Towards Comprehensive Proteome Sequencing Through Mass Spectrometry

By

Alicia L. Richards

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: 05/20/2016

The dissertation is approved by the following members of the Final Oral Committee:

Joshua J. Coon, Professor, Chemistry
Ying Ge, Professor, Chemistry
Randall Goldsmith, Professor, Chemistry
Lingjun Li, Professor, Chemistry
Michael R. Sussman, Professor, Biochemistry

TOWARDS COMPREHENSIVE PROTEOME SEQUENCING THROUGH MASS SPECTROMETRY

Alicia L. Richards

Under the supervision of Joshua J. Coon

At the University of Wisconsin-Madison

The research described in this dissertation presents strategies for improving the depth of coverage in MS-based proteomics, describing methods for achieving comprehensive identification of all expressed proteins in a given sample through both single-shot analysis and fractionation. An overview of mass spectrometry techniques employed to increase throughput and protein identifications of complex samples is presented in **Chapter 1**. In **Chapter 2**, antibody-based enrichment and isobaric tagging are used to investigate SIRT3 mediated changes in the mitochondrial acetylome across mouse liver, kidney, heart, muscle and brain tissues. **Chapter 3** describes a stable isotope labeling method to differentiate C-terminal peptide fragments, which, when labeled, appear as two peaks spaced by 36 mDa at high resolving powers, from N-terminal fragments, which appear as single peaks. This tool can assist in simplifying spectra prior to database searching and *de novo* sequence identification. **Chapters 4** and **5** describe a high-throughput workflow for comprehensive characterization of the yeast proteome in an expedited timeframe. This rapid proteome characterization is achieved with improved sample preparation, chromatographic separations, and by using a new Orbitrap hybrid mass spectrometer. In **Chapter 6**, this workflow is adapted for the analysis of 174 single gene yeast deletion strains to define the function of uncharacterized mitochondrial proteins. **Chapter 7** details the deep-sequencing of the human proteome through digestion with multiple enzymes and extensive fractionation of six distinct cell lines, providing one of the deepest and most comprehensive proteomic analysis of the human proteome to date.

Acknowledgements

I would like to thank my advisor, Professor Joshua J. Coon, for his mentorship throughout my time at UW-Madison. I am incredibly appreciative of the state-of-the-art resources available to the members of Coon Lab. As is hopefully obvious from the work presented here, no project has been too large-scale to undertake. I have had many opportunities for networking and for developing cutting-edge collaborations that I don't think I could have gotten anywhere else, and for that I am grateful.

The work presented here would not be possible without the help of past and present Coon Lab members. Thanks to Mike Westphall for keeping the lab running and for letting me tear instruments apart. Thanks to Alex Hebert for patiently answering my questions for the past five years. I was fortunate to work with Catie Minogue on many projects – some more successful than others – and I am thankful for her guidance and her friendship. Thanks to Nick Kwiecien for designing analysis tools that made much of this research possible, and for allowing me to distract him on a daily basis. A huge thank you to Molly McDevitt and Greg Potts for their friendship over the past five years. I would also like to thank Anna Merrill, Elyse Freiberger, Nick Riley, Paul Hutchins, Emily Wilkerson, Tim Rhoads, Matt Rush, Evgenia Shiskova, Harald Marx, Erin Weisenhorn, Kyle Connors, Katie Overmyer, Gary Wilson, Anji Trujillo, Dain Brademan, Vanessa Linke, Alan Higbee and Jason Russell. In my experience, it's very rare to enjoy spending time with your co-workers. Thank you for being the best group of co-workers I have ever had.

I have worked with many talented collaborators from the Denu, Pagliarini and Ané labs, specifically Kristin Dittenhafer-Reed, Jon Stefely and Dhileep Jayaraman. I would also like to

thank my committee members, Dr. Ying Ge, Dr. Randy Goldsmith, Dr. Lingjun Li, and Dr. Michael Sussman, for their service on my committee.

I was fortunate to receive funding through a National Institutes of Health-funded Genomic Sciences Training Program (5T32HG002760) and through the ACS Division of Analytical Chemistry and the Society for Analytical Chemists of Pittsburgh, which allowed me to pursue the work presented in this thesis. I am especially grateful for the training I received in GSTP, which exposed me to fields of research outside my own.

I would like to thank my undergraduate research supervisor, Dr. Sarah Trimpin at Wayne State University, who provided her undergraduate students with the same resources and opportunities as her graduate students. This was my first exposure to research, and solidified my decision to continue on to graduate school.

Thank you to my parents, and Jillian and Nick, who provide continuous encouragement, and who have supported me throughout my academic career. Finally, thank you to Chris Lietz for his support, encouragement and discussions over the past five years. Grad school can be difficult at times, and I was fortunate to have someone to go through it with. I am excited for our next steps together.

Table of Contents

Table of Contents (continued)

Table of Contents (continued)

Table of Contents (continued)

Chapter 7. Near Complete Sequencing of the Human Proteome

List of Figures

List of Figures (continued)

List of Figures (continued)

List of Tables

List of Abbreviations and Acronyms

4-HB	4-hydroxybenzoate
4-HBz	4-hydroxybenzaldehyde
ACN	Acetonitrile
AGC	Automatic gain control
ATP	Adenosine triphosphate
AUC	Area under the curve
BCA	Bicinchoninic acid protein assay
BEH	Bridged-ethylene hybrid
BSA	Bovine serum albumin
CAA	Chloroacetamide
CAD	Collisionally activated dissociation
C-CHPP	Chromosome-Centric Human Proteome Project
COMPASS	Coon OMSSA Proteomic Analysis Software Suite
CoQ	Coenzyme Q
Da	Dalton
DIA	Data independent analysis
DMSO	Dimethyl sulfoxide
DNA	Deoxynucleic acid
DTT	Dithiothreitol
ECD	Electron capture dissociation
ETD	Electron transfer dissociation
ESI	Electrospray ionization
FA	Formic acid

List of Abbreviations and Acronyms (continued)

FASP	Filter aided sample preparation
FDR	False discovery rate
FPPE	Formalin-fixed and paraffin-embedded
FT	Fourier transform
FTICR	Fourier transform ion cyclotron resonance
GC	Gas chromatography
GFP	Green fluorescent protein
GO	Gene ontology
GSEA	Gene set enrichment analysis
HCD	Higher-energy collisional dissociation
HF	Hydrofluoric acid
HILIC	Hydrophilic interaction liquid chromatography
HMGCS2	hydroxymethylglutaryl-CoA synthase 2
HPLC	High performance liquid chromatography
Hz	Hertz
IAA	Iodoacetamide
i.d.	Inner diameter
IMAC	Immobilized metal affinity chromatography
IT	Ion trap
iTRAQ	Isobaric tag for relative and absolute quantitation
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	Knock out
LC	Liquid chromatography
LFQ	Label free quantitation

List of Abbreviations and Acronyms (continued)

LIT	Linear ion trap
MALDI	Matrix-assisted laser desorption/ionization
MCNA	Molecule covariance network analysis
m	Mass
MeOH	Methanol
MRM	Multiple reaction monitoring
mRNA	Messenger RNA
MS	Mass spectrometry
MS^1	Survey or precursor scan
MS^2/MS/MS	Tandem mass spectrometry
ms	millisecond
MXP	Mitochondrial uncharacterized protein
m*/*z	mass-to-charge ratio
NETD	Negative electron transfer dissociation
NeuCode	Neutron encoding
NSI	Nanospray
o.d.	Outer diameter
OMSSA	Open Mass Spectrometry Search Algorithm
ORF	Open reading frame
OT	Orbitrap
OxPhos	Oxidative phosphorylation
pABA	*para*-amino benzoate
PIB	Proportional-integral derivative
PBS	Phosphate-buffered saline

List of Abbreviations and Acronyms (continued)

ppb	Parts per billion
PPHB	3-polyprenyl-4-hydroxybenzoate
PRM	Parallel reaction monitoring
psi	pounds per square inch
PSM	Peptide spectrum match
PTM	Post-translational modification
Q	Quadrupole
qIT	Quadrupole ion trap
QSSA	Quantitative site set functional score analysis
QTL	Quantitative trait loci
QTOF	Quadrupole time –of-flight
RD	Respiration-deficient
RDR	Respiration deficiency response
Rf	Radio frequency
RNA	Ribo-deoxy nucleic acid
RP	Reversed phase
SAX	Strong anion exchange
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
SILAC	Stable isotope labeling of amino acids in cell culture
SIRT3	Sirtuin 3
S/N	Signal-to-noise ratio
SNP	Single nucleotide polymorphism
SWATH	Sequential window acquisition of all theoretical fragment ion spectra

List of Abbreviations and Acronyms (continued)

TAP	Tandem affinity purification
TCEP	tris(2-carboxyethyl) phosphine
TFA	Trifluoroacetic acid
Th	Thomsson
TIC	Total ion chromatogram
TiO_2	Titanium dioxide
TMT	Tandem mass tags
TOF	Time-of-flight
Tyr	Tyrosine
v	Volume
WT	Wild-type
YPD	Yeast extract peptone dextrose
z	Charge

Chapter 1

Introduction

Portions of this chapter have been published:

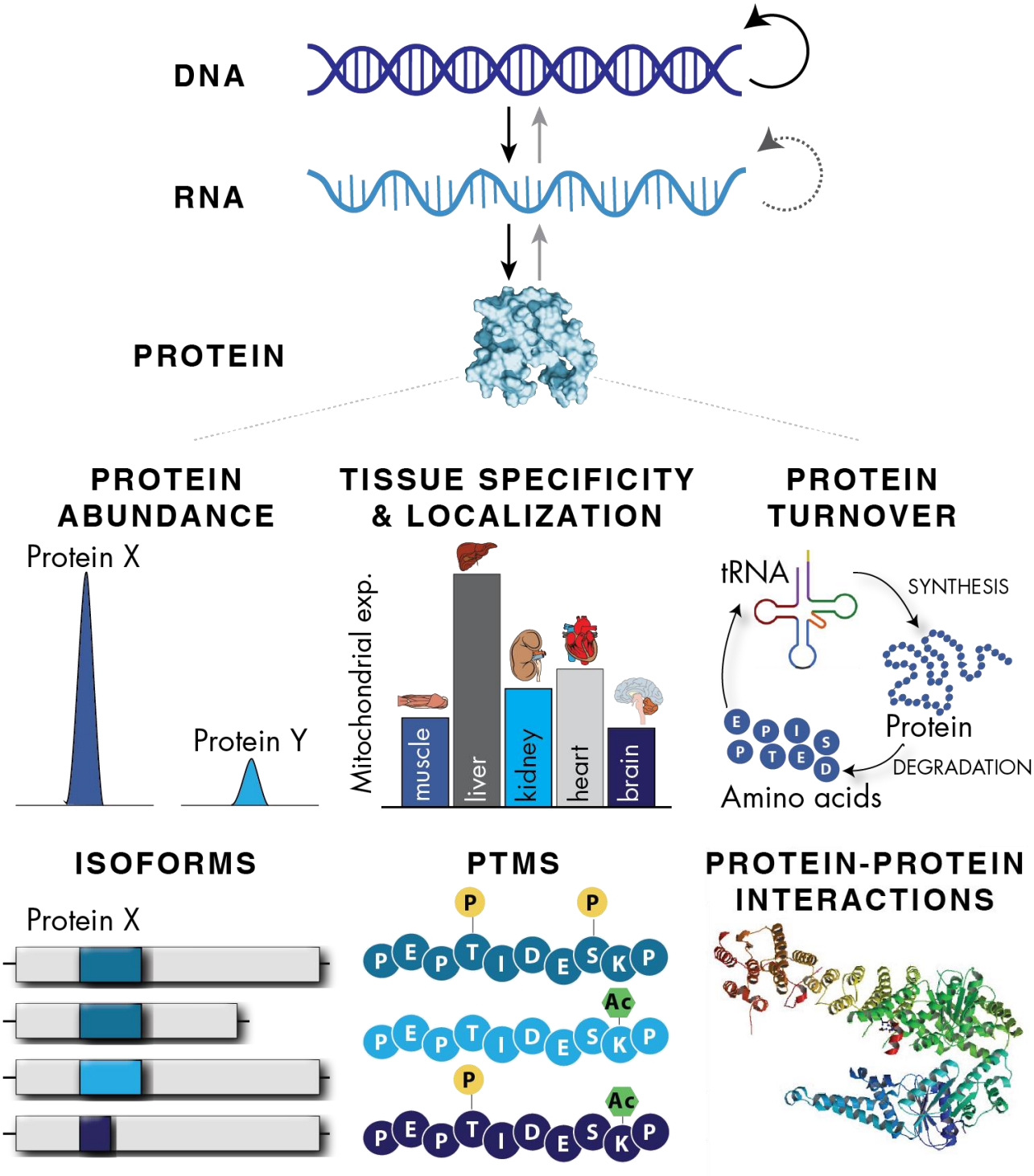
Richards AL, Merrill AE, Coon JJ. *Proteome Sequencing Goes Deep*. Current Opinion in Chemical Biology. **2014,** *24*, 11-17.

INTRODUCTION

Advances in mass spectrometry (MS) have transformed the scope and impact of protein characterization efforts. Identifying hundreds of proteins from rather simple biological matrices, such as yeast, was a daunting task just a few decades ago. Spurred by the advent of soft ionization methods electrospray (ESI)[1] and matrix-assisted laser desorption ionization (MALDI)[2] in the late 1980s, MS has become the central method for protein analysis. Since this time, the depth and rate at which a proteome can be characterized has steadily improved. Now, expression of more than half of the human protein coding genes can be confirmed in record time and from minute sample quantities. Access to proteomic information at such unprecedented depths has been fueled by strides in every stage of the high-throughput proteomics workflow — from sample processing to data analysis — and promises to revolutionize our understanding of the causes and consequences of proteome variation.

Genetic information flows from DNA to RNA to proteins (**Figure 1**). Genomic and transcriptomic technologies have matured such that a typical RNA-Seq experiment can identify and quantify nearly all transcribed gene products. Completion of the human genome project has inspired efforts to similarly map the entire human proteome. This is a non-trivial task, as mammalian proteomes are complex.[3] At the DNA level, evidence exists for ~20 300 protein-coding genes; however, post-translational modifications (PTMs) and RNA processing steps, including non-synonymous single nucleotide polymorphisms (nsSNPs) and alternative splicing events, all occur and exponentially increase the number of distinct proteoforms.[4-7] Further, proteomes are dynamic, and protein expression can vary over time, between cell types and location, and in response to external stimuli. Additionally, proteins can be expressed over a large dynamic range (**Figure 1**). The mammalian proteome spans at least seven orders of magnitude,[6,8] while the dynamic range of mammalian mRNA is estimated to cover three to four orders of magnitude.[9] Given that proteins more closely resemble phenotype than their encoding nucleic acid counterparts, they harbor unique biological details that can inform larger biological processes. Detection of 5,000 proteins in a proteomic experiment was a considerable achievement just a few years ago.[10-12] More recently, two groups identified over 10 000 protein groups in a single experiment. Through extensive protein and peptide fractionation (72 fractions) and

Figure 1. Flow of genetic information. (***Top***) Genetic information flows from DNA to RNA to protein. (***Bottom***) Proteomic analyses can be complicated by large difference in protein abundance, tissue specificity and cellular localization, and protein turnover. Isoforms and the presence of post-translational modifications can significantly increase the size of the proteome. Additionally, it may be important to preserve protein-protein interactions during analysis.



digestion with multiple enzymes, Nagaraj et al. identified 10, 255 protein groups from HeLa cells over 288 hours of instrument analysis.[13] A comparison with paired RNA-Seq data revealed nearly complete overlap between the detected proteins and the expressed transcripts. In that same year, a similar strategy enabled the identification of 10, 006 proteins from the U2OS cell line.[8]

A more comprehensive analysis of the human proteome can be achieved by applying similar technologies to largescale comparisons of multiple cell lines and tissues.[14-17] To date, three human proteome maps have been published. Kim and co-workers analyzed 30 human tissues and primary cells over 2000 LC–MS/MS experiments, resulting in the detection of 293 000 peptides with unique amino acid sequences and evidence for 17 294 gene products.[18] Wilhelm *et al*. amassed a total of 16,857 LC–MS/MS experiments from human cell lines, tissues, and body fluids. These experiments produced a total of 946 000 unique peptides, which map to 18 097 protein coding genes.[19] Together, these draft maps of the human proteome provide direct evidence for protein translation of over 90% of human genes. The Human Protein Atlas project takes a different approach, using >24,000 polyclonal antibodies to visualize ~17,000 proteins. Over the past decade, the project has produced >13,000,000 immunohistochemistry images, providing subcellular localization of proteins across multiple tissues.[20,21]

The Chromosome-Centric Human Proteome Project (C-CHPP) seeks to identify the protein products of all protein-coding human genes.[22] It is hypothesized that undetected proteins have features that make MS detection difficult, including low abundance, tissue or body fluid specificity, or amino acid sequences that may be incompatible with standard proteomic workflows.[23] Strategies employed by the >25 labs involved in the C-CHPP include the development of multiple reaction monitoring (MRM) assays for predicted proteins of interest, specialized enrichment steps to identify proteins of low abundance, and refining sample preparation to increase classes of under-represented proteins, including membrane proteins.[24] To date, raw files from over 1,000 experiments have been performed by the labs participating in this endeavor.

Although these studies have provided the deepest proteome coverage to date, complete proteome sequencing is not a trivial task, as the above analyses required a significant investment in researcher and

instrument time. For example, the study by Wilhem *et al.* required non-stop operation of a mass spectrometer for four straight years. New developments in mass spectrometer technology have increased the rate at which proteomes can be analyzed. Using such a device, we recently described a method that characterizes nearly every protein in yeast in just over one hour (4000 of the 4500 expressed yeast proteins).[25,26] Further improvements to sample preparation, MS instrumentation, and bioinformatics that have been key to obtaining comprehensive proteomic coverage. In order to maximize protein identifications, optimization can occur at every step in this process, as highlighted below.

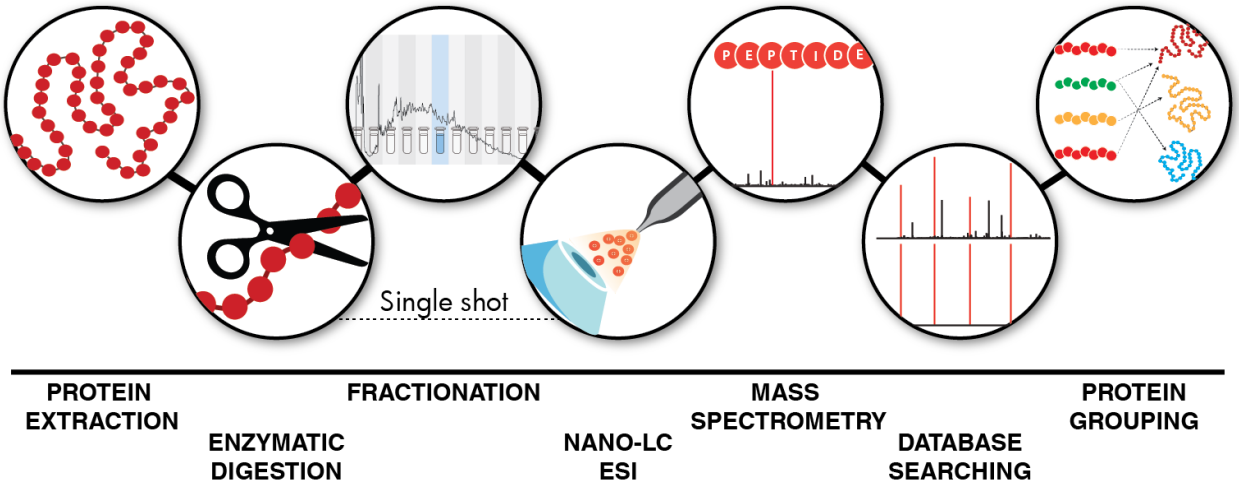
BOTTOM-UP PROTEOMICS

Bottom-up, or high-throughput proteomics, as outlined in **Figure 2**, has proven the most useful tool for comprehensive analyses. Here, proteins are extracted from lysed cells, enzymatically digested, and chromatographically separated by reversed phase (RP) nano-liquid chromatography (LC) prior to MS analysis. As peptides elute over a gradient of increasing organic solvent, they are ionized by a high voltage applied to the tip of the column using nano-ESI.[1] The mass spectrometer records the masses of eluting peptide cations every second or so. In between these so-called MS^1 scans the system isolates selected peptide precursors, dissociates them using collisions or chemical reactions, and records the masses of the pieces (i.e. MS/MS or tandem MS). Modern MS systems can measure peptide masses accurately to three decimal places while at the same time collecting tandem mass spectra at a blazing rate of over 20 Hz.[27-29] The hundreds of thousands of spectra generated from one of these experiments are then analyzed *in silico* using spectral matching algorithms.[30] Tandem MS spectra are compared to theoretical spectra of possible peptides within a specified mass tolerance. Identified peptides are then mapped to the fewest number of proteins that can explain all peptides. Quantitation strategies can be used to estimate protein expression levels across the dataset.

ADVANCES IN PROTEOMIC SAMPLE PREPARATION

Cell lysis. For any proteomic method, proteins must first be liberated from their host cells, via mechanical

Figure 2. Workflow for 'high-throughput' or 'bottom-up' proteomics. Workflow for 'high-throughput' or 'bottom-up' proteomics. Preparing proteomic samples for LC–MS/MS analysis requires protein extraction, proteolysis, and, optionally, peptide-level fractionation. Online LC separation of complex peptide mixtures introduces analytes into the mass spectrometer for precursor and fragment ion mass analysis. Tandem mass spectra are matched to theoretical spectra generated *in silico* to garner peptide sequences that are used for protein inference.



and/or chemical disruption, often into a denaturing solution. Reduction of disulfide bonds and alkylation of cysteine residues disrupts protein structure, leaving proteins amenable to site-specific cleavage with one or more proteases. This initial step — protein extraction and solubilization — is paramount, as it dictates which proteins will be accessible for eventual MS detection. Chaotropes, including urea, are often used to destroy a protein's ordered structure. Strong detergents, such as sodium dodecyl sulfate (SDS), are exceptional denaturants, but their removal, a requirement for efficient proteolytic digestion and sensitive mass-spectrometric analysis, is challenging. Standard filtration devices can be employed as proteomic reactors (filter-aided sample preparation, FASP), allowing dissolution of proteins in high concentrations of SDS which are then depleted before digestion.[31] Alternatively, an unbiased proteomic characterization can be achieved without SDS by digesting unclarified cellular lysate, a tactic that improves coverage of proteins harbored in poorly soluble membrane and nuclear organelles.[25,32,33] Protease and phosphatase inhibitors may be included during lysis to prevent proteolysis and de-phoshorylation of proteins, respectively.

Protein digestion. The maximum coverage obtainable for a protein is theoretically determined by its amino acid sequence and the cleavage specificity of the chosen proteolytic enzyme, typically trypsin. Trypsin is the most widely used enzyme in proteomics, as tryptic fragments have many characteristics making them amenable to analysis by mass spectrometry.[34] Following digestion, tryptic peptides are generally within the preferred mass range for analysis by MS. Tryptic peptides contain a basic residue (arginine or lysine) at the C-terminus, which enhances both ionization and gas phase fragmentation.

Despite the utility of trypsin, the majority of peptides trypsin generates contain 6 amino acids or fewer, which are too small for confident identification by MS, leading to gaps in proteome coverage.[35] As most proteins are identified using only a subset of their respective peptides, sequence coverage, or the percentage of the protein that is actually observed, is often low. Unfortunately, this coverage is likely insufficient to identify the myriad of unpredicted proteoforms arising from genetic alterations, including single nucleotide polymorphisms (SNPs), alternative splicing, and gene fusion.[7] Given this situation, even the current deepest proteomic datasets do not contain enough sequence data to identify such permutations.

A straightforward and long-recognized approach for boosting protein and proteome coverage is to digest a sample separately with multiple proteases.[35,36] In addition to trypsin, proteases employed in large-scale proteomics experiments include LysC, LysN, GluC, AspN, ArgC, and chymotrypsin. The cleavage specificity of each of these enzymes is outlined in **Table 1**. The combined use of trypsin and other enzymes has significantly increased the number of phosphorylation sites identified when compared to identifications made with trypsin only.[37,38] Digestion with five specific enzymes nearly doubled sequence coverage of the yeast proteome, from 24.5% average sequence coverage with trypsin alone, to 43.4% with all five proteases.[35] A similar boost in sequence coverage was seen for the human proteome; digestion of HeLa lysate with a combination of seven different proteases, coupled with extensive fractionation, identified ~8,500 proteins with >40% mean sequence coverage.[39] Recently, digestion with α-lytic proteases, semi-specific enzymes that preferentially cleave after aliphatic residues, increased trypsin-only protein identifications and sequence coverage by 24% and 101%, respectively.[40] Some downsides of such multi-protease strategies include heightened sample quantity and analysis time demands. Digesting with multiple enzymes sequentially instead of in parallel, however, can afford better coverage without the extra requirements.[39,41]

PTM analysis. MS allows the large-scale mapping of PTMs, which are vital in determining a protein's interactions with other proteins, localization, turnover and activity. Large-scale PTM studies have localized > 20,000 phosphorylation,[42] 10,000 ubiquitination,[43] and acetylation events[44] to specific amino acid residues from tissue or cell culture. Due to the low stoichiometry and sometimes transient nature of PTM peptides, a specific enrichment step is often required for detection. Immobilized metal affinity chromatography (IMAC)[45] or titanium dioxide (TiO_2),[46] where metal oxide chelators bind to negatively charged phosphate groups, are favored techniques for phosphopeptide analyses. Antibody-based immunoprecipitation is employed for other PTMs, including ubiquitinated and acetylated lysine residues.

Overall, recent developments in sample processing for high-throughput proteomics have emphasized simplification and scalability.[47] Furthermore, a robust workflow with minimal sample loss and

Table 1. Cleavage specificity of proteases commonly used in mass spectrometry experiments. Cleavage specificity is denoted by line. Amino acids in grey indicate less frequent site of cleavage.

Enzyme	Cleavage specificity
Trypsin	K|, R|
LysC	K|
LysN	|K
ArgC	|R
GluC	E|, D|
AspN	|D
Chymotrypsin	W|, Y|, F|, L|

X| = C-terminal cleavage specificity
|X = N-terminal cleavage specificity

contamination opens the door for applications with limited starting material. Using current technology, 9,500 proteins can be identified from just 100 nL of formalin-fixed and paraffin-embedded (FFPE) tissue.[48] Although it is impossible to ionize and sequence every peptide, efficient sample preparation, advances in MS instrumentation,[49-51] coupled with separation methodology,[52-55] and fragmentation techniques[56-59] have vastly increased the observable portion of the human proteome.

ADVANCES IN PEPTIDE SEPARATION AND MS INSTRUMENTATION

Separations. *In silico* tryptic digestion of the ~20,200 reviewed, canonical proteins of the human proteome (UniProt, downloaded 09/16/2015) predicts 678,007 peptides of suitable size for MS detection;[60] Note, this number does not include isoforms or post translationally or chemically modified peptides. Further, the number of possible peptides increases exponentially when considering peptides that arise from missed cleavages. A report by Mann and colleagues conservatively estimates that 100,000 peptides are introduced to the mass spectrometer during analyses of complex cell digests; however, only a fraction of these peptides are ever selected for MS/MS analysis.[61] Even following efficient solubilization and proteolysis, many proteins are only represented following detection of a few unique peptides.[6] This mainly stems signal suppression during the electrospray ionization — that is, peptides having higher overall basicity tend to preferentially ionize rendering the more acidic peptides undetected. The most successful approach to curb this problem is to reduce the number of unique peptide sequences present in the ionization source at one time, increasing the number of peptides converted to gas phase ions. By reducing sample complexity, more peptides can be identified within the given dynamic range of a mass analyzer. To this end, separation chemistries and implementations thereof are central to proteomic analysis. These can include high-pH RP,[62-64] ion exchange methods including strong cation exchange (SCX)[65] or strong anion exchange (SAX),[66] or hydrophilic interaction liquid chromatography (HILIC).[67] The extreme separation resolution provided by some of these platforms, such as the automated coupling of three physicochemically orthogonal stages of chromatography[68] and high-resolution isoelectric focusing,[69] is key to achieving genome-scale coverage of the proteome.

Chromatography. The sequencing speed of modern mass spectrometers is best harnessed when coupled to efficient, online peptide separation. High performance liquid chromatography (HPLC) systems operating at high pressures (>8000 psi)[53] and longer columns packed with small particles have become standard. The linear relationship between the number of identified peptides and peak capacity, the number of resolvable peaks across an elution, has been demonstrated.[52] Many recent workflows have focused on optimizing chromatographic separations rather than extensive fractionation for whole proteome analysis.[70] Forgoing sample pre-fractionation in favor of long columns packed with 2 mm particles, a recent study identified 4825 protein groups from the A375 cancer cell line over a three hour LC–MS/MS experiment.[32] A comparison of column lengths revealed that, for this particular instrument platform, a 50 cm column allowed identification of more proteins than either a 15 or 25 cm column at all gradient lengths tested, although a decrease in cumulative identified protein groups after three hours was reported. Note that the combination of long columns and small particles significantly raises column backpressure, necessitating either a UHPLC system capable of operating at pressures >10 000 psi, or a column heater. Silica monolithic columns, which can achieve separation efficiencies similar to traditional packed columns without a substantial increase in back pressure, have also been used.[55]

Instrumentation. Following the introduction of gas-phase ions to the mass spectrometer, they are separated by the mass analyzer according to their *m*/*z*. Mass analyzers employed in proteomics include quadrupoles (Q), linear ion traps (LIT), time-of-flight (TOF), and Orbitraps. Each type of mass analyzer has advantages and limitations, outlined in **Table 2**, making some more suited for specific applications. Improvements to instrumentation have focused on increasing resolving power, mass accuracy, analysis speed, and sensitivity. High resolving power and mass accuracy have changed the scope of high-throughput proteomics, allowing the identification of thousands of precursors present at any point during the gradient.[27,29,71,72] In lower resolution instruments, multiple peptides eluting at the same retention time make it difficult to obtain accurate mass and charge state information. High resolving power enables the identification of these closely spaced precursors, and is crucial in filtering possible peptide candidates during

Table 2. Overview of mass analyzers employed in proteomics research.

Mass analyzer	Separation principle	Resolving power	Mass accuracy	Speed
 Orbitrap	*m/z* - resonance frequency	10,000 - 480,000	<5 ppm	20 Hz (15K @ 200 *m/z*)
 Ion trap	*m/z* - resonance frequency	1,000 - 10,000	100 ppm	40 Hz
 Quadrupole	*m/z* - trajectory stability	100 - 1,000	100 ppm	40 Hz
 Time-of-flight	Velocity-flight time	1,000 - 60,000	5-10 ppm (with lockmass)	100 Hz (<20 Hz for proteomics)

database searching.[73] The Orbitrap mass analyser has emerged as a popular choice for proteomics research, as it combines high resolution and mass accuracy with speed and sensitivity. The Orbitrap is a spindle shaped device consisting of a central electrode and two outer electrodes.[74] Ions are first accumulated in the C-trap, an external Rf-only trapping device that pulses a defined number of ions into the Orbitrap.[75] Ions entering the Orbitrap undergo axial oscillations initiated by a strong electric field, while also rotating around the central electrode with a period proportional to $(m/z)^{1/2}$. These oscillating ions produce an image current, and are measured at high resolution[74] (480,000 at 200 *m*/*z* for a 1024 ms transient).[27] Since its commercial introduction in 2005, improvements to the Orbitrap design have significantly increased resolving power and acquisition speed. In 2011, the compact high field Orbitrap was introduced, providing an 18% increase in resolving power for the same detection period.[75]

Modern hybrid mass spectrometers, combining more than one mass analyzer, allow multiple experiments to be performed, and can couple highly accurate MS^1 scans with ultra-fast MS/MS sequencing rates. Increasing the sequencing speed of the mass spectrometer is advantageous. As the number of peptides introduced to and fragmented by the mass spectrometer increases, the percentage of the proteome identified in an analyses can also increase. This is illustrated in a recent study using a LIT-Orbitrap hybrid mass spectrometer,[75] which, compared to the previous generation instrument, achieves approximately twice the resolving power at the same scan speed, to analyze eleven human cell lines.[76] Across all cell lines, 11 731 proteins were identified, with an average of 10 361 proteins identified per cell line. The number of identified protein groups is comparable to a previous study from HeLa lysate,[13] but is generated in a fraction of the time (3 versus 12 days). The Orbitrap Fusion mass spectrometer, combining a quadrupole mass filter, a collision cell, a dual cell LIT and an Orbitrap mass analyzer, when operating at a MS/ MS acquisition speed of 20 Hz,[27] doubles the number of tryptic yeast peptides identified per second as compared with the Orbitrap Elite (19 versus 10 peptides/ second).[25]

Fragmentation. To produce peptide fragments, precursors can be isolated and collided with neutral gas molecules. With collisionally-activated dissociation (CAD), isolated precursors are accelerated by a

resonant RF electrical field and undergo numerous collisions with helium molecules in the ion trap, causing the kinetic energy of the molecules to be converted to vibrational energy.[77] As this energy increases above a certain threshold, the weakest peptide bonds break, and the precursor dissociates into b and y ions that can be used for identification. Higher-energy collisional dissociation (HCD) also produces b and y ions, with spectra more similar to those produced from QTOF or triple quadrupole mass spectrometers.[57] Here, ions are injected into the ion routing multipole (IRM) or HCD cell, where they collide with nitrogen molecules. These collisions impart more energy per collision than in CAD-based fragmentation, and can cleave multiple bonds, producing ions that can be analyzed in the higher resolution in the Orbitrap or at lower resolution in the ion trap. Compared with conventional IT analysis, HCD diminishes the"1/3 rule" of ion trap fragmentation, where ions below a certain mass cutoff are not detected, allowing the detection of isobaric tags present at low *m*/*z*.[78] Electron-based approaches such as electron-capture dissociation (ECD) or electron transfer dissociation (ETD) cleave the N-Cα bond, generating c and $z^{\cdot}$-type fragment ions. In ECD, low energy electrons are reacted with multiply charged positive ions in an FTICR cell.[79] With ETD, an anion provides the electron required for dissociation.[56] The negative mode analog of ETD, negative electron transfer dissociation (NETD) shows promise for identifying acidic portions of the proteome.[80]

DATA ANALYSIS

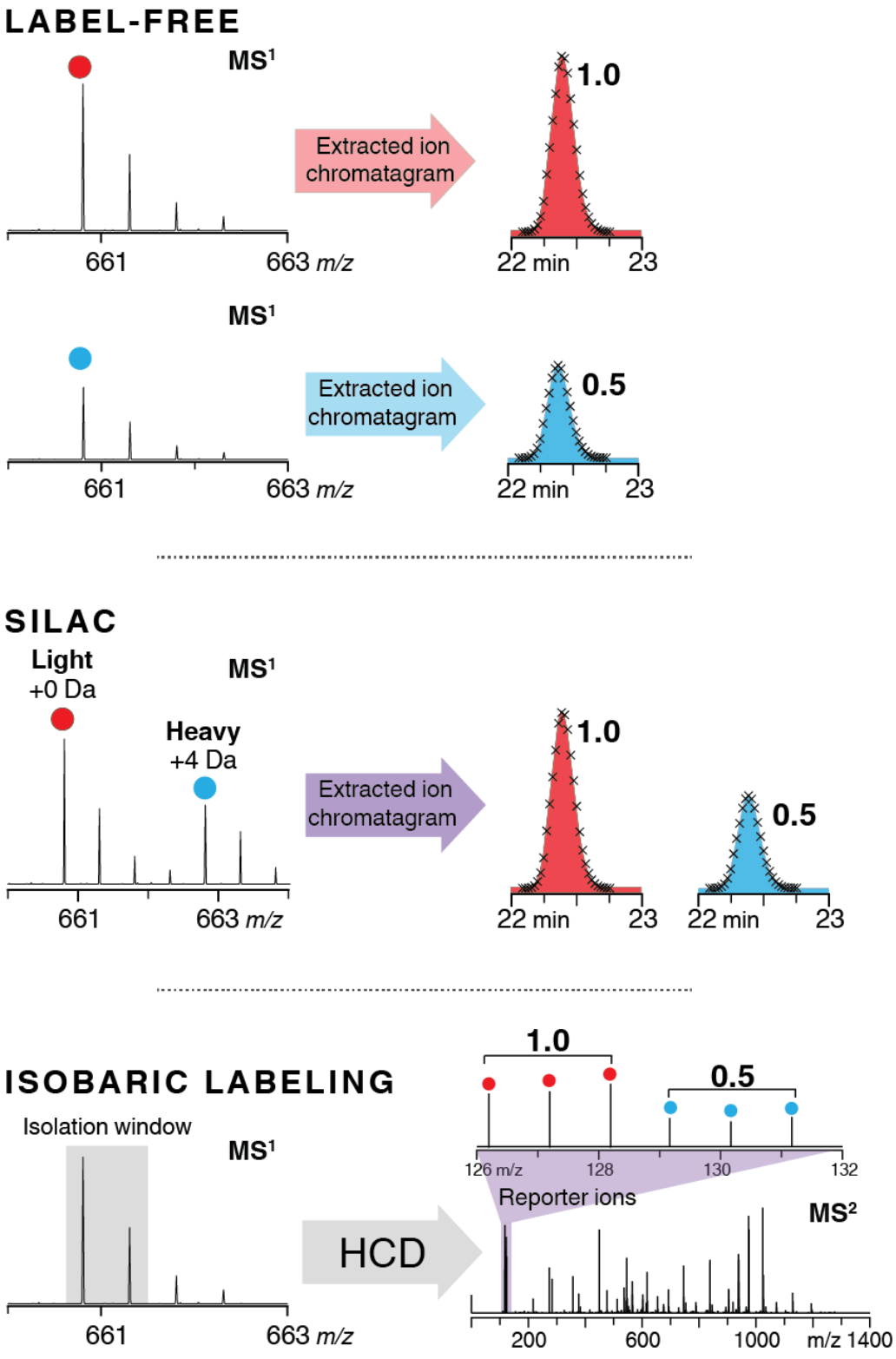
Identification and protein grouping. The presence of proteins within a given sample is inferred through peptide identifications. This can be achieved through *de novo* sequencing, where peptide sequences are determined directly from tandem mass spectra, independent of a database.[81] Far more common is database searching, where the fragments present in a tandem mas spectra are searched against theoretical spectra generated from the *in silico* digestion of the organism's protein database.[30] Confident hits are termed peptide spectrum matches (PSMs). The number of false identifications throughout the dataset is calculated by applying a false discovery rate (FDR), commonly using the target-decoy approach.[82] Here, data is searched against a concatenated database containing both the target sequences and theoretically impossible decoy sequences, created by shuffling or reversing the target amino acid sequences.[83] Random matches should

occur with equal frequency between the target and decoy sequences, allowing the number of false positives to be calculated based on the number of decoy hits. A score threshold, or FDR, is calculated for the dataset based on these distributions, and can be evaluated at either the PSM, peptide, or protein level. Several groups have proposed alternative strategies to the target-decoy approach for large datasets.[84,85] Following identification, peptides are assembled into their corresponding proteins. This task can become complicated as the same peptide can be present in more than one protein, leading to ambiguities in identification. Typically, the more peptides that can be successfully sequenced and mapped back to a corresponding protein, the more confident the protein identification.

Quantitation. Information for determining protein abundance is generally based on isotopic labeling or label-free approaches, as outlined in **Figure 3**. To facilitate large scale comparisons and increase reproducibility, multiplexed sample preparation, where samples are processed and analyzed in parallel, may be desirable. Stable isotopes are introduced through either metabolic or isotopic labeling, with quantitation performed at either the MS^1 or MS/MS level.[86-90] Isobaric tagging can be used to determine relative amounts of protein across a number of samples in a single experiment. Following protein digestion, samples are tagged mixed in equal ratios and analyzed via LC-MS/MS. These isobaric tags have the same nominal mass, but because of the number and placement of nitrogen and carbon isotopes, reveal distinctive reporter ion signals for each sample following MS/MS. The intensity of these signals allow comparisons of the relative protein abundance between samples. As quantitative information is obtained during the tandem MS scan, precursors must be selected for fragmentation in order to be quantified. Frequently, other precursor ions are co-isolated with the target peptide, contaminating the reporter ion spectra and effecting quantitative accuracy. Commercial tagging reagents allow up to 10 samples at a time to be compared, significantly increasing proteomic throughput.[68,91]

Non-isobaric labeling methods include SILAC and dimethyl labeling. With SILAC, either "light" or "heavy" versions of amino acids are metabolically incorporated during cell culture.[86,87] These samples are mixed, and when analyzed by LC-MS/MS, peptides are presented as both a "light" and "heavy" version.

Figure 3. Overview of different quantitation strategies used in mass spectrometry. (***top***) Label-free; (***middle***) SILAC; (***bottom***) isobaric labeling.



The area under the curve (AUC) of peptides from the MS^1 scan is compared to give relative quantitation between conditions. SILAC provides accurate quantitation, as co-eluting species are resolvable in the MS^1 scan, eliminating interference. In order to prevent isotopic clusters of the SILAC "light" and labeled partners from overlapping, a space of at least 4 Da between partners is required, which limits the multiplexing capability of SILAC. Additionally, each SILAC label increases the complexity of the MS^1 compared to spectra with a single version of each peptide, which can limit sampling depth and proteome coverage. NeuCode was recently introduced to overcome this challenge. NeuCode uses small (~36 mDa) mass defects present in stable isotopes to boost the multiplexing capabilities of SILAC without increasing spectral complexity.[92-94]

Label-free quantitation can use either MS or MS/MS-derived data for quantitation. Spectral counting measures the number of MS/MS fragments generated for each peptide, and is generally less accurate than AUC quantitation.[95] MS-based quantitation compares the AUC for the same peptide across multiple conditions. Label-free quantitation is an attractive option, as it requires no extra labelling steps or sample mixing. However, as samples are analyzed separately, reproducibility is important.[96] Peptidic features, including m/z, retention time and intensity are aligned across all samples. Computational approaches, including alignment and feature finding, are optimized to mitigate the effects of overlapping features due to co-elution or from poor S/N. Systems-level analyses have benefitted from the growing robustness and availability of informatics tools for label-free quantification strategies. Even highly fractionated proteomes can now be accurately compared in the absence of stable isotope labels.[96,97] Aided by highly accurate mass measurements, the confident transfer of peptide identifications between matching runs provides a 25–30% boost in the number of proteins quantified across multiple samples.[16] This feature makes label-free approaches very attractive for deep proteome quantification, though stable isotope labeling strategies are still more straightforward for the comparative analysis of low abundance PTMs.[98] Furthermore, statistical analysis of signatures at the peptide-level can reveal information regarding the presence and expression patterns of one or more proteomes, an approach that will be greatly empowered by high protein sequence coverage.[99]

Comparisons across large cohorts of samples are required for many large-scale applications, including KO library screening.[100,101] As comparing across thousands of samples has been beyond the capabilities of proteomics, these studies are traditionally performed using genomic technologies. However, correlation between mRNA and protein abundance can be poor,[102] and cannot provide evidence of PTMs. Label-free methods can theoretically compare unlimited numbers of samples, but as each sample is analyzed separately, analysis time can be high compared with labeling methodologies. However, as isobaric tagging requires additional clean-up steps to remove unreacted tag, and can introduce unknown modifications and inefficient fragmentation,[103] fractionation is required to increase identifications. Even with fractionation, it can be difficult to achieve proteomic depth comparable to what is routinely achieved through label-free quantitation, mitigating the decrease in analysis time obtained through multiplexing.[104]

Proteogenomics. For organisms with sequenced genomes, peptides detected by MS can assist in refining prediction-based gene annotations, a primary goal of the emergent proteogenomics field.[105] In addition to validating predicted genes, deep proteomic coverage can suggest novel protein-coding loci,[106] N-terminal signal peptides,[107] splice sites,[108] and nonsynonymous variants.[109] A long-term objective of proteogenomic mapping is to associate certain variations in protein sequence with disease states. One recent study combined a customized protein database with in-depth transcriptome and proteome profiling of livers from two inbred rat strains.[108] Interestingly, the results associated a genomic variant in the promoter region of a mis-annotated gene with the observed hypertensive phenotype of one strain, illustrating the advantages of such integrated approaches. Proteogenomic endeavors match tandem mass spectra to a database containing, ideally, all possible protein sequences encoded by an individual genome. This poses a computational challenge for large genomes with low protein-coding content, requiring extensive search-space reductions to boost sensitivity. A fresh strategy enabled unbiased proteogenomic mapping against the full human genome, along with deep proteome coverage, by blending isoelectric focusing for high-resolution peptide separation with accurate isoelectric point prediction for rational reduction of the search space.[69]

CONCLUSIONS

Breakthroughs in every stage of the shotgun proteomics workflow have collectively ushered in a new era of proteomics, one in which identification and quantification of complete proteomes can be routinely achieved.[110] Beyond propelling basic research, this age holds great potential for personalized medicine.[111] As deep cataloguing of protein expression becomes widespread, the spotlight will shift to extensive functional mapping of proteoforms and determining how their expression is regulated by genomic elements.[112] To this end, the complementary benefits of top-down,[113] targeted,[114] and antibody-based[115] approaches must be harnessed and effectively integrated. Finally, in light of the surging trend of deep proteomics, it is important to remember that, for some systems, meaningful biological insight can still be drawn from moderate depths of proteome coverage, which are becoming more accessible to proteomic researchers of all experience levels.

OVERVIEW OF PROJECTS

The research described here presents strategies for improving the depth of coverage in MS-based proteomics, describing methods for achieving comprehensive identification of all expressed proteins in a given sample through both single-shot analysis and fractionation. In **Chapter 2**, antibody-based enrichment and isobaric tagging are used to investigate SIRT3 mediated changes in the mitochondrial acetylome across mouse liver, kidney, heart, muscle and brain tissues (published, 2015). **Chapter 3** describes a stable isotope labeling method to differentiate C-terminal peptide fragments, which, when labeled, appear as two peaks spaced by 36 mDa at high resolving powers, from N-terminal fragments, which appear as single peaks (published, 2013). This tool can assist in simplifying spectra prior to database searching and *de novo* sequence identification. **Chapters 4** and **5** describe a high-throughput workflow for comprehensive characterization of the yeast proteome in an expedited timeframe. This rapid proteome characterization is achieved with improved sample preparation, chromatographic separations, and by using a new Orbitrap hybrid mass spectrometer (published, 2014; 2015). In **Chapter 6**, this workflow is adapted for the analysis of 174 single gene yeast deletion strains to define the function of uncharacterized mitochondrial proteins

(submitted, 2016). **Chapter 7** details the deep-sequencing of the human proteome through digestion with multiple enzymes and extensive fractionation of six distinct cell lines, providing one of the deepest and most comprehensive proteomic analysis of the human proteome to date (manuscript in preparation).

REFERENCES

(1) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64-71.

(2) Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. *Int J Mass Spectrom* **1987**, *78*, 53-68.

(3) Harrow, J.; Frankish, A.; Gonzalez, J. M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B. L.; Barrell, D.; Zadissa, A.; Searle, S.; Barnes, I.; Bignell, A.; Boychenko, V.; Hunt, T.; Kay, M.; Mukherjee, G.; Rajan, J.; Despacio-Reyes, G.; Saunders, G.; Steward, C.; Harte, R.; Lin, M.; Howald, C.; Tanzer, A.; Derrien, T.; Chrast, J.; Walters, N.; Balasubramanian, S.; Pei, B.; Tress, M.; Rodriguez, J. M.; Ezkurdia, I.; van Baren, J.; Brent, M.; Haussler, D.; Kellis, M.; Valencia, A.; Reymond, A.; Gerstein, M.; Guigo, R.; Hubbard, T. J. *Genome Res* **2012**, *22*, 1760-1774.

(4) Consortium, E. P. *PLoS Biol* **2011**, *9*, e1001046.

(5) Genomes Project, C.; Abecasis, G. R.; Altshuler, D.; Auton, A.; Brooks, L. D.; Durbin, R. M.; Gibbs, R. A.; Hurles, M. E.; McVean, G. A. *Nature* **2010**, *467*, 1061-1073.

(6) Zubarev, R. A. *Proteomics* **2013**, *13*, 723-726.

(7) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P. *Nat Methods* **2013**, *10*, 186-187.

(8) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. *Mol Syst Biol* **2011**, *7*, 549.

(9) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. *Nature* **2011**, *473*, 337-342.

(10) Burkard, T. R.; Planyavsky, M.; Kaupe, I.; Breitwieser, F. P.; Burckstummer, T.; Bennett, K. L.; Superti-Furga, G.; Colinge, J. *BMC Syst Biol* **2011**, *5*, 17.

(11) Wisniewski, J. R.; Zougman, A.; Mann, M. *J Proteome Res* **2009**, *8*, 5674-5678.

(12) Hubner, N. C.; Ren, S.; Mann, M. *Proteomics* **2008**, *8*, 4862-4872.

(13) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. *Mol Syst Biol* **2011**, *7*, 548.

(14) Phanstiel, D. H.; Brumbaugh, J.; Wenger, C. D.; Tian, S.; Probasco, M. D.; Bailey, D. J.; Swaney, D. L.; Tervo, M. A.; Bolin, J. M.; Ruotti, V.; Stewart, R.; Thomson, J. A.; Coon, J. J. *Nat Methods* **2011**, *8*, 821-827.

(15) Munoz, J.; Low, T. Y.; Kok, Y. J.; Chin, A.; Frese, C. K.; Ding, V.; Choo, A.; Heck, A. J. *Mol Syst Biol* **2011**, *7*, 550.

(16) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. *Mol Cell Proteomics* **2012**, *11*, M111 014050.

(17) Moghaddas Gholami, A.; Hahne, H.; Wu, Z.; Auer, F. J.; Meng, C.; Wilhelm, M.; Kuster, B. *Cell Rep* **2013**, *4*, 609-620.

(18) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. *Nature* **2014**, *509*, 575-581.

(19) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. *Nature* **2014**, *509*, 582-587.

(20) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; Wernerus, H.; Bjorling, L.; Ponten, F. *Nat Biotechnol* **2010**, *28*, 1248-1250.

(21) Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P. H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Ponten, F. *Science* **2015**, *347*, 1260419.

(22) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. *Nat Biotechnol* **2012**, *30*, 221-223.

(23) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S. *J Proteome Res* **2012**, *11*, 2005-2013.

(24) Horvatovich, P.; Lundberg, E. K.; Chen, Y. J.; Sung, T. Y.; He, F.; Nice, E. C.; Goode, R. J.; Yu, S.; Ranganathan, S.; Baker, M. S.; Domont, G. B.; Velasquez, E.; Li, D.; Liu, S.; Wang, Q.; He, Q. Y.; Menon, R.; Guan, Y.; Corrales, F. J.; Segura, V.; Casal, J. I.; Pascual-Montano, A.; Albar, J. P.; Fuentes, M.; Gonzalez-Gonzalez, M.; Diez, P.; Ibarrola, N.; Degano, R. M.; Mohammed, Y.; Borchers, C. H.; Urbani, A.; Soggiu, A.; Yamamoto, T.; Salekdeh, G. H.; Archakov, A.; Ponomarenko, E.; Lisitsa, A.; Lichti, C. F.;

Mostovenko, E.; Kroes, R. A.; Rezeli, M.; Vegvari, A.; Fehniger, T. E.; Bischoff, R.; Vizcaino, J. A.; Deutsch, E. W.; Lane, L.; Nilsson, C. L.; Marko-Varga, G.; Omenn, G. S.; Jeong, S. K.; Lim, J. S.; Paik, Y. K.; Hancock, W. S. *J Proteome Res* **2015**, *14*, 3415-3431.

(25) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2014**, *13*, 339-347.

(26) Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Nat Protoc* **2015**, *10*, 701-714.

(27) Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; Bui, H.; Specht, A.; Lange, O.; Denisov, E.; Makarov, A.; Horning, S.; Zabrouskov, V. *Anal Chem* **2013**, *85*, 11710-11714.

(28) Scheltema, R. A.; Hauschild, J. P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. *Mol Cell Proteomics* **2014**, *13*, 3698-3708.

(29) Kelstrup, C. D.; Jersie-Christensen, R. R.; Batth, T. S.; Arrey, T. N.; Kuehn, A.; Kellmann, M.; Olsen, J. V. *J Proteome Res* **2014**, *13*, 6187-6195.

(30) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J Am Soc Mass Spectrom* **1994**, *5*, 976-989.

(31) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. *Nat Methods* **2009**, *6*, 359-362.

(32) Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R. A. *Mol Cell Proteomics* **2013**, *12*, 3330-3338.

(33) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. *Nat Methods* **2014**, *11*, 319-324.

(34) Olsen, J. V.; Ong, S. E.; Mann, M. *Mol Cell Proteomics* **2004**, *3*, 608-614.

(35) Swaney, D. L.; Wenger, C. D.; Coon, J. J. *Journal of Proteome Research* **2010**, *9*, 1323-1329.

(36) Tsiatsiani, L.; Heck, A. J. R. *Febs J* **2015**, *282*, 2612-2626.

(37) Giansanti, P.; Aye, T. T.; van den Toorn, H.; Peng, M.; van Breukelen, B.; Heck, A. J. R. *Cell Reports* **2015**, *11*, 1834-1843.

(38) Gauci, S.; Helbig, A. O.; Slijper, M.; Krijgsveld, J.; Heck, A. J. R.; Mohammed, S. *Anal Chem* **2009**, *81*, 4493-4501.

(39) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. *Mol Cell Proteomics* **2014**, *13*, 1573-1584.

(40) Meyer, J. G.; Kim, S.; Maltby, D. A.; Ghassemian, M.; Bandeira, N.; Komives, E. A. *Mol Cell Proteomics* **2014**, *13*, 823-835.

(41) Wisniewski, J. R.; Mann, M. *Anal Chem* **2012**, *84*, 2631-2637.

(42) Sharma, K.; D'Souza, R. C. J.; Tyanova, S.; Schaab, C.; Wisniewski, J. R.; Cox, J.; Mann, M. *Cell Reports* **2014**, *8*, 1583-1594.

(43) Udeshi, N. D.; Svinkina, T.; Mertins, P.; Kuhn, E.; Mani, D. R.; Qiao, J. W.; Carr, S. A. *Molecular & Cellular Proteomics* **2013**, *12*, 825-831.

(44) Svinkina, T.; Gu, H. B.; Silva, J. C.; Mertins, P.; Qiao, J.; Fereshetian, S.; Jaffe, J. D.; Kuhn, E.; Udeshi, N. D.; Carr, S. A. *Molecular & Cellular Proteomics* **2015**, *14*, 2429-2440.

(45) Villen, J.; Gygi, S. P. *Nature Protocols* **2008**, *3*, 1630-1638.

(46) Thingholm, T. E.; Jorgensen, T. J. D.; Jensen, O. N.; Larsen, M. R. *Nature Protocols* **2006**, *1*, 1929-1935.

(47) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. *Nature Methods* **2014**, *11*, 319-U300.

(48) Wisniewski, J. R.; Dus, K.; Mann, M. *Proteom Clin Appl* **2013**, *7*, 225-233.

(49) Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. *Molecular & Cellular Proteomics* **2011**, *10*.

(50) Andrews, G. L.; Simons, B. L.; Young, J. B.; Hawkridge, A. M.; Muddiman, D. C. *Anal Chem* **2011**, *83*, 5442-5446.

(51) Helm, D.; Vissers, J. P. C.; Hughes, C. J.; Hahne, H.; Ruprecht, B.; Pachl, F.; Grzyb, A.; Richardson, K.; Wildgoose, J.; Maier, S. K.; Marx, H.; Wilhelm, M.; Becher, I.; Lemeer, S.; Bantscheff, M.; Langridge, J. I.; Kuster, B. *Molecular & Cellular Proteomics* **2014**, *13*, 3709-3715.

(52) Kocher, T.; Swart, R.; Mechtler, K. *Anal Chem* **2011**, *83*, 2699-2704.

(53) Cristobal, A.; Hennrich, M. L.; Giansanti, P.; Goerdayal, S. S.; Heck, A. J. R.; Mohammed, S. *Analyst* **2012**, *137*, 3541-3548.

(54) Hsieh, E. J.; Bereman, M. S.; Durand, S.; Valaskovic, G. A.; MacCoss, M. J. *J Am Soc Mass Spectr* **2013**, *24*, 148-153.

(55) Iwasaki, M.; Sugiyama, N.; Tanaka, N.; Ishihama, Y. *J Chromatogr A* **2012**, *1228*, 292-297.

(56) Syka, J. E. P.; Coon, J. J.; Schroeder, M. J.; Shabanowitz, J.; Hunt, D. F. *P Natl Acad Sci USA* **2004**, *101*, 9528-9533.

(57) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nature Methods* **2007**, *4*, 709-712.

(58) Swaney, D. L.; McAlister, G. C.; Coon, J. J. *Nature Methods* **2008**, *5*, 959-964.

(59) Frese, C. K.; Altelaar, A. F. M.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. *Journal of Proteome Research* **2011**, *10*, 2377-2388.

(60) Marx, H.; Lemeer, S.; Klaeger, S.; Rattei, T.; Kuster, B. *Journal of Proteome Research* **2013**, *12*, 2386-2398.

(61) Michalski, A.; Cox, J.; Mann, M. *Journal of Proteome Research* **2011**, *10*, 1785-1793.

(62) Batth, T. S.; Francavilla, C.; Olsen, J. V. *Journal of Proteome Research* **2014**, *13*, 6176-6186.

(63) Wang, H.; Sun, S. N.; Zhang, Y.; Chen, S.; Liu, P.; Liu, B. *J Chromatogr B* **2015**, *974*, 90-95.

(64) Wang, Y. X.; Yang, F.; Gritsenko, M. A.; Wang, Y. C.; Clauss, T.; Liu, T.; Shen, Y. F.; Monroe, M. E.; Lopez-Ferrer, D.; Reno, T.; Moore, R. J.; Klemke, R. L.; Camp, D. G.; Smith, R. D. *Proteomics* **2011**, *11*, 2019-2026.

(65) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nature Biotechnology* **2001**, *19*, 242-247.

(66) Ritorto, M. S.; Cook, K.; Tyagi, K.; Pedrioli, P. G.; Trost, M. *J Proteome Res* **2013**, *12*, 2449-2457.

(67) Bensaddek, D.; Nicolas, A.; Lamond, A. I. *International Journal of Mass Spectrometry* **2015**, *391*, 105-114.

(68) Zhou, F.; Lu, Y.; Ficarro, S. B.; Adelmant, G.; Jiang, W.; Luckey, C. J.; Marto, J. A. *Nat Commun* **2013**, *4*, 2171.

(69) Branca, R. M.; Orre, L. M.; Johansson, H. J.; Granholm, V.; Huss, M.; Perez-Bercoff, A.; Forshed, J.; Kall, L.; Lehtio, J. *Nat Methods* **2014**, *11*, 59-62.

(70) Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M. *Molecular & Cellular Proteomics* **2011**, *10*.

(71) Syka, J. E. P.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F. *Journal of Proteome Research* **2004**, *3*, 621-626.

(72) Hu, Q. Z.; Noll, R. J.; Li, H. Y.; Makarov, A.; Hardman, M.; Cooks, R. G. *J Mass Spectrom* **2005**, *40*, 430-443.

(73) Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Hernandez, P.; Appel, R. D.; Binz, P. A.; Goodlett, D. R. *J Am Soc Mass Spectr* **2008**, *19*, 891-901.

(74) Makarov, A. *Anal Chem* **2000**, *72*, 1156-1162.

(75) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A. *Molecular & Cellular Proteomics* **2012**, *11*.

(76) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. *Molecular & Cellular Proteomics* **2012**, *11*.

(77) Sleno, L.; Volmer, D. A. *J Mass Spectrom* **2004**, *39*, 1091-1112.

(78) Louris, J. N.; Cooks, R. G.; Syka, J. E. P.; Kelley, P. E.; Stafford, G. C.; Todd, J. F. J. *Anal Chem* **1987**, *59*, 1677-1685.

(79) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J Am Chem Soc* **1998**, *120*, 3265-3266.

(80) Riley, N. M.; Rush, M. J.; Rose, C. M.; Richards, A. L.; Kwiecien, N. W.; Bailey, D. J.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2015**, *14*, 2644-2660.

(81) Medzihradszky, K. F.; Chalkley, R. J. *Mass Spectrom Rev* **2015**, *34*, 43-63.

(82) Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. *Journal of Proteome Research* **2008**, *7*, 29-34.

(83) Elias, J. E.; Gygi, S. P. *Nat Methods* **2007**, *4*, 207-214.

(84) Savitski, M. M.; Wilhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. *Mol Cell Proteomics* **2015**, *14*, 2394-2404.

(85) Serang, O.; Kall, L. *J Proteome Res* **2015**, *14*, 4099-4103.

(86) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *Proc Natl Acad Sci U S A* **1999**, *96*, 6591-6596.

(87) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol Cell Proteomics* **2002**, *1*, 376-386.

(88) Thompson, A.; Schafer, J.; Kuhn, K.; Kienle, S.; Schwarz, J.; Schmidt, G.; Neumann, T.; Johnstone, R.; Mohammed, A. K.; Hamon, C. *Anal Chem* **2003**, *75*, 1895-1904.

(89) Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Mol Cell Proteomics* **2004**, *3*, 1154-1169.

(90) Frost, D. C.; Greer, T.; Li, L. *Anal Chem* **2015**, *87*, 1646-1654.

(91) McAlister, G. C.; Huttlin, E. L.; Haas, W.; Ting, L.; Jedrychowski, M. P.; Rogers, J. C.; Kuhn, K.; Pike, I.; Grothe, R. A.; Blethrow, J. D.; Gygi, S. P. *Anal Chem* **2012**, *84*, 7469-7478.

(92) Merrill, A. E.; Hebert, A. S.; MacGilvray, M. E.; Rose, C. M.; Bailey, D. J.; Bradley, J. C.; Wood, W. W.; El Masri, M.; Westphall, M. S.; Gasch, A. P.; Coon, J. J. *Molecular & Cellular Proteomics* **2014**, *13*, 2503-2512.

(93) Rhoads, T. W.; Prasad, A.; Kwiecien, N. W.; Merrill, A. E.; Zawack, K.; Westphall, M. S.; Schroeder, F. C.; Kimble, J.; Coon, J. J. *Molecular & Cellular Proteomics* **2015**, *14*, 2922-2935.

(94) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. *Nature Methods* **2013**, *10*, 332-+.

(95) Lundgren, D. H.; Hwang, S. I.; Wu, L. F.; Han, D. K. *Expert Rev Proteomic* **2010**, *7*, 39-53.

(96) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. *Molecular & Cellular Proteomics* **2014**, *13*, 2513-2526.

(97) Cox, J.; Mann, M. *Nature Biotechnology* **2008**, *26*, 1367-1372.

(98) D'Souza, R. C. J.; Knittle, A. M.; Nagaraj, N.; van Dinther, M.; Choudhary, C.; ten Dijke, P.; Mann, M.; Sharma, K. *Science Signaling* **2014**, *7*.

(99) Webb-Robertson, B. J. M.; Matzke, M. M.; Datta, S.; Payne, S. H.; Kang, J. Y.; Bramer, L. M.; Nicora, C. D.; Shukla, A. K.; Metz, T. O.; Rodland, K. D.; Smith, R. D.; Tardiff, M. F.; McDermott, J. E.; Pounds, J. G.; Waters, K. M. *Molecular & Cellular Proteomics* **2014**, *13*, 3639-3646.

(100) Cullen, L. M.; Arndt, G. M. *Immunol Cell Biol* **2005**, *83*, 217-223.

(101) Kemmeren, P.; Sameith, K.; van de Pasch, L. A. L.; Benschop, J. J.; Lenstra, T. L.; Margaritis, T.; O'Duibhir, E.; Apweiler, E.; van Wageningen, S.; Ko, C. W.; van Heesch, S.; Kashani, M. M.; Ampatziadis-Michailidis, G.; Brok, M. O.; Brabers, N. A. C. H.; Miles, A. J.; Bouwmeester, D.; van Hooff, S. R.; van Bakel, H.; Sluiters, E.; Bakker, L. V.; Snel, B.; Lijnzaad, P.; van Leenen, D.; Koerkamp, M. J. A. G.; Holstege, F. C. P. *Cell* **2014**, *157*, 740-752.

(102) Liu, Y.; Beyer, A.; Aebersold, R. *Cell* **2016**, *165*, 535-550.

(103) Pichler, P.; Kocher, T.; Holzmann, J.; Mazanek, M.; Taus, T.; Ammerer, G.; Mechtler, K. *Anal Chem* **2010**, *82*, 6549-6558.

(104) Sandberg, A.; Branca, R. M. M.; Lehtio, J.; Forshed, J. *Journal of Proteomics* **2014**, *96*, 133-144.

(105) Renuse, S.; Chaerkady, R.; Pandey, A. *Proteomics* **2011**, *11*, 620-630.

(106) Khatun, J.; Yu, Y. B.; Wrobel, J. A.; Risk, B. A.; Gunawardena, H. P.; Secrest, A.; Spitzer, W. J.; Xie, L.; Wang, L.; Chen, X.; Giddings, M. C. *Bmc Genomics* **2013**, *14*.

(107) Hartmann, E. M.; Armengaud, J. *Proteomics* **2014**, *14*, 2637-2646.

(108) Low, T. Y.; van Heesch, S.; van den Toorn, H.; Giansanti, P.; Cristobal, A.; Toonen, P.; Schafer, S.; Hubner, N.; van Breukelen, B.; Mohammed, S.; Cuppen, E.; Heck, A. J. R.; Guryev, V. *Cell Reports* **2013**, *5*, 1469-1478.

(109) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. *Journal of Proteome Research* **2014**, *13*, 228-240.

(110) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. *Molecular Cell* **2013**, *49*, 583-590.

(111) Munoz, J.; Heck, A. J. R. *Angew Chem Int Edit* **2014**, *53*, 10864-10866.

(112) Paik, Y. K.; Hancock, W. S. *Nature Biotechnology* **2012**, *30*, 1065-1067.

(113) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Molecular & Cellular Proteomics* **2013**, *12*, 3465-3473.

(114) Farrah, T.; Deutsch, E. W.; Omenn, G. S.; Sun, Z.; Watts, J. D.; Yamamoto, T.; Shteynberg, D.; Harris, M. M.; Moritz, R. L. *Journal of Proteome Research* **2014**, *13*, 60-75.

(115) Ponten, F.; Gry, M.; Fagerberg, L.; Lundberg, E.; Asplund, A.; Berglund, L.; Oksvold, P.; Bjorling, E.; Hober, S.; Kampf, C.; Navani, S.; Nilsson, P.; Ottosson, J.; Persson, A.; Wernerus, H.; Wester, K.; Uhlen, M. *Molecular Systems Biology* **2009**, *5*.

Chapter 2

SIRT3 Mediates Multi-tissue Coupling for Metabolic Fuel Switching

This chapter has been published:

Dittenhafer-Reed KE*, **Richards AL***, Fan J, Smallegan MJ, Siahpirani AF, Kemmerer ZA, Prolla TA, Roy S, Coon JJ, Denu JM. *SIRT3 Mediates Multi-tissue Coupling for Metabolic Fuel Switching*. Cell Metabolism. **2015,** *21*, 637-646.

*Authors contributed equally

ABSTRACT

SIRT3 is a member of the Sirtuin family of NAD+ - dependent deacylases and plays a critical role in metabolic regulation. Organism-wide SIRT3 loss manifests in metabolic alterations; however, the coordinating role of SIRT3 among metabolically distinct tissues is unknown. Using multi-tissue quantitative proteomics comparing fasted wild-type mice to mice lacking SIRT3, innovative bioinformatic analysis, and biochemical validation, we provide a comprehensive view of mitochondrial acetylation and SIRT3 function. We find SIRT3 regulates the acetyl-proteome in core mitochondrial processes common to brain, heart, kidney, liver, and skeletal muscle, but differentially regulates metabolic pathways in fuel-producing and fuel-utilizing tissues. We propose an additional maintenance function for SIRT3 in liver and kidney where SIRT3 expression is elevated to reduce the acetate load on mitochondrial proteins. We provide evidence that SIRT3 impacts ketone body utilization in the brain and reveal a pivotal role for SIRT3 in the coordination between tissues required for metabolic homeostasis.

INTRODUCTION

Mammals must maintain whole-body homeostasis with inconsistent fuel intake and have developed mechanisms to manage times of fuel deprivation. The production, utilization, and storage of carbohydrates, fatty acids, and protein are regulated in a tissue-dependent manner, as each organ has unique functions, metabolic pathways, and accessibility to fuel sources. Most tissues exhibit metabolic flexibility and are capable of using multiple types of fuels in response to shifts in nutrient availability. The regulation of fuel switching occurs at both the cellular and organismal level, and requires tissues to respond in concert to metabolic signals to maintain proper function.[1] As the mitochondria are the energy production centers of the cell, mitochondrial metabolism is vital in the adaptation to alterations in nutrient supply.

Protein acetylation is a post-translational modification (PTM) enriched in the mitochondria of multiple tissue types and is involved in numerous cellular processes through regulation of protein interactions, activity, and localization.[2,3] Mass spectrometry (MS)-based proteomics has identified over 2,500 unique acetyl sites in the mitochondria, a majority of which reside on metabolic enzymes.[2,4,5] Much of our knowledge on mitochondrial protein acetylation comes from studies on the role of the NAD+ - dependent deacetylase Sirtuin 3 (SIRT3) in liver mitochondria.[4,6] In the liver, SIRT3 regulates mitochondrial function and specifically modulates enzymatic activity of proteins involved in fatty acid oxidation, oxidative phosphorylation, ketone body synthesis, and the urea cycle.[4,7] Additionally, SIRT3 is thought to play a key role in regulating the production of reactive oxygen species.[7] SIRT3 is proposed to facilitate catabolism of fatty acids in the liver and the peripheral use of acetate during fasting.[8,9] However, molecular details of the role of SIRT3 in extra-hepatic tissues and in managing coordinated whole-body responses to fuel availability remain unclear.

To assess the role of SIRT3 across tissues and identify tissue specific SIRT3 substrates and regulated biological pathways, we employed a quantitative acetyl-proteomic method[4] to study SIRT3 in five tissues (brain, heart, kidney, liver, and skeletal muscle) from fasted mice that were wild-type (WT) or lacking SIRT3 (*Sirt3*$^{-/-}$). We provide a comprehensive multi-tissue quantitative acetyl-proteome analysis comparing multiple biological conditions. We identified 6,286 acetyl sites on 1,172 proteins of which nearly

4,000 sites were localized to mitochondrial proteins, providing a compendium of acetyl sites that are dynamically altered by SIRT3 in tissues that, until now, remained unexplored with respect to the function of SIRT3. A biological pathway analysis tailored to quantitative PTM data was developed and employed, allowing for an assessment of pathways regulated by acetylation and SIRT3. Bioinformatics analyses of our proteome data reveal fuel-producing (liver and kidney) and fuel-utilizing tissues (brain, heart, skeletal muscle) display unique, SIRT3-dependent alterations in their acetyl proteome, but also indicate SIRT3 regulates acetylation of proteins involved in common core mitochondrial processes among diverse tissues. We provide biochemical evidence that SIRT3 is required for utilization of ketone bodies to form acetyl-CoA in extra-hepatic tissues. We establish critical tissue-dependent roles for SIRT3 and reveal SIRT3 dictates multi-tissue coupling required for metabolic adaptation to nutrient availability.

RESULTS

Acetyl-Proteomics Quantifies Site-Specific Changes in Acetylation from Multiple Tissues in Mice Lacking SIRT3. We applied our quantitative acetyl-proteomic method to study the tissue-specific and SIRT3-dependent alterations in the proteome and acetylome (**Figure 1*A***).[4] Three 5-month-old WT and three germline *Sirt3*$^{-/-}$ mice were subjected to a 24-hr fast beginning the morning prior to tissue harvest. To minimize compensatory or secondary mechanisms that occur as mice age, we selected young mice to reveal primary tissue-specific biochemical alterations in the acetyl-proteome that have not yet manifested as an overt phenotype. We compared, in biological triplicate, the whole brain, heart, kidney, liver, and skeletal muscle proteome and acetyl-proteome in two conditions: WT and *Sirt3*$^{-/-}$ (**Figure 1*A***). We detected 6,286 acetyl sites and quantified 5,199 acetylation sites in five tissues (**Table 1**). Nearly half of the quantified acetyl sites (2,247) were localized on mitochondrial proteins,[10] achieving deep quantification of the mitochondrial proteome (668 proteins) (**Table 1**). To confirm differences in acetylation are due to increased acetylation, rather than a result of increased protein abundance, we analyzed the un-modified proteome of each tissue. Acetyl reporter ion intensities were corrected for protein amount in each condition, and used to calculate a protein normalized fold change between *Sirt3*$^{-/-}$ and WT. Of the acetyl sites quantified, 38%

Figure 1. Quantitative mapping of dynamic acetylation in multiple tissues. (***A***) Wild type and *Sirt3*$^{-/-}$ mouse brain, heart, kidney, liver, and skeletal muscle were compared in biological triplicate. Tissues were disrupted by bead milling and extracts were digested. Peptides were labeled with TMT reagents, combined and fractionated by strong cation exchange chromatography. Acetyl peptides were enriched by immunoprecipitation and analyzed via nano-RPLC MS/MS on an Orbitrap Elite. (***B***) Percentage of mitochondrial acetyl sites changing ≥ 2-fold in the *Sirt3*$^{-/-}$ condition for each tissue. Calculated as number of mitochondrial acetyl sites changing ≥ 2-fold per total number of quantified mitochondrial acetyl sites found in that tissue. (***C***) Venn diagram displaying overlapping acetyl sites between tissues.

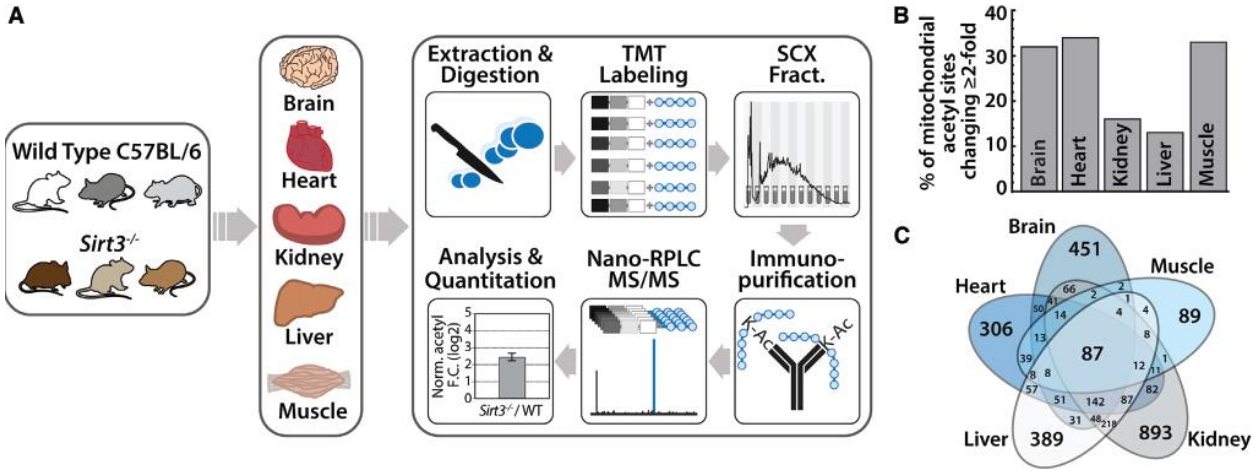


Table 1. Acetyl-proteomic metrics for each tissue. Acetyl sites with statistically significant fold changes ($p \leq 0.1$, Student's t test) between *Sirt3*$^{-/-}$ and WT are listed in parentheses next to total number of acetyl sites. Mitochondrial proteins categorized by inclusion in the MitoCarta inventory.[10]

Tissue	Brain	Heart	Kidney	Liver	Muscle
acetyl sites	1,247 (465)	1,159 (554)	1,370 (317)	2,117 (578)	393 (74)
mitochondrial acetyl sites	715 (339)	897 (499)	908 (255)	1400 (494)	207 (72)
acetyl proteins	523	338	500	647	152
mitochondrial acetyl proteins	220	230	268	331	90
total proteins	6,783	3,439	3,977	4,542	2,290
mitochondrial proteins	665	646	640	671	457
acetyl sites / protein	2.39	3.43	2.74	3.28	2.59

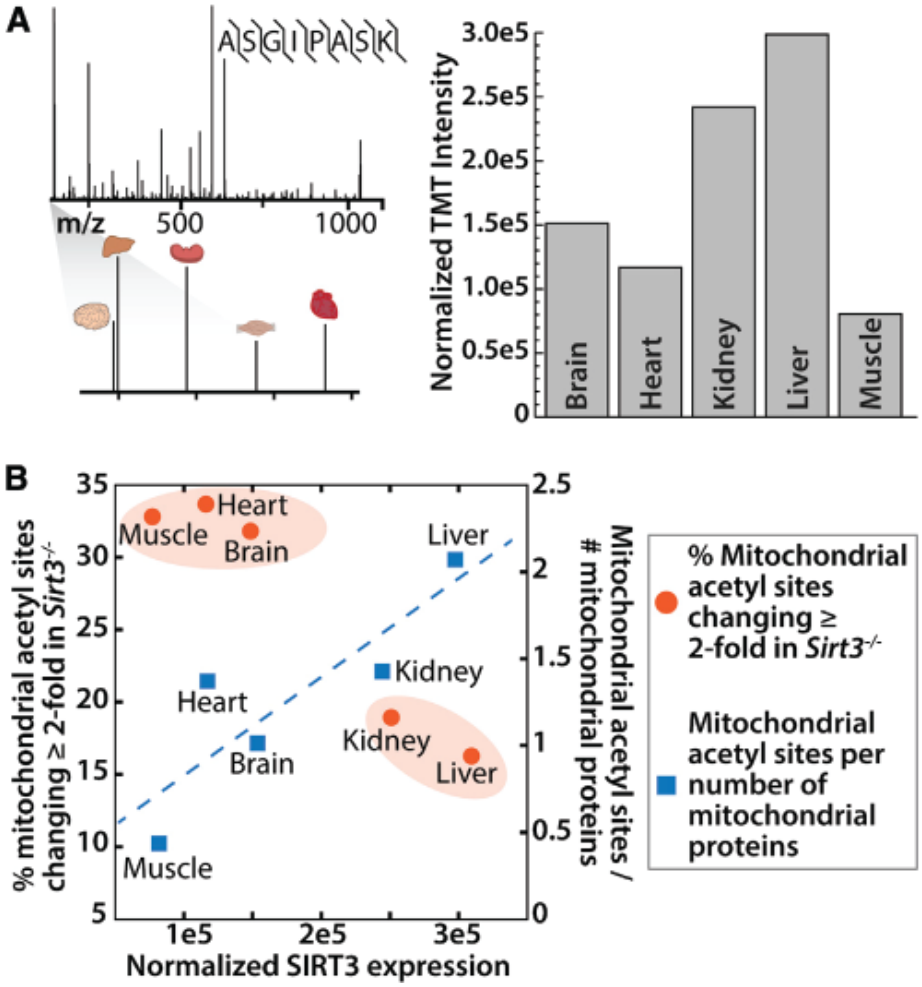
(1,990 acetyl sites) exhibited a statistically significant change (p % 0.1, Student's t test) in acetyl occupancy in mice lacking SIRT3. Our experimental approach analyzed the acetyl proteome and proteome of the entire tissue, allowing for an assessment of the potential effects of SIRT3 on acetylation outside of the mitochondria. An analysis of the distribution in fold change of acetyl sites between the *Sirt3*$^{-/-}$ and WT animals reveals a normal distribution for both mitochondrial and non-mitochondrial populations and a marked shift toward hyper acetylation of mitochondrial acetyl sites when compared to non-mitochondrial acetyl sites. There has been some debate on the role of SIRT3 outside of the mitochondria;[11] however, in the tissues tested, our study indicates the primary deacetylase activity of SIRT3 resides in the mitochondria. As previously reported for the liver,[4] SIRT3 does not considerably alter the mitochondrial or non-mitochondrial proteome; thus, changes in mitochondrial function are due primarily to alterations in acetylation, and not protein abundance.

The percent of mitochondrial acetyl sites changing greater than 2-fold in mice lacking SIRT3 was calculated (**Figure 1*B***). As SIRT3 is the only established mitochondrial deacetylase, hyper-acetylated lysine residues in *Sirt3*$^{-/-}$ mice are probable SIRT3 targets. The effect of SIRT3 on mitochondrial acetylation varies in a tissue-specific manner, with brain, heart, and muscle displaying the greatest percentage (33%) of acetyl isoforms increasing greater than 2-fold in the *Sirt3*$^{-/-}$ condition (**Figure 1*B***). Interestingly, liver and kidney contain the highest total number of mitochondrial acetyl sites (**Table 1**), but a lower percentage of acetyl sites that exhibit greater than 2-fold increases in acetylation in *Sirt3*$^{-/-}$. Liver and kidney contain a number of acetyl sites that change less dramatically (between 1- and 2-fold) in the absence of SIRT3. To determine how differences in SIRT3 protein expression across tissues dictate acetylation changes during fasting, we measured SIRT3 protein abundance by targeted MS using WT samples. Peptides from the brain, heart, kidney, liver, and skeletal muscle of one WT mouse were labeled with tandem mass tags (TMTs, Pierce) and combined in equal ratios. Parallel reaction monitoring (PRM)[12] was performed to detect a single, unique peptide for SIRT3 (ASGIPASK, amino acids 171–178). The relative TMT intensities for each tissue were normalized to total intensity across the TMT channel and plotted (**Figure 2*A***) and confirmed by western blotting of samples from three WT study animals. Liver and kidney have the highest

protein expression levels of SIRT3, followed by brain, heart, and skeletal muscle, comparable to previous literature reports.[13,14] Interestingly, the tissues with the highest levels of SIRT3 protein (liver and kidney) exhibit the lowest percentage of mitochondrial acetyl sites changing greater than 2-fold in the absence of SIRT3, but the greatest number of total mitochondrial acetyl sites identified per mitochondrial protein (**Figure 2*B***). The percentage of mitochondrial sites changing greater than 2-fold was used in this analysis to provide a normalization factor for the differences in total number of acetyl sites found per tissue. There is a positive linear correlation between SIRT3 expression and total number of mitochondrial acetyl sites identified in each tissue (**Figure 2*B***). However, when examining only sites that are dramatically altered, changing greater than 2-fold in *Sirt3*$^{-/-}$, there is segregation between liver and kidney, and muscle, heart, and brain (**Figure 2*B***). Overall, in tissues with lower SIRT3 expression (brain, heart, skeletal muscle) there are fewer, but more dynamic, changes in acetylation after a 24-hr fast when compared to the corresponding tissues of the *Sirt3*$^{-/-}$mice. These results provided a rich resource to explore the tissue specific functions of SIRT3 in whole-body metabolism.

Tissue-Specific Patterns of Acetyl Site Distribution. To determine whether acetylated proteins could be categorized into sub-populations based on their patterns across the five tissues, we implemented a probabilistic clustering algorithm based on a Gaussian mixture model.[4,15] The algorithm used the summed acetyl fold change for each unique UniProt identifier and generated five clusters (**Figure 3*A***) that segregate into two main groups: (1) acetylated proteins regulated by SIRT3 that exhibit hyper-acetylation in SIRT3-null mice and (2) proteins with minimal perturbations in acetylation due to SIRT3. We postulated that these two groups might represent functionally distinct proteins. An interrogation of the sub-cellular localization of proteins found in these groups using the functional annotation tool DAVID[16,17] against a background of all acetyl-proteins used in the clustering analysis revealed group 1 proteins are enriched for mitochondrial proteins ($p = 1.10 \times 10^{-38}$, Benjamini corrected), while group 2 proteins are dispersed throughout the cell (nucleus, $p = 2.70 \times 10^{-8}$; intracellular non-membrane-bound organelle, $p = 1.80 \times 10^{-6}$). Acetyl-proteins in group 1 are significantly enriched in metabolic processes ($p = 2.20 \times 10^{-5}$) in the mitochondria including

Figure 2. SIRT3 expression varies between tissues and corresponds to tissue acetylome alterations. (***A***) PRM mass spectrometry and TMT quantitation were employed to assess SIRT3 protein expression in each of the five tissues studied. TMT-labeled peptides from one WT animal from each tissue were compared in a single targeted MS experiment. Spectra from a represented peptide ASGIPASK with reporter ion magnification and protein fold change for each tissue. (***B***) A comparison of SIRT3 expression level (x-axis) with percentage of mitochondrial sites changing ≥ 2-fold in response to *Sirt3*$^{-/-}$ (y1) and mitochondrial acetyl sites per the number of mitochondrial proteins identified in that tissue (y2).

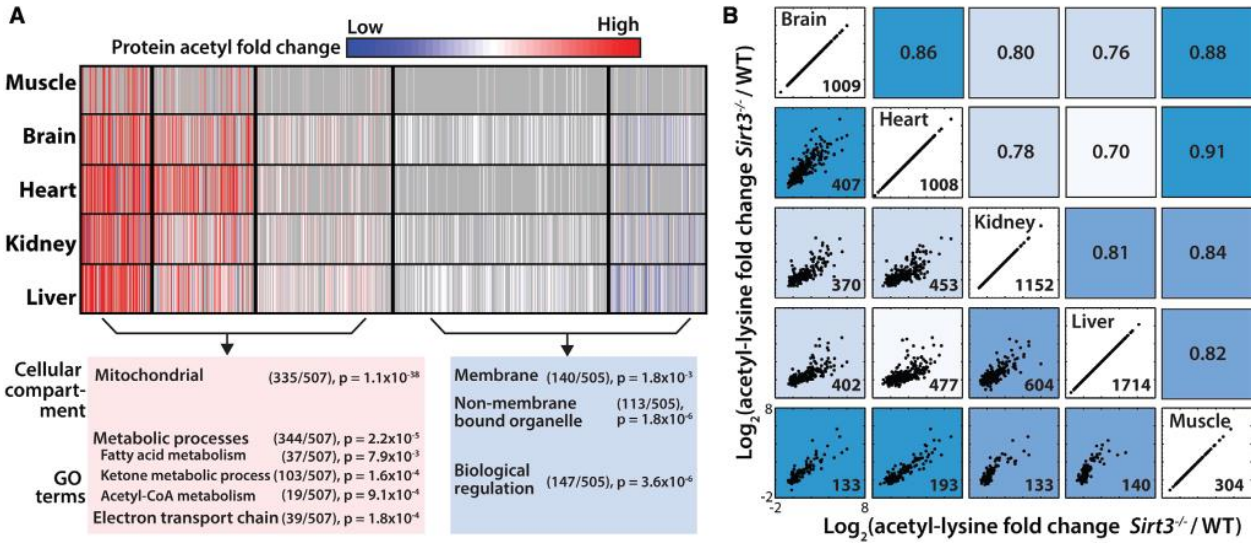


on a Gaussian mixture model.[4,15] The algorithm used the summed acetyl fold change for each unique UniProt identifier and generated five clusters (**Figure 3*A***) that segregate into two main groups: (1) acetylated proteins regulated by SIRT3 that exhibit hyper-acetylation in SIRT3-null mice and (2) proteins with minimal perturbations in acetylation due to SIRT3. We postulated that these two groups might represent functionally distinct proteins. An interrogation of the sub-cellular localization of proteins found carboxylic acid, fatty acid, ketone body, and acetyl-CoA metabolic processes, as well as the electron transport chain ($p = 1.20 \times 10^{-4}$). Group 2 proteins are enriched in regulation of biological processes ($p = 5.20 \times 10^{-7}$) and are not enriched in mitochondrial pathways. These results indicate group 1 proteins comprise potential SIRT3 targets. We also observed patterns between tissues within the clusters (**Figure 2*A***). To determine tissue similarities of proteins whose acetylation is regulated by SIRT3, we performed Euclidean distance hierarchical clustering on each cluster using an algorithm that was adapted to take into consideration missing data values resulting from incomplete overlap of acetyl sites and proteins (**Figure 1*C***) across all five tissues. Functionally similar tissues, like liver and kidney, and muscle and heart, are more closely related to each other in every cluster of acetyl-proteins. The segregation of fuel-utilizing (brain, heart, muscle) and fuel-producing tissues (liver and kidney) in the *Sirt3*$^{-/-}$ condition suggest SIRT3 regulates protein acetylation in a protein- and organ-dependent manner. We predicted that acetyl sites would also display tissue similarities and would further reveal tissue-dependent response to a loss of SIRT3 expression. To assess tissue similarities at the acetyl site level, normalized acetyl fold changes for sites found in common between any two tissues were plotted (**Figure 3*B***). The Pearson correlation coefficient was calculated and is presented on the right side of the diagonal of the scatter plots in **Figure 3*B***. The greater the Pearson correlation coefficient, the more similar acetyl sites are changing due to SIRT3 in any two tissues. Tissue similarities emerge that are consistent with our protein level clustering analysis, with heart, brain, and muscle more similar to each other than liver and kidney.

Pathway Analysis Tool Identifies Biological Pathways Regulated by Acetylation and SIRT3. The adaptation to diet and nutrient availability requires a coordinated change in mitochondrial metabolism.

These metabolic shifts are tissue specific, but require exquisite coupling of tissues to allow for fuel switching and metabolic homeostasis. The extent to which SIRT3 regulates divergent pathways of nutrient utilization and mediates metabolic coupling to respond to fuel availability is unknown. To identify biological pathways and understand the tissue-specific functions of SIRT3-controlled acetyl sites, we developed and applied a biological pathway analysis we termed quantitative site set functional score analysis (QSSA). Enrichment analysis based on gene ontologies is a steadfast companion in determining the biological significance of a particular physiological perturbation and is used to generate hypotheses on the functional relevance of the observed data. Most established enrichment methods were developed for gene expression data and later applied to proteomics data. Therefore the information contained in typical over-representation analysis and newer approaches such as gene set enrichment analysis (GSEA) is gene-product centric.[18] Using available methods with quantitative PTM data necessitates discarding site-specific modification information. Over-representation analysis, where an arbitrary cutoff often based on fold change is established, entails further discarding the magnitude of the effect and all information about proteins that fall below the cutoff. GSEA and its derivatives offer an improvement in that they allow for the incorporation of protein fold changes and inclusion of an entire proteomic dataset in assigning significance to each set of genes, but are still incapable of accounting for the wealth of quantitative MS data on PTMs. QSSA shifts the focus of the enrichment analysis from sets of genes to sets of PTM sites. We start from the same pre-determined categories of genes used in previous approaches (Gene Ontology, Kyoto Encyclopedia of Genes and Genomes [KEGG],[19] Reactome, etc.) and incorporate site-specific data by integrating protein sequence data from the UniProt database. QSSA is not a rigorous method for determining the statistical significance of enrichment, but in accordance with the exploratory nature of proteome-wide datasets, is a tailored approach incorporating the most salient factors in quantitative PTM data to facilitate generation of biological hypotheses. Our method is broadly applicable to diverse PTM datasets, including quantitative phospho-proteomics, and the scoring algorithm is not limited to the KEGG database. Any well-annotated gene ontology or molecular function database could be employed. Since our

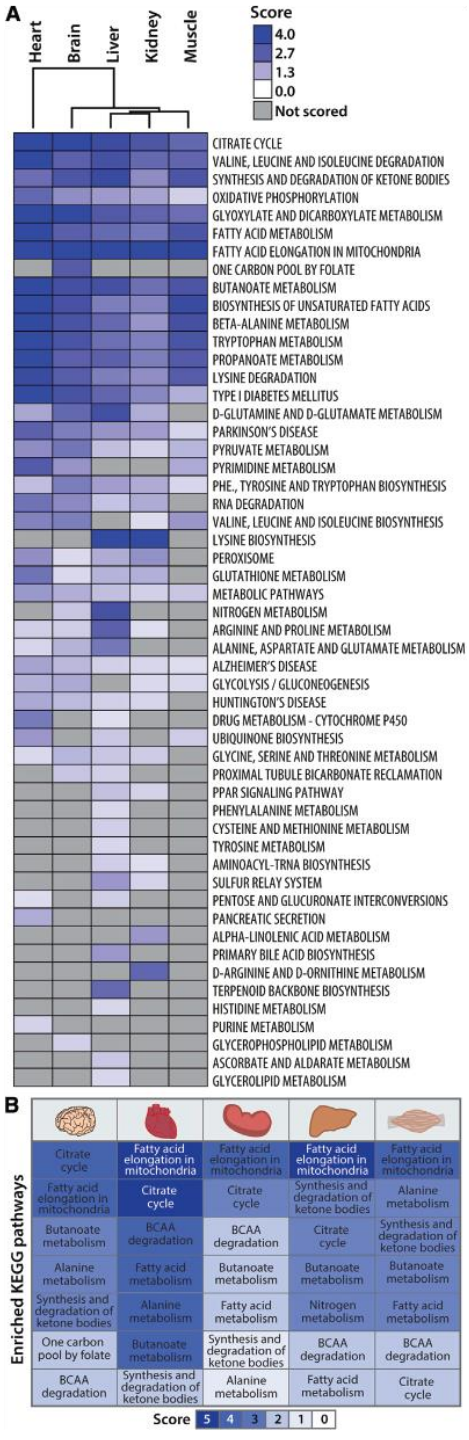
Figure 3. Tissue specific patterns of acetylated proteins and tissue correlation analysis. (***A***) Cluster analysis of acetyl-proteins identified 5 clusters that define 2 groups based on their overall trends between *Sirt3*$^{-/-}$ and WT. Acetyl-proteins are color coded according to their normalized fold change between *Sirt3*$^{-/-}$ and WT with cluster separations are denoted by black lines. Each row represents a tissue and each column represents an acetylated protein. Red: Increased acetylation, Blue: Decreased acetylation, Gray: Missing. Pathway analysis, number of acetyl proteins and p-values for enrichment are shown for two main groups of acetyl-proteins. (***B***) Normalized acetyl fold changes plotted for sites found in common between any two tissues. Scatter plots on the left side of the diagonal with the number of sites quantified and plotted in the lower right hand corner. Corresponding Pearson correlation coefficients are indicated on the right side of the diagonal. Tissues with acetyl sites behaving more similarly in response to *Sirt3*$^{-/-}$ are more highly correlated and are a darker shade of blue.



focus is lysine acetylation, we consider each pathway background as the set of lysines contained in proteins identified in our proteomics analysis belonging to a gene category. Acetylation may be important to the function of a biological process through the dynamic acetylation of multiple weakly regulated sites on a protein or through the regulation of one keystone site. To account for both of these possibilities, we assessed the acetyl coverage of a pathway and the fold change in acetylation of each acetyl site in the *Sirt3*$^{-/-}$ condition. To identify biological pathways that may be differentially regulated by SIRT3, QSSA was performed to calculate a standard score for each KEGG pathway, with higher scores representing pathways likely regulated by acetylation (**Figure 4**).

Biological pathway analysis indicated SIRT3 and acetylation regulate core mitochondrial pathways common to most tissues studied and pathways specific to individual or groups of metabolically similar tissues. Fundamental mitochondrial processes, including oxidative phosphorylation, the citrate cycle, and fatty acid metabolism, are enriched with acetyl sites sensitive to SIRT3 expression in brain, heart, kidney, liver, and skeletal muscle (**Figure 4*A***). These core pathways are essential for cellular function and represent common regulatory targets of SIRT3 across tissues. In support of this observation, of the 87 acetyl sites found in all tissues, nearly half are members of core mitochondrial pathways. These include multiple subunits of ATP synthase, complexes I and II of the electron transport chain, and numerous enzymes of the citrate cycle (citrate synthase, malate dehydrogenase, aconitase, succinate dehydrogenase, and isocitrate dehydrogenase). To illustrate the utility of the QSSA method, we directly compared KEGG pathway enrichment scores from QSSA with KEGG pathway fold enrichment scores from the commonly used functional annotation tool DAVID.[16,17] Validating the general applicability of QSSA, a number of high-scoring QSSA pathways were also identified in the DAVID analysis, including the citrate cycle and branched-chain amino acid catabolism. A side-by-side comparison of the scores provided for each pathway reveals that QSSA provides information on the extent of enrichment for many pathways for which DAVID provides no information, thus providing a more nuanced and fine-grained hypothesis-generating tool for quantitative MS-based PTM analysis. In fact, ketone body synthesis and utilization, a pathway for which

Figure 4. Novel pathway enrichment analysis tool identifies biological pathways regulated by SIRT3. (***A***) QSSA performed with all quantified acetyl sites identifies high scoring biological pathways enriched with acetyl sites and controlled by SIRT3. Hierarchical clustering groups biological pathways between fuel-utilizing and fuel-producing tissues. Dark blue: Highly scoring pathway likely regulated by SIRT3; Light blue: Low scoring pathway. (***B***) High scoring KEGG pathways are listed with their corresponding scores color coded according to scale.



we present detailed biochemical follow-up in the brain, was not identified as enriched using DAVID. Mitochondrial metabolism must also be regulated in a tissue-dependent manner to allow for metabolic flexibility. In response to a fast, fuel-producing pathways, such as fatty acid oxidation and ketone body generation, must be upregulated in hepatic tissues, while fuel-utilizing pathways, like acetyl-CoA production for the citrate cycle, are necessary in extra-hepatic tissues. To determine whether SIRT3 regulated biological processes in a tissue-dependent manner, calculated pathway scores were visualized by Euclidean distance hierarchical clustering (**Figure 4*A***), identifying segregation in pathway-scoring patterns between fuel-utilizing and fuel-producing tissues. We find a number of SIRT3-regulated pathways in brain, heart, and skeletal muscle that promote alternative fuel utilization, including degradation of short chain carbon molecules (propanoate and butanoate metabolism) and metabolism of amino acids (**Figure 4*B***). Skeletal muscle is a predominant location for branched-chain amino acid catabolism, providing succinyl-CoA for the citrate cycle and alanine that is shuttled to the liver to provide carbon for gluconeogenesis.[20] Both branched-chain amino acid degradation in the skeletal muscle and alanine metabolism in the liver and skeletal muscle were identified as probable metabolic pathways controlled by SIRT3, suggesting SIRT3 may be an important regulator of the glucose-alanine cycle during fasting metabolism. During fasting, another essential metabolic transition is the generation of ketone bodies by the liver and their utilization as an alternative fuel source by extra-hepatic tissues. Ketone body synthesis and utilization was identified as one of the highest candidate SIRT3- controlled pathways in brain, heart, skeletal muscle, and liver (***Figure 4B***). SIRT3 is known to regulate ketone body production in the liver through the deacetylation and activation of hydroxymethylglutaryl-CoA synthase 2 (HMGCS2);[21,22] however, whether SIRT3 mediates ketone body utilization in extra-hepatic tissues was unclear.

SIRT3 Mediates Cross-Tissue Coupling for Ketone Body Synthesis and Degradation. To provide evidence that SIRT3 differentially affects ketone body metabolism in hepatic and extra-hepatic tissues, we determined acetoacetate utilization and acetyl-CoA generation in the brain. This was accomplished by monitoring acetoacetate-dependent acetyl-CoA production in homogenates of brain cortices of 8-month-

old WT and *Sirt3*$^{-/-}$ mice that were fasted for 24 hr. The brain is an extra-hepatic tissue that relies solely on glucose and ketone bodies during a fast; therefore, the brain serves as the model tissue to study the effects of SIRT3-controlled protein acetylation on ketone body degradation. Using brain cortex homogenate from fasted WT or *Sirt3*$^{-/-}$ mice and quantitative LC-MS-based analysis, we monitored the production of acetylCoA from the addition of the ketone body acetoacetate (**Figures 5*A***). Succinyl-CoA and CoA were included in the reaction to ensure that flux through the pathway was not initially limited by the presence of required co-substrates. We find a 3-fold decrease in acetoacetate-dependent acetyl-CoA production in *Sirt3*$^{-/-}$ brain cortex homogenates at low succinyl-CoA levels (**Figure 5*A***), suggestive of a defect in acetyl-CoA production through the ketone body utilization pathway in mice lacking SIRT3 (**Figure 5*C***). We performed a targeted LC-MS-based metabolite analysis on WT and *Sirt3*$^{-/-}$ brain. Consistent with an alteration in ketone body utilization, a slight decrease in whole-cell acetyl-CoA was measured in the *Sirt3*$^{-/-}$ condition (**Figure 5*B***). Although this trend did not reach statistical significance, the observation is similar to our previously published report that acetyl-CoA levels were lower in liver of *Sirt3*$^{-/-}$ mice compared to WT.[4] The observation that reduced activity in ketone body utilization presented in **Figure 5*A*** is more dramatic than the slight reduction in whole-tissue acetyl-CoA levels (**Figure 5*B***) is not surprising and suggests that other compensating activities, including amino acid or fatty acid metabolism, reduction in acetyl-CoA-consuming processes, or other non-mitochondrial metabolic processes may prevent cellular acetyl-CoA from further depletion.

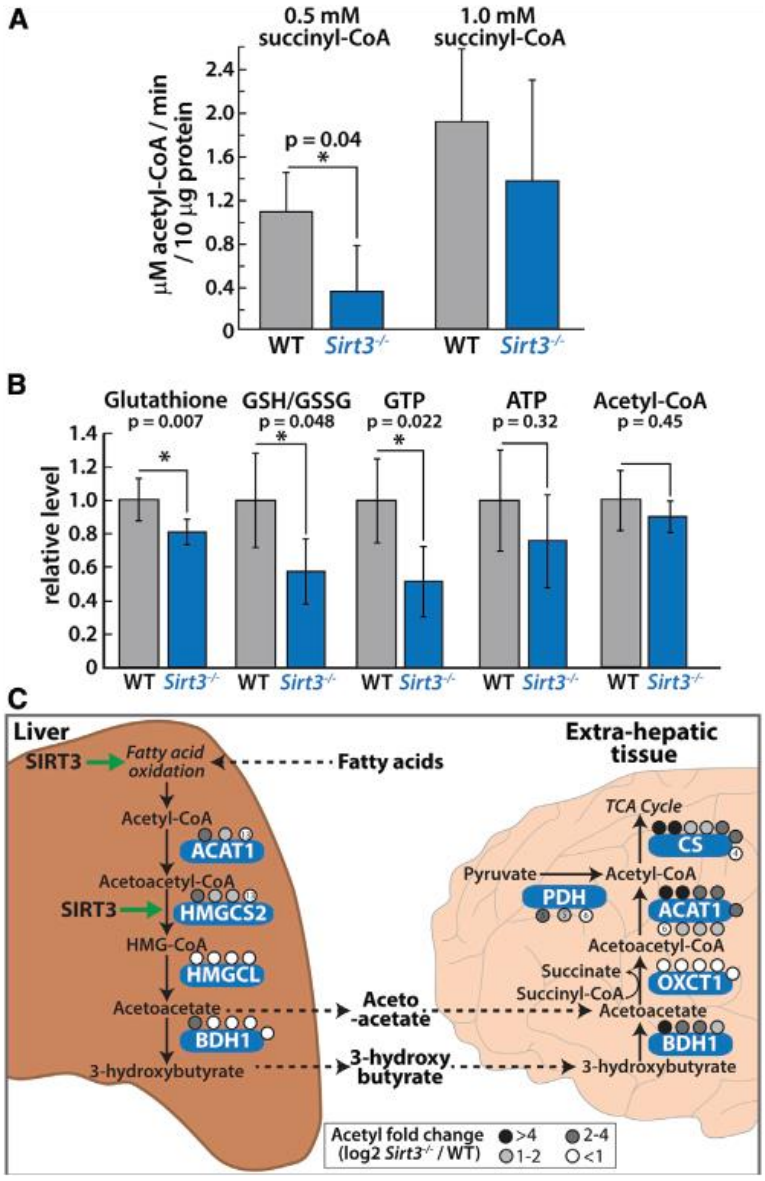
To confirm defects in ketone body generation in hepatic tissues, we measured a ketogenesis reaction in liver homogenates of WT and *Sirt3*$^{-/-}$ mice by monitoring the rate of acetylCoA-dependent CoA formation by LC-MS. In support of the hypothesis that SIRT3 is involved in tissue-specific fuel switching, we find a decrease in the activity of a ketogenesis pathway in *Sirt3*$^{-/-}$ liver. Consistent with our results, previous metabolomics analyses in plasma and liver identified alterations in the ketone body 3-hydroxybutyrate during fasting and in hydroxybutyryl carnitine levels[21-23] in mice lacking SIRT3.

Acetyl-CoA generated through ketone body degradation is combined with oxaloacetate to form citrate by the enzyme citrate synthase, entering the citrate cycle to provide reducing equivalents required

for adenosine triphosphate (ATP) production. In a separate reaction, we monitored citrate synthase activity by combining cortex homogenate with oxaloacetate and acetyl-CoA and measuring citrate formation over time to determine whether decreased acetyl-CoA production in the *Sirt3*$^{-/-}$ condition was due to accelerated utilization of acetyl-CoA by citrate synthase. This was not the case; in fact, we detected a decrease in oxaloacetate-dependent citrate formation in *Sirt3*$^{-/-}$, which is consistent with prior results suggesting citrate synthase is also a target of SIRT3.[4] However, we cannot rule out the possibility that acetyl-CoA or another co-substrate, like succinyl-CoA, is consumed by an undefined metabolic pathway in the *Sirt3*$^{-/-}$ condition and competes with the assayed activity.

Taken together, the decrease in acetoacetate-dependent generation of acetyl-CoA at low succinyl-CoA concentrations represents a deficiency in ketone body utilization in the brain of mice lacking SIRT3, supporting a critical role for SIRT3 in this pathway. Consistent with the metabolite flux analysis, two central enzymes in acetyl-CoA production are highly probable targets for SIRT3 regulation: succinyl-CoA: 3-ketoacidCoA transferase (OXCT) and acetyl-CoA acetyltransferase 1 (ACAT1). OXCT transfers the CoA from succinate to acetoacetate, forming acetoacetyl-CoA that is then converted by ACAT1 to two molecules of acetyl-CoA. Both enzymes are hyperacetylated in the brain of *Sirt3*$^{-/-}$ animals (**Figure 5*C***). Importantly, OXCT is only expressed in extra-hepatic tissues to avoid non-productive cycling of acetoacetate in the liver and is exclusively acetylated in the brain (five acetyl sites) and heart (three acetyl sites) in this analysis. Our previous report on ACAT1 enzymatic function identified a defect in catalysis of acetyl-CoA formation due to acetylation of K260 and K265,[5] providing direct enzymatic evidence that supports our current findings. Numerous acetyl sites, including K260 and K265, were identified in all tissues in this study. The observation that the magnitude of the defects in acetoacetate-dependent acetyl-CoA production were dependent on succinyl-CoA levels (**Figure 5*A***) suggested that hyper-acetylation of OXCT decreased catalytic efficiency, likely by affecting the V_{max}/K_m kinetic parameter for succinyl-CoA. Consistent with this observation, lysine 451 (K451), a site identified and quantified only in the brain, displays the largest magnitude acetyl fold change (67% increase, $p = 1.1 \times 10^{-4}$, Student's t test) in the *Sirt3*$^{-/-}$ condition, when compared to other acetyl sites quantified on OXCT. This acetyl residue lies within the CoA binding pocket

Figure 5. SIRT3 mediates cross-tissue link for ketone body production and utilization. (***A***) Average rate of acetoacetate-dependent acetyl-CoA production in brain cortex homogenates from four WT (grey) or four *Sirt3*$^{-/-}$ (blue) mice indicates a defect in acetyl-CoA production in *Sirt3*$^{-/-}$ at low succinyl-CoA concentrations. Error bars represent standard deviation (* indicates statistical significance, $p < 0.05$, Student's t test). (***B***) Targeted metabolite analysis of WT and *Sirt3*$^{-/-}$ brain cortex (15 month old, n=4 per condition) identifies altered energetic status in *Sirt3*$^{-/-}$ brain. (***C***) A model for the differential regulation of ketone body production and utilization by SIRT3 in hepatic and extra-hepatic tissues. In the liver, SIRT3 is required for the synthesis of the ketone bodies acetoacetate and 3-hydroxybutyrate, through regulation of HMGCS2 enzymatic activity and pathways that generate acetyl-CoA in response to nutrient deprivation. This counters the role of SIRT3 in the brain, where SIRT3 is necessary for the utilization of acetoacetate to form acetyl-CoA to be used for energy production.



of OXCT (PDB 3OXO), approximately 4 Å from the ribosyl-phosphate group of CoA, thus acetylation of this site would be predicted to disrupt the binding or orientation of succinyl-CoA and serve as a mechanism for regulating OXCT function. To provide in vitro evidence for the effect of acetylation, we performed a biochemical analysis of recombinant OXCT, demonstrating that acetylation of K451 resulted in low catalytic activity that could be rescued in part by pre-incubation with SIRT3. Using site-specific incorporation of acetyllysine, we expressed and purified recombinant OXCT that is fully acetylated at K451.[24] In vitro OXCT activity assays across several concentrations of succinyl-CoA were performed and revealed that SIRT3 induced a 2.5-fold increase in the catalytic activity of acetylated OXCT. Minimal changes in activity were observed with identically treated WT OXCT (data not shown). We note that upon expression and purification, acetylated OXCT displayed extremely low activity and likely formed an oligomeric structure that partially occluded the acetyl site. This necessitated limited proteolysis of acetylated OXCT to demonstrate that SIRT3 was capable of removing the acetyl mark, as indicated by LC-MS measurement of the reaction product O-acetyl-ADP-ribose.

The end product of ketone body utilization, acetyl-CoA, is the oxidation substrate for the citrate cycle, a critical pathway for energy generation in the mitochondria. To determine whether the deficiency in ketone body utilization in the brain and the alteration in mitochondrial acetylation manifest in an overt energetic state or redox state, we performed a targeted LC-MS-based metabolite analysis on brain cortex of WT and *Sirt3*$^{-/-}$ mice (**Figure 5*B***). We observed a significant decrease of GSH/GSSG and glutathione in *Sirt3*$^{-/-}$ mice, consistent with prior studies showing that SIRT3 loss leads to increased reactive oxygen species production and oxidative stress,[25] indicating that our current biological analysis reflects established phenotypes of SIRT3 knockout. Chemically, cellular energy is mainly produced in the form of two triphosphate compounds, GTP, which is directly formed in the citrate cycle by succinyl-CoA synthetase, and ATP, produced by oxidative phosphorylation. The high-energy cofactors, GTP and ATP, showed approximately a 50% and 20% decrease in the brains of mice lacking SIRT3 (**Figure 5*B***), respectively, with the change in GTP being statistically significant. Collectively, these data are consistent with a defect in the ability of *Sirt3*$^{-/-}$ mice to switch fuel under a prolonged fast, including decreased ketone body

generation in the liver and reduced ketone body utilization in extrahepatic tissues such as brain. The magnitude of decreased ATP and acetyl-CoA levels in the *Sirt3*$^{-/-}$ brain may be indicative of organism-wide compensatory mechanisms that act to maintain metabolic and energetic homeostasis. This study reveals multi-tissue homeostatic effects and captures early-stage biochemical dysregulation within five individual tissues resulting from whole-body loss of SIRT3 expression.

DISCUSSION

Through the regulation of protein acetylation, SIRT3 controls enzymatic activity and flux of numerous metabolic pathways. Our study provides evidence that SIRT3 possesses (1) common core functions in all tissues through the regulation of pathways like the citrate cycle and oxidative phosphorylation, (2) tissue-specific functions to allow for crosstalk between tissues that enables a response to nutrient shifts (e.g., ketone body generation in the liver and utilization in the brain), and (3) an additional maintenance function to minimize the effects of untargeted acetylation in energy-producing tissues (liver and kidney) where higher SIRT3 expression reflects the acetate burden of each tissue.

The effect of SIRT3 on mitochondrial acetylation varies in a tissue-specific manner. Brain, heart, and skeletal muscle exhibit a greater percentage of sites hyper-acetylated (> 23) in SIRT3 knockout mice, while liver and kidney contain the highest total number of mitochondrial acetyl sites, but a lower percentage of dynamic acetyl sites (**Figure 2**). In this 24-hr fasted phase, extrahepatic tissues rely on fatty acids and ketone bodies produced by increased fatty acid oxidation in the liver.[26] Brain, heart, and skeletal muscle must efficiently oxidize acetyl-CoA for energy production, while hepatic tissues produce large amounts of intra-mitochondrial acetyl-CoA. We propose that this necessary demand for acetyl-CoA production drives widespread protein acetylation in liver, creating a greater need for the maintenance deacetylase functions of SIRT3.

Our study and others reveal widespread acetylation of the mitochondrial proteome;[4,5,27] however, the mechanism of acetylation is as of yet unknown, with scarce evidence for enzymatic acetyltransferase activity residing in the mitochondria.[28] Protein acetylation might occur non-enzymatically with the levels

and extent of acetylation reflecting acetyl-CoA concentration and conditions of the organelle.[29] The slightly elevated pH (pH 7.9) of the mitochondrial matrix compared to the cytosol and the millimolar intra-mitochondrial concentrations of acetyl-CoA generate a favorable condition for non-enzymatic lysine acetylation.[30,31] Regardless, both enzyme-catalyzed and non-enzymatic acetylation reactions are expected to increase as acetyl-CoA increases, and therefore the tissue-specific levels of protein acetylation might reflect the distinct conditions of each tissue type. Moreover, it is important to highlight that functional (targeted) acetylation and spurious (untargeted or non-regulatory) acetylation can occur via enzyme-catalyzed or uncatalyzed reactions. Here we suggest that higher SIRT3 expression in liver and kidney is required to remove both targeted and spurious protein acetylation that is a consequence of the increased demand to produce acetyl-CoA under fasting. In the brain, heart, and skeletal muscle, acetyl-CoA is rapidly consumed for energy, leading to reduced levels of spurious acetylation and reduced need for high levels of SIRT3, which acts in a more targeted manner by primarily deacetylating regulatory acetylation sites. Consistent with this idea, acetyl-CoA concentrations in liver are estimated to be 3- to 5-fold higher compared to those in skeletal muscle and brain.[32-34]

Our study reveals that SIRT3 mediates metabolic coupling between fuel-producing and fuel-utilizing tissues through dynamic acetylation of numerous mitochondrial pathways identified with our multi-tissue mitochondrial acetylome and innovative bioinformatics pathway analysis tool. We provide evidence that SIRT3 promotes ketone body utilization in the brain by deacetylating and enhancing flux through OXCT and ACAT1, manifesting in bioeneregetic deficiencies in aged *Sirt3*$^{-/-}$ animals. While a picture is emerging for the importance of SIRT3 in fuel switching during adaptation to altered metabolic status, our multi-tissue, quantitative acetyl proteome compendium provides a rich resource to further explore the tissue-specific functions of SIRT3, and the comparative mitochondrial acetylome required to understand metabolic regulation and the etiology of human diseases associated with the inability to perform efficient fuel switching.

EXPERIMENTAL PROCEDURES

Sample Preparation, MS, and Data Analysis. Mouse brain, heart, kidney, liver, and skeletal muscle were isolated from 5-month, age-matched *Sirt3*$^{-/-}$ and WT mice that were fasted for 24 hr (9am–9am) and immediately frozen in liquid nitrogen. All animal studies were conducted at the AAALAC-approved animal facility in the Genetics and Biotechnology Center of the University of Wisconsin-Madison. Experiments were performed in accordance with protocols approved by the University of Wisconsin-Madison Institutional Animal Care and Use Committee. Samples were suspended in lysis buffer and lysed by glass bead milling (Retsch). Samples were denatured, reduced, alkylated, and digested with trypsin (Promega). TMT labeling of desalted peptides was performed according to the manufacturer's instructions (Thermo Pierce). Labeled peptides were fractionated by SCX chromatography, and acetylated peptides were enriched with pan-acetyl lysine antibody-agarose conjugate. Acetyl-enriched and nonenriched protein fractions were analyzed by online nano-reverse phase liquid chromatography (Waters) coupled to an Orbitrap Fusion (Q-OT-qIT, Thermo Scientific). All MS/MS data were analyzed using the Coon OMSSA Proteomics Software Suite (COMPASS).[35] Results were filtered to a 1% FDR.

SIRT3 Expression

Targeted MS. Unlabeled peptide fractions from brain, heart, kidney, liver, and skeletal muscle of a WT mouse were subject to TMT labeling. Following tagging, peptides were mixed at a 1:1:1:1:1 ratio according to BCA results. For targeted MS analysis, the Sirt3 peptide ASGIPASK (+2, m/z 594.864, amino acids 171–178) was isolated over a 1-Th window. Results were normalized to total signal intensity for each TMT tag.

Ketone Body Utilization Activity Assay in Brain Homogenate. Brain cortices isolated from 8-month-old, 24-hr-fasted WT and *Sirt3*$^{-/-}$ mice (n = 4) were placed in ice-cold PBS supplemented with deacetylase inhibitors (1 mM sodium butyrate, 1 mM trichostatin A) and homogenized 3 3 1,000 rpm with a glass homogenizer and Teflon pestle. The sample was centrifuged twice at 1,000 3 g for 10 min at 4°C to remove

insoluble material. The supernatant was used for the activity assay, and protein concentration was determined by BCA. To test acetoacetate utilization activity, homogenates were combined with 1 mM acetoacetate, 0.5 or 1 mM succinyl CoA, and 0.5 mM CoA in PBS at 37°C. A control reaction was performed without acetoacetate. Reaction time points were taken at 1, 3, 5, 8, 10, 12, 15, and 20 min by quenching the reaction mixture in nine volumes of ice-cold methanol. Citrate synthase activity was monitored by combining homogenates with 0.5 mM oxaloacetate and 0.5 mM acetyl-CoA in PBS at 37°C. A control reaction was performed without oxaloacetate. Reaction time points were taken at 1 and 10 min by immediately quenching the reaction mixture in nine volumes of ice-cold methanol. All samples were centrifuged at 21,000 x g for 10 min, and supernatants were diluted and further analyzed by reverse-phase separation on a Synergy Hydro-RP column (100 mm 3 2 mm, 2.5 mm particle size, Phenomenex) coupled by negative-mode electrospray ionization to an Orbitrap mass spectrometer, as previously described.[36]

Sample Preparation for Metabolite Analysis in Brain. Brain cortices isolated from 15-month, 24-hr-fasted WT and *Sirt3*$^{-/-}$ mice (n = 4) were placed in 0.8 ml cold (-20°C) 80:20 methanol:H_2O (v/v) and homogenized 3 3 1,000 rpm with a glass homogenizer and Teflon pestle. The sample was centrifuged at 14,000 x g for 10 min at 4°C. The supernatant was transferred to a new tube on ice. The pellet was re-extracted with 0.5 ml cold (-20°C) 80:20 methanol:H_2O (v/v) twice, and the supernatants were combined. The samples were dried under nitrogen and resuspended in H_2O for LC-MS analysis. The results were normalized to tissue weight.

Bioinformatics

Clustering. To reduce the overall percentage of missing values to enable clustering of the entire dataset, missing protein abundance values were imputed for acetyl isoforms that were identified and quantified, but lacked protein level quantitation. To impute values, the average normalized intensity for all proteins found in each tissue of each mouse was calculated. This value was then used to calculate a protein fold change, to which the acetyl fold change could be normalized. Acetyl site fold change values were collapsed onto

protein by taking the sum of the acetyl fold change for all acetyl sites identified and quantified for a unique UniProt identifier. To cluster the protein dataset (1,100 proteins in five tissues) we used a Gaussian mixture model clustering approach with consensus clustering. Due to the large number of missing values (85% of proteins having missing values in two or more tissues), prior to applying the clustering method, we interpolate the missing value with the mean of the non-missing values of the same protein. To cluster each K cluster, we used hierarchical clustering using the Euclidean metric for distance.

QSSA. The intersection of the KEGG pathway map[37] and proteins identified with < 1% FDR was used for the gene set background. Acetylation coverage for each (*p*) pathway was calculated as the number of acetyl sites identified (N_k) over the total number of lysines in the pathway (N_s), counted using protein sequences from UniProt. Dynamic response to SIRT3 was taken into account by calculating the mean magnitude fold change (FC) for each acetyl site quantified. To allow for combining acetylation coverage and fold change, the standard score of each quantity was taken. The overall pathway score was then calculated as the average of the individual Z scores:

$$Z\left(Z\left(\frac{N_s}{N_K}\right)+Z\left(\sum_{s\in p}|FC_s|\right)\right)$$

Tissues were clustered by hierarchical clustering by calculating Euclidean distance and centroid linkage in Cluster 3.0.[38] Clusters were visualized with Java TreeView.[39]

Statistical Analysis. Proteomic data are expressed as $\log_2$ fold change between *Sirt3*$^{-/-}$ and WT. Statistical significance was calculated by Welch's t test. Biochemical assay results are expressed as mean ± SD. Statistical significance was assessed by Student's t test. A p value of < 0.05 was considered statistically significant for biochemical assays. Graphs were prepared and statistical analyses were performed using Origin, GraphPad Prism 6, Excel, or MATLAB.

ACKNOWLEDGMENTS

We would like to thank Daniel Amador-Noguez for assistance with metabolite studies. This work was funded by NIA grant AG038679 to J.M.D. and T.A.P., NIH grant GM065386 to J.M.D., and NIH grant GM080148 to J.J.C. Additionally, S.R. was supported by a Sloan Foundation research fellowship. K.E.D.-R. was funded by an NSF Graduate Research Fellowship and an NIH traineeship (5T32GM08349). A.L.R. was supported by an NIH-funded Genomic Sciences Training Program (5T32HG002760).

REFERENCES

(1) Stipanuk, M. H.; Caudill, M.H. Biochemical, Physiological, and Molecular Aspects of Human Nutrition. **2013**, 3rd Edition.

(2) Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. *Science* **2009**, *325*, 834-840.

(3) Lundby, A.; Lage, K.; Weinert, B. T.; Bekker-Jensen, D. B.; Secher, A.; Skovgaard, T.; Kelstrup, C. D.; Dmytriyev, A.; Choudhary, C.; Lundby, C.; Olsen, J. V. *Cell Rep* **2012**, *2*, 419-431.

(4) Hebert, A. S.; Dittenhafer-Reed, K. E.; Yu, W.; Bailey, D. J.; Selen, E. S.; Boersma, M. D.; Carson, J. J.; Tonelli, M.; Balloon, A. J.; Higbee, A. J.; Westphall, M. S.; Pagliarini, D. J.; Prolla, T. A.; Assadi-Porter, F.; Roy, S.; Denu, J. M.; Coon, J. J. *Mol Cell* **2013**, *49*, 186-199.

(5) Still, A. J.; Floyd, B. J.; Hebert, A. S.; Bingman, C. A.; Carson, J. J.; Gunderson, D. R.; Dolan, B. K.; Grimsrud, P. A.; Dittenhafer-Reed, K. E.; Stapleton, D. S.; Keller, M. P.; Westphall, M. S.; Denu, J. M.; Attie, A. D.; Coon, J. J.; Pagliarini, D. J. *J Biol Chem* **2013**, *288*, 26209-26219.

(6) Rardin, M. J.; Newman, J. C.; Held, J. M.; Cusack, M. P.; Sorensen, D. J.; Li, B. A.; Schilling, B.; Mooney, S. D.; Kahn, C. R.; Verdin, E.; Gibson, B. W. *P Natl Acad Sci USA* **2013**, *110*, 6601-6606.

(7) Newman, J. C.; He, W. J.; Verdin, E. *J Biol Chem* **2012**, *287*, 42436-42443.

(8) He, W. J.; Newman, J. C.; Wang, M. Z.; Ho, L.; Verdin, E. *Trends Endocrin Met* **2012**, *23*, 467-476.

(9) Jing, E. X.; O'Neill, B. T.; Rardin, M. J.; Kleinridders, A.; Ilkeyeva, O. R.; Ussar, S.; Bain, J. R.; Lee, K. Y.; Verdin, E. M.; Newgard, C. B.; Gibson, B. W.; Kahn, C. R. *Diabetes* **2013**, *62*, 3404-3417.

(10) Pagliarini, D. J.; Calvo, S. E.; Chang, B.; Sheth, S. A.; Vafai, S. B.; Ong, S. E.; Walford, G. A.; Sugiana, C.; Boneh, A.; Chen, W. K.; Hill, D. E.; Vidal, M.; Evans, J. G.; Thorburn, D. R.; Carr, S. A.; Mootha, V. K. *Cell* **2008**, *134*, 112-123.

(11) Iwahara, T.; Bonasio, R.; Narendra, V.; Reinberg, D. *Mol Cell Biol* **2012**, *32*, 5022-5034.

(12) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2012**, *11*, 1475-1488.

(13) Lombard, D. B.; Alt, F. W.; Cheng, H. L.; Bunkenborg, J.; Streeper, R. S.; Mostoslavsky, R.; Kim, J.; Yancopoulos, G.; Valenzuela, D.; Murphy, A.; Yang, Y.; Chen, Y.; Hirschey, M. D.; Bronson, R. T.; Haigis, M.; Guarente, L. P.; Farese, R. V., Jr.; Weissman, S.; Verdin, E.; Schwer, B. *Mol Cell Biol* **2007**, *27*, 8807-8814.

(14) Shi, T.; Wang, F.; Stieren, E.; Tong, Q. *J Biol Chem* **2005**, *280*, 13560-13567.

(15) Zhu, J.; Zou, H.; Rosset, S.; Hastie, T. Tibshirani, R., Friedman, J.H. The Elements of Statistical Learning: Data Mining, Interference and Prediction. **2009**, 2nd Edition.

(16) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nucleic Acids Res* **2009**, *37*, 1-13.

(17) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. *Nat Protoc* **2009**, *4*, 44-57.

(18) Irizarry, R. A.; Wang, C.; Zhou, Y.; Speed, T. P. *Stat Methods Med Res* **2009**, *18*, 565-575.

(19) Kanehisa, M.; Goto, S. *Nucleic Acids Res* **2000**, *28*, 27-30.

(20) Suryawan, A.; Hawes, J. W.; Harris, R. A.; Shimomura, Y.; Jenkins, A. E.; Hutson, S. M. *Am J Clin Nutr* **1998**, *68*, 72-81.

(21) Hirschey, M. D.; Shimazu, T.; Goetzman, E.; Jing, E.; Schwer, B.; Lombard, D. B.; Grueter, C. A.; Harris, C.; Biddinger, S.; Ilkayeva, O. R.; Stevens, R. D.; Li, Y.; Saha, A. K.; Ruderman, N. B.; Bain, J. R.; Newgard, C. B.; Farese, R. V.; Alt, F.; Kahn, C. R.; Verdin, E. *Nature* **2010**, *464*, 121-U137.

(22) Shimazu, T.; Hirschey, M. D.; Hua, L.; Dittenhafer-Reed, K. E.; Schwer, B.; Lombard, D. B.; Li, Y.; Bunkenborg, J.; Alt, F. W.; Denu, J. M.; Jacobson, M. P.; Verdin, E. *Cell Metab* **2010**, *12*, 654-661.

(23) Hallows, W. C.; Yu, W.; Smith, B. C.; Devires, M. K.; Ellinger, J. J.; Someya, S.; Shortreed, M. R.; Prolla, T.; Markley, J. L.; Smith, L. M.; Zhao, S. M.; Guan, K. L.; Denu, J. M. *Mol Cell* **2011**, *41*, 139-149.

(24) Neumann, H.; Hancock, S. M.; Buning, R.; Routh, A.; Chapman, L.; Somers, J.; Owen-Hughes, T.; van Noort, J.; Rhodes, D.; Chin, J. W. *Mol Cell* **2009**, *36*, 153-163.

(25) Someya, S.; Yu, W.; Hallows, W. C.; Xu, J.; Vann, J. M.; Leeuwenburgh, C.; Tanokura, M.; Denu, J. M.; Prolla, T. A. *Cell* **2010**, *143*, 802-812.

(26) Bauer, M.; Hamm, A. C.; Bonaus, M.; Jacob, A.; Jaekel, J.; Schorle, H.; Pankratz, M. J.; Katzenberger, J. D. *Physiol Genomics* **2004**, *17*, 230-244.

(27) Kim, S. C.; Sprung, R.; Chen, Y.; Xu, Y. D.; Ball, H.; Pei, J. M.; Cheng, T. L.; Kho, Y.; Xiao, H.; Xiao, L.; Grishin, N. V.; White, M.; Yang, X. J.; Zhao, Y. M. *Mol Cell* **2006**, *23*, 607-618.

(28) Scott, I.; Webster, B. R.; Chan, C. K.; Okonkwo, J. U.; Han, K.; Sack, M. N. *J Biol Chem* **2014**, *289*, 2864-2872.

(29) Cai, L.; Sutter, B. M.; Li, B.; Tu, B. P. *Mol Cell* **2011**, *42*, 426-437.

(30) Garland, P. B.; Shepherd, D.; Yates, D. W. *Biochem J* **1965**, *97*, 587-594.

(31) Paik, W. K.; Pearson, D.; Lee, H. W.; Kim, S. *Biochim Biophys Acta* **1970**, *213*, 513-522.

(32) Allred, J. B.; Guy, D. G. *Anal Biochem* **1969**, *29*, 293-299.

(33) Cederblad, G.; Carlin, J. I.; Constantin-Teodosiu, D.; Harper, P.; Hultman, E. *Anal Biochem* **1990**, *185*, 274-278.

(34) Palladino, A. A.; Chen, J.; Kallish, S.; Stanley, C. A.; Bennett, M. J. *Mol Genet Metab* **2012**, *107*, 679-683.

(35) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. *Proteomics* **2011**, *11*, 1064-1074.

(36) Lu, W. Y.; Clasquin, M. F.; Melamud, E.; Amador-Noguez, D.; Caudy, A. A.; Rabinowitz, J. D. *Analytical Chemistry* **2010**, *82*, 3212-3221.

(37) Merico, D.; Isserlin, R.; Stueker, O.; Emili, A.; Bader, G. D. *Plos One* **2010**, *5*.

(38) de Hoon, M. J. L.; Imoto, S.; Nolan, J.; Miyano, S. *Bioinformatics* **2004**, *20*, 1453-1454.

(39) Saldanha, A. J. *Bioinformatics* **2004**, *20*, 3246-3248.

Chapter 3

Neutron-encoded Signatures Enable Product Ion Annotation from Tandem Mass Spectra

This chapter has been published:

Richards AL, Vincent CE, Guthals A, Rose CM, Westphall MS, Bandeira N, Coon JJ. *Neutron-encoded Signatures Enable Automated Product Ion Annotation*. Molecular & Cellular Proteomics. **2013,** *12*, 3812-3823.

ABSTRACT

We report the use of neutron-encoded (NeuCode) stable isotope labeling of amino acids in cell culture for the purpose of C-terminal product ion annotation. Two NeuCode labeling isotopologues of lysine, $^{13}C_6{}^{15}N_2$ and $^{2}H_8$, which differ by 36 mDa, were metabolically embedded in a sample proteome, and the resultant labeled proteins were combined, digested, and analyzed via liquid chromatography and mass spectrometry. With MS/MS scan resolving powers of ~50,000 or higher, product ions containing the C terminus (*i.e.* lysine) appear as a doublet spaced by exactly 36 mDa, whereas N-terminal fragments exist as a single *m*/*z* peak. Through theory and experiment, we demonstrate that over 90% of all y-type product ions have detectable doublets. We report on an algorithm that can extract these neutron signatures with high sensitivity and specificity. In other words, of 15,503 y-type product ion peaks, the y-type ion identification algorithm correctly identified 14,552 (93.2%) based on detection of the NeuCode doublet; 6.8% were misclassified (*i.e.* other ion types that were assigned as y-type products). Searching NeuCode labeled yeast with PepNovo$^+$ resulted in a 34% increase in correct *de novo* identifications relative to searching through MS/MS only. We use this tool to simplify spectra prior to database searching, to sort unmatched tandem mass spectra for spectral richness, for correlation of co-fragmented ions to their parent precursor, and for *de novo* sequence identification.

INTRODUCTION

The ability to make *de novo* sequence identifications directly from tandem mass spectra has long been a holy grail of the proteomic community. Such a capability would wean the field from its reliance upon sequenced genome databases. Even for organisms with fully annotated genomes, events such as single nucleotide polymorphisms, alternative splicing, gene fusion, and a host of other genomic transformations can result in altered proteomes. These alterations can vary from cell to cell and individual to individual. Thus, one could argue that the most valuable proteomic information, the individual and cellular proteome variation from the genome, remains elusive.[1] This problem has received considerable attention; that said, it is not easy to *de novo* correlate spectrum to sequence in a large-scale, automated fashion.[2-6] Improvements in mass accuracy have helped, but routine, reliable *de novo* sequencing without database assistance is not standard.[7-10]

A primary means to facilitate *de novo* spectral interpretation is the simple annotation of *m*/*z* peaks in tandem mass spectra as either N- or C-terminal. We and others have investigated this seemingly simple first step. Real-world spectra, however, are complex. Difficulties often arise in determining the charge state of the fragment or in differentiating between fragment ions and peaks arising from neutral loss, internal fragmentation, or spectral noise, both electronic and chemical. Several strategies have focused on product ion annotation. These approaches have included manipulation of the N-terminus basicity combined with electron transfer dissociation (ETD).[11-13] This approach can yield mostly N-terminal fragments for peptides having only two charges. However, it requires both ETD and the protease LysN. Other methods have used differential labeling of N- and C-terminal peptides to shift either one or the other product ion series, by either metabolic or chemical means.[14-18] Metabolic incorporation of amino acids is an efficient method of introducing distinctive labels that eliminates *in vitro* labeling, but this method requires that the sample be amenable to cell culture.[19,20] Additionally, it may be difficult to achieve complete labeling in complex systems. Several other approaches used to introduce heavy isotopes onto one terminus have been investigated, including trypsin digestion in ^{18}O water,[21-23] differential isotopic esterification,[24,25]

derivatization of the C-terminal carboxylate by p-bromophenethylamine,[8,26] N-terminal derivatization with sulfonic acid groups[27,28] and formaldehyde labeling via reductive amination.[29,30] These chemical modifications are introduced after cell lysis, often immediately prior to analysis. Although chemical labeling strategies can be used with a variety of samples, difficulties can arise from differences in labeling efficiency between samples, and often a clean-up step is required following labeling, which may lead to sample loss. No matter the labeling method, in this regime, the two precursors must be separately isolated, fragmented, and analyzed either together or separately. The recognition and selection of the broadly spaced doublet in real time also are necessary. These requirements have limited the utility of these approaches. Our own laboratory discovered that the c- and $^{\bullet}$z-type product ions generated from either electron capture dissociation or ETD have distinct chemical formulae and therefore can always be distinguished based on accurate mass alone.[30,31] The problem with this approach is that extremely high mass accuracy (<500 ppb) is required in order to distinguish these product ion types above ~600 Da in mass. Thus, the majority of the product ions within a spectrum cannot be readily mapped to either terminus with high confidence.

Despite these difficulties, we assert that robust *de novo* sequencing methodology would benefit greatly from a simple method that could be used to distinguish N- and C-terminal product ions with high accuracy and precision. Ideally, the approach would work regardless of the choice of proteolytic enzyme or dissociation method. Recently, we described a new technology for protein quantification called neutron encoding (NeuCode).[32] NeuCode embeds millidalton (mDa) mass differences into peptides and proteins by exploiting the mass defect induced by differences in the nuclear binding energies of the various stable isotopes of common elements such as C, N, H, and O. For example, consider the amino acid lysine, which has eight additional neutrons (+8 Da). One way to synthesize this amino acid is to add six ^{13}C atoms and two ^{15}N atoms (+8.0142 Da). Another isotopologue could be constructed by adding eight ^{2}H atoms (+8.0502). These two isotopologues differ by only 36 mDa; peptide precursors containing both of these amino acids would appear as a single, unresolved precursor *m*/*z* peak at a mass resolving power of less than ~100,000. However, under high resolving powers (*i.e.* greater than ~100,000 at *m*/*z* 400), this doublet is

resolved. We first developed this NeuCode concept in the context of metabolic labeling, akin to stable isotope labeling with amino acids in cell culture (SILAC), except that instead of the precursor partners being separated by 4 to 8 Da, they are separated by only 6 to 40 mDa. For quantitative purposes, NeuCode promises to deliver ultraplexed SILAC (>12) without increasing spectral complexity.

We reasoned that the isotopologues of Lys that permit NeuCode SILAC would generate a distinct fingerprint on C-terminal product ions. Specifically, peptides that have been labeled with NeuCode SILAC and digested with LysC uniformly contain Lys at the C terminus. Upon MS/MS, all C-terminal product ions should present as doublets (with duplex NeuCode), whereas N-terminal products will be detected as a single *m*/*z* peak. The very close *m*/*z* spacing of the NeuCode SILAC partners will ensure that each partner is *always* co-isolated and that the signatures are visible only upon high-resolving-power mass analysis. Here we investigate the combination of NeuCode SILAC and high-resolving-power MS/MS analysis to allow the straightforward identification of C-terminal product ions.

RESULTS

Theoretical Considerations. To test our hypothesis—that NeuCode SILAC labeling will permit the identification of C-terminal product ions—we sought to determine the MS/MS resolving power requirements. First, we examined the charge (*z*), mass (*m*), and *m*/*z* of all detected y-type ions from a library of 19,521 y-type ions extracted from 2,392 tandem mass spectra. We then determined the percentage of y-type ions that are resolved (full width at half-maximum) when labeled with lysine isotopologues differing by 36 mDa (**Figure 1**). This calculation takes into account the diversity of product ions *m*, *z*, and *m*/*z* that is typically observed in a shotgun experiment. These data demonstrate that at a resolving power of 60,000, ~90% of detected y-type ions were resolvable, and hence identifiable. At resolving powers above 100,000, virtually all NeuCode product ion doublets were detectable. Today it is routine to collect MS/MS spectra at resolving powers between 15,000 and 30,000 using both TOF and Orbitrap mass analyzers. Cutting-edge TOF analyzers are capable of 60,000 resolving power.[33,34] Orbitrap systems can achieve resolving powers in excess of one million,[35] but even low-end commercial systems offer up to 120,000 resolving power.

Fourier transform ion cyclotron resonance analyzers, of course, offer the highest resolving powers.[36-38] For the Fourier transform MS systems, increased resolving power is achieved by increasing the transient acquisition time. That said, such operation effectively reduces the tandem MS duty cycle for Orbitrap analyzers (~300 ms/scan for 15,000 resolving power and ~650 ms/scan for 120,000).[39] Below we describe experiments that tested these theoretical considerations. Furthermore, we explore the effect of a reduced duty cycle of peptide identifications.

Experimental Proof of Concept. We labeled a yeast cell culture with isotopologues of lysine, one containing six ^{13}C atoms and two ^{15}N atoms (+8.0142 Da) and the other with eight ^{2}H (+8.0502 Da), a 36-mDa difference, to test the premise that our NeuCode labeling strategy may facilitate direct product ion annotation. Following cell lysis, we digested the proteins with LysC and mixed the proteomes in a 1:1 ratio. The resulting peptides were analyzed via nHPLC-MS/MS on a quadrupole linear ion trap Orbitrap hybrid MS system (Orbitrap Elite). **Figure 1*A*** presents a scan sequence beginning with an MS^1 analysis. Expansion of a selected precursor *m*/*z* region (*m*/*z* 757; **Figure 1*B***) displays the isotopic cluster profile at resolving powers of either 30,000 (black trace) or 480,000 (red trace). Note that the NeuCode SILAC doublet is not detectable in the lower resolving power analysis but is easily distinguished upon high-resolving-power scanning. An example MS/MS spectrum, following beam-type collisional activated dissociation (HCD), is presented in **Figures 1*C*** and **1*D***. This scan confirms our guiding supposition that only y-type product ions appear as doublets, indicating the presence of a NeuCode labeled lysine within the product ion sequence. Of course, a small percentage of doublet-containing fragment ions could also arise from enzymatic cleavage at adjacent lysines, resulting in an N-terminal lysine, or from missed cleavage at KP residues. To confirm that the approach is not affected by the dissociation method, we performed a separate analysis using ETD dissociation (**Figure 2**). Again, only C-terminal fragments—$z^{\bullet}$-type ions in this case—existed as doublets and were readily distinguished. Manual validation of a number of these tandem mass spectra confirmed that C-terminal fragments, if present, contained the NeuCode doublet.

With these promising preliminary data, we sought to test the frequency of C-terminal fragment ions detected as doublets as a function of MS/MS resolving power ranging from 15,000 to 240,000. To accomplish this we analyzed the complex peptide mixture of NeuCode labeled yeast peptides using nHPLC-MS/MS with varied resolving powers. **Table 1** summarizes the number of MS/MS scans that were acquired over the 120-min separation across the various resolving powers. We note only a subtle drop in total scans from resolving powers of 15,000 to 60,000 (18,654 *versus* 16,230, respectively). Increasing the resolving power to 120,000 and 240,000 further reduced the total number of scans to 13,126 and 9,083. Next we performed a standard database search of these spectra to map them to sequence. Having the sequences of each precursor in hand, we wrote a custom algorithm to inspect each MS/MS spectrum for the presence or absence of a NeuCode doublet at the *m*/*z* value of each predicted fragment ion. For the spectra collected at 15,000 resolving power, the majority of y-type ions did not appear as doublets. Specifically, 26% of y-type products showed some evidence of a doublet; this number is higher than expected and is likely an overestimation, as manual inspection revealed that only the y_1 ion was consistently resolved. In contrast, 85% percent of y-type ions were detected as NeuCode doublets when analyzed at 60,000 resolving power. This result corresponds very well with our theoretical predictions (88.69%; **Figure 1**). Again, as predicted by theory, over 90% of all y-type ions were detected as NeuCode doublets at resolving powers in excess of 120,000 (**Table 1**). At all resolving powers, b-type ions were never detected as having a NeuCode doublet more than 2% of the time. From these data, we conclude that the vast majority of y-type ions can be detected with MS/MS resolving powers of ~60,000. For the Orbitrap system used here, the extended scan duration caused by this mode of analysis incurs only a subtle penalty in peptide identifications relative to the standard 15,000 resolving power (2,839 *versus* 2,546, respectively).

Product Ion Annotation. With confidence that our theoretical predictions were experimentally sound, we next sought to develop a methodology to annotate product ion type without knowledge of the peptide sequence. The premise was to search for NeuCode doublets quickly and with high precision. We developed

Figure 1. Resolution requirements for NeuCode labeled peptide analysis. ***(A)*****.** Theoretical calculations depicting the number of y-type ions that can be resolved at various mass resolving powers. At 36 mDa NeuCode spacing, approximately 90% of y-type ions will be resolved, and hence distinguished, at a resolving power of 60,000. Note the specified resolution is at *m*/*z* 400. NeuCode SILAC labeling permits identification of C-terminal product ions. Peptides were labeled with two isotopologues of Lys (36 mDa difference) and digested with LysC. MS^1 scan (***B***) with selected precursor shown in panel ***C***. Quantitative data can be concealed or revealed depending on resolving power (15,000, concealed; 240,000, revealed). Likewise, during MS/MS scanning product ions containing the C-terminus, y-type, in this case, appear as a doublet when analyzed under higher resolving power settings (>120,000, panels ***D*** and ***E***).

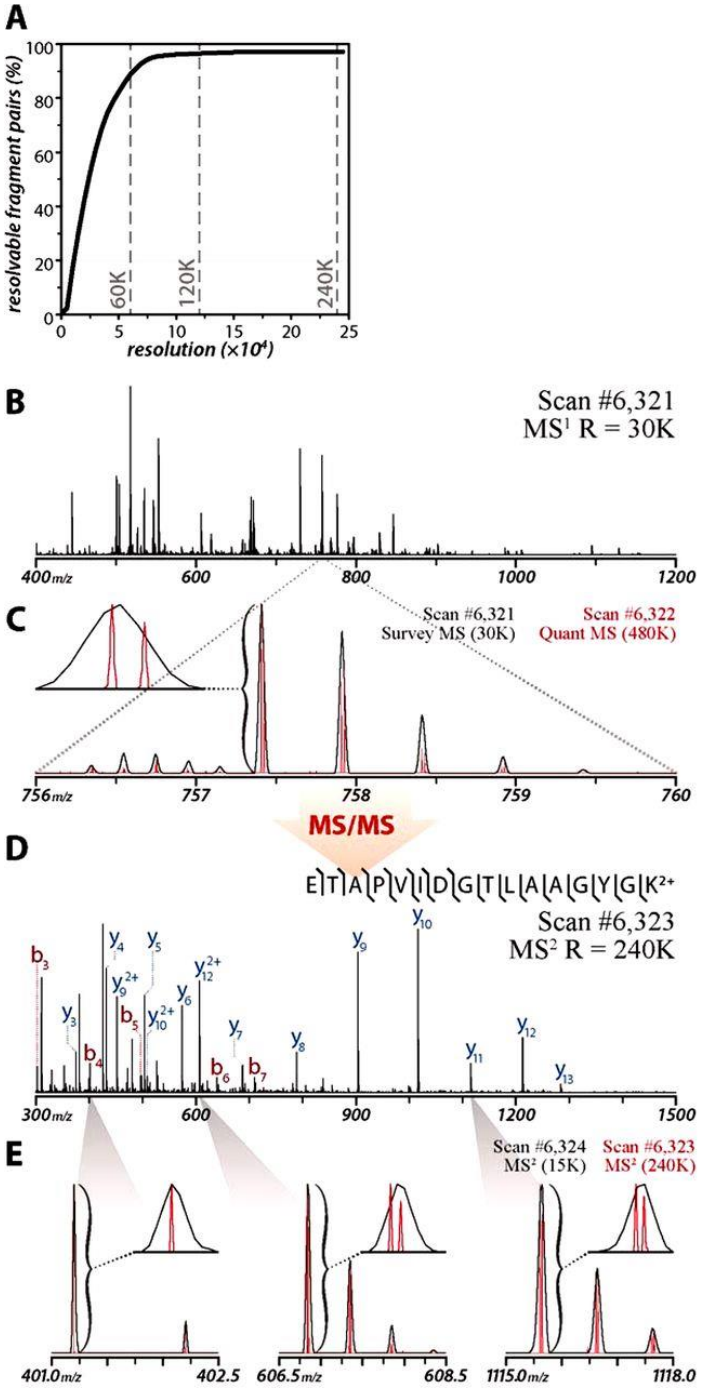


Figure 2. Appearance of NeuCode doublet peaks in MS^1 and MS^2 spectra. NeuCode SILAC labeling permits identification of C-terminal product ions and is indifferent to fragmentation method, in this case ETD. Peptides were labeled with two isotopologues of Lys (36 mDa difference) and digested with LysC. MS^1 scan (***A***) with ETD MS/MS analysis of a selected precursor shown in panel ***B***. Quantitative data can be concealed or revealed, depending on the resolving power (15,000, concealed; 240,000, revealed). Product ions containing the C-terminus, z$^\bullet$-type, in this case, appear as a doublet when analyzed under high resolving powers (>120,000, panels ***C*** and ***D***).

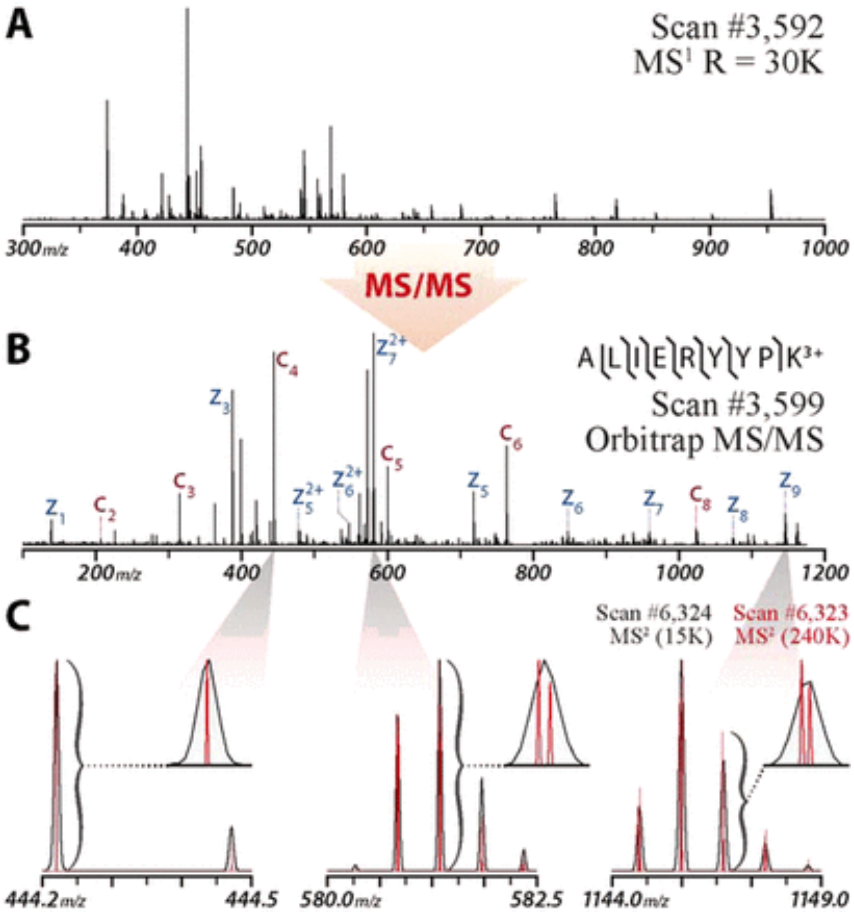
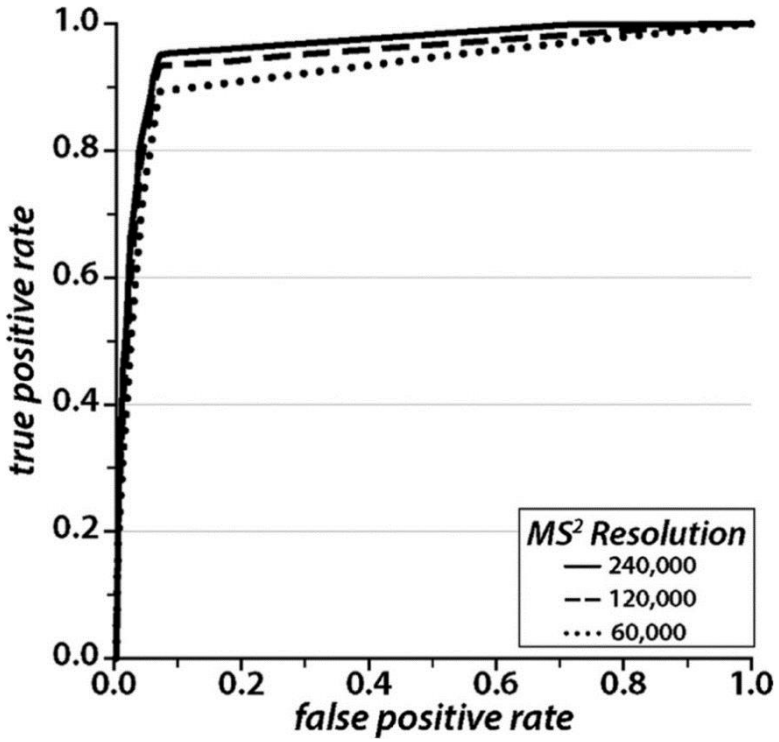


Table 1. **Comparison of NeuCode de novo metrics at various MS/MS resolutions at *m*/*z* 400**. *As y-type ions are not fully resolved at 15,000 resolution, the average mass of the isotopologue pairs was used for database searching. For all other resolutions, the mass of the light isotopologue was used.

MS/MS resolution	15,000	60,000	120,000	240,000
Number of scans	18,654	16,230	13,161	9,083
Unique peptides	2,028*	1,860	1,745	1,345
Peptide spectral matches	2,839*	2,546	2,466	1,831
Percentage of spectral fragments appearing as doublets				
y-ions	26.27	85.42	93.29	95.23
b-ions	0.45	1.35	1.79	2.04

a custom algorithm that considered NeuCode doublet *m*/*z* spacing and the total lysine count (determined from the MS^1 scan) to determine the product ion type. To avoid over assignment, spectra were also searched for corresponding isotopes and neutral loss peaks whenever a doublet pair was identified. The receiver operating characteristic curves for y-type ion prediction by use of the various datasets of MS/MS resolving power are presented in **Figure 3**. Note that we dismissed peptides that result from missed cleavage (~5% to 10%), as they contain internal lysine residues. This knowledge is gained via inspection of the NeuCode doublet in the prior MS^1 scan. The algorithm, which uses machine learning, achieved an overall accuracy of 94.3% with a specificity of 94.7% and a sensitivity of 93.3% from the highest resolving power dataset (240,000). At 120,000, that performance was almost identical (94.9% sensitivity and 92.1% specificity). At 60,000 resolving power, sensitivity slipped just a bit (88.9%), but specificity was comparable (93.1%). To put this into context, consider the 2,466 MS/MS spectra that were mapped to sequence with high confidence from the 120,000 resolving power dataset. In total, these spectra contained 15,503 y-type product ion peaks as detected by the search algorithm. Our y-type ion identification algorithm correctly identified 14,552 (93.2%) of these via detection of the NeuCode doublet. Also encouraging is the relatively low rate of false positive predictions: out of all remaining *m*/*z* peaks in the experimental spectra, only 6.8% were misclassified. Next we reasoned that our fragment ion annotation method should be unaffected by the dissociation method—that is, it should detect C-terminal ions regardless of the ion type. To test this idea, we collected a tandem mass spectral dataset by use of ETD, a dissociation method that is complementary to collisional activated dissociation.[40-43] Using an ETD dataset acquired at an MS^2 resolving power of 240,000, our overall accuracy in correctly identifying a $z^{\bullet}$-type was comparable to that achieved with y-type ions at 92.70%. Although the specificity of 93.00% was similar to what is achieved with HCD, the sensitivity of ETD was slightly lower at 82.80%. We suspect that a reason for this drop in sensitivity is that our learner was trained on HCD data and is more adept at predicting y-type than $z^{\bullet}$-type ions. Given these strong results and the minimal effect of the dissociation type, we envision utilizing these data for a variety of applications, including the facilitation of spectral pre-processing and *de novo* annotation.

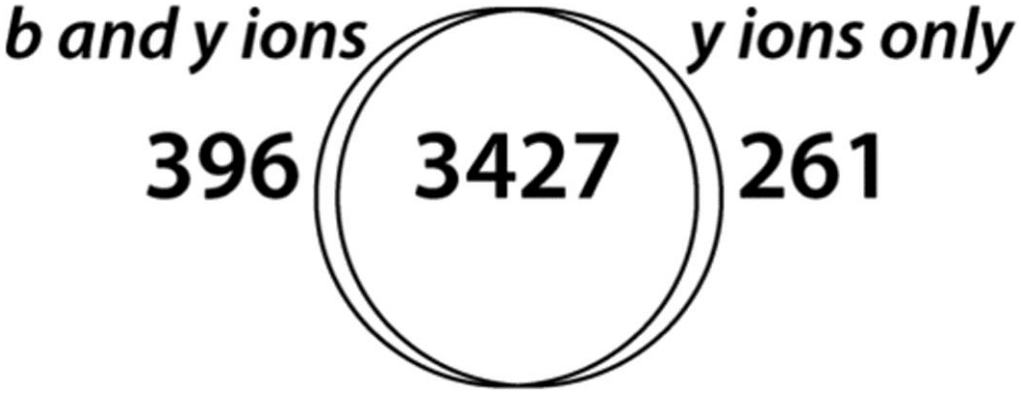
Figure 3. **Evaluation of algorithm performance in the context of different Orbitrap resolution settings.** ROC curve showing the ability of the algorithm to identify C-terminal product ions for peptides containing one lysine at resolving powers of 240,000, 120,000 or 60,000.



Spectral Identification with Product Ion Annotation. We reasoned that a direct means by which to confirm the quality of our product ion annotation method was to couple it with traditional database searching. Aside from testing our approach, product ion annotation could offer opportunities to improve traditional database spectral correlation algorithms by reducing search space and/or adding specificity. We provide here preliminary data to convey the promise of this approach. Briefly, we executed our annotation algorithm on a compendium of 23,442 tandem mass spectra that had been collected from peptides carrying the NeuCode SILAC labels. We then removed every *m*/*z* peak in each of the 23,442 spectra that was not annotated as a y-type ion. On average, only 2.3% of the *m*/*z* peaks in a tandem mass spectrum were retained following filtering. In other words, 98 out of 100 product ion peaks were dismissed. Note that a large quantity of these *m*/*z* peaks represent spectral noise or isotopic clusters and are typically removed prior to database searching.[16,44] Nevertheless, this process provides us with a simplification of the dataset that will speed searching and analysis, as only the most relevant *m*/*z* peaks are retained. These abridged DTA files were then searched in the conventional way, but for only y-type products, increasing the speed of the search, and potentially the specificity. The DTAs were also generated without filtering, and these files were searched to provide a control using normal parameters. Comparable results were obtained for both searches (**Figure 4**): 3,688 *versus* 3,823 identifications at a 1% FDR for the annotated and un-annotated spectra, respectively. Overall, the two searches jointly identified the vast majority (3,427) of the spectra, with each search identifying a handful of unique sequences. The majority of the spectra uniquely identified through database searching contained five or fewer y-type ions. These data suggest that we correctly annotated y-type ions. Moving forward, we expect significant benefits to be achieved with this NeuCode-facilitated approach (see below).

Spectral Quality Assessment. An immediate application for NeuCode product ion annotation is the assessment of tandem mass spectral quality, both pre- and post-database searching. Even with state-of-the-art MS instruments, often only 30% to 50% of MS/MS spectra are matched to sequence following a typical

Figure 4. **The utilization of NeuCode doublets to perform spectral peak filtering prior to automated database searching.** Venn diagram illustrating the overlap between a database search of all fragment ions and a database search of only annotated *y-type* fragments.

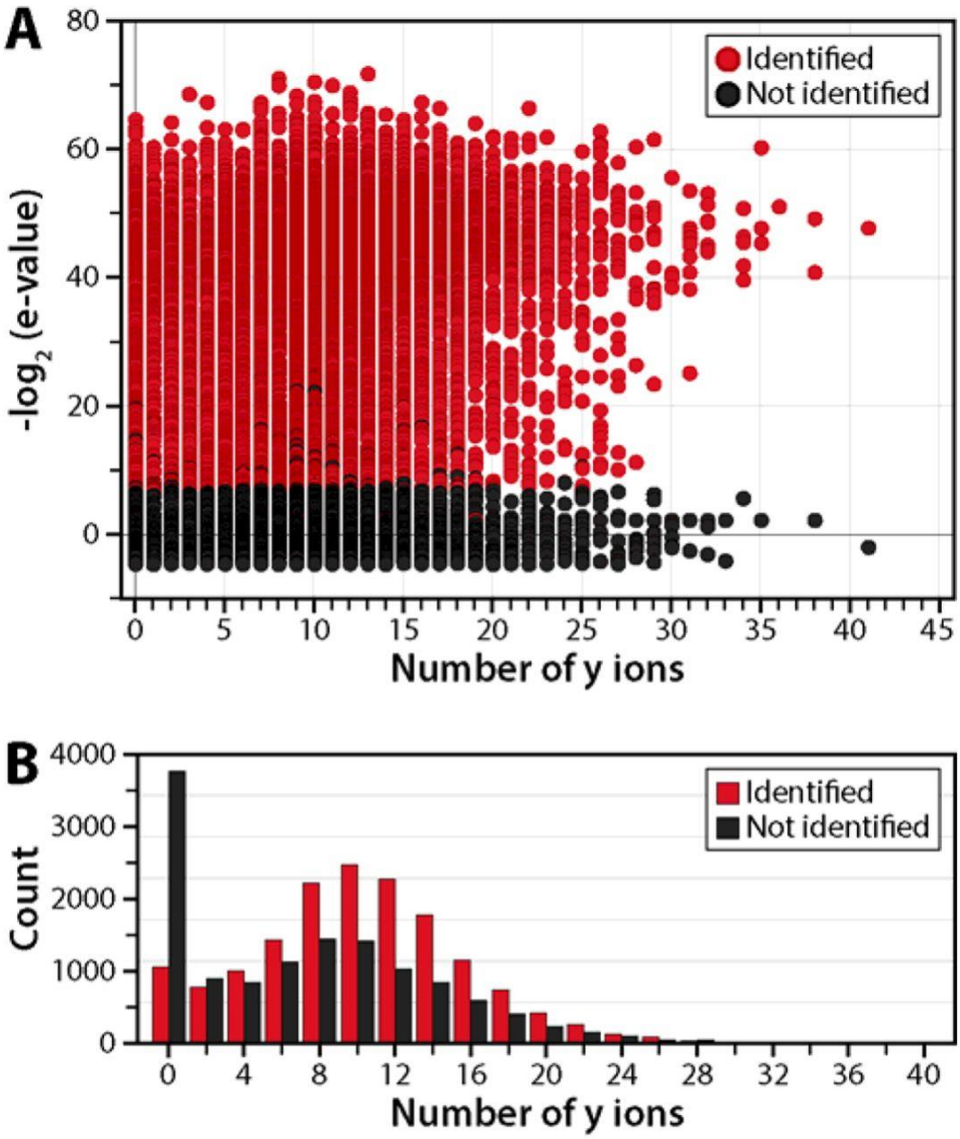


search. It is difficult, however, to know why these unmatched spectra did not correlate to a sequence in the database. Certainly, many of these are of dubious quality; more interesting, some might arise from precursors whose sequences either are not in the database or have been modified in ways that were not considered. It is likely all of these scenarios, among several others, are at play.

We reason that the y-type ion count could be a metric used to help determine spectral quality, similar to existing methods for assessing MS/MS quality, including the use of peak intensity[45] and sequence tags.[46-48] More specifically, we assert that unmatched spectra containing a high count of NeuCode doublets (*i.e.* y-type ions) may be peptidic in origin, and potentially salvageable. In this way the "y-type ion count" could be used as a metric by which to sort the proteomic wheat from the chaff. To test this idea, we used the compendium of 23,442 tandem mass spectra described above and searched them by use of the standard database correlation method.[49] Afterward we annotated all spectra (matched and unmatched) for y-type ions by use of our algorithm, described above. Next, we plotted the y-type ion count, as measured by our algorithm, *versus* the spectral match quality score (e-value, assigned by the search algorithm) for all spectra (**Figure 5*A***). Spectra not identified by database searching were assigned an e-value of 0. From these data we observed a similar distribution of y-type ion counts irrespective of a positive identification following traditional database searching. Note that there is a significant fraction of unidentified tandem mass spectra with no detectable y-type product ions (**Figure 5*B***). The bulk of these presumably result from nonpeptidic precursors, ineffective dissociation parameters, or precursors of very low abundance. Of the 9,858 unidentified spectra, 4,599 (46.7%) contained 8 or more y-type product ions.

These 4,599 spectra likely represent quality tandem mass spectra that contain considerable information but did not get mapped to sequence. We envisage triaging these spectra for broader searches or for *de novo* sequencing. To test this hypothesis, the applicability of NeuCode for *de novo* sequencing was benchmarked through the PepNovo$^+$ algorithm, as described above. A training dataset comprising 11 NeuCode yeast fractions analyzed over 90-min LC gradients resulted in 13,832 unique peptide identifications at a 1% FDR when searched with OMSSA against a UniProt yeast database (version 2011 07.[50] Raw files were converted into DTA spectra, with the in-house algorithm for selecting y-type ion

Figure 5. **Presence of precursor-derived NeuCode doublet fragment ions in tandem mass spectra that evade identification during database searching.** (***A***) Scatterplot showing the number of y ions present as a function of e-value. (***B***) Number of y-type ions present in the dataset.



doublets revised to increase the peak height of predicted y-type ions. This dataset was used to train PepNovo$^+$ predicted intensities and spectral features and compared with output of a traditional database search to establish scoring parameters. To compare the performance of NeuCode for *de novo* sequencing, raw files were also searched without any alterations to the DTA files—that is, by use of MS/MS without knowledge of which ions were y-type. In this dataset, 51.1% of the top NeuCode candidate raw files were also searched without any alterations to the DTA files—that is, by use of MS/MS without knowledge of which ions were y-type. In this dataset, 51.1% of the top NeuCode candidate sequences predicted by PepNovo$^+$ matched the database results. When the top five candidates were considered, this number increased to 67.9%. The identification of y-type ions provided a considerable boost in correct *de novo* identifications. These NeuCode results represent a 34% increase in correct identifications relative to the number obtained with MS/MS only, where 37.9% of spectra were correctly identified by PepNovo$^+$ (**Figure 6**). The percentage of correct sequences further increased when sequence tags were used rather than full *de novo* predictions, for this dataset peaking at 78.6% of NeuCode spectra mapped to the correct peptide sequence using amino acid tags of length 3. Once PepNovo$^+$ was trained, it was used to analyze our dataset of 23,442 tandem mass spectra. To increase the accuracy of PepNovo$^+$, only spectra with more than 15 fragment ions were considered. For this dataset, 61.6% of spectra were correctly ranked by PepNovo$^+$. When sequence tags were considered, this number increased to a high of 86.4%.

Correlation of Fragment to Precursor. There has been a recent trend toward data-independent acquisition routines in which larger precursor isolation windows are applied to dissociate and mass analyze product ions from multiple precursors in parallel. Sequential window acquisition of all theoretical fragment ion spectra (SWATH) is an example of one such approach.[51,52] These NeuCode doublets could likewise be used to group identified y-type ions from disparate precursors by matching the NeuCode ratios. To explore this idea, we performed an experiment in which we divided the MS^1 into sequential 25 *m*/*z* windows. During data-dependent MS/MS scanning, each selected precursor was isolated first with a 0.7 *m*/*z* window,

Figure 6. **Utilization of *y-type* ion predictions to assist *de novo* sequencing.** (***A***) Graph illustrating the percentage of correct *de novo* identifications with PepNovo$^+$ by use of either MS/MS or MS/MS with NeuCode y-type ions. (***B***) Top *de novo* predictions by use of PepNovo$^+$.

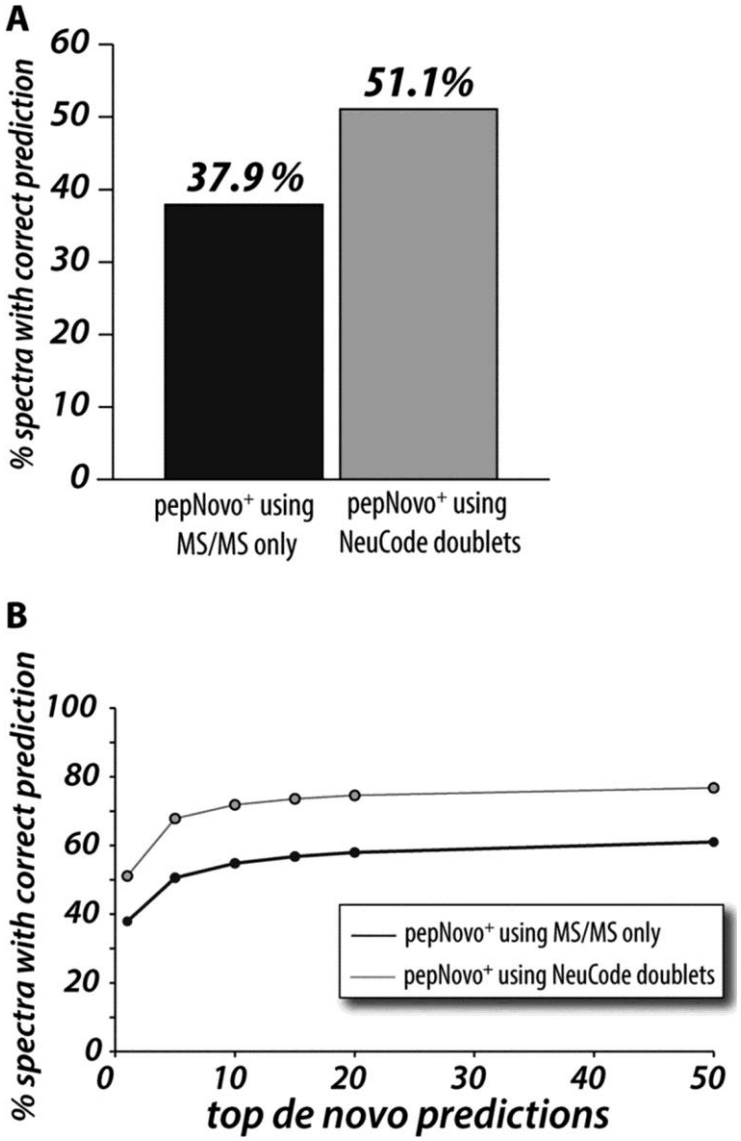
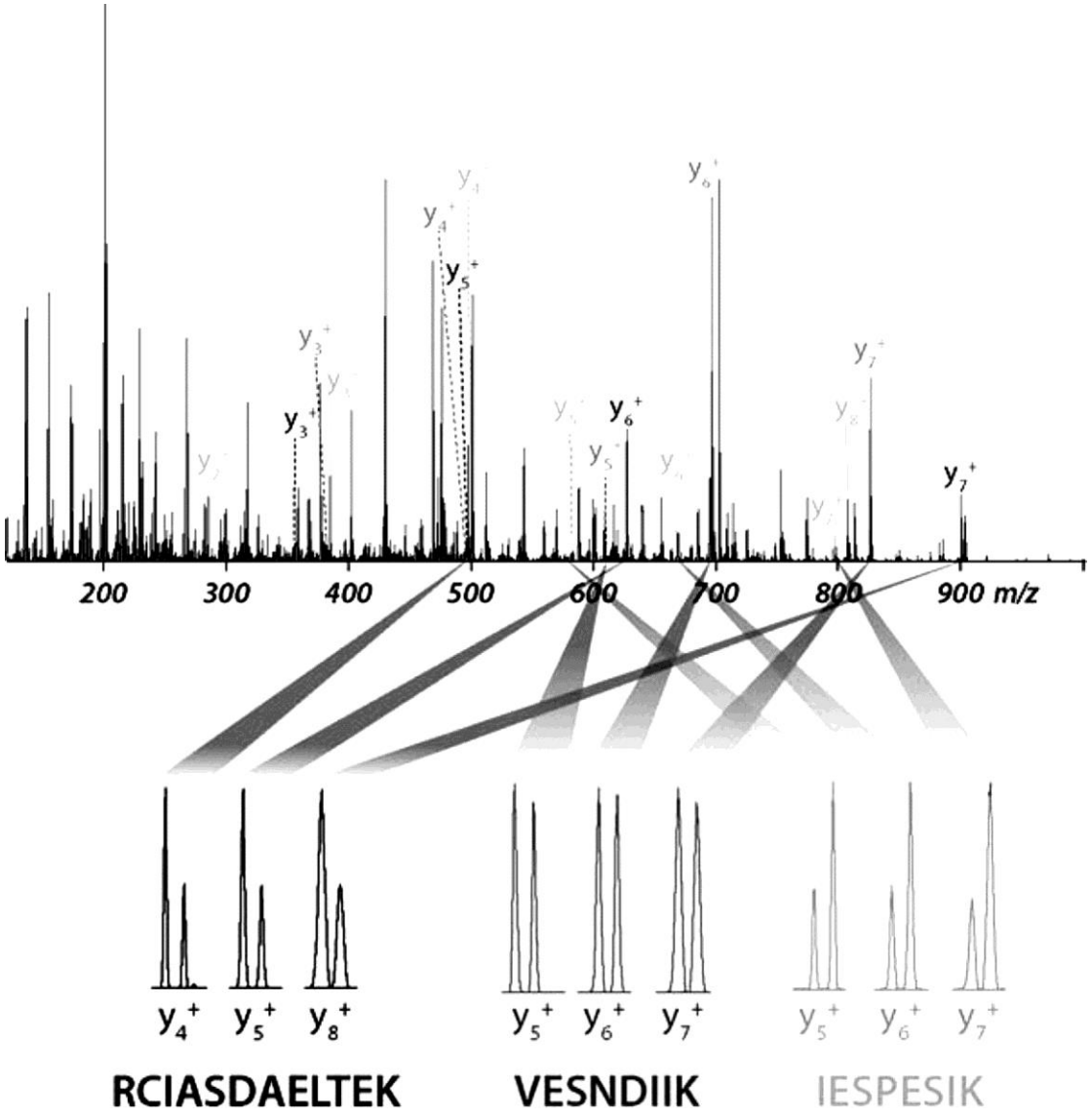


Figure 7. **Utilization of *y-type* ion predictions to distinguish fragment ions resulting from co-isolated and co-fragmented precursor species.** Identification of peptides ***i***. RCIASDAELTEK, ***ii***. VESNDIIK, and ***iii***. IESPESIK based on the relative abundances of lysine isotopologues with a precursor isolation window of 25 *m/z*.



dissociated, and mass analyzed. Immediately following that scan, a second isolation was performed, this time with a 25 *m*/*z* window. That MS/MS scan, which simulated a SWATH experiment, was mass analyzed with high resolving power. Based on the differences in the relative abundances between the NeuCode peptides, we reasoned that product ions reasoned that product ions could be mapped back to their corresponding precursor displaying the same ratio. **Figure 7** demonstrates this concept. A 25 *m*/*z* isolation window was applied to the precursor ion at

m/*z* 467.57. A list of identified peptides co-eluting during that scan was generated. We then searched for y-type ions corresponding to these peptides in the wide isolation MS^2 scans. Low *m/z* y-ions were avoided, as they are more likely to be nonspecific to a given peptide, resulting in distorted ratios. It was determined that peptides RCIASDAELTEK, VESNDIIK, and IESPESIK had sufficient y-type ions for identification. NeuCode y-type ion intensities were extracted, with average peak height ratios between the two isotopologues of 1.88, 1.10, and 0.47, respectively (**Figure 7**). Based on these data, we anticipate the possible use of NeuCode SILAC to assist in the simplification of product ion spectra from co-fragmented species.

DISCUSSION

Here we described a fresh approach that can be used to annotate product ion type. The primary advantage of the method is its use of closely spaced NeuCode amino acid isotopologues that ensure the co-isolation of labeled peptide precursors. Product ion annotation by use of NeuCode will require the collection of tandem mass spectra under higher resolving powers than typically used. That said, through theory and experiment we have demonstrated that a resolving power of ~50,000 is sufficient, which puts the technology within reach of all Fourier transform MS systems and higher-end TOF analyzers. The major downside of this requirement is the added transient acquisition time needed to achieve the higher resolving power. Here we show that only subtle penalties in total scan number and identifications are incurred relative to the results obtained with a standard MS/MS resolving power of 15,000, as opposed to the 60,000 needed here (18,654 *versus* 16,230 scans and 2,839 *versus* 2,546 identifications, respectively). Despite the

reduction in scans and, consequently, identifications, we are positioned to directly annotate product ion types with these improved data. Moving forward, we anticipate further improvements in instrument acquisition and resolving power so that this time penalty will likely become inconsequential. Note that just five years ago acquisitions of tandem MS/MS spectra at resolving powers exceeding 20,000 were far from routine. Finally, we note that methods do exist for the analysis of multiple product ion sets in parallel, providing another avenue by which to eliminate the time penalty of high-resolving-power data collection.

Through an in-house algorithm, we extracted NeuCode labeled y-type product ions from raw tandem mass spectra with excellent sensitivity and precision. We envision many avenues of research stemming from this core technology. First, with knowledge of C-terminal products, one can develop simple algorithms to extract N-terminal product information. Thus, we have a tractable method by which to eliminate the vast majority of spectral noise, both electronic and chemical. A second spoke likely leads to increased spectral searching rates. Specifically, with reductions in spectral complexity of over 90%, significant gains in spectral processing times should be achievable. We also imagine multiple routes to more rigorous methods by which to perform FDR filtering. Much as mass accuracy is used to filter decoy hits from true positive results, one could examine correct product ion assignment to candidate sequences as a filter to increase the number of spectral identifications at a fixed FDR rate. Yet another path is to use the added specificity of annotated product ions to increase the number of variable modifications considered.

Though our current algorithm predicts C-terminal product ions with high accuracy, we will continue its optimization so as to achieve a sensitivity and specificity closer to 100%. We hypothesize that information is lost in the form of peaks that do not show doublet splitting, either because they are b-type ions or because they are y-type ions with an undetectable doublet. A more intensive algorithm would consider some peaks as y-type product ions only and others as b- or possibly b- and y-type product ions, as in a standard search. Further improvements include the determination of complementary product pairs on the basis of the precursor mass and y-type products identified by peak splitting. This could potentially yield

very information-rich peak lists in which most fragments are known to be b- or y-type *a priori* and virtually all electronic and chemical noise is filtered out.

The final, and perhaps most enticing, use of the NeuCode product ion annotation approach is to facilitate large-scale *de novo* sequence analysis. Here we have provided preliminary evidence that y-type product ion identification can easily distinguish good spectra from bad. Further, we use this information to feed automated *de novo* analysis algorithms already in existence.

EXPERIMENTAL PROCEDURES

Sample preparation. *Saccharomyces cerevisiae* strain BY4741 Lys1Δ was grown in defined synthetic complete (SC, Sunrise Science, San Diego, CA) drop-out media with either heavy $6C^{13}/2N^{15}$ lysine (+8.0142 Da, Cambridge Isotopes, Tewksbury, MA), or heavy 8D (+8.0502 Da, Cambridge Isotopes). Cells were propagated to a minimum of 10 doublings. At mid-log phase, cells were harvested via centrifugation at 3,000 × *g* for 3 min and then washed three times with chilled double distilled H_2O. Cell pellets were resuspended in 5 ml lysis buffer (50 mM Tris pH 8, 8 M urea, 75 mM sodium chloride, 100 mM sodium butyrate, 1 mM sodium orthovanadate, protease and phosphatase inhibitor tablet), and protein was extracted via glass bead milling (Retsch, Haan, Germany). Protein concentration was measured via BCA (Pierce). Cysteines in the yeast lysate were reduced with 5 M dithiothreitol at ambient temperature for 30 min, alkylated with 15 mM iodoacetamide in the dark at ambient temperature for 30 min, and then quenched with 5 mM dithiothreitol. 50 mM tris (pH 8.0) was used to dilute the urea concentration to 4 M. Proteins were digested with LysC (1:50 enzyme: protein ratio) at ambient temperature for 16 h. The digestion was quenched with TFA and desalted with a tC18 Sep-Pak (Waters, Etten-Leur, The Netherlands). Samples were prepared by mixing $6C^{13}/2N^{15}$ (+8.0412 Da) and 8D (+8.0502 Da) labeled peptides in 1:1 ratios by mass. For strong cation exchange fractionation, peptides were dissolved in 400 µl of strong cation exchange buffer A (5 mM KH_2PO_4 and 30% acetonitrile; pH 2.65) and injected onto a polysulfoethylaspartamide column (9.4 mm × 200 mm; PolyLC) attached to a Surveyor LC quarternary pump (Thermo Electron, West Chester, PA) operating at 3 ml/min. Peptides were detected by photodiode array detector (Thermo Electron,

West Chester, PA). Fractions were collected every 2 min starting at 10 min into the following gradient: 0–2 min at 100% buffer A, 2–5 min at 0%–15% buffer B (5 mM KH_2PO_4, 30% acetonitrile, and 350 mM KCl (pH 2.65)), and 5–35 min at 15%–100% buffer B. Buffer B was held at 100% for 10 min. Finally, the column was washed with buffer C (50 mM KH_2PO_4 and 500 mM KCl (pH 7.5)) and water before recalibration. Fractions were collected by hand every 2 to 3 min starting at 10 min into the gradient and were lyophilized and desalted with a tC18 Sep-Pak (Waters).

LC-MS/MS. Samples were loaded onto a 15-cm-long, 75-μm capillary column packed with 5 μm Magic C18 (Michrom, Auburn, CA) particles in mobile phase A (0.2% formic acid in water). Peptides were eluted with mobile phase B (0.2% formic acid in acetonitrile) over a 120-min gradient at a flow rate of 300 nl/min. Eluted peptides were analyzed by an Orbitrap Elite mass spectrometer. For the nonfractionated samples, mass spectrometer instrument methods comprised one MS^1 scan followed by data-dependent MS^2 scans of the five most intense precursors. A survey MS^1 scan was performed by the Orbitrap at 30,000 resolving power to identify precursors to sample for tandem mass spectrometry, and this was followed by an additional MS^1 scan at 480,000 resolving power (at *m*/*z* 400; actual mass resolving power of 470,700). Data-dependent tandem mass spectrometry was performed via beam-type collisional activated dissociation (HCD) in the Orbitrap at a resolving power of 15,000, 60,000, 120,000, or 240,000 and a collision energy of 30. Preview mode was enabled, and precursors of unknown charge or with a charge of +1 were excluded from MS^2 sampling. For experiments comparing the duty cycle and resolving power required in order to distinguish *y*-ion doublets, MS^1 and MS^2 target ion accumulation values were set to 5×10^5 and 5×10^4, respectively. For all other experiments, MS^1 target accumulation values were set to 1×10^6 and MS^2 accumulation values were set to 4×10^5. Dynamic exclusion was set to 30 s for −0.55 *m*/*z* and +2.55 *m*/*z* of selected precursors. For ETD analysis, data-dependent top-five mass spectrometry was performed at a resolving power of 240,000 (*m*/*z* 400; actual MS^2 mass resolving power of 271,000).[53] ETD accumulation values were set to 1×10^6 for MS^1 target accumulation and 4×10^5 for MS^2 target accumulation. The fluoranthene reaction time was set to 100 ms. For the high-pH strong cation exchange

fractions, data-dependent tandem mass spectrometry was performed via HCD at a resolving power of either 60,000 or 120,000 and a collision energy of 30. Preview mode was enabled, and precursors of unknown charge or with a charge of +1 were excluded from MS^2 sampling. MS^1 targets were set to 1×10^6, and MS^2 accumulation values were set to 4×10^5. Dynamic exclusion was set to 45 s for −0.55 *m*/*z* and +2.55 *m*/*z* of selected precursors. Analysis by use of a wide isolation window was performed on an Orbitrap Fusion. MS^1 analysis was performed at 450,000 resolving power (*m*/*z* 200), and MS^2 analysis was performed at 120,000 resolving power (*m*/*z* 400). Data-dependent top-N mass spectrometry was performed, with precursors selected from sequential 25-Da windows. HCD was performed twice on the same precursor, first by use of a quadrupole isolation width of 0.7 *m*/*z* for peptide identification, and then using 25 *m*/*z* quadrupole isolation. Fragment ions were analyzed in the Orbitrap at a mass resolving power of 120,000 (*m*/*z* 400). MS^1 and MS^2 target accumulation values were set to 2×10^5 and 5×10^4, respectively.

Data Analysis. Thermo.raw files were converted to searchable DTA text files using the Coon OMSSA Proteomic Analysis Software Suite (COMPASS).[54] DTA files containing exclusively y-ions were generated using an in-house algorithm. DTA files were searched against the UniProt yeast database (version 132) with Lys-C specificity using the Open Mass Spectrometry Search Algorithm (OMSSA), version 2.1.9.[49] Methionine oxidation was searched as a variable modification. Cysteine carbamidomethylation and the mass shift imparted by the lysine isotopolgues were searched as fixed modifications. For MS^2 scans performed at a resolving power of 60,000, 120,000, or 240,000, a shift of +8.0142, representing the mass shift of the $^{13}C_6{}^{15}N_2$ isotopologue, was searched. For MS^2 scans performed at 15,000 resolving power, the average shift of the $^{13}C_6{}^{15}N_2$ and $^{8}H_2$ isotopologues (+8.0322) was searched. For all analyses, the precursor mass was obtained from the 480,000 MS scan. The precursor mass tolerance was defined as 50 ppm, and the fragment ion mass tolerance was set to 0.01 Da. Using the COMPASS software suite, obtained search results were filtered to 1% FDR based on E-values. y-ion doublets were extracted from raw files using an in-house algorithm. Briefly, an ensemble of three different machine learning models was used to score each MS/MS spectral peak for C-terminal product ion prediction. To train our ensemble learner to correctly

distinguish C-terminal product ion peaks from N-terminal product ion peaks and noise peaks within our experimental MS/MS spectra, we generated a representative training set of spectral data. Instances used for training and test sets were peaks acquired only from MS/MS spectra associated with a peptide identification. Peaks with a signal-to-noise value of less than 5 were not used. Feature information for each training/testing instance was extracted from raw spectral data. Seven MS/MS spectral features were selected to generate training and test set data: (1) "has doublet" (evaluated as "true" only if a spectral peak could be found at the predicted *m*/*z* of the peak's "heavy" partner), (2) "signal-to-noise" (discretized using a scale of 1–5 based on the peak's signal-to-noise value), (3) "is isotope," (4) "is neutral loss," (5) "number of isotopes," (6) "number of doublet isotopes," and (7) "has neutral loss."

To evaluate NeuCode SILAC labeling for automated *de novo* sequencing, PepNovo^{+}[8] was benchmarked on y-ion predicted spectra. First, a set of identified spectra from 13,832 unique peptides (>7,400 per precursor charge 2–3) was produced to train PepNovo^{+} so it could learn features such as the relative peak height ranks of b/y-ions and the probability of noise at each mass interval. These training spectra were acquired under the 11 NeuCode yeast strong cation exchange fractions prepared as described above. Thermo raw files were converted into mzXML format using ProteoWizard v2.2.2828 (with peak-picking turned on) and identified by MS-GF^{+} v9358[55] at a 1% spectrum-level FDR against the UniProt yeast database (plus isoforms), v20110729. A fixed modification of K^{+}8.0142 was imposed along with variable modifications of oxidized Met and deamidated Asn/Gln. All MS/MS scans were searched with a 50-ppm precursor mass tolerance, the high-accuracy LTQ instrument setting, the HCD fragmentation setting, and one allowed missed Lys-C cleavage.

Thermo.raw files were also converted into DTA spectra as before, except the in-house algorithm for selecting y-ion doublets was slightly altered to boost the peak height of predicted y-ions above that of other peaks (the cumulative peak height was equal to the sum of the monoisotopic doublet peaks, all isotopic doublet peaks, and two times the peak height of the base peak) and to convert their *m*/*z* to charge one. Remaining peaks not predicted to be y-ions were converted to charge one by a previously described MS/MS deconvolution tool.[56] Deconvoluted DTA spectra that originated from identified MS/MS scans were then

paired with the MSGF$^+$ peptide IDs and passed to PepNovo$^+$ for training. The resulting PepNovo$^+$ scoring model lacked the rank-boosting component,[57] which requires identified spectra from >100,000 unique peptides per precursor charge state and extensive modification of the PepNovo$^+$ source code to train. Still, the model was sufficient to perform *de novo* peptide sequencing on the y-ion predicted spectra. PepNovo$^+$ was also run on the raw MS/MS scans (mzXML spectra converted to MGF with all MS/MS peaks converted to charge one) by use of a previously trained HCD scoring model that also lacks the rank-boosting component.[40] The following PepNovo$^+$ parameters were set at all stages of training and benchmarking: fixed modification of K$^+$8.0142; variable modifications of oxidized Met and deamidated Asn; 0.01-Da fragment mass tolerance; use of spectrum precursor charge; and use of spectrum precursor *m*/z.

ACKNOWLEDGEMENTS

We thank A. J. Bureta for assistance with figure illustrations and A. E. Merrill and A. S. Hebert for culturing the yeast cells. We appreciate the intellectual stimulation provided by an anonymous reviewer to apply our approach to the correlation of product and precursor ions when multiple precursors are co-fragmented. This work was supported by National Institutes of Health Grant GM080148 to J.J.C. A.L.R. was supported by an NHGRI training grant to the Genomic Sciences Training Program (5T32HG002760). C.E.V. was supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM T15LM007359). C.M.R. was funded by an NSF Graduate Research Fellowship and NIH Traineeship (T32GM008505).

REFERENCES

(1) Seidler, J.; Zinn, N.; Boehm, M. E.; Lehmann, W. D. *Proteomics* **2010**, *10*, 634-649.

(2) Liska, A. J.; Shevchenko, A. *Proteomics* **2003**, *3*, 19-28.

(3) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. *Anal Chem* **2007**, *79*, 4870-4878.

(4) Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X. *J Proteome Res* **2006**, *5*, 3018-3028.

(5) Pitzer, E.; Masselot, A.; Colinge, J. *Proteomics* **2007**, *7*, 3051-3054.

(6) Taylor, J. A.; Johnson, R. S. *Anal Chem* **2001**, *73*, 2594-2604.

(7) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc Natl Acad Sci U S A* **2000**, *97*, 10313-10317.

(8) Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. *J Proteome Res* **2007**, *6*, 114-123.

(9) Pan, C.; Park, B. H.; McDonald, W. H.; Carey, P. A.; Banfield, J. F.; VerBerkmoes, N. C.; Hettich, R. L.; Samatova, N. F. *Bmc Bioinformatics* **2010**, *11*.

(10) Chi, H.; Sun, R. X.; Yang, B.; Song, C. Q.; Wang, L. H.; Liu, C.; Fu, Y.; Yuan, Z. F.; Wang, H. P.; He, S. M.; Dong, M. Q. *Journal of Proteome Research* **2010**, *9*, 2713-2724.

(11) Taouatas, N.; Drugan, M. M.; Heck, A. J. R.; Mohammed, S. *Nat Methods* **2008**, *5*, 405-407.

(12) Altelaar, A. F. M.; Navarro, D.; Boekhorst, J.; van Breukelen, B.; Snel, B.; Mohammed, S.; Heck, A. J. R. *P Natl Acad Sci USA* **2012**, *109*, 407-412.

(13) van Breukelen, B.; Georgiou, A.; Drugan, M. M.; Taouatas, N.; Mohammed, S.; Heck, A. J. R. *Proteomics* **2010**, *10*, 1196-1201.

(14) Gu, S.; Pan, S. Q.; Bradbury, E. M.; Chen, X. *Analytical Chemistry* **2002**, *74*, 5774-5785.

(15) Noga, M. J.; Asperger, A.; Silberring, J. *Rapid Commun Mass Sp* **2006**, *20*, 1823-1827.

(16) Hennrich, M. L.; Mohammed, S.; Altelaar, A. F. M.; Heck, A. J. R. *J Am Soc Mass Spectr* **2010**, *21*, 1957-1965.

(17) Munchbach, M.; Quadroni, M.; Miotto, G.; James, P. *Analytical Chemistry* **2000**, *72*, 4047-4057.

(18) Madsen, J. A.; Brodbelt, J. S. *Analytical Chemistry* **2009**, *81*, 3645-3653.

(19) Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. *P Natl Acad Sci USA* **1999**, *96*, 6591-6596.

(20) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol Cell Proteomics* **2002**, *1*, 376-386.

(21) Schnolzer, M.; Jedrzejewski, P.; Lehmann, W. D. *Electrophoresis* **1996**, *17*, 945-953.

(22) Shevchenko, A.; Chernushevich, I.; Ens, W.; Standing, K. G.; Thomson, B.; Wilm, M.; Mann, M. *Rapid Commun Mass Sp* **1997**, *11*, 1015-1024.

(23) Uttenweiler-Joseph, S.; Neubauer, G.; Christoforidis, A.; Zerial, M.; Wilm, M. *Proteomics* **2001**, *1*, 668-682.

(24) Goodlett, D. R.; Keller, A.; Watts, J. D.; Newitt, R.; Yi, E. C.; Purvine, S.; Eng, J. K.; von Haller, P.; Aebersold, R.; Kolker, E. *Rapid Commun Mass Sp* **2001**, *15*, 1214-1221.

(25) Goodlett, D. R.; Yi, E. C. *Trac-Trend Anal Chem* **2003**, *22*, 282-+.

(26) Kim, J. S.; Shin, M.; Song, J. S.; An, S.; Kim, H. J. *Anal Biochem* **2011**, *419*, 211-216.

(27) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Analytical Chemistry* **2003**, *75*, 156a-165a.

(28) Lee, Y. H.; Han, H.; Chang, S. B.; Lee, S. W. *Rapid Commun Mass Sp* **2004**, *18*, 3019-3027.

(29) Hsu, J. L.; Huang, S. Y.; Chow, N. H.; Chen, S. H. *Analytical Chemistry* **2003**, *75*, 6843-6852.

(30) Ji, C. J.; Li, L. *Journal of Proteome Research* **2005**, *4*, 734-742.

(31) Hubler, S. L.; Jue, A.; Keith, J.; McAlister, G. C.; Craciun, G.; Coon, J. J. *Journal of the American Chemical Society* **2008**, *130*, 6388-6394.

(32) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. *Nature Methods* **2013**, *10*, 332-+.

(33) Satoh, T.; Sato, T.; Tamura, J. *J Am Soc Mass Spectr* **2007**, *18*, 1318-1323.

(34) Klitzke, C. F.; Corilo, Y. E.; Siek, K.; Binkley, J.; Patrick, J.; Eberlin, M. N. *Energ Fuel* **2012**, *26*, 5787-5794.

(35) Denisov, E.; Damoc, E.; Lange, O.; Makarov, A. *Int J Mass Spectrom* **2012**, *325*, 80-85.

(36) Scigelova, M.; Hornshaw, M.; Giannakopulos, A.; Makarov, A. *Mol Cell Proteomics* **2011**, *10*.

(37) Kaiser, N. K.; Quinn, J. P.; Blakney, G. T.; Hendrickson, C. L.; Marshall, A. G. *J Am Soc Mass Spectr* **2011**, *22*, 1343-1351.

(38) Kaiser, N. K.; McKenna, A. M.; Savory, J. J.; Hendrickson, C. L.; Marshall, A. G. *Analytical Chemistry* **2013**, *85*, 265-272.

(39) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A. *Mol Cell Proteomics* **2012**, *11*.

(40) Guthals, A.; Clauser, K. R.; Frank, A. M.; Bandeira, N. *Journal of Proteome Research* **2013**, *12*, 2846-2857.

(41) Frese, C. K.; Altelaar, A. F. M.; Hennrich, M. L.; Nolting, D.; Zeller, M.; Griep-Raming, J.; Heck, A. J. R.; Mohammed, S. *Journal of Proteome Research* **2011**, *10*, 2377-2388.

(42) Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. *Mol Cell Proteomics* **2007**, *6*, 1942-1951.

(43) Swaney, D. L.; McAlister, G. C.; Coon, J. J. *Nature Methods* **2008**, *5*, 959-964.

(44) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. *Analytical Chemistry* **2007**, *79*, 5620-5632.

(45) Sadygov, R. G.; Cociorva, D.; Yates, J. R. *Nat Methods* **2004**, *1*, 195-202.

(46) Frank, A.; Tanner, S.; Bafna, V.; Pevzner, P. *Journal of Proteome Research* **2005**, *4*, 1287-1295.

(47) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J. L.; Chambers, M. C. *Journal of Proteome Research* **2008**, *7*, 3838-3846.

(48) Mann, M.; Wilm, M. *Analytical Chemistry* **1994**, *66*, 4390-4399.

(49) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X. Y.; Shi, W. Y.; Bryant, S. H. *Journal of Proteome Research* **2004**, *3*, 958-964.

(50) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H. Z.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. L. *Nucleic Acids Res* **2005**, *33*, D154-D159.

(51) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol Cell Proteomics* **2012**, *11*.

(52) Silva, J. C.; Denny, R.; Dorschel, C. A.; Gorenstein, M.; Kass, I. J.; Li, G. Z.; McKenna, T.; Nold, M. J.; Richardson, K.; Young, P.; Geromanos, S. *Analytical Chemistry* **2005**, *77*, 2187-2200.

(53) McAlister, G. C.; Phanstiel, D.; Good, D. M.; Berggren, W. T.; Coon, J. J. *Analytical Chemistry* **2007**, *79*, 3525-3534.

(54) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. *Proteomics* **2011**, *11*, 1064-1074.

(55) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J. R.; Pevzner, P. A. *Mol Cell Proteomics* **2010**, *9*, 2840-2852.

(56) Guthals, A.; Clauser, K. R.; Bandeira, N. *Mol Cell Proteomics* **2012**, *11*, 1084-1096.

(57) Frank, A. M. *Journal of Proteome Research* **2009**, *8*, 2241-2252.

Chapter 4

The One Hour Yeast Proteome

This chapter has been published:

Hebert AS*, **Richards AL***, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ. *The One Hour Yeast Proteome*. Molecular & Cellular Proteomics. **2014,** *13*, 339-347.

*Authors contributed equally

ABSTRACT

We describe the comprehensive analysis of the yeast proteome in just over one hour of optimized analysis. We achieve this expedited proteome characterization with improved sample preparation, chromatographic separations, and by using a new Orbitrap hybrid mass spectrometer equipped with a mass filter, a collision cell, a high-field Orbitrap analyzer, and, finally, a dual cell linear ion trap analyzer (Q-OT-qIT, Orbitrap Fusion). This system offers high MS^2 acquisition speed of 20 Hz and detects up to 19 peptide sequences within a single second of operation. Over a 1.3 h chromatographic method, the Q-OT-qIT hybrid collected an average of 13,447 MS^1 and 80,460 MS^2 scans (per run) to produce 43,400 ($\bar{x}$) peptide spectral matches and 34,255 ($\bar{x}$) peptides with unique amino acid sequences (1% false discovery rate (FDR)). On average, each one hour analysis achieved detection of 3,977 proteins (1% FDR). We conclude that further improvements in mass spectrometer scan rate could render comprehensive analysis of the human proteome within a few hours.

INTRODUCTION

The ability to measure differences in protein expression has become key to understanding biological phenomena.[1,2] Owing to cost, speed, and accessibility, transcriptomic analysis is often used as a proteomic proxy.[3,4] That said, mRNA is a genetic intermediary and cannot inform on the myriad of post-translational regulation processes.[5-7] For the past decade considerable effort has been invested in maturing proteomic technology to deliver information at a rate and cost commensurate to transcriptomic technologies.

Historically yeast, with its 6600 open reading frames, has been the preferred proteomic technology test-bed.[8] In 2003, Weissmann and colleagues measured approximate expression levels of each yeast gene using either GFP or TAP tags.[9] This seminal work established that ~4500 proteins are expressed during log-phase yeast growth. Subsequent mass spectrometry-based studies have confirmed this early estimate.[10-12] With this knowledge, we hereby define comprehensive proteome analysis as an experiment that detects ~90% of the expressed proteome (≥ 4000 proteins for yeast). Note others have used the term "nearly complete" for this purpose; we posit that comprehensive has identical meaning (*i.e.* including many, most, or all things).[13]

Initial MS-based proteomic analyses of yeast, each identifying up to a few hundred proteins, were conducted using a variety of separation and MS technologies.[14-16] Yates and co-workers reported the first large-scale yeast proteome study in 2001 with the identification of 1483 proteins following ~ 68 h of mass spectral analysis, *i.e.* 0.4 proteins were identified per minute.[17] Their method—two dimensional chromatography coupled with tandem mass spectrometry—has provided a template for large-scale protein analysis for the past decade.[18-20] By incorporating an offline first dimension of separation with more extensive fractionation (80 *versus* 15) Gygi *et al.* expanded on this work in 2003.[21] That said, the modest increase in identified proteins (1504) required 135 h of analysis, reducing the protein per minute count to 0.2. Armed with a faster hybrid mass spectrometer capable of accurate mass measurement, Mann and colleagues achieved detection of 2003 yeast proteins in an impressive 48 h (0.7 proteins/minute) in 2006.[22]

From these three pioneering studies we begin to see the impact of mass spectrometer acquisition rate on the depth and rate of proteome analysis. The most recent application of such technology to the yeast proteome, however, The Mann work used a hybrid linear ion trap-ion cyclotron resonance Fourier transform instrument (LTQ-FT) that delivered MS^2 scans at a rate of ~650 ms.[23] The earlier studies, *i.e.* Yates and Gygi, relied on the considerably slower scanning (1–3 s/scan) three-dimensional ion trap technology. In 2008, using the novel Orbitrap hybrid mass spectrometer, Mann and colleagues reported on the first comprehensive analysis of the yeast proteome by identifying nearly 4000 proteins.[10] Extensive fractionation and triplicate analysis of each fraction rendered the study a considerable time investment at ~144 analysis hours (0.5 proteins/minute). In 2010 our group achieved similar comprehensive analysis, but improved sequence coverage, using fractionation and multiple proteases.[24] That work, however, required even longer analysis time (0.2 proteins/min).

And that was the state-of-the-art as recently as three years ago. Doubtless we, the proteomics community, had achieved one momentous goal—comprehensive coverage of the yeast proteome. Still, obtaining this depth was not routine as it mandated days of MS analysis and a considerable amount of expert labor. In 2012, with new, even faster scanning, quadrupole-Orbitrap technology (Q-OT, Q-Exactive), Mann and colleagues dispatched the concept of fractionation, improved the quality of sample preparation, and placed emphasis on higher quality online separations.[25] With their streamlined method they achieved detection of just over 3900 yeast proteins following four hours of MS analysis. Even more impressive this strategy translated to the identification of 16.3 proteins per minute—a 33-fold improvement over the next best comprehensive study. This success was a remarkable achievement and illustrates that comprehensive proteomic technology can indeed be executed in a time efficient manner.

Time-of-flight hybrid systems, of course, can deliver very high MS^2 acquisition rates, up to 100 Hz in some reports. In 2011, Muddiman and colleagues reported yeast proteome analysis using a quadrupole-TOF system (*i.e.* TripleTOF) operating at a much lower rate (20 Hz) MS^2 scan rate.[26] Even at this reduced rate, only 16% of the spectra were mapped to unique sequences and 1112 unique proteins identified. Because of reduced MS^2 spectral quality (*i.e.* low signal-to-noise, S/N), even fewer unique peptide

identifications were achieved at higher MS^2 acquisition rates. Other studies using TOF technologies report similar results.[27,28] For maximal proteome depth, we conclude that increased scan speed must not come at the cost of reduced spectral quality. Recently, a new Orbitrap hybrid mass spectrometer having a mass filter, a collision cell, a high-field Orbitrap analyzer, and, finally, a dual cell linear ion trap analyzer was described (Q-OT-qIT, Orbitrap Fusion).[29,30] This system offers high MS^2 acquisition speed of 20 Hz—double that of the Q-OT system used by Mann and colleagues. We postulated that this fresh system, with its fast scan rate, could provide comprehensive proteome analysis in record time. To maximize performance we developed an optimized cellular lysis approach, employed trypsin digestion, and used dimethyl sulfoxide (DMSO, 5%) as an LC additive to increase abundance of acidic peptides and unify charge state.[31,32] Using this novel system we report the comprehensive analysis of the yeast proteome (4002 with 1% FDR) following 1.3 h of nLC-MS^2 analysis (70 min gradient). These experiments delivered an extraordinary 67 proteins per minute and demonstrate that complete analysis of the yeast proteome can be routinely performed in approximately one hour.

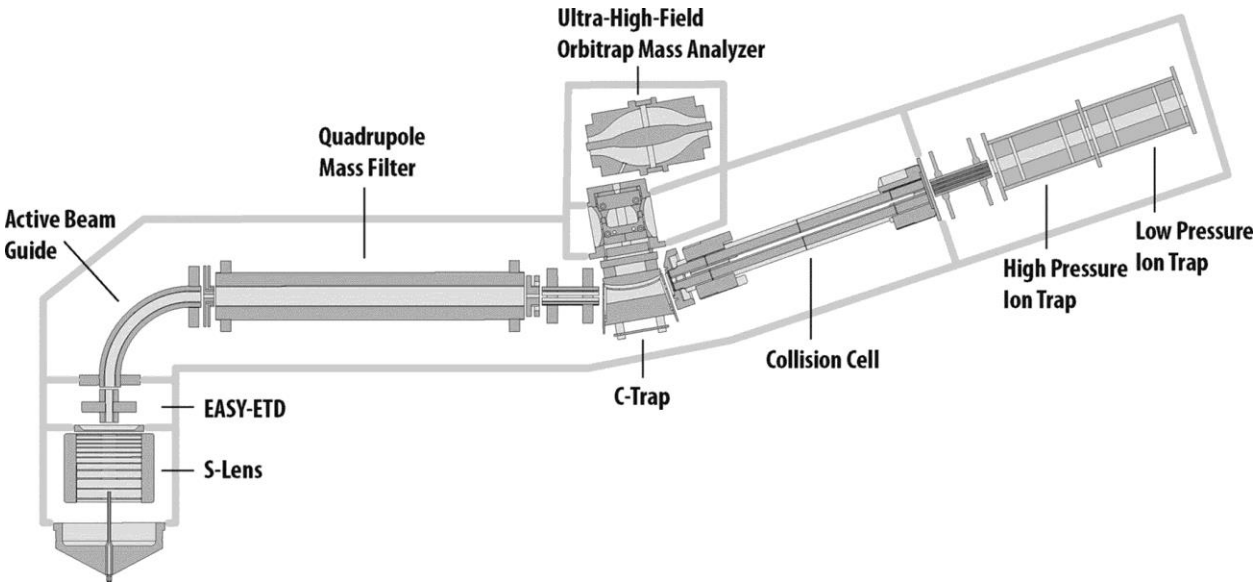
RESULTS

Considerable gains in the depth and rate of proteomic analysis have been realized over the past decade (*vide supra*). These improved results stem from routine use of high mass accuracy and resolution, but also from a steady increase in MS^2 acquisition rate. In the decade spanning the seminal Yates publication in 2001 and the single-shot proteome work of Mann *et al.* in 2012, MS^2 sampling rates rose from ~0.75 Hz to nearly 10 Hz.[17,25] Here we report on an even newer generation of mass spectrometer that comprises a mass resolving quadrupole, Orbitrap, collision cell, and linear ion trap (Q-OT-qIT, Fusion, **Figure 1**).[30] In this system, MS acquisition rate is not only boosted by the presence of a very fast scanning dual cell linear ion trap, but also by a control environment having multiple, independent processing units. The new system is equipped with a sophisticated control system that parallelizes the processes of ion injection, precursor isolation, fragmentation, and mass analysis to achieve a ~2× boost in acquisition rates. We reasoned that this Q-OT-

qIT configuration, with its 20 Hz MS^2 acquisition rate, could afford a considerable gain for rapid, whole proteome analysis.

The Q-OT-qIT System. To test this hypothesis we began by performing a parametric evaluation using a complex mixture of yeast tryptic peptides eluted into the system over a 70 min gradient. We examined several settings including collisional activation mode (*i.e.*HCD or trap CAD), MS^1 resolution, collision energy, maximum inject time, and dynamic exclusion settings. Detailed plots highlighting these results are included in Supplemental Information. Briefly, we found that MS^2 analysis using HCD followed by ion trap mass analysis (low-res HCD; 80,626 MS^2 events with 33,127 unique PSMs) generated more identifications compared with ion trap CAD with ion trap mass analysis (CAD; 75,973 MS^2 events with 31,820 unique PSMs). This is not surprising as HCD tends to offer more random backbone fragmentation and, with the Q-OT-qIT geometry, can be accomplished slightly faster. Operation of the system with an MS^1 resolving power setting of 60,000 (@ *m*/*z* 200) afforded a 20% increase in detected unique peptides over 15,000 resolving power. We conclude the boosted resolving power elevates precursor signal-to-noise (S/N) ratios, allowing for improved selection of low abundance precursors, and can potentially separate otherwise unresolved precursors so that multiple MS^2events can be acquired. Of course, in this scenario such closely spaced precursors would be co-isolated (0.7 *m*/*z* isolation width); however, the selected precursor *m/z* annotated in the MS^2 scan would be different and would facilitate identification from a chimeric MS^2 scan. Increasing MS^1 resolving power above 60,000 did not provide any apparent benefit for increased identifications. Thirty-five milliseconds was the optimal maximum injection time. Decreasing the maximum injection time to 30 msec, or increasing it to 45 msec caused a 10% decrease in peptide identifications. We found only slight variations in peptide identifications among dynamic exclusion settings of 30, 45, and 60 s. Quadrupole isolation widths from 0.5 to 1.5 *m*/*z* were examined, the best results were achieved at a value of 0.7 *m*/*z*. MS^1 and MS^2 automatic gain control (AGC) target values of 500,000 and 7000, respectively, produced the maximum number of peptide identifications.

Figure 1. Schematic of the Q-OT-qIT hybrid mass spectrometer (Fusion). The system differs from previous generations of quadrupole ion trap/Orbitrap hybrids by introduction of a resolving quadrupole mass filter and rearrangement of the geometry to place the linear ion traps to the rear of the collision cell. The reconfigured geometry relieves the linear ion trap of two of its former functions—precursor ion isolation and dissociation. The consequence is both improved and faster operation.



Lysis, Chromatography, and Additives. Yeast cell lysis is a critical step in achieving comprehensive proteome detection and must be executed with care. Detergents, such as SDS, or bead beating are typical approaches for yeast lysis.[33-35] The SDS method, used in the Mann yeast studies, has produced excellent results, but requires removal of the detergent before MS sampling. The bead beating method mixes glass beads and yeast cells in a buffer slurry, which is shaken for three, 1–4 min cycles, at 30 Hz. This approach, however, can be too gentle to sufficiently lyse the yeast and, in our hands, does not efficiently extract all proteins. We aimed to avoid use of detergents and investigated a more vigorous bead beating procedure. By simply extending the number of cycles to eight (4 min each) we achieved considerably improved results. Finally, we note increased identifications when lysates were not cleared of insoluble material. Zubarev and colleagues recently reported similar findings for mammalian cell culture samples.[36]

Previous single-shot yeast proteome analyses utilized long capillary LC columns (50 cm) and lengthy gradients (240 min).[25,36,37] We aimed to achieve comparable or better coverage using a much shorter gradient (70 min). We found that capillary LC columns of 30 cm packed with 1.7 μm BEH particles (Waters Corporation) operating at flow rates of 350–375 nL/minute offered consistent elution across the one hour gradient. To accommodate this flow rate, a home-built column heater was maintained at a temperature of 60 °C throughout the separation. Sample was loaded directly onto the column to avoid losses.

Recent work by Kuster *et al.* described that addition of ~5% DMSO to the mobile phase solvents boosted precursor S/N, providing up to a 20% increase in protein identifications.[31,32] We tested addition of DMSO to our chromatography solvents by comparing the number of yeast peptide and protein identifications obtained over our 70 min gradient either with or without DMSO. In our hands the presence of DMSO increased the average precursor signal, from $\sim 2.8 \times 10^7$ to 4.8×10^7 (arbitrary units) and increased the total ion current by 170%. This amplified signal afforded a 9% increase in unique peptide identifications and 5% more proteins. We conclude that DMSO can indeed improve performance, with no apparent downside, and included it for all subsequent experiments.

Whole Yeast Proteome Analysis. To test our supposition that the faster scanning Q-OT-qIT hybrid could deliver comprehensive yeast proteome analysis within ~1 h, we sequentially analyzed trypsin-digested, yeast cell lysate in quintuplicate. Each replicate began by loading ~1.4 μg of sample and followed by analysis over a 70 min gradient. Allowing for sample loading, column washing, and equilibration, the five consecutive analyses consumed ~8 h; however, actual instrument acquisition times were ~1.3 h per experiment. As anticipated, the Q-OT-qIT hybrid posted a considerable number of scans: on average 13,447 MS^1 and 80,460 MS^2 events per run. As a point of reference, state-of-the-art analysis in 2003, by Gygi *et al.*, recorded 162,000 MS^2events following 135 h of MS operation. The Q-OT-qIT hybrid delivered this number of scans in two and a half hours! Next, we analyzed the yeast sample on the most recently introduced quadrupole linear ion trap Orbitrap system (*i.e.* qIT-OT or Orbitrap Elite) using the identical chromatographic conditions.[38] That mass spectrometer produced only about a quarter of the MS^1 scans, as compared with the Q-OT-qIT (3635), and half the MS^2 events (39,447).

Figure 2 presents a series of MS^2 scans acquired by the Q-OT-qIT MS over a 1 s period. In this example, 22 precursors were selected for MS^2 analysis from MS^1scan #59,211. All 22 product ion spectra were acquired, individually, within 1 s and are presented in the lower portion of **Figure 2**. Nineteen of these 22 scans were mapped to sequence (1% FDR). On database searching, each one hour experiment (Q-OT-qIT system) produced 43,400 ($\bar{x}$) peptide spectral matches (PSMs) and 34,255 ($\bar{x}$) peptides with unique amino acid sequences (1% FDR, **Table 1**). Batched analysis of the five experiments yielded 47,624 unique peptides. In each analysis, over half of the 80,460 MS^2 scans were mapped to sequence (54%). Despite the swift Q-OT-qIT scan rate (~20 Hz), we conclude the system routinely delivers spectra of high quality.

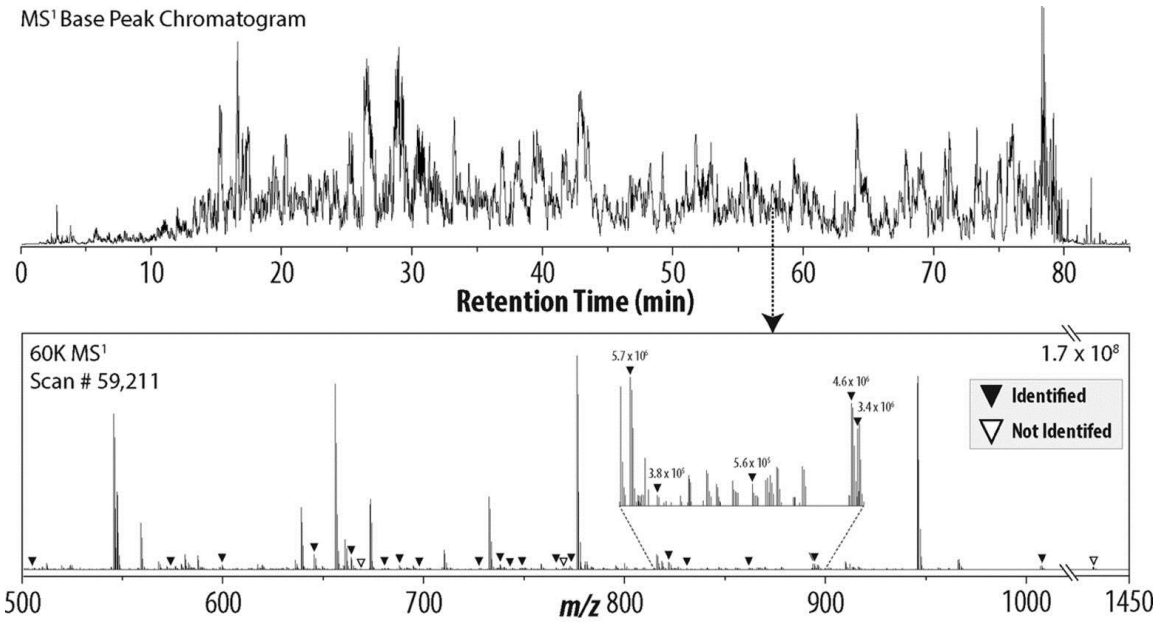
On average, each of the quintuplicate analyses achieved detection of 3,977 proteins (1% FDR), 13.5% (538, $\bar{x}$) of which originate from single peptides (**Figure 3*A***). Combination of the data reduces single peptide proteins to 460 while deepening coverage to 4395 protein groups. 3643 of these proteins (83%) were present in all five 1-h experiments and 3853 (88%) were found in four of five (**Figure 3*B***). Median sequence coverage was 18.4% and 23.7% for the individual and combined experiments, respectively (**Figure 3*C***), with a median of seven identified peptide sequences per protein. Yeast contains ~800 dubious

ORFs. These ORFs, which are believed not to encode a corresponding protein, are commonly used to verify the FDR of proteomic data sets. For the one hour experiments, between three and eight dubious ORFs were identified per dataset (**Table 1**), confirming these data are indeed well below the 1% FDR threshold.

To directly contrast the performance of the Q-OT-qIT hybrid to the most recent comprehensive yeast analysis we analyzed the same samples using a 240 min gradient. This longer method mimics the 2012 study of Mann and colleagues (*vide infra*). With the extended gradient conditions the Q-OT-qIT system identified 46,381 unique peptide sequences corresponding to 4392 protein groups (1% FDR), providing a median sequence coverage of 24.1% (average sequence coverage = 28.0%).

To estimate the dynamic range of the one hour experiments we compared our mass spectrometry-based identifications to those mapped by either tandem affinity (TAP) or green fluorescent protein (GFP) tagging experiments. From our one hour experiment data we identify 89% of proteins for which there is abundance data, including 73% of proteins present at less than 125 copies/cell (**Figure 3*D***).[9] We also note detection of 886 proteins lacking abundance data. Next, we benchmarked the dynamic range of our global analysis to a recent multiple reaction monitoring (MRM) study. There, Aebersold *et al.* targeted 152 yeast proteins, spanning the full concentration range including several proteins never observed in public proteomic data sets, using synthetic peptides and triple quadrupole MRM technology.[39] A single one hour Q-OT-qIT experiment, batched analysis of all five 1-h experiments, and our four hour analysis netted 122, 133, and 132 of the 152 Aebersold targets, respectively. The Aebersold work detected 137 of these protein targets likely following dozens of MRM experiments. We conclude that our one hour method, using the Q-OT-qIT hybrid, provides dynamic range and sensitivity comparable to state-of-the-art MRM studies, but with whole proteome depth. We also note that the system, with its quadrupole mass filter, offers considerable promise for parallel reaction monitoring.[40,41]

Figure 2. **Overview of Q-OT-qIT scan cycle.** At a retention time of 57.88 min scan #59,211, an MS^1, was acquired and presented several spectral features for MS^2 analysis. Triangles indicate the 22 precursors that were selected for subsequent MS^2 sampling—all of which were acquired within 1 s of scan #59,211. 19 of these 22 MS^2 spectra were subsequently mapped to sequence.

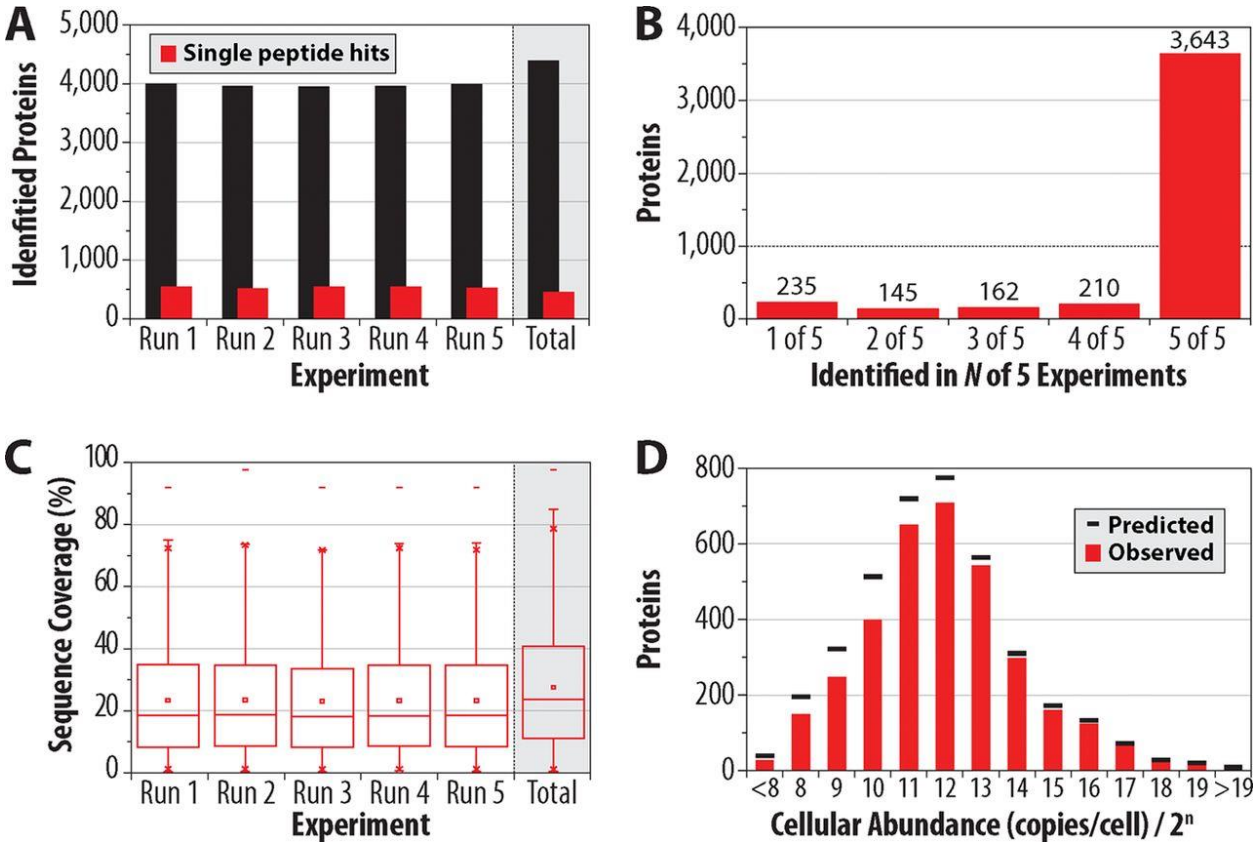


Scan #	RT	Identified Sequence	MS/MS Spectra
59,212	57.883	ALINNVSGYVIK	
59,213	57.885	YVVSFDELEAR	
59,214	57.885	IMINCFNECIDYAK	
59,215	57.886	SPIILQTSNGGAAYFAGK	
59,216	57.887	IVNEPTAAAIAYGLDK	
59,217	57.888	FELTGIPPAPR	
59,218	57.889	*n/a*	
59,219	57.889	GAFGTVVQAYLTQDKK	
59,220	57.890	ELQGELNFIEER	
59,221	57.891	LPLVGGHEGAGVVVGMEGENVK	
59,222	57.891	GIVIGSGAMSPLECR	
59,223	57.892	SPTGIVLMNMGGPSK	
59,224	57.893	*n/a*	
59,225	57.894	FQLEVTTPTAAAVGK	
59,226	57.894	AEVSSMPTLIFYK	
59,227	57.895	DAQVMIDLVNDK	
59,228	57.896	LNADFQKPWFATVNGQAR	
59,229	57.896	GAFGTVVQAYLTQDKK	
59,230	57.897	FDEKPDYLFLAR	
59,231	57.898	FLNNDLSAQFVK	
59,232	57.898	*n/a*	
59,233	57.899	SPQGLPYPIQIDK	

Table 1. Summary of the identification results for the quintuplicate one hour yeast proteome experiments using the Q-OT-qIT mass spectrometer. Note SGD stems from the Saccharomyces Genome Database (www.yeastgenome.org).

Experiment	PSMs	Peptides	Proteins	SGD verified	SGD un-characterized	SGD dubious
1	43,423	34,535	4,002	3,630	337	8
2	43,622	34,495	3,966	3,608	331	7
3	42,339	33,450	3,959	3,595	334	3
4	43,326	34,347	3,968	3,602	337	8
5	43,343	34,449	3,991	3,623	341	4
Total	**216,256**	**47,624**	**4,395**	**3,976**	**381**	**16**

Figure 3. **Performance metrics for one hour analysis, performed in quintuplicate, of a yeast trypic digest using the Q-OT-qIT hybrid.** On average, 3977 yeast proteins were identified in each experiment (1% FDR) with only 13.5% (538, $\bar{x}$) originating from single peptide identifications (***A***). 4395 proteins were detected across all experiments—3643 of which were present in all five one hour experiments (***B***). ***C***, presents the median sequence coverage for the individual and combined experiments. ***D***, displays the overlap in our identified proteins *versus* known expression level information derived from published tagging experiments.



DISCUSSION

The data presented above provides a deep view of the yeast proteome. Equally important is that this depth is achieved within an unprecedented time-scale. To understand how these gains were realized we plotted the number of identified peptide sequences per second as a function of elution time (**Figure 4*A***) for both the Q-OT-qIT (red) and the previous generation qIT-OT (black) hybrids. Remarkably, this plot reveals that the Q-OT-qIT hybrid (Orbitrap Fusion) routinely identifies ~8 peptides per second with occasions where up to 19 peptide sequences are detected in a single 1 s window. These stunning metrics are approximately double that achieved by the qIT-OT (Orbitrap Elite). Further, the speed of the Q-OT-qIT allows for deeper MS^2 sampling of the MS^1. The mean precursor depth sampled by the Q-OT-qIT is the 349^{th} most abundant *m*/*z* peak in the MS^1 scan whereas the qIT-OT system achieves an average depth of only 202 (**Figure 4*B***). The Q-OT-qIT frequently samples *m*/*z* peaks that are the 800th or weaker in intensity rank.

To further examine the effect of Q-OT-qIT scan speed we plotted the cumulative number of unique peptide sequence identifications as a function of retention time (**Figure 4*C***, I/L ambiguity removed). Almost linearly across the 70 min gradient and wash period the Q-OT-qIT (solid red) accumulates unique peptide identifications at a rate considerably faster than the qIT-OT (solid black)—8.3 *versus* 4.3 unique sequences per second, respectively (linear fit between 10 and 80 min). Use of a four hour gradient on this same system slightly reduces the slope (3.7 unique sequences per second, linear fit between 10 and 200 min, dotted red), but allows for considerably deeper analysis and again outpaces the qIT-OT system (2.3 unique sequences per second, dotted black). For reference we plotted these same data for the 2012 Mann *et al*. study using the Q-OT hybrid (Q-Exactive, dotted blue). That system posts the shallowest slope (1.7 unique sequences per second) and approximately half the number of unique peptide sequences as compared with the Q-OT-qIT (Orbitrap Fusion) for the same analysis time. Note that Mann used a Lys-C digest in this work. Panel D of **Figure 4** compares unique protein identification rates for the same data sets. Again the Q-OT-qIT system is the top performer, even when comparing the one hour analysis (solid red) to four hour experiments on either of the other systems (dotted black and blue).

Here we described new mass spectrometer technology that is capable of achieving comprehensive yeast proteome coverage within an unprecedented time-scale. Doubtless over the past decade many improvements to sample preparation, chromatography, and MS hardware have contributed to making this achievement possible. Among all these, we attribute increased mass spectrometer scan speed as the primary reason for the acceleration in proteome analysis speed and depth. **Figure 5** illustrates the pace of protein identifications for several large-scale yeast proteomic analyses as a function of the mass spectrometer MS^2 scan rate. Note the rapid ascent in protein identification rates scales correlates with increasing MS^2 scan rate.

The correlation depicted in **Figure 5** was somewhat surprising to us as we expected that ionization suppression of lower abundance peptides would become increasingly dominant as complex peptide mixtures comprising whole proteomes are separated over shorter gradients - *i.e.* from four to one hour.[42,43] In other words, as the separation duration of the online chromatography is compressed, increased co-elution must occur. With increased co-elution one might expect that, regardless of the MS speed or sensitivity, ionization suppression would prevent a considerable fraction of peptides from becoming gas-phase ions - a requisite for MS detection. The results shown here refute this hypothesis and confirm that further improvements in MS sensitivity and speed will continue to reduce whole proteome analysis time, most likely to less than one hour for relatively simple proteomes like yeast.

Finally, we conclude that comprehensive analysis of mammalian proteomes within several hours is now within our technical reach. Consider that recent estimates suggest between 10,000 and 12,000 proteins are expressed at any given time for human cells in culture.[44-46] That is only approximately three to four times the complexity of the yeast proteome. Thus, our current efforts are aimed at achieving comprehensive coverage of mammalian system within just a few hours of analysis. Looking forward, one more doubling of MS^2 acquisition rate, *i.e.* from 20 to 40 Hz, has potential to deliver detection of the whole human proteome in just one to two hours. And, given the history and rate of MS innovation, such capability is likely only a few years away.

Figure 4. Analytical metrics of yeast proteome analysis using the Q-OT-qIT (Fusion) as compared with qIT-OT (Orbitrap Elite) and Q-OT (Q-Exactive) hybrids. The Q-OT-qIT (**panel *A***, red) achieves identification of up to 19 peptides per second as compared with 10 with the qIT-OT system (***A***, black). Peak depth is likewise considerably higher on account of the faster MS^2 scanning rate of the Q-OT-qIT system (***B***). ***C***, plots the pace of unique yeast peptide identifications for the three instruments. For the one hour analysis, the Q-OT-qIT posts almost twice as many unique peptide identifications as compared with the qIT-OT. Similar data, except for unique proteins, is shown in ***D***.

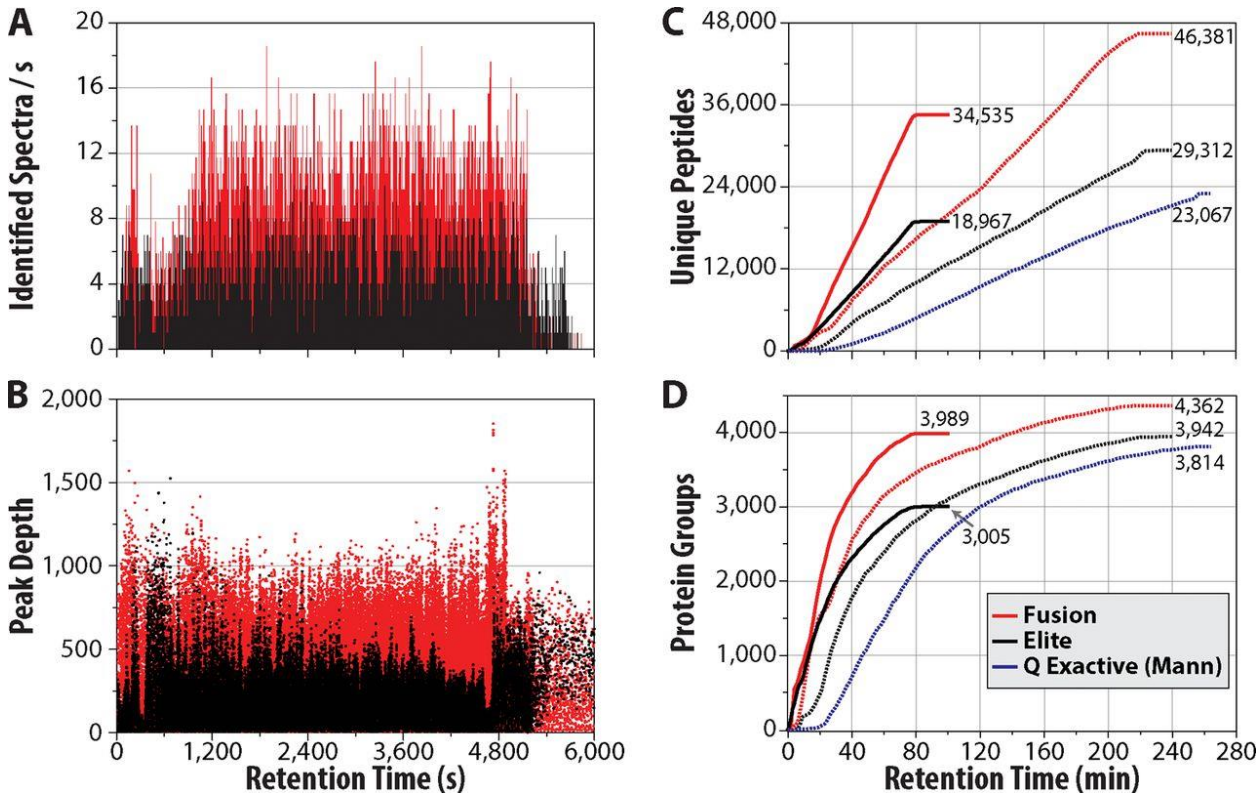
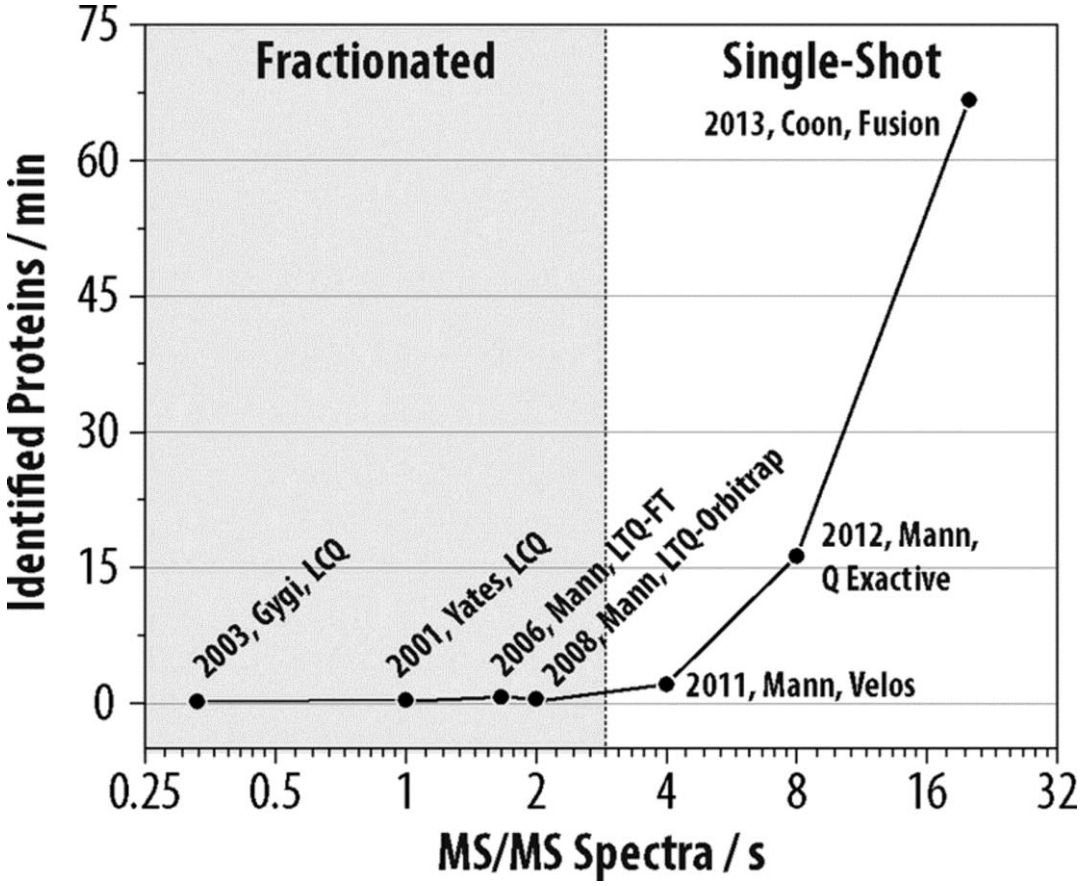


Figure 5. Rate of protein identifications as a function of mass spectrometer scan rate for selected large-scale yeast proteome analyses over the past decade. Each data point is annotated with the year, corresponding author, and type of MS system used.



EXPERIMENTAL PROCEDURES

Yeast Culture and Lysis. *Saccharomyces cerevisiae* strain BY4741 was grown in yeast extract peptone dextrose media (1% yeast extract, 2% peptone, 2% dextrose). Four liters of media was divided between four two-liter flasks and inoculated with a starter culture (OD_{600} = 2.58). Cells were allowed to propagate for ~12 generations (20 h) to an OD_{600} ~2 (average of 2.18). The cells were harvested by centrifugation at 5000 rpm for 5 min, supernatant decanted, resuspended in chilled NanoPure water and all pellets were pooled together. The cells were washed two more times and centrifuged for the final pelleting at 5000 rpm for 10 min. A pellet corresponding to 5% of the total cells grown, was resuspended in lysis buffer composed of 50 mMTris pH8, 8 M urea, 75 mM sodium chloride, 100 mM sodium butyrate, protease (Roche) and phosphatase inhibitor tablet (Roche). Yeast cells were lysed by glass bead milling (Retsch). Briefly, 2 ml of acid washed glass beads were combined with 2.5 ml of resuspended yeast cells in a stainless steel container and shaken 8 times at 30 hz for 4 min with a 1 min rest in between.

Protein Digestion. Lysate protein concentration was measured by BCA (Thermo Pierce). Protein was reduced by addition of 5 mM dithiothreitol and incubated for 45 min at 55 °C. The mixture was cooled to room temperature, followed by alkylation of free thiols by addition of 15 mM iodoacetamide in the dark for 30 min. The alkylation reaction was quenched with 5 mM dithiothreitol. Urea concentration was diluted to 1.5 Mwith 50 mM Tris pH 8.0. Proteolytic digestion was performed by addition of Trypsin (Promega, Madison, WI), 1:50 enzyme to protein ratio, and incubated at ambient temperature overnight. An additional 1:50 bolus of trypsin was added in the morning and incubated at ambient temperature for 1 h. The digestion was quenched by addition of TFA and desalted over a tC18 Sep-Pak (Waters, Milford, MA).

nLC-MS^2 Analysis. Reversed phase columns were prepared in-house. Briefly, a 75–360 μm inner-outer diameter bare-fused silica capillary, with a laser pulled electrospray tip, was packed with 1.7 μm diameter, 130 Å pore size, Bridged Ethylene Hybrid C18 particles (Waters) to a final length of 35 cm. The column was installed on a nanoAcquity UPLC (Waters) using a stainless steel ultra-high pressure union formatted

for 360 μm outer diameter columns (IDEX) and heated to 60 °C for all runs. Mobile phase buffer A was composed of water, 0.2% formic acid, and 5% DMSO. Mobile phase B was composed of acetonitrile, 0.2% formic acid, and 5% DMSO. Samples were loaded onto the column for 12 min at 0.35 μl/min. Mobile phase B increases to 4% in the first 0.1 min then to 12% B at 32 min, 22% B at 60 min, and 30% B at 70 min, followed by a 5 min wash at 70% B and a 20 min re-equilibration at 0%B.

Eluting peptide cations were converted to gas-phase ions by electrospray ionization and analyzed on a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo). Survey scans of peptide precursors from 300 to 1500 *m*/*z* were performed at 60K resolution (at 200 *m*/z) with a 5×10^5 ion count target. Tandem MS was performed by isolation at 0.7 Th with the quadrupole, HCD fragmentation with normalized collision energy of 30, and rapid scan MS analysis in the ion trap. The MS^2 ion count target was set to 10^4 and the max injection time was 35 ms. Only those precursors with charge state 2–6 were sampled for MS^2. The dynamic exclusion duration was set to 45 s with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. The instrument was run in top speed mode with 5 s cycles, meaning the instrument would continuously perform MS^2 events until the list of nonexcluded precursors diminishes to zero or 5 s, whichever is shorter. Elite runs were performed with Survey scans of peptide precursors from 300 to 1500 *m*/*z* 60K resolution (at 200*m*/*z*) with a 1×10^6 ion count target. Tandem MS was performed by isolation at 1.8 Th with the ion-trap, CAD fragmentation with normalized collision energy of 35, and rapid scan MS analysis in the ion trap. The data dependent top 20 precursors were selected for MS^2. MS^2 ion count target was set to 5×10^3 and the max injection time was 125 ms. Only those precursors with charge state +2 or higher were sampled for MS^2. The dynamic exclusion duration was set to 40 s with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on.

Data Analysis. The raw data was processed using Proteome Discoverer (version 1.4.0.288, Thermo Fischer Scientific). MS^2 spectra were searched with SEQUEST engine against a database of 6632 yeast open reading frames (ORFs)[1](www.yeastgenome.com, February 3, 2011).[47] Peptides were generated from

a tryptic digestion with up to two missed cleavages, carbamidomethylation of cysteines as fixed modifications, and oxidation of methionines and protein N-terminal acetylation as variable modifications. Precursor mass tolerance was 20 ppm and product ions were searched at 0.35 Da tolerances. Peptide spectral matches (PSM) were validated using percolator based on q-values at a 1% FDR.[48] With proteome Discoverer, peptide identifications were grouped into proteins according to the law of parsimony and filtered to 1% FDR.[49] For cumulative protein group identification, PSMs passing the FDR were exported to a text file and processed by a modified version of Protein Hoarder (version 2.4.1).[50] The PSMs were iteratively processed in successive 1 min windows and grouped into proteins using the law of parsimony at a 1% FDR.

ACKNOWLEDGEMENTS

We thank Graeme McAlister, Steve Gygi, Jae Schwartz, John Syka, Jens Griep-Raming, Vlad Zabrouskov, Mike Senko, and Jesse Canterbury for helpful discussions. We are grateful to Anna Merrill for yeast production and for critical reading of the manuscript. We thank Audrey Gasch for assistance with yeast growth. This work was supported by the National Institutes of Health (R01 GM080148) and the National Science Foundation (0701846). A.L.R. gratefully acknowledges support from a National Institutes of Health-funded Genomic Sciences Training Program (5T32HG002760).

REFERENCES

(1) Walther, T. C.; Mann, M. *J Cell Biol* **2010**, *190*, 491-500.

(2) Mallick, P.; Kuster, B. *Nat Biotechnol* **2010**, *28*, 695-709.

(3) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. *Science* **1995**, *270*, 467-470.

(4) DeRisi, J. L.; Iyer, V. R.; Brown, P. O. *Science* **1997**, *278*, 680-686.

(5) Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R. *Mol Cell Biol* **1999**, *19*, 1720-1730.

(6) Grimsrud, P. A.; Swaney, D. L.; Wenger, C. D.; Beauchene, N. A.; Coon, J. J. *Acs Chem Biol* **2010**, *5*, 105-119.

(7) Hebert, A. S.; Dittenhafer-Reed, K. E.; Yu, W.; Bailey, D. J.; Selen, E. S.; Boersma, M. D.; Carson, J. J.; Tonelli, M.; Balloon, A. J.; Higbee, A. J.; Westphall, M. S.; Pagliarini, D. J.; Prolla, T. A.; Assadi-Porter, F.; Roy, S.; Denu, J. M.; Coon, J. J. *Mol Cell* **2013**, *49*, 186-199.

(8) Cherry, J. M.; Hong, E. L.; Amundsen, C.; Balakrishnan, R.; Binkley, G.; Chan, E. T.; Christie, K. R.; Costanzo, M. C.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hirschman, J. E.; Hitz, B. C.; Karra, K.; Krieger, C. J.; Miyasato, S. R.; Nash, R. S.; Park, J.; Skrzypek, M. S.; Simison, M.; Weng, S.; Wong, E. D. *Nucleic Acids Res* **2012**, *40*, D700-D705.

(9) Ghaemmaghami, S.; Huh, W.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737-741.

(10) de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. *Nature* **2008**, *455*, 1251-U1260.

(11) Wu, R. H.; Dephoure, N.; Haas, W.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P. *Mol Cell Proteomics* **2011**, *10*.

(12) Webb, K. J.; Xu, T.; Park, S. K.; Yates, J. R. *J Proteome Res* **2013**, *12*, 2177-2184.

(13) Mixner, R. *Libr J* **2012**, *137*, 100-100.

(14) Figeys, D.; Ducret, A.; Yates, J. R.; Aebersold, R. *Nat Biotechnol* **1996**, *14*, 1579-1583.

(15) Shevchenko, A.; Jensen, O. N.; Podtelejnikov, A. V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann, M. *P Natl Acad Sci USA* **1996**, *93*, 14440-14445.

(16) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R. *Nat Biotechnol* **1999**, *17*, 676-682.

(17) Washburn, M. P.; Wolters, D.; Yates, J. R. *Nature Biotechnology* **2001**, *19*, 242-247.

(18) Beausoleil, S. A.; Jedrychowski, M.; Schwartz, D.; Elias, J. E.; Villen, J.; Li, J. X.; Cohn, M. A.; Cantley, L. C.; Gygi, S. P. *P Natl Acad Sci USA* **2004**, *101*, 12130-12135.

(19) Zanivan, S.; Gnad, F.; Wickstrom, S. A.; Geiger, T.; Macek, B.; Cox, J.; Fassler, R.; Mann, M. *J Proteome Res* **2008**, *7*, 5314-5326.

(20) Ledvina, A. R.; Beauchene, N. A.; McAlister, G. C.; Syka, J. E. P.; Schwartz, J. C.; Griep-Raming, J.; Westphall, M. S.; Coon, J. J. *Anal Chem* **2010**, *82*, 10068-10074.

(21) Peng, J. M.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. *Journal of Proteome Research* **2003**, *2*, 43-50.

(22) de Godoy, L. M. F.; Olsen, J. V.; de Souza, G. A.; Li, G. Q.; Mortensen, P.; Mann, M. *Genome Biol* **2006**, *7*.

(23) Syka, J. E. P.; Marto, J. A.; Bai, D. L.; Horning, S.; Senko, M. W.; Schwartz, J. C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D. F. *Journal of Proteome Research* **2004**, *3*, 621-626.

(24) Swaney, D. L.; Wenger, C. D.; Coon, J. J. *Journal of Proteome Research* **2010**, *9*, 1323-1329.

(25) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. *Mol Cell Proteomics* **2012**, *11*.

(26) Andrews, G. L.; Simons, B. L.; Young, J. B.; Hawkridge, A. M.; Muddiman, D. C. *Anal Chem* **2011**, *83*, 5442-5446.

(27) Cristobal, A.; Hennrich, M. L.; Giansanti, P.; Goerdayal, S. S.; Heck, A. J. R.; Mohammed, S. *Analyst* **2012**, *137*, 3541-3548.

(28) Geromanos, S. J.; Hughes, C.; Ciavarini, S.; Vissers, J. P. C.; Langridge, J. I. *Anal Bioanal Chem* **2012**, *404*, 1127-1139.

(29) Yaul, F. M.; Chandrakasan, A. P. *Isscc Dig Tech Pap I* **2014**, *57*, 198-+.

(30) Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q. Y.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; Bui, H.; Specht, A.; Lange, O.; Denisov, E.; Makarov, A.; Horning, S.; Zabrouskov, V. *Anal Chem* **2013**, *85*, 11710-11714.

(31) Meyer, J. G.; Komives, E. A. *J Am Soc Mass Spectr* **2012**, *23*, 1390-1399.

(32) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Medard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat Methods* **2013**, *10*, 989-+.

(33) Wisniewski, J. R.; Zougman, A.; Nagaraj, N.; Mann, M. *Nat Methods* **2009**, *6*, 359-U360.

(34) Dephoure, N.; Gygi, S. P. *Sci Signal* **2012**, *5*.

(35) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. *Nature Methods* **2013**, *10*, 332-+.

(36) Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R. A. *Mol Cell Proteomics* **2013**, *12*, 3330-3338.

(37) Thakur, S. S.; Geiger, T.; Chatterjee, B.; Bandilla, P.; Frohlich, F.; Cox, J.; Mann, M. *Molecular & Cellular Proteomics* **2011**, *10*.

(38) Michalski, A.; Damoc, E.; Lange, O.; Denisov, E.; Nolting, D.; Muller, M.; Viner, R.; Schwartz, J.; Remes, P.; Belford, M.; Dunyach, J. J.; Cox, J.; Horning, S.; Mann, M.; Makarov, A. *Mol Cell Proteomics* **2012**, *11*.

(39) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. *Cell* **2009**, *138*, 795-806.

(40) Peterson, A. C.; Russell, J. D.; Bailey, D. J.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2012**, *11*, 1475-1488.

(41) Gallien, S.; Duriez, E.; Crone, C.; Kellmann, M.; Moehring, T.; Domon, B. *Mol Cell Proteomics* **2012**, *11*, 1709-1723.

(42) King, R.; Bonfiglio, R.; Fernandez-Metzler, C.; Miller-Stein, C.; Olah, T. *J Am Soc Mass Spectr* **2000**, *11*, 942-950.

(43) Annesley, T. M. *Clin Chem* **2003**, *49*, 1041-1044.

(44) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. *Mol Syst Biol* **2011**, *7*.

(45) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. *Molecular Systems Biology* **2011**, *7*.

(46) Mann, M.; Kulak, N. A.; Nagaraj, N.; Cox, J. *Mol Cell* **2013**, *49*, 583-590.

(47) Eng, J. K.; Mccormack, A. L.; Yates, J. R. *J Am Soc Mass Spectr* **1994**, *5*, 976-989.

(48) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. *J Proteome Res* **2009**, *8*, 3176-3181.

(49) Nesvizhskii, A. I.; Aebersold, R. *Mol Cell Proteomics* **2005**, *4*, 1419-1440.

(50) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. *Proteomics* **2011**, *11*, 1064-1074.

Chapter 5

One-hour Proteome Analysis in Yeast

This chapter has been published:

Richards AL*, Hebert AS*, Ulbrich A, Bailey DJ, Coughlin EE, Westphall MS, Coon JJ. *One Hour Proteome Analysis in Yeast*. Nature Protocols. **2015,** *10*, 701-714.

*Authors contributed equally

ABSTRACT

Recent advances in chromatography and mass spectrometry have made rapid and deep proteomic profiling possible. To maximize the performance of a novel Orbitrap hybrid mass spectrometer, we have developed a protocol that combines improved sample preparation, including optimized cellular lysis by extensive bead beating, and chromatographic conditions, specifically, 30 cm capillary columns packed with 1.7 µm BEH material (Waters Corporation), and the manufacture of a column heater to accommodate flow rates of 350-375 nL/min, that increase the number of proteins identified across a single LC-MS/MS separation, reducing the need for extensive sample fractionation. This strategy allowed the identification of up to 4,002 proteins (at 1% FDR) in yeast (*Saccharomyces cerevisiae* strain BY4741) over 1.3 hours of LC-MS/MS analysis. Quintuplicate analysis of technical replicates reveals 83% overlap at the protein level, demonstrating the reproducibility of this procedure. This protocol, which includes cell lysis, overnight tryptic digestion, sample analysis, and database searching, takes approximately 24 hours to complete. If yeast growth is required, this procedure takes approximately 4 days to complete. Aspects of this protocol, including chromatographic separation and instrument parameters, can be adapted for the optimal analysis of other organisms.

INTRODUCTION

Ideally, mass spectrometry (MS) techniques would permit the rapid identification of every protein in the proteome. Historically, deep, sampling of proteomes has been laborious and difficult to achieve. A major limitation has been the tandem MS (i.e., MS/MS) sampling speed. MS/MS acquisition rates have ranged from 1 to 10 Hz and often could not keep pace with the number of unique, co-eluting peptide species found in complex mixtures. The resulting under-sampling led to stochastic precursor sampling and meant that there was a lot of variability in the proteins identified from different HPLC runs of the same sample.[1-3] Identifying low abundance proteins in such experiments was difficult because these low level species were often not sampled, when the peaks were close to large, wide peaks associated with more abundant peptides.[3] A common solution is fractionation of the sample prior to liquid chromatography (LC)- tandem mass spectrometry (MS/MS) analysis.[3] While effective, this approach is time consuming for both the analyst and the instrument, limiting the throughput of proteomics.

Advances in the MS/MS acquisition rate of mass spectrometers have increased the percentage of proteomes identifiable in a single experiment, while decreasing instrument analysis time. A significant boost in acquisition speed was realized with the introduction of a novel Orbitrap hybrid mass spectrometer consisting of a mass filter, a collision cell, a high-field Orbitrap and a dual cell linear ion trap analyzer (Q-OT-qIT, Orbitrap Fusion).[4,5] Here, precursor ions are selected by the quadrupole and fragmented by higher energy collisional dissociation (HCD), collisional-activated dissociation (CAD), or electron transfer dissociation (ETD) and are m/z analyzed either by the ion trap or Orbitrap mass analyzer. Precursor selection using the quadrupole mass filter allows the ion trap and Orbitrap analyzers to operate in parallel, significantly improving acquisition speed over previous generation instruments. When product ions are analyzed in the ion trap, the Fusion enables MS/MS scan rates of over 20 Hz.[2]

As the first eukaryote with a sequenced genome,[6] *Saccharomyces cerevisiae* is commonly used as a model organism is systems biology. Several proteomics studies have identified the estimated 4,500 proteins[7] that are expressed in yeast under standard laboratory conditions.[8-12] Building on the successes of these earlier studies, we sought to improve both the depth of proteome coverage and the speed at which

proteins can be identified. Although the very fast acquisition speed of the Fusion greatly increases the number of peptides selected for MS/MS sequencing, we also sought to increase peptide identifications through improved sample preparation and chromatography, and through the optimization of instrument parameters. Specifically, improvements include lysing yeast cells by extensive bead beating, the use of a 30 cm capillary column packed with 1.7 µm BEH material (Waters Corporation) operating at flow rates of 350-375 nL/minute, and the addition of dimethyl sulfoxide (DMSO) to the mobile phase. Similar to previously published reports,[13,14] we find this additive substantially increases the number of peptides identified in each run.

This optimized sample preparation, coupled with MS analysis performed in the Orbitrap and low resolution HCD MS/MS performed by ion-trap rapid scan, allowed us to identify up to 34,535 yeast peptides over a 70-minute LC-MS/MS run. This translates to up to 4,002 proteins, or identification of ~90% of the expressed yeast proteome. Further, this system is capable of identifying 67 proteins per minute.[4] This protocol describes cellular lysis, trypsin digestion, column packing, LC gradient setup, and Orbitrap Fusion instrument parameters that lead to comprehensive yeast proteome coverage with minimal instrument time. If beginning from yeast lysis, this protocol can be completed in 24 hours. This protocol is easily adaptable to other organisms; for samples requiring less forceful lysing (for example, human cell lines), the bead beating steps may be omitted. This protocol is organized in the following manner: In the **Experimental Design** section, parameters which have afforded the best results and **Limitations** of the procedure are discussed. A list of all **Materials** required, including a detailed **Reagent Setup** section, which provides instruction for the preparation of all solvents and stock solutions, and **Equipment Setup**, which provides a detailed procedure for **fabrication of reversed phase analytical columns** and LC-MS/MS separation, is included. A **Procedure** for yeast lysis, digestion, reduction and alkylation, LC-MS/MS analysis and database searching is detailed in step-wise order, and includes an in-depth section on the **optimization of mass analyzer settings**. The **Timing** of all steps in the protocol is detailed, followed by a discussion of

Anticipated Results. A **Troubleshooting Table** addresses possible problems that may occur when following this procedure and suggested solutions.

EXPERIMENTAL DESIGN

Although the protocol may be modified based on the sample of interest or available instrumentation, we note that the following details have provided us with the best results:

Yeast lysis. A vigorous bead beating procedure for yeast cell lysis, where proteins are lysed via a ball mill, produces the optimal extracted protein mixture for downstream bottom up processing. Yeast pellets are milled at 30 Hz eight times for four minutes each, as these lysates produced more complete proteome coverage compared with either lysing at 30 Hz for fewer rounds or lysing for at a lower frequency. As others have noted,[15] we observe a further increase in the number of peptides identified when the cellular debris is not cleared prior to analysis. Preserving the cellular debris resulted in a 5% increase in protein identifications when compared with an identical analysis where precipitated lysate was discarded prior to digestion. Gene ontology analysis reveals an enrichment in membrane proteins when the cellular debris is included during digestion, resulting in more complete proteome coverage. Following digestion, samples are centrifuged and only the supernatant is processed.

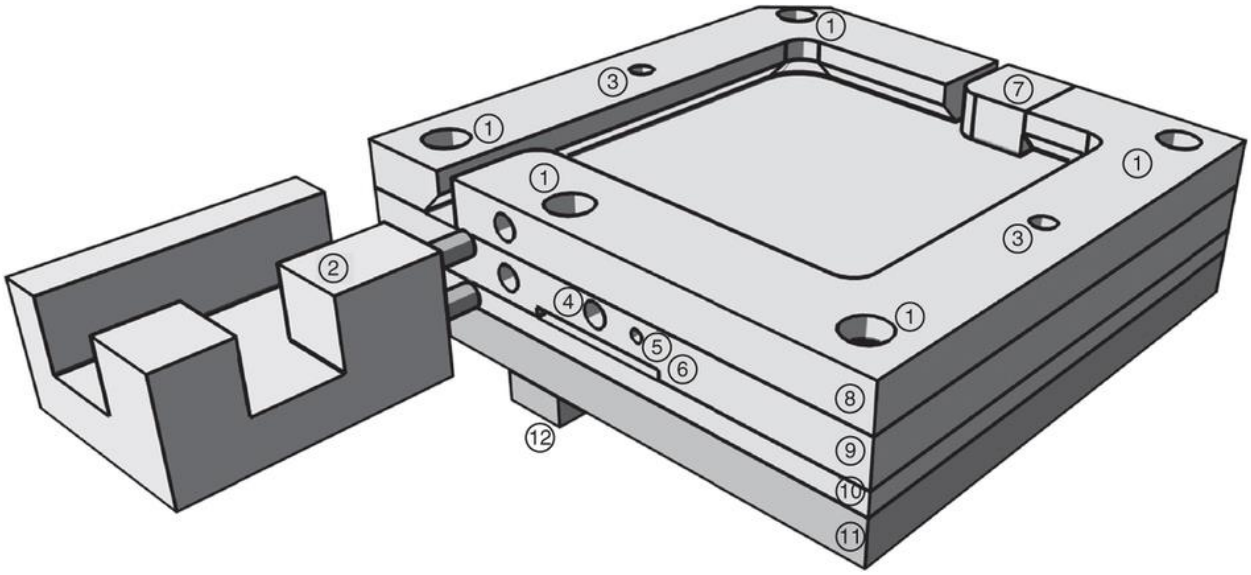
Chromatography column and column heater. Ultra-high performance nano-liquid chromatography setups employing 35 to 40 cm columns packed with 1.7 µm C18 particles produced optimal chromatography for increased peptide identifications. 1.7 µm particles improve chromatographic resolution peaks, and increase peak capacity as compared to packing materials made up of larger particles. This allows for more unique identifications. This setup requires a column heater to operate at suitable pressures, but higher temperatures provide additional chromatographic improvements. We found it economical to manufacture a column heater in house (**Figure 1**), but there are several excellent models commercially

available, such as the PRSO-V1 from Sonation GmbH or several options from Phoenix S&T, among other options. If a column heater is not available, we recommend using 3 µm C18 particles or packing a shorter column (~15 cm) with 1.7 µm particles. This is because both smaller particle size and increased column length can provide more theoretical plates, and thus, better chromatographic separation.

Commercial heaters may be more suitable for some research groups since manufacture requires basic familiarity with Computer aided design drawing software, electronics, and metal machining. In our laboratory, the grounded column oven was manufactured from aluminum and polycarbonate and relies on a resistive heater that is driven by a solid-state relay connected to a proportional-integral derivative (PID) controller receiving feedback from a resistance temperature detector. We set the PID controller to allow temperature control from 25˚C to 80˚C. Note that the design is intrinsically safe against runaway overheating in the case of a closed relay failure since the 20 W heating element is too weak to heat the oven to the glass transition temperature of polycarbonate (~145˚C). The oven fits completely within the Flex Ion Source without any changes to the source and can be controlled from the instrument computer via a USB interface. Columns 10 cm or longer fit into the 5.1 x 5.1 cm heater cavity whose edges are undercut by 3.5 mm to securely retain columns securely and the emitter tip is held in place by a spring-loaded closure.

Chromatography mobile phase. As previously[16] noted, the addition of 5% DMSO to mobile phase A and B increases the ionization efficiency of peptides, resulting in an increase in intensity of peptide precursors.[13,14] To examine this shift towards higher precursor intensities, duplicate injections of 1 µg of tryptic yeast peptides were analyzed with and without the addition of 5% DMSO in the mobile phases. A total of 24,067 peptides were common across all four separations; the addition of DMSO increased the average precursor signal from ~$1.8x10^7$ to $3.3x10^7$ (arbitrary units). This shift towards higher intensities contributes to an increase in unique peptide identifications, from an average of 32,481 to 34,246 when DMSO is included. Note that this advantage may be different for each instrument type and may also depend on other instrumental variables such as duty cycle, column width, and/or ionization source. Using the

Figure 1. Structure of in-house manufactured column heater. Labeled rendering shows the column oven without a lid. (1) Holes for 4-40 x 3/4 –inch screws to attach the four layers to one another. (2) Polycarbonate micro-tee retainer. (3) Holes to accept pins from lids (not shown). (4) Hole for resistance thermometer. (5) Hole for grounding wore. (6) Slot for foil heater connections. (7) Spring-loaded column tip retainer. (8) Undercut top layer. (9) Layer containing thermometer, heater (against bottom), and grounding wire. (10) Aluminum layer between heater and polycarbonate base plate. (11) Polycarbonate base plate. (12) Polycarbonate attachment to nanospray (NSI) source.



highest purity DMSO commercially available, we have observed no differences in the cleaning cycles or maintenance of our instruments.

Mass spectrometry. A survey scan in the Orbitrap, followed by rapid scan ion-trap HCD MS/MS affords the highest number of unique peptide identifications. The Q-OT-qIT geometry of the Orbitrap Fusion permits ion injection and HCD fragmentation, but not CAD fragmentation, to occur simultaneously with ion trap mass analysis. This translates to an approximately 10% reduction in duty cycle and thus a 6% increase in MS/MS scans for HCD compared to CAD fragmentation. On average ion-trap analyses take 30-35 ms, and setting a low max injection time of 25 to 35 ms ensures ion injection and fragmentation are optimally parallelized with the ion-trap analysis. When optimized, ion trap MS/MS reduces duty cycle by ~20% compared to OT MS/MS, while maintaining an identification rate above 50% for all MS/MS scans. Additionally, quadrupole isolation of 0.7 Th width around the target precursors works optimally for ion trap MS/MS, presumably by reducing co-isolation of off-target precursors, while maintaining adequate flux for MS/MS analysis.

Although the number of MS/MS scans taken at MS resolving power settings of 30,000, 60,000, and 120,000 (at *m/z* 200) is comparable, the most unique peptide identifications were observed when using the 60,000 or 120,000 MS^1 scans. Further, although more MS/MS events and PSMs are recorded at an MS^1 resolving power of 15,000, an increase in the resolving power to 60,000 gives an approximately 20% boost in unique peptide identifications. We reason that with increased resolving power comes a corresponding increase in the number of observable peptidic features, allowing deeper sampling of the proteome. An increase in MS resolving power above 120,000, however, yields no further improvement in identifications and the 450,000 resolving power slows down the duty cycle enough to reduce the number of identifications.

Setting the MS^1 AGC to 5e5, the highest allowable, provided the highest number of scans and unique peptide identifications.

Optimal values for some of the parameters described above will differ with instrumentation. Specifically, optimal AGC values differ between the Orbitrap Elite and the Orbitrap Fusion, and should be

adjusted accordingly. Additionally, these parameters should not be treated as independent variables; changing any of these parameters may impact how well other parameters work. For example, if the MS/MS max inject time was set to 100 ms, then one would likely observe little difference between Orbitrap and ion trap MS/MS performance.

LIMITATIONS

Results may vary greatly depending on type, generation, and operational state of the mass spectrometer used as well as that of the nano-LC and chromatography setup. Any added multiplexing labels, for isobaric tags including TMT and iTRAQ, will have a negative impact on the observed proteomic coverage, as this will require Orbitrap analysis.

The results are likely to change significantly depending on the protease used. For example, when using LysC as a protease, we observe an ~ 30% decrease in unique peptide identifications. Similar or greater decreases in identifications are expected when using other proteases, including GluC, chymotrypsin and AspN. Proteomic coverage will vary with the amount of peptide sample being analyzed. Typically, injecting much less than 1 µg of sample reduces sampling depth. Like-wise, injecting much more than ~1.5 µg can reduce column lifetime.

Organisms of higher complexity, containing greater numbers of protein-encoding genes than yeast (for example, mammalian or plant species), will have an increased dynamic range of protein abundance. With such organisms, it may be unrealistic to achieve full proteome coverage in a single run, necessitating off-line fractionation. However, following the protocols outlined in this manuscript will help minimize the number of fractions and instrument time required.

MATERIALS

Reagents

• Acetonitrile HPLC grade (ACN) (Fisher Scientific, cat. no. A955-4)

• Agar (Sigma, cat. no. 9002-18-0)

• Ammonium formate (Sigma, cat. no. 516961)

• Bicinchoninic acid (BCA) protein assay kit (Pierce, cat. no. 23227)

• Calcium chloride (Sigma, cat. no. 223506-500g)

• Complete protease inhibitors cocktail tablets mini, EDTA-free (Roche, cat. no. 11 846 170 001)

• Dextrose (Fisher Scientific, cat. no. D16500)

• Dimethylsulfoxide (DMSO) ≥99.9% (Sigma Aldrich, cat. no. 472301)

• Dithiothreitol (DTT) (Sigma, cat. no. 43819)

• Formic acid (FA) (Thermo Scientific, cat. no. 28905)

• Glycerol (Sigma, cat. no. G5516)

• Iodoacetamide (IAA) (Acros Organic, cat. no. 122270250)

• Hydrofluoric acid (HF), 48% (Sigma, cat no. 7664-39-3)

• Methanol HPLC grade (Fisher Scientific, cat. no. A454SK-4)

• Nicotinamide (Sigma, cat no. 98-92-0)

• Peptone (Fisher Scientific, cat. no. DF0118-17-0)

• PhosStop phosphatase inhibitors cocktail tablets (Roche, cat. no. 4906837001)

• RP-HPLC solvent A = 0.2% FA, 5% DMSO in HPLC grade water (vol/vol)

• RP-HPLC solvent B = 0.2% FA, 5% DMSO in ACN (vol/vol)

• Sep-pak C18 cartridge, 50 mg sorbent (Waters, cat. No. WAT054955)

• Sequencing grade modified trypsin (Promega, cat. no. V5111 or V5113)

• Sodium butyrate (Sigma, cat no. 156-54-7)

• Sodium chloride (Sigma, cat. no. S3014-500g)

• Trifluoroacetic acid (TFA) (Thermo Scientific, cat. no. 28904)

• Tris base (EP154-1, cat. no. EP154-1)

• Urea (Sigma, cat. no. U5378-1kg)

• Water HPLC grade (Sigma, cat. no. 270733-4 L)

• Water, nanopure

• Yeast extract (granulated) (Fisher Scientific, cat. no. BP9727)

Equipment

• Autoclave

• Autodesk Inventor 2014

• Beaker, recommended 1 L

• Bunsen burner

• Centrifuge, refrigerated, capable of centrifugation at 4,000xg (Thermo Scientific, cat. no. 75004380)

• Centrifuge buckets, 50 mL conical (Thermo Scientific, cat. no. 75-003-683)

• Ceramic scoring wafer (Restek, cat. no. 20116)

• Compressed helium tank

• Cuvettes

• Filter units, 250 mL capacity (Millipore, cat. no. SCGP-U02-RE)

• Flask, recommended 4 L

• Glass beads 425-600 um (Sigma, cat. no. G8772-500G)

• Graduated cylinder, recommended 4 L

• Heatblock (VWR)

• High pressure micro tee union for 360 µm OD tubing (IDEX Health and Science, cat. no. UH-750)

• HPLC system capable of nanoliter per minute flow rates (nanoAcquity UPLC; Waters Corporation)

• Incubation shaker

• Incubator

• Laser-based micropipette puller (Sutter Instrument Co., cat. no. P-2000/F)

• LTQ Orbitrap Fusion mass spectrometer (Thermo Fisher, cat. no. IQLAAEGAAPFADBMBCX)

• Magnetic stir plate (IKA, cat. no. 0003907500)

• Magnetic stir bar, miniature (VWR, cat. no. 58948-400)

• Magnetic stir bar (VWR, cat. no. 58947-114)

• Microscope (Zeiss, cat. no. 495005-0004-000)

• MM4000 Mixer Mill (Retsch, cat. no. 20.745.0001)

• Petri dishes, sterile, 9 cm diameter (Sigma, cat. no. SIAL506CC0SnV)

• pH paper strips, pH range 0-2.5 (EMD Millipore Corporation, cat. no. 1.09540.0001)

• pH paper strips, pH range 6.5-10 (EMD Millipore Corporation, cat. no. 1.09543.0001)

• Pressure bomb, connected to helium tank (NextAdvance, cat. no. PC77)

• Screw top grinding jar (Retsch, cat. no. 01.462.0231)

• Serological pipettes, 1 mL (Fisher Scientific, cat. no. 13-678-11B)

• Serological pipettes, 5 mL (Fisher Scientific, cat. no. 13-678-11D)

• Spectrophotometer

• SpeedVac (Savant Refrigerated Vapor Trap; Thermo Scientific, cat. no. RVT5105-115)

• Test tubes (Sigma, cat. no. CL56982516X)

• Test tube rack

• Tube rocker (VWR, cat. no. 10159-754)

• TX-400 swinging bucket centrifuge rotor (Thermo Scientific, cat. no. 75-003-181)

• Vortex (Vortex Genie 2; Scientific Industries, cat. no. SI-0236)

• 1.5 mL micro centrifuge tubes (Sarstedt, cat.no. 72.692.005)

• 2 mL micro centrifuge tubes (Sarstedt, cat. no. 72.694.005)

• 2 mL cryogenic vial (Corning, cat. no. 430662)

• 50 mL conical centrifuge tube (Thermo Scientific, cat. no. 14-432-22)

• 75 x 360 µm fused silica (Polymicro Technologies, cat no. TSP075375)

• 1.7 µm BEH C18 Packing Material (Waters Corporation, cat. no. 186002350)

• 3.5 µm BEH C18 Packing Material (Waters Corporation, cat. no. 186003034)

Column heater

- ¼” and 1/8” aluminum plate stock
- ¼” polycarbonate plate stock

- 1” polycarbonate tube stock
- 1/8” Stainless steel dowel pins (4)
- PID controller (Model 32B, Dwyer Instruments, cat. no. 32B-23)
- Adhesive backing polyimide film heater (Omega Engineering, cat. no. KH-202/5-P)
- 120V solid-state DC controlled relay (Opto22 P120D2 or equivalent, Newark, cat. no. 18M9301)
- DIN 5-pin Deltron plug (Newark, cat. no. 69K6079)
- DIN 5-pin Deltron jack (Newark, cat. no. 69K6045)
- RS-485 to USB converter (gridconnect, ATC-820, cat. no. GC-ATC-820)
- Love Link 3 control software (Dwyer Instruments, cat. no. LOVELINK III)
- Line switch
- Resistive temperature detector (Omega Engineering, cat. no. PR-10-2-100-1/8-2-E)
- 22 Ga wire, solder, machine screws (4-40)
- Power inlet receptacle, fused (Digikey, cat. no. Q303-ND)
- Fuses, 500 mA (2)
- Box (Digikey, cat. no. HM1123-ND)

REAGENT SETUP

Yeast cells. This procedure starts by extracting a yeast cell pellet. This could be obtained by following standard procedures, by using the protocol described in http://www.nature.com/doifinder/10.1038/protex.2015.030 or by purchasing a commercially available yeast protein extract (e.g., Promega, cat. no. V7341). Growing yeast cells requires dedicated equipment, including an autoclave, spectrophotometer, incubator and incubator shaker. If these are not available, we recommend using a commercially available yeast extract. As the steps described in http://www.nature.com/doifinder/10.1038/protex.2015.030 can yield large quantities of yeast protein, they will not need to be performed each time this protocol is followed. Times and expected outcomes are for *S. cerevisiae* strain BY4741 grown in yeast extract peptone dextrose agar (YPD) medium.

DTT and IAA solutions

Prepare DTT stock solution by dissolving DTT in water to a final concentration of 0.25 M. DTT is oxygen sensitive and should be prepared fresh prior to use. Prepare IAA stock solution by dissolving IAA in water to a final concentration of 0.25 M. IAA is light sensitive and should be prepared fresh prior to use.

Lysis buffer

Lysis buffer is prepared in water, and includes 8 M urea, 75 mM NaCl, 50 mM Tris (pH 8), 75 mM NaCl, one tablet of protease inhibitor cocktail (cOmplete mini, Roche) per 10 ml of lysis buffer, one tablet of PhosStop phosphatase inhibitors cocktail tablets (Roche) per 10 ml of lysis buffer, 100 mM Na butyrate, 10 mM nicotinamide. Fresh lysis buffer should be prepared immediately before use. Concentrated stocks of 1 M Na butyrate and nicotinamide can be prepared in water and stored at -80°C.

Sep-pak solvents

Wash buffer: 0.1% (vol/vol) trifluoroacetic acid in water. Elution buffer: 50% (vol/vol) acetonitrile, 0.1% (vol/vol) trifluoroacetic acid in water, followed by 75% (vol/vol) acetonitrile and 0.1% (vol/vol) trifluoroacetic acid in water.

BSA peptide mixture

A mixture of tryptic bovine serum albumin (BSA) peptides can be prepared and used to monitor elution times and column degradation. Peak shape should be monitored for instances of tailing and fronting. An ideal peak shape should be Gaussian, with a retention time < 30 s (for a 60-min LC gradient). Stock solutions of BSA can be digested, aliquoted and stored at -80°C. Prepare a 1 µg/µL solution of BSA in 50 mM tris and vortex for 1 minute. Add 250 mM stock solution of DTT to a final concentration of 5 mM DTT. Incubate the sample for 45 minutes at 57°C to reduce disulfide bonds. Incubate with 15 mM IAA for 30 minutes in the dark at ambient temperature. Incubate the solution with 5 mM DTT for 15 minutes at ambient temperature to quench any remaining IAA. Add trypsin at a 1:50 enzyme: protein ratio and incubate

for ~16 hours at room temperature. To quench the digestion, add a minimum amount of 10% (vol/vol) TFA to reduce the sample to pH <2. Dry down the tryptic BSA peptides in the Speed-Vac. For MS/MS analysis, the tryptic BSA solution is resuspended in 0.2% (vol/vol) FA, with 100-300 fmol injected on column.

LC-MS/MS (Liquid chromatography coupled to tandem mass spectrometry) solvents

Mobile phase A: 0.2% (vol/vol) formic acid, and 5% (vol/vol) dimethyl sulfoxide in HPLC grade water. Mobile phase B: 0.2% (vol/vol) formic acid, and 5% (vol/vol) dimethyl sulfoxide in acetonitrile.

EQUIPMENT SETUP

Column fabrication. The set-up of a pressure bomb, attached to a tank of compressed helium, is presented in **Figure 2*C***. A commercially available pressure bomb is recommended in the Equipment section.

1. Analytical columns with integrated emitter tips are manufactured using 360 µm outer diameter (OD) x 75 µm inner diameter (ID) fused silica capillary tubing. To form the emitter tip, use a butane lighter to remove approximately 2.54 cm of polyimide coating 5 cm from one end of the fused silica.
2. Clean this area with methanol to remove any remaining charred polyimide coating before inserting the silica into the puller (**Figure 2*A***).
3. Using the manufacturer's recommended settings for the laser puller, pull a 15 µm tip, avoiding the polyimide coating. A recommended program for the Sutter Instrument Company P-2000 laser-based micropipette puller is described in **Supplemental Table 1.** However, as the optimum settings for individual laser-based pullers may differ, this should be considered a guideline.
4. Inspect the tip under a microscope; if the tip is closed, etch with HF to create an opening. In the fume hood, transfer 50 µL HF to a microfuge tube and fill another tube with 0.5 M ammonium formate. Dip the electrospray tip in HF for 1-2 minutes. Quench any remaining HF present by immersing the tip in ammonium formate for 30 seconds. Thoroughly rinse the tip with water. Once

the HF has been removed from the tip, Use a pressure bomb placed on a magnetic stir plate to rinse the analytical column with several column volumes of methanol (**Figure 2*C***).

5. Place a 1.5 mL conical tube containing 1 mL of methanol into the pressure bomb
6. Insert analytical column, with pulled tip pointing upwards, through the high pressure bomb. Be careful when handling the column, as the tip is fragile and easily damaged. Slowly adjust the position of the column until it touches the bottom of the tube, then raise it so that the bottom of the column so it is 1 to 2 mm above the bottom of the tube.
7. Set the pressure regulator on the helium tank to 1,000 psi.
8. Increase the pressure in the pressure bomb by slowly turning the valve to the open position.
9. Rinse the column with several column volumes of methanol (30 s-1 min).
10. Release the pressure by turning the valve to the closed position.
11. Prepare a slurry of 3.5 µm BEH packing material by transferring the packing material to a glass vial containing a mini magnetic stir bar. Add acetonitrile to the vial. As the 3.5 µm BEH packing material is added only to keep the smaller 1.7 µm material in the column, a very dilute slurry is recommended (0.1 mg 3.5 µm packing material, 1 mL acetonitrile).
12. Place the glass vial containing the 3.5 µm BEH slurry into the pressure bomb (**Figure 2*D*)**.
13. Insert analytical column, with pulled tip pointing upwards, through the ferrule and in to the high pressure bomb, so that it is almost touching the bottom of the vial. Turn on the magnetic stir plate.
14. Increase the pressure in the pressure bomb by slowly turning the valve to the open position.
15. Fill ~5 mm of the tip with 3.5 µm BEH slurry. If desired, a lower pressure of helium (200-300 psi) can be used for this step. Packing the tip can be accomplished by opening the valve for 1-2 seconds, then slowly releasing the pressure. Run methanol through the column to push all packing material towards the top. If necessary, pack with more 3.5 µm BEH material. A light source (lamp, flashlight) placed beside the column can be used to visually monitor its packing.
16. Prepare a slurry of 1.7 µm BEH packing material by transferring the packing material (~0.5 mg) to a glass vial containing a mini magnetic stir bar. Add 1 mL of acetone to the vial.

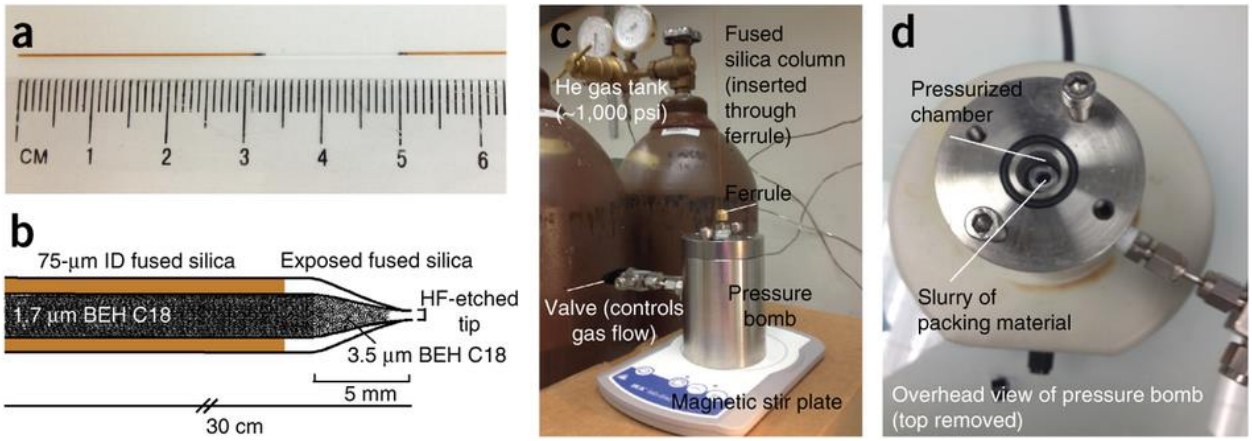
17. Fill the analytical column with 1.7 µm BEH packing material to a length of ~35 cm.
18. Release the pressure by turning the valve to the closed position slowly to avoid column unpacking. Use a light source to check for any gaps in the packing material. If gaps are present, run methanol through the column to push packing material upwards.
19. Dry the column for 10-15 minutes on the pressure bomb with air.

Reversed phase analytical columns.

These are prepared in-house using the protocol listed under **Column fabrication**. In our setup, a 25 µm i.d. capillary is plumbed directly to the injection port of a nanoAcquity UPLC (Waters) and to an in house prepared reversed phase column through an ultra-high pressure stainless steel union formatted for 360 µm o.d. capillaries (IDEX). When attaching an analytical column, set the LC flow rate to 0 µL/min. Care should be taken to tighten the union enough to prevent leaks, but not so tight as to crush the fused silica. When the column is attached, switch the LC flow rate to 100 µL/min at 100% solvent A; at this flow rate, pressure should be between 4,000-6,000 psi. Continue flowing solvent A for 10-15 minutes. Ramp the flow rate down to 0 µL/min. Remove the analytical column and inspect the end under a light source; it is common for packing material to shift upwards at higher pressures. If necessary, trim the end of the column so that the entire length of the column is filled with packing material. Reconnect the column, and equilibrate it by flowing 100% A for twenty minutes at a flow rate of 100 µL/min. The 25 µm capillary should be kept as short as possible while still allowing the column to reach the MS inlet (<50 cm). Note, use of an excessively long 25 µm capillary or a wider i.d. capillary will cause peptides to come out later because the delivery of the gradient will be delayed. To maintain a suitable backpressure throughout the analysis (<10,000 psi), the analytical column is heated to 60°C throughout the run.

***Sample injection and separation method*.** For sample injection, first the pre-programmed volume of sample is placed in the sample loop. Second, trapping is performed at a flow rate of 0.325-0.375 µL/min minute in 100% solvent A for 13 minutes. The total trapping volume should be at least the sample injection volume

Figure 2. Column fabrication. (***A***). To pull an emitter tip, first cut an appropriate length (35-40 cm) of 360 µm OD x 75 µm ID fused silica. Remove approximately 2 cm of polyimide coating 3-4 cm from one end of the fused silica. (***B***) The column is packed with 1.7 µm BEH C18 packing material according to the protocol. (***C, D***) Setup of the pressure bomb.



plus 1 µL. During trapping, the flow path includes the sample loop, 25 µm capillary, and column. To avoid sample loss, analytes are loaded directly on to the analytical column during trapping (e.g. no pre-column is used). Note, one must remove all salts prior to injection when the pre-column is omitted. During the separation, the percentage of solvent B is adjusted as described below and 2 kV is applied directly to the stainless steel union. Following the gradient the column is washed with 70% B for 5 minutes and then equilibrated in 100% A for 15 min.

With this setup and a 75 µm i.d. bare-fused silica capillary column packed to 35 cm with 1.7 µm BEH packing material, a pressure of 7,000 to 9,000 psi is expected for 100% solvent A at a flow 350 µL/min. The pressure will depend on a number of factors. Decreasing the diameter of the electrospray tip orifice, the diameter of the packing material, or the column inner diameter will cause increased back pressure, as will increasing the column length.

The following RP chromatography gradient was used in our laboratory. We chose a step-wise gradient over a linear gradient because it resulted in more peptide identifications, presumably by providing better separation of the large group of hydrophilic peptides that elute early in the gradient.

Time interval (minutes)	**Gradient (%B)**	**Flow rate (µL/minute)**
0	0	350
0.1	4	350
32	12	350
60	22	350
70	30	350
75	70	350
80	70	350
100	0	350

PROCEDURE

1 | Vigorously resuspend the yeast pellet with chilled lysis buffer. The pellet can first be pipetted up and down, followed by vortexing for ~ 1 minute to ensure even resuspension. A target protein concentration of ~2 mg/mL is recommended.

2 | Add 2 mL of glass beads to each mixer mill jar. Glass beads can be measured in a 2 mL centrifuge tube, then transferred to the mixer mill jar.

3 | Add 2 mL of yeast lysate to each mixer mill jar.

4 | Fasten jar into mixer mill. Mill for 4 minutes at 30 Hz, followed by a 1 minute rest. Repeat this seven additional times.

5 | Transfer lysate from the mixer mill jar to a 2 mL centrifuge tube.

Yeast lysate should be kept at 4°C; it can be stored for up to one week.

6 | Following the manufacturer's protocol, perform a BCA assay to determine protein concentration.

Protein reduction and alkylation

7 | Prepare a 250 mM stock of DTT. Add the appropriate volume of 250 mM DTT to the yeast lysate for a final concentration of 5 mM DTT. Incubate the sample for 45 minutes at 57°C to reduce disulfide bonds. Do not perform the reduction step above 60°C or for much longer than 45 minutes, as urea can degrade into isocyanic acid, which reacts with the primary amines of the protein analytes. To test the extent of this reaction at 57°C, we searched our data with carbamylation of lysine and n-term peptides set as variable modifications. 1.7% of unique peptides were carbamylated, a very small portion of the dataset.

8 | Allow the yeast lysate to cool to room temperature. To alkylate cysteines, incubate the lysate with 15 mM IAA for 45 min in the dark at room temperature. IAA is light sensitive. IAA stock solution must be prepared fresh and kept in the dark, and the alkylation reaction must be performed in the dark.

9 | Incubate the lysate with 5 mM DTT for 15 min at ambient temperature to quench any remaining IAA.

Trypsin digestion

10 | Dilute the lysate with 50 mM tris (pH 8) until the urea concentration is 1.5 M. Before adding trypsin, ensure the pH of the lysate in 1.5 M urea is ~8. To determine the pH, pipet 0.2 µL of lysate on to pH paper; match the color on the paper with the provided chart.

11 | Add trypsin at a 1:50 enzyme to protein mass ratio and incubate for ~16 hours at ambient temperature with gentle rocking.

12 | After 16 hours, add a second bolus of trypsin at a 1:50 enzyme to protein ratio and incubate on a rocker for 1-2 hours.

Desalting (Sep-pak)

13 | Desalt the peptides prior to analysis to remove urea and salts. Select a size of the Sep-pak cartridge that corresponds with the amount of starting protein material. Use ~20 times more bulk material than protein sample. For example, for 4 mg of protein, use a 100 mg Sep-pak. The below steps are for use of either a 50 mg or 100 mg Sep-pak. For 500 mg Sep-paks, increase all volumes listed by five.

14 | Add a minimum amount of 10% (vol/vol) TFA to reduce the sample to pH <2. Confirm approximate pH with pH paper. Pipet ~ 0.2 µL of the peptide solution on to pH paper; adjust the pH of the peptide solution with 10% (vol/vol) TFA until the pH paper reading matches the desired pH. Centrifuge the sample for 1 minute at 8,000*g* to pellet any insoluble material.

15 | Wash and condition the Sep-pak cartridge by adding 3 mL of 100% (vol/vol) ACN, followed by 1 ml of 75% (vol/vol) ACN/0.1% TFA, and 1 mL of 50% (vol/vol) ACN/0.1% TFA.

16 | Equilibrate the cartridge with 3 mL of 0.1% (vol/vol) TFA.

17 | Slowly, load the supernatant of the acidified sample on to the Sep-pak.

18 | Wash the sample with 3 mL of 0.1% (vol/vol) TFA.

19 | Move the Sep-pak to a 2-mL microfuge tube. Elute the sample with 0.6 mL of 50% (vol/vol) ACN/0.1% TFA, followed by 0.6 mL of 75% (vol/vol) ACN/0.1% TFA.

20 | Freeze the eluate at -80°C and concentrate in Speed-Vac. When complete the sample should be a fluffy white powder. Samples can be stored at -80°C for several months in this form.

21 | Create an instrument method with the following parameters:

Method parameter	Value
MS^1 detector type	Orbitrap
MS^1 resolution	60,000
MS^1 AGC target	5e5
Precursor charge states	2-6
Dynamic exclusion	
Exclusion count	1
Exclusion duration	15-45 s
Exclusion width	±10 ppm
Data dependent mode – Top speed	3 to 5 s
Precursor priority	Most intense
MS^2 isolation mode	Quadrupole
Isolation window	0.7 *m/z*
Activation type	HCD
Collision energy	30%
MS^2 detector type	Ion trap
MS^2 scan rate	Rapid
MS^2 AGC target	1e4
MS^2 max inject time	25 to 35 ms

If the results obtained in this procedure are not as expected or otherwise suboptimal, it might be possible to solve this problem by optimizing the parameters shown in this table using the advice in the section

Optimizing the settings of the mass analyzer.

22 | Resuspend samples to a concentration of 1 µg/µL in 0.2% FA.

23 | Inject 1.4 µL of the sample onto the LC-MS/MS system.

Database searching

24 | In our laboratory raw data is processed using Proteome Discoverer (version 1.4.0.288, Thermo Fischer Scientific), although other software suites are also available. All MS/MS spectra are searched with the SEQUEST search engine against a database of 6,632 yeast ORFs (database downloaded from www.yeastgenome.com, February 3, 2011). Regardless of the software used, set the enzyme specificity to trypsin, with up to two missed cleavages permitted. Set carbamidomethylation of

cysteines as a fixed modification, and oxidation of methionines and protein N-terminal acetylation as variable modifications. Search precursor and product ion mass tolerances at 20 ppm and 0.35 Da, respectively.

25 | Validate PSMs using percolator based on q-values at a 1% false discovery rate (FDR).[17]

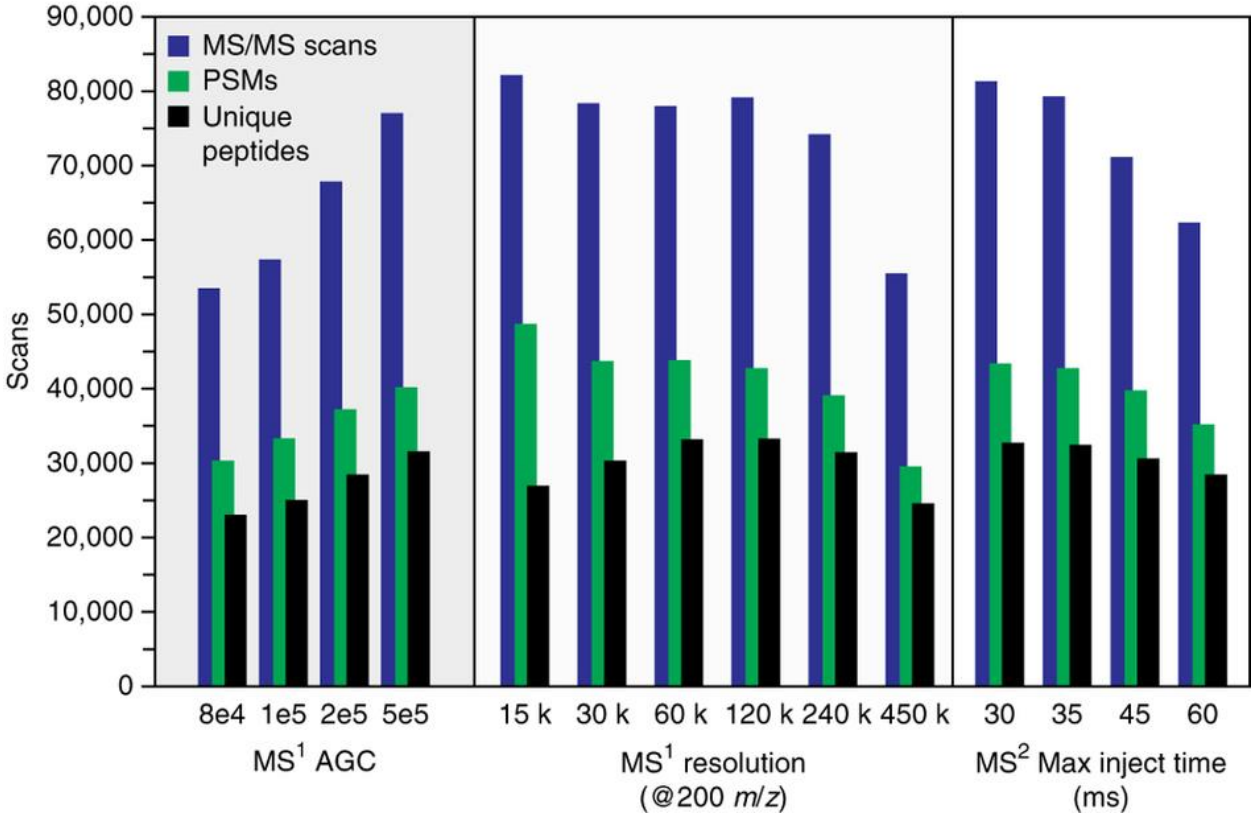
Optimizing the settings of the mass analyzer

The recently introduced Orbitrap Fusion combines an Orbitrap mass analyzer with a quadrupole mass filter, a collision cell, and a dual pressure linear ion trap (LIT). The unique geometry of the Fusion allows the processes of ion injection, precursor isolation, peptide fragmentation and detection to be parallelized, significantly increasing the MS/MS acquisition rate. The optimization of various instrument parameters have been discussed in previous publications.[4,5]

In the data-dependent method described in step 21, survey scans of are performed at an MS resolving power of 60,000 (at 200 *m/z*), precursor peptides are isolated for fragmentation in the quadrupole at a width of 0.7 *m/z*, fragmented in the collision cell followed by mass analysis of the fragments in the ion trap. The instrument is operated in top speed mode with a cycle time of 5 s, meaning that MS/MS events are continuously performed on precursor ions for either 5 s, or until no acceptable precursors remain. Between 15 and 80 min of the LC-MS/MS separation, an average of nine MS/MS scans are taken between MS scans; this number can vary considerably across the gradient, however, with up to 55 MS/MS events triggered from a single MS scan. When operating in this mode, the instrument is capable of scanning rates more than 20 Hz. Of course, other combinations of isolation events, fragmentation methods and detection options are available, but these may affect the duty cycle of the instrument, and in turn, the MS/MS acquisition rate.

We found that optimizing MS^1 AGC target, MS^1 resolving power, and MS^2 maximum injection time had the greatest effect on increasing the number of identified peptides (**Figure 3**). The MS^1 AGC target defines the number of charges introduced to the Orbitrap for the survey scan. To test the effect of MS^1 AGC, we analyzed several settings between 8e4 and 5e5, the highest AGC setting offered on the instrument.

Figure 3. Mass spectrometer settings impact identifications. Effect of MS^1 AGC target, resolution and MS/MS max inject time on number of MS/MS scans, PSMs and unique peptides.



Increasing the number of charges introduced to the mass spectrometer uniformly increases the number of MS/MS scans, PSMs and unique identified peptides.

An increase in the number of unique peptides is also afforded from selecting an MS^1 resolving power of 60,000. Decreasing the resolution correlates linearly with decreased transient collection time; this allows more MS/MS scans to be taken at 15,000 then at other resolving powers with longer transients. However, these extra scans do not translate to more unique peptide identifications. Although the number of MS/MS scans is comparable at resolving powers of 30,000, 60,000 and 120,000, the number of unique peptides was greatest at resolving powers of 60,000 and 120,000, presumably because the higher resolving power allows better resolution of peptidic features, allowing deeper sampling of the proteome. No further increase in unique peptides is observed above 120,000, as any benefits from increased resolving power are outweighed by longer transient time, which results in fewer MS/MS scans, PSMs and peptides.

Increasing the MS/MS max inject time, or the maximum amount of time allowed for MS/MS analysis, predictably results in a decrease in the number of MS/MS scans. This decrease also translates to fewer unique peptide identifications at 45 and 60 ms MS/MS max inject times. As analysis in the ion trap requires 30-35 ms, lower max inject times ensure full parallelization of the mass spectrometer.

TIMING

Column fabrication: 2 h

Yeast protein resuspension and cell lysis: Steps 1-3, 5 min; Steps 4 and 5, 45 min per round of lysis

Determination of protein concentration by BCA assay: Step 6, 40 min

Protein reduction and alkylation: Step 7, 50 min; Step 8, 35 min; Step 9, 15 min

Trypsin digestion: Step 10, 5 min; Step 11, 16 h; Step 12, 2 h

Peptide desalting: Step 13, 5 min; Step 14, 30 min; Step 15, 10 min; Step 16, 2 min; Step 17, 10 min; Step 18, 10 min; Steps 19 and 20, 3-5 h

LC-MS/MS analysis, including sample loading on column, LC-MS/MS gradient, and column washing and equilibration: Steps 21-23, 2 h

Database searching: Steps 24 and 25, 2 h (searching speed may vary depending on the computational platform)

ANTICIPATED RESULTS

In developing this protocol, yeast was analyzed as described in the **Materials** and **Procedure** sections. To test the reproducibility between runs, we performed the experiment in technical quintuplicate. On average, 80,460 MS/MS scans were taken per analyses. Following database searching, this translates to an average of 43,400 PSMs and 34,255 unique peptides. In other words, 54% of MS/MS scans were mapped to a peptide; of these identified peptides, 79% were unique. These peptide identifications yielded an average of 3,977 proteins at 1% FDR. The number of MS/MS scans, PSMs and unique peptides for this data set is presented in **Tables 1**.

Key to these high identification rates is maintaining a consistent number of identifications across the entire LC-MS/MS gradient. **Figure 4*A*** shows the cumulative number of peptide identifications as a function of retention time for each of the five replicates. Cumulative peptide and protein identifications were determined using the Coon OMSSA Proteomic Analysis Software Suite (COMPASS).[18] PSMs passing a 1% FDR cut-off were exported to a text file and processed by a modified version of Protein Hoarder (version 2.4.1). These PSMs were iteratively processed in successive one minute windows and grouped into proteins using the law of parsimony at a 1% FDR.

For all replicates, the rate of identification is almost linear over the 70-minute gradient and 10-minute wash period. The average rate of unique peptide identifications per second is 8.3. Unique protein identifications for each run are plotted in **Figure 4*B***. The consistency of identifications is further illustrated in **Figure 5*A***, where the average number of unique peptides identified across the LC-MS/MS method is plotted for the five replicates. For all runs, we identify between 2,000-2,700 unique peptides per each 5-minute range from 15 minutes to 80 minutes. Expectedly, there is an initial spike in the number of protein identifications at the beginning of the gradient (**Figure 5*B***), with a steady decrease in new protein groups throughout the run. We also examined the effect of gradient length on the number on peptide and protein

identifications. Using the same sample preparation and instrument parameters described above, yeast was analyzed over 30, 45, 70, 120, 180 and 240 LC-MS/MS separations. The number of MS/MS scans, PSMs and unique peptides for these data sets are presented in **Table 2.** Cumulative unique peptide identifications are plotted in **Figure 6*A***. Similar rates of identification are achieved for the 30, 45, and 70 minute separations, as evidenced by the slope of the line. Predictably, longer gradients result in more peptide identifications. However, the number of unique peptides decreases significantly with increasing gradient length. For example, in the 180 minute analysis, 48,970 unique peptides are identified, while 51,211 unique peptides were identified in 240 minutes, an increase of just 2,241 peptides in 60 minutes of analysis time. The number of proteins for each LC-MS/MS gradient is plotted in **Figure 6*B***.

Figure 4. Yeast peptide and protein identifications for all replicates. (***A***) Plots the number of cumulative unique yeast peptide identifications for five technical replicates across the LC-MS/MS gradient. Panel (***B***) plots the corresponding proteins across the gradient. Cumulative peptide and protein identifications were determined using the Coon OMSSA Proteomic Analysis Software Suite (COMPASS)[18]. PSMs passing a 1% FDR cut-off were exported to a text file and processed by a modified version of Protein Hoarder (version 2.4.1). These PSMs were iteratively processed in successive one minute windows and grouped into proteins using the law of parsimony at a 1% FDR.

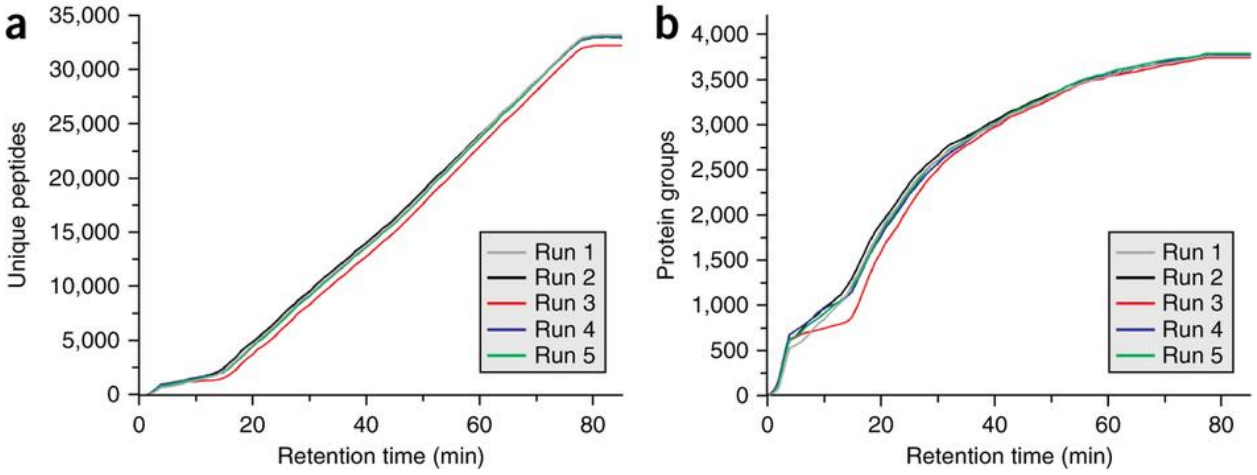


Figure 5. Unique peptides and proteins identified over the LC-MS/MS gradient. (***A***) Plots the number of unique yeast peptides identified in five minute bins for five technical replicates across the LC-MS/MS gradient. Panel (***B***) plots the corresponding proteins across the gradient.

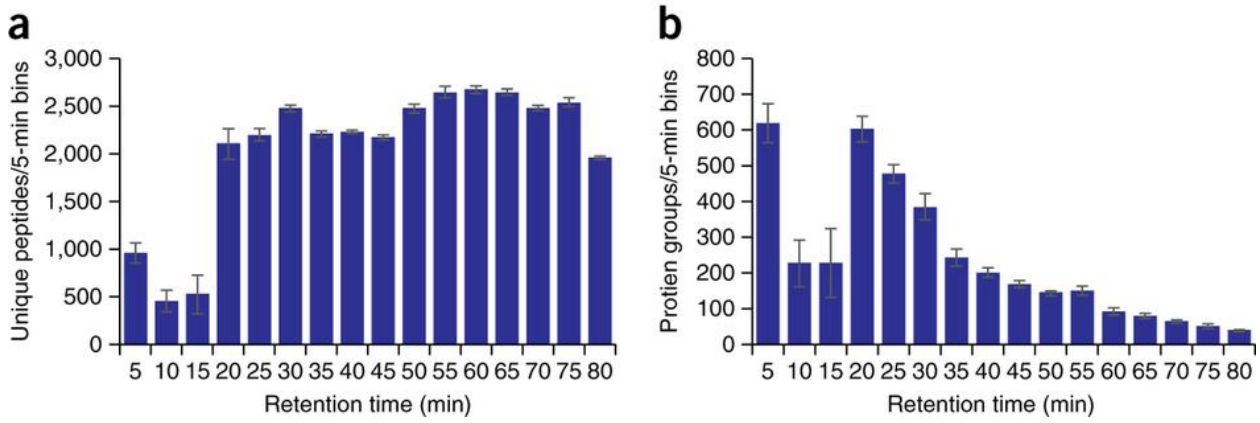


Figure 6. Effect of gradient length on peptide and protein identifications. (***A***) Plots the number of cumulative unique yeast identifications for LC-MS/MS gradients of 30, 45, 70, 120, 180 and 240 min. Panel (***B***) plots the corresponding protein identifications for each gradient.

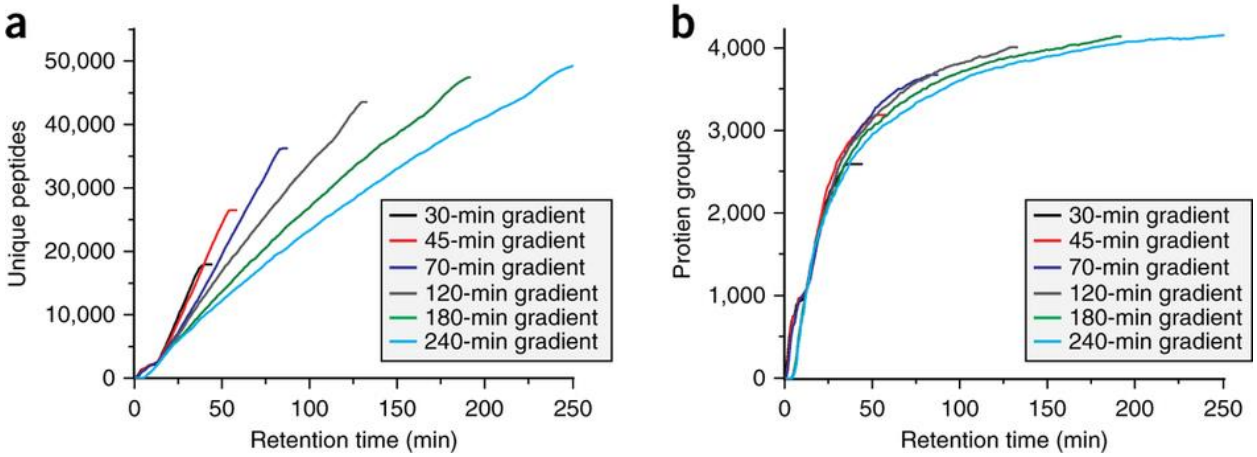


Table 1. Number of MS/MS scans, PSMs and unique peptides for five replicate yeast experiments (Data presented in **Figure 4**).

Experiment	MS/MS scans	PSMs	Unique peptides
1	80,900	43,423	34,535
2	81,252	43,622	34,495
3	78,001	42,339	33,450
4	80,916	43,326	34,347
5	81,235	43,343	34,449

Table 2. Number of MS/MS scans, PSMs and unique peptides for gradient length experiments (Data presented in **Figure 6**).

Gradient length(min)	MS/MS scans	PSMs	Unique peptides
30	45,632	27,700	18,152
45	60,656	35,421	26,931
70	86,267	52,872	36,900
120	124,014	74,813	44,743
180	169,181	96,121	48,970
240	212,088	115,077	51,211

Supplementary Table 1. Recommended laser puller settings. HEAT is the laser power output. FILAMENT is the scan length of the heat supplying laser; the manufacturer's protocol suggests a filament setting of 0 for fused silica. If tips are pulled too thin, the VELOCITY setting should be increased, while decreasing the HEAT setting. DELAY sets the time at which the fused silica is pulled relative to laser deactivation. PULL controls the force with which the fused silica is pulled. A PULL setting of 0 is recommended for fused silica.

Line 1:	heat = 260	filament (fil) = 0	velocity (vel) =10	delay (del) = 128	pull = 0
Line 2:	heat = 260	filament (fil) = 0	velocity (vel) =20	delay (del) = 128	pull = 0
Line 3:	heat = 250	filament (fil) = 0	velocity (vel) =15	delay (del) = 140	pull = 0
Line 4:	heat = 270	filament (fil) = 0	velocity (vel) =15	delay (del) = 128	pull = 0

TROUBLESHOOTING

Steps	Problem	Possible reason	Solution
Column fabrication, Steps 4, 6-9	Liquid is not spraying from analytical column	Column is not fully open	Examine the column under a microscope; if it is not fully open, etch the tip by dipping it in HF for 1-2 minutes (see **Column fabrication** section for details. HF is extremely dangerous; follow appropriate precautions). The column should also be examined for any obstructions (dust, particles) that may prevent it from packing
		Open end of the column is clogged	The end of the column can become damaged or clogged when inserting it through the ferrule leading to the high pressure bomb. After insertion, cut approximately 0.2 cm off the end of the column using a ceramic scoring wafer
Column fabrication, Step 15	Column is packing too quickly with 3.5 µm packing material	Slurry of 3.5 µm packing material is too concentrated	Make a more dilute solution of packing material by increasing the amount of ACN
Column fabrication, Step 17	Column is not packing		Turn helium off, then back on, using the valve connected to the pressure bomb. This will sometimes cause the column to begin packing again
		Open end of the column is clogged	The end of the column can become clogged with packing material. Cut approximately 0.2 cm off the end of the column using a ceramic scoring wafer
		Slurry of 1.7 µm packing material is too concentrated	Ensure that all of the packing material is able to go into solution when stirred; if not, increase the amount of ACN in the slurry. (Note that ACN will evaporate over time)
		Open end of the column does not reach the packing material slurry	Ensure that the end column is fully submerged in the packing material slurry
	Gaps in packing material	Poor flow of packing material; blockage	Increase the pressure of the helium
			Connect to UPLC and flow 100% mobile phase B through the column at 0.3 µl/min for 15 min. Discard column if the gap does not fill

Reversed-phase Chromatography (**Equipment Setup**)	High back pressure during LC-MS/MS separation	Analytical column may be clogged	Replace the column
		High-pressure union is clogged	Over time, the high pressure union can become blocked with packing material or fused silica. Remove the union. When held up to the light, you should be able to see through it. If the union is blocked, wash it with ethanol to remove debris. If necessary, replace the union
		Analytical column is attached too tight at union	If the analytical column is attached too tight, the end can become crushed and block the union. Remove the column and trim the end with a scoring wafer. To avoid this problem, after threading the column through the ferrule, trim the end with a ceramic scoring wafer. The column should protrude ~ 1 mm out of the ferrule. Tighten the analytical column as much as possible, without crushing the tubing. Once the analytical column is connected, gently pull on it. If it comes loose, use slightly more force when tightening the union
		25 µm line clogged	Verify by removing analytical column; if there is not a drastic decrease in pressure, replace the 25 µm line
	High back pressure during sample loading (>8,500 psi)	Impurities in the sample	Ensure that the sample was properly desalted prior to analysis
Chromatography	Very low back pressure	Column not properly attached	Verify whether liquid is flowing from the tip; if not, make sure that the 25 µm line and the analytical column are properly connected at the union. If no liquid flow is observed from the 25 µm line, check its connection to the LC system
	Broad peaks (>1 min)	Column was packed > 1 cm with 3.5 µm packing material	Discard column (see step 11 from **Column fabrication**)
		Gaps in packing material in column	Flow 100% solvent A through the column at 300-350 uL/min (with column heater). Inspect the column; if there are no more gaps and the packed column is of suitable length, trim the unpacked end from the column using a ceramic scoring wafer. If the gaps persist, discard the column

		Sample overloading	Reduce the amount of material injected on column
	Late-eluting sample	Column degradation	Inject 100-300 fmol of tryptic BSA peptides on column; if peaks have shifted significantly, discard column. Similarly, if TIC and base peak intensity are significantly lower than usual and the instrument is working as expected, discard column
		Trapping sample for too long	Decrease trapping time according to instructions in '**Reversed-phase chromatography**' section (**Equipment Setup**)
		Column is not packed all the way to the end	Inspect the column using a light source. If it is not packed to the end, remove the unpacked portion using a ceramic scoring wafer
		25 µm line too long	If possible, trim this line to under 50 cm
		Pumps not delivering correct amount of solvent	Measure the flow rate of pumps A and B; if they are not delivering the correct amount of solvent, the gradient may need to be adjusted to accommodate the actual flow rates. Contact the LC vendor for maintenance help
	Run to run variation	Column is not fully equilibrated	Increase the time of the equilibration period at the end of the LC gradient
		Pumps not delivering correct amount of solvent	Measure the flow rate of pumps A and B; if they are not delivering the correct amount of solvent, the gradient may need to be adjusted to accommodate the actual flow rates. Contact the LC vendor for maintenance help
		Column degradation	Inspect chromatogram for wide peaks. This can also be tested by injecting 100-300 fmol tryptic BSA peptides on column. If peaks are wider than ~40 s, discard column
Step 8	Amidation reaction	Urea degradation during reduction (**Step 7**)	Perform the reduction step (**Step 7**) at a lower temperature (i.e., at ambient temperature)
Step 14	Protein precipitates out of solution upon acidification with TFA		Use a less concentrated solution of TFA; add TFA slowly, checking the pH after each addition to avoid over acidification

Step 24	Fewer identifications than expected; database search gives more semi-tryptic peptides than tryptic peptides	Active endogenous proteases	Collect yeast at an earlier time point. If this is not an option, try digesting with a higher enzyme: protein ratio for a shorter period of time (for example, 1:5 ratio for 4 h) (**Steps 11,12**)
Step 24	Large number (>50%) missed cleavages		Increase enzyme to protein ratio to 1:10 (**Steps 11,12**)
Step 24	Fewer than expected identifications	Instrument is out of calibration	Follow the recommended calibration schedule provided in the calibration console. If MS mass error > 6 p.p.m., recalibrate
Step 24	Fewer than expected identifications; decrease in TIC intensity over time.	Quadrupole may require maintenance	Over time, the quadrupole and other parts of the instrument may require cleaning or maintenance. Contact your service engineer for instructions

ACKNOWLEDGEMENTS

We are grateful to Anna Merrill for yeast production. We thank Audrey Gasch for assistance with yeast growth. This work was supported by the National Institutes of Health (R01 GM080148) and the National Science Foundation (0701846). A.L.R. gratefully acknowledges support from a National Institutes of Health-funded Genomic Sciences Training Program (5T32HG002760).

REFERENCES

(1) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A. J.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; Carr, S. A.; Clauser, K. R.; Jaffe, J. D.; Kowalski, K. A.; Neubert, T. A.; Regnier, F. E.; Schilling, B.; Tegeler, T. J.; Wang, M.; Wang, P.; Whiteaker, J. R.; Zimmerman, L. J.; Fisher, S. J.; Gibson, B. W.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Stein, S. E.; Tempst, P.; Paulovich, A. G.; Liebler, D. C.; Spiegelman, C. *J Proteome Res* **2010**, *9*, 761-776.

(2) Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd. *Anal Chem* **2004**, *76*, 4193-4201.

(3) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd. *Nat Biotechnol* **2001**, *19*, 242-247.

(4) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2014**, *13*, 339-347.

(5) Senko, M. W.; Remes, P. M.; Canterbury, J. D.; Mathur, R.; Song, Q.; Eliuk, S. M.; Mullen, C.; Earley, L.; Hardman, M.; Blethrow, J. D.; Bui, H.; Specht, A.; Lange, O.; Denisov, E.; Makarov, A.; Horning, S.; Zabrouskov, V. *Anal Chem* **2013**, *85*, 11710-11714.

(6) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. *Science* **1996**, *274*, 546, 563-547.

(7) Ghaemmaghami, S.; Huh, W. K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737-741.

(8) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. *Nature* **2008**, *455*, 1251-1254.

(9) Wu, R.; Dephoure, N.; Haas, W.; Huttlin, E. L.; Zhai, B.; Sowa, M. E.; Gygi, S. P. *Mol Cell Proteomics* **2011**, *10*, M111 009654.

(10) Webb, K. J.; Xu, T.; Park, S. K.; Yates, J. R., 3rd. *J Proteome Res* **2013**, *12*, 2177-2184.

(11) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. *Mol Cell Proteomics* **2012**, *11*, M111 013722.

(12) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. *Nature Methods* **2014**, *11*, 319-U300.

(13) Meyer, J. G.; E, A. K. *J Am Soc Mass Spectrom* **2012**, *23*, 1390-1399.

(14) Hahne, H.; Pachl, F.; Ruprecht, B.; Maier, S. K.; Klaeger, S.; Helm, D.; Medard, G.; Wilm, M.; Lemeer, S.; Kuster, B. *Nat Methods* **2013**, *10*, 989-991.

(15) Pirmoradian, M.; Budamgunta, H.; Chingin, K.; Zhang, B.; Astorga-Wells, J.; Zubarev, R. A. *Mol Cell Proteomics* **2013**, *12*, 3330-3338.

(16) Huang da, W.; Sherman, B. T.; Lempicki, R. A. *Nat Protoc* **2009**, *4*, 44-57.

(17) Elias, J. E.; Gygi, S. P. *Nat Methods* **2007**, *4*, 207-214.

(18) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. *Proteomics* **2011**, *11*, 1064-1074.

Chapter 6

Mitochondrial Protein Functions Revealed by Global Mass Spectrometry Profiling

Portions of this chapter are part of a manuscript that has been submitted:

Stefely JA*, Kwiecien NW*, Freiberger EC, **Richards AL**, Jochem A, Rush MJP, Ulbrich A, Robinson K, Hutchins PD, Veling MT, Guo X, Kemmerer ZA, Connors KJ, Trujillo EA, Sokol J, Westphall MS, Hebert AS, Pagliarini DJ, Coon JJ. *Mitochondrial Protein Functions Revealed by Global Mass Spectrometry Profiling*, **2016**.

*Authors contributed equally

ABSTRACT

Mitochondria are complex organelles linked to diverse human diseases, often through incompletely characterized proteins and pathways. Here, using systematic mass spectrometry profiling, we quantified 3,500 biochemical phenotypes across 174 single gene deletion yeast strains to define functions for poorly characterized mitochondrial proteins. By profiling proteome, lipidome, and metabolome changes resulting from diverse mitochondrial perturbations, we revealed correlations between related genes, identified a universal respiration deficiency response that includes potential mitochondrial disease biomarkers, and defined gene-specific perturbations. Integrating our multi-omic dataset of 3.5 million biomolecule measurements revealed functionally-predictive molecule covariance networks. We leveraged these data to examine uncharacterized features of coenzyme Q (CoQ) biosynthesis and defined previously unrecognized roles for Hfd1p, Aro9p, and Aro10p in producing the CoQ precursor, 4-hydroxybenzoate. Collectively, our results provide molecular insight into mitochondrial biology and, more broadly, establish a high-throughput, multi-omic approach for quantifying diagnostic phenotypes and defining protein functions.

INTRODUCTION

Mitochondria are dynamic organelles associated with over 150 human diseases.[1-3] Technical advances in mass spectrometry and systems biology enabled the yeast and mammalian mitochondrial proteomes to be defined.[4-6] However, functional annotation of these proteins lags behind, creating a bottleneck in biomedical research.[7] Nearly 300 of the ~1,200 mammalian mitochondrial proteins are uncharacterized,[8] referred to here as "mitochondrial uncharacterized (x) proteins" (MXPs). Furthermore, given that numerous mitochondrial pathways are not fully understood, we posit that many partially characterized proteins have essential, yet undiscovered roles. A specific example of an incomplete mitochondrial pathway addressed here involves the biosynthesis of ubiquinone (coenzyme Q, CoQ), an essential lipid required for oxidative phosphorylation (OxPhos). While CoQ was discovered 60 years ago,[9,10] multiple enzymes required for its synthesis remain unidentified. Systematic strategies for discovering protein function could fill these gaps, among many others.

Protein functions can be informed by perturbing a biological system and measuring phenotype changes. The yeast *Saccharomyces cerevisiae* is a widely used model organism for elucidating the functions of eukaryotic genes and the proteins they encode. Phenotype analysis of single gene deletion ("Δ*gene*") yeast defined genes that affect fitness, drug resistance, morphology, and respiration,[11-13] and mRNA profiling more broadly characterized responses to environmental changes[14] and genetic perturbations.[15,16] Yet, identifying functionally diagnostic phenotype changes remains a challenge. Often, relevant individual phenotypes cannot be predicted *a priori*, necessitating broad profiling to capture meaningful changes. Even with extensive phenotype coverage, complications can arise when numerous changes occur in response to a perturbation, making it difficult to distinguish functionally relevant changes from a pervasive, non-specific background. We reasoned that advances in high-throughput phenotype profiling strategies could overcome these problems and provide new insight into the functions of genes and proteins.

Using high resolution mass spectrometry we recently accelerated the speed of proteome analysis in yeast to the rate of one proteome per hour.[17,18] Further, we constructed a high resolution GC-

Quadrupole/Orbitrap hybrid mass spectrometer for the analysis of small molecule metabolites by gas chromatography.[19-21] With these technologies, we reasoned that proteome, metabolome, and lipidome analyses of hundreds, or even thousands, of yeast cultures could be performed within several weeks time, providing data on the scale and speed of genomic technologies, but on the actual effector biomolecules. Here, to define functions for MXPs, we used mass spectrometry to analyze the proteomes, metabolomes, and lipidomes of 174 yeast strains in biological triplicate across two metabolic conditions—generating a unique multi-omic dataset of phenotype changes with over 3.5 million biomolecule measurements. To exemplify how this global profiling approach can be used to annotate protein functions, we examine multiple MXPs, with an emphasis on gaps in the mitochondrial CoQ biosynthesis pathway.

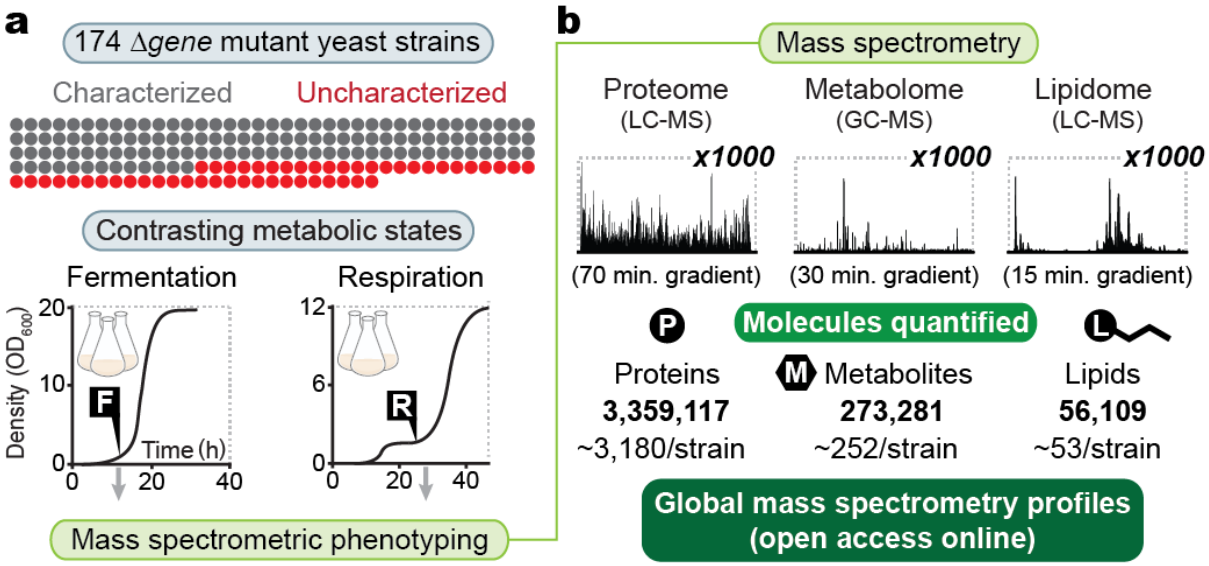
RESULTS

Global mass spectrometry profiling

To generate functionally informative profiles of biochemical phenotype changes, we relied on diverse genetic perturbations, contrasting metabolic states, and quantitative mass spectrometry (**Figure 1**). We analyzed 174 Δ*gene* yeast strains—covering 124 characterized genes and 50 genes encoding MXPs—spanning a broad range of mitochondrial functions (**Extended Data Figure 1*A***). 144 of these genes are conserved in humans and 60 are homologs of genes implicated in human disease. Three biological replicates of each strain were grown under both fermentation and respiration culture conditions (**Figure 1*A* and Extended Data Figure 1*B***).

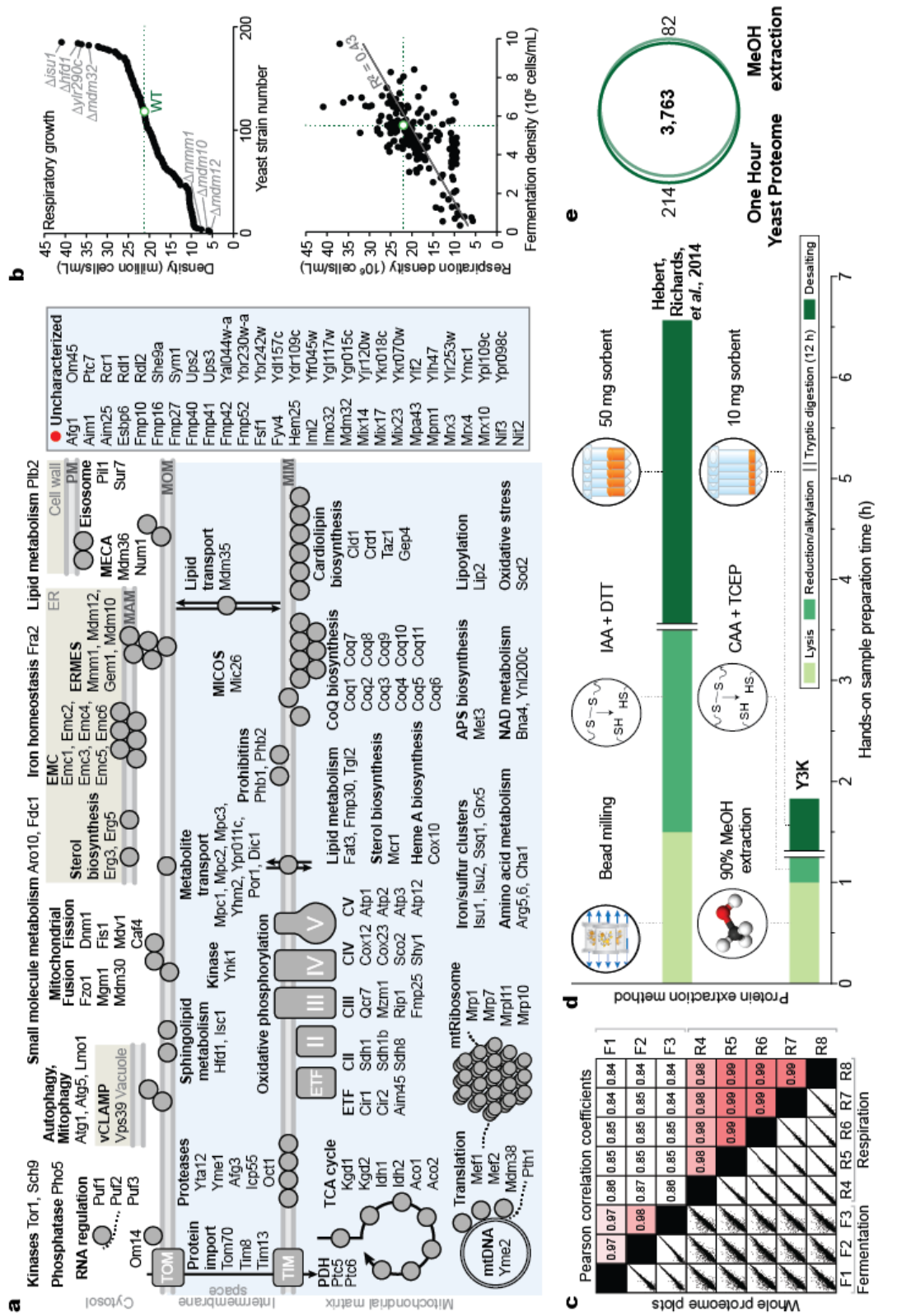
In order to profile diverse yeast strains during respiratory growth—when mitochondrial OxPhos is highly active—we first needed to develop a distinct respiration condition suitable for large scale investigation. Early log phase fermentation cultures repress mitochondrial respiration, high glucose cultures grown past the diauxic shift are too biologically dynamic to allow reproducible sampling across a large scale study,[22,23] and cultures containing solely non-fermentable sugars preclude growth of respiration deficient yeast. To overcome these problems, we developed a culture system that includes low glucose (1 g/L) and high glycerol (30 g/L), which enables a short fermentation phase followed by a longer respiration

Figure 1. Global mass spectrometry profiles. (***A***) Yeast lacking characterized or uncharacterized genes were grown under contrasting metabolic conditions. Proteins, lipids, and water-soluble metabolites were extracted, identified, and quantified by high resolution mass spectrometry. (***B***) Over 3,000 mass spectrometry-based proteomic, metabolomic, and lipidomic experiments were conducted, yielding over 3,500,000 biomolecule measurements. The resultant global profiles of molecule perturbations can be explored online through a free visualization suite.

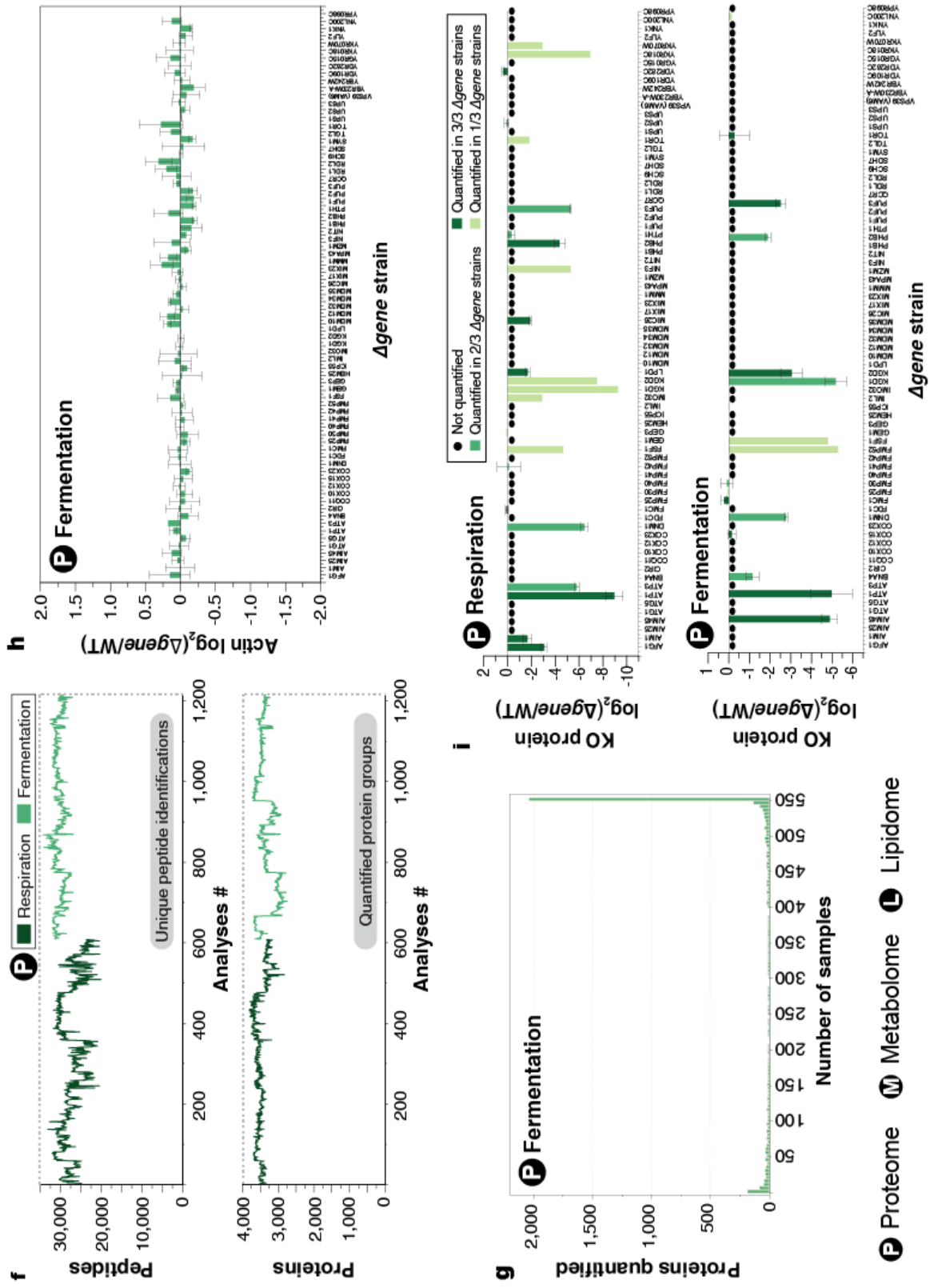


Extended Data Figure 1. **Workflow optimization and target strain characteristics.** (***A***) Proteins encoded by the individual genes knocked out of the 174 yeast strains investigated in this study, shown in the context of biological pathways. APS, adenosine-5'-phosphosulfate; CII–CV, oxidative phosphorylation 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. (***B***) Density of yeast cultures in the respiratory growth condition (mean, n = 3) plotted in strain rank order (***top***) or against fermentation culture density (mean, n = 3) (***bottom***). (***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***) Overview of the yeast protein extraction method optimized for this study compared to previous work. (***E***) Proteins identified using the extraction method optimized for this study compared to the proteins identified in our previous work. (***F***) Number of unique peptides (*top*) and quantified proteins (**bottom**) in each analysis across all respiration and fermentation datasets. (***G***) Number of proteins quantified per number of samples. (***H***) Abundance measurements of actin a selection of Δ*gene* strains in the fermentation dataset. (***I***) Fold changes of select proteins from their corresponding Δ*gene* strains ($\log_2$(Δ*gene*/WT) for the respiration (***top***) and fermentation (***bottom***) datasets. (***J***) Number of proteins, lipids, and metabolites quantified per Δ*gene* strain under fermentation and respiration conditions (mean ± s.d., n = 3). (***K***) Violin plots depicting the range of fold changes in molecule abundance ($\log_2$(Δ*gene*/WT)) across all molecule classes and metabolic states. (***L***) Density plots of the distribution of coefficients of variation (CVs) (%) for each molecule measured in biological triplicate across all mutants and growth conditions. (***M***) Venn diagrams depicting the average overlap of molecules quantified within in a single Δ*gene* strain across fermentation and respiration growth conditions. (***N***) Average profile overlap between different Δ*gene* strains. (***O***) Mass spectrometry (MS) experiments conducted per day (***top***) and phenotypes (molecules) quantified per day (***bottom***) for proteomics, lipidomics, and metabolomics.

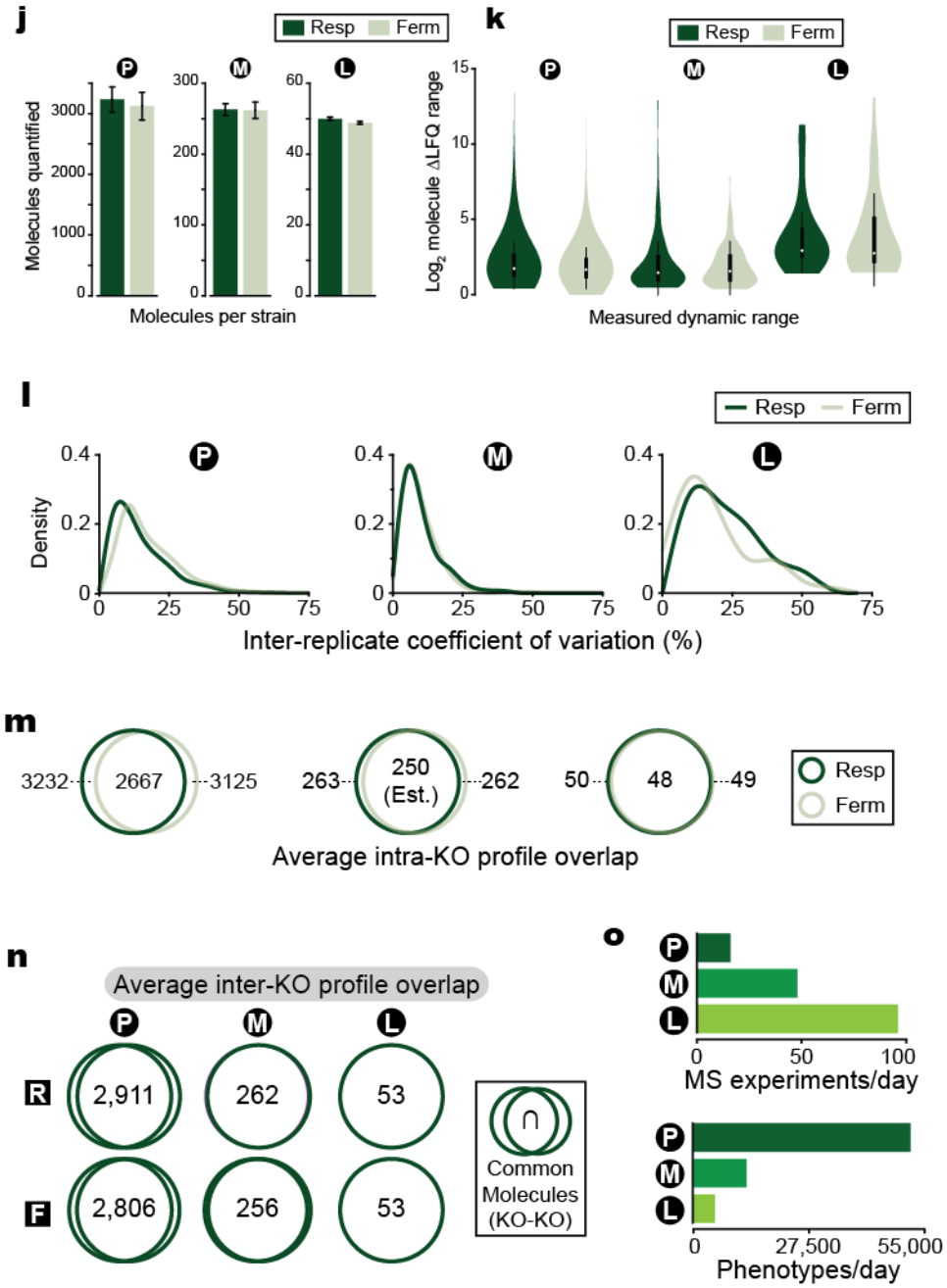
Extended Data Figure 1 (*continued*).



Extended Data Figure 1 (*continued*).



Extended Data Figure 1 (*continued*).



phase. This respiration condition affords steady growth and a stable biological state—as reflected by a proteome that is stable over multiple hours (**Extended Data Figure 1*C***)—and, thus, an essential window for reproducible sample harvesting.

All 1,044 yeast cultures were analyzed by high resolution mass spectrometry (MS), yielding quantitative measurements of 4,040 proteins, 411 metabolites, and 53 lipids (averaging 3,180 proteins, 252 metabolites, and 53 lipids per strain, per growth condition) (**Figure 1*B***). Key to our success was streamlining procedures for proteome extraction and preparation to under two hours of hands-on time per culture (**Extended Data Figure 1*D***). Using our previously developed high throughput methodology, we analyzed 16 yeast proteomes per day per MS system. Altogether, 55,000 proteins were quantified per day across these 16 samples. For each sample, peptides were separated via a high-resolution nano-UPLC interfaced to an Orbitrap Fusion mass spectrometer that provides MS/MS sampling rates > 20 Hz. We used an ion intensity-based label free algorithm (MaxLFQ) to measure the relative expression of the vast majority of detected proteins (>85%). Using this experimental approach, we compare the proteomes of all KO strains to wild-type control cells.

To reduce the hands-on processing time required to prepare each sample for MS analysis, we developed a system of cellular disruption, protein digestion and peptide desalting requiring ~two hours per set of twenty samples (**Extended Data Figure 1*D***). Briefly, yeast pellets were resuspended in lysis buffer (8M urea, 100 mM tris, pH = 8.0). Methanol (MeOH) was added to 90% total volume, and proteins were pelleted through centrifugation. In addition to effectively lysing the cells, the MeOH precipitation eliminated endongenous, non-specific protease activity, which was problematic in some of the Δ*gene* strains. The use of tris(2-carboxyethyl) phosphine (TCEP) and chloroacetamide (CAA) allowed reduction and alkylation to take place simultaneously, further reducing processing time.[24]

We compared the proteins identified from replicate analysis of yeast lysed via 90% MeOH precipitation to the total proteins identified in our previous work,[17] where lysis was performed through an extended bead-beating procedure. Following MaxQuant searching, 3,845 proteins were identified from the MeOH dataset; 3,977 proteins were identified from our previous dataset (**Extended Data Figure 1*E***). Of

the 4,059 total protein groups identified, 3,763, or 93%, were common to both sample preparation techniques, indicating very high overlap between both preparation methods. To determine which proteins we are "missing" in our analyses, meaning they are not routinely identified in our one-hour MS experiments, we compared the proteins identified in either experiment against the list of expected yeast proteins in the Uniprot database. To determine if a specific class of proteins was under-represented in our analyses, we performed gene ontology (GO) enrichment on the subset of proteins that were not identified.[25] In both sample preparation methods, the missing proteins were enriched for cellular processes related to membrane proteins, including those intrinsic to membrane and integral to membrane. Membrane proteins are notoriously difficult to identify via MS, owing to their hydrophobicity and low abundance compared with other proteins. Again, similar results were obtained for both lysis methods, suggesting our methanol extraction is able to achieve comparable results with the bead beating methodology, in a fraction of the time.

Comparing the proteomes across hundreds of analyses required high reproducibility across the datasets. The number of unique peptides and quantified proteins was tracked across all analyses (**Extended Data Figure 1*F***). Although these analyses were performed on two separate instruments, over a period of several months, and comprised two unique growth conditions, identifications remained consistent across all samples. Altogether, this dataset comprises over 1.5 million PSMs and 162,859 unique peptides. Analysis of all respiration datasets identified 812,651 PSMs and 135,106 unique peptides. An average of 27,485 unique peptides were identified per respiration sample; from these peptides, an average of 3,504 proteins were quantified per sample. Across the fermentation datasets, we identified 721,566 PSMs and 128,511 unique peptides. Each sample in the fermentation dataset yielded an average of 30,358 unique peptides and 3,350 quantified proteins. Note, the number of quantified proteins are lower in the final dataset, as proteins having unexpected misregulation were omitted from the final comparisons. The distribution of the overlap in quantified proteins across the 554 samples in the fermentation dataset is presented in **Extended Data Figure 1*G***. 1,665 were present across all 554 samples, with 2,019 proteins were present in at least 550 of the samples. The same 3,000 proteins were quantified in 407 of the 554

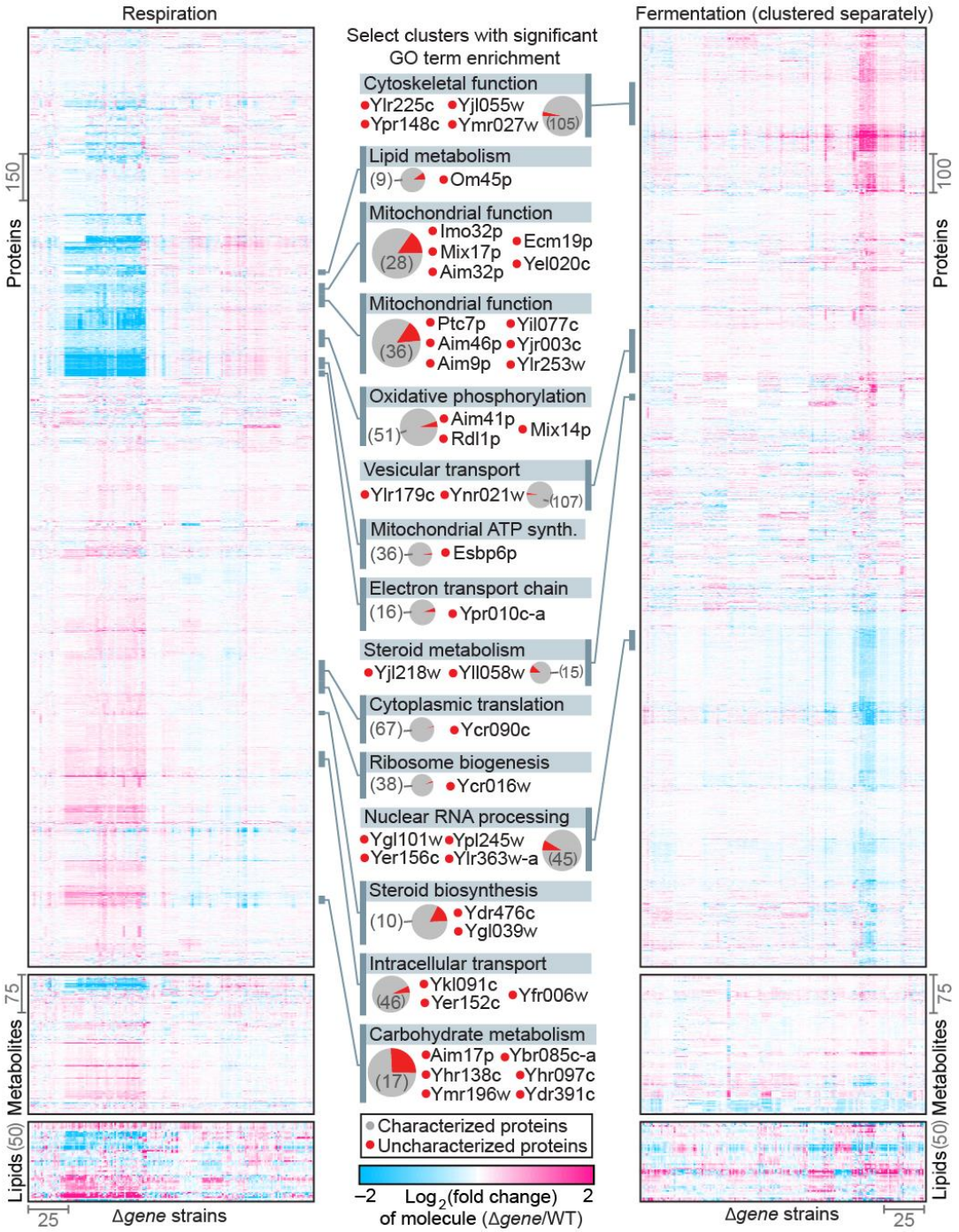
analyses, or across 73% of all WT and Δ*gene* strains. Finally, 84 proteins were quantified in only 1 sample. We believe this number is artificially high, as database searching and protein grouping was performed in batches of 60, rather than across the entire dataset.

Label-free quantitation of the identified proteins was performed using the MaxQuantLFQ algorithm. Briefly, quantitation is performed by taking the area under the curve (AUC) of a given peptide in the MS scan. To compare fold changes across datasets, the protein label-free quantitation (LFQ) values for Δ*gene* strain were compared to LFQ values from a WT control. To assess the accuracy of this approach across a large dataset, we examined the expression of actin across all analyzed samples. Actin is a highly expressed cytoskeletal housekeeping protein whose abundance should be consistent across all samples. As expected, actin was quantified in all WT and Δ*gene* strains across the dataset. The fold change (Δ*gene*/WT) for select Δ*gene* strains in the fermentation dataset is shown in **Extended Data Figure 1*H***. Across all Δ*gene* strains, expression levels are centered around zero, meaning there was little change in expression between all WT and Δ*gene* strains, and also between Δ*gene* strains. We also examined the fold changes (Δ*gene*/WT) obtained for proteins that were knocked out (**Extended Data Figure 1*I***). In almost all cases, these proteins exhibited a very large negative fold change or were not detected in our analysis, indicating the expected loss of deleted proteins.

Using a label-free quantitation approach negated the need for a chemical tagging step, further increasing experimental throughput. We observed a wide dynamic range across all profiled omes, with select molecule abundances spanning more than 3 orders of magnitude (**Extended Data Figure 1*K***). 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 (**Extended Data Figure 1*L–O***). An open access web portal for interactive visualization, exploration, and comparison of all 3.5 million biomolecule measurements within the dataset is freely available online.

A high-level view of the entire dataset reveals significant perturbations in both fermentation and respiration across all three omes, with more pronounced perturbations in respiration (**Figure 2**). Hierarchical

Figure 2. **A global view of protein-lipid-metabolite perturbation profiles.** Hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT quantified by mass spectrometry). The center column annotates select clusters with significant functional (GO term) enrichments. Pie charts indicate proteins in clusters encoded by characterized (***grey***) or uncharacterized (***red***) genes.



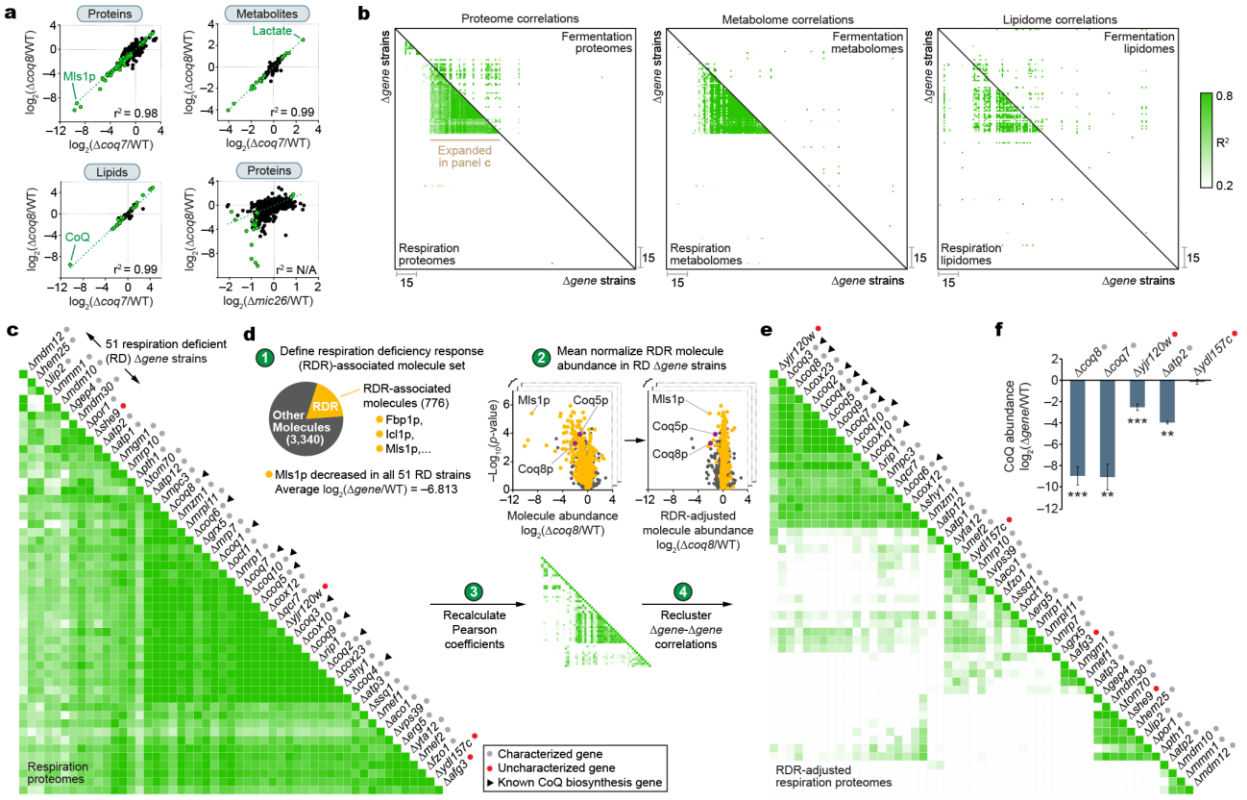
clustering reveals groups of functionally related molecules (along the y-axis) and groups of functionally related Δ*gene* strains (along the x-axis). Protein clusters show significant gene ontology (GO) term enrichments for diverse processes and include both characterized and uncharacterized proteins (**Figure 2**). For example, the MXPs Esbp6p and Ypr010c-a are clustered with proteins involved in mitochondrial ATP synthesis and electron transport chain function, respectively, suggesting that they function in these pathways. Thus, global mass spectrometry profiling reveals molecular fingerprints with the potential to catalyze annotation of uncharacterized proteins and genes through "guilt by association" along both the molecule-molecule and Δ*gene*–Δ*gene* (strain–strain) axes.

Regression analysis of phenotype changes

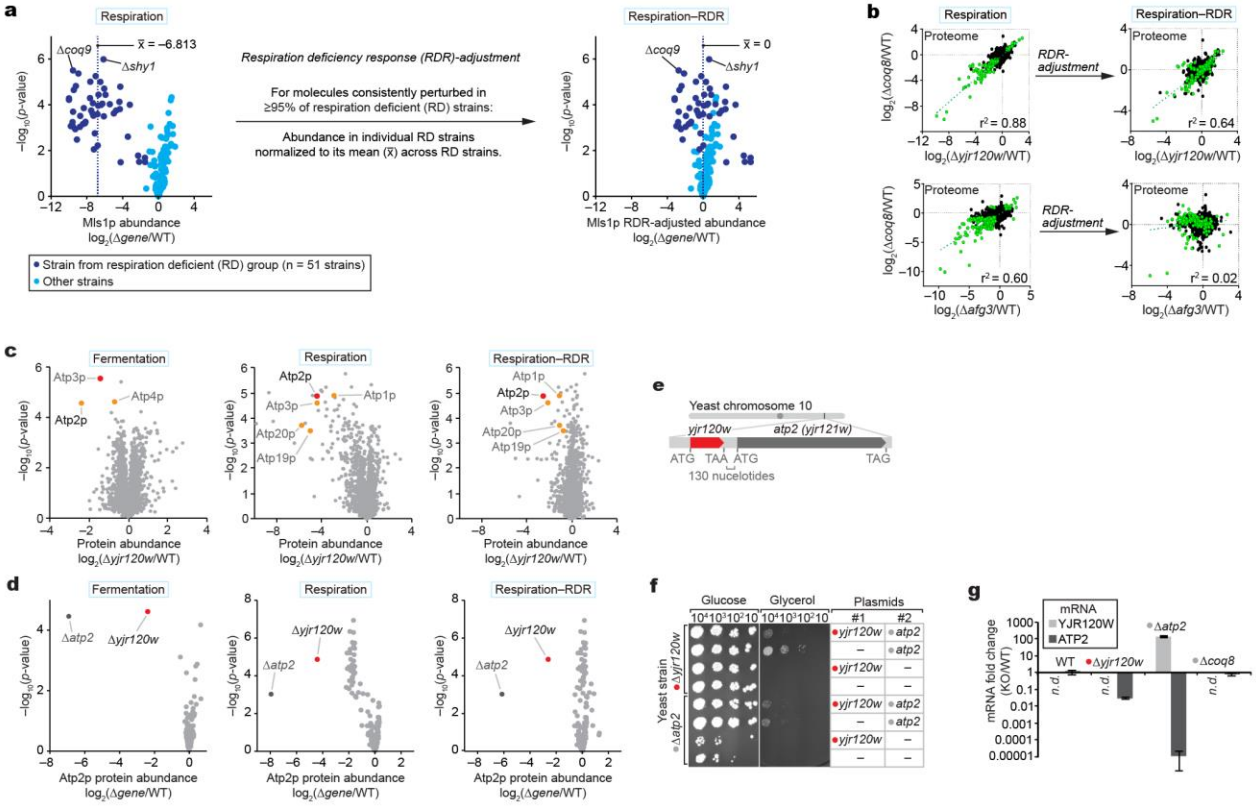
We examined Δ*gene*–Δ*gene* correlations through pairwise comparisons of Δ*gene* perturbation profiles. Deletion of functionally related genes, such as the CoQ biosynthesis genes *coq7* and *coq8*, caused phenotype changes with strong linear correlations across all three omes, while those of unrelated genes, such as *coq8* and *mic26*, lack meaningful correlations (Fig. 3a). Examining correlations across all strains in the study reveals numerous functional relationships, with stronger correlations observed in respiration (**Figure 3*B***).

A group of respiration-deficient (RD) strains showed robust correlations across all three omes (**Figure 3*B, C***), suggesting that they share universal biomarkers of mitochondrial dysfunction and reflecting their similar broad biological functions in mitochondrial OxPhos. This RD profile offers important new insights into the general nature of mitochondrial dysfunction (see below); however, to assess more specific biochemical roles for individual MXPs, we normalized for this common "respiration deficiency response" (RDR) found across RD strains (**Figure 3*D***). 776 molecules were identified as being consistently perturbed across RD strains, and the individual RD strain measurements of these RDR-associated molecules were mean normalized to reveal characteristic deviations from the general RDR (visualized as "respiration–RDR" plots) (**Figure 3*D*** and **Extended Data Figure 3*A, B***). Importantly, this procedure enables identification of Δ*gene*-specific changes buried in a background of non-specific, RDR-associated changes.

Figure 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*-value < 0.05). (***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***) Expanded view of highly correlated strains in the respiration proteomes correlation map. (***D***) Procedure for subtraction of a common respiration deficiency response (RDR). (***E***) Re-clustered respiration proteome strain-strain correlation map after RDR-adjustment. (***F***) CoQ abundance changes in select Δ*gene* strains (mean ± s.d., n = 3).



Extended Data Figure 3. **Molecular fingerprint regression analysis and subtraction of shared responses to reveal deeper biochemical insight.** (***A***) Respiration deficiency response (RDR)-abundance adjustment of a representative molecule (Mls1p) by subtraction of the average fold change in abundance (mean $\log_2$(Δ*gene*/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 ($|\log_2(FC)| > 0.7$, p-value < 0.05) prior to RD-abundance adjustment. (***C***) Relative protein abundances (mean $\log_2$(Δ*yjr120w*/WT), n = 3) versus statistical significance ($-\log_{10}$(p-value)) as quantified by mass spectrometry. (***D***) Relative Atp2p protein abundance (mean $\log_2$(Δ*gene*/WT), n = 3) versus statistical significance ($-\log_{10}$(p-value)) across all mutants in the study. (***E***) Genomic organization of *yjr120w* and *atp2*. (***F***) Serial dilutions of yeast transformed with the indicated plasmids grown on agar plates with glucose (to enable fermentation) or glycerol (to force respiration). (***G***) 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.



For example, prior to RDR normalization, the decrease in Coq8p—an atypical kinase required for CoQ biosynthesis[26]—in Δ*coq8* yeast is hidden by RDR-associated proteins with large abundance changes (**Figure 3*D***). RDR normalization not only makes the decrease in Coq8p stand out, but a significant decrease in Coq5p also becomes readily apparent, suggesting a strong functional link between these two proteins (**Figure 3*D***). Accordingly, we recently demonstrated that mammalian COQ8A interacts with and stabilizes COQ5 *in vivo* (Floyd et al., 2016, *in revision*) (Stefely et al., 2016, *submitted*). Recalculating correlation coefficients with respiration–RDR plots strikingly reduces correlations between more functionally disparate genes, and reclustering Δ*gene*–Δ*gene* correlations reveals new clusters of genes with similar biochemical correlations reveals new clusters of genes with similar biochemical functions (**Figure 3*D, E***). For example, known CoQ biosynthesis genes were brought into a tighter cluster that also includes the uncharacterized gene *yjr120w* (Fig. 3c, e), suggesting that *yjr120w* might support CoQ biosynthesis. Consistently, we observed CoQ deficiency in Δ*yjr120w* yeast (**Figure 3*F***).

To examine the molecular basis for the CoQ deficiency of Δ*yjr120w* yeast, we inspected our proteomics dataset, which revealed significant decreases in ATP synthase proteins, especially Atp2p (**Extended Data Figure 3*C, D***). Compared to all other strains in our study, the large decrease in Atp2p is unique to Δ*yjr120w* and Δ*atp2*. A relationship between *yjr120w* and *atp2* is also suggested by their genetic proximity (**Extended Data Figure 3*E***). Consistently, plasmid overexpression of *atp2* rescues the Δ*yjr120w* respiratory growth defect (**Extended Data Figure 3*F***), demonstrating a functional relationship between *atp2* and *yjr120w in vivo*. A decrease in *atp2* mRNA in the Δ*yjr120w* strain is likely a component of the underlying mechanism (**Extended Data Figure 3*G***). Interestingly, CoQ deficiency was also observed in Δ*atp2* yeast (**Figure 3*F***). Collectively, these results show that specific ATP synthase subunits support CoQ biosynthesis and, more broadly, demonstrate that global mass spectrometry profiling can reveal functional links between perturbed genes.

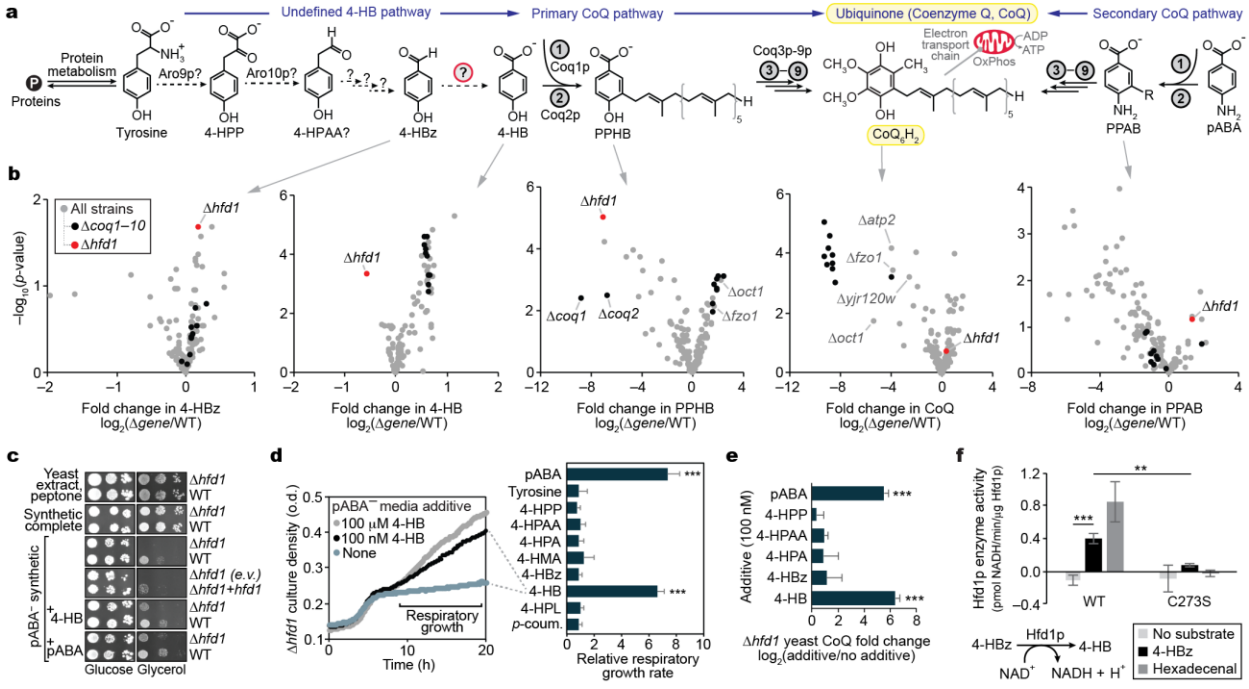
Multi-omic biochemical pathway analysis connects Hfd1p to CoQ biosynthesis

Multi-omic profiling enables parallel measurements of enzymes, substrates, and products in biochemical pathways, thereby providing an opportunity to identify enzyme-substrate relationships. Here, we tested this approach by examining undefined features of the CoQ biosynthesis pathway, which include multiple missing enzymes (**Figure 4*A*** and **Extended Data Figure 4*A***). Significant CoQ deficiency and accumulation of the early CoQ intermediate 3-polyprenyl-4-hydroxybenzoate (PPHB) was observed for known *coq* genes and, interestingly, some genes not previously linked to CoQ function, such as *oct1* and *fzo1* (**Figure 4*B***). Consistently, disruption of the mammalian *fzo1* homolog, *Mfn2*, was recently shown to cause CoQ deficiency,[27] suggesting an evolutionarily conserved role for mitochondrial fusion in CoQ biosynthesis.

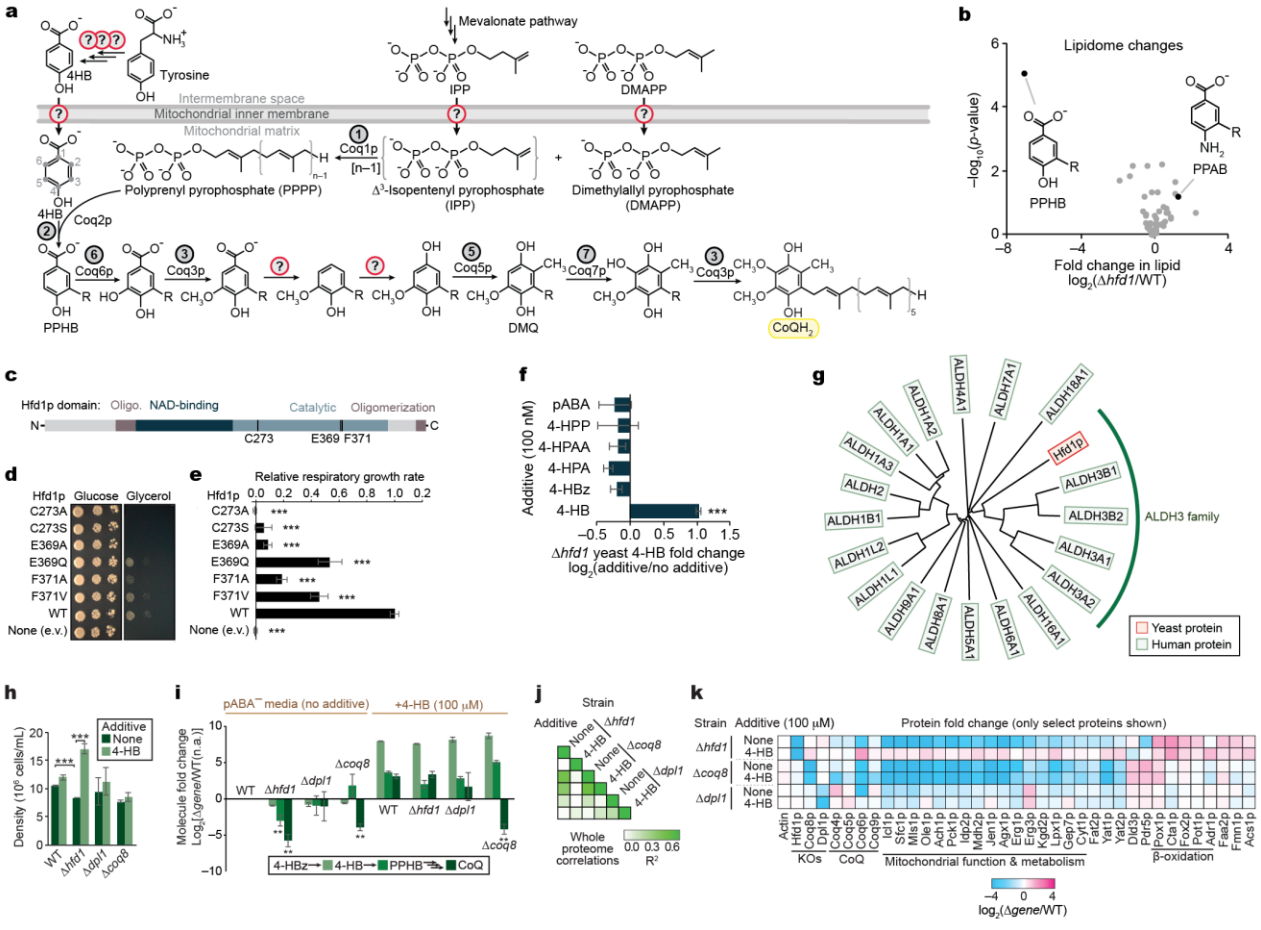
Since the 1960s, we have known that mammals can convert tyrosine (Tyr) into 4-hydroxybenzoate (4-HB) for CoQ biosynthesis,[28-30] however, the biochemical pathway remains undefined in both mammals and yeast (**Figure 4*A***). Our metabolomics and lipidomics data show that Δ*hfd1* yeast are significantly deficient in 4-HB and PPHB (**Figure 4*B*** and **Extended Data Figure 4*B***). Notably, the small decrease in 4-HB in the Δ*hfd1* yeast stands out when compared to all other strains in the study, highlighting an important benefit of analyzing many Δ*gene* strains in parallel. Despite the PPHB deficiency, Δ*hfd1* yeast have normal CoQ abundance (**Figure 4*B***), likely because of increased flux through an alternative *para*-amino benzoate (pABA)-dependent CoQ pathway,[31,32] as suggested by elevation of the aminated analog of PPHB (PPAB) in Δ*hfd1* yeast (**Figure 4*B***). Because Hfd1p is predicted to be an aldehyde dehydrogenase,[33] we hypothesized that it catalyzes dehydrogenation of 4-hydroxybenzaldehyde (4-HBz) to form 4-HB. Consistently, 4-HBz is elevated in Δ*hfd1* yeast (**Figure 4*B***).

We used a chemical-genetic approach to test our proposed role for Hfd1p in dehydrogenation of 4-HBz to form 4-HB. Most yeast medias contain either 4-HB (a component of "yeast extract" in rich media) or pABA (a component of "yeast nitrogen base" in standard synthetic media), enabling yeast to bypass the Tyr-to-4-HB pathway for CoQ biosynthesis. Here, we used a specially formulated synthetic media lacking pABA ($pABA^{-}$), which allowed us to define the aromatic precursors present. Δ*hfd1* yeast exhibited a striking lack of respiratory growth on media lacking 4-HB and pABA, a phenotype rescued by addition of

Figure 4. **Hfd1p supports production of 4-hydroxybenzoate for coenzyme Q biosynthesis.** (***A***) CoQ biosynthesis pathways, highlighting the undefined Tyr-to-4-HB pathway. 4-HPP, 4-hydroxyphenylpyruvate; 4-HPAA, 4-hydroxyphenylacetaldehyde; 4-HBz, 4-hydroxybenzaldehyde; 4-HB, 4-hydroxybenzoate; PPHB, 3-polyprenyl-4-hydroxybenzoate; PPAB, 3-polyprenyl-4-aminobenzoate; pABA, para-aminobenzoate. (***B***) Relative abundances of 4-HB, PPHB, CoQ, and PPAB (mean, n = 3) versus statistical significance across all mutants in the study. (***C***) Serial dilutions of yeast grown on variable solid medias. E.v., empty vector; +*hfd1*, *hfd1* plasmid transformed. (***D***) Relative respiratory growth rates of Δ*hfd1* yeast in pABA^{-} synthetic media with the additives shown (mean ± s.d., n = 3). 4-HPA, 4-hydroxyphenylacetate; 4-HMA, 4-hydroxymandelate; 4-HPL, 4-hydroxyphenyllactate; *p*-coum., para-coumarate. (***E***) Relative abundances of CoQ in Δ*hfd1* yeast cultured in pABA^{-} media with the additives shown (mean $\log_2$(additive/unsupplemented) ± s.d., n = 3). (***F***) Enzyme activity of recombinant Hfd1p *in vitro* (mean ± s.d., n = 4 independent protein preparations). **P* < 0.05; ***P* < 0.01; ****P* < 0.001.



Extended Data Figure 4. **Hfd1p supports production of 4-HB for CoQ biosynthesis.** (***A***) Ubiquinone (CoQ) biosynthesis pathway, highlighting undefined steps. 4-HB, 4-hydroxybenzoate; PPHB, 3-polyprenyl-4-hydroxybenzoate; DMQ, demethoxy-CoQ. (***B***) Relative lipid abundances (mean $\log_2$(Δ*hfd1*/WT), n = 3) versus statistical significance (–$\log_{10}$(*p*-value)) as quantified by mass spectrometry. (***C***) Protein domain structures of Hfd1p, highlighting residues involved in catalysis. (***D***) Serial dilutions of Δ*hfd1* 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 Δ*hfd1* yeast transformed with plasmids encoding the indicated Hfd1p variants and grown in pABA$^-$ synthetic liquid media. (***F***) Relative abundances of 4-HB in Δ*hfd1* yeast cultured in pABA$^-$ media with the additives shown (mean $\log_2$(additive/unsupplemented) ± s.d., n = 3). (***G***) Phylogenetic tree of human ALDH superfamily members and yeast Hfd1p. (***H***) Density of yeast (at time point of harvest, 25 h) cultured in pABA$^-$ media with or without 4-HB (mean ± s.d., n = 3). (***I***) Relative abundances of 4-HB, PPHB, and CoQ compared to WT yeast cultured in unsupplemented pABA$^-$ media (mean $\log_2$(Δ*gene*/WT with no additive) ± s.d., n = 3) as quantified by mass spectrometry. (***J***) Whole proteome correlation map (as in Figure 3) for yeast grown in pABA$^-$ media with or without 4-HB (mean, n = 3). (***K***) Relative abundances of select proteins as quantified by mass spectrometry (mean $\log_2$(Δ*gene*/WT), n = 3) analysis of yeast cultured in pABA$^-$ media with or without 4-HB. **P* < 0.05; ***P* < 0.01; ****P* < 0.001.



pABA or 4-HB (**Figure 4*C***). The pABA$^-$ Δ*hfd1* phenotype is also rescued by WT Hfd1p expressed from a plasmid, but not by Hfd1p with catalytic residue mutations (**Figure 4*C*** and **Extended Data Figure 4*C*–*E***). Testing a panel of potential intermediates in the Tyr-to-4-HB pathway revealed that 4-HB and pABA but not 4- HBz —can rescue the pABA$^-$ respiratory growth and CoQ production of Δ*hfd1* yeast (**Figure 4*D, EE*** and **Extended Data Figure 4*F***) supporting a role for Hfd1p in dehydrogenation of 4-HBz to produce 4-HB. To directly test this proposed activity, we purified recombinant Hfd1p for enzyme assays. WT Hfd1p catalyzes NADH-dependent dehydrogenation of 4-HBz (**Figure 4*F***), but a C273S point mutant, which cannot rescue the pABA$^-$ respiratory phenotype of Δ*hfd1* yeast (**Extended Data Figure 4*D, E***), does not (**Figure 4*F***). Collectively, this work demonstrates that Hfd1p catalyzes dehydrogenation of 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 at least 19 human homologs with diverse metabolic functions.[34] Based on pNADH hylogenetic analyses, Hfd1p is most similar to the human ALDH3 family (**Extended Data Figure 4*G***). While ALDH3A2 (FALDH) mutations are known to cause Sjogren-Larsson Syndrome[35] due to a defect in fatty aldehyde metabolism, 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 hexadecenal.[33] Consistently, we observed Hfd1p activity *in vitro* with hexadecenal, similar to that observed with 4-HBz (**Figure 4*F***). 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 (**Extended Data Figure 4*H, I***). 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 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 (**Extended Data Figure 4*J, K***). Collectively, these results demonstrate a major cellular function for the aldehyde

dehydrogenase Hfd1p in the Tyr-to-4-HB pathway and suggest that one of its human homologs (ALDH3A1, A2, B1, or B2) likely plays a similar role. This work also lays out an approach for future studies to define the numerous additional undefined steps in the Tyr-to-4-HB pathway and, more broadly, demonstrates how multi-omic mass spectrometry profiling can be used to annotate an elusive biochemical pathway.

CONCLUSION

One of the next great challenges in biomedical research is to annotate the functions of all human genes and proteins[7]. Mitochondria encapsulate a vital portion of this challenge, as they are associated with diverse human diseases and myriad cellular activities,[1-4] but still contain many incompletely defined pathways.[8] Here, we employed high resolution quantitative mass spectrometry to elucidate functions for uncharacterized mitochondrial proteins (MXPs). We annotated proteins that support mitochondrial CoQ biosynthesis—including three enzymes (Hfd1p, Aro9p, and Aro10p) in the previously undefined Tyr-to-4-—a protein that likely supports mitochondrial ribosome function (Yor020w-a), and a gene that supports mitochondrial ATP synthase function (*yjr120w*). Our work also defines a universal cellular response to respiration deficiency and highlights potential new biomarkers for human mitochondrial disease. Moreover, this work reveals strong functional associations for numerous additional uncharacterized proteins and provides a molecular foundation for future biochemical work on myriad mitochondrial pathways. To make our multi-omic dataset of over 3.5 million biomolecule measurements accessible to the entire biomedical research community, we developed a suite of visualization tools that are freely available online (y3kproject.org). More broadly, this work establishes a high-throughput approach for quantifying diagnostic phenotypes and defining functions of genes and proteins in any organism.

EXPERIMENTAL PROCEDURES

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[12] or made in-house using a *KanMX* deletion cassette to match those in the consortium collection. Gene deletions were confirmed by either proteomics (significant decrease in the encoded protein) or a PCR assay.

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[36]. YPD or YPGD media (100 mL media at ambient temperature in a sterile 250 mL Erlenmeyer flask) was inoculated with $2.5x10^6$ 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. For growth curve analyses, media glucose concentration was determined with a Glucose (HK) Assay Kit (Sigma) as described[36].

LC-MS/MS proteomics. $1x10^8$ 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 methodologies[17,18].

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 (Version 1.5.0.25).[37] Searches were performed against a target-decoy database[38] of reviewed yeast proteins plus isoforms (Uniprot, downloaded January 20, 2013) using the Andromeda search algorithm.[39] 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[40] 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 ≥ 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. 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 dataset. 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.

LC-MS lipidomics. $1x10^8$ 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. Frozen yeast pellets ($1x10^8$ cells) were thawed on ice and mixed with glass beads (0.5 mm diameter, 100 µL). $CHCl_3$/MeOH (1:1, v/v, 4 °C) (900 µL) was added and vortexed (2 × 30 s). HCl (1 M, 200 µL, 4 °C) was

added and vortexed (2 × 30 s). The samples were centrifuged (5,000 *g*, 2 min, 4 °C) to complete phase separation. 400 µL of the organic phase was transferred to a clean tube and dried under $Ar_{(g)}$. The organic residue was reconstituted in ACN/IPA/H_2O (65:30:5, v/v/v) (100 µL) for LC-MS analysis.

LC-MS analysis was performed on an Ascentis Express C18 column held at 35 °C (150 mm × 2.1 mm × 2.7 µm particle size; Supelco) using an Accela LC Pump (500 µL/min flow rate; Thermo). Mobile phase A consisted of 10 mM ammonium acetate in ACN/H_2O (70:30, v/v) containing 250 µL/L acetic acid. Mobile phase B consisted of 10 mM ammonium acetate in IPA/ACN (90:10, v/v) with the same additives. Initially, mobile phase B was held at 50% for 1.5 min and then increased to 95% over 6.5 min where it was held for 2 min. The column was then reequilibrated for 3.5 min before the next injection. 10 µL of sample were injected by an HTC PAL autosampler (Thermo). The LC system was coupled to a Q Exactive mass spectrometer (Build 2.3 SP2) by a HESI II heated ESI source kept at 325 °C (Thermo). The inlet capillary was kept at 350 °C, sheath gas was set to 35 units, and auxiliary gas to 15 units, and the spray voltage was set to 3,000 V. Several scan functions were used to achieve optimal data acquisition for different lipid classes. For phospholipids, MS^1 data was acquired from 1–9 min at a resolving power of 35,000 with the AGC target set to $1x10^6$, mass range to 500–900 Th, and maximum injection time to 250 ms. For fatty acids and lyso species, MS^1 data was acquired from 0–3 min at a resolving power of 17,500 with the AGC target set to $5x10^5$, mass range to 220–600 Th, and maximum injection time to 100 ms. For cardiolipins, MS^1 data was acquired from 6.5–9.5 min at a resolving power of 17,500 with the AGC target set to $5x10^5$, mass range to 1320–1500 Th, and maximum injection time to 250 ms. For cytidine diacylglycerols, MS^1 data was acquired from 1–4.5 min at a resolving power of 17,500 with the AGC target set to $5x10^5$, mass range to 920–1050 Th, and maximum injection time to 250 ms. Quantitation for all of these species was performed by integrating the MS^1 peak areas of either the $[M–H]^-$ or $[M+Ac]^-$ ions. Coenzyme Q_6 and demethoxycoenzyme Q_6 were monitored from 4.7 to 5.8 min by tandem mass spectrometry using the 591.44 → 197.08 Th and 561.43 → 167.07 Th transitions at a normalized collision energy of 27 units, a resolving power of 17,500, a maximum injection time of 250 ms, and an isolation width of 1.5 Th. For some follow-

up studies MS^1 spectra were acquired from 200–1550 m/z supplemented with scheduled targeted scan modes to quantify key CoQ intermediates in their optimal polarity.

Data analysis. Peaks were automatically integrated using TraceFinder software (Thermo) and all integrations were checked manually. Missing values from undetected peaks were imputed and 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 5% of measured measured peak intensities was generated. Missing values were filled in with values drawn from this distribution at random. Total measured ion current from peaks quantified within replicate MS analyses was normalized to a corresponding WT control using a two-step procedure. First, to account for differences in cardiolipin extraction efficiency, summed cardiolipin intensities were normalized to equal the summed intensity of corresponding cardiolipin species in the WT control. All other lipid intensities were then normalized to the equal the summed intensity of non-cardiolipin species in the same control. Replicate lipid intensities from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$(strain), n = 3). Mean intensities were then normalized to WT (mean $\log_2$(Δ*gene*/WT), n = 3) and a 2-tailed t-test (homostatic) was performed to obtain *P* values.

GC-MS metabolomics. $1x10^8$ 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*.

Data analysis. All metabolomic data processing was done using a software suite developed in-house that is available at *https://github.com/coongroup*. Following data acquisition raw EI-GC/MS spectral data was deconvolved into "features" and then grouped into individual spectra containing only product ions stemming from a singular parent. Feature groups from samples and background were compared and those found in both were removed from further analyses. Compound identifications for the metabolites analyzed were assigned by comparing deconvolved high-resolution spectra against unit-resolution reference spectra present in the NIST 12 MS/EI Library as well as to authentic standards run in-house. To calculate spectral similarity between experimental and reference spectra a weighted dot product calculation was used. Metabolites lacking a confident identification were classified as 'Unidentified metabolites' and appended with a corresponding retention time to create a unique identifier. Data from fermentation and respiration cultures were normalized separately. Extracted peak heights were used to represent metabolite abundance. A median *in silico* intensity profile of all detected features was generated by calculating and storing the median intensity for each separate feature. All replicate MS analyses were normalized through a total ion current normalization procedure which scaled all measured metabolite intensities to equal the sum of all corresponding metabolite intensities in the *in silico* intensity profile.

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).

Regression analysis of phenotype changes. *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 in the Supplemental Methods.

Respiration Deficient 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 protein perturbation profile signature. Molecules which were consistently perturbed across ≥ 95% of RD Δ*gene* strains where a quantitative measurement was reported were classified as RDR-associated (776). 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 Δgene–Δgene 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 and in the Supplemental Methods.

Hfd1p 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.).

Yeast growth assays. Δ*hfd1* 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 Δ*hfd1* 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 $5x10^6$ 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.

Quantitation of CoQ and 4-HB in pABA⁻ Δhfd1 yeast cultures. 2.5x10^6 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), 1x10^8 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 protein purification. Expression and purification of recombinant Hfd1p$^{C\Delta25}$ (Hfd1p lacking its C-terminal 25 amino acids, which comprise a putative transmembrane domain) was performed essentially as described[26], with minor variations. Briefly, PIPE cloning was used to generate a pVP68K vector encoding Hfd1$^{C\Delta25}$ fused to an 8His-cytoplasmically-targeted maltose-binding protein with a linker including a tobacco etch virus protease recognition site (8His-MBP-[TEV]-Hfd1p$^{C\Delta25}$). This construct was expressed in *E. coli* (BL21(DE3)-RIPL strain) by autoinduction. Proteins were purified using cobalt IMAC resin, tobacco etch virus (TEV) cleavage, and a second subtractive IMAC purification to remove 8His-MBP. For these Hfd1p preparations, buffers contained 50 mM HEPES (pH 7.5) instead of 50 mM KH_2PO_4, 20 mM Tris-HCl (pH 7.2).

Hfd1p enzymology. Hfd1p enzyme activity assays were conducted in groups of three replicate 100 μL reactions, each containing 8 μg Hfd1p$^{C\Delta25}$, 1 mM NAD^+, and 200 μM substrate (4-HBz or (2E)-hexadecenal (Avanti 857459P)) 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). Reported activity represents the mean of four separate Hfd1p$^{C\Delta25}$ purifications.

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 ≤ 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 ($\Delta gene/WT < 0.6$ and $P < 0.05$) were considered "respiration deficient" (RD) (41 RD strains). Strains with borderline respiration growth ($0.6 \leq \Delta gene/WT < 0.8$) were categorized as "borderline respiration" (14 strains). Strains with respiration growth rates near WT or better than WT ($0.8 \leq \Delta gene/WT$) were categorized as respiration competent (RC) (111 RC strains).

For PCA, average $\log_2(\Delta gene/WT)$ 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 decreeing in RD strains (negative $\log_2(RD/RC)$). Proteins with Bonferroni-corrected $P < 1x10^{-20}$ 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 $\log_2(\Delta gene/WT)$ 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.

SUPPLEMENTAL INFORMATION

Supplemental Methods

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 datasets, and column-wise cluster order for fermentation and respiration datasets was generated using only protein fold change profiles. Column ordering was then applied to metabolite and lipid fold change datasets 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 (FDR), and reported where $P < 0.05$.

For the analysis of Δ*gene*–Δ*gene* correlations, clustering was performed on respiration protein perturbation profile correlation data and the resultant ordering was applied to Δ*gene*–Δ*gene* correlation datasets 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 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 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.

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).

mRNA quantitation. BY4742 WT, Δ*coq8,* Δ*atp2,* and Δ*yjr120w* yeast were grown overnight in 3 mL YEPD. From the overnight culture, $2.5x10^6$ 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 ΔΔ*Ct* method.

Supplemental Hfd1p biochemistry methods. *Hfd1p phylogenetics*. The amino acid sequences of the 19 known *Homo sapiens* ALDH proteins[34] and *S. cerevisiae* Hfd1p (NP_013828.1) were aligned by MUSCLE[41], analyzed by ClustalW2 Phylogeny[42], and visualized in iTOL[43].

Mass spectrometry profiling of pABA⁻ yeast cultures (WT, Δ*hfd1,* Δ*dpl1,* and Δ*coq8)*. 2.5x10^6 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, 1x10^8 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.

Supplemental Yor020w-a, Aro9p, Aro10p, and Aim18p related methods. *Proteomic analysis of* Δ*yor020w-a* yeast. 2.5x10^6 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, 1x10^8 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*. 2.5x10^6 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, 1x10^8 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.

ACKNOWLEDGEMENTS

This work was supported by a Searle Scholars Award, a Shaw Scientist Award, and by NIH grants U01GM94622, R01DK098672 and R01GM112057 (to D.J.P.); NIH Ruth L. Kirschstein NRSA F30AG043282 (to J.A.S.); ACS Division of Analytical Chemistry Award and the Society for Analytical Chemists of Pittsburgh (to A. L. R.).

REFERENCES

(1) Nunnari, J.; Suomalainen, A. *Cell* **2012**, *148*, 1145-1159.

(2) Koopman, W. J.; Willems, P. H.; Smeitink, J. A. *N Engl J Med* **2012**, *366*, 1132-1141.

(3) Vafai, S. B.; Mootha, V. K. *Nature* **2012**, *491*, 374-383.

(4) Pagliarini, D. J.; Calvo, S. E.; Chang, B.; Sheth, S. A.; Vafai, S. B.; Ong, S. E.; Walford, G. A.; Sugiana, C.; Boneh, A.; Chen, W. K.; Hill, D. E.; Vidal, M.; Evans, J. G.; Thorburn, D. R.; Carr, S. A.; Mootha, V. K. *Cell* **2008**, *134*, 112-123.

(5) Calvo, S. E.; Clauser, K. R.; Mootha, V. K. *Nucleic Acids Res* **2016**, *44*, D1251-1257.

(6) Sickmann, A.; Reinders, J.; Wagner, Y.; Joppich, C.; Zahedi, R.; Meyer, H. E.; Schonfisch, B.; Perschil, I.; Chacinska, A.; Guiard, B.; Rehling, P.; Pfanner, N.; Meisinger, C. *Proc Natl Acad Sci U S A* **2003**, *100*, 13207-13212.

(7) Green, E. D.; Guyer, M. S.; National Human Genome Research, I. *Nature* **2011**, *470*, 204-213.

(8) Pagliarini, D. J.; Rutter, J. *Genes Dev* **2013**, *27*, 2615-2627.

(9) Crane, F. L.; Hatefi, Y.; Lester, R. L.; Widmer, C. *Biochim. Biophys. Acta.* **1957**, *25*, 220-221.

(10) Morton, R. A. *Nature* **1958**, *182*, 1764-1767.

(11) Winzeler, E. A.; Shoemaker, D. D.; Astromoff, A.; Liang, H.; Anderson, K.; Andre, B.; Bangham, R.; Benito, R.; Boeke, J. D.; Bussey, H.; Chu, A. M.; Connelly, C.; Davis, K.; Dietrich, F.; Dow, S. W.; El Bakkoury, M.; Foury, F.; Friend, S. H.; Gentalen, E.; Giaever, G.; Hegemann, J. H.; Jones, T.; Laub, M.; Liao, H.; Liebundguth, N.; Lockhart, D. J.; Lucau-Danila, A.; Lussier, M.; M'Rabet, N.; Menard, P.; Mittmann, M.; Pai, C.; Rebischung, C.; Revuelta, J. L.; Riles, L.; Roberts, C. J.; Ross-MacDonald, P.; Scherens, B.; Snyder, M.; Sookhai-Mahadeo, S.; Storms, R. K.; Veronneau, S.; Voet, M.; Volckaert, G.; Ward, T. R.; Wysocki, R.; Yen, G. S.; Yu, K.; Zimmermann, K.; Philippsen, P.; Johnston, M.; Davis, R. W. *Science* **1999**, *285*, 901-906.

(12) Giaever, G.; Chu, A. M.; Ni, L.; Connelly, C.; Riles, L.; Veronneau, S.; Dow, S.; Lucau-Danila, A.; Anderson, K.; Andre, B.; Arkin, A. P.; Astromoff, A.; El-Bakkoury, M.; Bangham, R.;

Benito, R.; Brachat, S.; Campanaro, S.; Curtiss, M.; Davis, K.; Deutschbauer, A.; Entian, K. D.; Flaherty, P.; Foury, F.; Garfinkel, D. J.; Gerstein, M.; Gotte, D.; Guldener, U.; Hegemann, J. H.; Hempel, S.; Herman, Z.; Jaramillo, D. F.; Kelly, D. E.; Kelly, S. L.; Kotter, P.; LaBonte, D.; Lamb, D. C.; Lan, N.; Liang, H.; Liao, H.; Liu, L.; Luo, C.; Lussier, M.; Mao, R.; Menard, P.; Ooi, S. L.; Revuelta, J. L.; Roberts, C. J.; Rose, M.; Ross-Macdonald, P.; Scherens, B.; Schimmack, G.; Shafer, B.; Shoemaker, D. D.; Sookhai-Mahadeo, S.; Storms, R. K.; Strathern, J. N.; Valle, G.; Voet, M.; Volckaert, G.; Wang, C. Y.; Ward, T. R.; Wilhelmy, J.; Winzeler, E. A.; Yang, Y.; Yen, G.; Youngman, E.; Yu, K.; Bussey, H.; Boeke, J. D.; Snyder, M.; Philippsen, P.; Davis, R. W.; Johnston, M. *Nature* **2002**, *418*, 387-391.

(13) Steinmetz, L. M.; Scharfe, C.; Deutschbauer, A. M.; Mokranjac, D.; Herman, Z. S.; Jones, T.; Chu, A. M.; Giaever, G.; Prokisch, H.; Oefner, P. J.; Davis, R. W. *Nat Genet* **2002**, *31*, 400-404.

(14) DeRisi, J. L.; Iyer, V. R.; Brown, P. O. *Science* **1997**, *278*, 680-686.

(15) Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J.; Stoughton, R.; Armour, C. D.; Bennett, H. A.; Coffey, E.; Dai, H.; He, Y. D.; Kidd, M. J.; King, A. M.; Meyer, M. R.; Slade, D.; Lum, P. Y.; Stepaniants, S. B.; Shoemaker, D. D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S. H. *Cell* **2000**, *102*, 109-126.

(16) Kemmeren, P.; Sameith, K.; van de Pasch, L. A.; Benschop, J. J.; Lenstra, T. L.; Margaritis, T.; O'Duibhir, E.; Apweiler, E.; van Wageningen, S.; Ko, C. W.; van Heesch, S.; Kashani, M. M.; Ampatziadis-Michailidis, G.; Brok, M. O.; Brabers, N. A.; Miles, A. J.; Bouwmeester, D.; van Hooff, S. R.; van Bakel, H.; Sluiters, E.; Bakker, L. V.; Snel, B.; Lijnzaad, P.; van Leenen, D.; Groot Koerkamp, M. J.; Holstege, F. C. *Cell* **2014**, *157*, 740-752.

(17) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2014**, *13*, 339-347.

(18) Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Nat Protoc* **2015**, *10*, 701-714.

(19) Peterson, A. C.; Balloon, A. J.; Westphall, M. S.; Coon, J. J. *Anal Chem* **2014**, *86*, 10044-10051.

(20) Peterson, A. C.; Hauschild, J. P.; Quarmby, S. T.; Krumwiede, D.; Lange, O.; Lemke, R. A.; Grosse-Coosmann, F.; Horning, S.; Donohue, T. J.; Westphall, M. S.; Coon, J. J.; Griep-Raming, J. *Anal Chem* **2014**, *86*, 10036-10043.

(21) Kwiecien, N. W.; Bailey, D. J.; Rush, M. J.; Cole, J. S.; Ulbrich, A.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Anal Chem* **2015**, *87*, 8328-8335.

(22) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. *Cell* **2009**, *138*, 795-806.

(23) Casanovas, A.; Sprenger, R. R.; Tarasov, K.; Ruckerbauer, D. E.; Hannibal-Bach, H. K.; Zanghellini, J.; Jensen, O. N.; Ejsing, C. S. *Chem Biol* **2015**, *22*, 412-425.

(24) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. *Nat Methods* **2014**, *11*, 319-324.

(25) Huang da, W.; Sherman, B. T.; Lempicki, R. A. *Nat Protoc* **2009**, *4*, 44-57.

(26) Stefely, J. A.; Reidenbach, A. G.; Ulbrich, A.; Oruganty, K.; Floyd, B. J.; Jochem, A.; Saunders, J. M.; Johnson, I. E.; Minogue, C. E.; Wrobel, R. L.; Barber, G. E.; Lee, D.; Li, S.; Kannan, N.; Coon, J. J.; Bingman, C. A.; Pagliarini, D. J. *Mol Cell* **2015**, *57*, 83-94.

(27) Mourier, A.; Motori, E.; Brandt, T.; Lagouge, M.; Atanassov, I.; Galinier, A.; Rappl, G.; Brodesser, S.; Hultenby, K.; Dieterich, C.; Larsson, N. G. *J Cell Biol* **2015**, *208*, 429-442.

(28) Olson, R. E.; Bentley, R.; Aiyar, A. S.; Dialameh, G. H.; Gold, P. H.; Ramsey, V. G.; Springer, C. M. *J Biol Chem* **1963**, *238*, 3146-3148.

(29) Bentley, R. R., V. G.; Springer, C. M.; Dialameh, G. H.; Olson, R. E. *Biochem Biophys Res Commun* **1961**, *5*, 443-446.

(30) Booth, A. N. M., M. S.; Robbins, D. J.; Emerson, O. H.; Jones, F. T.; Deeds, F. *J Biol Chem* **1960**, *235*, 2649-2652.

(31) Pierrel, F.; Hamelin, O.; Douki, T.; Kieffer-Jaquinod, S.; Muhlenhoff, U.; Ozeir, M.; Lill, R.; Fontecave, M. *Chem Biol* **2010**, *17*, 449-459.

(32) Marbois, B.; Xie, L. X.; Choi, S.; Hirano, K.; Hyman, K.; Clarke, C. F. *J Biol Chem* **2010**, *285*, 27827-27838.

(33) Nakahara, K.; Ohkuni, A.; Kitamura, T.; Abe, K.; Naganuma, T.; Ohno, Y.; Zoeller, R. A.; Kihara, A. *Mol Cell* **2012**, *46*, 461-471.

(34) Jackson, B.; Brocker, C.; Thompson, D. C.; Black, W.; Vasiliou, K.; Nebert, D. W.; Vasiliou, V. *Hum Genomics* **2011**, *5*, 283-303.

(35) De Laurenzi, V.; Rogers, G. R.; Hamrock, D. J.; Marekov, L. N.; Steinert, P. M.; Compton, J. G.; Markova, N.; Rizzo, W. B. *Nat Genet* **1996**, *12*, 52-57.

(36) Hebert, A. S.; Merrill, A. E.; Stefely, J. A.; Bailey, D. J.; Wenger, C. D.; Westphall, M. S.; Pagliarini, D. J.; Coon, J. J. *Mol Cell Proteomics* **2013**, *12*, 3360-3369.

(37) Cox, J.; Mann, M. *Nat Biotechnol* **2008**, *26*, 1367-1372.

(38) Elias, J. E.; Gygi, S. P. *Nat Methods* **2007**, *4*, 207-214.

(39) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. *J Proteome Res* **2011**, *10*, 1794-1805.

(40) Cox, J.; Hein, M. Y.; Luber, C. A.; Paron, I.; Nagaraj, N.; Mann, M. *Mol Cell Proteomics* **2014**, *13*, 2513-2526.

(41) Edgar, R. C. *Nucleic Acids Res* **2004**, *32*, 1792-1797.

(42) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. *Bioinformatics* **2007**, *23*, 2947-2948.

(43) Letunic, I.; Bork, P. *Nucleic Acids Res* **2011**, *39*, W475-478.

Chapter 7

Near Complete Sequencing of the Human Proteome

Manuscript under preparation.

Richards AL, Hebert AS, Merrill AE, Kwiecien NW, Bailey DJ, Westphall MS, Coon JJ. *Near Complete Sequencing of the Human Proteome*. **2016**.

ABSTRACT

Advances in MS have transformed the scope and impact of protein characterization efforts. This was exemplified by the recent publication of draft maps of the human proteome. In a typical high throughput MS experiment, most proteins are identified from a small subset of their respective peptides. Accordingly, sequence coverage is often low (~20-25%). Unfortunately, this coverage is likely insufficient to identify the diverse proteoforms and post-translational modifications present within a sample. Utilizing a new hybrid mass spectrometer equipped with a mass filter, a collision cell, a high-field Orbitrap analyzer, and a dual cell linear ion trap analyzer (Q-OT-qIT, Orbitrap Fusion), we present strategy to achieve unprecedented proteomic depth and coverage. Our method combines this new MS technology with powerful chromatographic separations, digestion enzymes with orthogonal specificity, and MS/MS fragmentation using multiple dissociation techniques. We have applied this approach to whole proteome analyses of six human cell lines. More than 12,000 proteins have been identified from each line, with median sequence coverage greater than 65%. To our knowledge these data represent the deepest and most comprehensive proteomic analysis of the human proteome to date.

INTRODUCTION

In a typical high throughput mass spectrometry (MS) experiment, proteins of interest are digested with trypsin, separated over a reversed phase (RP) liquid chromatography (LC) gradient, and sequenced by tandem mass spectrometry (MS/MS).[1,2] MS technology has matured so much that complete proteomes of simple organisms can be analyzed in ~ 1 hour of instrument time.[3,4] State-of-the-art application of this approach can result in the identification of over 10,000 proteins from mammalian cell lines or tissues.[5-8] From multiple MS experiments covering various tissues and cell types, draft maps of the human proteome have provided evidence for the translation of over 90% of known protein-coding genes.[9,10] Though MS has approached near comprehensive identification of all expressed proteins, the majority of amino acids residues that make up these proteins are never observed. In all bottom-up proteomics studies, protein identifications are inferred from a subset of their respective peptides.

The human proteome is complex. Although the human genome contains around 20,000 protein-coding genes,[11] it is estimated that alternative splicing events, where mRNA fragments are combined in different arrangements, can yield over 100,000 unique transcripts.[12] Other alterations, including single nucleotide polymorphisms (SNPs) and alternative reading frames, can further increase proteomic complexity. Evidence of these events can be obtained from genomic and transcriptomic data, but it's currently unknown how many of these occurrences are actually translated in to proteins. Sequence coverage in a proteomics experiment is likely insufficient to fully characterize all proteoforms[13] present within a sample. Given this situation, even the current deepest proteomic datasets do not contain enough sequence data to identify all such permutations.

Trypsin is the preferred enzyme for MS-based proteomics. It cleaves C-terminal to lysine and arginine and produces peptides of length and charge distributions most amenable to MS/MS. However, even with assistance of powerful chromatographic separations, not all portions of the proteome are accessible with trypsin. *In silico* tryptic digestion of the ~20,200 reviewed, canonical proteins of the human proteome (downloaded from Uniprot on 09/16/2015) predicts 678,007 peptides of suitable size for MS detection (6 – 35 amino acids). These peptides are comprised of 7,461,806 amino acid residues.

Consequently, exclusive use of trypsin will necessarily lead to gaps in sequence coverage, as this number only represents approximately 65% of the proteome's 11,313,328 total amino acid residues.[14,15]

Digestion with other enzymes in addition to trypsin has increased the amino acid coverage of individual proteins, phosphorylation sites,[16,17] and whole proteomes.[18-22] MacCoss *et al.* combined tryptic digestion with the non-specific proteases subtilisin and elastase to detect novel PTMs in simple protein mixtures.[23] Heck *et al.* have demonstrated the utility of the multi-enzyme approach for mapping phosphorylation sites. Of the 18,430 unique phosphosites catalogued in Jurkat cells following digestion with trypsin, LysC, AspN, GluC and chymotrypsin, only 27% were identified by more than one enzyme.[24] The Mann laboratory analyzed HeLa peptides following digestion with trypsin, LysC and GluC, identifying 10,255 proteins and achieving nearly complete overlap of expressed transcripts collected from paired RNA-Seq data.[5] Our own lab showed that digestion of the yeast proteome with LysC, ArgC, AspN and GluC in addition to trypsin nearly doubled sequence coverage, from 24.5% average sequence coverage with trypsin alone, to 43.4% with all five proteases.[20] Similar increases were seen by Mirzaei *et al*. for the HeLa proteome, where digestion of HeLa lysate with a combination of seven different proteases (single, double and triple digestion combinations of trypsin, chymotrypsin, elastase, GluC, LysC, AspN and ArgC), coupled with extensive fractionation, identified ~8,500 proteins with >40% mean sequence coverage.[19]

Despite the promise of the multi-protease strategy for increasing sequence coverage, no study has approached near comprehensive identification of all residues across an entire proteome. We reason that modern fast-scanning instruments and optimized chromatographic separation methods, when combined with extensive fractionation, should identify the majority of amino acid residues across all detected proteins. To generate deep sequencing data, we investigated six human cell lines – hES1, HeLa S3, GM12878, K562, HepG2 and HUVEC. Each cell line was digested with six specific proteases (separately, trypsin, LysC, LysN, chymotrypsin, GluC and AspN) and fractionated via high pH reversed-phase (RP) chromatography. The resultant peptides were analyzed on an Orbitrap hybrid mass spectrometer (Orbitrap Fusion, Orbitrap Lumos). Peptides derived from all enzymes were sequenced using higher-energy collisional dissociation (HCD), and a subset were sequenced with electron transfer dissociation (ETD) and collisionally-activated

dissociation (CAD). We collected approximately 19,366,000 high resolution mass spectra and 161,641,087 ion trap MS/MS scans from 2,395 raw files. Each cell line yielded more than 12,000 protein identifications and a median sequence coverage greater than 65%. To our knowledge, these data represent the deepest and most comprehensive proteomic analysis of the human proteome to date.

RESULTS

Deep sequencing of the yeast proteome. We first tested our deep-sequencing strategy on yeast. It is estimated that approximately 4,500 of yeast's 6,600 open reading frames (ORFs) are expressed under standard laboratory growth conditions. Our previous yeast multi-enzyme study coupled offline fractionation and digestion with five enzymes – trypsin, LysC, ArgC, GluC and AspN - with either CAD or ETD analysis using a decision-tree algorithm optimized to increase the likelihood that fragment spectra would be matched to sequence. From the combined datasets, 3,918 proteins were identified with 43.4% average sequence coverage. We were interested in revisiting this experiment using a faster scanning mass spectrometry system, with the goal of boosting sequence coverage beyond our previous study. Here, yeast was digested with six enzymes – trypsin, LysC, which cleaves C-terminal to lysine, LysN, which cleaves N-terminal to lysine, GluC, which cleaves C-terminal to glutamic acid, AspN, which cleaves N-terminal to aspartic acid, and chymotrypsin, which cleaves C-terminal to several hydrophobic residues, including phenylalanine, tyrosine, tryptophan and leucine – separated into 20 fractions, and each fraction was sequenced with both HCD and ETD.

Table 1 summarizes the general results of these experiments. Yeast digested with trypsin outperformed all other datasets, identifying 127,123 unique peptides, followed by LysC (85,434), chymotrypsin (76,406), LysN (70,435), GluC (59,583) and AspN (57,081). Batched analysis of these digests provided 426,813 unique peptides, a 236% increase compared to analysis with trypsin alone. Combined analysis of all digestion and fragmentation conditions identified 5,048 proteins, which were identified with 79.4% median sequence coverage.

Table 1. Multi-enzyme analysis of the yeast proteome. Summary of PSM, unique peptide, protein identifications and sequence coverage for the yeast proteome following separate digestion with six different proteases.

PROTEASE	Peptide spectral matches (PSMS)	Unique peptides	Proteins	Median sequence coverage
trypsin R,K|	579,350	127,123	4,995	52.47%
LysC K|	471,282	85,434	4,919	45.56%
LysN |K	327,823	70,435	4,679	37.50%
chymotrypsin F,Y,W,L|	233,147	76,406	4,457	34.12%
GluC E|	323,927	59,583	4,503	32.28%
AspN |D	303,947	57,081	4,445	26.29%
ALL	**2,263,129**	**426,813**	**5,048**	**79.43%**

We next examined the number of non-redundant amino acid residues identified from these unique peptides. The 5,048 proteins were composed of 2,624,302 non-redundant amino acid residues. Using our multi-enzyme approach, we found evidence for 1,891,786 of these residues. The number of residues identified by each digestion and fragmentation method is plotted in **Figure 1*A***. As expected, trypsin identified the highest number of non-redundant amino acid residues, with 1,302,591. Summation of the non-redundant amino acid residues across all six enzymatic datasets increased this number to 1,891,786, a 45% increase compared to analysis with trypsin alone.

The overlap of non-redundant amino acid residue identifications between yeast tryptic peptides and peptides derived from all other enzymatic digestions is presented in **Figure 1*B*.** Of the 1,891,786 total non-redundant amino acid identifications, 637,915 of these amino acid residues were not present in the trypsin dataset. To visualize the impact the addition of alternate enzymes has on sequence coverage, we plotted the sequence coverage of all yeast proteins identified in our dataset (**Figure 1*C***). Proteins are rank-ordered according to increasing percent tryptic sequence coverage (red). The sequence coverage obtained for each protein from combining all enzymatic digests is plotted in black, and includes 53 proteins not present in the trypsin dataset. The median protein sequence coverage from tryptic peptides is 52.5%, which is increased to 79.4% through the combination of all six enzymes. On average, each protein saw a 21.1% increase in sequence coverage following digestion with six proteases compared with digestion with trypsin alone.

Figure 2 further reveals the impact of digestion with multiple enzymes on protein sequence coverage. Here, sequence coverage is plotted according to protein expression levels obtained by Ghaemmaghami *et al.* from TAP-tagging and GFP experiments (**Figure 2*A*)**.[25] For proteins present at 50,000 copies per cell or greater, the average sequence coverage of the combined enzymatic digestions was over 90% (black line). Even proteins with low expression levels (<100 copies per cell) still achieved ~60% sequence coverage following digestion with all six enzymes. This is illustrated for protein YLR0573, which is expressed at 72 copies per cell. The sequence coverage obtained with each enzyme is illustrated in **Figure 2*B***; trypsin identified 53.2% of the amino acids present in the protein, LysC identified 52.2%, LysN identified 43.0%,

Figure 1. The use of multiple proteases increases sequence coverage of the yeast proteome. (***A***) Number of non-redundant amino acids identified in each digestion and fragmentation method for the yeast proteome. The 5,048 proteins identified express 2,624,302 non-redundant amino acids (***black bar***). Of those, 1,891,786 were identified by our dataset (***dark grey bar***). (***B***) Overlap in non-redundant amino acid identifications between the trypsin dataset and all other datasets. (***C***) All identified yeast proteins are plotted according to sequence coverage. Black dots represent the total sequence coverage obtained from all enzymes. The red dots represent the sequence coverage obtained from trypsin alone.

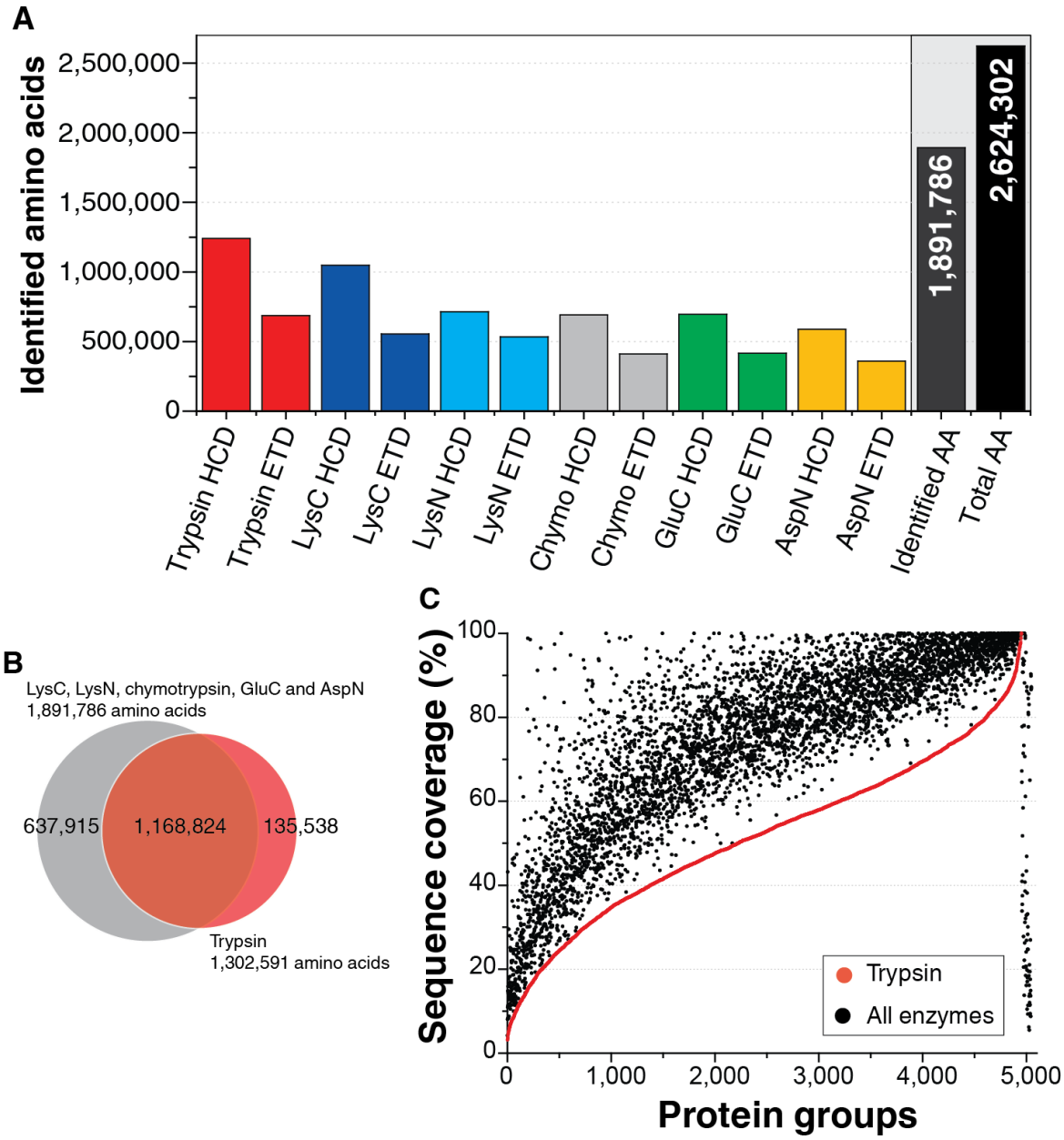
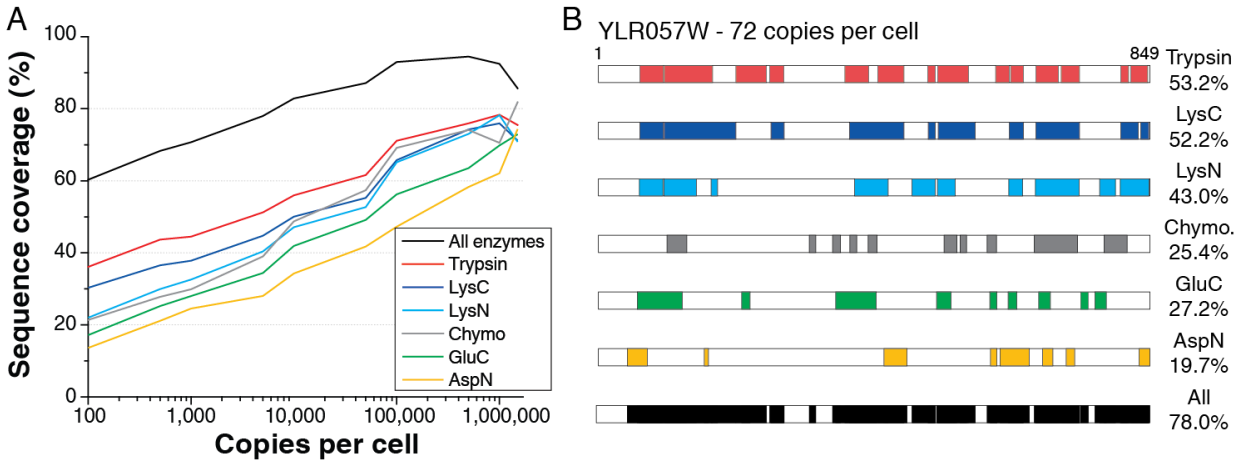


Figure 2. Sequence coverage scales with protein abundance. (***A***) Sequence coverage for all enzymes combined (***black line***) and individual enzymes (***colored lines***) as a function of protein abundance. (***B***) Sequence coverage of YLR057W, a protein expressed by yeast at 72 copies per cell. Colored sections show the portions identified by individual enzymes, with the black sections representing the combined sequence coverage from all enzymes.



chymotrypsin identified 25.4%, GluC identified 27.2%, and AspN identified 19.7%. The combined sequence coverage was 78.0%, with each enzyme identifying unique regions of the protein.

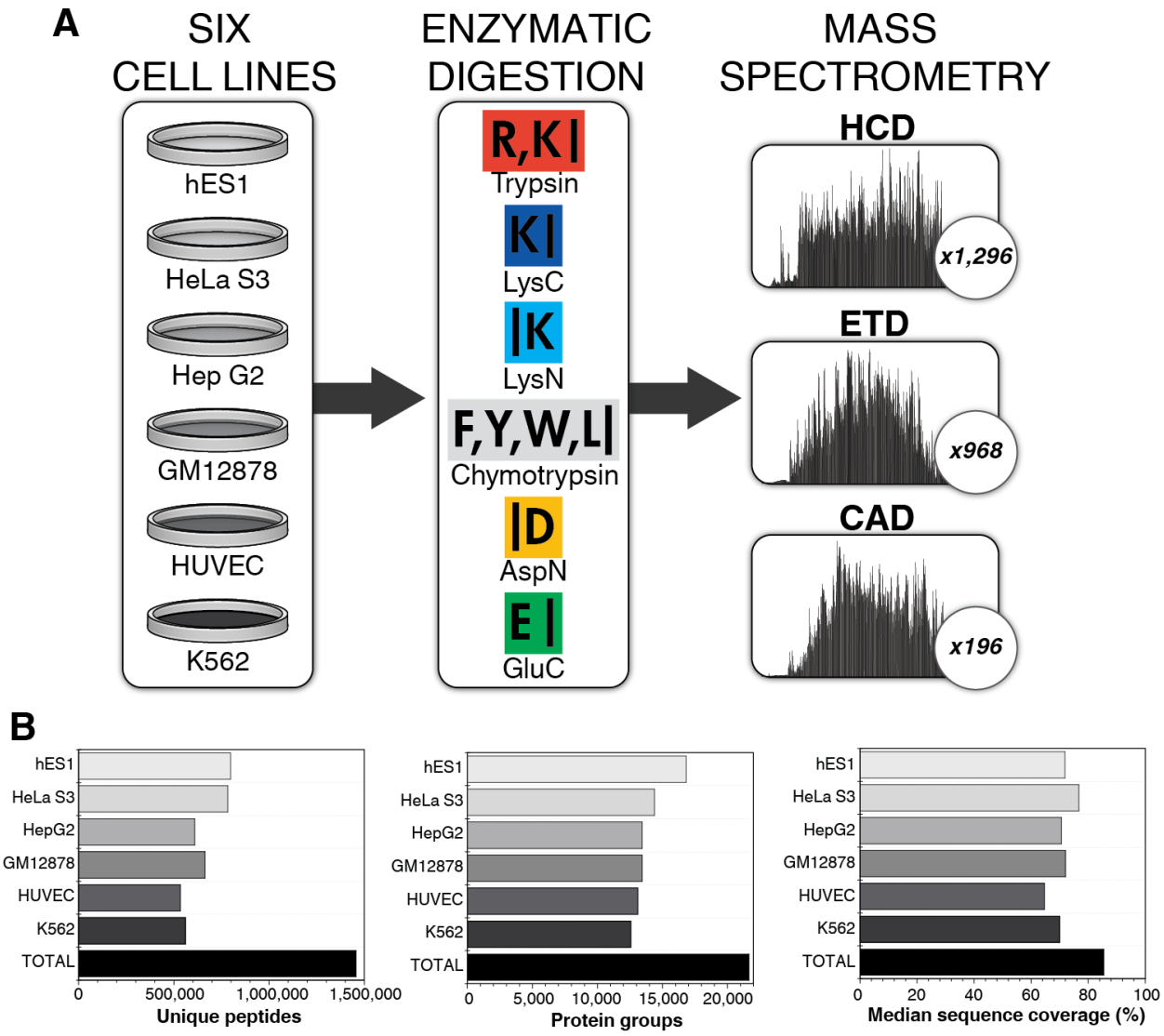
Deep sequencing of human cell lines. We next sought to apply this deep sequencing strategy to a more complex sample: the human proteome. Six diverse human cell lines were selected for comprehensive analysis; hES1, an embryonic stem cell line, HeLa S3, derived from cervical carcinoma, HepG2, from liver carcinoma, GM12878, a blood lymphoblastoid line, K562, from chronic myeloid leukemia, and HUVEC, from umbilical vein epithelial cells. Aliquots of each cell line were separately digested with trypsin, LysC, LysN, chymotrypsin, GluC or AspN. To maximize identifications, peptides were heavily fractionated (24 – 80 fractions), and analyzed on either an Orbitrap Fusion (Thermo) or an Orbitrap Lumos (Thermo). Peptides were sequenced using collisional activation methods (CAD, HCD) and electron-transfer dissociation (ETD) (**Figure 3*A***). All fractions were analyzed via HCD. LysC, LysN, AspN, GluC and chymotrypsin fractions were analyzed using ETD, and chymotrypsin fractions were analyzed using CAD.

An overview of the results obtained for each cell line is plotted in **Figure 3*B***. For all cell lines, we obtained very deep proteomic coverage, identifying an average of 661,205 unique peptides per cell line. These peptides correspond to an average of 13,963±1,394 proteins per cell line. The highest number of protein identifications was obtained for the hES1 cell line (16,832), followed by HeLa S3 (14,393), HepG2 (13,435), GM12878 (13,447), HUVEC (13,099) and K562 (12,569).

Analysis of all cell lines together identified 20,236,295 PSMs, 1,457,329 unique peptides and 21,669 protein groups at a peptide and protein false FDR of 1%. These proteins correspond to 13,072 genes, which comprise 65% of predicted protein-coding genes. The average number of unique peptides per protein was 105 (median = 67). No proteins were identified from only one unique peptide; only 393 proteins, or 1.8% of the total proteins, were identified by ten or fewer unique peptides. These 21,669 protein groups have a median sequence coverage of 85.5%.

Across all HCD experiments, 963, 084 unique peptides were identified. These peptides map to 18,215 protein groups, which were identified with 77.7% median sequence coverage. ETD analysis identified

Figure 3. Deep sequencing workflow and results. (*A*) Six human cell lines were separately digested with six enzymes, fractionated and analyzed with HCD, ETD or CAD using an Orbitrap mass spectrometer. (***B***) Summary of the unique peptides, proteins and median sequence coverage obtained for all six human cell lines.



646,566 unique peptides and 14,993 protein groups with 62.8% median sequence coverage, while 242,445 unique peptides and 12,524 protein groups were identified with 41.8% median sequence coverage from the chymotrypsin CAD fractions. **Table 2** summarizes the results from these experiments.

Despite being analyzed with only HCD, the trypsin dataset produced the largest number of unique peptides (412,859), followed by chymotrypsin (340,617), LysC (231,594), LysC (210,721), GluC (196,202) and AspN (173,326). The identifications generated by each enzyme are summarized in **Table 3**. The performance of chymotrypsin can be partially explained by it being the only dataset analyzed by all three fragmentation methods. Notably, each enzyme dataset identified over 10,000 protein groups. Trypsin produced the largest number, identifying 17,960 proteins with 58.9% median sequence coverage from 305 raw files. 13,639 protein groups were identified with 52.5% median sequence coverage from the chymotrypsin dataset, followed by 13,517 proteins with 47.3% sequence coverage from the LysC dataset, 12,742 proteins with 45.6% sequence coverage with LysN, 12,005 proteins with 38.1% median sequence coverage with GluC, and 11,588 proteins with 34.1% median sequence coverage from AspN.

The 21,669 human proteins present in our dataset were composed of 12,130,151 non-redundant amino acid residues. Our multi-enzyme approach was able to confirm 9,083,367, or 74.8% of these residues. In total, the unique peptides identifies in the combined tryptic datasets from all cell lines produced 5,736,116 non-redundant amino acids. The number of residues identified by each digestion and fragmentation method across all cell lines is plotted in **Figure 4*A***. The overlap between the tryptic amino acid residues and amino acid residues identified by all other enzymes is plotted in **Figure 4*B*.** The addition of alternate enzymes to trypsin added 3,622,870 amino acids which were not present in the tryptic dataset. The impact of these additional amino acids on protein sequence coverage can be seen in **Figure 4*C***. The sequence coverage of human proteins identified using trypsin is plotted in red, with sequence coverage obtained for each protein from the combination of all enzymatic datasets plotted in black. Median sequence coverage for proteins identified following tryptic digestion is 58.9%; median sequence coverage increases to 85.5% following digestion with all six enzymes. Each protein saw an average increase in sequence coverage of 20.1% from digestion with six enzymes compared to digestion with trypsin alone.

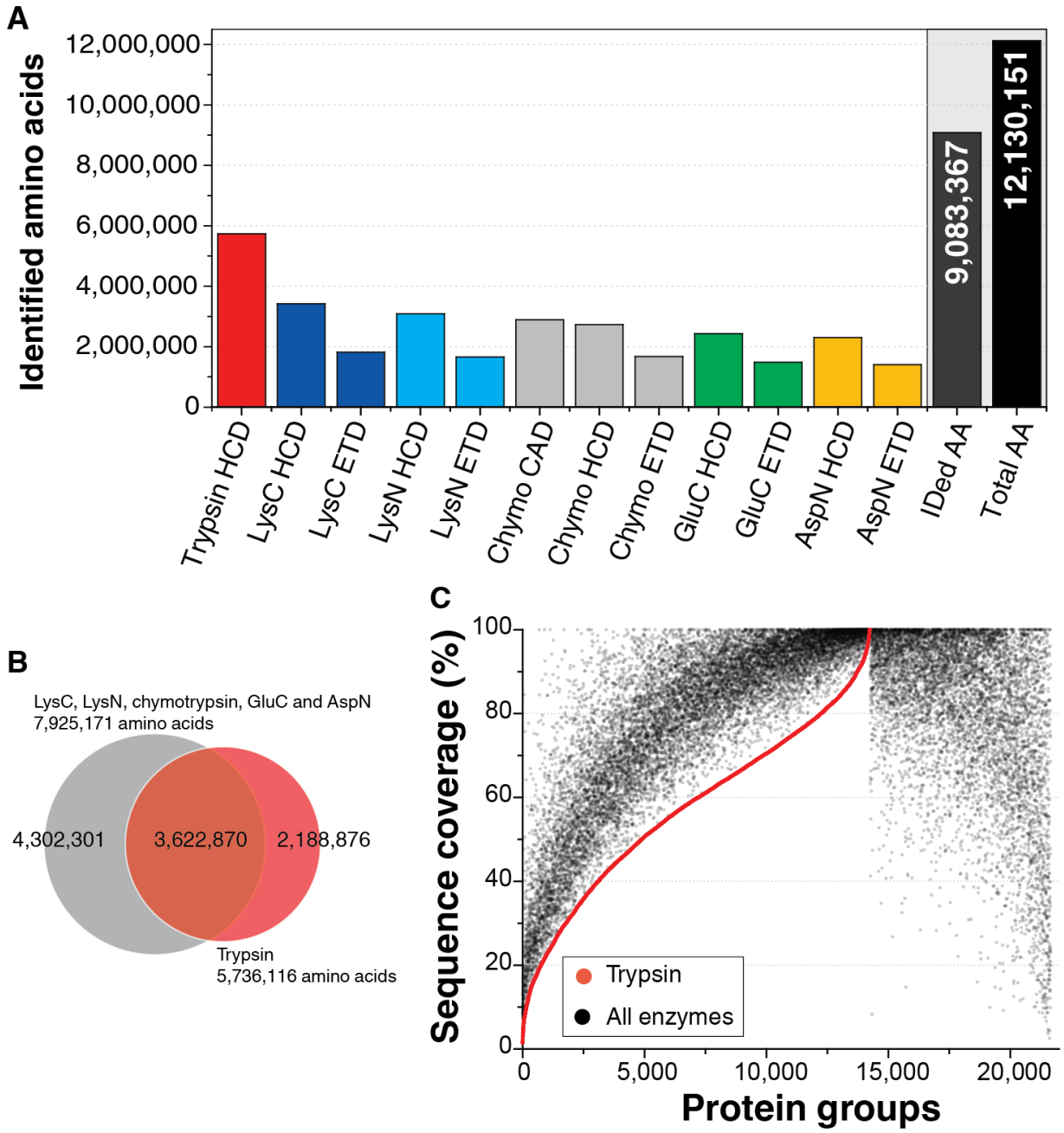
Table 2. Summary of unique peptides, proteins and the median sequence coverage obtained across all six human cell lines following analysis with HCD, ETD and CAD. Note, all enzymatic digests were analyzed with HCD, all enzymatic digests except trypsin were analyzed with ETD. Only the chymotrypsin digests were analyzed with CAD.

	LC-MS/MS experiments	Unique peptides	Proteins	Med. sequence coverage
HCD	1,296	1,215,239	20,914	81.2 %
ETD	968	646,556	15,121	61.0 %
CAD	196	233,250	12,327	40.5 %

Table 3. Summary of unique peptides, proteins and the median sequence coverage obtained across six human cell lines following digestion with the specified enzyme. Note, only HCD was performed on the trypsin dataset. HCD and ETD were performed for all other enzymes. Chymotrypsin was additionally analyzed with CAD.

	LC-MS/MS experiments	Unique peptides	Proteins	Med. sequence coverage
Trypsin	305	412,859	17,690	58.9 %
LysC	408	210,721	13,517	47.3 %
LysN	406	231,594	12,742	45.6 %
Chymo	584	340,617	13,639	52.5 %
GluC	381	196,202	12,005	38.1 %
AspN	381	173,326	11,588	34.1 %
TOTAL	**2,460**	**1,443,636**	**21,616**	**85.5 %**

Figure 4. The use of multiple proteases increases sequence coverage of the human proteome. (***A***) Number of non-redundant amino acids identified in each digestion and fragmentation method for the human proteome. The 21,669 proteins identified express 12,130,151 non-redundant amino acids (***black bar***). Of those, 9,083,367 were identified by our dataset (***dark grey bar***). (***B***) Overlap in non-redundant amino acid identifications across all human cell lines between the trypsin dataset and all other datasets. (***C***) All identified human proteins are plotted according to sequence coverage. Black dots represent the total sequence coverage obtained from all enzymes. The red dots represent the sequence coverage obtained from trypsin alone.



A comparison of HCD and ETD identifications for AspN, GluC, chymotrypsin, LysN and LysC digests is presented in **Figure 5**. **Figure 5*A*** shows the distribution in sequence coverage for proteins analyzed with HCD or ETD for the specified enzymes across all cell lines. HCD provided the greatest median sequence coverage, ranging from 29.6% for the AspN digested peptides to 43.0% for the LysC digested peptides. Sequence coverage for peptides analyzed with ETD ranged from 19.5% for peptides digested with AspN to 28.5% for the chymotrypsin datasets. Although the chymotryptic peptides had the highest sequence coverage of all the enzymatic digests analyzed with ETD, this combination identified fewer protein groups (10,178) than most other enzymatic digests analyzed with ETD (10,065). Encouragingly, all digestion and fragmentation combinations identified over 10,000 protein groups.

The addition of ETD added a number of new non-redundant amino acid residues. 8,582,243 non-redundant amino acid residues were identified from the HCD datasets. 4,683,174 non-redundant amino acid residues were identified in the ETD datasets. Note, tryptic peptides were not analyzed with ETD. 3,475,777 of these amino acid residues were identified by both fragmentation methods, meaning HCD contributed 5,106,466 unique amino acid residues, and ETD contributed 1,207,397 unique amino acid residues across all datasets (**Figure 5*B***). The overlap between peptides identified with HCD and peptides identified with ETD is illustrated in **Figure 5*C*** for the HeLa S3 cell line. Across all enzymes, approximately 40% of peptides identified by ETD were not present in the HCD datasets. The addition of these new peptides translates to an approximately 5% increase in sequence coverage compared with analysis by HCD alone (**Figure 5*D***).

Sequence coverage. The large number of unique peptides identifies in this dataset allowed us to obtain very deep sequence coverage of the human proteome. Sequence coverage for the 21,669 proteins identified across all six cell lines is plotted in the histogram in **Figure 6*A***. One quarter of the identified 21,669 proteins, 5,499, had sequence coverage of 95% or greater, including 961 proteins with 100% sequence coverage. Over half of all proteins, 12,998, were identified with >80% sequence coverage. Only 263

Figure 5. Human proteome identifications using HCD and ETD. (***A***) Sequence coverage and number of proteins identified by the specified enzyme following analysis with either HCD or ETD. (***B***) Overlap in non-redundant amino acid identifications across all human cell lines between HCD (***blue***) and ETD (***grey***). (***C***) Overlap in unique peptides identified between HCD (***blue***) and HCD (***grey***) for the HeLa S3 dataset. (***D***) Increase in median sequence coverage of the HeLa S3 proteome for each enzyme following the addition of the ETD dataset to the HCD dataset.

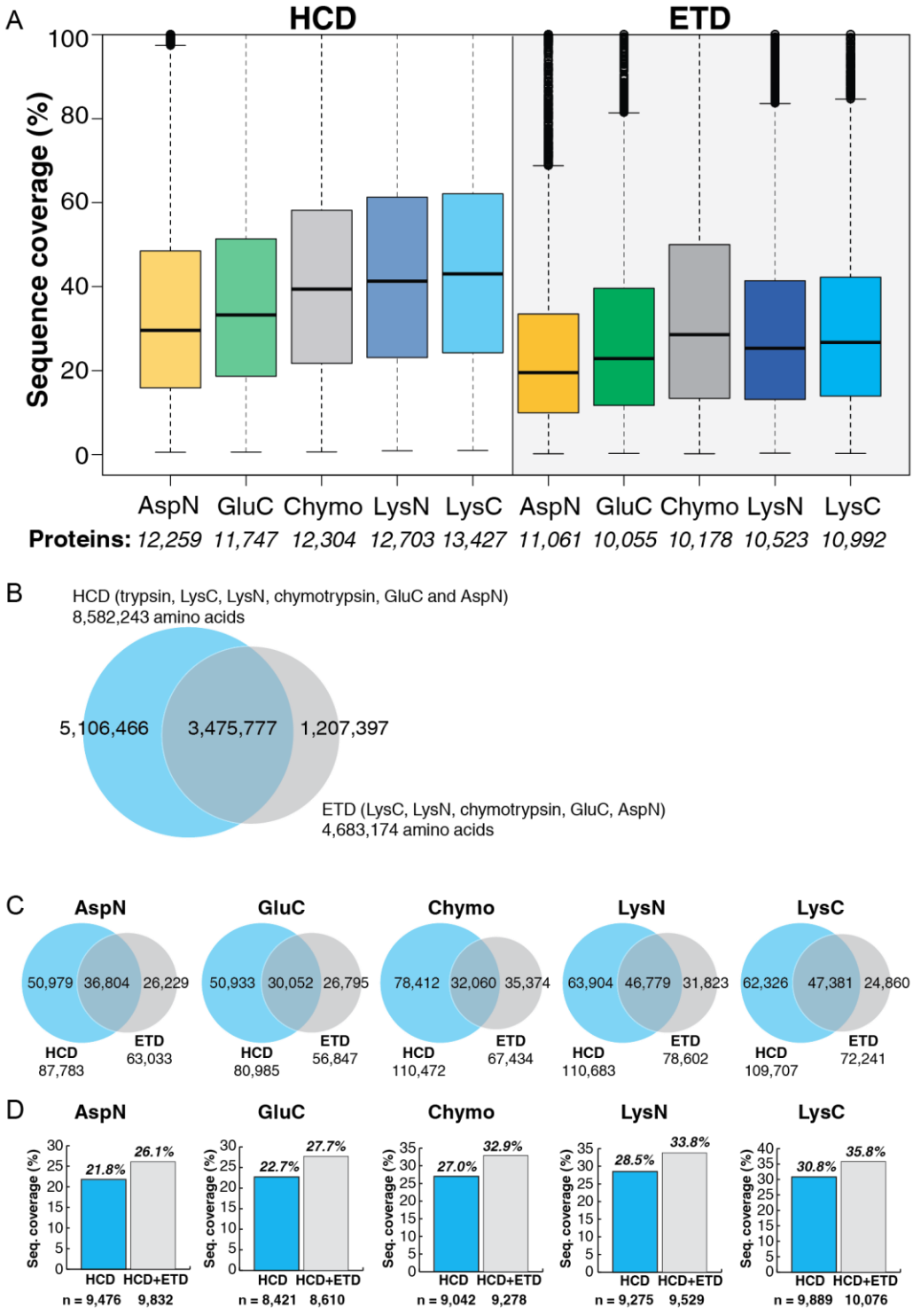
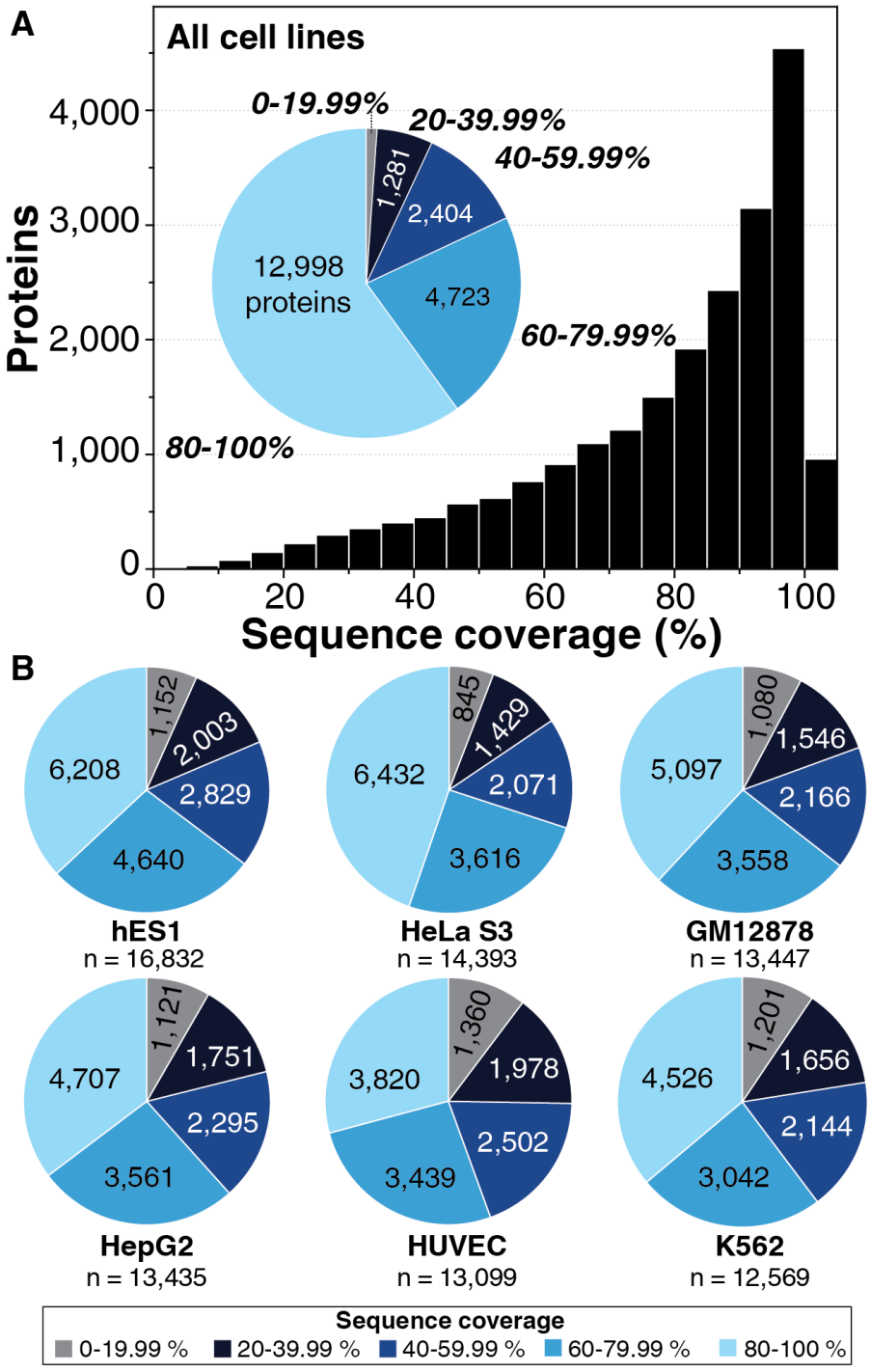


Figure 6. Protein sequence coverage. (***A***) Sequence coverage obtained for the 21,669 proteins from the combined analysis of six human cell lines. 12,998, or 65% of identified proteins, had greater than 80% sequence coverage. Only ~230 proteins were identified with less than 20% sequence coverage (***inset***). (***B***) Distribution of sequence coverage of each cell line.

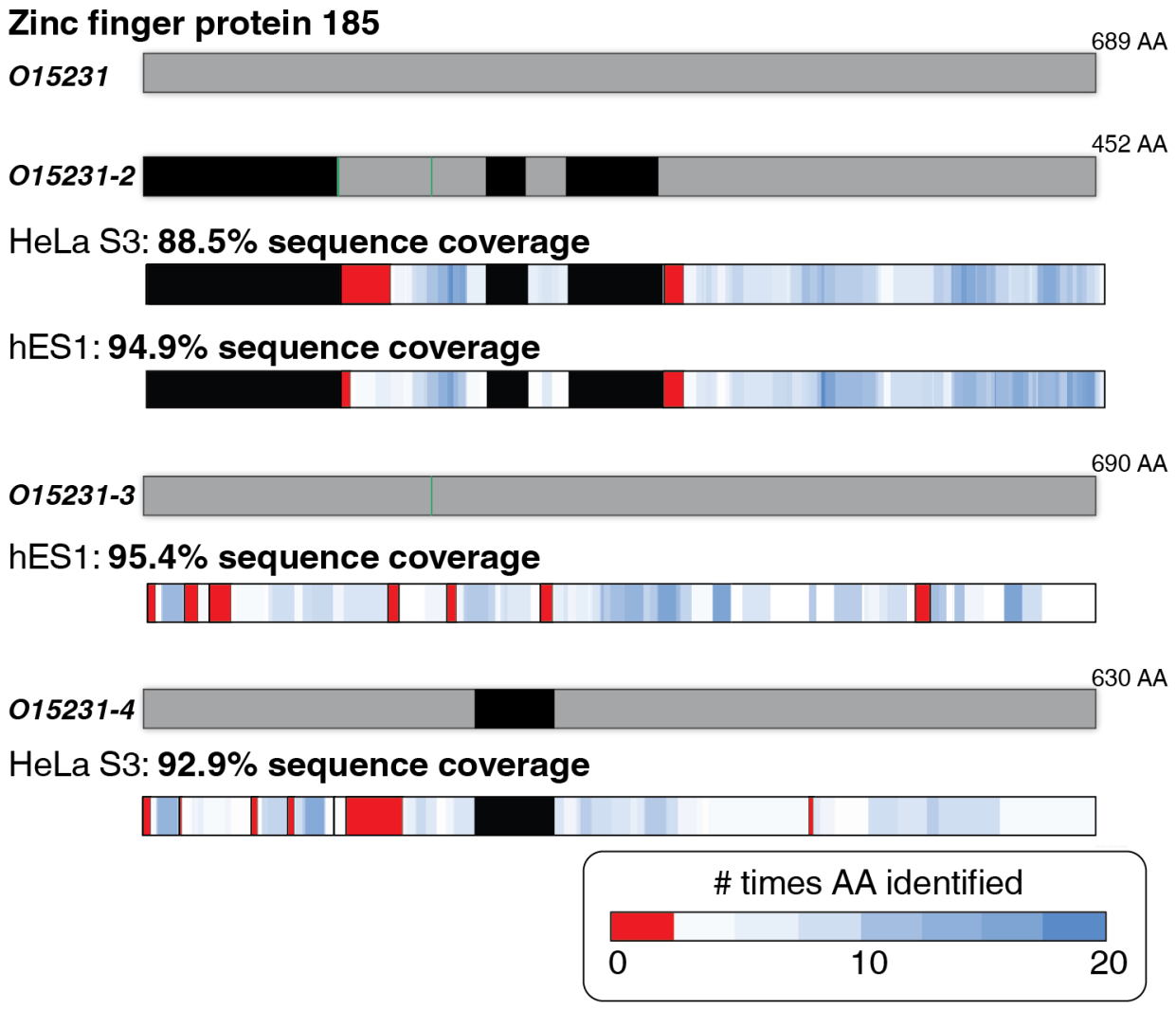


proteins were identified with sequence coverage less than 20%. On average, these 263 proteins with low sequence coverage were identified by 19 unique peptides.

Within each cell line, we identified over 500,000 unique peptides. Analysis of the hES1 cell line produced the largest number of unique peptides (797,583; median sequence coverage 71.8%), followed by HeLa S3 (782,604; median sequence coverage 76.7%), GM12878 (662,418; median sequence coverage 72.0%), HepG2 (609,683; median sequence coverage 70.6%), K562 (560,143; median sequence coverage 70.0%) and HUVEC (534,301; median sequence coverage 60.6%). Due to the large number of unique peptides, high sequence coverage was obtained for each cell line. In all cases, less than a quarter of the identified proteins had sequence coverage below 40%. A significant fraction of proteins in all cell lines (>30%) had greater than 80% sequence coverage (**Figure 6*B***).

Isoform identification. We can use this high sequence coverage to distinguish between protein isoforms. Present in the dataset are three isoforms of zinc finger protein 185, whose canonical sequence is 689 amino acid residues in length. Each isoform is either missing amino acid residues or has different or additional amino acids compared to the canonical sequence. From HeLa S3 cells, we identified 82 unique peptides derived from isoform 2. Multiple peptides were identified from each digestion condition – 24 unique peptides were identified following digestion with trypsin (67.5% sequence coverage), 11 unique peptides were identified following digestion with LysC (30.0% sequence coverage), 8 unique peptides were identified following digestion with LysN (23.7% sequence coverage), 15 unique peptides were identified following digestion with chymotrypsin (57.6% sequence coverage), 24 unique peptides were identified following digestion with GluC (63.0% sequence coverage), and 8 unique peptides were identified following digestion with AspN (31.6% sequence coverage) - providing 88.0% sequence coverage. Isoform 2 was also identified in the hES1 dataset, with 94.9% sequence coverage. Owing to this high sequence coverage, we identify several overlapping regions of the protein, which allows us to unequivocally identify regions of the protein that differ from their canonical sequence. This concept is demonstrated in **Figure 7**, where the number of times an expressed amino acid was identified is illustrated. The darker blue regions represent a

Figure 7. Isoform detection. High sequence coverage was used to distinguish between protein isoforms of zinc finger protein 185. Isoform 2 is missing amino acid residues 1-140, 246-275 and 305-372 (***black sections***) that are present in the canonical isoform. Additionally, amino acid residues YKKL at positions 141-144 in the canonical form are present as amino acid residues MQRQ and there is an additional amino acid residue (S) at position 206 as compared to the canonical sequence (***green sections***). Isoform 2 was expressed, and identified with high sequence coverage, in HeLa S3 and hES1 cell lines. Isoform 3, which expresses an additional amino acid residue (S) at position 206 as compared to the canonical sequence (***green section***) was identified in the hES1 cell line, and isoform 4, which is missing amino acids residues 246-304 (***black sections***) that are present in the canonical form, is expressed in the HeLa S3 cell line.



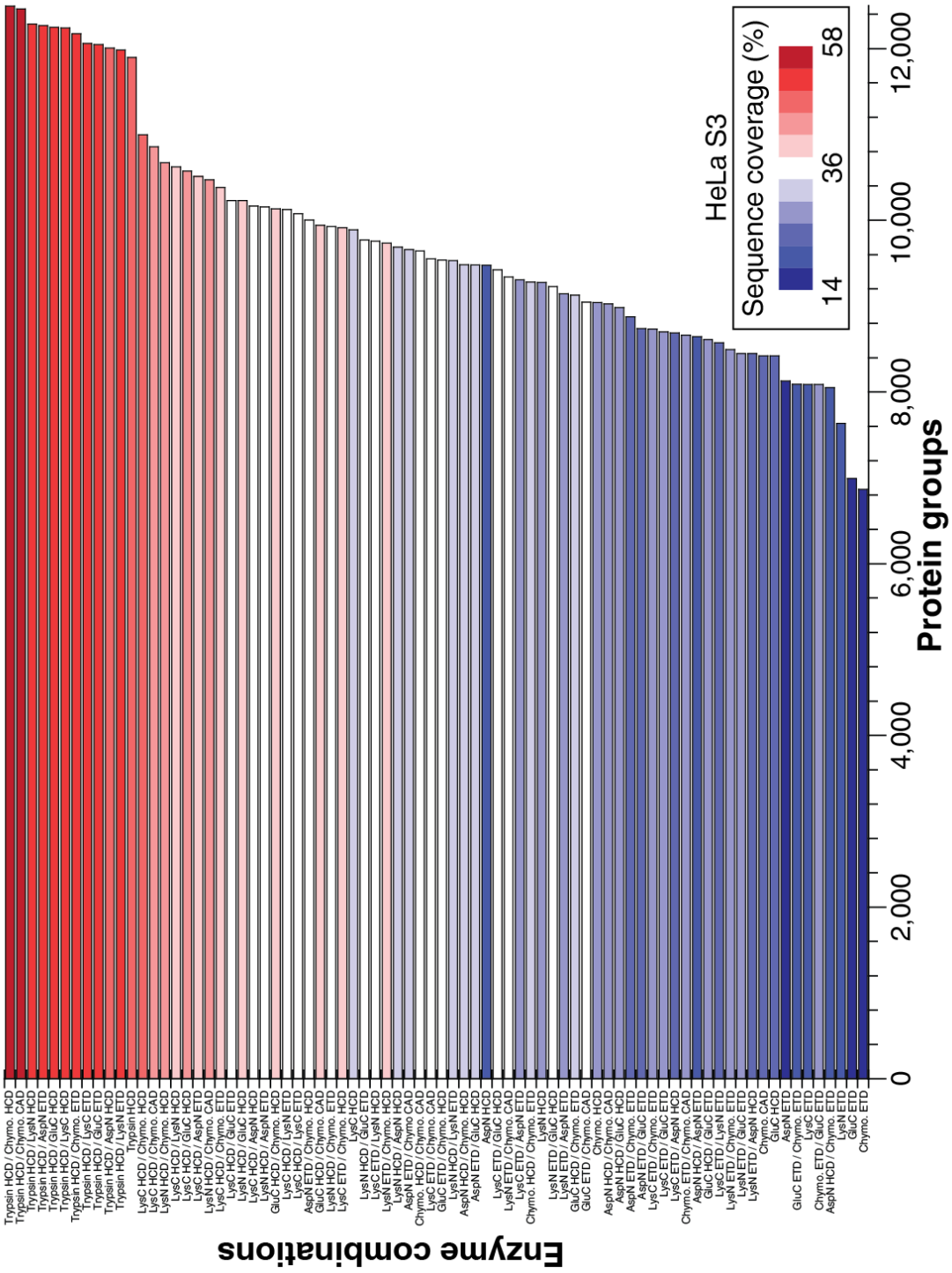
high number of identifications for a particular amino acid, while the red regions represent segments of the protein that were not accounted for in our dataset. Our dataset also contains evidence for two other isoforms of zinc finger protein 185, hES1 also expresses isoform 3 (identified with 95.4% sequence coverage), while HeLa S3 also expresses isoform 4 (identified with 92.9% sequence coverage).

DISCUSSION AND CONCLUSIONS

The large number of unique peptides identified in this study was key to deep protein sequence coverage. To our knowledge, the over 1.4 million unique peptides presented here is the largest collection of unique peptides from a MS experiment. The most complete previous analyses of the human proteome identified significantly fewer unique peptides. For example, Kim et al. found evidence for 293,000 unique tryptic peptides from more than 2,000 LC-MS/MS experiments on thirty diverse human tissue samples, resulting in a median sequence coverage of ~28% for products from 17,924 genes. From our 305 analyses of tryptic digestions from the six cell lines, we identified 412,859 unique peptides. These peptides mapped to 17,960 proteins, providing a median sequence coverage of 58.9%. Across 16,587 LC-MS/MS analyses of human cell lines, tissues and body fluids, Wilhelm *et al.* identified 569,233 unique peptides. The majority of these peptides are tryptic, although the dataset does include peptides resulting from digestion with LysC, GluC and chymotrypsin. In a fraction of the analyses, we were able to identify a similar number of tryptic peptides, alluding to the quality of our dataset.

Finally, the analyses presented here represent a significant time commitment that may not be feasible for a typical high-throughput proteomics experiment. With this in mind, we sought to determine the combination of two digestion and fragmentation methods that would provide the highest number of identified proteins and the greatest protein sequence coverage. A comparison of enzyme performance is presented in **Figure 8.** This data was taken from the analysis of HeLa S3 cells; for two enzyme combinations, each bar represents 60 LC-MS/MS analyses. For single enzyme digests, each bar represents 30 LC-MS/MS analyses. The color of the bar represents the obtained sequence coverage, which ranged from 14% (dark blue) to 55.2% (dark red). The combination of trypsin and any other enzyme provided high

Figure 8. Enzyme performance. Comparison of different digestion conditions and fragmentation methods for the HeLa S3 proteome. Results are ordered according to protein group identifications. Each bar is colored according to the median sequence coverage of the identified proteins, which ranged from 14% (dark blue) to 55.2% (dark red).



sequence coverage. Trypsin combined with chymotrypsin provided both the highest number of proteins identified and the highest sequence coverage (12,493 proteins with 55.2% median sequence coverage, HCD; 12,463 proteins with 54.3% median sequence coverage). As chymotrypsin cleaves after a number of hydrophobic residues, its performance is orthogonal to trypsin. In many cases, digestion with trypsin alone outperformed digestion with two alternate enzymes.

Here we analyzed six human cell lines at a proteomic depth previously unachieved by mass spectrometry. This coverage relied on digestion with multiple enzymes, extensive fractionation and multiple fragmentation techniques. In total, we provide evidence for 21,669 proteins analyzed to a depth of 85.5% sequence coverage. This dataset provides enormous potential for the large-scale identification of alternative translation products, which may not be present in tryptic datasets.

EXPERIMENTAL PROCEDURES

Yeast Culture and Lysis. *Saccharomyces cerevisiae* strain BY4741 was grown in yeast extract peptone dextrose media (1% yeast extract, 2% peptone, 2% dextrose). Four liters of media was divided between four two-liter flasks and inoculated with a starter culture (OD_{600} = 1.17). Cells were allowed to propagate for ~18 h to an average OD_{600} of 1.31. The cells were harvested by centrifugation at 5000 rpm for 5 min, the supernatant was decanted, and the pellets were resuspended in chilled NanoPure water. The cells were washed two more times and centrifuged for the final pelleting at 5000 rpm for 10 min. A pellet corresponding to 5% of the total cells grown, was resuspended in lysis buffer containing 8 M urea, 50 mM tris (pH 8), 75 mM sodium chloride, 100 mM sodium butyrate, protease (Roche) and phosphatase inhibitor tablet (Roche). Yeast cells were lysed by glass bead milling (Retsch). Briefly, 2 ml of acid washed glass beads were combined with 2.5 ml of resuspended yeast cells in a stainless steel container and shaken 8 times at 30 Hz for 4 min with a 1 min rest in between. . Lysate protein concentration was measured by BCA (Thermo Pierce).

Cell culture and lysis. HeLa S3 cells (ATCC CCL-22; ATCC, Manassas, VA) were grown at 37°C with 5% CO_2 in F-12K medium (ATCC) supplemented with 10% fetal bovine serum (FBS) and antibiotics. HUVEC cells (Lonza CC-2517; Lonza, Walkersville, MD) were grown at 37°C with 5% CO_2 in Endothelial Growth Media (EGM) supplemented with EGM Complete Media (Lonza) and antibiotics. HepG2 cells (ATCC HB-8065; ATCC) were grown at 37°C with 5% CO_2 in Eagle's Minimum Essential Medium (EMEM, ATCC) supplemented with 10% FBS and antibiotics. K562 cells (ATCC CCL-243; ATCC) were grown at 37°C with 5% CO_2 in Iscove's Modified Dulbecco's Medium (IMDM, ATCC) supplemented with 10% FBS and antibiotics. GM12878 cells (GM12878 K Order 104598; Coriell Institute for Medical Research, Camden, NJ) supplemented with 15% FBS and RPMI-1640 medium (Sigma Aldrich). Cells were harvested at >70% confluency through centrifugation at 300xg for 5 minutes at 4°C. The supernatant was removed, and cells were washed with phosphate-buffered saline (PBS) and centrifuged at 300xg for 5 minutes at 4°C. The resulting pellet was stored at -80°C. Cell pellets were resuspended in lysis buffer containing 8 M urea, 50 mM tris (pH 8), 5 mM $CaCl_2$, 30 mM NaCl, and protease (Roche) and phosphatase (Roche) inhibitor tablets. The pellet was lysed by four rounds of sonication at 4°C, alternating between 20 seconds on and 20 seconds off. Lysate protein concentration was measured by BCA (Thermo Pierce).

Digestion. Protein was reduced by addition of 5 mM dithiothreitol and incubated for 45 min at 55 °C. The mixture was cooled to room temperature, followed by alkylation of free thiols by addition of 15 mM iodoacetamide in the dark for 30 min. The alkylation reaction was quenched with 5 mM dithiothreitol. For tryptic digestion, a 1 mg protein aliquot was digested overnight with 20 µg trypsin (Promega, Madison, WI) at room temperature in 1 M urea. For LysC digestion, a 1 mg protein aliquot was digested overnight with 20 µg LysC (Wako, Richmond, VA) at room temperature in 4 M urea. For LysN digestion, a 1 mg protein aliquot was digested for four hours with 20 µg LysN (Thermo Pierce) at 37°C in 4 M urea. For GluC digestion, a 1 mg protein aliquot was digested overnight with 25 µg GluC (Roche Diagnostics, Indianapolis, IN) at room temperature in 0.5 M urea. For chymotrypsin digestion, a 1 mg protein aliquot was digested overnight with 12.5 µg of chymotrypsin resuspended in 0.2% FA (Promega, Madison, WI) in

1 M urea. For digestion with AspN, a 1 mg protein aliquot was incubated with 6 µg AspN (Roche Diagnostics, Indianapolis, IN) at room temperature overnight. Each digest was quenched by the addition of TFA and desalted on a 100 mg C_{18} Sep-Pak cartridge (Waters, Milford, MA).

Fractionation. High-pH RP fractionation was performed either using a Surveyor LC quarternary pump or a Dionex UltiMate 3000. Fractionation was performed at a flow rate of 1.0 mL/min using a 5 µm column packed with C18 particles (250-mm by 4.6-mm, Phenomenex) on a Surveyor LC quarternary pump. Samples were resuspended in buffer A and separated using the following gradient: 0-2 min, 100% buffer A and separated by increasing buffer B over a 60-minute gradient at a flow rate of 0.8 mL/minute (buffer A: 20 mM ammonium formate, pH 10; buffer B: 20 mM ammonium formate, pH 10, in 80% ACN). Flow rate was increased to 1.5 mL/minute during equilibration. Fractionation was performed at a flow rate of 0.45 mL/min using a 1.7 µm column packed with BEH particles (50-mm by 1-mm, Waters) on a Dionex Ultimate 3000 pump (Thermo). Samples were resuspended in buffer A and separated by increasing buffer B over a 45-minute gradient at a flow rate of 0.45 mL/minute (buffer A: 20 mM ammonium bicarbonate; buffer B: 20 mM ammonium bicarbonate in 80% ACN). Trypsin digested H1-hESC cells were first fractionated via strong cation exchange fractionation. Peptides were dissolved in 400 µl of strong cation exchange buffer A (5 mM KH_2PO_4 and 30% acetonitrile; pH 2.65) and injected onto a polysulfoethylaspartamide column (9.4 mm × 200 mm; PolyLC) attached to a Surveyor LC quarternary pump (Thermo Electron, West Chester, PA) operating at 3 ml/min. Fractions were collected every 2 min starting at 10 min into the following gradient: 0–2 min at 100% buffer A, 2–5 min at 0%–15% buffer B (5 mM KH_2PO_4, 30% acetonitrile, and 350 mM KCl (pH 2.65)), and 5–35 min at 15%–100% buffer B. Buffer B was held at 100% for 10 min. Fractions were collected from 8-12 minutes, 12-14 minutes, 14-16 minutes and 16-25 minutes. Each of these four SCX fractions was further fractionated by high-pH RP fractionation on a Surveyor LC quarternary pump, as described above.

LC-MS/MS. Samples were resuspended in 0.2% formic acid (FA) and separated via reversed phase (RP) chromatography. Peptides were injected on to a RP column prepared in-house. Approximately 35 cm of 75 μm-360 μm inner-outer diameter bare-fused silica capillary, each with a laser pulled electrospray tip, were packed with 1.7 μm diameter, 130 Å pore size, Bridged Ethylene Hybrid C18 particles (Waters). Columns were fitted on to either a nanoAcquity (Waters) or Dionex (Thermo) and heated to 60 °C using a home-built column heater. Mobile phase buffer A was composed of water and 0.2% formic acid. Mobile phase B was composed of 70% ACN, 0.2% formic acid, and 5% DMSO. Each sample was separated over a 100-min gradient, including time for column re-equilibration. Flow rates were set at 300-350 μl/min.

Peptide cations were electrosprayed into a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo) or a Thermo Orbitrap Lumos (Q-OT-qIT, Thermo). All fractions were analyzed using HCD. Precursor scans were performed from 300 to 1,500 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos. Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole, fragmented by HCD with a normalized collision energy of 30, and analyzed using turbo scan in the ion trap. For some analyses, precursors above 500 *m/z* were fragmented by HCD using the described conditions, while precursors below 500 *m/z* were fragmented by CAD with a normalized collision energy of 30. The maximum injection time for MS^2 analysis was normally set at either 25 or 35 ms, but was set higher for some analyses, with an ion count target of 10^4. Precursors with a charge state of 2-8 were sampled for MS^2. Dynamic exclusion time was set at 15 seconds, with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

LysC, LysN, AspN, GluC and chymotrypsin fractions were analyzed using ETD. To maximize identifications, precursor scans were performed from 200 to 800 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos. Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole. Precursors were fragmented by ETD using custom reaction times; +3: 40 ms, +4: 22 ms, +5: 14 ms, +6: 10

ms, +2: 70 ms. EThcD was performed on +2 precursors, at 25% supplemental activation collision energy. Precursor ions were selected for fragmentation based on charge state in the following order: +3, +4, +5, +6, +2. Fragment ions were analyzed in the ion trap. Dynamic exclusion time was set at 15 seconds, with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

Chymotrypsin fragments were analyzed using CAD. Precursor scans were performed from 300 to 1,500 *m/z* at either 60K or 120K resolution (at 400 *m/z*). A 5 x 10^5 ion count target was used on the Orbitrap Fusion, a 1 x 10^6 ion count target was used on the Orbitrap Lumos. Precursors selected for tandem MS were isolated at 0.7 Th with the quadrupole, fragmented by CAD with a normalized collision energy of 30, and analyzed using turbo scan in the ion trap. The maximum injection time for MS^2 analysis was normally set at either 25 or 35 ms, but was set higher for some analyses, with an ion count target of 10^4. Precursors with a charge state of 2-8 were sampled for MS^2. Dynamic exclusion time was set at 15 seconds, with a 10 ppm tolerance around the selected precursor and its isotopes. Monoisotopic precursor selection was turned on. Analyses were performed in top speed mode with either 3 or 5 second cycles.

Database searching. Thermo.raw files were converted to searchable DTA text files using the Coon OMSSA Proteomic Analysis Software Suite (COMPASS).[26] DTA files were searched against a target-decoy version[27] of the UniProt human database (downloaded August 2012) using version 2.1.9 of the Open Mass Spectrometry Search Algorithm (OMSSA).[28] Methionine oxidation was searched as a variable modification, and cysteine carbamidomethylation was searched as a fixed modification. The precursor, charge-reduced precursor, and neutral loss peaks were removed from ETD spectra. The precursor mass tolerance was defined as 50 ppm, and the fragment ion mass tolerance was set to 0.35 Da. Tryptic peptides were searched with C-terminal arginine and lysine specificity, LysC digested peptides were searched with C-terminal lysine cleavage specificity, LysN digested peptides were searched with N-terminal lysine cleavage specificity, chymotrypsin digested peptides were searched with C-terminal tryptophan, leucine, tyrosine and phenylalanine cleavage specificity, GluC digested peptides were searched with C-terminal

glutamic acid cleavage specificity, and AspN digested peptides were searched with N-terminal aspartic acid specificity. Obtained results were filtered to 1% peptide and protein level FDR using E-values.

ACKNOWLEDGEMENTS

This work was supported by an ACS Division of Analytical Chemistry Award and the Society for Analytical Chemists of Pittsburgh to ALR and an NLM training grant (#5T15LM007359) to AEM.

REFERENCES

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198-207.

(2) Nagaraj, N.; Kulak, N. A.; Cox, J.; Neuhauser, N.; Mayr, K.; Hoerning, O.; Vorm, O.; Mann, M. *Mol Cell Proteomics* **2012**, *11*, M111 013722.

(3) Hebert, A. S.; Richards, A. L.; Bailey, D. J.; Ulbrich, A.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2014**, *13*, 339-347.

(4) Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J. *Nat Protoc* **2015**, *10*, 701-714.

(5) Nagaraj, N.; Wisniewski, J. R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Paabo, S.; Mann, M. *Mol Syst Biol* **2011**, *7*, 548.

(6) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. *Mol Cell Proteomics* **2012**, *11*, M111 014050.

(7) Munoz, J.; Low, T. Y.; Kok, Y. J.; Chin, A.; Frese, C. K.; Ding, V.; Choo, A.; Heck, A. J. *Mol Syst Biol* **2011**, *7*, 550.

(8) Moghaddas Gholami, A.; Hahne, H.; Wu, Z.; Auer, F. J.; Meng, C.; Wilhelm, M.; Kuster, B. *Cell Rep* **2013**, *4*, 609-620.

(9) Kim, M. S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; Thomas, J. K.; Muthusamy, B.; Leal-Rojas, P.; Kumar, P.; Sahasrabuddhe, N. A.; Balakrishnan, L.; Advani, J.; George, B.; Renuse, S.; Selvan, L. D.; Patil, A. H.; Nanjappa, V.; Radhakrishnan, A.; Prasad, S.; Subbannayya, T.; Raju, R.; Kumar, M.; Sreenivasamurthy, S. K.; Marimuthu, A.; Sathe, G. J.; Chavan, S.; Datta, K. K.; Subbannayya, Y.; Sahu, A.; Yelamanchi, S. D.; Jayaram, S.; Rajagopalan, P.; Sharma, J.; Murthy, K. R.; Syed, N.; Goel, R.; Khan, A. A.; Ahmad, S.; Dey, G.; Mudgal, K.; Chatterjee, A.; Huang, T. C.; Zhong, J.; Wu, X.; Shaw, P. G.; Freed, D.; Zahari, M. S.; Mukherjee, K. K.; Shankar, S.; Mahadevan, A.; Lam, H.; Mitchell, C. J.; Shankar, S. K.; Satishchandra, P.; Schroeder, J. T.; Sirdeshmukh, R.; Maitra, A.; Leach, S. D.; Drake, C. G.; Halushka, M. K.; Prasad, T. S.; Hruban, R. H.; Kerr, C. L.; Bader, G. D.; Iacobuzio-Donahue, C. A.; Gowda, H.; Pandey, A. *Nature* **2014**, *509*, 575-581.

(10) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; Mathieson, T.; Lemeer, S.; Schnatbaum, K.; Reimer, U.; Wenschuh, H.; Mollenhauer, M.; Slotta-Huspenina, J.; Boese, J. H.; Bantscheff, M.; Gerstmair, A.; Faerber, F.; Kuster, B. *Nature* **2014**, *509*, 582-587.

(11) Pennisi, E. *Science* **2012**, *337*, 1159, 1161.

(12) Nilsen, T. W.; Graveley, B. R. *Nature* **2010**, *463*, 457-463.

(13) Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P. *Nat Methods* **2013**, *10*, 186-187.

(14) Tsiatsiani, L.; Heck, A. J. *FEBS J* **2015**, *282*, 2612-2626.

(15) Trevisiol, S.; Ayoub, D.; Lesur, A.; Ancheva, L.; Gallien, S.; Domon, B. *Proteomics* **2016**, *16*, 715-728.

(16) Ruprecht, B.; Roesli, C.; Lemeer, S.; Kuster, B. *Proteomics* **2016**.

(17) Gauci, S.; Helbig, A. O.; Slijper, M.; Krijgsveld, J.; Heck, A. J.; Mohammed, S. *Anal Chem* **2009**, *81*, 4493-4501.

(18) Wisniewski, J. R.; Mann, M. *Anal Chem* **2012**, *84*, 2631-2637.

(19) Guo, X.; Trudgian, D. C.; Lemoff, A.; Yadavalli, S.; Mirzaei, H. *Mol Cell Proteomics* **2014**, *13*, 1573-1584.

(20) Swaney, D. L.; Wenger, C. D.; Coon, J. J. *J Proteome Res* **2010**, *9*, 1323-1329.

(21) Riley, N. M.; Rush, M. J.; Rose, C. M.; Richards, A. L.; Kwiecien, N. W.; Bailey, D. J.; Hebert, A. S.; Westphall, M. S.; Coon, J. J. *Mol Cell Proteomics* **2015**, *14*, 2644-2660.

(22) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Scalf, M.; Smith, L. M. *J Proteome Res* **2014**, *13*, 228-240.

(23) MacCoss, M. J.; McDonald, W. H.; Saraf, A.; Sadygov, R.; Clark, J. M.; Tasto, J. J.; Gould, K. L.; Wolters, D.; Washburn, M.; Weiss, A.; Clark, J. I.; Yates, J. R., 3rd. *Proc Natl Acad Sci U S A* **2002**, *99*, 7900-7905.

(24) Giansanti, P.; Aye, T. T.; van den Toorn, H.; Peng, M.; van Breukelen, B.; Heck, A. J. *Cell Rep* **2015**, *11*, 1834-1843.

(25) Ghaemmaghami, S.; Huh, W. K.; Bower, K.; Howson, R. W.; Belle, A.; Dephoure, N.; O'Shea, E. K.; Weissman, J. S. *Nature* **2003**, *425*, 737-741.

(26) Wenger, C. D.; Phanstiel, D. H.; Lee, M. V.; Bailey, D. J.; Coon, J. J. *Proteomics* **2011**, *11*, 1064-1074.

(27) Elias, J. E.; Gygi, S. P. *Nat Methods* **2007**, *4*, 207-214.

(28) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. *J Proteome Res* **2004**, *3*, 958-964.