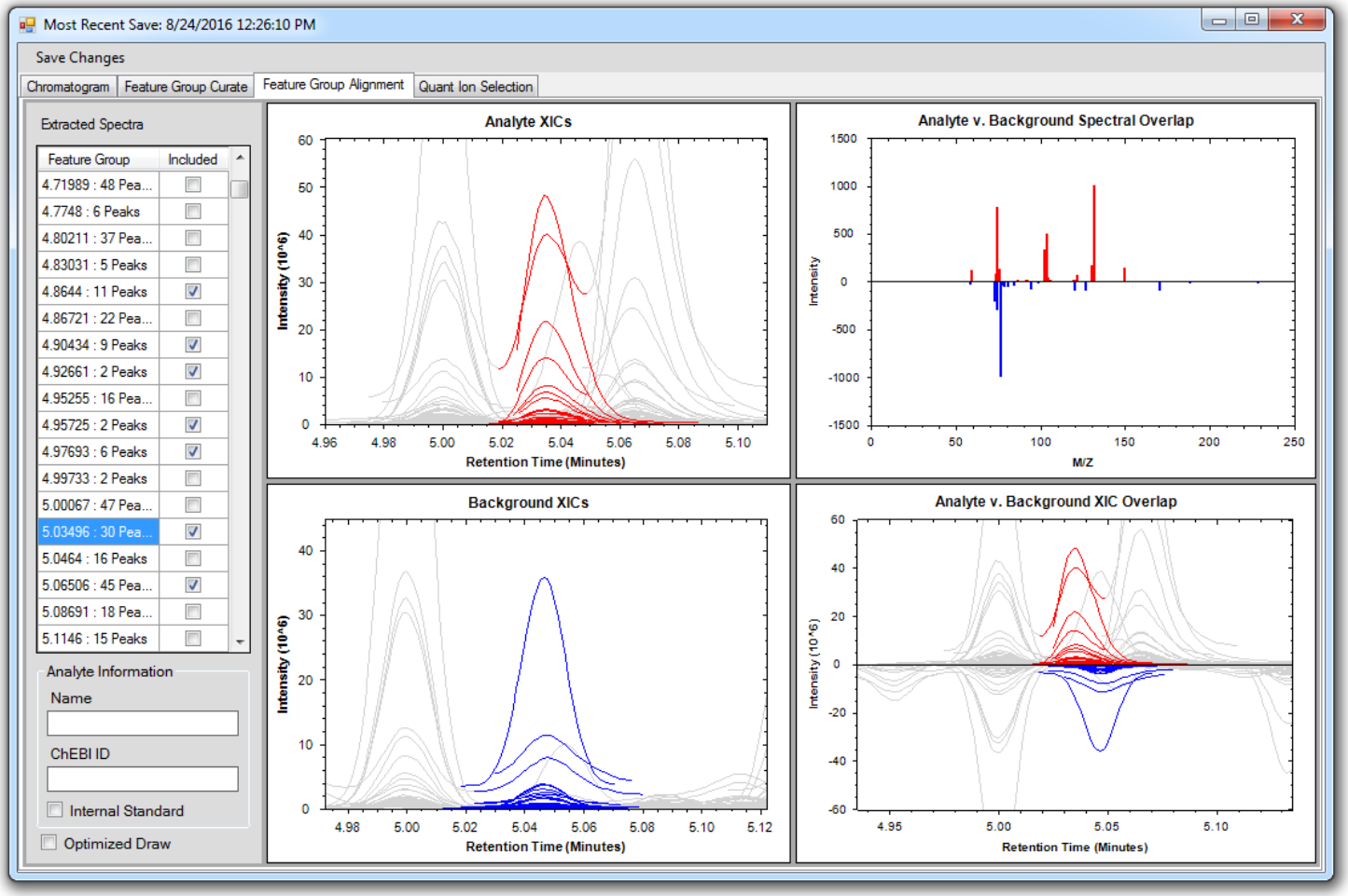


Figure 3.5: Deconvolution Studio–Feature group target selection. `Feature Groups` can be selected as targets for downstream quantitation under this tab. All `Feature Groups` from the 'analyte' .gcmaster file are displayed in the list along the left-hand side of the GUI. Clicking any `Feature Group` updates all four graph panes. The upper-left, and lower-left panes show a selected 'analyte' `Feature Group` and selected 'background' `Feature Group`, respectively, along with other closely eluting `Features`. The lower-right hand pane shows a reflection of the data these two panes, and the upper-right hand pane shows a reflection of mass spectra from the selected `Feature Groups`. Users can choose whether to use a selected 'analyte' `Feature Group` for quantitation by checking the appropriate box under the 'Included' column. Further, users may provide a name, and ChEBI identifier to the chosen `Feature Group` under 'Analyte Information' box along the left-hand side of the GUI.

and background files are shown, with selected analyte and background Feature Groups indicated in red and blue, respectively. Finally, in the upper right hand corner, a reflection of fragmentation spectra from both chosen Feature Groups is displayed. Collectively, these plots should provide users with the requisite information to determine whether a Feature Group is unique to the analyte file, and should therefore be targeted for quantitation. In this tab, we also provide users with the ability to provide both a name and ChEBI identifier for targeted Feature Groups if desired.

Quant Ion Selection. For quantitation purposes, we have developed our pipeline to utilize intensity measurements from a single ion to represent the abundance of a particular molecule. Intuitively, one might want to choose the most abundant Feature from a Feature Group (read, molecule) as this has the best chance of being observed in times of lowered abundance. However, we recognize that many fragments are shared between molecules, which can obscure the origin of signal when dealing with closely eluting compounds. Instead, we assert that it is preferable to choose a fragment with *m/z* dissimilar to other closely eluting fragments.

Under the 'Quant Ion Selection' tab **(Figure 3.6)**, all Feature Groups specified as targets for quantitation are listed. By clicking on any Feature Group, the associated spectrum is displayed. All ions from the Feature Group's consensus spectrum are

displayed in an associated list—labeled 'All Ions.' The ion selected for quantitative purposes is displayed in a separate list—labeled 'Selected Ion'. Individual ions can be moved between these lists, however only one ion is allowed on the 'Selected Ions' list at any given time. The ion chosen for quantitation is differentially colored in the displayed consensus spectrum, and a chromatographic profile is shown in a separate pane.

To determine uniqueness of the chosen quant ion, a user can load a Thermo Raw MS data file. Following upload, chromatographic traces extracted directly from the raw MS file—using a ± 10 ppm tolerance—and are displayed with a 15 second span surrounding the fragment's apex RT. Ideally, the only observable signal at this *m/z* would be derived from the fragment of interest. Such is not always the case, but it is desirable that there is some sort of baseline separation between fragment signals. Another component which must be considered when selecting a quant ion is signal-to-noise (S/N). If a very lowly-abundant fragment is chosen for quantitative purposes, this will likely preclude the detection of a large decrease in abundance for the particular molecule. Similarly, if a fragment intensity is exceedingly high, detection of large abundance increases will be challenging. Here we have provided users with multiple data views which inform the quant ion selection process. All changes in quant ion selection are stored locally until a user

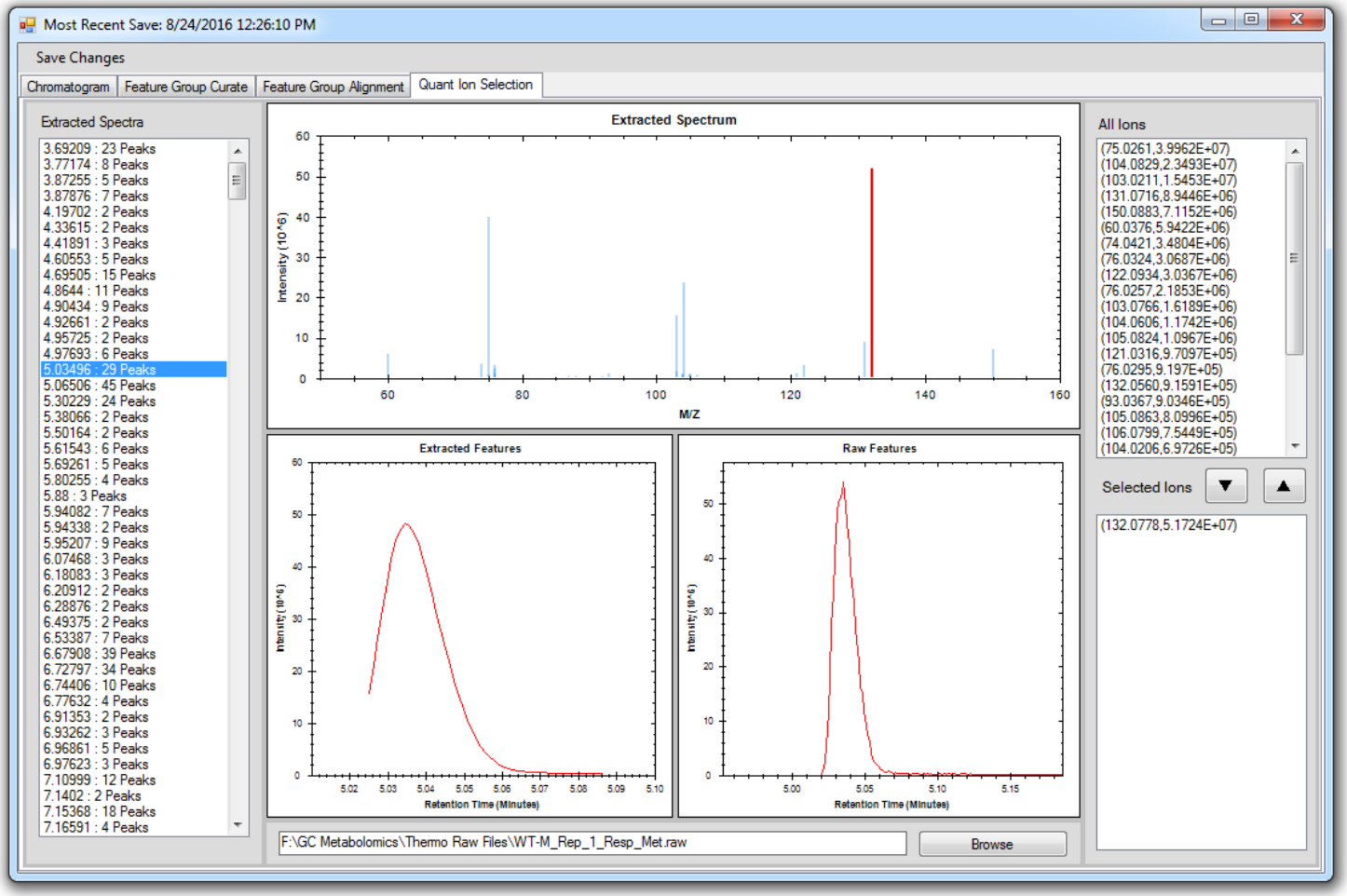


Figure 3.6: Deconvolution Studio–Quant ion selection. Ions to be used for quantitation can be selected for each targeted `Feature Group` under this tab. All `Feature Groups` from the 'analyte' .gcmaster file are displayed in the list along the left-hand side of the GUI. Clicking any `Feature Group` updates all three graph panes, and the associated ion lists along the right-hand side of the GUI. A mass spectrum is displayed, with quant ion highlighted along the top. The quant ion's chromatographic elution profile—as extracted from Deconvolution Studio—is shown in the lower-left pane. Optional inclusion of an associated Raw MS data file will trigger display of the quant ion's elution profile over an extended time range, prior to any smoothing. Users can select an ion for quantitation by moving values between the 'All Ions' and 'Selected Ions' lists.

clicks 'Save Changes' at which these modifications are stored in the .gcmaster file.

Experiment Builder.

Design. One of the main objectives of our pipeline is to produce outputs which can be immediately visualized and explored. To this end, inclusion of statistical analyses in our processing is desirable. In order to enable meaningful statistical testing, the organizational hierarchy of the experiment must be known so that data can be properly compared. One of the inputs in our downstream tool GC-Quant, is a text file (*.gcexp) which describes how each MS data file is related to the others. This file can be created using our tool, Experiment Builder. To facilitate easy editing, the .gcexp file is written in a semicolon-delimited format, which lends itself to editing in our tool as well as in Microsoft Excel. Furthermore, these files can easily be generated programmatically. The general structure of each line in this file is as follows: .gcfeat file path, replicate name, condition name, control condition. This structure provides all information required for statistical analysis to be carried out. We provide a simple GUI, developed in C# .NET, which facilitates construction of this file by means of an editable and interactive table where individual files can be added and associated fields changed with ease **(Figure 3.7)**.

We have designed a flexible and generic experimental hierarchy structure, which

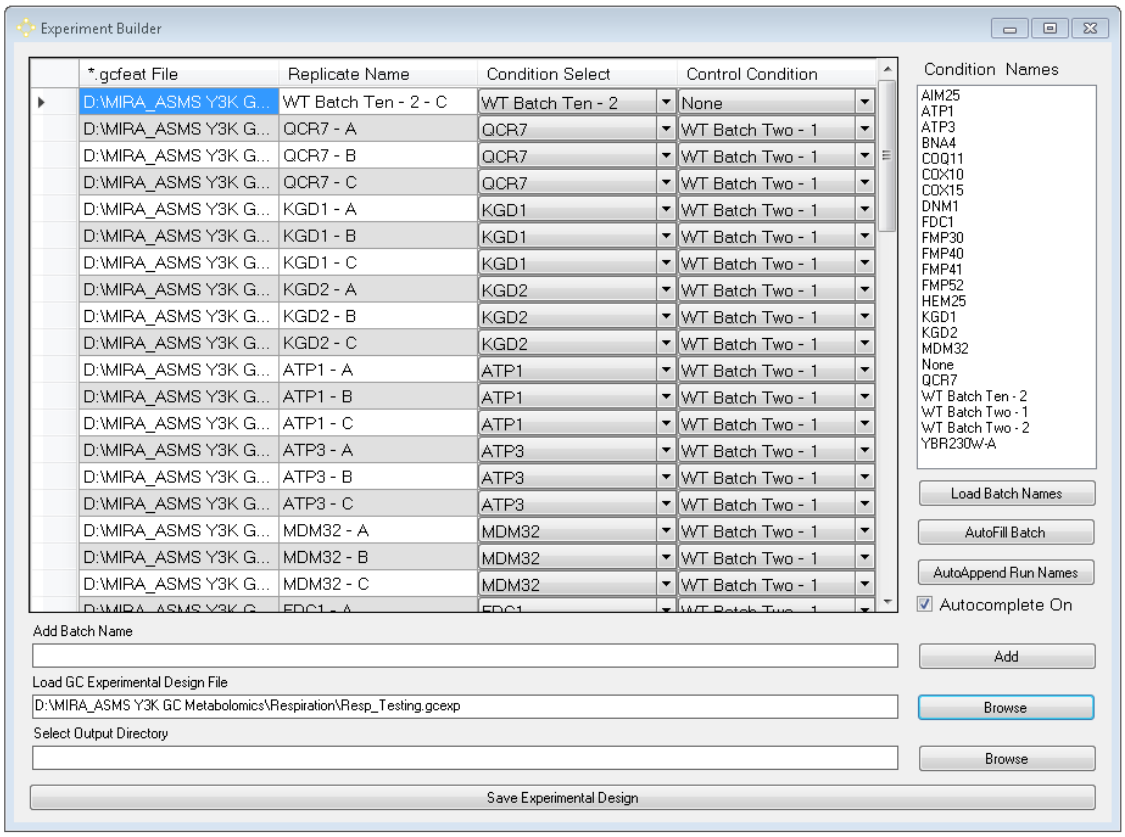


Figure 3.7: Experiment Builder. The GUI program used for creation and editing of a text file (*.gcexp) describing the hierarchical organization of deconvolved MS data files to be quantified. Here, users drag and drop deconvolved .gcfeat files, provide replicate names to each file, as well as 'condition' and 'control' mappings. Edited data can be saved to a new file by clicking 'Save Experimental Design' button.

we require users to utilize for describing the organization of their data. We assume that individual analyses can be grouped into 'conditions,' which encapsulate replicate samples that have undergone similar treatments prior to analysis. These 'conditions' can contain either single, or multiple replicate MS experiments. Further, we note that often these 'conditions' will represent an experimental control which can be used for normalization and comparative purposes. The file structure outlined above is designed such that individual MS data files can be mapped into the described structure.

Function. Experiment Builder is the third software tool in our pipeline and serves to create a text file (.gcexp) containing information on the hierarchical organization of replicate MS experiments (.gcfeat files), to be used in our downstream quantitation tool. The generated text file contains semicolon-delimited values and can be constructed and edited within the provided Experiment Builder GUI, or programmatically generated using custom scripts. Users can drag-and-drop .gcfeat files into the Experiment Builder GUI and each will be added as a new row in the displayed data table. This table consists of four columns—*.gcfeat File, Replicate Name, Condition Select, and Control Condition Select—all of which can be edited. For each added .gcfeat file, users will specify a replicate name under the associated column. Then, users will map individual replicates to 'conditions.' Conditions can be added

by entering a name in the 'Add Condition Name' textbox and clicking the 'Add' button. The right-hand 'Condition Names' list is then automatically populated with the new entry, and it is available for subsequent mapping. The 'Condition Select' column contains dropdown lists with the names of all user-specified conditions. For each replicate entry, users will select a condition name from this list. We have added an 'AutoAppend Run Names' button which attempts to map similarly named replicates with the same condition to expedite the .gcexp file creation process.

After all replicates have been mapped to appropriate conditions, users will indicate control condition mappings. We recognize that often several conditions will be profiled in a given project, and that many conditions will map to a single control. These control conditions can be specified by selecting a single condition from the dropdown lists under the 'Control Condition Select' column. After performing this procedure once, all other replicates associated with a non-control condition are automatically mapped to the specified control, and the associated fields are updated appropriately. Control conditions have no associated control, and, as such, are left blank. Once all .gcfeat files have been added, and mappings completed, users can export their work by selecting an output directory, and clicking the 'Save Experimental Design' button. This triggers creation of a new timestamped .gcexp file which is ready for use in downstream applications.

GC-Quant

Design. GC-Quant is the fourth program in our pipeline which serves to perform all quantitation and normalization procedures across a set of user-provided MS data files. This program accepts a .gcmaster file (Deconvolution Studio output) and a .gcexp file (Experiment Builder output) as inputs. The former provides information about which chromatographic features should be extracted from all files for quantitative analysis. The latter provides a file location for each deconvolved .gcfeat file, as well as, information about the hierarchical organization of replicate MS experiments. All inputs can be easily imported into our functional GC-Quant GUI developed in C# .NET **(Figure 3.8)**. Upon execution of this tool, all previously specified .gcfeat files are loaded into the program and listed in the GC-Quant progress window. During execution, progress is reported back in real-time and displayed in the GUI. After all underlying processes complete, a .gcresults file is created which contains normalized quantitative data and results from statistical analyses. This file can be imported into our developed data visualizer—GC-Viewer—where all data can be readily explored. The algorithms employed by GC-Quant to ensure that meaningful quantitative information is appropriately extracted and transformed are described in detail here.

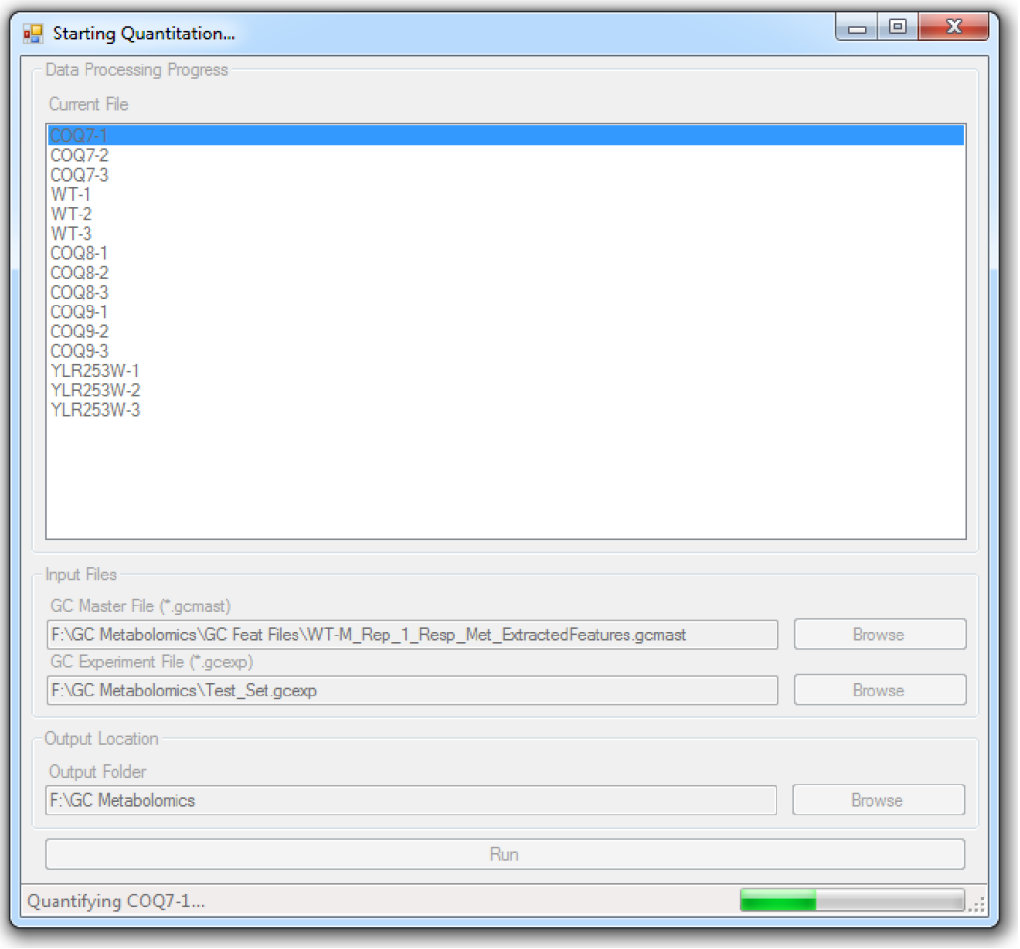


Figure 3.8: GC-Quant. The GUI program used for quantitation, normalization, and statistical analysis of all user-selected MS data files. Users provide by a .gcmaster (target list information) and .gcexp (file location hierarchical organization) files as inputs. Upon launch all files to be quantified are displayed in the 'Data Processing Progress' window, and progress is updated in real-time.

Chromatographic Realignment. One of the most substantial benefits afforded by GC/MS metabolomic analysis, is a high-degree of chromatographic reproducibility. Typically, only very small shifts in retention of individual molecules are noted, even across extended periods of data acquisition. Chromatographic realignment algorithms are frequently employed by MS quantitation packages to enable comparison of molecular abundances for a conserved set of features across MS data files. This high level of chromatographic similarity in GC/MS studies alleviates the computational burden of realignment—which is often challenging in orthogonal LC/MS analyses—to compensate for systematic shifts or warping. That said, we have developed our quantitation tool to account for deviations in molecular retention, to ensure that our quantitative results are robust.

The GC-Quant tool utilizes a .gcmaster file as an input—which is generated from a single representative file—that provides information about what features should be quantified across all files. This .gcmaster file also contains the TIC chromatogram from the associated MS data file used for its creation. We utilize this TIC chromatogram as a reference for all realignment procedures (i.e., all replicate MS files are aligned to this chromatogram prior to quantitation). To account for any warping, we perform all alignments on a molecule-by-molecule basis. Briefly, for each molecule listed as a target in the .gcmaster file, a one minute segment of

the TIC chromatogram—centered around the molecule's apex RT—is extracted. Within each analyte file to be quantified, a TIC segment covering the same retention time range is also extracted. We anticipate that these two segments will contain similar prominent features which we exploit for our realignment purposes. We aim to calculate a characteristic time offset (.gcmaster elution time – expected analyte elution time) for the current molecule. We begin by rastering the analyte TIC segment across the .gcmaster TIC segment—in 0.005 minute increments—and calculating a dot product score for overlapping sections at each point. This rastering process is carried from -0.15 minutes to +0.15 minutes by default. The offset position which yields a maximal dot product score is stored as the characteristic RT offset for the selected molecule within the analyte file. This value is later used for feature location purposes. Figure shows TIC chromatogram segments from WT yeast samples before and after chromatographic realignment **(Figure 3.9)**.

Metabolite Quantitation. The deconvolution processes performed by Deconvolution Engine, negate the need to go into individual raw MS data files to locate and extract abundances for targeted metabolites. Instead, each .gcfeat file already inherently contains all chromatographic `Features` which could possibly be used for quantitation. Hence, all that is necessary to extract abundances for targeted metabolites, is to identify `Features` in a given .gcfeat file which have matching

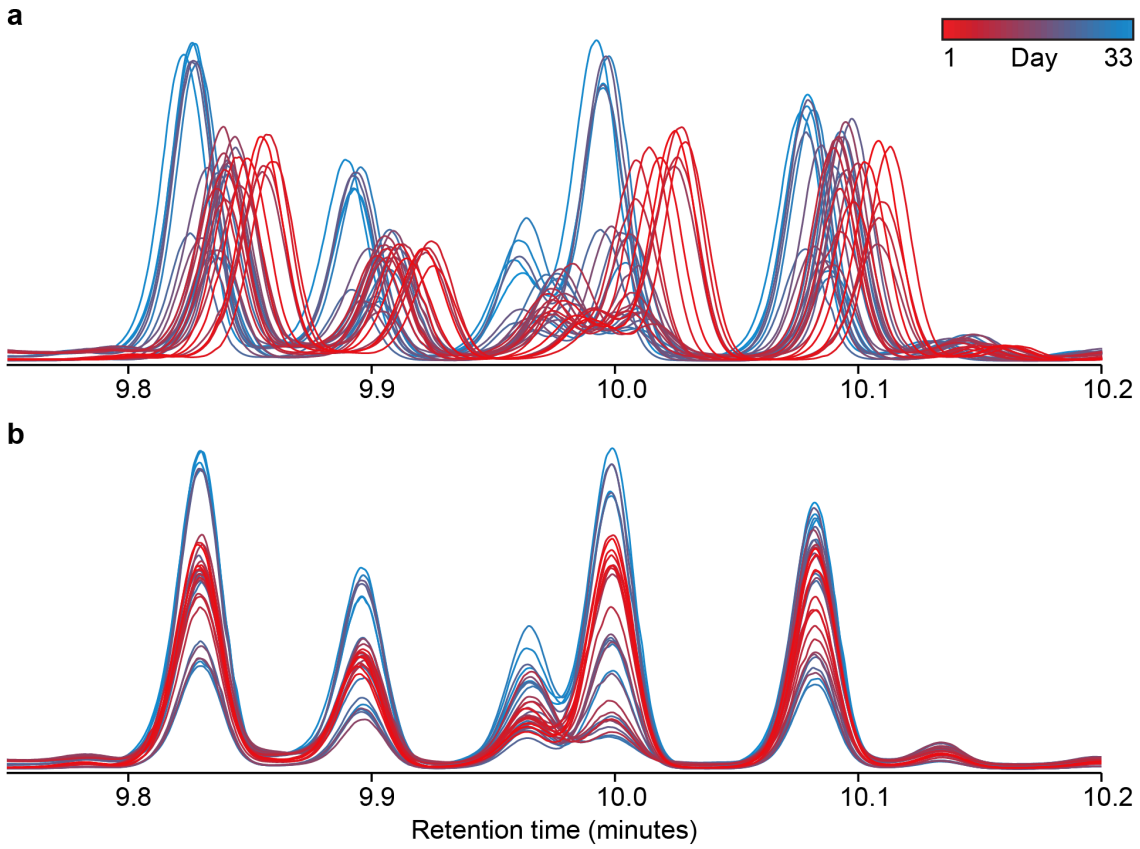


Figure 3.9: Chromatographic realignment. TIC chromatogram segments from 33 raw GC-Orbitrap MS data files acquired across consecutive days before and after realignment. **a.)** Chromatographic profiles extracted directly from raw MS data files. **b.)** Chromatographic profiles following calculation and application of a characteristic RT offset.

m/z and RT values. These matched `Features` can then be normalized and used for comparative analyses.

At this point in the execution, a RT offset for each targeted molecule has been calculated within each analyte .gcfeat file. For each molecule, the GC-Quant algorithm locates all metabolites which have a RT within a ± 0.075 minute span of the expected RT (.gcmaster RT – calculated offset), and a *m/z* within a ± 10 ppm tolerance of the specified quant ion. If no matching `Features` are found, the algorithm moves to the next molecule. If one matching `Feature` is found, it is associated with the targeted molecule, and stored locally for downstream normalization. If more than one matching `Feature` is found—which is rare given the restrictive RT and *m/z* tolerances used—the `Feature` having a RT closest to the expected RT is designated as the correct match, and also stored locally. This process is repeated iteratively and to exhaustion for all targeted metabolites, across all user specified .gcfeat files. Upon completion all metabolite abundances are $\log_2$ transformed.

Normalization. We have designed our pipeline to be used for relative quantitation rather than absolute quantitation. The tacit assumption of this approach is that overall metabolite levels are similar between samples, but that the relative abundances of these species will differ. In our normalization procedure, we first create a 'virtual replicate analysis.' For each targeted metabolite, we identify the median quantified

abundance, considering all replicates where the metabolite was located. This value is then assigned as the associated abundance for the selected metabolite within our virtual replicate. We repeat this process for all targeted metabolites. This virtual replicate is used as a reference for normalization of all profiled replicates. For each replicate experiment, we calculate the total abundance of all located metabolites, as well as the total abundance of the same set of metabolites within our virtual replicate. Then, using a TIC-based normalization approach, we scale the abundance of each replicate metabolite equivalently such that the total abundance is equal to that of the virtual replicate. This process is repeated for iteratively for all user specified .gcfeat files.

Statistical Analysis. At this point in execution, each user specified replicate MS analysis contains normalized quantitative information for all metabolites that could be located. In order to expedite data analysis, we perform some basic statistics within the GC-Quant program, which are subsequently stored in a provided result file (.gcresult). Granted that we require information about the mapping of replicate analyses into conditions, we can perform statistical analyses at this level. For each metabolite, we calculate the average condition abundance and standard deviation, using replicate abundance measurements. If a control has been specified, we calculate fold changes for each molecule by subtracting $\log_2$ control averages

from $\log_2$ condition averages. Additionally, we calculate a *p*-value for each fold change (two-sided Student's t-test; homostatic), which reflects the significance of the measured perturbation. All calculated values are written to an output file and are used for automatic generation of data visualizations in our GC-Viewer utility.

Output File. We export results from GC-Quant in a SQLite database file (*.gcresults) which can be loaded into GC-Viewer and explored or ported to text-based formats using any SQLite data viewer. This output file contains all qualitative and quantitative experimental information in a series of well-defined data tables. A single entry is stored for each replicate MS experiment in the table named 'Replicate_Table,' which contains replicate name, a unique replicate identifier, associated .gcfeat file path, condition name, a unique condition identifier, control condition name, and control condition identifier. Similarly, a single entry is stored for each condition in the 'Condition_Table' table which contains condition name, a unique condition identifier, a comma-delimited list of associated replicate identifiers, a comma-delimited list of associated replicate names, a control condition identifier, and a control condition name. Replicate quantitative data is stored under in a table called 'ReplicateQuant_Table.' Here, each quantified molecule, within each replicate, is stored with associated replicate and condition identifiers—as mentioned above—along with an apex RT, RT offset, quant ion *m/z*, and apex intensity. Aggre-

gated condition quantitative data is stored in a table called 'ConditionQuant_Table.' In this table, we store a semicolon-delimited string of all replicate intensities, an average intensity, a standard deviation, a control-normalized intensity ($\log_2$ fold change), and a *p*-value. Following insertion of data into data tables, indexes are added where appropriate to expedite queries employed by GC-Viewer.

GC-Viewer

Design. GC-Viewer is the fifth and final software tool in our pipeline which enables users to visualize and explore their quantified metabolomics data. This standalone software utility accepts .gcresults files—outputs from GC-Quant—as inputs and converts all associated data into interactive plots. GC-Viewer is a functional GUI (developed in C# .NET) with graph panes (ZedGraph) embedded across multiple tabs **(Figure 3.10)**. These individual graph panes can be used to automatically generate interactive visualizations which enable users to rapidly identify measurements and trends of interest. To facilitate communication of results from our pipeline, all generated visualizations can be exported to SVG files which can be manipulated in graphic editing packages such as Adobe Illustrator or Microsoft Powerpoint. Here, we describe the visualizations which can be generated within GC-Viewer, as well as the interactivity which this tool supports.

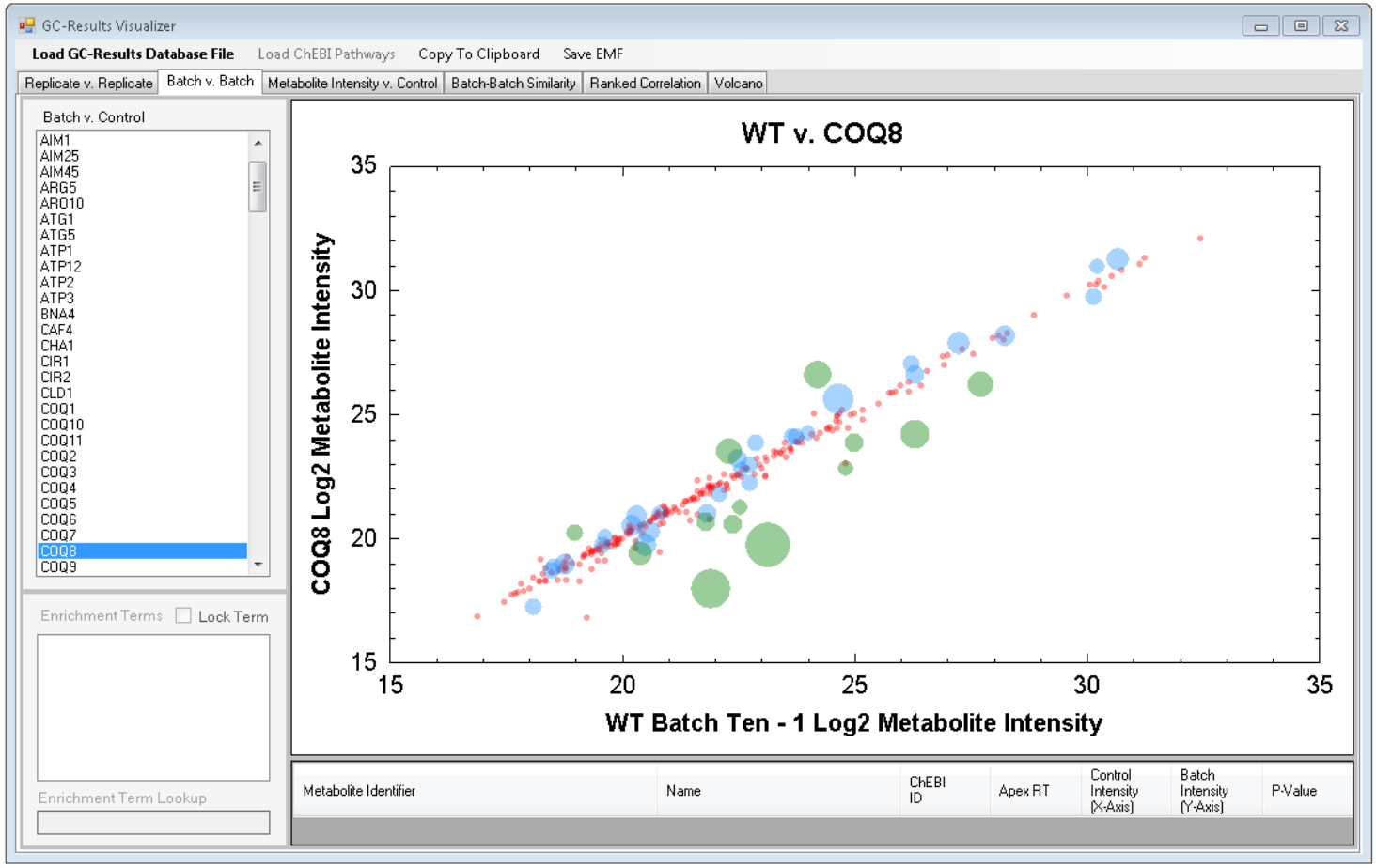


Figure 3.10: GC-Viewer. The GUI program used for visualization of quantitative and statistical results from GC-Quant. This program accepts a *.gcresults file which can be loaded by clicking the 'Load GC-Results Database File' button. Visualizations are automatically created and displayed across all six tabs. Generated plots can be saved to a file or copied to the clipboard by clicking the appropriate buttons.

Function. GC-Viewer viewer consists of six separate visualization tabs, each of which can be used to create a unique view of a quantitative metabolomics data set **(Figure 3.11)**. Upon launching the tool, users are prompted to select a .gcresults file for further analysis. Once a file is selected, all underlying data is loaded into the software tool, and input selection lists are populated appropriately across all tabs. Under the 'Replicate vs. Replicate' tab, users can compare metabolomic profiles from any two replicate MS analyses by selecting desired condition and nested replicate options. These data are displayed as a scatter plot with $\log_2$ metabolite abundances along the x- and y-axes. A Pearson correlation coefficient (R^2) is calculated as a metric of similarity between replicate profiles, and reported in the graph pane.

Under the 'Condition vs. Control' tab, users can explore molecular perturbations within any condition relative to its specified control. Data are displayed here as a bubble plot with control intensities along the x-axis, and condition intensities along the y-axis. Each data point is colored according to fold change and statistical significance (p-value>0.05 = red, p-value<0.05 and $|FC|<1$ = blue, p-value<0.05 and $|FC|>1$ = green). Additionally, data point size is scaled with increasing significance (read, decreasing p-value). Double-clicking any data point will cause the associated data to be added to a table at the bottom of the tab. We provide a similar view of the

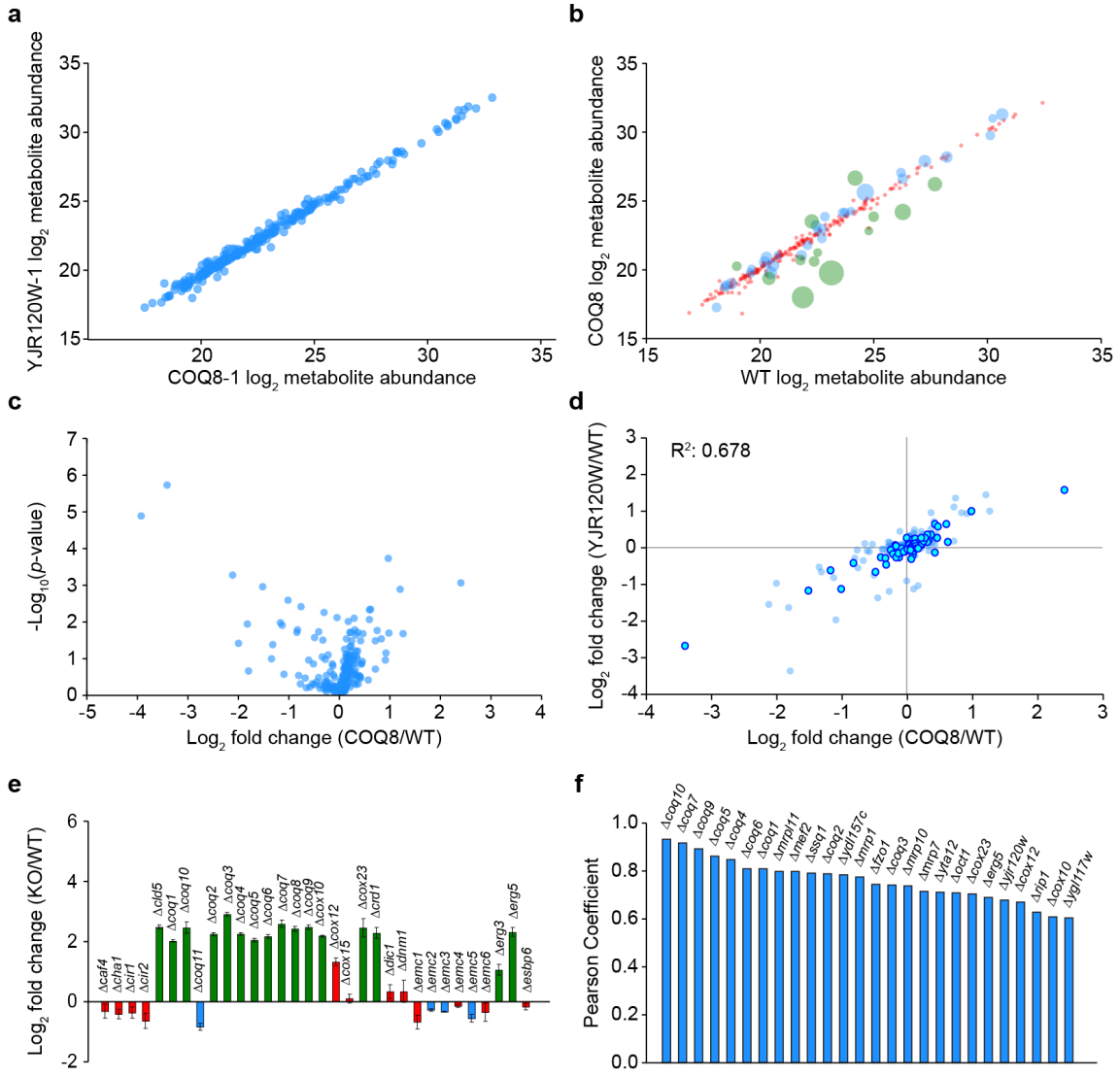


Figure 3.11: GC-Viewer visualizations. Representative visualizations generated by GC-Viewer. **a.)** Replicate vs. Replicate. log_2-transformed metabolite abundances from replicates of COQ8 KO (x-axis) and YJR120W KO (y-axis). **b.)** KO vs. WT. A bubble plot showing average log_2 abundances (n=3) from WT (x-axis) and COQ8 KO (y-axis). Data point size and color are used to indicate fold change and statistical significance. **c.)** KO vs. WT-Volcano Plot. A volcano plot showing average log_2 molecule fold change (n=3 mean log_2[KO-WT]) along the x-axis, and statistical significance (-log10[*p*-value]) along the y-axis.

Figure 3.11: d.) KO vs. KO Correlation. A scatter plot showing average $\log_2$ molecule fold change (n=3 mean $\log_2$[KO-WT]) from COQ8 KO (x-axis) and YJR120W KO (y-axis). Highlighted data points reflect those where p<0.05 across both conditions. The reported Pearson coefficient (R^2) is calculated by fitting a line to all data points. **e.)** Molecule Fold Change. Bar chart showing average $\log_2$ fold change (n=3 mean $\log_2$[KO-WT]) for a single molecule (lactic acid) across KOs, along with standard deviation (error bars). Color reflects statistical significance and fold change. **f.)** KO Correlations. Bar chart reflecting Pearson correlation coefficients (as calculated in **d.**) for all KOs against the COQ8 KO strain.

data under the 'Volcano' tab. Here, complete metabolite perturbation profiles are displayed as a volcano plot—fold change along the x-axis, and statistical significance (p-value) along the y-axis—for each user-selected condition.

Users can analyze fold changes for individual molecules across all conditions under the 'Metabolite Intensity vs. Control' tab. Here, fold changes for a selected molecule across all conditions are represented as a bar chart, with error bars indicating ± 1 standard deviation. Bars are colored according to fold change and statistical significance using the same scheme employed in the 'Condition vs. Control' tab. Any two profiled conditions can be compared under the 'Condition vs. Condition' tab. Here, users will select two conditions for comparison. Abundance fold changes from all molecules profiled across both conditions are displayed as a scatter plot with one condition along the x-axis, and the other along the y-axis. Molecules which are significantly changing (p<0.05) in both conditions are highlighted in light blue. A Pearson correlation coefficient—calculated using all data points—is reported

as a metric of similarity between the two strains. This coefficient is calculated on the fly and displayed in the graph pane. Extended data from any molecule can be displayed in an associated data table by double-clicking a data point of interest.

To enable users to quickly identify which conditions are most closely related to a selected condition, we have constructed a view under the 'Ranked Correlations' tab wherein all Pearson correlation coefficients are displayed in descending order as a bar chart. Users can select a Pearson coefficient cutoff and only correlations exceeding that value will be displayed. All of the described plots can be copied to the clipboard, or saved as an SVG file. We support export functionality to make it easier for users to generate manuscript-ready figures from our data analysis pipeline.

Highlighted Results

The described metabolite quantitation pipeline was developed out of necessity to enable metabolomic profiling of yeast knockout ('Δ*gene*') strains analyzed as part of a large-scale multi-omic profiling experiment (Y3K; described in detail in Chapter 5)[11]. Here, 174 Δ*gene* strains were grown in biological triplicate under two separate growth conditions, and profiled using quantitative proteomic, lipidomic, and metabolomic MS techniques. Considering only the metabolomic portion of

this work, these efforts generated upwards of 1,000 GC-Orbitrap MS data files. This constitutes a rich data set for software development and testing. From these data we anticipate that extracted quantitative metabolomic profiles between replicates of the same Δ*gene* strain would be similar. In **(Figure 3.12)** we highlight correlations between metabolite profiles from three replicates (grown under respiration conditions) of three separate Δ*gene* strains. All of the deleted genes code for proteins involved in disparate biochemical pathways. As such, we anticipate that intra-replicate correlations will be stronger than inter-replicate correlations, an expectation which aligns with observations. Across the entire Y3K data set, we report a median coefficient of variation (CV) of 9.98%, considering all metabolites profiled. In our opinion, this exceptionally small variation speaks not only to the reproducibility of the acquired data, but also to the performance of our quantitation tools.

Aside from testing pipeline performance on the basis of intra-replicate profile similarities, the Y3K data set affords the ability to test for correlations between functionally similar Δ*gene* strains. Many of the Δ*gene* strains profiled in Y3K were knockouts of genes coding for proteins involved in similar biochemical processes. For instance, we analyzed knockouts of all COQ genes (COQ1-COQ11), each of which codes for a protein involved in synthesis of the essential lipid Coenzyme

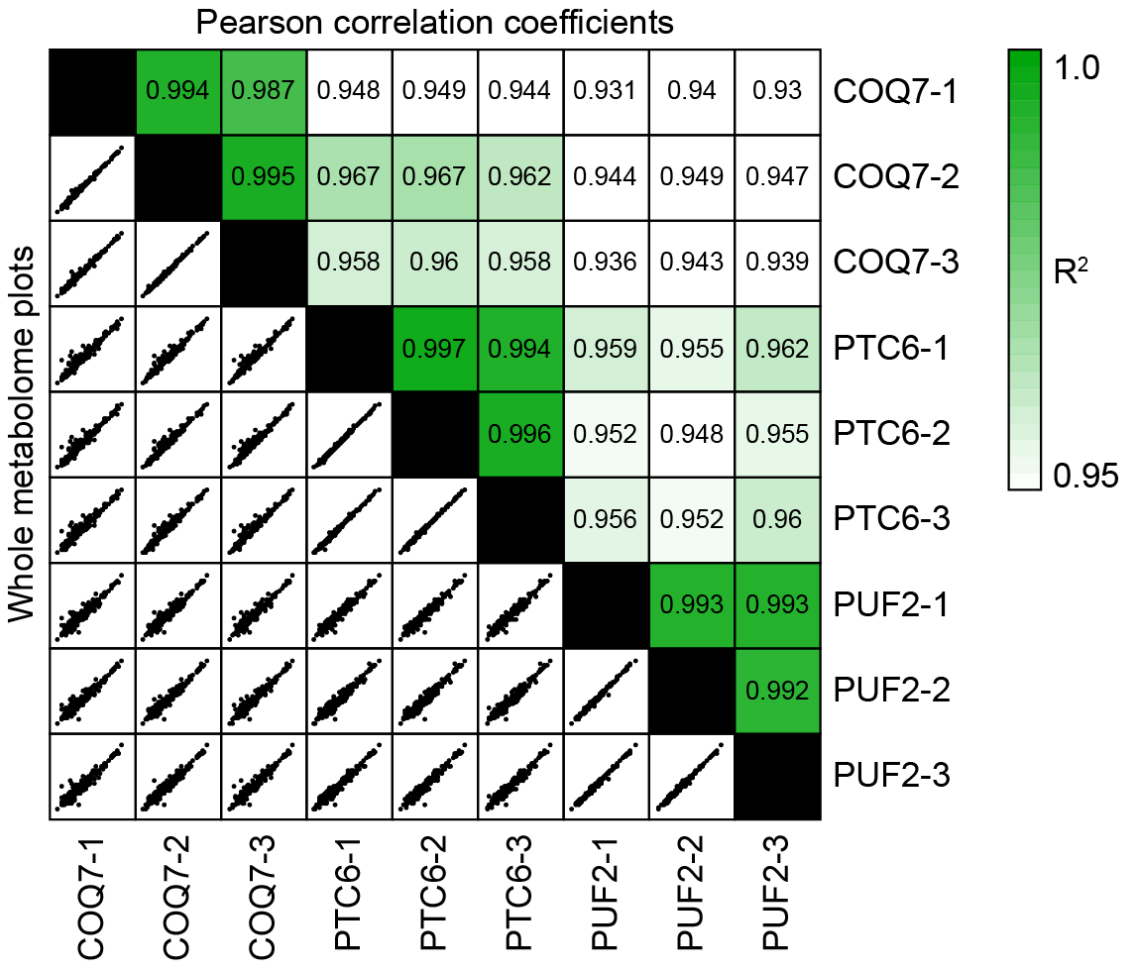


Figure 3.12: Replicate metabolite profile correlations. Replicate vs. replicate metabolite profile correlations. Here, $\log_2$-transformed metabolite profile comparisons are displayed from three replicates of COQ7, PTC6, and PUF2 each below the diagonal. Pearson correlation coefficients (R^2) are displayed for each comparison above the diagonal.

Q (CoQ). Elimination of proteins involved in conserved biochemical processes induces similar cellular responses which can be observed across omes. We note that within the complimentary protein and lipid data, similar profiles were observed between functionally related knockouts. The metabolomic profiles extracted using our pipeline reflect these correlations, and we report strong correlative agreement between all three omes.

The aforementioned global analyses demonstrate the exemplary performance of our developed software tools. In a narrower scope, we also found that these metabolomic data were exceptionally useful for hypothesis generation and testing. In the Y3K study, we elucidated a novel function of the incompletely characterized protein Hfd1p. It was hypothesized, and subsequently confirmed, that this protein facilitates the conversion of the metabolite 4-hydroxybenzaldehyde (4-HBz) to 4-hydroxybenzoic acid (4-HB). Speculation of this putative function arose as 4-HB was observed to be uniquely—and significantly (p<0.05; Student's t-test)—downregulated in the Δ*hfd1* knockout strain. This characteristic genotype–phenotype relationship was specific enough to merit follow-up testing which led to a new assignment of protein function. Outside of Y3K, this pipeline has been used for other metabolomic profiling efforts, notably recent work from Stefely et al[12]. Here, Stefely and colleagues utilized these processing tools to highlight that

central carbon metabolism metabolite levels in Coq8a$^{-/-}$ were not significantly altered from WT. While this result is admittedly less climactic compared to the Y3K results described here, these data were nonetheless informative to the larger biochemical conclusions drawn in that work. Also, we draw attention to the fact that our developed tools remained highly performant when presented with data from mouse, an organism which is markedly more complex than yeast.

Future Directions.

The software suite described here capitalizes on the unparalleled mass accuracy and sensitivity afforded by the recently commercialized GC-Orbitrap platform, and enables metabolomic profiling on a grand scale. We have developed user-friendly tools for converting raw MS data into meaningful quantitative values, and even provide a convenient interface through which to explore and compare these measurements. All of these tools have been designed to run on PCs and are modularized such that users can interact with, and edit data at separate stages in processing. Collectively, these tools provide a comprehensive data analysis solution for extracting novel biochemical insight from GC-Orbitrap metabolomics data.

While this pipeline has been of great value to our lab's research efforts, we recognize opportunities for improvement. We constructed this pipeline with the

intention that users would analyze data which adheres to a set experimental design (i.e., a control condition is always present). In many cases users will wish to analyze data sets lacking an associated control, which would preclude usage of our developed suite. Addition of functionality to allow comparison of conserved metabolite features across MS data files, without need for a control, is welcome. Implementing this functionality requires developing new algorithms for improved feature selection across a large set of files. Manually comparing single files against associated blanks to identify targets for quantitation—as is done in our current pipeline—becomes untenable when data sets grow large. Thus, it is imperative that the underlying feature selection routines utilize all provided data to automatically construct these target lists—while discriminating against noise features—without requiring user input of any kind.

Although GC/MS profiling affords many advantages with regards to reproducible chromatography and fragmentation, LC/MS remains the preferred technology for metabolomic analysis. To extend the utility of our developed pipeline, it is desirable that we support LC/MS analyses as well. LC/MS data acquisition is fundamentally different from GC/MS in that intact precursors are monitored (MS^1) prior to selection for MS^2 analysis. However, the same untargeted quantitative profiling techniques can be applied. Here, we would seek to extract, quantify, and

compare abundance profiles from intact species. We would likely employ similar feature grouping procedures to account for adducts and characteristic loss species. Assignment of molecular identifies would be performed in much the same way, however all compound-informative fragmentation information would be derived from MS^2 spectra.

Finally, we recognize that metabolomic studies are often conducted on a variety of instruments from numerous vendors. In order to make our tools more broadly useful, it is imperative that we support data from multiple vendors. The most obvious way to provide this kind of support is to further develop our tools to accept universal MS data formats such as mzML[13] and mzXML[14]. These universal formats have been developed as community standards by a consortium of MS users. Furthermore, there are a number of freeware tools which support conversion of data acquired on instruments from nearly all MS vendors[15–17]. By enabling our tools to accept mzML or mzXML inputs we can provide functional software solutions to a much larger audience of metabolomic researchers.

References

[1] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear

peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, pp. 779–787, 2006.

[2] M. F. Clasquin, E. Melamud, and J. D. Rabinowitz, "LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine," *Current Protocols in Bioinformatics*, 2012.

[3] A. C. Peterson, G. C. McAlister, S. T. Quarmby, J. Griep-Raming, and J. J. Coon, "Development and characterization of a GC-enabled QLT-orbitrap for high-resolution and high-mass accuracy GC/MS," *Analytical Chemistry*, vol. 82, pp. 8618–8628, 2010.

[4] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, 2014.

[5] A. Peterson and A. Balloon, "Development of a GC/Quadrupole-Orbitrap mass spectrometer, part II: new approaches for discovery metabolomics," *Analytical …*, vol. 86, pp. 10044–51, Oct. 2014.

[6] N. W. Kwiecien, D. J. Bailey, M. J. P. 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, pp. 8328–35, Aug. 2015.

[7] O. Fiehn, J. Kopka, R. N. Trethewey, and L. Willmitzer, "Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry.," *Analytical chemistry*, vol. 72, pp. 3573–3580, 2000.

[8] a. Dempster, "A new Method of Positive Ray Analysis," 1918.

[9] W. Bleakney, "A new method of positive ray analysis and its application to the measurement of ionization potentials in mercury vapor," *Physical Review*, vol. 34, pp. 157–160, 1929.

[10] D. R. Knapp, "Handbook of Analytical Derivatization Reactions," *John Wiley Sons New York*, p. 741, 1979.

[11] 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*, pp. 1–11, Sept. 2016.

[12] J. a. Stefely, F. Licitra, L. Laredj, A. G. Reidenbach, Z. a. Kemmerer, A. Grangeray, T. Jaeg-Ehret, C. E. Minogue, A. Ulbrich, P. D. Hutchins, E. M. Wilkerson, Z. Ruan, D. Aydin, A. S. Hebert, X. Guo, E. C. Freiberger, L. Reutenauer, A. Jochem, M. Chergova, I. E. Johnson, D. C. Lohman, M. J. P. Rush, N. W. Kwiecien, P. K. Singh, A. I. Schlagowski, B. J. Floyd, U. Forsman, P. J. Sindelar, M. S. Westphall, F. Pierrel, J. Zoll, M. Dal Peraro, N. Kannan, C. a. Bingman, J. J. Coon, P. Isope, H. Puccio, and D. J. Pagliarini, "Cerebellar Ataxia and Coenzyme Q Deficiency through Loss of Unorthodox Kinase Activity.," *Molecular cell*, vol. 63, pp. 608–20, Aug. 2016.

[13] E. Deutsch, "mzML: A single, unifying data format for mass spectrometer output," 2008.

[14] P. G. A. Pedrioli, J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu, and R. Aebersold, "A common open representation of mass spectrometry data and

its application to proteomics research," *Nature Biotechnology*, vol. 22, pp. 1459–1466, 2004.

[15] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, "ProteoWizard: Open source software for rapid proteomics tools development," *Bioinformatics*, vol. 24, pp. 2534–2536, 2008.

[16] 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. Aebersold, "A guided tour of the Trans-Proteomic Pipeline," 2010.

[17] 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, pp. 918–920, 2012.

Chapter 4

MITOCHONDRIAL PROTEIN FUNCTIONS ELUCIDATED BY MULTI-OMIC MASS SPECTROMETRY PROFILING

This chapter has been published:

Stefely JA*, **Kwiecien NW***, Freiberger EC, Richards AL, Jochem A, Rush MJP, Ulbrich A, Robinson KP, Hutchins PD, Veling MT, Guo X, Kemmerer ZA, Connors KJ, Trujillo EA, Sokol J, Marx H, Westphall MS, Hebert AS, Pagliarini DJ, Coon JJ. *Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling.* Nature Biotechnology. **2016**, doi:10.1038/nbt.3683

* Authors contributed equally

Abstract

Mitochondrial dysfunction is associated with many human diseases, including cancer and neurodegeneration, that are often linked to proteins and pathways that are not well-characterized. To begin defining the functions of such poorly characterized proteins, we used mass spectrometry to map the proteomes, lipidomes and metabolomes of 174 yeast strains, each lacking a single gene related to mitochondrial biology. 144 of these genes have human homologs, 60 of which are associated with disease and 39 of which are uncharacterized. We present a multi-omic data analysis and visualization tool that we use to find covariance networks that can predict molecular functions, correlations between profiles of related gene deletions, gene-specific perturbations that reflect protein functions, and a global respiration deficiency response. Using this multi-omic approach, we link seven proteins including Hfd1p and its human homolog ALDH3A1 to mitochondrial coenzyme Q (CoQ) biosynthesis, an essential pathway disrupted in many human diseases. This Resource should provide broad molecular insights into mitochondrial protein functions.

Introduction

High resolution mass spectrometry (MS) has become the primary analysis tool for many classes of biomolecules, including proteins, metabolites, and lipids. Major advancements in MS technology—particularly in the rate and depth of analysis—have enabled dozens of proteomes, metabolomes, and lipidomes to be analyzed in a single day[1–3]. Studies of bacteria demonstrated that parallel measurement of multiple molecule classes can synergistically enhance the biological insight afforded[4,5]. Recently, proteomics has been integrated with transcriptomics and genomics in mice[6,7]. However, large-scale, comprehensive (i.e., proteome-wide), multi-omic data acquisition, integration, and visualization tools remain underdeveloped, often lagging behind genomics in terms of coverage, speed, and broad accessibility for end users. Given the interdependence of proteins, lipids, and metabolites, we reasoned that coordinated analysis across all three biomolecule classes could afford new insight into eukaryotic biology. In particular, we hypothesized that this multi-omic profiling strategy, when coupled with genetic and environmental perturbations, could enable functional predictions for uncharacterized proteins.

We applied this strategy to study mitochondria, dynamic organelles whose dysfunction is associated with over 150 human diseases including cancer, diabetes, Parkinson's, and numerous genetic disorders[8–10]. While the yeast and mammalian

mitochondrial proteomes were recently defined[11–13], functional annotation of these proteins lags behind[14], impeding biomedical research on the many diseases impacted by mitochondrial metabolism. Of the ~1,200 mammalian mitochondrial proteins, nearly 300 are "mitochondrial uncharacterized (x) proteins" (MXPs)[15,16] that have no well-established biochemical function within mitochondria. Here, toward defining functions for MXPs, we performed over 3,000 MS experiments in parallel to analyze the proteomes, metabolomes, and lipidomes of 174 single-gene deletion ("Δ*gene*") Saccharomyces cerevisiae yeast strains in biological triplicate across two metabolic conditions, fermentation and respiration **(Fig. 4.1a)**. To facilitate development of biological hypotheses based on the resultant "yeast-three-thousand (Y3K)" data set **(Fig. 4.1b)**, we also developed a multi-omic data visualization approach (highlighted in **Fig. 4.1c** and online at http://y3kproject.org/). Our data establish many new connections between MXPs and proteins with well-established functions by virtue of gene-specific phenotypes or shared global biomolecular changes that result from the loss of each protein's expression. We leveraged a subset of these connections to address the incomplete mitochondrial pathway that generates ubiquinone (coenzyme Q, CoQ), an essential lipid required for oxidative phosphorylation (OxPhos) and linked to diseases ranging from severe infantile multisystemic disease to isolated myopathy and aging[17,18].

Results

Multi-omic mass spectrometry profiling. The 174 Δ*gene* yeast strains we analyzed covered 124 characterized genes that were selected to span a broad range of pathways to assist functional mapping, and 50 uncharacterized genes that encode MXPs **(Fig. 4.1a and Supplementary Fig. S4.1a)**. In selecting these targets, we prioritized genes with human homologs (144/174 genes) and those associated with disease (60/144 genes) based on primary literature analysis and online database gene annotation (e.g., omim.org). Inclusion of characterized genes, some of which could be considered as only partially characterized, also provided the ability to connect them to previously unrecognized functions. Each strain was grown in biological triplicate under two contrasting growth conditions, a standard fermentation culture condition and a carefully optimized respiration culture condition that stimulates mitochondrial function **(Fig. 4.1a, Supplementary Fig. S4.1b–e, and Supplementary Note 1)**—yielding six separate cultures per yeast strain.

Altogether we grew more than 1,050 yeast cultures (including WT cultures), each of which was analyzed using three separate high-resolution MS-based proteomic, metabolomic, and lipidomic techniques. These 3,000+ MS experiments yielded quantitation of 4,040 proteins, 411 metabolites, and 53 lipids (averaging 3,180 proteins, 252 metabolites, and 53 lipids per culture)—over 3.5 million biomolecule

measurements in total **(Fig. 4.1a and Supplementary Fig. S4.2a,b)**. Key to our approach was streamlining procedures for proteome extraction and preparation to under two hours of hands-on time **(Supplementary Fig. S4.2c)**. Use of label-free quantitation negated the need for a chemical tagging step and further increased throughput. We observed a wide dynamic range across all profiled omes, with some molecule abundances spanning more than three orders of magnitude **(Supplementary Fig. S4.2d)**. Additionally, we observed remarkable reproducibility between replicate cultures, with a median coefficient of variation of 12.7% considering all profiled biomolecules, and high overlap of molecules quantified across cultures **(Supplementary Fig. S4.2e–g)**.

A high-level view of the Y3K data set shows significant perturbations across all three omes, with more pronounced perturbations in respiration **(Fig. 4.1b and Supplementary Fig. S4.3a)**. Hierarchical clustering revealed groups of functionally related molecules (along the y-axis) and groups of functionally related Δ*gene* strains (along the x-axis). Protein clusters show significant gene ontology (GO) term enrichments for diverse processes and include both characterized and uncharacterized proteins **(Supplementary Fig. S4.3b)**. For example, the uncharacterized proteins Esbp6p and Ypr010c-a cluster with proteins involved in mitochondrial ATP synthesis and electron transport chain function, respectively **(Supplementary**

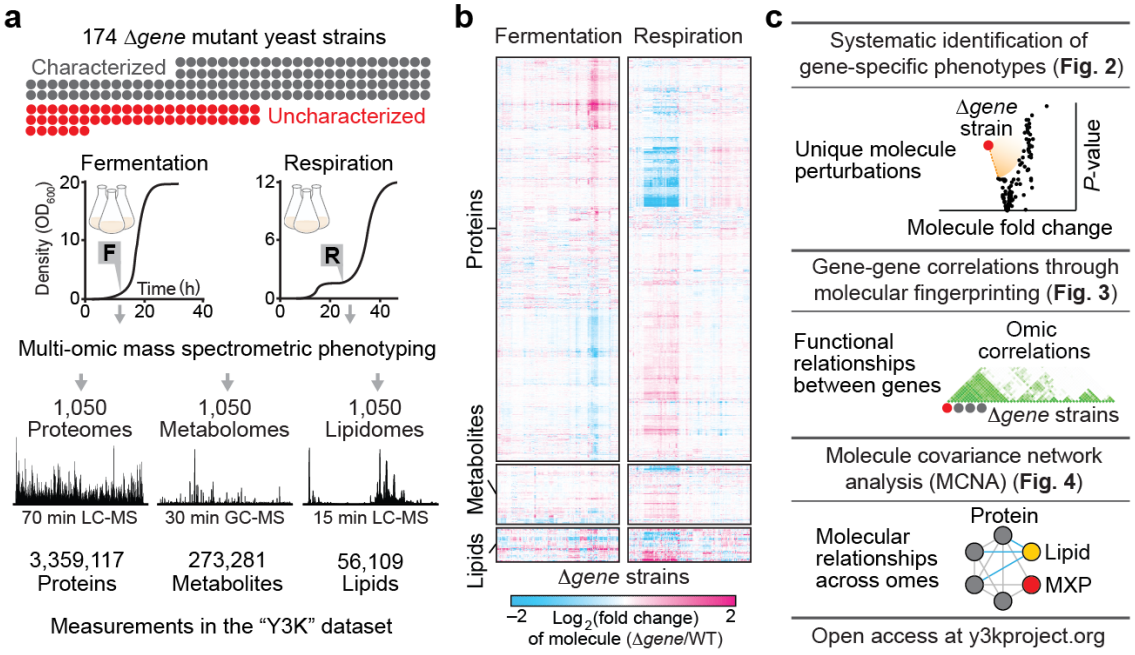
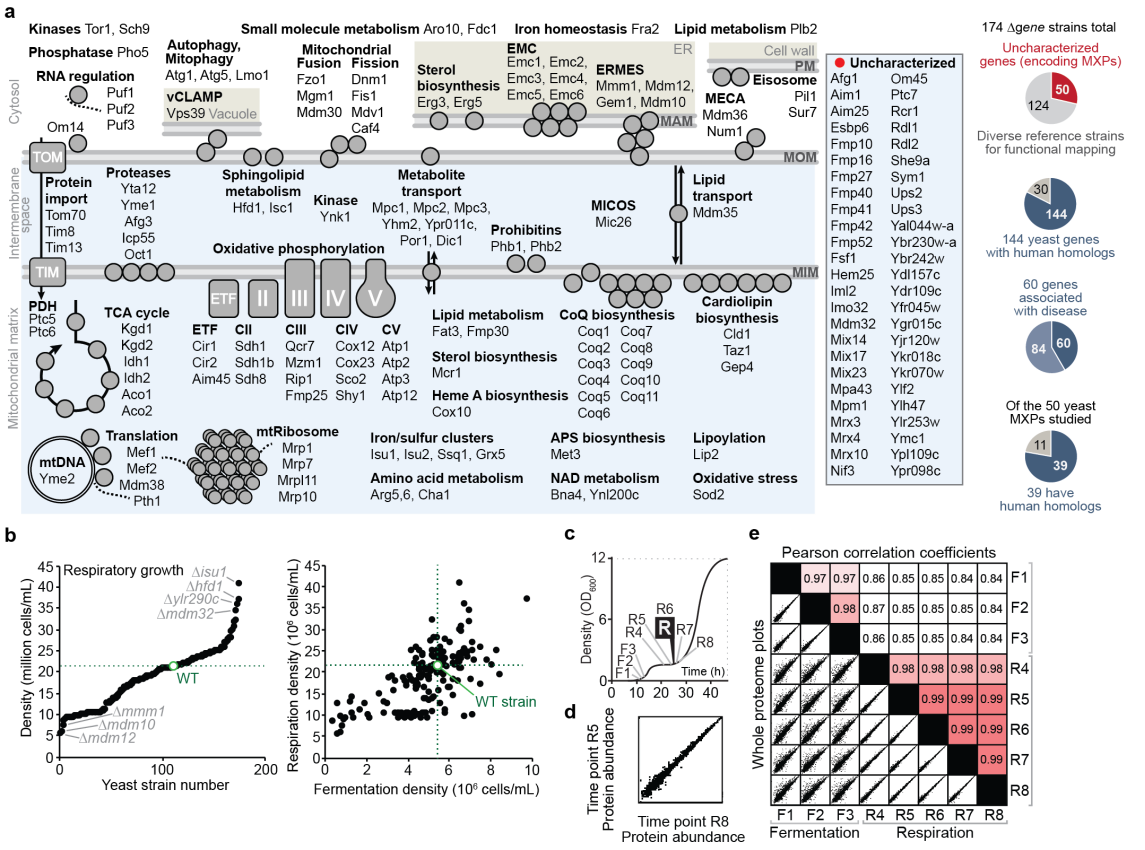


Figure 4.1: Multi-omic mass spectrometry profiling and data visualization. Multi-omic mass spectrometry profiling and data visualization. Overviews of **(a)** the experimental design and high resolution quantitative MS analysis, **(b)** the Y3K data set, shown as hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT as quantified by MS, mean, n = 3; $P < 0.05$, two-sided Student's t-test), and **(c)** the multi-omic data analysis and visualization tools developed here.

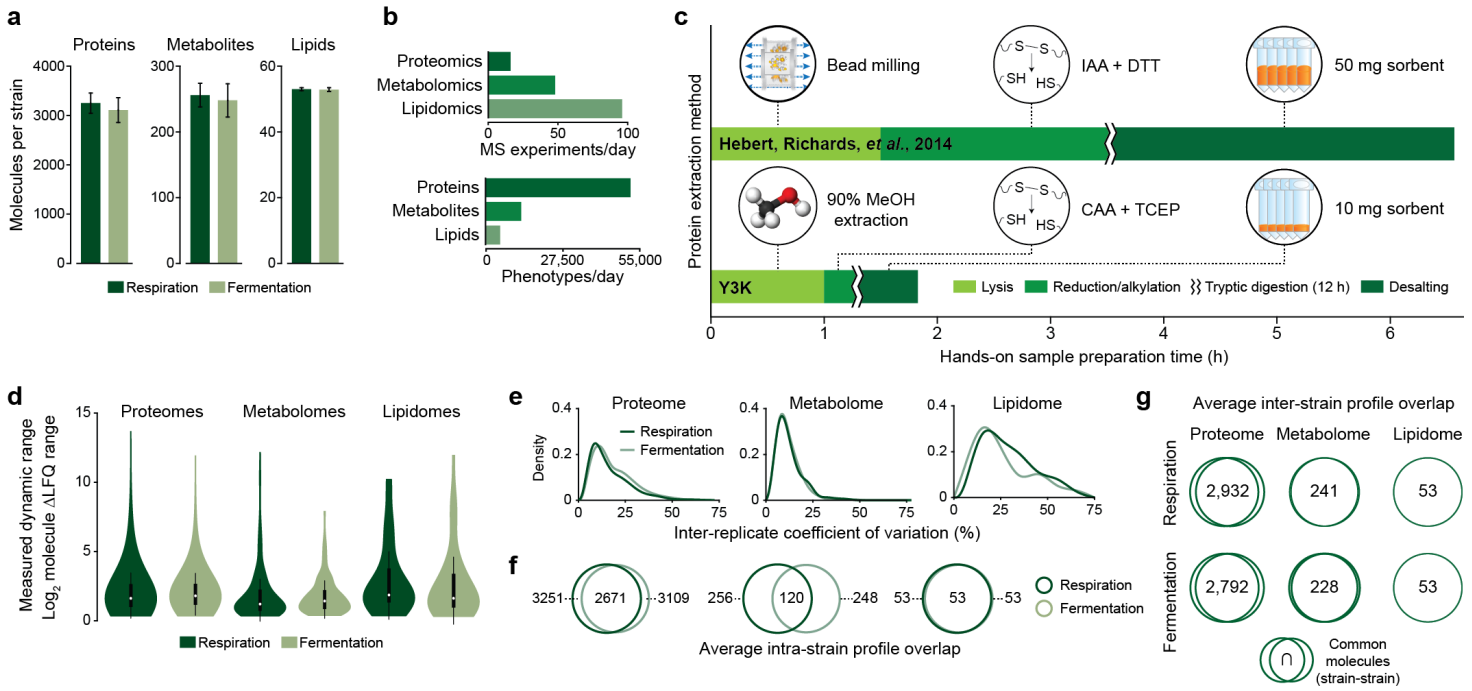


Supplementary Figure S4.1: Δ*Gene* target strain characteristics and respiration culture optimization. (a) Proteins encoded by the individual genes knocked out of the 174 yeast strains investigated in this study, shown in the context of biological pathways. APS, adenosine-5′-phosphosulfate; CII–CV, oxidative phosphorylation complexes II–V; ER, endoplasmic reticulum; EMC, ER membrane complex; ERMES, ER-mitochondria encounter structure; ETF, electron transfer flavoprotein complex; MAM, mitochondria-associated membrane; MECA, mitochondria-ER-cortex anchor; MICOS, mitochondrial contact site and cristae organizing system; MIM, mitochondrial inner membrane; MOM, mitochondrial outer membrane; mtDNA, mitochondrial DNA; mtRibosome, mitochondrial ribosome; NAD, nicotinamide adenine dinucleotide; PDH, pyruvate dehydrogenase; TCA, tricarboxylic acid cycle; vCLAMP, vacuole and mitochondria patch. The pie charts show the total number of characterized and uncharacterized genes profiled (top); the total number of profiled genes that have human homologs (upper middle); of these genes with human homologs, the number of profiled genes that are also associated with disease (lower middle); and of the uncharacterized genes profiled, the number of genes that have human homologs (bottom).

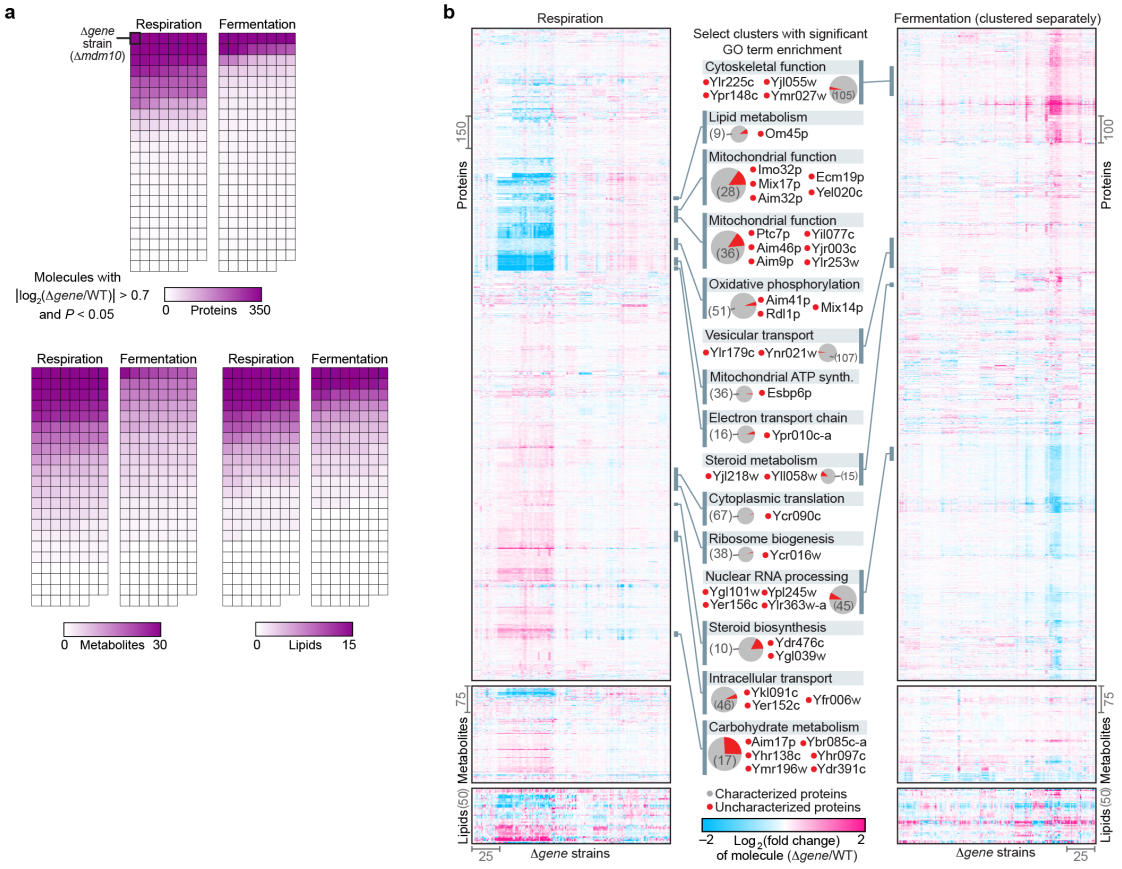
Supplementary Figure S4.1: (b) Density of yeast cultures in the respiratory growth condition (mean, n = 3) plotted in strain rank order (left) or against fermentation culture density (mean, n = 3) (right). **(c)** Optical density at 600 nm (OD_{600}) of yeast cultures (media with 3% [w/v] glycerol and 0.1% [w/v] glucose) indicating time points at which yeast were harvested during fermentation (F1–F3) or respiration (R4–R8). Time point R6 (25 h) was selected for the respiration culture condition of the larger study. **(d)** Whole-proteome plot of protein abundances at time points R5 and R8. **(e)** Pairwise whole proteome plot comparisons (as in d) across all eight time points (lower left) and linear regression analysis of each comparison (r^2, Pearson correlation coefficients) (upper right).

Fig. S4.4). Here, we leverage analyses from three different vantage points, each of which can be recapitulated with our online data visualization suite, exploiting unique biological perspectives afforded by a multi-omic data set of diverse genetic perturbations **(Fig. 4.1c)**.

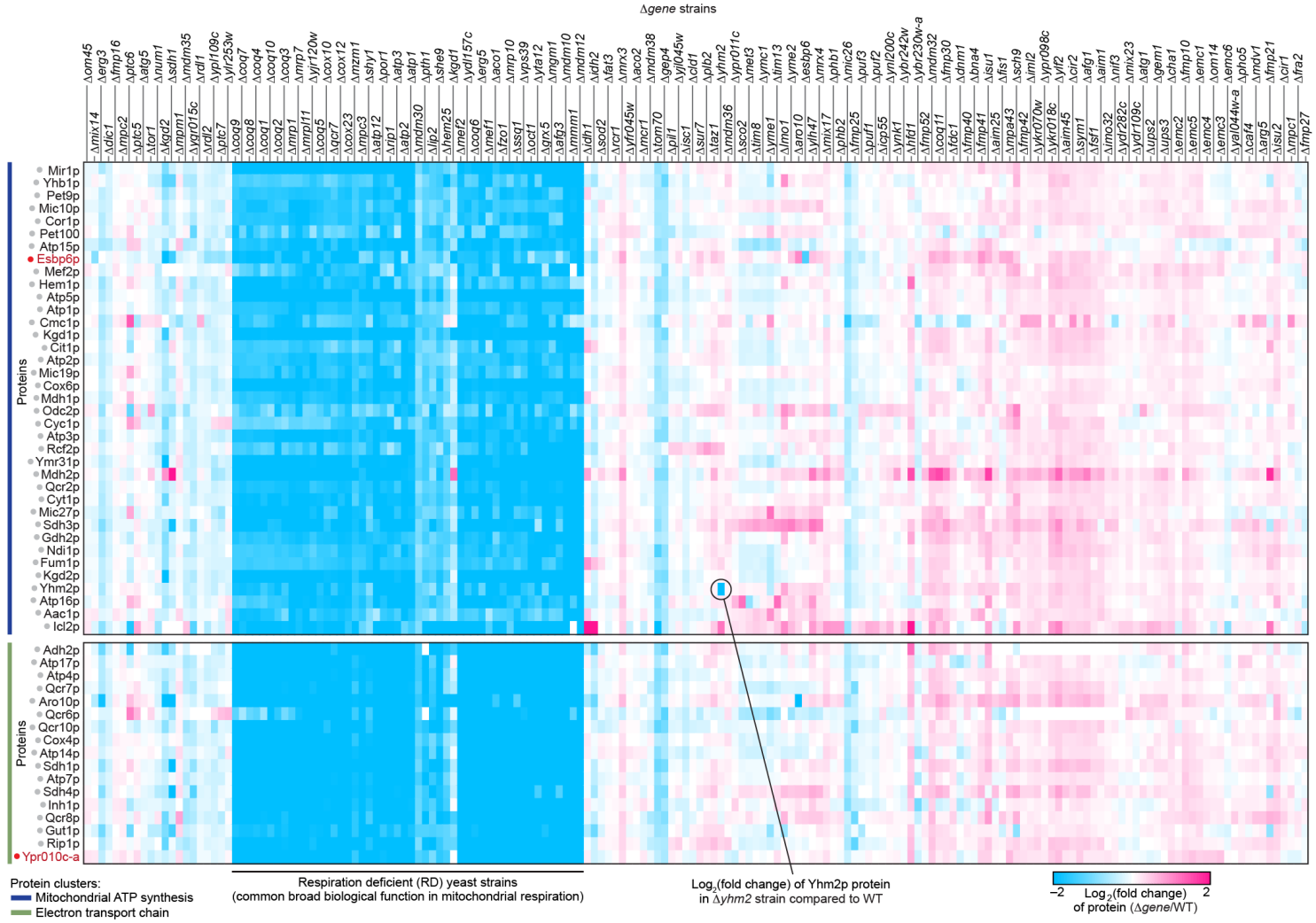
Identification of gene-specific phenotypes. First, we systematically surveyed the Y3K data set for significant molecule perturbations unique to just one or two of the strains in the study **(Fig. S4.2a)**. This unbiased search revealed 714 Δ*gene*-specific phenotypes **(Fig. 4.2a and Supplementary Note 2)**, which can reveal functional relationships. For example, the electron transfer flavoprotein (ETF) subunit Aim45p was uniquely decreased in just two Δ*gene* strains: the Δ*aim45* strain, and the Δ*cir1* strain, which lacks the second ETF heterodimer subunit **(Fig. 4.2b)**. Numerous additional Δ*gene*-specific phenotypes were used to generate biological hypotheses **(Supplementary Figs. S4.5 and S4.6)**. We decided to investigate one of these



Supplementary Figure S4.2: Mass spectrometry analysis metrics and quality assessment. **(a)** Proteins, lipids, and metabolites quantified per Δ*gene* strain (mean ± s.d., n = 3). **(b)** MS experiments conducted per day (top) and phenotypes (molecules) quantified per day (bottom) for proteomics, lipidomics, and metabolomics. **(c)** Overview of the yeast protein extraction method optimized for this study compared to previous work. **(d)** Violin plots depicting the range of fold changes in molecule abundance ($\log_2$[Δ*gene*/WT]) across all molecule classes and metabolic states. **(e)** Density plots of the distribution of coefficients of variation (CVs) (%) for each molecule measured in biological triplicate across all mutants and growth conditions. **(f)** Venn diagrams depicting the average overlap of molecules quantified within individual Δ*gene* strains across fermentation and respiration growth conditions. **(g)** Average profile overlap between different Δ*gene* strains.



Supplementary Figure S4.3: Features of protein-lipid-metabolite perturbation profiles. (a) Heat maps depicting the number of molecules significantly perturbed within each Δ*gene* strain ($P < 0.05$; two-sided Student's t-test). **(b)** Hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT quantified by MS; $P < 0.05$; two-sided Student's t-test). The center column annotates select clusters with significant functional (GO term) enrichments ($P < 0.05$; Fisher's exact test followed by Benjamini-Hochberg FDR correction for multiple hypothesis testing). Pie charts indicate proteins in clusters encoded by characterized (gray) or uncharacterized (red) genes.



Supplementary Figure S4.4: Expanded view of two protein clusters from the respiration Y3K data set heat map (respiration profiles). Heat map indicates relative abundance of proteins in Δ*gene* strains compared to WT as quantified by MS. See **Supplementary Figure S4.3** for the full heat map.

observations at biochemical depth: a Δ*hfd1*-specific decrease in 4-hydroxybenzoate (4-HB), the CoQ headgroup precursor **(Fig. 4.2c)**.

Though it has been known for decades that mammals can convert tyrosine (Tyr) into 4-HB for CoQ biosynthesis[19,20], the biochemical pathway has remained undefined in mammals and yeast **(Fig. 4.2c)**. The Y3K data set reveals Δ*hfd1* yeast to be significantly deficient in both the metabolite 4-HB ($P < 0.001$) and the lipid CoQ intermediate 3-polyprenyl-4-hydroxybenzoate (PPHB) ($P < 10^{-5}$) **(Fig. 4.2c and Supplementary Fig. S4.7a)**. Despite the PPHB deficiency, Δ*hfd1* yeast have normal CoQ abundance **(Fig. 4.2c)**, likely because of increased flux through an alternative para-amino-benzoate (pABA)- dependent CoQ pathway[21,22], as suggested by elevation of the aminated analog of PPHB (PPAB) in Δ*hfd1* yeast **(Fig. 4.2c)**. This is in contrast to terminal CoQ biosynthesis genes (*coq3–coq9*), and some genes not previously linked to CoQ function (e.g. *oct1* and *fzo1*), whose deletion causes significant ($P < 0.05$) CoQ deficiency and accumulation of PPHB **(Fig. 4.2c)**. Because Hfd1p is predicted to be an aldehyde dehydrogenase[23], we hypothesized that it catalyzes dehydrogenation of 4-hydroxybenzaldehyde (4-HBz) to form 4-HB. Consistently, 4-HBz is elevated in Δ*hfd1* yeast **(Supplementary Fig. S4.7b)**.

We used chemical-genetics to test the proposed Hfd1p activity. Most culture

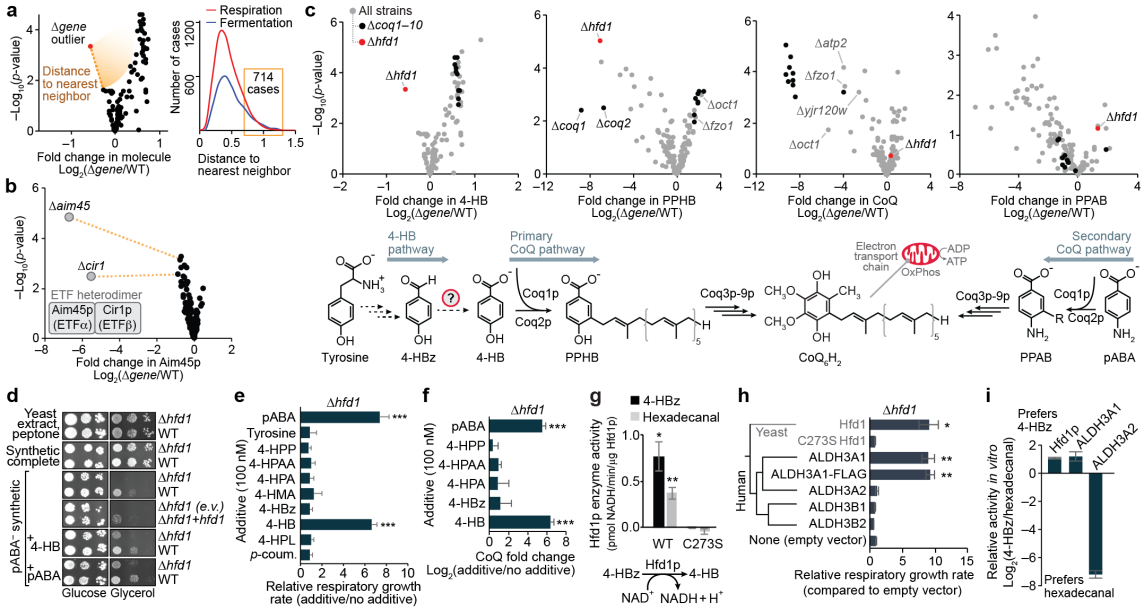
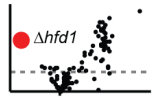
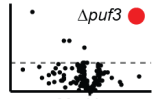
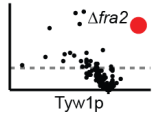
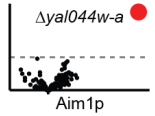
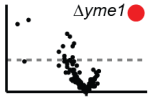
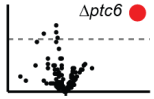
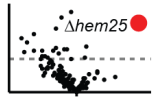
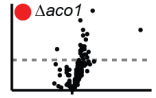


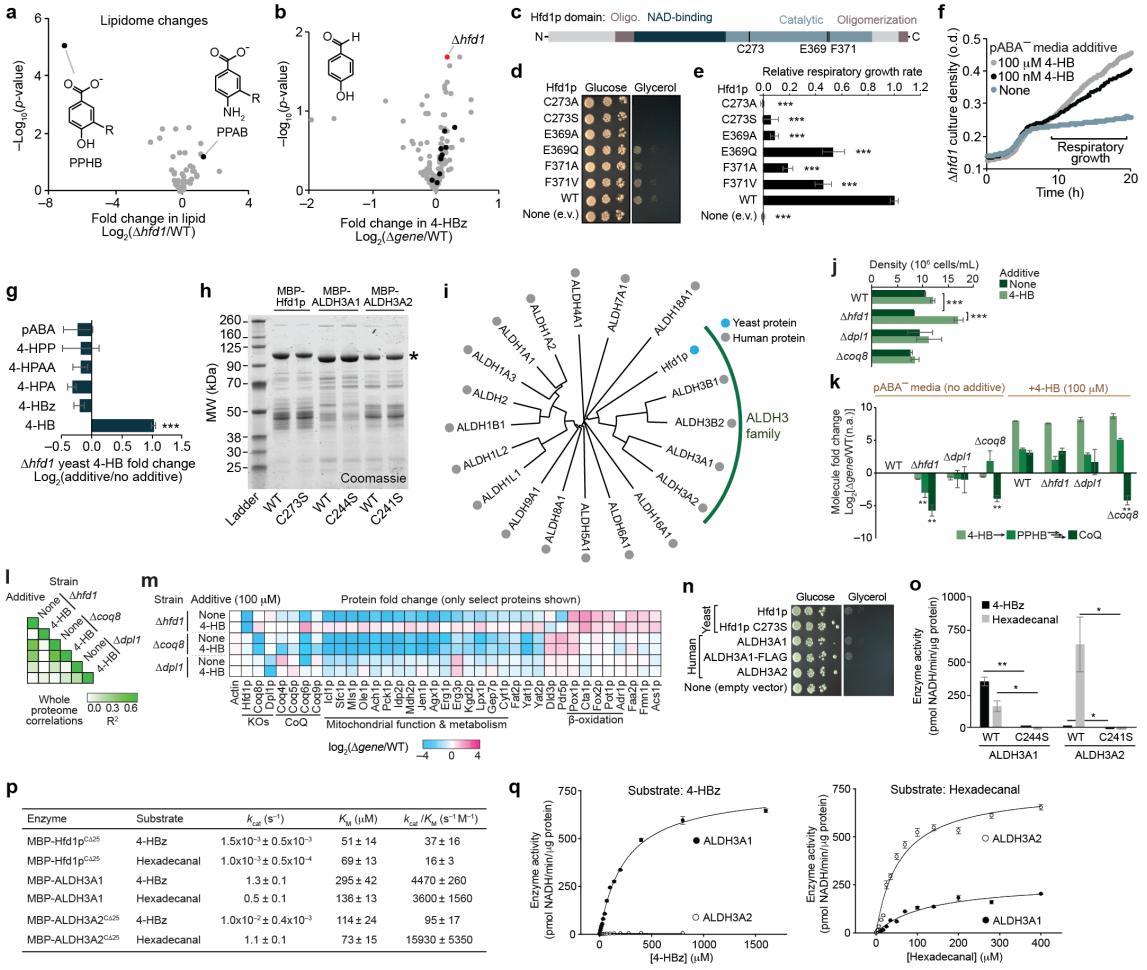
Figure 4.2: Δ*Gene*-specific phenotype detection links Hfd1p to production of 4-hydroxybenzoate for coenzyme Q biosynthesis. **(a)** Overview of the Δ*gene*-specific phenotype detection approach and number of Δ*gene*-specific phenotypes identified in the respiration and fermentation data sets (distance to nearest neighbor on a normalized scale, see Supplementary Note 2). **(b)** Relative abundance of Aim45p (mean, n = 3) versus statistical significance across strains. **(c)** Relative abundances of 4-HB, PPHB, CoQ, and PPAB (mean, n = 3) versus statistical significance across Δ*gene* strains. **(d)** Serial dilutions of yeast grown on variable solid medias. E.v., empty vector; +hfd1, hfd1 plasmid transformed. **(e)** Relative respiratory growth rates of Δ*hfd1* yeast in pABA⁻ synthetic media with the additives shown (mean ± s.d, n = 3). 4-HPP, 4-hydroxyphenylpyruvate; 4-HPAA, 4-hydroxyphenylacetaldehyde; 4-HPA, 4-hydroxyphenylacetate; 4-HMA, 4-hydroxymandelate; 4-HPL, 4-hydroxyphenyllactate; p-coum., para-coumarate. **(f)** Relative CoQ abundance in Δ*hfd1* yeast cultured in pABA⁻ media with the additives shown (mean ± s.d., n = 3). **(g)** Enzyme activity of recombinant MBP-Hfd1$^{C\Delta 25}$ in vitro against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(h)** Phylogenetic relationship between yeast Hfd1p and the human ALDH3 family, and relative respiratory growth rates of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured in pABA⁻ synthetic media (mean ± s.d, n = 4). **(i)** Relative activity of the dehydrogenases shown against 4-HBz compared to hexadecanal (mean ± s.e.m., n = 3). *P < 0.05; **P < 0.01; ***P < 0.001 (two-sided Student's t-test for all panels).

Supplementary Figure S4.5: Subsets of the Δ*gene*-specific phenotypes identified in this study. Relative abundances of individual molecules (mean $\log_2$[Δ*gene*/WT], n = 3) (x-axes) versus statistical significance ($-\log_{10}$[*p*-value]; two-sided Student's t-test) (y-axes) as quantified by MS. The plots shown represent a subset of molecules identified as 'Δ*gene*-specific phenotypes' through an unbiased survey of the Y3K data set (see Fig. 2a). The array here is limited to the most robust outliers (based on both statistical significance and fold-change, see Supplementary Note 2 and Methods)—the top 20 upregulated proteins, the top 20 downregulated proteins, the top 10 metabolites, and the top 4 or 5 lipids—excluding 'knocked out proteins' (e.g. Fmp52p in the Δ*fmp52* strain) and excluding a given Δ*gene* strain after it appeared twice on the rank list. Biological hypotheses surrounding gene-phenotype relationship were generated for the starred plots (see Supplementary Fig. 6).

Volcano plot	Biological hypothesis generated by Y3K gene-phenotype relationship
 4-hydroxybenzoic acid (4-HB)	Hypothesis: Hfd1p, a predicted dehydrogenase, catalyzes production of 4-HB. (This hypothesis is tested in detail in this manuscript, see Fig. 2)
 Mrpl4p	Hypothesis: The mRNA encoding Mrpl4p, a mitochondrial ribosome subunit, is directly bound by Puf3p, an mRNA-binding protein that enhances degradation of select mRNAs.
 Tyw1p	Hypothesis: Tyw1p, an iron-sulfur protein, is upregulated in response to iron overload due to loss of Fra2p, a negative regulator of the iron regulon.
 Aim1p	Hypothesis: Yal044w-a, a BolA-like protein, is upregulated to compensate for the loss of Aim1p, a homologous BolA-like protein.

Volcano plot	Biological hypothesis generated by Y3K gene-phenotype relationship
 Coa1p	Hypothesis: Coa1p, a mitochondrial inner membrane OxPhos Complex IV assembly factor, is a substrate for the mitochondrial inner membrane protease Yme1p (i-AAA protease).
 Ykl069w	Hypothesis: Ykl069w, a putative methionine-R-sulfoxide reductase, is upregulated in response to oxidative stress due to loss of Ptc6p, a mitochondrial type 2C protein phosphatase.
 Erg20p	Hypothesis: Erg20p, farnesyl pyrophosphate synthetase, is upregulated to compensate for a defect in Heme A (a farnesylated form of heme) biosynthesis due to loss of Hem25p.
 Glutamine	Hypothesis: Intracellular glutamine is decreased due to loss of Aco1p, aconitase, which enables glutamine biosynthesis and inhibition of which enhances glutaminolysis. Consistently, glutamine auxotrophy of the Δ*aco1* strain has already been reported.

Supplementary Figure S4.6: Examples of hypotheses that can be generated from a subset of the Δ*gene*-specific phenotypes identified in this study. Subset of Δ*gene*-specific phenotypes identified in the Y3K data set. Volcano plots indicate relative molecule abundances (mean $\log_2$[Δ*gene*/WT], n = 3) (x-axes) versus statistical significance ($-\log_{10}$[*p*-value]; two-sided Student's t-test) (y-axes) as quantified by MS. Hypotheses were developed to describe each Δ*gene*-phenotype relationship reported here.



Supplementary Figure S4.7: Hfd1p supports production of 4-HB for CoQ biosynthesis. (a) Relative lipid abundances (mean, n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS. **(b)** Relative abundances of 4-HBz (mean, n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) across all Δgene strains in the study. **(c)** Protein domain structures of Hfd1p, highlighting residues involved in catalysis. **(d)** Serial dilutions of*Δhfd1* yeast transformed with plasmids encoding the indicated Hfd1p variants grown on pABA⁻ synthetic solid medias with glucose or glycerol. **(e)** Relative respiratory growth rates of*Δhfd1* yeast transformed with plasmids encoding the indicated Hfd1p variants and grown in pABA⁻ synthetic liquid media. **(f)** Growth curves showing the respiratory growth of*Δhfd1* yeast in pABA⁻ synthetic media with the additives shown. **(g)** Relative 4-HB abundance in*Δhfd1* yeast cultured in pABA⁻ media with the additives shown (mean $\log_2$[additive/unsupplemented] ± s.d., n = 3). **(h)** SDS-PAGE analysis (Coomassie stained gel) of protein fractions from an isolation of MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (WT and catalytically dead mutant for each). **(i)** Phylogenetic tree of human ALDH superfamily members and yeast Hfd1p.

Supplementary Figure S4.7: (j) Density of yeast (upon harvest) cultured in pABA$^-$ media ± 4-HB (mean ± s.d., n = 3). **(k)** Relative abundances of 4-HB, PPHB, and CoQ compared to WT yeast cultured in pABA$^-$ media (mean $\log_2$[Δgene/WT with no additive] ± s.d., n = 3) as quantified by MS. **(l)** Whole proteome correlation map for yeast grown in pABA$^-$ media ± 4-HB (mean, n = 3). **(m)** Relative abundances of select proteins as quantified by MS (mean $\log_2$[Δgene/WT], n = 3) analysis of yeast cultured in pABA$^-$ media ± 4-HB. **(n)** Serial dilutions of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured on solid pABA$^-$ synthetic media plates. **(o)** Enzyme activity of MBP-ALDH3A1 or MBP-ALDH3A2(CΔ25) against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(p)** Table of enzyme kinetic parameters for MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (mean ± s.e.m., n = 3). **(q)** Representative enzyme kinetic curves for MBP-ALDH3A1 and MBP-ALDH3A2(CΔ25). *P < 0.05; **P < 0.01; ***P < 0.001 (two-sided Student's t-test).

media contain either 4-HB (in yeast extract) or pABA (in standard yeast nitrogen base), enabling yeast to bypass the Tyr-to-4-HB pathway, so we used a defined medium lacking pABA and 4-HB ("pABA$^-$"). Δ*hfd1* yeast exhibited striking respiration deficiency on pABA$^-$ media, a phenotype rescued by pABA, 4-HB, or WT Hfd1p, but not by Hfd1p with mutations to putative catalytic residues[24] **(Fig. 4.2d and Supplementary Fig. S4.7c–e)**. Testing a panel of potential intermediates in the pathway revealed that 4-HB, but not 4-HBz, can rescue the respiratory growth and CoQ production of Δ*hfd1* yeast **(Fig. 4.2e,f and Supplementary Fig. S4.7f,g)**, supporting a role for Hfd1p in dehydrogenation of 4-HBz. To directly test this activity, we purified recombinant Hfd1p for enzyme assays **(Supplementary Fig. S4.7h)**. WT Hfd1p catalyzes NAD^+-dependent dehydrogenation of 4-HBz, but a

C273S (catalytic residue) point mutant does not **(Fig. 4.2g)**. Together, these results demonstrate that Hfd1p dehydrogenates 4-HBz to produce 4-HB for CoQ biosynthesis.

Hfd1p is a member of the ancient aldehyde dehydrogenase (ALDH) superfamily, which is found across all three superkingdoms of life and includes 19 human homologs with diverse functions[25]. Based on phylogenetic analyses, Hfd1p is most similar to the human ALDH3 family **(Supplementary Fig. S4.7i)**. ALDH3A2 (FALDH) mutations cause Sjögren–Larsson Syndrome[26] due to defective fatty aldehyde metabolism. However, the endogenous functions of ALDH3A1, B1, and B2 remain obscure, and which of these human ALDH3 functions are conserved in Hfd1p has not been completely defined. Previous work showed that sphingolipid metabolism is perturbed in Δ*hfd1* yeast due to a defect in dehydrogenation of hexadecanal, and this defect can be rescued by ALDH3A2, but not by ALDH3A123[27]. However, a separate sphingolipid pathway defect (Δ*dpl1*) does not disrupt the 4-HB-CoQ pathway **(Supplementary Fig. S4.7j–m and Supplementary Note 3)**, suggesting that the two pathways are otherwise independent. Consistent with the idea that Hfd1p is a dual-function protein that supports both sphingolipid metabolism and CoQ biosynthesis, we observed Hfd1p activity in vitro with hexadecanal, similar to that observed with 4-HBz **(Fig. 4.2g)**. However, in contrast to

rescue of the sphingolipid metabolism defect, we found that ALDH3A1, but not ALDH3A2, rescues the pABA$^-$ respiratory growth phenotype of Δ*hfd1* yeast **(Fig. 4.2h and Supplementary Fig. S4.7n)**. Moreover, while ALDH3A2 shows a strong substrate preference for hexadecanal over 4-HBz, Hfd1p and ALDH3A1 show a preference for 4-HBz **(Fig. 4.2i and Supplementary Fig. S4.7o–q)**. These results suggest that the dual functions of yeast Hfd1p have diverged in human ALDH3A1 and ALDH3A2. Collectively, these results demonstrate a major cellular function for the aldehyde dehydrogenase Hfd1p in the Tyr-to-4-HB pathway and strongly suggest that ALDH3A1 plays a similar role in human CoQ biosynthesis.

Regression analysis of global perturbation profiles. While molecular changes unique to a given Δ*gene* strain can be functionally informative, similarities between Δ*gene* strains can also assist characterization. In our second analysis approach, we examined Δ*gene*–Δ*gene* correlations through pairwise comparisons of global Δ*gene* perturbation profiles. Deletion of functionally related genes, such as the cytochrome c oxidase genes *cox12* and *cox23*, caused highly similar whole proteome perturbations **(Fig. 4.3a)**. Notably, highly correlated phenotype changes were also observed in Δ*cox12* and Δ*cox23* metabolomes and lipidomes **(Fig. 4.3a)**. However, deletion of unrelated genes, such as *cox12* and *mic26*, generated uncorrelated phenotype changes **(Fig. 4.3a)**. Examination of Δ*gene*–Δ*gene* correlations

across the entire study indicated numerous functional relationships, with stronger correlations observed in respiration **(Fig. 4.3b)**.

A group of respiration-deficient (RD) strains showed robust correlations across all three omes **(Fig. 4.3b)**, reflecting their similar broad biological functions in mitochondrial OxPhos and suggesting that they share a universal "respiration deficiency response" (RDR). Multi-omic principle component and GO term analyses revealed a coordinated RDR that provides biological insight into respiration defects—a common feature of many diseases including cancer—and suggests that a multi-omic biomarker fingerprint could afford a specific diagnostic for mitochondrial disease **(Fig. 4.3c–f, Supplementary Fig. S4.8, and Supplementary Note 4)**. However, stress responses such as the RDR also pose a barrier to biochemical investigations because they can obscure functionally-informative phenotypes. To assess more specific biochemical roles for individual proteins, we normalized for the RDR across RD strains **(Supplementary Fig. S4.9 and Supplementary Note 5)**. Across all of our RD strains, 776 molecules were identified as being consistently perturbed. The individual measurements of these RDR-associated molecules were mean normalized ("RDR-adjusted") to reveal characteristic deviations from the general RDR and to enable visualization of Δ*gene*-specific changes.

Recalculating Δ*gene*–Δ*gene* correlation coefficients with RDR-adjusted plots

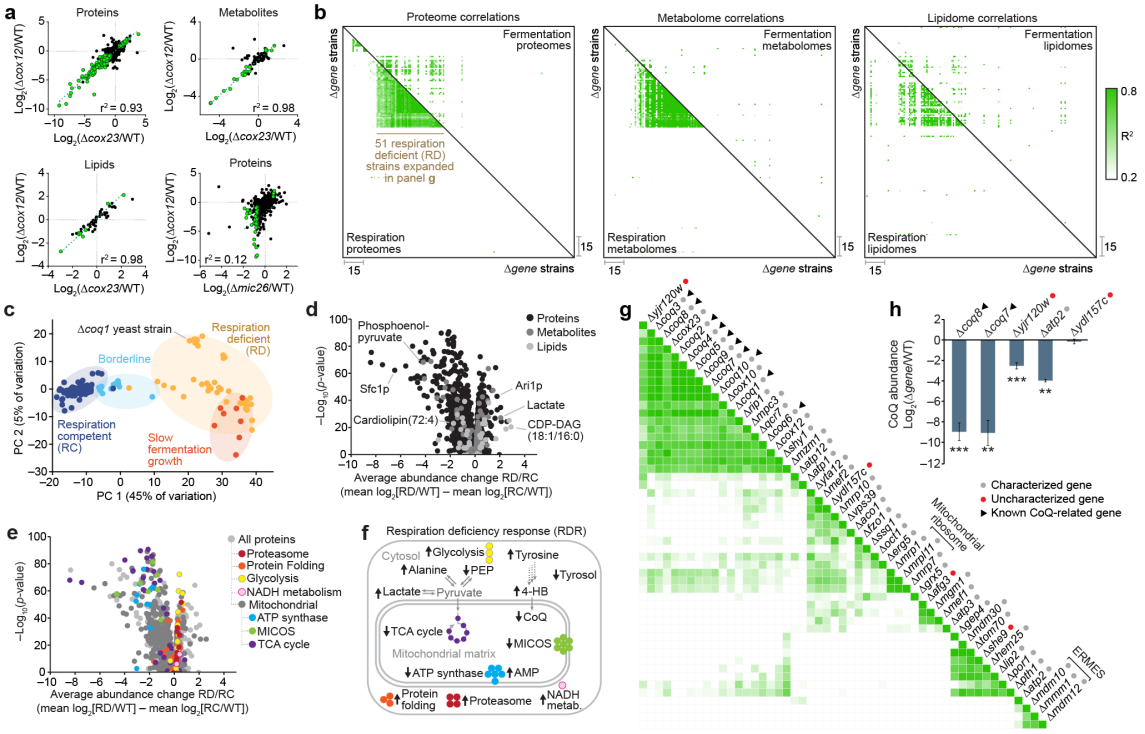
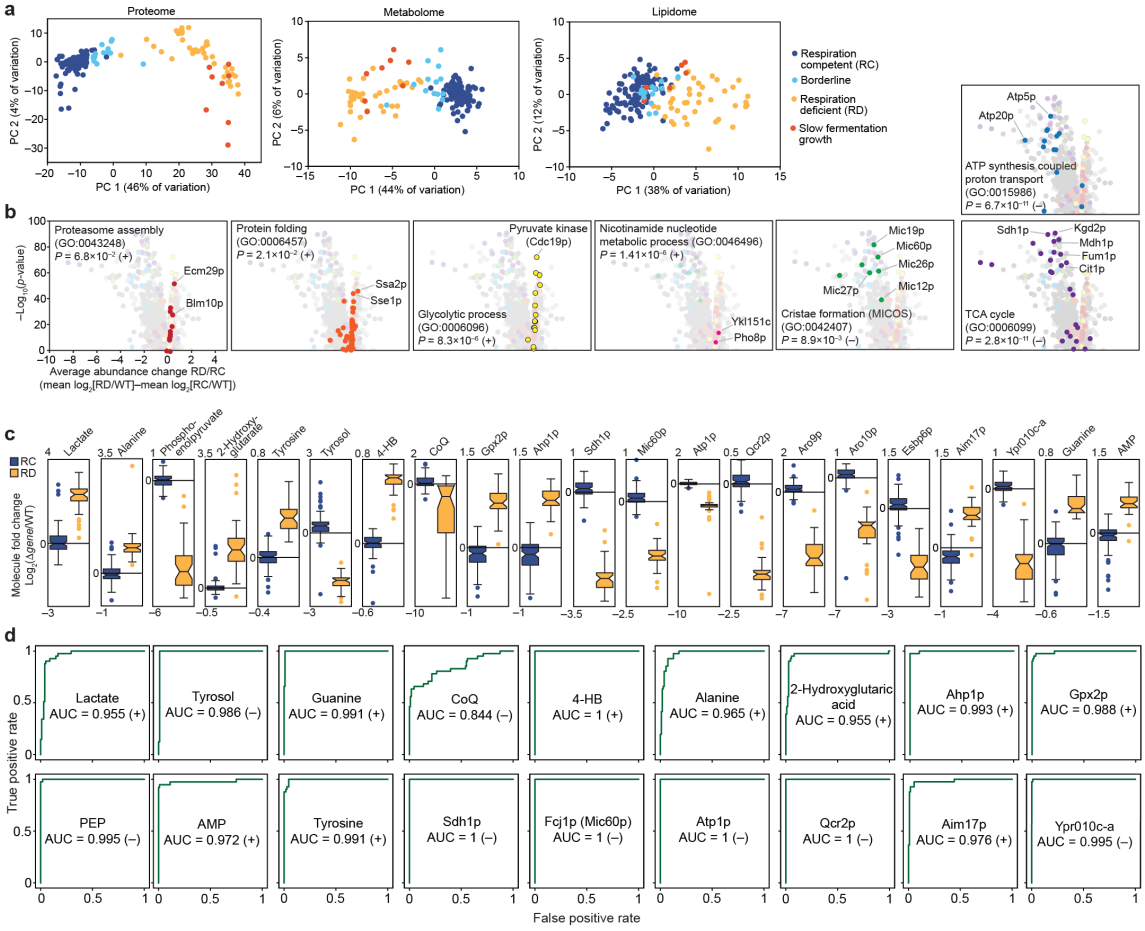
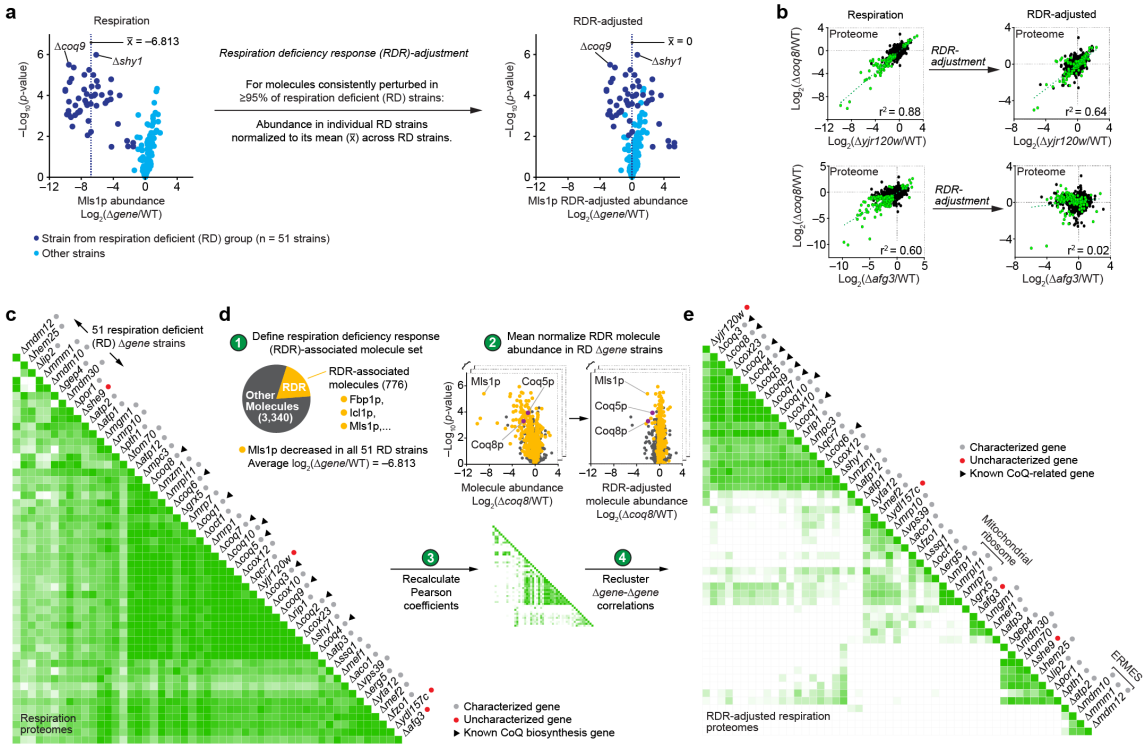


Figure 4.3: Functional correlations through perturbation profile regression analysis. **(a)** Plots comparing relative molecule abundances between pairs of Δ*gene* strains. Strain-strain similarity assessed by linear regression analysis of Δ*gene* perturbation profiles. Green points indicate molecules significantly perturbed in both mutants ($|\log_2[FC]| > 0.7$, $P < 0.05$; two-sided Student's t-test). **(b)** Maps of Pearson correlation coefficients (r^2) for pairs of Δ*gene* perturbation profiles across omes and metabolic conditions. Strains are clustered based on respiration proteome correlations, and this strain order is held consistent across all 6 maps. **(c)** Projection of respiration competent (RC) and deficient (RD) strains onto the plane defined by principal component (PC) axes 1 and 2 (full multi-omic respiration data set). **(d)** Average fold change in molecule abundances (mean $\log_2$[RD strains/RC strains]) versus statistical significance ($-\log_{10}$[*p*-value, Bonferroni corrected two-sided t-test]). **(e)** RD versus RC proteome perturbation volcano plot (as in d) showing select functional groups (GO terms) significantly enriched in either upregulated or downregulated proteins. **(f)** Scheme of RDR pathways. **(g)** Re-clustered respiration proteome strain-strain correlation map following RDR- adjustment. (h) CoQ abundance changes in select Δ*gene* strains (mean ± s.d., n = 3); $^{**}P < 0.01$; $^{***}P < 0.001$ (two-sided Student's t-test).



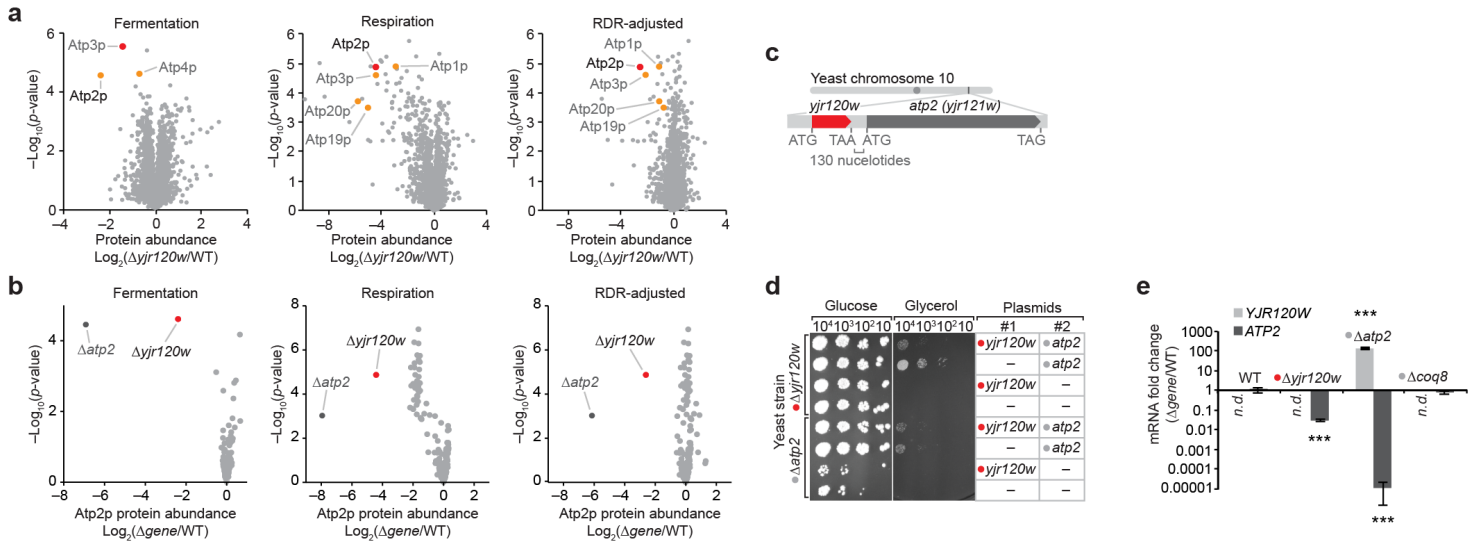
Supplementary Figure S4.8: Identification of respiration deficiency response pathways and potential biomarkers. **(a)** Projection of RC and RD strains onto the planes defined by principal component (PC) axes 1 and 2 for separate proteome, metabolome, and lipidome PC analyses. **(b)** RD versus RC proteome perturbation volcano plots (as in Fig. 3e) showing select functional groups (GO terms) significantly enriched (Bonferroni corrected *p*-values shown in figure) in either upregulated or downregulated proteins. **(c)** Box plots depicting median molecule fold changes for RC and RD strains ($\log_2$[RD or RC average/WT]) (n = 111 for RC, 41 for RD). Notch indicates 95% c.i. **(d)** Receiver operating characteristic (ROC) curves for select molecules depicting the false positive rates and true positive rates for prediction of respiration deficiency associated with particular molecule fold changes. AUC, area under the curve.



Supplementary Figure S4.9: Subtraction of shared responses to reveal deeper biochemical insight. **(a)** RDR-abundance adjustment of a representative molecule (Mls1p) by subtraction of the average fold change in abundance (mean $\log_2[\Delta gene/\text{WT}]$, n = 3) across respiration deficient (RD) strains. This adjustment was only performed within RD strains. **(b)** Plots comparing relative protein abundances between pairs of Δ*gene* strains. Linear regression analysis of pairs of perturbation profiles before (left) and after (right) RD-abundance adjustment. Green points indicate molecules significantly perturbed in both mutants ($|\log_2(\text{FC})| > 0.7$; $P < 0.05$; two-sided Student's t-test) prior to RDR-adjustment. **(c)** Expanded view of highly correlated strains in the respiration proteomes correlation map (see Fig. 3b). **(d)** Procedure for normalization of the RDR. **(e)** Re-clustered respiration proteome strain-strain correlation map following RDR-adjustment (also shown in Fig. 3g).

strikingly reduces correlations between more functionally disparate genes **(Supplementary Fig. S4.9c–e)**. Reclustering Δ*gene*–Δ*gene* correlations reveals new clusters of genes with similar biochemical functions **(Fig. 4.3g)**. For example, known CoQ biosynthesis genes were brought into a tighter cluster that also includes the uncharacterized gene *yjr120w* **(Fig. 4.3g)**, suggesting that *yjr120w* might support CoQ biosynthesis. Consistently, we observed CoQ deficiency in Δ*yjr120w* yeast **(Fig. 4.3h)**, the molecular basis of which we determined to include loss of Atp2p, an ATP synthase subunit **(Supplementary Fig. S4.10 and Supplementary Note 6)**. These results show that specific ATP synthase subunits support CoQ biosynthesis and, more broadly, demonstrate how global mass spectrometry profiling can reveal functional links between genes.

Molecule covariance network analysis. Similarly, in our third analysis approach, we leveraged the multi-omic nature of our mass spectrometry profiles to determine pairwise covariance between proteins, metabolites, and lipids. This approach is similar to mRNA coexpression profiling, which can be used to predict gene function[28–30], but it integrates three complementary classes of molecules. Perturbations for functionally related molecules, such as the protein Coq4p and the lipid CoQ intermediate PPHB, show strong positive or negative correlations, while those of unrelated molecules, such as Coq4p and Rpb4p, lack correlations **(Fig. 4.4a)**. Cor-



Supplementary Figure S4.10: Molecular perturbations of yeast lacking *yjr120w*. **(a)** Relative protein abundances (mean $\log_2$[Δ*yjr120w*/WT], n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS. (b) Relative Atp2p protein abundance (mean $\log_2$[Δ*gene*/WT], n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) across all mutants in the study. (c) Genomic organization of *yjr120w* and *atp2*. (d) Serial dilutions of yeast transformed with the indicated plasmids grown on agar plates with glucose (to enable fermentation) or glycerol (to force respiration). (e) Fold changes in mRNA abundances (mean Δ*gene*/WT, n = 3) as quantified by real time polymerase chain reaction (RT-PCR) analysis. Yjr120w mRNA was not detected (n.d.) in WT yeast, so imputation of this missing value was used to calculate the fold increase in yjr120w mRNA shown for the Δ*atp2* strain. **P* < 0.05; ***P* < 0.01; ****P* < 0.001 (two-sided Student's t-test).

related molecules include proteins in complexes, such as the cytosolic TRiC/CCT chaperonin complex (Cct2p and Cct7p), and enzyme-product pairs (e.g. Ura1p and orotic acid) **(Fig. 4.4a)**.

Examining correlations across all 4,505 molecules in the Y3K data set through this multi-omic molecule covariance network analysis (MCNA) reveals numerous functional relationships, which can be visualized as networks of molecules (nodes) and correlations (edges) **(Fig. 4.4b and Supplementary Fig. S4.11a)**. After applying strict correlation thresholds (Bonferroni-adjusted p-value < 0.001, $|\rho| \geqslant 0.58$), 237,342 edges remain among 2,382 nodes in the respiration data set **(Supplementary Fig. S4.11a–f)**. Many edges were observed between RDR-associated molecules **(Supplementary Fig. S4.11g)**, reflecting their common relationship to mitochondrial metabolism. As described above for Δ*gene* correlations, we deepened the molecular insight of the MCNA by RDR-adjustment, which reduced overall connectivity and increased the selectivity of functionally related molecule sub-networks **(Supplementary Fig. S4.11g)**. For example, the selectivity of the mitochondrial ribosome sub-network increased 16-fold **(Supplementary Fig. S4.11h)**. These RDR-adjusted networks associated the MXP Yor020w-a with the mitochondrial ribosome **(Supplementary Fig. S4.11g)**. To test this association, we examined the proteome of Δ*yor020w-a* yeast, which showed a significant decrease in the mitochondrial

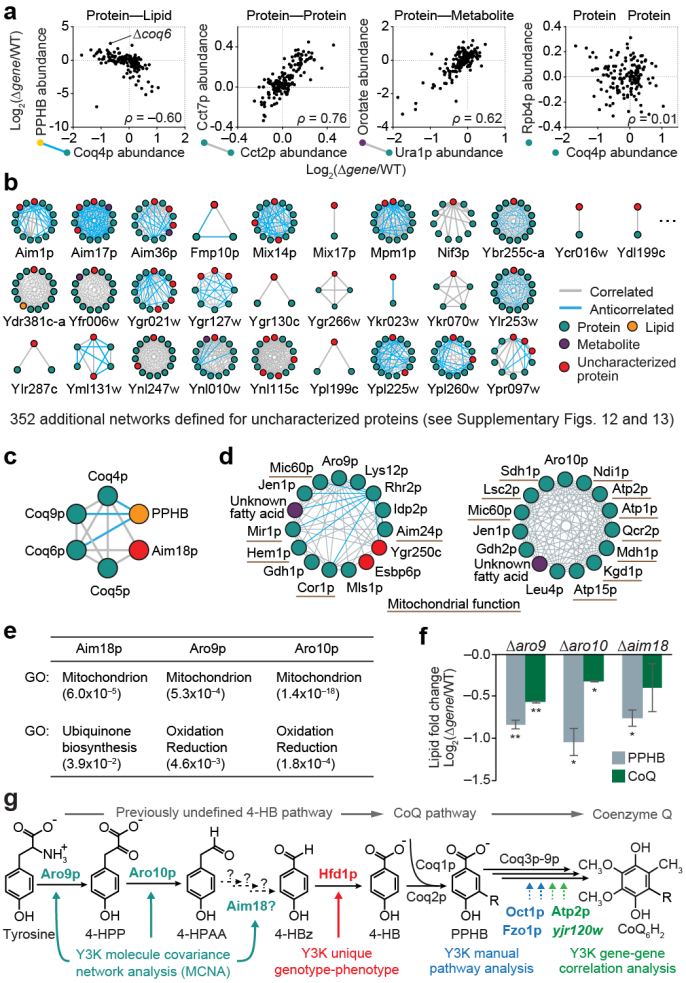


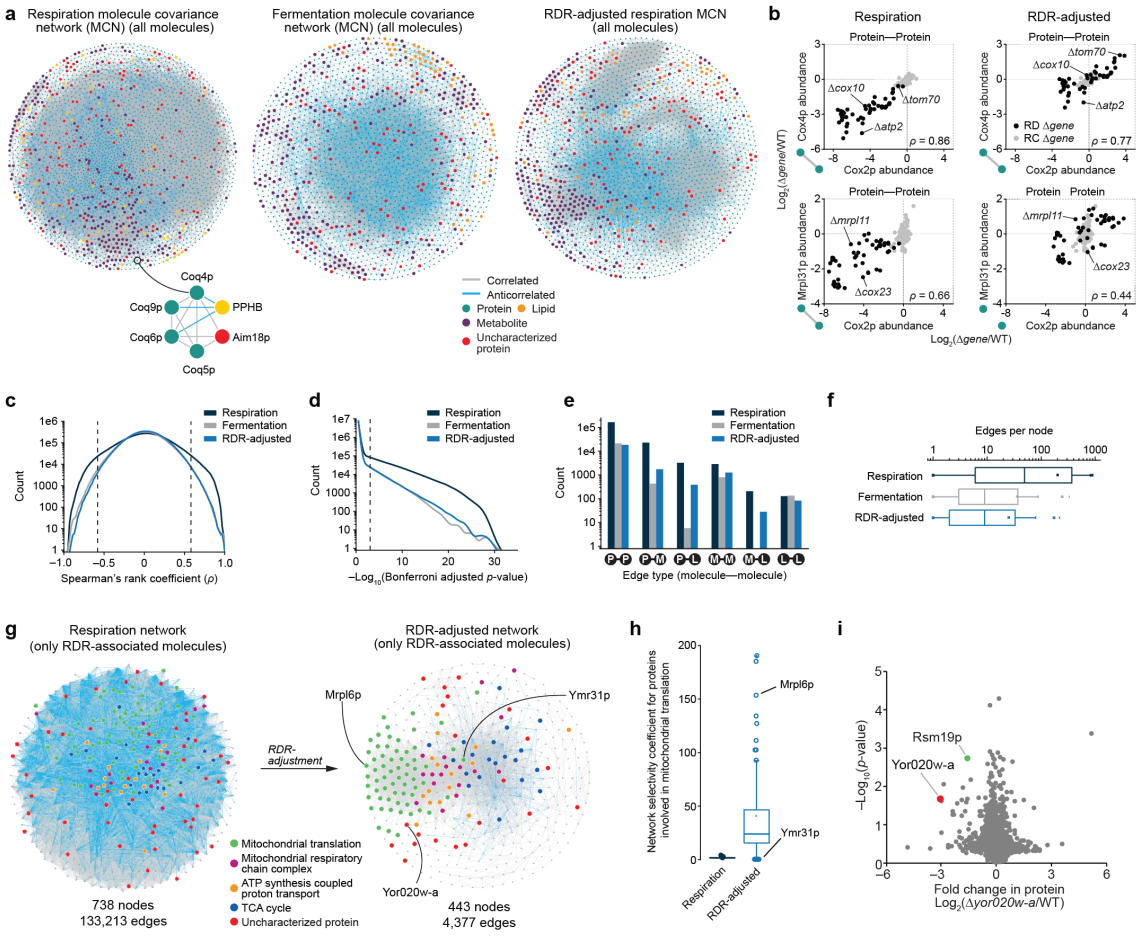
Figure 4.4: Multi-omic molecule covariance network analysis assists functional characterization. (a) Relative abundances of molecule pairs across Δ*gene* strains. Covariance assessed by Spearman's rank coefficient (ρ). **(b)** Nearest neighbor molecule covariance networks for a representative subset of uncharacterized proteins. **(c)** Network for Coq4p in the RDR-adjusted respiration data set. **(d)** Networks showing the 14 molecules most strongly correlated to Aro9p or Aro10p in the RDR-adjusted respiration data set. **(e)** GO term analyses of the Aim18p, Aro9p, and Aro10p networks (*p*-values). **(f)** Relative abundances of CoQ and PPHB (mean $\log_2$[Δ*gene*/WT], n = 2) in Δ*aro9*, Δ*aro10*, and Δ*aim18* strains compared to WT yeast cultured in pABA$^-$ media; **P* < 0.05; ***P* < 0.01 (two-sided Student's t-test). **(g)** Y3K-enabled characterization of proteins that support the CoQ pathway.

ribosome protein Rsm19p **(Supplementary Fig. S4.11i)**, suggesting that Yor020w-a is linked to mitochondrial translation.

Hundreds of additional uncharacterized proteins were linked to characterized molecules by our MCNA, providing a foundation for generating hypotheses about their functions **(Fig. 4.4b, Supplementary Figs. S4.12 and S4.13)**. For example, the MXP Aim18p was linked to a network of CoQ biosynthesis proteins, and Aro9p and Aro10p were linked to numerous mitochondrial proteins that support OxPhos **(Fig. 4.4c–e)**. Based on domain homology and predicted enzymatic functions, we hypothesized that Aim18p, Aro9p, and Aro10p could function in the Tyr- to-4-HB pathway **(Supplementary Fig. S4.14 and Supplementary Note 7)**. Consistently, when cultured in a pABA$^-$ media, Δ*aim18*, Δ*aro9*, and Δ*aro10* yeast are deficient in both CoQ and PPHB **(Fig. 4.4f)**. This work shows how global mass spectrometry profiling can be used to generate biological hypotheses and characterize protein functions through distinct multi-omic data analysis approaches **(Fig. 4.4g)**.

Discussion

A constant challenge in biology is to comprehensively monitor and understand the molecular effects of a defined alteration (e.g., a disease mutation, a drug treatment, or a gene deletion). Mass spectrometry (MS) has become central to answering this



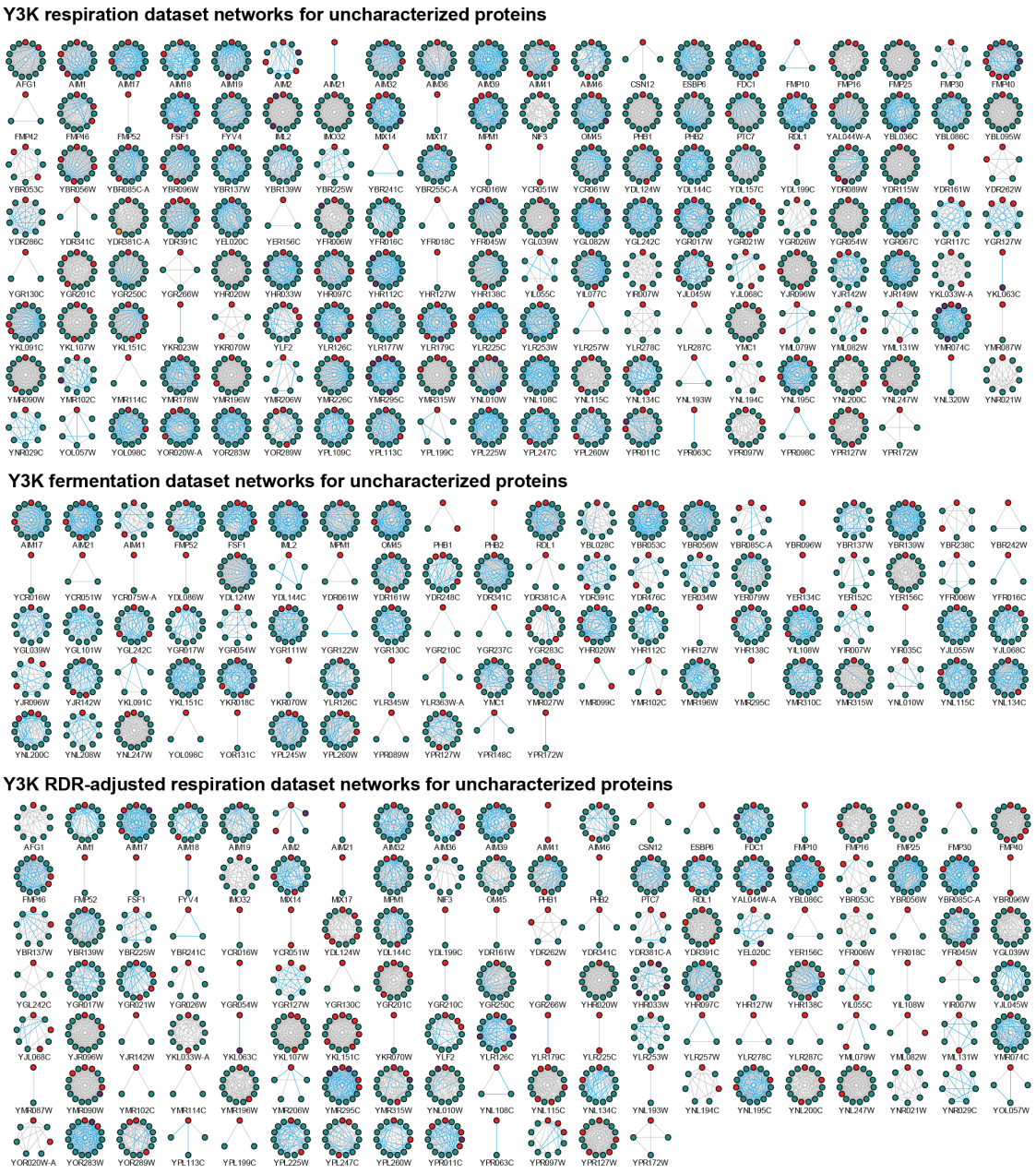
Supplementary Figure S4.11: Features of multi-omic molecule covariance networks. Network of all covariant molecules observed in each data set ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). **(b)** Regression analysis of pairs of RDR-associated molecules before and after RDR adjustment using Spearman's rank coefficient (ρ). Points corresponding to RD and RC Δ*gene* strains are indicated. **(c)** Distribution of calculated Spearman coefficients for all pairwise molecule covariance comparisons (ρ cutoff at ±0.58 used throughout the study is indicated). **(d)** Distribution of Bonferroni-adjusted *p*-values from all pairwise molecule comparisons (*p*-value cutoff at 0.001 used throughout the study is indicated). **(e)** Bar chart indicating number of protein–protein (P–P), protein–metabolite (P–M), protein–lipid (P–L), metabolite–metabolite (M–M), metabolite–lipid (M–L), and lipid–lipid (L–L) edges in each data set. **(f)** Box plots indicating the number of edges per node in the respiration, fermentation, and RDR-adjusted networks.

Supplementary Figure S4.11: (g) Network of all covariant RDR-associated molecules ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test) generated using the respiration (left) and RDR-adjusted (right) data sets. Nodes are highlighted according to GO category. **(h)** Box plots indicating the molecule covariance network (MCN) specificity coefficient for all nodes involved in mitochondrial translation in both the respiration and RDR-adjusted respiration RDR-associated molecule networks (shown in panel G). **(i)** Relative protein abundances (mean $\log_2$[Δ*yor020w-a*/WT], n = 2) versus statistical significance ($-\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS.

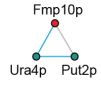
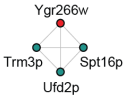
challenge.

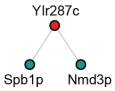
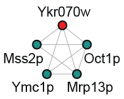
Here, we leveraged a subset of our multi-omic data set to investigate gaps in knowledge of CoQ biosynthesis. Despite CoQ's essential function in the mitochondrial electron transport chain, role as a key cellular antioxidant, and link to numerous human diseases (e.g., ataxias, myopathies, and nephrotic syndromes), multiple steps in CoQ biosynthesis remain uncharacterized[17,31,32]. In particular, enzymes involved in the initial stage of CoQ biosynthesis— wherein the headgroup precursor 4-HB is produced—were previously undefined in mammals and yeast.

Our Δ*gene*-specific phenotype detection approach suggested a role for the ancient aldehyde dehydrogenase superfamily member Hfd1p in 4-HB biosynthesis. Biochemical and genetic studies confirmed this role for Hfd1p in yeast and further demonstrated that the human homolog ALDH3A1 can also catalyze production of 4-HB *in vivo* and *in vitro* **(Fig. 4.2)**, thereby highlighting ALDH3A1 as a candidate disease gene for primary CoQ deficiency.

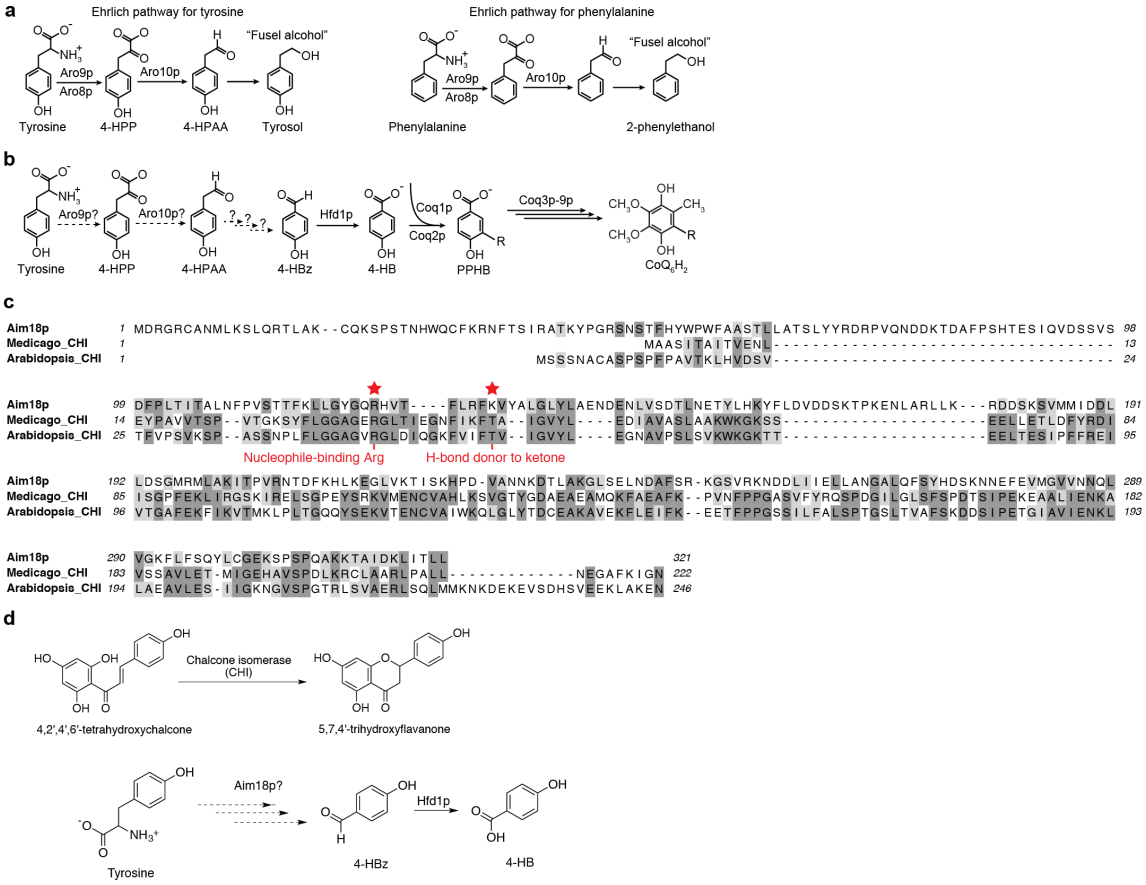


Supplementary Figure S4.12: Molecule covariance networks for uncharacterized proteins. 'Nearest neighbor' molecule covariance networks for all uncharacterized proteins observed across the respiration, fermentation, and RDR-adjusted respiration data sets ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). If more than 14 correlated molecules were present in a given covariance network, only the top 14 correlated molecules (nearest neighbors) are displayed.

Molecule covariance network	GO terms (for larger networks) or brief protein descriptions (for smaller networks)	Biological hypothesis generated by Y3K MCNA
Respiration networks		
Ydr381c-a 	Mitochondrion ($2.6x10^{-13}$) Mitochondrial respiratory chain complex III ($4.7x10^{-7}$)	Hypothesis: Ydr381c-a, an uncharacterized protein, supports mitochondrial OxPhos Complex III.
Ygr130c  Eis1p Tfs1p	Eis1p: Eisosome component Tfs1p: Phospholipid binding protein that regulates protein kinase signaling pathways	Hypothesis: Ygr130c regulates eisosome activity by linking it to a protein kinase signaling network
Yfr006w 	Golgi apparatus ($4.5x10^{-3}$) Intracellular protein transport ($7.8x10^{-5}$)	Hypothesis: Yfr006w, a putative peptidase, supports intracellular peptide transport.
Ybr255c-a 	Mitochondrion ($2.0x10^{-14}$) Oxidation-reduction process ($1.7x10^{-4}$)	Hypothesis: Ybr255c-a, an uncharacterized protein, supports mitochondrial OxPhos.
Mix17p  Mas2p	Mas2p: Subunit of the mitochondrial processing protease	Hypothesis: Mix17p, an uncharacterized mitochondrial protein, is involved in mitochondrial protein import and processing.
Ykr023w  Mgm101	Mgm101: Mitochondrial DNA repair protein.	Hypothesis: Ykr023w, an uncharacterized mitochondrial protein, supports mitochondrial genome maintenance.
Fmp10p  Ura4p Put2p	Put2p: delta-1-pyrroline-5-carboxylate dehydrogenase	Hypothesis: Fmp10p, an uncharacterized mitochondrial protein, is involved in proline-to-glutamate metabolism.
Ydl199c  Hxt5p	Hxt5p: Hexose transporter with high affinity for glucose.	Hypothesis: Ydl199c, a predicted membrane transporter, is involved in hexose transport.
Ygr266w  Trm3p Spt16p Ufd2p	Spt16p: subunit of FACT complex that regulates chromatin structure.	Hypothesis: Ygr266w, an uncharacterized mitochondrial outer membrane protein that is upregulated due to DNA replication stress, affects Spt16p activity.

Molecule covariance network	GO terms (for larger networks) or brief protein descriptions (for smaller networks)	Biological hypothesis generated by Y3K MCNA
Respiration networks (continued)		
Ylr253w 	Mitochondrion ($1.4x10^{-25}$) Oxidation-reduction process ($2.5x10^{-7}$)	Hypothesis: Ylr253w, an uncharacterized mitochondrial kinase, supports mitochondrial OxPhos.
Ylr287c  Spb1p Nmd3p	Spb1p: Ribosome processing Nmd3p: Ribosome processing	Hypothesis: Ylr287c supports cytosolic ribosome maturation.
Ynl247w 	Nucleolus ($1.3x10^{-8}$) Ribosome biogenesis ($4.4x10^{-8}$)	Hypothesis: Ynl247w supports cytosolic ribosome maturation.
Ykr070w  Mss2p Oct1p Ymc1p Mrp13p	Mss2p: CIV assemblyfactor Oct1p: Mitochondrial intermediate peptidase Ymc1p: Putative mitochondrial glycine transporter that supports heme biosynthesis.	Hypothesis: Ykr070w, an uncharacterized mitochondrial protein, supports mitochondrial Complex IV function.
Fermentation networks		
Fsf1p 	Mitochondrion ($9.8x10^{-8}$) Metabolic process ($3.7x10^{-3}$)	Hypothesis: Fsf1p supports mitochondrial metabolism
Ybl028c 	Nucleolus ($1.6x10^{-2}$) 90S preribosome ($2.0x10^{-2}$)	Hypothesis: Ybl028c supports cytosolic ribosome maturation.
Ybr096w Hem2p	Hem2p: Aminolevulinate dehydratase	Hypothesis: Ybr096w supports heme biosynthesis
Yer079w 	Nucleolus ($3.9x10^{-10}$) Ribosome biogenesis ($2.8x10^{-8}$)	Hypothesis: Yer079w is a negative regulator of ribosome biogenesis

Supplementary Figure S4.13: Examples of hypotheses that can be generated from a subset of the molecule covariance network analyses in this study. Nearest neighbor molecule covariance networks from uncharacterized proteins containing more than four connected nodes were tested for GO term enrichment using a Fisher's exact test with Benjamini–Hochberg FDR adjustment to account for multiple hypothesis testing. Networks containing four or fewer connected nodes were analyzed manually for functionally related molecules. Based on these MCNA results, biological hypotheses about the functions of the uncharacterized proteins shown were developed.



Supplementary Figure S4.14: Hypothesized pathways for Aro9p, Aro10p, and Aim18p. (a) Putative biochemical functions of Aro9p and Aro10p in catabolism of tyrosine and phenylalanine. **(b)** Predicted functions for Aro9p and Aro10p in the Tyr-to-4-HB-to-CoQ pathway. **(c)** Protein sequence alignments of Aim18p (*S. cerevisiae*) and chalcone isomerases (CHI) from *Medicago* and *Arabidopsis* highlighting conservation of putative catalytic residues (starred residues). **(d)** Example of a CHI catalyzed reaction (upper scheme) and the hypothesized pathway of Aim18p action (lower scheme).

Distinct Y3K data set analyses placed additional proteins into the CoQ biosynthesis pathway. MCNA showed unexpected connections between Aro9p, Aro10p, and mitochondrial OxPhos proteins, which helped place Aro9p and Aro10p into the Tyr-to-4-HB pathway **(Fig. 4.4)**. Similarly, links between Aim18p and known CoQ biosynthesis enzymes also connected Aim18p to CoQ biosynthesis. Furthermore, Y3K gene-gene correlation analyses and manual pathway analyses linked CoQ biosynthesis to other proteins whose molecular functions in this pathway are not yet fully defined (e.g. Atp2p, Fzo1p, and Oct1p). Disruption of the mammalian Fzo1p homolog, MFN2—a protein essential for mitochondrial fusion that harbors causative mutations in Charcot-Marie-Tooth disease[33]—was recently shown to cause CoQ deficiency through an unclear molecular mechanism[34]. Our results suggest that this unexpected relationship between MFN2 and CoQ biosynthesis is evolutionarily conserved, and establish yeast as a model system for further probing its mechanism.

Our Y3K data set provides many additional leads for further biochemical studies of numerous metabolic pathways that impact human health and disease, and we expect that the open access web utility (http://y3kproject.org/) will enable others to generate their own hypotheses. With demand for multi-omic data set analysis approaches increasing, we also hope that our multifaceted, data visualization

website will serve as a useful model for future studies.

We anticipate that the multi-omic Y3K data set will provide a resource for broader systems biology inquiries. For example, our definition of the yeast respiration deficiency response (RDR) **(Fig. 4.3)** may assist studies of how cells broadly respond to defects in OxPhos, which are observed in diverse diseases including many cancers. Our RDR work also suggests that a multi-omic fingerprint of numerous molecules could provide a highly specific biomarker panel.

Methods

Yeast strains and cultures. The parental (WT) *Saccharomyces cerevisiae* strain for this study was the haploid MATalpha BY4742. Single gene deletion (Δ*gene*) derivatives of BY4742 were either obtained through the gene deletion consortium[30] or made in-house using a *KanMX* deletion cassette to match those in the consortium collection. All gene deletions were confirmed by either proteomics (significant decrease in the encoded protein) or a PCR assay. Δ*gene* strains made in-house were also confirmed by gene sequencing.

Single lots of yeast extract ('Y') (Research Products International, RPI), peptone ('P') (RPI), agar (Fisher), dextrose ('D') (RPI), glycerol ('G') (RPI), and G418 (RPI) were used for all medias. YP and YPG solutions were sterilized by automated

autoclave. G418 and dextrose were sterilized by filtration (0.22 μm pore size, VWR) and added separately to sterile YP or YPG. YPD+G418 plates contained yeast extract (10 g/L), peptone (20 g/L), agar (15 g/L), dextrose (20 g/L), and G418 (200 mg/L). YPD media (fermentation cultures) contained yeast extract (10 g/L), peptone (20 g/L), and dextrose (20 g/L). YPGD media (respiration cultures) contained yeast extract (10 g/L), peptone (20 g/L), glycerol (30 g/L) and dextrose (1 g/L).

Yeast from a –80 °C glycerol stock were streaked onto YPD+G418 plates and incubated (30 °C, ~60 h). Starter cultures (3 mL YPD) were inoculated with an individual colony of yeast and incubated (30 °C, 230 rpm, 10–15 h). A WT culture was included with each set of Δ*gene* strain cultures (usually 19 Δ*gene* cultures and 1 WT culture). Cell density was determined by optical density at 600 nm (OD_{600}) as described[35]. YPD or YPGD media (100 mL media at ambient temperature in a sterile 250 mL Erlenmeyer flask) was inoculated with 2.5×10^6 yeast cells and incubated (30 °C, 230 rpm). Samples of the YPD cultures were harvested 12 h after inoculation, a time point that corresponds to early fermentation (logarithmic) growth. Samples of YPGD cultures were harvested 25 h after inoculation, a time point that corresponds to early respiration growth.

Liquid chromatography tandem mass spectrometry (LC-MS/MS) proteomics. 1×10^8 yeast cells were harvested by centrifugation (3,000 *g*, 3 min, 4 °C), the supernatant

was removed, and the cell pellet was flash frozen in $N_{2(l)}$ and stored at –80 °C. Yeast pellets were resuspended in 8 M urea, 100 mM tris (pH = 8.0). Yeast cells were lysed by the addition of methanol to 90%, followed by vortexing (~30 s). Proteins were precipitated by centrifugation (12,000 *g*, 5 min). The supernatant was discarded, and the resultant protein pellet was resuspended in 8 M urea, 10 mM tris(2-carboxyethyl)phosphine (TCEP), 40 mM chloroacetamide (CAA) and 100 mM tris (pH = 8.0). Sample was diluted to 1.5 M urea with 50 mM tris and digested with trypsin (Promega) (overnight, ~22 °C) (1:50, enzyme:protein). Samples were desalted using Strata X columns (Phenomenex Strata-X Polymeric Reversed Phase, 10 mg/mL). Strata X columns were equilibrated with one column volume of 100% acetonitrile (ACN), followed by 0.2% formic acid. Acidified samples were loaded on column, followed by washing with three column volumes of 0.2% formic acid or 0.1% TFA. Peptides were eluted off the column by the addition of 500 μL 40% ACN with either 0.2% formic acid or 0.1% TFA and 500 μL 80% ACN with either 0.2% formic acid or 0.1% TFA. Peptide concentration was measured using a quantitative colorimetric peptide assay (Thermo). LC-MS/MS analyses were performed using previously described methodologies1, 2.

LC/MS data analysis. Raw data files were acquired in batches of 60 (3 biological replicates of 19 Δ*gene* strains and 1 WT strain) with time between LC-MS analyses

minimized to reduce run-to-run variation. Batches of raw data files were subsequently processed using MaxQuant[36] (Version 1.5.0.25). Searches were performed against a target-decoy[37] database of reviewed yeast proteins plus isoforms (UniProt, downloaded January 20, 2013) using the Andromeda[37] search algorithm. Searches were performed using a precursor search tolerance of 4.5 ppm and a product mass tolerance of 0.35 Da. Specified search parameters included fixed modification for carbamidomethylation of cysteine residues and a variable modification for the oxidation of methionine and protein N-terminal acetylation, and a maximum of 2 missed tryptic cleavages. A 1% peptide spectrum match (PSM) false discovery rate (FDR) and a 1% protein FDR was applied according to the target-decoy method. Proteins were identified using at least one peptide (razor + unique). Proteins were quantified using MaxLFQ with an LFQ minimum ratio count of 2. LFQ intensities were calculated using the match between runs feature, and MS/MS spectra were not required for LFQ comparisons. Missing values were imputed where appropriate for proteins quantified in $\geqslant$ 50% of MS data files in a batch. Proteins not meeting this requirement were omitted from subsequent analyses. Imputation was performed on a replicate-by-replicate basis. For each replicate MS analysis a normal distribution with mean and standard deviation equivalent to that of the lowest 1% of measured LFQ intensities was generated. Missing values were filled in with

values drawn from this distribution at random. Approximately 4.05% and 4.53% of quantitative measurements were imputed in the respiration and fermentation proteomic data sets, respectively. Replicate protein LFQ values from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$[strain], n = 3). Average Δ*gene* LFQ intensities were normalized against their appropriate WT control (mean $\log_2$[Δ*gene*/WT], n = 3) and a 2-tailed t-test (homostatic) was performed to obtain *P* values.

To control for batch-specific effects, proteins having unexpected and characteristic misregulation across a majority of Δ*gene* strains processed together were identified and omitted from the data set. For each protein quantified within a batch of Δ*gene* strains a distribution of protein fold-changes (intra-batch) was generated. The analogous distribution of protein fold-changes from all other Δ*gene* strains processed separately (inter-batch) was created. These two distributions were compared against each other using a Kolmogorov-Smirnov test (2-tailed) to obtain *P* values. If a significant difference existed at $P < 0.05$ (Bonferroni-adjusted) protein abundance measurements were omitted from the batch in question. This process of comparing intra-batch and inter-batch protein fold change distributions was carried iteratively and to exhaustion and resulted in the omission of an average 165 proteins/Δ*gene* strain (~4.8% of quantified proteins) for respiration, and 188 proteins/Δ*gene* strain

(~5.9%) for fermentation.

Gas chromatography-mass spectrometry (GC-MS) metabolomics. 1×10^8 yeast cells yeast cells were isolated by rapid vacuum filtration onto a nylon filter membrane (0.45 μm pore size, Millipore) using a Glass Microanalysis Filter Holder (Millipore), briefly washed with phosphate buffered saline (1 mL), and immediately submerged into ACN/MeOH/H_2O (2:2:1, v/v/v, 1.5 mL, pre-cooled to –20 °C) in a plastic tube. The time from sampling yeast from the culture to submersion in cold extraction solvent was less than 30 s. Tubes with the extraction solvent, nylon filter, and yeast were stored at –80 °C prior to analysis.

Tubes with yeast extract (also still containing insoluble yeast material and the nylon filter) were thawed at room temperature for 45 min., vortexed (~15 s), and centrifuged at room temperature (6400 rpm, 30 s) to pellet insoluble yeast material. Yeast extract (25 μL aliquot) and internal standards (25 μL aqueous mixture of isotopically labelled alanine-2,3,3,3-d_4, adipic acid-d_{10}, and xylose-$^{13}C_5$ acid, 5 ppm in each) were aliquoted into a 2 mL plastic tube and dried by vacuum centrifuge (~1 hr). The dried metabolites were resuspended in pyridine (25 μL) and vortexed. 25 μL of N-methyl-N-trimethylsilyl]trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) was added, and the sample was vortexed and incubated (60 °C, 30 min). Samples were then transferred to a glass autosampler

vials and analyzed using a GC/MS instrument comprising a Trace 1310 GC coupled to a Q Exactive Orbitrap mass spectrometer. For the yeast metabolite extracts a linear temperature gradient ranging from 50 °C to 320 °C was employed spanning a total runtime of 30 minutes. Analytes were injected onto a 30 m TraceGOLD TG-5SILMS column (Thermo) using a 1:10 split at a temperature of 275 °C and ionized using electron ionization (EI). The mass spectrometer was operated in full scan mode using a resolution of 30,000 ($m/\Delta m$) relative to 200 m/z.

GC/MS data analysis. The resulting GC-MS data were processed using an in-house developed software suite (https://github.com/coongroup/Y3K-Software). Briefly, all m/z peaks are aggregated into distinct chromatographic profiles (i.e., feature) using a 10 ppm mass tolerance. These chromatographic profiles are then grouped according to common elution apex (i.e., feature group). The collection of features (i.e., m/z peaks) sharing a common elution apex, therefore, represent an individual EI-MS spectrum of a single eluting compound. The EI-MS spectra were then compared against a matrix run and a background subtraction was performed. Remaining EI-MS spectra are then searched against the NIST 12 MS/EI library and subsequently subjected to a high resolution filtering (HRF) technique as described elsewhere. EI-MS spectra that were not identified were assigned a numeric identifier. Feature intensity, which was normalized using total metabolite signal,

was used to estimate metabolite abundance. Following initial processing, raw data files were re-analyzed to extract metabolite signals which were not successfully deconvolved and registered as missing values in the data set. This process provided measurements for ~1.87%, and 2.25% of metabolites quantified in the respiration and fermentation data sets, respectively. Remaining missing values were imputed using the same imputation strategy as described in the proteomic data processing section. Quantitative values imputed using this process account for ~0.17% and 0.13% of metabolites in the respiration and fermentation data sets, respectively.

Replicate metabolite intensities from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$[strain], n = 3). Average Δ*gene* metabolite intensities were normalized against their appropriate WT control (mean $\log_2$[Δ*gene*/WT], n = 3) and a 2-tailed t-test was performed to obtain *P* values. To account for batch-specific effects the same Kolmogorov–Smirnov testing approach as described in the proteomic data processing section was used. Distributions of inter-batch and intra-batch metabolite fold changes were compared iteratively and those that were significantly different at $P < 0.05$ (Bonferroni-adjusted) resulted in metabolite abundance measurements being omitted from the batch in question (~15 metabolites/Δ*gene* strain (~5.0%) from respiration and ~21 metabolites/Δ*gene* strain (~5.9%) from fermentation).

Δ*Gene*-specific phenotype detection. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2$[Δ*gene*/WT]) and negative $\log_2$ fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(*p*-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}$(*p*-value) were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at a Euclidean distance of 0.70.

Regression analysis of Δ*gene*–Δ*gene* perturbation profiles. For all pairwise combinations of Δ*gene* strains from the same growth condition linear regression analysis was conducted on protein, lipid, and metabolite perturbation profiles, respectively. Fold change measurements (mean $\log_2$[Δ*gene*/WT], n = 3) from molecules where

FC > 0.7 and $P < 0.05$ were used and a minimum of 20 proteins, 10 metabolites, and 5 lipids, respectively, were required. These measurements were fit to a line and the associated Pearson correlation coefficient was reported. Coefficients carrying negative signs were set to 0. For pairs of Δ*gene* strains lacking a sufficient number of molecules that met the aforementioned criteria, the Pearson coefficient was reported as 0. Hierarchical clustering of Δ*gene*–Δ*gene* correlations was performed as described below.

Respiration deficiency response (RDR) abundance adjustment. All Δ*gene* strains grown under respiration conditions were classified as respiration deficient (RD) (51) or respiration competent (RC) (123) based on observation of a common perturbation profile signature. For all molecules profiled within RD Δ*gene* strains an RDR score was calculated. This metric represents the proportion of RD Δ*gene* strains over which the molecule was consistently perturbed, relative to all RD Δ*gene* strains where the molecule was quantified. Considering all RD Δ*gene* strains, 776 molecules produced an RDR score > 0.95 (consistently perturbed across more than 95% of RD Δ*gene* strains where quantified) and were subsequently classified as RDR-associated. For each RDR-associated molecule, individual RD Δ*gene* strain measurements were mean normalized and stored. These RDR-adjusted measurements were then used in described respiration–RDR analyses.

Regression analysis of RDR-adjusted Δ*gene*–Δ*gene* perturbation profiles. For all RD Δ*gene* strains linear regression analysis was performed pairwise on RDR-adjusted protein perturbation profiles. Fold change measurements from molecules where FC > 0.7 and $P < 0.05$ (p-value prior to RDR adjustment) were used and a minimum of 20 proteins was required. Correlations and clustering were otherwise conducted as described above.

Hierarchical clustering. All hierarchical clustering performed in this study was done in Perseus. For all clustering operations Spearman correlation was used with average linkage, preprocessing with k-means, and the number of desired clusters set to 300 for both rows and columns.

For clustering of Δ*gene* perturbation profiles, clustering was performed separately for fermentation and respiration data sets, and column-wise cluster order for fermentation and respiration data sets was generated using only protein fold change profiles. Column ordering was then applied to metabolite and lipid fold change data sets from the corresponding growth condition and row-wise clustering was conducted. GO term enrichment was performed in Perseus. P values were obtained from a Fisher's exact test, adjusted for multiple hypothesis testing[38] and reported where $P < 0.05$.

For the analysis of Δ*gene*–Δ*gene* correlations, clustering was performed on respi-

ration protein perturbation profile correlation data and the resultant ordering was applied to Δ*gene*–Δ*gene* correlation data sets from all other omes and growth conditions for parallel visual display. The same clustering process was carried out for the analysis of Δ*gene*–Δ*gene* correlations of RD Δ*gene* strains following RDR-adjustment.

Generation of Δ*gene* strains and cloning of genes and mutants for follow-up studies. *S. cerevisiae* (BY4742) gene deletion strains for *hfd1*, *atp2*, *ypr010c-a*, and *yjr120w* were generated using a PCR deletion strategy in which the open reading frames were replaced by a KanMX cassette from the pFA6a-kanMX6 plasmid. Briefly, KanMX was amplified with primers containing sequence homologous to sequence just upstream of the ATG and just downstream from the terminal codon for each ORF. Amplicons were transformed into BY4742, and yeast were plated onto YEPD plates containing 100 μg/mL G418. Knockouts were confirmed by PCR and sequencing.

To generate plasmid yeast gene constructs, *S. cerevisiae hfd1*, *atp2*, and *yjr120w* were amplified by Accuprime Pfu polymerase (Invitrogen, USA) with primers generating a SpeI site (forward) and SalI (reverse) (BamHI forward and EcoRI reverse for *yjr120w*). The *hfd1*, *atp2*, and *yjr120w* amplicons and the yeast expression vectors p426GPD and p423GPD were digested with SpeI and SalI or BamHI and EcoRI. *Hfd1* and *yjr120w* were ligated to p426GPD, *atp2* was ligated to p423GPD,

and each ligation was transformed into DH5α *E. coli*. Plasmid minipreps were performed and recombinants were confirmed by sequencing. *Hfd1* mutants were generated via standard site-directed mutagenesis, and mutations were confirmed by sequencing.

To generate plasmid human gene constructs, *Homo sapiens ALDH3A1* and *ALDH3A2* were amplified by Accuprime Pfu polymerase with primers generating a SpeI site (forward) and SalI (reverse). The *ALDH3A1* and *ALDH3A2* amplicons and the yeast expression vector p426GPD were digested with SpeI and SalI. *ALDH3A1* and *ALDH3A2* were ligated to p426GPD and each ligation was transformed into DH5α *E. coli*. Plasmid minipreps were performed and recombinants were confirmed by sequencing.

***Yjr120w* molecular biology studies—yeast growth assays.** Δ*atp2* and Δ*yjr120w* yeast were transformed with p426GPD plasmids (either encoding for Yjr120w or empty vector) and p423GPD (either encoding for Atp2p or empty vector) and grown on Ura⁻, His⁻ plates containing 2% glucose. Starter cultures were inoculated with individual colonies of yeast and incubated (30 °C, ~16 h, 230 rpm). To assay Δ*atp2* and Δ*yjr120w* yeast growth on agar plates, serial dilutions of yeast from a starter culture were prepared in Ura⁻, His⁻ media lacking glucose. 10-fold serial dilutions of yeast cells were dropped onto Ura⁻, His⁻ agar media plates containing either

glucose (2%, w/v) or glycerol (3%, w/v) and incubated (30 °C, 4 d).

***Yjr120w* molecular biology studies—mRNA quantitation.** BY4742 WT, Δ*coq8*, Δ*atp2*, and Δ*yjr120w* yeast were grown overnight in 3 mL YEPD. From the overnight culture, 2.5×10^6 cells were used to inoculate 100 mL YPGD media. 1 mL of culture was collected after 25 hours and total RNA was isolated using Masterpure Yeast RNA Purification Kit (Epicentre). 1 μg of RNA was reverse transcribed using Superscript III first strand synthesis kit (Thermo). Using the resultant cDNA as template, set up QPCR reactions: 2 μL cDNA, 12.5 μL Power Sybr Green Master Mix (Thermo), and 300 nmol/L forward and reverse primers. Primers amplifying the following targets were used: *atp2*, *yjr120w*, and *ubc6* (reference gene). QPCR cycled as follows: After an initial 2 minute incubation at 50 °C, template was denatured at 95 °C for 10 minutes, cycled 40 times: 95 °C for 15 s, 60 °C for 1 minute. RNA abundance was calculated using the $\Delta\Delta Ct$ method.

Hfd1p and ALDH3A1 biochemical studies—media lacking pABA. A specially formulated synthetic media lacking pABA ('pABA$^-$') was used for numerous follow-up studies in this project. This media consisted of CSM Mixture; Complete, 790 mg/L (# DCS0019, Formedium LTD, Hunstanton, U.K.) and yeast nitrogen base without amino acids and para-amino benzoic acid, 6.9 g/L (# CYN4102, Formedium

LTD, Hunstanton, U.K.).

Hfd1p and ALDH3A1 biochemical studies—yeast growth assays. *Δhfd1* yeast transformed with p426GPD plasmids encoding for Hfd1p variants were grown on uracil drop-out (Ura^-) synthetic media plates containing glucose (2%, w/v). Individual colonies of yeast were used to inoculate starter cultures of synthetic media lacking pABA ($pABA^-$) but containing 20 g/L glucose. To assay WT and *Δhfd1* yeast growth on agar plates, serial dilutions of yeast from a starter culture were prepared in $pABA^-$ media lacking glucose. 10^4, 10^3, or 10^2 yeast cells were dropped onto agar media plates containing either glucose (2%, w/v) or glycerol (3%, w/v) and incubated (30 °C, 4 d). The base medias for the agar plates consisted of either YEP (rich media), synthetic complete, $pABA^-$, $pABA^-$ supplemented with 100 μM 4-hydroxybenzoic acid, or pABA– supplemented with 100 μM pABA.

To assay yeast growth in liquid media, yeast from a $pABA^-$ starter culture were swapped into $pABA^-$ media with glucose (0.1%, w/v) and glycerol (3%, w/v) (base medium) at an initial density of 5×10^6 cells/mL. To interrogate the rescue efficacy of various compounds, 100 nM (final concentrations) of pABA, tyrosine, 4-HPP, 4-HPAA, 4-HPA, 4-HMA, 4-HBz, 4-HB, 4HPL, or *p*-coumarate were added to the base medium. The cultures were incubated in a sterile 96 well plate with an optical, breathable coverseal (shaking at 1140 rpm). Optical density readings (OD_{600}) were

obtained every 10 min. Respiratory growth rates were determined by fitting a linear equation to the respiratory growth phase and determining the slope of the line. Relative respiratory growth rates were determined by comparing cultures with additives to those without additive.

Hfd1p and ALDH3A1 biochemical studies—Quantitation of CoQ and 4-HB in pABA$^-$ Δ*hfd1* yeast cultures. 2.5×10^6 Δ*hfd1* yeast cells from a pABA$^-$ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA$^-$ media with glucose (0.1%, w/v), glycerol (3%, w/v), and potential rescue compound (100 nM pABA, 4-HPP, 4-HPAA, 4-HPA, 4-HBz, 4-HB, or none). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h (analogous to the primary respiration culture system used for this study), 1×10^8 yeast cells were harvested for lipidomic or metabolomic analyses, and CoQ and 4-HB were quantified by mass spectrometry as described above. These cultures and analyses were conducted in biological triplicate.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p phylogenetics. The amino acid sequences of the 19 known *Homo sapiens* ALDH proteins25 and *S. cerevisiae* Hfd1p (NP_013828.1) were aligned by MUSCLE[39], analyzed by ClustalW2 Phylogeny[40], and visualized in iTOL[41].

Hfd1p and ALDH3A1 biochemical studies—Mass spectrometry profiling of pABA$^-$ yeast cultures (WT, Δ*hfd1*, Δ*dpl1*, and Δ*coq8*). 2.5×10^6 yeast cells from a pABA$^-$ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA$^-$ media with glucose (0.1%, w/v), glycerol (3%, w/v), and rescue compound (100 μM 4-HB or none). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for lipidomic, metabolomics, and proteomic analyses by mass spectrometry as described in the main Methods section. These cultures and analyses were conducted in biological triplicate.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p, ALDH3A1, and ALDH3A2 expression and purification. PIPE cloning was used to generate pVP68K vectors encoding ALDH3A1, Hfd1p$^{C\Delta25}$, or ALDH3A2$^{C\Delta25}$ (Hfd1p or ALDH3A2 lacking their C-terminal 25 amino acids, which comprise putative transmembrane domains) fused to an 8His-cytoplasmically-targeted maltose-binding protein with a linker including a tobacco etch virus protease recognition site (8His-MBP-[TEV]-ALDH3A1, 8His-MBP-[TEV]-Hfd1p$^{C\Delta25}$, or 8His-MBP-[TEV]-ALDH3A2$^{C\Delta25}$). These constructs were expressed in *E. coli* (BL21[DE3]-RIPL strain) by autoinduction. Cells were isolated and resuspended in lysis buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 1 mg/mL lysozyme (Sigma), pH 7.5). Cells were lysed by sonication (4 °C, 2 × 20 s), and the lysate was clarified by centrifugation (15,000

g, 30 min, 4 °C). The clarified lysate was mixed with cobalt IMAC resin (Talon resin) and incubated (4 °C, 1 h). The resin was pelleted by centrifugation (700 g, 2 min, 4 °C) and washed three times (10 resin bed volumes each) with wash buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 10 mM imidazole, pH 7.5). His-tagged protein was eluted with elution buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 100 mM imidazole, pH 7.5). The eluted protein was concentrated with a 50-kDa MW-cutoff spin filter (Merck Millipore Ltd.) and exchanged into storage buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, pH 7.5). Protein concentrations were determined by absorbance at 280 nm. The MBP-fusion proteins were aliquoted, frozen in $N_{2(l)}$, and stored at –80 °C.

Hfd1p and ALDH3A1 biochemical studies—Hfd1p, ALDH3A1, and ALDH3A2 enzymology. Enzyme activity assays were conducted in groups of three replicate 100 μL reactions, each containing MBP-fusion protein (0.2–25 μg), 1 mM NAD^+, and 200 μM substrate (4-HBz or hexadecanal (Avanti 857458M)) in an aqueous buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% Triton X-100). NADH production was observed by monitoring fluorescence (356 nm excitation, 460 nm emission) over a 30–60 minute period with a Cytation 3 Imaging Reader (BioTek). K_M and k_{cat} values were determined by measuring reaction rates in the linear range at

varying substrate (4-HBz or hexadecanal) concentrations. Curve fitting to generate Michaelis-Menten parameters was performed using SigmaPlot (Systat Software, San Jose, CA). Reported activity represents the mean of three separate protein purifications.

Molecule Covariance Network Analysis For all pairwise combinations of molecules quantified within a particular growth condition, regression analysis was conducted using fold change measurements from all Δ*gene* strains having a measurement for both molecules in the pair. Spearman's regression analysis was performed to obtain correlation coefficients (ρ). From these test statistics P values were calculated using a two-sided Student's t-test. All P values were corrected for multiple hypothesis testing (Bonferroni) and correlations where $|\rho| \geqslant 0.58$ and $P < 0.001$ were reported. For RDR-adjusted regression analysis, the RDR adjustment procedure was carried out as described in the 'Respiration deficiency response (RDR) abundance adjustment' section (above). All pairs of covariant molecules are visualized as networks generated using the Gephi open graph visualization platform (version 0.9.0). Complete respiration, fermentation and RDR-adjusted respiration network layouts were generated using the Fruchterman–Reingold graph-drawing algorithm with area set to 10,000 and gravity set to 30. Gene Ontology terms were obtained from the Saccharomyces Genome Database (SGD). To calculate network selectivity

the following equation was used:

$$S_{MCN} = [E_{Obs,In}/E_{Tot,In}]/[(E_{Obs,Out} + 1)/E_{Tot,Out}]$$

Where S_{MCN} represents the selectivity coefficient for the molecule covariance network (MCN) surrounding an individual node of interest, $E_{Obs,In}$ is the number edges observed within a pathway of interest, $E_{Tot,In}$ is the number of total possible edges within the pathway of interest, $E_{Obs,Out}$ is the number of edges observed to molecules outside the pathway of interest, and $E_{Tot,Out}$ is the number total possible edges to molecules outside the pathway of interest.

Gene ontology (GO) term enrichment analysis was performed using a Fisher's exact test with subsequent Benjamini-Hochberg FDR adjustment39 to account for multiple hypothesis testing.

Proteomic analysis of Δ*yor020w-a* yeast 2.5×10^6 yeast cells from a $pABA^-$ (2% w/v glucose) starter culture (Δ*yor020w-a* or WT) were used to inoculate 100 mL of $pABA^-$ media with glucose (0.1%, w/v) and glycerol (3%, w/v). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for proteomic analyses by mass spectrometry as described in the main Methods section. These cultures and analyses were conducted in biological duplicate.

Quantitation of CoQ and PPHB in pABA$^-$ Δ*aro9*, Δ*aro10*, Δ*aim18*, and WT yeast cultures 2.5×10^6 yeast cells from a pABA$^-$ (2% w/v glucose) starter culture were used to inoculate 100 mL of pABA$^-$ media with glucose (0.1%, w/v) and glycerol (3%, w/v). These 100 mL cultures were incubated (30 °C, 230 rpm). After 25 h, 1×10^8 yeast cells were harvested for lipid analysis, and CoQ and PPHB were quantified by mass spectrometry as described in the Main methods section. These cultures and analyses were conducted in biological duplicate.

Respiration deficiency response analysis The densities of Δ*gene* cultures were compared to those of WT cultures (2-tailed T-test). Strains with slow growth in fermentation cultures (Δ*gene*/WT $\leqslant$ 0.2 and P < 0.05) were categorized as 'slow fermentation growth' strains (8 strains). Remaining strains were grouped into three categories based on their growth rates in respiration cultures. Strains with significantly decreased respiration growth (Δ*gene*/WT < 0.6 and P < 0.05) were considered respiration deficient (RD) (41 RD strains). Strains with borderline respiration growth (0.6 $\leqslant$ Δ*gene*/WT < 0.8) were categorized as 'borderline respiration' (14 strains). Strains with respiration growth rates near WT or better than WT (0.8 $\leqslant$ Δ*gene*/WT) were categorized as respiration competent (RC) (111 RC strains).

For PCA, average $\log_2$(Δ*gene*/WT) values for each protein, metabolite, and lipid measured in the respiration condition were analyzed using Perseus PCA software.

PCA projections were exported from Perseus.

For volcano plot analyses, average $\log_2$(RD/RC) values were calculated as [mean $\log_2$(RD Δ*gene* strains/WT)] – [mean $\log_2$(RC Δ*gene* strains/WT)]. A t-test (2-tailed, homostatic) was performed to obtain *P* values. *P* values were corrected for multiple hypothesis testing by multiplying each *P* value obtained by the number of biomolecules included in this analysis (4,116) (Bonferroni correction).

For GO term analyses, proteins were separated as increasing in RD strains (positive $\log_2$[RD/RC]) or decreasing in RD strains (negative $\log_2$[RD/RC]). Proteins with Bonferroni-corrected $P < 1 \times 10^{-20}$ were collected from each group and subjected to GO term enrichment analysis (http://geneontology.org/page/go-enrichment-analysis). Select GO terms were highlighted because they were significantly enriched (Bonferroni corrected $P < 0.05$) in proteins that were reduced (–) or increased (+) in RD strains. Boxplots of select molecules were generated using matplotlib in python to compare particular molecules across all RD and RC strains.

For ROC analysis, RD strains were considered positive examples whereas RC cells were considered negative examples. Using the $\log_2$(Δ*gene*/WT) values for individual biomolecules as a discriminator, ROCs were generated by calculating false positive rate (FPR) and true positive rate (TPR) for values that fall above a particular cutoff for molecules that are increased in RD strains relative to WT and

below that cutoff for molecules that are decreased in RD strains relative to WT. A + sign indicates that an increase in that molecule is predictive of RD whereas a – sign indicates that a reduction in that molecule is predictive of RD.

Supplementary Notes

Development of a stable and reproducible respiration culture condition. To profile diverse yeast strains during respiratory growth, when mitochondrial OxPhos is highly active, we first needed to develop a distinct respiration condition suitable for large-scale investigation. Early log phase fermentation cultures repress mitochondrial respiration, cultures containing solely non-fermentable sugars preclude growth of respiration deficient yeast, and high glucose cultures grown past the diauxic shift are too biologically dynamic to allow reproducible sampling across a large-scale study[42,43]. To overcome these problems, we developed a culture system that includes low glucose (1 g/L) and high glycerol (30 g/L), enabling a short fermentation phase followed by a longer respiration phase. This respiration condition affords steady growth and a stable biological state—as reflected by a proteome that is constant over multiple hours **(Supplementary Fig. S4.1c–e)**—and, thus, an essential window for reproducible sample harvesting.

Δ*Gene*-specific phenotype detection. To identify Δ*gene*-specific phenotypes, we broadly surveyed our data for characteristic outlier abundance measurements. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2$[Δ*gene*/WT]) and negative $\log_2$fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(*p*-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}$(*p*-value) were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. The results of this analysis and representative examples are highlighted **(Fig. 4.2, Supplementary Figs. S4.5 and S4.6)**. We observed maximal Euclidean distances across a range of 0.006 to 1.25. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at 0.70 and report 714 molecules (4.6% of considered cases across both culture

conditions) which exceed this threshold. This procedure provided a useful 'first pass' analysis and afforded a truncated set of leads, which were used to develop biological hypotheses.

Lack of effect of Dpl1p disruption on the Tyr-to-4-HB-to-CoQ pathway. To test the idea that the CoQ biosynthesis and sphingolipid catabolism pathways are independent, we examined Δ*dpl1* yeast, which lack a known dihydrosphingosine phosphate lyase. Δ*dpl1* yeast show neither a pABA$^-$ respiratory growth phenotype nor CoQ deficiency **(Supplementary Fig. S4.7j,k)**. These results demonstrate that disruption of the Tyr-to-4-HB pathway in Δ*hfd1* yeast is not downstream of a defect in sphingolipid metabolism. Furthermore, proteome analyses showed that Δ*hfd1* cultured without 4-HB and pABA are similar to Δ*coq8* yeast—but not Δ*dpl1* yeast—and adding 4-HB to Δ*hfd1* cultures returns their proteomes to WT-like profiles **(Supplementary Fig. S4.7l,m)**.

Quantitative definition of the respiration deficiency response (RDR). To quantitatively define the RDR, we categorized strains as respiration deficient (RD) or competent (RC) and examined differences between these two groups. Principal component analysis of the Y3K respiration data set revealed marked separation of RD and RC strains **(Fig. 4.3c and Supplementary Fig. S4.8a)**. The underlying

phenotype changes that distinguish RD and RC strains include proteins, lipids, and metabolites **(Fig. 4.3d)**. RDR perturbations include significant decreases in ATP synthase, TCA cycle, and MICOS proteins **(Fig. 4.3e,f and Supplementary Fig. S4.8b)**, likely to decrease allocation of useless proteome mass to dysfunctional mitochondria[44]. Importantly, the RDR also includes a positive response, and numerous proteins—including protein folding, NADH metabolism, and proteasome assembly proteins—are significantly upregulated in RD strains **(Fig. 4.3e,f)**. Numerous individual molecules—including lactate, alanine, 2-hydroxyglutarate, tyrosol, 4-HB, Gpx2p, and Ahp1p, among many others—are significantly perturbed in RD strains and strongly predictive of respiration deficiency **(Supplementary Fig. ??c,d)**. Our quantitative assessment of the RDR highlights biochemical features of the cellular response to defects in mitochondrial respiration, and suggests that a multi-omic assessment of proteins, lipids, and metabolites could afford a highly specific biomarker panel for diseases affected by OxPhos deficiency.

RDR normalization procedure. Δ*gene* strains were classified as RD (51) or respiration competent (RC) (123) based on observation of a common perturbation profile signature in the respiration culture condition. For each molecule we calculated an RDR score. This metric represents the proportion of RD Δ*gene* strains over which the molecule was consistently perturbed, relative to all RD Δ*gene* strains

where the molecule was quantified. Across all RD Δ*gene strains*, 776 molecules were identified as having an RDR score > 0.95 (consistently perturbed across more than 95% of RD Δ*gene* strains where quantified) and classified as RDR-associated. The individual measurements of these RDR-associated molecules were then mean normalized ('RDR-adjusted') using abundance values from RD Δ*gene* strains. This normalization procedure revealed characteristic deviations from the general RDR **(Supplementary Fig. S4.9)**. Importantly, this procedure enables visualization of Δ*gene*-specific changes. For example, prior to RDR normalization, the expected decrease in Coq8p in Δ*coq8* yeast is obscured by RDR-associated proteins with large abundance changes **(Supplementary Fig. S4.9d)**. RDR normalization not only uncovers the decrease in Coq8p, but a significant decrease in Coq5p, a functionally-related CoQ biosynthesis protein, also becomes readily apparent **(Supplementary Fig. S4.9d)**.

Molecular defects of Δ*yjr120w* yeast. To examine the molecular basis for the CoQ deficiency of Δ*yjr120w* yeast, we inspected our proteomics data set, which revealed significant decreases in ATP synthase proteins, especially Atp2p **(Supplementary Fig. S4.10a)**. Compared to other strains, the large decrease in Atp2p is unique to Δ*yjr120w* and Δ*atp2* **(Supplementary Fig. S4.10b)**. A relationship between *yjr120w* and *atp2* is also suggested by their genetic proximity **(Supplementary Fig. S4.10c)**.

Plasmid overexpression of *atp2* rescues the Δ*yjr120w* respiratory growth defect **(Supplementary Fig. S4.10d)**, indicating a functional relationship between *atp2* and *yjr120w in vivo*. A decrease in *atp2* mRNA in the Δ*yjr120w* strain is a component of the underlying mechanism **(Supplementary Fig. S4.10e)**. Interestingly, CoQ deficiency was also observed in Δ*atp2* yeast **(Fig. 4.3h)**.

Predicted enzymatic functions of Aim18p, Aro9p, and Aro10p. Since 1907, yeast have been known to catabolize amino acids into fusel (German for 'bad liquor') alcohols through the Ehrlich pathway[45,46], but the physiological roles for the enzymes involved—such as Aro9p and Aro10p—are not fully understood. Aro9p and Aro10p were previously thought to provide a simple catabolic route for extracting nitrogen from aromatic amino acids[47] **(Supplementary Fig. S4.14a)**, but our MCNA unexpectedly indicated strong correlations between Aro9p, Aro10p, and proteins involved in mitochondrial respiration **(Fig. 4.4d,e)**, suggesting a more complicated biological function that supports OxPhos. We hypothesized that this function might be in the Tyr-to-4-HB-to-CoQ pathway **(Supplementary Fig. S4.14b)**, given the putative enzymatic activities of Aro9p and Aro10p in tyrosine and phenylalanine metabolism. Consistently, when cultured in pABA$^-$ media, Δ*aro9* and Δ*aro10* yeast are deficient in CoQ and PPHB **(Fig. 4.4f)**.

Aim18p is a protein of undefined molecular function that has been detected

in mitochondria[48] and potentially linked to mitochondrial inheritance (Altered Inheritance of Mitochondria, 'AIM') by large-scale studies in yeast[49]. Protein sequence alignments show that Aim18p contains a chalcone-flavone isomerase (CHI)-like domain **(Supplementary Fig. S4.14c)**, whose homologs in plants typically function on aromatic small molecules (chalcones) **(Supplementary Fig. S4.14d)**[50–52]. Given the potential for this protein domain to catalyze modifications of aromatic small molecules, we hypothesized that Aim18p might function in the Tyr-to-4-HB pathway to produce the CoQ headgroup (Supplementary Fig. 14d). Consistently, when cultured in pABA$^-$ media, we observed deficiency of PPHB in Δ*aim18* yeast **(Fig. 4.4f)**.

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, pp. 339–347, 2014.

[2] 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, pp. 701–714, 2015.

[3] A. C. Peterson, J. P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S.

Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/quadrupole-orbitrap mass spectrometer, Part I: Design and characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, 2014.

[4] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, "Multiple high-throughput analyses monitor the response of E. coli to perturbations.," *Science (New York, N.Y.)*, vol. 316, pp. 593–7, 2007.

[5] J. M. Buescher, W. Liebermeister, M. Jules, M. Uhr, J. Muntel, E. Botella, B. Hessling, R. J. Kleijn, L. Le Chat, F. Lecointe, U. Mäder, P. Nicolas, S. Piersma, F. Rügheimer, D. Becher, P. Bessieres, E. Bidnenko, E. L. Denham, E. Dervyn, K. M. Devine, G. Doherty, S. Drulhe, L. Felicori, M. J. Fogg, A. Goelzer, A. Hansen, C. R. Harwood, M. Hecker, S. Hubner, C. Hultschig, H. Jarmer, E. Klipp, A. Leduc, P. Lewis, F. Molina, P. Noirot, S. Peres, N. Pigeonneau, S. Pohl, S. Rasmussen, B. Rinn, M. Schaffer, J. Schnidder, B. Schwikowski, J. M. Van Dijl, P. Veiga, S. Walsh, A. J. Wilkinson, J. Stelling, S. Aymerich, and U. Sauer, "Global network reorganization during dynamic adaptations of

Bacillus subtilis metabolism.," *Science (New York, NY)*, vol. 335, pp. 1099–1103, 2012.

[6] E. G. Williams, Y. Wu, P. Jha, S. Dubuis, P. Blattmann, C. A. Argmann, S. M. Houten, T. Amariuta, W. Wolski, N. Zamboni, R. Aebersold, and J. Auwerx, "Systems proteomics of liver mitochondria function.," *Science (New York, N.Y.)*, vol. 352, p. aad0189, 2016.

[7] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, and M. Daniel, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, pp. 500–505, 2016.

[8] J. Nunnari and A. Suomalainen, "Mitochondria: In sickness and in health," 2012.

[9] W. J. Koopman, P. H. Willems, J. A. M. Smeitink, and D. Ph, "Monogenic mitochondrial disorders," *The New England journal of medicine*, vol. 366, pp. 1132–1141, 2012.

[10] S. B. Vafai and V. K. Mootha, "Mitochondrial disorders as windows into an ancient organelle.," *Nature*, vol. 491, pp. 374–83, 2012.

[11] D. J. Pagliarini, S. E. Calvo, B. Chang, S. A. Sheth, S. B. Vafai, S. E. Ong, G. A. Walford, C. Sugiana, A. Boneh, W. K. Chen, D. E. Hill, M. Vidal, J. G. Evans,

D. R. Thorburn, S. A. Carr, and V. K. Mootha, "A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology," *Cell*, vol. 134, pp. 112–123, 2008.

[12] S. E. Calvo, K. R. Clauser, and V. K. Mootha, "MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins," *Nucleic Acids Research*, vol. 44, pp. D1251–D1257, 2016.

[13] A. Sickmann, J. Reinders, Y. Wagner, C. Joppich, R. Zahedi, H. E. Meyer, B. Schönfisch, I. Perschil, A. Chacinska, B. Guiard, P. Rehling, N. Pfanner, and C. Meisinger, "The proteome of Saccharomyces cerevisiae mitochondria.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 13207–12, 2003.

[14] E. D. Green and M. S. Guyer, "Charting a course for genomic medicine from base pairs to bedside.," *Nature*, vol. 470, pp. 204–13, Feb. 2011.

[15] D. J. Pagliarini and J. Rutter, "Hallmarks of a new era in mitochondrial biochemistry," 2013.

[16] B. J. Floyd, E. M. Wilkerson, M. T. Veling, C. E. Minogue, C. Xia, E. T. Beebe, R. L. Wrobel, H. Cho, L. S. Kremer, C. L. Alston, K. A. Gromek, B. K. Dolan, A. Ulbrich, J. A. Stefely, S. L. Bohl, K. M. Werner, A. Jochem, M. S. Westphall,

J. W. Rensvold, R. W. Taylor, H. Prokisch, J.-J. P. Kim, J. J. Coon, and D. J. Pagliarini, "Mitochondrial Protein Interaction Mapping Identifies Regulators of Respiratory Chain Function.," *Molecular cell*, vol. 63, pp. 621–32, 2016.

[17] C. M. Quinzii and M. Hirano, "Coenzyme Q and mitochondrial disease," 2010.

[18] A. Kalén, E. L. Appelkvist, and G. Dallner, "Age-related changes in the lipid compositions of rat and human tissues.," *Lipids*, vol. 24, pp. 579–584, 1989.

[19] R. Bentley and V. Ramsey, "The origin of the benzoquinone ring of coenzyme Q 9 in the rat," *Biochemical and …*, vol. 5, no. 6, pp. 443–446, 1961.

[20] S. Merle, D. J. Robbins, and H. Emerson, "Phenolic Acid Metabolites of Tyrosine," vol. 236, no. 9, 1960.

[21] F. Pierrel, O. Hamelin, T. Douki, S. Kieffer-Jaquinod, U. Mühlenhoff, M. Ozeir, R. Lill, and M. Fontecave, "Involvement of mitochondrial ferredoxin and para-aminobenzoic acid in yeast coenzyme q biosynthesis," *Chemistry and Biology*, vol. 17, pp. 449–459, 2010.

[22] B. Marbois, L. X. Xie, S. Choi, K. Hirano, K. Hyman, and C. F. Clarke, "para-aminobenzoic acid is a precursor in coenzyme Q6 biosynthesis in Saccharomyces cerevisiae," *Journal of Biological Chemistry*, vol. 285, pp. 27827–27838, 2010.

[23] K. Nakahara, A. Ohkuni, T. Kitamura, K. Abe, T. Naganuma, Y. Ohno, R. A. Zoeller, and A. Kihara, "The Sj??gren-Larsson Syndrome Gene Encodes a Hexadecenal Dehydrogenase of the Sphingosine 1-Phosphate Degradation Pathway," *Molecular Cell*, vol. 46, pp. 461–471, 2012.

[24] Z. J. Liu, Y. J. Sun, J. Rose, Y. J. Chung, C. D. Hsiao, W. R. Chang, I. Kuo, J. Perozich, R. Lindahl, J. Hempel, and B. C. Wang, "The first structure of an aldehyde dehydrogenase reveals novel interactions between NAD and the Rossmann fold.," *Nature structural biology*, vol. 4, pp. 317–326, 1997.

[25] B. Jackson, C. Brocker, D. C. Thompson, W. Black, K. Vasiliou, D. W. Nebert, and V. Vasiliou, "Update on the aldehyde dehydrogenase gene (ALDH) superfamily.," *Human genomics*, vol. 5, pp. 283–303, 2011.

[26] V. De Laurenzi, G. R. Rogers, D. J. Hamrock, L. N. Marekov, P. M. Steinert, J. G. Compton, N. Markova, and W. B. Rizzo, "Sjögren-Larsson syndrome is caused by mutations in the fatty aldehyde dehydrogenase gene.," *Nature genetics*, vol. 12, pp. 52–7, 1996.

[27] T. Kitamura, T. Naganuma, K. Abe, K. Nakahara, Y. Ohno, and A. Kihara, "Substrate specificity, plasma membrane localization, and lipid modification of

the aldehyde dehydrogenase ALDH3B1.," *Biochimica et biophysica acta*, vol. 1831, pp. 1395–401, 2013.

[28] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. a. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional Discovery via a Compendium of Expression Profiles," *Cell*, vol. 102, pp. 109–126, 2000.

[29] P. Kemmeren, K. Sameith, L. A. L. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. Van Wageningen, C. W. Ko, S. Van Heesch, M. M. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. C. H. Brabers, A. J. Miles, D. Bouwmeester, S. R. Van Hooff, H. Van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. Van Leenen, M. J. A. Groot Koerkamp, and F. C. P. Holstege, "Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors," *Cell*, vol. 157, pp. 740–752, 2014.

[30] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel,

M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. a. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, "Functional profiling of the Saccharomyces cerevisiae genome.," *Nature*, vol. 418, pp. 387–391, 2002.

[31] L. N. Laredj, F. Licitra, and H. M. Puccio, "The molecular genetics of coenzyme Q biosynthesis in health and disease," 2014.

[32] U. C. Tran and C. F. Clarke, "Endogenous synthesis of coenzyme Q in eukaryotes," *Mitochondrion*, vol. 7, 2007.

[33] S. Züchner, I. V. Mersiyanova, M. Muglia, N. Bissar-Tadmouri, J. Rochelle, E. L. Dadali, M. Zappia, E. Nelis, A. Patitucci, J. Senderek, Y. Parman, O. Evgrafov, P. D. Jonghe, Y. Takahashi, S. Tsuji, M. a. Pericak-Vance, A. Quattrone, E. Battaloglu, A. V. Polyakov, V. Timmerman, J. M. Schröder, and J. M. Vance,

"Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A.," *Nature genetics*, vol. 36, pp. 449–451, 2004.

[34] A. Mourier, E. Motori, T. Brandt, M. Lagouge, I. Atanassov, A. Galinier, G. Rappl, S. Brodesser, K. Hultenby, C. Dieterich, and N. G. Larsson, "Mitofusin 2 is required to maintain mitochondrial coenzyme Q levels," *Journal of Cell Biology*, vol. 208, pp. 429–442, 2015.

[35] A. S. Hebert, A. E. Merrill, J. a. Stefely, D. J. Bailey, C. D. Wenger, M. S. Westphall, D. J. Pagliarini, and J. J. Coon, "Amine-reactive neutron-encoded labels for highly plexed proteomic quantitation.," *Molecular & cellular proteomics : MCP*, vol. 12, pp. 3360–9, 2013.

[36] 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, pp. 1367–72, 2008.

[37] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "Andromeda: A peptide search engine integrated into the MaxQuant environment," *Journal of Proteome Research*, vol. 10, pp. 1794–1805, 2011.

[38] T. Author, Y. Benjamini, Y. Hochberg, and Y. Benjaminit, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling

the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

[39] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, 2004.

[40] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. Mcgettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, pp. 2947–2948, 2007.

[41] I. Letunic and P. Bork, "Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy," *Nucleic Acids Research*, vol. 39, 2011.

[42] P. Picotti, B. Bodenmiller, L. N. Mueller, B. Domon, and R. Aebersold, "Full Dynamic Range Proteome Analysis of S. cerevisiae by Targeted Proteomics," *Cell*, vol. 138, pp. 795–806, 2009.

[43] A. Casanovas, R. R. Sprenger, K. Tarasov, D. E. Ruckerbauer, H. K. Hannibal-Bach, J. Zanghellini, O. N. Jensen, and C. S. Ejsing, "Quantitative analysis of proteome and lipidome dynamics reveals functional regulation of global lipid metabolism," *Chemistry and Biology*, vol. 22, pp. 412–425, 2015.

[44] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J. R. Williamson, and T. Hwa, "Overflow metabolism in Escherichia coli results from efficient proteome allocation," *Nature*, vol. 528, pp. 99–104, 2015.

[45] F. Ehrlich, "Über die Bedingungen der Fuselölbildung und über ihren Zusammenhang mit dem Eiweissaufbau der Hefe," *Berichte der deutschen chemischen Gesellschaft*, vol. 40, pp. 1027–1047, 1907.

[46] L. H. Hazelwood, J.-M. G. Daran, A. van Maris, J. T. Pronk, and J. R. Dickinson, "The Ehrlich Pathway for Fusel Alcohol Production: a Century of Research on Saccharomyces cerevisiae Metabolism," *Applied and Environmental Microbiology*, vol. 74, pp. 2259–2266, 2008.

[47] M. M. Kneen, R. Stan, A. Yep, R. P. Tyler, C. Saehuan, and M. J. McLeish, "Characterization of a thiamin diphosphate-dependent phenylpyruvate decarboxylase from Saccharomyces cerevisiae," 2011.

[48] J. Reinders, R. P. Zahedi, N. Pfanner, C. Meisinger, and A. Sickmann, "Toward the complete yeast mitochondrial proteome: Multidimensional separation techniques for mitochondrial proteomics," *Journal of Proteome Research*, vol. 5, pp. 1543–1554, 2006.

[49] D. C. Hess, C. Myers, C. Huttenhower, M. A. Hibbs, A. P. Hayes, J. Paw, J. J. Clore, R. M. Mendoza, B. S. Luis, C. Nislow, G. Giaever, M. Costanzo, O. G. Troyanskaya, and A. A. Caudy, "Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis," *PLoS Genetics*, vol. 5, 2009.

[50] M. Gensheimer and A. Mushegian, "Chalcone isomerase family and fold: no longer unique to plants," *Protein Sci*, vol. 13, pp. 540–544, 2004.

[51] M. N. Ngaki, G. V. Louie, R. N. Philippe, G. Manning, F. Pojer, M. E. Bowman, L. Li, E. Larsen, E. S. Wurtele, and J. P. Noel, "Evolution of the chalcone-isomerase fold from fatty-acid binding to stereospecific catalysis.," *Nature*, vol. 485, pp. 530–3, May 2012.

[52] J. M. Jez, M. E. Bowman, R. a. Dixon, and J. P. Noel, "Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase.," *Nature structural biology*, vol. 7, pp. 786–791, 2000.

Chapter 5

DEVELOPMENT OF WEB-BASED DATA VISUALIZATION TOOLS AND A PLATFORM FOR CODELESS GENERATION OF CUSTOM DATA ANALYSIS WEB PORTALS

Portions of this chapter have been published:

Marx H*, Minogue CV*, Jayaraman D, Richards AL, **Kwiecien NW**, Sihapirani AF, Rajasekar S, Maeda J, Garcia K, Del Valle-Echevarria AR, Volkening JD, Westphall MS, Roy S, Sussman MR, Ané JM, Coon JJ. *A proteomic atlas of the legume Medicago truncatula and its nitrogen-fixing endosymbiont Sinorhizobium meliloti.*
Nature Biotechnology. **2016**, doi:10.1038/nbt.3681.

Stefely JA*, **Kwiecien NW***, Freiberger EC, Richards AL, Jochem A, Rush MJP, Ulbrich A, Robinson KP, Hutchins PD, Veling MT, Guo X, Kemmerer ZA, Connors KJ, Trujillo EA, Sokol J, Marx H, Westphall MS, Hebert AS, Pagliarini DJ, Coon JJ. *Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling.*
Nature Biotechnology. **2016**, doi:10.1038/nbt.3683.

* Authors contributed equally

Introduction

Recent advances in mass spectrometry (MS) profiling technologies have afforded substantial increases in both speed of data acquisition and experimental throughput[1,2]. These advances have opened the door to large-scale MS-based profiling studies where the analysis of hundreds, or even thousands, of samples is considered routine. However, the creation of increasingly larger data sets presents a new challenge in both processing and interpretation of results. Currently, publically available tools for the analysis of large MS data sets—particularly multi-omic data sets—remain poorly developed. There is a great need in the MS community for software solutions which facilitate rapid exploration, integration, and dissemination of data. Online data analysis and visualization tools have become increasingly popular in other areas of science[3–6], and stand to alleviate many of the issues associated with analysis of large MS data sets. Functional web-based utilities are an efficient means to share results with collaborators, and minimize the burden of file transfer and version control. Online web-portals expedite data lookups and enable users to rapidly access and explore all measurements within a data set. Further, web-based visualizations are desirable as they enable quantitative data to be automatically synthesized into plots and graphs which are more easily understood and interpretable.

Currently, the process of developing a web-based interface for MS data exploration is tedious and time consuming. Construction of these tools requires computer programming and web development expertise which many researchers lack. As a result, online data analysis portals are rarely developed, and if so, they are built by individual labs on a project-specific basis. Tools designed for non-programmers which serve to convert MS data into a format that can be uploaded into the cloud, interfaced with interactive visualizations, and shared with collaborators, do not exist and need to be developed. We have recently started work on a new platform—targeted towards non-programmers—which enables codeless generation of online MS data exploration portals. Using this platform, researchers are able to create project-specific sites and upload results directly to the web from generic spreadsheets of MS data (i.e., peak tables). Then, users can select specific visualizations through which they wish to explore their data that are automatically embedded into their custom site. Once created, users are free to share these web tools with collaborators around the world at the click of a mouse.

Here we present two custom web visualization portals designed to serve as complimentary utilities for resources recently published in *Nature Biotechnology*. Additionally, we report on the development of a web-based platform for codeless generation of project-specific, online data visualization portals.

The Medicago Protein Compendium

Medicago truncatula is a premier model legume for the study of symbiotic relationships between plants and microbes. *M. truncatula* forms a complex symbiotic relationship with the nitrogen-fixing soil bacteria *Sinorhizobium meliloti*. Root nodules form on the plant wherein these bacteria fix nitrogen from the soil, and transfer it back to the host[7,8]. This symbiosis largely mitigates the need for fertilization of *M. truncatula* to support plant growth. Agricultural solutions which afford increasingly higher yields, and more sustainable crops at lower cost, are necessary to support the world's growing population. A reduction in crop fertilization requirements would doubtless be of great benefit in this regard. *M. truncatula's* close phylogenetic relationship to many agricultural crops, and diminished need for fertilization, makes this an extremely relevant and important system to study.

The Wisconsin Medicago Group is a consortium of researchers from the University of Wisconsin-Madison. This group is dedicated to applying novel technologies to elucidate the biochemical mechanisms which regulate the complex plant-bacterium symbiosis between *M. trunctula* and *S. meliloti*. In 2013, work began on a large-scale MS analysis to comprehensively profile *M. truncatula* and its nitrogen fixing symbiont *S. meliloti* at the proteomic, phosphoproteomic, and acetylomic level. These data were collected over many months, and produced a

data set containing 23,013 protein groups (19,679 from *M. truncatula*, 3,334 from *S. meliloti*), 20,120 protein phosphorylation sites, and 734 lysine acetylation sites. This data set (The Medicago Protein Compendium) represents the most extensive profile of *M. truncatula* acquired to date, and was recently published as a resource in Nature Biotechnology[9]. In order to make these data more readily accessible to biological and agricultural researchers, we developed a complimentary online data exploration tool (http://compendium.medicago.wisc.edu). All data from the compendium can be rapidly queried using a simple and intuitive lookup utility and subsequently visualized through interactive plots and tables. Here we will briefly describe the functionality of this developed web tool.

Data Accession. All protein data from the Medicago Protein Compendium can be readily queried using lookup tools on the site's 'Search Data' page **(Figure 5.1)**. Here we provide a search form which supports queries built using multiple optional parameters ('Accession', 'Organism', 'Description', and 'Gene'). A user can select the parameter they wish to search on from an associated dropdown list, and add or remove terms by clicking the '+' and '-' buttons. If multiple search parameters are provided, a concatenated query is built and submitted as a request to the remote database.

It is of note, that by performing an empty search, all database entries are re-

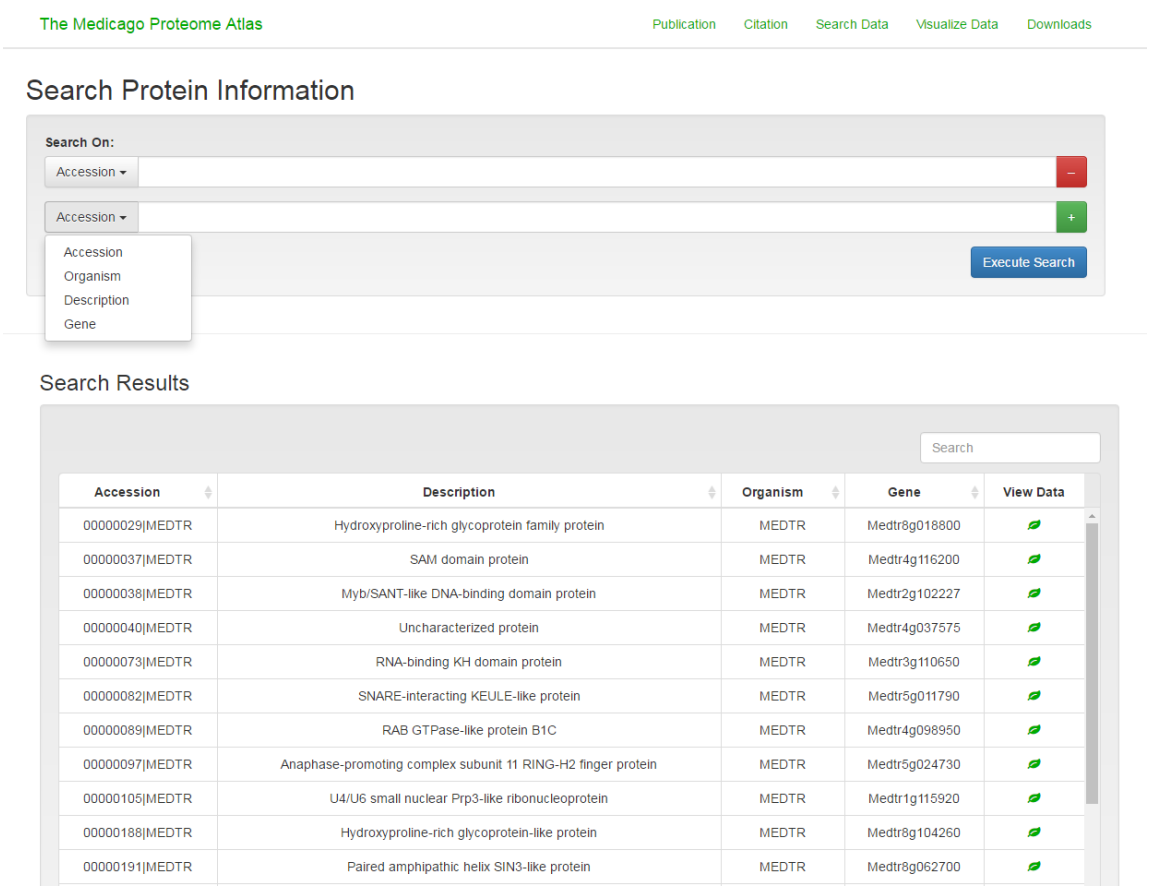


Figure 5.1: The Medicago Protein Compendium–Data lookup. Data accession page supporting dynamic protein queries from multiple, optional, search parameters. Returned results are displayed in the data table which can be filtered through the 'search' text box. All returned results contain a clickable link which redirects to an associated data visualization page.

turned. Queried results are added to the attached data table which supports filtering, sorting, and pagination of returned data. For each returned result, a clickable link which redirects to a data visualization page for the specified protein is provided. These data visualization pages contain all protein-specific quantitative and qualitative data, including phosphorylation and acetylation data, if available.

Data Visualization. By clicking any returned protein entry in the 'Search Data' table, a user is redirected to a 'Data Visualization' page containing all quantitative and qualitative data from the selected molecule. At the top of the page, the selected protein is listed along with its unique database identifier **(Figure 5.2)**. To provide users with information about detected sequence coverage, a collapsible panel displaying the entire protein amino acid sequence is shown. Here, identified peptide regions are highlighted in yellow. Along the top of this panel a calculation of percent sequence coverage (amino acids observed/total amino acids) is displayed **(Figure 5.2)**.

All quantitative and qualitative data associated with the selected protein can be observed in the data table located in the center of the page **(Figure 5.2)**. The 'Protein Databases' tab displays the corresponding protein entry—with accession and description—in each of the five databases utilized in the study (Augustus, Ensembl, JCVI, RefSeq, and Uniprot). The 'Peptides' tab displays qualitative information

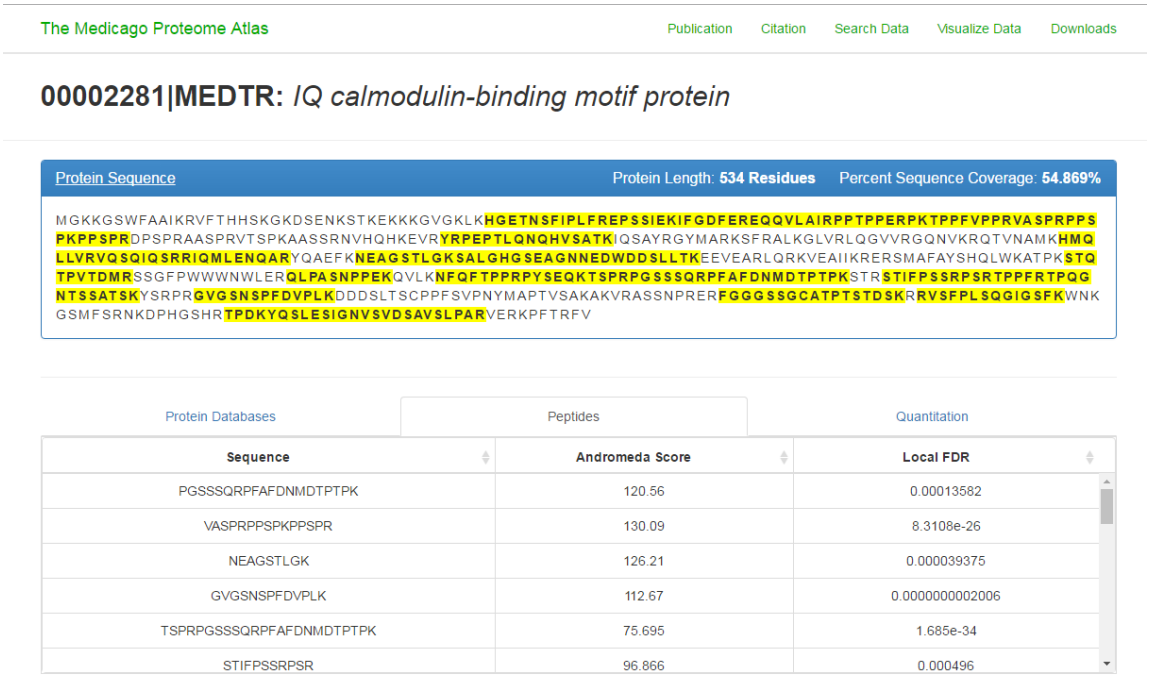


Figure 5.2: The Medicago Protein Compendium–Qualitative data visualization. Data Visualization page displaying all qualitative data associated with IQ calmodulin-binding motif protein. Molecule accession code, and parent organism are specified at the top of the page. The protein's amino acid protein sequence is shown in the blue panel with identified peptide regions highlighted in yellow. Quantitative and qualitative data are available here in tabular format, and can be selectively accessed by clicking the 'Protein Databases,' 'Peptides,' and 'Quantitation' tabs.

(Andromeda score and local FDR) for each protein-specific peptide identified in the study. Finally, the 'Quantitation' tab shows quantitative data for the selected protein across all tissues, with an indication of which labeling method (TMT or LFQ) was employed to obtain the measurement.

Quantitative protein, phosphorylation, and acetylation data is displayed below in the form of an interactive bar chart and heat map **(Figure 5.3)**. All quantitative phosphorylation and acetylation data was measured using a TMT labeling strategy. This approach was advantageous as it yielded relative abundances of modified sites across all tissues, except for Apical Meristem which was not profiled. Relative abundance changes for each modified site are displayed as columns in the heat map with the modified residue/position indicated along the x-axis, and tissue along the y-axis. Each pixel reflects abundance changes—relative to the mean—for a single tissue. Alongside the heat map is a bar chart reflecting unmodified protein abundance changes within each tissue. Note, the row-wise ordering of tissues is maintained between the heat map and bar chart. Users can toggle between quantitation methods (LFQ and TMT) using the associated dropdown and the bar chart will update automatically. Each pixel and bar in these two visualizations supports tooltips which display additional information on mouse hover.

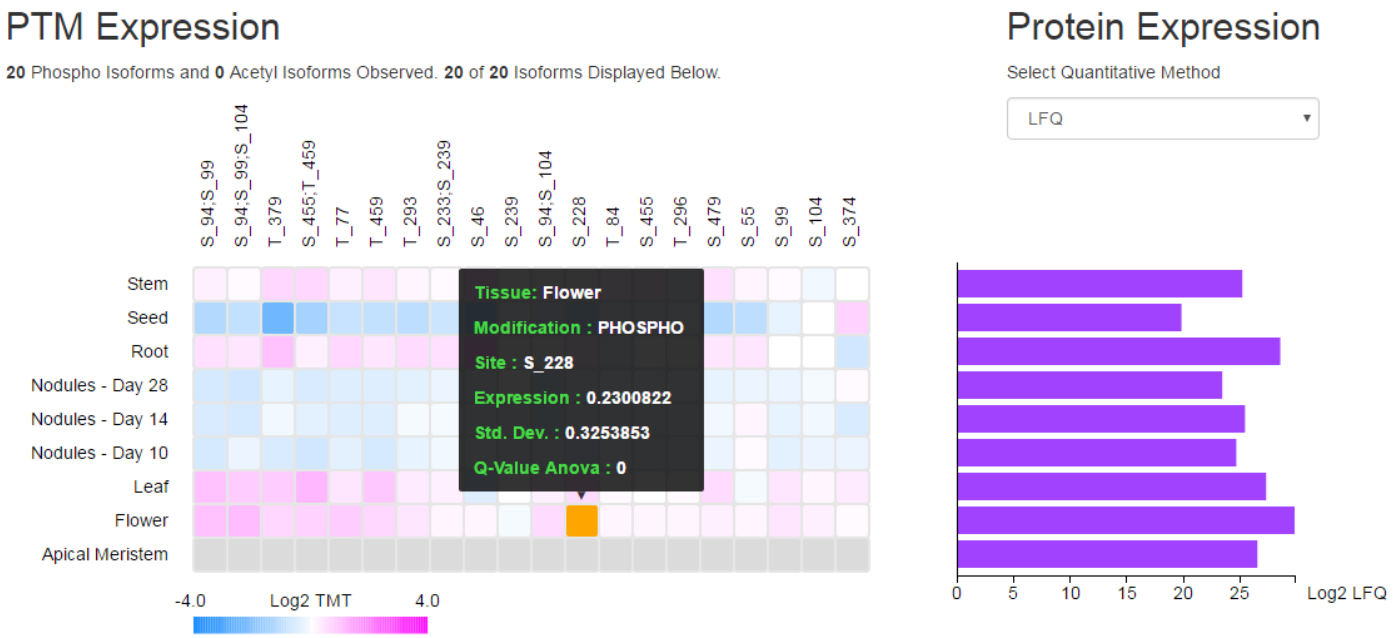


Figure 5.3: The Medicago Protein Compendium–Quantitative data visualization. Interactive data visualizations highlighting changes in modified and unmodified protein abundances. The heat map **(left)** indicates mean-normalized fold changes ($\log_2$) in phosphorylated and acetylated forms of the selected protein. Modified residues are indicated along the top of the map. Each row contains data from a single tissue. A tooltip containing extended measurement information is displayed by hovering over any pixel. Unmodified protein abundance measurements are shown in the bar chart **(right)**. Users can specify a quantitative method (either 'LFQ' or 'TMT') from the dropdown menu, and the associated data will be displayed. Each bar stems from a single tissue, and is aligned with its associated row in the heat map.

The Y3K Project Online

As discussed in **Chapter 4**, the Y3K project was a large-scale multi-omic study wherein 174 single gene deletion ($\Delta gene$) yeast strains were comprehensively profiled at the proteomic, metabolomic, and lipidomic level, under two distinct growth conditions[10]. This effort yielded more than 3,000 individual MS data files, and produced a data set containing over 3.5 million biomolecule measurements. Analysis and integration of this large multi-omic data set presented a unique challenge, as tools to facilitate our analyses simply did not exist. In order to enable our team to rapidly explore, interact with, and compare these data at depth, we constructed a unique web portal (http://y3kproject.org). This online tool presented our data through a number of interactive and easily interpretable data visualizations. We employed three different analyses in our study, which linked seven new proteins to Coenzyme Q (CoQ) biosynthesis. Importantly, each of these analyses can be recapitulated online using our web resource. Here, we will briefly describe the design of the Y3K website, as well as the data visualizations and analyses which this portal supports.

Design. The Y3K web portal was constructed using the Twitter Bootstrap framework, and supports numerous data visualizations developed using the D3.js JavaScript

charting library[11]. All data is stored on a remote server in a MySQL database which is appropriately indexed to facilitate rapid data lookups. The Y3K site was designed to be used primarily as a data exploration tool, and we provide seven different interactive visualizations on the 'Data Visualization' page. In addition to these interactive plots, we provide ample information about the design of the project, composition of the data set, and research labs involved with Y3K under the 'About Y3K' page. All data from our data set can be queried in tabular format under the 'Lookup Data' tab. By entering search terms into two simple input fields (Molecule/Strain and Growth Condition) any of the Y3K data can be accessed and downloaded directly from an internet browser. In an effort to make this web resource as useful as possible for all researchers, we have included an extensive 'Tutorial' page. On this page, we list eight common analyses and tasks which a user may want to perform, and include a detailed description of how to carry out each of these routines with an illustrated reference. Finally, in order to provide continued support for this web tool to the community, we have added a 'Contact Us' form to each page. This form allows a user to send an email to the Y3K website administrators, alerting them to previously undetected glitches and bugs, or to request additional usage information.

Visualizations. All data from the Y3K data set can be accessed through numerous interactive visualizations. These plots were designed to provide users with a number of vantage points through which to view these data, and have greatly facilitated our own analyses. All visualizations can be updated without refresh by selecting inputs from corresponding dropdown menus, buttons, and lists. In order to provide as much information as possible, all data points in these plots support tooltips which display extended information about measurements and calculated results. Furthermore, we have incorporated active legends to all visualizations, which update in real-time in accordance with the data actively being displayed.

KO vs WT. We provide a view of all profiled molecule perturbations—within a particular ome and growth condition—for each Δ*gene* strain under the 'KO vs WT' tab. Here, these data are presented as volcano plots with molecule fold change (n=3, mean $\log_2$[KO-WT]) displayed along the x-axis, and statistical significance (*p*-value; two-tailed Student's t-test) displayed along the y-axis **(Figure 5.4a)**. Data points are differentially colored and sized according to both fold change and statistical significance ($p>0.05$ and $|FC|<1$ = gray; $p<0.05$ and $|FC|<1$ = blue; $p<0.05$ and $|FC|>1$ = green). These plots can be updated automatically by selecting a different ome, growth condition, or Δ*gene* strain using the appropriate form inputs. Specific fold change and *p*-value data, as well as molecule metadata (names and identifiers),

can be observed in a tooltip by hovering over any data point. Double-clicking any data point pushes the associated quantitative data and qualitative metadata to a table located at the bottom of the tab.

KO vs. KO. We enable users to directly compare molecular abundance changes between any two Δ*gene* strains in the 'KO vs. KO' tab. Here, users can select any two Δ*gene* strains from a single ome and growth condition. Upon selection, measured fold changes (n=3, mean $\log_2$[KO-WT]) for molecules quantified across both Δ*gene* strains are displayed as a scatter plot **(Figure 5.4b)**. Additionally, a line of best fit is appended to the plot, and a Pearson correlation coefficient (calculated using all shared molecules) is reported. The x- and y-axes inherently fail to capture *p*-value information, which we provide in associated data point tool tips. To better convey the *p*-value dimension, we have implemented a dynamic data point highlighting scheme. Here, users can select different highlighting options ('Shared', 'Unique', 'Correlated', 'Anticorrelated', or 'Highlighting Off') from a dropdown menu. Each of these options selectively colors different data points within the KO–KO scatter plot based on observed fold change and *p*-values. For instance, by selecting 'Shared,' all data points which have an absolute fold change > 0.7 and *p*-value< 0.05 in both Δ*gene* strains will be highlighted in green. Alternatively, by selecting 'Unique,' molecules which meet these criteria in only one of the two strains will be colored in

red or blue (x-axis strain and y-axis strain, respectively). All highlighting schemes are described in an associated legend which updates in accordance with a user's selection.

KO–KO Correlations. The previously described 'KO vs KO' tab provides valuable information about phenotypic similarities between any two Δ*gene* strains. However, these comparisons are made one at a time, mitigating a user's ability to rapidly identify pairs of similar Δ*gene* strains. In the 'KO–KO Correlations' tab, all calculated Pearson correlation coefficients—from a single strain—can be viewed simultaneously as a bar chart **(Figure 5.4c)**. These bars are ranked by descending correlation coefficient or in alphabetical order with respect to Δ*gene* name. These bars can be hovered over to identify the correlated strain, as well as the specific Pearson correlation coefficient.

ΔMolecule Across KOs. Changes in molecular abundance in response to a diverse set of perturbations (gene knockouts) can explored for individual molecules in the 'ΔMolecule Across KOs' tab. Here, users can select a single molecule—from an easily filtered list—and growth condition of interest. All perturbation data associated with the selected molecule is then displayed in the form of a volcano plot, with measured fold change (n=3, mean $\log_2$[KO-WT]) along the x-axis, and

statistical significance (*p*-value; two-tailed Student's t-test [homostatic]) along the y-axis **(Figure 5.4d)**. Each data point is associated with a single Δ*gene* strain which can be identified by hovering over the point. Through this perspective, users can rapidly identify Δ*gene* strains associated with large abundance changes of their selected molecule. This view is particularly useful for researchers interested in a small set of molecules, as they can very quickly access specific information about their targets of interest.

Molecule vs Molecule. In order to identify covariance between molecules across profiled omes, we employed an analysis technique called Molecule Covariance Network Analysis (MCNA). MCNA is similar to coexpression analyses frequently employed in large-scale mRNA studies[12–14]. Using this approach we identified 237,342 pairs of coregulated molecules (Bonferroni-adjusted *p*-value < 0.001, $|\rho| \geqslant 0.58$), which can be explored in the 'Molecule vs Molecule' tab. Here, users can select any molecule of interest from a filterable list. Upon selection of a molecule, all other correlated molecules are displayed in a data table with complimentary information—number of strains where both molecules were quantified and a correlation coefficient (Spearman's rho). Selection of a coregulated molecule from this secondary table will trigger display of a molecule–molecule correlation plot. Here, measured abundance fold changes (n=3, mean $\log_2$[KO-WT]) from both molecules are displayed as a

scatter plot **(Figure 5.4e)**, where each data point corresponds to a single Δ*gene* strain. Tooltips are displayed by hovering over individual data points.

Molecule Covariance Networks. As previously described, the 'Molecule vs Molecule' tab displays individual pairs of covariant molecules as scatter plots of measured fold changes across Δ*gene* strains. Alternatively, this data can be viewed as a network, which displays many molecule–molecule correlations simultaneously—a view which is provided in the 'Molecule Covariance Networks' tab **(Figure 5.4f)**. We have created covariance networks for each profiled molecule. Here, all positively or negatively correlated molecules—(Bonferroni-adjusted p-value < 0.001, $|\rho| \geqslant 0.58$)—are ranked in descending order ($|\rho|$). The most highly-correlated molecules (maximum of 24) are then used to create a nearest-neighbor network, where each molecule is represented as a single node and colored according to ome. Edges are drawn between pairs of correlated or anti-correlated molecules, weighted according to correlation coefficient ($|\rho|$), and colored based on the sign of the correlation. Each of these networks has been pre-computed and stored as a JSON object on the remote server. This front-end processing step improves performance of network visualization by eliminating computationally intensive on-the-fly queries and calculations. Users are able to select any profiled molecule and an interactive nearest neighbor network will be generated automatically. To facilitate exploration of indi-

vidual correlations, users can double click any node and unconnected edges will disappear. A second double-click will restore all edges to the displayed network.

RD vs. RC. All profiled strains in our Y3K study were classified as either respiration deficient (RD) or competent (RC) based on observed growth rate (relative to WT) and presentation of a conserved multi-omic signature—the respiration deficiency response (RDR). The RDR is comprised of 776 molecules which were characteristically perturbed across a large plurality ($\geqslant$ 95%) of RD gene strains, where quantified. This signature reflects a generalized stress response stemming from respiration defects—a common feature of many diseases. The composition of this response is of interest as it affords novel biological insight into the underpinnings of respiration deficiency. In order to enable users to explore molecules characteristically perturbed as part of the RDR we constructed a two-part visualization under the 'RD vs RC' tab. All molecule fold changes are stratified into RD and RC groups and visualized as box plots in the left panel **(Figure 5.4g)**. On the right, a volcano plot showing fold change (mean $\log_2$[RD/WT]–mean $\log_2$[RC/WT]) and statistical significance (-$\log_{10}$[Bonferroni-adjusted *p*-value]; two-sided Student's t-test [homostatic]) is displayed **(Figure 5.4h)**. Any molecule selected will result in the RD/RC box plots updating, and will be highlighted in green in this volcano plot.

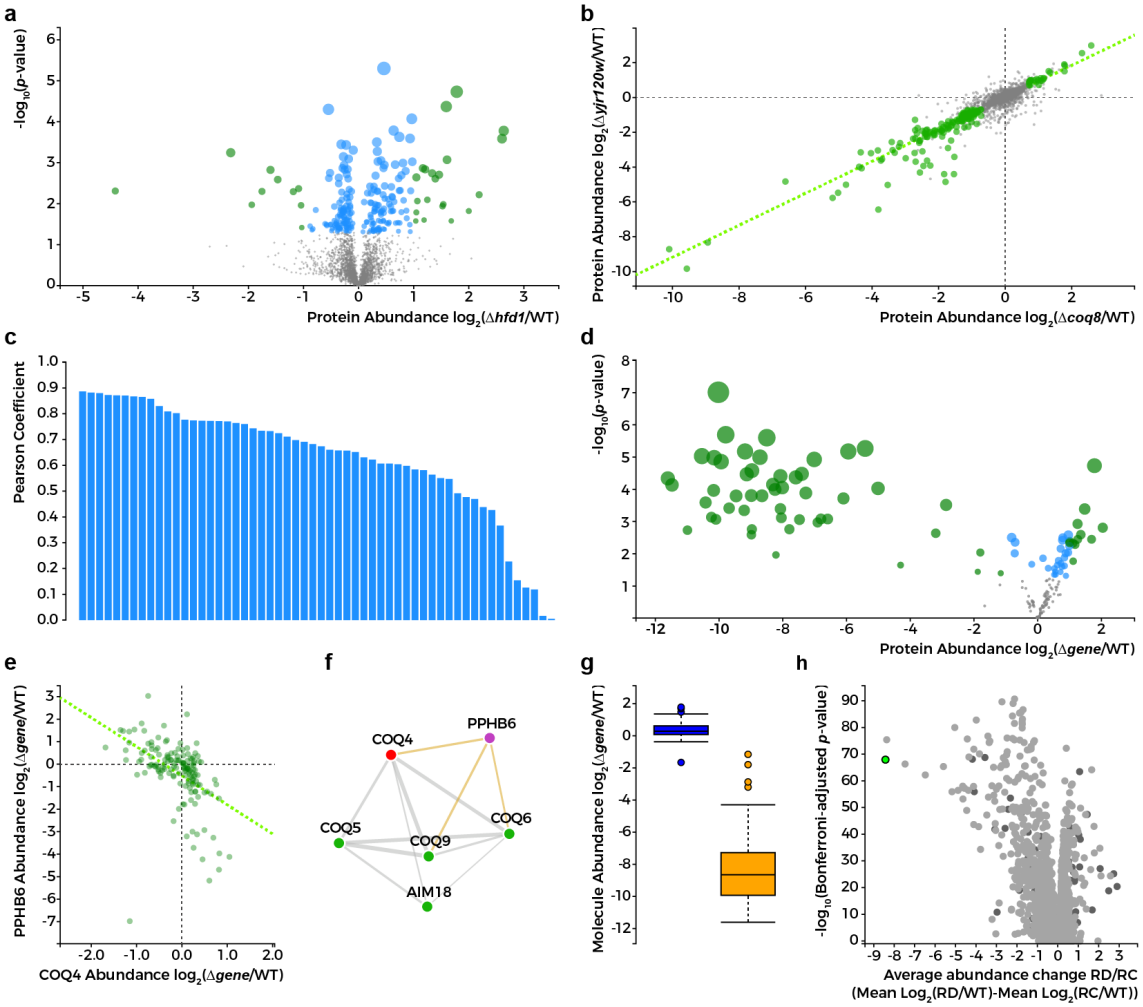


Figure 5.4: The Y3K Project Online–Interactive data visualizations. Representative examples of all data visualizations which can be automatically generated in the Y3K project web portal (http://y3kproject.org). **a.)** KO vs WT. Volcano plot showing all molecule perturbations from a single Δ*gene* strain. Fold change (mean $\log_2$[Δ*gene*/wt], n=3) is shown along the x-axis, and statistical significance (-$\log_{10}$[*p*-value]; two-sided Student's t-test) is shown along the y-axis. **b.)** KO vs. KO. Scatter plot showing fold changes (mean $\log_2$[Δ*gene*/wt], n=3) from all molecules profiled across two Δ*gene* strains. **c.)** KO–KO Correlations. Bar chart indicating phenotypic similarities between a user-selected Δ*gene* strain, and all other profiled strains. **d.)** ΔMolecule Across KOs. Volcano plot showing measured changes in molecule abundance (mean $\log_2$[Δ*gene*/wt], n=3), and statistical significance (-$\log_{10}$[*p*-value]; two-sided Student's t-test) for a single molecule across all profiled Δ*gene* strains

Figure 5.4: .e.) Molecule vs. Molecule. Scatter plot displaying fold changes (mean $\log_2$[Δ*gene*/wt], n=3) for two molecules across all Δ*gene* strains where both molecules were quantified. **f.)** Molecule Covariance Network. Nearest neighbor molecule covariance network showing correlations between molecules covariant with a user-selected molecule. **g.)** Box plots showing measured fold changes for a single molecule acrossl respiration competent (RC) Δ*gene* strains (blue) and respiration deficient (RD) Δ*gene* strains (orange). **h.)** Volcano plot showing average fold change in molecule abundance (mean $\log_2$[RD strains/RC strains]) against statistical significance (-$\log_{10}$[*p*-value, Bonferroni corrected two-sided t-test]) between all RC and RD strains.

A Platform for Codeless Generation of Custom Data Analysis Web Portals

The web tools designed and developed in support of the aforementioned projects have undoubtedly bolstered the data analysis process and public profile of the studies. However, we acknowledge that a great deal of work has gone into building and deploying these web portals. Their construction required extensive knowledge and application of several programming languages (JavaScript, PHP, C#, etc.) and coding libraries (D3.js, Angular.js, underscore.js, etc.), as well as expertise in relational database design (MySQL) which many researchers lack. Collectively, this creates a bottleneck to the construction and publication of similar web resources. In order to extend the benefit which these data exploration solutions can provide, to a broad audience, it is essential that we develop complimentary tools to facilitate the upload, organization, processing, and eventual online visualization of MS-based "omic" data sets. It is imperative that these tools be designed to accommodate

non-programmers, and provide a code-free development environment. Here, we describe the work-to-date on a platform designed to meet these demands. This platform facilitates upload of generic spreadsheets containing processed mass spectrometry results (peak tables) and enables on-the-fly hierarchical organization of data (i.e., grouping of replicate experiments, selection of control experiments, etc.). Following data upload, platform users can select individual visualizations to add to their custom web portal from a menu of options. Based on these selections, a complete webpage is constructed with all associated functionality embedded. These custom web portals can then be shared with collaborators and other researchers—at the discretion of the creator—via a developed user permissions sharing scheme.

The following sections will describe succinctly the process of creating and sharing of custom project web portals using a prototype version of our platform **(Figure 5.5)**. Specifically, we will provide an overview of 1) the data upload process, 2) data processing and database entry, 3) visualization selection, 4) web portal sharing, and 5) developed features and functionality of these codeless web portals.

Data Upload. In order to achieve the overarching goal of our platform—rapid data upload and visualization—it is critical that all uploaded data be stored using a well-defined organizational structure. Storage of data in a predictable manner enables reuse of code for queries and visualizations across projects. Specifically, this

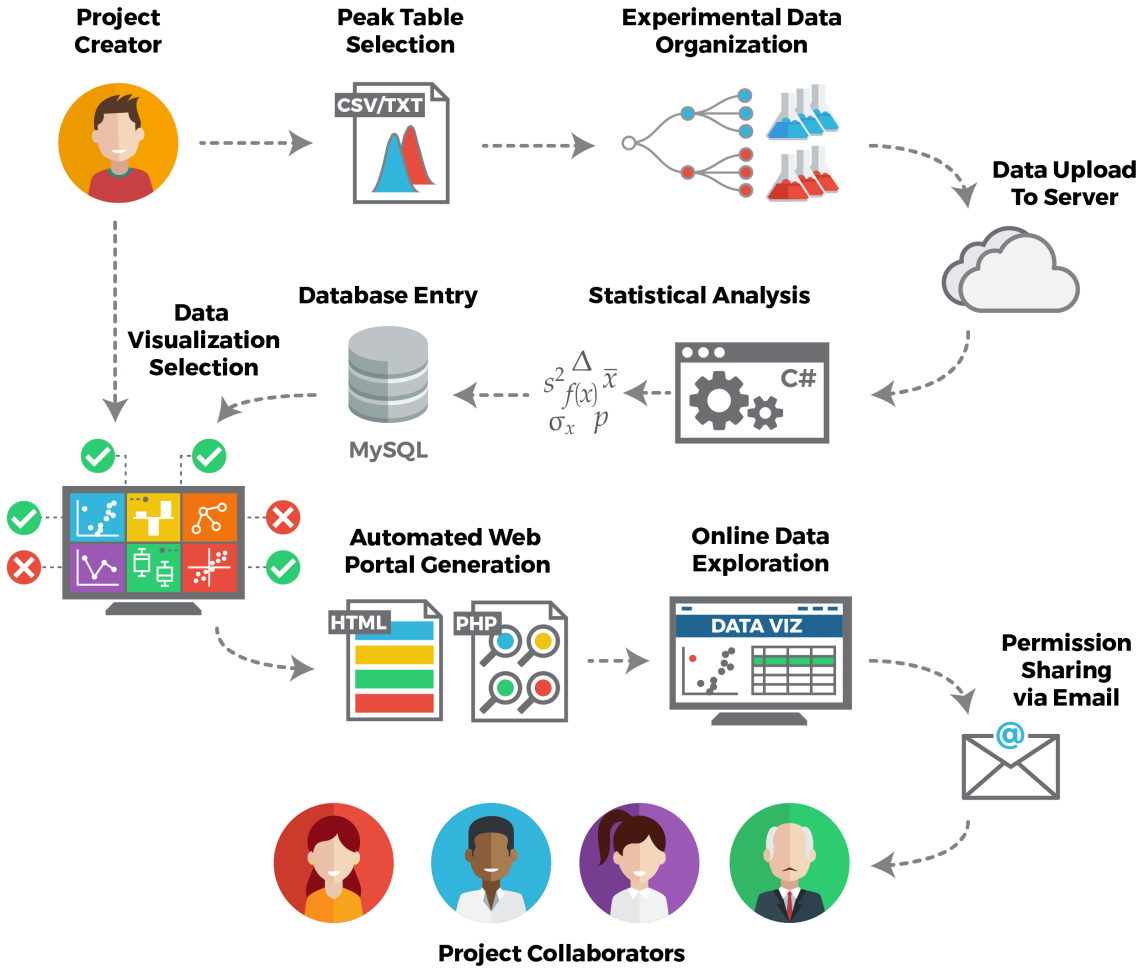


Figure 5.5: Custom data visualization portal creation workflow. An overview of all steps involved in creating a custom data visualization portal using our prototype platform. Users will create individual projects and provide MS data in the form of peak tables, and define a hierarchical organization of that data. These data are uploaded to a remote webserver where they are analyzed through a battery of statistical tests and results are stored in a central MySQL database. The project creator can then select visualizations through which to view these data, and a custom webpage with these interactive plots embedded is automatically generated. The user is then free to explore their data portal and can provide access permissions to collaborators.

characteristic storage will allow creation of functionally similar web portals—built around unique data sets—using identical blocks of code. For our purposes, all data is uploaded to a single central relational database (MySQL) with a static table structure. In order to ensure that user data is uploaded properly to this database, a number of considerations must be made. First, it is necessary that our upload functionality be flexible enough to handle data from a broad array of experiment types (proteomics, metabolomics, lipidomics, etc.). Second, we must account for the hierarchical organization of data and accommodate various experimental designs. Finally, it is essential that we develop this upload functionality to accept data processed using a variety of quantitative software packages.

There exist a large number of publically available software tools for extraction of quantitative information from raw MS data files. For proteomic analysis, software packages such as MaxQuant[15], COMPASS[16], Proteome Discoverer, Skyline[17], and Spectronaut can be used to identify and quantify profiled peptide species, and then aggregate these quantitative values into consensus protein abundances. For metabolomic applications, tools such as XCMS[18], Maven[19], Compound Discoverer, and our in-house high-resolution GC/MS processing suite **(Chapter 3)**, can be used to identify and quantify profiled small molecules. For lipidomic applications, packages such as LipidSearch or TraceFinder can be used to extract similar quan-

titative information. Each of these software tools are algorithmically unique and offer a range of added functionality to the user. However, all perform the same basic task—quantitation of profiled molecules—and produce structurally similar output peak tables. Almost all peak tables produced by MS processing software contain columns corresponding to replicate MS experiments, and rows corresponding to profiled molecules. Within each replicate column, individual cells indicate the abundance/intensity of the corresponding molecule as measured in that MS experiment. Frequently, these peak tables also contain descriptive information about the molecules profiled (molecule names, descriptions, identifiers, etc.). We capitalize on this conserved output structure in our platform, and have designed a file upload architecture which accepts MS data from nearly all MS quantitation software packages.

To upload quantitative MS data, a user will first create a new Project where all project-associated data will be stored. Next, the user will navigate to the 'Data Upload' tab where they can upload generic peak tables in either comma, semicolon, or tab-delimited format. After selecting a peak table of interest, the file is read locally, and all column headers are displayed in the browser. The user is then prompted to categorize individual columns as either 'Unique Identifiers,' 'Feature Metadata,' or 'Quantitative Data' **(Figure 6.1)**.

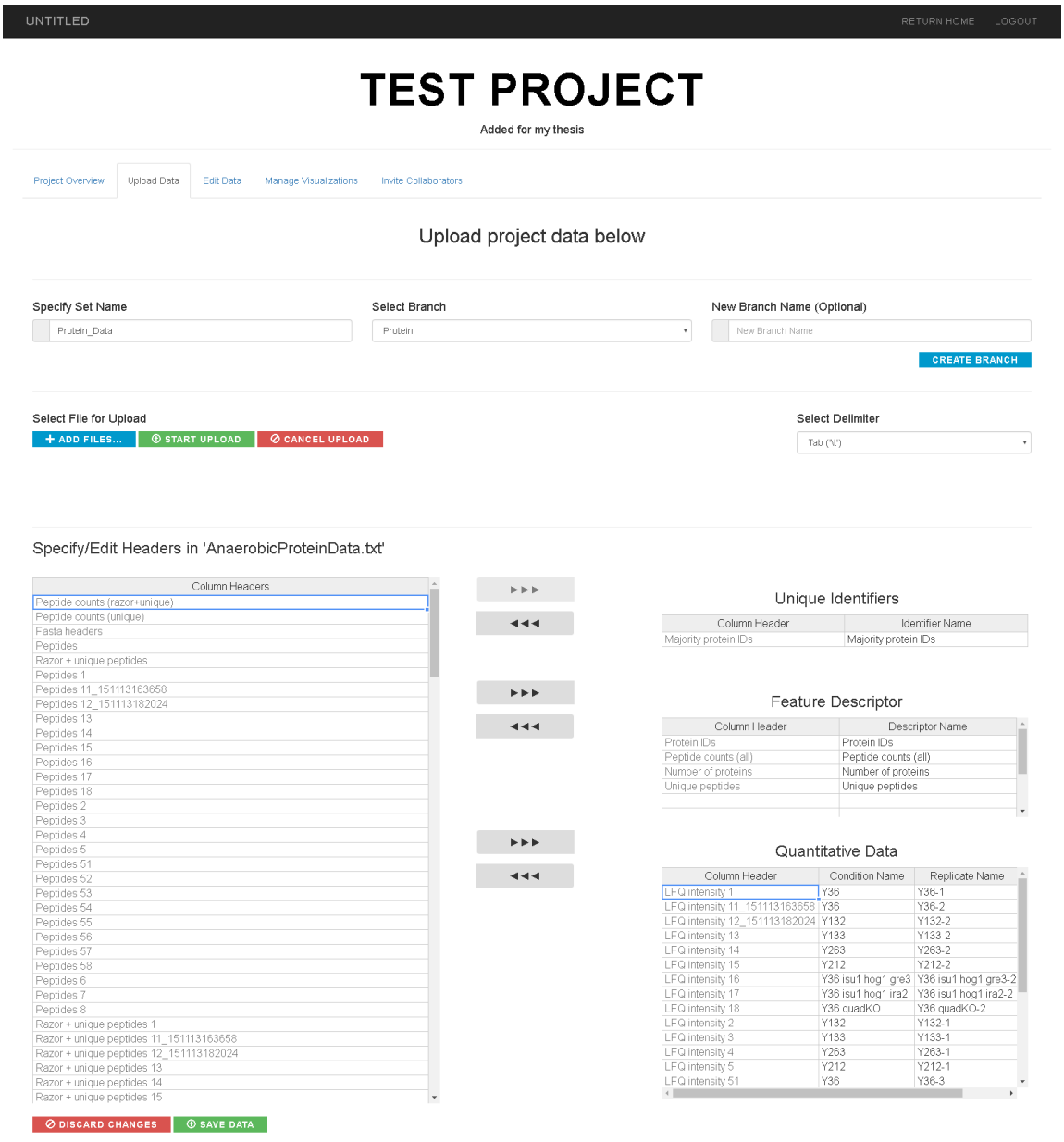


Figure 5.6: Data Visualization Portal–Data upload. Data upload page from our prototype platform. Users select peak tables and all column headers are subsequently read and displayed. Individual column headers can be specified as 'Unique Identifiers,' 'Feature Metadata,' and 'Quantitative Data' by moving them into the appropriate table. Inside the 'Quantitative Data' table, users can provide a replicate name for each column, as well as a condition mapping. Prior to upload, users will select a branch where the data should be added, and provide a name for the current data set.

Briefly, 'Unique Identifiers' are values which uniquely represent profiled molecules inside peak tables. For instance, in proteomic analyses, standard gene names or Uniprot IDs are commonly used to identify distinct molecules profiled across experiments. 'Feature Metadata' columns, are those which contain additional molecule descriptors that would be useful to reference during data exploration in the finished web portal. These columns may contain extended molecule descriptions, aliases, or alternate identifiers, among other values. Finally, 'Quantitative Data' columns contain quantitative measurements of individual molecules from replicate MS experiments. It is anticipated that each column corresponds to exactly one MS experiment.

Users can organize columns by moving the corresponding headers into their appropriate table in the upload data tab **(Figure 6.1)**. Before finalizing the data upload, a few pieces of additional information are requested which will help define a hierarchical organization of the user-provided data that is essential for enabling downstream statistical testing and visualization generation.

We have developed a generic tree-based structure which can be used to describe the organization of MS profiling experiments of all types **(Figure 5.7)**. These generic trees consist of nodes at 5 levels, described here from the top down. The root node or 'Project' node at level 1 corresponds to the project that the user has created. Nodes

at level 2 (‘Branch’ nodes) contain distinct sets of data which will be compared in downstream analyses. For example, a multi-omic profiling experiment may contain separate Branch nodes for protein, metabolite, and lipid data. Nodes at level 3 (‘Set’ nodes) contain all of the data from a single uploaded peak table. Nodes at level 4 (‘Condition’ nodes) contain all replicate data corresponding to a single condition or treatment. For example, if 20 knockout yeast strains were analyzed at the protein level in biological triplicate, 20 ‘Condition’ nodes would be created, each corresponding to one of the knockout strains. It is worth noting that ‘Condition’ nodes can also be specified as controls, which will then be used for normalization purposes in downstream statistical processing. Finally, nodes at level 5 (‘Replicate’ nodes) contain all of the data from a single replicate MS experiment. Using the previous example of 20 knockout yeast strains profiled in biological triplicate, 60 ‘Replicate’ nodes would be created, each of which would be connected to its one parent ‘Condition’ node. We contend that this structure can be used to describe the organizational hierarchy of nearly any MS-based profiling experiment. Furthermore, this design can be exploited to logically organize quantitative MS data into easily queryable data structures.

Once values have been specified for all required fields, the user is allowed to upload their data to a remote server. Before this process begins, the user is

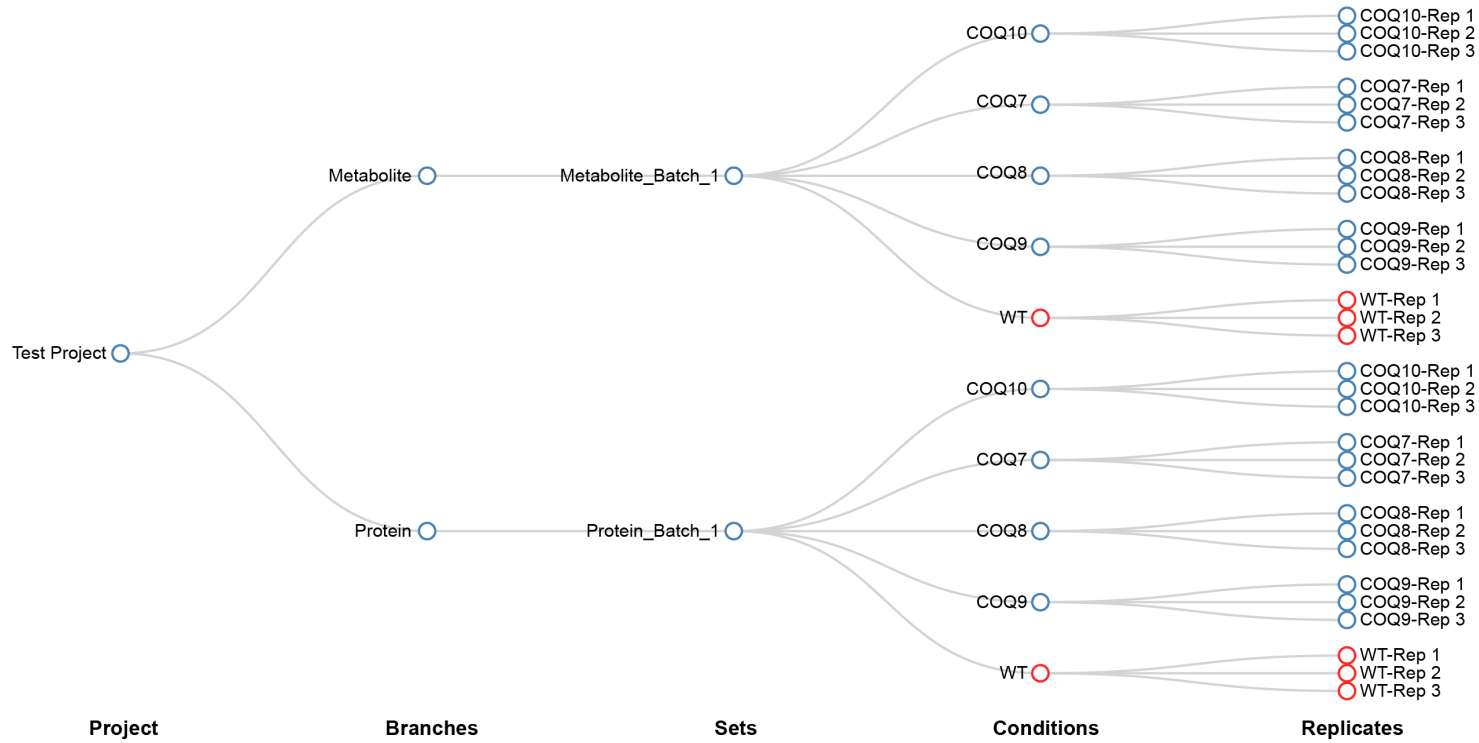


Figure 5.7: Hierarchical data organization. Example hierarchical organization of data. Here COQ7, COQ8, COQ9, COQ10, and WT replicates were analyzed in triplicate at the proteomic and metabolomic level, as part of the project called 'Test Project.' Protein and metabolite data are separated into separate branches, each of which contains one set of data. A condition is created for all strains, each of which has three associated replicates. WT replicates served as an experimental control and are colored red accordingly.

presented with a new dialog showing the hierarchical organization of columns in the provided data file (in tree-form) and prompted to confirm the organization. Upon confirmation, the user-defined column organization is stored in the central MySQL database, and the specified peak table is transferred to the webserver.

Data Processing and Database Entry. At this point in the web portal generation process, a user has created a new 'Project' and uploaded at least one peak table with appropriate organizational information to the webserver. In order to make individual measurements accessible on-command, it is required that the data be entered into a logically-structured database (MySQL is used here) to enable subsequent querying and visualization.

For each uploaded file, entries for all column headers have been stored in a MySQL data table with an indication of whether that column contains 'Unique Identifiers,' 'Feature Metadata,' 'Quantitative Data,' or if it should be ignored during subsequent processing. Each 'Quantitative Data' column is assumed to arise from a single replicate MS experiment. For these entries, the user-provided replicate name along with its associated condition has been stored. Using these information we can extract quantitative values from the user-provided peak table, and insert them into the MySQL database. This task is performed by a client-side script developed in C# .NET, which is executed upon user confirmation of their peak table structure.

The developed script performs two basic functions: addition of unique identifiers and quantitative values, and statistical and descriptive analysis of the uploaded data followed by insertion of results to the database. Together, these processes work to convert user-uploaded data files into a format which lends itself to web-based data accession and visualization. Both processes are described in more detail below.

Insertion of Unique Identifiers and Quantitative Values. For indexing and querying purposes, all unique molecule, replicate, condition, and sets are assigned a unique numerical identifier. These numerical identifiers facilitate expedient reference to, and querying of, specific subsets of data. Further, numerical identifiers are easy to manage across projects contained within a singular database.

First, unique molecule identifiers are read from the user-uploaded file, and added to the database where appropriate. Here, all existing molecule identifiers associated with the project are queried and stored locally. Each molecule identifier contained in the uploaded file is cross-referenced against the list of existing molecule identifiers. The named identifiers which have not yet been added to the database are assigned a unique numerical identifier—incremented from the last identifier added—and inserted into the database. This cross-referencing procedure prevents duplicate entries from being added, and serves to associate measurements across uploaded data files. Next, each named replicate, condition, and set is added to

the database, similarly, with unique numerical identifiers. It is of note that these names are checked on initial data upload to avoid collisions with existing replicate, condition, and set names. Given that we are guaranteed to avoid collisions at this step, no cross-referencing against existing entries is required. Once all molecule, condition, replicate, and set identifiers have been added to the database with appropriate numerical identifiers, the script extracts quantitative data from the uploaded file and inserts it into the database. Each quantitative measurement is added to the database with a reference to its corresponding molecule and replicate identifier. It is also of note that user options to ignore 0, empty, or null values are provided during the file upload step, along with an option to $\log_2$-transform values. This organization of measurements and identifiers in the database enables rapid querying of measurements filtered by molecule name, replicate name, condition name, set name, or any combination thereof.

Statistical and Descriptive Analysis of Uploaded Data. The described database insertion procedure for unique identifiers and quantitative data produces data table entries wherein individual measurements from uploaded peak tables are associated with single replicates. Aggregation of replicate measurements at the 'Condition' level enables calculation of average molecule abundance, standard deviation, and coefficient of variation (CV). Comparison of averaged abundances against control

'Conditions'—or alternatively, against all other replicates in the uploaded data set—enables calculation of fold changes and statistical significance (i.e., *p*-values). These fold change and *p*-values can be further exploited in various computational processes such as unique genotype–phenotype scanning and principal components analysis (PCA). Here we have automated all of these calculations and computational processes.

First, all replicate measurements are grouped together at the 'Condition' level. Using these grouped measurements we calculate and store average abundance, standard deviation, and coefficient of variation. Calculations of fold change and *p*-value are performed two ways for each profiled molecule. First, we utilize a mean-normalization strategy wherein averaged abundances from each condition are normalized to the average abundance considering all replicate measurements. Here, *p*-values are calculated by performing a two-sided Student's t-test (homostatic) using grouped replicate values from each 'Condition' and all other replicate values (non-inclusive) as inputs. In the case that a particular 'Condition' has been indicated as a control, we also calculate control-normalized fold changes and *p*-values. Similarly, average abundance measurements from each condition are normalized against the average abundance of the associated control condition. *P*-values are calculated by performing a two-sided Student's t-test (homostatic) using grouped

replicate values from each condition, and grouped replicate values from the control condition as inputs. In all cases, we adjust calculated *p*-values using the Benjamini-Hochberg FDR and Bonferroni adjustment procedures, to account for multiple hypothesis testing. These adjusted values are stored locally in addition to the unadjusted *p*-value.

Following calculation of descriptive statistics, fold changes, and *p*-values, we also perform outlier (unique genotype–phenotype scanning) and PCA analyses. These analyses (or variations thereof) are frequently employed by systems biologists as they rapidly identify characteristic molecule–condition relationships, and inform functional similarity between profiled conditions and replicates. In both cases, mean-normalized values and control-normalized values (if available) are used as inputs. The genotype–phenotype scanning procedure (as described in **Chapter 4**) produces a Euclidean distance calculation, an outlier condition identifier, and an enum type indicating whether the outlier measurement was increased or decreased from the mean. The PCA procedure is performed twice. First, using averaged molecule fold changes from all 'Conditions', and then using molecule abundances from all replicates. In both cases the analysis produces scaled vectors and variance fractions corresponding to each numbered principal component as outputs. All of these data are stored locally.

Upon completion of all described processing, calculated values are stored in logically organized data tables in the MySQL database. In all cases, columns indicating associated molecule, replicate, condition, and set identifiers are included to expedite queries and minimize the need for complex data joins. These data are now well formatted for the purposes of online visualization. Further, this conserved database structure greatly facilitates the reuse of developed code, as data from multiple projects are stored in exactly the same manner.

User Visualization Selection. At this point in the portal generation process, users have uploaded peak tables complete with associated organizational information, data from these files has been inserted into appropriate MySQL data tables, and results from statistical and descriptive analyses has been stored. All of these tasks were achieved with minimal user interaction. In fact, all that was requested of a user is that they upload a file, organize and appropriately name data columns, and confirm the data organization. We highlight the fact that *absolutely no new code generation* has been required of the user, only simple tasks which can be completed in minutes for modestly sized data sets.

For the purposes of automatic visualization generation, we note that most data plots have well-defined inputs. For example, volcano plots contain data points having a fold change (x-axis) and significance component (y-axis; typically -$\log_{10}$[*p*-

value]). Similarly, bar charts contain bars having numerical (y-axis) and positional components (x-axis), and often error bars showing variance or standard deviation in the numerical component (y-axis). Granted that these, and other, visualizations have standard inputs, we can make them available to users by simply requesting that they specify what values should be used.

Following data upload, users can navigate to the 'Manage Visualizations' tab where a menu of visualization options is presented. This menu displays all of the visualizations which are available to be integrated into a user's custom web analysis portal. Among the offered visualizations are Volcano Plot–Full Perturbation Profile, Bar Chart–Molecule Perturbations, Scatter Plot–Condition vs. Condition, PCA–Conditions, PCA–Replicates, and Outlier Analysis. It is worth noting that these visualizations reflect an early development set which will continue to grow. Those listed were chosen as they are commonly used by researchers to explore "omics" data.

For each listed visualization a description is included along with an input selection dropdown menu and an "on/off" toggle. The input selection list indicates what values the user wishes to use to generate a particular visualization. For instance, under the 'Bar Chart–Molecule Perturbations' input selection the following options are listed: Control-Normalized, Mean-Normalized, and Log_2 Intensity.

Each of these options reflects one way in which changes in molecule abundance across Conditions can be visualized. After selecting the appropriate inputs for desired visualizations, the user will toggle those visualizations to the 'On' state and save their selections by clicking the 'Save Changes' button. Clicking the 'Save Changes' button stores the user's selections in a data table, and triggers an event which automatically builds a new webpage with all visualizations embedded.

Briefly, a new directory is created for each project on the webserver. The files contained in this directory are used to render the web portal in a browser, perform server-side database queries, and display all returned results in the form of interactive visualizations. This compartmentalized structure is desirable as it greatly simplifies restructuring of individual project websites and makes the process of website migration to a new webserver or domain straightforward. The event fired by clicking 'Save Changes' launches an executable (C# .NET) which updates all of the files in the project directory. First, a new webpage is generated by concatenating segments of developed code (stored in the root folder of the webserver) to create a fully functional interface with different visualizations embedded. For instance, if a user selected 'Volcano Plot—Full Perturbation Profile' and 'PCA–Replicates' only code segments required to display these two plots would be added to the new webpage. This concatenation process is extremely quick to execute (<1 sec-

ond) and completely overwrites the existing webpage allowing users to re-select visualizations and recreate custom web portals *ad infinitum*.

In addition to generating a new webpage, individual PHP files containing server-side query commands are updated. For each visualization, the user-selected inputs are specified and PHP files containing relevant commands are updated to reflect the requested data. This process of automatically updating server-side queries simplifies client-side operations by masking the underlying data requests. Further, restricting query results to include only those values desired by the user reduces the volume of data returned by a single query and bolsters performance. Additionally, preparing and storing all MySQL queries on the server-side of the application affords an added layer of security in prevention against SQL injection attacks.

Immediately after these text files are updated in the project directory the new web portal can be explored by the user. As mentioned, this portal can be updated to include different visualizations, and tailored specifically for the user's project.

Web Portal Sharing. At this point in the web portal creation process, the user has uploaded data with associated organization information, all data has been processed and added to appropriate data tables, the user has selected visualizations (and inputs) of interest, and a new web portal has been created. Again, we note that all of this has been achieved by requiring a user to perform simple tasks, and

no new code has to be generated on their part. At this point, the only user who can view the newly generated web portal is the creator. This user also has the exclusive ability to grant collaborating users access to the project's portal.

Briefly, we have designed a permissions scheme consisting of three levels. Level one (read-only access) provides users with the ability to view developed web portals and download associated data. Level two (read/edit access) provides users with all level one permissions, in addition to the ability to add, remove, and edit uploaded data. Level three (project owner access) provides users with all level one and two permissions, in addition to the ability to grant collaborating users access to the web portal, as well as the power to delete the web portal entirely. These permissions were designed to afford the appropriate amount of control to individual users so they can effectively utilize and share these tools while maintaining privacy and security of their data.

By default, the only user who is initially granted access to a project's web portal—other than administrators—is the creator. The creator is the default 'Project Owner' which grants them all of the permissions listed above. Project owners are the only users who have the ability to provide access to other collaborators, affording complete discretion over who can utilize their data. Project access can be managed under the 'Manage Visualizations' tab. Upon navigating here, users can enter

collaborator email addresses along with a short message, and specify an appropriate permission level for the collaborator in question. Clicking 'Send Invite' triggers a series of events culminating in invitations being sent to the specified collaborators. In this process, a unique 20-digit alphanumeric code is randomly generated (there exist ~7.04×10^{35} possible combinations) for each invited collaborator and stored in the central database, along with information about the project and the project owner-specified permission level. Then, each collaborator is automatically emailed their specific 20-digit alphanumeric code, along with the owner's message and a link to the website. After navigating to the site and either logging in (for current users) or creating an account, invited users can select 'Accept Invitation' where they are prompted to enter their emailed 20-digit code. This code is then checked against the data table containing all sent codes for a matching entry. If a match is found, the user is granted access to the associated project—with the appropriate permissions—and the project is added to a list of web portals which they are able to interact with and explore.

Developed Features and Functionality. Collectively, the previously described processes have enabled the creation of project-specific web portals without requiring any coding on the part of users. All user tasks are straightforward and can be performed in a web browser without requiring the download or installation of

any software tools. We contend that this platform will have widespread utility for mass spectrometrists and other biological researchers who routinely interact with MS-based "omics" data. To make these developed web portals as useful as possible we have developed a number of unique features and functionality including interactive visualizations, informative tooltips, real-time chart editing, data and chart downloads, and data table lookups. These developed features are described briefly below.

Interactive Visualizations. All user-selected visualizations have been created using the D3.js JavaScript library[11]. D3 (Data Driven Documents) is a well-developed and well-supported library for displaying data in the form of interactive charts in web applications. This library is widely used and has a strong community following which is a valuable resource that makes it easier to develop, maintain, and debug code for displaying data. All developed visualizations are designed to be highly interactive and users can easily select data to visualize by choosing options from lists of valid inputs. Animations are employed when transitioning between data sets for all visualizations, and data point indexes are used to maintain object consistency (i.e., the same data point represents the same molecule, condition, etc. before and after a transition). This object consistency provides useful visual cues to the user about how individual measurements differ between selected conditions,

for instance.

Informative Tooltips. All developed visualizations support informative tooltips. Within each plot, hovering over an individual data point displays a tooltip which contains relevant metadata about the molecule, measurement, condition, etc. in question. These tooltips enable users to rapidly identify data points of interest and retrieve more information about that measurement. In contrast to static plots, this display of tool tips greatly increases the amount of information which can be transferred to a user per unit time, expediting data analysis.

Real-Time Chart Editing. All visualizations were designed to employ differential coloring schemes and data filters. For instance, generated volcano plots use a three color scheme to indicate significance and fold change. For these plots, we have built controls for dynamically resetting fold change and p-value cutoffs which will automatically recolor data points in the chart appropriately. We have also employed chart filters to selectively display only those data points which meet certain fold change and p-value thresholds. Additionally, we have included an option to set fixed scales to axes in generated plots. Changes to these settings are automatically rendered in the associated visualization providing users an additional level of control over their data.

Data and Chart Downloads. We acknowledge that for communicating data to other scientists—either in communications or publications—it is often helpful to include individual plots. To make web-portal generated plots more useful for researchers we have added an option to download all data associated with a particular plot. Here, by clicking a button a user can download a tab-delimited file containing all of the numerical and textual information used to generate a particular visualization. From this file, users can port data into their preferred plotting software package and generate new graphs for sharing. Additionally, we have built in an option to download generated plots directly in .SVG format. This file can be imported directly into graphics editors (such as Adobe Illustrator or Microsoft PowerPoint) and each plotted object can be manipulated to adjust color, size, opacity, etc. to the user's liking.

Data Table Lookups. While visual display of data is highly useful, it is often advantageous to explore data in numerical format. Within each developed web portal a data table lookup option is provided by default. Individual query terms corresponding to all project-specific replicates, conditions, and profiled molecules have been created and stored in the database. By selecting any one of these query terms, all associated data is returned and displayed in tabular format. This data table is interactive and supports filtering, sorting, column hiding, and pagination

of results to make it easier to navigate to individual pieces of data. All queried results can also be downloaded in tab-delimited format for plotting and analysis at the discretion of the user.

Conclusions and Future Directions

Mass spectrometry has positioned itself as one of the premier tools for interrogating biological systems and elucidating novel biochemical insight. More and more, large-scale MS studies are being undertaken to answer increasingly complex biological questions. It is imperative that we continue to develop software solutions to meet the analysis needs of these experiments. Online data analysis and exploration tools are an attractive solution, but remain underdeveloped. Both the Medicago Proteome Compendium and Y3K Project web portals have enabled researchers to dig deeper into these massive data sets, and have brought them to larger biological community in a format which is readily accessible from any internet browser.

There are undoubtedly great challenges posed by data curation, web portal creation, and deployment of developed web tools which create a bottleneck for researchers without advanced programming skills. However, these online tools can greatly expedite data analysis and enable researchers to investigate their results at greater depth. Our work on a codeless web portal generation platform is a step

in the right direction. Continued development of tools designed for use by non-programmers widens the aforementioned bottleneck and will make online data analysis a much more tractable option for the entire research community.

As we continue development on our platform we intend to build in additional functionality including gene ontology (GO) enrichment, coexpression analysis, and hierarchical clustering. We aim to bolster the overall web design to make exploration and analysis more intuitive for all users. We recognize that data security and privacy are concerns for all scientists, and, as such, must ensure that all data is well-protected while remaining accessible to collaborators worldwide. We note that increasingly, researchers are analyzing data on numerous devices including mobile and tablet and it is critical that we support these platforms. Finally, we believe it is absolutely essential that we promote data sharing between individuals so that multiple data sets can leveraged against one another to synergistically enhance the insight afforded. Our platform is uniquely positioned in that we host multiple data sets from a singular location making inter-project data comparison a very manageable option.

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, pp. 339–347, 2014.

[2] 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, pp. 701–714, 2015.

[3] W. Kent, C. Sugnet, and T. Furey, "The human genome browser at UCSC," *Genome Biology*, pp. 996–1006, 2002.

[4] G. Dennis and B. Sherman, "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 9, 2003.

[5] B. Giardine, C. Riemer, and R. Hardison, "Galaxy: a platform for interactive large-scale genome analysis," *Genome Biology*, vol. 15, pp. 1451–5, Oct. 2005.

[6] J. Severin, M. Lizio, J. Harshbarger, H. Kawaji, C. O. Daub, Y. Hayashizaki, N. Bertin, and A. R. R. Forrest, "Interactive visualization and analysis of large-scale sequencing datasets using ZENBU.," *Nature biotechnology*, vol. 32, pp. 217–9, Mar. 2014.

[7] T. T. Xiao, S. Schilderink, S. Moling, E. E. Deinum, E. Kondorosi, H. Franssen, O. Kulikova, A. Niebel, and T. Bisseling, "Fate map of Medicago truncatula root nodules.," *Development (Cambridge, England)*, vol. 141, pp. 3517–28, 2014.

[8] A. C. Timmers, M. C. Auriac, and G. Truchet, "Refined analysis of early symbiotic steps of the Rhizobium-Medicago interaction in relationship with microtubular cytoskeleton rearrangements.," *Development (Cambridge, England)*, vol. 126, pp. 3617–3628, 1999.

[9] H. Marx, C. E. Minogue, D. Jayaraman, A. L. Richards, N. W. Kwiecien, A. F. Sihapirani, 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*, Oct. 2016.

[10] 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*, pp. 1–11, Sept. 2016.

[11] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

[12] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. a. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional Discovery via a Compendium of Expression Profiles," *Cell*, vol. 102, pp. 109–126, 2000.

[13] P. Kemmeren, K. Sameith, L. A. L. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. Van Wageningen, C. W. Ko, S. Van Heesch, M. M. Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. C. H. Brabers, A. J. Miles, D. Bouwmeester, S. R. Van Hooff, H. Van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. Van Leenen, M. J. A. Groot Koerkamp, and F. C. P. Holstege, "Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors," *Cell*, vol. 157, pp. 740–752, 2014.

[14] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel,

M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. a. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, "Functional profiling of the Saccharomyces cerevisiae genome.," *Nature*, vol. 418, pp. 387–391, 2002.

[15] 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, pp. 1367–72, 2008.

[16] 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, pp. 1064–1074, 2011.

[17] B. MacLean, D. M. Tomazela, N. Shulman, M. Chambers, G. L. Finney, B. Frewen, R. Kern, D. L. Tabb, D. C. Liebler, and M. J. MacCoss, "Skyline: An open source document editor for creating and analyzing targeted proteomics

experiments," *Bioinformatics*, vol. 26, pp. 966–968, 2010.

[18] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Analytical Chemistry*, vol. 78, pp. 779–787, 2006.

[19] M. F. Clasquin, E. Melamud, and J. D. Rabinowitz, "LC-MS data processing with MAVEN: A metabolomic analysis and visualization engine," *Current Protocols in Bioinformatics*, 2012.

Chapter 6

THE YEAST CONTROLLER: A WEB-BASED QUALITY CONTROL TOOL FOR MONITORING PERFORMANCE OF LC/MS SYSTEMS

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

Kwiecien NW, Brademan DR, Hebert AS, Westphall MS, Coon JJ. *A Web-Based Quality Control Tool for Monitoring Performance of LC/MS Systems.* **2016**.

Introduction

Proper maintenance and monitoring of liquid chromatography-coupled mass spectrometry (LC/MS) system performance is critical for enabling high-throughput profiling in cutting-edge MS research labs. Instrument downtime and time spent troubleshooting MS performance issues are undesirable, as these periods preclude the collection of valuable experimental data. The process of identifying the root cause of suboptimal instrument performance and subsequent restorative maintenance can be burdensome, particularly when the source of issues is obscure. Routine analysis of quality control (QC) samples is useful for providing continual snapshots of instrument performance, and for signaling declines in expected operation.

In our own lab, we regularly analyze tryptic digests of whole yeast proteomes ('yeast controls'), to track performance of dedicated proteomic profiling LC/MS systems. Each yeast control data file is searched against a target-decoy protein database[1] and the number of unique peptides identified (at 1% FDR) is reported as a metric of performance. Granted that identification of as many unique peptides as possible per experiment is one of the overarching goals or proteomic analysis, this simple value is useful for assessing overall system performance. However, this single metric fails to capture many components which contribute to and influence

system operation. For instance, no information about peptide elution peak widths, systematic mass error, or the distribution of peptide precursors throughout the chromatographic separation can be ascertained from the number of unique peptides alone. This information would be useful to have during times of lowered instrument performance, as deviations here can be used to diagnose specific problems.

For our own QC purposes, individual yeast control data files are manually searched by instrument operators, and unique peptide identification counts are logged in an informal record. Although this process is tedious for users, it has proven useful in helping to maintain consistency of LC/MS performance. It is of note that the described data analysis routine is completely static, and therefore, opens the door to automatic processing. Although a number of automated QC data analysis tools for monitoring LC/MS performance have been developed[2–6], none have been widely adopted by the proteomics community. Ideal QC data processing software would completely eliminate the need for hands-on analysis following data collection. These software solutions should maintain historical records of all processed QC data files, facilitate rapid data lookups and visualizations, and support comparisons between QC data files across instruments and even labs. Furthermore, it is desirable that these tools employ intelligent algorithms to assist in diagnosing instrument issues during times of diminished performance.

To bolster the means by which we analyze, store, and utilize yeast control data in our own lab we have constructed a web-based data deposition, processing, and visualization tool—The Coon Lab Yeast Controller. This web-based tool supports drag-and-drop uploads of raw MS data files, performs all data analysis operations, and logs numerous metrics of performance in a central database. All uploaded data can be immediately visualized within a web-browser through a convenient dashboard interface following processing. The Yeast Controller affords a substantial time-savings to instrument operators, keeps an accurate historical record of individual instrument performance, and can be used to more rapidly identify errant trends to streamline troubleshooting. The Yeast Controller has been designed specifically with our lab in mind and tailored to our QC needs. That said, the basic framework developed here should be applicable to QC procedures employed by many outside MS labs.

Design and Functionality

The Yeast Controller is comprised of three major components: 1.) a front-end web interface for data deposition and visualization, 2.) a back-end database containing all instrument information, user information, and processed QC analysis results, and 3.) a set of software scripts to handle processing of user-provided raw data

and insertion of results into the database. Here we will describe the design and capabilities of this web-based tool with a focus on data upload, data processing, and data visualization.

Data Upload. The Yeast Controller uses a central back-end database (MySQL) for storage of all processed QC results, in addition to information about individual MS systems and laboratory instrument users. This design is convenient as it stores all data in a single location which can be accessed by multiple users concurrently. Within this database, a unique 'Instrument' profile has been added for each of our dedicated proteomic analysis LC/MS systems (one LTQ Velos, one Orbitrap Elite, and four Orbitrap Fusion Lumos systems) along with a set of platform-optimized data processing parameters. All lab members who conduct proteomics experiments and analyze yeast controls are invited to create individual user profiles. These profiles are password-protected and allow for upload of QC data files to specific instruments.

To upload raw QC data files, users login to the Yeast Controller portal and navigate to the 'Data Upload' page **(Figure 6.1)**. Here, users will first select the instrument which generated the data file from a drop down list. In an effort to prevent uploaded files from being improperly associated with the wrong instrument, we provide instrument-specific upload privileges on a user-by-user basis. Users are

then prompted to enter information about chromatographic separation methods. No information about the employed LC method is stored within a Thermo raw data file, but this information is useful for troubleshooting performance issues. Users can create and save new LC methods by clicking the 'Create New Chromatographic Method' button **(Figure 6.1)**, providing their method a unique name, and entering the following information: LC model, column length, column temperature, packing material, particle diameter, inner diameter, gradient length, flow rate, buffer A composition, and buffer B composition **(Figure 6.2)**. We note that these methods often will not change which eliminates a source of variation and is optimal for temporal monitoring of instrument performance. These user-created methods are added to a drop down list of LC methods which are available to all Yeast Controller users.

After selecting an instrument and LC method, users can drag-and-drop Thermo raw data files into the browser window, which are automatically populated in an on-screen list. We provide a 'Send Email on Completion' option which triggers an automatic email alert to the user upon processing completion. This message reports total processing time as well as relevant metrics of performance (identified peptides, proteins, etc.) Upon clicking 'Start Upload' the file is transferred to a local web server and a new entry is added to a processing queue indicating the complete

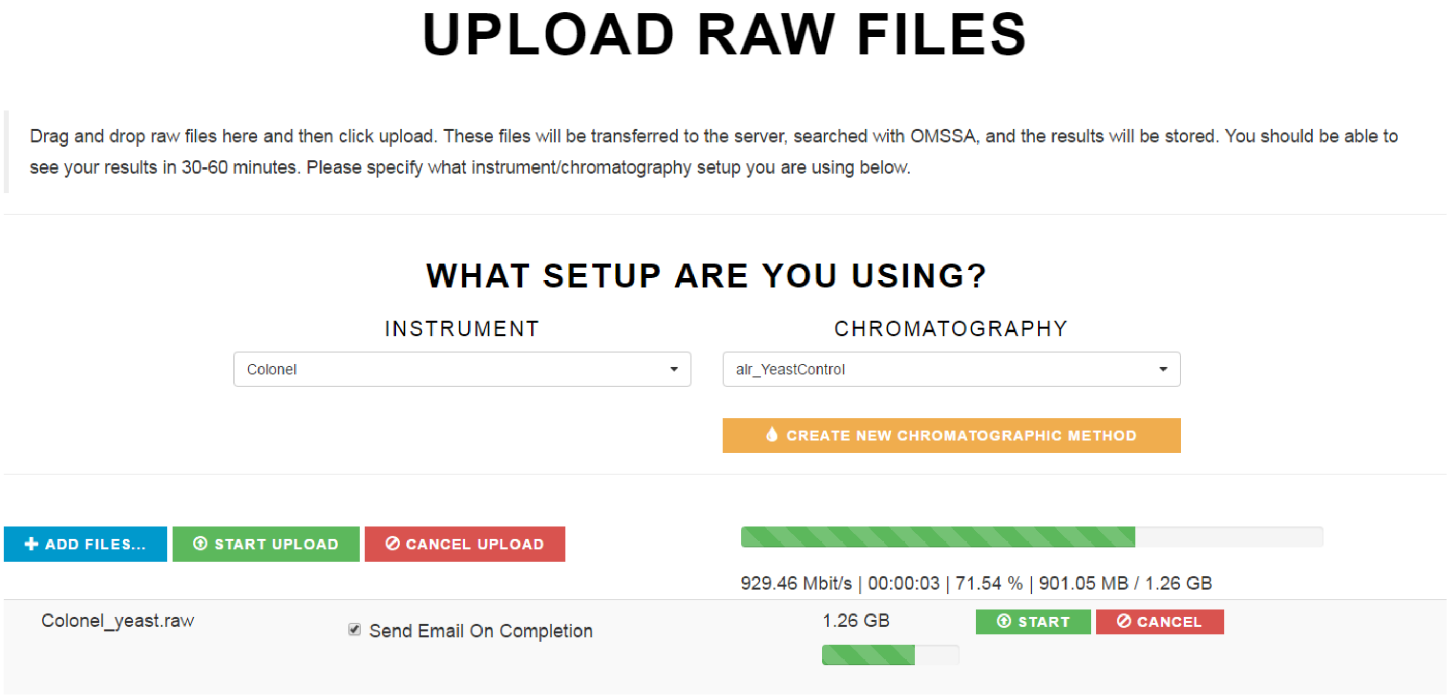


Figure 6.1: Yeast controller data upload. The data upload form which facilitates transfer of QC data files (Thermo .raw) to a local server for automated data processing. Users will indicate the instrument used for QC data acquisition, and the employed chromatographic method from the dropdown lists prior to upload. Drag-and-drop file selection is supported, and all provided files are listed in a tabular format. Users can choose to receive email notifications containing QC analysis results upon processing completion by checking the associated box. Upload progress is reported in real-time during file transfer, as shown here.

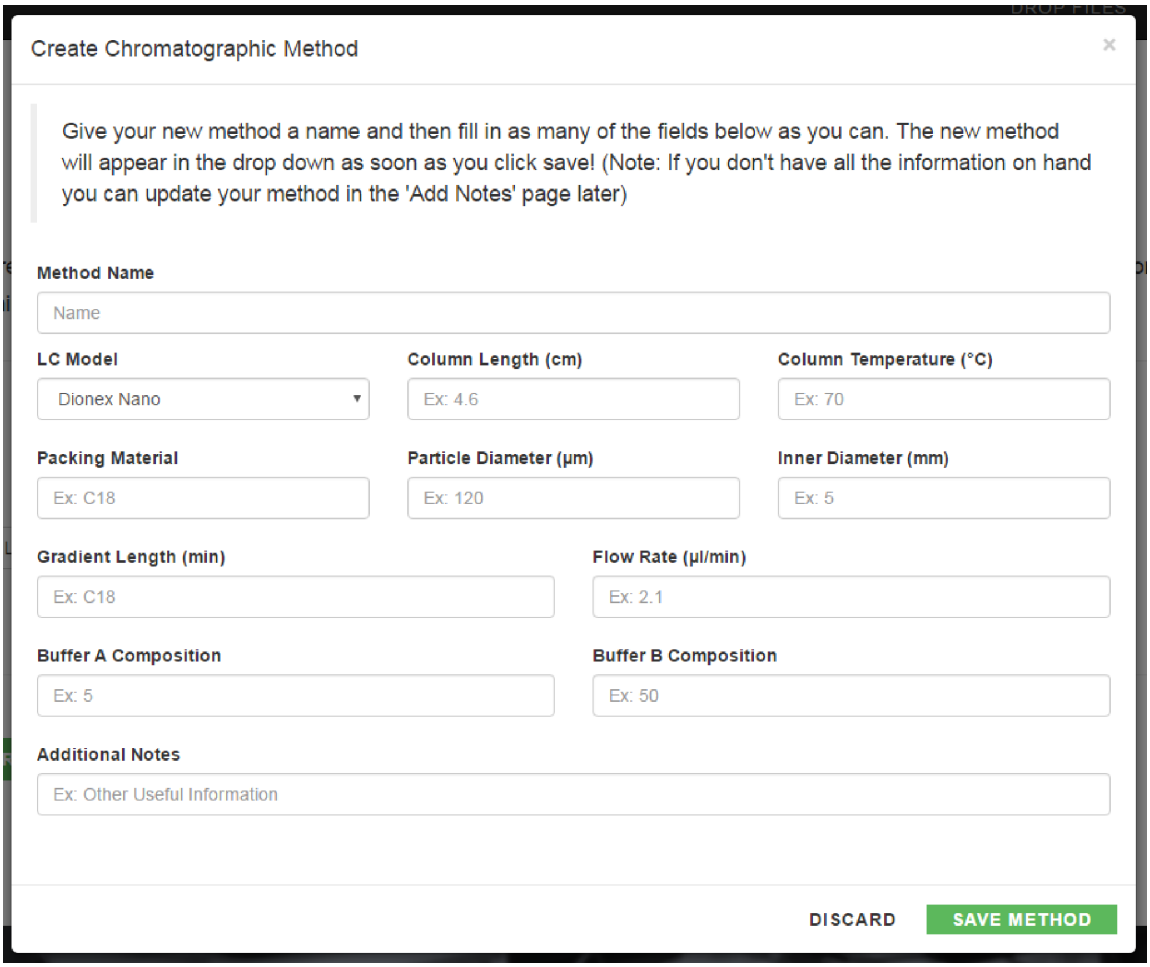


Figure 6.2: Chromatographic method creation form. This form is used to create new chromatographic method entries in the central Yeast Controller database. These methods are listed in the data upload page, accessible to all users, and can be associated with individual QC experiments prior to file upload. The included fields represent all information needed to adequately describe an employed LC method.

file path, instrument of origin, and uploading user. This entry signals to server-side processing scripts that a new file has been uploaded, and needs to be analyzed.

Data Processing. The traditional steps in processing a raw QC file are as follows: format conversion (.raw to .dta), peptide-spectral matching against a concatenated target/decoy protein database done using OMSSA[7], result filtering to a 1% FDR level, aggregation of peptides into protein groups, and extraction of chromatographic peak width information from all identified peptide species. The Yeast Controller has been designed to perform all of these processes automatically and sequentially without requiring external input from users.

As mentioned previously, a new entry is added to a process queue for each uploaded QC file. A background script (C# .NET) runs continuously on the local webserver and checks this queue for new entries every 5 seconds. Once a new entry is found, MS^2 spectra are immediately extracted and written to DTA files (traditional SEQUEST data file format which is compatible with the OMSSA search algorithm) using the DTA Generator program in COMPASS[8]. We split all MS^2 spectra across a number of DTA files, each containing 2000 individual spectra. This enables parallel searching of smaller sets of MS^2 spectra—with OMSSA distributed across multiple CPUs—which reduces overall processing time, relative to searching all MS^2 spectra in a single operation. Our lab has a private cluster consisting of 17 nodes (read,

CPUs) which are dedicated for processing of OMSSA searches. These 17 nodes are also connected to a larger campus-wide distributed computing network as a part of the HTCondor program. HTCondor is a system which monitors CPU usage from computers around campus and recruits inactive machines for processing of computational tasks[9,10]. We utilize the HTCondor computing network for our data searching processes which affords substantial performance benefits.

After all DTA files are written, each is submitted as a single job to HTCondor for database searching against a concatenated target-decoy yeast protein database using OMSSA. Upon completion, each task will return a file containing database search results, and the job will be removed from a list of submitted tasks. To avoid race conditions after each job is returned (i.e., a file read begins before the file write has completed), the size of each returned file is monitored to check for completion of file writes. After each returned file's size is observed not to change for a period of at least two minutes, it is assumed that the task and file write have both successfully completed.

All returned result files are merged into a single CSV file containing scored PSMs from both target and decoy peptide sequences. This file is moved to a new time-stamped directory on the webserver which holds the raw QC data file, and will store processed results. The combined CSV results file is then filtered to a

1% FDR level using the FDR Optimizer program from COMPASS. This program exports a list of PSMs and parsimony peptides (target and decoy) identified at 1% FDR, which are similarly stored in the QC file directory. Identified peptides are then aggregated into protein groups using the Protein Hoarder program, also a part of the COMPASS suite. This program outputs a list of consensus protein groups in addition to a summary file containing select figures of merit.

Maintenance of LC/MS system performance also requires monitoring of the liquid chromatography system. To provide users with diagnostic information about LC performance, we report descriptive statistics about peptide elution peak widths. To do so, we have developed an in-house algorithm for extracting chromatographic features from identified peptides using characteristic *m/z* and RT. For each peptide where a chromatographic elution profile can be successfully extracted, the peptide's elution time at full width half maximum (FWHM), apex S/N, and apex RT is written to a file stored in the QC file's target directory.

At this point in processing all relevant data searching and extraction routines have been completed and associated result data written to files. In order to make this information queryable and available for comparative analyses, it must be added to the central Yeast Controller database. A comprehensive entry is made for each QC file containing the following information: file ID (unique identifier for the

QC file), instrument, uploading user, time collected, gradient length, PSM count, peptide count, protein count, MS^1 count, MS^2 count, average PPM error, standard deviation in PPM error, average protein sequence coverage, standard deviation in protein sequence coverage, median apex S/N, standard deviation in apex S/N, and a five-point summary of peak widths. An entry for each detected peptide is also added to the database with the following information: file ID, MS^2 scan number, sequence, mass, charge, measured *m/z*, theoretical *m/z*, ppm error, apex S/N, apex RT, and peak width. We have found in many instances that it is a useful practice to compare distributions of values against historical distributions to check for characteristic deviations. The following distributions are stored as ordered strings in four separate entries for each QC file: peptide *m/z* (25 Th bins), peptide RT (3 minute bins), peptide S/N (1 unit bins), precursor *m/z* values (25 Th bins). This preprocessing step negates the need for an on-the-fly query/binning procedure which is a boon to visualization performance.

After all data has been inserted to the central Yeast Controller database it can be immediately queried and visualized in-browser. At this point, if a user has requested an email notification, a message is sent containing of the number of identified peptides and proteins, as well as an overall performance metric—percentage of maximum peptides detected historically. This alert system was developed for con-

venience of users. It enables researchers to go about other tasks during processing, but still receive up-to-the-minute information about LC/MS system performance.

At the time of this writing, the Yeast Controller database contained QC information from 597 raw data files which yielded 43,650,411 individual MS data scans, and 9,844,986 peptide spectral matches. It is also of note that no raw QC data files have been removed from the server. MS data file sizes are rapidly increasing and storage of hundreds of raw files requires ample storage space. However, keeping these files intact allows new data analyses to be added in the future, which will be immediately bolstered by availability of a historical data set.

Data Visualization To enable rapid accession and comparison of current and historical QC data, we have built a front-end web-based data visualization portal for exploring these data. This web portal was constructed using the Twitter Bootstrap framework and integrates interactive visualizations built using the D3.js JavaScript library[11]. Our tool presents all data through a convenient dashboard style interface where instrument-specific QC data can be explored openly by all members of the Coon Research lab **(Figure 6.3)**.

The visualization dashboard displays relevant figures of merit—peptides identified, proteins identified, and % peak performance (current peptide count/highest peptide count)—from the most recent QC file processed for each lab instrument at

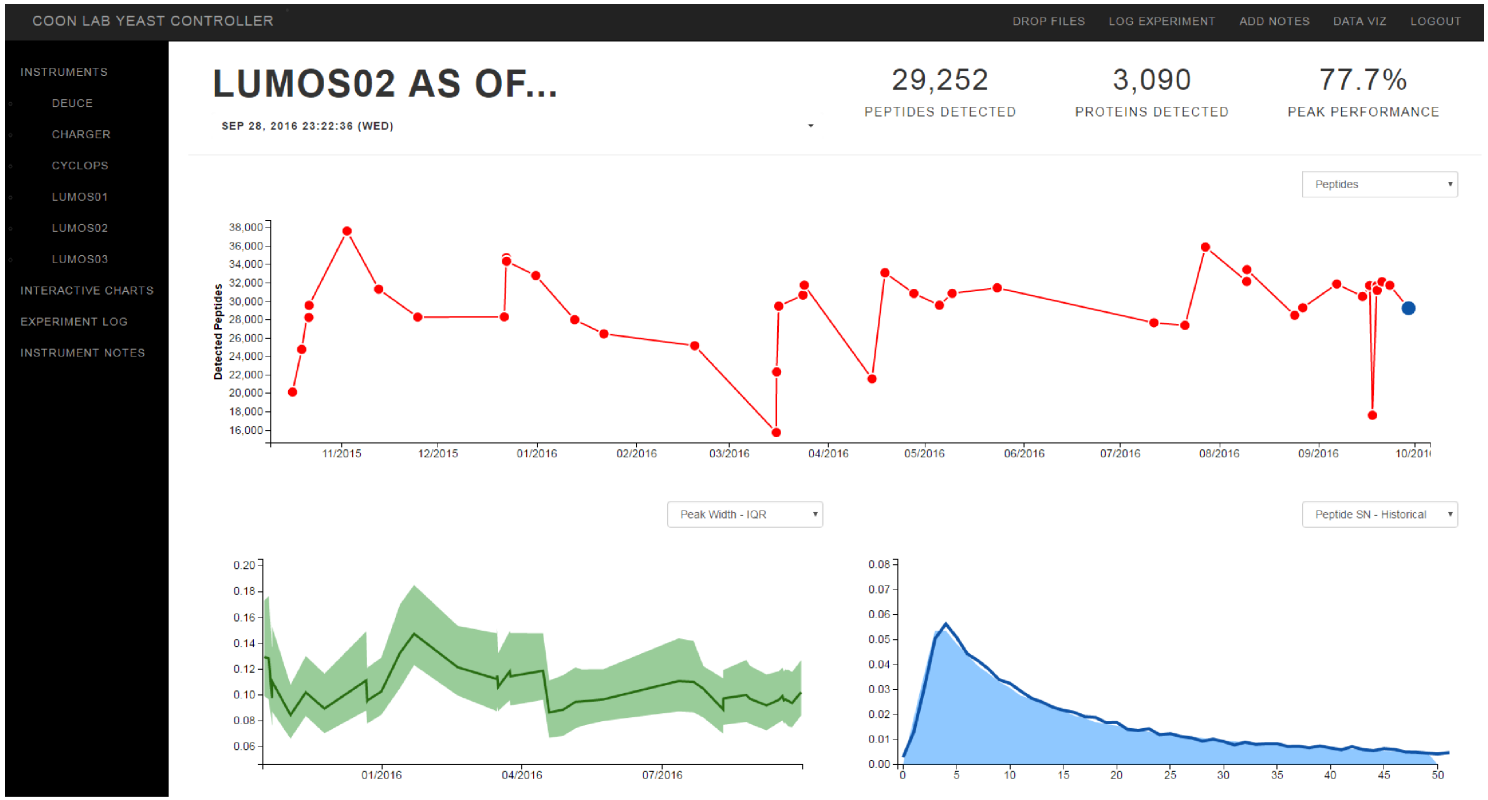


Figure 6.3: Data visualization dashboard. The dashboard interface used to explore all Yeast Controller QC data. All data associated with an instrument can be viewed by selecting the appropriate option from the left-hand menu. For each instrument, performance metrics (peptide identification count, protein identification count, %of peak performance) from the most recent QC data file are displayed at the top of the page. Three additional plots are provided. These plots can be used to view historical records of performance metrics, and to compare data distributions from individual QC files against previously acquired files meeting certain performance thresholds.

the top of page on login. Three separate graph panes below are used to simultaneously display different performance data, each of which can be adjusted by selecting options from a dropdown list **(Figure 6.4)**. The largest pane at the top of the page shows a running plot of either peptide identifications, protein identifications, MS^2 scans acquired, MS^1 scans acquired, or average ppm mass error by date of QC file acquisition. By hovering over individual data points, users can view tooltips which include added data about the corresponding QC file. By clicking these data points, users can select individual QC runs which they wish to retrieve more data from.

The lower right-hand pane displays current distributions relative to historical distributions **(Figure 6.5)**. These historical distributions consist of data from QC analyses collected during times of high system performance (90+% of all-time peak performance). The utility in this view is that it enables users to compare data from singular files against data from multiple files collected when instruments were performing as expected. This view makes visual detection of deviations in trends away from what is expected much easier, and greatly facilitates localization of system performance issues. Finally, the last pane can be used to display historical distributions of peptide peak widths **(Figure 6.6)**. Here, average peptide peak (FWHM) is shown along with the IQR of all peptide peak widths. This is a useful metric to monitor as increases in peak width can indicate systematic peak

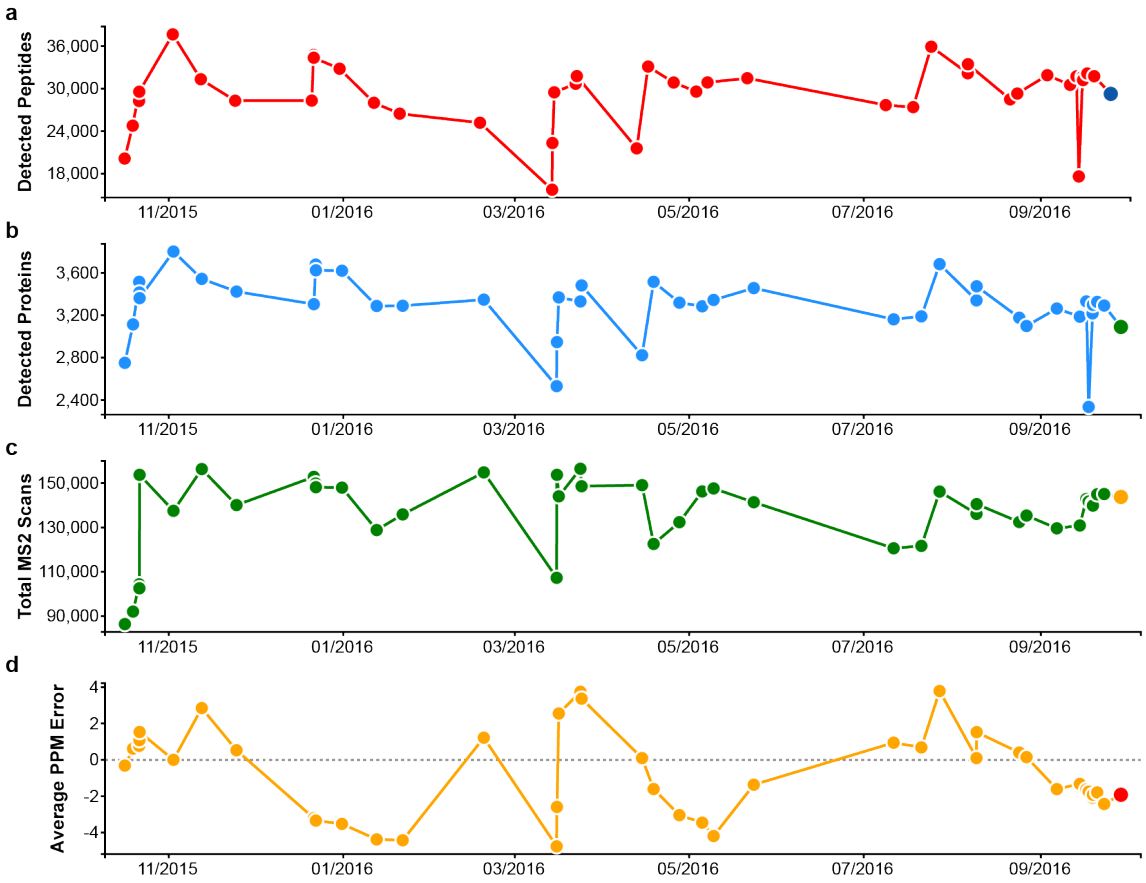


Figure 6.4: Historical record of QC performance metrics. Four example graphs showing QC performance metrics over time from data files collected on a single instrument. These graphs can be automatically generated without refresh in our data visualization dashboard by selecting options from a dropdown. Here, **a.)** peptide identification counts, **b.)** protein identification counts, **c.)** MS^2 scans acquired, and **d.)** average ppm mass error from each QC data file are displayed along the y-axis, with file acquisition date shown along the x-axis.

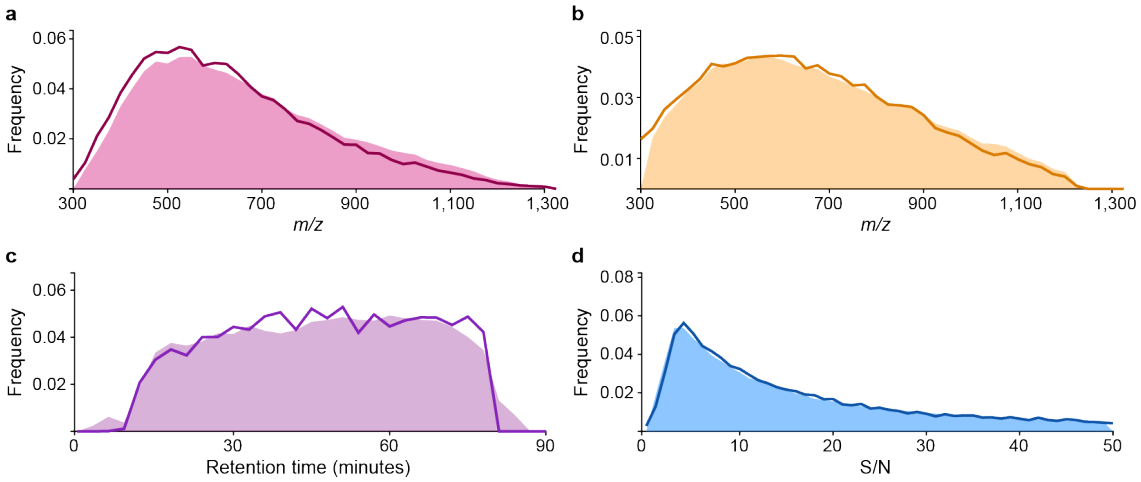


Figure 6.5: Comparison of QC data distributions. Four example distribution comparisons, which can be automatically generated in our dashboard interface, are shown. Here the distributions being compared consist of values taken from a single QC data file (dark line) or values taken from all QC data files where peptide identification counts were at least 90% of the highest recorded count (light fill). Frequency distribution comparisons of **a.)** peptide *m/z* values, **b.)** MS^2 precursor *m/z* values, **c.)** peptide RT, and **d.)** peptide S/N are shown here.

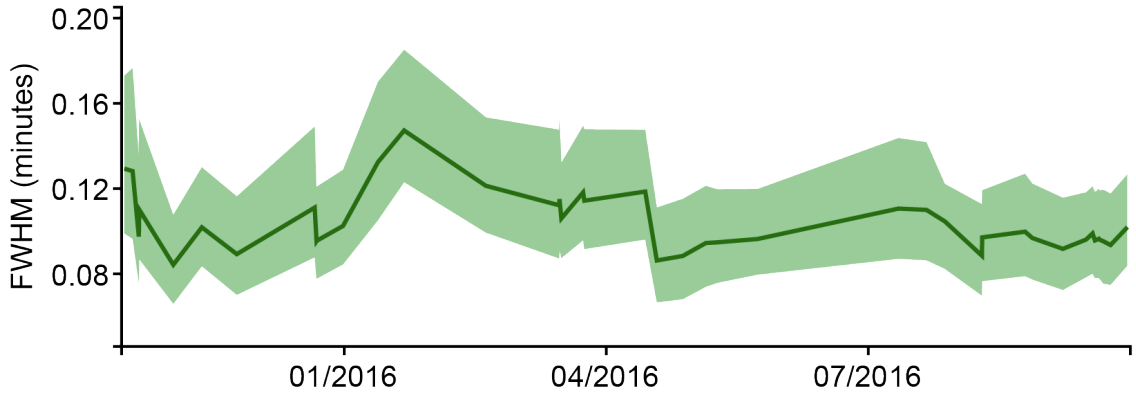


Figure 6.6: Chromatographic peak width display. Display of peptide peak widths across a set of QC data files collected on a single instrument. For each QC file, average peptide peak width (FWHM) is included as a point along the dark green line, and the IQR of all peptide peak widths is displayed in the light green band. These values are plotted against file acquisition date, which is shown along the x-axis.

broadening and potential column failure.

To provide an added level of user control, users are allowed to delete files which they have uploaded. Occasionally, users will recognize substantial issues with QC data files rendering them unfit for inclusion in a historical record to be used for comparative purposes. Only the uploading user is given this power to prevent tampering or mistaken deletion from other users.

Conclusions and Future Directions

Internally, the Yeast Controller has dramatically changed how our lab processes and interacts with QC data. This utility has afforded substantial time savings on the part of individual instrument users and made data more accessible and interpretable to all lab members. That said, there are a number of improvements to this in-house tool that stand to be made. First, additional data analyses and corresponding visualizations are welcome. No analyses comparing populations of detected peptides have been implemented to date. It would be interesting to compare features such as peptide isoelectric point, length, charge, etc. to determine if the composition of profiled molecules can provide insight into or correlate with deviations in instrument performance.

The current implementation of the Yeast Controller does not support any in-

telligent algorithms designed to provide information to users about the source of instrument performance issues. This is perhaps the most fertile ground for addition of functionality. It would be of great benefit for all users if our online system could provide information on the source of instrument issues with fine granularity. Development of this capability requires extensive correlation analyses wherein the root-cause of known performance issues are associated with low-performance QC data files. These low-performance files then need to be analyzed to identify characteristic features and deviations in trends. Detection of such features and trends in future files can then be associated with the previously identified performance issue causes. Lastly, it would be useful to develop functionality to support the addition of user notes on general instrument maintenance and individual QC data files that can be referenced at a later time. Outside of individual notes, development and inclusion of more comprehensive tutorial materials would be greatly beneficial.

Although traditionally rooted in proteomics, our lab has recently begun to diverge by moving into lipidomic and metabolomic analysis. This lateral shift in experimental LC/MS and GC/MS methodologies is accompanied by a need for orthogonal QC monitoring approaches. The current implementation of the Yeast Controller only supports QC processing for LC/MS systems dedicated to proteomics, which are set up for analysis of yeast control samples. Consequently, it

cannot be used for processing small molecule assays. Development of web based tools for upload, processing, and visualization of metabolomic and lipidomic QC assays would be of benefit to lab members who routinely conduct small molecule profiling experiments.

The Yeast Controller has been a valuable resource for our lab and stands to benefit the larger proteomics community if made publicly available. Currently, there is a deficiency of high-quality and open-access resources available for web-based monitoring of LC/MS QC data. The Yeast Controller could be developed as a community solution to enable widespread performance monitoring and data sharing throughout the proteomics community. Our web-based approach facilitates rapid transfer of information which could be leveraged to support labs not currently maximizing their LC/MS system's performance, and help to optimize data acquisition methodologies. An additional step which might be taken to make this community resource more attractive, is the provision of a standard sample for QC analysis. Provision of QC standard samples from a single stock would enable researchers around the globe to more directly compare data between systems by eliminating a prominent source of variation.

Logistically, there are a number of challenges associated with converting the Yeast Controller into a publicly available utility. First, file transfers and data stor-

age will quickly become burdensome as MS file sizes are rapidly expanding. For reference, an average yeast control QC file from the latest generation Orbitrap Fusion Lumos system is ~1.5-2 GB in size (90 minute LC gradient). Comparatively, a similar file from an earlier generation Orbitrap Elite system is ~500 MB. Adequate server space would need to be acquired based on estimated upload traffic, and unused files must be removed on a routine basis. Data privacy and protection of files is a major concern for individual labs. We must have a well-developed security system in place to ensure that no data could be tampered with or downloaded without permission. Finally, we must develop functionality to promote data sharing and comparative analysis between labs, in an effort to provide users with the information needed to bolster data collection methods. We recognize that system performance can vary greatly between labs, and believe it is a worthwhile endeavor to help provide all MS users with the knowledge and tools they need to maximize their LC/MS system's performance. By promoting information sharing between members of the community, individual labs stand to capitalize on each other's knowledge and expertise, and improve the means by which they routinely acquire data.

References

[1] 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, pp. 207–214, 2007.

[2] 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*, Sept. 2016.

[3] C. Bielow, G. Mastrobuoni, and S. Kempa, "Proteomics Quality Control: Quality Control Software for MaxQuant Results," *Journal of Proteome Research*, vol. 15, pp. 777–787, Mar. 2016.

[4] 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, pp. 2360–2366, May 2015.

[5] R. M. Taylor, J. Dance, R. J. Taylor, and J. T. Prince, "Metriculator: quality assessment for mass spectrometry-based proteomics," *Bioinformatics*, vol. 29, pp. 2948–2949, Nov. 2013.

[6] 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, pp. 5540–5547, Nov. 2012.

[7] 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, pp. 958–964, 2004.

[8] 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, pp. 1064–1074, 2011.

[9] D. Thain, T. Tannenbaum, and M. Livny, "Distributed computing in practice: The Condor experience," 2005.

[10] M. Livny, J. Basney, R. Raman, and T. Tannenbaum, "Mechanisms for High Throughput Computing," *SPEEDUP Journal*, vol. 11, pp. 36–40, 1997.

[11] M. Bostock, V. Ogievetsky, and J. Heer, "D3: Data-Driven Documents," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011.

COLOPHON

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