METHODS FOR LARGE-SCALE QUANTITATIVE PROTEOMICS IN SYSTEMS BIOLOGY

by

Elyse C. Freiberger

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

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: 21 February 2018

The dissertation is approved by the following members of the Final Oral Committee:
Joshua J. Coon, Professor, Biomolecular Chemistry
David J. Pagliarini, Associate Professor, Biochemistry
Peter Lewis, Assistant Professor, Biomolecular Chemistry
Lloyd M. Smith, Professor, Chemistry

ACKNOWLEDGMENTS

First and foremost I have to thank Professor Josh Coon. The research program he has put together is world-class, and I've loved my time working with him and the other Coon lab members. The breadth and depth of knowledge and equipment Prof. Coon has assembled is remarkable – it's an environment where grad students are given the space and resources to thrive, and I'm grateful to have been a part of the group. I'd also like to thank Mike Westphall. His hilarious outlook on life, ability to fix any piece of equipment or instrumentation we could possibly need, and uncanny way of explaining complicated physics concepts to very confused chemists make him an invaluable colleague I'm so happy to have worked alongside.

I want to thank all my labmates who have helped me along the way with their helpful discussions (and patience answering all my questions, especially in the beginning). In particular, I want to thank Alex Hebert, the first person to train me both on how to do proteomics, and how to manage the often-finicky instrumentation. Thanks also to Greg Potts, my first desk mate, and Kevin Schauer, my last, for tolerating my occasionally surly attitude. I'm grateful to have worked with Alicia Richards, who made my first foray into large-scale proteomics much more fun than it had any right to be, and to Nick Riley and Nick Kwiecien, who were great friends and made navigating the culture of such a large lab much easier. I've really enjoyed my time here at UW, particularly because of two people I met in the very first months I spent on campus. The first is Kim Krautkramer, whose

color commentary on grad school life and academia always made me smile. The second is Evgenia Shishkova. I'm so thankful we ended up in the same lab, doing the same kind of work; I don't think the good days would have been half as fun, nor the bad days nearly as tolerable, without her there.

Proteomics is a highly interdisciplinary field, and none of my work would have been possible without the efforts of my excellent collaborators in the Pagliarini lab, Attie lab, and Mehle lab within UW as well as Ray Deshaies at Caltech, Gary Churchill and Dan Gatti at the Jackson Laboratories, and Bob Kennedy at the University of Michigan. I'd also like to thank Professors Dave Pagliarini, Lloyd Smith, and Peter Lewis, for their guidance and service on my doctoral committee. While I was an undergrad at Purdue, I worked for the late, great D. James Morre learning the fundamentals of experimental biochemistry. Jim instilled in me a love of scientific research so enduring I never missed a day in the lab (even as I was missing several days a week in class…), and convinced me to my planned career path from law school to medicinal chemistry. As I was finishing up my time in college and deciding between two interesting job offers, Jim told me to take the mass spec job because he thought it a very marketable skill to learn. I took his advice and in the ten years since have never regretted it.

I was hired in to that mass spec job by Anthony Lee, then a group leader at Abbott Laboratories (now AbbVie). He and the resident mass spec expert Jianwei Shen were excellent mentors to me in my early industry career, and I am so grateful for their support and advice when I decided to go back to school.

I cannot thank my parents enough for their unending support over the past 3+ decades. They have always encouraged me to be curious and imaginative. They were always supportive of me following my dreams (with the gentle caveat that my dreams should always end in gainful employment) and have always encouraged me to further my education in pursuit of those goals.

Finally, I am unendingly thankful for the love and support of my husband, Jon. He has been so patient throughout this often stressful, hectic, and overwhelming process. I cannot wait to move back to Chicago with you and resume a much more normal life.

So it goes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

Δ*gene*	Single-gene deletion
3-MT	3-Methoxytyramine
ACN	Acetonitrile
AP-MS	Affinity purification – mass spectrometry
CAA	Chloroacetamide
CC	Collaborative Cross
Comt	Catechol-O-methyltransferase
CoQ	Coenzyme Q
CV	Coefficient of variation [(σ/μ) x 100%]
CX-MS	Cross-linking – mass spectrometry
Da	Dalton, unit of mass equal to the approximate mass of one proton or neutron
DAVID	Database for Annotation Visualization and Integrated Discovery
Dbh	Dopamine β-hydroxylase
Ddc	DOPA decarboxylase
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DO	Diversity outbred
DOPAC	34-dihydroxyphenylacetic acid
eQTL	Transcript quantitative trait locus

ER	Endoplasmic reticulum
ERAD	Endoplasmic reticulum-associated degradation
ESI	Electrospray ionization
FC	Fold change
FDR	False discovery rate
FT-ICR	Fourier-transform ion cyclotron resonance
GC	Gas chromatography
GLP-1	Glucagon-like peptide-1
GO	Gene ontology
GSIS	Glucose-stimulated insulin secretion
GWAS	Genome-wide association study
HF/HS diet	High fat/high sucrose Western-style diet
HVA	Homovanillic acid
iBAQ	Intensity-based absolute quantification
IP	Immunoprecipitation
iTRAQ	Isobaric tags for relative and absolute quantitation
KEGG	Kyoto Encyclopedia of Genes and Genomes
L-DOPA	L-34-Dihydroxyphenylalanine
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LFQ	Label-free quantification

LOD	Logarithm of odds
lQTL	Lipid quantitative trait locus
MALDI	Matrix-assisted laser desorption ionization
Mao	Monoamine oxidase
MAP	Mixing after purification
Mb	Mega base pairs
MCNA	Molecule covariance network analysis
ME	Module eigengene
mQTL	Metabolite quantitative trait locus
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS^1	Precursor mass analysis
MS^2	Tandem mass spectrum
MS/MS	Tandem mass spectrum
MXP	Mitochondrial uncharacterized protein
m/z	Mass-to-charge ratio
oGTT	Oral glucose tolerance test
OxPhos	Oxidative phosphorylation
PA	Palmitate
PAM	Purification after mixing

PC1	First principal component
PCA	Principal component analysis
Pnmt	Phenylethanolamine N-methyltransferase
PQC	Protein quality control
pQTL	Protein quantitative trait locus
PRM	Parallel reaction monitoring
PTM	Post-translational modification
QTL	Quantitative trait locus
RC	Respiration competent
RDR	Respiration deficiency response
SAM	S-adenosylmethionine
SD	Standard deviation
SEC	Size exclusion chromatography
SEM	Standard error of the mean
SILAC	Stable isotope labeling by amino acids in cell culture
SILAM	Stable isotope labeling by amino acids in mammals
SIM	Selected ion monitoring
SNP	Single nucleotide polymorphism
SRM	Single reaction monitoring
T2D	Type 2 diabetes

TCA	Cycle tricarboxylic acid cycle
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
Th	Tyrosine hydroxylase
Th	Thomson, a unit of measurement equal to mass divided by charge
TIC	Total ion current
TMT	Tandem mass tags
UPS	Ubiquitin-proteasome system
VCP	Valosin-containing protein
WGCNA	Weighted gene co-expression network analysis
WT	Wild-type
XIC	Extracted ion chromatogram
Y3K	Yeast 3,000 (dataset from chapter 3)
	Mouse strain abbreviations:
129	129S1/SvImJ
B6	C57BL/6J
CAST	CAST/EiJ
NOD	NOD/ShiLtJ
NZO	NZO/HILtJ
PWK	PWK/PhJ

WSB WSB/EiJv

ABSTRACT

The research described in this dissertation presents strategies for acquiring high quality large-scale mass spectrometry-based quantitative proteomics data that have been applied to model systems as diverse as Brewers' yeast, cultured mammalian cells, and isolated murine pancreatic islets. Background on proteomics technologies, an overview of the requirements for quantitative proteomics experiments, and major concerns specific to proteomics on a large scale are presented in **Chapter 1**. A study of valosin-containing protein interactors using size-exclusion chromatography and a novel quantification scheme is described in **Chapter 2**. A comprehensive analysis of 174 single-gene deletion yeast strains wherein proteomic data is integrated with metabolic and lipidomic results is described in **Chapter 3**. In **Chapter 4**, pancreatic islet proteomes across eight inbred laboratory mouse strains are quantified and the data are correlated with phenotypic information and *ex vivo* insulin secretion results. **Chapter 5** details method optimization and analysis of 383 mouse islet proteomes using the diversity outbred mouse resource for the largest protein quantitative trait locus mapping study completed to date.

Chapter 1

INTRODUCTION AND BACKGROUND

Introduction

Prior to the advent of fast, accurate genome profiling, the study of complex biological systems was limited to targeted experiments exploring a few genes, gene products, or metabolites at a time, often times in *ex vivo* or *in vitro* analyses, and integration of results into the larger context of the organism could be difficult. The development of functional genomic techniques in the 1990s paved the way for a more holistic approach to complement the focused biochemical assays that allowed scientists to compare complete sets of genomic information within and between whole biological systems or populations[1,2]. Following closely behind was the field of transcriptomics, the measurement of complete sets of mRNA in a system, and from that point, as science journalist James Gorman writes, "An epidemic of neologia ensued"[3,4]. Metabolomics, lipidomics, glycomics, fluxomics, and interactomics (the studies of all the metabolites, lipids, carbohydrates, changes in molecular dynamics, and molecular interactions in a system, respectively) are just some of the burgeoning 'omic' fields to have emerged in the past few decades[3,5]. With the goal of a complete understanding of cellular function in a system, the next logical step after mapping the transcriptome is to map the proteome – the 'ome' field interested in proteins – since proteins are the molecules responsible for driving expression of the phenotypes coded for in the genome. Because transcriptomics has more mature technologies, with next-generation sequencing able to measure transcripts at a depth near completeness, it may be tempting to think that we can learn all we need about protein expression via their transcript information[4].

However, the relationship between transcript and protein abundance is not always direct; experimentally, the correlation coefficients between proteins and mRNA in the same tissue has been consistently calculated at ~ $R^2 = 0.4$[6–10]. Furthering the study of proteomics offers unique perspectives on biological function, and while the technologies to measure proteomes are younger and less developed, the field is moving quickly toward complete proteome characterization in a matter of hours, via mass spectrometry-based analyses[11–13].

Mass spectrometry-based proteomics

Mass spectrometry has its genesis at the turn of the 20th century; J. J. Thomson first used cathode ray tubes to measure charge-to-mass ratios in 1897, and is credited with discovering the electron. Modern mass spectrometers were developed independently by Arthur Dempster in 1918 and Thomson protégé F. W. Aston in 1919[14]. However, for nearly seventy years, mass spectrometry remained the exclusive purview of scientists working with single atoms and small molecules[15]. Mass spectrometers require that analytes be in the gas phase for detection, and as of the mid-1980s no method had yet been discovered that could move larger biological molecules such as proteins and nucleic acids into the gas phase without causing severe degradation. The concurrent development of two so-called 'soft' ionization methods – electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) – that became available in 1988 opened the world of bio-molecules up to mass spectrometric analysis[16,17].

Proteomics methods can be divided into two broad categories based on whether proteins

are kept intact for analysis, or are digested into peptides. Methods requiring intact proteins are termed 'top-down'. These analyses require special consideration due to the size of the molecule, the number of different charge states in which each protein is present, and the variety of PTM combinations[18]. 'Bottom-up' refers to methods in which proteins are digested to peptides for sequencing and quantification using protease enzymes[19]. All analyses in this work are of the bottom-up. The protease used most commonly for MS-based proteomics experiments is trypsin, which cleaves C-terminal to lysine and arginine residues, in part because it guarantees a nitrogen-containing side chain in each peptide that can carry a positive charge (in addition to the N-terminus), reducing the mass-to-charge ratio of the peptide ions. Because of the distribution of lysines and arginines throughout the proteome, tryptic peptides tend to be of a size that is particularly amenable to mass spectrometric analysis (between 700 and 1200 Da), and trypsin itself is easily modified in commercial preparations to increase robustness and catalytic efficiency[20].

Many steps in the standard sample preparation workflow for a bottom-up proteomics analysis do not change appreciably from experiment to experiment (**Fig. 1.1**). Firstly, and regardless of sample source, the proteins must be denatured so that the protease can access cleavage sites. This can be done effectively using a combination of the chaotropes guanidine and urea. Guanidine is added to the pellet then samples are boiled which lyses the cells (if whole cells are used), denatures the proteins, and quickly quenches any native protease activity that might degrade the proteins in undesirable ways. However, many proteases, including trypsin, are not active enough in guanidine to use it during the enzymatic

incubation[21]. So, the next step is to remove the guanidine using a methanol precipitation. The resulting clean pellet can then be resuspended in urea, which keeps the proteins denatured but is more trypsin-compatible. Concomitantly with the urea resuspension, cysteine residues are reduced and alkylated to destroy any remaining vestiges of tertiary structure[22,23]. The protease of choice is then added and incubated with the proteins until the digestion is complete. Following enzymatic digestion, the sample must be desalted to remove the urea and buffering salts as they can contaminate the mass spectrometer inlet and ruin the chromatographic column. Cartridges containing reverse-phase chromatographic media are used, and the resulting eluate is dried down and the peptides resuspended in an LC-MS-compatible solvent such as dilute formic acid for injection on the analytical system[21–23].

Instrument set-up used for bottom-up proteomics analysis generally begins with a nano-flow liquid chromatography system used with a reverse-phase C18 column[22,23]. After separation and ionization, peptides can be detected using a number of different mass analyzers. Low-resolution analyzers include ion traps and quadrupole mass filters, which can reliably resolve peaks to $\pm$0.1-1 Th[24–26]. This is also often called unit resolution, since one can generally be confident in the mass within one Th unit. Time-of-flight, Orbitrap, and FT-ICR are common high resolution mass analyzers in order of increasing resolution capacity[27–30]. These mass analyzers can be used alone or in tandem with each other. In this work, the instruments used contained three mass analyzers: a quadrupole mass filter, a dual-pressure ion trap, and an orbitrap . This combination of analyzers allow for parallelization

of different analyses, which greatly increases the number of analyses possible during a single experiment[31].

A few different acquisition schema can be used for bottom-up proteomics. Data-independent (DIA) methods start with MS^1 scans that measure intact peptide *m/z*, then isolate all precursors in a specified time window for fragmentation[32]. Data-dependent acquisition (DDA) methods also start with MS^1 scans of intact peptides, but then include a series of MS^2 scans in which individual peptides are isolated and fragmented so the fragments can be directly matched to an individual precursor mass[19]. When using hybrid MS instrumentation, MS^1 scans are generally high-resolution, and depending on experimental needs, MS^2 scans can be either high or low resolution. High resolution MS^2 scans result in more confident amino acid sequence information (which can be important when trying to localize PTMs), whereas unit resolution MS^2 scans are much faster so more peptides can be identified (which is often preferable in whole-proteome experiments)[13].

Proteins are identified within these tandem MS data using database searching. For any organism whose genome is sequenced, all potential amino acid sequences can be catalogued. We use these sets of potentially-expressed proteins as templates for our MS results; we can predict where our proteases will cleave and generate a list of all possible peptides and peptide masses, then predict all possible fragmentation products from those peptides. We computationally match each fragmented peptide to an amino acid sequence first using intact masses, then by fragmentation pattern. All detected peptides in our samples are matched to potential sequences, and confidence values are calculated[13,19]. Data quality

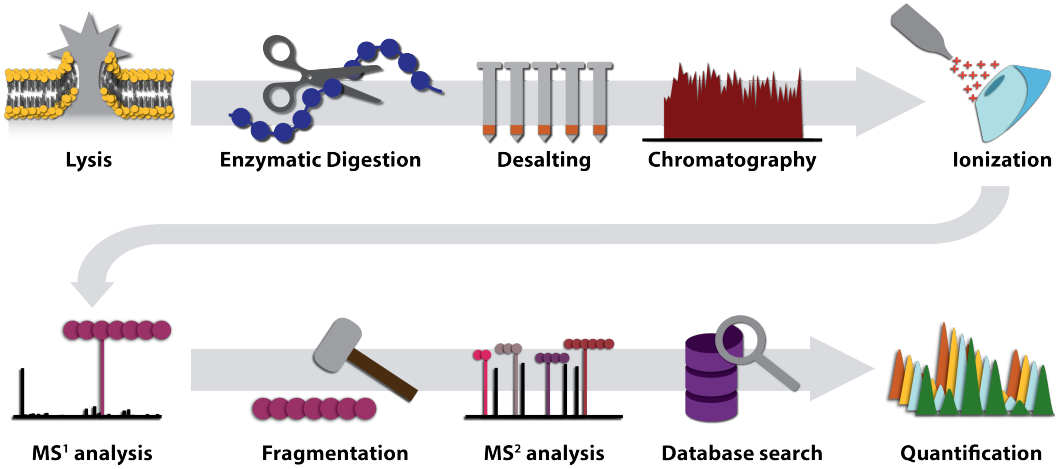


Figure 1.1: Bottom-up proteomics workflow. Graphical representation of all the steps needed for sample preparation, MS analysis, and protein identification in a standard bottom-up proteomics experiment.

is then assessed using the target-decoy method. Target-decoy searching was developed by Elias and Gygi in 2007 as a way to estimate false-positive rates in peptide and protein identifications. This estimation is done by searching the data against a decoy database in addition to the known protein sequence database[33]. In our case, the decoy database is generated by reversing the amino acid sequences in the target database; this retains amino acid frequency and distribution. Any peptides matched to the decoy database are known to be false positives, and are used to set a significance cutoff based on the desired FDR. The assumption being that the number of known false positives correlates with the number of unknown false positives—peptides that are matched incorrectly to sequences from the target database.

Once peptide identifications are assigned, those peptides can be collapsed into protein identifications. Due to overlap in amino acid sequences between proteins, particularly between protein isoforms, determining which proteins are actually present in a pool of known peptides is not necessarily straightforward. The issues surrounding this process are collectively referred to as the "protein inference problem"[34]. Occam's razor is a driving philosophy with regard to protein assembly in bottom-up proteomics; the goal is to find the fewest number of proteins that describe all the peptides in your data. Sometimes, however, proteins are so similar that one identification cannot be chosen over the other. In MaxQuant, the search and quantification software used in the balance of this work, if two or more proteins can be explained by the exact same set of peptides, that information is retained in the output and instead of a single protein ID, they become a 'protein group'[23,35].

Quantitative proteomics

In response to many biological questions, we are interested in determining not only what proteins are present, but also how much of each detected protein is present. Broadly speaking, there are two different strategies one can take when trying to quantify proteomes. The first is to use absolute quantification, wherein one is measuring the precise amount of each protein contained in each individual sample. This approach is complicated by the fact that MS intensity measurements alone are not reliable proxies for analyte concentration due to variations in ionization efficiency between molecular species. Instead, synthetic peptides can be used either to generate standard curves for each protein of interest or as spiked-in internal standards[36]. This method can be more time-consuming and low-throughput than a standard bottom-up experiment because of the need to identify peptides of interest, generate the synthetic standards, then analyze using targeted methods such as SRM/MRM, PRM, or SIM[36]. Absolute quantitation can be approximated in bottom-up experiments using the iBAQ algorithm, which uses summed intensities of all peptides mapped to a specific protein normalized to the number of theoretically-observable peptides that could be generated from that protein[35,37].

The second strategy for quantifying proteins via mass spectrometry is to measure the ratios of individual proteins between two or more experimental conditions, called relative quantification (**Fig. 1.2**). In this case, we can use intensity as a proxy for expression level because we are comparing the same molecules to one another and any ionization efficiency

issues are rendered moot. There are several common relative quantification strategies, including chemical tagging, metabolic tagging, and label-free methods. Chemical tagging such as tandem mass tags (TMT) or isobaric tags for relative and absolute quantitation (iTRAQ) involves isobaric labeling at the peptide level, post-digest[10,38,39]. Samples can then be pooled and analyzed simultaneously. Tags are fragmented during MS^2 or MS^3 analysis into reporter ions, whose relative intensities represent the quantitative measurement. Metabolic tagging comprises methods like SILAC/SILAM and NeuCode that involve incorporation of heavy-isotope-labeled amino acids during protein synthesis[40–42]. Samples grown in different labels are then pooled, and their relative MS^1 intensities represent the quantitative measurement. Tagging methods in general are beneficial in that multiple samples can be processed together, reducing the number of individual instrumental experiments needed, and alleviating technical variation due to sample preparation workflows, which is particularly useful when enriching for low-level PTMs or adding other post-digest preparative steps. Metabolic tagging can also be used to determine rates of protein turnover, since heavy-labeled peptides will only be generated for newly-synthesized proteins[43,44]. However, because they must be incorporated into culture media or animal feed over lengthy growth periods, their use is limited to cell culture or fast-growing model organisms. Chemical tagging can be applied to a broader range of sample types, but suffers from issues of dynamic-range suppression due to co-isolation of different peptides during the fragmentation steps[45]. Both chemical tagging and metabolic tagging also result in fewer protein IDs per single-shot injection than in comparable untagged experiments.

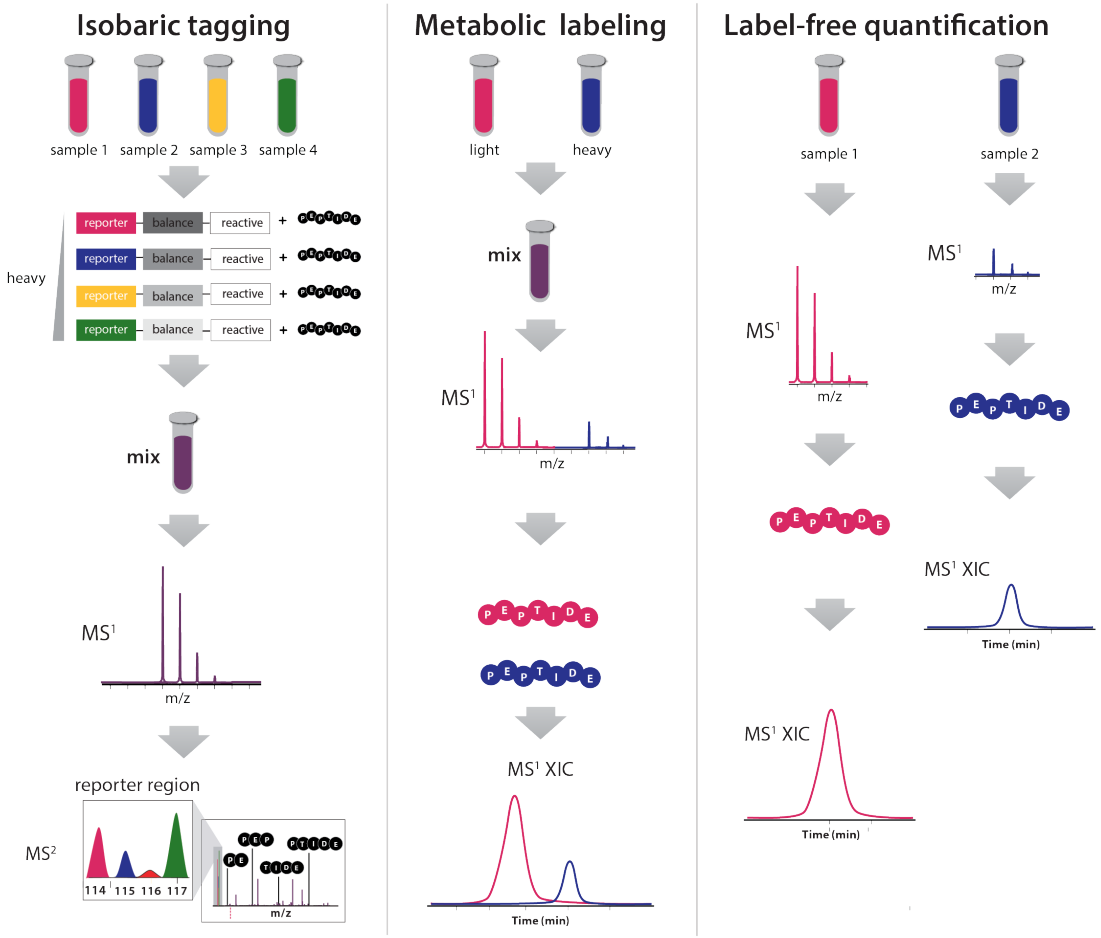


Figure 1.2: Relative protein quantification. The three major categories of relative protein quantification. Isobaric tagging includes TMT and iTRAQ schema, metabolic labeling includes SILAC, SILAM, and NeuCode SILAC, and label free quantification requires no tagging methods, but does require post-acquisition computational support for normalization such as is found in the MaxQuant software[35,37].

The last strategy for relative quantification in proteomics, and the one that will be used in the substance of this work, is label-free quantification. In this case, we are referring specifically to intensity-based quantification. Spectral counting, where proteins are quantified using the number of spectra mapped back to that protein, while useful in some circumstances, can only be considered semi-quantitative[46]. Intensity-based relative quantification is desirable in situations where instrument time isn't a limiting factor, because it does not require additional tagging steps during sample preparation, can be used with any sample type, and does not impact the number of protein IDs generated from each experiment. The major concern with label-free quantification is reproducibility; peptide intensities can be affected by instrument performance, chromatographic quality, and variations in sample preparation. Aside from ensuring as much parity in sample handling between injections as possible, this technical variation can be difficult to manage. The MaxLFQ algorithm within Juergen Cox and Matthias Mann's MaxQuant software handles these issues by normalizing the measured intensities to the total ion current (TIC) for each MS experiment[35,37]. Resulting normalized intensities are termed LFQ intensities, and the maximum reliably-quantifiable LFQ ratios are in excess of 100-fold.

Specific concerns for large-scale quantitative proteomics experiments

Increasing the number of samples analyzed in an 'omics' study can yield real benefits to the systems biology questions one might be exploring, including increased statistical significance and broader scope, but such large studies come with significant technical

concerns that can confound any anticipated benefits if not appropriately addressed[47]. As the time needed to prepare and analyze the samples increases, batch effects (technical variations due to sample handling) become more problematic. In some cases, variations due to batch effects can completely mask biological variation in raw results[48]. **Figure 1.3** shows principal component analysis of a large-scale proteomics dataset before and after normalization; the way in which the non-normalized data segregates by batch is striking and very common if data collection continues across a long period of time. Sources of batch effects can be found all along the analytical workflow, any time samples are prepared together in smaller subsets of the whole: when organisms are raised or cultured, during tissue extraction, during sample lysis and digestion, and even during the instrumental analysis. One very clear source of batch effect variation is during chromatographic separation; in many cases, samples run on the same nano-LC column are more similar than samples run on different columns, irrespective of biological condition.

Batch effects can be mitigated, though not completely avoided, by thoughtful experimental design. In particular, one should try to avoid compounding the batch effects by shuffling the batches throughout sample handling. For example, if at all possible samples should not be prepared in the same subsets in which they were grown or cultured, and for label-free quantification especially sample order should be randomized prior to MS analysis. Regardless of what efforts are taken to minimize technical variation during analysis, some batch effects will always need to be addressed in post-processing. Such measures include normalizing to a standard prepared and run with the samples of interest, use of

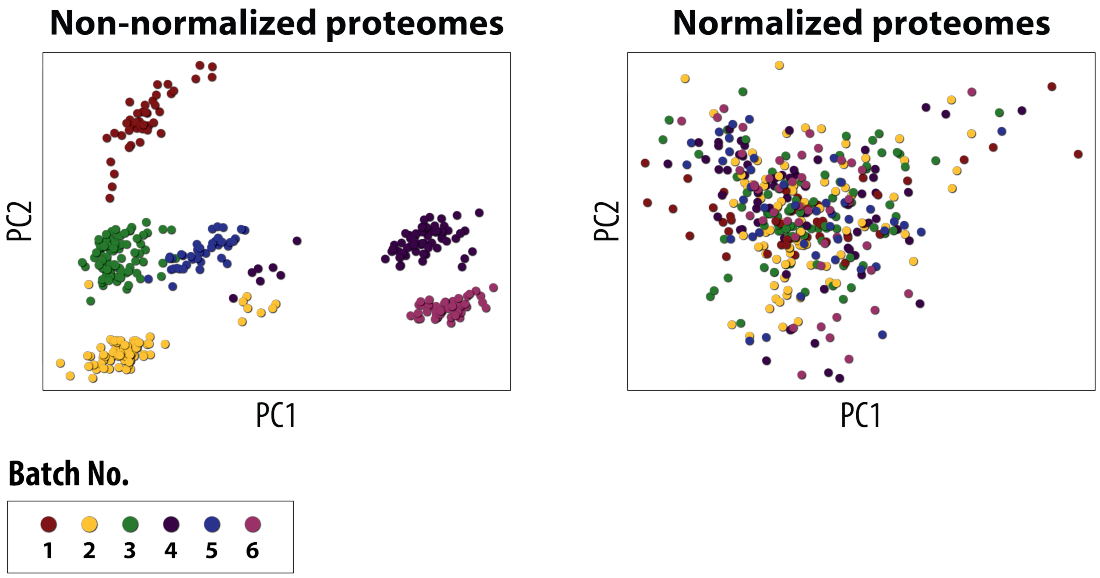


Figure 1.3: Batch effects in large-scale proteomics experiments. Representative large-scale proteomics data in PCA plots. These data include 380 proteomes collected in six batches across six different chromatographic columns and six months. Proteins were quantified using MaxQuant, and batches were normalized using ComBat[49].

more computationally complicated approaches such as the ComBat algorithm (which uses empirical Bayes methods to adjust data by batch)[49], and inclusion of batch information as covariates during statistical analyses.

Quantitative trait locus mapping

Quantitative trait locus (QTL) mapping is an old technique finding new applications in the age of large-scale 'omics' analysis; it is a statistical method used to link variations at specific loci on the genome with variation in a specific quantified phenotype. The first true QTL study was published in 1923 by Karl Sax, linking variations in pigmentation to loci in intercrossed strains of green beans[50]. Since then, this method has become widely used as a genetic tool. The only requirements are known set of markers across the genome, and a measurable phenotype. In the case of 'omics' analyses, each measurement stands alone as an individual phenotype. So, in our proteomics study, each protein is analyzed independently against the marker map. The output in our case is a logarithm of odds, or LOD, score. The LOD score corresponds to the likelihood of linkage between the genetic marker and the measured phenotype, generally calculated by taking the $\log_{10}$ of the ratio of the likelihood of the alternative hypothesis (linkage) being true over the likelihood of the null hypothesis (no linkage) being true – higher LOD scores suggest stronger evidence for linkage[51]. See **Figure 1.4a** and **b** for examples of how QTL mapping data can be plotted.

Once QTLs with significant LOD scores have been tabulated, the data can be further processed in a number of ways. The first is to identify likely candidate gene drivers of

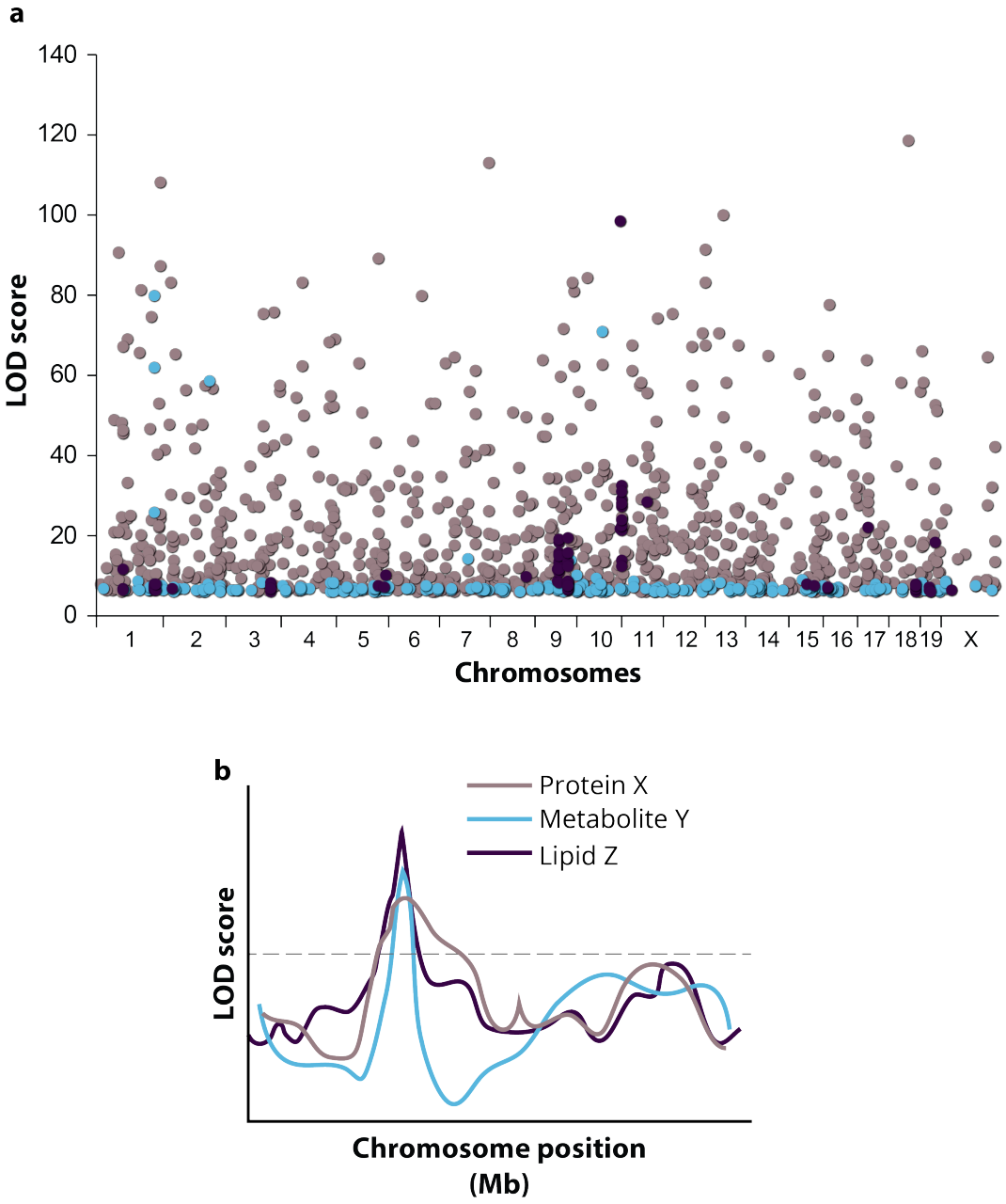


Figure 1.4: Representative multi-omic QTL data. Example of how multiples 'omes' of QTL data is plotted. All data are simulated. **A** Manhatten plot showing LOD apices for three 'omes' across all chromosomes. **B** Zooming in on a small subsection of a chromosome, plotting LOD scores against chromosomal position for a protein, a metabolite, and a lipid. These three molecules have apices that overlap and all exceed the significance threshold. They are good candidates for being functionally related.

the QTL effect; depending on the mapping resolution, this list can be narrowed to a sub-Mb window on the chromosome[52]. Overlapping QTLs, or QTLs mapping to the same chromosomal region, may also be of interest as they are more likely to be interacting functionally. In protein and transcript QTL datasets, mediation analysis may be used to determine more confidently which QTL phenotypes are interrelated. Finally, an association study may be used on narrow linkage windows to determine which specific SNPs are potential drivers of the QTL effects, information that can direct follow-up knock-out or point mutation experiments to confirm hypothesized regulatory interactions[53,54].

References

[1] E. A. Susaki and H. R. Ueda, "Whole-body and Whole-Organ Clearing and Imaging Techniques with Single-Cell Resolution: Toward Organism-Level Systems Biology in Mammals," *Cell Chemical Biology*, vol. 23, pp. 137–157, Jan 2016.

[2] R. A. Ankeny, "Sequencing the genome from nematode to human: changing methods, changing science," *Endeavour*, vol. 27, pp. 87–92, jun 2003.

[3] J. Gorman, "'Ome,' the Sound of the Scientific Universe Expanding,"*The New York Times: Science*, May 3, 2012.

[4] R. Lowe, N. Shirley, M. Bleackley, S. Dolan, and T. Shafee, "Transcriptomics technologies," *PLOS Computational Biology*, vol. 13, p. e1005457, may 2017.

[5] H. G. Stunnenberg and N. C. Hubner, "Genomics meets proteomics: identifying the culprits in disease.," *Human genetics*, vol. 133, pp. 689–700, jun 2014.

[6] Q. Tian, S. B. Stepaniants, M. Mao, L. Weng, M. C. Feetham, M. J. Doyle, E. C. Yi, H. Dai, V. Thorsson, J. Eng, D. Goodlett, J. P. Berger, B. Gunter, P. S. Linseley, R. B. Stoughton, R. Aebersold, S. J. Collins, W. A. Hanlon, and L. E. Hood, "Integrated genomic and proteomic analyses of gene expression in Mammalian cells.," *Molecular & cellular proteomics : MCP*, vol. 3, pp. 960–9, oct 2004.

[7] C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva, "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line," *Molecular Systems Biology*, vol. 6, p. 400, aug 2010.

[8] E. Lundberg, L. Fagerberg, D. Klevebring, I. Matic, T. Geiger, J. Cox, C. Älgenäs, J. Lundeberg, M. Mann, and M. Uhlen, "Defining the transcriptome and proteome in three functionally different human cell lines," *Molecular Systems Biology*, vol. 6, p. 450, dec 2010.

[9] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, pp. 337–342, may 2011.

[10] S. P. Gygi, B. Rist, T. J. Griffin, J. Eng, and R. Aebersold, "Proteome analysis of low-

abundance proteins using multidimensional chromatography and isotope-coded affinity tags.," *Journal of proteome research*, vol. 1, no. 1, pp. 47–54.

[11] N. M. Riley, A. S. Hebert, and J. J. Coon, "Proteomics Moves into the Fast Lane," *Cell Systems*, vol. 2, pp. 142–143, mar 2016.

[12] A. Bensimon, A. J. Heck, and R. Aebersold, "Mass Spectrometry–Based Proteomics and Network Biology," *Annual Review of Biochemistry*, vol. 81, pp. 379–405, jul 2012.

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

[14] J. Griffiths, "A Brief History of Mass Spectrometry," *Analytical Chemistry*, vol. 80, pp. 5678–5683, aug 2008.

[15] J. H. Beynon, "The history of mass spectrometry and the search for zero," *Biological Mass Spectrometry*, vol. 8, pp. 380–383, sep 1981.

[16] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules.," *Science (New York, N.Y.)*, vol. 246, pp. 64–71, oct 1989.

[17] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," *Analytical Chemistry*, vol. 60, pp. 2299–2301, oct 1988.

[18] J. R. Yates and N. L. Kelleher, "Top Down Proteomics," *Analytical Chemistry*, vol. 85, pp. 6151–6151, jul 2013.

[19] M. Mann, N. A. Kulak, N. Nagaraj, and J. Cox, "The Coming Age of Complete, Accurate, and Ubiquitous Proteomes," *Molecular Cell*, vol. 49, pp. 583–590, feb 2013.

[20] D. L. Swaney, C. D. Wenger, and J. J. Coon, "Value of using multiple proteases for large-scale mass spectrometry-based proteomics.," *Journal of proteome research*, vol. 9, pp. 1323–9, mar 2010.

[21] J. L. Proc, M. A. Kuzyk, D. B. Hardie, J. Yang, D. S. Smith, A. M. Jackson, C. E. Parker, and C. H. Borchers, "A quantitative study of the effects of chaotropic agents, surfactants, and solvents on the digestion efficiency of human plasma proteins by trypsin.," *Journal of proteome research*, vol. 9, pp. 5422–37, oct 2010.

[22] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, pp. 339–347, Jan 2014.

[23] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, pp. 701–714, apr 2015.

[24] R. A. Yost and C. G. Enke, "Triple quadrupole mass spectrometry for direct mixture

analysis and structure elucidation," *Analytical Chemistry*, vol. 51, pp. 1251–1264, oct 1979.

[25] P. E. Miller and M. B. Denton, "The quadrupole mass filter: Basic operating concepts," *Journal of Chemical Education*, vol. 63, p. 617, jul 1986.

[26] J. C. Schwartz, M. W. Senko, and J. E. P. Syka, "A two-dimensional quadrupole ion trap mass spectrometer," *Journal of the American Society for Mass Spectrometry*, vol. 13, pp. 659–669, jun 2002.

[27] A. Makarov, "Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis," 2000.

[28] R. H. Perry, R. G. Cooks, and R. J. Noll, "Orbitrap mass spectrometry: Instrumentation, ion motion and applications," *Mass Spectrometry Reviews*, vol. 27, pp. 661–699, nov 2008.

[29] J. E. P. Syka, J. A. Marto, D. L. Bai, S. Horning, M. W. Senko, J. C. Schwartz, B. Ueberheide, B. Garcia, S. Busby, T. Muratore, J. Shabanowitz, and D. F. Hunt, "Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications.," *Journal of proteome research*, vol. 3, no. 3, pp. 621–6.

[30] I. V. Chernushevich, A. V. Loboda, and B. A. Thomson, "An introduction to quadrupole-

time-of-flight mass spectrometry," *Journal of Mass Spectrometry*, vol. 36, pp. 849–865, aug 2001.

[31] M. W. Senko, P. M. Remes, J. D. Canterbury, R. Mathur, Q. Song, S. M. Eliuk, C. Mullen, L. Earley, M. Hardman, J. D. Blethrow, H. Bui, A. Specht, O. Lange, E. Denisov, A. Makarov, S. Horning, and V. Zabrouskov, "Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates," *Analytical Chemistry*, vol. 85, pp. 11710–11714, dec 2013.

[32] A. Doerr, "DIA mass spectrometry," *Nature Methods*, vol. 12, pp. 35–35, Jan 2015.

[33] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature Methods*, vol. 4, pp. 207–214, mar 2007.

[34] A. I. Nesvizhskii and R. Aebersold, "Interpretation of Shotgun Proteomic Data," *Molecular & Cellular Proteomics*, vol. 4, pp. 1419–1440, oct 2005.

[35] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nature Biotechnology*, vol. 26, pp. 1367–1372, dec 2008.

[36] A. Otto, J. Rg Bernhardt, M. Hecker, U. Vö Lker, and D. Rte Becher, "Proteomics: From relative to absolute quantification for systems biology approaches 3," *Systems Biology of Bacteria*, vol. 39, pp. 81–106, 2012.

[37] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.," *Molecular & cellular proteomics : MCP*, vol. 13, pp. 2513–26, sep 2014.

[38] L. Dayon and J.-C. Sanchez, "Relative Protein Quantification by MS/MS Using the Tandem Mass Tag Technology," in *Methods in molecular biology (Clifton, N.J.)*, vol. 893, pp. 115–127, 2012.

[39] A. Thompson, J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms," *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003. PMID: 12713048.

[40] A. S. Hebert, A. E. Merrill, D. J. Bailey, A. J. Still, M. S. Westphall, E. R. Strieter, D. J. Pagliarini, and J. J. Coon, "Neutron-encoded mass signatures for multiplexed proteome quantification," *Nature Methods*, vol. 10, pp. 332–334, apr 2013.

[41] A. E. Merrill, A. S. Hebert, M. E. MacGilvray, C. M. Rose, D. J. Bailey, J. C. Bradley, W. W. Wood, M. El Masri, M. S. Westphall, A. P. Gasch, and J. J. Coon, "NeuCode Labels for Relative Protein Quantification," *Molecular & Cellular Proteomics*, vol. 13, pp. 2503–2512, sep 2014.

[42] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and

M. Mann, "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.," *Molecular & cellular proteomics : MCP*, vol. 1, pp. 376–86, may 2002.

[43] M. K. Doherty, D. E. Hammond, M. J. Clague, S. J. Gaskell, and R. J. Beynon, "Turnover of the Human Proteome: Determination of Protein Intracellular Stability by Dynamic SILAC," *Journal of Proteome Research*, vol. 8, pp. 104–112, Jan 2009.

[44] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, pp. 337–342, may 2011.

[45] S. Y. Ow, M. Salim, J. Noirel, C. Evans, I. Rehman, and P. C. Wright, "iTRAQ Underestimation in Simple and Complex Mixtures: "The Good, the Bad and the Ugly"," *Journal of Proteome Research*, vol. 8, pp. 5347–5355, nov 2009.

[46] D. H. Lundgren, S.-I. Hwang, L. Wu, and D. K. Han, "Role of spectral counting in quantitative proteomics," *Expert Review of Proteomics*, vol. 7, pp. 39–53, feb 2010.

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

[48] J. Gregori, L. Villarreal, O. Méndez, A. Sánchez, J. Baselga, and J. Villanueva, "Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics," *Journal of Proteomics*, vol. 75, pp. 3938–3951, jul 2012.

[49] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, pp. 118–127, Jan 2007.

[50] T. I. Sax, Karl, "THE ASSOCIATION OF SIZE DIFFERENCES WITH SEED-COAT PATTERN AND PIGMENTATION IN PHASEOLUS VULGARIS," *Genetics*, vol. 8, 1923.

[51] R. W. Doerge, "MULTIFACTORIAL GENETICS MAPPING AND ANALYSIS OF QUANTITATIVE TRAIT LOCI IN EXPERIMENTAL POPULATIONS," *Nature Reviews Genetics*, vol. 3, pp. 43–52, Jan 2002.

[52] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill, "High-resolution genetic mapping using the Mouse Diversity outbred population.," *Genetics*, vol. 190, pp. 437–47, feb 2012.

[53] D. Gatti, J. E. French, and K. Schughart, "QTL Mapping and Identifi cation of Candidate Genes in DO Mice: A Use Case Model Derived from a Benzene Toxicity Experiment," *Methods in Molecular Biology*, vol. 1488.

[54] A. Darvasi, A. Weinreb, V. Minke, J. I. Wellert, and M. Soller, "Detecting Marker-QTL Linkage and Estimating QTL Gene Effect and Map Location Using a Saturated Genetic Map,"

Chapter 2

VCP-ADAPTOR INTERACTIONS ARE EXCEPTIONALLY DYNAMIC AND SUBJECT TO DIFFERENTIAL MODULATION BY A VCP INHIBITOR

ECF designed and ran MS experiments, analyzed data, and made figures for the SEC-MS sections of this work.

This chapter has been published:

Liang Xue, Emily E. Blythe, **Elyse C. Freiberger**, Jennifer Mamrosh, Alexander S. Hebert, Justin M. Reitsma, Sonja Hess, Joshua J. Coon, and Raymond Deshaies. *VCP-adaptor interactions are exceptionally dymanic and subject to differential modulation by a VCP inhibitor.* Molecular and Cellular Proteomics. **2016**, *9*, 2970-2986.

Abstract

Protein quality control (PQC) plays an important role in stemming neurodegenerative diseases and is essential for the growth of some cancers. Valosin-containing protein (VCP)/p97 plays a pivotal role in multiple PQC pathways by interacting with numerous adaptors that link VCP to specific PQC pathways and substrates and influence the post-translational modification state of substrates. However, our poor understanding of the specificity and architecture of the adaptors, and the dynamic properties of their interactions with VCP hinders our understanding of fundamental features of PQC and how modulation of VCP activity can best be exploited therapeutically. In this study we use multiple mass spectrometry-based proteomic approaches combined with biophysical studies to characterize the interaction of adaptors with VCP. Our results reveal that most VCP–adaptor interactions are characterized by rapid dynamics that in some cases are modulated by the VCP inhibitor NMS873. These findings have significant implications for both the regulation of VCP function and the impact of VCP inhibition on different VCP–adaptor complexes.

Introduction

Protein Quality Control is thought to play an important role in human health, and mutations in key regulators of PQC lead to neurodegenerative disease[1–4]. PQC also is a factor in cancer. During tumorigenesis, many genomic changes occur, including aneuploidy, that place a high demand on cellular mechanisms for coping with proteotoxic stress, including PQC[5–7].

The importance of PQC in cancer is highlighted by the use of proteasome inhibitors to treat multiple myeloma (MM)[8,9]. The success of proteasome inhibitors as a therapy for MM has spurred interest in developing a deeper understanding of the significance of PQC to the pathogenesis of cancer, and in identifying other critical mediators of PQC that might serve as alternative targets for therapy of cancer[8].

VCP (also known as p97), a homohexameric AAA ATPase, participates in multiple PQC pathways, including ribosome-, mitochondria-, and endoplasmic reticulum-associated degradation (ERAD)[10–15] and mediates degradation of proteins that misfold due to stress from heat or oxygen radicals[16]. VCP also has been implicated in processing of protein aggregates and stress granules via autophagy[17–20]. The implication of VCP in PQC mediated by both the ubiquitin proteasome system (UPS) and autophagy suggests that it may serve as a critical node that orchestrates cellular PQC.

The function of VCP is modulated by 'primary' binding proteins, of which about two dozen well-validated partners are known[21–23]. Of particular interest is a set of adaptors that are thought to serve as specificity factors that link substrates to VCP. These adaptors often contain a VCP interaction motif such as a UBX or PUB domain, and occasionally contain an ubiquitin-binding domain[24,25]. The N-terminal region (N domain) of VCP can potentially bind 13 different UBX domain adaptors as well as UFD1L–NPLOC4[24]. In several cases, an adaptor plays an important role in linking VCP to a specific substrate[23,26–29]. However, as a whole the adaptors remain poorly understood. Relatively few adaptor–substrate pairs are known, and in addition little is known about how exactly the adaptors work, including the

dynamics of their recruitment and dissociation, and whether these processes are regulated by substrates or other factors. Besides substrate adaptors, VCP also binds enzymes that are thought to act upon VCP substrates. These include ubiquitin chain-extending and chain-trimming enzymes that bind to internal regions on VCP, as well as peptide-N-glycanase which binds near the C-terminus[30].

VCP depletion is toxic to cancer cells[31,32] but is well-tolerated by primary hepatocytes[33] and skeletal muscle *in vivo*[34], suggesting that the VCP network may be a good target for cancer therapy[8]. To explore the potential of VCP as a target for cancer therapy, we developed DBeQ and ML240, which are reversible, competitive inhibitors of VCP ATPase activity[31,35]. Subsequent optimization of ML240 yielded CB-5083[36], which is currently being tested in phase I clinical trials. CB-5083 is a potent inhibitor of the PQC functions of VCP, and as a consequence triggers a massive unfolded protein response that culminates in activation of apoptosis[37]. Recently, it was reported that the inhibitory action of ML240 is blunted by the VCP adaptor NSFL1C/p47[38]. This suggests that VCP inhibitors may have selective effects on different complexes, and motivates efforts to better understand the assembly state of VCP i cells, and the impact of VCP inhibitors on its assembly state.

New VCP functions and substrates have been sought through the application of affinity purification-mass spectrometry to identify proteins that bind to either VCP or its adaptors[21,23,28,39]. These efforts have identified multiple new functions for VCP, linking it to cullin–RING ubiquitin ligases, endosomal sorting, and ciliary biogenesis. However, this approach is potentially hampered by the dynamics of adaptor–VCP and substrate–adaptor

interactions, which remain poorly understood. Here, we take a multi-pronged strategy that combines immunoprecipitation (IP)-mass spectrometry, cross-linking, and size exclusion chromatography - mass spectrometry to study VCP-interacting proteins. These studies revealed that the interaction of VCP with its adaptors is exceptionally dynamic and can be modulated by chemical inhibitors, which we have verified and quantified by direct biophysical studies.

Experimental Procedures

Mamalian cell culture HEK293 cells (ATCC) and BJ fibroblasts were maintained in Dulbecco's Modified Eagle Medium (DMEM) (Sigma) supplemented with 10% heat inactivated FBS, 100 μg/ml streptomycin, and 100 IU/ml penicillin in 5% CO2 at 37ºC. Cells were washed with PBS, trypsinized, collected, and frozen at -80°C for further use.

SILAC labeling of cells For SILAC experiments with HEK293T cells, DMEM lacking arginine and lysine was supplemented with 10% dialyzed fetal bovine serum (FBS), 1% L-glutamine, 1% Pen/Strep and 1 mM sodium pyruvate. Media used for cells grown in "heavy" was supplemented with 50 mg/L of $^{13}C_6$ $^{15}N_2$-lysine and $^{13}C_6$-arginine (Cambridge Isotope Laboratories) and 10 mg/L of unlabeled proline, while media used for the growth of "light" cells was supplemented with 50 mg/L of unlabeled lysine and arginine and 10 mg/L of proline. Cells were grown in their respective media until incorporation of heavy amino acids reached 98%, as determined by MS/MS analysis of derivatized amino acid

hydrolysate.

CRISPR/Cas9-assisted chromosomal knock in. A donor plasmid (RDB3052) was designed to introduce the FLAG epitope tag sequences into the 5′ end of the VCP coding sequence by homologous recombination. The selection cassette and sequences from the 5′ end of the VCP gene comprising the FLAG sequence flanked by 300bp upstream and downstream of the VCP 7 initiation codon were connected by the 2A sequence[40]. The modified PX330 plasmid[41] (RDB3053) was used to cut the chromosomal VCP locus at the 5′ end of the open reading frame. The CRISPR protocol was then followed[42] and mutant cells were selected for growth in the presence of 100 μg/ml Zeocin for one week. Surviving colonies were assayed by PCR and Western blotting.

In-cell Cross-linking. In-cell cross-linking was performed using Dithiobis [succinimidyl propionate] (DSP) (Thermo Scientific). For each experiment, DSP was freshly prepared as a 200 mM stock solution in dimethyl sulfoxide (DMSO) and diluted to the appropriate final concentrations in phosphate-buffered saline, pH 7.4 (PBS, Fisher Scientific). Cells were washed twice with PBS at room temperature to remove residual medium and incubated with DSP for 20 minutes at room temperature. After removal of the DSP, the cross-linking reaction was quenched by incubating the cells with 25 mM Tris-HCl (pH 7.4) at room temperature for 10 minutes. The quenching solution was then removed and the cells were frozen and stored at -80°C.

Experimental Design and Statistical Rationale. SILAC experiment for HEK293 cells treated with NMS873 or MG132 include total 4 samples. Each SILAC comparison was done in biological duplicate, and then the mass labels were swapped and the experiment repeated in duplicate again. Label free experiment for BJ fibroblast and HEK293 treated with NMS873 or MG132 includes 6 samples in total. Each treatment including control was done separately with biological duplicates. SEC-MS experiment for HEK293 cells treated with NMS873 includes 3 biological replicates, which contains 16 MS runs each. Two-tailed Student's t-test was applied to analyze chromatographic peak shift and intensity change. Statistical process for individual experiment would be discussed in following sessions.

Size exclusion chromatography for protein lysate fractionation. Cells were collected and lysed by sonication in buffer B (25mM Tris-HCl pH 7.4, 150mM NaCl, 1:10000 diluted protease inhibitor cocktail (Roche)) on ice. The chilled lysates were then centrifuged in a SS34 rotor at 13,200xg for 20 minutes at 4°C. The cleared supernatant was filtered through a 0.45 μm syringe filter before chromatographic separation. All the samples were normalized based on protein concentration. Size exclusion chromatography was performed on a 24 ml Superose 6 column using an AKTA FPLC system and standard purification templates (GE Healthcare Biosciences Uppsala, Sweden). The column was equilibrated and then developed in buffer B until the absorption of the flow-through returned to baseline. Sixty fractions of 400 μl each were collected and every third fraction from #11 through #58 was processed for further mass spectrometry analysis.

Protein digestion of SEC fractions. Frozen elution fractions from the NMS-873 and VCP knockdown experiments were obtained, along with wild-type HEK293 fractions prepared alongside each perturbation. Experiments were performed in biological triplicate. To mitigate significant non-tryptic peptide activity, urea was added to each fraction prior to thawing for a final concentration of 8M. Samples were sonicated until urea was completely dissolved, then incubated with 10 mM tris(2-carboxyethyl)phosphine (TCEP) and 40 mM chloroacetamide (CAA) at room temperature for 10 minutes. Fractions were then pooled by threes, to create twelve final fractions per replicate. Pooled fractions were concentrated on centrifugal filters (Amicon Ultra 0.5 mL Ultracel membrane10 kDa) to 100 μL. Samples were diluted with 50 mM Tris pH 8 to 1.5 M urea, then digested overnight at room temperature with 3 μg trypsin (Promega). Samples were acidified with 0.1% trifluoroacetic acid (TFA), then desalted using SepPak C18 solid-phase extraction (SPE) cartridges (Waters). SPE cartridges were equilibrated with one column volume of 100% acetonitrile (ACN), followed by 0.1% TFA. Acidified samples were loaded on column, followed by washing with three column volumes of 0.1% TFA. Peptides were eluted off the column by the addition of 0.7 mL of 40% ACN with formic acid 0.1% TFA and 0.5 mL 80% ACN with 0.1% TFA. Eluates were dried overnight in a SpeedVac concentrator (Thermo), and resuspended in 75 μL of 0.2% formic acid.

Mass spectrometry analysis of SEC fractions. Peptides were injected on to a reverse-phase column prepared in-house. Approximately 35 cm of 75 μm-360 μm inner-outer

diameter barefused silica capillary, each with a laser pulled electrospray tip, were packed with 1.7 μm diameter, 130 Å pore size, Bridged Ethylene Hybrid (BEH) C18 particles (Waters). Columns were fitted on to a nanoAcquity (Waters) and heated to 55-65°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 100% 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 converted to gas-phase ions by electrospray ionization and analyzed on a Thermo Orbitrap Fusion (Q-OT-qIT, Thermo). Precursor scans were performed from 300 to 1,500 *m/z* at 60K resolution (at 400 *m/z*) using a 5 x 105 AGC target. Precursors selected for tandem MS were isolated at 1 Th with the quadrupole, fragmented by HCD with a normalized collision energy of 30, and analyzed using rapid scan in the ion trap. For some analyses, precursors above 500 *m/z* and signal to noise ratio higher than 1.5 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 35 ms, with an AGC target of 104. Precursors with a charge state of 2-6 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 5 second cycles. The raw files were searched directly against the *Homo sapiens* database with nonredundant entries (20,198 entries) using Andromeda on Maxquant (Version 1.5.2.8). The proteome database also includes four SEC elution standards (bovine

thyroglobulin, bovine γ globulin, chicken ovalbumin, and horse myoglobin). Searches were performed using a precursor search tolerance of 4.5 ppm and a product mass tolerance of 0.35 Da. Search criteria included a static modification of +57.0214 Da on cysteine residues, dynamic modification of +15.9949 Da on oxidized methionine and N-terminal acetylation of + 42.0106 Da. Searches were performed with full tryptic digestion and allowed a maximum of two missed cleavages on peptides analyzed by the sequence database. False discovery rates (FDR) were set to 1% for each analysis.

SEC chromatogram analysis. To determine how each treatment (VCP knockdown and incubation with NMS-873) affected how proteins eluted off the SEC column, we calculated an apex shift metric. We are defining "apex" as the fraction in which the majority of the protein elutes. Apex shift was calculated by generating an average fraction number weight against the percent of the total protein that eluted in each fraction. Fractions were numbered 1 through 12, with fraction 1 containing the largest MW proteins/complexes, and fraction 12 containing the smallest. Percent protein per fraction was calculated by first summing the LFQ intensities across the fractions for each protein, then dividing the LFQ intensity in each fraction by the total LFQ intensity in all fractions. The average apex for each protein was then calculated by multiplying each fraction number by the percent protein per fraction, then summing the results. The sum then represents the fraction number in which the majority of the protein elutes, taking into account the spread of the protein across all fractions as well. To calculate average apex shift, we first averaged the apices across biological replicates,

then subtracted the mean apex fraction of the wild-type from the mean apex fraction of the corresponding treatment: in doing so, a positive apex shift indicated elution of a protein at a lower molecular weight after treatment than in the wild-type, and a negative shift indicated elution at a higher molecular weight after treatment. Pvalues were calculated from the apex fraction numbers across replicates using a two-tailed Student's t-test. Fold-change calculations were made on total protein across all fractions; LFQ intensities were summed across fractions, $\log_2$ transformed, and averaged, then the mean $\log_2$ intensities from the wild-type samples were subtracted from the mean $\log_2$ intensities in their corresponding treatment samples. P-values were calculated from the $\log_2$ LFQ intensities across replicates using a two-tailed Student's t-test.

Western blotting to evaluate fractionation of VCP adaptors upon size exclusion chromatography. Samples from cell lysate were prepared by size exclusion chromatography as before, with the addition of 0.05% Triton X-100 in the lysis buffer. Samples were run on a 4-20% gradient SDS-PAGE Gel (Novex), transferred to a nitrocellulose membrane, blocked with 5% milk in TBS-T, incubated with the appropriate primary and secondary antibodies, and developed using Immobilon Western Chemiluminescent HRP Substrate (Millipore) with film. Antibodies used are anti-VCP rabbit polyclonal (Santa Cruz), anti-UBXN7 rabbit polyclonal (Millipore), anti-UBXN6 mouse monoclonal (gift from Dale Haines), and anti-UFD1L mouse monoclonal (BD Biosciences), goat anti-rabbit HRP conjugate (Santa Cruz), and goat anti-mouse HRP conjugate (Santa Cruz).

Immunoprecipitation/Western blotting/mass spectrometry. Cells were collected and lysed by sonication in buffer A (25mM Tris-HCl pH 7.4, 150mM NaCl, 1mM EDTA, 5% glycerol, 1% NP-40, and 1x protease inhibitor cocktail (Roche)) on ice. Lysates were cleared of debris by centrifugation at 16000 g for 5 mins and normalized based on protein concentration. One mg of lysate was incubated with 0.2 µl antibodies (Anti-VCP, Abcam ab113 or Anti-FLAG, Sigma F1804) at 4°C with rotation for the specified times. The reactions were then supplemented with 20 µl of Protein A/G magnetic beads (Thermo Scientific, 88802) and incubated for 5 minutes at room temperature. The beads were washed twice with 400 µl of buffer A followed by two washes with 25mM Tris-HCl pH 7.4. Immunoprecipitates were subsequently processed for Western blotting or mass spectrometry as described below. To detect immunoprecipitated proteins by Western blotting, bound proteins were eluted by boiling the Protein A/G beads in SDS loading buffer supplemented with 50 mM DTT for 5 min. The eluents were separated on a 10% SDS-polyacrylamide gel and transferred onto a nitrocellulose membrane. The membranes were probed using antibodies against proteins of interest. To detect immunoprecipitated proteins by mass spectrometry, a buffer containing 5 mM dithiothreitol, 8 M urea, and 50 mM Tris-HCl, pH 8 was added to the beads. Reduced cysteines were then alkylated by treating the sample with 15 mM iodoacetamide for 30 min. at room temperature, in the dark. After reduction and alkylation the sample was digested with LysC at a volume ratio of 1:100 (enzyme:sample) for 2h at room temperature. After LysC digestion, the urea was diluted to 2M and trypsin was added at a volume ratio of 1:50 (enzyme:sample) and allowed to digest overnight at 25°C. The next day, tryptic peptides

were desalted using a Sep-pak C18 column (Waters) and lyophilized to dryness.

Mass spectrometric data acquisition. Peptide samples were dissolved in 80 µl of 0.2% formic acid and injected into an EASY II nano-UPLC (Thermo Scientific) system. Reverse phase chromatography was performed using a 15 cm silica analytical column with a 75 µm inner diameter packed in-house with reversed phase ReproSil-Pur C18AQ 3 µm resin (Dr Maisch GmbH, Amerbuch-Entringen, Germany). The mobile phase buffer consisted of 0.2% formic acid in mass spectrometry grade water with an eluting buffer of 0.2% formic acid in 80% CH3CN (Buffer B) run over a linear gradient (8-35% Buffer B, 120 min) and using a flow rate of 350 nL/min. The HPLC system was coupled online with a high-resolution Orbitrap mass spectrometer (LTQ-Orbitrap Elite; Thermo Scientific). The mass spectrometer was operated in the data-dependent mode in which a full-scan MS (from m/z 300-1700 with the resolution of 60,000 at m/z 400) was followed by 20 MS/MS scans of the most abundant ions using collisioninduced dissociation (CID). Ions with a charge state of +1 were excluded and the dynamic exclusion time was set to 60 sec after two fragmentations.

Database search and quantification. The raw files were searched directly against the *Homo sapiens* database with non-redundant entries (20,198 entries; human Swiss-Prot downloaded on Jan, 2014) using Andromeda on Maxquant (Version 1.5.2.8). Peptide precursor mass tolerance was set to 10 ppm, and MS/MS tolerance was set to 0.8 Da. Search criteria included a static modification of +57.0214 Da on cysteine residues, a dynamic modification

of +15.9949 Da on oxidized methionine. Searches were performed with full tryptic digestion and allowed a maximum of two missed cleavages on peptides analyzed by the sequence database. False discovery rates (FDR) were set to 1% for each analysis. For protein-protein interaction analysis, published genetic and physical interactions with VCP were obtained from the BioGRID[43] database, version 3.4.135.

Recombinant Protein Expression and Purification. Full-length VCP was amplified by PCR from human VCP pET_15T[38] and ligated into pET24b using NdeI/SalI to produce a non14 cleavable C-terminal His-tagged construct (RDB3219). For FRET studies, VCP coding sequences were amplified by PCR and ligated into a modified pET28a vector to produce a construct with a non-cleavable C-terminal His-tag and an N-terminal ybbR tag with a short linker (MDSLEFIASKLAGGGS). Human ND1L (1-480) pET24b construct (RDB2945) is previously described[38]. The construct for full-length NSFL1C with a non-cleavable N-terminal His-tag[44] was obtained through Addgene (#21268), and site-directed mutagenesis was used to make a NSFL1C-T370C mutation. Proteins were expressed and purified as described previously[38], with the exception that NSFL1C was expressed in TOP10 cells for 3 hours at 37 °C. For FRET, NSFL1C-T370C was incubated with tetramethylrhodamine-5-maleimide (ThermoFisher) prior to gel filtration to produce NSFL1CTAMRA. For ybbR labeling, Cy5-CoA conjugate and Sfp enzyme were made as described[45]. Thirty micromolar ybbR-VCP was incubated for at least 3 hours at room temperature with 60 μM Cy5-CoA conjugate and 12 μM Sfp in 50 mM HEPES pH 7.4, 10 mM $MgCl_2$ prior to gel filtration. All

proteins were purified on a Superose 6 gel filtration column.

FRET Measurements. All FRET measurements were carried out in 20 mM HEPES, pH 7.4, 100 mM KCl, 3 mM $MgCl_2$, 1 mM TCEP, and 1 mg/mL ovalbumin (Sigma). Nucleotides were optionally present at 2 mM, and inhibitors were optionally present at 15 μM. Equilibrium binding assays were carried out on a FluoroLog-3 (Jobin Yvon), with excitation at 540 nm and emission scan 555-750 nm. Stopped-flow experiments were carried out on a Kintek SF-300X instrument with excitation at 540 nm and a 580/20 emission filter. Data were analyzed using Prism 6 (GraphPad).

Results

Study of VCP complex assembly using size exclusion chromatography–mass spectrometry To investigate VCP's association with its adaptors and other binding partners, we sought to employ size exclusion chromatography (SEC) in a buffer containing no detergent and low salt, with the idea that gentle conditions might preserve the integrity of native protein complexes. For these experiments, we prepared lysates from HEK293 cells that were untreated, supplemented with NMS873 for 6 hours, or depleted of VCP by doxycycline induction of shRNA (**Supplementary Fig. S2.1a**). The lysates were fractionated on a gel filtration column, and every third fraction was analyzed by shotgun mass spectrometry. To evaluate the fractionation behavior of the 7,919 proteins identified in 3 biological replicates of both the control vs. NMS873-treated or control vs. VCP knock-down experiments, we

developed an algorithm that estimated the relative amount of each protein by label-free quantification, as well as the mean fractionation position (apex) on the size exclusion column. A histogram summarizing the apex shifts for all proteins identified in the perturbation experiments as well as an untreated vs. untreated control is shown in **Fig. 2.1a**. The variability in the fractionation behavior of known VCP adaptors, number of identified proteins in each experiment, and changes in protein level observed upon depletion or inhibition of VCP are plotted in **Supplementary Fig. S2.1b-d**. The fractionation behavior of all 8,000 proteins is consolidated in **Supplemental Table S1** and fully listed in **Supplemental Table S2**.

We observed that for both NMS873 treatment and VCP knockdown, there was a bias for proteins increasing in molecular weight (i.e. eluting in any earlier fraction, possibly due to accumulation in larger protein complexes) and decreasing in abundance, with the bias being stronger for the former (**Fig. 2.1b, c**). The known functions of VCP in protein complex disassembly[12,13] suggested we might observe a bias for proteins increasing in molecular weight and abundance following VCP inhibition, as these proteins would remain trapped in complexes (including complexes with VCP in the case of NMS873 treatment) and unavailable to the proteasome for degradation. Some of these proteins may associate with membranes, chromatin, or other sedimentable structures, which, given that our lysis buffer did not include detergent or nucleases, could explain their unexpected apparent decrease in abundance. Proteins that shifted in molecular weight in response to VCP inhibition do not share biological functions (**Supplementary Fig. S2.2a**). Few of the proteins we

identified as changing in molecular weight and abundance following VCP modulation had been previously reported as VCP-associated (**Supplementary Fig. S2.2b, c**). The proteome elution pattern in our SEC experiments generally agrees with a previous dataset generated by Kirkwood and co-workers[46].

Using the dataset in Supplemental Table S1, we generated theoretical chromatograms for VCP and a subset of its adaptors. For this investigation we focused on UFD1L–NPLOC4 and the putative substrate-recruiting UBX domain adaptors that bind the N-terminal domain, of which twelve were identified (theoretical chromatograms for these proteins are shown in **Fig. 2.2a-d** and **Supplementary Fig. S2.3a-c** and the reproducibility of their fractionation across triplicates is shown in **Supplementary Fig. S2.4**). Evaluation of their fractionation behavior in untreated vs. perturbed cells revealed that they could be segregated into four distinct classes as follows.

Class I adaptors, comprising UBXN2A and ASPSCR1, exhibited strong co-fractionation with VCP in untreated cells, and were not affected by NMS873 treatment (**Supplementary Fig. S2.3a**). Class I adaptors appear to have a more stable association with VCP than other adaptors. Moreover, UBXN2A levels were reduced below detection upon VCP depletion (**Supplementary Fig. S2.3a**), suggesting that its stability was dependent upon assembly with VCP. ASPSCR1 was not affected by VCP depletion, but it should be noted that only 80% depletion of VCP was achieved and thus VCP may still have been present in sufficient amounts to saturate ASPSCR1.

UBXN4, UBXN8, and FAF2 define Class II VCP adaptors, which were constitutively

assembled in complexes of higher MW than the VCP peak regardless of whether or not VCP was depleted or inhibited with NMS873 (**Supplementary Fig. S2.3b**; note that UBXN8 was not detected in the experiment with VCP-depleted cells). Thus, the high MW forms of these proteins presumably arose from assembly with proteins other than VCP. Interestingly, UBXN4 and UBXN8 are ER membrane proteins, yet our analysis was performed on cell lysates prepared without detergent. It is not known whether the forms of UBXN4/8 detected here are in small vesicles or are proteolytic fragments released from the ER.

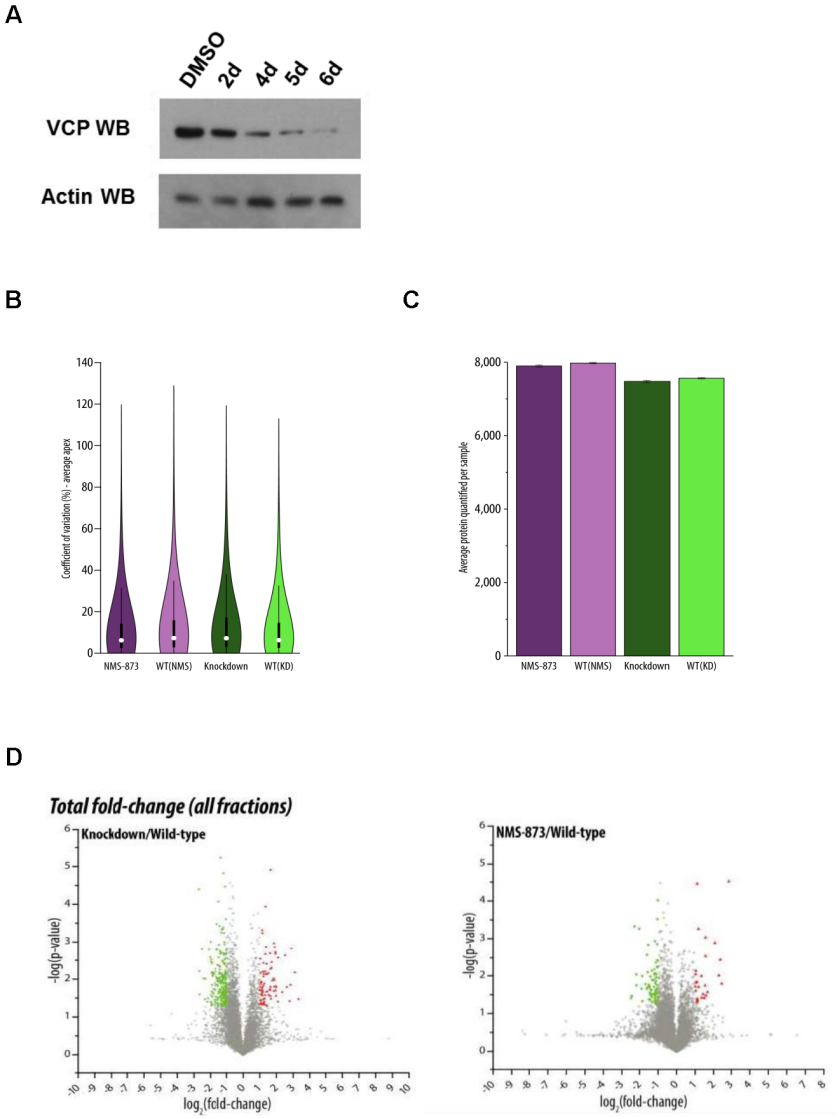
The Class III adaptors UBXN2B, NSFL1C, and UBXN1, migrated at MWs lower than the VCP peak, and were not affected by either chemical inhibition or depletion of VCP **Supplementary Fig. S2.3c**).

The Class IV adaptors UFD1L, NPLOC4, UBXN7, and FAF1 migrated primarily at MWs lower than the VCP peak in unperturbed cells **Fig. 2.2a**), but a substantial fraction of each was recruited into high MW complexes upon inhibition of VCP with NMS873 **Fig. 2.2b**. Formation of these higher MW complexes was not an indirect consequence of a reduction of VCP activity, because a similar shift in MW was not observed in cells depleted of VCP (**Fig. 2.2c,d**). This suggests that NMS873 stabilized formation of high MW complexes between these adaptors and VCP.

The only adaptor that did not fit neatly into Classes I-IV is UBXN6. In the presence of NMS873, UBXN6 shifted to higher MW like UBXN7, but in the depletion experiment it partitioned into high and low MW complexes (**Fig. 2.2a-d**). We note that the theoretical chromatograms for VCP, UBXN6, UBXN7, and UFD1L match the fractionation of these

proteins as determined by western blotting of SEC fractions (**Supplementary Fig. S2.5**). We also note that retrospective analysis of theoretical chromatograms derived from a dataset generated by Kirkwood et al.[46] (**Supplementary Fig. S2.6**) yielded results for unperturbed cells similar to ours with the exception of UBXN6, which exhibited a high MW peak in our dataset but not theirs, and UBXN4 and UBXN8, which were not detected in their dataset, most likely because they did not analyze very high MW fractions (e.g., as shown in **Supplementary Fig. S2.6** Kirkwood et al. did not capture the entire VCP peak, whereas we did, as shown in **Supplementary Fig. S2.3a-c**).

VCP–adaptor interactions are exceptionally dynamic The results of the SEC-mass spectrometry took us by surprise, because even well-studied adaptors like UFD1L–NPLOC4, UBXN7, and NSFL1C showed little or no co-fractionation with VCP in unperturbed cells. This suggested to us that VCP might exhibit much more dynamic association with its adaptors than was previously appreciated. To address this question, we performed conventional affinity purification-mass spectrometry experiments. The first approach we took was to immunoprecipitate VCP from HEK293 cell lines in which the genomic locus of VCP was modified using CRISPR/Cas9 technology[47] to encode VCP with a FLAG epitope appended to its N-terminus (HEK293$^{\mathrm{FLAGVCP}}$ cells). Although we successfully obtained cell lines that expressed endogenous VCP bearing an N-terminal FLAG tag ($^{\mathrm{FLAG}}$VCP, **Supplementary Fig. S2.7a**), we were unable to obtain cell lines in which all alleles of VCP were tagged, suggesting that this might be a lethal event. Shotgun mass spectrometry following im-



Supplementary Figure S2.1: Analysis of VCP complexes by SEC-MS. A, Time course of efficiency of VCP knockdown induced by doxycycline (1 µg/ml) using HEK293 cell line that has a stably integrated, doxycycline-inducible VCP shRNA (DTC204). At the indicated time after transfection, cell lysate was prepared and evaluated by SDS-PAGE and Western blotting with anti-VCP. The 5 day point was chosen as the condition for all MS experiments, since on day 6 a large amount of cell death occurred. **B**, Violin plots showing the distribution of coefficients of variation (n=3) of apex measurements for proteins in each sample. **C**, Number of proteins quantified in each experiment (mean ± SEM, n=3). **D**, Total abundance change (defined as [LFQintensity]treated divided by [LFQintensity]control) for all detected proteins in response to VCP knockdown (left panel) or NMS873 treatment (right panel).

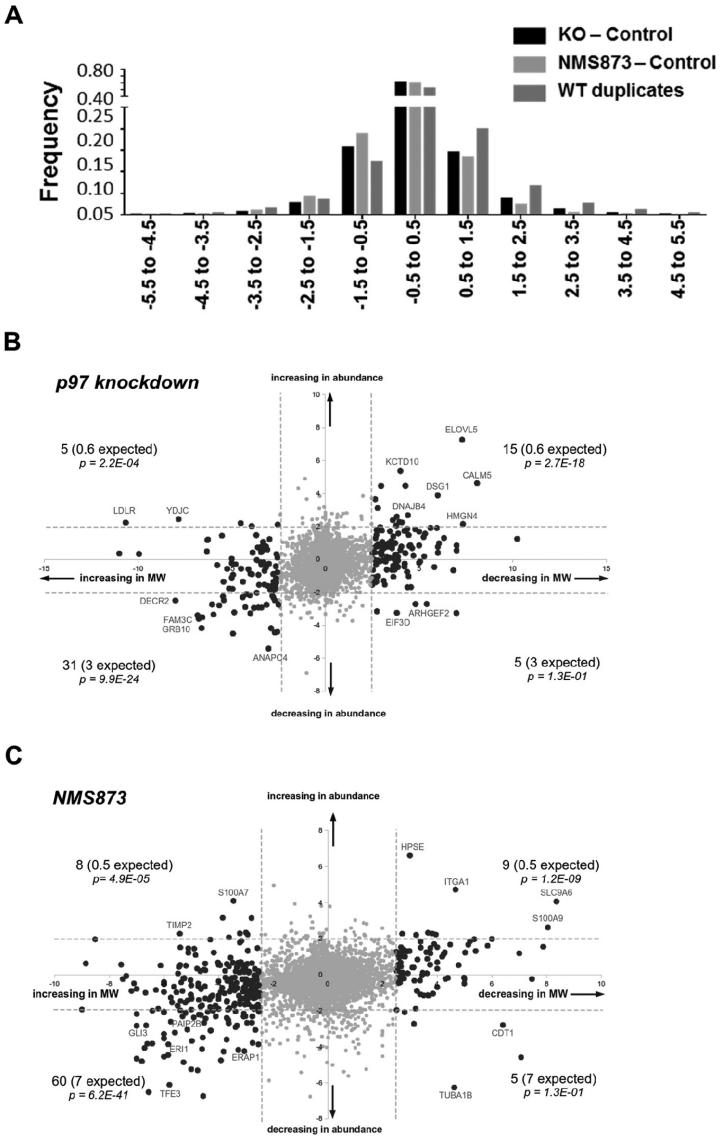
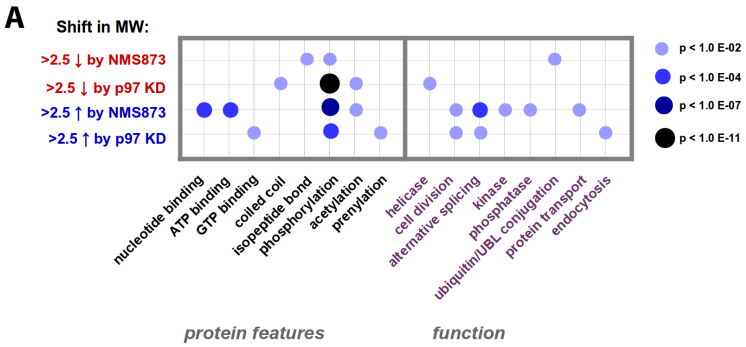


Figure 2.1: Analysis of the fractionation behavior of VCP and its adaptors by SEC-mass spectrometry in response to modulation of VCP activity. **A**, The relative LFQ intensity of peptides from VCP and Class IV adaptors in SEC fractions from untreated HEK293 cells were averaged to generate an estimate of the level of these proteins in each fraction. These levels were then plotted. Quantitative data on the behavior of triplicates for each protein are provided in Supplemental Fig. 4. **B**, same as panel A, except that cells were treated with 10 μM NMS873 for 6 hours prior to lysis. **C**, Same as panel A, except that cells were transfected with control shRNA 96 hours prior to lysis. **D**, Same as panel A, except that cells were transfected with shRNA against VCP sequences 96 hours prior to lysis.



B

>2.5 increase in MW >4 fold increase in abundance	>2.5 increase in MW >4 fold decrease in abundance	>2.5 decrease in MW >4 fold increase in abundance	>2.5 decrease in MW >4 fold decrease in abundance
APLP2	ANAPC4	CALML5	ANK3
LDLR	ANKRD13A	DNAJB4	ARFGEF2
RAB33B	BBS7	DSG1	CAMKK1
SHQ1	BCL11A	ELOVL5	EIF3D
YDJC	BECN1	EPB41L5	TBK1
	CDC16	GALNT7	
	CDC42	GNG5	
	COG8	HIST1H1D	
	CSNK1A1	HMGN4	
	DECR2	KCTD10	
	EDEM3	RNH1	
	FAM160B1	SDAD1	
	FAM3C	SPAG5	
	GRB10	STK11IP	
	ITSN2	WDR91	
	KIAA0391		
	MED31		
	MFN2		
	NGRN		
	NKIRAS2		
	**PTPN9		
	RABGEF1		
	RMND5A		
	SATB2		
	SELH		
	SH3BP1		
	SLC30A9		
	TCP11L1		
	TEX264		
	TTC28		
	TUBAL3		

C

>2.5 increase in MW >4 fold increase in abundance	>2.5 increase in MW >4 fold decrease in abundance	>2.5 decrease in MW >4 fold increase in abundance	>2.5 decrease in MW >4 fold decrease in abundance
AZGP1	AP5S1	ELP5	CDT1
CTDSPL2	ARHGAP8	HPSE	DSC1
DACH2	ARHGEF6	ITGA1	MMAA
IRF6	BCKDHB	S100A9	RSRC1
RBKS	BCL2L1	SGOL1	TUBA1B
RTEL1	BCL2L12	SLC9A6	
S100A7	BCL7B	SON	
TIMP2	C10orf76	TTLL4	
	C11orf83	ZNF771	
	C5orf51		
	C7orf55		
	CBWD2		
	CCDC28B		
	CISD3		
	CRYGD		
	CYB5R4		
	DDX59		
	DSCC1		
	DYNC2LI1		
	ERAP1		
	ERI1		
	GLI3		
	IFIT5		
	INPP4A		
	ITPK1		
	KDM1B		
	KPTN		
	LENG9		
	LIN9		
	LRRC42		
	LZTS2		
	MALSU1		
	MAP2K7		
	MID1IP1		
	MOB3B		
	MTO1		
	NEK3		
	OSBPL2		
	PAIP2B		
	PGM5		
	POLR1C		
	PTPN14		
	**PTPN9		
	RAB34		
	RARS2		
	RCHY1		
	RPS6KB2		
	SCAI		
	SDCBP		
	SFR1		
	SGK3		
	SKA1		
	SNX21		
	STK39		
	TFE3		
	TRADD		
	TRMT13		
	UBE2E1		
	XRCC3		
	ZFPM1		

Supplementary Figure S2.2: Analysis of VCP complexes by SEC-MS. A, Time course of efficiency of VCP knockdown induced by doxycycline (1µg/ml) using HEK293 cell line that has a stably integrated, doxycycline-inducible VCP shRNA (DTC204). At the indicated time after transfection, cell lysate was prepared and evaluated by SDS-PAGE and Western blotting with anti-VCP. The 5 day point was chosen as the condition for all MS experiments, since on day 6 a large amount of cell death occurred. **B**, Violin plots showing the distribution of coefficients of variation (n=3) of apex measurements for proteins in each sample. **C**, Number of proteins quantified in each experiment (mean ± SEM, n=3). **D**, Total abundance change (defined as [LFQintensity]treated divided by [LFQintensity]control) for all detected proteins in response to VCP knockdown (left panel) or NMS873 treatment (right panel).

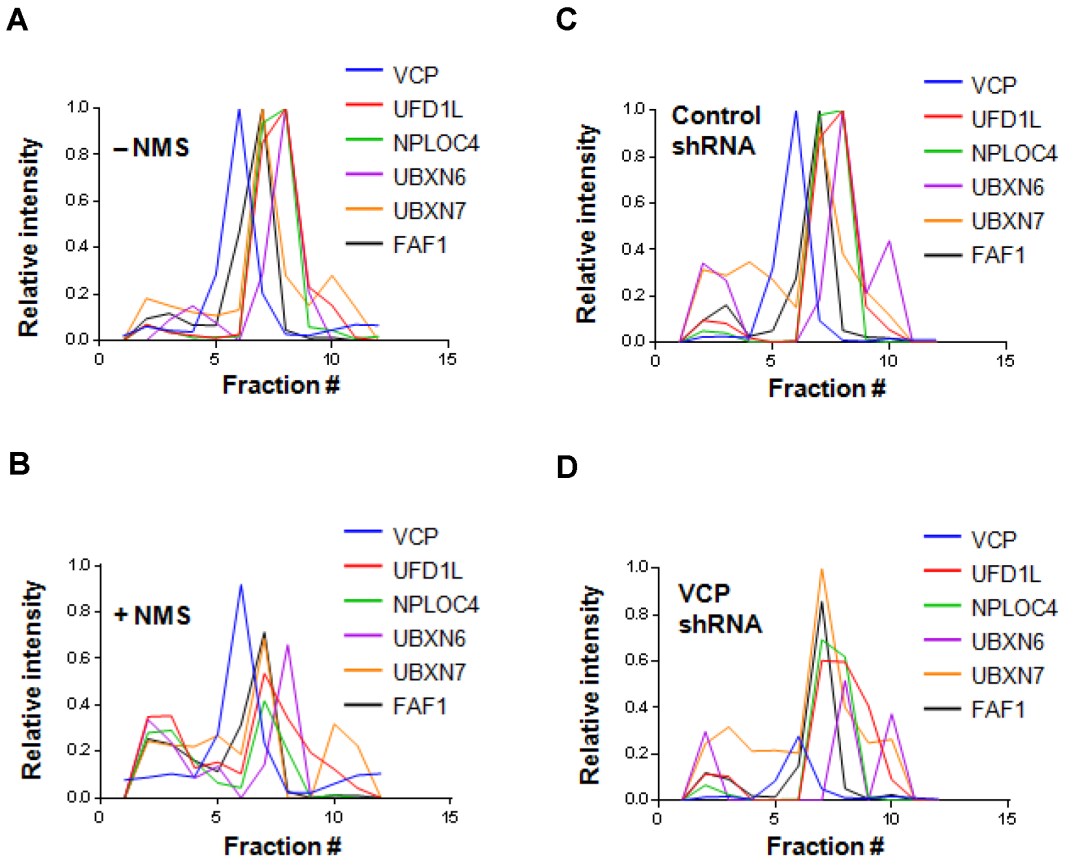
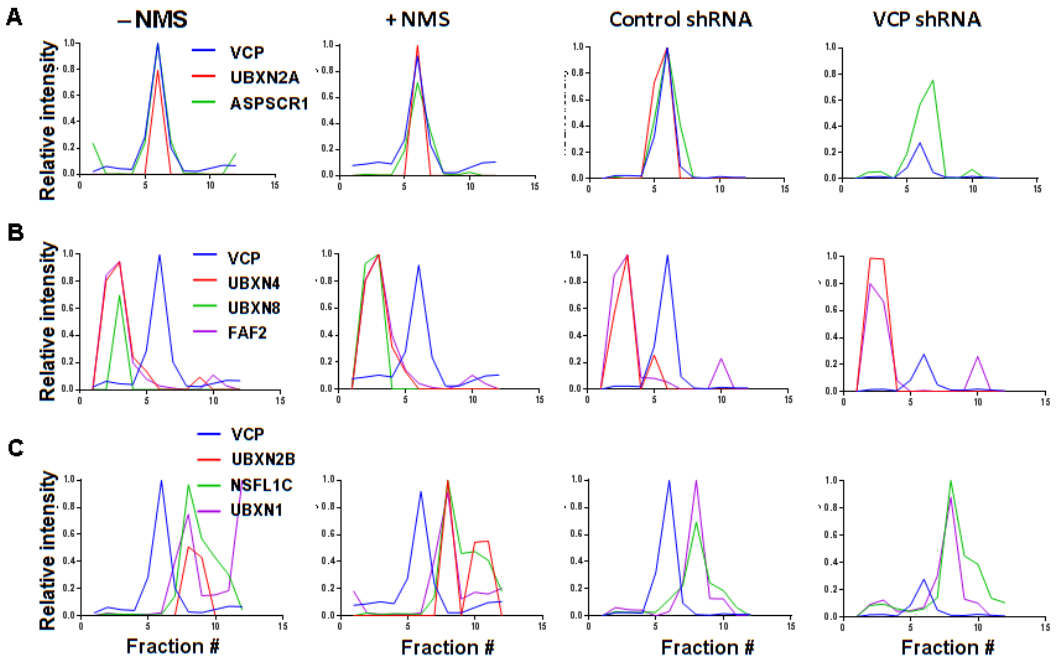
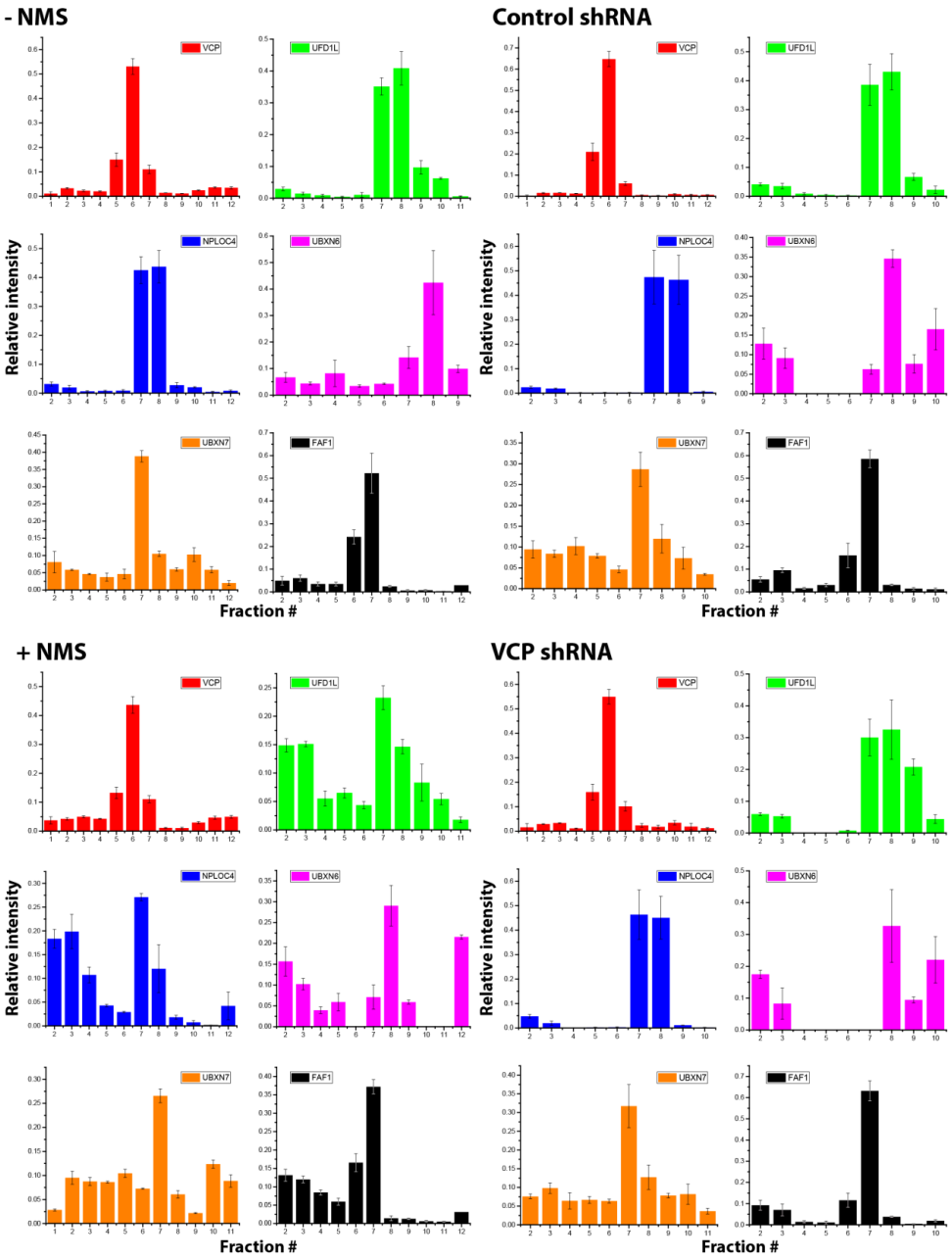


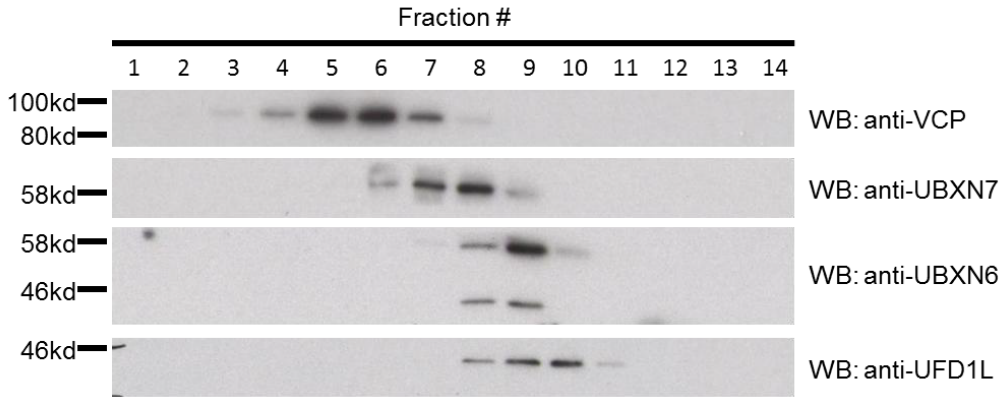
Figure 2.2: Analysis of the fractionation behavior of VCP and its adaptors by SEC-mass spectrometry in response to modulation of VCP activity. **A**, The relative LFQ intensity of peptides from VCP and Class IV adaptors in SEC fractions from untreated HEK293 cells were averaged to generate an estimate of the level of these proteins in each fraction. These levels were then plotted. Quantitative data on the behavior of triplicates for each protein are provided in Supplemental Fig. 4. **B**, same as panel A, except that cells were treated with 10 μM NMS873 for 6 hours prior to lysis. **C**, Same as panel A, except that cells were transfected with control shRNA 96 hours prior to lysis. **D**, Same as panel A, except that cells were transfected with shRNA against VCP sequences 96 hours prior to lysis.



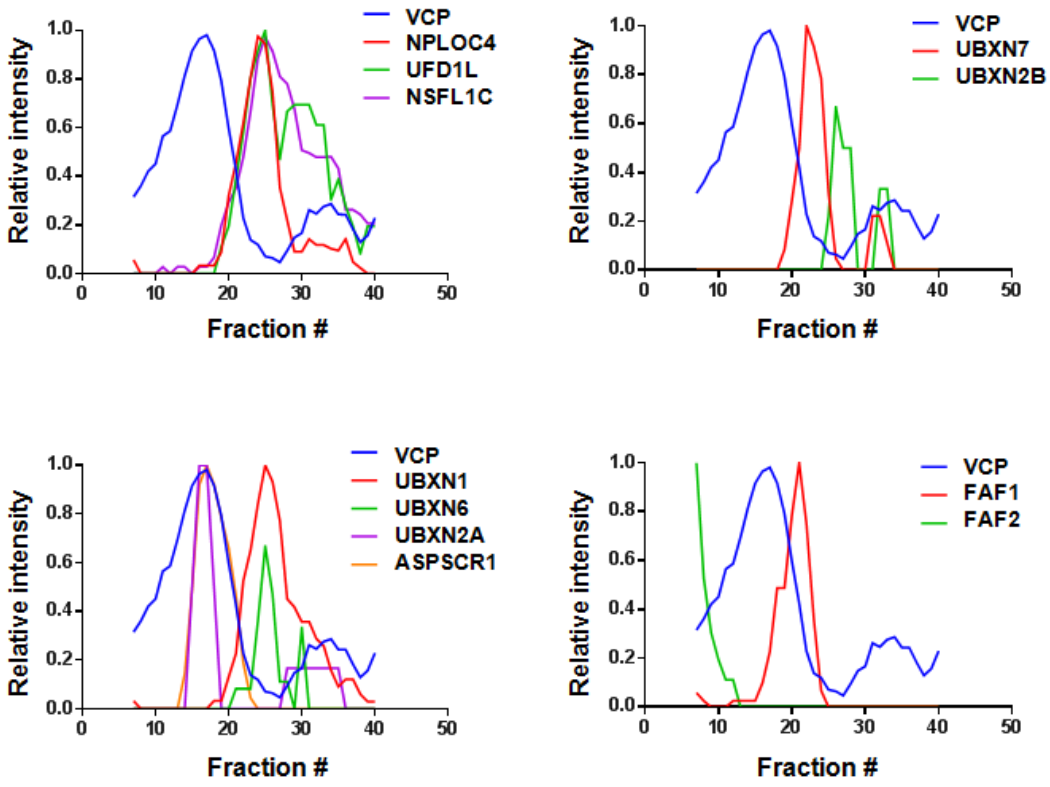
Supplementary Figure S2.3: Fractionation of VCP adaptors by SEC. SEC fractionation behavior of different classes of VCP adaptors upon perturbation of VCP activity by NMS873 or shRNAmediated depletion. **A**, Class I adaptors co-fractionated with VCP. **B**, Class II adaptors were constitutively assembled in complexes of higher MW than the VCP peak. **C**, Class III adaptors fractionated at MWs lower than VCP and were not affected by either chemical inhibition or depletion of VCP.



Supplementary Figure S2.4: Fractionation of VCP adaptors by SEC. The relative LFQ intensities of VCP and adaptors in SEC fractions from HEK293 cells (bars represent standard error of the mean, n=3).



Supplementary Figure S2.5: Fractionation of VCP adaptors by SEC. HEK293 lysate from untreated cells was chromatographed on a Superose 6 column, and 1 mL fractions were analyzed via western blot.



Supplementary Figure S2.6: Fractionation of VCP adaptors by SEC. *In silico*-generated SEC chromatograms of representative VCP adaptors in the Kirkwood et. al. dataset.

munoprecipitation (IP) of endogenous ^{FLAG}VCP retrieved with anti-FLAG resin consistently yielded hundreds of proteins. To determine how many of these proteins were likely to be specific VCP binding partners as opposed to nonspecific contaminants, we performed a control experiment in which lysate from HEK293^{FLAGVCP} cells grown in medium containing 'heavy' lysine and arginine was mixed with lysate from untagged HEK293 cells grown with 'light' lysine and arginine (referred to here as a 'purification after mixing' (PAM)-SILAC protocol)[48]. Anti-FLAG IP for 60 minutes followed by mass spectrometry revealed that a large fraction of the identified proteins had log2 H/L ratios between -1 and 1 (**Supplemental Table S3**), which is a range that many false positives fall in due to analytical error[49]. However, given the poor co-fractionation of adaptors observed in the SECmass spectrometry experiments as well as results from unrelated studies in which we observed a remarkable level of exchange of F-box proteins during IP of Cul1 (J.M.R., unpublished data), we wondered whether the small number of proteins with H/L $\log_2$ ratios >1 might be due to exceptionally dynamic association of VCP with its physiological partners. To assess this possibility, we repeated the analysis, but in this case we performed a 'mixing after purification' (MAP)-SILAC experiment[48] in which tagged 'heavy' and untagged 'light' samples were subjected to IP separately and then mixed immediately before LC-MS/MS analysis **Fig. 2.3a**. This revealed that a large fraction of the proteins identified in the IP from HEK293^{FLAGVCP} lysate had a H/L ratio >1 and thus were enriched relative to the IP from lysate of control cells. Reducing the IP duration from 60 minutes to 30 or even 5 minutes in the PAM-SILAC protocol modestly reduced the recovery of ^{FLAG}VCP and adap-

tors (**Supplementary Fig. S2.3b, c**), but did not significantly attenuate exchange (**Fig. 2.2b**, **Supplemental Table S3**). To investigate this further, we repeated the experiment but with shorter IP intervals ranging from 10 minutes down to the time it took to invert a tube five times to mix its contents ("tube flipping"), prior to collecting the immune complexes in a 1 minute incubation with protein A/G resin. Following the IP step, mass spectrometry was performed and the SILAC ratios of a subset of the known VCP adaptors were evaluated. For this analysis, we focused our attention on UFD1L–NPLOC4 and the UBX domain adaptors that bind the N-terminal domain and are thought to promote substrate recruitment. The extremely short IPs reduced the efficiency of ^{FLAG}VCP recovery (although less than might be generally assumed; Supplemental Fig. S3B, C). Despite the use of extremely short IP intervals, we were able to unambiguously measure SILAC ratios for ^{FLAG}VCP and six of its putative substrate adaptors across all time points. Whereas the SILAC ratios of ^{FLAG}VCPP peptides were high, suggesting that protomers within VCP hexamers did not exchange rapidly, the peptides recovered from VCP adaptors had ratios close to 1 suggesting very rapid equilibration of the 'heavy' and 'light' adaptors with 'heavy' ^{FLAG}VCP (**Fig. 2.2c**, Supplemental Table S4). A notable exception is ASPSCR1, which was one of the two adaptors that showed strong co-fractionation with VCP by SEC (**Supplementary Fig. S2.3a**). Thus, we conclude that the association of ^{FLAG}VCP with most N domain-binding adaptors (and potentially the substrates that bind to those adaptors) is extraordinarily dynamic, and thus what one observes by SEC or in a conventional IP is probably not an accurate reflection of the complexes that exist in the cell. Although the IP and SEC experiments

both pointed to the same conclusion, we had some concern that the epitope tag used for the IP experiments might destabilize the association of ^{FLAG}VCP with its adaptors. To address this issue, we first compared the proteins recovered after the ^{FLAG}VCP IP to those recovered after IP of endogenous untagged VCP with an anti-VCP antibody that binds to a C-terminal epitope on VCP (note that most of the putative substraterecruitment adaptors bind to the N-terminal domain of VCP, which is far from the location of the antibody epitope in the crystal structure). In this and other experiments (data not shown), we consistently identified more proteins (**Supplementary Fig. S2.7d**, **Supplemental Table S5**), including known VCP adaptors (**Supplementary Fig. S2.7e**), when we used the anti-VCP antibody, suggesting that the N-terminal FLAG tag might indeed reduce the binding of some adaptors. Thus, for future experiments, we employed the anti-VCP antibody.

VCP–adaptor interaction is largely due to direct binding To evaluate whether untagged VCP also exhibits dynamic association with its adaptors, we needed to develop a method to block association of adaptors with endogenous VCP during the IP step, since it was not possible to perform a conventional PAM-SILAC experiment. Because the epitope recognized by the anti-VCP antibody is in the C-terminal region, we expressed and purified a truncated form of VCP that incorporates the N and D1 domains as well as a linker between the D1 and D2 domains (ND1L). ND1L should bind UFD1L–NPLOC4 and VCP adaptors with a UBX domain[50] but as expected, it did not bind the anti-VCP antibody (**Fig. 2.4a**). Thus, ND1L added to cell lysates should behave as a passive 'sponge' that soaks up any adaptors that

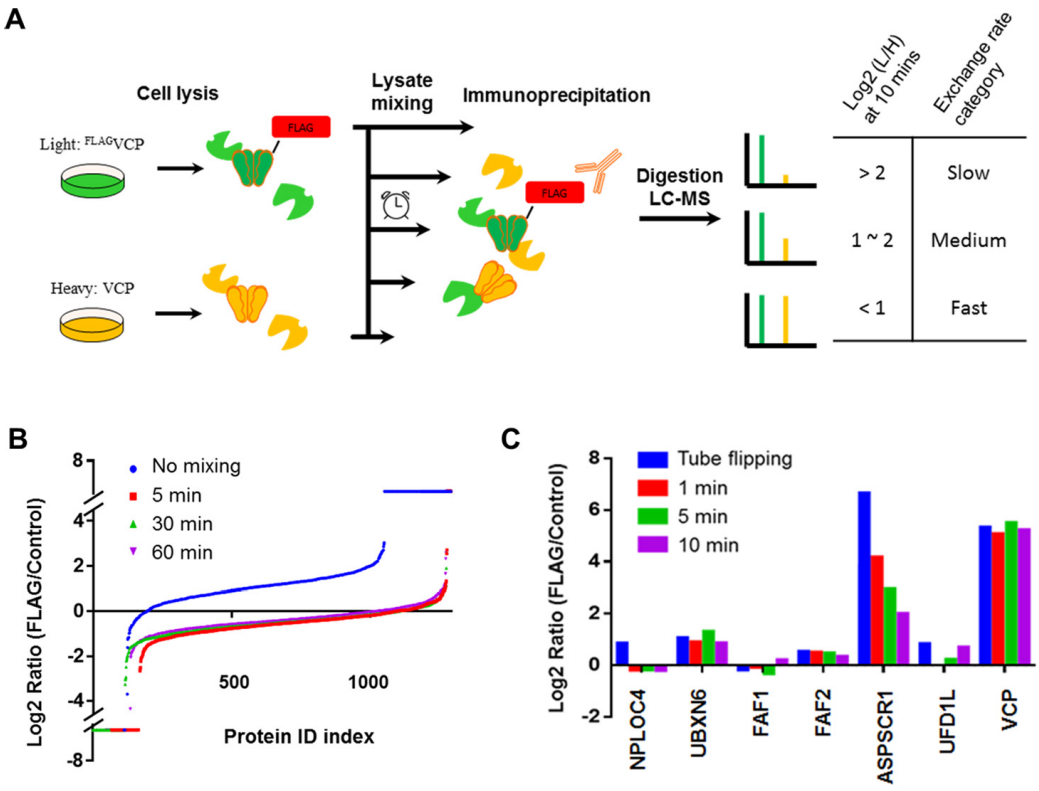
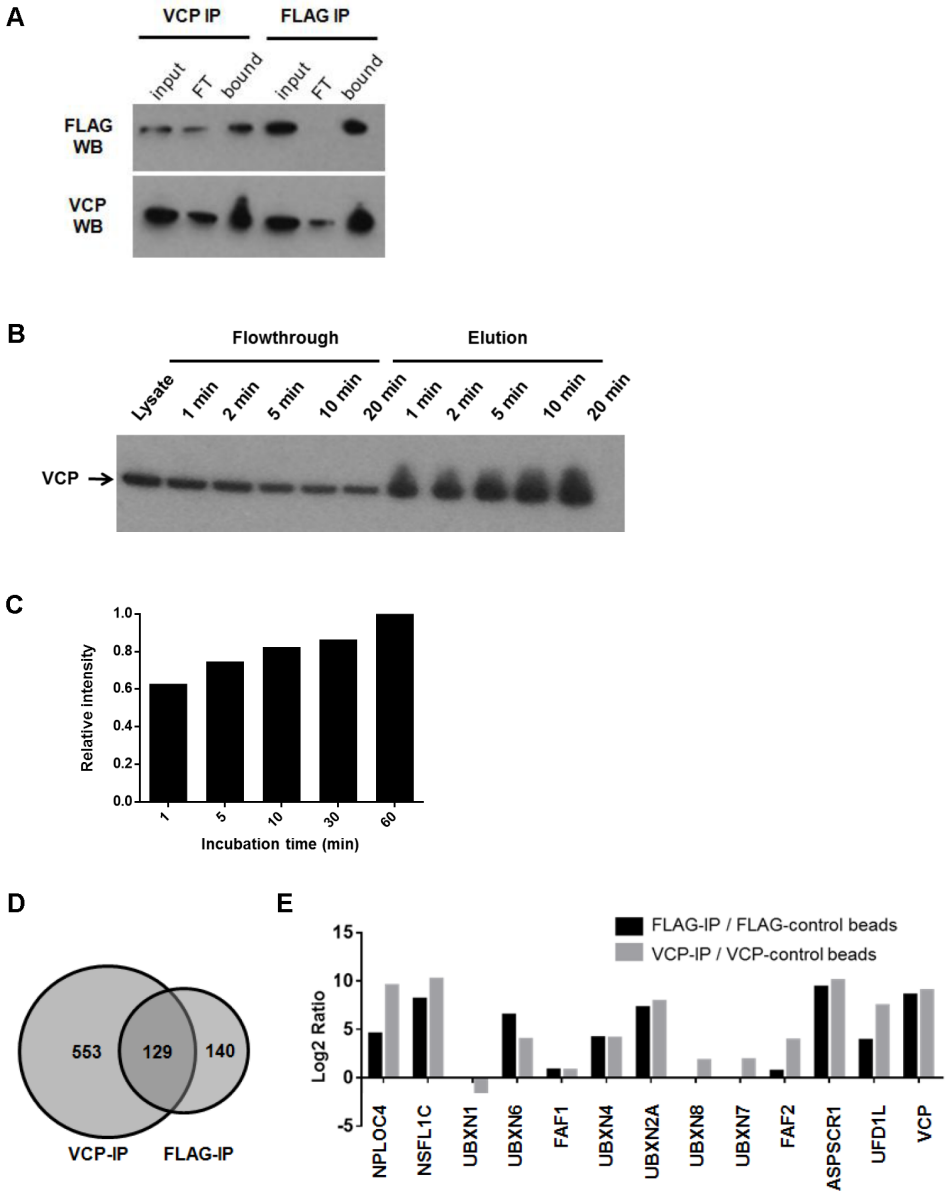


Figure 2.3: VCP–adaptor complexes undergo rapid exchange during IP from cell lysate. **A**, Schematic workflow of IP-MS experiment for studying VCP adaptor exchange. **B**, Global protein exchange during IP. HEK293FLAGVCP and HEK293 cells were grown in medium containing 'heavy' or 'light' lysine plus arginine, respectively. The cell lysates were mixed and incubated with anti-FLAG antibody for the indicated amount of time prior to collecting immune complexes on protein A/G resin for 5 min. For "No mixing", samples were mixed after the IP step. n=1 for all treatments. **C**, A repeat of the experiment in B, except that shorter IP times were evaluated and the antibody capture interval was reduced to 1 min. "Tube flipping" corresponds to 5 inversions of the capped tube. H/L ratios are reported for VCP and the indicated adaptor proteins. n=1 for all treatments.



Supplementary Figure S2.7: Control experiments for VCP immunoprecipitation. A, Western blotting shows the expression of endogenous VCP bearing an N-terminal FLAG tag in HEK293 cells. Lysates of HEK293^{FLAGVCP} cells were incubated with anti-VCP or anti-FLAG antibody. The input, flow through (FT) and bound fractions were separated by SDS-PAGE and blotted with anti-VCP or anti-FLAG antibody. Note that the FLAG tag did not cause a perceptible shift in the mobility of VCP. WB, Western blot.

Supplementary Figure S2.7: B, VCP was recovered in IPs as short as 6 min. The time indicated above each lane is the duration of the incubation with antibody prior to adding protein A/G beads for an additional 5 min to capture immune complexes. As the IP time increased, the amount of VCP detected by Western blot decreased in the flow through and increased in the elution. **C**, Label-free quantification of VCP recovery in IPs of varying duration for the experiments in **Fig. 2.3a, b**. **D**, Venn diagram shows the overlap in protein identifications comparing FLAG IP from HEK293FLAGVCP cells with IP of untagged endogenous VCP from wild type HEK293 cells. n=1. **E**, Enrichment of VCP adaptors in VCP pull-downs. Two experiments were conducted in parallel. In the first experiment (black bars), HEK293^{FLAGVCP} and HEK293 cells were grown in medium containing 'heavy' or 'light' lysine plus arginine, respectively. Cell lysates from both cultures were individually subjected to IP with anti-FLAG and following the IP step the samples were mixed and analyzed by mass spectrometry. In the second experiment (gray bars), Lysate from the 'heavy' HEK293 cells was subjected to IP with anti-VCP whereas lysate from the 'light' HEK293 cells was subjected to mock IP (no antibody was present before protein A/G beads added for capture). Following the IP step the samples were mixed and analyzed by mass spectrometry. The H/L ratios for known VCP-interacting proteins are shown. n=1.

are either not bound to VCP or dissociate from VCP during the IP step, thereby preventing them from re-binding to VCP in the cell lysate. However, ND1L should have no effect on the koff of adaptors that are bound to VCP at the time of cell lysis. It should also not be recovered in the IP step, thereby allowing us to capture a record of the VCP complexes that preexisted in cells and survived the IP step (**Fig. 2.4b**). Titration of 'light' ND1L into 'heavy' cell lysate prior to IP with anti-VCP for 1 minute followed by recovery of immune complexes on protein A/G resin for 5 minutes reduced the number of proteins recovered (**Fig. 2.4c**, **Supplementary Fig. S2.8**), but had no effect on the recovery of 'heavy' VCP (**Fig. 2.4d**, **Supplemental Table S6**; for this experiment, only C-terminal peptides were counted to rigorously exclude the possibility that they arose from ND1L). Notably, addition of ND1L caused a strong reduction in the recovery of VCP adaptors, with the

curious exception of ASPSCR1, and to a lesser extent UBXN2A (**Fig. 2.4d**). ASPSCR1 was also resistant to exchange in the PAM-SILAC protocol (**Fig. 2.3c**) and both ASPSCR1 and UBXN2A showed good cofractionation with VCP in the SEC analysis (**Supplementary Fig. S2.3a**). This experiment suggests two important conclusions: (i) ND1L does not dislodge stably bound adaptors from VCP, and (ii) most adaptors that bind the N-terminal domain exhibit extremely dynamic association with untagged endogenous VCP and dissociate during a 6 minute IP.

In-cell cross-linking significantly increases the recovery of VCP adaptors Given that many VCP–adaptor complexes were extremely prone to dissociation and exchange in cell lysate we reasoned that it is not possible to capture a reliable snapshot of the VCP complexes that existed in a cell using conventional chromatography or IP methods. Therefore, we employed a cross-linking agent to 'freeze' VCP complexes, to enable identification of protein interactions that occur in cells. To trap physiological VCP complexes, we added the cross-linking agent DSP to intact cells to stabilize complexes prior to cell lysis. Optimal cross-linking conditions were determined by assessing the formation of high molecular weight (MW) VCP cross-linked complexes as a function of DSP concentration. The lowest concentration of DSP that yielded near-quantitative cross-linking of VCP was 0.8 mM (**Fig. 2.5a**), and this was therefore employed for subsequent experiments. When IP with anti-VCP was performed for 2 hours, most of the cross-linked VCP was recovered in the bound fraction (**Supplementary Fig. S2.9**). SILAC mass spectrometry analysis revealed that cross-

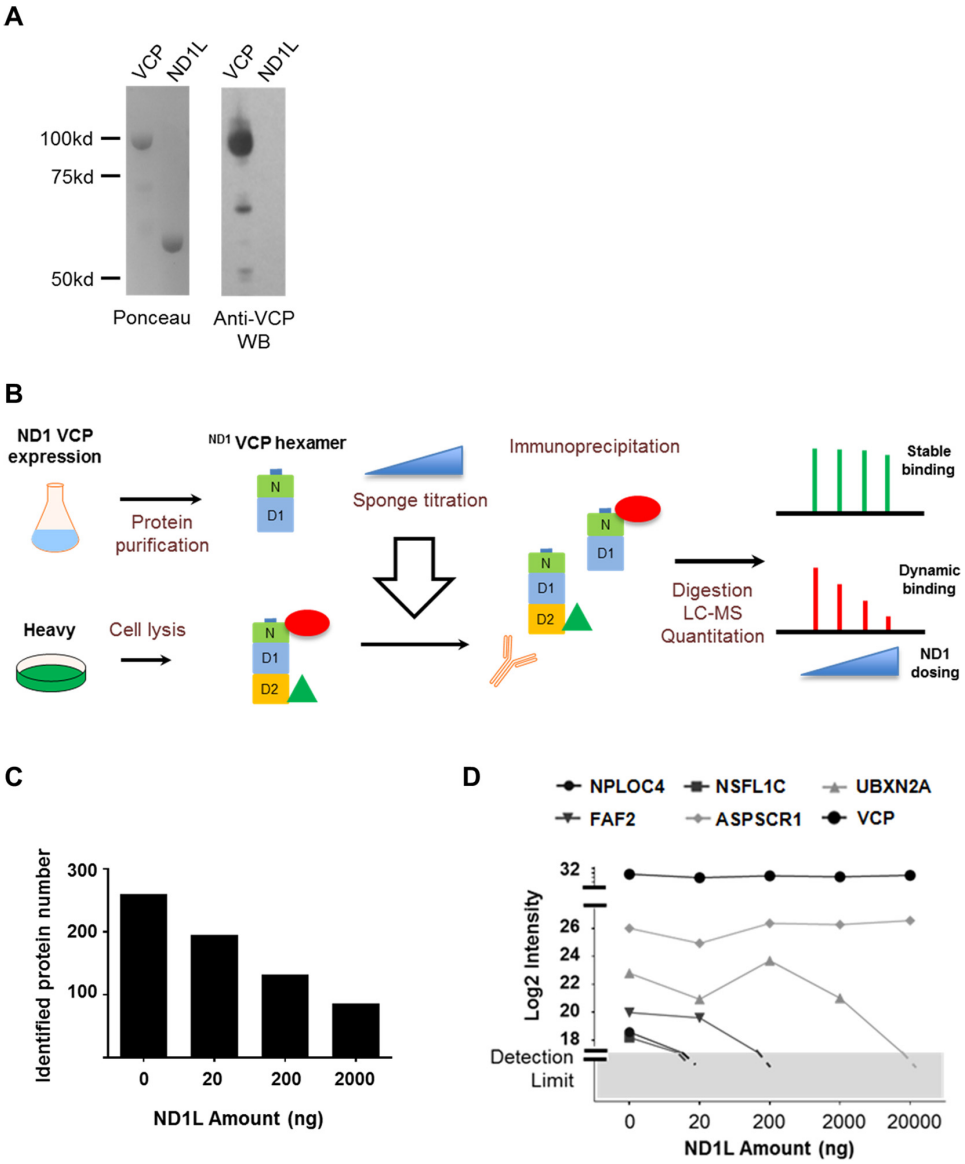
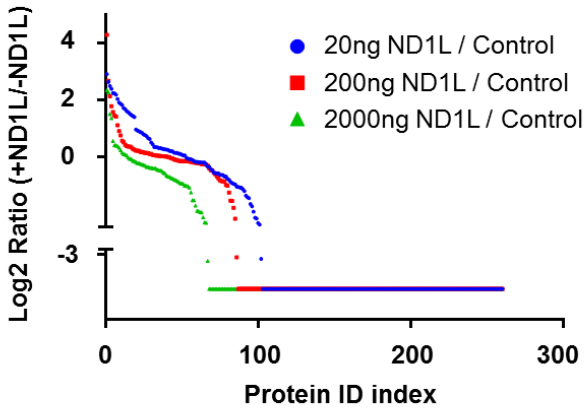


Figure 2.4: VCP adaptors undergo dynamic exchange during IP of untagged endogenous VCP. A, VCP antibody does not bind ND1L. Full-length VCP and the ND1L fragment were fractionated by SDS-PAGE and subjected to Western blotting with anti-VCP. Left panel: Ponceau S staining of the nitrocellulose filter. Right panel: Western blot of the same filter. **B**, Schematic workflow of "Sponge" experiment. **C**, Effect of ND1L competitor on recovery of VCP-interacting proteins. The indicated amounts of purified ND1L were added to 1 milligram of HEK293 cell lysate, which was then subjected to IP for 1 minute with anti-VCP followed by 5 min with protein A/G resin prior to mass spectrometry. n=1 for all treatments. **D**, The relative amounts of individual VCP adaptors identified in the experiment in panel B are plotted. Relative protein amounts were estimated by LFQ value from MaxQuant.



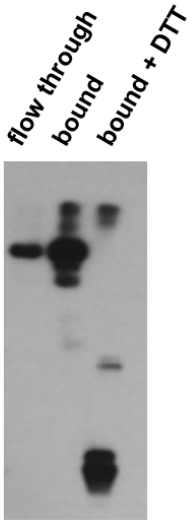
Supplementary Figure S2.8: Effect of ND1L 'sponge' on recovery of VCP binding proteins during immunoprecipitation. Titration of increasing amounts of 'light' ND1L into 'heavy' cell lysate progressively reduced the number of proteins recovered by IP with a VCP antibody that does not bind ND1L.

linking with 0.8 mM DSP yielded much higher relative levels of multiple VCP adaptors (with the exception of the slowly exchanging UBXN2A and ASPSCR1, which already bound relatively stably in the absence of cross-linker) than 0.1 mM DSP (**Fig. 2.5b**). Notably, the adaptor UBXN2B/p37, as well as validated VCP interactors NGLY1, PLAA, and UBE4B were only identified when cells treated with DSP were used for IP with anti-VCP. Given the superior ability of the cross-linking method to identify VCP adaptors, we performed a global SILAC mass spectrometry analysis in HEK293 cells to identify proteins whose association with VCP was modulated by either the allosteric VCP ATPase inhibitor NMS873[32] or the proteasome inhibitor MG132. The rationale behind these experiments is that NMS873 might trap VCP in a state that favors binding of certain adaptor proteins and substrates. By contrast, MG132 should prevent the degradation of VCP substrates destined for the proteasome, allowing them to accumulate on VCP. Each SILAC comparison was done in duplicate, and then the mass labels were swapped and the experiment repeated in duplicate again. Comparison of biological replicates in each case confirmed the robustness of our methodology (**Supplementary Fig. S2.10a, b**; for a full list of IDs, see **Supplemental Table S7**). A similar analysis was done on BJ fibroblasts, although label-free mass spectrometry was used instead of SILAC (see **Supplementary Fig. S2.10c, d** for comparison of replicates and **Supplemental Table S8** for a full list of IDs). For both experiments, the majority of proteins with >2 fold alteration in association with VCP showed increased (as opposed to reduced) binding in the presence of NMS873 or MG132, and many of the proteins that showed enhanced binding in the presence of NMS873 were similarly enriched by MG132 treatment,

suggesting that VCP-interacting proteins are often regulated by the proteasome (**Fig. 2.6a, b**). In general, these VCP-interacting proteins appear to be functionally linked to protein biogenesis and protein quality control, as they are highly enriched for factors involved in nonsense-mediated decay, translation, amino acid metabolism, and co-translational signal recognition particle (SRP) targeting to membranes (**Fig. 2.6c**). Examination of the VCP adaptors revealed that NMS873 had a profound effect on their assembly with VCP: of the 11 N domain-binding substrate adaptors identified in both replicates of the HEK293 experiment, two decreased in amount by at least 2-fold and five increased in amount by at least 2-fold upon inhibition of VCP (**Fig. 2.6d**). Qualitatively similar behavior was observed in the experiment with BJ fibroblasts (**Fig. 2.6d**). The recovery of increased UBXN7 and FAF2 in association with VCP upon NMS873 treatment is all the more remarkable given that NMS873 treatment reduces the steady-state level of these proteins[23]. Inhibition of VCP by NMS873 could stabilize association of some adaptors by at least two conceivable mechanisms. In the first case, NMS873 might stabilize a conformation of VCP that binds more tightly to certain adaptors. In the second case, NMS873 might block degradation of the adaptor or VCP substrates targeted to the UPS, which upon accumulation bind to their cognate adaptors and stabilize their association with VCP. We reasoned that the second hypothesis could be tested by repeating the experiment in the presence of the proteasome inhibitor MG132, which should also block degradation of adaptors and VCP substrates. In this case, we predict that MG132 would enhance adaptor association with VCP. However, we only clearly observed this effect for FAF1 (**Fig. 2.6d**), and to a significantly lesser extent

with UBXN7 (both in HEK293 cells). For the other 3 adaptors whose VCP association was enhanced by NMS873, MG132 had a relatively negligible effect in both the HEK293 cells and BJ fibroblasts. This suggests that NMS873 might act primarily via a direct conformational effect on VCP.

Bioinformatic analyses highlight similar responses to VCP inhibition in different cell types and emphasize the benefit of cross-linking To evaluate the relative performance of the different approaches taken during the course of this work, we performed comparative bioinformatic analyses of our datasets. **Fig. 2.7a** charts the behavior of proteins with known VCP binding motifs (UBX, VBM, PUB, VIM, and PUL domains) as well as a handful of other well-studied VCP-interacting proteins in AP-MS experiments performed with and without cross-linking. Some cell type-specific effects are seen (the most dramatic being the strong association of ERAD components AMFR, SYVN1, and DERL2 with VCP in cross-linked BJ cells treated with NMS873), but for the most part the results are similar across cell types, with NMS873 consistently enhancing VCP association of some adaptors while diminishing the association of others. Notably, both NMS873 and MG132 enhanced association of ubiquitin with VCP (annotated here as RPS27A). Adaptors generally behaved similarly to other adaptors in their subcategory in response to a variety of conditions, including VCP and proteasomal inhibition, as well as during different types of immunoprecipitation. From this, we speculated we might be able to identify candidate substrates of these adaptors by querying for proteins with a similar pattern of variation across the different experiments. **Fig.**



Supplementary Figure S2.9: Crosslinked VCP complex can be purified by immunoprecipitation. When immunoprecipitated with anti-VCP for 2 hours, most of the cross-linked VCP was recovered in the bound fraction. VCP cross-links were largely resolved upon treating the bound fraction with the reducing agent DTT.

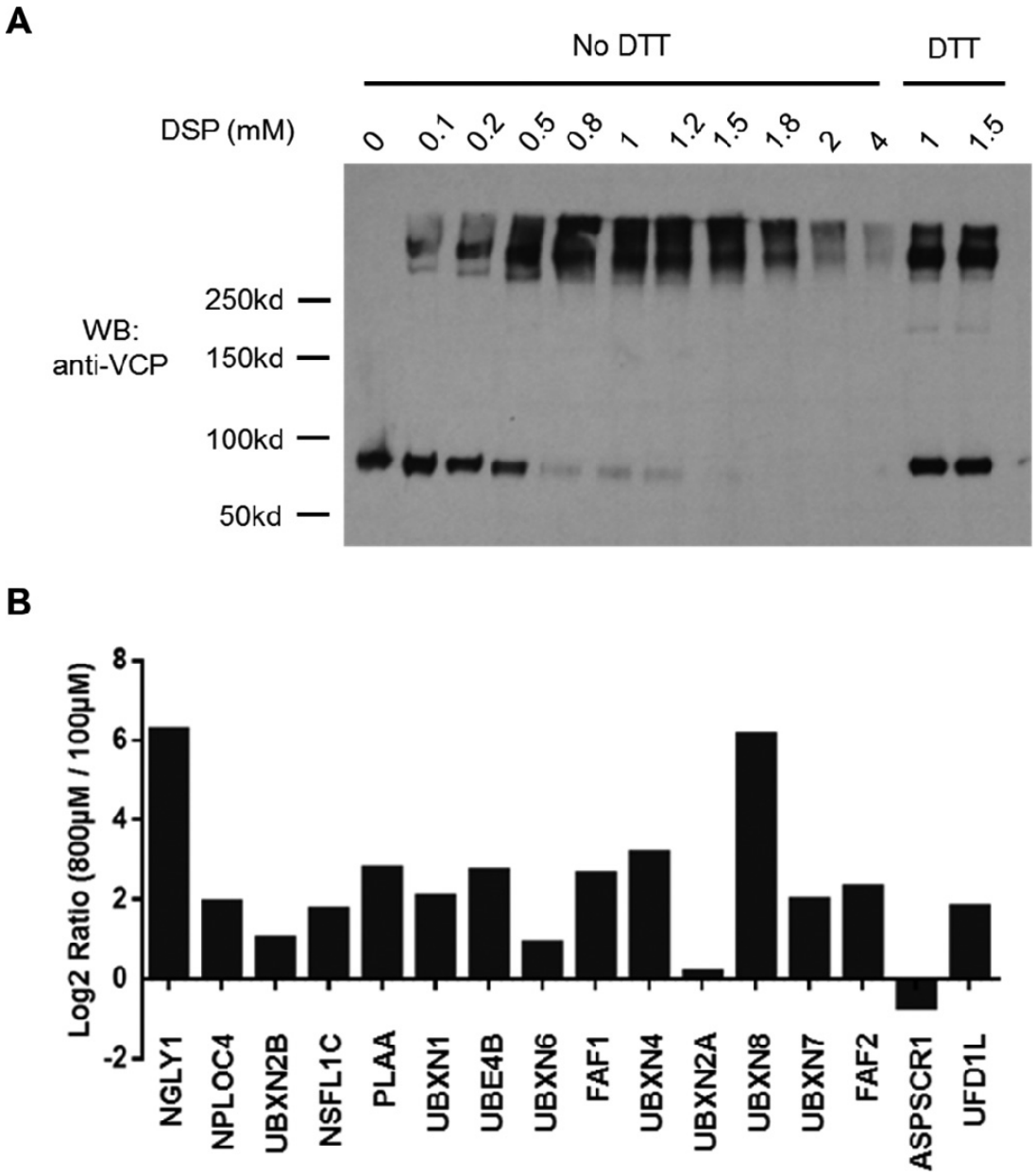
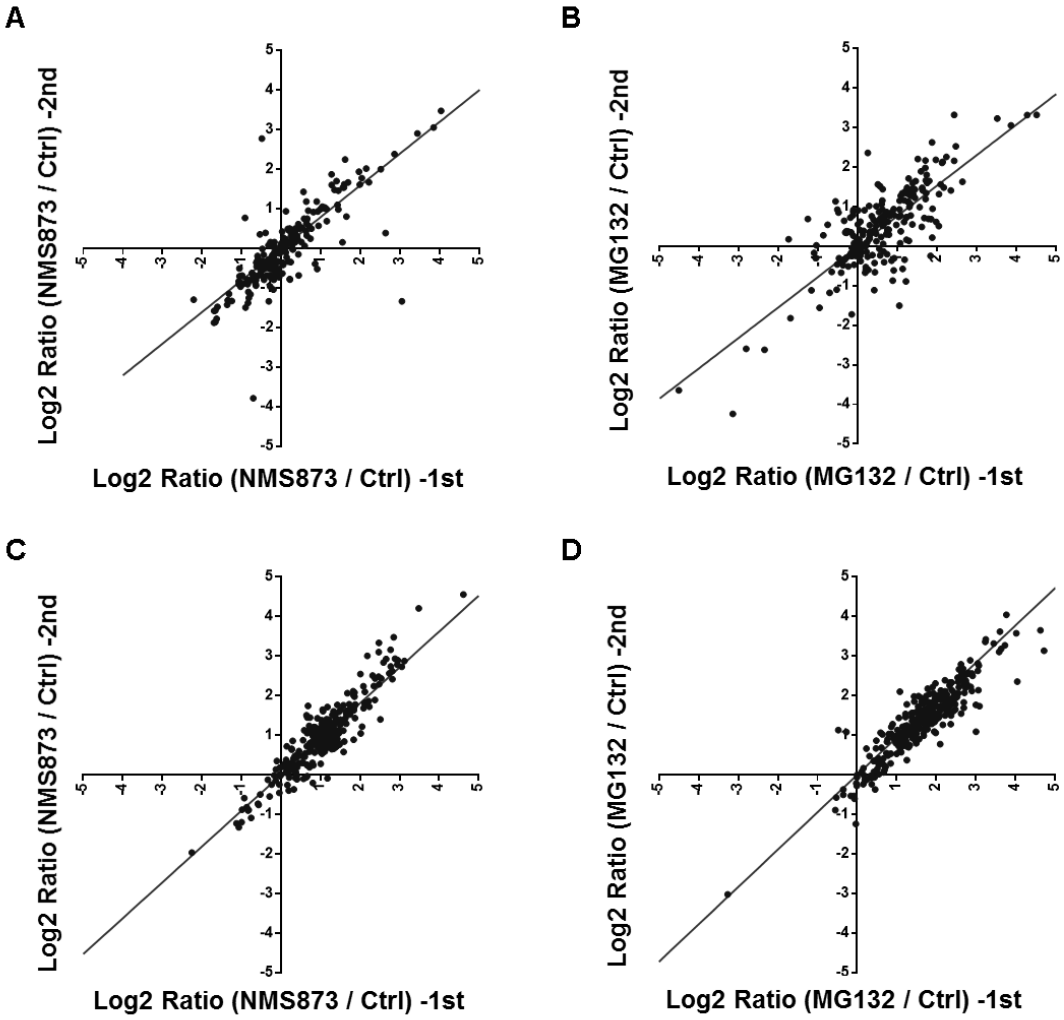


Figure 2.5: The crosslinker DSP stabilizes the interaction of VCP with most of its adaptors. A, DSP was added in the indicated amounts to HEK293 cells in the absence or presence of DTT for 30 min. prior to cell lysis. Cross-linking of VCP was evaluated by fractionation of lysates by SDS-PAGE followed by immunoblotting with anti-VCP. **B**, SILAC analysis was performed with 'heavy'- and 'light'-labeled cells treated with 800 μM and 100 μM DSP, respectively. The H/L ratio is reported for the indicated VCP-binding proteins. n=1 for all treatment.



Supplementary Figure S2.10: Chemical inhibition of VCP modulates its repertoire of associated adaptor proteins in HEK293 cells and BJ fibroblasts. **A**, Label swap SILAC experiments were performed in which cells were either mock-treated or supplemented with 10 µM NMS873 for 6 hours. Cells were treated with 800 µM DSP for 30 minutes prior to cell lysis, mixing of cell lysates, IP with anti-VCP, and mass spectrometry analysis. The ratios for each protein identified in the replicate experiments are plotted on the x and y axes. **B**, same as panel A, except that cells were treated with or without 10 µM MG132 for 2 hours. **C**, Duplicate mass spectrometry experiments with label-free quantification were performed in which cells were either mocktreated or supplemented with 10 µM NMS873 for 6 hours. Cells were treated with 800 µM DSP for 30 minutes prior to cell lysis, mixing of cell lysates, IP with anti-VCP, and mass spectrometry analysis. The ratios for each protein identified in the replicate experiments are plotted on the x and y axes. **D**, same as panel C, except that cells were treated with or without 10 µM MG132 for 2 hours.

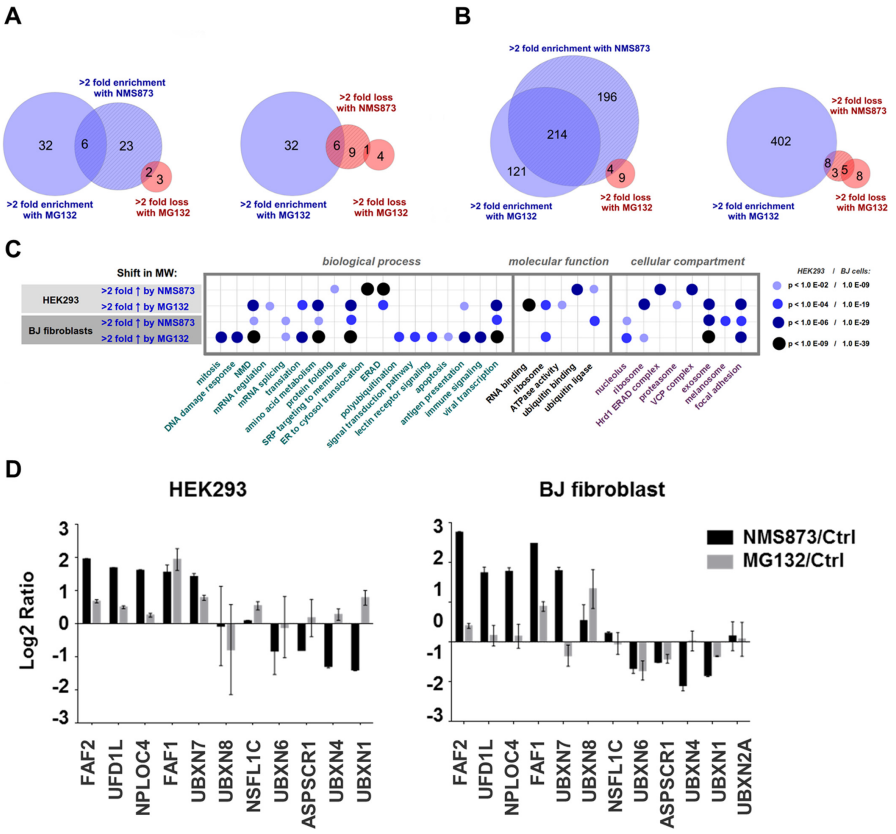


Figure 2.6: Chemical inhibition of VCP modulates its repertoire of associated adaptor proteins. **A**, Effect of NMS873 and MG132 on VCP-associated proteins in HEK293 cells. Duplicate labelswap SILAC experiments (4 analyses in total) were performed in which cells were either mocktreated or supplemented with 10 μM NMS873 for 6 hours, or 10 μM MG132 or vehicle ("control") for 2h. Cells were treated with 800 μM DSP for 30 mins prior to cell lysis, mixing of cell lysates, IP with anti-VCP, and SILAC mass spectrometry analysis. Number of proteins meeting the indicated criteria are depicted.**B**, same as panel **A**, except that BJ fibroblasts were used, all cells were grown in 'light' medium, and relative protein amounts were determined by label-free quantification. **C**, Significant enrichment for gene ontology classes as calculated by PANTHER. All terms with significance below a certain threshold (p< E-02 for HEK293 cells; p <E-09 for BJ cells) are displayed. **D**, Effect of NMS873 or MG132 treatment on recovery of the indicated adaptor proteins in the VCP IPs from HEK293 cells (left panel) and BJ fibroblasts (right panel) is plotted. Only adaptors detected in both replicates for a given cell type are shown. Error bars represent standard error of the mean ratio, n=2.

2.7b (see **Supplementary Fig. S2.11** for gene names) provides a graphical representation of the proteins that significantly covaried with each UBX domain adaptor. Some interesting relationships worthy of further investigation were noted, including significant covariation of UBXN4 with proteasome subunits. To get a better understanding of how different cell types and cell treatments (crosslinking or no cross-linking) compared to each other, we constructed Venn diagrams showing the overlap between different experiments. The upper panel in **Fig. 2.7c** shows that 93% of the proteins in HEK293 cells that showed increased binding to VCP upon NMS873 treatment followed by cross-linking, showed the same behavior in similarly-treated BJ fibroblasts. By contrast, the overlap between the datasets from cross-linked and non-cross-linked HEK293 cells was poor (**Fig. 2.7c**, lower panel). Given that the percentage of hits that is independently validated (by cross-referencing with the BioGRID ([43]) databases) is much higher for the cross-linked dataset, many of the additional hits found in the non-cross-linked sample may be due to higher nonspecific background. In **Fig. 2.7d**, the behavior of the 29 proteins that overlap in the top panel of **Fig. 2.7c** is shown across all experiments performed during the course of this work. Taken together, these data suggest that cross-linking yields datasets with the highest enrichment of validated binding partners (**Supplemental Table S9**). An added benefit of cross-linking is that it enriches for interactions that occur within the cell as opposed to in the lysate, where natural compartment barriers have been breached.

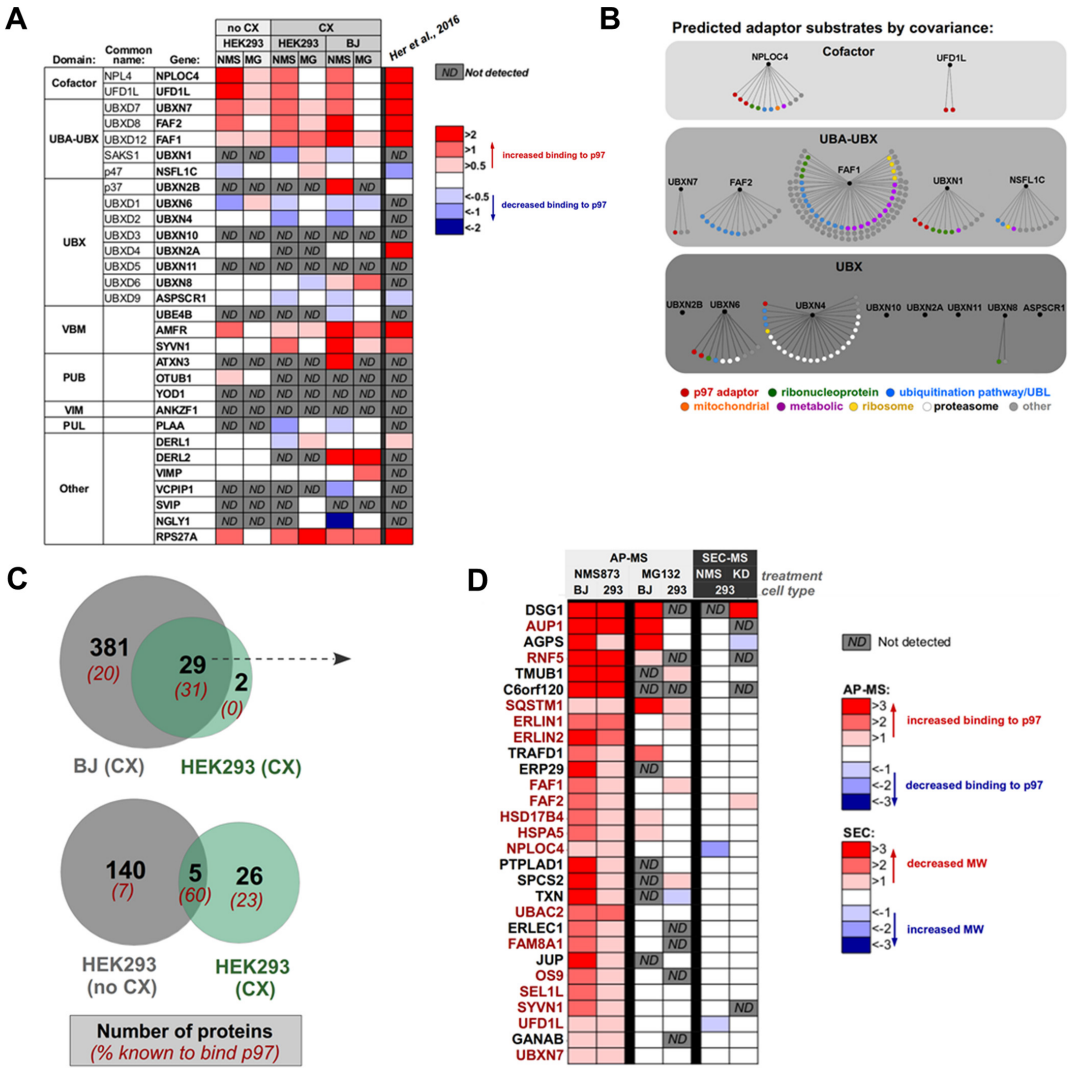


Figure 2.7: Summary of mass spectrometry results. A, Summary of the impact of MG132 or NMS873 treatment on VCP binding proteins highlighted by Her et al.[51], in the presence or absence of cross-linking. The final column shows data from[51]. The cross-linking data is from **Fig. 2.6**, whereas the non-cross-linking experiment (**Supplemental Table S10**) was done separately with HEK293 cells using a similar protocol except that label-free quantification was used. **B**, Prediction of potential substrates for UFD1L–NPLOC4 and UBX domain proteins by covariance analyses. Significance of covariance between UBX domain adaptors and all other proteins identified in the AP-MS experiments reported here was calculated through Pearson correlation values. All proteins with $p<0.05$ are displayed as a colored circle indicating their functional categorization. List of proteins for each adaptor can be found in **Supplementary Fig. S2.11** and **Supplemental Table S10**.

Predicted adaptor substrates by covariance:
(Accompanies Fig. 7B)

NPLOC4:
CHCHD3, FAF2, NEDD8, NOP58, RALY, RCN1, STT3A, TXN, UBAC2, UBXN7, UFD1L, VDAC1
UFD1L:
NPLOC4, UBXN7
UBXN7:
RCN1, TXN, UFD1L
FAF2:
ACTG1, ANXA5, ANXA6, CALM1, CUL2, CUL4B, ERLIN1, ERLIN2, PPP1CA, RNF5, SEL1L, SYVN1, TMUB1, UBAC2
FAF1:
ACSL3, ARL6IP5, ARMCX3, ASPH, ATL3, ATP6AP2, BAG2, BCAP31, BRIX1, BSG, C6orf120, CCT8, COPS3, CPT1A, CUL1, CUL3, CUL4B, CYB5B, DDB1, DDX1, DERL2, DNAJB11, DNAJB12, DNM1L, DYNLL1, ERLEC1, ERLIN1, FADS2, FAM8A1, FKBP8, GANAB, GPX8, HERPUD1, HNRNPAB, IMP3, ITGB1, KIAA1967, KPRP, MAGT1, MARCKS, MBOAT7, MCM3, MCM4, MYH14, NACA, NHP2L1, NOL6, OCIAD1, OS9, P4HB, PIGK, POLR2A, PPIB, PPP1CA, PTRH2, QARS, RBM8A, RBX1, RCN2, RDH11, RNF170, RNF5, RPL18A, RPL22, RPL37A, RPS10, SAMM50, SDCBP, SERPINH1, SF1, SKP1, SLC25A11, SNRPF, SPTLC1, SRPRB, SRSF10, SSR3, TCEB1, THRAP3, TMCO1, TMED7, TMED9, TMEM259, TMUB2, TOM1, TOR1AIP1, TP53, TRA2B, TUBB8, UBAP2L, UBE2G2, UBE2K, UGGT1, WDR26, YIPF5, YWHAG
UBXN1:
ALYREF, ASPSCR1, COMT, DSG1, HNRNPF, HNRNPH3, HNRPDL, ILF3, PLAA, SYNCRIP, UBXN4, UBXN6
NSFL1C:
GANAB, GRN, HLA-A, NACA, PPIB, RBX1, RCC1, RPS29, SEC61B, SUMO1, TMEM66
UBXN6:
DNAJB2, DSG1, FUS, PSMA2, PSMA7, PSMB3, SUMO1, TMEM66, UBXN1, UBXN2A
UBXN4:
GNB2L1, PCBP2, PLAA, PSMA4, PSMA5, PSMA6, PSMA7, PSMB1, PSMB2, PSMC1, PSMC2, PSMC4, PSMC6, PSMD1, PSMD11, PSMD12, PSMD13, PSMD2, PSMD3, PSMD7, PSMD8, RPS18, UBE4A, UBQLN1, UBQLN2, UBXN1
UBXN8:
HNRNPH3, TMEM66
UBXN2B, UBXN10, UBXN2A, UBXN11, ASPSCR1:

Supplementary Figure S2.11: Potential substrates for UFD1L–NPLOC4 and UBX domainproteins as determined by covariance analyses..

Figure 2.7: C, Overlap of AP-MS data for cross-linked HEK293 cells with cross-linked BJ fibroblasts and non-cross-linked HEK293 cells. Venn diagram depicts overlap of proteins with >2 fold enrichment by NMS873 for each cell type/condition. Numbers in black represent absolute number of proteins, while numbers in red represent percentage of these proteins reported as binding to VCP (as indexed by BioGRID). The dotted arrow signifies that the 29 proteins that overlap in the upper Venn diagram were used to populate the heatmap in panel **D**. **D**, Proteins with >2 fold enrichment by NMS873 in both HEK293 cells and BJ fibroblasts are depicted, as is the behavior of these same proteins in the SEC-mass spectrometry experiments.

Dynamic association of NSFL1C with VCP is an intrinsic property of the respective complexes Both our SEC and IP mass spectrometry experiments led to the unexpected conclusion that adaptor proteins that are known to function with VCP form extremely dynamic complexes with VCP that undergo rapid dissociation and exchange in cell lysates. We now sought to both validate these findings and place them on a more quantitative footing. To this end, we developed a quantitative FRET assay that allowed us to measure dynamic association of NSFL1C with VCP in vitro. Using the crystal structure of NSFL1C bound to the N-domain of VCP as a guide[52], we first mutagenized the C-terminal amino acid (Thr370) of NSFL1C to cysteine, and then reacted the purified recombinant protein with maleimide–TAMRA to generate NSFL1C^{TAMRA} (**Fig. 2.8a**). For VCP, we used the ybbR tagging method ([45]) to attach a Cy5 tag at its N-terminus (^{Cy5}VCP; **Fig. 2.8a**). Upon mixing NSFL1C^{TAMRA} and ^{Cy5}VC and exciting with 540 nm light, we observed a significant reduction in TAMRA fluorescence coupled to an increase in Cy5 emission (**Fig. 2.8b**). This FRET signal was due to specific interaction of NSFL1C^{TAMRA} and ^{Cy5}VCP, because it was competed by addition of excess unlabeled VCP (**Fig. 2.8b**). By titrating ^{Cy5}VCP, we

estimated a K_d of 65 nM for interaction of the two proteins in the absence of nucleotide (**Fig. 2.8c** and Table 1). This affinity is 10-fold tighter than what was reported from isothermal titration calorimetry studies[53,54], but is close to the affinity measured in a pair of surface plasmon resonance (SPR) studies (20-31 nM)[50,55]. However, there are significant problems with measuring NSFL1C–VCP interactions by SPR due to the oligomeric nature of both proteins. In addition, none of the studies cited above addressed the crucial issue of the dynamics of NSFL1C–VCP interaction. To investigate binding dynamics, we measured the kon for complex formation and k_{off} for complex dissociation in the absence of nucleotide ('apo') or in the presence of ADP or ATPγS with and without VCP inhibitors. Examples of k_{off} and k_{on} measurements in the absence of nucleotide are shown in **Fig. 2.8d** and **e**, respectively, and a plot of kobs vs. [^{Cy5}VCP] that was used to estimate k_{on} is shown in **Fig. 2.8f**. As shown in **Supplementary Fig. S2.12**, k_{on} values were essentially invariant, ranging from 8^{-11} x 107 M^{-1} sec^{-1} regardless of nucleotide state. Meanwhile, k_{off} showed slightly more variation, ranging from 2.5 sec^{-1} in the presence of ATP and NMS873 to 9.5 sec^{-1} in the presence of ADP. Consistent with the lack of an effect of NMS873 on co-IP of NSFL1C with VCP from lysates of BJ cells treated with DSP (**Fig. 2.6d**), addition of NMS873 in the presence of ATP had less than a 2-fold effect onk_{off} (**Supplementary Fig. S2.12**). These results confirm that purified NSFL1C^{TAMRA} exhibited extremely dynamic association with purified ^{Cy5}VCP in accordance with the behavior of these proteins in HEK293 cell extracts, and that their association was relatively insensitive to modulation by NMS873.

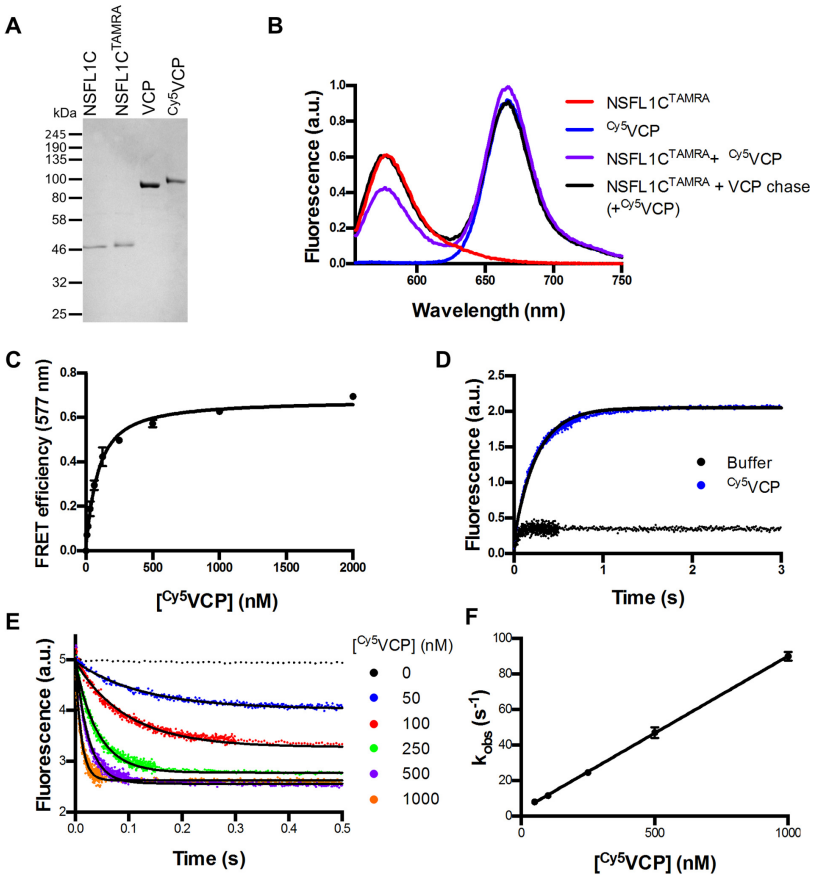


Figure 2.8: Biochemical characterization of VCP–adaptor interactions. A, SDS-PAGE gel of purified recombinant proteins used in this study. **B**, Fluorescence emission spectra of 16.7 nM NSFL1CTAMRA trimer, 83.3 nM Cy5VCP hexamer, and a mixture excited at 540 nm shows a ~20% loss of TAMRA donor fluorescence in the presence of Cy5VCP. Loss of fluorescence is prevented by preincubation of NSFL1CTAMRA with 833 nM unlabeled VCP hexamer. C, Equilibrium titration of 50 nM NSFL1CTAMRA trimer with Cy5VCP hexamer. Fit to a quadratic binding equation yields a KD of 65 $\pm$ 7 nM. Error bars represent $\pm$SD, with n = 3. **D**, Change in donor fluorescence of 50 nM NSFL1CTAMRA trimer preincubated with 300 nM Cy5VCP hexamer upon addition of 3 μM unlabeled VCP hexamer in the absence of nucleotide. Curve was fit to a single exponential to give koff of 4.04 s^{-1}. **E**, Change in donor fluorescence of 50 nM NSFL1CTAMRA trimer upon addition of Cy5VCP hexamer at various concentrations in the absence of nucleotide. Curves were fit to a single exponential. **F**, Exponential fits measured in panel **E** plotted against the concentration of Cy5VCP hexamer. Linear slope gives kon of 8.68 x 107 $M^{-1}s^{-1}$. Error bars represent $\pm$SD, with n $\geqslant$ 5.

Table 1. Kinetic and equilibrium binding constants for NSFL1C^{TAMRA} and ^{Cy5}VCP.

Nucleotide State	Inhibitor	K_D (eq) (nM)	k_{on} ($x10^7$ M^{-1} s^{-1})	k_{off} (s^{-1})	K_D (k_{off}/k_{on}) (nM)
Apo	-	65 ± 7	8.68 ± 0.09	4.04 ± 0.09	47
ATPγS	-	n.d.	10.5 ± 0.1	3.30 ± 0.02	31
ADP	-	n.d.	8.30 ± 0.07	9.50 ± 0.08	114
Apo	CB-5083	n.d.	8.0 ± 0.1	4.80 ± 0.03	60
ATP	NMS-873	n.d.	11.01 ± 0.01	2.50 ± 0.03	28

Supplementary Figure S2.12: Kinetic and equilibrium binding constants for NSFL1C^{TAMRA} and ^{Cy5}VCP..

Discussion

We set out in this study with the goal of using size exclusion chromatography coupled to mass spectrometry to characterize VCP–adaptor interactions and identify new VCP-interacting proteins. However, in the course of carrying out these experiments, we made the surprising finding that VCP shows little or no co-fractionation with adaptors that are well-validated to bind VCP. Similar results were observed in conventional immunoprecipitation mass spectrometry experiments, in which SILAC was used to quantify the degree to which VCP-binding proteins undergo exchange during purification. Characterization of the interaction between VCP and its adaptors has led to two important conclusions. First, under unperturbed conditions, VCP dissociates rapidly from all adaptors in cell lysate except ASPSCR1 and probably UBXN2A. We were able to reproduce this finding using purified VCP and NSFL1C, indicating that rapid exchange is likely to be an intrinsic property of VCP–adaptor interactions. Second, inhibition of VCP ATPase activity with the allosteric inhibitor NMS873 has markedly distinct effects on the interaction of different adaptors with VCP. Each of these findings has significant implications, which are discussed in more detail in the following paragraphs.

A cautionary note for analysis of VCP-interacting proteins Because the interaction of VCP with its adaptors is so dynamic, lists of VCP-interacting proteins obtained through conventional immunoprecipitation protocols[21,23,28] are most likely dominated by proteins

that formed interactions with VCP in cell lysate. Whereas many of these interactions are nonetheless likely to be physiological, some of them might not be, due to the loss of compartmentation and other forms of regulation that accompanies cell lysis. To some extent adaptor dynamics can be suppressed by NMS873 as discussed in the next section, but it must be appreciated that NMS873 has differential effects on adaptors and thus does not preserve the natural state that existed in cells. The only strategy we have identified that is likely to adequately suppress adaptor exchange is to treat cells with cross-linking agents prior to cell lysis and IP, but that can carry its own risks. In the absence of cross-linking, few proteins were identified that were also found in our cross-linking dataset, and the majority of these proteins have not been reported to bind VCP (**Fig. 2.7c**, full list can be found in **Supplemental Table S11**). Therefore, we suspect that VCP-interacting proteins recovered in the absence of crosslinking may not accurately represent the physiological associations of VCP. We did not address whether association of proteins with VCP adaptors is also subject to the same dynamic behavior documented here for VCP–adaptor interactions, but this possibility should be borne in mind. An analysis of the effects of NMS873 on VCP-interacting proteins appeared while we were finalizing our manuscript[56]. Although their mass spectrometry findings are generally consistent with our own (see **Fig. 2.7a** for a comparison), their work focused primarily on mechanisms of VCP resistance to NMS873 and consequently there are several important differences with the work reported here. First, their analyses were performed exclusively under native (i.e. non-cross-linked) conditions that favor interactions formed in cell lysate. Second, our coverage of known VCP interactors

was substantially more comprehensive. Finally, Her et al. did not evaluate the extent of VCP–adaptor exchange in cell lysate or with purified proteins.

Dynamics of VCP-adaptor interactions The extremely rapid off-rates that we documented for NSFL1C–VCP complexes pose an interesting question: can VCP–adaptor complexes adequately perform their tasks in cells? VCP has been reported to have a turnover rate ranging from 0.75-5.2 ATP per second per hexamer[38,57–59]. Thus, each protomer turns over less than 1 ATP per second. By comparison, the slowest k_{off} that we measured for uninhibited VCP–NSFL1C complexes is 3.3 sec^{-1}. Comparing these numbers suggests that, at most, 1 or 2 ATP is hydrolyzed per hexamer during the lifespan of a VCP–NSFL1C complex, and this is likely to be a maximal estimate because NSFL1C moderately slows down ATP hydrolysis by VCP[55,60]. It seems unlikely that this level of ATP hydrolysis would be sufficient to sustain the proposed segregase function of VCP. We speculate that there are mechanisms that modulate the lifespan of VCP–adaptor complexes. Consistent with this idea, NSFL1C forms a tight complex with VCP and co-purifies with it from rat liver extracts[61], whereas we did not observe co-migration of these proteins during SEC of HEK293 cell extracts. Thus, there may exist factors in liver cells that stabilize VCP–NSFL1C complexes. This could include covalent modifications on VCP, of which many have been reported[62], covalent modifications on the adaptor[63], or engagement of substrate by the adaptor. In contrast to the behavior of most adaptors, ASPSCR1 (also known as UBXD9, TUG, or ASPL) exhibited relatively stable binding to VCP in both the IP and SEC protocols. It was reported previously that the vast

majority of ASPSCR1 is bound to VCP (consistent with our results in the SEC experiment, **Supplementary Fig. 2.3a**), and disassembles VCP hexamers into monomers[64]. However, the physiological significance of this remodeling was not determined. UBXN2A exhibited a similar behavior to ASPSCR1, but in addition its expression appeared to depend on the presence of VCP (**Supplementary Fig. 2.3a**), suggesting that UBXN2A stability may be dependent upon its assembly with VCP. Interestingly, although ASPSCR1 has been linked to GLUT4 trafficking[65] and UBXN2A to mortalin[51] and nicotinic acetylcholine receptors[66], neither protein has been shown to link specific substrates to VCP. Given the distinct VCP binding dynamics and stoichiometries displayed by UBXN2A and ASPSCR1, these proteins may not function as substrate adaptors but instead may sequester or recycle VCP hexamers.

Effects of ATPase inhibitors on VCP-adaptor association The reversible, competitive VCP ATPase inhibitor CB-5083 is in clinical development for treatment of cancer. We still know relatively little about the full range of physiological functions and regulation of VCP, which creates a challenge for thinking about the disease settings in which VCP inhibition is likely to be most efficacious. From this perspective, it is of interest that the reversible, allosteric VCP ATPase inhibitor NMS873[32] exhibited strong differential effects on the association of adaptors with VCP. An intriguing idea is that, in principal, the reciprocal should also be true – adaptors whose binding to VCP is stabilized by an inhibitor should also stabilize binding of the inhibitor to VCP. This raises the possibility that different VCP–adaptor complexes might show differential sensitivity to small molecule inhibition,

which has been suggested in prior studies with NSFL1C[38]. Preferential inhibition of particular VCP–adaptor complexes could be an effective strategy to focus inhibition on VCP complexes that are of particular importance to the survival of cancer cells that have constitutively high levels of unfolded protein response[67].

Footnotes

We thank the Shu-ou Shan lab for Sfp plasmid, T.F. Chou for full-length human VCP in pET15, and S. Bulfer and M. Arkin for ND1L construct. We thank Shu-ou Shan and members of the Shan lab for helpful discussion on the FRET assay, Xun Wang for assistance of VCP tagging by CRISPR technique, and R. Verma for helpful suggestions and comments on the manuscript. We also thank Tanya Porras-Yakushi for help in editing the manuscript for clarity, Annie Moradian and Roxana Eggleston-Rangel for mass spectrometry support and Michael J. Sweredoski for bioinformatics assistance. This work was supported by the Gordon and Betty Moore Foundation, through Grant GBMF775, the Beckman Institute and the NIH through Grant 1S10RR029594 (to SH). J.M.R. was supported by a postdoctoral fellowship from NIH (F32 GM112308-03). R.J.D. is an Investigator of the Howard Hughes Medical Institute and this work was funded by HHMI. E.E.B. is supported by an NIH Training Grant (T32 GM007616).

References

[1] W. E. Balch, R. I. Morimoto, A. Dillin, and J. W. Kelly, "Adapting proteostasis for disease intervention.," *Science (New York, N.Y.)*, vol. 319, pp. 916–919, feb 2008.

[2] B. Bukau, J. Weissman, and A. Horwich, "Molecular chaperones and protein quality control.," *Cell*, vol. 125, pp. 443–451, may 2006.

[3] S. M. Doyle, O. Genest, and S. Wickner, "Protein rescue from aggregates by powerful molecular chaperone machines.," *Nature reviews. Molecular cell biology*, vol. 14, pp. 617–629, oct 2013.

[4] A. L. Goldberg, "Protein degradation and protection against misfolded or damaged proteins.," *Nature*, vol. 426, pp. 895–899, dec 2003.

[5] J. Luo, N. L. Solimini, and S. J. Elledge, "Principles of cancer therapy: oncogene and non-oncogene addiction.," *Cell*, vol. 136, pp. 823–837, mar 2009.

[6] S. Santaguida and A. Amon, "Short- and long-term effects of chromosome mis-segregation and aneuploidy.," *Nature reviews. Molecular cell biology*, vol. 16, pp. 473–485, aug 2015.

[7] E. Arias and A. M. Cuervo, "Chaperone-mediated autophagy in protein quality control.," *Current opinion in cell biology*, vol. 23, pp. 184–189, apr 2011.

[8] R. J. Deshaies, "Proteotoxic crisis, the ubiquitin-proteasome system, and cancer therapy.," *BMC biology*, vol. 12, p. 94, nov 2014.

[9] B. A. Teicher and K. C. Anderson, "CCR 20th anniversary commentary: In the beginning, there was PS-341.," mar 2015.

[10] R. Verma, R. S. Oania, N. J. Kolawa, and R. J. Deshaies, "Cdc48/p97 promotes degradation of aberrant nascent polypeptides bound to the ribosome.," *eLife*, vol. 2, p. e00308, jan 2013.

[11] O. Brandman, J. Stewart-Ornstein, D. Wong, A. Larson, C. C. Williams, G.-W. Li, S. Zhou, D. King, P. S. Shen, J. Weibezahn, J. G. Dunn, S. Rouskin, T. Inada, A. Frost, and J. S. Weissman, "A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress.," *Cell*, vol. 151, pp. 1042–1054, nov 2012.

[12] Q. Defenouillere, Y. Yao, J. Mouaikel, A. Namane, A. Galopier, L. Decourty, A. Doyen, C. Malabat, C. Saveanu, A. Jacquier, and M. Fromont-Racine, "Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 5046–5051, mar 2013.

[13] E. B. Taylor and J. Rutter, "Mitochondrial quality control by the ubiquitin-proteasome system.," *Biochemical Society transactions*, vol. 39, pp. 1509–1513, oct 2011.

[14] Y. Ye, H. H. Meyer, and T. A. Rapoport, "The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol.," *Nature*, vol. 414, pp. 652–656, dec 2001.

[15] H. Meyer and C. C. Weihl, "The VCP/p97 system at a glance: connecting cellular function to disease pathogenesis.," *Journal of cell science*, vol. 127, pp. 3877–3883, sep 2014.

[16] B. Medicherla and A. L. Goldberg, "Heat shock and oxygen radicals stimulate ubiquitin-dependent degradation mainly of newly synthesized proteins.," *The Journal of cell biology*, vol. 182, pp. 663–673, aug 2008.

[17] J.-S. Ju, R. A. Fuentealba, S. E. Miller, E. Jackson, D. Piwnica-Worms, R. H. Baloh, and C. C. Weihl, "Valosin-containing protein (VCP) is required for autophagy and is disrupted in VCP disease.," *The Journal of cell biology*, vol. 187, pp. 875–888, dec 2009.

[18] E. Tresse, F. A. Salomons, J. Vesa, L. C. Bott, V. Kimonis, T.-P. Yao, N. P. Dantuma, and J. P. Taylor, "VCP/p97 is essential for maturation of ubiquitin-containing autophagosomes and this function is impaired by mutations that cause IBMPFD.," *Autophagy*, vol. 6, pp. 217–227, feb 2010.

[19] J. R. Buchan, R.-M. Kolaitis, J. P. Taylor, and R. Parker, "Eukaryotic stress granules are cleared by autophagy and Cdc48/VCP function.," *Cell*, vol. 153, pp. 1461–1474, jun 2013.

[20] J.-S. Ju, S. E. Miller, P. I. Hanson, and C. C. Weihl, "Impaired protein aggregate handling and clearance underlie the pathogenesis of p97/VCP-associated disease.," *The Journal of biological chemistry*, vol. 283, pp. 30289–30299, oct 2008.

[21] G. Alexandru, J. Graumann, G. T. Smith, N. J. Kolawa, R. Fang, and R. J. Deshaies, "UBXD7 binds multiple ubiquitin ligases and implicates p97 in HIF1alpha turnover.," *Cell*, vol. 134, pp. 804–816, sep 2008.

[22] A. Buchberger, H. Schindelin, and P. Hanzelmann, "Control of p97 function by cofactor binding.," *FEBS letters*, vol. 589, pp. 2578–2589, sep 2015.

[23] M. Raman, M. Sergeev, M. Garnaas, J. R. Lydeard, E. L. Huttlin, W. Goessling, J. V. Shah, and J. W. Harper, "Systematic proteomics of the VCP-UBXD adaptor network identifies a role for UBXN10 in regulating ciliogenesis.," *Nature cell biology*, vol. 17, pp. 1356–1369, oct 2015.

[24] C. Schuberth and A. Buchberger, "UBX domain proteins: major regulators of the AAA ATPase Cdc48/p97.," *Cellular and molecular life sciences : CMLS*, vol. 65, pp. 2360–2371, aug 2008.

[25] S. Elsasser and D. Finley, "Delivery of ubiquitinated substrates to protein-unfolding machines.," *Nature cell biology*, vol. 7, pp. 742–749, aug 2005.

[26] R. Verma, R. Oania, R. Fang, G. T. Smith, and R. J. Deshaies, "Cdc48/p97 mediates UV-dependent turnover of RNA Pol II.," *Molecular cell*, vol. 41, pp. 82–92, jan 2011.

[27] A. Riemer, G. Dobrynin, A. Dressler, S. Bremer, A. Soni, G. Iliakis, and H. Meyer, "The p97-Ufd1-Npl4 ATPase complex ensures robustness of the G2/M checkpoint by facilitating CDC25A degradation.," *Cell cycle (Georgetown, Tex.)*, vol. 13, no. 6, pp. 919–927, 2014.

[28] D. Ritz, M. Vuk, P. Kirchner, M. Bug, S. Schutz, A. Hayer, S. Bremer, C. Lusk, R. H. Baloh, H. Lee, T. Glatter, M. Gstaiger, R. Aebersold, C. C. Weihl, and H. Meyer, "Endolysosomal sorting of ubiquitylated caveolin-1 is regulated by VCP and UBXD1 and impaired by VCP disease mutations.," *Nature cell biology*, vol. 13, pp. 1116–1123, aug 2011.

[29] J. He, Q. Zhu, G. Wani, N. Sharma, and A. A. Wani, "Valosin-containing Protein (VCP)/p97 Segregase Mediates Proteolytic Processing of Cockayne Syndrome Group B (CSB) in Damaged Chromatin.," *The Journal of biological chemistry*, vol. 291, pp. 7396–7408, apr 2016.

[30] S. Jentsch and S. Rumpf, "Cdc48 (p97): a "molecular gearbox" in the ubiquitin pathway?," *Trends in biochemical sciences*, vol. 32, pp. 6–11, jan 2007.

[31] T.-F. Chou, S. J. Brown, D. Minond, B. E. Nordin, K. Li, A. C. Jones, P. Chase, P. R. Porubsky, B. M. Stoltz, F. J. Schoenen, M. P. Patricelli, P. Hodder, H. Rosen, and R. J. Deshaies, "Reversible inhibitor of p97, DBeQ, impairs both ubiquitin-dependent and autophagic protein clearance pathways.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 4834–4839, mar 2011.

[32] P. Magnaghi, R. D'Alessio, B. Valsasina, N. Avanzi, S. Rizzi, D. Asa, F. Gasparri, L. Cozzi, U. Cucchi, C. Orrenius, P. Polucci, D. Ballinari, C. Perrera, A. Leone, G. Cervi, E. Casale, Y. Xiao, C. Wong, D. J. Anderson, A. Galvani, D. Donati, T. O'Brien, P. K. Jackson, and A. Isacchi, "Covalent and allosteric inhibitors of the ATPase VCP/p97 induce cancer cell death.," *Nature chemical biology*, vol. 9, pp. 548–556, sep 2013.

[33] P. Acharya, M. Liao, J. C. Engel, and M. A. Correia, "Liver cytochrome P450 3A endoplasmic reticulum-associated degradation: a major role for the p97 AAA ATPase in cytochrome P450 3A extraction into the cytosol.," *The Journal of biological chemistry*, vol. 286, pp. 3815–3828, feb 2011.

[34] R. Piccirillo and A. L. Goldberg, "The p97/VCP ATPase is critical in muscle atrophy and the accelerated degradation of muscle proteins.," *The EMBO journal*, vol. 31, pp. 3334–3350, aug 2012.

[35] T.-F. Chou, K. Li, K. J. Frankowski, F. J. Schoenen, and R. J. Deshaies, "Structure-activity relationship study reveals ML240 and ML241 as potent and selective inhibitors of p97 ATPase.," *ChemMedChem*, vol. 8, pp. 297–312, feb 2013.

[36] H.-J. Zhou, J. Wang, B. Yao, S. Wong, S. Djakovic, B. Kumar, J. Rice, E. Valle, F. Soriano, M.-K. Menon, A. Madriaga, S. Kiss von Soly, A. Kumar, F. Parlati, F. M. Yakes, L. Shawver, R. Le Moigne, D. J. Anderson, M. Rolfe, and D. Wustrow, "Discovery of a First-in-Class, Potent, Selective, and Orally Bioavailable Inhibitor of the p97 AAA ATPase (CB-5083).," *Journal of medicinal chemistry*, vol. 58, pp. 9480–9497, dec 2015.

[37] D. J. Anderson, R. Le Moigne, S. Djakovic, B. Kumar, J. Rice, S. Wong, J. Wang, B. Yao, E. Valle, S. Kiss von Soly, A. Madriaga, F. Soriano, M.-K. Menon, Z. Y. Wu, M. Kampmann, Y. Chen, J. S. Weissman, B. T. Aftab, F. M. Yakes, L. Shawver, H.-J. Zhou, D. Wustrow, and M. Rolfe, "Targeting the AAA ATPase p97 as an Approach to Treat Cancer through Disruption of Protein Homeostasis.," *Cancer cell*, vol. 28, pp. 653–665, nov 2015.

[38] T.-F. Chou, S. L. Bulfer, C. C. Weihl, K. Li, L. G. Lis, M. A. Walters, F. J. Schoenen, H. J. Lin, R. J. Deshaies, and M. R. Arkin, "Specific inhibition of p97/VCP ATPase and kinetic analysis demonstrate interaction between D1 and D2 ATPase domains.," *Journal of molecular biology*, vol. 426, pp. 2886–2899, jul 2014.

[39] C.-C. Yu, J.-C. Yang, Y.-C. Chang, J.-G. Chuang, C.-W. Lin, M.-S. Wu, and L.-P. Chow, "VCP phosphorylation-dependent interaction partners prevent apoptosis in Helicobacter pylori-infected gastric epithelial cells.," *PloS one*, vol. 8, no. 1, p. e55724, 2013.

[40] A. L. Szymczak, C. J. Workman, Y. Wang, K. M. Vignali, S. Dilioglou, E. F. Vanin, and D. A. A. Vignali, "Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide-based retroviral vector.," may 2004.

[41] L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang, "Multiplex genome engineering using CRISPR/Cas systems.," *Science (New York, N.Y.)*, vol. 339, pp. 819–823, feb 2013.

[42] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, and F. Zhang, "Genome engineering using the CRISPR-Cas9 system.," *Nature protocols*, vol. 8, pp. 2281–2308, nov 2013.

[43] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets.," *Nucleic acids research*, vol. 34, pp. D535–9, jan 2006.

[44] R. M. Bruderer, C. Brasseur, and H. H. Meyer, "The AAA ATPase p97/VCP interacts with its alternative co-factors, Ufd1-Npl4 and p47, through a common bipartite binding mechanism.," *The Journal of biological chemistry*, vol. 279, pp. 49609–49616, nov 2004.

[45] J. Yin, A. J. Lin, D. E. Golan, and C. T. Walsh, "Site-specific protein labeling by Sfp phosphopantetheinyl transferase.," *Nature protocols*, vol. 1, no. 1, pp. 280–285, 2006.

[46] K. J. Kirkwood, Y. Ahmad, M. Larance, and A. I. Lamond, "Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics.," *Molecular & cellular proteomics : MCP*, vol. 12, pp. 3851–3873, dec 2013.

[47] O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. Mikkelson, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, and F. Zhang, "Genome-scale CRISPR-Cas9 knockout screening in human cells.," *Science (New York, N.Y.)*, vol. 343, pp. 84–87, jan 2014.

[48] X. Wang and L. Huang, "Identifying dynamic interactors of protein complexes by quantitative mass spectrometry.," *Molecular & cellular proteomics : MCP*, vol. 7, pp. 46–57, jan 2008.

[49] M. Mann, "Functional and quantitative proteomics using SILAC.," dec 2006.

[50] W. S. Chia, D. X. Chia, F. Rao, S. Bar Nun, and S. Geifman Shochat, "ATP binding to p97/VCP D1 domain regulates selective recruitment of adaptors to its proximal N-domain.," *PloS one*, vol. 7, no. 12, p. e50490, 2012.

[51] S. Sane, A. Abdullah, D. A. Boudreau, R. K. Autenried, B. K. Gupta, X. Wang, H. Wang, E. H. Schlenker, D. Zhang, C. Telleria, L. Huang, S. C. Chauhan, and K. Rezvani, "Ubiquitin-like (UBX)-domain-containing protein, UBXN2A, promotes cell death by interfering with the p53-Mortalin interactions in colon cancer cells.," *Cell death & disease*, vol. 5, p. e1118, mar 2014.

[52] I. Dreveny, H. Kondo, K. Uchiyama, A. Shaw, X. Zhang, and P. S. Freemont, "Structural basis of the interaction between the AAA ATPase p97/VCP and its adaptor protein p47.," *The EMBO journal*, vol. 23, pp. 1030–1039, mar 2004.

[53] F. Beuron, I. Dreveny, X. Yuan, V. E. Pye, C. McKeown, L. C. Briggs, M. J. Cliff, Y. Kaneko, R. Wallis, R. L. Isaacson, J. E. Ladbury, S. J. Matthews, H. Kondo, X. Zhang, and P. S. Freemont, "Conformational changes in the AAA ATPase p97-p47 adaptor complex.," *The EMBO journal*, vol. 25, pp. 1967–1976, may 2006.

[54] P. Hanzelmann, A. Buchberger, and H. Schindelin, "Hierarchical binding of cofactors to the AAA ATPase p97.," *Structure (London, England : 1993)*, vol. 19, pp. 833–843, jun 2011.

[55] X. Zhang, L. Gui, X. Zhang, S. L. Bulfer, V. Sanghez, D. E. Wong, Y. Lee, L. Lehmann, J. S. Lee, P.-Y. Shih, H. J. Lin, M. Iacovino, C. C. Weihl, M. R. Arkin, Y. Wang, and T.-F. Chou, "Altered cofactor regulation with disease-associated p97/VCP mutations.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. E1705–14, apr 2015.

[56] N.-G. Her, J. I. Toth, C.-T. Ma, Y. Wei, K. Motamedchaboki, E. Sergienko, and M. D. Petroski, "p97 Composition Changes Caused by Allosteric Inhibition Are Suppressed by an On-Target Mechanism that Increases the Enzyme's ATPase Activity.," *Cell chemical biology*, vol. 23, pp. 517–528, apr 2016.

[57] C. Song, Q. Wang, and C.-C. H. Li, "ATPase activity of p97-valosin-containing protein (VCP). D2 mediates the major enzyme activity, and D1 contributes to the heat-induced activity.," *The Journal of biological chemistry*, vol. 278, pp. 3648–3655, feb 2003.

[58] H. Niwa, C. A. Ewens, C. Tsang, H. O. Yeung, X. Zhang, and P. S. Freemont, "The role of the N-domain in the ATPase activity of the mammalian AAA ATPase p97/VCP.," *The Journal of biological chemistry*, vol. 287, pp. 8561–8570, mar 2012.

[59] B. DeLaBarre, J. C. Christianson, R. R. Kopito, and A. T. Brunger, "Central pore residues

mediate the p97/VCP activity required for ERAD.," *Molecular cell*, vol. 22, pp. 451–462, may 2006.

[60] H. H. Meyer, H. Kondo, and G. Warren, "The p47 co-factor regulates the ATPase activity of the membrane fusion protein, p97.," *FEBS letters*, vol. 437, pp. 255–257, oct 1998.

[61] H. Kondo, C. Rabouille, R. Newman, T. P. Levine, D. Pappin, P. Freemont, and G. Warren, "p47 is a cofactor for p97-mediated membrane fusion.," *Nature*, vol. 388, pp. 75–78, jul 1997.

[62] C. Mori-Konya, N. Kato, R. Maeda, K. Yasuda, N. Higashimae, M. Noguchi, M. Koike, Y. Kimura, H. Ohizumi, S. Hori, and A. Kakizuka, "p97/valosin-containing protein (VCP) is highly modulated by phosphorylation and acetylation.," *Genes to cells : devoted to molecular & cellular mechanisms*, vol. 14, pp. 483–497, apr 2009.

[63] K. Uchiyama, E. Jokitalo, M. Lindman, M. Jackman, F. Kano, M. Murata, X. Zhang, and H. Kondo, "The localization and phosphorylation of p47 are important for Golgi disassembly-assembly during the cell cycle.," *The Journal of cell biology*, vol. 161, pp. 1067–1079, jun 2003.

[64] C. M. Orme and J. S. Bogan, "The ubiquitin regulatory X (UBX) domain-containing protein TUG regulates the p97 ATPase and resides at the endoplasmic reticulum-golgi

intermediate compartment.," *The Journal of biological chemistry*, vol. 287, pp. 6679–6692, feb 2012.

[65] J. P. Belman, E. N. Habtemichael, and J. S. Bogan, "A proteolytic pathway that controls glucose uptake in fat and muscle.," *Reviews in endocrine & metabolic disorders*, vol. 15, pp. 55–66, mar 2014.

[66] K. Rezvani, Y. Teng, Y. Pan, J. A. Dani, J. Lindstrom, E. A. Garcia Gras, J. M. McIntosh, and M. De Biasi, "UBXD4, a UBX-containing protein, regulates the cell surface number and stability of alpha3-containing nicotinic acetylcholine receptors.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 29, pp. 6883–6896, may 2009.

[67] T.-F. Chou and R. J. Deshaies, "Development of p97 AAA ATPase inhibitors.," *Autophagy*, vol. 7, pp. 1091–1092, sep 2011.

Chapter 3

MITOCHONDRIAL PROTEIN FUNCTIONS ELUCIDATED BY MULTI-OMIC MASS SPECTROMETRY PROFILING

ECF designed and ran MS experiments and analyzed data for the proteomics sections of this work.

Portions of this chapter have been published:

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

* Authors contributed equally

A note about this chapter

This project was a large collaboration, so I want to direct attention to the work relevant to this dissertation and add context for the depth and quality of the proteomics data collection. The proteomics work comprised two sets of ~525 yeast samples, including 174 knockout strains grown in triplicate, and corresponding wild-type samples; one set of samples was grown in a standard fermentation culture, the other in the culture medium optimized for respiration. Expanding upon the results presented in **Supplementary Fig S3.2**, methods for sample preparation were modified from those presented in Hebert et al. 2014 to improve throughput[1]. The lysis step previously involved extensive bead milling, which was time-consuming and resulted in sample loss from transfer in and out of the glass bead aliquot. In this study, cells were lysed by reconstitution in urea and extraction in 90% methanol (MeOH). Reagents used for reduction and alkylation of the cysteine residues were changed from dithiothreitol (DTT – for reduction) and iodoacetamide (IAA – for alkylation) to TCEP (for reduction) and chloroacetamide (CAA – for alkylation). TCEP and CAA can be added concurrently, whereas DTT and IAA must be incubated separately, because the –SH group on DTT will compete with the cysteine residues for alkylation. Also, IAA is light sensitive while CAA solution is not, so storage and incubation conditions are less laborious. The final optimized step was the desalting process; buffer salts and other excipients that may contaminate the LC-MS system must be removed prior to analysis. We had previously used Waters C18 Sep-Pak desalting cartridges with sorbent masses of 50 mg. In the updated

process, we used Phenomenex Strata-X Polymeric reverse phase cartridges with sorbent masses of 10 mg. The decreased sorbent mass requires proportionally less solvent for loading, washing, and eluting the peptides, and so requires much less time from priming the cartridges to drying down the eluate. Additionally, the polymeric reverse phase sorbent is more resistant to drying effects, and so results in more reproducible eluates across a large batch of samples. Ultimately, we reduced the hands-on sample preparation time from over 6 hrs to under 2 hrs, which allowed the collection of 12 proteomes and over 50,000 phenotypes a day (**Supp Fig S3.2b,c**).

Samples were grown in batches of 19 knockouts and one wild-type (for 60 total samples per growth condition per batch). Samples were digested and analyzed in these same batches. A total of twelve batches were analyzed, and the time from the beginning of batch one to the end of batch twelve spanned more than seven months. All batches were searched and quantified using MaxQuant, and due to computational limitations, batches were all searched separately. To account for significant batch effects arising from each batch being grown, prepped, analyzed, and quantified in the same groups, each protein was normalized to the wild-type replicates from the same batch rather than to a pool of wild-type replicates from across the study. The majority of the technical variation from each step of the workflow should have been captured in these wild-type samples since they saw the same conditions throughout the process as did the knockouts. As such, all quantitative values in the body of this work are presented in fold-change of a protein in the knockout over the wild-type. These fold-change values are $\log_2$ normalized prior to further

informatics analysis. Data quality was assessed and is summarized in **Supplemental Fig 3.2a, d, e, f, and g**. Because knockouts were not compared between growth conditions, quality metrics were compiled for fermentation and respiration samples separately. We averaged ~3,100 proteins quantified per sample (**Supplemental Fig. 3.2a**) with an overlap of 2,800-2,900 molecules between strains. This metric is particularly important because MaxLFQ analysis results in relative quantitative values, so to draw any conclusions about protein expression level we must compare the same protein across multiple samples. Our quantitative reproducibility is plotted in **Supp Fig 3.2e**, with a median %CV (std/mean * 100) of between 7 and 10 depending on growth conditions. Finally, the dynamic range of fold-changes we were able to measure spanned from 2x to over 8,000x, with a majority of measurements falling between 2x and 32x. The body of this chapter contains discussion of how these data were used for further informatics analysis. For specific examples, please see **Figs. 3.2c, 3.3, and 3.4**, and **Supplemental Figs S3.3, S3.4, S3.5, S3.6, S3.8, S3.9, S3.11, S3.12, S3.13**.

Abstract

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

Introduction

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

We applied this strategy to study mitochondria, dynamic organelles whose dysfunction is associated with over 150 human diseases including cancer, diabetes, Parkinson's, and numerous genetic disorders[8–10]. While the yeast and mammalian mitochondrial proteomes were recently defined[11–13], functional annotation of these proteins lags behind[14], impeding biomedical research on the many diseases impacted by mitochondrial metabolism. Of the

~1,200 mammalian mitochondrial proteins, nearly 300 are "mitochondrial uncharacterized (x) proteins" (MXPs)[15,16] that have no well-established biochemical function within mitochondria. Here, toward defining functions for MXPs, we performed over 3,000 MS experiments in parallel to analyze the proteomes, metabolomes, and lipidomes of 174 single-gene deletion ("Δ*gene*") Saccharomyces cerevisiae yeast strains in biological triplicate across two metabolic conditions, fermentation and respiration **(Fig. 3.1a)**. To facilitate development of biological hypotheses based on the resultant "yeast-three-thousand (Y3K)" dataset **(Fig. 3.1b)**, we also developed a multi-omic data visualization approach (highlighted in **Fig. 3.1c** and online at http://y3kproject.org/). Our data establish many new connections between MXPs and proteins with well-established functions by virtue of gene-specific phenotypes or shared global biomolecular changes that result from the loss of each protein's expression. We leveraged a subset of these connections to address the incomplete mitochondrial pathway that generates ubiquinone (coenzyme Q, CoQ), an essential lipid required for oxidative phosphorylation (OxPhos) and linked to diseases ranging from severe infantile multisystemic disease to isolated myopathy and aging[17,18].

Results

Multi-omic mass spectrometry profiling. The 174 Δ*gene* yeast strains we analyzed covered 124 characterized genes that were selected to span a broad range of pathways to assist functional mapping, and 50 uncharacterized genes that encode MXPs **(Fig. 3.1a and Supplementary Fig. S3.1a)**. In selecting these targets, we prioritized genes with human

homologs (144/174 genes) and those associated with disease (60/144 genes) based on primary literature analysis and online database gene annotation (e.g., omim.org). Inclusion of characterized genes, some of which could be considered as only partially characterized, also provided the ability to connect them to previously unrecognized functions. Each strain was grown in biological triplicate under two contrasting growth conditions, a standard fermentation culture condition and a carefully optimized respiration culture condition that stimulates mitochondrial function **(Fig. 3.1a, Supplementary Fig. S3.1b–e, and Supplementary Note 1)**—yielding six separate cultures per yeast strain.

Altogether we grew more than 1,050 yeast cultures (including WT cultures), each of which was analyzed using three separate high-resolution MS-based proteomic, metabolomic, and lipidomic techniques. These 3,000+ MS experiments yielded quantitation of 4,040 proteins, 411 metabolites, and 53 lipids (averaging 3,180 proteins, 252 metabolites, and 53 lipids per culture)—over 3.5 million biomolecule measurements in total **(Fig. 3.1a and Supplementary Fig. S3.2a,b)**. Key to our approach was streamlining procedures for proteome extraction and preparation to under two hours of hands-on time **(Supplementary Fig. S3.2c)**. Use of label-free quantitation negated the need for a chemical tagging step and further increased throughput. We observed a wide dynamic range across all profiled omes, with some molecule abundances spanning more than three orders of magnitude **(Supplementary Fig. S3.2d)**. Additionally, we observed remarkable reproducibility between replicate cultures, with a median coefficient of variation of 12.7% considering all profiled biomolecules, and high overlap of molecules quantified across cultures **(Supplementary**

Fig. S3.2e–g).

A high-level view of the Y3K dataset shows significant perturbations across all three omes, with more pronounced perturbations in respiration **(Fig. 3.1b and Supplementary Fig. S3.3a)**. Hierarchical clustering revealed groups of functionally related molecules (along the y-axis) and groups of functionally related Δ*gene* strains (along the x-axis). Protein clusters show significant gene ontology (GO) term enrichments for diverse processes and include both characterized and uncharacterized proteins **(Supplementary Fig. S3.3b)**. For example, the uncharacterized proteins Esbp6p and Ypr010c-a cluster with proteins involved in mitochondrial ATP synthesis and electron transport chain function, respectively **(Supplementary Fig. S3.4)**. Here, we leverage analyses from three different vantage points, each of which can be recapitulated with our online data visualization suite, exploiting unique biological perspectives afforded by a multi-omic dataset of diverse genetic perturbations **(Fig. 3.1c)**.

Identification of gene-specific phenotypes. First, we systematically surveyed the Y3K dataset for significant molecule perturbations unique to just one or two of the strains in the study **(Fig. S3.2a)**. This unbiased search revealed 714 Δ*gene*-specific phenotypes **(Fig. 3.2a and Supplementary Note 2)**, which can reveal functional relationships. For example, the electron transfer flavoprotein (ETF) subunit Aim45p was uniquely decreased in just two Δ*gene* strains: the Δ*aim45* strain, and the Δ*cir1* strain, which lacks the second ETF heterodimer subunit **(Fig. 3.2b)**. Numerous additional Δ*gene*-specific phenotypes were

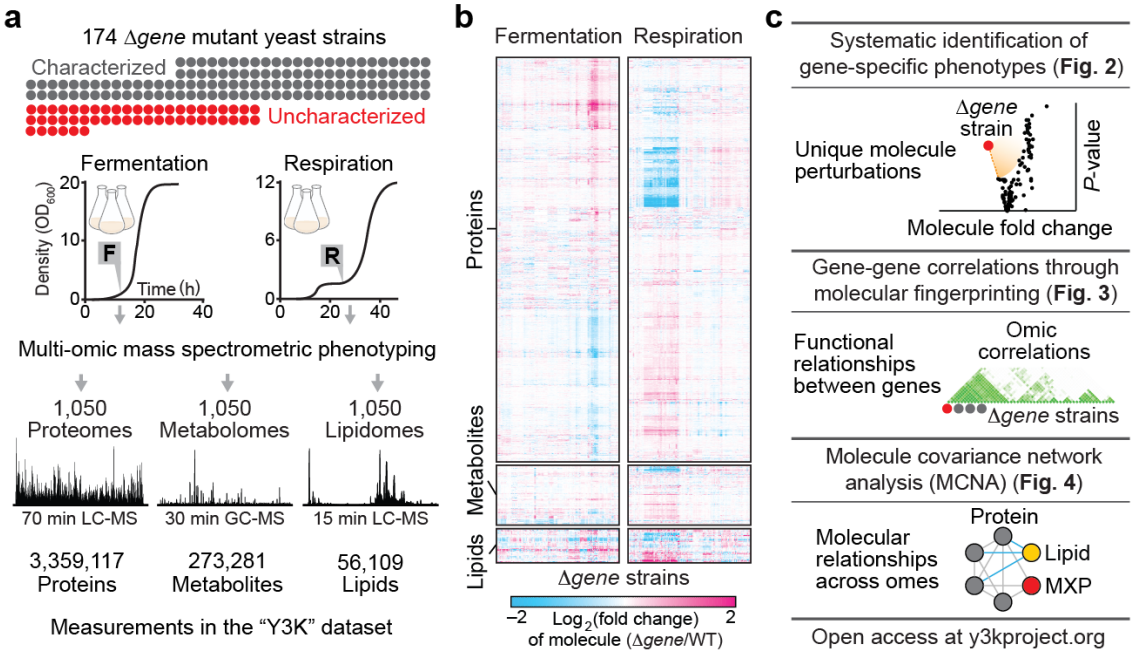
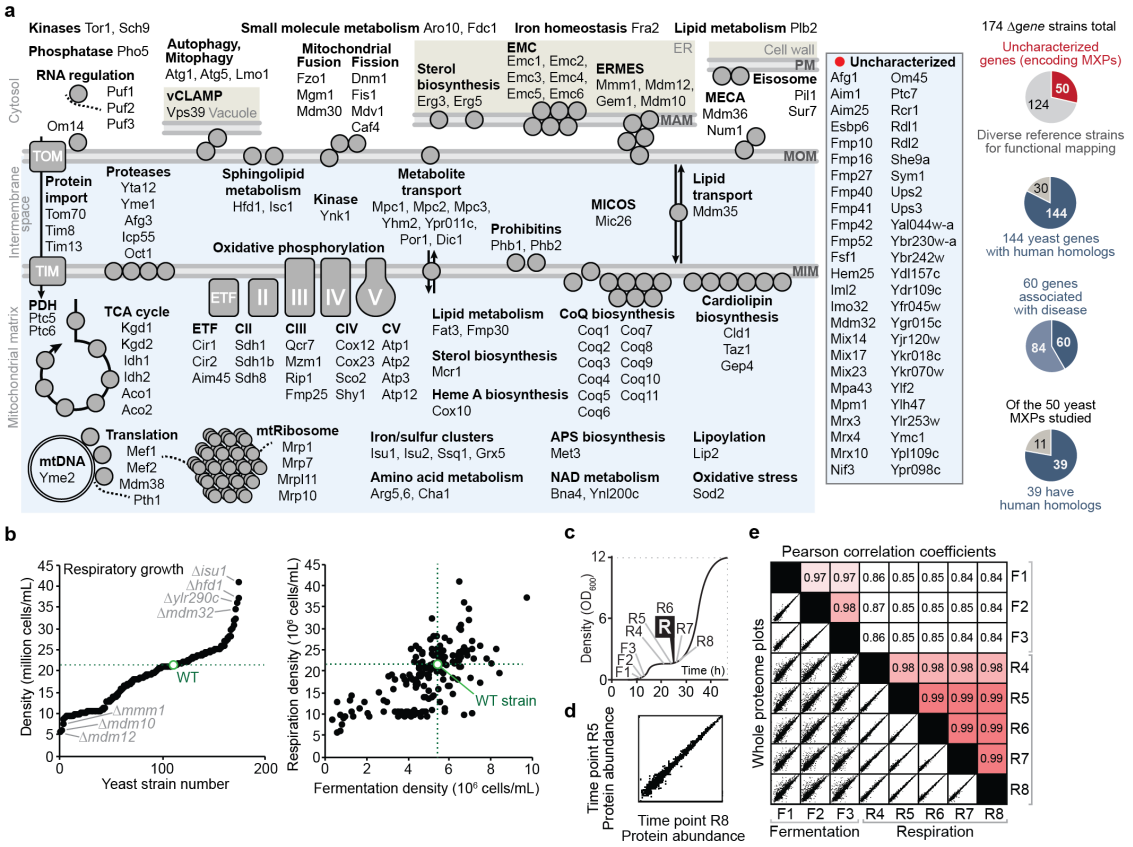


Figure 3.1: Multi-omic mass spectrometry profiling and data visualization. Multi-omic mass spectrometry profiling and data visualization. Overviews of **(a)** the experimental design and high resolution quantitative MS analysis, **(b)** the Y3K dataset, shown as hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT as quantified by MS, mean, n = 3; $P < 0.05$, two-sided Student's t-test), and **(c)** the multi-omic data analysis and visualization tools developed here.

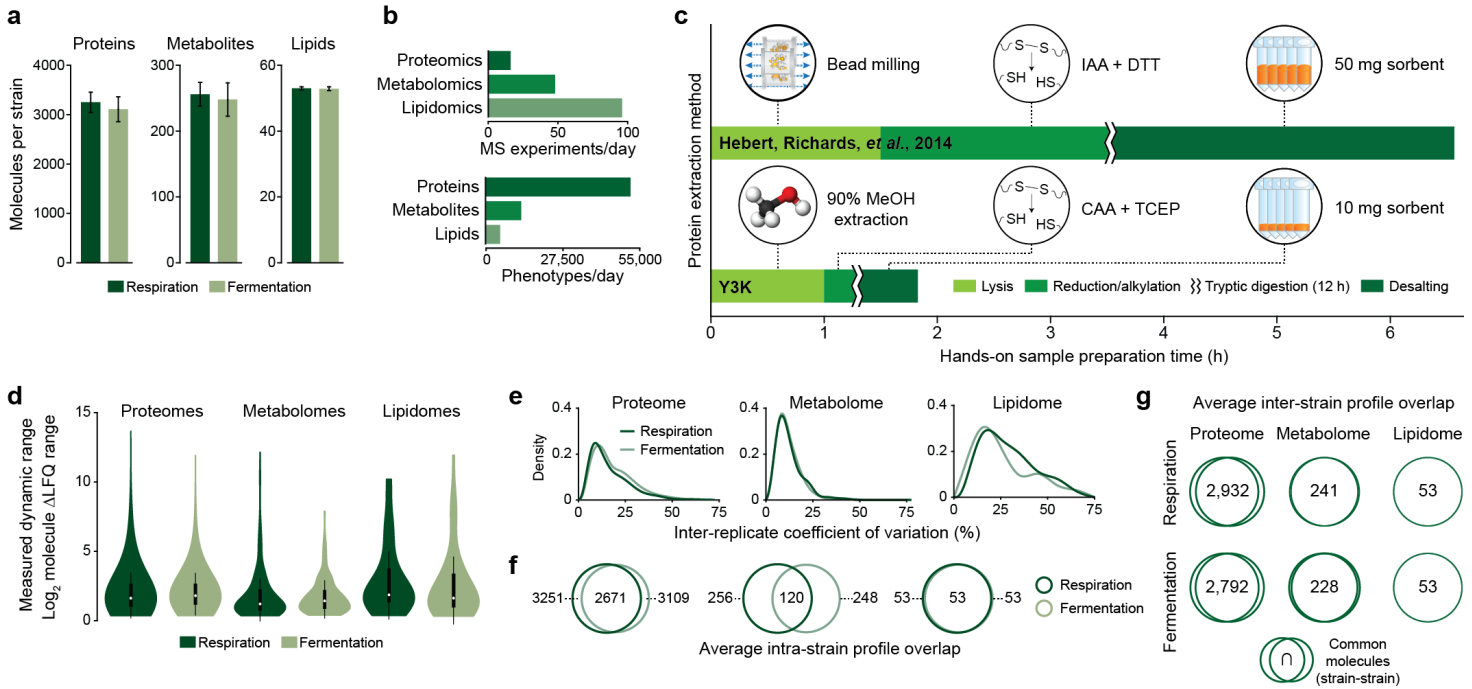


Supplementary Figure S3.1: *ΔGene* target strain characteristics and respiration culture optimization. (a) Proteins encoded by the individual genes knocked out of the 174 yeast strains investigated in this study, shown in the context of biological pathways. APS, adenosine-5′-phosphosulfate; CII–CV, oxidative phosphorylation complexes II–V; ER, endoplasmic reticulum; EMC, ER membrane complex; ERMES, ER-mitochondria encounter structure; ETF, electron transfer flavoprotein complex; MAM, mitochondria-associated membrane; MECA, mitochondria-ER-cortex anchor; MICOS, mitochondrial contact site and cristae organizing system; MIM, mitochondrial inner membrane; MOM, mitochondrial outer membrane; mtDNA, mitochondrial DNA; mtRibosome, mitochondrial ribosome; NAD, nicotinamide adenine dinucleotide; PDH, pyruvate dehydrogenase; TCA, tricarboxylic acid cycle; vCLAMP, vacuole and mitochondria patch. The pie charts show the total number of characterized and uncharacterized genes profiled (top); the total number of profiled genes that have human homologs (upper middle); of these genes with human homologs, the number of profiled genes that are also associated with disease (lower middle); and of the uncharacterized genes profiled, the number of genes that have human homologs (bottom).

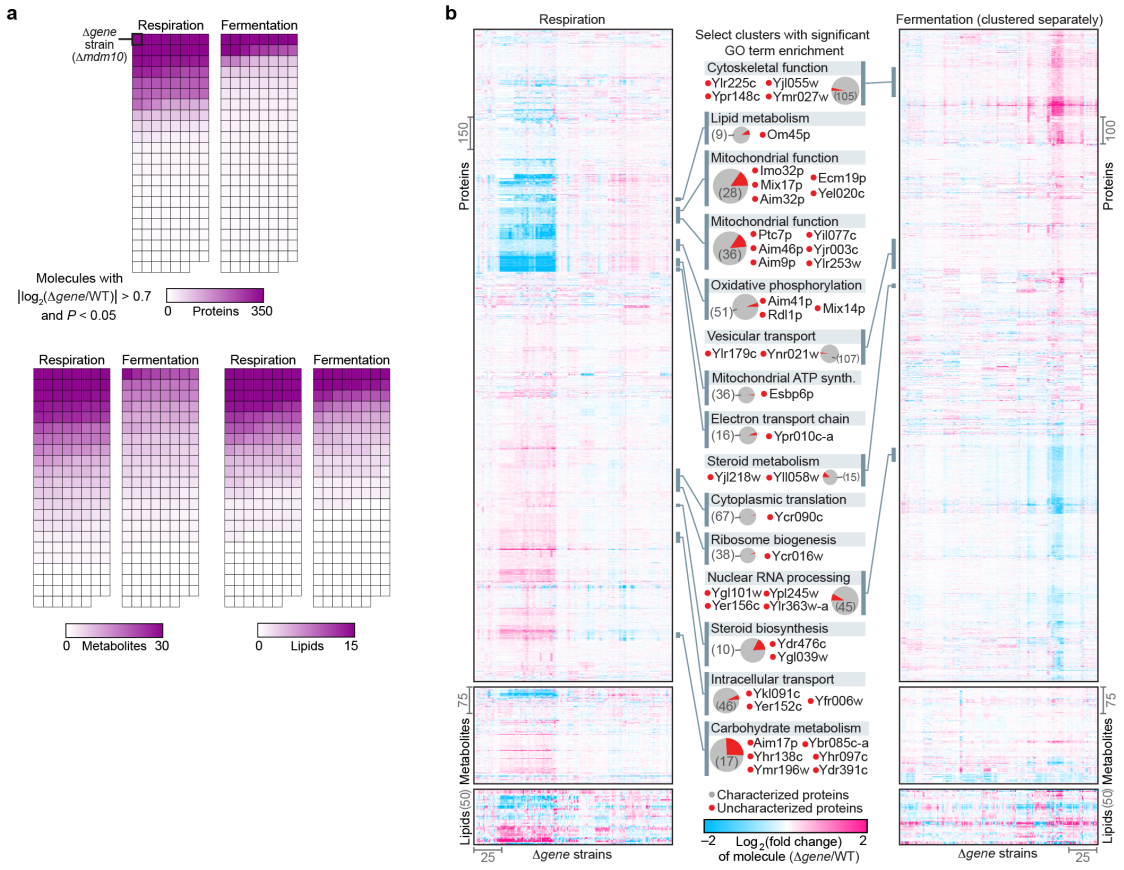
Supplementary Figure S3.1: (b) Density of yeast cultures in the respiratory growth condition (mean, n = 3) plotted in strain rank order (left) or against fermentation culture density (mean, n = 3) (right). **(c)** Optical density at 600 nm (OD_{600}) of yeast cultures (media with 3% [w/v] glycerol and 0.1% [w/v] glucose) indicating time points at which yeast were harvested during fermentation (F1–F3) or respiration (R4–R8). Time point R6 (25 h) was selected for the respiration culture condition of the larger study. **(d)** Whole-proteome plot of protein abundances at time points R5 and R8. **(e)** Pairwise whole proteome plot comparisons (as in d) across all eight time points (lower left) and linear regression analysis of each comparison (r^2, Pearson correlation coefficients) (upper right).

used to generate biological hypotheses **(Supplementary Figs. S3.5 and S3.6)**. We decided to investigate one of these observations at biochemical depth: a Δ*hfd1*-specific decrease in 4-hydroxybenzoate (4-HB), the CoQ headgroup precursor **(Fig. 3.2c)**.

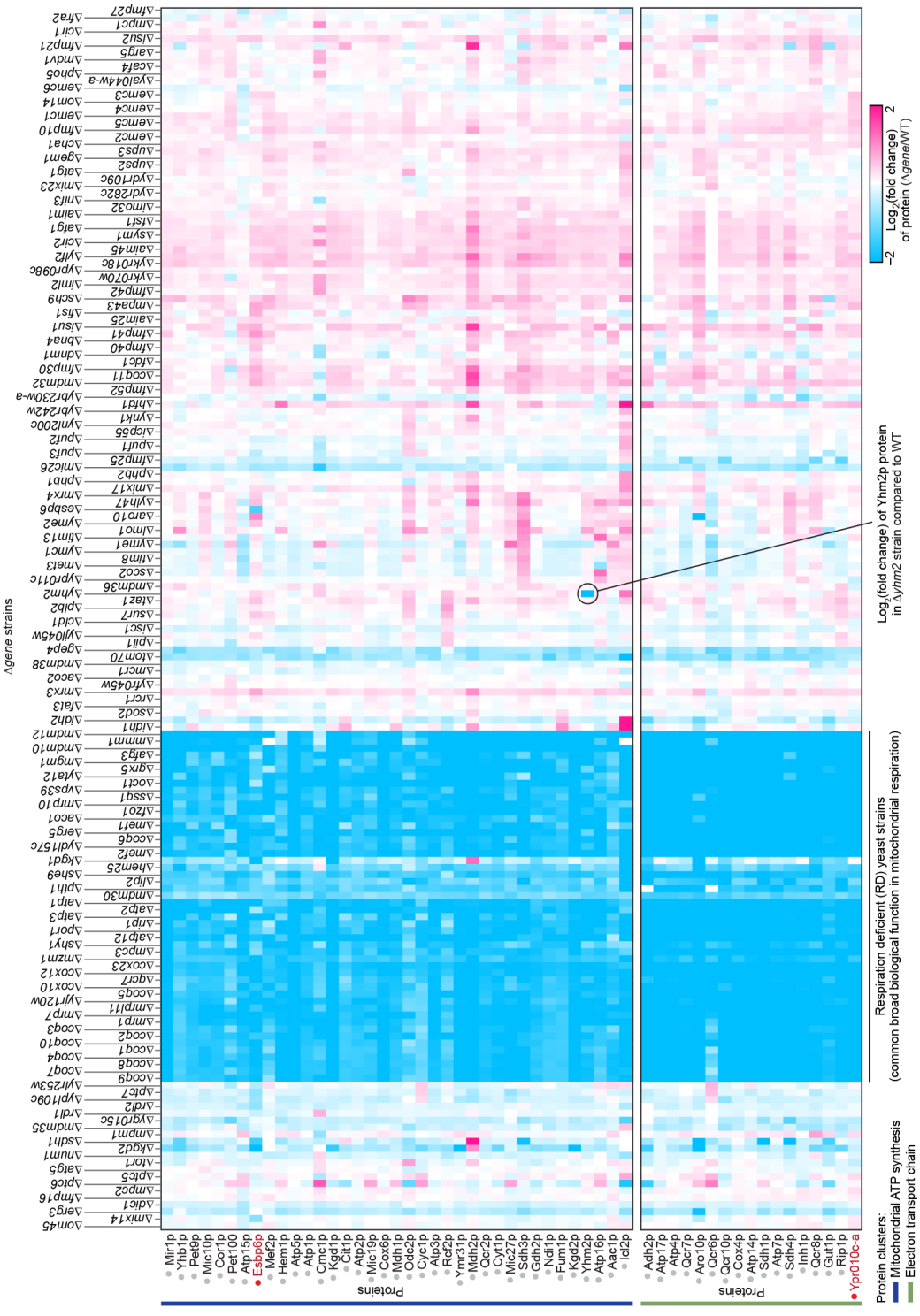
Though it has been known for decades that mammals can convert tyrosine (Tyr) into 4-HB for CoQ biosynthesis[19,20], the biochemical pathway has remained undefined in mammals and yeast **(Fig. 3.2c)**. The Y3K dataset reveals Δ*hfd1* yeast to be significantly deficient in both the metabolite 4-HB ($P < 0.001$) and the lipid CoQ intermediate 3-polyprenyl-4-hydroxybenzoate (PPHB) ($P < 10^{-5}$) **(Fig. 3.2c and Supplementary Fig. S3.7a)**. Despite the PPHB deficiency, Δ*hfd1* yeast have normal CoQ abundance **(Fig. 3.2c)**, likely because of increased flux through an alternative para-amino-benzoate (pABA)- dependent CoQ pathway[21,22], as suggested by elevation of the aminated analog of PPHB (PPAB) in Δ*hfd1* yeast **(Fig. 3.2c)**. This is in contrast to terminal CoQ biosynthesis genes (*coq3–coq9*), and some genes not previously linked to CoQ function (e.g. *oct1* and *fzo1*), whose deletion causes significant ($P < 0.05$) CoQ deficiency and accumulation of PPHB **(Fig. 3.2c)**. Because Hfd1p is predicted to be an aldehyde dehydrogenase[23], we hypothesized that it catalyzes



Supplementary Figure S3.2: Mass spectrometry analysis metrics and quality assessment. (a) Proteins, lipids, and metabolites quantified per Δ*gene* strain (mean ± s.d., n = 3). **(b)** MS experiments conducted per day (top) and phenotypes (molecules) quantified per day (bottom) for proteomics, lipidomics, and metabolomics. **(c)** Overview of the yeast protein extraction method optimized for this study compared to previous work. **(d)** Violin plots depicting the range of fold changes in molecule abundance ($\log_2$[Δ*gene*/WT]) across all molecule classes and metabolic states. **(e)** Density plots of the distribution of coefficients of variation (CVs) (%) for each molecule measured in biological triplicate across all mutants and growth conditions. **(f)** Venn diagrams depicting the average overlap of molecules quantified within individual Δ*gene* strains across fermentation and respiration growth conditions. **(g)** Average profile overlap between different Δ*gene* strains.



Supplementary Figure S3.3: Features of protein-lipid-metabolite perturbation profiles. **(a)** Heat maps depicting the number of molecules significantly perturbed within each Δ*gene* strain ($P < 0.05$; two-sided Student's t-test). **(b)** Hierarchical clusters of Δ*gene* strains and significantly perturbed molecules (relative abundances compared to WT quantified by MS; $P < 0.05$; two-sided Student's t-test). The center column annotates select clusters with significant functional (GO term) enrichments ($P < 0.05$; Fisher's exact test followed by Benjamini-Hochberg FDR correction for multiple hypothesis testing). Pie charts indicate proteins in clusters encoded by characterized (gray) or uncharacterized (red) genes.



Supplementary Figure S3.4: Expanded view of two protein clusters from the respiration Y3K dataset heat map (respiration profiles). Heat map indicates relative abundance of proteins in Δ*gene* strains compared to WT as quantified by MS. See **Supplementary Figure S3.3** for the full heat map.

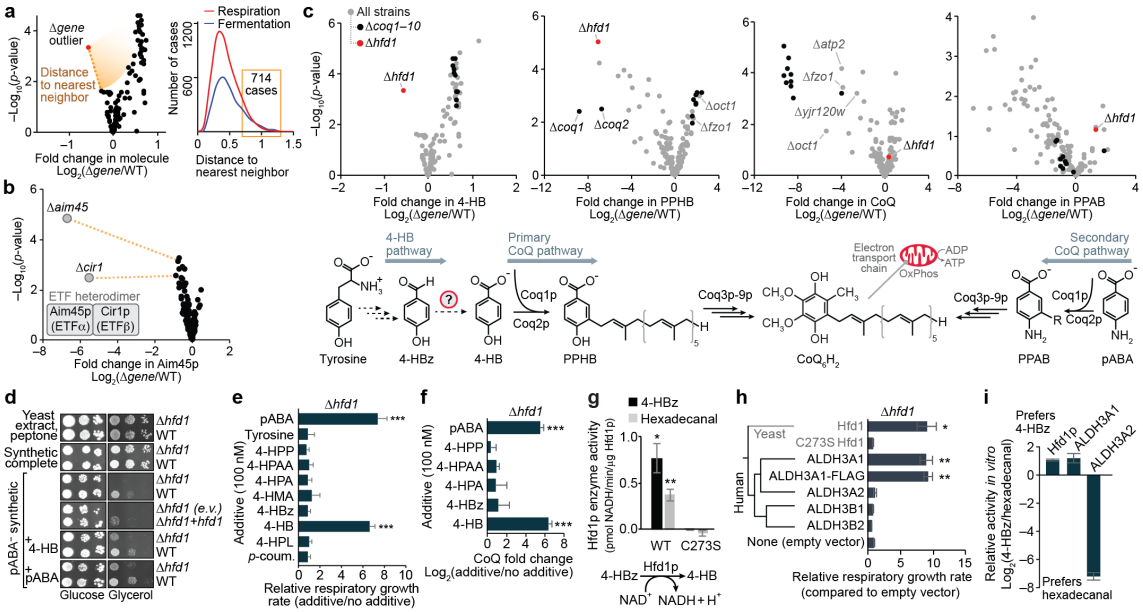
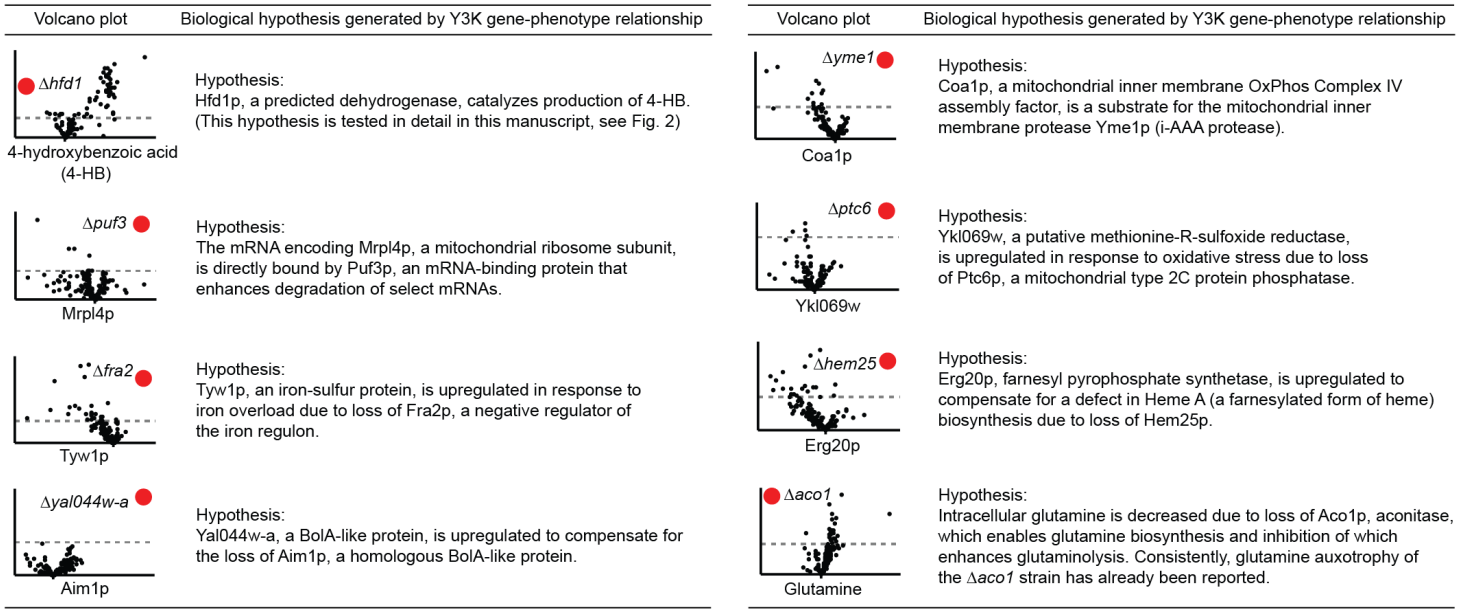


Figure 3.2: Δ*Gene*-specific phenotype detection links Hfd1p to production of 4-hydroxybenzoate for coenzyme Q biosynthesis. (a) Overview of the Δ*gene*-specific phenotype detection approach and number of Δ*gene*-specific phenotypes identified in the respiration and fermentation datasets (distance to nearest neighbor on a normalized scale, see Supplementary Note 2). **(b)** Relative abundance of Aim45p (mean, n = 3) versus statistical significance across strains. **(c)** Relative abundances of 4-HB, PPHB, CoQ, and PPAB (mean, n = 3) versus statistical significance across Δ*gene* strains. **(d)** Serial dilutions of yeast grown on variable solid medias. E.v., empty vector; +hfd1, hfd1 plasmid transformed. **(e)** Relative respiratory growth rates of Δ*hfd1* yeast in pABA$^-$ synthetic media with the additives shown (mean ± s.d, n = 3). 4-HPP, 4-hydroxyphenylpyruvate; 4-HPAA, 4-hydroxyphenylacetaldehyde; 4-HPA, 4-hydroxyphenylacetate; 4-HMA, 4-hydroxymandelate; 4-HPL, 4-hydroxyphenyllactate; p-coum., para-coumarate. **(f)** Relative CoQ abundance in Δ*hfd1* yeast cultured in pABA$^-$ media with the additives shown (mean ± s.d., n = 3). **(g)** Enzyme activity of recombinant MBP-Hfd1$^{C\Delta25}$ in vitro against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(h)** Phylogenetic relationship between yeast Hfd1p and the human ALDH3 family, and relative respiratory growth rates of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured in pABA$^-$ synthetic media (mean ± s.d, n = 4). **(i)** Relative activity of the dehydrogenases shown against 4-HBz compared to hexadecanal (mean ± s.e.m., n = 3). **P* < 0.05; ***P* < 0.01; ****P* < 0.001 (two-sided Student's t-test for all panels).

Supplementary Figure S3.5: Subsets of the Δ*gene*-specific phenotypes identified in this study. Relative abundances of individual molecules (mean $\log_2$[Δ*gene*/WT], n = 3) (x-axes) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) (y-axes) as quantified by MS. The plots shown represent a subset of molecules identified as 'Δ*gene*-specific phenotypes' through an unbiased survey of the Y3K dataset (see Fig. 2a). The array here is limited to the most robust outliers (based on both statistical significance and fold-change, see Supplementary Note 2 and Methods)—the top 20 upregulated proteins, the top 20 downregulated proteins, the top 10 metabolites, and the top 4 or 5 lipids—excluding 'knocked out proteins' (e.g. Fmp52p in the Δ*fmp52* strain) and excluding a given Δ*gene* strain after it appeared twice on the rank list. Biological hypotheses surrounding gene-phenotype relationship were generated for the starred plots (see Supplementary Fig. 6).

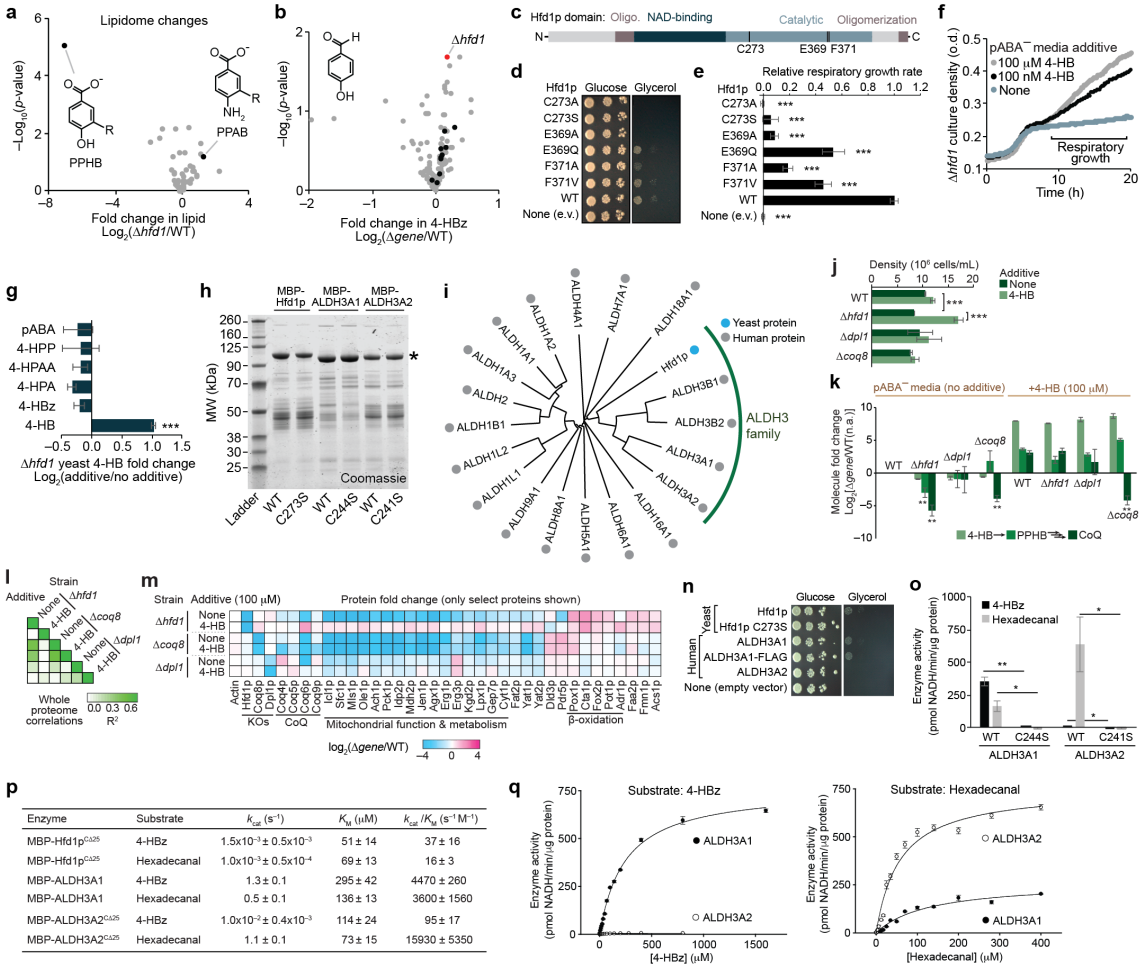


Supplementary Figure S3.6: Examples of hypotheses that can be generated from a subset of the Δ*gene*-specific phenotypes identified in this study. Subset of Δ*gene*-specific phenotypes identified in the Y3K dataset. Volcano plots indicate relative molecule abundances (mean $\log_2$[Δ*gene*/WT], n = 3) (x-axes) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) (y-axes) as quantified by MS. Hypotheses were developed to describe each Δ*gene*-phenotype relationship reported here.

dehydrogenation of 4-hydroxybenzaldehyde (4-HBz) to form 4-HB. Consistently, 4-HBz is elevated in Δ*hfd1* yeast **(Supplementary Fig. S3.7b)**.

We used chemical-genetics to test the proposed Hfd1p activity. Most culture media contain either 4-HB (in yeast extract) or pABA (in standard yeast nitrogen base), enabling yeast to bypass the Tyr-to-4-HB pathway, so we used a defined medium lacking pABA and 4-HB ("pABA$^-$"). Δ*hfd1* yeast exhibited striking respiration deficiency on pABA$^-$ media, a phenotype rescued by pABA, 4-HB, or WT Hfd1p, but not by Hfd1p with mutations to putative catalytic residues[24] **(Fig. 3.2d and Supplementary Fig. S3.7c–e)**. Testing a panel of potential intermediates in the pathway revealed that 4-HB, but not 4-HBz, can rescue the respiratory growth and CoQ production of Δ*hfd1* yeast **(Fig. 3.2e,f and Supplementary Fig. S3.7f,g)**, supporting a role for Hfd1p in dehydrogenation of 4-HBz. To directly test this activity, we purified recombinant Hfd1p for enzyme assays **(Supplementary Fig. S3.7h)**. WT Hfd1p catalyzes NAD^+-dependent dehydrogenation of 4-HBz, but a C273S (catalytic residue) point mutant does not **(Fig. 3.2g)**. Together, these results demonstrate that Hfd1p dehydrogenates 4-HBz to produce 4-HB for CoQ biosynthesis.

Hfd1p is a member of the ancient aldehyde dehydrogenase (ALDH) superfamily, which is found across all three superkingdoms of life and includes 19 human homologs with diverse functions[25]. Based on phylogenetic analyses, Hfd1p is most similar to the human ALDH3 family **(Supplementary Fig. S3.7i)**. ALDH3A2 (FALDH) mutations cause Sjögren–Larsson Syndrome[26] due to defective fatty aldehyde metabolism. However, the endogenous functions of ALDH3A1, B1, and B2 remain obscure, and which of these human ALDH3



Enzyme	Substrate	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($s^{-1}M^{-1}$)
MBP-Hfd1p^{CΔ25}	4-HBz	$1.5x10^{-3} \pm 0.5x10^{-3}$	51 ± 14	37 ± 16
MBP-Hfd1p^{CΔ25}	Hexadecanal	$1.0x10^{-3} \pm 0.5x10^{-4}$	69 ± 13	16 ± 3
MBP-ALDH3A1	4-HBz	1.3 ± 0.1	295 ± 42	4470 ± 260
MBP-ALDH3A1	Hexadecanal	0.5 ± 0.1	136 ± 13	3600 ± 1560
MBP-ALDH3A2^{CΔ25}	4-HBz	$1.0x10^{-2} \pm 0.4x10^{-3}$	114 ± 24	95 ± 17
MBP-ALDH3A2^{CΔ25}	Hexadecanal	1.1 ± 0.1	73 ± 15	15930 ± 5350

Supplementary Figure S3.7: Hfd1p supports production of 4-HB for CoQ biosynthesis. **(a)** Relative lipid abundances (mean, n = 3) versus statistical significance ($-\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS. **(b)** Relative abundances of 4-HBz (mean, n = 3) versus statistical significance ($-\log_{10}$[*p*-value]; two-sided Student's t-test) across all Δgene strains in the study. **(c)** Protein domain structures of Hfd1p, highlighting residues involved in catalysis. **(d)** Serial dilutions of 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)** Growth curves showing the respiratory growth of Δ*hfd1* yeast in pABA$^-$ synthetic media with the additives shown. **(g)** Relative 4-HB abundance in Δ*hfd1* yeast cultured in pABA$^-$ media with the additives shown (mean $\log_2$[additive/unsupplemented] ± s.d., n = 3). **(h)** SDS-PAGE analysis (Coomassie stained gel) of protein fractions from an isolation of MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (WT and catalytically dead mutant for each). **(i)** Phylogenetic tree of human ALDH superfamily members and yeast Hfd1p.

Supplementary Figure S3.7: (j) Density of yeast (upon harvest) cultured in pABA$^-$ media ± 4-HB (mean ± s.d., n = 3). **(k)** Relative abundances of 4-HB, PPHB, and CoQ compared to WT yeast cultured in pABA$^-$ media (mean $\log_2$[Δgene/WT with no additive] ± s.d., n = 3) as quantified by MS. **(l)** Whole proteome correlation map for yeast grown in pABA$^-$ media ± 4-HB (mean, n = 3). **(m)** Relative abundances of select proteins as quantified by MS (mean $\log_2$[Δgene/WT], n = 3) analysis of yeast cultured in pABA$^-$ media ± 4-HB. **(n)** Serial dilutions of Δ*hfd1* yeast transformed with plasmids encoding the proteins shown and cultured on solid pABA$^-$ synthetic media plates. **(o)** Enzyme activity of MBP-ALDH3A1 or MBP-ALDH3A2(CΔ25) against 4-HBz (200μM) or hexadecanal (200μM) (mean ± s.e.m., n = 3). **(p)** Table of enzyme kinetic parameters for MBP-Hfd1p(CΔ25), MBP-ALDH3A1, and MBP-ALDH3A2(CΔ25) (mean ± s.e.m., n = 3). **(q)** Representative enzyme kinetic curves for MBP-ALDH3A1 and MBP-ALDH3A2(CΔ25). *P < 0.05; **P < 0.01; ***P < 0.001 (two-sided Student's t-test).

functions are conserved in Hfd1p has not been completely defined. Previous work showed that sphingolipid metabolism is perturbed in Δ*hfd1* yeast due to a defect in dehydrogenation of hexadecanal, and this defect can be rescued by ALDH3A2, but not by ALDH3A123[27]. However, a separate sphingolipid pathway defect (Δ*dpl1*) does not disrupt the 4-HB-CoQ pathway **(Supplementary Fig. S3.7j–m and Supplementary Note 3)**, suggesting that the two pathways are otherwise independent. Consistent with the idea that Hfd1p is a dual-function protein that supports both sphingolipid metabolism and CoQ biosynthesis, we observed Hfd1p activity in vitro with hexadecanal, similar to that observed with 4-HBz **(Fig. 3.2g)**. However, in contrast to rescue of the sphingolipid metabolism defect, we found that ALDH3A1, but not ALDH3A2, rescues the pABA$^-$ respiratory growth phenotype of Δ*hfd1* yeast **(Fig. 3.2h and Supplementary Fig. S3.7n)**. Moreover, while ALDH3A2 shows a strong substrate preference for hexadecanal over 4-HBz, Hfd1p and ALDH3A1 show a preference for 4-HBz **(Fig. 3.2i and Supplementary Fig. S3.7o–q)**. These results

suggest that the dual functions of yeast Hfd1p have diverged in human ALDH3A1 and ALDH3A2. Collectively, these results demonstrate a major cellular function for the aldehyde dehydrogenase Hfd1p in the Tyr-to-4-HB pathway and strongly suggest that ALDH3A1 plays a similar role in human CoQ biosynthesis.

Regression analysis of global perturbation profiles. While molecular changes unique to a given Δ*gene* strain can be functionally informative, similarities between Δ*gene* strains can also assist characterization. In our second analysis approach, we examined Δ*gene*–Δ*gene* correlations through pairwise comparisons of global Δ*gene* perturbation profiles. Deletion of functionally related genes, such as the cytochrome c oxidase genes *cox12* and *cox23*, caused highly similar whole proteome perturbations **(Fig. 3.3a)**. Notably, highly correlated phenotype changes were also observed in Δ*cox12* and Δ*cox23* metabolomes and lipidomes **(Fig. 3.3a)**. However, deletion of unrelated genes, such as *cox12* and *mic26*, generated uncorrelated phenotype changes **(Fig. 3.3a)**. Examination of Δ*gene*–Δ*gene* correlations across the entire study indicated numerous functional relationships, with stronger correlations observed in respiration **(Fig. 3.3b)**.

A group of respiration-deficient (RD) strains showed robust correlations across all three omes **(Fig. 3.3b)**, reflecting their similar broad biological functions in mitochondrial OxPhos and suggesting that they share a universal "respiration deficiency response" (RDR). Multi-omic principle component and GO term analyses revealed a coordinated RDR that provides biological insight into respiration defects—a common feature of many diseases

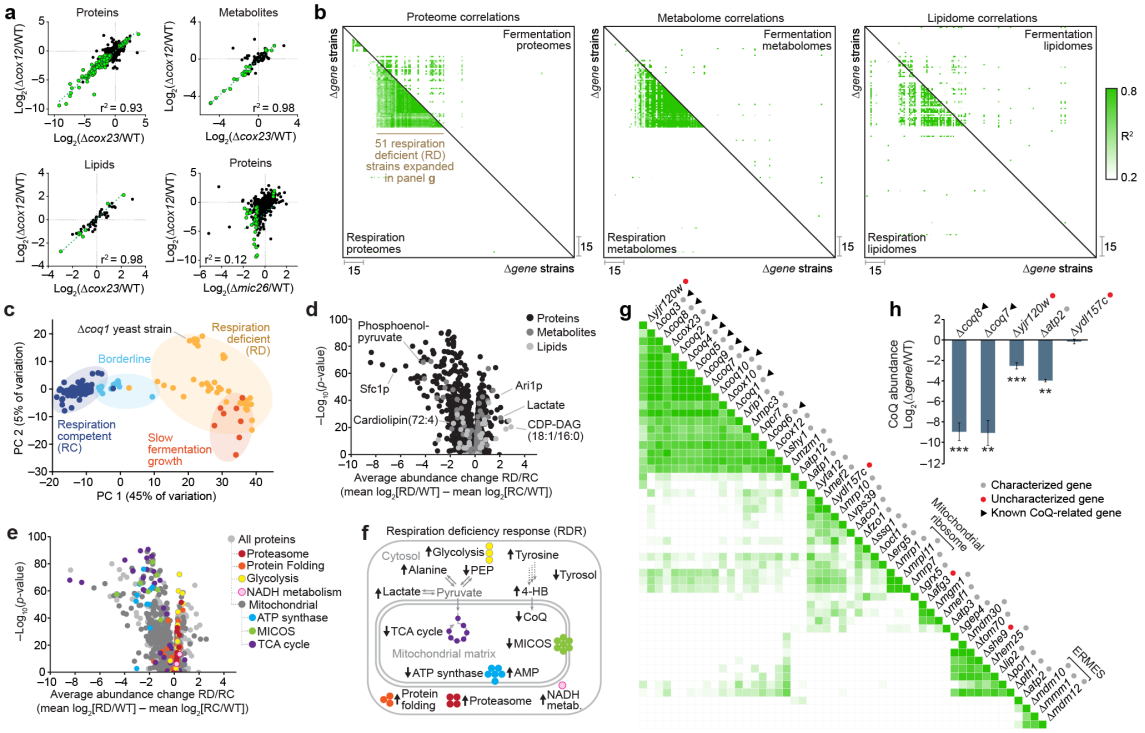
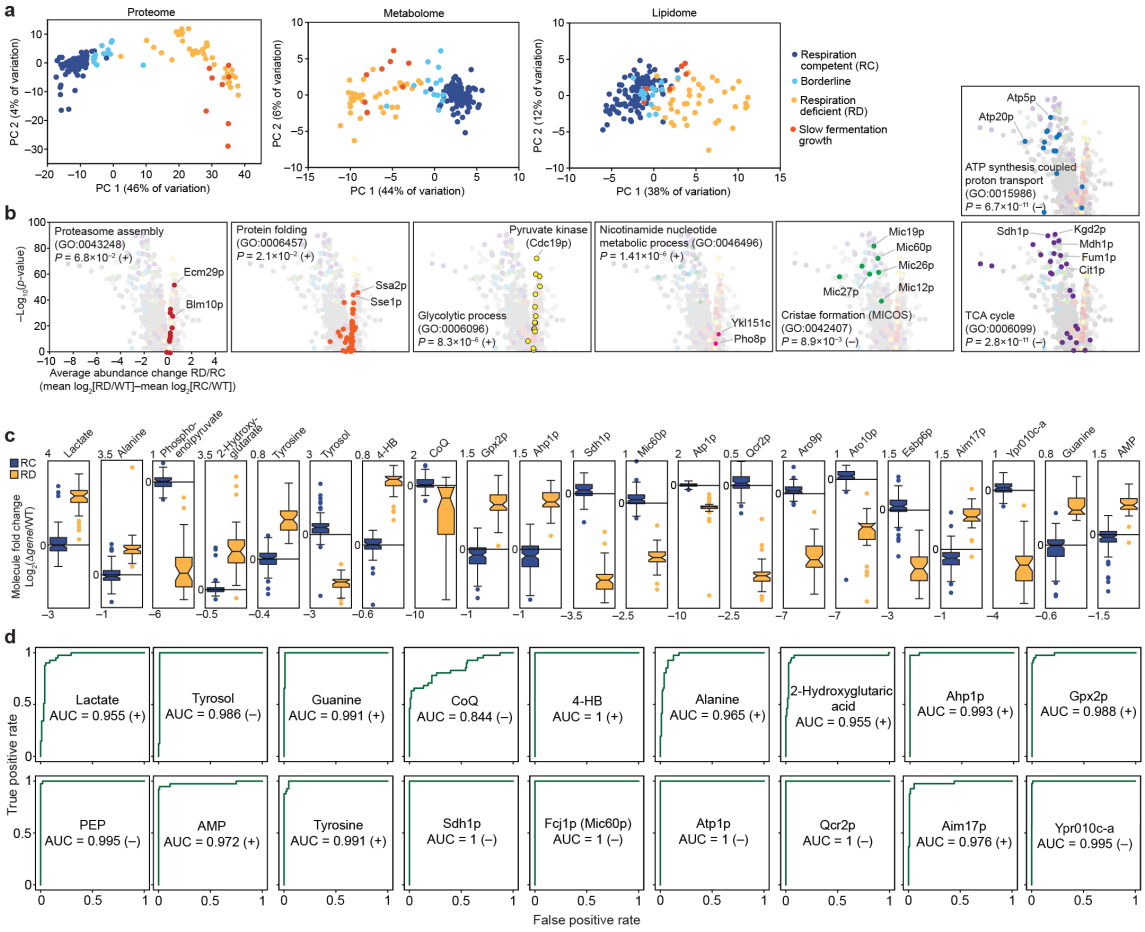


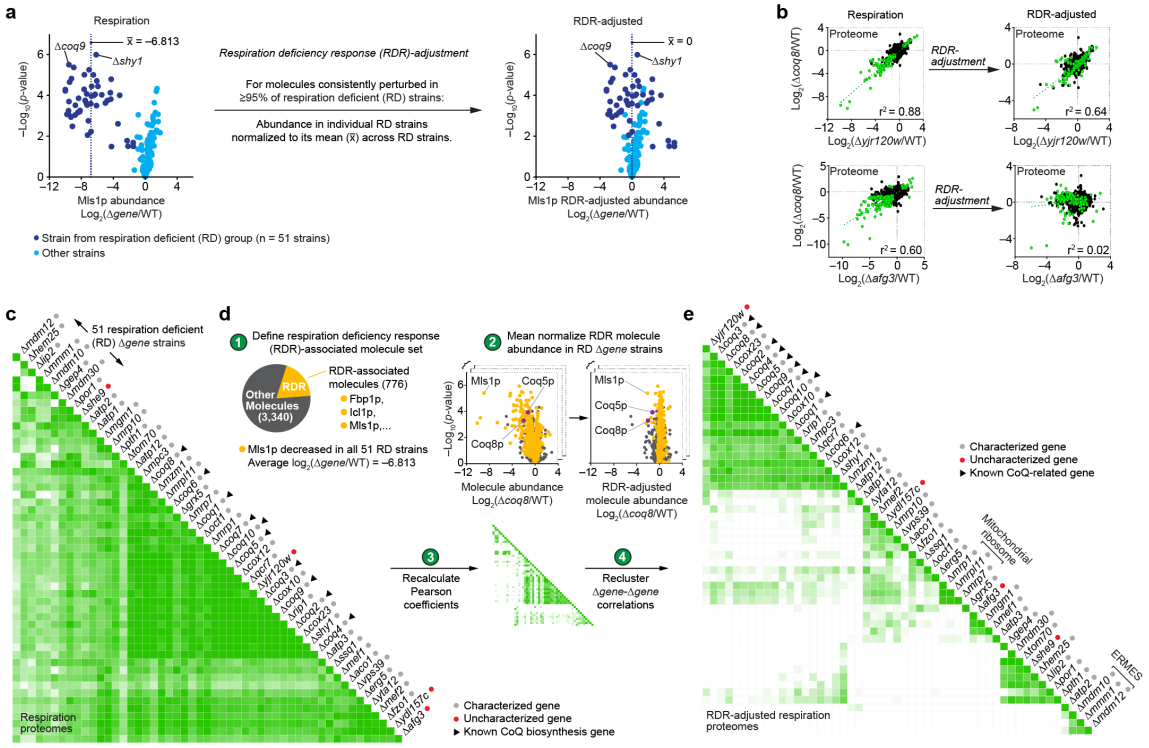
Figure 3.3: Functional correlations through perturbation profile regression analysis. (a) Plots comparing relative molecule abundances between pairs of Δ*gene* strains. Strain-strain similarity assessed by linear regression analysis of Δ*gene* perturbation profiles. Green points indicate molecules significantly perturbed in both mutants ($|\log_2[FC]| > 0.7$, $P < 0.05$; two-sided Student's t-test). **(b)** Maps of Pearson correlation coefficients (r^2) for pairs of Δ*gene* perturbation profiles across omes and metabolic conditions. Strains are clustered based on respiration proteome correlations, and this strain order is held consistent across all 6 maps. **(c)** Projection of respiration competent (RC) and deficient (RD) strains onto the plane defined by principal component (PC) axes 1 and 2 (full multi-omic respiration dataset). **(d)** Average fold change in molecule abundances (mean $\log_2$[RD strains/RC strains]) versus statistical significance ($-\log_{10}$[p-value, Bonferroni corrected two-sided t-test]). **(e)** RD versus RC proteome perturbation volcano plot (as in d) showing select functional groups (GO terms) significantly enriched in either upregulated or downregulated proteins. **(f)** Scheme of RDR pathways. **(g)** Re-clustered respiration proteome strain-strain correlation map following RDR- adjustment. (h) CoQ abundance changes in select Δ*gene* strains (mean ± s.d., n = 3); $^{**}P < 0.01$; $^{***}P < 0.001$ (two-sided Student's t-test).

including cancer—and suggests that a multi-omic biomarker fingerprint could afford a specific diagnostic for mitochondrial disease **(Fig. 3.3c–f, Supplementary Fig. S3.8, and Supplementary Note 4)**. However, stress responses such as the RDR also pose a barrier to biochemical investigations because they can obscure functionally-informative phenotypes. To assess more specific biochemical roles for individual proteins, we normalized for the RDR across RD strains **(Supplementary Fig. S3.9 and Supplementary Note 5)**. Across all of our RD strains, 776 molecules were identified as being consistently perturbed. The individual measurements of these RDR-associated molecules were mean normalized ("RDR-adjusted") to reveal characteristic deviations from the general RDR and to enable visualization of Δ*gene*-specific changes.

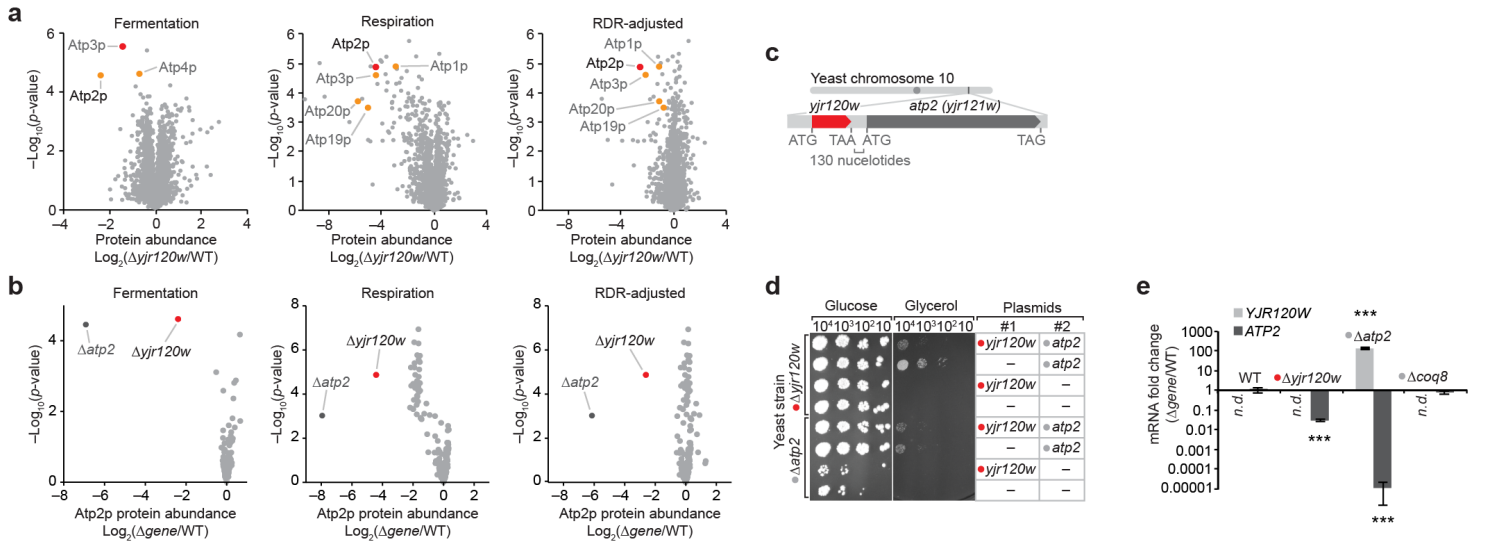
Recalculating Δ*gene*–Δ*gene* correlation coefficients with RDR-adjusted plots strikingly reduces correlations between more functionally disparate genes **(Supplementary Fig. S3.9c–e)**. Reclustering Δ*gene*–Δ*gene* correlations reveals new clusters of genes with similar biochemical functions **(Fig. 3.3g)**. For example, known CoQ biosynthesis genes were brought into a tighter cluster that also includes the uncharacterized gene *yjr120w* **(Fig. 3.3g)**, suggesting that *yjr120w* might support CoQ biosynthesis. Consistently, we observed CoQ deficiency in Δ*yjr120w* yeast **(Fig. 3.3h)**, the molecular basis of which we determined to include loss of Atp2p, an ATP synthase subunit **(Supplementary Fig. S3.10 and Supplementary Note 6)**. These results show that specific ATP synthase subunits support CoQ biosynthesis and, more broadly, demonstrate how global mass spectrometry profiling can reveal functional links between genes.



Supplementary Figure S3.8: Identification of respiration deficiency response pathways and potential biomarkers. (a) Projection of RC and RD strains onto the planes defined by principal component (PC) axes 1 and 2 for separate proteome, metabolome, and lipidome PC analyses. **(b)** RD versus RC proteome perturbation volcano plots (as in Fig. 3e) showing select functional groups (GO terms) significantly enriched (Bonferroni corrected p-values shown in figure) in either upregulated or downregulated proteins. **(c)** Box plots depicting median molecule fold changes for RC and RD strains ($\log_2$[RD or RC average/WT]) (n = 111 for RC, 41 for RD). Notch indicates 95% c.i. **(d)** Receiver operating characteristic (ROC) curves for select molecules depicting the false positive rates and true positive rates for prediction of respiration deficiency associated with particular molecule fold changes. AUC, area under the curve.



Supplementary Figure S3.9: Subtraction of shared responses to reveal deeper biochemical insight. (a) RDR-abundance adjustment of a representative molecule (Mls1p) by subtraction of the average fold change in abundance (mean $\log_2[\Delta gene/\text{WT}]$, n = 3) across respiration deficient (RD) strains. This adjustment was only performed within RD strains. **(b)** Plots comparing relative protein abundances between pairs of Δ*gene* strains. Linear regression analysis of pairs of perturbation profiles before (left) and after (right) RD-abundance adjustment. Green points indicate molecules significantly perturbed in both mutants ($|\log_2(FC)| > 0.7$; $P < 0.05$; two-sided Student's t-test) prior to RDR-adjustment. **(c)** Expanded view of highly correlated strains in the respiration proteomes correlation map (see Fig. 3b). **(d)** Procedure for normalization of the RDR. **(e)** Re-clustered respiration proteome strain-strain correlation map following RDR-adjustment (also shown in Fig. 3g).



Supplementary Figure S3.10: Molecular perturbations of yeast lacking *yjr120w*. **(a)** Relative protein abundances (mean $\log_2$[Δ*yjr120w*/WT], n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS. (b) Relative Atp2p protein abundance (mean $\log_2$[Δ*gene*/WT], n = 3) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) across all mutants in the study. (c) Genomic organization of *yjr120w* and *atp2*. (d) Serial dilutions of yeast transformed with the indicated plasmids grown on agar plates with glucose (to enable fermentation) or glycerol (to force respiration). (e) Fold changes in mRNA abundances (mean Δ*gene*/WT, n = 3) as quantified by real time polymerase chain reaction (RT-PCR) analysis. Yjr120w mRNA was not detected (n.d.) in WT yeast, so imputation of this missing value was used to calculate the fold increase in yjr120w mRNA shown for the Δ*atp2* strain. **P* < 0.05; ***P* < 0.01; ****P* < 0.001 (two-sided Student's t-test).

Molecule covariance network analysis. Similarly, in our third analysis approach, we leveraged the multi-omic nature of our mass spectrometry profiles to determine pairwise covariance between proteins, metabolites, and lipids. This approach is similar to mRNA coexpression profiling, which can be used to predict gene function[28–30], but it integrates three complementary classes of molecules. Perturbations for functionally related molecules, such as the protein Coq4p and the lipid CoQ intermediate PPHB, show strong positive or negative correlations, while those of unrelated molecules, such as Coq4p and Rpb4p, lack correlations **(Fig. 3.4a)**. Correlated molecules include proteins in complexes, such as the cytosolic TRiC/CCT chaperonin complex (Cct2p and Cct7p), and enzyme-product pairs (e.g. Ura1p and orotic acid) **(Fig. 3.4a)**.

Examining correlations across all 4,505 molecules in the Y3K dataset through this multi-omic molecule covariance network analysis (MCNA) reveals numerous functional relationships, which can be visualized as networks of molecules (nodes) and correlations (edges) **(Fig. 3.4b and Supplementary Fig. S3.11a)**. After applying strict correlation thresholds (Bonferroni-adjusted p-value < 0.001, $|\rho| \geqslant 0.58$), 237,342 edges remain among 2,382 nodes in the respiration dataset **(Supplementary Fig. S3.11a–f)**. Many edges were observed between RDR-associated molecules **(Supplementary Fig. S3.11g)**, reflecting their common relationship to mitochondrial metabolism. As described above for Δ*gene* correlations, we deepened the molecular insight of the MCNA by RDR-adjustment, which reduced overall connectivity and increased the selectivity of functionally related molecule sub-networks **(Supplementary Fig. S3.11g)**. For example, the selectivity of the mitochondrial ribosome

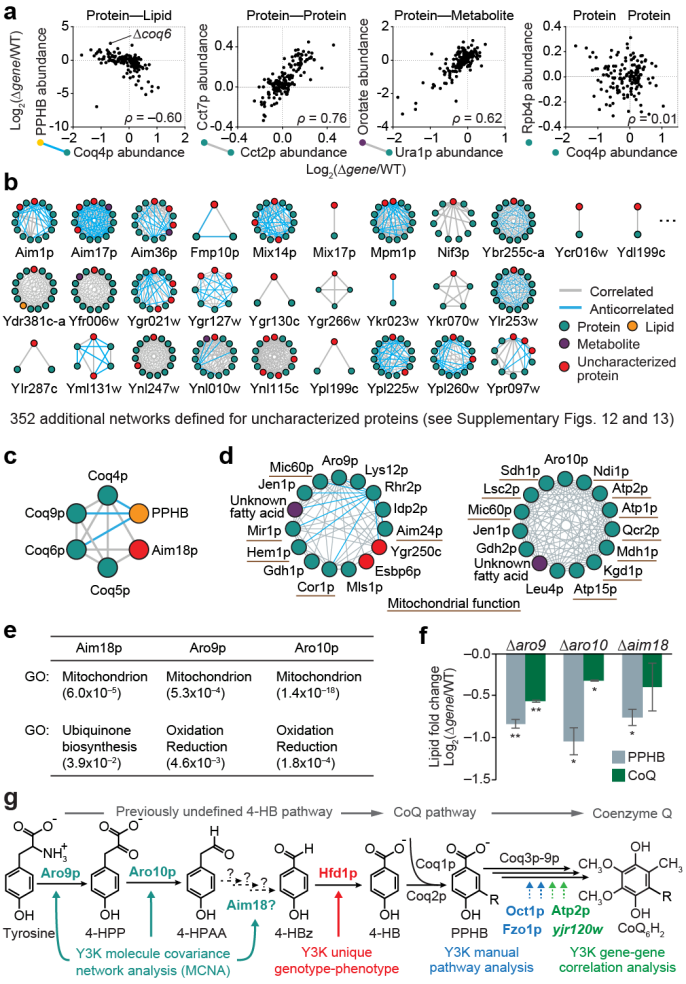


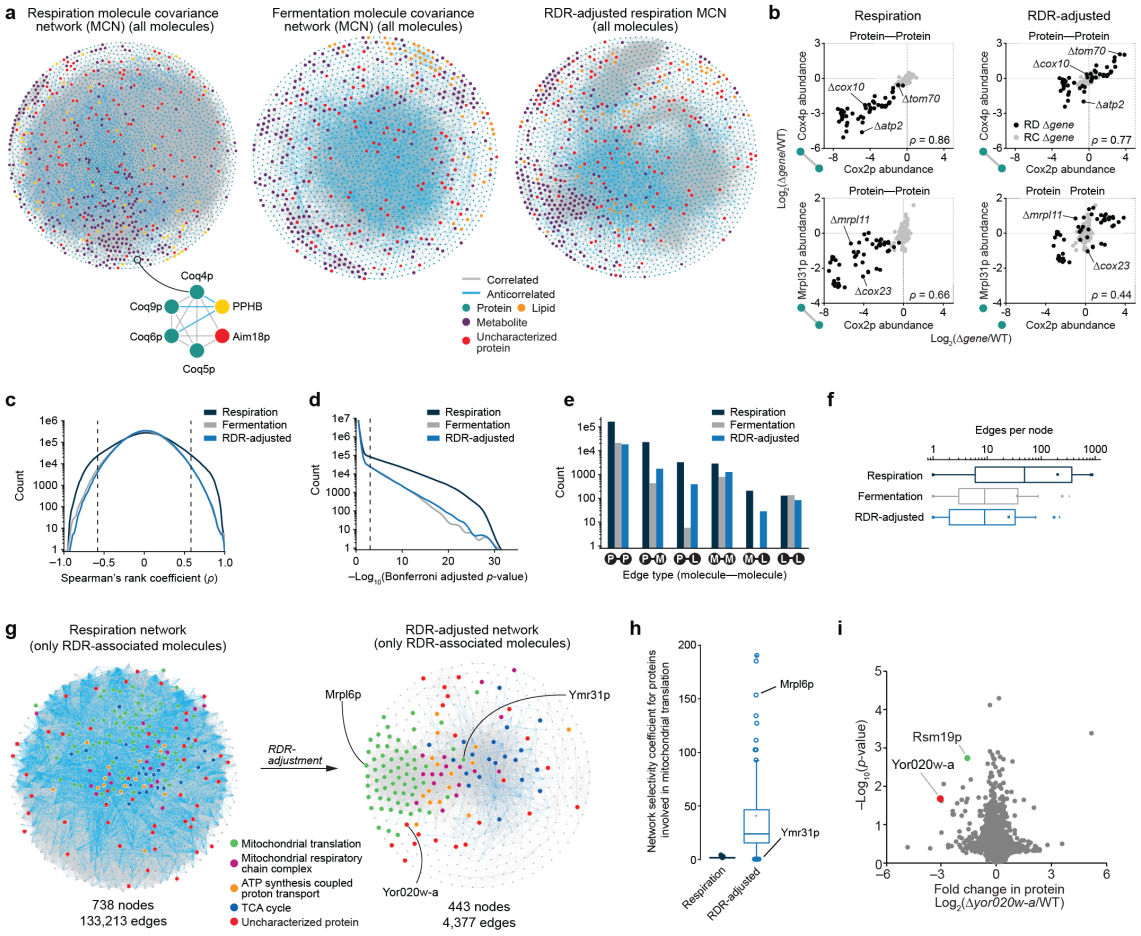
Figure 3.4: Multi-omic molecule covariance network analysis assists functional characterization. **(a)** Relative abundances of molecule pairs across Δ*gene* strains. Covariance assessed by Spearman's rank coefficient (ρ). **(b)** Nearest neighbor molecule covariance networks for a representative subset of uncharacterized proteins. **(c)** Network for Coq4p in the RDR-adjusted respiration dataset. **(d)** Networks showing the 14 molecules most strongly correlated to Aro9p or Aro10p in the RDR-adjusted respiration dataset. **(e)** GO term analyses of the Aim18p, Aro9p, and Aro10p networks (*p*-values). **(f)** Relative abundances of CoQ and PPHB (mean $\log_2$[Δ*gene*/WT], n = 2) in Δ*aro9*, Δ*aro10*, and Δ*aim18* strains compared to WT yeast cultured in pABA$^-$ media; $^*P < 0.05$; $^{**}P < 0.01$ (two-sided Student's t-test). **(g)** Y3K-enabled characterization of proteins that support the CoQ pathway.

sub-network increased 16-fold **(Supplementary Fig. S3.11h)**. These RDR-adjusted networks associated the MXP Yor020w-a with the mitochondrial ribosome **(Supplementary Fig. S3.11g)**. To test this association, we examined the proteome of Δ*yor020w-a* yeast, which showed a significant decrease in the mitochondrial ribosome protein Rsm19p **(Supplementary Fig. S3.11i)**, suggesting that Yor020w-a is linked to mitochondrial translation.

Hundreds of additional uncharacterized proteins were linked to characterized molecules by our MCNA, providing a foundation for generating hypotheses about their functions **(Fig. 3.4b, Supplementary Figs. S3.12 and S3.13)**. For example, the MXP Aim18p was linked to a network of CoQ biosynthesis proteins, and Aro9p and Aro10p were linked to numerous mitochondrial proteins that support OxPhos **(Fig. 3.4c–e)**. Based on domain homology and predicted enzymatic functions, we hypothesized that Aim18p, Aro9p, and Aro10p could function in the Tyr- to-4-HB pathway **(Supplementary Fig. S3.14 and Supplementary Note 7)**. Consistently, when cultured in a pABA$^-$ media, Δ*aim18*, Δ*aro9*, and Δ*aro10* yeast are deficient in both CoQ and PPHB **(Fig. 3.4f)**. This work shows how global mass spectrometry profiling can be used to generate biological hypotheses and characterize protein functions through distinct multi-omic data analysis approaches **(Fig. 3.4g)**.

Discussion

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



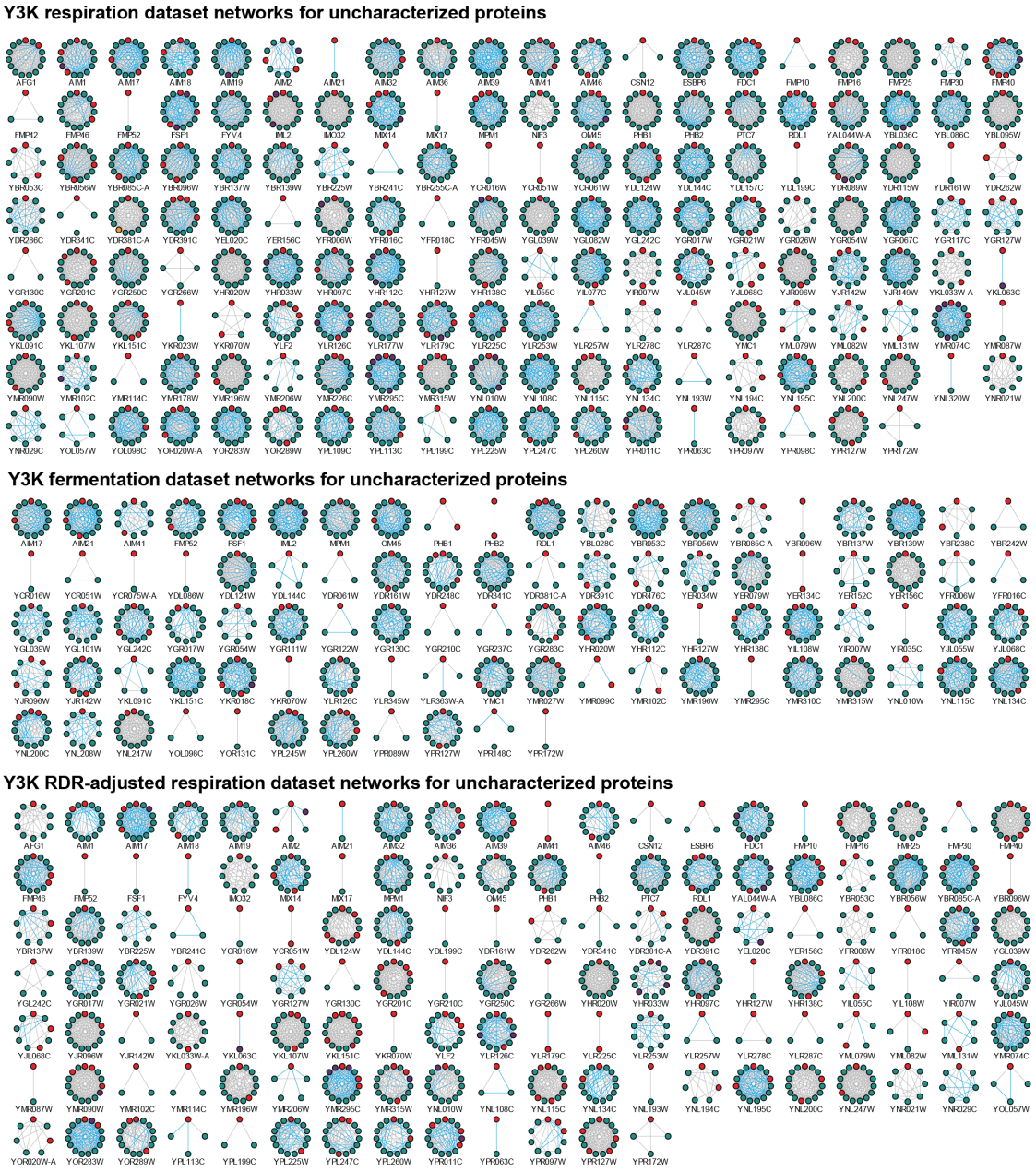
Supplementary Figure S3.11: Features of multi-omic molecule covariance networks. Network of all covariant molecules observed in each dataset ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). **(b)** Regression analysis of pairs of RDR-associated molecules before and after RDR adjustment using Spearman's rank coefficient (ρ). Points corresponding to RD and RC Δ*gene* strains are indicated. **(c)** Distribution of calculated Spearman coefficients for all pairwise molecule covariance comparisons (ρ cutoff at ±0.58 used throughout the study is indicated). **(d)** Distribution of Bonferroni-adjusted *p*-values from all pairwise molecule comparisons (*p*-value cutoff at 0.001 used throughout the study is indicated). **(e)** Bar chart indicating number of protein–protein (P–P), protein–metabolite (P–M), protein–lipid (P–L), metabolite–metabolite (M–M), metabolite–lipid (M–L), and lipid–lipid (L–L) edges in each dataset. **(f)** Box plots indicating the number of edges per node in the respiration, fermentation, and RDR-adjusted networks.

Supplementary Figure S3.11: (g) Network of all covariant RDR-associated molecules (| ρ | $\geqslant$ 0.58, Bonferroni-adjusted P < 0.001; two-sided Student's t-test) generated using the respiration (left) and RDR-adjusted (right) datasets. Nodes are highlighted according to GO category. **(h)** Box plots indicating the molecule covariance network (MCN) specificity coefficient for all nodes involved in mitochondrial translation in both the respiration and RDR-adjusted respiration RDR-associated molecule networks (shown in panel G). **(i)** Relative protein abundances (mean $\log_2$[Δ*yor020w-a*/WT], n = 2) versus statistical significance (–$\log_{10}$[*p*-value]; two-sided Student's t-test) as quantified by MS.

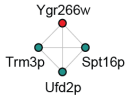
Here, we leveraged a subset of our multi-omic dataset to investigate gaps in knowledge of CoQ biosynthesis. Despite CoQ's essential function in the mitochondrial electron transport chain, role as a key cellular antioxidant, and link to numerous human diseases (e.g., ataxias, myopathies, and nephrotic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

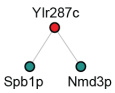
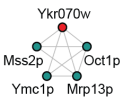
Our Δ*gene*-specific phenotype detection approach suggested a role for the ancient aldehyde dehydrogenase superfamily member Hfd1p in 4-HB biosynthesis. Biochemical and genetic studies confirmed this role for Hfd1p in yeast and further demonstrated that the human homolog ALDH3A1 can also catalyze production of 4-HB *in vivo* and *in vitro* **(Fig. 3.2)**, thereby highlighting ALDH3A1 as a candidate disease gene for primary CoQ deficiency.

Distinct Y3K dataset analyses placed additional proteins into the CoQ biosynthesis pathway. MCNA showed unexpected connections between Aro9p, Aro10p, and mitochon-

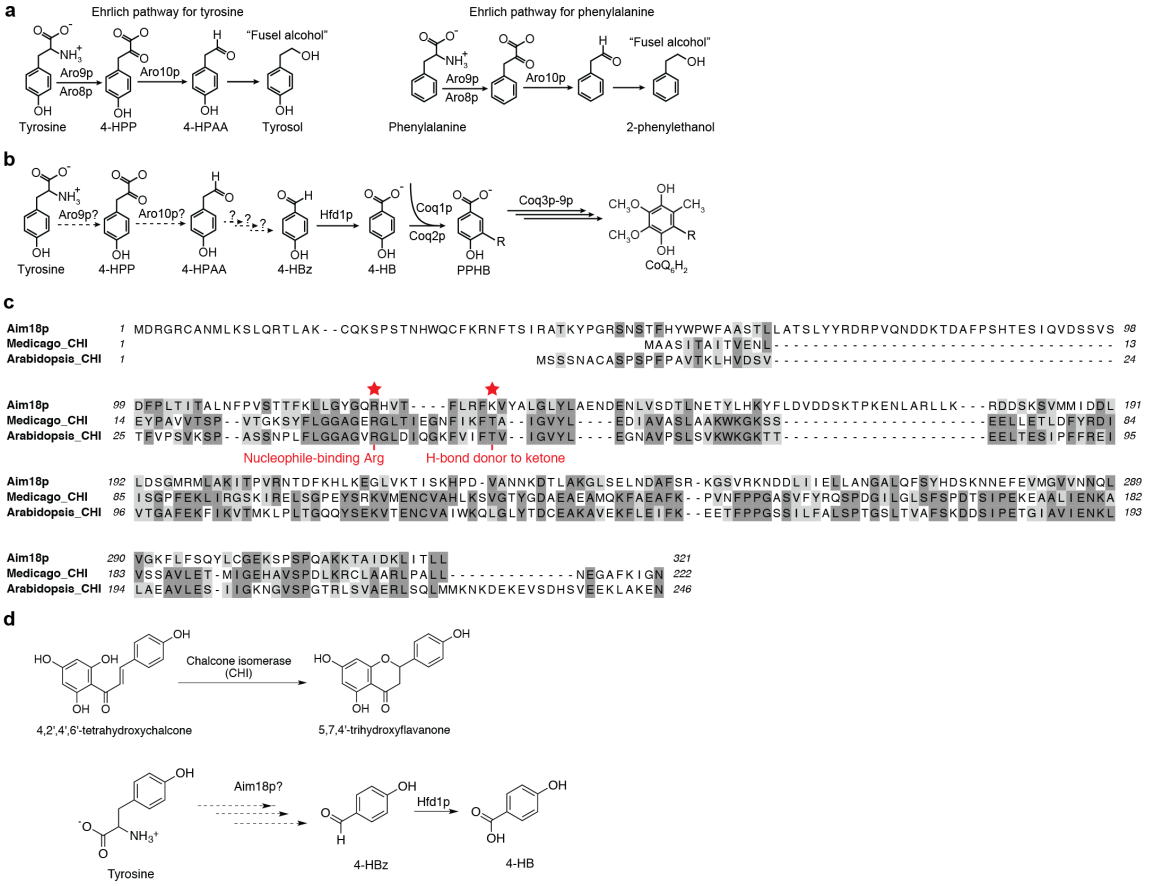


Supplementary Figure S3.12: Molecule covariance networks for uncharacterized proteins. 'Nearest neighbor' molecule covariance networks for all uncharacterized proteins observed across the respiration, fermentation, and RDR-adjusted respiration datasets ($|\rho| \geqslant 0.58$, Bonferroni-adjusted $P < 0.001$; two-sided Student's t-test). If more than 14 correlated molecules were present in a given covariance network, only the top 14 correlated molecules (nearest neighbors) are displayed.

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

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

Supplementary Figure S3.13: Examples of hypotheses that can be generated from a subset of the molecule covariance network analyses in this study. Nearest neighbor molecule covariance networks from uncharacterized proteins containing more than four connected nodes were tested for GO term enrichment using a Fisher's exact test with Benjamini–Hochberg FDR adjustment to account for multiple hypothesis testing. Networks containing four or fewer connected nodes were analyzed manually for functionally related molecules. Based on these MCNA results, biological hypotheses about the functions of the uncharacterized proteins shown were developed.



Supplementary Figure S3.14: Hypothesized pathways for Aro9p, Aro10p, and Aim18p. **(a)** Putative biochemical functions of Aro9p and Aro10p in catabolism of tyrosine and phenylalanine. **(b)** Predicted functions for Aro9p and Aro10p in the Tyr-to-4-HB-to-CoQ pathway. **(c)** Protein sequence alignments of Aim18p (*S. cerevisiae*) and chalcone isomerases (CHI) from *Medicago* and *Arabidopsis* highlighting conservation of putative catalytic residues (starred residues). **(d)** Example of a CHI catalyzed reaction (upper scheme) and the hypothesized pathway of Aim18p action (lower scheme).

drial OxPhos proteins, which helped place Aro9p and Aro10p into the Tyr-to-4-HB pathway **(Fig. 3.4)**. Similarly, links between Aim18p and known CoQ biosynthesis enzymes also connected Aim18p to CoQ biosynthesis. Furthermore, Y3K gene-gene correlation analyses and manual pathway analyses linked CoQ biosynthesis to other proteins whose molecular functions in this pathway are not yet fully defined (e.g. Atp2p, Fzo1p, and Oct1p). Disruption of the mammalian Fzo1p homolog, MFN2—a protein essential for mitochondrial fusion that harbors causative mutations in Charcot-Marie-Tooth disease[33]—was recently shown to cause CoQ deficiency through an unclear molecular mechanism[34]. Our results suggest that this unexpected relationship between MFN2 and CoQ biosynthesis is evolutionarily conserved, and establish yeast as a model system for further probing its mechanism.

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

We anticipate that the multi-omic Y3K dataset will provide a resource for broader systems biology inquiries. For example, our definition of the yeast respiration deficiency response (RDR) **(Fig. 3.3)** may assist studies of how cells broadly respond to defects in OxPhos, which are observed in diverse diseases including many cancers. Our RDR work also suggests that a multi-omic fingerprint of numerous molecules could provide a highly

specific biomarker panel.

Methods

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

Single lots of yeast extract ('Y') (Research Products International, RPI), peptone ('P') (RPI), agar (Fisher), dextrose ('D') (RPI), glycerol ('G') (RPI), and G418 (RPI) were used for all medias. YP and YPG solutions were sterilized by automated autoclave. G418 and dextrose were sterilized by filtration (0.22 μm pore size, VWR) and added separately to sterile YP or YPG. YPD+G418 plates contained yeast extract (10 g/L), peptone (20 g/L), agar (15 g/L), dextrose (20 g/L), and G418 (200 mg/L). YPD media (fermentation cultures) contained yeast extract (10 g/L), peptone (20 g/L), and dextrose (20 g/L). YPGD media (respiration cultures) contained yeast extract (10 g/L), peptone (20 g/L), glycerol (30 g/L) and dextrose (1 g/L).

Yeast from a –80 °C glycerol stock were streaked onto YPD+G418 plates and incubated (30 °C, ~60 h). Starter cultures (3 mL YPD) were inoculated with an individual colony of yeast and incubated (30 °C, 230 rpm, 10–15 h). A WT culture was included with each set

of Δ*gene* strain cultures (usually 19 Δ*gene* cultures and 1 WT culture). Cell density was determined by optical density at 600 nm (OD_{600}) as described[35]. YPD or YPGD media (100 mL media at ambient temperature in a sterile 250 mL Erlenmeyer flask) was inoculated with 2.5×10^6 yeast cells and incubated (30 °C, 230 rpm). Samples of the YPD cultures were harvested 12 h after inoculation, a time point that corresponds to early fermentation (logarithmic) growth. Samples of YPGD cultures were harvested 25 h after inoculation, a time point that corresponds to early respiration growth.

Liquid chromatography tandem mass spectrometry (LC-MS/MS) proteomics. 1×10^8 yeast cells were harvested by centrifugation (3,000 *g*, 3 min, 4 °C), the supernatant was removed, and the cell pellet was flash frozen in $N_{2(l)}$ and stored at –80 °C. Yeast pellets were resuspended in 8 M urea, 100 mM tris (pH = 8.0). Yeast cells were lysed by the addition of methanol to 90%, followed by vortexing (~30 s). Proteins were precipitated by centrifugation (12,000 *g*, 5 min). The supernatant was discarded, and the resultant protein pellet was resuspended in 8 M urea, 10 mM tris(2-carboxyethyl)phosphine (TCEP), 40 mM chloroacetamide (CAA) and 100 mM tris (pH = 8.0). Sample was diluted to 1.5 M urea with 50 mM tris and digested with trypsin (Promega) (overnight, ~22 °C) (1:50, enzyme:protein). Samples were desalted using Strata X columns (Phenomenex Strata-X Polymeric Reversed Phase, 10 mg/mL). Strata X columns were equilibrated with one column volume of 100% acetonitrile (ACN), followed by 0.2% formic acid. Acidified samples were loaded on column, followed by washing with three column volumes of 0.2% formic acid or 0.1% TFA. Peptides

were eluted off the column by the addition of 500 μL 40% ACN with either 0.2% formic acid or 0.1% TFA and 500 μL 80% ACN with either 0.2% formic acid or 0.1% TFA. Peptide concentration was measured using a quantitative colorimetric peptide assay (Thermo). LC-MS/MS analyses were performed using previously described methodologies1, 2.

LC/MS data analysis. Raw data files were acquired in batches of 60 (3 biological replicates of 19 Δ*gene* strains and 1 WT strain) with time between LC-MS analyses minimized to reduce run-to-run variation. Batches of raw data files were subsequently processed using MaxQuant[36] (Version 1.5.0.25). Searches were performed against a target-decoy[37] database of reviewed yeast proteins plus isoforms (UniProt, downloaded January 20, 2013) using the Andromeda[37] search algorithm. Searches were performed using a precursor search tolerance of 4.5 ppm and a product mass tolerance of 0.35 Da. Specified search parameters included fixed modification for carbamidomethylation of cysteine residues and a variable modification for the oxidation of methionine and protein N-terminal acetylation, and a maximum of 2 missed tryptic cleavages. A 1% peptide spectrum match (PSM) false discovery rate (FDR) and a 1% protein FDR was applied according to the target-decoy method. Proteins were identified using at least one peptide (razor + unique). Proteins were quantified using MaxLFQ with an LFQ minimum ratio count of 2. LFQ intensities were calculated using the match between runs feature, and MS/MS spectra were not required for LFQ comparisons. Missing values were imputed where appropriate for proteins quantified in $\geqslant$ 50% of MS data files in a batch. Proteins not meeting this requirement were omitted

from subsequent analyses. Imputation was performed on a replicate-by-replicate basis. For each replicate MS analysis a normal distribution with mean and standard deviation equivalent to that of the lowest 1% of measured LFQ intensities was generated. Missing values were filled in with values drawn from this distribution at random. Approximately 4.05% and 4.53% of quantitative measurements were imputed in the respiration and fermentation proteomic datasets, respectively. Replicate protein LFQ values from corresponding Δ*gene* or WT strains were pooled, $\log_2$ transformed, and averaged (mean $\log_2$[strain], n = 3). Average Δ*gene* LFQ intensities were normalized against their appropriate WT control (mean $\log_2$[Δ*gene*/WT], n = 3) and a 2-tailed t-test (homostatic) was performed to obtain *P* values.

To control for batch-specific effects, proteins having unexpected and characteristic misregulation across a majority of Δ*gene* strains processed together were identified and omitted from the 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.

Gas chromatography-mass spectrometry (GC-MS) metabolomics. 1×10^8 yeast cells yeast cells were isolated by rapid vacuum filtration onto a nylon filter membrane (0.45 μm pore size, Millipore) using a Glass Microanalysis Filter Holder (Millipore), briefly washed with phosphate buffered saline (1 mL), and immediately submerged into ACN/MeOH/H2O (2:2:1, v/v/v, 1.5 mL, pre-cooled to –20 °C) in a plastic tube. The time from sampling yeast from the culture to submersion in cold extraction solvent was less than 30 s. Tubes with the extraction solvent, nylon filter, and yeast were stored at –80 °C prior to analysis.

Tubes with yeast extract (also still containing insoluble yeast material and the nylon filter) were thawed at room temperature for 45 min., vortexed (~15 s), and centrifuged at room temperature (6400 rpm, 30 s) to pellet insoluble yeast material. Yeast extract (25 μL aliquot) and internal standards (25 μL aqueous mixture of isotopically labelled alanine-2,3,3,3-d_4, adipic acid-d_{10}, and xylose-$^{13}C_5$ acid, 5 ppm in each) were aliquoted into a 2 mL plastic tube and dried by vacuum centrifuge (~1 hr). The dried metabolites were resuspended in pyridine (25 μL) and vortexed. 25 μL of N-methyl-N-trimethylsilyl]trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) was added, and the sample was vortexed and incubated (60 °C, 30 min). Samples were then transferred to a glass autosampler vials and analyzed using a GC/MS instrument comprising a Trace 1310 GC coupled to a Q Exactive Orbitrap mass spectrometer. For the yeast metabolite extracts a linear temperature gradient ranging from 50 °C to 320 °C was employed spanning a total runtime of 30

minutes. Analytes were injected onto a 30 m TraceGOLD TG-5SILMS column (Thermo) using a 1:10 split at a temperature of 275 °C and ionized using electron ionization (EI). The mass spectrometer was operated in full scan mode using a resolution of 30,000 ($m/\Delta m$) relative to 200 m/z.

GC/MS data analysis. The resulting GC-MS data were processed using an in-house developed software suite (https://github.com/coongroup/Y3K-Software). Briefly, all m/z peaks are aggregated into distinct chromatographic profiles (i.e., feature) using a 10 ppm mass tolerance. These chromatographic profiles are then grouped according to common elution apex (i.e., feature group). The collection of features (i.e., m/z peaks) sharing a common elution apex, therefore, represent an individual EI-MS spectrum of a single eluting compound. The EI-MS spectra were then compared against a matrix run and a background subtraction was performed. Remaining EI-MS spectra are then searched against the NIST 12 MS/EI library and subsequently subjected to a high resolution filtering (HRF) technique as described elsewhere. EI-MS spectra that were not identified were assigned a numeric identifier. Feature intensity, which was normalized using total metabolite signal, was used to estimate metabolite abundance. Following initial processing, raw data files were re-analyzed to extract metabolite signals which were not successfully deconvolved and registered as missing values in the dataset. This process provided measurements for ~1.87%, and 2.25% of metabolites quantified in the respiration and fermentation datasets, respectively. Remaining missing values were imputed using the same imputation strategy

as described in the proteomic data processing section. Quantitative values imputed using this process account for ~0.17% and 0.13% of metabolites in the respiration and fermentation datasets, respectively.

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

Δ*Gene*-specific phenotype detection. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2$[Δ*gene*/WT]) and negative $\log_2$ fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(p-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}$(*p*-value) were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all

neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at a Euclidean distance of 0.70.

Regression analysis of Δ*gene*–Δ*gene* perturbation profiles. For all pairwise combinations of Δ*gene* strains from the same growth condition linear regression analysis was conducted on protein, lipid, and metabolite perturbation profiles, respectively. Fold change measurements (mean $\log_2$[Δ*gene*/WT], n = 3) from molecules where FC > 0.7 and $P < 0.05$ were used and a minimum of 20 proteins, 10 metabolites, and 5 lipids, respectively, were required. These measurements were fit to a line and the associated Pearson correlation coefficient was reported. Coefficients carrying negative signs were set to 0. For pairs of Δ*gene* strains lacking a sufficient number of molecules that met the aforementioned criteria, the Pearson coefficient was reported as 0. Hierarchical clustering of Δ*gene*–Δ*gene* correlations was performed as described below.

Respiration deficiency response (RDR) abundance adjustment. All Δ*gene* strains grown under respiration conditions were classified as respiration deficient (RD) (51) or respiration

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

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

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

For clustering of Δ*gene* perturbation profiles, clustering was performed separately for

fermentation and respiration 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[38] 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 cloning of genes and mutants for follow-up studies. *S. cerevisiae* (BY4742) gene deletion strains for *hfd1*, *atp2*, *ypr010c-a*, and *yjr120w* were generated using a PCR deletion strategy in which the open reading frames were replaced by a KanMX cassette from the pFA6a-kanMX6 plasmid. Briefly, KanMX was amplified with primers containing sequence homologous to sequence just upstream of the ATG and just downstream from the terminal codon for each ORF. Amplicons were transformed into BY4742, and yeast were plated onto YEPD plates containing 100 μg/mL G418. Knockouts were confirmed by PCR and sequencing.

To generate plasmid yeast gene constructs, *S. cerevisiae hfd1*, *atp2*, and *yjr120w* were

amplified by Accuprime Pfu polymerase (Invitrogen, USA) with primers generating a SpeI site (forward) and SalI (reverse) (BamHI forward and EcoRI reverse for *yjr120w*). The *hfd1*, *atp2*, and *yjr120w* amplicons and the yeast expression vectors p426GPD and p423GPD were digested with SpeI and SalI or BamHI and EcoRI. *Hfd1* and *yjr120w* were ligated to p426GPD, *atp2* was ligated to p423GPD, and each ligation was transformed into DH5α *E. coli*. Plasmid minipreps were performed and recombinants were confirmed by sequencing. *Hfd1* mutants were generated via standard site-directed mutagenesis, and mutations were confirmed by sequencing.

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

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

media lacking glucose. 10-fold serial dilutions of yeast cells were dropped onto Ura^-, His^- agar media plates containing either glucose (2%, w/v) or glycerol (3%, w/v) and incubated (30 °C, 4 d).

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

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

Hfd1p and ALDH3A1 biochemical studies—yeast growth assays. 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 5×10^6 cells/mL. To interrogate the rescue efficacy of various compounds, 100 nM (final concentrations) of pABA, tyrosine, 4-HPP, 4-HPAA, 4-HPA, 4-HMA, 4-HBz, 4-HB, 4HPL, or *p*-coumarate were added to the base medium. The cultures were incubated in a sterile 96 well plate with an optical, breathable coverseal (shaking at 1140 rpm). Optical density readings (OD_{600}) were obtained every 10 min. Respiratory growth rates were determined by fitting a linear equation to the respiratory growth phase and determining the slope of the line. Relative respiratory growth rates were determined by comparing cultures with additives to those without additive.

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

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

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

Hfd1p and ALDH3A1 biochemical studies—Hfd1p, ALDH3A1, and ALDH3A2 expression and purification. PIPE cloning was used to generate pVP68K vectors encoding ALDH3A1, Hfd1p$^{C\Delta25}$, or ALDH3A2$^{C\Delta25}$ (Hfd1p or ALDH3A2 lacking their C-terminal 25 amino acids, which comprise putative transmembrane domains) fused to an 8His-cytoplasmically-targeted maltose-binding protein with a linker including a tobacco etch virus protease recognition site (8His-MBP-[TEV]-ALDH3A1, 8His-MBP-[TEV]-Hfd1p$^{C\Delta25}$, or 8His-MBP-[TEV]-ALDH3A2$^{C\Delta25}$). These constructs were expressed in *E. coli* (BL21[DE3]-RIPL strain) by autoinduction. Cells were isolated and resuspended in lysis buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 1 mg/mL lysozyme (Sigma), pH 7.5). Cells were lysed by sonication (4 °C, 2 × 20 s), and the lysate was clarified by centrifugation (15,000 g, 30 min, 4 °C). The clarified lysate was mixed with cobalt IMAC resin (Talon resin) and incubated (4 °C, 1 h). The resin was pelleted by centrifugation (700 g, 2 min, 4 °C) and washed three times (10 resin bed volumes each) with wash buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 10 mM imidazole, pH 7.5). His-tagged protein was eluted with elution buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, 100 mM imidazole, pH 7.5). The eluted protein was concentrated with a 50-kDa MW-cutoff spin filter (Merck Millipore Ltd.) and exchanged into storage buffer (50 mM HEPES, 300 mM NaCl, 10% glycerol, 5 mM BME, 0.25 mM PMSF, pH 7.5). Protein concentrations were determined by absorbance at 280 nm. The MBP-fusion proteins were aliquoted, frozen in $N_{2(l)}$, and stored at –80 °C.

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

Molecule Covariance Network Analysis For all pairwise combinations of molecules quantified within a particular growth condition, regression analysis was conducted using fold change measurements from all Δ*gene* strains having a measurement for both molecules in the pair. Spearman's regression analysis was performed to obtain correlation coefficients (ρ). From these test statistics P values were calculated using a two-sided Student's t-test. All P values were corrected for multiple hypothesis testing (Bonferroni) and correlations where $|\rho| \geqslant 0.58$ and $P < 0.001$ were reported. For RDR-adjusted regression analysis, the RDR adjustment procedure was carried out as described in the 'Respiration deficiency response (RDR) abundance adjustment' section (above). All pairs of covariant molecules are visualized as networks generated using the Gephi open graph visualization platform

(version 0.9.0). Complete respiration, fermentation and RDR-adjusted respiration network layouts were generated using the Fruchterman–Reingold graph-drawing algorithm with area set to 10,000 and gravity set to 30. Gene Ontology terms were obtained from the Saccharomyces Genome Database (SGD). To calculate network selectivity the following equation was used:

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

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

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

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

mass spectrometry as described in the main Methods section. These cultures and analyses were conducted in biological duplicate.

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

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

For PCA, average $\log_2$(Δ*gene*/WT) values for each protein, metabolite, and lipid mea-

sured in the respiration condition were analyzed using Perseus PCA software. PCA projections were exported from Perseus.

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

For GO term analyses, proteins were separated as increasing in RD strains (positive $\log_2$[RD/RC]) or decreasing in RD strains (negative $\log_2$[RD/RC]). Proteins with Bonferroni-corrected $P < 1\times10^{-20}$ were collected from each group and subjected to GO term enrichment analysis (http://geneontology.org/page/go-enrichment-analysis). Select GO terms were highlighted because they were significantly enriched (Bonferroni corrected $P < 0.05$) in proteins that were reduced (–) or increased (+) in RD strains. Boxplots of select molecules were generated using matplotlib in python to compare particular molecules across all RD and RC strains.

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

is predictive of RD whereas a – sign indicates that a reduction in that molecule is predictive of RD.

Supplementary Notes

Development of a stable and reproducible respiration culture condition. To profile diverse yeast strains during respiratory growth, when mitochondrial OxPhos is highly active, we first needed to develop a distinct respiration condition suitable for large-scale investigation. Early log phase fermentation cultures repress mitochondrial respiration, cultures containing solely non-fermentable sugars preclude growth of respiration deficient yeast, and high glucose cultures grown past the diauxic shift are too biologically dynamic to allow reproducible sampling across a large scale study[42,43]. To overcome these problems, we developed a culture system that includes low glucose (1 g/L) and high glycerol (30 g/L), enabling a short fermentation phase followed by a longer respiration phase. This respiration condition affords steady growth and a stable biological state—as reflected by a proteome that is constant over multiple hours **(Supplementary Fig. S3.1c–e)**—and, thus, an essential window for reproducible sample harvesting.

Δ*Gene*-specific phenotype detection. To identify Δ*gene*-specific phenotypes, we broadly surveyed our data for characteristic outlier abundance measurements. For each profiled molecule (in both respiration and fermentation growth conditions) we separated potential Δ*gene*-specific measurements into two groups: positive $\log_2$ fold change ($\log_2$[Δ*gene*/WT])

and negative $\log_2$fold change. These two sets were then plotted individually with $\log_2$ fold change and $-\log_{10}$(p-value [two-sided Student's t-test]) along the x- and y- axes, respectively. Data were normalized such that the largest $\log_2$ fold change and largest $-\log_{10}$(*p*-value) were set equal to 1. Considering the three largest fold changes where $P < 0.05$, we calculated the Euclidean distance to all neighboring data points and stored the smallest result. A requirement was imposed that all considered 'neighbors' have a smaller fold change than the data point being considered. It is anticipated that data points corresponding to Δ*gene*-specific phenotypes will be outliers in the described plots and have large associated nearest-neighbor Euclidean distances. The described routine yielded three separate distances, the largest of which was stored for further analysis. The results of this analysis and representative examples are highlighted **(Fig. 3.2, Supplementary Figs. S3.5 and S3.6)**. We observed maximal Euclidean distances across a range of 0.006 to 1.25. We set a cutoff for classification as a 'Δ*gene*-specific phenotype' at 0.70 and report 714 molecules (4.6% of considered cases across both culture conditions) which exceed this threshold. This procedure provided a useful 'first pass' analysis and afforded a truncated set of leads, which were used to develop biological hypotheses.

Lack of effect of Dpl1p disruption on the Tyr-to-4-HB-to-CoQ pathway. To test the idea that the CoQ biosynthesis and sphingolipid catabolism pathways are independent, we examined Δ*dpl1* yeast, which lack a known dihydrosphingosine phosphate lyase. Δ*dpl1* yeast show neither a pABA$^-$ respiratory growth phenotype nor CoQ deficiency **(Supplementary**

Fig. S3.7j,k). These results demonstrate that disruption of the Tyr-to-4-HB pathway in Δ*hfd1* yeast is not downstream of a defect in sphingolipid metabolism. Furthermore, proteome analyses showed that Δ*hfd1* cultured without 4-HB and pABA are similar to Δ*coq8* yeast—but not Δ*dpl1* yeast—and adding 4-HB to Δ*hfd1* cultures returns their proteomes to WT-like profiles **(Supplementary Fig. S3.7l,m)**.

Quantitative definition of the respiration deficiency response (RDR). To quantitatively define the RDR, we categorized strains as respiration deficient (RD) or competent (RC) and examined differences between these two groups. Principal component analysis of the Y3K respiration dataset revealed marked separation of RD and RC strains **(Fig. 3.3c and Supplementary Fig. S3.8a)**. The underlying phenotype changes that distinguish RD and RC strains include proteins, lipids, and metabolites **(Fig. 3.3d)**. RDR perturbations include significant decreases in ATP synthase, TCA cycle, and MICOS proteins **(Fig. 3.3e,f and Supplementary Fig. S3.8b)**, likely to decrease allocation of useless proteome mass to dysfunctional mitochondria[44]. Importantly, the RDR also includes a positive response, and numerous proteins—including protein folding, NADH metabolism, and proteasome assembly proteins—are significantly upregulated in RD strains **(Fig. 3.3e,f)**. Numerous individual molecules—including lactate, alanine, 2-hydroxyglutarate, tyrosol, 4-HB, Gpx2p, and Ahp1p, among many others—are significantly perturbed in RD strains and strongly predictive of respiration deficiency **(Supplementary Fig. S3.8c,d)**. Our quantitative assessment of the RDR highlights biochemical features of the cellular response to defects in

mitochondrial respiration, and suggests that a multi-omic assessment of proteins, lipids, and metabolites could afford a highly specific biomarker panel for diseases affected by OxPhos deficiency.

RDR normalization procedure. Δ*gene* strains were classified as RD (51) or respiration competent (RC) (123) based on observation of a common perturbation profile signature in the respiration culture condition. For each molecule we calculated an RDR score. This metric represents the proportion of RD Δ*gene* strains over which the molecule was consistently perturbed, relative to all RD Δ*gene* strains where the molecule was quantified. Across all RD Δ*gene strains*, 776 molecules were identified as having an RDR score > 0.95 (consistently perturbed across more than 95% of RD Δ*gene* strains where quantified) and classified as RDR-associated. The individual measurements of these RDR-associated molecules were then mean normalized ('RDR-adjusted') using abundance values from RD Δ*gene* strains. This normalization procedure revealed characteristic deviations from the general RDR **(Supplementary Fig. S3.9)**. Importantly, this procedure enables visualization of Δ*gene*-specific changes. For example, prior to RDR normalization, the expected decrease in Coq8p in Δ*coq8* yeast is obscured by RDR-associated proteins with large abundance changes **(Supplementary Fig. S3.9d)**. RDR normalization not only uncovers the decrease in Coq8p, but a significant decrease in Coq5p, a functionally-related CoQ biosynthesis protein, also becomes readily apparent **(Supplementary Fig. S3.9d)**.

Molecular defects of Δ*yjr120w* yeast. 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 **(Supplementary Fig. S3.10a)**. Compared to other strains, the large decrease in Atp2p is unique to Δ*yjr120w* and Δ*atp2* **(Supplementary Fig. S3.10b)**. A relationship between *yjr120w* and *atp2* is also suggested by their genetic proximity **(Supplementary Fig. S3.10c)**. Plasmid overexpression of *atp2* rescues the Δ*yjr120w* respiratory growth defect **(Supplementary Fig. S3.10d)**, indicating a functional relationship between *atp2* and *yjr120w in vivo*. A decrease in *atp2* mRNA in the Δ*yjr120w* strain is a component of the underlying mechanism **(Supplementary Fig. S3.10e)**. Interestingly, CoQ deficiency was also observed in Δ*atp2* yeast **(Fig. 3.3h)**.

Predicted enzymatic functions of Aim18p, Aro9p, and Aro10p. Since 1907, yeast have been known to catabolize amino acids into fusel (German for 'bad liquor') alcohols through the Ehrlich pathway[45,46], but the physiological roles for the enzymes involved—such as Aro9p and Aro10p—are not fully understood. Aro9p and Aro10p were previously thought to provide a simple catabolic route for extracting nitrogen from aromatic amino acids[47] **(Supplementary Fig. S3.14a)**, but our MCNA unexpectedly indicated strong correlations between Aro9p, Aro10p, and proteins involved in mitochondrial respiration **(Fig. 3.4d,e)**, suggesting a more complicated biological function that supports OxPhos. We hypothesized that this function might be in the Tyr-to-4-HB-to-CoQ pathway **(Supplementary Fig. S3.14b)**, given the putative enzymatic activities of Aro9p and Aro10p in tyrosine and

phenylalanine metabolism. Consistently, when cultured in pABA$^-$ media, Δ*aro9* and Δ*aro10* yeast are deficient in CoQ and PPHB **(Fig. 3.4f)**.

Aim18p is a protein of undefined molecular function that has been detected in mitochondria[48] and potentially linked to mitochondrial inheritance (Altered Inheritance of Mitochondria, 'AIM') by large-scale studies in yeast[49]. Protein sequence alignments show that Aim18p contains a chalcone-flavone isomerase (CHI)-like domain **(Supplementary Fig. S3.14c)**, whose homologs in plants typically function on aromatic small molecules (chalcones) **(Supplementary Fig. S3.14d)**[50–52]. Given the potential for this protein domain to catalyze modifications of aromatic small molecules, we hypothesized that Aim18p might function in the Tyr-to-4-HB pathway to produce the CoQ headgroup (Supplementary Fig. 14d). Consistently, when cultured in pABA$^-$ media, we observed deficiency of PPHB in Δ*aim18* yeast **(Fig. 3.4f)**.

References

[1] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, pp. 339–347, 2014.

[2] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, pp. 701–714, 2015.

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

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

[5] J. M. Buescher, W. Liebermeister, M. Jules, M. Uhr, J. Muntel, E. Botella, B. Hessling, R. J. Kleijn, L. Le Chat, F. Lecointe, U. Mäder, P. Nicolas, S. Piersma, F. Rügheimer, D. Becher, P. Bessieres, E. Bidnenko, E. L. Denham, E. Dervyn, K. M. Devine, G. Doherty, S. Drulhe, L. Felicori, M. J. Fogg, A. Goelzer, A. Hansen, C. R. Harwood, M. Hecker, S. Hubner, C. Hultschig, H. Jarmer, E. Klipp, A. Leduc, P. Lewis, F. Molina, P. Noirot, S. Peres, N. Pigeonneau, S. Pohl, S. Rasmussen, B. Rinn, M. Schaffer, J. Schnidder, B. Schwikowski, J. M. Van Dijl, P. Veiga, S. Walsh, A. J. Wilkinson, J. Stelling, S. Aymerich, and U. Sauer, "Global network reorganization during dynamic adaptations of Bacillus subtilis metabolism.," *Science (New York, NY)*, vol. 335, pp. 1099–1103,

2012.

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

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

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

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

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

[11] D. J. Pagliarini, S. E. Calvo, B. Chang, S. A. Sheth, S. B. Vafai, S. E. Ong, G. A. Walford, C. Sugiana, A. Boneh, W. K. Chen, D. E. Hill, M. Vidal, J. G. Evans, D. R. Thorburn, S. A. Carr, and V. K. Mootha, "A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology," *Cell*, vol. 134, pp. 112–123, 2008.

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

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

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

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

[16] B. J. Floyd, E. M. Wilkerson, M. T. Veling, C. E. Minogue, C. Xia, E. T. Beebe, R. L. Wrobel, H. Cho, L. S. Kremer, C. L. Alston, K. A. Gromek, B. K. Dolan, A. Ulbrich, J. A. Stefely, S. L. Bohl, K. M. Werner, A. Jochem, M. S. Westphall, J. W. Rensvold, R. W. Taylor, H. Prokisch, J.-J. P. Kim, J. J. Coon, and D. J. Pagliarini, "Mitochondrial Protein Interaction Mapping Identifies Regulators of Respiratory Chain Function.," *Molecular cell*, vol. 63, pp. 621–32, 2016.

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

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

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

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

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

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

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

[24] Z. J. Liu, Y. J. Sun, J. Rose, Y. J. Chung, C. D. Hsiao, W. R. Chang, I. Kuo, J. Perozich, R. Lindahl, J. Hempel, and B. C. Wang, "The first structure of an aldehyde dehy-

drogenase reveals novel interactions between NAD and the Rossmann fold.," *Nature structural biology*, vol. 4, pp. 317–326, 1997.

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

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

[27] T. Kitamura, T. Naganuma, K. Abe, K. Nakahara, Y. Ohno, and A. Kihara, "Substrate specificity, plasma membrane localization, and lipid modification of the aldehyde dehydrogenase ALDH3B1.," *Biochimica et biophysica acta*, vol. 1831, pp. 1395–401, 2013.

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

[29] P. Kemmeren, K. Sameith, L. A. L. Van De Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. Van Wageningen, C. W. Ko, S. Van Heesch, M. M.

Kashani, G. Ampatziadis-Michailidis, M. O. Brok, N. A. C. H. Brabers, A. J. Miles, D. Bouwmeester, S. R. Van Hooff, H. Van Bakel, E. Sluiters, L. V. Bakker, B. Snel, P. Lijnzaad, D. Van Leenen, M. J. A. Groot Koerkamp, and F. C. P. Holstege, "Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors," *Cell*, vol. 157, pp. 740–752, 2014.

[30] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. a. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, "Functional profiling of the Saccharomyces cerevisiae genome.," *Nature*, vol. 418, pp. 387–391, 2002.

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

[32] U. C. Tran and C. F. Clarke, "Endogenous synthesis of coenzyme Q in eukaryotes,"

Mitochondrion, vol. 7, 2007.

[33] S. Züchner, I. V. Mersiyanova, M. Muglia, N. Bissar-Tadmouri, J. Rochelle, E. L. Dadali, M. Zappia, E. Nelis, A. Patitucci, J. Senderek, Y. Parman, O. Evgrafov, P. D. Jonghe, Y. Takahashi, S. Tsuji, M. a. Pericak-Vance, A. Quattrone, E. Battaloglu, A. V. Polyakov, V. Timmerman, J. M. Schröder, and J. M. Vance, "Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot-Marie-Tooth neuropathy type 2A.," *Nature genetics*, vol. 36, pp. 449–451, 2004.

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

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

[36] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.," *Nature biotechnology*, vol. 26, pp. 1367–72, 2008.

[37] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann, "An-

dromeda: A peptide search engine integrated into the MaxQuant environment," *Journal of Proteome Research*, vol. 10, pp. 1794–1805, 2011.

[38] T. Author, Y. Benjamini, Y. Hochberg, and Y. Benjaminit, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

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

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

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

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

[43] A. Casanovas, R. R. Sprenger, K. Tarasov, D. E. Ruckerbauer, H. K. Hannibal-Bach, J. Zanghellini, O. N. Jensen, and C. S. Ejsing, "Quantitative analysis of proteome

and lipidome dynamics reveals functional regulation of global lipid metabolism," *Chemistry and Biology*, vol. 22, pp. 412–425, 2015.

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

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

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

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

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

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

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

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

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

Chapter 4

ISLET PROTEOMICS REVEALS GENETIC VARIATION IN DOPAMINE PRODUCTION RESULTING IN ALTERED INSULIN SECRETION

ECF designed experiments, ran samples, analyzed data, and co-authored this work.

This chapter is under review:

Kelly A. Mitok*, **Elyse C. Freiberger***, Kathryn L. Schueler, Mary E. Rabaglia, Donald S. Stapleton, Nicholas W. Kwiecien, Paige A. Malec, Alexander S. Hebert, Aimee T. Broman, Robert T. Kennedy, Mark P. Keller, Joshua J. Coon, Alan D. Attie *Islet proteomics reveals genetic variation in dopamine production resulting in altered insulin secretion.* Journal of Biological Chemistry. **2018**, *Under Review*.

* Authors contributed equally

Abstract

The mouse is a critical model in diabetes research, but most research in mice has been limited to a small number of mouse strains and limited genetic variation. We used the eight founder strains of the Collaborative Cross (CC) (C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), PWK/PhJ (PWK), WSB/EiJ (WSB), CAST/EiJ (CAST)) to investigate the genetic dependence of diabetes-related metabolic phenotypes and insulin secretion. We found that strain background is associated with an extraordinary range in body weight, plasma glucose, insulin, and triglycerides, and insulin secretion. Our whole-islet proteomic analysis of the eight mouse strains demonstrates that genetic background exerts a strong influence on the islet proteome. We computed protein modules consisting of highly correlated proteins that enrich for biological pathways and provide a searchable database of the islet protein expression profiles. To validate the data resource, we identified tyrosine hydroxylase (Th), a key enzyme in catecholamine synthesis, as a protein that is highly expressed in β-cells of PWK and CAST islets. We show that CAST islets synthesize elevated levels of dopamine, which suppresses insulin secretion. Prior studies, using only the B6 strain, concluded that adult mouse islets do not synthesize L-3,4-dihydroxyphenylalanine (L-DOPA), the product of Th and precursor of dopamine. Thus, the choice of the CAST strain, guided by our islet proteomic survey, was crucial for these discoveries. In summary, we provide a valuable data resource to the research community, and show that proteomic analysis identified a strain-specific pathway by which

dopamine synthesized in β-cells inhibits insulin secretion.

Introduction

Approximately 50% of the variation in the risk of type 2 diabetes (T2D) in humans is due to genetic factors[1]. Most of the candidate genes identified in genome-wide association studies for T2D affect pancreatic islet function either directly or indirectly from other tissues[2,3]. Humans have a wide range in insulin secretory capacity and insulin secretion shows high heritability[4,5]. However, model organisms are necessary for detailed mechanistic studies to understand how specific genes affect insulin secretion. The mouse has been indispensable in diabetes research[6]. The phenotype spectrum present in the wide array of mouse strains offers the opportunity to discover gene action in relation to diabetes traits. However, most research in mice has been limited to a small number of mouse strains covering limited genetic variation. The majority of mouse gene knockout studies have been performed in C57BL/6J (B6), with most of the remaining studies done in FVB and 129/Sv. Often, a gene deletion results in "no phenotype", but the absence of a discernible phenotype could be due to the strain background suppressing the phenotype of the gene deletion. In 2002, the Collaborative Cross (CC) project was initiated to produce recombinant inbred strains from eight genetically diverse founder strains; five classical inbred mouse strains C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), and three wild-derived strains PWK/PhJ (PWK), WSB/EiJ (WSB), and CAST/EiJ (CAST)[7]. Collectively, these eight strains contain as much genetic variation as the entire human

population; ~40 million single nucleotide polymorphisms, and numerous insertions and deletions. Approximately 75% of the genetic variation is contributed by the three wild-derived strains. The great diversity across the strains offers an opportunity to evaluate the influence of genetic variation on metabolic phenotypes without the need to create transgenic mice. We assessed the variability of diabetes-related metabolic phenotypes, conducted whole-islet proteomics, and measured isolated islet insulin secretory responses from the eight CC founder strains and both sexes. Our data show a wide range of diabetes-related metabolic phenotypes among the strains and indicate that genetic background exerts a strong influence on the islet proteome, which can be causally linked to differences in insulin secretion among the strains. Further, the data show that modules of highly correlated proteins are driven by specific strains and enrich for biological pathways. We discovered that β-cells of PWK and CAST mice uniquely have elevated levels of tyrosine hydroxylase (Th), the first step in the catecholamine synthesis pathway. We show that increased Th in CAST islets leads to enhanced dopamine production, resulting in blunted insulin secretion. Our findings demonstrate the utility of exploiting the wide genetic diversity in the CC founder mouse strains available to the research community.

Results

Genetic diversity drives diabetes-related phenotypic variability We assessed the variability of diabetes-related metabolic phenotypes of the eight genetically diverse Collaborative Cross (CC) founder mouse strains; C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ

(NOD), NZO/HILtJ (NZO), PWK/PhJ (PWK), WSB/EiJ (WSB), and CAST/EiJ (CAST), which were metabolically challenged with a Western-style diet high in fat and sucrose (HF/HS diet; 44.6% kcal from fat, 40.7% kcal from sucrose) for 16 weeks. This resulted in a large range in diet-induced weight gain and insulin resistance. All mice were obtained from the Jackson Laboratory, housed in the same vivarium and maintained on the same diet throughout the study. We observed an extraordinary range in diabetes-related metabolic phenotypes among the eight mouse strains and between the sexes, reflecting their genetic diversity. Body weight (**Fig. 4.1a, b**), fasting plasma glucose (**Fig. 4.1c, d**), insulin (**Fig. 4.1e, f**), and triglyceride (**Supplementary Fig. S4.1**) all showed strain- and sex-dependent differences over the course of the 16-week HF/HS dietary challenge. Body weight was lowest in the three wild-derived strains (CAST, PWK, and WSB), and highest in NZO, with NZO females becoming severely obese, reaching 80 grams by 20 weeks of age. Food intake correlated with body weight in some, but not all strains (**Supplementary Fig. S4.2**). For example, the NZO male mice outweighed other males and consistently had the highest food intake. Female 129 mice consumed the least amount of food, but did not have the lowest body weight. Fasting glucose levels remained within a normal range in all mice (90-180 mg/dl for HF/HS fed mice), except for NZO males. Fasting insulin levels, a marker of insulin resistance, however, showed dramatic strain- and sex-dependent variation; there was ~100-fold difference in fasting plasma insulin between the most insulin-resistant (NZO), and most insulin-sensitive (CAST) strains, for the females at 20 weeks old and ~10-fold difference in the males. Male NZO were the only mice to become overtly diabetic (fasting

glucose $>$ 300 mg/dl), and did not survive the full 16-week dietary challenge. In contrast, female NZO mice were severely obese, yet maintained euglycemia by increasing insulin. We previously evaluated the dynamic changes in plasma glucose and insulin in the eight male strains on regular chow or HF/HS diet during an oral glucose tolerance test (oGTT). We found remarkable strain-dependent variation in whole-body glucose homeostasis and circulating insulin[8]. In particular, male CAST mice demonstrated a rapid and transient rise in plasma glucose and insulin during the oGTT. Further, CAST was the only strain resistant to HF/HS diet-induced changes in all metabolic phenotypes. These results suggest that male CAST mice utilize unique physiological pathways to regulate glucose-stimulated insulin secretion and whole-body glucose homeostasis.

Islet insulin and glucagon secretory response is determined by genetic background To evaluate the relationship between genetic diversity and islet function, we isolated islets from both sexes of each mouse strain that were maintained on the HF/HS diet for 16 weeks. The number of islets isolated per mouse (**Fig. 4.1g**), insulin content per islet (**Fig. 4.1h**), and glucagon content per islet (**Supplementary Fig. S4.3a**) all varied greatly. In several strains (B6, A/J, WSB, CAST, NZO and PWK), $>$ 400 islets were collected per mouse. 129 and NOD mice had fewer islets. NZO male mice yielded the fewest islets overall (~50 pooled from four mice), a consequence of their extreme diabetes. It is likely that other factors, such as effectiveness of pancreatic digestion by collagenase, affect the number of islets isolated per mouse. However, the small number of islets isolated from the severely diabetic animals

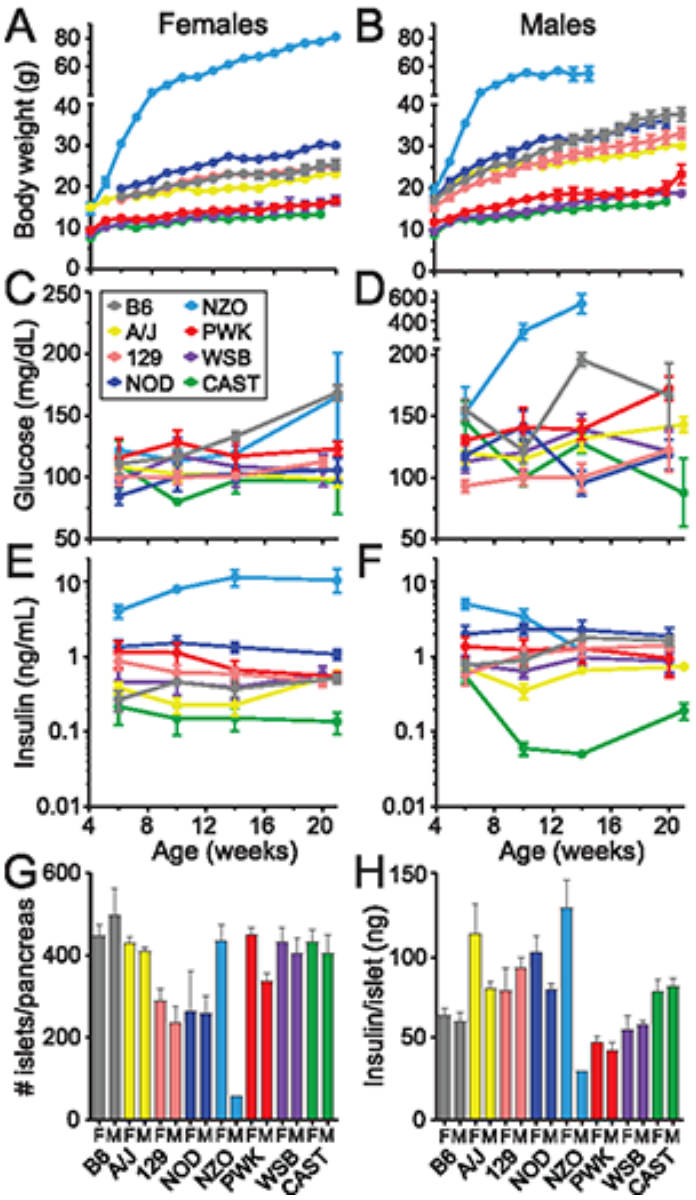
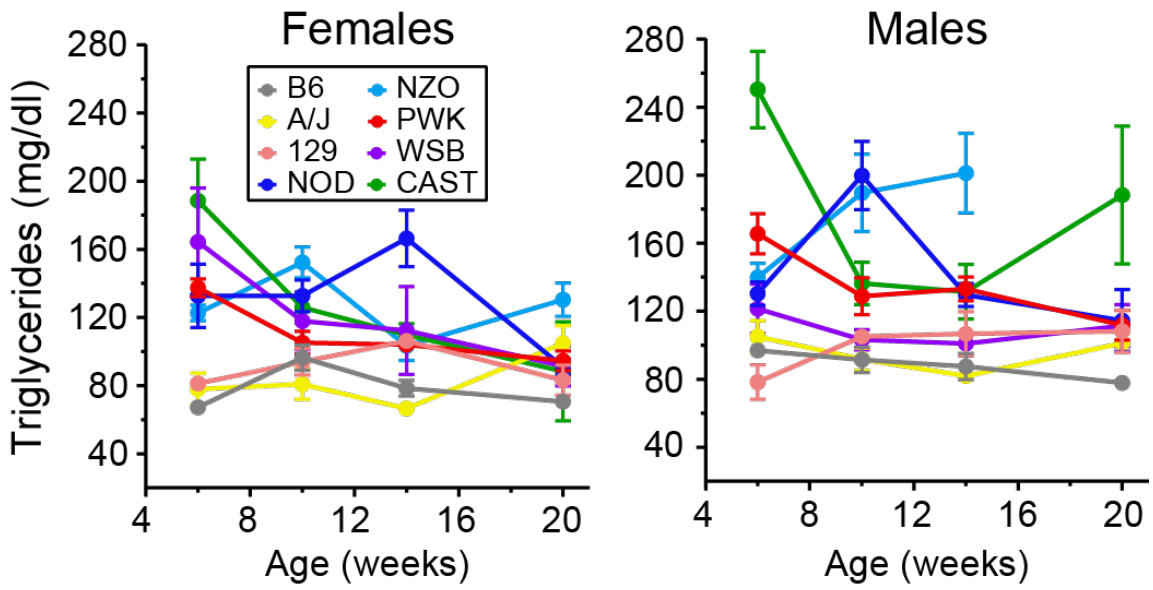
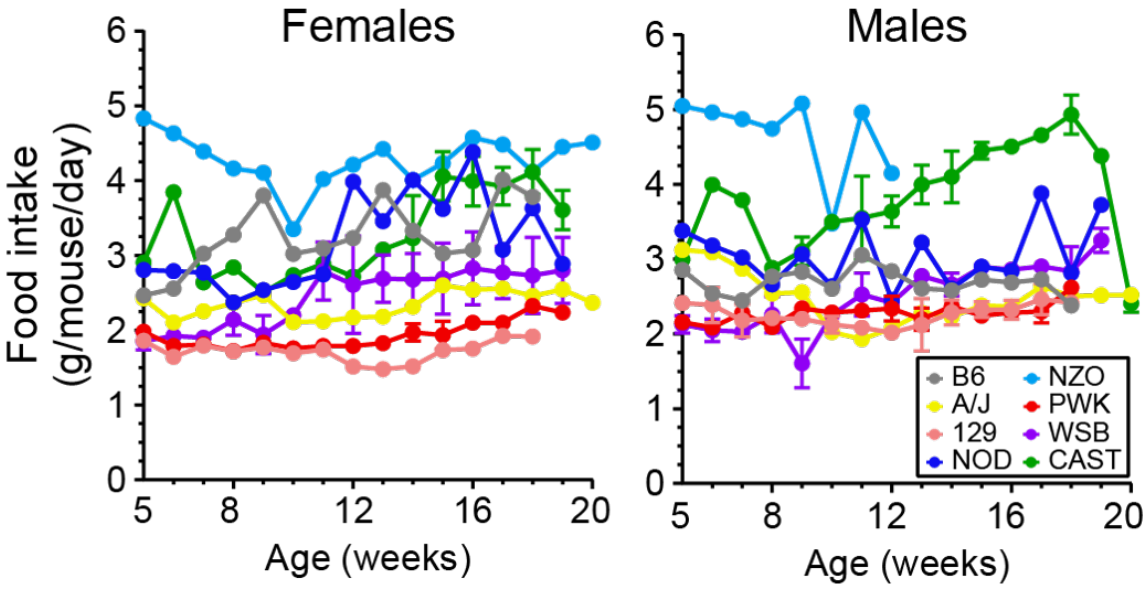


Figure 4.1: Diabetes-related metabolic phenotypes vary with genetic background. Male and female mice of the eight CC founder strains (C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), PWK/PhJ (PWK), WSB/EiJ (WSB), and CAST/EiJ (CAST)) were maintained on a HF/HS diet beginning at 4 weeks of age. Body weight (**A**, **B**), fasting plasma glucose (**C**, **D**), and insulin (**E**, **F**) were measured at multiple time points during the dietary challenge. Number of islets per pancreas (**G**), and insulin content per islet (**H**) were determined for all mice at 22 weeks of age, except for NZO male mice, which were sacrificed at 14 weeks of age due to severe hyperglycemia. Body weight, plasma glucose, and insulin levels were determined after a 3 – 4 hour fast. Data are mean $\pm$ SEM, N $\geqslant$ 3 mice/sex/strain.



Supplementary Figure S4.1: Fasting plasma triglycerides vary with genetic background. Male and female mice of the CC founder strains (B6, A/J, 129, NOD, NZO, PWK, WSB, and CAST) were maintained on a HF/HS diet beginning at 4 weeks of age. Fasting plasma triglycerides were measured at multiple time points during the dietary challenge. Levels were determined after a 3 – 4 hour fast. Data are mean ± SEM, N > 3 mice/sex/strain.



Supplementary Figure S4.2: Food intake varies with genetic background. Male and female mice of the CC founder strains (B6, A/J, 129, NOD, NZO, PWK, WSB, and CAST) were maintained on a HF/HS diet beginning at 4 weeks of age. Beginning at 4 weeks of age, food intake was calculated by weighing the food remaining after one week, subtracting it from the amount fed, then dividing by number of mice per cage and days to get g/mouse/day. Data are mean ± SEM, N > 3 mice/sex/strain.

suggests that the islet number measurement is also related to the physiological state of the mouse at the time the isolation was performed. Islets from A/J and NZO females had the highest insulin content, whereas islets from WSB, PWK and NZO males had the lowest. Islets from female PWK mice had the highest glucagon content, whereas islets from male A/J mice had the lowest. There was a sex-effect on glucagon content, with islets from male mice generally having lower glucagon content than islets from female mice. Comparison of the patterns across the strains shows that glucagon content per islet was not strongly correlated with insulin content per islet. To evaluate the relationship between genetic background and insulin secretion, we measured secretion in response to a variety of insulin secretagogues: glucose (3.3, 8.3 and 16.7 mM), the incretin hormone glucagon-like peptide-1 (GLP-1, 100 nM), the fatty acid palmitate (PA, 0.5 mM), amino acids (0.5 mM L-leucine, 2 mM L-glutamine and 1.25 mM L-alanine), and a depolarizing concentration of KCl (40mM). Glucose stimulates insulin secretion upon transport into the β-cell via Glut2, becomes phosphorylated by glucokinase, and enters the glycolytic pathway. This process induces a rise in the ATP/ADP ratio and closure of ATP-sensitive potassium channels, followed by membrane depolarization, opening of voltage-dependent calcium channels, influx of calcium ions, and fusion of insulin-containing granules with the plasma membrane, resulting in insulin secretion. Fatty acids, GLP-1 (via GLP-1 receptor), and amino acids can augment this process through "amplification pathways"[9]. KCl, a non-metabolic insulin secretagogue, stimulates insulin secretion by opposing K+ efflux from the cell, resulting in membrane depolarization and calcium entry. The insulin secretory

responses to the secretagogues varied greatly among the strains and between the sexes (**Fig. 4.2**). Insulin secretion is usually represented by one of three metrics: total insulin secreted (total secretion) (**Fig. 4.2** panel 1), fold-change in insulin secreted over basal (fold-change) (**Fig. 4.2** panel 2), and insulin secreted as a percent of insulin content (% of content) (**Fig. 4.2** panel 3). Each measure provides different information. Total insulin secreted depicts how much insulin was secreted from the islets, fold-change in insulin secreted over basal illustrates the robustness of the insulin secretory response, and insulin secreted as a percent of insulin content reports the contribution of insulin content to the amount of secreted insulin. In all strains, 16.7 mM glucose plus palmitate (G16.7 + PA) elicited the largest insulin secretory response. With some exceptions, the remaining secretagogues had decreased potency in the following rank order: 3.3 mM glucose plus KCl (G3.3 + KCl), 16.7 mM glucose (G16.7), sub-maximal glucose with amino acids (G8.3 + AA), sub-maximal glucose with GLP-1 (G8.3 + GLP-1), sub-maximal glucose alone (G8.3), and low glucose (G3.3). Inter-strain variability was apparent, particularly in response to more moderate secretagogues (G8.3, G8.3 + GLP-1, G8.3 + AA, G16.7, G3.3 + KCl). At the two extreme insulin secretory conditions (G3.3 and G16.7 + PA), we observed the most consistent secretion responses across all strains, suggesting that basal release and release in response to a strong stimulus can overcome genetic influences. Islets from NZO mice secreted the greatest amount of total insulin in response to all secretagogues, including G3.3, the basal condition (**Fig. 4.2** panel 1). These results suggested that NZO islets have secrete high levels of insulin under non-stimulatory conditions. When normalizing insulin secretion to insulin content, NZO

islets appeared to demonstrate superb secretory capacity (**Fig. 4.2** panel 3). This trend, however, is driven by the low insulin content in these mice (**Fig. 4.1h**). Indeed, NZO islets showed reduced responsiveness to several secretagogues, including G16.7 + PA, compared to the other strains, when represented as fold-change over basal (**Fig. 4.2** panel 2), showing that the majority of insulin secreted from NZO islets is basal, unregulated secretion. In addition to strain, sex exerted a strong influence on insulin secretion in some (B6, CAST, 129, PWK, NZO), but not all strains, suggesting strain-by-sex interactions. Strain-by-sex interactions became more apparent when insulin secretion was represented as fold-change over basal; fold-change in insulin secretion was much lower for females than males in several strains, including B6, 129 and CAST and to a lesser degree, in PWK and NZO. These data show that genetic background has a strong influence on insulin secretion in response to a variety of secretagogues, both metabolic and non-metabolic. In addition to insulin, we measured glucagon secretion from the isolated islets in response to KCl. Islets from PWK, NZO, and NOD mice secreted the highest amount of glucagon, and islets from B6, A/J, and 129 secreted the least (**Supplementary Fig. S4.3b**). However, when glucagon secretion was expressed as a percent of content, these strain-differences were reduced, demonstrating that glucagon content strongly influences the amount of glucagon secreted (**Supplementary Fig. S4.3c**).

Whole-islet proteomics reveals strain- and sex-dependent differences We measured the islet proteomes of the eight HF/HS-fed CC founder strains from both sexes, using

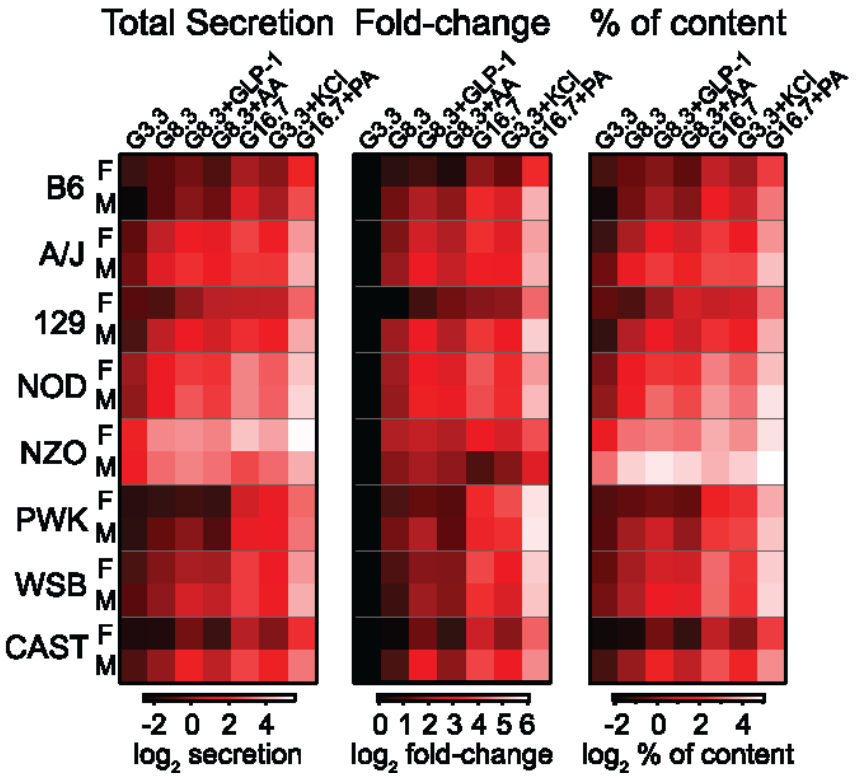
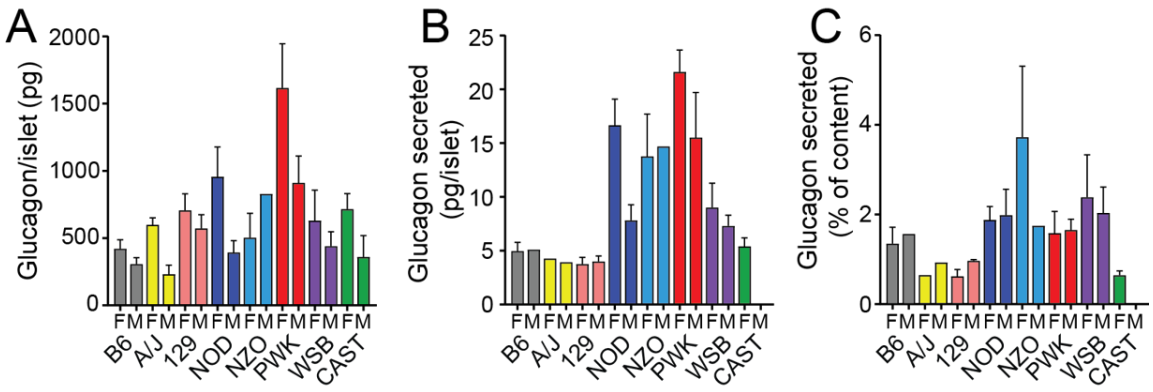


Figure 4.2: The insulin secretory response of isolated islets is influenced by genetic background. Heat maps illustrate three metrics of the insulin secretory response of cultured islets; total amount of insulin secreted, secretion as fold over basal, and secretion as a percent of islet insulin content. The following conditions were used to stimulate insulin secretion from islets of the HF/HS diet-fed CC founder strains at 20 weeks of age: 3.3mM glucose (G3.3, basal conditions), 8.3mM glucose (G8.3), 8.3mM glucose+100nM GLP-1 (G8.3+GLP-1), 8.3mM glucose+1.25mM L-alanine, 2mM L-glutamine, and 0.5mM L-leucine (G8.3+AA), 16.7mM glucose (G16.7), 3.3mM glucose+40mM KCl (G3.3+KCl), and 16.7mM glucose+0.5mM palmitic acid (G16.7+PA). Values represent average secretory responses for > 3 mice/sex/strain, except NZO male mice, where a pool of islets from 4 mice were used.



Supplementary Figure S4.3: The islet glucagon content and secretory response is influenced by genetic background. A Total glucagon per islet (pg) **B** total glucagon secreted from isolated islets in response to 3.3mM glucose+40mM KCl (G3.3+KCl) and **C** glucagon secreted from isolated islets in response to G3.3+KCl represented as a percent of islet glucagon content were determined for all mice at 20 weeks of age, except for NZO male mice due to severe hyperglycemia at 14 weeks of age. Data are mean ±SEM, N > 3 mice/-sex/strain for glucagon content. N varies for glucagon secretion, as some samples were off the curve low on our glucagon assay.

high-resolution mass spectrometry coupled with nano-flow liquid chromatography[10–18]. This did not include male NZO mice, as these animals yielded too few islets as a result of severe diabetes. Our analysis yielded an average detection of 23,148 unique peptides (**Supplementary Fig. S4.4a**), corresponding to 4,705 quantified proteins per sample (**Supplementary Fig. S4.4b**). We quantified 5,255 total proteins, and 4,775 across all eight strains (**Supplementary Fig. S4.4c** and **Table S1**), yielding >90% overlap among the samples, which permitted across-strain comparisons. Quantitative reproducibility was excellent with a median coefficient of variation of 16.7% across all samples (**Supplementary Fig. S4.4d**). To identify strain-dependent patterns in the islet proteome, we computed the Z-score for all identified proteins, followed by unsupervised hierarchical clustering (**Fig. 4.3**). The Z-score indicates how many standard deviations a data point (in this case protein abundance) is from the mean (the mean abundance of that protein across all samples). All 5,255 proteins were used in the clustering, and those that were not detected in a sample are colored grey and denoted N/A. Clustering resulted in the samples grouping strongly by strain and sex (vertical axis). CAST, NZO, PWK, and B6 grouped perfectly based on both strain and sex. All A/J mice grouped together and nearly grouped based on sex. The WSB mice clustered into two groups nearly according to sex. All but one NOD and one 129 mouse grouped by strain and sex. One male NOD and one female 129 grouped with the male WSB mice. The protein clustering (horizontal axis) resulted in sub-sets of strain-specific up- and down-regulation of protein abundance that were significantly enriched for Gene Ontology (GO) terms (**Fig. 4.3**). The largest set contained 1,190 proteins and

enriched for GO term "protein transport" ($P < 10^{-60}$). These proteins were downregulated in CAST and male WSB, and upregulated in NZO female islets. Many of these proteins are involved in vesicle fusion (Vamps, Syntaxins). The upregulation of these proteins in the NZO female islets suggests that an increase in vesicle transport and fusion could explain the high basal, unregulated secretion from these islets (**Fig. 4.2**, panel 1). There was a smaller set of 47 proteins associated with antigen processing and presentation ($P < 10^{-12}$) that were exclusively upregulated in a subset of NOD mice, which included histocompatibility 2 class II antigen A alpha (H2-Aa), A beta 1 (H2-Ab1), E beta (H2-Eb1), and CD74 antigen (Cd74). These proteins have previously been shown to be enriched in intra-islet myeloid cells (21,22). It is possible that one or more of these proteins is involved in the autoimmune-mediated death of β-cells in NOD, a model for type 1 diabetes. Two clusters of proteins showed differential regulation in NZO only; ER proteins (upregulated; $P < 10^{-12}$) and mitochondrial proteins (downregulated; $P < 10^{-40}$). Some of the proteins associated with mitochondrial function that are decreased in the NZO female islets may be involved in the relatively poor insulin secretory response that we observed in NZO islets (**Fig. 4.2**, panel 2). Our data show that genetic background exerts a strong influence on the islet proteome that can likely be causally linked to differences in insulin secretion.

Islet proteome co-expression modules enrich for physiological pathways The results presented in **Fig. 4.3** prompted us to use a weighted gene co-expression network analysis (WGCNA) approach[19,20] to compute co-expression modules consisting of highly correlated

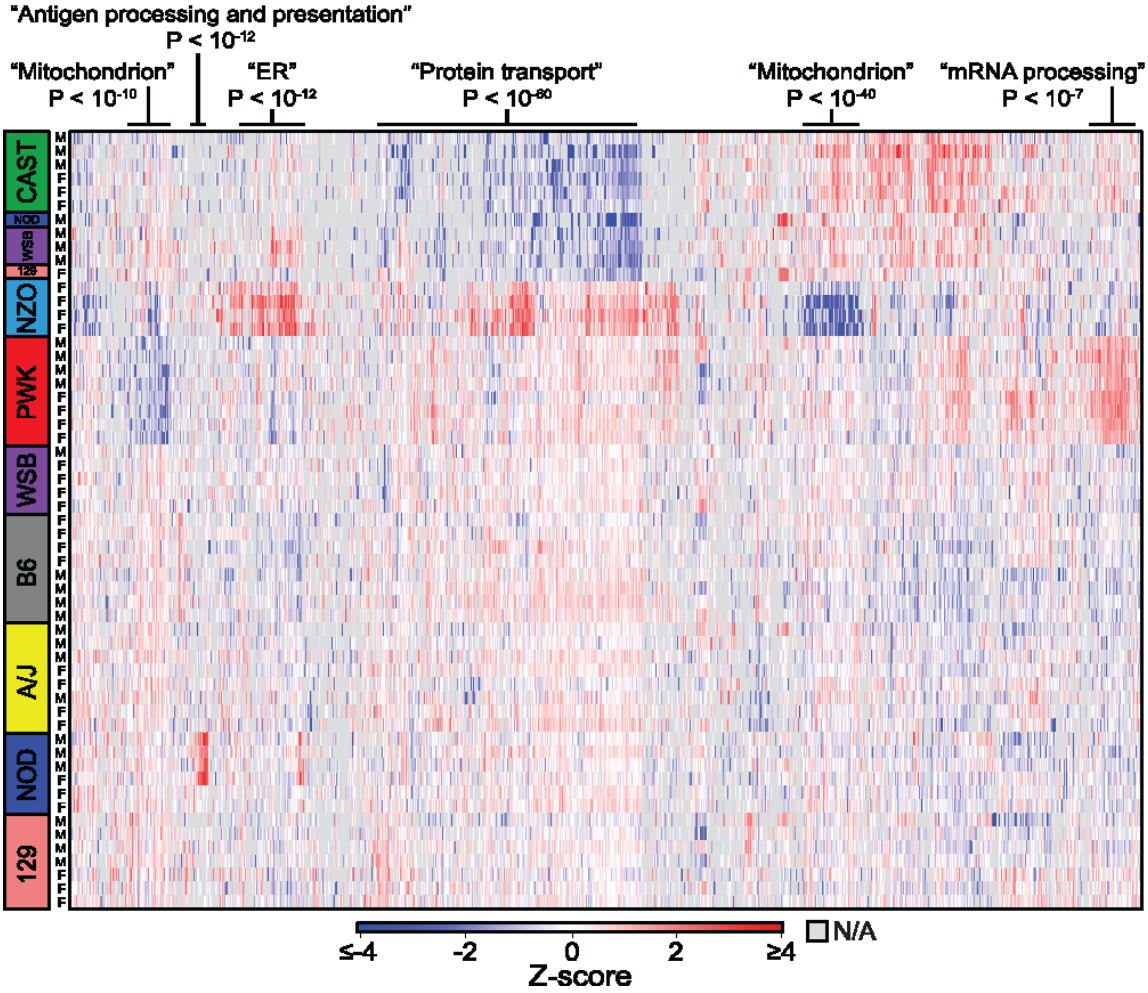
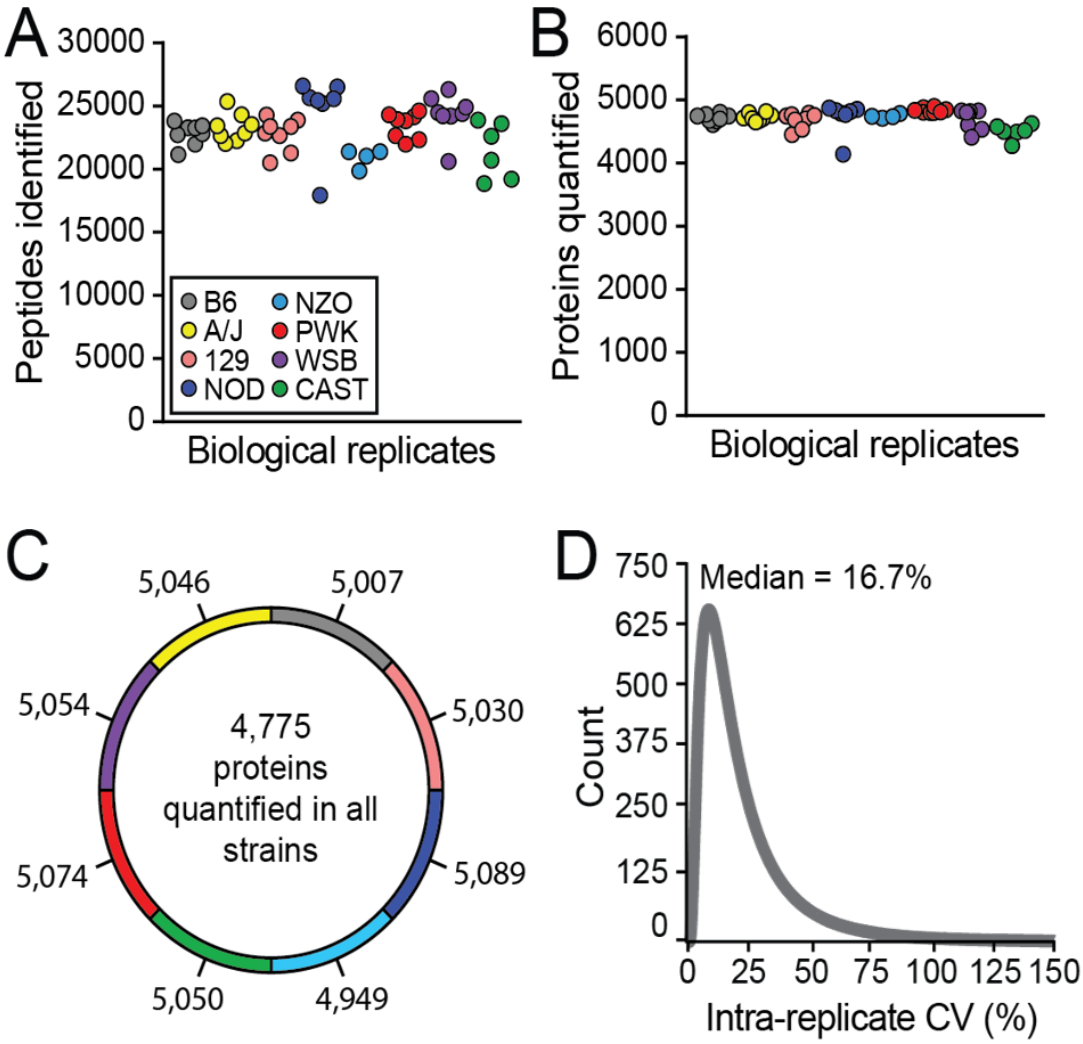


Figure 4.3: Whole-islet proteome is strongly influenced by genetic background. Proteomics analysis was conducted on islets of all eight of the founder strains and both sexes, with the exception of the male NZO (yielded too few islets as a result of the severe diabetes induced by the HF/HS diet) using high-resolution mass spectrometry coupled with nano-flow liquid chromatography. Samples were quantified via MaxLFQ. Z-scores were calculated across the proteins to standardize the quantitative output, then the data was hierarchically clustered both across proteins and across strains via Perseus. All 5,255 proteins were used in the clustering and proteins that were no detected in a given sample are colored grey and denoted N/A. The protein clustering (horizontal axis) resulted in some striking areas of strain-specific up- and down-regulation of protein abundance. Using a distance threshold of 0.82, we defined 100 clusters for the purposes of Gene Ontology (GO) enrichment. Of these, six major clusters displayed both significant enrichment for GO terms and marked differential expression across strains, and these are highlighted in the clustering heat map.



Supplementary Figure S4.4: The islet glucagon content and secretory response is influenced by genetic background. **A** Total glucagon per islet (pg) **B** total glucagon secreted from isolated islets in response to 3.3mM glucose+40mM KCl (G3.3+KCl) and **C** glucagon secreted from isolated islets in response to G3.3+KCl represented as a percent of islet glucagon content were determined for all mice at 20 weeks of age, except for NZO male mice due to severe hyperglycemia at 14 weeks of age. Data are mean ±SEM, N > 3 mice/-sex/strain for glucagon content. N varies for glucagon secretion, as some samples were off the curve low on our glucagon assay.

protein subgroups (**Table S2**). Proteomics experiments can provide more information than a list of differentially expressed proteins. WGCNA can be used to analyze this higher-level information by considering relationships between measured proteins, which can be assessed by correlations between expression profiles. WGCNA starts with thousands of proteins, identifies co-expression modules, and uses correlation between an expression profile and a sample trait to identify important proteins for further validation (see Experimental Methods for computational details). When grouping proteins into co-expression modules, we did not utilize information about functional annotation. Among the 5,255 proteins identified from our whole-islet proteomics experiment, ~83% were uniquely assigned to a co-expression protein module. The WGCNA approach computed 20 co-expression modules from the proteomics data, which are identified by a color name (**Table S3**). The modules contained varying numbers of proteins, ranging from 49 to 1,396. A cluster dendrogram shows the modules as downward branches (**Fig. 4.4**). The depth of the branches indicates the overall correlation between proteins in a module, with deeper branches having greater correlation. **Table S2** lists all modules and their protein membership. We and others have shown that highly correlated transcripts (in this case, proteins) are often associated with common physiological pathways[21–25]. GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) were used to determine if the modules contained proteins that enriched for specific biological pathways. Remarkably, all modules were significantly enriched (Z-score > 3) with one or more GO and/or KEGG terms. In the cluster dendrogram, module branches are labeled with a general description of the overall GO terms or KEGG

pathways enriched in each module (**Fig. 4.4**). All significantly enriched categories for the modules are included in **Table S4**. For each module, we computed a module eigengene (ME) (first principal component (PC1)) to describe the pattern of protein abundance among all of the CC founder mice. The ME can be considered a representative of the protein expression profiles in a module. MEs for all modules are shown in **Supplementary Figs. S4.5, S4.6, S4.7, and S4.8**, and illustrate the protein abundance pattern across the strain-sex combinations. The variance described by the MEs ranged from ~31% (blue module) to ~47% (lightcyan module) (**Table S3**). Proteins that were not identified within a co-expression module were put into the grey module (904 proteins), which had a variance described by the ME of ~6%. The variance described by the MEs is the percent variance among the proteins within a module that is explained by the PC1, or the ME. Typically, they can be ~30% or greater, and much higher than the percent variance describing the ME for the grey module. This shows that proteins in the non-grey modules have highly coordinated expression. The top-enriched module was lightcyan, which enriched for the GO term "cytosolic ribosome" (Z = 36.3). The lightcyan module contains 83 proteins, including ribosomal protein S2 (Rps2), ribosomal protein L18 (Rpl18) and many other Rpl and Rps proteins. Proteins in this module were most highly upregulated in islets from both sexes of CAST, as well as female WSB , and most highly downregulated in islets from female B6 and male 129 (**Supplementary Fig. S4.5**). The abundance of these ribosomal proteins may reflect the amount of protein turnover in these islets. Other modules that highly enriched for physiological pathways included tan, which enriched for the GO terms "response to

interferon-gamma" (Z = 15.6) and "immune response" (Z = 13.05) and the KEGG pathways "Staphylococcus aureus infection" (Z = 15.8) and "antigen processing and presentation" (Z = 10.0). This module contains histocompatibility 2 class II antigen A alpha (H2-Aa), A beta 1 (H2-Ab1), E beta (H2-Eb1), CD74 antigen (Cd74) and interferon-induced guanylate-binding protein 2 (Gbp2). This module describes the cluster of proteins highly abundant in NOD islets shown in **Fig. 4.3**. The midnightblue module was highly enriched for the GO term "serine-type endopeptidase activity" (Z = 15.6), and the KEGG pathway "pancreatic secretion" (Z = 14.5). This module contains pancreatic lipase (Pnlip), pancreatic lipase-related protein 2 (Pnliprp2), and pancreatic colipase (Clps), which reflects the unavoidable contamination of acinar tissue in isolated islet preparations. The ME for midnightblue shows that proteins in this module are upregulated in 129 and downregulated in A/J islets, which may reflect the amount of contaminating acinar tissue in the islet preparations from these strains (**Supplementary Fig. S4.6**). The turquoise module contains 1,396 proteins, which enriched for the GO term "Golgi vesicle transport" (Z = 8.6) and the KEGG pathway "SNARE interactions in vesicular transport" (Z = 7.0) and included adaptor related protein complex 1 gamma 1 subunit (Ap1g1), Rab8a, Sec22b, vesicle associated membrane protein 7 (Vamp7), syntaxin 6 (Stx6), and many other Rabs, Secs, Vamps, and Stxs. Proteins in the turquoise module are upregulated in female NZO and downregulated in male WSB and female CAST islets, and describe the "protein transport" cluster in **Fig. 4.3**. Islets from the NZO mice had high basal secretion and poor insulin secretory response to glucose when insulin secretion was presented as fold over basal (**Fig. 4.2**), suggesting that upregulation

of proteins involved in vesicle transport and SNARE interactions results in a high rate of non-regulated basal secretion. The red module, enriched in the KEGG pathways "oxidative phosphorylation" (Z = 10.4) and "citrate cycle" (TCA cycle) (Z = 8.8), has an ME with the opposite pattern to that of turquoise. Proteins in the red module are downregulated in female NZO and upregulated in male WSB and female CAST islets. The red module includes ATP synthase H+ transporting mitochondrial F1 (Atp5a1) and other mitochondrial ATP synthase subunits, cytochrome C oxidase subunit 5A (Cox5a), isocitrate dehydrogenase 3 (NAD(+)) gamma (Idh3g), and pyruvate carboxylase (Pcx). This decrease in proteins in the TCA cycle and OxPhos pathway in the NZO islets could also explain their lack of response to glucose-stimulated insulin secretion. These results demonstrate that a network analysis yields robust protein sets (modules) that enrich for biological function, demonstrate striking strain- and sex-dependent patterns of protein abundance, and are describable by an ME that captures a large portion of the variance across the samples.

Islet proteome co-expression modules correlate with diabetes-related phenotypes To determine the potential physiological significance of the islet modules, we asked if the modules were correlated with the diabetes-related metabolic phenotypes we measured in the CC founder mice. Using MEs, we computed the correlation between the modules and several whole-body physiological traits (e.g. plasma insulin), and the insulin and glucagon secretory responses measured from the isolated islets (**Fig. 4.5**). Because islets isolated from the same mice were used for both the secretion studies and proteomics anal-

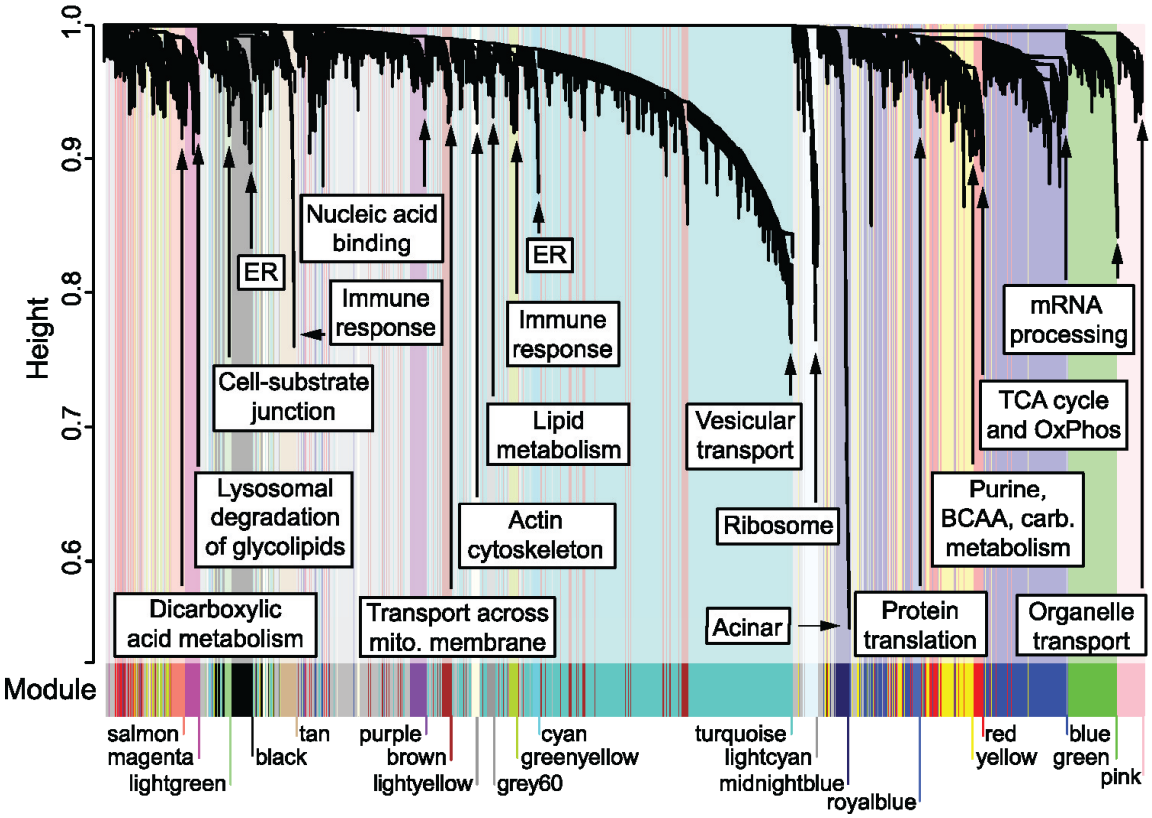
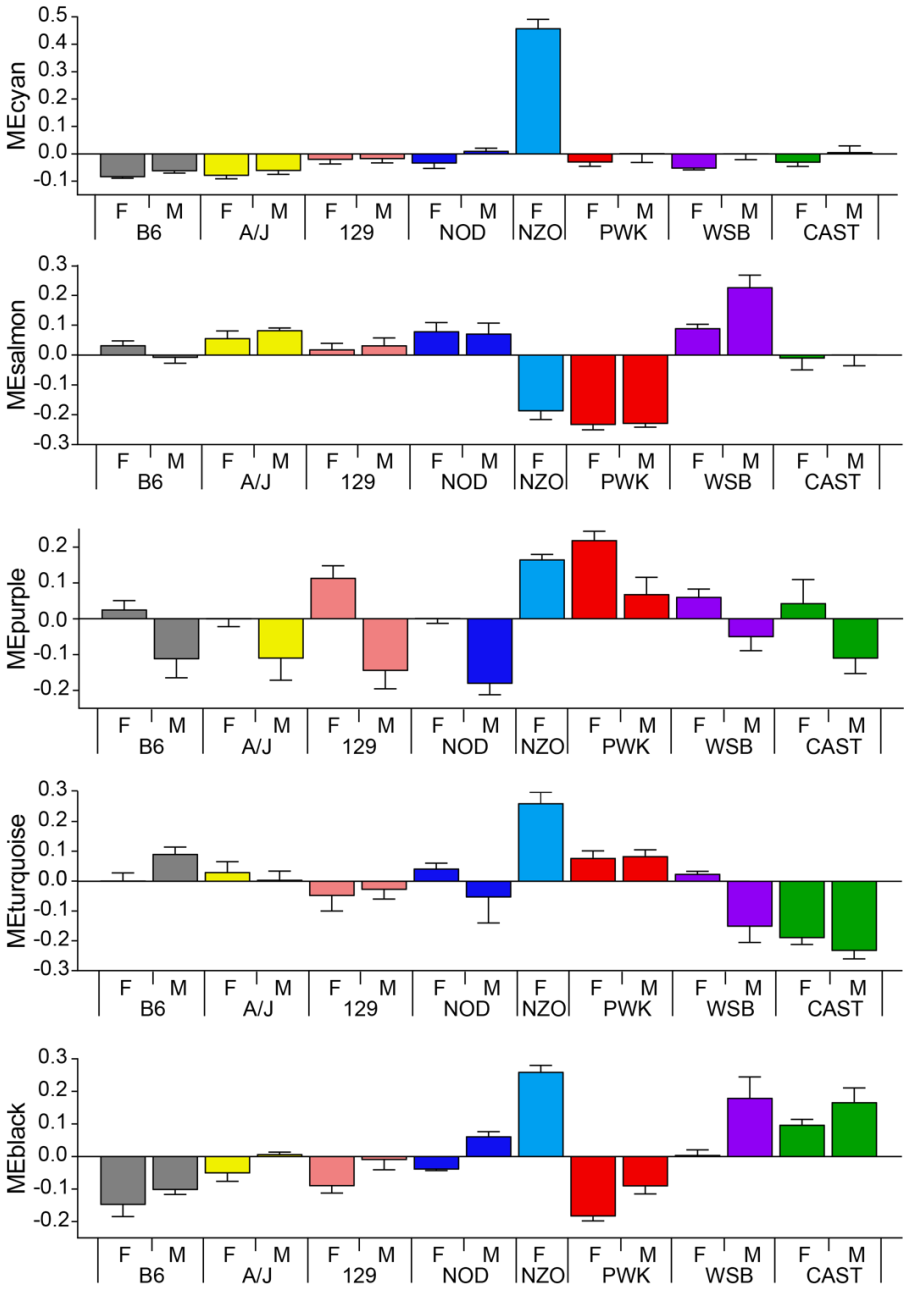
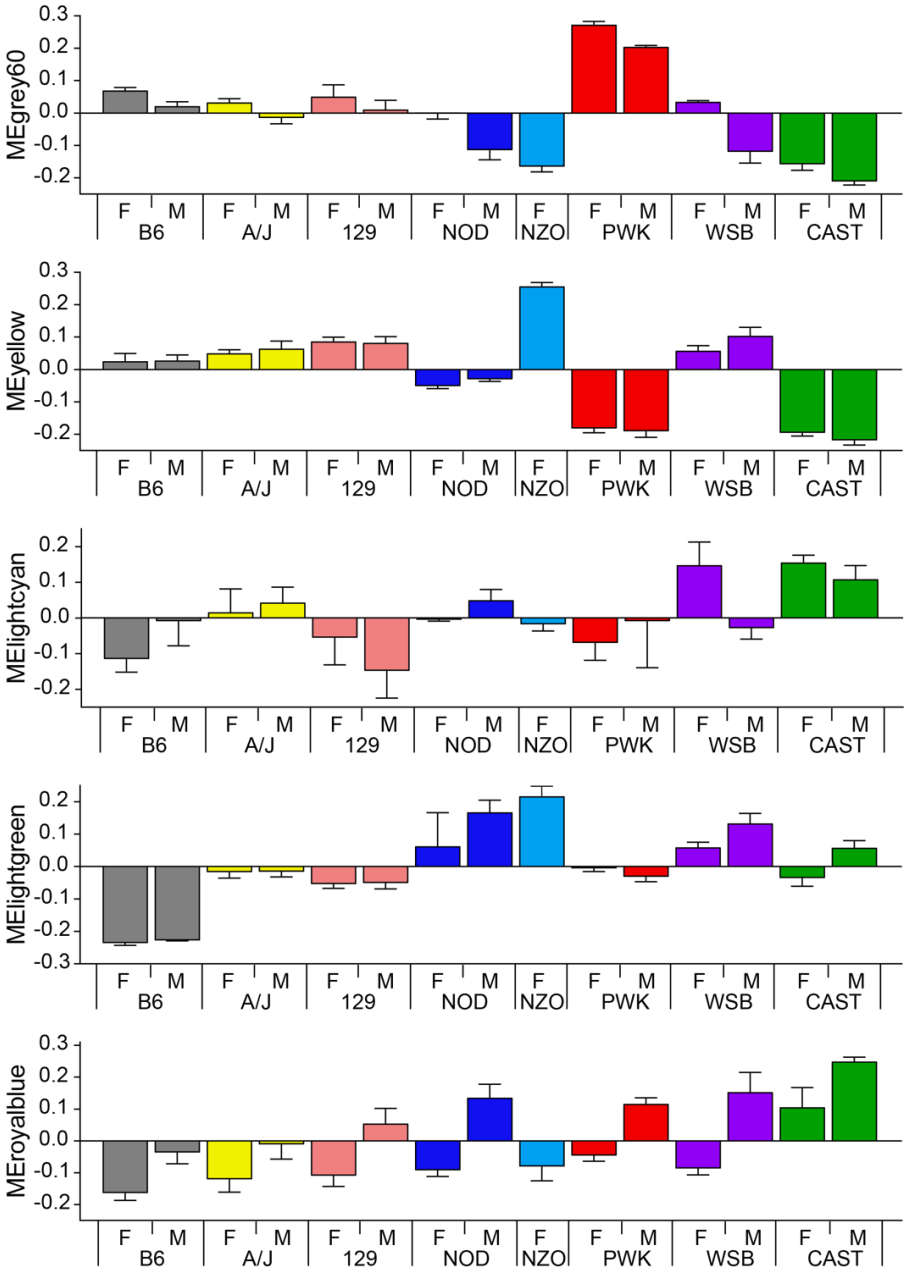


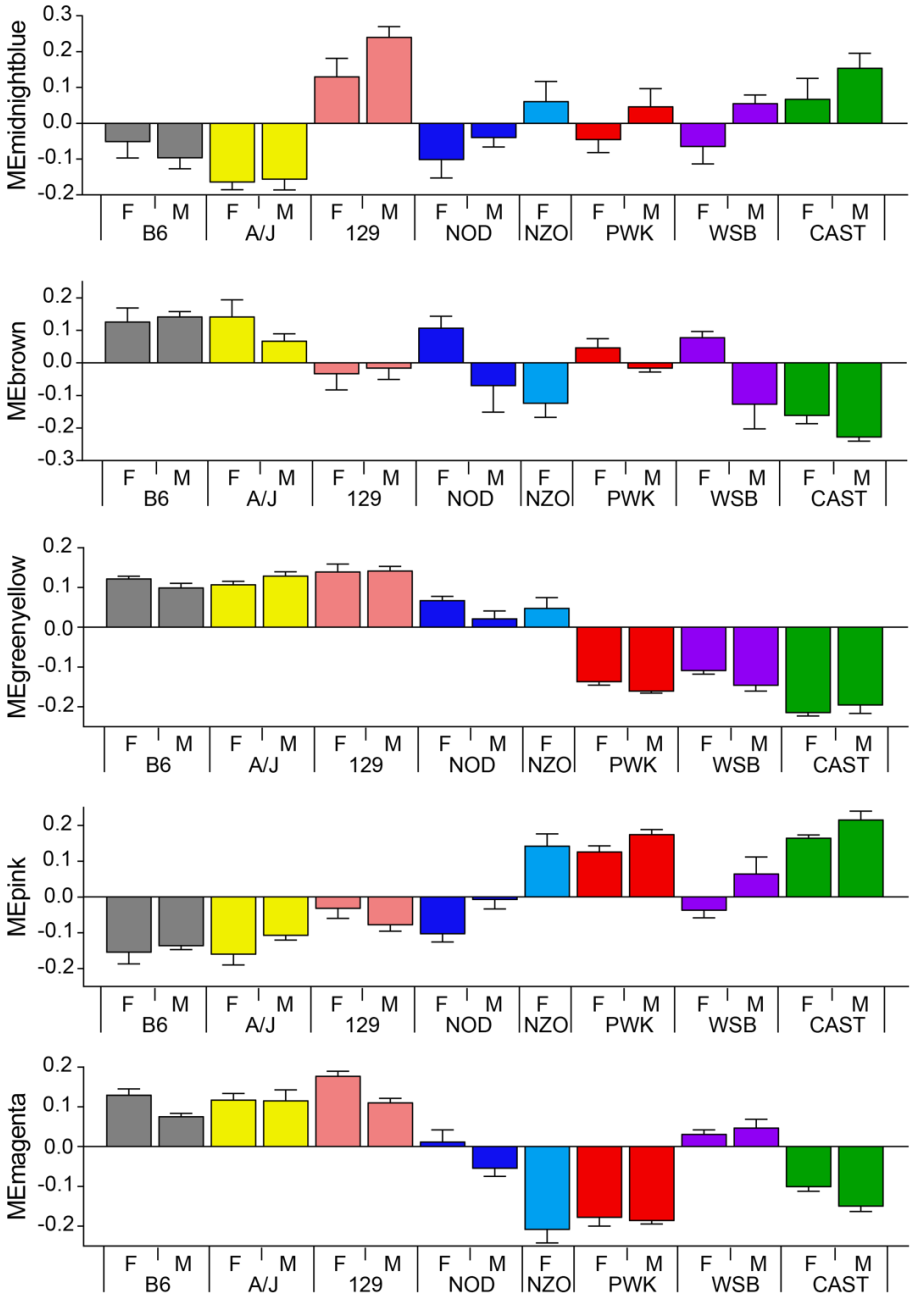
Figure 4.4: Islet proteome co-expression modules enrich for physiological functions. WGCNA-based unsupervised clustering of the whole-islet proteome of the CC founder strains identifies co-expression modules of highly correlated proteins as shown in the cluster dendrogram. Modules, denoted by color, are indicated by the downward branches. Enrichment analysis for Gene Ontology (GO) and the Kyoto Encyclopedia for Genes and Genomes (KEGG) was performed to determine if modules contained proteins that over represented specific biological pathways. All modules were significantly enriched with one or more GO/KEGG terms ($Z > 3$). Branches are labeled with general descriptive terms; complete lists of all GO/KEGG terms for the modules are included in **Supplementary Table 4.**



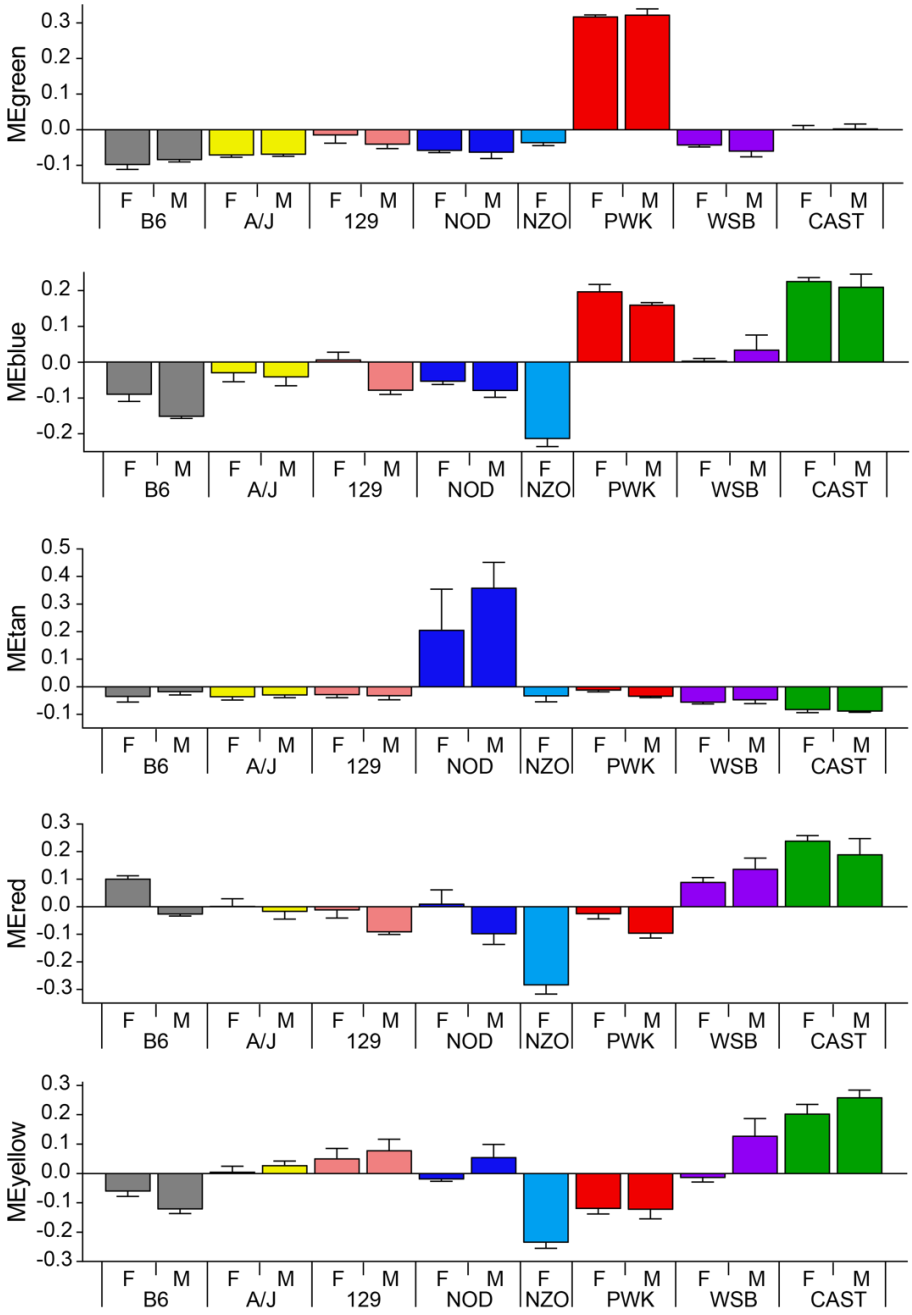
Supplementary Figure S4.5: Graphs of the module eigengenes across the CC founder strains for each module.



Supplementary Figure S4.6: Graphs of the module eigengenes across the CC founder strains for each module.



Supplementary Figure S4.7: Graphs of the module eigengenes across the CC founder strains for each module.



Supplementary Figure S4.8: Graphs of the module eigengenes across the CC founder strains for each module.

ysis, we were able to directly compare the protein and insulin secretion measurements. Several modules were significantly correlated with more than one phenotype (**Fig. 4.5**). For example, the lightgreen module, enriched for the GO terms "cell-substrate adherens junction" (Z = 8.8) and "focal adhesion" (Z = 8.8), showed the strongest positive correlation with insulin secretion in response to all secretagogue classes, as well as positive correlation with triglyceride (TG) and to a lesser extent, plasma insulin. Proteins in the lightgreen module include integrin linked kinase (Ilk), melanoma cell adhesion molecule (Mcam), thy-1 cell surface antigen (Thy1), and lectin galactoside-binding soluble 1 (Lgals1), and are upregulated in islets from NOD and NZO and downregulated in islets from B6 (**Supplementary Fig. S4.5**). This suggests that the high total insulin secretion from the NZO and NOD islets and low insulin secretion from the B6 islets shown in **Fig. 4.2** could be due to an increase or decrease, respectively, in proteins involved in cell-substrate junctions. The cyan and black modules, enriched in endoplasmic reticulum proteins, also showed a similar correlation pattern to these phenotypes. The cyan module enriched for the KEGG pathway "protein processing in endoplasmic reticulum" (Z = 3.3) and includes DnaJ heat shock protein family (Hsp40) member C1 (DnaJc1), protein disulfide isomerase family A member 15 (Txndc5), signal sequence receptor subunit 1 (Ssr1), and SEC13 homolog nuclear pore and COPII coat complex component (Sec13). The ME for cyan shows that proteins in this module are highly upregulated in islets from female NZO mice (**Supplementary Fig. S4.7**). The black module enriched for the GO terms "endoplasmic reticulum chaperone complex" (Z = 13.7) and "response to endoplasmic reticulum stress" (Z = 6.8) and includes

protein disulfide isomerase family A member 4 (Pdia4), heat shock protein 90 beta family member 1 (Hsp90b1), calreticulin (Calr), and endoplasmic reticulum lectin 1 (Erlec1), and are also upregulated in islets from female NZO mice and are downregulated in female B6 and PWK islets. This suggests that the high insulin secretion from the NZO islets and low insulin secretion from the B6 and PWK islets shown in Figure 2 could be due to an increase or decrease, respectively, in proteins involved in the ER stress response. Modules enriched for the GO terms "aerobic respiration" (Z = 10.7) (red), "mRNA processing" (Z = 9.8) (blue), and "lipid metabolic process" (Z = 3.1) (grey60) showed the strongest negative correlation to insulin secretion in response to all secretagogue classes, as well as a negative correlation to body weight and plasma insulin. The magenta module was the most highly negatively correlated with glucagon secretion in response to KCl and enriched for the GO term "glycosphingolipid metabolic process" (Z = 8.8) and the KEGG pathway "lysosome" (Z = 6.6). Proteins in the magenta module include galactosylceramidase (Galc), GM2 ganglioside activator (Gm2a), hexosaminidase subunits a and b (Hexa and Hexb), and prosaposin (Psap). The ME for magenta shows that proteins in this module are generally upregulated in B6, A/J, and 129 islets, and downregulated in female NZO, PWK, and CAST islets (**Supplementary Fig. S4.6**).

Tyrosine hydroxylase is highly abundant in β-cells of PWK and CAST islets Previously, we showed that CAST mice were resistant to the HF/HS diet, and demonstrated remarkably rapid insulin and glucose responses during an oGTT[8]. Further, a preliminary survey of

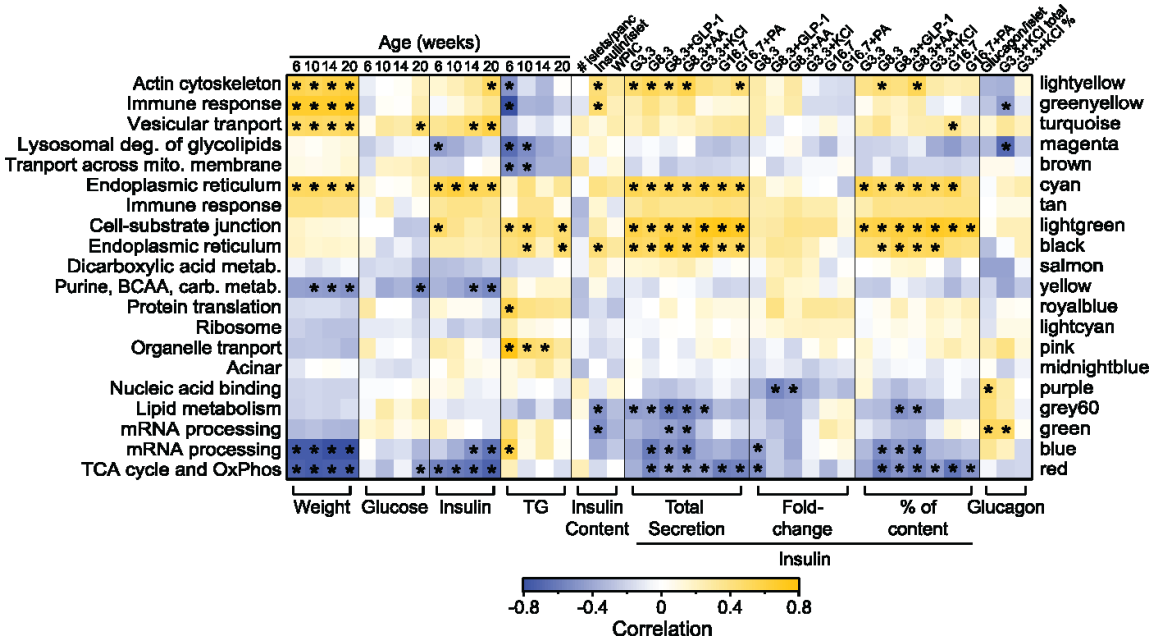


Figure 4.5: Islet proteome co-expression modules correlate with physiological phenotypes. Heat map illustrates the correlation between module eigengenes (MEs) and diabetes-related phenotypes (body weight, fasting plasma glucose, fasting plasma insulin, fasting plasma triglycerides (TGs), number of islets per pancreas, insulin content per islet, whole-pancreas insulin content, three metrics for insulin secretion, glucagon content per islet, and two metrics of glucagon secretion in response to G3.3+KCl) measured from the CC founder strains. Correlations computed from normalized quantile ranks; * $P \leqslant 0.001$.

the islet phosphoproteome in islets from CAST mice showed that serine 31 on tyrosine hydroxylase (Th) was phosphorylated in response to glucose. Th activity is regulated by phosphorylation at specific residues[26]. ERK1/2-mediated phosphorylation at serine 31 stabilizes the enzyme, and stimulates catalytic activity[27,28]. Th is the first step in the catecholamine synthesis pathway, converting L-tyrosine to L-DOPA. Catecholamines are potent inhibitors of insulin secretion[29–31]. Our survey of the islet proteome in the CC founder mice shows that Th is ~70-fold higher in islets from CAST and PWK mice compared to the other strains (**Fig. 4.6b**). Th is present in the blue module. The ME for blue shows that the blue module proteins are generally upregulated in PWK and CAST islets and downregulated in female NZO and male B6 islets (**Fig. 4.6a**). GO terms that describe Th in this module include "cell body" (Z = 3.4), "axon" (Z = 2.4), and "catecholamine metabolic process" (Z = 2.1) (**Table S4**). These GO terms for the blue module also include neural cell adhesion molecule 1 (Ncam1) and neuropeptide Y (Npy), among others. Ncam1 has been shown to be required for cell-type segregation and normal ultrastructure in pancreatic islets[32]. Npy is a secreted neuropeptide that influences many physiological processes, including cortical excitability, stress response, and food intake[33], and also inhibits insulin secretion[34]. These other proteins in the blue module should also be elevated in islets from PWK and CAST islets like Th. Indeed, Npy is 4-fold higher in PWK islets and 12-fold higher in CAST islets over B6 islets. These proteins in the blue module are associated with inhibition of insulin secretion, consistent with the negative correlation between the ME of the blue module and insulin secretion in **Fig. 4.5**. In pancreatic sections from B6, PWK, and

CAST mice, we determined the proportion of β-cells and α-cells that were Th-positive in B6, CAST or PWK islets using immunohistochemistry (**Fig. 4.6c,d**). PWK and CAST islets had ~35-fold more Th-positive β-cells per islet area ($P < 0.0001$), compared to B6 islets. There was no statistically significant difference in the number of Th-positive α-cells or Th-positive unidentified cells (neither β nor α) per islet area across the strains. In summary, our data shows that β-cells from CAST and PWK mice have greatly elevated Th protein, suggesting that islets from these mice utilize catecholamines as an additional regulatory mechanism for insulin secretion that is absent in strains that have low Th levels.

Increased dopamine production in CAST islets is associated with decreased insulin secretion Catecholamine synthesis begins with Th converting L-tyrosine to L-DOPA, which in turn becomes dopamine via DOPA decarboxylase (Ddc). Dopamine can then become norepinephrine via dopamine β-hydroxylase (Dbh) and norepinephrine can become epinephrine via phenylethanolamine N-methyltransferase (Pnmt). Excess dopamine is metabolized by two enzymes, Comt, producing 3-methoxytyramine (3-MT), and Mao, producing 3,4-dihydroxyphenylacetic acid (DOPAC). These two intermediates are further metabolized to produce homovanillic acid (HVA) by Mao or Comt, respectively. Interestingly, our proteomic data revealed that Ddc is highly expressed among all eight mouse strains and sexes, Comt is more abundant in CAST islets than other strains, and Mao is equally abundant among the eight mouse strains and sexes (**Table S1**). Dbh and Pnmt were not detected in any of the strains or sexes. Thus, we hypothesized that elevated Th activity

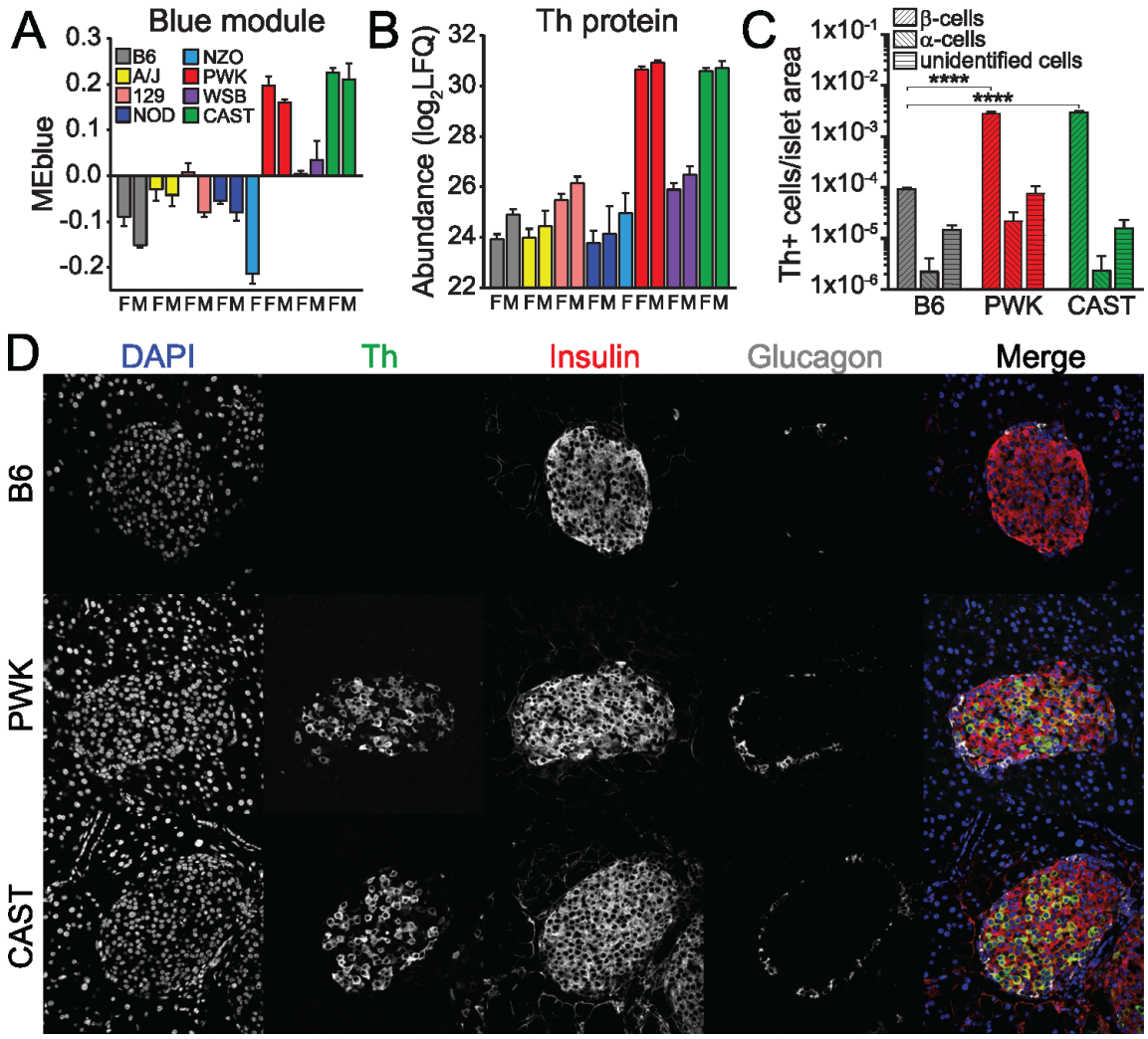


Figure 4.6: Tyrosine hydroxylase is highly expressed in β-cells of PWK and CAST islets. **A** The module eigengene (ME) for the blue module illustrates that proteins in the module, which includes Th, are highly expressed in PWK and CAST islets. **B** The abundance of islet Th protein in the islet samples. **D** Immunohistochemistry for Th (green), insulin (red), and glucagon (cyan) in pancreatic sections from male B6, CAST, and PWK mice maintained on the HF/HS diet. **C** Quantification of Th-immunoreactivity in B6, PWK, and CAST islets. The number of Th positive (Th+) β-cells (insulin positive), α-cells (glucagon positive), or unidentified cells (neither insulin nor glucagon positive) per islet area. Data are mean ± SEM, n=3 mice per strain and 40 islets per mouse. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7. **** $P \leqslant 0.0001$.

in CAST islets would yield high levels of dopamine and its various metabolites. Using mass spectrometry, we quantified intermediates of the dopamine biosynthetic pathway in islets isolated from B6 and CAST mice. L-tyrosine, the precursor to L-DOPA, was not significantly different between B6 and CAST islets (**Fig. 4.7a**). Although there was a trend for L-DOPA to be elevated in CAST islets, it was not significantly different between B6 and CAST islets (**Fig. 4.7a**). It is likely that newly synthesized L-DOPA is rapidly converted to dopamine via the high levels of Ddc in the islets of all strains. Indeed, CAST islets had ~5-fold higher levels of dopamine ($P < 0.001$) (**Fig. 4.7c**). 3-MT and DOPAC were also elevated, consistent with enhanced synthesis of these catecholamine intermediates (**Fig. 4.7a**). HVA was not detected in B6 and CAST islets maintained under normal conditions (**Fig. 4.7a**). To bypass the strain difference in Th activity, we incubated B6 islets with L-DOPA, the product of Th. L-DOPA is transported into cells via the cell surface large amino acid transporter (Laat)[35,36], which is equally abundant in all eight strains (**Table S1**). Pre-incubating B6 islets with L-DOPA led to a dramatic increase in islet levels of dopamine, as well as its metabolites, mimicking what we observed with CAST islets (**Fig. 4.7a**). These results strongly suggest that the elevated levels of dopamine in CAST islets are due to the increased abundance and activity of Th. We hypothesized that the increased dopamine levels in CAST islets, or that achieved in B6 islets by pre-incubation with L-DOPA, would result in reduced glucose-stimulated insulin secretion (GSIS). We measured GSIS in islets isolated from B6 and CAST. In order to enhance the suppressive autocrine effect of secreted dopamine on insulin secretion, we incubated 15 islets in 125 μL of secretion media for these experiments.

In response to high glucose (16.7 mM), insulin secretion from CAST islets was reduced by ~60% compared to B6 islets (P = 0.008) (**Fig. 4.7b**). Pre-incubating B6 islets with L-DOPA (50 μM, 45 mins) mimicked the response observed in CAST islets; insulin secretion from B6 islets was reduced by ~40% in response to L-DOPA pre-incubation. The addition of 1 μM dopamine suppressed insulin secretion from B6 islets by ~50%, confirming the autocrine negative feedback previously reported (33,41,42). Interestingly, addition of exogenous dopamine (1 μM) did not cause an additional suppression of secretion from CAST islets, suggesting that endogenously produced dopamine was sufficient to suppress the insulin secretory response. In summary, β-cells of CAST islets express high levels of Th, the first step in catecholamine synthesis, resulting in elevated levels of dopamine. In response to glucose, dopamine is co-secreted with insulin, establishing a negative autocrine feedback that blunts the secretory response.

Discussion

In this study, we used the eight genetically diverse CC founder mouse strains fed a HF/HS diet to assess the contribution of genetic variation to diabetes-related phenotypes. We found that genetic diversity strongly influenced a host of metabolic phenotypes, including body weight, fasting plasma glucose and insulin, and insulin secretion from isolated islets in response to a range of metabolic stimuli. The eight strains displayed a wide range in insulin resistance, as judged by the level of fasting plasma insulin. We previously found that of the eight strains, CAST is the only strain completely resistant to HF/HS diet-induced

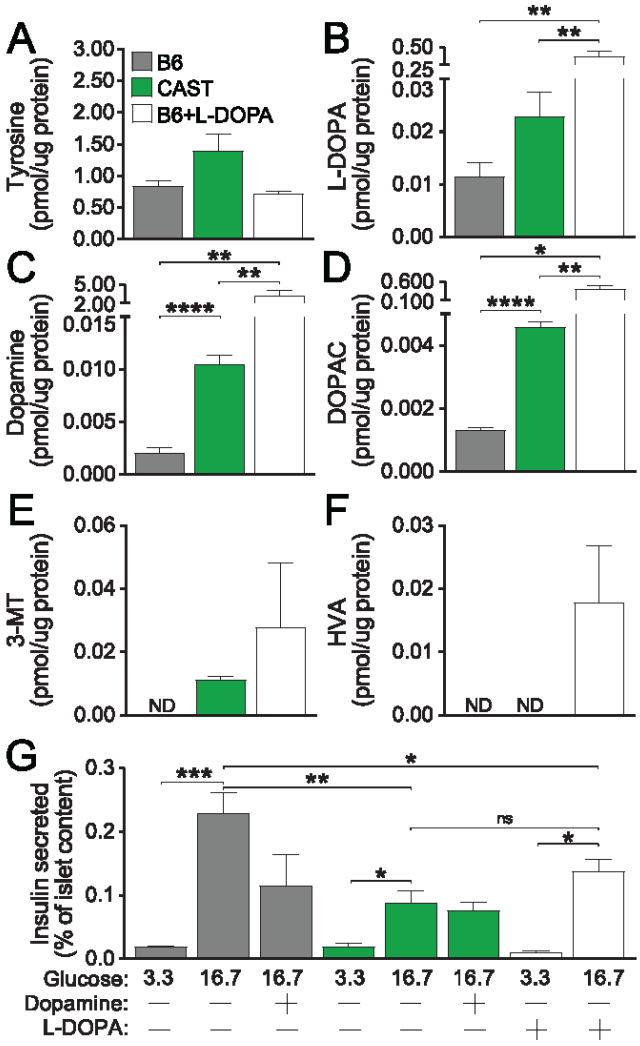


Figure 4.7: Increased dopamine synthesis in CAST islets is associated with decreased insulin secretion. A Abundance of dopamine-related metabolites in B6 and CAST islets from 20 week old HF/HS diet fed mice determined by mass spectrometry. Data are mean ± SEM, n = 4 for B6 and CAST and n = 2 for B6 + L-DOPA. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7. * $P \leqslant 0.05$, ** $P \leqslant 0.01$. **B** Insulin secretion from B6 and CAST islets, and B6 islets pre-incubated with 50 μM L-DOPA (B6+L-DOPA). Insulin secretion was stimulated with 16.7 mM glucose in the absence or presence of 1 μ M dopamine. Insulin secretion from B6 islets that were pre-incubated with L-DOPA was similar to insulin secretion from CAST islets. Data are mean ± SEM, n = 4 for B6 and CAST and n = 2 for B6 + L-DOPA. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7. * $P \leqslant 0.05$, ** $P \leqslant 0.01$.

changes in glucose homeostasis during an oral glucose tolerance test (oGTT), consistent with the high insulin sensitivity of the CAST mice[8]. At the other extreme, the NZO mice are the most insulin resistant, and essentially HF/HS diet-intolerant. NZO male mice become severely hyperglycemic, resulting in death by 14 weeks of age. Thus, genetic variation in the CC founder strains results in a range of phenotypes from complete resistance to lethality in response to the Western-style dietary challenge. We evaluated the relationship between genetic diversity and islet function. The number of islets isolated per mouse, the insulin and glucagon content per islet, and the islet insulin and glucagon secretory response varied widely among the mice. Ranking the mice based on their insulin secretory response to different classes of secretagogues illustrates how genetic background drives the insulin secretory response. When focusing on the fold-change in insulin secretion in response to the different secretagogues, islets from NOD males had the highest insulin secretion in response to G8.3 + GLP-1, whereas islets from B6 females had the lowest response. In contrast, in response to G16.7 + PA, islets from PWK males had the highest insulin secretory response and islets from NZO males had the lowest response. In response to the non-metabolic secretagogue KCl, islets from PWK females secreted the most insulin while islets from B6 females secreted the least. Characterization of islet protein composition is key to unlocking the molecular details of diabetes pathophysiology. One drawback is that islet scarcity has confounded extensive analyses and has required pooling islets from multiple animals. However, recent developments in mass spectrometric technologies have improved sensitivity and permit deep profiling of single animals without the need for

pooling, pre-fractionation, or heavy-isotope quantitative tagging[17,37]. The CC founder strains have been extensively studied individually, and Chick et al. completed proteome profiling of liver tissues across all eight strains[38], but our study represents the first in-depth analysis of the islet proteomes and insulin secretion phenotypes across the whole cohort. To identify islet proteins that underlie the strain- and sex-dependent differences in insulin and glucagon secretion, we conducted proteomics on islets collected from each strain and sex (excluding NZO males) and identified and quantified 5,255 total proteins. The largest murine islet proteome characterized to date contained approximately 6,800 proteins (46). These analyses, however, required extensive peptide fractionation, resulting in 24 hours of analysis per proteome; our methods quantified a proteome of ~75% the size in ~8% of the analysis time. Unsupervised clustering of the islet proteome revealed that the mice clustered based on strain and sex. This shows that genetic background has a strong influence on the islet proteome that can be linked to differences in insulin secretion. We used the WGCNA approach to compute co-expression protein modules consisting of highly correlated proteins. We found that these protein groups enriched for biological pathways and correlated with the diabetes-related phenotypic measures. Correlations can lead to hypotheses that can be tested for causality. For example, two modules (black and cyan) enriched for proteins that are critical for endoplasmic reticulum (ER) homeostasis and positively correlate with insulin secretion, plasma insulin, and body weight. Because the ER is involved in protein folding, modification, and trafficking to the Golgi, ER homeostasis is critical in β-cells[39–42]. Proteins in the black module include hypoxia up-regulated 1

(Hyou1), protein disulfide isomerase associated 4 (Pdia4), and heat shock protein 90 beta (Grp94) member 1 (Hsp90b1), all reported to be upregulated in islets under conditions that elicit ER stress[43–45]. Other potentially novel proteins in the black module may be important for ER Genetic variation in dopamine regulates insulin secretion 10 stress-induced changes in islet function and/or health. The lightgreen module was enriched for proteins involved in cell-substrate junctions and positively correlated with insulin secretion. It is known that adherens junctions between β-cells is required for proper insulin secretion[46? –50]. Proteins in the lightgreen module include annexin A1 (Anxa1) and paxilin (Pxn), both reported to be important in insulin secretion[51,52]. There may be novel proteins in the lightgreen module important for cell-cell communication through adherens junctions and focal adhesions. The grey60 module enriches for the GO terms "phosphoric ester hydrolase activity" and "lipid metabolic process", and negatively correlates with insulin secretion. Proteins in the grey60 module include carnitine palmitoyltransferase 2 (Cpt2), TAM41 mitochondrial translocator assembly and maintenance homolog (Tamm41), and acyl-CoA synthetase short-chain family member 2 (Acss2). Testable hypotheses can be generated about the function of these proteins in negatively regulating insulin secretion. Interestingly, 5 of the 20 modules (salmon, brown, grey60, yellow, and red) were most highly enriched in distinct mitochondrial-associated pathways, enriching for mitochondrial dicarboxylic acid metabolism (salmon), transport across the mitochondrial membrane (brown), mitochondrial lipid metabolism (grey60), mitochondrial purine nucleoside, branched-chain amino acid, and carbohydrate metabolism (yellow), and TCA cycle/OxPhos (red). Each of these

modules consist of different proteins that have distinct expression patterns across the strains and sexes. The presence of these mitochondrially-enriched modules is consistent with the importance of mitochondrial function in islets. Mitochondrial proteins appear to be down-regulated in the islets of NZO mice, which show the greatest total insulin secretion of all the strains in Figure 2 panel 1. However, when represented as fold change over basal secretion (**Fig. 4.2** panel 2), islets from the NZO mice show a clear deficit in regulated insulin secretion. This shows that the NZO mice have a high non-stimulatory basal insulin secretion, and a poor stimulated insulin secretion, for which mitochondrial oxidative metabolism is important[53–55]. A caveat to performing omics studies on whole islets is that changes in islet omics may reflect differences in compositions of islet cell types. Mouse islets are composed of 60-80% β-cells producing insulin and amylin in a central core and a layer of other endocrine cells surrounding the core, which is composed of 15-20% α-cells producing glucagon, <10% δ-cells producing somatostatin, <5% PP-cells producing pancreatic polypeptide, and <1% ϵ cells producing ghrelin (68-71). Also, axon endings can remain within islets after isolation, and fragments of acinar and ductal cells can remain attached to the islets, and their abundance could plausibly be strain-specific. Not only can these compositions be altered by genetic background, but different regions of the pancreas within the same mouse can have islets with different endocrine cell contents[56,57]. Further, recent papers employing single-cell RNAseq and mass spectrometric studies on islet cells have found heterogeneity within islet cell types[58–63]. There are strain- and sex-specific differences in the abundances of the major islet hormones. These differences could be plausibly due to

differences in islet cell type composition and/or hormone content per cell; fluorescence activated cell sorting (FACS) purification of the different cell types in each of the strains would need to be conducted to investigate this. A recent report by Cruciani-Guglielmacci et al.[64] compared the variation in body weight, glucose homeostasis, insulin secretion, and islet gene expression across six different mouse strains (C57BL/6J, DBA/2J, A/J, AKR/J, 129S2/SvPas, and BALB/cJ), all maintained on either regular chow or a high fat/high sucrose (HFHS) diet. Three of these strains (C57BL/6J, A/J, and 129) were included in our current and previous studies[8]. Like our study, striking strain-specific differences in diabetes-related phenotypes were observed in response to the HFHS diet. The HFHS diet resulted in obesity, glucose intolerance and insulin resistance in DBA/2J and AKR/J mice, whereas, these same phenotypes were separable in BALB/cJ, which only showed evidence of glucose intolerance. A major difference between the report by Cruciani-Guglielmacci and colleagues and our studies is the inclusion of CAST mice, which were completely resistant to HF/HS dietary challenge; they showed no change in body weight, glucose homeostasis, or insulin dynamics. Thus, including the wild-derived strains, which contain greater genetic diversity than the classical inbred strains, yielded a higher level of phenotypic diversity. In addition to surveying diabetes-related physiological phenotypes, Cruciani-Guglielmacci and colleagues performed islet transcriptomics on mice maintained on either regular chow or HFHS diet, enabling them to identify transcripts that were diet-responsive in each of the six strains studied. In contrast, our study surveyed whole-islet proteomics in eight mouse strains, all maintained on the HFHS diet. Interestingly, Cruciani-Guglielmacci and

colleagues showed that the islet transcriptional profile was more closely related to genetic background than dietary conditions; length of time on a particular diet, or diet composition. In both studies, WGCNA-based clustering was used to compute islet gene modules (transcriptomics or proteomics), and gene set analysis was conducted on the modules to identify enriched biological pathways. Pathways that were enriched within modules from both studies included cell-substrate junction, immune response, lipid metabolism, actin cytoskeleton, ribosome (biosynthesis), tricarboxylic acid cycle, oxidative phosphorylation, carbohydrate metabolism, and antigen processing and presentation. Some pathways were enriched in only one study (e.g., DNA repair and replication, vesicular transport), and may reflect post-transcriptional regulatory mechanisms, including protein turnover. Unfortunately, Elovl2, a gene validated to play a role in the regulation of insulin secretion by Cruciani-Guglielmacci et al. was not included in the 5,255 proteins that were identified in our study. Future studies would be required to directly assess the genetic dependence of Elovl2 protein abundance differences in the eight Collaborative Cross founder strains, and to what extent these differences play a role in differential insulin secretion among these strains. Driven by our preliminary finding that glucose promotes phosphorylation of serine 31 on Th in CAST islets, we asked if Th was differentially abundant across the strains. Our proteomic survey showed that Th was expressed far more highly in PWK and CAST islets. It has been known for over 40 years that mouse islets can synthesize and secrete dopamine, but seemingly only after supplementing them with its precursor, L-DOPA[65,66]. Mouse β-cells contain all of the components necessary to synthesize dopamine

from L-DOPA. The large aromatic amino acid transporter (Laat) on the surface of the β-cell rapidly transports L-DOPA into the cell. L-DOPA is decarboxylated into dopamine by DOPA decarboxylase (Ddc)[67]. Dopamine is packaged into insulin granules via the vesicular monoamine transporter 2 (Vmat2)[68,68], resulting in co-secretion of dopamine with insulin in response to a stimulus. Dopamine acts in an autocrine fashion to inhibit insulin secretion by binding to dopamine receptors on the surface of the β-cells[29,69,70]. Further, β-cells express monoamine oxidases (Mao) and catechol-O-methyltransferase (Comt), which degrade excess cytoplasmic dopamine[71]. Tyrosine hydroxylase, the enzyme that converts L-tyrosine to L-DOPA, is the only dopamine biosynthetic enzyme thought to be essentially absent in mouse β-cells. However, these conclusions were drawn from studies that utilized B6 mice[29,65–72]. Therefore, the consensus has been that for mouse β-cells to synthesize and secrete dopamine, they must first import L-DOPA. In the central nervous system, dopamine is secreted from neurons and functions as a neurotransmitter, although it is not released into the bloodstream. Peripheral, non-neuronal production of L-DOPA results in nanomolar levels of circulating L-DOPA[73]. One source of this circulating L-DOPA is thought to be intestinal cells, which express high levels of Th[73]. Although some speculate that β-cells import L-DOPA from the circulation[70], evidence that circulating L-DOPA is taken up by the β-cells is elusive. Here, with a survey that included several strains, we show that β-cells from CAST mice express high levels of Th, leading to the synthesis of dopamine. This de novo synthesized dopamine is associated with reduced insulin secretion, which can be mimicked in islets from B6 mice by pre-incubating the islets in L-DOPA, the product of

Th activity. Inhibition of Th in isolated human islets greatly increases insulin secretion[74], suggesting that like CAST islets, human islets synthesize dopamine de novo. Therefore, strains of mice with islet cells that express Th are more appropriate to study this pathway and extend the implications to humans. Why would CAST β-cells synthesize a molecule that potently inhibits insulin secretion? One explanation is that they require an additional mechanism to reduce insulin secretion because of their high level of insulin sensitivity. We show that CAST mice are extremely insulin sensitive, requiring the lowest plasma insulin of all the strains to maintain euglycemia, and being resistant to HF/HS diet. Islets in CAST mice may employ an autocrine dopamine-mediated break on insulin secretion to ensure a brief rise in insulin, followed by a suppression of secretion (as occurs during an oGTT), to avoid hypoglycemia. The B6 mouse strain has become the most widely used mouse model for studying human physiology, as well as the most common strain used in gene editing. However, when a gene alteration fails to produce a phenotype, it is possible that the B6 strain was not the best choice to study the gene's function[75,76], rather than concluding that mice are not appropriate models to study human pathophysiology[77,78]. The "absence of a phenotype" in a single mouse strain can be the result of strain-to-strain variation. The phenotype variation between mouse strains motivates the search for comparable variation across the human population. In our current study, we saw dramatic strain variation in Th expression in mouse islets, which determines the ability to produce dopamine de novo. Based on these results, we predict that there may be genetic variation in humans in the contribution of β-cell-derived dopamine to the regulation of insulin secretion.

A data resource for the research community Our work provides a resource to identify the presence or absence of specific biological pathways and proteins in the islets of the eight genetically and phenotypically diverse CC founder mouse strains. Both collaborating labs have made available searchable databases of our islet proteomics data from the eight strains of mice[79,80]. At http://diabetes.wisc.edu/cc_founder.php under the "Whole islet proteomics" link is a user-friendly web interface that allows the user to enter a gene symbol. If the query was one of the 5,255 proteins identified by our whole-islet proteomic survey, the average abundance of that protein will be displayed across the 15 experimental groups (8 strains of each sex, except NZO males). In addition, we have incorporated a protein-to-protein correlation tool that can be used to identify groups of proteins with highly correlated expression profiles. Lists of correlated proteins can be directly uploaded to the Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.ncifcrf.gov), an NIH-funded bioinformatics resource that provides functional annotation to large gene lists. As an example of the database utility, searching for glucagon (Gcg) and then clicking the "Plot" bar graph icon under "Actions" reveals a striking strain- and sex-dependent pattern of protein abundance. A correlation analysis can then be performed by clicking the "Correlation" icon under "Actions" to determine if other proteins show a similar pattern as Gcg. After setting the desired options and submitting, clicking "Show" for a correlated protein will produce a graph of the correlation between that protein and Gcg across the samples. Clicking "View Details", selecting all of the proteins in the list, and selecting "Heat Map", generates a heat map of the Z-scores of the proteins across the samples, where

proteins, mice, or both, can be hierarchically clustered (Fig. S6C). Returning to the protein list and clicking "DAVID" automatically uploads the correlated list to the DAVID functional annotation tool website. Selecting "Functional Annotation Clustering" reveals that proteins that correlated with Gcg across the strains enrich for chaperone ($P = 2.7x10^{-8}$) and RNA binding ($P = 1.7x10^{-4}$). Proteins within each of these annotation clusters can be identified by clicking on the blue bar for each enrichment term. The website allows for molecules of interest to be queried to generate simple abundance column plots across strains. Individual strains can be compared to one another to determine significant changes across the dataset, and outlier analysis can be performed to identify significant changers specific to individual strains. Our aim is to provide a valuable tool to the scientific community to support further biological inquiry and guide future studies. Our study provides an extremely valuable tool to help determine the appropriate strain and sex in which to study a specific biological pathway, or to knockout a gene.

Experimental procedures

Animals Animal care and study protocols were approved by the University of Wisconsin-Madison Animal Care and Use Committee. Mice were housed within the Biochemistry Department vivarium and maintained on a 12 hour light/dark cycle (6am-6pm). The eight Collaborative Cross founder strains (C57BL/6J (B6), A/J, 129S1/SvImJ (129), NOD/ShiLtJ (NOD), NZO/HILtJ (NZO), PWK/PhJ (PWK), WSB/EiJ (WSB), and CAST/EiJ (CAST)) were obtained from The Jackson Laboratory (Bar Harbor, ME, USA), and bred at the University

of Wisconsin-Madison Biochemistry Department, except for CAST and NZO. Mice were group-housed by strain and sex (2-5 mice/cage) except for CAST that required individual housing. Mice were housed under a temperature and humidity-controlled conditions, and received ad libitum access to water and food. Beginning at 4 weeks of age, mice were maintained on a high-fat/high-sucrose diet (HF/HS) (TD.08811, Envigo Teklad Custom Diet, 44.6% kcal from fat, 14.7% kcal from protein, 40.7% kcal from carbohydrate). Mice were sacrificed at 22 weeks of age, except for NZO males that were sacrificed at 14 weeks, due to high mortality attributable to severe diabetes.

Reagents Collagenase Type XI (C7657), BSA (A4503), Ficoll Type 400-DL (F9378), FBS (12306C), dopamine (H8502), L-DOPA (D9628), ascorbic acid (A5960), and all general chemicals were purchased from Sigma. Dextrose (D16) was purchased from Fisher Scientific. Hanks' balanced salt solution (HBSS) (14065056) and RPMI medium 1640 (11879-020) were from Thermo Scientific.

In vivo measurements Body weight was measured weekly beginning at 4 weeks of age. Blood was collected by retro-orbital bleed following a 4 hour fast (8am-noon) at 6, 10, and 14 weeks of age, and a 3 hour fast (5am-8am) at sacrifice (22 weeks of age), and used to measure plasma glucose, insulin, and triglyceride (TG). Glucose was measured by the glucose oxidase method using a commercially available kit (TR15221, Thermo Scientific). Insulin was measured by radioimmunoassay (RIA; SRI-13K, Millipore). TG was measured

using a commercially available kit (TR22421, Thermo Scientific). If plasma insulin was off the low end of the standard curve for the assay (some CAST male mice), the value of the lowest standard on the assay was reported (0.1 ng/mL). Beginning at 4 weeks of age, food intake was calculated by weighing the food remaining after one week, subtracting it from the amount fed, then dividing by number of mice per cage and days to get g/mouse/day.

Islet isolation For all experiments that include islet isolation, intact pancreatic islets were isolated from mice using a collagenase digestion procedure as previously described[81]. Islets were hand-picked and counted under a stereomicroscope to minimize contaminating acinar tissue.

Insulin and glucagon secretion measurements After isolation, islets were placed in recovery media (RPMI 1640, 11.1 mM glucose, Anti/Anti antibiotics, 10% FBS) for 2 hours at 37°C and 5% CO_2. All insulin secretion media (3.3 mM glucose (G3.3), 8.3 mM glucose (G8.3), 8.3 mM glucose plus 100 nM GLP-1 (G8.3 + GLP-1), 8.3 mM glucose plus 1.25 mM L-alanine, 2 mM L-glutamine, 0.5 mM L-leucine (G8.3 + AA), 16.7 mM glucose (G16.7), 3.3 mM glucose plus 40 mM KCl (G3.3 + KCl), and 16.7 mM glucose plus 0.5 mM palmitate (G16.7 + PA), was made in Krebs Ringer Buffer (KRB: 118.41 mM NaCl, 4.69 mM KCl, 1.18 mM $MgSO_4$, 1.18 mM KH_2PO_4, 2 mM $NaHCO_3$, 5 mM HEPES, 2.52 mM $CaCl_2$, pH 7.4) containing 0.5% BSA, G16.7 + PA), which contained 0.67% BSA from the PA that was conjugated to BSA. For each mouse, 50 average sized islets were transferred from the recovery media to a 35-mm

petri dish containing 3 mL pre-incubation media (KRB + 0.5% BSA + 3.3 mM glucose). The rest of the islets from each mouse were washed twice with PBS, snap frozen in liquid nitrogen, stored at -80°C, and then used for the whole-islet proteomics (see Whole-islet proteomics on islets from the eight CC founder strains). For NZO male mice, islets from 4 mice were pooled to have enough for the secretion measurements. There were not enough islets to conduct proteomics on the NZO males. Islets were returned to the 37°C incubator for a 45-minute pre-incubation period. 100 μL of each secretagogue incubation media was placed in six wells of a 96-well plate. At the end of the pre-incubation period, individual islets were transferred to individual wells containing the incubation media, alternating the transfer between all seven incubation conditions, to ensure similar sized islets were distributed between all seven secretion conditions. Half way through the islet transfers, three islets were placed in 1 mL acid EtOH for measuring insulin and glucagon content. At the end of the 45-minute incubation period, the media was transferred to a 96-well polypropylene storage plate. Media was frozen at -20°C until analyzed. Basal secretion of insulin (G3.3) was measured using the Sensitive Insulin RIA (Millipore, SRI-13K). Secretion for all other conditions, and insulin content, was determined by insulin ELISA (in-house developed using a pair of anti-insulin/proinsulin antibodies (clones D6C4 and D3E7-BT) purchased from Fitzgerald (10R-I136a and 61R-I136BBT) Glucagon RIA (Millipore, GL-32K) was used to measure glucagon content and secretion. If secreted glucagon was off the low end of the standard curve for the assay, the value of the lowest standard on the assay was reported (2 pg/islet). To generate **Fig. 4.2**, insulin secreted per islet was calculated three

different ways: as total insulin secreted per islet (not normalized to anything (panel 1), as insulin secreted per islet as fold over basal secretion (normalized to basal insulin secretion at 3.3mM glucose) (panel 2), and as insulin secreted per islet as a percent of insulin content per islet (normalized to insulin content per islet) (panel 3). For each mouse, the average value from the six individual average sized islets for each secretion condition (six technical replicates per condition) was calculated. Each data point (one colored square of the heat map in **Fig. 4.2**) was then calculated by averaging these values among the same strain/sex combination. Each strain/sex combination had $n \geqslant 3$ mice.

Whole-islet proteomics on islets from the eight CC founder strains

Proteomic sample preparation After islets from each mouse were isolated and allowed to recover for 2 hours in recovery media (see Insulin and glucagon secretion measurements), 50 islets from each mouse were used for the secretion measurements and the rest of the islets were washed twice with PBS, snap frozen in liquid nitrogen, stored at -80°C, and then used for the whole-islet proteomics. Islets from each mouse were lysed by boiling in 6 M guanidine, and protein concentration was determined using the Pierce BCA Protein Assay Kit (Thermo). 50 μg of protein was aliquoted from each sample, precipitated with 90% methanol, mixed, and centrifuged at 12,000 x g for 5 mins. The supernatants were discarded, and the protein pellets resuspended in 8 M urea, 10 mM tris(2-carboxyethyl)phosphine (TCEP), 40 mM chloroacetamide (CAA) and 100 mM tris (pH = 8). Lysates were diluted to

1.6 M urea with 50 mM tris (pH = 8), and digested overnight at room temperature with trypsin (Promega) at a ratio of 1:50 enzyme to protein. Samples were desalted using Strata X columns (Phenomenex Strata-X Polymeric Reversed Phase 10 mg/mL). Desalting columns were equilibrated with 1 mL 100% acetonitrile (ACN) followed by 1 mL of 0.2% formic acid. Samples were acidified with TFA and loaded onto the equilibrated Strata X columns, which were then washed with 1 mL 0.2% formic acid. Peptides were eluted into clean tubes with 1 mL 80% ACN, dried, and reconstituted in 0.2% formic acid. Peptide concentration was measured prior to MS analysis using the Pierce Quantitative Colorimetric Peptide Assay (Thermo).

LC-MS/MS analysis 2 μg of islet peptides were loaded onto a reversed phase nano-LC column for chromatographic separation prior to MS analysis. Columns are prepared in-house from 35 cm of 75 μm ID, 360 μm OD fused-silica capillary tubing with polyimide coating with a laser-pulled electrospray tip. They were packed with 1.7 μm diameter, 130 Å pore size, bridged ethylene hybrid C18 particles (Waters). Columns were fitted onto an UltiMate 3000 UHPLC system (Thermo) and heated to 55°C using a home-built column heater. Mobile phase buffer A was composed of 0.2% formic acid. Mobile phase B was composed of 70% ACN, 0.2% formic acid. Samples were separated over a 120-minute gradient, including time for column re-equilibration. Flow rates were set at 325 nL/min. Peptide cations were converted to gas-phase ions by electrospray ionization and analyzed on an Orbitrap Fusion Lumos (Q-OT-qIT, Thermo). Precursor scans were collected from

300 to 1,350 *m/z* at 60k resolution (at 400 *m/z*) using a 1e6 AGC target. Precursors selected for MS/MS analysis were isolated at 0.7 Th with the quadrupole mass filter, fragmented by HCD with a collision energy of 25. The maximum injection time for MS/MS analysis was 15 ms with an AGC target of 3e4. Only precursors from charge state 2-8 were selected. Dynamic exclusion time was set to 5 s, with a mass tolerance of 25 ppm. Analyses were performed in top-speed mode with a cycle time of 2 s.

Hierarchical clustering MaxLFQ proteinGroups output was further processed using the Perseus software platform (Version 1.5.6.0)[82]. Label-free quantification (LFQ) values were log2 transformed. Z-scores were calculated for each protein across all samples ($(x - \mu)/\sigma$), and the data was hierarchically clustered in an unsupervised manner using Pearson's correlations as the distance metric. In the resulting dendrogram, clusters were defined using a distance threshold of 0.82. Uniprot accession numbers from each cluster were used for gene ontology enrichment analysis via the DAVID tool; reported p-values have been corrected for multiple tests using the Benjamini-Hochberg method.

Generating co-expression modules We used a previously developed method to identify protein co-expression modules (weighted correlation network analysis or WGCNA)[19,20]. An extensive overview of WGCNA, including numerous tutorials, can be found at Prof. Steve Horvath's UCLA genetics department website[83] Since proteomics were conducted on 3 or more mice from each strain/sex combination, any protein that was detected in 2 or

fewer mice/samples was considered not included in the module calculation. This resulted in the exclusion of 27 proteins out of the 5,255 identified. For those proteins that were detected in 3 or more samples (5,228 proteins), any samples where it was not detected we entered a zero value, followed by rank transformation of all values. An adjacency matrix was constructed for these proteins. Each entry in the matrix is the absolute Pearson's correlation, adjusted so that the overall network is approximately scale-free. Connection strength between two proteins (x_i and x_j) in the network was determined according to the adjacency function, $a_{ij} = |0.5 + 0.5 \times cor(x_i, x_j)|^{\beta}$, using the estimated power parameter β of 12, resulting in a weighted network[19,84]. This yields a "signed" co-expression network that preserves the directionality of the correlation between the protein pairs, yielding values that range from 0 to 1. We note that this allows for all correlations to be used, unlike approaches that invoke arbitrary thresholds. For a discussion of the advantage of weighted versus unweighted networks, see[19] and references therein. For the WGCNA, suggestions in Prof. Horvath's tutorial were followed, with a power parameter of 12, Pearson correlation, and signed modules[85]. The minimum number of proteins to make up a module was set at 30.

Calculating correlations between MEs and physiological traits and GO/KEGG enrichment The proteins were clustered into modules by color and the ME was calculated as the first principle component for the proteins in the module[19]. The first principle component estimate for each module was then used along with Pearson's correlation to correlate the modules with the clinical traits. Normalized ranks of the clinical trait values were used

when calculating the correlation. We used a previously developed method for GO/KEGG enrichment of co-expression modules[86].

Immunohistochemistry Immunohistochemistry was performed on 40 islets in pancreas sections from each of three B6, CAST, and PWK male mice, which were 20 weeks of age and maintained on the HF/HS diet. Mice were euthanized by CO_2 asphyxiation, perfused with 4% paraformaldehyde through the heart, and the pancreas removed, embedded in paraffin, and sectioned. Briefly, paraffin-embedded pancreas sections were de-waxed in xylenes, rehydrated in decreasing percentages of ethanol, and boiled in antigen retrieval solution (VectorLabs, H3300). After the sections were cooled and washed with PBS, sections were blocked with 10% normal donkey serum in PBS for 1 hour at RT. Primary antibody solution in 1% normal donkey serum in PBS was incubated overnight at 4°C. After a PBS wash, secondary antibody solution in 1% normal donkey serum in PBS was incubated 1 hour at RT. After a PBS wash, slides were allow to dry and mounted in mounting media (VectorLabs, H-1000). Primary antibodies used were: polyclonal guinea pig anti-insulin (Agilent, A056401-2), monoclonal mouse anti-glucagon antibody (Sigma, G2654), and polyclonal rabbit anti-tyrosine hydroxylase (Millipore, AB152). Secondary antibodies used were: goat anti-rabbit AlexaFluor 488 (Thermo Scientific, A-11008), chicken anti-mouse AlexaFluor 647 (Thermo Scientific, A-21463), and donkey anti-guinea pig Cy3 (Jackson ImmunoResearch, 706-165-148). All primary and secondary antibodies were used at a dilution of 1:500. DAPI was added to the secondary antibody solution at a concentration of

1.3 μg/mL to view nuclei. Images were acquired on a Nikon A1R+ point scanning confocal system that uses photomultiplying tubes at room temperature with a Nikon 40X Pan Apo oil immersion lens with a numerical aperture of 1.3. Acquisition software is NIS-Elements Ar. Post-processed using ImageJ software by adjusting the brightness and contrast. The area of each islet was measured, and the number of Th+/insulin+/glucagon- (Th+ β-cells), Th+/insulin-/glucagon+ (Th+ α-cells), and Th+/insulin-/glucagon- (Th+ unidentified cells) cells were counted. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7.

Measuring dopamine-related metabolites in B6 and CAST islets 100 islets each from four B6 and four CAST male mice (20 weeks of age on HF/HS diet) were recovered in culture media (RPMI 1640, 1.7 mM glucose, 10% FBS) for 2 hours at 37°C and 5% CO_2. Islets were then washed twice with PBS and frozen as a pellet in liquid nitrogen for storage at -80°C. Metabolites were extracted from the frozen islets by the addition of 50 μL of 80% (v/v) ice cold acetonitrile, followed by sonication. The mixture was centrifuged for 5 min at 12,100 x g. The supernatant was derivatized with benzoyl chloride as previously described[87]. Briefly, by sequential addition of 10 μL 100 mM sodium carbonate, 10 μL 2% (v/v) benzoyl chloride in acetonitrile, and 10 μL of the internal standard solution. The resulting solution was diluted with 50 μL of water. The internal standard solution consisted of metabolites derivatized with $^{13}C_6$ benzoyl chloride in 20% (v/v) acetonitrile with 1% (v/v) sulfuric acid. Protein content was determined using a Pierce BCA Protein Assay kit (Thermo Fisher

Scientific, Walther, MA), and metabolite concentrations were normalized to protein content. Calibration standards were prepared in water, diluted in acetonitrile to match the sample composition, and derivatized. Samples were analyzed using a Waters nanoAcquity UPLC coupled to an Agilent 6410B triple quadrupole mass spectrometer. An Acquity HSS T3 C18 (1 mm x 100 mm, 1.8 μm, 100 Å pore size) column was used and the injection volume was 5 μl. Mobile phase A was 10 mM ammonium formate with 0.15% formic acid. Mobile phase B was acetonitrile. The flow rate was 100 μL/min, and the gradient used was: initial, 0% B; 0.1 min, 17% B; 0.5 min, 17% B; 3 min, 25% B; 3.3 min, 56% B; 4.9 min, 70% B; 5 min, 100% B; 6 min, 100% B; 6.1 min, 0% B; 8 min, 0% B. Electrospray ionization was used in positive mode and the capillary was at 4 kV. The nebulizer pressure was 15 psi, the drying gas was at 11 L/min, and the gas temperature was 350°C. Detection was performed in dynamic MRM mode, and the MRM conditions are listed in Table S5. Automated peak integration was performed with Agilent MassHunter Workstation Quantitative Analysis for QQQ, version B.05.00. All peaks were inspected to ensure proper integration. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7.

B6 and CAST islet insulin secretion Islets from four B6 and four CAST male mice (20 weeks of age on HF/HS diet) were recovered in culture media (RPMI 1640, 3.3 mM glucose, 10% FBS) at 37°C in 5% CO_2 for 2 hours. Islets were then pre-incubated in 95% O_2/5% CO_2-gassed KRB (118.41 mM NaCl, 4.69 mM KCl, 1.18 mM $MgSO_4$, 1.18 mM KH_2PO_4, 25 mM $NaHCO_2$, 5 mM HEPES, 2.52 mM $CaCl_2$) supplemented with 0.5% BSA and 3.3

mM glucose (B6 and CAST), or 3.3 mM glucose plus 50 μM L-DOPA (B6+L-DOPA) and 100 μM ascorbic acid for 45 minutes at 37°C in 5% CO_2. After pre-incubation, 15 islets were transferred to an Eppendorf tube containing 125 μL of secretion media (KRB with 0.5% BSA and 3.3 mM glucose or 16.7 mM glucose with or without 1 μM dopamine) and placed in a 37°C water bath for 45 minutes. After the secretion, the secretion media was removed from the islets and the islets were lysed in NP-40 lysis buffer (100 mM Tris, pH 8.0, 300 mM NaCl, 10 mM NaF, 2 mM Na_3VO_4, 2% NP-40 alternative, cOmplete mini EDTA-free protease inhibitor cocktail (Roche)). Insulin in the media, and lysates was measured using an in-house developed insulin ELISA[25]. Statistics were performed using unpaired, parametric, two-tailed t-tests in GraphPad Prism 7.

References

[1] A. S. Dimas, V. Lagou, A. Barker, J. W. Knowles, R. Magi, M.-F. Hivert, A. Benazzo, D. Rybin, A. U. Jackson, H. M. Stringham, C. Song, A. Fischer-Rosinsky, T. W. Boesgaard, N. Grarup, F. A. Abbasi, T. L. Assimes, K. Hao, X. Yang, C. Lecoeur, I. Barroso, L. L. Bonnycastle, Y. Bottcher, S. Bumpstead, P. S. Chines, M. R. Erdos, J. Graessler, P. Kovacs, M. A. Morken, N. Narisu, F. Payne, A. Stancakova, A. J. Swift, A. Tonjes, S. R. Bornstein, S. Cauchi, P. Froguel, D. Meyre, P. E. H. Schwarz, H.-U. Haring, U. Smith, M. Boehnke, R. N. Bergman, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. Quertemous, L. Lind, T. Hansen, O. Pedersen, M. Walker, A. F. H. Pfeiffer, J. Spranger, M. Stumvoll, J. B. Meigs, N. J. Wareham, J. Kuusisto, M. Laakso, C. Langenberg, J. Dupuis, R. M. Watanabe,

J. C. Florez, E. Ingelsson, M. I. McCarthy, I. Prokopenko, and MAGIC Investigators, "Impact of Type 2 Diabetes Susceptibility Variants on Quantitative Glycemic Traits Reveals Mechanistic Heterogeneity," *Diabetes*, vol. 63, pp. 2158–2171, Jun 2014.

[2] C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Müller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikram, B.-J. Kim, Y. Kim, Y. J. Kim, M.-S. Kwon, J. Lee, S. Lee, K.-H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B.-G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stančáková, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S.-H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C.-Y. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farook, S. P. Fowler, B. I. Freedman, M. Gris-

wold, D. E. Hale, P. J. Hicks, C.-C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S.-T. Tan, H. A. Taylor, F. Thameem, G. Wilson, T. Y. Wong, P. R. Njølstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jørgensen, T. Jørgensen, C. Ladenvall, J. M. Justesen, A. Käräjämäki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. Hrabé de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O'Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A.-C. Syvänen, R. N. Bergman, D. Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Mägi, L. Lind, Y. Farjoun, K. R.

Owen, A. L. Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J.-Y. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M. Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burtt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler, and M. I. McCarthy, "The genetic architecture of type 2 diabetes," *Nature*, vol. 536, pp. 41–47, Aug 2016.

[3] L. Marullo, J. S. El-Sayed Moustafa, and I. Prokopenko, "Insights into the Genetic Susceptibility to Type 2 Diabetes from Genome-Wide Association Studies of Glycaemic Traits," *Current Diabetes Reports*, vol. 14, p. 551, Nov 2014.

[4] R. A. Scott, L. J. Scott, R. Mägi, L. Marullo, K. J. Gaulton, M. Kaakinen, N. Pervjakova, T. H. Pers, A. D. Johnson, J. D. Eicher, A. U. Jackson, T. Ferreira, Y. Lee, C. Ma, V. Steinthorsdottir, G. Thorleifsson, L. Qi, N. R. Van Zuydam, A. Mahajan, H. Chen, P. Almgren, B. F. Voight, H. Grallert, M. Müller-Nurasyid, J. S. Ried, N. W. Rayner, N. Robertson, L. C. Karssen, E. M. van Leeuwen, S. M. Willems, C. Fuchsberger, P. Kwan, T. M. Teslovich, P. Chanda, M. Li, Y. Lu, C. Dina, D. Thuillier, L. Yengo, L. Jiang, T. Sparso, H. A. Kestler, H. Chheda, L. Eisele, S. Gustafsson, M. Frånberg, R. J. Strawbridge, R. Benediktsson, A. B. Hreidarsson, A. Kong, G. Sigurðsson, N. D. Kerrison, J. Luan, L. Liang, T. Meitinger, M. Roden, B. Thorand, T. Esko, E. Mihailov,

C. Fox, C.-T. Liu, D. Rybin, B. Isomaa, V. Lyssenko, T. Tuomi, D. J. Couper, J. S. Pankow, N. Grarup, C. T. Have, M. E. Jørgensen, T. Jørgensen, A. Linneberg, M. C. Cornelis, R. M. van Dam, D. J. Hunter, P. Kraft, Q. Sun, S. Edkins, K. R. Owen, J. R. Perry, A. R. Wood, E. Zeggini, J. Tajes-Fernandes, G. R. Abecasis, L. L. Bonnycastle, P. S. Chines, H. M. Stringham, H. A. Koistinen, L. Kinnunen, B. Sennblad, T. W. Mühleisen, M. M. Nöthen, S. Pechlivanis, D. Baldassarre, K. Gertow, S. E. Humphries, E. Tremoli, N. Klopp, J. Meyer, G. Steinbach, R. Wennauer, J. G. Eriksson, S. Mannistö, L. Peltonen, E. Tikkanen, G. Charpentier, E. Eury, S. Lobbens, B. Gigante, K. Leander, O. McLeod, E. P. Bottinger, O. Gottesman, D. Ruderfer, M. Blüher, P. Kovacs, A. Tonjes, N. M. Maruthur, C. Scapoli, R. Erbel, K.-H. Jöckel, S. Moebus, U. de Faire, A. Hamsten, M. Stumvoll, P. Deloukas, P. J. Donnelly, T. M. Frayling, A. T. Hattersley, S. Ripatti, V. Salomaa, N. L. Pedersen, B. O. Boehm, R. N. Bergman, F. S. Collins, K. L. Mohlke, J. Tuomilehto, T. Hansen, O. Pedersen, I. Barroso, L. Lannfelt, E. Ingelsson, L. Lind, C. M. Lindgren, S. Cauchi, P. Froguel, R. J. Loos, B. Balkau, H. Boeing, P. W. Franks, A. Barricarte Gurrea, D. Palli, Y. T. van der Schouw, D. Altshuler, L. C. Groop, C. Langenberg, N. J. Wareham, E. Sijbrands, C. M. van Duijn, J. C. Florez, J. B. Meigs, E. Boerwinkle, C. Gieger, K. Strauch, A. Metspalu, A. D. Morris, C. N. Palmer, F. B. Hu, U. Thorsteinsdottir, K. Stefansson, J. Dupuis, A. P. Morris, M. Boehnke, M. I. McCarthy, I. Prokopenko, and DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, "An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans," *Diabetes*, vol. 66, pp. 2888–2902, Nov 2017.

[5] A. R. Wood, A. Jonsson, A. U. Jackson, N. Wang, N. van Leewen, N. D. Palmer, S. Kobes, J. Deelen, L. Boquete-Vilarino, J. Paananen, A. Stančáková, D. I. Boomsma, E. J. de Geus, E. M. Eekhoff, A. Fritsche, M. Kramer, G. Nijpels, A. Simonis-Bik, T. W. van Haeften, A. Mahajan, M. Boehnke, R. N. Bergman, J. Tuomilehto, F. S. Collins, K. L. Mohlke, K. Banasik, C. J. Groves, M. I. McCarthy, E. R. Pearson, A. Natali, A. Mari, T. A. Buchanan, K. D. Taylor, A. H. Xiang, A. P. Gjesing, N. Grarup, H. Eiberg, O. Pedersen, Y.-D. Chen, M. Laakso, J. M. Norris, U. Smith, L. E. Wagenknecht, L. Baier, D. W. Bowden, T. Hansen, M. Walker, R. M. Watanabe, L. M. 't Hart, R. L. Hanson, T. M. Frayling, and T. M. Frayling, "A Genome-Wide Association Study of IVGTT-Based Measures of First-Phase Insulin Secretion Refines the Underlying Physiology of Type 2 Diabetes Variants," *Diabetes*, vol. 66, pp. 2296–2309, Aug 2017.

[6] A. D. Attie, G. A. Churchill, and J. H. Nadeau, "How mice are indispensable for understanding obesity and diabetes genetics," *Current Opinion in Endocrinology & Diabetes and Obesity*, vol. 24, pp. 83–91, Apr 2017.

[7] G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe, W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann,

K. Hunter, H.-C. Hsu, F. A. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb, R. F. Kooy, B. Llamas, F. Lammert, J.-M. Lassalle, P. R. Lowenstein, L. Lu, A. Lusis, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, B. Paigen, K. Paigen, A. A. Palmer, H.-J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. Sillanpää, L. D. Siracusa, H.-W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. Threadgill, L. A. Toth, W. Valdar, F. P.-M. de Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire, N. Yi, D. Zhang, M. Zhang, F. Zou, and Complex Trait Consortium, "The Collaborative Cross, a community resource for the genetic analysis of complex traits," *Nature Genetics*, vol. 36, pp. 1133–1137, Nov 2004.

[8] J. H. Kreznar, M. P. Keller, L. L. Traeger, M. E. Rabaglia, K. L. Schueler, D. S. Stapleton, W. Zhao, E. I. Vivas, B. S. Yandell, A. T. Broman, B. Hagenbuch, A. D. Attie, and F. E. Rey, "Host Genotype and Gut Microbiome Modulate Insulin Secretion and Diet-Induced Metabolic Phenotypes," *Cell Reports*, vol. 18, pp. 1739–1750, Feb 2017.

[9] M. Komatsu, M. Takei, H. Ishii, and Y. Sato, "Glucose-stimulated insulin secretion: A newer perspective.," *Journal of diabetes investigation*, vol. 4, pp. 511–6, Nov 2013.

[10] J. M. Baughman, C. M. Rose, G. Kolumam, J. D. Webster, E. M. Wilkerson, A. E. Merrill, T. W. Rhoads, R. Noubade, P. Katavolos, J. Lesch, D. S. Stapleton, M. E. Rabaglia, K. L. Schueler, R. Asuncion, M. Domeyer, J. Zavala-Solorio, M. Reich, J. DeVoss, M. P. Keller, A. D. Attie, A. S. Hebert, M. S. Westphall, J. J. Coon, D. S. Kirkpatrick, and A. Dey, "NeuCode Proteomics Reveals Bap1 Regulation of Metabolism," *Cell Reports*, vol. 16, pp. 583–595, Jul 2016.

[11] K. E. Dittenhafer-Reed, A. L. Richards, J. Fan, M. J. Smallegan, A. Fotuhi Siahpirani, Z. A. Kemmerer, T. A. Prolla, S. Roy, J. J. Coon, and J. M. Denu, "SIRT3 Mediates Multi-Tissue Coupling for Metabolic Fuel Switching," *Cell Metabolism*, vol. 21, pp. 637–646, Apr 2015.

[12] B. J. Floyd, E. M. Wilkerson, M. T. Veling, C. E. Minogue, C. Xia, E. T. Beebe, R. L. Wrobel, H. Cho, L. S. Kremer, C. L. Alston, K. A. Gromek, B. K. Dolan, A. Ulbrich, J. A. Stefely, S. L. Bohl, K. M. Werner, A. Jochem, M. S. Westphall, J. W. Rensvold, R. W. Taylor, H. Prokisch, J.-J. P. Kim, J. J. Coon, and D. J. Pagliarini, "Mitochondrial Protein Interaction Mapping Identifies Regulators of Respiratory Chain Function," *Molecular Cell*, vol. 63, pp. 621–632, Aug 2016.

[13] J. L. Horton, O. J. Martin, L. Lai, N. M. Riley, A. L. Richards, R. B. Vega, T. C. Leone, D. J. Pagliarini, D. M. Muoio, K. C. Bedi, K. B. Margulies, J. J. Coon, D. P. Kelly, and D. P. Kelly, "Mitochondrial protein hyperacetylation in the failing heart.," *JCI insight*, vol. 2, Feb 2016.

[14] K. A. Overmyer, C. R. Evans, N. R. Qi, C. E. Minogue, J. J. Carson, C. J. Chermside-Scabbo, L. G. Koch, S. L. Britton, D. J. Pagliarini, J. J. Coon, and C. F. Burant, "Maximal Oxidative Capacity during Exercise Is Associated with Skeletal Muscle Fuel Selection and Dynamic Changes in Mitochondrial Protein Acetylation," *Cell Metabolism*, vol. 21, pp. 468–478, Mar 2015.

[15] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, pp. 701–714, Apr 2015.

[16] N. M. Riley, A. S. Hebert, and J. J. Coon, "Proteomics Moves into the Fast Lane," *Cell Systems*, vol. 2, pp. 142–143, Mar 2016.

[17] E. Shishkova, A. S. Hebert, and J. J. Coon, "Now, More Than Ever, Proteomics Needs Better Chromatography," *Cell Systems*, vol. 3, pp. 321–324, Oct 2016.

[18] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, pp. 1191–1197, Nov 2016.

[19] B. Zhang and S. Horvath, "A General Framework for Weighted Gene Co-Expression

Network Analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, p. Article17, Jan 2005.

[20] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, Dec 2008.

[21] M. Carlson, B. Zhang, Z. Fang, P. Mischel, S. Horvath, and S. Nelson, "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks," *BMC Genomics*, vol. 7, p. 40, Mar 2006.

[22] P. S. Gargalovic, M. Imura, B. Zhang, N. M. Gharavi, M. J. Clark, J. Pagnon, W.-P. Yang, A. He, A. Truong, S. Patel, S. F. Nelson, S. Horvath, J. A. Berliner, T. G. Kirchgessner, and A. J. Lusis, "Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 12741–12746, Aug 2006.

[23] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis, and S. Horvath, "Integrating Genetic and Network Analysis to Characterize Genes Related to Mouse Weight," *PLoS Genetics*, vol. 2, no. 8, p. e130, 2006.

[24] S. Horvath, B. Zhang, M. Carlson, K. V. Lu, S. Zhu, R. M. Felciano, M. F. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A. C. Scheck, L. M. Liau, H. Wu, D. H. Geschwind, P. G. Febbo, H. I. Kornblum, T. F. Cloughesy, S. F. Nelson, and P. S. Mischel, "Analysis

of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target," *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17402–17407, 2006.

[25] M. P. Keller, Y. Choi, P. Wang, D. Belt Davis, M. E. Rabaglia, A. T. Oler, D. S. Stapleton, C. Argmann, K. L. Schueler, S. Edwards, H. A. Steinberg, E. Chaibub Neto, R. Kleinhanz, S. Turner, M. K. Hellerstein, E. E. Schadt, B. S. Yandell, C. Kendziorski, and A. D. Attie, "A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility," *Genome Research*, vol. 18, pp. 706–716, Feb 2008.

[26] I. Tekin, R. Roskoski, N. Carkaci-Salli, and K. E. Vrana, "Complex molecular regulation of tyrosine hydroxylase," *Journal of Neural Transmission*, vol. 121, pp. 1451–1481, Dec 2014.

[27] J. W. Haycock, N. G. Ahn, M. H. Cobb, and E. G. Krebs, "ERK1 and ERK2, two microtubule-associated protein 2 kinases, mediate the phosphorylation of tyrosine hydroxylase at serine-31 in situ.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, pp. 2365–9, Mar 1992.

[28] P. R. Dunkley, L. Bobrovskaya, M. E. Graham, E. I. Von Nagy-Felsobuki, and P. W. Dickson, "Tyrosine hydroxylase phosphorylation: regulation and consequences," *Journal of Neurochemistry*, vol. 91, pp. 1025–1043, Oct 2004.

[29] B. Rubí, S. Ljubicic, S. Pournourmohammadi, S. Carobbio, M. Armanet, C. Bartley, and P. Maechler, "Dopamine D2-like Receptors Are Expressed in Pancreatic Beta Cells

and Mediate Inhibition of Insulin Secretion," *Journal of Biological Chemistry*, vol. 280, pp. 36824–36832, Nov 2005.

[30] J. M. Feldman and H. E. Lebovitz, "Mechanism of epinephrine and serotonin inhibition of insulin release in the golden hamster in vitro.," *Diabetes*, vol. 19, pp. 480–6, Jul 1970.

[31] R. L. Sorenson, R. P. Elde, and V. Seybold, "Effect of norepinephrine on insulin, glucagon, and somatostatin secretion in isolated perifused rat islets.," *Diabetes*, vol. 28, pp. 899–904, Oct 1979.

[32] F. Esni, I. B. Täljedal, A. K. Perl, H. Cremer, G. Christofori, and H. Semb, "Neural cell adhesion molecule (N-CAM) is required for cell type segregation and normal ultrastructure in pancreatic islets.," *The Journal of cell biology*, vol. 144, pp. 325–37, Jan 1999.

[33] F. Reichmann and P. Holzer, "Neuropeptide Y: A stressful review," *Neuropeptides*, vol. 55, pp. 99–109, Feb 2016.

[34] T. A. Schwetz, A. Ustione, and D. W. Piston, "Neuropeptide Y and somatostatin inhibit insulin secretion through different mechanisms," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 304, pp. E211–E221, Jan 2013.

[35] T. Kageyama, M. Nakamura, A. Matsuo, Y. Yamasaki, Y. Takakura, M. Hashida, Y. Kanai, M. Naito, T. Tsuruo, N. Minato, and S. Shimohama, "The 4F2hc/LAT1

complex transports L-DOPA across the blood-brain barrier.," *Brain research*, vol. 879, pp. 115–21, Oct 2000.

[36] A. Ustione, D. W. Piston, and P. E. Harris, "Minireview: Dopaminergic regulation of insulin secretion from the pancreatic islet.," *Molecular endocrinology (Baltimore, Md.)*, vol. 27, pp. 1198–207, Aug 2013.

[37] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, pp. 339–347, jan 2014.

[38] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, pp. 500–505, Jun 2016.

[39] S. H. Back, S.-W. Kang, J. Han, and H.-T. Chung, "Endoplasmic Reticulum Stress in the β -Cell Pathogenesis of Type 2 Diabetes," *Experimental Diabetes Research*, vol. 2012, pp. 1–11, 2012.

[40] S. H. Back and R. J. Kaufman, "Endoplasmic Reticulum Stress and Type 2 Diabetes," *Annual Review of Biochemistry*, vol. 81, pp. 767–793, Jul 2012.

[41] M.-K. Kim, H.-S. Kim, I.-K. Lee, and K.-G. Park, "Endoplasmic Reticulum Stress and Insulin Biosynthesis: A Review," *Experimental Diabetes Research*, vol. 2012, pp. 1–7, Mar 2012.

[42] S. Z. Hasnain, J. B. Prins, and M. A. McGuckin, "Oxidative and endoplasmic reticulum stress in β-cell dysfunction in diabetes.," *Journal of molecular endocrinology*, vol. 56, pp. R33–54, Feb 2016.

[43] A. El Ouaamari, J.-Y. Zhou, C. W. Liew, J. Shirakawa, E. Dirice, N. Gedeon, S. Kahraman, D. F. De Jesus, S. Bhatt, J.-S. Kim, T. R. W. Clauss, D. G. Camp, R. D. Smith, W.-J. Qian, and R. N. Kulkarni, "Compensatory Islet Response to Insulin Resistance Revealed by Quantitative Proteomics," *Journal of Proteome Research*, vol. 14, pp. 3111–3122, Aug 2015.

[44] O. Omikorede, C. Qi, T. Gorman, P. Chapman, A. Yu, D. M. Smith, and T. P. Herbert, "ER stress in rodent islets of Langerhans is concomitant with obesity and β-cell compensation but not with β-cell dysfunction and diabetes.," *Nutrition & diabetes*, vol. 3, p. e93, Oct 2013.

[45] R. Roat, V. Rao, N. M. Doliba, F. M. Matschinsky, J. W. Tobias, E. Garcia, R. S. Ahima, and Y. Imai, "Alterations of Pancreatic Islet Structure, Metabolism and Gene Expression in Diet-Induced Obese C57BL/6J Mice," *PLoS ONE*, vol. 9, p. e86815, Feb 2014.

[46] V. Cirulli, "Cadherins in islet β-cells: more than meets the eye.," *Diabetes*, vol. 64, pp. 709–11, Mar 2015.

[47] J. K. Johansson, U. Voss, G. Kesavan, I. Kostetskii, N. Wierup, G. L. Radice, and H. Semb,

"N-cadherin is dispensable for pancreas development but required for β-cell granule turnover," *genesis*, vol. 48, pp. 374–381, Apr 2010.

[48] A. C. Hauge-Evans, P. E. Squires, S. J. Persaud, and P. M. Jones, "Pancreatic beta-cell-to-beta-cell interactions are required for integrated responses to nutrient stimuli: enhanced Ca2+ and insulin secretory responses of MIN6 pseudoislets.," *Diabetes*, vol. 48, pp. 1402–8, Jul 1999.

[49] G. Parnaud, V. Lavallard, B. Bedat, D. Matthey-Doret, P. Morel, T. Berney, and D. Bosco, "Cadherin Engagement Improves Insulin Secretion of Single Human β-Cells," *Diabetes*, vol. 64, pp. 887–896, Mar 2015.

[50] G. J. Rogers, M. N. Hodgkin, and P. E. Squires, "E-Cadherin and Cell Adhesion: a Role in Architecture and Function in the Pancreatic Islet," *Cellular Physiology and Biochemistry*, vol. 20, no. 6, pp. 987–994, 2007.

[51] D. Rondas, A. Tomas, M. Soto-Ribeiro, B. Wehrle-Haller, and P. A. Halban, "Novel mechanistic link between focal adhesion remodeling and glucose-stimulated insulin secretion.," *The Journal of biological chemistry*, vol. 287, pp. 2423–36, jan 2012.

[52] C. L. Rackham, A. E. Vargas, R. G. Hawkes, S. Amisten, S. J. Persaud, A. L. Austin, A. J. King, and P. M. Jones, "Annexin A1 is a key modulator of Mesenchymal Stromal Cell mediated improvements in islet function," *Diabetes*, vol. 65, p. db150990, Oct 2015.

[53] P. A. Antinozzi, H. Ishihara, C. B. Newgard, and C. B. Wollheim, "Mitochondrial metabolism sets the maximal limit of fuel-stimulated insulin secretion in a model pancreatic beta cell: a survey of four fuel secretagogues.," *The Journal of biological chemistry*, vol. 277, pp. 11746–55, Apr 2002.

[54] S. Malmgren, D. G. Nicholls, J. Taneera, K. Bacos, T. Koeck, A. Tamaddon, R. Wibom, L. Groop, C. Ling, H. Mulder, and V. V. Sharoyko, "Tight Coupling between Glucose and Mitochondrial Metabolism in Clonal β-Cells Is Required for Robust Insulin Secretion," *Journal of Biological Chemistry*, vol. 284, pp. 32395–32404, Nov 2009.

[55] A. Wiederkehr and C. B. Wollheim, "Mitochondrial signals drive insulin secretion in the pancreatic β-cell," *Molecular and Cellular Endocrinology*, vol. 353, pp. 128–137, Apr 2012.

[56] D. Baetens, F. Malaisse-Lagae, A. Perrelet, and L. Orci, "Endocrine pancreas: three-dimensional reconstruction shows two types of islets of langerhans.," *Science (New York, N.Y.)*, vol. 206, pp. 1323–5, Dec 1979.

[57] G. Wieczorek, A. Pospischil, and E. Perentes, "A comparative immunohistochemical study of pancreatic islets inlaboratory animals (rats, dogs, minipigs, nonhuman primates)," *Experimental and Toxicologic Pathology*, vol. 50, pp. 151–172, jan 1998.

[58] E. Bader, A. Migliorini, M. Gegg, N. Moruzzi, J. Gerdes, S. S. Roscioni, M. Bakhti, E. Brandl, M. Irmler, J. Beckers, M. Aichler, A. Feuchtinger, C. Leitzinger, H. Zis-

chka, R. Wang-Sattler, M. Jastroch, M. Tschöp, F. Machicao, H. Staiger, H.-U. Häring, H. Chmelova, J. A. Chouinard, N. Oskolkov, O. Korsgren, S. Speier, and H. Lickert, "Identification of proliferative and mature β-cells in the islets of Langerhans," *Nature*, vol. 535, pp. 430–434, Jul 2016.

[59] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai, "A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure.," *Cell systems*, vol. 3, pp. 346–360.e4, Oct 2016.

[60] C. Dorrell, J. Schug, P. S. Canaday, H. A. Russ, B. D. Tarlow, M. T. Grompe, T. Horton, M. Hebrok, P. R. Streeter, K. H. Kaestner, and M. Grompe, "Human islets contain four distinct subtypes of β cells," *Nature Communications*, vol. 7, p. 11756, Jul 2016.

[61] N. R. Johnston, R. K. Mitchell, E. Haythorne, M. P. Pessoa, F. Semplici, J. Ferrer, L. Piemonti, P. Marchetti, M. Bugliani, D. Bosco, E. Berishvili, P. Duncanson, M. Watkinson, J. Broichhagen, D. Trauner, G. A. Rutter, and D. J. Hodson, "Beta Cell Hubs Dictate Pancreatic Islet Responses to Glucose," *Cell Metabolism*, vol. 24, pp. 389–401, Sep 2016.

[62] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. Ämmälä, and R. Sandberg, "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes.," *Cell metabolism*, vol. 24, pp. 593–607, Oct 2016.

[63] Y. J. Wang, M. L. Golson, J. Schug, D. Traum, C. Liu, K. Vivek, C. Dorrell, A. Naji, A. C. Powers, K.-M. Chang, M. Grompe, and K. H. Kaestner, "Single-Cell Mass Cytometry Analysis of the Human Endocrine Pancreas," *Cell Metabolism*, vol. 24, pp. 616–626, Oct 2016.

[64] C. Cruciani-Guglielmacci, L. Bellini, J. Denom, M. Oshima, N. Fernandez, P. Normandie-Levi, X. P. Berney, N. Kassis, C. Rouch, J. Dairou, T. Gorman, D. M. Smith, A. Marley, R. Liechti, D. Kuznetsov, L. Wigger, F. Burdet, A.-L. Lefèvre, I. Wehrle, I. Uphues, T. Hildebrandt, W. Rust, C. Bernard, A. Ktorza, G. A. Rutter, R. Scharfmann, I. Xenarios, H. Le Stunff, B. Thorens, C. Magnan, and M. Ibberson, "Molecular phenotyping of multiple mouse strains under metabolic challenge uncovers a role for Elovl2 in glucose-induced insulin secretion," *Molecular Metabolism*, vol. 6, pp. 340–351, Apr 2017.

[65] L. Cegrell, "The occurrence of biogenic monoamines in the mammalian endocrine pancreas.," *Acta physiologica Scandinavica. Supplementum*, vol. 314, pp. 1–60, 1968.

[66] L. E. Ericson, R. Håkanson, and I. Lundquist, "Accumulation of dopamine in mouse pancreatic B-cells following injection of L-DOPA. Localization to secretory granules and inhibition of insulin secretion.," *Diabetologia*, vol. 13, pp. 117–24, Apr 1977.

[67] P. Lindström, "Aromatic-L-amino-acid decarboxylase activity in mouse pancreatic islets.," *Biochimica et biophysica acta*, vol. 884, pp. 276–81, Nov 1986.

[68] Y. Saisho, P. E. Harris, A. E. Butler, R. Galasso, T. Gurlo, R. A. Rizza, and P. C. Butler, "Relationship between pancreatic vesicular monoamine transporter 2 (VMAT2) and insulin expression in human pancreas," *Journal of Molecular Histology*, vol. 39, pp. 543–551, Oct 2008.

[69] I. García-Tornadú, A. M. Ornstein, A. Chamson-Reig, M. B. Wheeler, D. J. Hill, E. Arany, M. Rubinstein, and D. Becu-Villalobos, "Disruption of the Dopamine D2 Receptor Impairs Insulin Secretion and Causes Glucose Intolerance," *Endocrinology*, vol. 151, pp. 1441–1450, Apr 2010.

[70] A. Ustione and D. W. Piston, "Dopamine Synthesis and D3 Receptor Activation in Pancreatic β-Cells Regulates Insulin Secretion and Intracellular [Ca ²⁺] Oscillations," *Molecular Endocrinology*, vol. 26, pp. 1928–1940, Nov 2012.

[71] I. Lundquist, G. Panagiotidis, and A. Stenström, "Effect of L-dopa administration on islet monoamine oxidase activity and glucose-induced insulin release in the mouse.," *Pancreas*, vol. 6, pp. 522–7, Sep 1991.

[72] A. Raffo, K. Hancock, T. Polito, Y. Xie, G. Andan, P. Witkowski, M. Hardy, P. Barba, C. Ferrara, A. Maffei, M. Freeby, R. Goland, R. L. Leibel, I. R. Sweet, and P. E. Harris, "Role of vesicular monoamine transporter type 2 in rodent insulin secretion and glucose metabolism revealed by its specific antagonist tetrabenazine," *Journal of Endocrinology*, vol. 198, pp. 41–49, May 2008.

[73] D. S. Goldstein, G. Eisenhofer, and I. J. Kopin, "Sources and Significance of Plasma Levels of Catechols and Their Metabolites in Humans," *Journal of Pharmacology and Experimental Therapeutics*, vol. 305, pp. 800–811, Jun 2003.

[74] N. Simpson, A. Maffei, M. Freeby, S. Burroughs, Z. Freyberg, J. Javitch, R. L. Leibel, and P. E. Harris, "Dopamine-Mediated Autocrine Inhibitory Circuit Regulating Human Insulin Secretion in Vitro," *Molecular Endocrinology*, vol. 26, pp. 1757–1772, Oct 2012.

[75] A. R. Osterburg, P. Hexley, D. M. Supp, C. T. Robinson, G. Noel, C. Ogle, S. T. Boyce, B. J. Aronow, and G. F. Babcock, "Concerns over interspecies transcriptional comparisons in mice and humans after trauma.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, p. E3370, Sep 2013.

[76] B. Barrington, "Are Mice Reliable Models for Human Disease Studies? | Understand Nutrition." *Understand Nutrition*, February 14, 2013. Available: `https://understandnutrition.org/2013/02/14/are-mice-reliable-models-for-human-disease-studies/` [Accessed: 6 February 2018].

[77] J. Seok, H. S. Warren, A. G. Cuenca, M. N. Mindrinos, H. V. Baker, W. Xu, D. R. Richards, G. P. McDonald-Smith, H. Gao, L. Hennessy, C. C. Finnerty, C. M. López, S. Honari, E. E. Moore, J. P. Minei, J. Cuschieri, P. E. Bankey, J. L. Johnson, J. Sperry, A. B. Nathens, T. R. Billiar, M. A. West, M. G. Jeschke, M. B. Klein, R. L. Gamelli, N. S. Gibran, B. H.

Brownstein, C. Miller-Graziano, S. E. Calvano, P. H. Mason, J. P. Cobb, L. G. Rahme, S. F. Lowry, R. V. Maier, L. L. Moldawer, D. N. Herndon, R. W. Davis, W. Xiao, R. G. Tompkins, and L. S. C. R. P. Inflammation and Host Response to Injury, Large Scale Collaborative Research Program, "Genomic responses in mouse models poorly mimic human inflammatory diseases.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, pp. 3507–12, Feb 2013.

[78] G. Kolata, "Mice Fall Short as Test Subjects for Some of Humans' Deadly Ils," *The New York Times: Science*, February 11, 2013.

[79] Coon Lab Founder Mice Proteomics. Available: `http://coonlabdata.com/founder_mice/` Accessed: 6 February 2018

[80] Attie Lab Diabetes Database. Available: `http://diabetes.wisc.edu/cc_founder.php` Accessed: 6 February 2018

[81] M. E. Rabaglia, M. P. Gray-Keller, B. L. Frey, M. R. Shortreed, L. M. Smith, and A. D. Attie, "α-Ketoisocaproate-induced hypersecretion of insulin by islets from diabetes-susceptible mice," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 289, pp. E218–E224, Aug 2005.

[82] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox, "The Perseus computational platform for comprehensive analysis of (prote)omics data," *Nature Methods*, vol. 13, pp. 731–740, Sep 2016.

[83] S. Horvath, "Weighted Gene Co-expression Network Analysis," 2009. Available: https://labs.genetics.ucla.edu/horvath/htdocs/CoexpressionNetwork/. [Accessed: 6 February 2018].

[84] J. Dong and S. Horvath, "Understanding network concepts in modules," *BMC Systems Biology*, vol. 1, p. 24, Jun 2007.

[85] P. Langfelder and S. Horvath, "Tutorial for the WGCNA package for R I. Network analysis of liver expression data in female mice 2.c Dealing with large data sets: block-wise network construction and module detection," 2014. Available: https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/FemaleLiver-02-networkConstr-blockwise.pdf. [Accessed: 6 February 2018].

[86] M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist, "Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis," Aug 2007.

[87] J.-M. T. Wong, P. A. Malec, O. S. Mabrouk, J. Ro, M. Dus, and R. T. Kennedy, "Benzoyl chloride derivatization with liquid chromatography–mass spectrometry for targeted metabolomics of neurochemicals in biological samples," *Journal of Chromatography A*, vol. 1446, pp. 78–90, May 2016.

Chapter 5

GENETIC CONTROL OF THE MOUSE ISLET PROTEOME

ECF designed experiments, ran samples, analyzed data, and authored this work.

Elyse C. Freiberger, Daniel M. Gatti, Vanessa Linke, Nicholas W. Kwiecien, Edna A. Trujillo, Kathryn L. Schueler, Mary E. Rabaglia, Donald S. Stapleton, Alexander S. Hebert, Mark P. Keller, Alan D. Attie, Gary A. Churchill, Joshua J. Coon.

Abstract

Multi-omics analysis is a powerful means of understanding disease at the systems level; however, because it involves many different techniques working in tandem, applying these methods to sample-limited experimental models can be incredibly challenging. In particular, pancreatic islets are a key tissue for diabetes research, but they are among the rarest cells in the body, comprising only 1-2% of the pancreas. Laboratory mouse strains yield ~250 islets per animal, confounding the ability to perform multiple 'omics' analyses on islets from a single mouse. We have taken advantage of advances in LC-MS/MS technology along with optimized sample preparation workflows to develop a rapid, reliable approach to quantifying the mouse islet proteome that requires only 25 islets per mouse (~25,000 cells). We applied these methods to 383 diversity oubred (DO) mouse islet aliquots to generate a protein quantitative trait loci (pQTL) mapping dataset that is unique in its focus and scale. We quantified of 8,200 proteins across all samples, and identified 2,432 significant pQTLs, comprising 1,119 local pQTls and 1,313 distant pQTLs. These data were integrated with islet eQTL analysis on the same animal cohort to describe novel regulatory pathways for protein expression and to identify potential allelic drivers of these effects.

Introduction

As the speed and depth of mass spectrometry-based 'omics' technologies continue to improve, the breadth of possibilities for their application also expands[1]. Our lab and

others have pushed the boundaries of proteomic, metabolomic, and lipidomic analyses to the point that tens to hundreds of comprehensive 'omes' can be measured each day, which now allows for larger scale experiments to be undertaken than were previously possible[2–6]. In proteomics specifically, we have developed methods to analyze nearly complete proteomes in less than two hours; prior methods took nearly five times as long to reach the same depth[4,5]. We have since applied that method to a large-scale yeast knockout study geared toward annotation of unknown protein functions that also incorporated metabolomics and lipidomics information[7]. Other studies have been combining mass-spectrometry-based 'omics' analysis with the more mature 'omics' analyses of genomics and transcriptomics[2,3,8]. Genomic information can be correlated with other 'omes' using statistical techniques such as genome-wide association studies (GWAS) or quantitative trait locus (QTL) mapping[2,9–12]. QTL mapping has been a common tool in plant biology and agricultural research[13–16] but recently has become popular in the study of mammalian 'omes' due in part to the development of the Diversity Outbred (DO) mouse population by Gary Churchill and colleagues at the Jackson Laboratories. These mice were bred from eight inbred laboratory strains and better approximate human populations in both genetic variability and heterozygosity[17–21] (**Supp. Fig. S5.1a**). The breeding scheme also affords a finer resolution for QTL mapping than was possible with other mouse models – QTL resolution is dependant on the number of known recombination events and the genetic marker density. A recent study by Chick et al. investigates the interplay between proteomics, transcriptomics, and genomics in 192 DO mouse livers[2]. This work was the first

description of large-scale proteomics being correlated with genetics using QTL mapping; the authors detected nearly 2,900 significant protein QTLs (pQTLs). Here we present a more focused, larger-scale pQTL study using the DO mouse resource. With a cohort of 383 mice fed a high-fat, high-sucrose diet, we have geared our analyses toward connecting genetic loci with phenotypes related to diabetes etiology and metabolic dysfunction. To that end, we present a proteomics analysis that is second of four 'omic' QTL studies to be assembled into a comprehensive picture of genetic regulation of pancreatic islets function (**Fig. 5.1a**. Pancreatic islets are a key tissue used for diabetes research, but they present a particular challenge to multi-omics analysis as they are among the rarest cells in the body, comprising only 1-2% of the pancreas[22–25]. Laboratory mouse strains yield ~250 islets per animal, confounding the ability to perform multiple "omics" analyses on islets from a single mouse. We have recently developed a rapid, reliable approach to quantifying the mouse islet proteome that requires only 25 islets per mouse (~ 25,000 cells)[26]. The remainder of the islets isolated from each animal was used for extensive transcriptomic and secretion phenotype analysis. We used our optimized label-free workflow to quantify the islet proteomes, then subjected a percentage of the proteins to QTL mapping based on a detection-frequency cutoff. The pQTL results were segregated by their proximity to the coding gene of the associated protein, and assessed for possible regulatory mediators. Transcriptomic and eQTL analyses for the same islet samples were described by Keller et al.[27]; these data were used first to determine the extent of the overlap between the transcriptome and proteome, and then to identify QTL mediation by both RNA and protein.

The balance of these results will be integrated with plasma metabolomic and lipidomic QTL measurements to generate one of the largest multi-omics resources to date.

Results

Proteomics data collection Prior to the genesis of this proteomics study, the majority of islets from our cohort of DO mice had already been aliquoted for other analyses (e.g. transcriptomics, isolated islet secretion studies). For each mouse only 25 islets remained, which corresponds to ~13 μg of protein on average (**Fig. 5.1b**). This is about an order of magnitude less protein than we'd normally require for a standard discovery proteomics study. Protein loss during sample preparation is a major concern when working with such small amounts of starting material - with every transfer step, the proportion of the sample that is lost to incidental retention on the tubes or pipette tips is proportionally larger. Islets specifically complicate the issue as well, because they are composed of at least five different cell types. Across the pancreas, the distribution of those cell types is fairly constant, but as the population of islets is reduced, the likelihood of an outlier islet skewing the apparent distribution increases. Such inconsistent cell-type composition could negatively affect reproducibility in the protein quantification. To increase feasibility and reproducibility of a sample-limited, large-scale proteomics study, we modified our standard proteomics analytical workflow for optimized sample retention and throughput (**Supp. Fig. S5.1b**). We moved from a guanidine boil plus methanol extraction lysis step to lysing in urea using careful sonication; with such low starting material, the methanol extraction becomes

impossible on a large scale due to lack of a consistently visible protein pellet. This approach has some limitations, as temperature must be monitored closely to prevent carbamylation of the proteins. It also may prove too gentle for tissues that contain more connective tissue, but because the islets have been manually isolated and separated from cellular debris, the lack of heat or extensive mechanical lysis is not detrimental. Following lysis, samples were moved into clean 96-well plates for the remainder of the sample preparation steps. The use of 96-well plates simplifies sample identity notation, and allows for use of multi-channel pipettes, which increases the throughput enough that one analyst can prepare all 383 samples in a week's time. The final protein mass yield from the 25-islet aliquots ranged from ~1 to 55 μg, with the majority yielding between 6 and 22 μg (**Fig. 5.1b**). For label-free quantification, constant protein load on-column is important for reproducibility, so in order to allow at least two injections per sample (in case of instrument failure), we elected to inject 1 μg of peptides per analysis, instead of our standard 2 μg. We also extended our chromatographic gradient from 90 minutes to 120 minutes to maximize identifications. Because of the low amount of starting material, pre-fractionation to increase proteomic depth was not feasible.

Quantified proteomes Following data collection and analysis, we assessed the quality of the quantitative results to confirm its suitability for the QTL mapping pipeline. The first metric of interest is proteomic depth; we quantified a total of 8,234 proteins across all samples, and averaged 5,504 quantified proteins per sample (**Fig. 5.1c**). The total number of

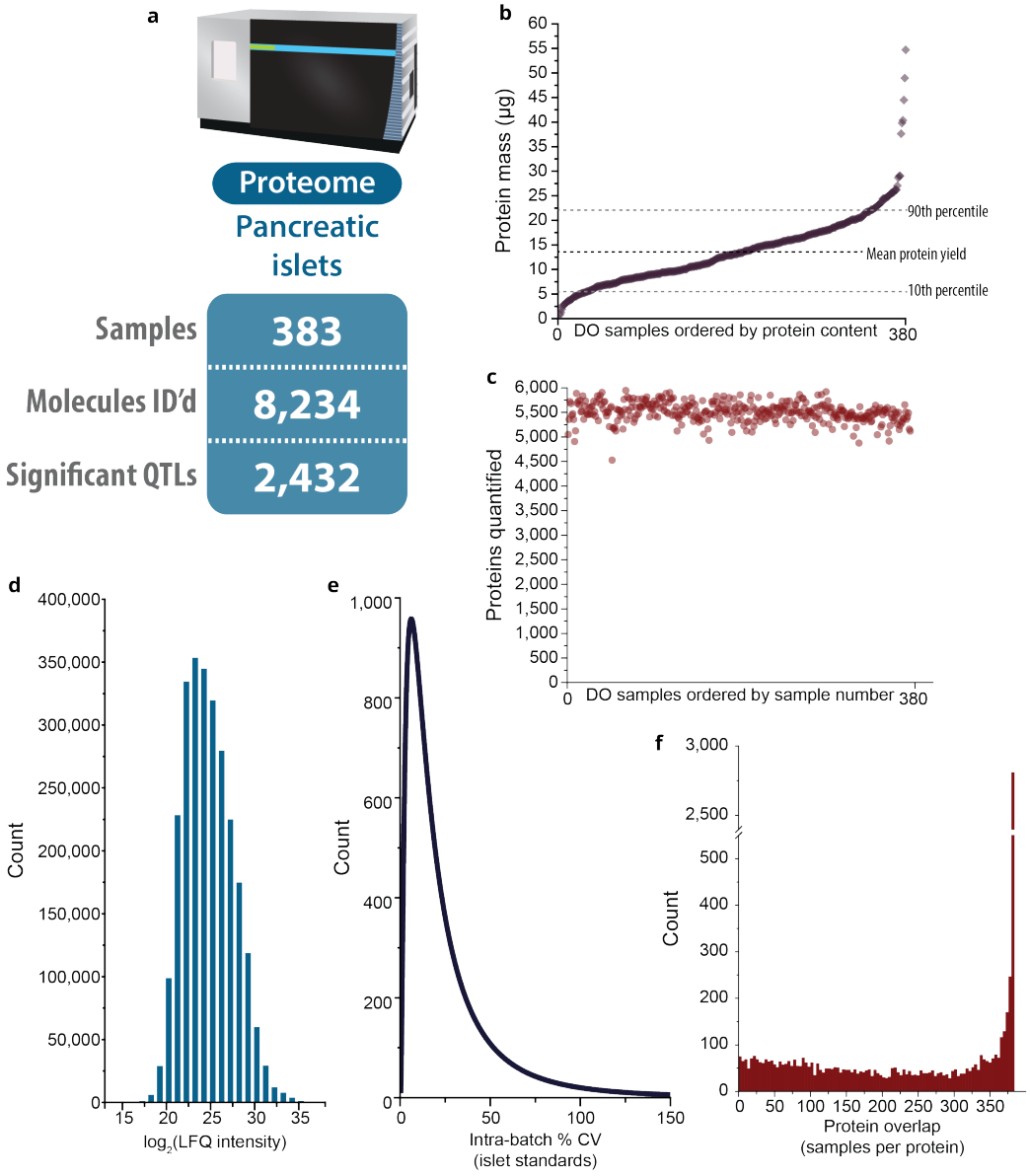
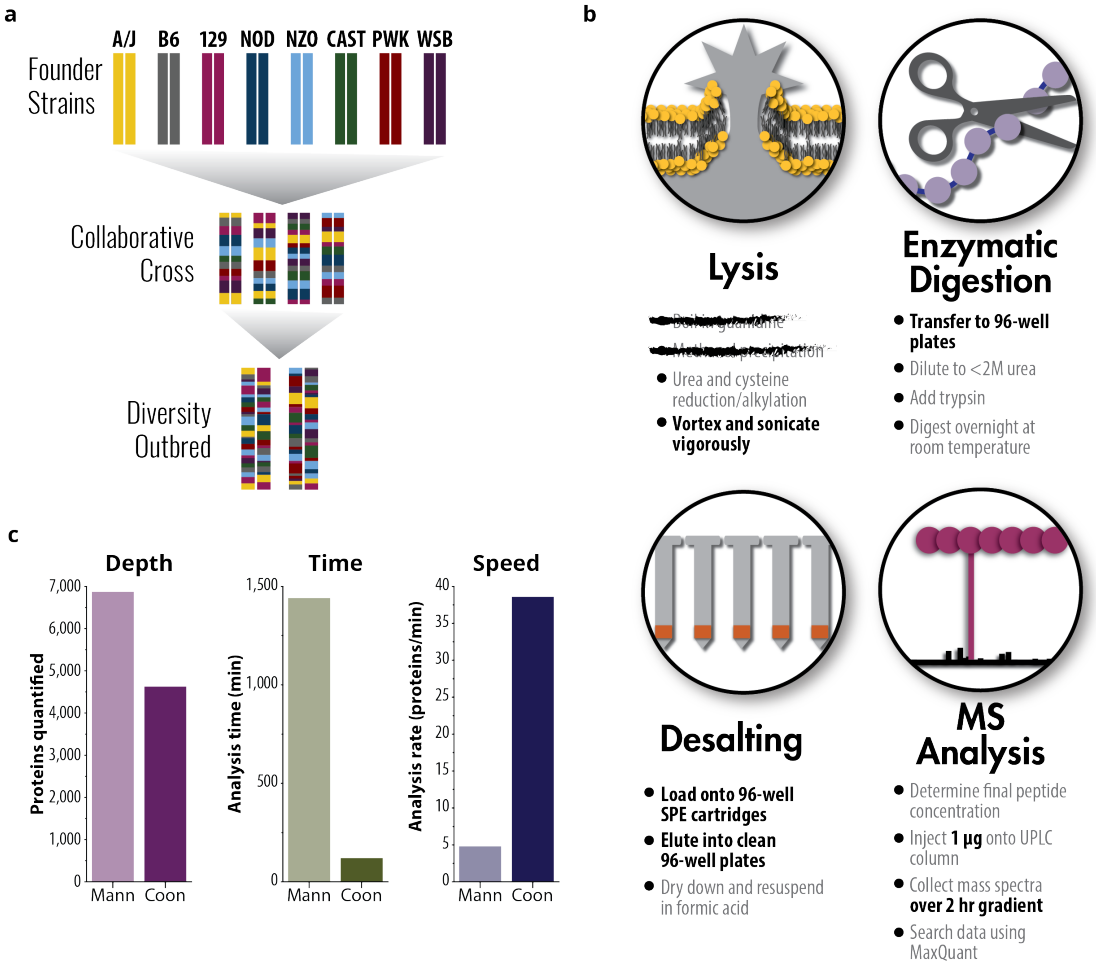
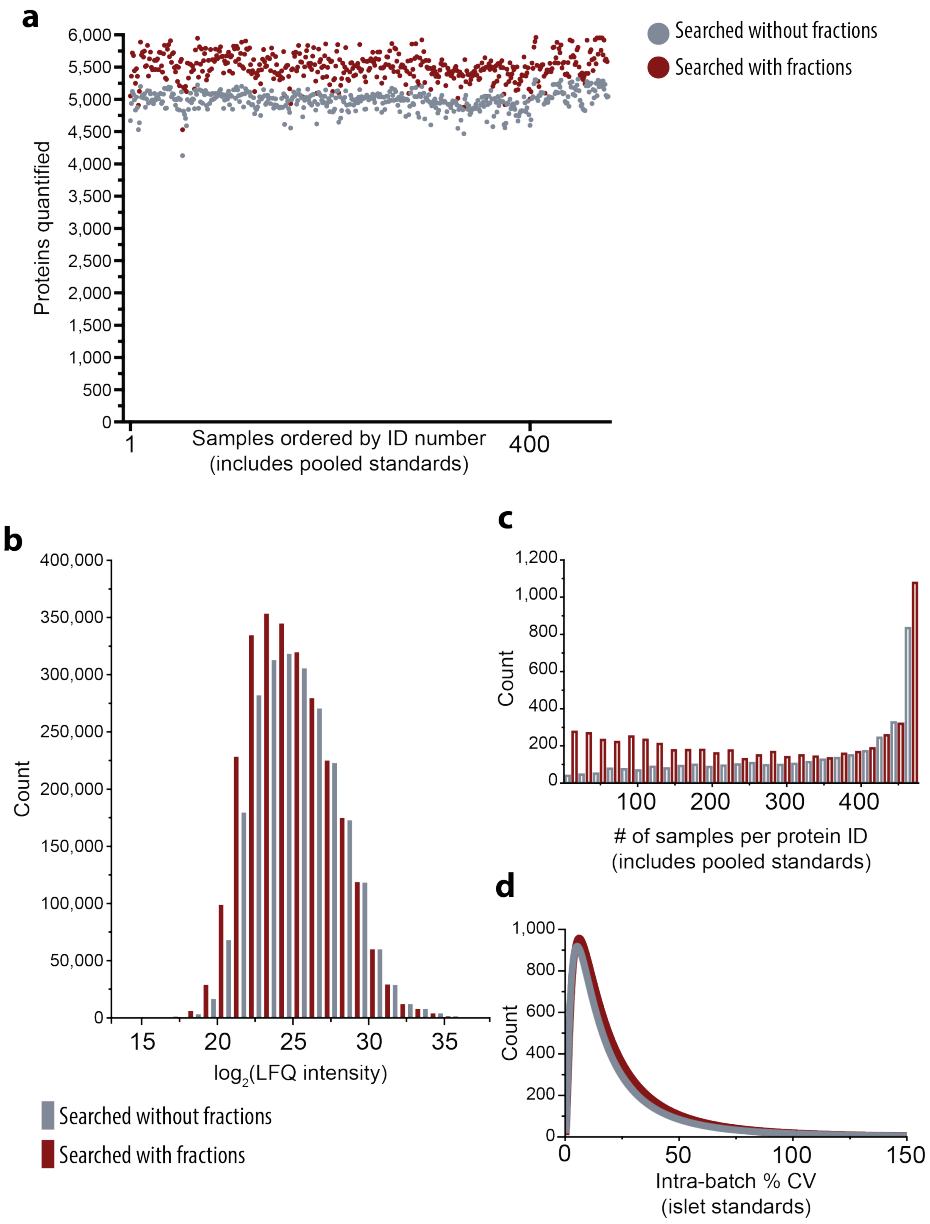


Figure 5.1: Experimental metrics. **A** Summary of the proteomics data including sample number, quantified proteins, and significant pQTls ($P < 0.1$). **B** Distribution of protein yield from all DO mouse 25-islet aliquots, ordered from lowest to highest yield (µg). The mean yield is 13.4 µg, 10th percentile is 6.3 µg, and 90th percentile is 22.2 µg. C. Distribution of $\log_2$-normalized label-free quantification (LFQ) intensities across all quantified proteins. **D**. Density plot of coefficients of variation between pooled islet standards within batches. Median %CV = 16.6. **E** Distribution of protein overlap: x-axis represents the number of samples in which each protein is quantified.



Supplementary Figure S5.1: Summary of methods. **A** Schematic of the breeding scheme for DO mouse generation, through the inbred Collaborative Cross population (adapted from Churchill 2012[19]). **B** Summary of the workflow optimizations used for quantification of sample-limited islet proteomes. **C** Comparison of the optimized high-throughput workflow with the deepest islet proteome information collected to date[28].

quantified proteins was optimized by taking advantage of the MaxQuant match-between-runs parameter, which quantifies a detected MS^1 feature in all samples (within a mass and retention-time tolerance) even in the absence of MS^2 data, as long as that same feature is identified via MS^2 in at least one sample across the dataset. We also benchmarked the dynamic range of our measurements (**Fig. 5.1d**), and quantitative reproducibility (**Fig. 5.1e**). Because each DO mouse is genetically unique, our data had not biological replicates. Instead, we used a standard sample of islets pooled from each of the eight founder mouse strains as a proxy to confirm the quality of the quantification. Measurements from these samples were used in **Fig. 5.1d** to generate the coefficients of variation. The last major concern when addressing the data suitability is overlap; because we have adopted a relative quantification workflow, our ability to probe the proteome at depth is only as good as our overlap in identifications. **Fig. 5.1f** shows a histogram of the number of samples in which each protein is identified, and the majority fall toward the right of the plot, at the highest level of overlap. To maximize the number of identified features, we included in the search a set of deeply-fractionated mouse islets generated from a pool of islets from all eight founder mouse strains. Proteins from these samples were only identified and not included in the quantification. Addition of these data into the search parameters increased our quantified proteins in the actual DO samples by 20% (**Supp. Fig. S5.2a**), without sacrificing quantitative quality (**Supp. Fig. S5.2c,d**). The major gain in identifications happens at lower intensities, where it is more difficult to get quality MS^2 information in single-shot analyses (**Supp. Fig. S5.2b**).



Supplementary Figure S5.2: Comparison of protein quantification with and without inclusion of pre-fractionated islet proteome library in the search. Output searched without fractions (blue) and searched with fractions (red). All analyses include both DO mouse islet samples and single-shot pooled islet standards. **A** Effect of fractions on proteins quantified. DO samples are ordered by sample number. **B** Effect of fractions on the distribution of quantified protein intensities. **C** Effect of fractions on overlap of protein identifications. **D** Effect of fractions on % CV between pooled islet standards.

QTL mapping For the pQTL mapping, we set a detection frequency threshold of 50% for each protein; that is, we only used proteins quantified in at least 50% of the samples for the analysis, which resulted in an input of 5,433 proteins. We detected 2,432 total significant pQTLs at P < 0.1 from 2,031 unique proteins (**Fig. 5.1a**, **Fig. 5.2a**). Of these results, 1,119 pQTLs were local (*cis*) to their coding gene, and 1,313 were distant (*trans*) (**Fig. 5.2b**). For the trans pQTLs, we tested for mediation against both protein and transcripts to identify the drivers of those effects — the local pQTLs are assumed to be driven by mutations at their coding genes, whereas the distant pQTLs must be regulated by some other genes/gene products. For examples of these results, see the **Synuclein regulation** and **Tyrosine hydroxylase** sections.

Comparison with other DO mouse QTL datasets The DO mouse population has been used for other 'omics' QTL mapping, including islet transcriptomics, and liver proteomics. We compared results from our islet pQTL study to these studies[2,27]. The islet eQTL study used islet aliquots from same cohort of mice as our islet pQTL study; as expected the number of transcripts quantified far outnumbers the proteins quantified, but of the proteins quantified, 97% were also quantified in the transcript study (**Fig. 5.3a**). However, only 22% of the significant pQTLs were also detected as eQTLs (**Fig. 5.3b**). Of these 535 overlapping QTLs, the majority are local. That is, the QTL maps to the coding gene of the transcript or protein (**Fig. 5.3c** center). In the pQTL study, a majority of significant QTLs are distant (**Fig. 5.3c** right), while most of the eQTLs are local (**Fig. 5.3c** left). Comparing our islet

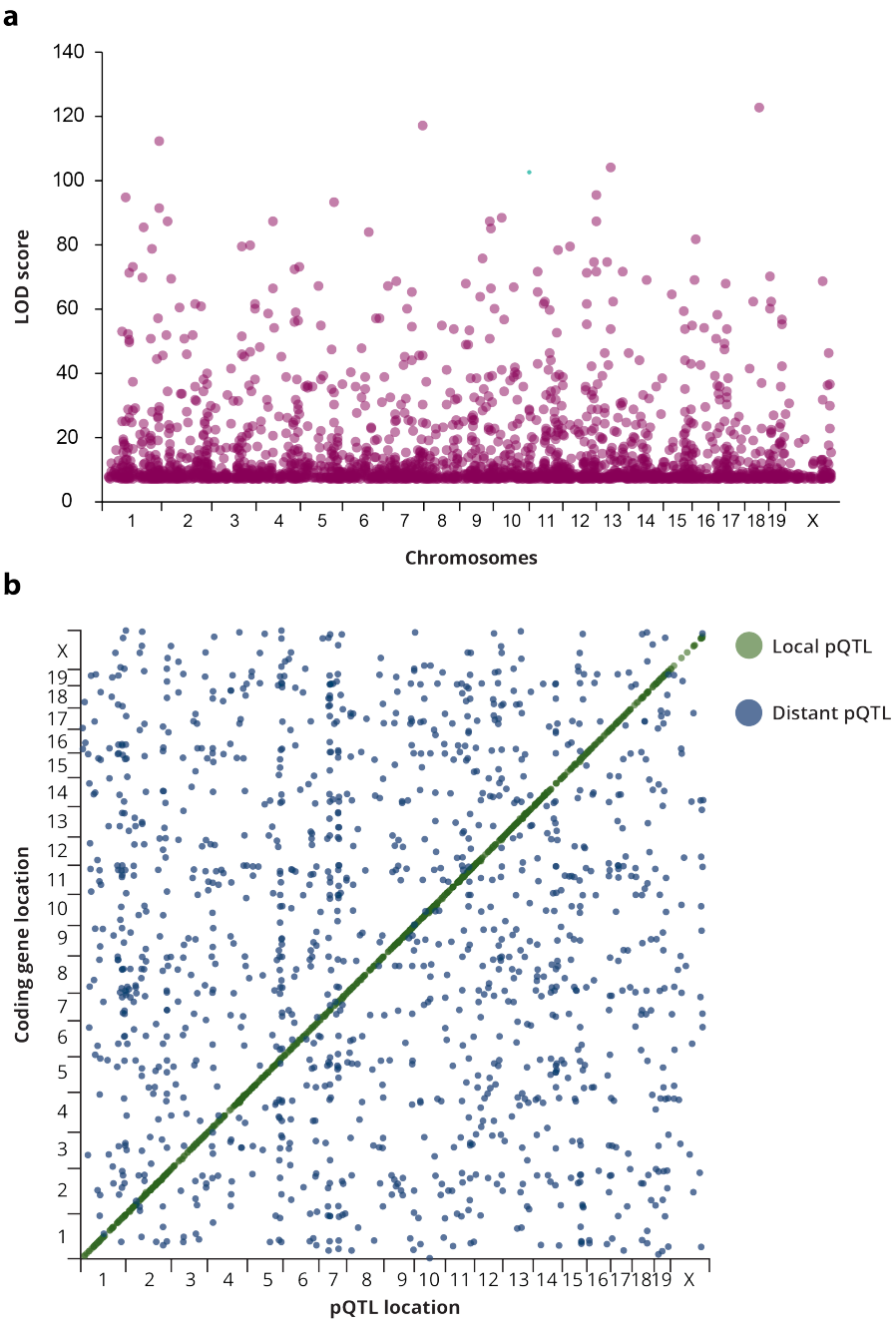


Figure 5.2: QTL mapping. **A** Manhattan plot showing all significant pQTLs detected in the islet proteomics experiment. **B** Scatter plot showing the relationship between pQTL location and protein coding gene location. Local pQTLs cluster along the center line in green, distant pQTLs are distributed throughout the remainder of the plot in blue.

proteomics to the liver proteomics study by Chick et al. on a DO mouse cohort, the liver study had more quantified protein input into the QTL analysis (5,433 vs 6,706), and 80% of the proteins quantified in islets were also quantified in liver (**Fig. 5.4a**). Of the 2,433 significant islet pQTLs, only 17% were also identified as significant liver pQTLs (textbfFig. 5.4b). Again, the majority of the overlapping QTLs are local (textbfFig. 5.4c center). In the balance of the pQTL studies, more distant QTLs were identified in the islet than in the liver, even though more QTLs total were identified in the liver (textbfFig. 5.4c right and left).

Synuclein regulation One of highest LOD scores of a distant QTL belongs to the synuclein alpha pQTL at chromosome 13 (LOD = 53.66). The synuclein family of proteins (alpha, beta, and gamma synuclein), which are poorly-characterized functionally, have primarily been studied in neurons. Specific interactions between members of the synuclein family, particularly in the islet, have remained unclear, but we are detecting the significant, distant Snca pQTL as overlapping with both a local Sncb pQTL (LOD = 104.07), and distant Sncg pQTL (LOD = 14.74) (**Fig. 5.5a**). Because Snca and Sncb have greater than 60% sequence homology, we tracked our peptide coverage to confirm that we are in fact able to differentiate between the two proteins and are not just detecting an artifact of the search algorithm (**Supp. Fig. S5.3**) For all three synuclein pQTLs, founder effects show a pronounced expression increase in mice with CAST-linked mutations, which suggests all three synucleins are interacting in some way within the islet (**Fig. 5.5b**). Mediation analysis confirms the regulation of Snca and Sncg levels by Sncb at both the protein and transcript level (**Fig.**

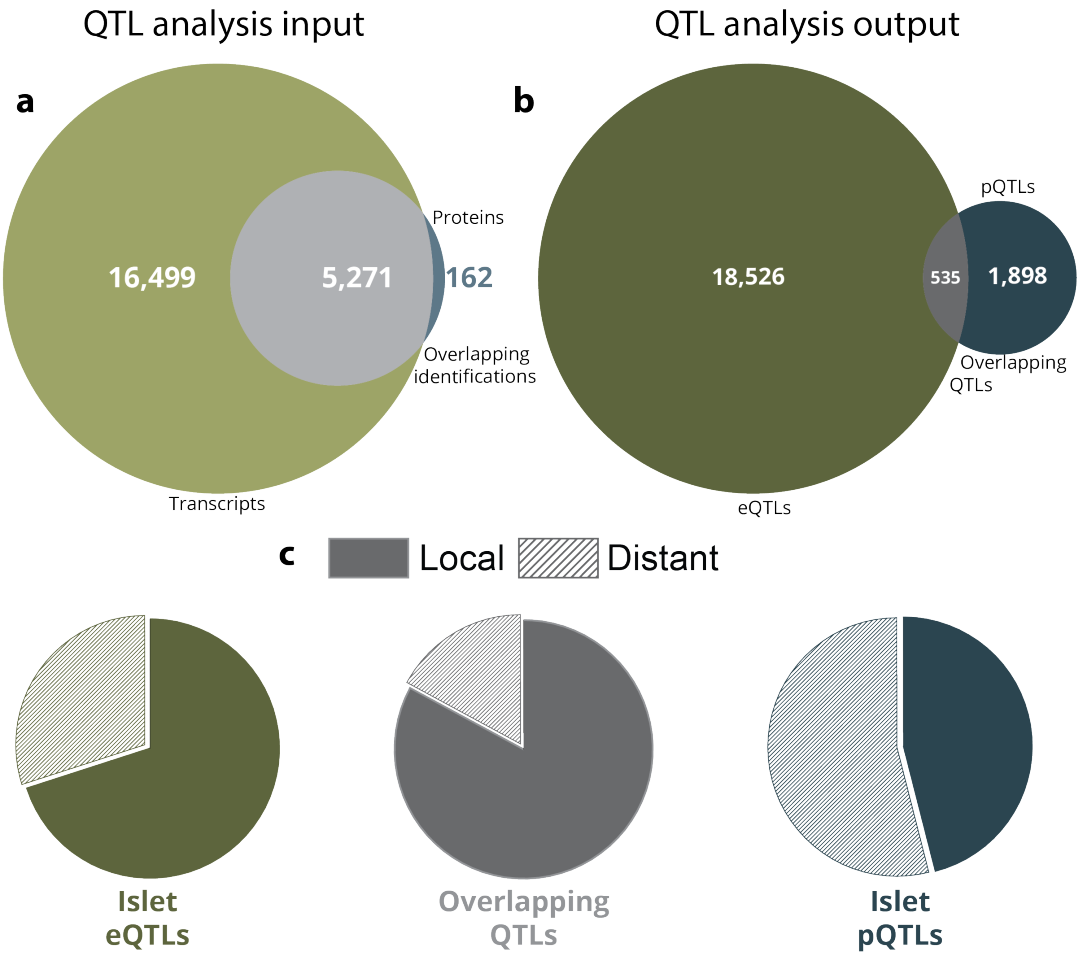


Figure 5.3: Overlap with islet eQTL mapping. **A** Comparison of gene product identities between islet transcriptome (green) and islet proteome (blue) analyses. 21,777 transcripts and 5,433 proteins were used for QTL mapping. Of those, the overlap is 5,271 gene products. **B**. 19,061 significant eQTLs and 2,433 significant pQTLs were identified. Of these, only 535 mapped the same gene products to the same loci. **C** The breakdown of local versus distant QTLs for the eQTL dataset (green), pQTL dataset (blue), and the 535 overlapping QTLs (gray). Transcriptomic and eQTL data from Keller et al.[27].

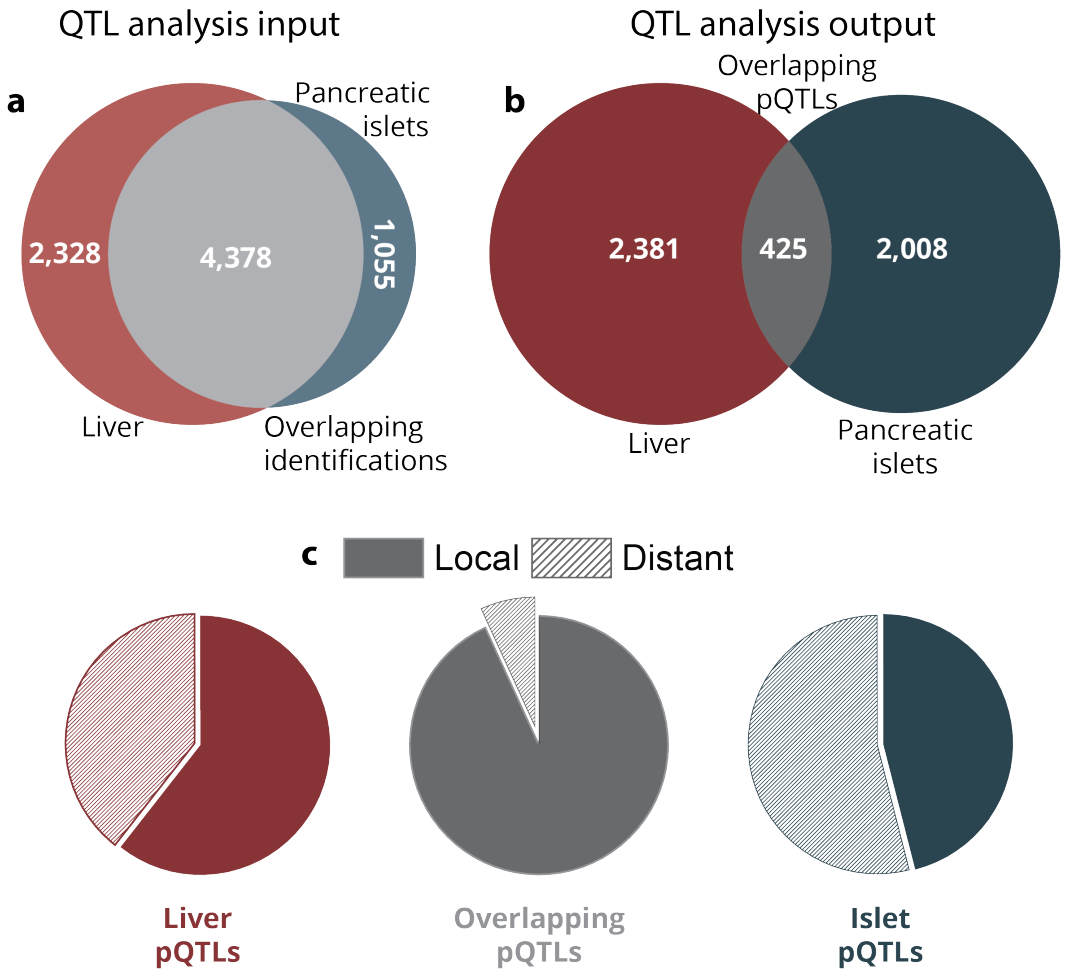


Figure 5.4: Overlap with liver pQTL mapping. **A** Comparison of protein identities between liver proteome (red) and islet proteome (blue) analyses. 6,706 liver proteins and 5,433 islet proteins were used for QTL mapping. Of those, the overlap in identifications is 4,378 proteins. **B** 2,806 significant liver pQTLs and 2,433 significant islet pQTLs were identified. Of these, only 425 mapped the same proteins to the same loci. **C** The breakdown of local versus distant QTLs for the liver dataset (red), islet dataset (blue), and the 425 overlapping QTLs (gray). Liver proteomics and pQTL data from Chick et al.[2].

5.5c,d, **Supp. Fig. S5.4a,b**). SNP association analysis generates significance values (LODs) for each known SNP within the area of the locus of interest. In this case, we are looking at a 4 Mb window centered on the synuclein locus; the resulting data suggests specific allelic variants that could drive the pQTL phenomena. **Supp. Fig. S5.5, Supp. Fig. S5.6**, and **Supp. Fig. S5.7** plot all SNP LODs, with the top scoring points highlighted in pink. Founder effects are reiterated in the top panels, while gene identities that overlap with the locus are plotted in the bottom panels. The table in **Supp. Fig. S5.8** shows the SNPs that are high-scoring with relation to all three overlapping synuclein pQTLs.

Tyrosine hydroxylase In a previous study we identified strain-specific tyrosine hydroxylase (Th) expression in islets as a potential driver of differences in insulin secretion. In the pQTL analysis, Th maps to its coding gene on chromosome 7 with a very high LOD score (LOD = 116.96) (**Fig. 5.6a**). The Th local pQTL correlates with an increase in Th expression in DO mice containing PWK- and CAST-specific alleles at that locus (**Fig. 5.6b**). These strain effects correspond to the expression patterns seen in PWK and CAST strains in the previous study (**Supp. Fig. S5.9**). However, several distant pQTLs map to the same chromosomal location, none of which are previously known to interact with Th (**Fig. 5.6a**). Using mediation analysis, we can determine whether the regulatory driver of these distant QTLs is Th or another nearby gene. Results of the mediation analysis shows Mat2a, Mat2b, and Dnajc12 distant pQTLs are mediated by Th at both the transcript and protein level (**Fig. 5.6c,d**). These proteins also show similar expression changes in PWK and CAST founder

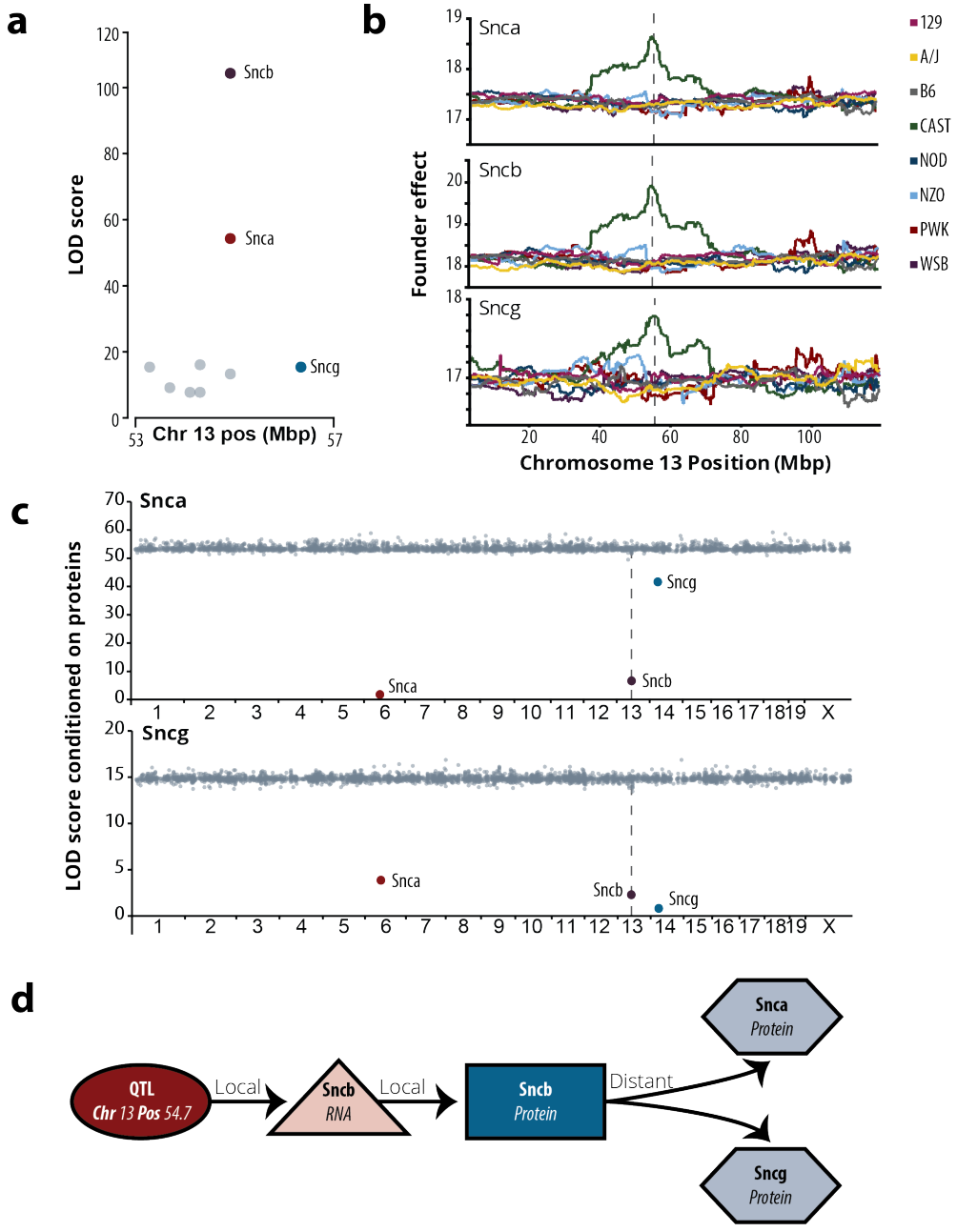
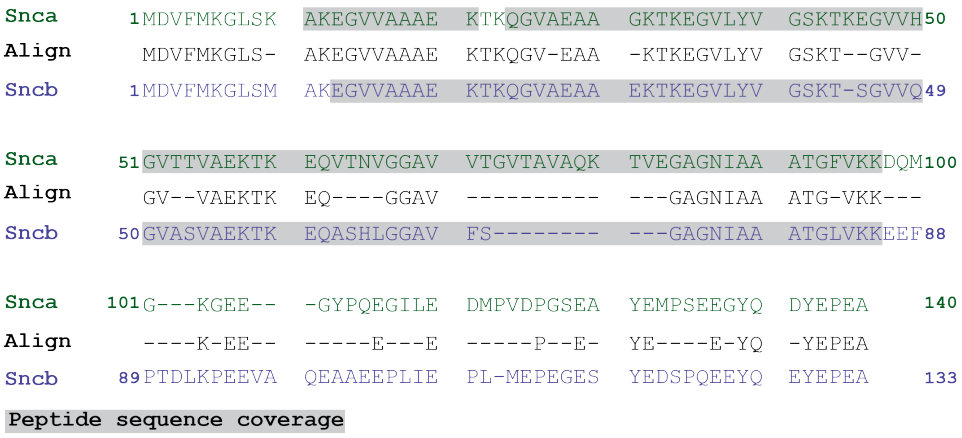
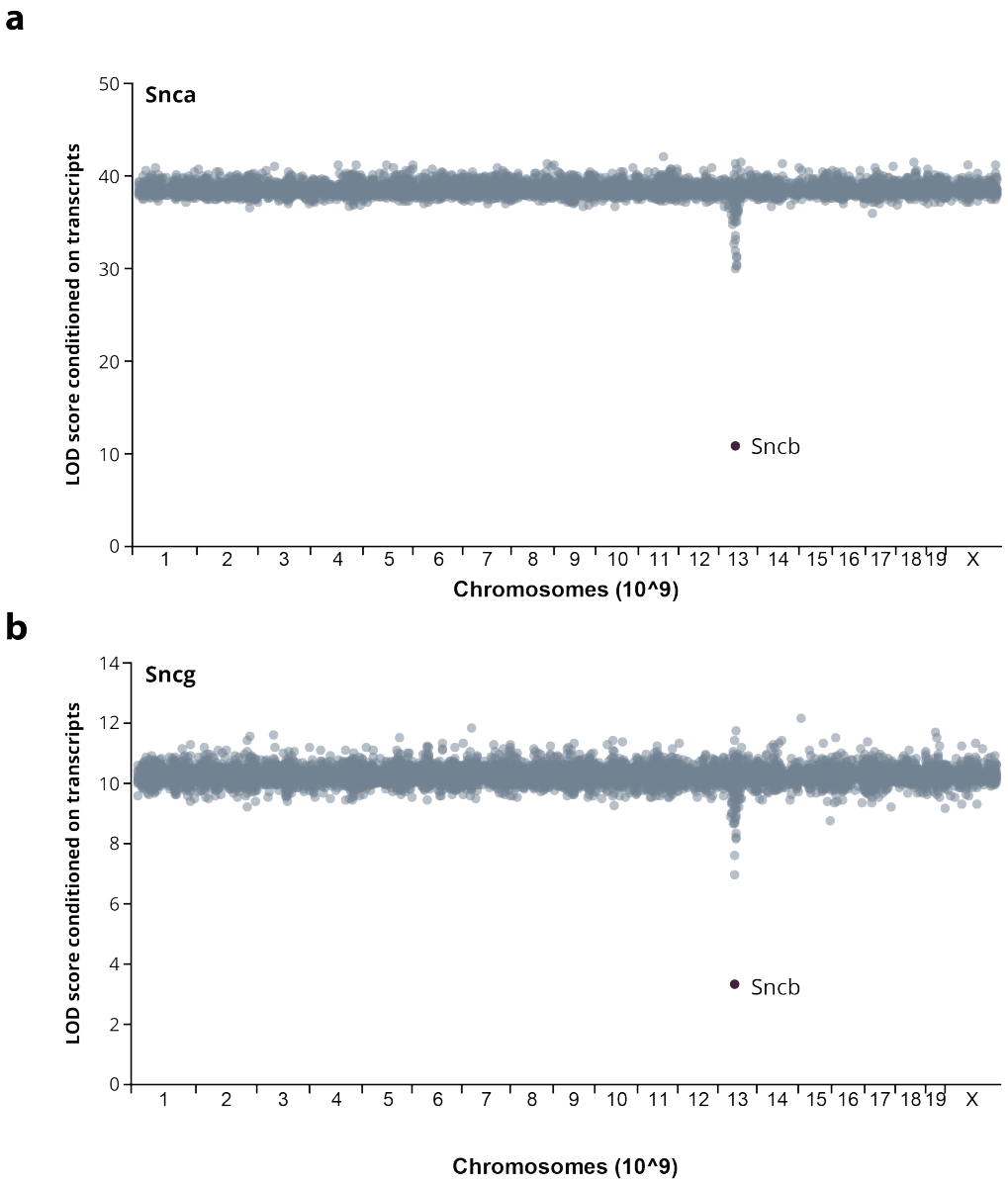


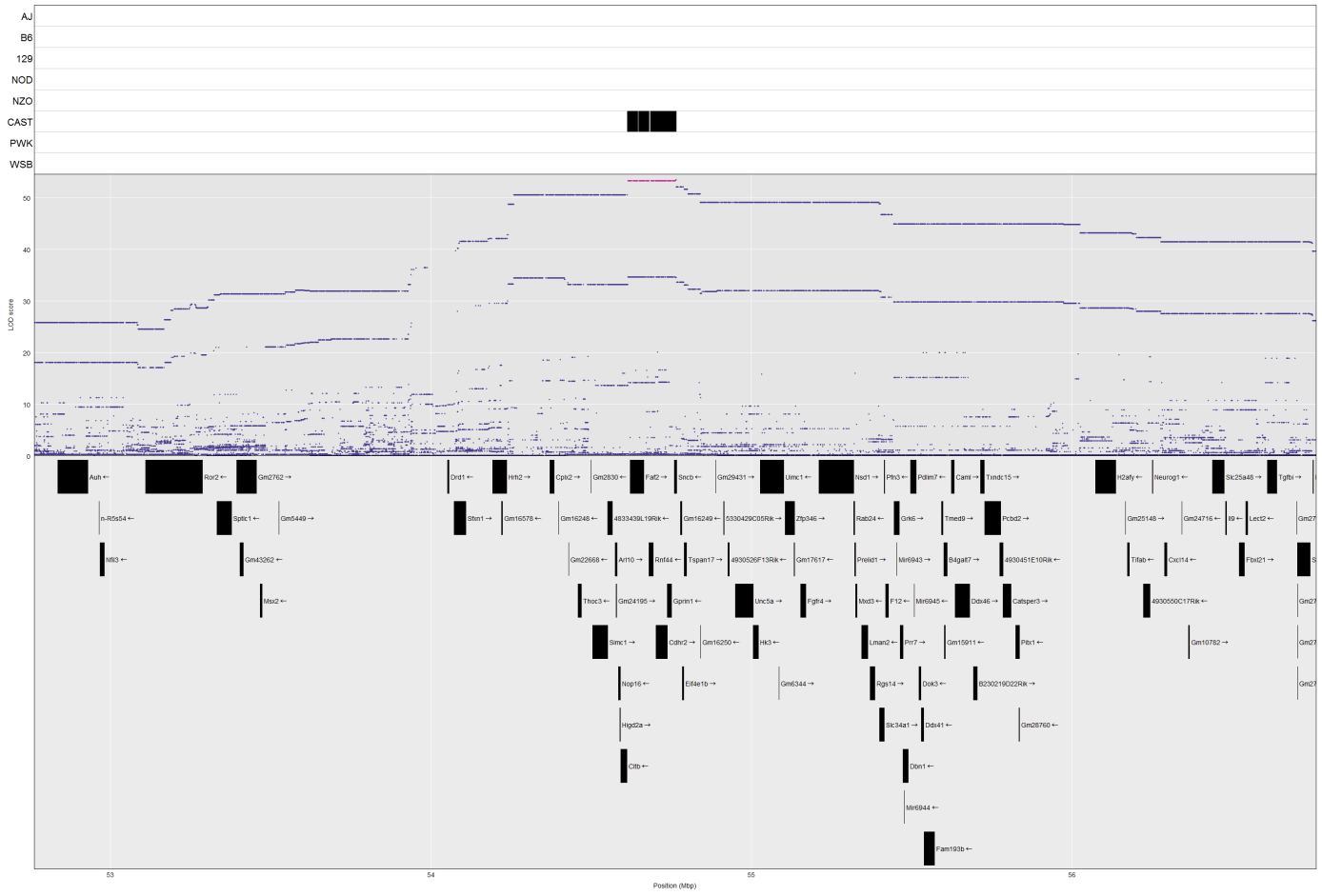
Figure 5.5: Synuclein locus at chromosome 13. **A** Significant pQTLs mapping to the synuclein locus: Sncb local pQTL (purple), Snca distant pQTL (red), Sncg distant pQTL (blue), and unrelated pQTLs not implicated in synculein by mediation analysis or founder effect regulation (gray). **B** Founder effect plots for Snca (top), Sncb (middle), and Sncg (bottom) around the locus of interest.



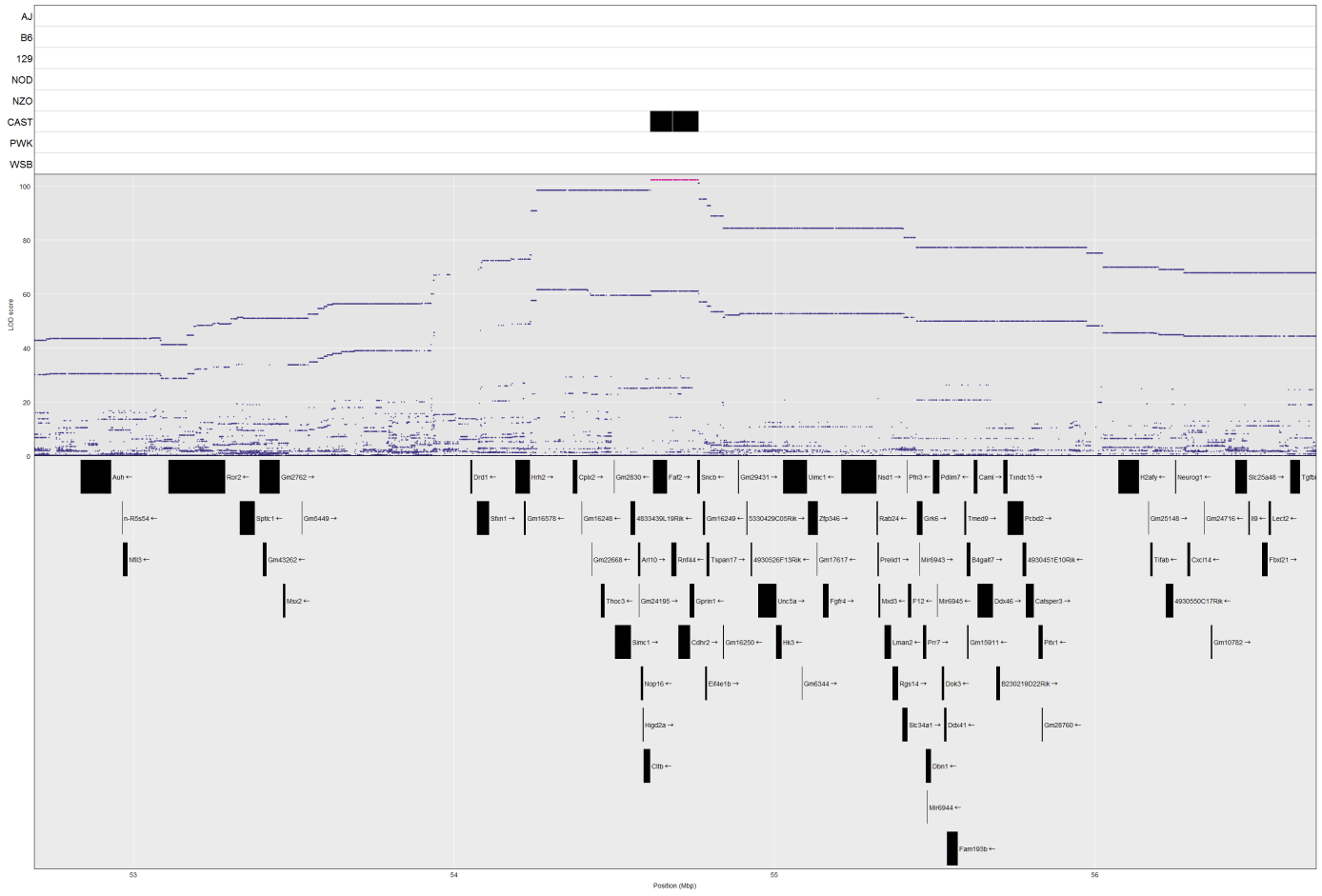
Supplementary Figure S5.3: Peptide coverage of Snca and Sncb. Snca sequence (green) versus Sncb sequence (purple). Overlap in the sequences in represented in the center (black). Peptides quantified in our data are highlighted in gray and cover several areas of differentiation.



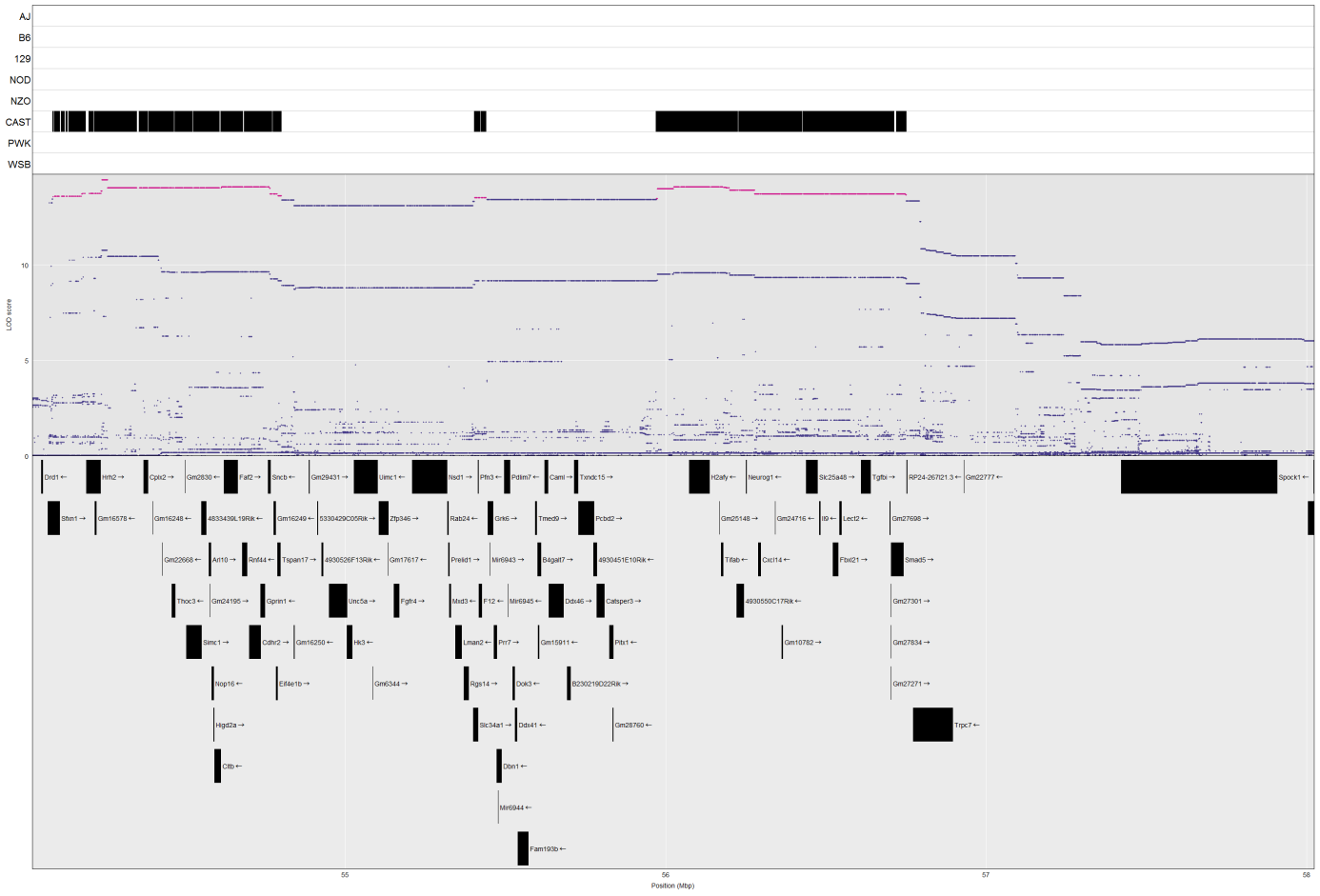
Supplementary Figure S5.4: Mediation of synucleins by transcript.. Mediation plots for Snca (**A**) and Sncg (**B**) with LOD scores conditioned on transcripts. LODs conditioned on Sncb (purple) are highlighted. At the locus of interest, Sncb causes the greatest drop in LOD score.



Supplementary Figure S5.5: SNP associations of Snca pQTL at chromosome 13. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.



Supplementary Figure S5.6: SNP associations of Sncb pQTL at chromosome 13. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.



Supplementary Figure S5.7: SNP associations of Sncg pQTL at chromosome 13. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.

SNP	Alleles	Pos (Mb)
13:54614538_T/C	T|C	54.614538
rs46262394	G|A	54.617248
rs46485412	G|A	54.621885
rs45635103	C|T	54.62906
13:54633274_G/T	G|T	54.633274
rs265344836	C|T	54.651728
rs236168112	C|T	54.691859
rs213573667	C|G	54.705446
rs243001327	C|T	54.708521
rs51693628	C|T	54.763209

Supplementary Figure S5.8: SNP associations at the synuclein locus chromosome 13. This table includes the SNPs denoted as significant in SNP association analysis for all three synucleins at the chromosome 13 locus. Included are SNP ID, allele effect, and position in Mb.

Figure 5.5: C Mediation plots for Snca (top) and Sncg (bottom) with LOD scores conditioned on proteins. LODs conditioned on Snca (red), Sncb (purple), and Sncb (blue) are highlighted. At the locus of interest, Sncb causes the greatest drop in LOD score. When Snca or Sncg are conditioned on themselves, the LOD score characteristically drops to zero. **D** Model for medation of Snca and Sncg through Sncb transcript and protein at chromosome 13.

effect analyses (**Fig. 5.6b**). In the protein mediation plots for Mat2a, Mat2b, and Dnajc12, all three proteins show characteristic LOD drops for each other in addition to Th (**Fig. 5.6c**); since none of these are coded at the locus of interest, the are not mediating the QTL significance, but are co-regulated with one another (through Th). We again performed a SNP association analysis, highlighting the significant SNPs in the middle panels, recapitulating founder effects in the top panel, and listing gene identities in the bottom panels (**Supp. Fig. S5.10, Supp. Fig. S5.11, Supp. Fig. S5.12**, and **Supp. Fig. S5.13**). The table in **Supp. Fig. S5.14** charts specific SNPs that are high-scoring across all four pQTLs overlapping at the Th locus.

Discussion

These data represent the largest protein QTL study to-date in terms of cohort size, and has pushed the boundaries of starting material required for comprehensive bottom-up proteomics. We have optimized protocols both for throughput and sample-retention to allow a single analyst to prepare and process hundreds of low-level tissue samples and generate high-quality, in-depth quantitative proteomes. Compared to the largest islet proteome described to date, we quantify at 70% of the depth in 8% of the analysis time,

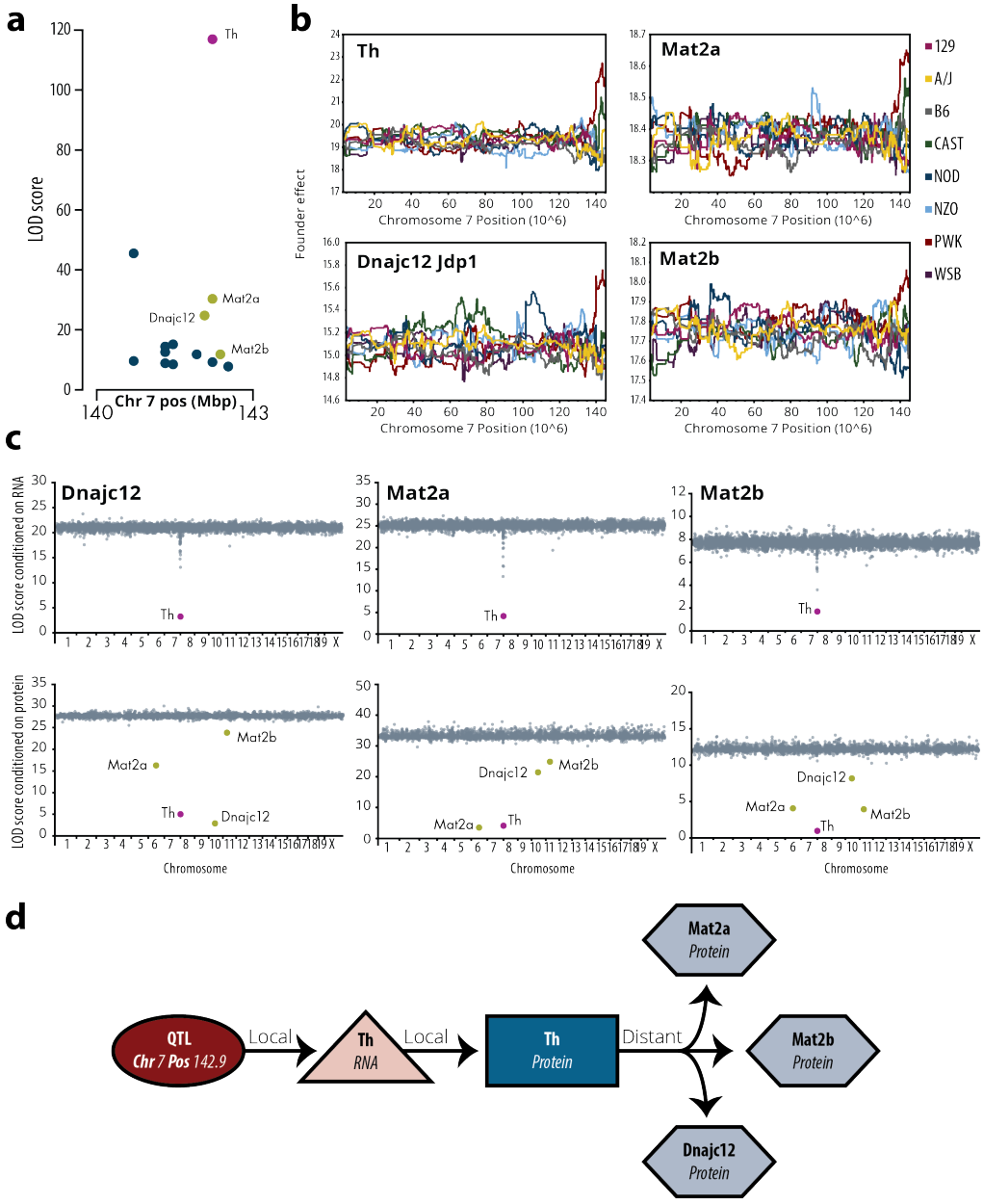
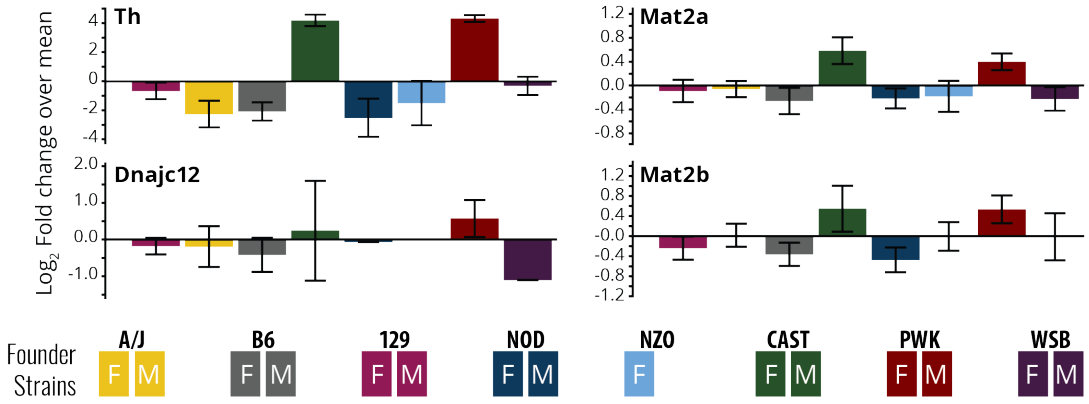
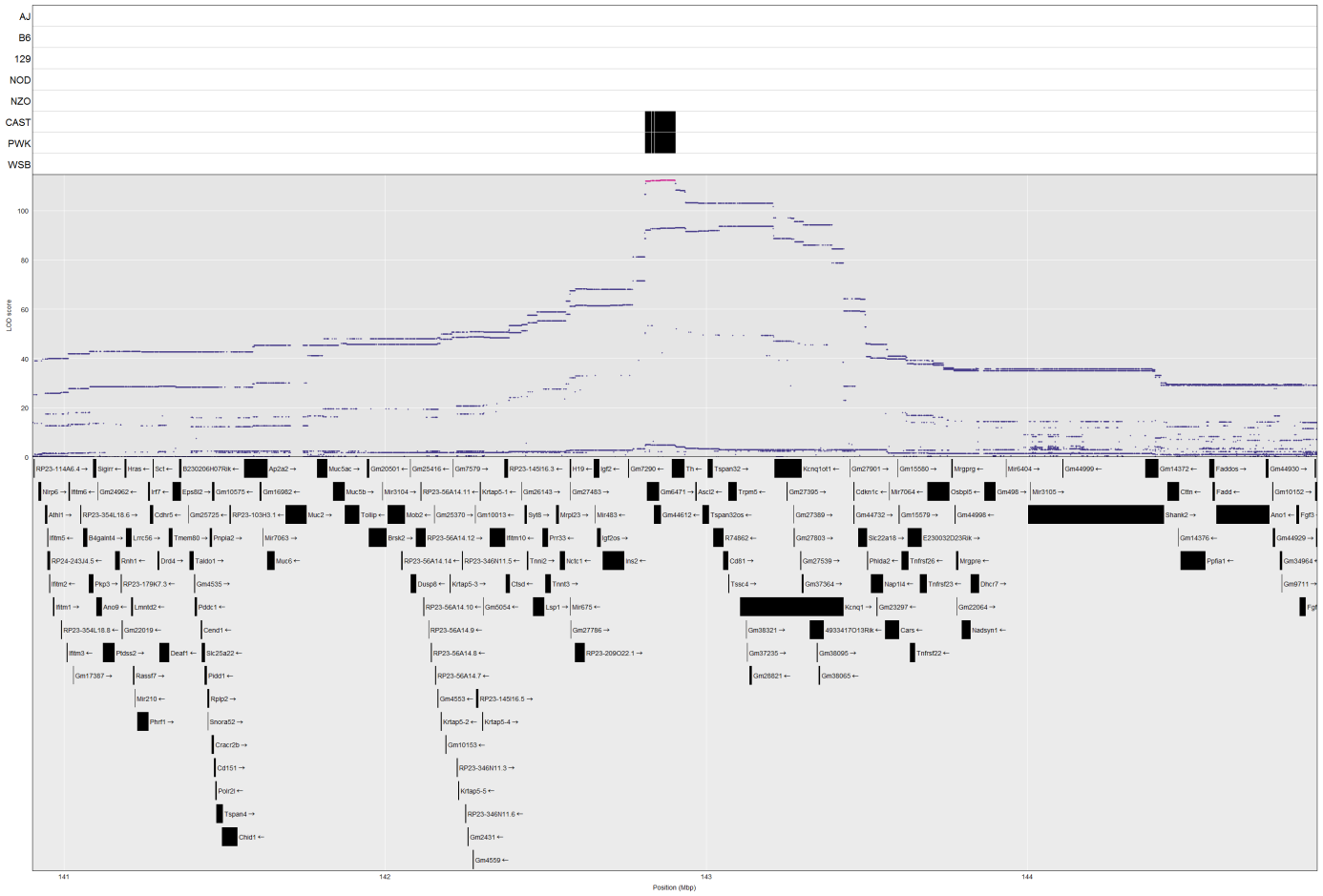


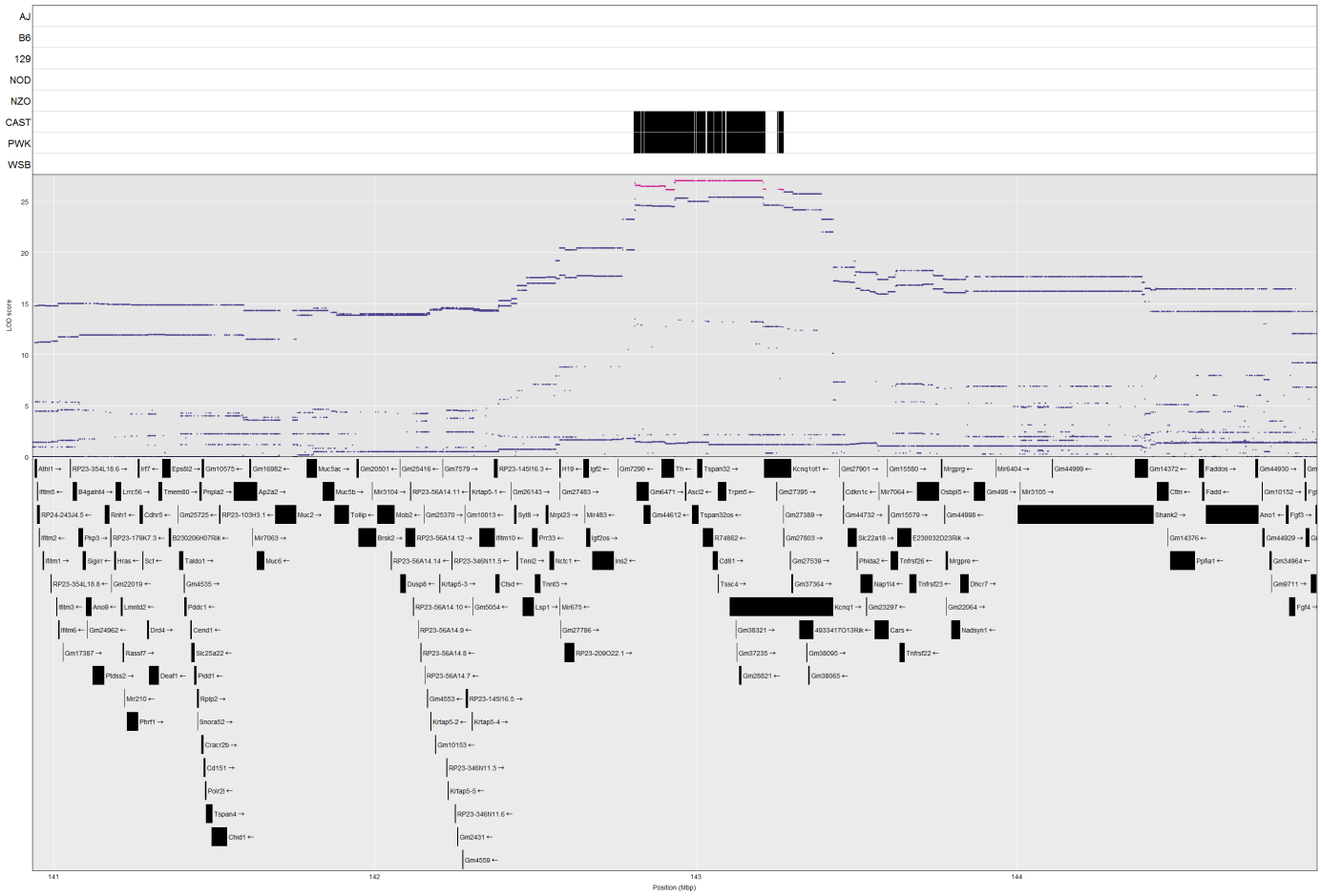
Figure 5.6: Tyrosine hydroxylase locus at chromosome 7. **A** Significant pQTLs mapping to the Th locus: Th local pQTL (pink), Mat2a, Mat2b, and Dnajc12 distant pQTLs (green), and unrelated pQTLs not implicated as tyrosine hyroxylase interactors by mediation analysis or founder effect regulation (blue). **B** Founder effect plots for Th (upper left), Mat2a (upper right), Dnajc12 (lower left), and Mat2b (bottom right) around the locus of interest..



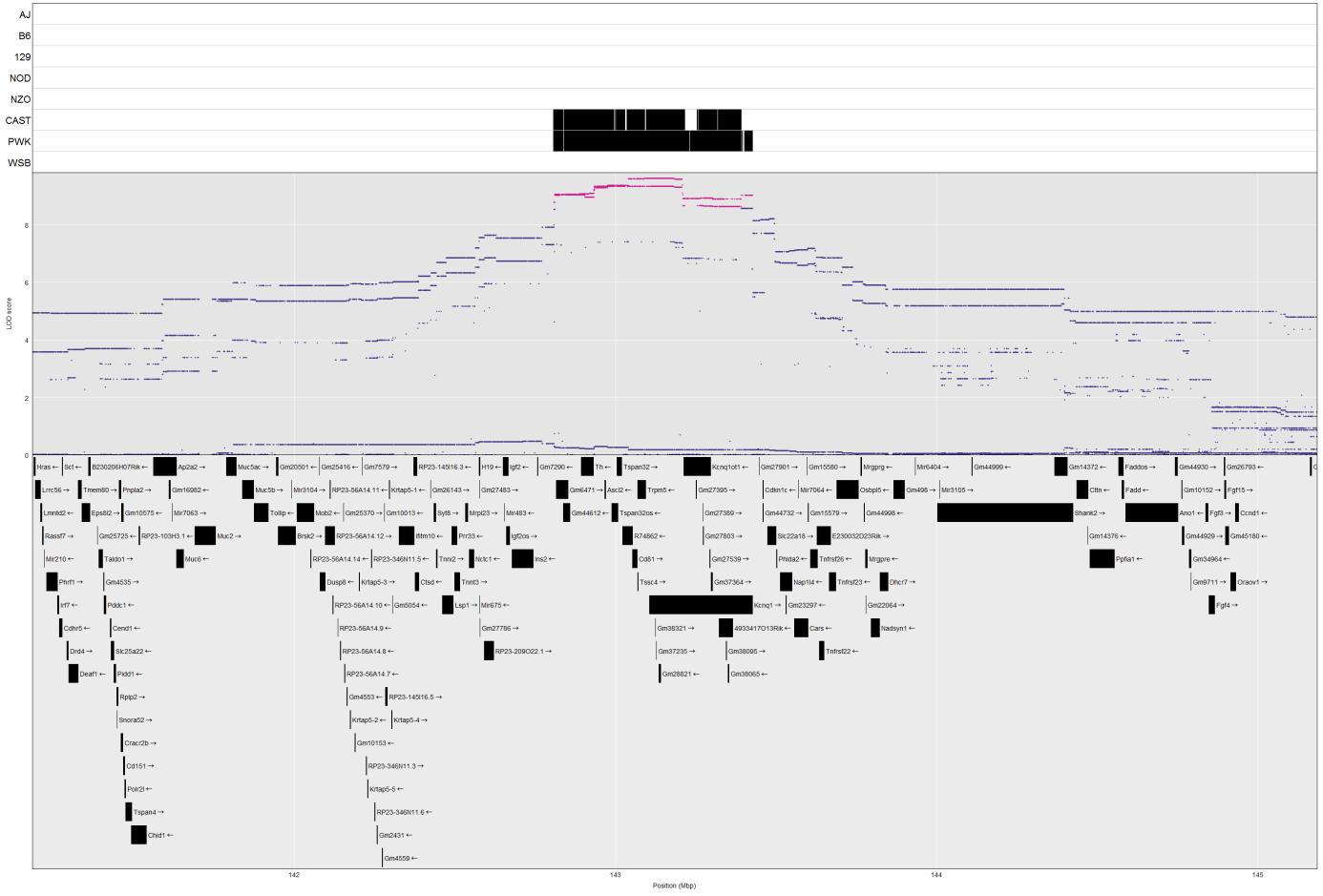
Supplementary Figure S5.9: Founder mouse islet proteomics for proteins at Th locus. Protein expression changes in Th (top left), Mat2a (top right), Dnajc12 (bottom left), and Mat2b (bottom right) in the eight founder mouse strains. N = 8 (4 male, 4 female) except NZO (n=4, female only). Error bars ±SD.[29].



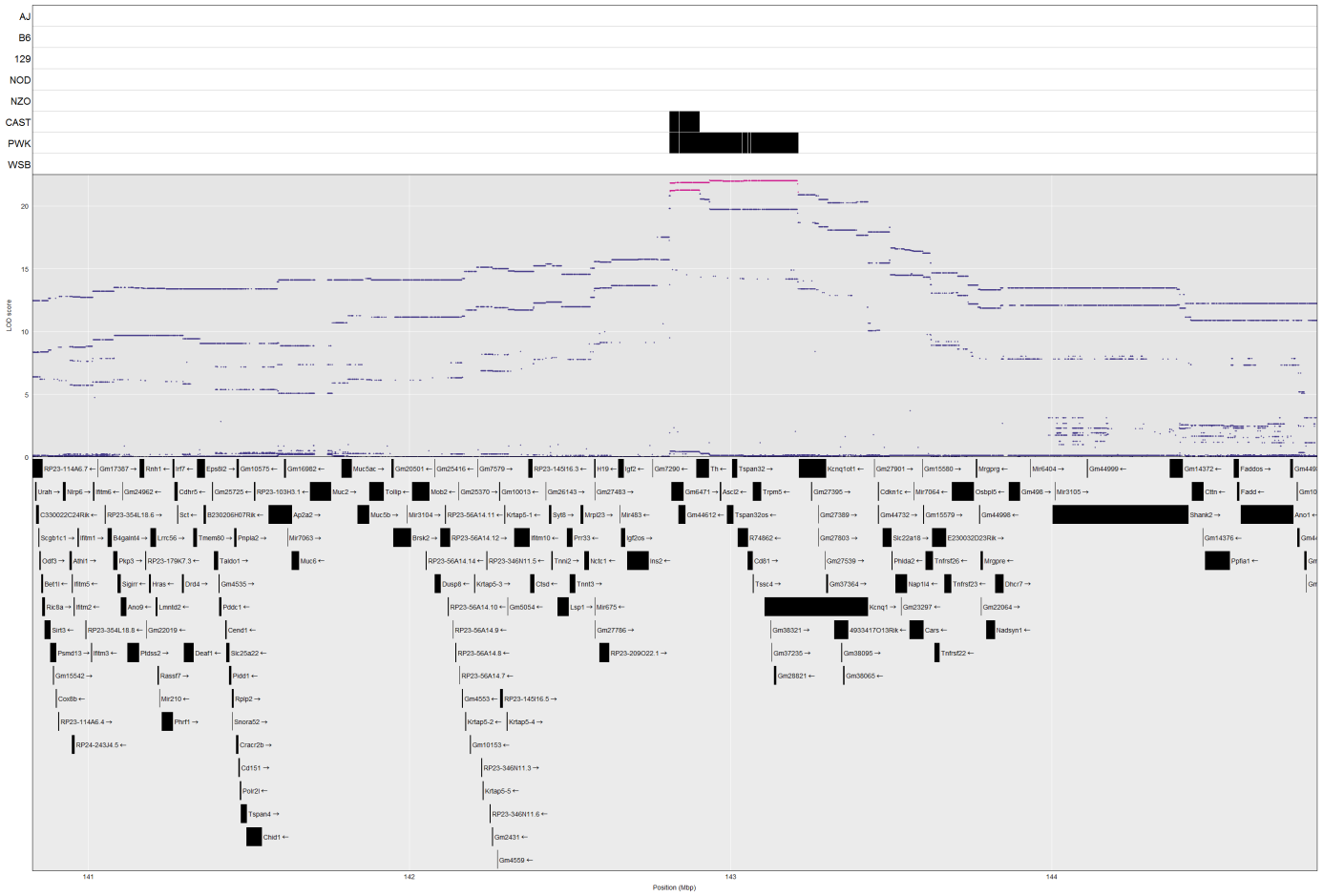
Supplementary Figure S5.10: SNP associations of Th pQTL at chromosome 7. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.



Supplementary Figure S5.11: SNP associations of Mat2a pQTL at chromosome 7. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.



Supplementary Figure S5.12: SNP associations of Mat2b pQTL at chromosome 7. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.



Supplementary Figure S5.13: SNP associations of Dnajc12 pQTL at chromosome 7. The top panel shows the founders associated with significant SNPs. Significant SNPs are highlighted (pink) in the second panel. Gene names and lengths found in this location on the genome are outlined in the bottom panel.

SNP	Alleles	Pos (Mb)
rs220539846	C|T	142.811086
rs239419706	A|G	142.811399
rs33824286	A|G	142.82729
rs262058187	C|A	142.832157
rs33824151	C|T	142.856708
rs579617249	T|C	142.857115
rs263134361	C|T	142.863655
rs219757258	G|C	142.863749
rs33822307	C|T	142.888042
rs263421536	G|A	142.888352
rs234511411	A|C	142.902887

Supplementary Figure S5.14: SNP associations at the tyrosine hydroxylase locus chromosome 7. This table includes the SNPs denoted as significant in SNP association analysis for all four interacting proteins at the chromosome 7 locus. Included are SNP ID, allele effect, and position in Mb.

Figure 5.6: C Mediation plots for Dnajc12 (left), Mat2a (center), and Mat2b (right), with LOD scores conditioned on both transcripts (top) and proteins (bottom). LODs conditioned on Th (pink) are highlighted in both the transcript and protein plots, and LODs conditioned on Mat2a, Mat2b, and Dnajc12 (green) are also highlighted in the protein plots. At the locus of interest, Th causes the greatest drop in LOD score. When Mat2a, Mat2b, or Dnajc12 are conditioned on themselves, the LOD score characteristically drops near zero. **D** Model for medation of Mat2a, Mat2b, and Dnajc12 through Th transcript and protein at chromosome 7.

using a fraction of the material[28]. When comparing our islet pQTL with both eQTL in the same tissue (**Fig. 5.3**), and the same QTL type in a different tissue (**Fig. 5.4**), we have shown that our islet pQTL analysis provides a unique view of how the genetic landscape affects protein expression. In fact, while the overlap in identified gene products between transcript and protein is higher than in the Chick et al. liver study, the number of overlapping QTLs is much lower (though the percentage of local vs. distant QTLs among the overlap is similar)[2]. High-level statistical analyses of the significant pQTLs such as mediation and SNP association can identify co-regulated molecules and drive follow-up biochemical studies to confirm hypothesized interactions. We have present two such loci of interest; the dataset as a whole is rich with other possibilities. Because the focus is on pancreatic islets, and the mice were metabolically challenged with a HF/HS diet, one might assume that the significant pQTLs are all related to type II diabetes or other metabolic disorders, but the statistical methods are agnostic to disease state, so any dysfunction or differential protein regulation in the pancreatic islets could potentially be captured[30–33].

One significant cluster of pQTLs detected in this study comprises the three members of

the synuclein family. These proteins are best studied in neurons, where they are implicated in the progression of both Alzheimer's disease and Parkinson's disease[34]. It is perhaps unsurprising, then, that they show strong locus-specific expression changes in pancreatic islets, since incidence of diabetes is well-correlated with an increased likelihood of comorbidity with neurodegenerative disorders through a yet unknown mechanism or mechanisms[35,36]. There is evidence that alpha-synuclein is involved in the regulation of insulin secretion via interaction with KATP channels on the insulin granule[37], and in humans, low serum alpha-synuclein positively correlates with increased insulin resistance[38]. If the trend follows in mice, the increased expression of synucleins in CAST-like mice could help explain the strain's particular insulin sensitivity[29]. A previous study has proposed beta-synuclein as a regulator of alpha-synuclein aggregation in neurons, which agrees our data supporting beta-synuclein as a regulator of not only alpha-synuclein but also gamma-synuclein within the islets[39]. Because all three synucleins have human homologues[34], further investigation into the mechanism and effects of these proteins in mouse islets using the SNPs suggested by our association data as a starting point could provide information specifically relevant to human disease etiology.

Following a lead generated in our previous islet proteomics experiments, we investigated overlapping pQTLs at the tyrosine hydroxylase locus (**Fig. 5.6**). Differential expression of Mat2a, Mat2b, and Dnajc12 in CAST- and PWK-like mice was shown to be mediated by tyrosine hydroxylase at the protein and transcript level. These interactions are completely novel; the only known interactions are between Mat2a and Mat2b. We have previously

described Th's role in the regulation of insulin secretion via an alternate route of dopamine synthesis that is specific to PWK and CAST strains[29]. Mat2a is the catalytic subunit of methionine adenosyltransferase A, which synthesizes the cofactor S-adenosylmethionine (SAM) from methionine and ATP; Mat2b is the regulatory subunit[40]. SAM is required for the inactivation of catecholamines such as dopamine via catechol O-methyltransferase (COMT) – COMT transfers a methyl group from SAM onto dopamine to form 3-methoxytyramine (3-MT), which we measure directly in a CAST mice in a previous study. In CAST mouse islets, which contain higher levels of dopamine than standard B6 mouse islets, and in the high-dopamine positive control, 3-MT was present in measurable amounts, while it was not detectable in the B6 contro[29]. The role of Dnajc12 relative to tyrosine hydroxylase is less clear; it is a chaperone protein of the Hsp40 protein family involved in protein folding and complex assembly, but no specificity has yet been described[41]. Perhaps it is involved in the formation of the methionine adenosyltransferase complex. Further study of these proteins and how they interact is needed to elucidate specific mechanisms of action, beginning with perturbation of the associated SNPs (**Supp. Fig. S5.14**).

Future directions

We have so far integrated QTL dataset from islet proteomics with transcript QTLs from the same tissue. To complete the multi-omics story, we plan to integrate comprehensive plasma metabolomic and lipidomic QTL analysis to determine how gene expression in the islet affects secreted biomolecules. The plasma lipid data comprises 1,733 features,

641 of which are identified, and includes 648 significant lipid QTLs (lQTLs) ($P < 0.1$); the plasma metabolite data comprises 332 features, 63 of which are identified, and includes approximately 19 significant metabolite QTLs (mQTLs) (LOD$\geqslant$8). Much like with distant pQTLs and eQTLs, regulation of mQTLs and lQTLs can be investigated using mediation analysis to determine genetic drivers of the effects. We will be conditioning the small-molecule QTLs on both islet transcripts and islet proteins. Potential outcomes could include identification of novel regulatory pathways, functional annotation of genes, and correlation of enzymes with the small molecules whose synthesis or degradation they catalyze. The dataset as a whole would represent a resource to direct further study, a way to generate leads for follow-up biochemical assays, knockout models, and more targeted 'omics' analyses.

Methods

Animal husbandry and islet isolation The cohort of Diversity Outbred (DO) mice were obtained from the Jackson Labs (stock no. 009376). For animal husbandry and islet isolation methods, see Keller et al. Materials and Methods[27]. For this work, waves 2 through 5 were used.

Sample preparation Islet samples from all batches were randomized and prepared in batches of ~88. Along with each batch of DO islet samples, between 6 and 8 islet standards were prepared. Islet standards were generated by pooling islets suspended in 8 M urea/10 mM tris(2-carboxyethyl)phosphine (TCEP)/40 mM chloroacetamide (CAA)/100mM tris

(pH = 8) from each of the eight DO founder strains. Individual aliquots of the pooled islets were portioned out and digested along with each DO batch for quality control purposes.

DO islet aliquots were resuspended in 6 M urea with 10mM TCEP, 40mM CAA and 100mM tris (pH = 8), and lysed via sonication in an ice bath. DO and pooled standard lysates were transferred to 96-well plates then diluted to 1.6 M urea with 50mM tris (pH = 8). Samples were digested overnight at room temperature with trypsin (Promega) at a ratio of 1:50 enzyme to protein, then desalted using Strata X 96-well cartridges (Phenomenex Strata-X Polymeric Reversed Phase 10 mg/1 mL). Desalting plates were equilibrated with 1 mL 100% acetonitrile (ACN) followed by 1 mL of 0.2% formic acid. Samples were acidified with TFA and loaded onto the equilibrated Strata X columns, which were then washed with 0.75 mL 0.2% formic acid. Peptides were eluted into clean 96-well plates with 0.75 mL 80% ACN, dried, and reconstituted in 0.2% formic acid. Peptide concentration was measured prior to MS analysis using the Pierce Quantitative Colorimetric Peptide Assay (Thermo).

LC-MS/MS analysis See "Islet proteomics reveals genetic variation in dopamine production resulting in altered insulin secretion" chapter for LC-MS/MS methods; in this work, 1 μg of peptides were loaded on column for each sample.

Pooled islet pre-fractionation A 100 μg aliquot of the pooled islet standard was digested overnight with 1:50 trypsin and desalted on a 60 mg StrataX polymeric reverse phase column. Peptides were dried down and reconstituted in 100 μL 0.2% formic acid. These

peptides were separated across an XBridge Peptide BEH C18 column (130 Å, 3.5 μm, 4.6 mm x 150 mm, Waters). Mobile phase A consisted of 10 mM ammonium formate pH 10 and mobile phase B consisted of 10 mM ammonium formate pH in 80% methanol. The flow rate was 0.8 mL/min for a 25 minute gradient, during which 12 fractions were collected using a 1260 Infinity II HPLC configured with an analytical-scale fraction collector (Agilent). Fractions were transferred to clean vials and dried in a SpeedVac (Thermo). Samples were resuspended in 0.2% formic acid and each fraction was analyzed across the same 120 min. LC-MS/MS method used in the above DO sample

Database searching Raw data was searched in MaxQuant (version 1.5.6.5) against a Mus musculus uniprot database including isoforms (downloaded 15 May 2017). Samples that were prepared and run together (including the islet standard fractions) were segregated into separate parameter groups. FDR was set to 1% at both the PSM and the protein level. A minimum of one peptide was required for quantification. All DO samples and single-shot standards were quantified using MaxLFQ; fractionated standards were not quantified. The match-between-runs feature was used, and MS/MS spectra were not required for quantification[42–46].

pQTL mapping The dataset was reduced to include only proteins identified in $\geqslant$ 50% of the samples, comprising 5,433 proteins. Data were log transformed and missing values were imputed using a Bayesian PCS missing value estimator (bpca) in the R package pcaMethods.

The data were normalized to remove batch effects using the ComBat algorithm with batch and sex as covariates. Normalized data were mapped with the R package QTL2, which uses a Haley-Knott regression analysis, with sex, generation, and animal husbandry batch (not sample preparation batch) as additive covariates[47–49].

Statistical significance thresholds were calculated using permutation testing via the QTL2 R package. For each protein, 1,000 permutations were performed. Significant pQTLs were called 'local' if they map within +/- 10 Mb of the midpoint of their coding gene. Anything mapping outside that metric were called 'distant'[49].

Mediation analysis was performed using the R package r/intermediate with input of either the complete proteome dataset or the complete transcriptome dataset as potential mediators and individual significant distant pQTLs as targets. Z-scores are calculated from the resulting conditioned LOD scores, and molecules are called potential mediators if they precipitate a drop in LOD score of at least six[27,50–52].

References

[1] N. Riley, A. Hebert, and J. Coon, "Proteomics Moves into the Fast Lane," *Cell Systems*, vol. 2, pp. 142–143, mar 2016.

[2] J. M. Chick, S. C. Munger, P. Simecek, E. L. Huttlin, K. Choi, D. M. Gatti, N. Raghupathy, K. L. Svenson, G. A. Churchill, and S. P. Gygi, "Defining the consequences of genetic variation on a proteome-wide scale," *Nature*, vol. 534, pp. 500–505, jun 2016.

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

[4] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "The One Hour Yeast Proteome," *Molecular & Cellular Proteomics*, vol. 13, pp. 339–347, jan 2014.

[5] A. L. Richards, A. S. Hebert, A. Ulbrich, D. J. Bailey, E. E. Coughlin, M. S. Westphall, and J. J. Coon, "One-hour proteome analysis in yeast," *Nature Protocols*, vol. 10, pp. 701–714, apr 2015.

[6] A. C. Peterson, J.-P. Hauschild, S. T. Quarmby, D. Krumwiede, O. Lange, R. A. S. Lemke, F. Grosse-Coosmann, S. Horning, T. J. Donohue, M. S. Westphall, J. J. Coon, and J. Griep-Raming, "Development of a GC/Quadrupole-Orbitrap Mass Spectrometer, Part I: Design and Characterization," *Analytical Chemistry*, vol. 86, pp. 10036–10043, oct 2014.

[7] J. A. Stefely, N. W. Kwiecien, E. C. Freiberger, A. L. Richards, A. Jochem, M. J. P. Rush, A. Ulbrich, K. P. Robinson, P. D. Hutchins, M. T. Veling, X. Guo, Z. A. Kemmerer, K. J. Connors, E. A. Trujillo, J. Sokol, H. Marx, M. S. Westphall, A. S. Hebert, D. J. Pagliarini, and J. J. Coon, "Mitochondrial protein functions elucidated by multi-omic mass spectrometry profiling," *Nature Biotechnology*, vol. 34, pp. 1191–1197, nov 2016.

[8] T. C. Burch, G. Isaac, C. L. Booher, J. S. Rhim, P. Rainville, J. Langridge, A. Baker, and J. O. Nyalwidhe, "Comparative Metabolomic and Lipidomic Analysis of Phenotype Stratified Prostate Cells," *PLOS ONE*, vol. 10, p. e0134206, aug 2015.

[9] P. Y. Kao, K. H. Leung, L. W. Chan, S. P. Yip, and M. K. Yap, "Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions," *Biochimica et Biophysica Acta (BBA) - General Subjects*, vol. 1861, pp. 335–353, feb 2017.

[10] M. R. van der Sijde, A. Ng, and J. Fu, "Systems genetics: From GWAS to disease pathways," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, pp. 1903–1909, oct 2014.

[11] L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, S. Ecker, A. Datta, D. Richardson, F. Burden, D. Mead, A. L. Mann, J. M. Fernandez, S. Rowlston, S. P. Wilder, S. Farrow, X. Shao, J. J. Lambourne, A. Redensek, C. A. Albers, V. Amstislavskiy, S. Ashford, K. Berentsen, L. Bomba, G. Bourque, D. Bujold, S. Busche, M. Caron, S.-H. Chen, W. Cheung, O. Delaneau, E. T. Dermitzakis, H. Elding, I. Colgiu, F. O. Bagger, P. Flicek, E. Habibi, V. Iotchkova, E. Janssen-Megens, B. Kim, H. Lehrach, E. Lowy, A. Mandoli, F. Matarese, M. T. Maurano, J. A. Morris, V. Pancaldi, F. Pourfarzad, K. Rehnstrom, A. Rendon, T. Risch, N. Sharifi, M.-M. Simon, M. Sultan, A. Valencia, K. Walter, S.-Y. Wang, M. Frontini, S. E. Antonarakis, L. Clarke, M.-L. Yaspo, S. Beck, R. Guigo, D. Rico, J. H. Martens, W. H. Ouwehand, T. W. Kuijpers,

D. S. Paul, H. G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, and N. Soranzo, "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells," *Cell*, vol. 167, pp. 1398–1414.e24, nov 2016.

[12] J. Adamski and K. Suhre, "Metabolomics platforms for genome wide association studies—linking the genome to the metabolome," *Current Opinion in Biotechnology*, vol. 24, pp. 39–47, feb 2013.

[13] P. NOSIL, D. J. FUNK, and D. ORTIZ-BARRIENTOS, "Divergent selection and heterogeneous genomic divergence," *Molecular Ecology*, vol. 18, pp. 375–402, feb 2009.

[14] B. C. Collard and D. J. Mackill, "Marker-assisted selection: an approach for precision plant breeding in the twenty-first century," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, pp. 557–572, feb 2008.

[15] J. HOLLAND, "Genetic architecture of complex traits in plants," *Current Opinion in Plant Biology*, vol. 10, pp. 156–161, apr 2007.

[16] T. Würschum, "Mapping QTL for agronomic traits in breeding populations," *Theoretical and Applied Genetics*, vol. 125, pp. 201–210, jul 2012.

[17] E. J. Chesler, D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, V. M. Philip, B. H. Voy, C. T. Culiat, D. W. Threadgill, R. W. Williams, G. A. Churchill, D. K. Johnson, and K. F. Manly, "The Collaborative Cross at Oak Ridge National Laboratory:

developing a powerful resource for systems genetics," *Mammalian Genome*, vol. 19, pp. 382–389, jun 2008.

[18] G. A. Churchill, D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie, J. Beatty, W. D. Beavis, J. K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K. W. Broman, K. J. Buck, E. Buckler, M. Burmeister, E. J. Chesler, J. M. Cheverud, S. Clapcote, M. N. Cook, R. D. Cox, J. C. Crabbe, W. E. Crusio, A. Darvasi, C. F. Deschepper, R. W. Doerge, C. R. Farber, J. Forejt, D. Gaile, S. J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N. L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H.-C. Hsu, F. A. Iraqi, B. Ivandic, H. J. Jacob, R. C. Jansen, K. J. Jepsen, D. K. Johnson, T. E. Johnson, G. Kempermann, C. Kendziorski, M. Kotb, R. F. Kooy, B. Llamas, F. Lammert, J.-M. Lassalle, P. R. Lowenstein, L. Lu, A. Lusis, K. F. Manly, R. Marcucio, D. Matthews, J. F. Medrano, D. R. Miller, G. Mittleman, B. A. Mock, J. S. Mogil, X. Montagutelli, G. Morahan, D. G. Morris, R. Mott, J. H. Nadeau, H. Nagase, R. S. Nowakowski, B. F. O'Hara, A. V. Osadchuk, G. P. Page, B. Paigen, K. Paigen, A. A. Palmer, H.-J. Pan, L. Peltonen-Palotie, J. Peirce, D. Pomp, M. Pravenec, D. R. Prows, Z. Qi, R. H. Reeves, J. Roder, G. D. Rosen, E. E. Schadt, L. C. Schalkwyk, Z. Seltzer, K. Shimomura, S. Shou, M. J. Sillanpää, L. D. Siracusa, H.-W. Snoeck, J. L. Spearow, K. Svenson, L. M. Tarantino, D. Threadgill, L. A. Toth, W. Valdar, F. P.-M. de Villena, C. Warden, S. Whatley, R. W. Williams, T. Wiltshire, N. Yi, D. Zhang, M. Zhang, F. Zou, and Complex Trait Consortium, "The Collaborative Cross, a community resource for

the genetic analysis of complex traits," *Nature Genetics*, vol. 36, pp. 1133–1137, nov 2004.

[19] G. A. Churchill, D. M. Gatti, S. C. Munger, and K. L. Svenson, "The diversity outbred mouse population," *Mammalian Genome*, vol. 23, pp. 713–718, oct 2012.

[20] R. W. Logan, R. F. Robledo, J. M. Recla, V. M. Philip, J. A. Bubier, J. J. Jay, C. Harwood, T. Wilcox, D. M. Gatti, C. J. Bult, G. A. Churchill, and E. J. Chesler, "High-precision genetic mapping of behavioral traits in the diversity outbred mouse population," *Genes, Brain and Behavior*, vol. 12, pp. 424–437, jun 2013.

[21] K. L. Svenson, D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill, "High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population," *Genetics*, vol. 190, pp. 437–447, feb 2012.

[22] A. J. F. King, "The use of animal models in diabetes research.," *British journal of pharmacology*, vol. 166, pp. 877–94, jun 2012.

[23] K. Srinivasan and P. Ramarao, "Animal models in type 2 diabetes research: an overview.," *The Indian journal of medical research*, vol. 125, pp. 451–72, mar 2007.

[24] W. T. Cefalu, "Animal Models of Type 2 Diabetes: Clinical Presentation and Pathophysiological Relevance to the Human Condition," *ILAR Journal*, vol. 47, pp. 186–198, jan 2006.

[25] C. Ionescu-Tirgoviste, P. A. Gagniuc, E. Gubceac, L. Mardare, I. Popescu, S. Dima, and M. Militaru, "A 3D map of the islet routes throughout the healthy human pancreas," *Scientific Reports*, vol. 5, p. 14634, dec 2015.

[26] J. Jo, M. Y. Choi, and D.-S. Koh, "Size distribution of mouse Langerhans islets.," *Biophysical journal*, vol. 93, pp. 2655–66, oct 2007.

[27] M. P. Keller, P. Simecek, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, A. T. Broman, D. M. Gatti, M. Vincent, R. Bacher, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie, "Genetic architecture of gene regulation in pancreatic islets from diversity outbred mice,," *Genetics*, vol. Under review, 2018.

[28] L. F. Waanders, K. Chwalek, M. Monetti, C. Kumar, E. Lammert, and M. Mann, "Quantitative proteomic analysis of single pancreatic islets.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 18902–7, nov 2009.

[29] K. A. Mitok, E. C. Freiberger, K. L. Schueler, M. E. Rabaglia, D. S. Stapleton, N. W. Kwiecien, P. A. Malec, A. S. Hebert, A. T. Broman, R. T. Kennedy, M. P. Keller, J. J. Coon, and A. D. Attie, "Islet proteomics reveals genetic variation in dopamine production resulting in altered insulin secretion," *Journal of Biological Chemistry*, vol. Under review, 2018.

[30] A. Heydemann, "An Overview of Murine High Fat Diet as a Model for Type 2 Diabetes Mellitus," *Journal of Diabetes Research*, vol. 2016, pp. 1–14, jul 2016.

[31] M. Sumiyoshi, M. Sakanaka, and Y. Kimura, "Chronic intake of high-fat and high-sucrose diets differentially affects glucose intolerance in mice.," *The Journal of nutrition*, vol. 136, pp. 582–7, mar 2006.

[32] Z.-H. Yang, H. Miyahara, J. Takeo, and M. Katayama, "Diet high in fat and sucrose induces rapid onset of obesity-related metabolic syndrome partly through rapid response of genes involved in lipogenesis, insulin signalling and inflammation in mice," *Diabetology & Metabolic Syndrome*, vol. 4, p. 32, jul 2012.

[33] T. Ishimoto, M. A. Lanaspa, C. J. Rivard, C. A. Roncal-Jimenez, D. J. Orlicky, C. Cicerchi, R. H. McMahan, M. F. Abdelmalek, H. R. Rosen, M. R. Jackman, P. S. MacLean, C. P. Diggle, A. Asipu, S. Inaba, T. Kosugi, W. Sato, S. Maruyama, L. G. Sánchez-Lozada, Y. Y. Sautin, J. O. Hill, D. T. Bonthron, and R. J. Johnson, "High-fat and high-sucrose (western) diet induces steatohepatitis that is dependent on fructokinase," *Hepatology*, vol. 58, pp. 1632–1643, nov 2013.

[34] C. Lavedan, "The synuclein family.," *Genome research*, vol. 8, pp. 871–80, sep 1998.

[35] O. Ojo and J. Brooke, "Evaluating the Association between Diabetes, Cognitive Decline and Dementia.," *International journal of environmental research and public health*, vol. 12, pp. 8281–94, jul 2015.

[36] E. Cereda, M. Barichella, C. Pedrolli, C. Klersy, E. Cassani, R. Caccialanza, and G. Pez-

zoli, "Diabetes and risk of Parkinson's disease: a systematic review and meta-analysis.," *Diabetes care*, vol. 34, pp. 2614–23, dec 2011.

[37] X. Geng, H. Lou, J. Wang, L. Li, A. L. Swanson, M. Sun, D. Beers-Stolz, S. Watkins, R. G. Perez, and P. Drain, "α-Synuclein binds the K _{ATP} channel at insulin-secretory granules and inhibits insulin secretion," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 300, pp. E276–E286, feb 2011.

[38] G. Rodriguez-Araujo, H. Nakagami, Y. Takami, T. Katsuya, H. Akasaka, S. Saitoh, K. Shimamoto, R. Morishita, H. Rakugi, and Y. Kaneda, "Low alpha-synuclein levels in the blood are associated with insulin resistance," *Scientific Reports*, vol. 5, p. 12081, dec 2015.

[39] M. Hashimoto, E. Rockenstein, M. Mante, M. Mallory, and E. Masliah, "β-Synuclein Inhibits α-Synuclein Aggregation: A Possible Role as an Anti-Parkinsonian Factor," *Neuron*, vol. 32, pp. 213–223, oct 2001.

[40] C. L. Quinlan, S. E. Kaiser, B. Bolaños, D. Nowlin, R. Grantner, S. Karlicek-Bryant, J. L. Feng, S. Jenkinson, K. Freeman-Cook, S. G. Dann, X. Wang, P. A. Wells, V. R. Fantin, A. E. Stewart, and S. K. Grant, "Targeting S-adenosylmethionine biosynthesis with a novel allosteric inhibitor of Mat2A," *Nature Chemical Biology*, vol. 13, pp. 785–792, may 2017.

[41] M. W. W. W. Mouse Genome Database (MGD) at the Mouse Genome Informatics

website, The Jackson Laboratory, Bar Harbor, "Dnajc12 MGI Mouse Gene Detail - MGI:1353428 - DnaJ heat shock protein family (Hsp40) member C12."

[42] J. Cox and M. Mann, "MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification," *Nature Biotechnology*, vol. 26, pp. 1367–1372, dec 2008.

[43] J. Cox, A. Michalski, and M. Mann, "Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors," *Journal of The American Society for Mass Spectrometry*, vol. 22, pp. 1373–1380, aug 2011.

[44] C. Schaab, T. Geiger, G. Stoehr, J. Cox, and M. Mann, "Analysis of high accuracy, quantitative proteomics data in the MaxQB database.," *Molecular & cellular proteomics : MCP*, vol. 11, p. M111.014068, mar 2012.

[45] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, and M. Mann, "Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.," *Molecular & cellular proteomics : MCP*, vol. 13, pp. 2513–26, sep 2014.

[46] S. Tyanova, T. Temu, A. Carlson, P. Sinitcyn, M. Mann, and J. Cox, "Visualization of lc-ms/ms proteomics data in maxquant," *PROTEOMICS*, vol. 15, no. 8, pp. 1453–1456, 2015.

[47] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[48] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig, "pcamethods – a bioconductor package providing pca methods for incomplete data," *Bioinformatics*, vol. 23, pp. 1164–1167, 2007.

[49] K. W. Broman, *qtl2*, 2017. http://kbroman.org/qtl2, https://github.com/rqtl/qtl2.

[50] P. Simecek, S. Munger, and G. Churchill, *intermediate: eQTl/pQTL Mediation analysis*, 2017. R package version 0.2-4.

[51] R. M. Baron and D. A. Kenny, "The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations.," *Journal of personality and social psychology*, vol. 51, pp. 1173–82, dec 1986.

[52] M. S. Fritz and D. P. MacKinnon, "Required Sample Size to Detect the Mediated Effect," *Psychological Science*, vol. 18, pp. 233–239, mar 2007.

Chapter 6

CONCLUSIONS

In this work I have described the application of quantitative proteomics methods to four different biological questions. The first used pre-fractionation with size exclusion chromatography to probe the effects of both a knockdown model and a chemical inhibitor on VCP interactors; the study spanned two separate 72-sample experiments (12 SEC fractions per condition in biological triplicate) and required post-processing of the label-free quantification results using weighted averaging tailored specifically to capture an accurate picture of the mass shifts characteristic of protein-protein binding and unbinding. The second application spans the largest sample set, over 1,000 yeast samples for proteomics analysis – 174 knockout strains across two growth conditions in biological triplicate – with an eye toward functional annotation of unknown genes. The scope of this undertaking required

an enormous amount of planning, sample preparation, instrument time, and computational resources. The data were ultimately integrated with lipidomic and metabolomic data on the same gene knockouts which led to novel functional descriptions for several previously-orphaned genes as well as generation of a valuable resource to drive future research.

The smallest dataset described herein is the founder mouse strain islet proteomics study, which comprised only 58 samples (8 mouse strains, two sexes, and $\geqslant$ biological triplicate). These data were collected initially to determine whether the mouse strains were distinguishable based on islet proteome alone; we found that not only did the proteomes sort by strain via unsupervised clustering, but that strain-specific trends in protein expression levels were evident in the resulting heat maps. The final study also focused on mouse islet proteomics, but in this casewais geared toward correlations with genomic changes across the population. We analyzed 383 DO mouse islet proteomes and work is ongoing to describe the complete set of quantitative trait linkages in these data along with transcript, lipid, and metabolite QTLs through a variety of bioinformatics approaches.

Current capacities for large-scale proteomics studies tend to be limited by the interplay between analysis time and desire for proteomic depth. Technology has not yet reached the point where whole mammalian proteomes are quantifieable in a single-shot analysis; depth can be increased by pre-fractionation with an orthogonal chromatographic solid-phase, but the trade-off is a sharp increase in analytical time per sample. The greater the sample size, the more impactful that increase is to the study outlook as a whole. However, depending

on the biological system and question being investigated, whole proteome coverage may not be required to make meaningful conclusions. For example, in the diversity outbred islet study, due to the scarcity and biological variability of pancreatic islets, we were able to use single-shot analyses because at an average of ~5,500 proteins quantified per sample, we were still detecting between 3 and 5 times more proteins than in most prior islet studies. We detected a broad enough range of expressed proteins to move our understanding of islet function forward an appreciable amount.

The day is coming when fast, complete proteome measurements will be a matter of course, and perhaps very large-scale studies will be more commonplace and less of a material and time commitment, but improvements across the board to instrument sensitivity, chromatographic separations, and especially to informatic tools would be required to get to that point. Until then, the major concern should not be "Can I?" when it comes to undertaking a larger-scale proteomics study, but "Should I?". Given no restrictions on instrument usage or personnel commitment, we are at the point that nearly any conceivable large-scale experiment is theoretically possible. But the reality is, more data is not always better data. A clear end goal that would tangibly benefit from a broader scope of conditions, or from an increased number of replicates to strengthen statistical power, should be a requirement. Organizing such datasets remains huge hurdle, and to attempt to do so in the absence of a specific experimental intention would be a Sisyphean task. That's not to say they should be avoided altogether: the Y3K project would not have been nearly as successful if so wide a net had not been cast to capture as many gene functions within the

knockout pool, and the lack of biological replicates within the DO resource necessitates a large sample size to capture any statistically significant linkage events. These are examples where very large-scale experiments have paid off beautifully, but I hope to also have shown that given an appropriate hypothesis, slightly more manageable sample sizes can also yield high-quality and impactful data.

COLOPHON

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