UNDERSTANDING MAMMALIAN BIOLOGY AND DISEASE THROUGH TISSUE-SPECIFIC PROTEIN-PROTEIN INTERACTION NETWORKS

by

Michael Skinnider

B.ArtsSc., McMaster University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Genome Science and Technology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Understanding mammalian biology and disease through tissue-specific protein-protein interaction networks

submitted by Michael Skinnider in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Genome Science and Technology

Examining Committee:

Dr. Leonard Foster, Professor, Department of Biochemistry & Molecular Biology, UBC
Supervisor

Dr. Pamela Hoodless, Professor, Department of Medical Genetics, UBC
Supervisory Committee Member

Dr. Joerg Gsponer, Associate Professor, Department of Biochemistry & Molecular Biology, UBC
Supervisory Committee Member

Dr. Gabriela Cohen-Freue, Associate Professor, Department of Statistics, UBC
University Examiner

Dr. Paul Pavlidis, Professor, Department of Psychiatry, UBC
University Examiner

Additional Supervisory Committee Members:

Dr. Sara Mostafavi, Assistant Professor, Department of Medical Genetics, UBC
Supervisory Committee Member

Abstract

Biological functions are mediated by the dynamic organization of DNA, RNA, proteins, and other biomolecules in complex networks of interactions. Efforts to chart the network of biologically relevant macromolecular interactions—the "interactome"—therefore occupy a central position in the endeavour to understand the biochemical basis of human physiology, and its perturbation in disease. However, existing interactome maps are incomplete, even for well-studied organisms. Moreover, the dynamics of the interactome in response to cellular stimuli and across normal physiological contexts remain incompletely understood.

This thesis considers the application of a quantitative proteomic approach, protein correlation profiling (PCP), to map the interactome in its native physiological context. I explore computational methods for the analysis of PCP data, and describe their application to infer a dataset of protein-protein interactions from seven mouse tissues.

In Chapter 2, I studied the dominant paradigm used to analyze PCP data, which entails the use of supervised machine-learning methods to infer interaction networks from these complex datasets. I found that one widely used strategy needlessly biases network inference towards highly studied proteins and away from novel interactions between functionally un-connected proteins.

In Chapter 3, I applied the methods studied in Chapter 2 to a newly collected, *in vivo* PCP dataset. I used the same machine-learning approach to infer tissue-specific protein-protein interaction networks for seven mouse tissues. I then analyzed these tissue interactome networks to uncover insights about protein function, network evolution, and human disease.

Collectively, the work described in this thesis provides a framework to understand the rewiring of the protein-protein interaction network across physiological conditions using PCP.

Lay Summary

Proteins interact with one another to perform normal cellular functions, and disruption of these interactions results in many human diseases. Consequently, defining the complete network of protein-protein interactions in humans and other organisms has been a longstanding goal of biomedical research. Many experimental methods have been developed to identify new protein-protein interactions. However, most rely on genetically manipulated cell lines, or artificial expression of proteins in yeast. As a result, whether the interactions identified by these methods actually take place within human tissues has remained unclear. This dissertation employs a method known as protein correlation profiling to systematically map networks of protein-protein interactions in seven mouse tissues. Computational methods to infer networks of interacting proteins are discussed, and applied to infer tissue-specific networks of protein-protein interactions. These networks are shown to offer new perspectives on cellular biology.

Preface

The research program described in this dissertation was identified and designed by myself and my supervisor, Leonard Foster. I performed all parts of the research and analysis of the research data, except as noted below.

A version of chapter 2 has been published. **Skinnider, M.A.,** Stacey, R.G., Foster, L.J. (2018) Genomic data integration systematically biases interactome mapping. *PLoS Computational Biology*, 14, e1006474. I conducted all of the experiments described in this publication and drafted the manuscript. Greg Stacey contributed to experimental design. Leonard Foster supervised the project. All authors contributed to editing the manuscript.

A version of chapter 3 has been submitted. **Skinnider, M.A.***, Scott, N.E.*, Prudova, A., Kerr, C.H., Stoynov, N., Stacey, R.G., Chan, Q.W.T., Rattray, D., Gsponer, J., Foster, L.J. An atlas of protein-protein interactions across mouse tissues. I conducted all bioinformatic and statistical analyses, and drafted the manuscript. Nichollas Scott performed SEC-PCP-SILAM. Anna Prudova, Craig Kerr, Queenie Chan, and Dave Rattray performed immunoprecipitations. Craig Kerr additionally performed ribosome purification. Nikolay Stoynov performed mass spectrometry. Greg Stacey contributed to bioinformatic analysis. Joerg Gsponer and Leonard Foster co-supervised the project. All authors contributed to editing the manuscript.

Experiments conducted in chapter 3 were approved by the University of British Columbia Animal Care Committee (A13-0094).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AP-MS	Affinity purification-mass spectrometry
AUC, AUROC	Area under the receiver operating characteristic curve
CF-MS	Co-fractionation mass spectrometry
EGF	Epidermal growth factor
FDR	False discovery rate
FPKM	Fragments per kilobase million
GO	Gene Ontology
IEF	Isoelectric focusing
IEX	Ion exchange chromatography
IP	Immunoprecipitation
LC	Literature curation
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LFQ	Label-free quantification
PCA	Protein complementation assay
PCP	Protein correlation profiling
PMID	PubMed identifier
PPI	Protein-protein interaction
RF	Random forest
ROC	Receiver operating characteristic curve
SEC	Size exclusion chromatography
SILAC	Stable isotope labelling in cell lines

SILAM	Stable isotope labelling in mammals
SVM	Support vector machine
Y2H	Yeast two-hybrid

Acknowledgements

First, I am deeply grateful for the mentorship of my supervisor, Leonard Foster. The freedom Leonard gave me to develop and pursue my own scientific ideas profoundly shaped my experience of completing a PhD. I would not have enjoyed my time as a graduate student nearly as much without his patient and generous support.

I thank my committee members, Pamela Hoodless, Joerg Gsponer, and Sara Mostafavi, for their advice. I also thank Torsten Nielsen, Lynn Raymond, and Jane Lee from the MD/PhD program for their guidance.

I owe a special debt to Nichollas Scott for collecting the dataset described in Chapter 3, without which this thesis would not exist.

I am very grateful for all the sources of funding that have supported my research over the past six years. I would particularly like to acknowledge support from the Canadian Institutes of Health Research and the UBC Faculty of Medicine for the work presented in this thesis.

Most of all, I thank my family, Jessie, and Alpha.

Dedication

For Jessie

Chapter 1: Introduction

1.1 Protein-protein interaction networks

Proteins are central protagonists of life at the molecular level. Cells depend on proteins to form a stable cytoskeleton, to wind and unwind DNA, to communicate with other cells, and to produce new proteins, among many other functions. These tasks, however, are rarely carried out by individual proteins acting in isolation. Instead, proteins work cooperatively with other proteins to carry out key cellular functions. Frequently, this cooperation takes the form of a physical interaction with one or more other proteins. Collectively, the complete set of protein-protein interactions that occur within a cell form a large and complex network (Schwikowski et al., 2000). This network is also dynamic, rewiring itself to carry out normal cellular functions such as cell division, and to allow the cell to respond to stimuli, such as intercellular signalling or infection (Ideker and Krogan, 2012).

Around the turn of the 21st century, mapping the complete network of protein-protein interactions in humans and select model organisms (most notably yeast) emerged as a major focus of biomedical research (Bartel et al., 1996; Fromont-Racine et al., 1997; Gavin et al., 2002; Ho et al., 2002; Ito et al., 2000, 2001; Uetz et al., 2000; Walhout et al., 2000). A major driver of this interest was the advent of genome sequencing technology, which had begun to produce comprehensive inventories of amino acid sequences encoded in the genomes of diverse organisms. The functions of many of these newly discovered proteins, however, were initially unknown. At the same time, several independent studies noted that proteins with similar known functions displayed a statistical tendency to interact with one another in large-scale protein-protein interaction networks (Oliver, 2000; Schwikowski et al., 2000). These observations

implied that placing proteins of unknown function into the context of their physical interactions with other proteins could provide a high-throughput approach to elucidate their cellular functions. One widely used analogy suggested that whereas genome sequencing had provided a 'parts list' for cellular function, systematic mapping of cellular interaction networks would provide a 'wiring diagram' of the connections responsible for the execution of these emergent functions (VanBogelen et al., 1999). Further impetus for systematic interaction mapping projects came from the observation that many disease-causing mutations exert their pathological effects by disrupting protein-protein interactions (Sahni et al., 2015; Wang et al., 2012). These findings opened up the possibility that understanding precisely which interactions might be disrupted by a given mutation could uncover new therapeutic approaches to treat disease, or new molecular tools to improve the diagnosis of genetic disorders.

1.2 Experimental identification of protein-protein interactions

The goals of understanding cellular function, and its perturbation in disease, have motivated an enormous investment of resources into mapping protein-protein interactions over the past two decades. These efforts have produced systematic interactome maps for many eukaryotes, including human (Ewing et al., 2007; Hein et al., 2015; Huttlin et al., 2015, 2017; Rolland et al., 2014; Rual et al., 2005; Stelzl et al., 2005; Wang et al., 2011), yeast (Babu et al., 2012; Gavin et al., 2002, 2006; Ho et al., 2002; Ito et al., 2000, 2001; Krogan et al., 2006; Schwikowski et al., 2000; Tarassov et al., 2008; Uetz et al., 2000; Yu et al., 2008), fly (Formstecher et al., 2005; Giot et al., 2003; Guruharsha et al., 2011), worm (Li et al., 2004; Simonis et al., 2009), and plant (Arabidopsis Interactome Mapping Consortium, 2011; Jones et al., 2014). Historically, the majority of these efforts have relied on one of two biochemical techniques to experimentally map

protein-protein interactions: that is, yeast two-hybrid (Y2H) or affinity purification-mass spectrometry (AP-MS). The Y2H assay involves tagging two proteins of interest with either half of a reporter protein, typically a transcription factor, and expressing the tagged proteins in yeast. A physical interaction between the two proteins within the yeast nucleus leads to the reconstitution of the transcription factor, enabling transcription of a reporter gene. Conversely, in AP-MS, the tagged protein is expressed in cells from the organism of interest, typically under systems that aim to approximate its normal *in vivo* expression. The tag is used to purify the protein of interest (the 'bait') using affinity capture, and mass spectrometry-based proteomics is subsequently employed to identify and quantify the co-purified proteins (the 'prey'). These two techniques yield complementary pictures of the interactome, with Y2H designed to probe direct biophysical interactions between predefined protein pairs, and AP-MS aiming to recover all constituents of multi-protein complexes (Luck et al., 2017; Wodak et al., 2013).

A third and complementary approach to these large-scale systematic screens is to carefully review many thousands of papers that have reported protein-protein interactions on a smaller scale. This approach, often referred to as 'literature curation' (LC), can produce interaction maps of comparable quality to those detected by systematic screens (Cusick et al., 2009; Reguly et al., 2006; Salwinski et al., 2009). However, a downside of curation approaches is that the resulting interactome maps exhibit strong biases towards proteins that many scientists have chosen to study, resulting in a large 'blank space' in the resulting interaction network that is not present in systematic interactomes mapped using high-throughput methods (Rolland et al., 2014).

The labour and cost required to systematically screen for interactions at the proteome scale using Y2H or AP-MS, and the inherent biases of approaches based on literature curation,

have also prompted the development of computational methods to predict protein-protein interactions (Keskin et al., 2016). These computational methods employ a wide range of features to predict the presence or absence of an interaction between any given pair of proteins, ranging from molecular dynamics simulations (Cunningham et al., 2020; Smith and Sternberg, 2002), to amino acid sequence-based machine learning (Shen et al., 2007), to three-dimensional protein structures (Meyer et al., 2018; Wang et al., 2012), to leveraging functional genomics datasets such as gene expression or genome sequencing data (Fortelny et al., 2017; Jansen et al., 2003; Ramani et al., 2008), among many others. Computational methods have also been developed to specifically predict context-specific protein-protein interaction networks. The dominant approach among these methods has been to incorporate context-specific gene or protein expression to refine a 'static' interactome network (de Lichtenberg et al., 2005). A related approach consists of using large-scale, tissue-specific gene or protein expression datasets to identify protein pairs with correlated patterns of abundance, which provides a basis for the inference of functional association, if not necessarily physical interaction (Pierson et al., 2015; Saha et al., 2017).

More recently, thermal proximity co-aggregation (TPCA) has emerged as a complementary tool for monitoring protein-protein interactions in a more native physiological context (Becher et al., 2018; Dai et al., 2018; Jarzab et al., 2020; Mateus et al., 2020; Tan et al., 2018). TPCA is based on the cellular thermal shift assay (CETSA), in which soluble protein abundance is measured in lysates after heating them to varying degrees (Jafari et al., 2014; Martinez Molina and Nordlund, 2016). This heating causes proteins to unfold and precipitate, leading to the appearance of a melting curve when the remaining soluble protein is quantified. The addition of a ligand to the protein of interest produces a shift in this curve, which motivated the original development of CETSA to study drug-target interactions. The discovery that proteins

in the same macromolecular complex also tend to display correlated melting curves (Tan et al., 2018) led to the proposal that the assay could be adapted to monitor protein-protein interactions. To date, however, TPCA has not been shown to enable the *de novo* inference of novel interactions.

While Y2H and AP-MS have historically been the dominant biochemical techniques used to experimentally map protein-protein interaction networks, several weaknesses of these assays have been noted. A practical constraint is that applying either method on the scale of the human proteome, with approximately 20,000 protein-coding genes, is extremely laborious. A second issue is that the introduction of a protein tag into the native protein sequence has the potential to disrupt biologically relevant protein-protein interactions, or alter the subcellular localization of the target protein observed *in vivo* (Kim et al., 2019; Werner et al., 2009). A third issue, related to the second, is that both methods are limited in their capacity to capture interactions under physiologically relevant contexts. The Y2H assay involves expressing fusion proteins in the yeast nucleus, a very different environment than that of the human cell, and one in which post-translational modifications that may mediate conditional interactions are notably absent (Grossmann et al., 2015). AP-MS more accurately reflects native cellular context, with proteins assayed in the host species and often expressed at *in vivo* levels, but its reliance on genetically transformed cell lines represents a limitation to understanding the interactome of specific cell types, tissues, or physiological contexts. Finally, both methods are challenging to apply in species that are not amenable to genetic manipulation, such as polyploid plants (McWhite et al., 2020). Some of these limitations have been overcome by other assays for detecting protein-protein interactions, such as the bimolecular fluorescence complementation assay (Kerppola,

2008), which can be applied to monitor interactions in living cells. However, these assays have not been successfully applied to systematically screen for interactions at the proteome scale.

1.1 Protein correlation profiling

To address these shortcomings, several new approaches to mapping protein-protein interaction networks have emerged in recent years, some of which have been applied on a large scale (Rattray and Foster, 2019; Richards et al.; Snider et al., 2015). One such technique, which is the primary focus of this thesis, is protein correlation profiling (PCP, also known as co-fractionation mass spectrometry or CF-MS). PCP was originally described as a method to map the subcellular localization of proteins within the cell (Andersen et al., 2003; Foster et al., 2006). In this formulation, cellular lysates were separated using density gradient centrifugation in order to produce fractions enriched for particular cellular organelles. Each of these fractions was subsequently analyzed using mass spectrometry to identify the proteins within that fraction, and quantify their relative abundance. This allowed the cellular localization of poorly understood proteins to be characterized in high throughput, via comparison to established protein markers of particular cellular compartments.

In 2012, two studies showed that adapting the method used to fractionate cellular extracts, but maintaining the workflow of mass spectrometric analysis of individual fractions, could enable the mapping of protein-protein interactions and protein complexes instead of cellular organelles (Havugimana et al., 2012; Kristensen et al., 2012). Kristensen *et al.* used size exclusion chromatography (SEC) to separate protein complexes based on their hydrodynamic volume, whereas Havugimana *et al.* used ion exchange chromatography (IEX) and isoelectric focusing (IEF) to separate protein complexes based on their charge and isoelectric point,

respectively. In both studies, correlations between quantitative protein abundances across a gradient of size, charge, or isoelectric point were high for proteins known to interact, allowing the identification of novel protein-protein interactions between protein pairs with similarly high correlations. Surprisingly, both studies showed that this approach could be used to generate maps of the interactome of similar quality to those achieved using much more labor-intensive approaches.

In addition to mapping protein-protein interactions under a single cellular condition, Kristensen *et al.* used isotopic labeling to simultaneously quantify proteins from two different populations of cells: one unstimulated, and the second stimulated with epidermal growth factor (Kristensen et al., 2012). This experiment heralded one of the primary advantages of PCP: that is, its ability to capture dynamic or context-specific rearrangements in the interactome. This advantage of PCP stands in marked contrast to Y2H and AP-MS, which each monitor protein-protein interactions within a single cellular context of questionable physiological relevance. Accordingly, PCP has been applied to monitor rearrangements in the interactome during the innate immune response (Kerr et al., 2020), apoptosis (Scott et al., 2017), the cell cycle (Heusel et al., 2020), *Salmonella* infection (Scott et al., 2015), and the circadian cycle (Gorka et al., 2019), among other contexts. These studies have demonstrated substantial 'rewiring' of the protein-protein interaction network in response to normal cellular stimuli. In turn, these observations raise the question of whether such rewiring might be widespread even between normal physiological conditions—for instance, between different cell types or healthy tissues—despite having gone unnoticed by classical experimental techniques.

The power and flexibility of PCP for protein-protein interaction network mapping come at the expense of new challenges in computational data analysis. Inferring a network of

interacting proteins from a raw PCP dataset requires pinpointing a tiny minority of interacting proteins from a space that grows quadratically with the total number of proteins quantified. Moreover, within this very large search space, a considerable degree of 'chance' co-elution can occur, whereby similar patterns of protein abundance are observed for two proteins for reasons other than their participation in a common protein complex. Early studies applied various different measures of association to distinguish co-eluting proteins within individual replicates, including the Euclidean distance, Pearson correlation, or mutual information (Gazestani et al., 2016; Kirkwood et al., 2013; Kristensen et al., 2012; Scott et al., 2015). However, the availability of multiple biological replicates, or even datasets collected using distinct methods for cellular fractionation, drove the development of more sophisticated approaches to integrate multiple sources of information reflecting protein-protein interactions (Havugimana et al., 2012; Wan et al., 2015). Supervised machine learning has emerged within the field as the standard strategy to this end. In this paradigm, a statistical model (or classifier) is trained to identify interacting protein pairs, using features computed from each replicate as input and a training set constructed from known protein complexes. The classifier is trained in cross-validation to avoid leaking information between the training and test data, and allow for the possibility that some known complexes may not be assembled in a given dataset. Many variations on this general theme have been described, including the structure of the cross-validation procedure, the number and identities of the features computed for each replicate, and the choice of classifier. A more fundamental difference is between approaches that seek to integrate evidence exclusively from experimental PCP data, and approaches that incorporate external datasets. In the latter paradigm, PCP data collected from some biological system of interest is concatenated with existing genomic data that reflects the likelihood of a protein-protein interaction, such as mRNA

coexpression or co-evolution in sequenced genomes. This approach has been widely used in the field (Havugimana et al., 2012; Kastritis et al., 2017; Larance et al., 2016; Pourhaghighi et al., 2020; Wan et al., 2015), with proponents suggesting the incorporation of external genomic datasets can offset the likelihood of 'chance' co-elution (Havugimana et al., 2012).

1.2 Overview of the thesis

The objectives of the work presented in this thesis were twofold: first, to reveal the strengths and weaknesses of the various computational methods that are currently used to analyze PCP data, with a particular focus on methods that integrate external genomic datasets; and second, to apply these computational methods to a newly collected PCP dataset to derive insights into physiological interaction rewiring across mammalian tissues. Below, I briefly review the central aims and key findings of the following chapters.

In Chapter 2, I detail a systematic investigation of the impact of incorporating external genomic datasets in the procedure used to infer networks from PCP data. I used a supervised machine-learning approach (Stacey et al., 2017) to infer protein-protein interaction networks from several published PCP datasets. I then obtained various different genomic datasets that are correlated with the likelihood of a protein-protein interaction from public sources, and concatenated these to the experimental data. The properties of the resulting protein-protein interaction networks were then subjected to a systematic analysis. I found that networks inferred using external genomic data typically exhibited a greater degree of 'functional connectivity:' that is, the tendency for proteins with a shared function to physically interact. However, I used a time-split validation experiment (Sheridan, 2013) to demonstrate that this functional connectivity did not necessarily correlate with the likelihood of experimentally validating an interaction. In

other words, protein-protein interactions inferred from the PCP data alone were just as likely to later be 'discovered' as those inferred from both PCP and external genomic data. This finding implies that interactions discovered from PCP data alone may represent novel connections between proteins that were previously not known to be functionally related. Conversely, I showed that the procedure of integrating external genomic data biased network inference towards well-studied proteins, and disrupted the identification of co-complex interactions. I concluded this work by comparing individual types of external genomic data. I found a subset of data types, including mRNA coexpression (Ramani et al., 2008) and phylogenetic profiles (Pellegrini et al., 1999), that provide the best trade-off between increased functional connectivity and decreased ability to discover novel interactions when integrated into network inference. These specific data types may be useful in scenarios when integrating external datasets is felt to be necessary.

Whereas Chapter 2 investigates the computational methods used to infer protein-protein interaction networks from PCP data, Chapter 3 details the application of these methods to a specific PCP dataset. The adaptation of PCP for *in vivo* experiments enabled the first comprehensive survey of protein-protein interactions across healthy mammalian tissues. I applied a supervised machine-learning approach to infer protein-protein interaction networks for seven mouse tissues from this unprecedented resource. These tissue-specific networks provided an opportunity to address a number of biological questions regarding protein-protein interactions for the first time. For instance, I showed that tissue-specific interactions are particularly likely to be novel, and we show that these also disproportionately involve proteins of unknown function, or for which no interactions were previously known, thereby placing many poorly understood proteins into a functional context. Moreover, I used this resource of experimentally derived interactions to benchmark methods for tissue-specific interactome prediction. This analysis

underscored an important deficiency of these approaches: namely, that changes in protein abundance across tissues are insufficient to predict *in vivo* interactome rewiring. Next, I explored modes of evolution in mouse tissue interactomes. My analysis contrasted an evolutionarily ancient 'core' of the interactome present in all tissues with evolutionarily recent, 'accessory' modules present in individual tissues, and I identified systematic suppression of cross-talk between these modules. I characterized the properties of proteins whose interaction partners are disproportionately rewired across tissues, identifying structural features that facilitate rewiring, and finding that these rewired proteins are subject to multiple convergent programs of tight cellular regulation. Finally, I showed empirically—for the first time—that genes implicated in diseases that selectively impact specific tissues are more tightly interconnected in the interactome networks of those tissues. Collectively, my analysis of this unprecedented resource of PCP data provide a foundation for understanding the organization of the physiological interactome.

Chapter 4 concludes the thesis by discussing some of the strengths and limitations of the research described herein. Looking towards the future, I review some of the key challenges facing the field, and describe some promising research directions to address these challenges.

Chapter 2: Genomic data integration systematically biases interactome mapping[1]

2.1 Introduction

Biological functions are mediated by the dynamic organization of proteins and other biomolecules, including DNA, RNA, and metabolites, into complex networks of interactions (Barabási and Oltvai, 2004). Perturbations of these networks are implicated in human disease (Sahni et al., 2015). Consequently, efforts to chart the network of biologically relevant protein-protein interactions (the "interactome") occupy a central position in the endeavour to understand the biochemical basis of human physiology and disease pathobiology (Barabási et al., 2011; Vidal et al., 2011). Nearly two decades of study have produced initial systematic interactome maps of humans (Hein et al., 2015; Huttlin et al., 2015, 2017; Rolland et al., 2014; Wan et al., 2015) and model organisms. However, traditional methods for interactome mapping, such as yeast two-hybrid (Y2H) or affinity purification-mass spectrometry (AP-MS) require the introduction of tags into all proteins of interest in order to provide a measurable readout (Gavin et al., 2006; Uetz et al., 2000). Such tags are laborious to introduce, and may disrupt the native interactions or localization of the protein (Werner et al., 2009). Furthermore, these methods cannot easily be applied to identify temporal rearrangements in the interactome, instead yielding

[1] A version of chapter 2 has been published. **Skinnider, M.A.,** Stacey, R.G., Foster, L.J. (2018) Genomic data integration systematically biases interactome mapping. *PLoS Computational Biology*, 14, e1006474.

static pictures of the cellular protein interaction network (Kristensen and Foster, 2013; Kristensen et al., 2012).

In response to interest in assembling complete maps of human and model organism interactomes, and in identifying changes in protein-protein interactions in response to perturbation, a number of experimental techniques have emerged to increase the throughput and resolution of interactome mapping using co-migration, also referred to as protein correlation profiling (PCP). Recently, we described an approach that combines PCP with stable isotope labeling by amino acids in cell culture (PCP-SILAC) and size exclusion chromatography (SEC), and applied this method to identify rearrangements in the interactome of HeLa cells following stimulation with epidermal growth factor (EGF) (Kristensen et al., 2012). More extensive fractionation methods have also been employed to identify co-migrating proteins across a wide range of biochemical conditions (Havugimana et al., 2012; Wan et al., 2015). Importantly, although neither SEC-PCP-SILAC nor orthogonal co-migration approaches yield direct evidence of physical protein-protein interactions, they provide a basis for inference of co-complex membership based on correlated protein abundance across conditions designed to separate protein complexes based on their size or other biochemical properties.

Co-migration methods for interactome mapping quantify thousands of proteins across a large number of fractions. Discriminating interacting from non-interacting protein pairs within the resulting complex and noisy proteomic datasets represents a significant computational challenge. Consequently, a number of published computational pipelines incorporate additional sources of evidence supporting the presence or absence of a physical interaction, derived from external genomic datasets, in machine-learning classifiers. Diverse sources of publicly available functional genomics data supporting functional or physical association have been incorporated

into published classifiers: for instance, mRNA co-expression, protein co-evolution, or gene co-citation in published literature abstracts. Published protein-protein interactions, either from previous high-throughput studies or compiled from small-scale experiments, may also be incorporated as sources of evidence, as may interactions between orthologous proteins in other model organisms ("interologs") (Yu et al., 2004).

The popularity of incorporating external genomics datasets in co-migration data analysis attests to the widespread belief that this methodology increases the quality of the resulting interaction networks, by enabling the classifier to more accurately discriminate between true and spurious interactions. Consequently, this strategy has been employed by several large-scale interactome mapping efforts, e.g. (Havugimana et al., 2012; Kastritis et al., 2017; Larance et al., 2016; Wan et al., 2015). However, despite its broad use, the effects of genomic data integration on the global properties of the protein-protein interaction networks recovered from proteomic data has not been rigorously assessed. A major goal of large-scale interactome mapping projects is to discover novel interactions, yet it seems intuitively likely that incorporating information about known interactions, or functional associations, could decrease the power of a classifier to reveal truly novel interactions within experimental datasets. Moreover, publicly available genomics datasets are often biased towards well-studied proteins, and it is unclear whether this bias is propagated into the composition and topology of the resulting interaction networks. Finally, the precise effects of each external functional genomics data type have not been rigorously documented, and individual datasets have been integrated in a largely *ad hoc* manner. The question of which individual datasets should be integrated to optimize network quality therefore remains open. Although co-migration methods can significantly increase the throughput of interactome mapping, they nevertheless require substantial investment of time and

resources to generate. It is therefore critical to ensure that the resulting datasets are analyzed in a manner that balances power to discover novel interactions with the desire to prioritize true positives for further experimental validation.

In the present study, we rigorously evaluate the effects of incorporating external genomic datasets on the quality, topology, and novelty of protein-protein interaction networks recovered from mass spectrometric data. We first apply our framework to analyze co-migration datasets we have produced using SEC-PCP-SILAC. As a baseline, we predict interactions using PrInCE (Stacey et al., 2017), a naive Bayes classifier trained exclusively on dataset-derived features: e.g., the Euclidean distance or Pearson correlation between two chromatograms. Our intent is not to argue that our own experimental and computational pipelines represent universally optimal techniques for detecting protein-protein interactions using the principle of co-migration. Rather, we believe that our methods are sufficiently representative that our conclusions generalize more broadly to other experimental and computational approaches. In support of this argument, we present evidence that our results are qualitatively unchanged when the naive Bayes classifier used to predict interactions in PrInCE is replaced by a support vector machine, or when training classifiers on an alternative selection of data-derived features. In addition, we extend our analysis to recently published co-migration datasets generated by others, using orthogonal experimental methods, to demonstrate that our results apply to data analysis of co-migration experiments in general. We find that, while incorporating external genomic datasets increases the power of the resulting networks to predict protein function, it leads to a substantial decrease in the proportion of novel interactions discovered. Because novel interactions could represent either true undetected interactions or false positive associations, we apply a time-split approach (Sheridan, 2013) to estimate the proportion of true positives among putative novel interactions. Importantly,

we find that novel interactions predicted with or without external genomic datasets are equally likely to be discovered by subsequent studies, suggesting that genomic data integration impedes discovery of novel interactions between proteins without previously known functional associations. Although we find that no single source of external data significantly improves the functional connectivity of inferred networks without introducing a bias towards known interactions, we identify a subset of features that balance these two objectives. Our results reveal a widespread and unappreciated limitation in a methodology used to process proteomic datasets, with implications for efforts to map the human interactome.

2.2 Materials and methods

2.2.1 Co-migration datasets

Our initial analysis incorporated three sets of previously published co-migration experiments performed within our own laboratory using SEC coupled to PCP-SILAC (SEC-PCP-SILAC). These experiments mapped rearrangements in the interactome of HeLa cells to stimulation with EGF (Kristensen et al., 2012) and to infection with *Salmonella enterica* (Scott et al., 2015), as well as rearrangements in the interactome of Jurkat T cells during Fas-mediated apoptosis (Scott et al., 2017). In addition, we analyzed a dataset mapping the interactomes of seven mouse tissues, using PCP coupled to stable isotope labelling in mammals (PCP-SILAM) (Skinnider et al., 2018a). Within each experiment, each condition was analyzed separately (for example, stimulated vs. unstimulated cells), for a total of 13 conditions. In addition, we analyzed a dataset generated using SEC coupled to label-free quantification (LFQ) rather than SILAC (referred to herein as "SEC-LFQ") (Kirkwood et al., 2013), and two datasets generated using extensive biochemical co-fractionation to separate protein complexes (Havugimana et al., 2012; Wan et al.,

2015), in order to evaluate the robustness of our conclusions to experimental methodology. These datasets were obtained from the supplementary material of the corresponding publications, with the exception of the Wan et al. dataset, which was obtained from the supporting website (http://metazoa.med.utoronto.ca; only human experiments were analyzed here). Protein isoforms and proteins quantified in three or fewer fractions were filtered from the Kirkwood et al. data (Kirkwood et al., 2013), leaving a total of 4,519 proteins for further analysis. Proteins quantified in three or fewer fractions, and proteins quantified in only a single experiment, were filtered from the Wan et al. data (Wan et al., 2015), and mapped to UniProt identifiers, leaving a total of 3,895 proteins for further analysis. A complete list of the datasets used in this study is given below in **Table 2.1**.

Dataset name	**Reference**	**Description**	**Protein quantification**
Kristensen	Kristensen et al., 2012	HeLa cells stimulated with epidermal growth factor (EGF) and subjected to size exclusion chromatography	SILAC
Apoptosis	Scott et al., 2017	Jurkat T cells stimulated with Fas and undergoing apoptosis, subjected to size exclusion chromatography	SILAC
Havugi-mana	Havugimana et al., 2012	Nuclear and cytoplasmic extracts from HEK293 and HeLa cells, subjected to ion exchange chromatography and isoelectric focusing	Label-free quantification
HeLa	Scott et al., 2015	HeLa cells infected with *Salmonella enterica* and subjected to size exclusion chromatography	SILAC
Kirkwood	Kirkwood et al., 2013	U2OS cells subjected to size exclusion chromatography	Label-free quantification
Tissues	Skinnider et al., 2018	Seven mouse tissues, subjected to size exclusion chromatography	SILAM
Wan	Wan et al., 2015	CB660, G166, and HEK293 cells subjected to ion exchange chromatography	Label-free quantification

Table 2.1 PCP datasets used to evaluate the impact of genomic data integration.

2.2.2 External genomics datasets

We evaluated the impact of integrating nine external genomic features at the protein pair level into classifiers designed to identify protein-protein interactions (PPIs) from PCP data, including: (i) messenger RNA (mRNA) co-expression, (ii) phylogenetic profiles, (iii) domain-domain interactions, (iv) co-citation in literature abstracts, (v) gene fusion, (vi) gene proximity, and three datasets of previously known PPIs curated from (vii) small-scale experiments, (viii) two-hybrid (2H) screens, or (ix) co-immunoprecipitation (co-IP) experiments. These genomic datasets were chosen in order to mimic as closely as possible the selection of features that have been incorporated into previous genome-scale integrated functional networks (Lee et al., 2011; Mostafavi et al., 2008; Taşan et al., 2015), and used to train classifiers designed to identify PPIs within co-migration datasets (Havugimana et al., 2012; Kastritis et al., 2017; Larance et al., 2016; Lee et al., 2011; Wan et al., 2015). For example, both Havugimana et al. (Havugimana et al., 2012) and Wan et al. (Wan et al., 2015) incorporated co-evolution, literature co-citation, mRNA co-expression, gene neighborhoods, and PPIs derived from AP/MS, Y2H, and literature curation as features in machine-learning classifiers to analyze biochemical co-fractionation datasets. Similarly, Kastritis et al. (Kastritis et al., 2017) incorporated gene neighborhood, gene fusion, co-expression, and literature co-occurrence to identify PPIs in a thermophilic eukaryote. Likewise, Larance et al. included high-throughput and small-scale interactions from PPI databases in their classifier to identify membrane protein complexes in a combined cross-linking and co-migration experiment (Larance et al., 2016), while Crozier et al. incorporated gene neighborhood, gene fusion, mRNA co-expression, literature co-occurrence, and experimental and literature-curated protein-protein interactions from STRING to identify protein complexes in *Trypanosoma brucei* from co-migration data (Crozier et al., 2017). We withheld functional data

derived from the Gene Ontology (GO) in order to validate the functional connectivity of the resulting interaction networks, as detailed further below.

RNA expression data from healthy human and mouse tissues was obtained from the Bgee database (SIB Swiss Institute of Bioinformatics Members, 2016), and coexpression was calculated from normalized (FPKM) expression values using the Pearson correlation coefficient. Phylogenetic profiles were constructed by mapping human or mouse proteins to 272 other species using InParanoid (Ostlund et al., 2010), and calculating the Pearson correlation coefficient between binarized presence/absence values (Fortelny et al., 2017). Interacting domains identified within three-dimensional structure data were obtained from the 3did database (Mosca et al., 2014), and Pfam domain annotations for the human and mouse proteomes were obtained from the UniProt web server to identify protein pairs possessing known domain-domain interactions (The UniProt Consortium, 2017). Co-citation, gene fusion, and gene proximity scores were obtained from STRING, version 10.5 (Szklarczyk et al., 2017).

To assemble comprehensive datasets of previously published interactions, we systematically compiled interactions from sixteen databases: BIND (Alfarano et al., 2005), BioGRID (Chatr-Aryamontri et al., 2017), DIP (Salwinski et al., 2004), HINT (Das and Yu, 2012), HIPPIE (Alanis-Lobato et al., 2017), HPRD (Keshava Prasad et al., 2009), IID (Kotlyar et al., 2016), InBioMap (Li et al., 2017), MatrixDB (Launay et al., 2015), Mentha (Calderone et al., 2013), MINT (Licata et al., 2012), MPPI (Pagel et al., 2005), NetPath (Kandasamy et al., 2010), PINA (Cowley et al., 2012), Reactome (Fabregat et al., 2016), and WikiPathways (Kutmon et al., 2016). From pathway databases, only the subset of information cataloguing physical PPIs was retained. When available, experimental methods supporting each interaction were recorded using the Molecular Interactions ontology (Hermjakob et al., 2004). Two-hybrid interactions were

selected as the subset of interactions supported by the MI evidence codes containing the string 'two hybrid', while co-IP interactions were selected as the subset of interactions supported by evidence codes containing the string 'coimmunoprecipitation'. In addition, all PubMed identifiers (PMIDs) supporting each interaction were recorded and used to link the interaction to all the year(s) in which it was detected using XML files distributed by PubMed. All gene and protein identifiers were mapped to UniProt accessions using identifier mapping files distributed by UniProt (The UniProt Consortium, 2017). We assembled comprehensive databases of known interactions from human, mouse, worm, fly, and yeast using this procedure, mapping orthologs with InParanoid (Ostlund et al., 2010).

We used the resulting comprehensive catalog of interactions in three ways. First, we used interactions between orthologous protein pairs, detected by either (i) literature curation of small-scale experiments (LC), (ii) yeast two-hybrid (Y2H), or (iii) co-immunoprecipitation (IP) as external genomic features, integrating them into machine-learning classifiers to recover protein interaction networks from co-migration data. We defined small-scale experiments as publications reporting 25 or fewer PPIs. To minimize circularity in interaction prediction, in all cases we excluded interactions from the species of interest (i.e., human or mouse) when training classifiers on co-migration data, including only interologs detected in different species.

Second, we used all published interactions from the species of interest (i.e., human or mouse) to calculate the proportion of interactions in each recovered network that were known (i.e., previously reported in any database), defining the remaining interactions as putatively novel. In this experiment, we used two different datasets for two time-split validation experiments, excluding interactions detected prior to 2017 or prior to 2016, respectively, to calculate the total proportion of previously known interactions.

Third, we used the withheld sets of interactions to estimate the proportion of putatively novel interactions that actually represent true positives by performing a time-split validation (Sheridan, 2013). In time-split validation, rather than partitioning a set of examples into training and test folds randomly, the training set is partitioned based on the date or time at which each exampled was acquired. In this setting, for example, we withhold all interactions detected in 2016 or later as the validation fold. Our rationale in selecting this validation scheme was to provide a reasonable simulation of prospective interactome mapping projects, in which investigators have access to all previously published interactions and are interested in the proportion of predicted interactions that represent true positives. This technique was pioneered in the field of cheminformatics, where it was shown to more accurately reflect the accuracy of predictions in a prospective setting (Chen et al., 2012; Sheridan, 2013). However, time-split validation generalizes to any dataset for which each observation is time-stamped, and can reduce the impact of literature contamination or other unanticipated biases in the training dataset, thereby providing a more accurate estimation of true accuracy in comparison to the overly optimistic approach of random-split cross-validation (Sheridan, 2013). We used two different datasets to conduct two different time-split validation experiments, splitting our dataset of known interactions at those discovered in 2017 or later and in 2016 or later, respectively.

2.2.3 Protein interaction prediction from co-migration data

We first predicted protein interactions using our own computational pipeline, PrInCE (Stacey et al., 2017), a naive Bayes classifier that learns exclusively from dataset-derived features. By default, PrInCE incorporates six features, including the Euclidean distance between protein profiles; the Pearson correlation coefficient between profiles and its P-value; the Pearson

correlation coefficient calculated from 'cleaned' profiles, with missing values replaced by Gaussian noise and single missing values imputed; and the distance in fractions between the maximum values of each profile. PrInCE also attempts to deconvolve each profile into a mixture of one to five Gaussians and includes the minimum Euclidean distance between any pair of Gaussians from the two profiles as a sixth feature. These features were used as input to classifiers designed to handle data from our own SEC-PCP-SILAC experiments, as well as SEC coupled to spectral counts (SEC-LFQ) (Kirkwood et al., 2013). In our analysis of the co-fractionation datasets, we emulated the feature selection of Havugimana *et al*. (Havugimana et al., 2012), calculating weighted cross-correlation and noise model correlation for each fractionation experiment, as well as a co-apex score across all fractionation experiments, as previously described. The Euclidean distance between MS1 intensity profiles was also included as a feature when analyzing the Wan et al. dataset, to emulate the authors' original analysis (Wan et al., 2015).

In all cases, the selected features were used as input to a naive Bayes classifier, optionally supplemented with combinations of one or more features derived from the external genomic datasets described above. We evaluated all possible combinations of zero to nine external genomic features for each co-migration dataset in turn, up to a limit of ten randomly selected combinations for each number of external genomic features. The naive Bayes classifier was trained on the CORUM database (Ruepp et al., 2010) using ten-fold cross-validation, taking intra-complex interactions as true positives and inter-complex interactions as true negatives. Protein pairs were ranked by their median classifier scores across all ten folds, and the top 10,000 pairs were selected to form a predicted interaction network for each experiment, in order to control for the effect of interactome size on our conclusions. We additionally generated

networks by controlling the false discovery rate, filtering networks with fewer than 2,000 interactions to minimize spurious associations caused by very small networks. In addition, we confirmed that our results were qualitatively unchanged when using alternative classifiers, including the random forests (RFs) implementation in the R package 'ranger', and the support vector machines (SVMs) implementation in the R package 'LiblineaR'. Due to increased computational demands, only five-fold cross-validation was used for SVM and RF classifiers.

2.2.4 Functional connectivity

Gene Ontology (GO) terms annotated to each protein were retrieved from the UniProt-GOA database (Dimmer et al., 2012). GO annotations supported only by evidence codes ND, IPI, IEA, and/or NAS were filtered. We used the 'EGAD' R package (Ballouz et al., 2017) to perform all functional connectivity calculations in three-fold cross-validation, considering only GO terms annotated to between 0.5% and 5% of the proteins in the resulting network in order to exclude very broad or specific terms.

2.2.5 Interaction novelty

We calculated the proportion of interactions within each network that had been previously reported by assembling comprehensive catalogs of human and mouse interactions, as described above. To gain insight into the likelihood that novel interactions detected by each method represented true positives, we withheld human interactions that had not been reported prior to 2017, and calculated the proportion of 'novel' interactions predicted from each human dataset that had been discovered in 2017, in a time-split validation approach (Sheridan, 2013). Interactions supported only by the PMIDs of previously published experiments analyzed here

(Havugimana et al., 2012; Kirkwood et al., 2013; Kristensen et al., 2012; Scott et al., 2015, 2017) were excluded from these analyses to eliminate circularity. To identify individual combinations of external genomic datasets that significantly increased the proportion of putative novel interactions that were subsequently discovered in our time-split validation scheme, we compared each combination of external genomic features to the baseline networks recovered solely from dataset-derived features using a Brunner–Munzel test, followed by Bonferroni correction to control the family-wise error rate.

2.2.6 Other outcomes

The 'igraph' package (Csardi and Nepusz, 2006) in R was used to calculate the global clustering coefficient of each network. The number of publications referencing each protein was obtained from the 'gene2pubmed' file distributed by the NCBI (Maglott et al., 2007).

2.2.7 Statistical analysis

Spearman's ρ was used to evaluate associations between number of genomic features and outcomes. To test for differences in functional connectivity between networks recovered with individual features and baseline networks, a two-sided Brunner–Munzel test was performed for each network, and the P-values were aggregated using Fisher's method. To test for differences in proportion of novel interactions between individual feature and baseline networks, P-values from proportion tests performed for each network were aggregated using Fisher's method. Differences between networks in bias towards highly studied proteins or global clustering coefficients were assessed using two-sided Brunner–Munzel tests.

2.3 Results

Published computational pipelines for analysis of co-migration datasets make use of variable numbers and types of external genomic datasets to identify interactions. We sought to systematically evaluate the effects of genomic data integration on the properties of the interactomes recovered from raw proteomics data. The framework for our study was, therefore, as follows: We first integrated variable numbers of external genomic datasets into machine-learning classifiers and used each classifier to predict interactions from 16 co-migration datasets, generated using SEC-PCP-SILAC, SEC coupled to label-free quantification (SEC-LFQ), or biochemical co-fractionation. We then analyzed the following properties for each network: (i) its biological coherence, defined in the following section; (ii) the proportion of interactions within the network that were novel; (iii) the clustering coefficient of the network, a measure of its ability to recover fully connected protein complexes; and (iv) the degree of bias within the network towards highly studied proteins. While none of these outcomes should be taken as a simple indicator of network quality in isolation, in combination they capture relevant properties of the analytical approach used to recover interaction networks from co-migration data. Next, we decomposed the effect of individual genomic features by adding external datasets one at a time and evaluated the same properties of the resulting networks. In addition, since functional genomics datasets exhibit varying degrees of incompleteness, we evaluated the robustness of the recovered networks to variations in the completeness of the training data. Finally, we confirmed our results were robust to the choice of classifier. An overview of our study design is provided in **Figure 2.1**.

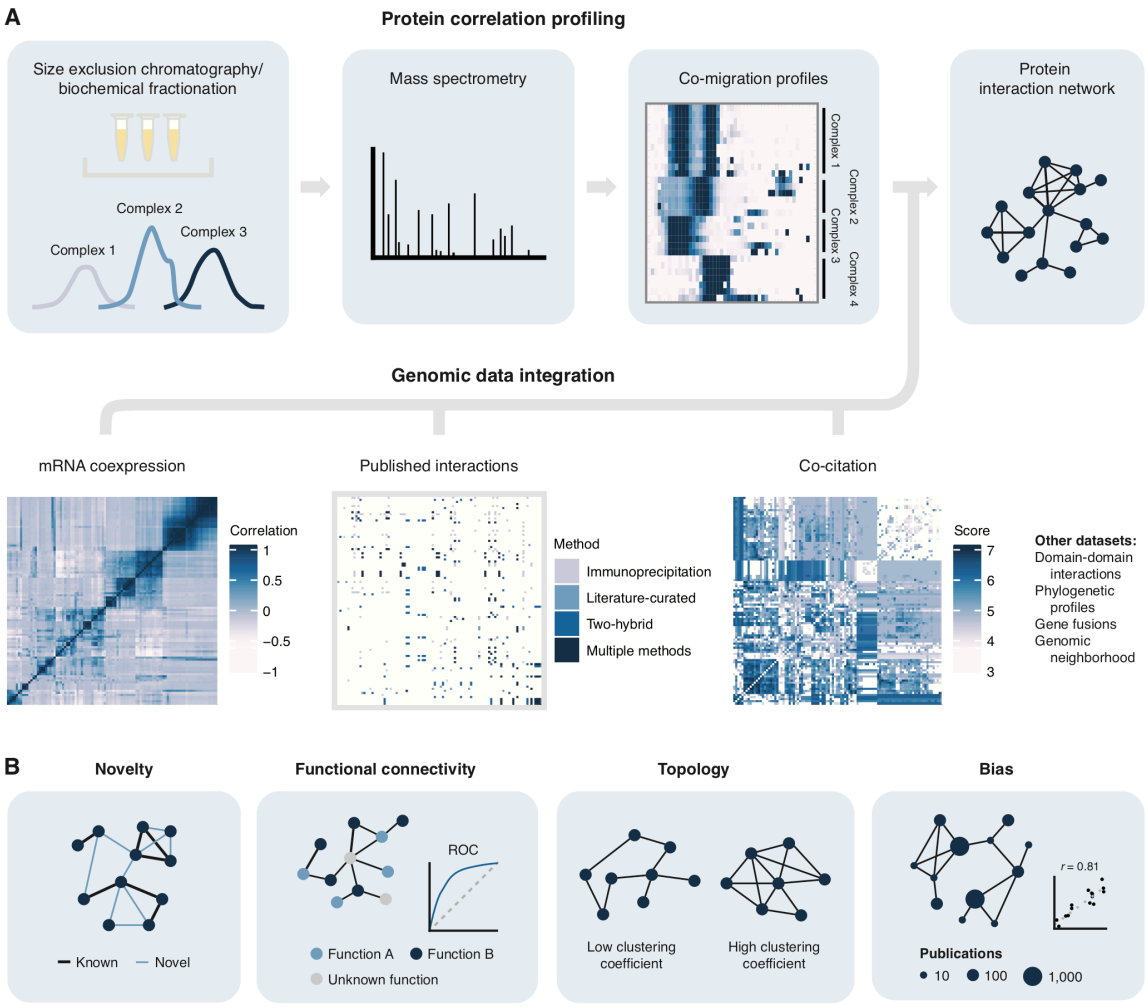


Figure 2.1 Genomic data integration for interactome mapping. (A) Overview of co-migration workflow for interactome mapping and integration of external genomic datasets. (B) Primary outcomes analyzed in the present study.

We began by analyzing the functional connectivity of each network, defined as the degree to which the function of any given protein can be predicted from those of its interacting partners, based on the principle of 'guilt by association' (Ballouz et al., 2015; Oliver, 2000). In this analysis, we labeled each protein with its annotated Gene Ontology (GO) terms, then withheld a subset of those labels. We then asked how accurately these withheld GO terms can be

predicted on the basis of the interaction network alone, by assigning a score to each protein-GO term pair that reflects the proportion of the protein's interacting partners that are annotated with the same term. This procedure was repeated three times, and the mean area under the receiver operating characteristic (ROC) curve (AUROC) was calculated for each GO term. The resulting distribution of AUROCs provides a quantitative overview of the biological coherence of the network, with higher AUROCs characteristic of a network in which proteins with a given function tend to be connected to other proteins with the same function. Importantly, this measure of a network's biological coherence is directly aligned with a key task for which protein-protein interaction (PPI) networks have been used (that is, protein function prediction) (Gillis et al., 2014; Wang and Marcotte, 2010).

We predicted protein interaction networks by supplementing dataset-derived features with combinations of zero to nine external genomic features and calculated the median AUROC for each network across all GO terms. The number of genomic features used to train each classifier was strongly and significantly correlated with the functional connectivity of the resulting networks (**Figure 2.2**; Spearman's $\rho = 0.40$, $P = 3.2 \times 10^{-51}$). This conclusion held for both datasets generated using SEC-PCP-SILAC ($\rho = 0.49$, $P = 5.5 \times 10^{-63}$), as well as SEC-LFQ datasets ($\rho = 0.37$, $P = 6.6 \times 10^{-3}$); no significant correlation was observed for the biochemical co-fractionation datasets ($\rho = 0.02$, $P = 0.81$). The trend was most pronounced when adding one or a handful of genomic features to the mass spectrometric data, and appeared to saturate quickly with the addition of more than 3-5 features, depending on the dataset. Thus, using conventional measures of a network's biological coherence, interaction networks constructed using genomic data integration outperform those constructed using co-migration data alone.

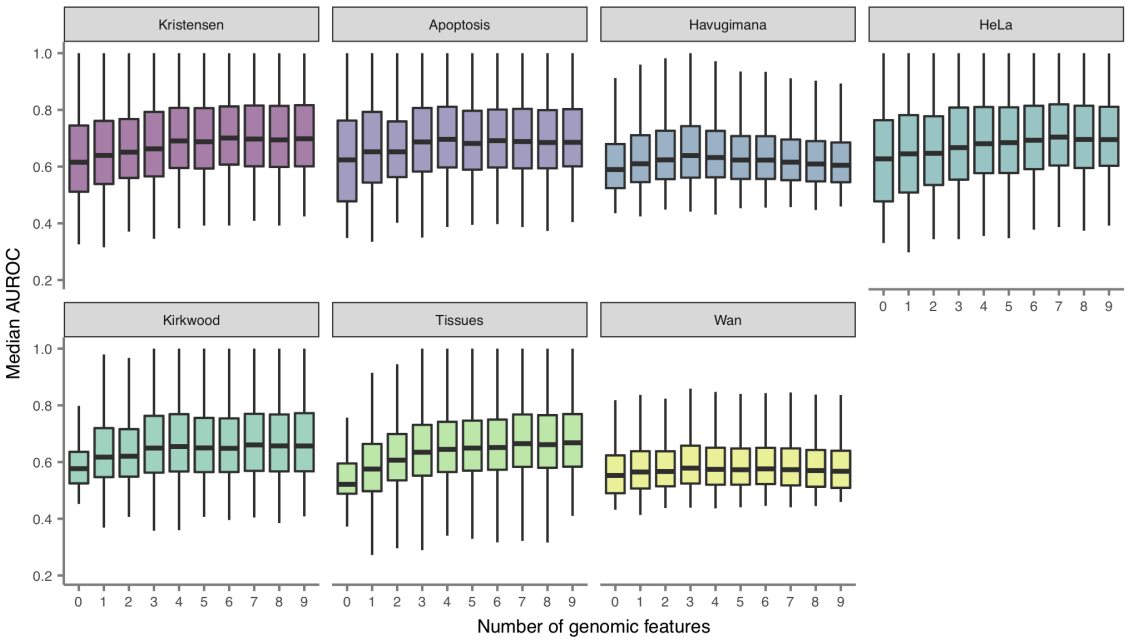


Figure 2.2 Functional connectivity of co-migration interactomes. PPI networks were predicted from co-migration data using machine-learning classifiers trained on raw co-migration data alone, or co-migration data supplemented with combinations of one to nine external genomic features. For each network, the functional connectivity was calculated as the distribution of AUROCs for prediction of protein function by neighbor voting in three-fold cross-validation.

Although genomic data integration increased the accuracy of protein function prediction from interactome topology, incorporating information about known functional associations or physical interactions could decrease the power of a classifier to discover novel interactions. Given that a primary goal of high-throughput interactome mapping projects is to discover novel interactions, we therefore undertook a systematic effort to compile known PPIs across seventeen databases, and used these catalogs of known interactions to calculate the proportion of previously known interactions in each network generated from co-migration data. We observed a negative correlation between the number of external genomic datasets integrated into the classifier and the

proportion of novel interactions within the network (**Figure 2.3**; Spearman's $\rho = -0.34$, $P = 1.9 \times 10^{-36}$). This trend remained significant for both SEC-PCP-SILAC ($\rho = -0.38$, $P = 4.0 \times 10^{-36}$) and biochemical co-fractionation datasets ($\rho = -0.43$, $P = 1.5 \times 10^{-8}$), although not the SEC-LFQ dataset ($\rho = 0.014$, $P = 0.90$). These observations suggest that the increased functional connectivity of networks recovered by integrating external genomic datasets may come at the expense of decreasing power to discover novel interactions.

Putative novel interactions could represent truly undiscovered interactions within an experimental dataset, but they could also reflect spurious or noisy associations within the data. Thus, a critical outcome in evaluating genomic data integration is what fraction of putative novel interactions actually represent true positives. To estimate this proportion, we performed a time-split validation experiment (Sheridan, 2013). We withheld interactions reported only in 2017 (n = 34,237) when identifying known interactions, then calculated the proportion of putatively novel interactions that were subsequently identified in 2017 within the human co-migration datasets. If putative novel interactions recovered by genomic data integration are more likely to represent true positives, then this procedure should be associated with a higher rate of interaction discovery in the validation set. However, surprisingly, we observed the inverse relationship, with a moderate *negative* correlation between genomic data integration and the likelihood of a putatively novel interaction later being discovered (Fig 3B; Spearman's $\rho = -0.13$, $P = 5.6 \times 10^{-3}$). We confirmed this moderate negative trend for SEC-PCP-SILAC ($\rho = -0.18$, $P = 1.1 \times 10^{-4}$) and SEC-LFQ ($\rho = -0.64$, $P = 1.2 \times 10^{-10}$) datasets individually, but found the reverse trend for the biochemical co-fractionation datasets ($\rho = 0.21$, $P = 0.0066$). However, in aggregate, the absence of a positive effect of genomic data integration on the proportion of putative novel interactions that were subsequently validated suggests that the effect of this

procedure on the biological relevance of the recovered interactions is negligible. We therefore conclude that genomic data integration decreases power to discover true, condition-specific interactions between proteins that are not yet linked by a known functional association. To ensure our results were insensitive to the temporal cutoff used in our time-split validation scheme, we repeated these analyses by withholding interactions discovered in 2016 or later. This analysis confirmed the absence of a positive trend ($\rho = -0.057$, $P = 0.13$; **Figure 2.4**). Finally, we confirmed that our results were robust to controlling networks based on the false discovery rate, rather than number of interactions (**Figure 2.5**), finding that genomic data integration decreased the proportion of novel interactions within these networks ($\rho = -0.31$, $P = 6.4 \times 10^{-22}$), but did not have a significant effect on the proportion of interactions that were subsequently discovered under either time split ($\rho = -0.039$, $P = 0.45$ and $\rho = -0.029$, $P = 0.57$ for 2017 and 2016, respectively).

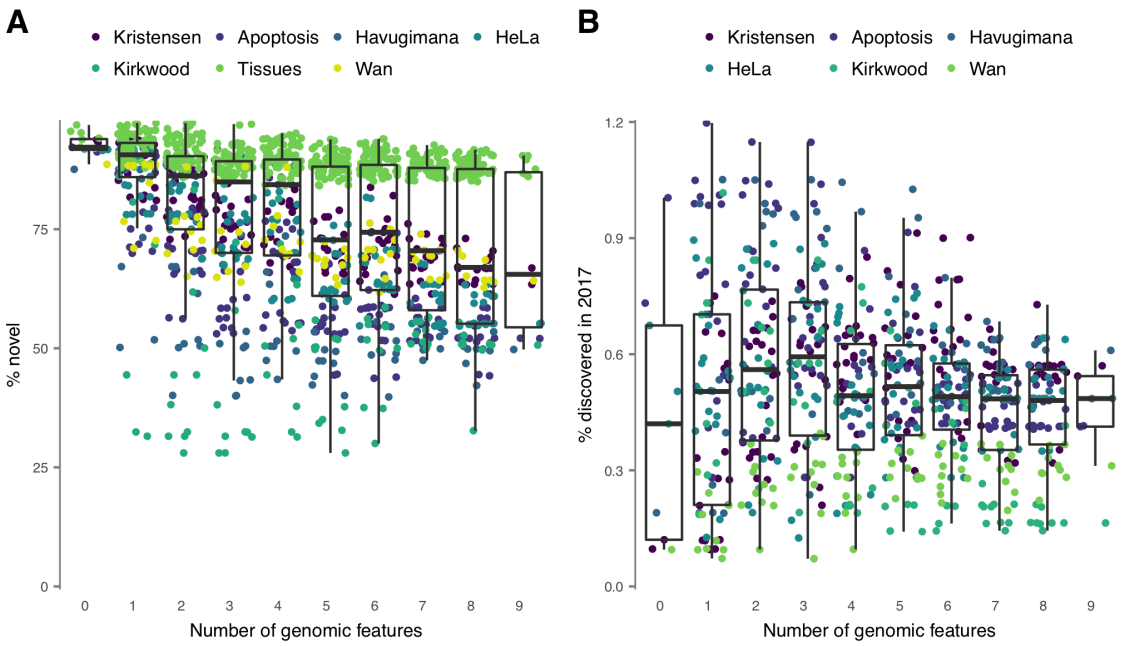


Figure 2.3 Novel interactions in co-migration interactomes. (A) Comprehensive databases of human and mouse protein-protein interactions were compiled and used to calculate the proportion of novel interactions within interaction networks recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets. Each point represents a network derived from a machine-learning classifier incorporating a different combination of external genomic features. (B) Interactions discovered in 2017 only were withheld from the database of known interactions and used to estimate the proportion of true positives among putative novel interactions by time-split validation. The 'Tissues' dataset is omitted as insufficient interactions were available to conduct time-split validation in mouse.

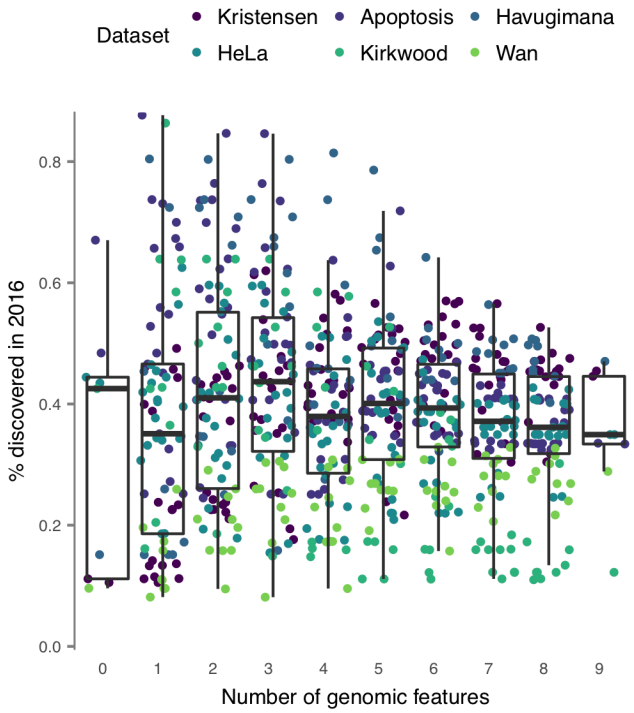


Figure 2.4 Novel interactions in PCP interactomes, with alternate time split. Interactions discovered in 2016 or later were withheld from the database of known interactions and used to estimate the proportion of true positives among putative novel interactions by time-split validation.

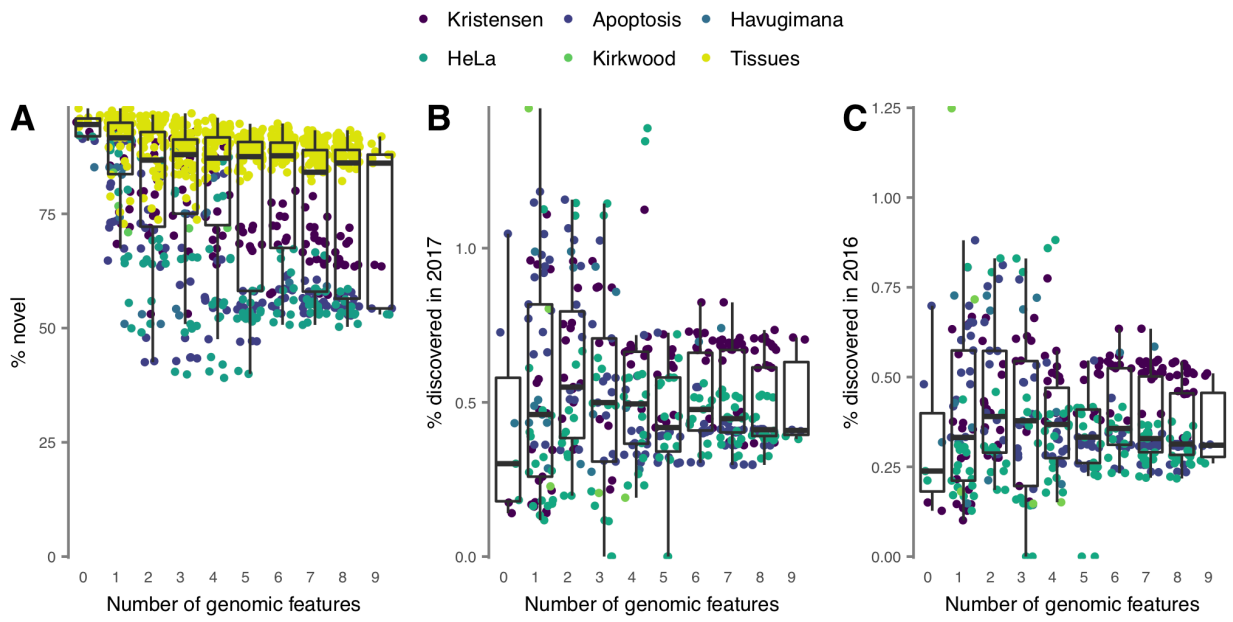


Figure 2.5 Novel interactions in false discovery rate-controlled PCP interactomes. (A) Proportion of novel interactions within interaction networks at 50% precision recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets. (B–C) Proportion of true positives among putative novel interactions by time-split validation in false discovery rate-controlled networks, using interactions discovered in 2017 or later (B) and 2016 or later (C) to estimate the proportion of true positives.

We note as a caveat to this conclusion that, when we analyzed the subset of networks presented in **Figure 2.3B** that were generated with combinations of zero and three features, we detected a positive correlation between the number of external genomic features integrated and the proportion of interactions that were subsequently discovered ($\rho = 0.17$, $P = 0.0055$). This result might be interpreted to suggest that a limited degree of genomic data integration can have a positive effect on the quality of the recovered interaction network. However, this post-hoc analysis should be interpreted cautiously, particularly since the correlation was no longer statistically significant when analyzing networks generated with combinations of up to four features ($\rho = 0.028$, $P = 0.59$). We additionally asked whether any specific combinations of

external genomic datasets consistently had a positive effect on the proportion of putative novel interactions that were subsequently discovered, relative to baseline networks recovered without external genomic datasets, but found that no combination had a significant effect after multiple hypothesis testing correction (all $P > 0.05$).

We next considered another potential bias introduced by genomic data integration: in particular, many genomic datasets are characterized by bias toward highly studied proteins. For example, literature-curated interactions from small-scale studies disproportionately involve well-studied proteins (Rolland et al., 2014). Similarly, interacting protein domains require three-dimensional structural templates to define, and therefore introduce bias towards proteins that have been studied using the techniques of structural biology. Incorporating these biased datasets into computational pipelines to analyze co-migration data introduces the possibility that biases towards highly studied proteins will be propagated into the resulting interaction networks. Such biases have been described in the context of a large 'uncharted zone' in interactomes derived solely from small-scale, literature-curated experiments, wherein the products of many human disease genes associate with few or no interacting partners, while a smaller number of highly studied proteins are densely connected (Rolland et al., 2014). We developed a quantitative metric of this bias within each co-migration interactome by calculating the Spearman correlation between the number of interactions predicted for each protein, and the number of publications in which it has been mentioned. Networks derived from co-migration data alone displayed minimal bias towards highly studied proteins (median Spearman's $\rho = 0.059$), but the degree of bias increased sharply with the number of external genomic datasets incorporated (**Figure 2.6A**; Spearman's $\rho = 0.32$, $P = 8.3 \times 10^{-32}$), a correlation that remained significant across two of three experimental methods (SEC-PCP-SILAC, $\rho = 0.36$, $P = 1.9 \times 10^{-32}$; SEC-LFQ, $\rho = 0.56$, $P = 6.1$

$\times 10^{-8}$; biochemical co-fractionation, $\rho = -0.10$, $P = 0.19$). Thus, in addition to containing a lower proportion of novel interactions overall, interaction networks predicted using genomic data integration display a greater bias towards highly studied proteins, whose functions are more likely to already be well understood.

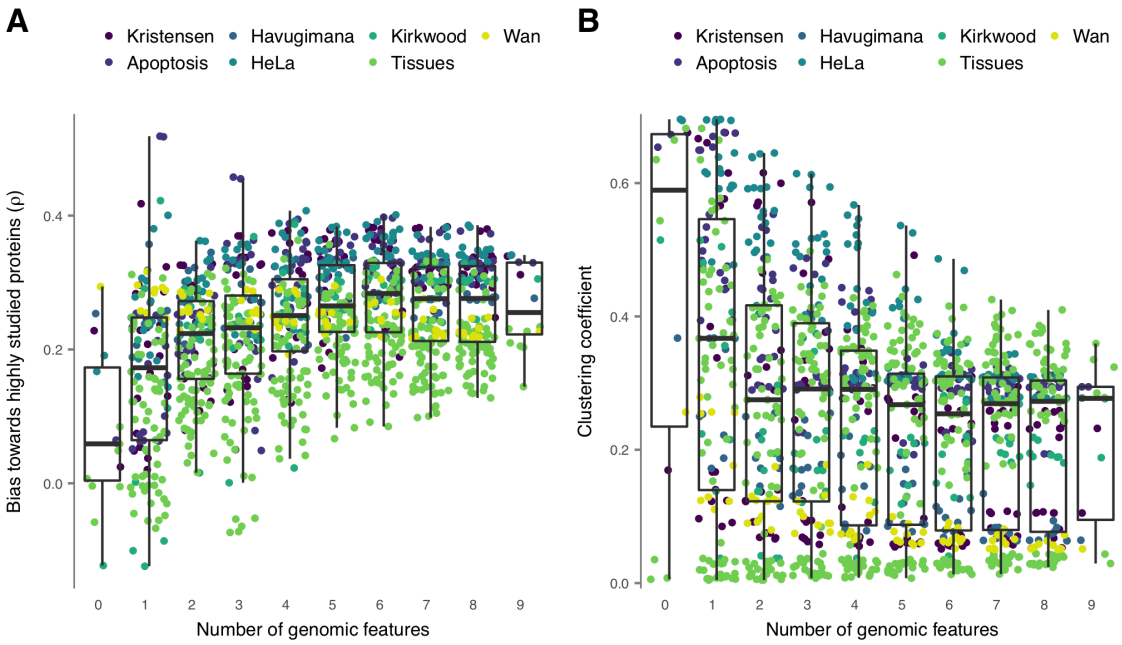


Figure 2.6 Bias towards highly studied proteins and protein complex recovery in co-migration interactomes. (A) Correlation between protein degree (number of interacting partners) and number of publications describing that protein in interaction networks recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets. (B) Global clustering coefficients of protein-protein interaction networks recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets.

Many proteins assemble into not only pairwise interactions, but higher-order, multi-protein complexes. In graph theoretic terms, these complexes can be described as cliques, as every protein in the complex co-migrates with every other protein in the complex. A global

topological measure that reflects the tendency of interacting proteins within a network to form cliques, and therefore complexes, would thus provide an orthogonal, high-level assessment of the biological relevance of the network, given that the aim of all the co-migration experiments analyzed here was to identify protein complexes co-migrating across one or more separation gradients. Such a measure is provided by the clustering coefficient of a network, which measures the probability that any two proteins connected to a given third protein are themselves connected in the interaction network (Barabási and Oltvai, 2004); a network with a high clustering coefficient is thus one with a global topology characteristic of protein complexes. We calculated the clustering coefficients for each network as a measure of the tendency of the classifier to preferentially identify protein complexes, and found that, although clustering coefficients varied considerably across datasets, genomic data integration was consistently associated with decreased clustering coefficients across all three methods (**Figure 2.6B**; Spearman's $\rho = -0.20$, $P = 1.2 \times 10^{-12}$; SEC-PCP-SILAC, $\rho = -0.23$, $P = 7.7 \times 10^{-14}$; SEC-LFQ, $\rho = -0.14$, $P = 0.22$; biochemical co-fractionation, $\rho = -0.77$, $P = 5.3 \times 10^{-33}$). Importantly, the clustering coefficients of the resulting interaction networks reflect an outcome independent of investigator biases inherent to the Gene Ontology and PPI databases. Our analysis of interaction network topology therefore suggests that genomic data integration impedes the identification of co-eluting complexes within co-migration datasets.

In all analyses described above, we made use of a naive Bayes classifier to distinguish interacting and non-interacting protein pairs. However, published computational pipelines have employed a variety of machine-learning methods. To confirm that our results were insensitive to the precise design of the computational pipeline, we repeated these experiments using support vector machines (SVMs) to classify interactions. We found our results were qualitatively

unchanged (**Figure 2.7**), indicating our conclusions are robust to the exact statistical techniques used to integrate known functional associations with co-migration data to map interactions.

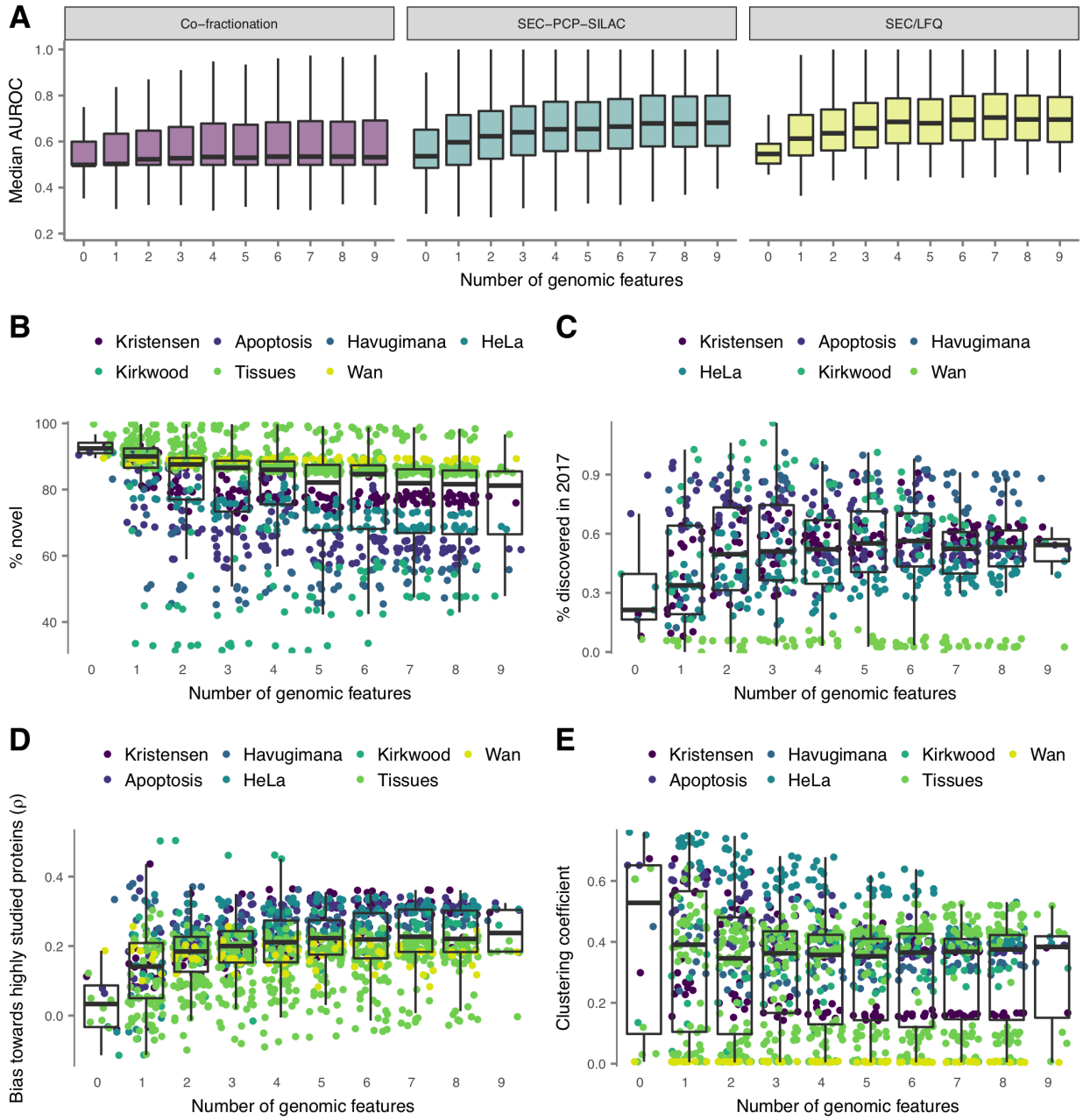


Figure 2.7 Conclusions are robust to statistical framework used to predict interactions from PCP data. (A–E) Results obtained using support vector machines, instead of naive Bayes classifiers, to predict interactomes from 16

co-elution datasets. (A) Functional connectivity of PCP interactomes, predicted using machine-learning classifiers trained on raw co-migration data alone or supplemented with combinations of one to nine external genomic features (Spearman's $\rho = 0.46$, $P = 4.2 \times 10^{-68}$). (B) Proportion of novel interactions in PCP interactomes recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets ($\rho = -0.33$, $P = 1.2 \times 10^{-34}$). (C) Interactions discovered in 2017 only were withheld from the database of known interactions and used to estimate the proportion of true positives among putative novel interactions by time-split validation ($\rho = 0.082$, $P = 0.028$). (D) Bias towards highly studied proteins in PCP interactomes, as quantified by Spearman correlation between protein degree and number of publications describing that protein, in interaction networks recovered from co-migration data alone, or supplemented with combinations of one to nine external genomic datasets ($\rho = 0.29$, $P = 1.1 \times 10^{-26}$). (E) Recovery of co-eluting complexes in PCP interactomes, as quantified by the global clustering coefficients of interactions recovered from co-migration data alone or supplemented with combinations of one to nine external genomic datasets ($\rho = -0.085$, $P = 2.4 \times 10^{-3}$).

Our results thus far indicate that, when predicting interaction networks from co-migration data, increasing genomic data integration is associated with increased functional connectivity, at the expense of decreasing power to discover novel interactions or reveal complete protein complexes. A remaining question is whether the trends described here apply to all genomic datasets, or whether individual datasets can be isolated as drivers of improved functional connectivity or decreased novelty. A single feature capable of increasing the biological coherence of the resulting networks, while leaving the proportion of novel interactions largely unchanged, would be highly desirable for co-migration data analysis. To address this question, we incorporated individual genomic features into classifiers, and evaluated the functional connectivity and novelty of the resulting networks. Eight of the nine features tested here resulted in a significant increase in functional connectivity relative to a baseline of exclusively dataset-derived features (**Figure 2.8A**; all $P \leq 1.6 \times 10^{-12}$, Fisher's method), with the sole exception

being gene fusion (P > 0.99). However, the same eight features resulted in highly significant increases in the proportion of previously known interactions (**Figure 2.8B**; all $P \leq 7.0 \times 10^{-26}$, Fisher's method, except P > 0.99 for gene fusion). The magnitude of these effects were variable, with the largest decreases in the proportion of novel interactions observed upon integration of co-citation data or previously published interactions. In contrast, smaller changes were observed for mRNA coexpression and phylogenetic profiles, suggesting these constitute less biased sources of genome-wide functional associations (**Figure 2.8C**).

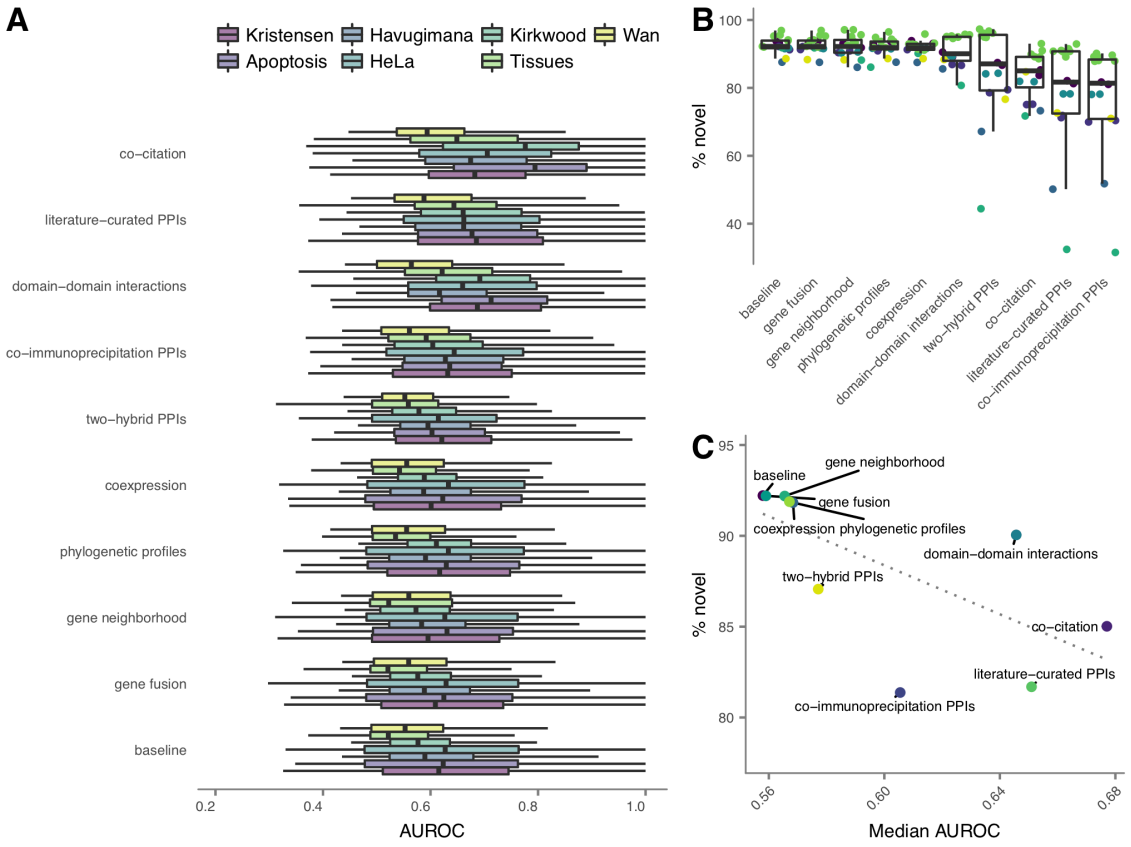


Figure 2.8 Effect of individual genomic features on functional connectivity and novelty in co-migration interactomes. (A) Functional connectivity and (B) proportion of novel interaction networks recovered from co-migration data alone, or in combination with individual external genomic features. (C) Median proportion of novel

interactions vs. median functional connectivity of networks recovered in combination with individual genomic features. Dotted line shows ordinary least squares linear regression.

We also analyzed the effect of individual features on network bias towards highly studied proteins, network clustering coefficients, and the proportion of putatively novel interactions later discovered in time-split validation (**Figure 2.9**). Notably, while none of coexpression, phylogenetic profiles, gene fusion, or gene neighborhood had a statistically significant impact on network bias towards highly studied proteins (**Figure 2.9A**; all $P \geq 0.64$, Brunner–Munzel test), all sources of interologous PPIs increased bias towards well-studied proteins (all $P \leq 5.7 \times 10^{-3}$), as did domain-domain interactions ($P = 5.5 \times 10^{-4}$); as expected, literature co-citation had the greatest effect ($P = 2.8 \times 10^{-9}$). In contrast, no individual feature had a significant impact on the clustering coefficient of the network (**Figure 2.9B**; all $P \geq 0.051$, Brunner–Munzel test). This finding suggests that integration of several features is required to significantly impact protein complex recovery from co-migration data. Finally, we analyzed the proportion of putative novel interactions that were subsequently discovered in our time-split validation scheme, a proxy for the proportion of putative novel interactions that correspond to true positives. Our results were identical regardless of the time point used to split the training dataset: domain-domain interactions and interologous literature-curated or co-immunoprecipitation PPIs reproducibly increased the proportion of putative novel interactions that were later discovered (all $P \leq 3.5 \times 10^{-4}$, Fisher's method). Surprisingly, co-citation data reproducibly led to a significant *decrease* in the proportion of true positive interactions ($P \leq 1.6 \times 10^{-11}$); the remaining features did not have a significant effect in isolation ($P \geq 0.33$). Taken together, these analyses reinforce the notion that integration of different functional genomic features has distinct and characteristic

impacts on the properties of recovered interaction networks and provide additional support for the use of mRNA co-expression and phylogenetic profiles in cases where genomic data integration is felt to be necessary.

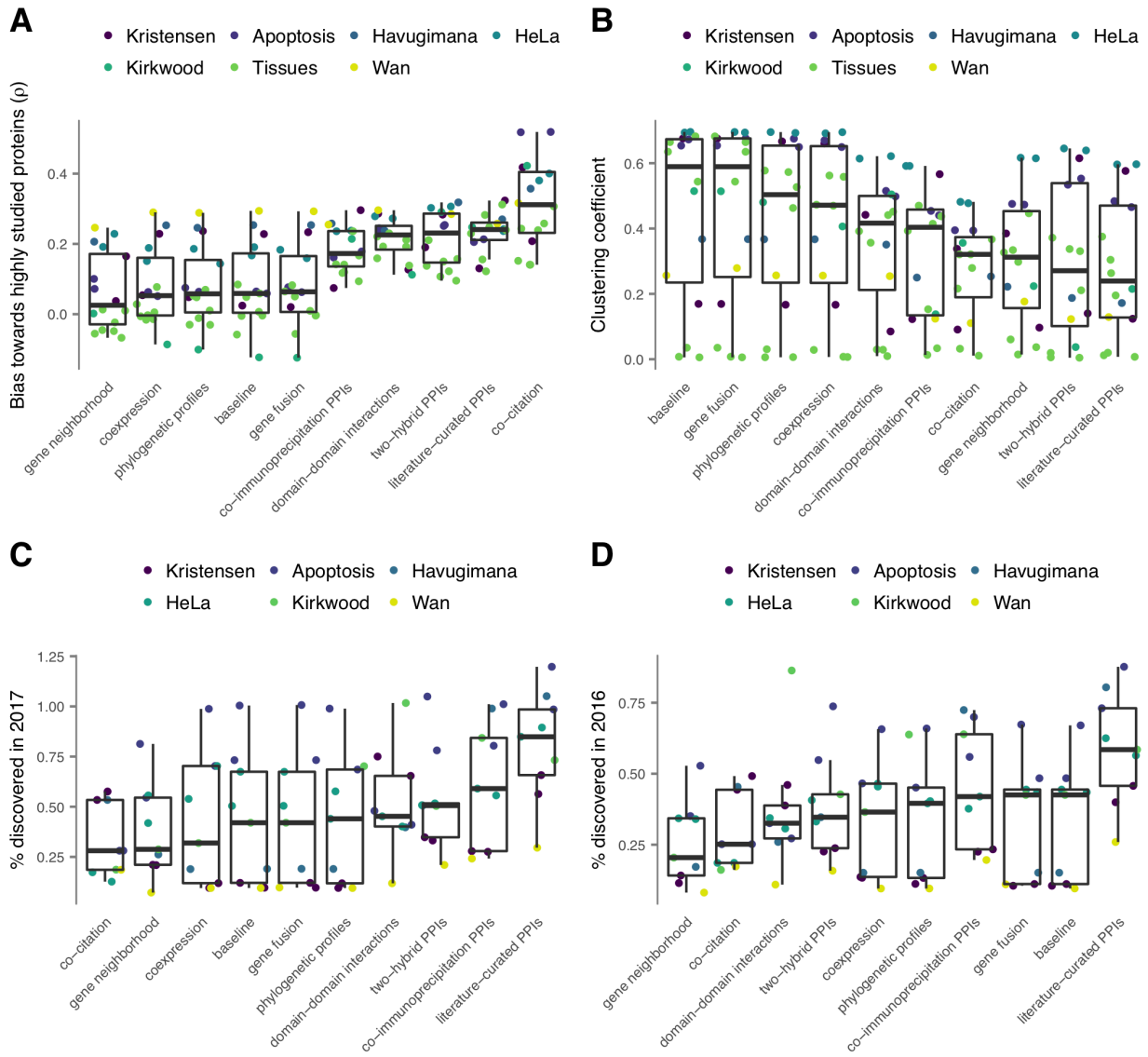


Figure 2.9 Effect of individual genomic features on bias, protein complex recovery, and novelty. (A) Bias towards highly studied proteins in interaction networks recovered with individual external genomic features, compared to baseline. (B) Global clustering coefficients of interaction networks recovered with individual external

genomic features, compared to baseline. (C–D) Proportion of true positives among putative novel interactions in interaction networks recovered with individual external genomic features, compared to baseline and estimated using time-split validation.

A final question we aimed to address is how stable co-migration interactomes are to slight perturbations in external genomic datasets. Many functional genomics datasets are incomplete: for example, the identification of domain-domain interactions relies on limited three-dimensional structural information, and existing interactome maps of model organisms are limited in scope. We simulated incompleteness of external genomic features by randomly withholding 20% of the annotations for each feature and predicted interaction networks with the resulting incomplete datasets. This procedure was repeated five times, and the Jaccard indices between all pairs of interaction networks were calculated as a measure of robustness to incomplete data. Networks displayed variable degrees of stability to perturbations in external genomic datasets, with four of nine features resulting in a median Jaccard index less than 0.75 between networks derived from the same co-migration data (**Figure 2.10**). This finding is concerning, as it suggests that integration of these features impedes robust, stable identification of interacting protein pairs. In contrast, co-expression, phylogenetic profiles, and all three sources of interologous PPIs provided comparatively greater robustness to incomplete data, suggesting these datasets yield more useful information for high-throughput interactome mapping.

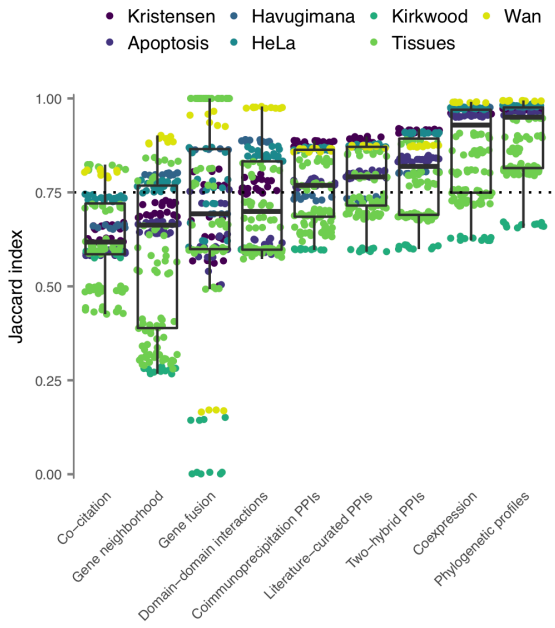


Figure 2.10 Robustness of predicted interaction networks to incomplete genomic data. The effect of incompleteness in external genomic data was analyzed by randomly withholding 20% of each genomic feature in five-fold cross-validation. The Jaccard index between networks derived from the same co-migration dataset was calculated as a measure of robustness. The dotted line corresponds to a Jaccard index of 0.75.

2.4 Discussion

Elucidating the complete interactome of humans and model organisms has been a longstanding goal of modern high-throughput biology, yet conventional methods to reveal PPIs are labour-intensive and not amenable to studying the interactome under physiological conditions or in response to stimulus. To address this gap, a new generation of methods for interactome mapping using co-migration have been developed and successfully applied in recent years (Havugimana et al., 2012; Kastritis et al., 2017; Kristensen et al., 2012; Wan et al., 2015). However, best practices for analysis of the unique proteomic datasets generated by these methods remain poorly defined. We conducted an extensive analysis of one widely used approach, which integrates

known functional associations with co-migration data to define interactomes using machine learning. Our results reveal that interaction networks recovered from raw data with this technique perform better on conventional benchmarks, such as their ability to predict protein function. However, this improved performance is associated with a decreased ability to identify novel interacting protein pairs within experimental data. Importantly, novel interactions could represent either true, undiscovered interactions, or false positive protein pairs. Our analysis suggests that classifiers trained without external data are, for the most part, discovering truly novel and potentially condition-specific interactions that lack known functional associations. In addition, we find that integration of external genomic datasets propagates bias towards highly-studied proteins into interaction networks and precludes accurate recognition of co-eluting complexes. Taken together, these findings suggest that genomic data integration biases the resulting protein interaction networks towards known functional associations, while impeding the discovery of true physical interactions between protein pairs that lack known functional associations in existing datasets.

The concept of functional association has proven tremendously useful for predicting gene function (Mostafavi et al., 2008) and interpreting genome-wide association study data (Lee et al., 2011; Taşan et al., 2015), among other genome-scale applications. However, a key point highlighted by our analysis is that functional association is not synonymous with physical interaction. Conversely, many apparently interacting proteins lack known functional associations. As we show, this distinction is critical in interpreting high-throughput efforts to map the interactome by co-migration, given that a primary goal of such projects is to discover previously unknown physical interactions. In fact, interactions substantiated solely from raw data appear no less likely than interactions substantiated by functional associations to later be

discovered. This observation attests to the high quality of co-migration data. However, it also highlights the noisiness of existing functional annotations: for instance, mRNA co-expression is known to be systematically biased by the spatial proximity of gene pairs (Kustatscher et al., 2017), while shared evolutionary trajectories can indicate participation in related functional pathways or processes rather than direct physical interaction (Li et al., 2014). Perhaps as a result, genomic data integration appears to decrease the power of machine-learning approaches to identify co-eluting protein complexes, as judged by a topological metric independent of investigator biases within known interaction databases or functional annotations in the Gene Ontology. Our results thus reveal that emphasis of prior functional associations comes at the expense of ability to reveal previously unknown biology. Conversely, the lack of a known functional association between pairs of putatively interacting proteins should not be used to rule out novel interactions.

Our analysis suggests that integration of external genomic datasets during the process of network construction from co-migration data impedes the discovery of novel interactions or protein complexes. However, there is clearly value in integrating genomic data after network construction, in order to validate the overall quality of the experimental network, or to compare datasets generated by different experimental methods (von Mering et al., 2002). Notably, such comparative analysis is made more difficult by the fact that many published studies have integrated publicly available functional genomics datasets with their mass spectrometric data to derive the final protein interaction networks. Consequently, these published interaction networks combine experimentally detected interactions with previously known functional associations, making it difficult to tease apart the relative contributions of each to the network. Thus, genomic data integration also represents a barrier to the comparative analysis of existing datasets.

Clearly, the scope of our analysis is not exhaustive: one can envision the addition of further external genomic datasets into classifiers designed to reconstruct interactome networks from raw co-migration proteomic data, or alternative methods of representing the features considered in this analysis (for instance, incorporating evidence of protein co-evolution at the sequence level, instead of genomic co-occurrence). However, in selecting the particular external genomic datasets that were considered within this study, our aim was to reflect those that have been incorporated in published co-migration projects, in order to ensure the conclusions of our study are relevant to the analytical protocols that have actually been implemented within the field to date. In this respect, it is noteworthy that no individual combination of external genomic datasets had a significant positive effect on the proportion of putative novel interactions that were subsequently discovered, under our time-split validation scheme. Moreover, by dissecting the individual contribution of each external genomic feature to the properties of the resulting interaction networks, we provide an extensive resource to understand the individual impacts of the nine external genomic features evaluated herein.

On the basis of our analysis, we argue that bioinformatic methods for interpretation of co-migration data should rely exclusively on dataset-derived features if a primary aim of a research effort is to discover novel interactions. We believe that this recommendation is particularly relevant to efforts to map the interactome across physiological conditions or in response to perturbation. It is tempting to additionally speculate that re-analysis of published co-migration data, using solely data-derived features, has the potential to reveal previously unknown interactions and complexes (Drew et al., 2017). If integration of additional data is deemed essential to maximize the success of labour-intensive downstream validation, only one or a few features should be included: we found that integration of even a single feature reliably resulted in

a significant decrease in the proportion of novel interactions recovered. Our analysis suggests mRNA coexpression, phylogenetic profiling, and domain-domain interactions are among the most appropriate features for this purpose, as they yield significant increases in functional connectivity at the expense of relatively modest increases in the proportion of novel interactions (**Figure 2.8C**). Finally, as similar genomic data integration techniques have been employed to map the human interactome using AP-MS, our findings are likely to generalize to mass spectrometry-based efforts to map the interactome beyond co-migration.

2.5 Conclusions

We find that a commonly used computational technique for analysis of co-migration data hinders the discovery of novel interactions from raw proteomics datasets, accentuates bias towards extensively studied proteins, and impedes accurate recovery of co-eluting complexes. Importantly, interactions predicted from raw data alone appear no less likely to later be experimentally discovered. This leads us to recommend that bioinformatics pipelines for interactome mapping from co-migration datasets should rely exclusively on this raw data, when possible. When integration of additional features is deemed essential, mRNA co-expression, phylogenetic profiling, and domain-domain interactions balance increased functional connectivity and decreased novelty. Prioritizing high-quality co-migration data over noisy functional associations will support efforts to reveal complete maps of PPIs in humans and other organisms.

Chapter 3: An atlas of protein-protein interactions across mouse tissues

3.1 Introduction

Cellular functions are mediated by the dynamic association of individual proteins into complexes, signalling pathways, and other macromolecular assemblies. The biological functions of many proteins depend on specific physical interactions with other proteins, and disruption of these interactions can result in disease (Sahni et al., 2015; Wang et al., 2012). Defining the complete map of functional protein-protein interactions in a given organism (the interactome) has therefore been a longstanding goal of the post-genomic era, with a view to better understanding protein function, cellular processes, and ultimately the relationship between genotype and phenotype (Vidal et al., 2011). To this end, high-throughput methods have been developed to map interactomes at the proteome scale, including yeast two-hybrid (Y2H), affinity purification-mass spectrometry (AP-MS), protein complementation assay (PCA), and protein correlation profiling (PCP). These methods have been applied successfully to generate high-quality maps of the interactomes of humans (Hein et al., 2015; Huttlin et al., 2015, 2017; Luck et al., 2020; Rolland et al., 2014) and other metazoans (Guruharsha et al., 2011; Simonis et al., 2009).

Widely used methods for mammalian interactome mapping rely on heterologous expression or genetically manipulated cell lines, and as a consequence, existing interactome maps provide limited insight into which interactions occur in specific cell types or tissues, or under pathophysiologically relevant conditions (Snider et al., 2015). Targeted interactome mapping in tissue- or cell type-specific contexts has revealed rewiring of protein interactions in select human diseases (Pankow et al., 2015; Shirasaki et al., 2012), yet fundamental questions

regarding the organization of the interactome across mammalian tissues remain unanswered. Whereas large-scale efforts to profile the transcriptome, proteome, and epigenome of human tissues have been undertaken (Melé et al., 2015; Roadmap Epigenomics Consortium et al., 2015; Uhlén et al., 2015), a comparable resource at the interactome level is lacking.

We previously developed a high-throughput method for interactome mapping by combining size exclusion chromatography (SEC) with protein correlation profiling–stable isotope labeling by amino acids in cell culture (PCP-SILAC) (Kristensen et al., 2012). Here, we couple PCP with stable isotope labeling of mammals (SILAM) (Krüger et al., 2008; McClatchy et al., 2007) to map the interactomes of seven mouse tissues. This resource provides the first global view of protein-protein interactions in a mammalian tissue-specific context. Our study additionally provides the first systematic interactome map in mouse, revealing over 120,000 novel interactions and more than doubling the size of the known mouse interactome. The resulting *in vivo* interactome maps uncover widespread rewiring of protein interactions across physiological contexts.

3.2 Materials and methods

3.2.1 Generation of SILAM tissues

Protocols for the generation of SILAM labelled and unlabelled tissues were approved by the University of British Columbia Animal Care Committee in accordance with international guidelines (protocol number: A13-0094). SILAM colonies and unlabelled littermate controls were generated according to the approach of Krüger et al. (Krüger et al., 2008; Zanivan et al., 2012). Briefly, two sets of female and male littermate C57BL/6 mice (Charles River Laboratories) aged 8 weeks were segregated into separate cages and female mice fed either a

SILAC chow diet (8 g $^{13}C_6$-Lysine/kg, Cambridge Isotope Laboratories, Andover, MA) or unlabelled matched chow diet differing only in the incorporation of $^{13}C_6$-Lysine. After 10 weeks of SILAC feeding, male mice were introduced to F0 SILAM females to allow mating. The resulting F1 SILAM mice were allowed to develop to 8 weeks and subjected to another round of mating to generate F2 SILAM mice. F2 SILAM mice were allowed to develop to 8 weeks prior to tissue isolation and protein complex extraction. In parallel to the generation of SILAM mice, unlabelled mice were derived from littermates of the F0 SILAM breed and allowed to develop to 8 weeks. Animals were housed with littermates of the same sex, up to 4 per cage, with environmental enrichment according to the University of British Columbia Animal Care requirements.

3.2.2 SILAM incorporation monitoring

SILAM incorporation rates were monitored in female mice during the generation of F1 and F2 litters. For blood sample collection, mice were anesthetized with isoflurane and ~20 µL of blood was collected by tail snipping every month during SILAM feeding. To measure isotope incorporation, the medulla oblongata was collected during animal termination of F0, F1 and F2 mice. Blood and medulla oblongata samples for SILAM incorporation analysis were snap frozen in liquid nitrogen and stored at –80°C. Samples were boiled in 1% sodium deoxycholate, digested, and quantified as described below.

3.2.3 Tissue harvesting

Tissues were harvested from 8-week old C57BL/6 male mice. Mice were terminally anesthetized with isoflurane and after the loss of corneal reflexes moved to a chilled surgical platform. To

limit blood contamination within tissues and inhibit phosphatase and protease activity, the heart was exposed by Y-incision and mice perfused with 50 mL of ice-cold size exclusion chromatography (SEC) mobile phase [50 mM KCl, 50 mM $NaCH_3COO$, pH 7.2, containing 2x cOmplete protease inhibitor cocktail without EDTA (Roche) and 2x Halt protease and phosphatase inhibitor cocktail (Thermo Scientific, San Jose CA)]. Perfusate was introduced into the left ventricle by précising the ventricle wall with a needle while the right ventricle was cut to allow drainage. Upon complete blanching of the liver, the seven tissues of interest (heart, brain, thymus, liver, kidney, skeletal muscle, and lung) were removed, rinsed with ice-cold SEC buffer and placed into ice-cold SEC containing 2x protease inhibitor cocktail without EDTA and 2x protease and phosphatase inhibitor cocktail. Tissues were further cut into smaller pieces to enhance accessibility of inhibitors and placed on ice.

3.2.4 Preparation of cytoplasmic complexes

Complex preparation of dissected tissues and size exclusion chromatography were performed as described previously (Kristensen et al., 2012), with minor modifications. Briefly, tissues samples were lysed using a Dounce homogenizer with 200 strokes of a loose pestle followed by 200 strokes with a tight pestle. Lysates were ultracentrifuged at 100,000 relative centrifugal force (r.c.f.) for 15 min at 4°C to remove insoluble material and to partially deplete highly abundant ribosomes. Large molecular weight complexes were then concentrated using 100,000 Da molecular weight cut-off spin columns (Sartorius Stedim, Goettingen, Germany). Five milligrams of total filtered protein was then immediately loaded onto a chromatography system consisting of two 300 × 7.8 mm BioSep-4000 Columns (Phenomenex, Torrance, CA) equilibrated with SEC mobile phase and separated into 80 fractions by a 1200 Series semi-

preparative HPLC (Agilent Technologies, Santa Clara, CA) at a flow rate of 0.5 mL/min at 8°C. The HPLC consisted of a G1310A isocratic pump, G7725i manual injector, and G1364 fraction collector with a G1330 thermostat. Fractions 1 to 55 corresponded to molecular weights from 2 MDa to 100 KDa, as determined by the use of common SEC standards thyroglobulin, apoferritin and bovine serum albumin (Sigma-Aldrich). Each tissue was separated independently by SEC for both labelled and unlabelled samples. Fractions 1 to 55 from the seven heavy-labelled tissue preparations were pooled and served as an internal reference allowing the comparison between and across all samples. The pooled reference was spiked into each of the corresponding light fractions at a volume of 1:0.75 (light to heavy).

3.2.5 In-solution digestion

Individual PCP-SILAM samples were prepared using in-solution digestion as previously described (Rogers et al., 2010). Briefly, sodium deoxycholate was added to each fraction to a final concentration of 1.0% (w/v) and samples boiled for 5 min. Boiled samples were allowed to cool to RT then reduced for 1 hour with 10 mM dithiothreitol (DTT) at room temperature. Samples were then alkylated for 1 h with 20 mM iodoacetamide (IAA) in the dark at room temperature and excess IAA quenched with 40 mM DTT for 20 min. Two micrograms of Lys-C (Wako) were added to each fraction and samples were incubated overnight at 37°C with shaking. Samples were acidified to pH < 3 with acetic acid to precipitate deoxycholic acid, which was then removed by centrifugation at 16,000 r.c.f. for 20 min. To ensure the removal of particulate matter, peptide digests were further clarified using Unifilter 800 Whatman filter plates (GE Healthcare Life Sciences). The resulting peptide supernatant was purified using self-made Stop-and-go-extraction tips (StageTips) (Rappsilber et al., 2007) composed of C18 Empore material

(3M) packed in to 200 µL pipette tips. Prior to addition of the peptide solution, StageTips were conditioned with methanol, followed by 80% MeCN, 0.1% formic acid (Buffer B), then 0.1% formic acid (Buffer A). Peptide supernatants were loaded onto columns and washed with three bed volumes of Buffer A. Peptide samples were stored directly on Stage Tip at 4°C until required, when they were eluted with Buffer B directly into a HPLC autosampler plate and dried using a vacuum concentrator. An alternative in-solution digestion method was used for a subset of sample sets. For these, each fraction was denatured with 6 M urea / 2 M thiourea, reduced for 30 min with 10 mM DTT at room temperature, alkylated for 30 min with 20 mM IAA in the dark at room temperature. A mixture of 1.5 µg of Lys-C, 4 mM DTT, and 50 mM ammonium bicarbonate was added to each fraction to dilute urea to 2 M and incubated overnight at room temperature with shaking. The peptides were purified with Stage Tips, eluted, and dried using a vacuum concentrator.

3.2.6 Liquid chromatography and mass spectrometry analysis

Prior to LC-MS/MS analysis, samples were resuspended in 15 µL Buffer A. One of three LC-MS/MS setups was used: one subset was acquired using an EASY-nLC1000 system (Thermo Scientific) coupled to a Q-Exactive mass spectrometer (Thermo Scientific). Another subset used a Dionex Ultimate 3000 UHPLC (Thermo Scientific) coupled to a Q-Exactive plus mass spectrometer (Thermo Scientific). The final subset used an EASY-nLC1000 (Thermo Scientific) coupled to a quadrupole–time of flight mass spectrometer (Impact II; Bruker Daltonics). LC-MS/MS was accomplished using a two-column system in which samples were concentrated prior to separation on an in-house packed 2 cm long, 100 µm inner diameter fused silica fritted trap column containing 5 µm Aqua C18 beads (Phenomenex) and then separated using an in-house

packed C18 analytical 75 μm inner diameter column composed of 35 cm ReproSil-Pur C18 AQ 1.9 μm (Dr. Maisch, Ammerbuch-Entringen, Germany) column with an integrated spray tip (6–8 μm-diameter opening, pulled on a P-2000 laser puller from Sutter Instruments) that is held at 50°C by an in-house built column heater or a PepMap100 C18 20 mm × 75 μm trap and a PepMap C18 500 mm × 75 μm analytical column (Thermo Fisher Scientific). Samples were concentrated onto the trap for 5 min using 100% Buffer A at 5 L/min after which the gradient was altered from 100% Buffer A to 40% Buffer B over 180 min at 250 nL/min with the eluting peptides infused directly into the mass spectrometer via nESI. An alternative gradient was shortened to 150 min separation. The Q-Exactive and Q-Exactive Plus were operated in a data-dependent manner using Xcalibur (Thermo Scientific) with the top ten most intense multiply-charged ions above a 5% underfill ratio from MS1 scans (resolution 70,000; 350-2,000 m/z, AGC target of 3 × 106) selected for HCD MS-MS events (resolution 17.5k AGC target of 1 × 106 with a maximum injection time of 60 or 120 ms, NCE 28 with 20% stepping) with 25 s dynamic exclusion enabled. The Impact II was operated in a data-dependent auto-MS/MS mode with inactive focus fragmenting the 20 most abundant ions (one at the time at 18 Hz rate) after each full-range scan from m/z 200 to m/z 2000 at 5 Hz rate. The isolation window for MS/MS was between 2 and 3 units depending on the parent ion mass to charge ratio, and the collision energy ranged from 23 to 65 eV depending on ion mass and charge. Parent ions were then excluded from MS/MS for the next 0.4 min and reconsidered if their intensity increased more than five times. Singly charged ions were excluded from fragmentation.

3.2.7 Co-immunoprecipitation

For talin immunoprecipitations, brain and liver tissues collected from 3 mice (B10, male, 10-week old), were rinsed in ice cold in size-exclusion chromatography (SEC) mobile phase (50 mM Tris, 50 mM KCl, 50 mM $NaCH_3COO$, pH 7.2) including protease inhibitors without EDTA (Roche) and phosphatase inhibitors (1 mM sodium orthovanadate, 5 mM sodium pyrovanadate and 0.5 mM pervanadate) and placed on ice. Tissues (100-200 mg) were individually disrupted in a Dounce homogenizer (2 min, tight dounce) in 2 ml of ice-cold size-exclusion chromatography (SEC) mobile phase (50 mM Tris, 50 mM KCl, 50 mM $NaCH_3COO$, pH 7.2) including protease inhibitors without EDTA (Roche) and phosphatase inhibitors (1 mM sodium orthovanadate, 5 mM sodium pyrovanadate and 0.5 mM pervanadate). The resulting lysates were clarified by centrifugation (15 min; 4°C; 16000 r.c.f.) and 1 mL of each supernatant was used for immunoprecipitation using C-9 mouse monoclonal talin antibody (sc-365875, Santa Cruz) following the manufacturer's protocol. Briefly, the lysates were pre-cleared for 30 min using mouse IgG and protein L-agarose (sc-2336, Santa Cruz), incubated for 30 min with 2 mg of primary antibody, followed by 30 min incubation with protein L-agarose. The bound beads were washed twice with ice-cold SEC buffer and the protein eluted by boiling with the Laemmli buffer. The resulting proteins were subjected to SDS-PAGE, in gel-digested, purified on STAGE tips and analyzed my mass spectrometry.

3.2.8 Ribosome isolation

Tissues were harvested from two 10-week old C57BL/6 male mice and rinsed two times in PBS. Tissues were then minced using a scalpel and suspended in 3 mL of Ribosome Homogenization Buffer (50 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 25 mM KCl, 0.2 M sucrose). Tissue samples

were homogenized for 1 min using a loose Dounce homogenizer followed by 2 min in a tight Dounce homogenizer. Lysates were clarified through centrifugation using a Sorvall Surespin 630 rotor at 20,000 r.c.f. for 10 min at 4°C. The crude lysates were then layered over a sucrose cushion (50 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 25 mM KCl, 2 M sucrose) at a 1:1 (v/v) ratio. Subsequently, samples were ultracentrifuged at 100,000 r.c.f. for 24 h at 4°C. The resulting pellet was resuspended in Ribosome Homogenization Buffer. Protein concentration was determined via NanoDrop (Thermo Fisher). Equals amount of protein from each tissue sample was then denatured with 6M urea, reduced with 1 µg of DTT, alkylated with 2.5 µg of iodoacetamide, followed by adding four volumes of 50 mM NH_4HCO_3 and subjected to trypsin digestion overnight at room temperature. Peptide samples were cleaned via STAGE tips and analyzed by LC-MS/MS. The raw data was searched using MaxQuant version 1.5.3.30 with variable modifications: acetylation (K), deamidation (NQ), and oxidation (M); fixed modifications: carbamidomethyl (C); and label-free quantitation with a minimum ratio count of 1. The resulting LFQ intensities were normalized to the median ribosomal protein intensity in each replicate.

3.2.9 Proteomic data analysis

MaxQuant (versions 1.5.5.1 and 1.6.5.0) (Cox and Mann, 2008) was used for identification and quantification of the resulting experiments. Database searching was carried out against the UniProt *Mus musculus* database (downloaded August 17, 2014; 49,235 entries), augmented with common contaminants, with the following search parameters: carbamidomethylation of cysteine as a fixed modification, oxidation of methionine, and acetylation of protein N-termini. The digestion mode was either semi-specific LysC or specific LysC with a maximum of two missed

cleavages. A multiplicity of two was used, denoting the SILAM amino acid combinations (light lysine and heavy lysine respectively). The precursor mass tolerance was set to 7 parts-per-million (ppm) and MS/MS of 10 ppm, with a maximum false discovery rate of 1% for protein identifications and peptide-spectrum matches, which was filtered to 1% at the peptide and protein level after peptide score correction. To enhance the identification of peptides between fractions and replicates, the match between runs option was enabled with a precursor match window set to 2 min and an alignment window of 10 min. Potential contaminants, reverse hits, and proteins identified only by modified peptides were removed. Protein groups were mapped to gene symbols, retaining only the chromatogram with the greatest number of non-missing observations in cases where multiple protein groups mapped to a single symbol. These steps yielded a matrix reflecting the abundance of 7,225 proteins across 770 SEC fractions.

To evaluate the reproducibility of PCP-SILAM, we computed the Spearman correlations between each biological replicate from the same tissue, and compared these to correlations between biological replicates in published, cell-line-based PCP-SILAC data from three studies (Kerr et al., 2020; Scott et al., 2015, 2017), collectively representing a total of 26 replicate pairs. We additionally performed hierarchical clustering of all fourteen PCP-SILAM replicates, using the Spearman correlation as the similarity measure.

To assess the proteome coverage afforded by PCP-SILAM, we compared proteins detected in each tissue to those detected in previous deep proteomic studies of the same murine tissues, including both in-depth studies of individual tissues (Azimifar et al., 2014; Deshmukh et al., 2015; Sharma et al., 2015) and a larger-scale survey of mouse tissues (Geiger et al., 2013). Protein groups quantified in at least one replicate were mapped to gene symbols. Only whole tissue lysates, and not purified cell types, were included in this analysis.

To evaluate the statistical power of PCP-SILAM to resolve known protein complexes, we reproduced a receiver operating characteristic (ROC) analysis (Romanov et al., 2019) of large-scale shotgun proteomics datasets (Battle et al., 2015; Carlyle et al., 2017; Chick et al., 2016; Geiger et al., 2012, 2013; Guo et al., 2015; Khan et al., 2013; Kustatscher et al., 2019; Parker et al., 2019; Wu et al., 2013), additionally comparing to matched transcriptomic (Battle et al., 2015; Carlyle et al., 2017; Chick et al., 2016; van Heesch et al., 2019; Khan et al., 2013) and translatomic (Battle et al., 2015; van Heesch et al., 2019) data when possible. Briefly, the Pearson correlation was calculated between each gene or protein pair in every dataset, and pairs quantified in fewer than ten overlapping samples were filtered to mitigate the impact of spurious correlations. Then, these gene or protein pairs were labelled as true positives and true negatives using the dataset of protein complexes manually curated by (Ori et al., 2016), calling intra-complex pairs as true positives and inter-complex pairs as true negatives. Finally, we used these labels to calculate the area under the receiver operating characteristic curve (AUC), using the 'AUC' R package. We additionally compared the AUC in *in vivo* PCP-SILAM data to *in vitro* data from four previous PCP-SILAC studies (Kerr et al., 2020; Scott et al., 2015, 2017; Stacey et al., 2018) using the same procedure, observing no significant difference.

3.2.10 Interactome network inference

PrInCE (Stacey et al., 2017), an open-source pipeline for co-fractionation mass spectrometry data analysis, was used to reconstruct high-confidence interactomes from PCP-SILAM profiles. PrInCE first performs basic data filtering and preprocessing, then applies a machine-learning approach to rank interactions, and returns as output a list of interactions at a given precision threshold. Briefly, single missing values are imputed as the mean of neighboring intensities,

proteins detected in fewer than five fractions are filtered, and co-fractionation profiles are smoothed by a sliding average with a width of five fractions. Next, mixtures of one to five Gaussians are fitted to each profile, and model selection is performed using the bias-corrected Akaike information criterion (Hurvich and Tsai, 1989). Then, six measures of distance or similarity are calculated for each protein pair, each of which reflects the likelihood of a physical interaction between those two proteins based on their mass spectrometric profiles. These six features include: the Pearson correlation coefficient (calculated separately for both the raw and smoothed chromatograms) as well as its P value in the raw chromatograms; the Euclidean distance between chromatograms; the number of fractions separating the maximum values of each co-fractionation profile; and the Euclidean distance between the closest pair of fitted Gaussians. Missing values in the feature matrix are imputed with the median, plus or minus a small amount of random noise, sampled from a normal distribution with a mean of zero and a standard deviation equal to 5% of the standard deviation of the relevant feature. A classifier is subsequently trained on a reference set of interactions in 10-fold cross-validation, taking pairs of proteins within the same complex as true positives and pairs of proteins in different complexes as true negatives, and predictions are made for the entire test set (and all proteins not found in any training set complex) during each fold of cross-validation. Protein pairs are then ranked as candidate interactions based on their median classifier score across all ten cross-validation folds. PrInCE implements a number of classifiers for use in network inference; here, we used heterogeneous classifier fusion, as implemented in the PrInCE Bioconductor package, to train an ensemble of classifiers, including naive Bayes, random forest, logistic regression, and support vector machine models, then aggregated predictions by taking the mean rank across all four classifiers. This strategy improves the robustness of the network inference procedure by down-

weighting spurious protein pairs that are highly ranked only by a single model while up-weighting pairs that are ranked near the top by all classifiers.

The interaction score calculated for each protein pair is then converted to a measure of precision (or, equivalently, a false discovery rate) for each interaction by calculating the ratio of true positives to true positives and true negatives among interactions at that probability or higher. Finally, a list of interactions is output at a user-specified precision. We applied PrInCE to each of the two replicates in each tissue separately, then combined and re-scored interactions detected in either replicate by the application of Fisher's method to the false discovery rates of each interaction in either replicate. Only protein pairs co-eluting in at least five overlapping fractions were considered as potential interactors. Interactions detected within each tissue at 95% precision or higher (equivalent to a 5% false discovery rate) were retained for further analysis.

Like previous machine-learning approaches to network inference from co-fractionation mass spectrometry (Havugimana et al., 2012; Wan et al., 2015), PrInCE requires a resource of known protein complexes to train the classifier, such as those provided by the CORUM database (Giurgiu et al., 2019) or the manually curated dataset of Ori et al., 2016. Previously, however, we observed that several of these protein complexes may be degraded or disassembled under PCP assay conditions (Stacey et al., 2018). We found that using these protein complexes to train a classifier impaired the ability of that classifier to distinguish interacting and non-interacting protein pairs, as they displayed little statistical evidence of co-elution. However, we found that the accuracy of network inference could be substantially improved by using only a subset of CORUM protein complexes that had consistently co-eluted in previously published PCP studies to train the classifier (Stacey et al., 2018), and we therefore used that same subset in this study.

3.2.11 Literature-curated, high-throughput, and mouse tissue interactomes

Interaction data from the BioPlex (Huttlin et al., 2015), BioPlex 2 (Huttlin et al., 2017), HI–II–14 (Rolland et al., 2014), QUBIC (Hein et al., 2015), and HuRI (Luck et al., 2020) screens were downloaded from the supporting information of the respective publications and mapped to gene symbols. In addition, a master database of 82,602 experimentally detected mouse interactions was compiled by merging interactions from BIND (Alfarano et al., 2005), BioGRID (Oughtred et al., 2019), CORUM (Giurgiu et al., 2019), DIP (Salwinski et al., 2004), HINT (Das and Yu, 2012), IntAct (Orchard et al., 2014), iRefIndex (Razick et al., 2008), mentha (Calderone et al., 2013), and MINT (Licata et al., 2012). Each interaction was associated with one or more publication(s) in which the interaction was reported; interactions without a traceable publication were assumed to be reported in a single publication. Identifiers were mapped to gene symbols, and self-interactions were removed. The database was then partitioned into interactions reported in at least one, two, three, or four publications. Finally, manually curated protein complexes were obtained from (Ori et al., 2016). This yielded a total of 17 interactome networks.

Human and mouse GO annotations were obtained from the UniProt-GOA. Annotations with the evidence codes “IPI” (inferred from physical interaction), “IEA” (inferred from electronic annotation), or “ND” (no data), or the qualifier “NOT”, were removed.

For each network, we computed a series of indices that reflect the concordance of the network with other large-scale genomic datasets. First, we calculated the functional coherence of the network, using the EGAD R package (Ballouz et al., 2017), as previously described (Skinnider et al., 2018b, 2019). In this analysis, each gene in the network is annotated with its known functions (that is, GO terms), and a subset of these labels is then randomly withheld. A simple neighbor-voting algorithm (Schwikowski et al., 2000) is then used to predict functions for

the withheld proteins. That is, for each protein in the held-out set, a score is assigned for each GO term that reflects the proportion of that protein's interacting partners annotated with the GO term of interest. The process is repeated in three-fold cross-validation, and the AUC is calculated for each GO term, thereby quantifying the accuracy of protein function predictions made based on the topology of the network. As an alternative measure of functional connectivity, we computed the graph assortativity of each GO term, which quantifies the tendency of proteins annotated to a particular GO term to interact with other proteins annotated to the same term, and observed similar results. In both cases, we discarded GO terms annotated to less than 20 or more than 200 proteins, in order to mitigate the impact of very specific or very broad GO terms on the results (Skinnider et al., 2018, 2019). Second, we computed the coexpression of each interacting protein pair, quantified using the Pearson correlation coefficient, in two large-scale proteomics datasets (Kustatscher et al., 2019; Lapek et al., 2017). Third, for each interacting protein pair, we computed the similarity of subcellular colocalization, again using the Pearson correlation coefficient to quantify similarity in two alternative subcellular proteomics datasets (Geladaki et al., 2019; Orre et al., 2019). The distribution of each metric within each network was visualized as a spectrum, using a random sample of 1,000 interacting protein pairs for the co-expression and co-localization spectra. Networks were arranged by the proportion of GO terms or proteins with less than random functional coherence (that is, AUC < 0.5 or assortativity < 0), coexpression, or colocalization (correlation < 0).

3.2.12 Analysis of novel mouse interactions

The combined database of unique mouse interactions described above was used to define known interactions and identify interactome orphans as proteins in the mouse tissue interactomes for

which no interactions were previously deposited in any interaction database (Kotlyar et al., 2015). The statistical significance of the overlap was calculated using the hypergeometric test, defining the population size to be the total potential number of interactions between all unique proteins present in either the PCP-SILAM dataset or the literature-curated dataset. Experimentally determined binding affinities of known protein-protein interactions were determined from the PDBbind database (Liu et al., 2015), after mapping human proteins to their mouse orthologs, and the binding affinities of interactions detected by PCP-SILAM were compared to the background of all affinities in the database. Proteins of unknown function were defined as proteins without any GO term annotation in the UniProt GOA database. To characterize functional differences in the PCP-SILAM and literature-curated interactomes, we adapted the approach of (van Leeuwen et al., 2016). For each GO term or pair of GO terms in the GO slim, we computed the proportion of interactions between proteins annotated with the GO term(s) of interest in either the literature-curated interactome or the union of PCP-SILAM interactomes. We then computed the total possible number of interactions in either network, based on the total number of proteins annotated with the GO term(s) of interest in the network, and divided this by the total number of possible protein pairs to obtain the background proportion. We then used these proportions to calculate the enrichment for each network separately, and calculated the difference in enrichment between the literature-curated and PCP-SILAM interactomes as an odds ratio. Statistical significance was assessed via a Z test, with Benjamini-Hochberg correction. Markov clustering was performed using the R package 'MCL' on the union of the literature-curated and PCP-SILAM interactomes, allowing self-loops and otherwise with default parameters. Each cluster was subsequently categorized based on whether the set of proteins therein was connected by literature-curated interactions only, PCP-SILAM

interactions only, or both. To produce **Figure 3.5G**, the number of publications for each gene was calculated using the NCBI file gene2pubmed. Proteins were then divided into 40 evenly sized bins based on the number of publications in which they were mentioned, and the number of interactions between proteins in each pair of bins was calculated, following the methodology of Rolland et al., 2014. Protein abundance in mouse NIH3T3 mouse fibroblasts was obtained from (Schwanhäusser et al., 2011) and used to organize the mouse interactome by protein abundance following an analogous procedure.

3.2.13 Comparison to predicted tissue interactomes

Predicted interactomes for six tissues, based on tissue-specific gene expression, were obtained from IID (Kotlyar et al., 2016); a predicted interactome was not available for thymus. Mouse tissue-specific gene coexpression networks were constructed for six tissues; an insufficient number of samples were available for skeletal muscle. Microarray samples of healthy mouse tissues from the Affymetrix GeneChip Mouse Genome 430 2.0 platform were identified using Bgee (Bastian et al., 2020) and downloaded from ArrayExpress. Samples were processed using BrainArray Custom CDF (Dai et al., 2005) version 21.0.0 and normalized using MAS5 (Hubbell et al., 2002). Probes called as present in fewer than 20% of samples for each tissue were removed. ComBat (Johnson et al., 2007) was used to adjust for batch effects, using each experiment as a batch, following best practices (Vandenbon et al., 2016). Finally, coexpression networks were constructed by taking the top 0.5% of connections, using the Pearson correlation as the similarity measure.

Enrichment for experimentally detected interactions was calculated as the ratio of overlap between tissue interactomes and tissue-specific coexpression networks and predicted tissue

interactomes to the overlap when tissue interactomes were randomly rewired 1,000 times using a degree-preserving algorithm (Maslov and Sneppen, 2002). The number of iterations for the edge rewiring algorithm was set to 6.9 times the number of edges in each network (Ray et al., 2012). Analysis of network topology was performed using the R package 'igraph' (Csardi and Nepusz, 2006). Hub proteins were defined as the top 10% most connected proteins in each network (Batada et al., 2006). The tendency for a protein to interact with different partners across tissues was quantified as the mean Jaccard index across all tissue pairs, with a lower Jaccard index reflecting greater rewiring of protein interactions and a higher Jaccard index reflecting relatively stable interactions across tissues. Enrichment or depletion for interactions at each level of specificity was calculated by calculating the ratio of interactions observed at each tissue specificity relative to random expectation, using the same rewired interactomes as above. No interactions were observed in five or more tissues within randomized networks.

3.2.14 Evolutionary analysis of tissue-specific interactions

The evolutionary conservation of interactions detected by PCP-SILAM was evaluated by comparing interactions at each level of tissue-specificity (that is, interactions detected in one tissue, two tissues, three tissues, and so on) to (i) literature-curated interactions in model organisms, and (ii) systematic screens for protein-protein interactions in humans. Evolutionary conservation of mouse interactions in Saccharomyces cerevisiae, Caenorhabditis elegans, and Drosophila melanogaster was calculated using model organism-specific interactions from BioGRID (Oughtred et al., 2019). Mouse proteins were mapped to their one-to-one orthologs in each organism, as well as human, using InParanoid (Sonnhammer and Östlund, 2015). Human protein-protein interaction screens were obtained and preprocessed as described above. We also

sought to evaluate the evolutionary novelty of tissue-specific interactions using unbiased, genome-wide datasets. To this end, the difference in evolutionary rates between interacting protein pairs was calculated as in (Fraser et al., 2002), and phylogenetic profiles were constructed using the InParanoid database, with the similarity in phylogenetic profile of a protein pair defined as the Pearson correlation in the presence or absence of each protein across all species (Fortelny et al., 2017). Estimates of protein evolutionary age were obtained from ProteinHistorian (Capra et al., 2012). To ensure that significant correlations between tissue specificity and evolutionary conservation were not affected by higher false-positive rates for interactions detected in fewer tissues, we performed two additional sets of analyses. First, we removed all the interactions that were detected in only a single tissue and repeated all analyses using interactions found in anywhere from 2 to 7 tissues. Because these interactions were independently detected at least twice, they are highly unlikely to be false positives. Second, we calculated partial Kendall correlations, controlling for the precision at which each interaction was detected in each tissue. We found the results remained statistically significant in all cases.

To analyze cross-talk between tissue-specific and housekeeping proteins, we calculated tissue specificity of each individual protein as the number of tissues in which it was detected in at least one PCP-SILAM fraction. We then calculated the statistical significance of the over- or under-representation of interactions between housekeeping proteins and proteins detected in one to seven tissues based on the randomly rewired interactomes described above, using Bonferroni correction to correct for multiple hypothesis testing. That is, we calculated the number of interactions in the liver interactome between proteins detected exclusively in the liver and proteins found in all seven tissues, in both real and randomly rewired networks, and expressed this as a Z score. We then repeated a similar calculation for proteins found in two to seven

tissues in the liver interactome, and then repeated this entire procedure for the six remaining tissues. To assess statistical significance across all seven tissues, Z scores were aggregated using Stouffer's method, weighting each Z score by the number of interactions detected in that tissue.

3.2.15 Analysis of interaction rewiring across tissues

We developed a quantitative index reflecting the degree of interaction rewiring for each protein across the seven mouse tissues in our study based on the Jaccard index of its interacting partners, as depicted schematically in Figure 6A. Briefly, for each of the 21 tissue pairs in turn, we computed the Jaccard index for each protein present in both networks, defined as the number of interactions present in both tissues (the intersection) divided by the total number of interactions present in either tissue (the union). We then calculated the mean Jaccard index for each protein over all tissue pairs.

We then asked whether we could leverage this index to identify biochemical determinants of interaction rewiring. To this end, enrichment analysis of rewired proteins was performed using data from the following sources. Intrinsically disordered proteins were identified using IUPred (Dosztányi et al., 2005), with proteins containing more than 30% disordered residues categorized as intrinsically disordered (Gsponer et al., 2008). Proteins containing linear motifs were identified using ANCHOR (Mészáros et al., 2009). Protein phosphorylation data was obtained from (Huttlin et al., 2010), and the tissue specificity of each phosphorylation site was quantified by calculating the Gini coefficient of the spectral counts in each tissue, a measure of inequality that has previously been used to quantify the tissue specificity of gene expression (O'Hagan et al., 2018). We tested for differences in the median Jaccard index across these categories using the Brunner-Munzel test, a nonparametric test of stochastic equality, as implemented in the R

package ‘lawstat.’ The association between phosphorylation tissue specificity was tested using the Spearman rank correlation. Partial Spearman correlations, controlling for intrinsic disorder, were calculated using the ‘ppcor’ R package.

We then investigated whether interaction rewiring resulted in the formation of tissue-specific interactions involved in cell-cell signalling. We tested for associations between interaction tissue specificity and the presence of a protein kinase, transcription factor, or cell surface receptor using the Kendall rank correlation. Transcription factors were obtained from the Mouse TF Atlas (Zhou et al., 2017). Cell surface protein receptors were obtained from (Ramilowski et al., 2015). Protein kinases were obtained from the UniProt protein kinase index (https://www.uniprot.org/docs/pkinfam). Edge betweenness centrality, a topological metric that reflects information flow through a network and is defined as the number of shortest paths between any two proteins that pass through a given edge, was calculated in the aggregate PCP-SILAM interactome formed by the union of unique interactions across all seven tissues, using the R package ‘igraph’ (Csardi and Nepusz, 2006).

We subsequently asked whether cellular strategies were in place to regulate the availability of rewired proteins. Protein and mRNA abundance, half-lives, and translation and transcription rates were obtained from (Schwanhäusser et al., 2011), who used parallel metabolic pulse labelling to quantify the entire cascade of gene expression in mouse NIH3T3 fibroblasts using absolute units (e.g., molecules of RNA per cell). We then computed associations with the mean Jaccard index using the Spearman rank correlation, and again performed partial correlation analyses using the ‘ppcor’ R package. To evaluate the impact of technical limitations in the detection or quantification of low-abundance proteins on these findings, we performed two additional Jaccard index calculations. First, we reasoned that restricting our analysis to proteins

that were detected in all seven tissues would mitigate the impact of sporadic protein identification. Accordingly, we filtered the tissue interactomes to include only interactions between these 'housekeeping' proteins, then repeated the Jaccard index calculation. Second, we divided the proteins identified in this study into three bins of lowly, moderately, and highly abundant proteins, based on their summed total intensity across all fractions. We then removed proteins from the bottom two bins from the tissue interactomes and repeated the Jaccard index calculation only for the top tercile of highly abundant proteins.

Last, we analyzed the relationship between interaction rewiring and disease. Mouse disease genes were obtained from the Mouse Genome Database (Smith et al., 2018). Tissue-specific disease genes, and their associated tissues, were obtained from (Basha et al., 2020), and network interconnectivity was calculated using the mean shortest path as previously described (Menche et al., 2015), restricting the analysis to disease-tissue pairs in which at least four disease genes were present in the corresponding tissue interactome.

3.2.16 Data and software availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (Vizcaíno et al., 2014) via the PRIDE partner repository (Vizcaíno et al., 2016) with the dataset identifier PXD022309. In addition, processed chromatograms for each tissue have been deposited to the EMBL-EBI BioStudies database (Sarkans et al., 2018), with accession S-BSST152.

3.3 Results

3.3.1 Quantitative *in vivo* interactome profiling of mouse tissues

We profiled the interactome of seven mouse tissues, including brain, heart, skeletal muscle (gastrocnemius), lung, kidney, liver, and thymus, using PCP-SILAM (**Figure 3.1A**). The *in vivo* interactome of each tissue was preserved by extracting and separating protein complexes under non-denaturing conditions and in the presence of excess protease and phosphorylase inhibitors. A total of 55 SEC fractions were collected from each tissue in biological duplicate, from both $^{13}C_6$-labeled (heavy) and unlabeled (light) mice. The heavy fractions were then pooled to generate a global reference mixture, which was spiked into all 770 light fractions. Each fraction was then subjected to liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis, and the resulting dataset was processed using MaxQuant (Cox and Mann, 2008).

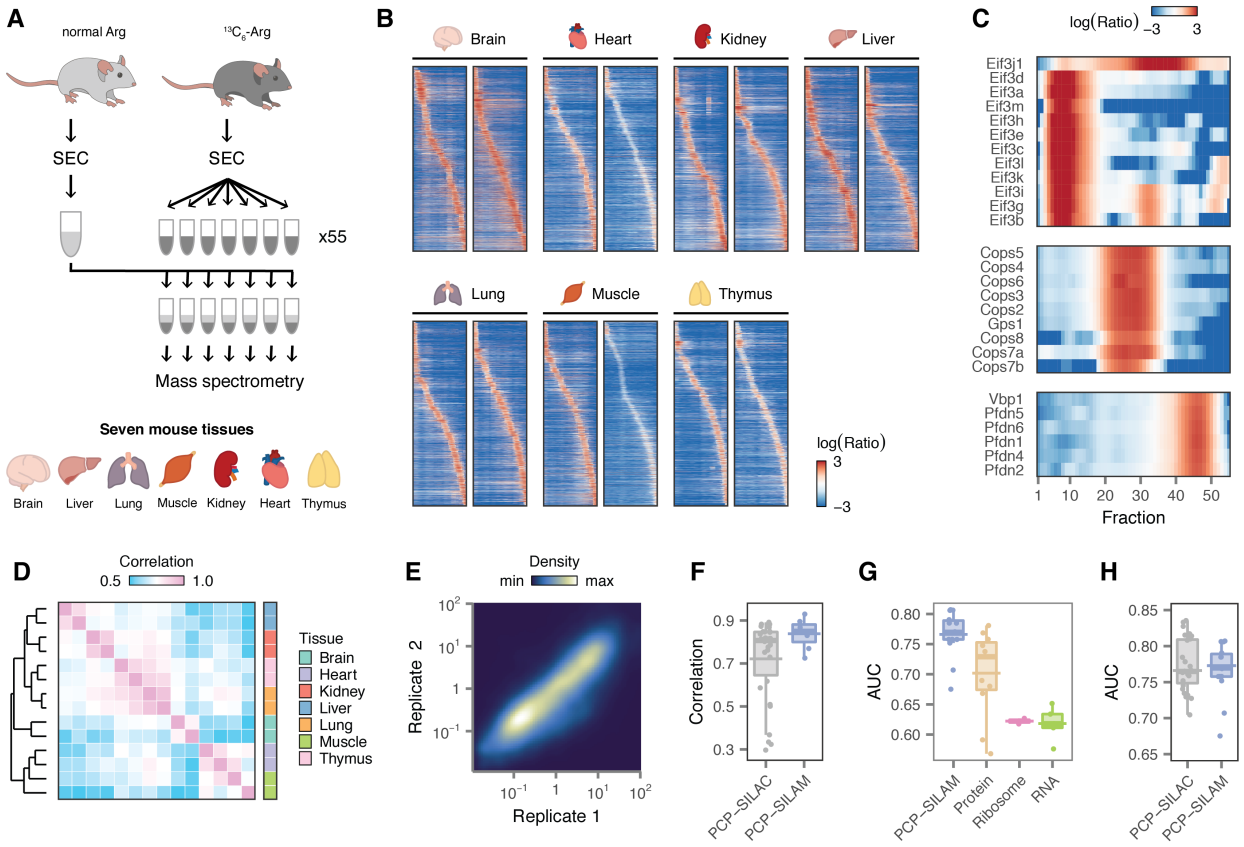


Figure 3.1 Quantitative interactome profiles of seven mouse tissues with PCP-SILAM. (A) PCP-SILAM workflow for interactome mapping in mouse tissues. (B) Elution profiles of 7,225 proteins across 770 SEC fractions spanning seven mouse tissues. (C) Enlarged elution profiles of representative protein complexes (top, eIF3 complex; middle, COP9 signalosome; bottom, prefoldin complex). (D) Hierarchical clustering of PCP-SILAM replicates. (E) Reproducibility of PCP-SILAM protein quantification. (F) Rank correlations between biological replicates in PCP-SILAM and in published PCP-SILAC data (Kerr et al., 2020; Scott et al., 2015, 2017). Horizontal lines show the mean correlation. (G) Recovery of known protein complexes by patterns of co-abundance in PCP-SILAM data, as compared to large-scale proteomics, transcriptomics, and ribosome profiling datasets. Horizontal lines in (F-G) show the mean AUC. (H) Recovery of known protein complexes in PCP-SILAM and published PCP-SILAC data.

A total of 7,225 unique proteins were detected across all fractions (**Figure 3.1B**). The set of identified proteins encompassed many well-known protein complexes, as exemplified by the

prefoldin complex, the COP9 signalosome, or the eIF3 complex (**Figure 3.1C**). Notably, PCP-SILAM chromatograms differentiated non-constitutive subunits, such as the loosely bound eIF3j subunit, and complex isoforms, such as the eIF3b-eIF3g-eIF3i submodule (Valášek et al., 2017). Consistent with the presence of the SILAM global reference, we observed near-saturation in the quantitation of novel proteins with additional PCP fractions (**Figure 3.2A**). Reproducibility between biological replicates was high, with samples clustering by tissue rather than by batch, and quantitatively comparable to cell-line-based PCP-SILAC (**Figures 3.1D-F and 3.2A**).

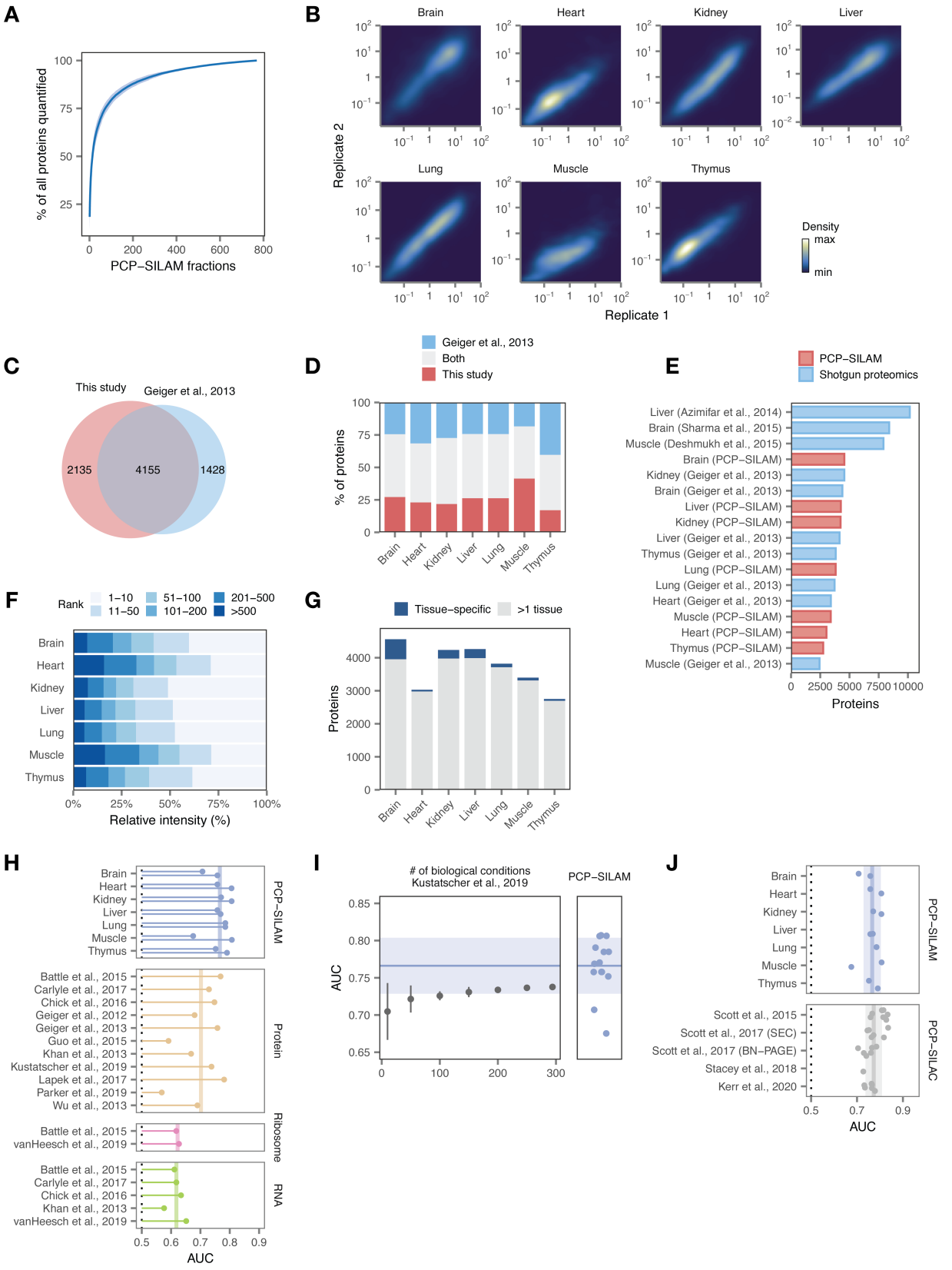

A
% of all proteins quantified
PCP−SILAM fractions
B
Brain
Heart
Kidney
Liver
Lung
Muscle
Thymus
Replicate 2
Replicate 1
Density
max
min
C
This study
Geiger et al., 2013
2135
4155
1428
D
Geiger et al., 2013
Both
This study
% of proteins
E
PCP−SILAM
Shotgun proteomics
Liver (Azimifar et al., 2014)
Brain (Sharma et al., 2015)
Muscle (Deshmukh et al., 2015)
Brain (PCP−SILAM)
Kidney (Geiger et al., 2013)
Brain (Geiger et al., 2013)
Liver (PCP−SILAM)
Kidney (PCP−SILAM)
Liver (Geiger et al., 2013)
Thymus (Geiger et al., 2013)
Lung (PCP−SILAM)
Lung (Geiger et al., 2013)
Heart (Geiger et al., 2013)
Muscle (PCP−SILAM)
Heart (PCP−SILAM)
Thymus (PCP−SILAM)
Muscle (Geiger et al., 2013)
Proteins
F
Rank
1−10
11−50
51−100
101−200
201−500
>500
Relative intensity (%)
G
Tissue−specific
>1 tissue
Proteins
H
PCP−SILAM
Protein
Ribosome
RNA
Battle et al., 2015
Carlyle et al., 2017
Chick et al., 2016
Geiger et al., 2012
Geiger et al., 2013
Guo et al., 2015
Khan et al., 2013
Kustatscher et al., 2019
Lapek et al., 2017
Parker et al., 2019
Wu et al., 2013
vanHeesch et al., 2019
AUC
I
# of biological conditions
Kustatscher et al., 2019
PCP−SILAM
AUC
J
PCP−SILAM
PCP−SILAC
Scott et al., 2015
Scott et al., 2017 (SEC)
Scott et al., 2017 (BN−PAGE)
Stacey et al., 2018
Kerr et al., 2020
AUC

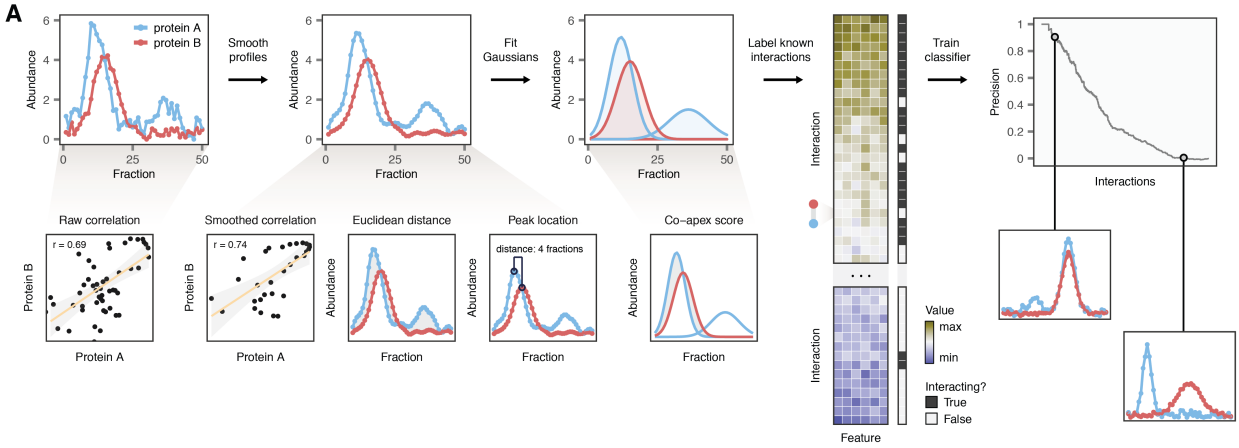
Figure 3.2 Quantitative profiling of mouse tissue interactomes by PCP-SILAM. (A) Saturation of unique proteins identified with additional PCP fractions, consistent with the presence of the SILAM global reference. The mean and standard deviation of 100 bootstraps are shown. (B) Reproducibility of PCP-SILAM protein quantification within individual tissues. SILAM intensities for each protein in each fraction are shown for each pair of replicates as a two-dimensional density plot. (C) Unique proteins identified across all seven mouse tissues by this study and (Geiger et al., 2013). (D) Unique proteins identified within each mouse tissue by this study and (Geiger et al., 2013). (E) Total number of proteins quantified in this study and previous deep proteomic analyses of mouse tissues (Azimifar et al., 2014; Deshmukh et al., 2015; Geiger et al., 2013; Sharma et al., 2015). (F) Distribution of protein abundances within tissue interactomes. Summed ion intensities of ranked proteins are presented as a percentage of the total summed ion intensities. (G) Proteins identified only in a single tissue (tissue-specific), or in at least one other tissue, by PCP-SILAM. (H) Recovery of known protein complexes based on co-abundance in PCP-SILAM fractions or large-scale proteomic, transcriptomic, or translatomic datasets, for each individual dataset shown in **Figure 3.1G**. (I) Recovery of known protein complexes based on co-abundance in PCP-SILAM or co-expression in subsets of between 10 and 300 experiments from the ProteomeHD resource (Kustatscher et al., 2019). (J) Recovery of known protein complexes in PCP-SILAM and published, cell-line-based PCP-SILAC datasets.

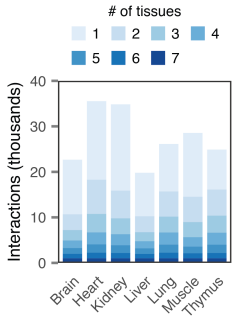
The total proteome coverage achieved by PCP-SILAM was slightly lower than in previous deep proteomic analyses of murine tissues (Azimifar et al., 2014; Deshmukh et al., 2015; Geiger et al., 2013; Sharma et al., 2015) (**Figure 3.2C-E**), likely due to less extensive fractionation, and because not all components of the proteome participate in macromolecular interactions amenable to analysis by SEC. Consistent with shotgun proteomics studies (Geiger et al., 2013), the tissue interactome profiles were dominated by highly abundant proteins, which made up ~75% of total protein mass (**Figure 3.2F**), and relatively few proteins are observed to be exclusively quantified within a single tissue (range, 1.8%–17.4%; **Figure 3.2G**) (Kim et al., 2014).

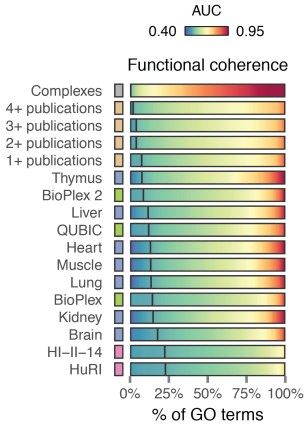
As an initial assessment of the quality of our data, we compared PCP-SILAM to large-scale proteomics, transcriptomics, and ribosome profiling datasets for their ability to recover known protein complexes (Romanov et al., 2019). We used receiver operating characteristic (ROC) curve analysis to quantify the degree to which the observed patterns of co-abundance in each dataset could separate pairs of proteins within the same complex from a vast background of non-interacting protein pairs. Patterns of co-abundance across SEC fractions consistently proved more informative than those in large-scale transcriptome, translatome, or proteome datasets (**Figures 3.1G and 3.2H**), with a mean area under the ROC curve (AUC) of 0.77, compared to 0.70 for eleven large-scale proteomics studies ($p = 0.016$, paired t-test), 0.62 for five RNA-seq studies ($p = 4.9 \times 10^{-5}$), and 0.62 in a pair of ribosome profiling studies ($p = 2.0 \times 10^{-8}$). Notably, the mean AUC in PCP-SILAM data was higher than that achieved in a meta-analysis of 5,288 SILAC proteomics experiments from 294 different biological conditions (Kustatscher et al., 2019) (**Figure 3.2I**). These findings illustrate the primary advantage of PCP over large-scale proteome 'co-regulation' networks (Kustatscher et al., 2019; Lapek et al., 2017): namely, that separation over the SEC column should specifically discriminate pairs of proteins in the same protein complex from indirectly associated pairs, whose abundance fluctuates in a correlated manner due to involvement in a common underlying biological process. No significant difference in AUC was observed between PCP-SILAM and cell-line-based PCP-SILAC (**Figures 3.1H** and **3.2J**; $p = 0.58$), indicating that adaptation for *in vivo* interactome profiling did not compromise data quality, despite the greater complexity of entire tissues.

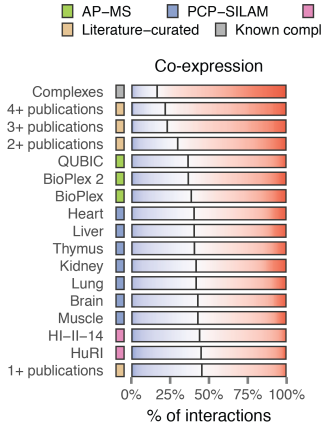
3.3.2 Inference of high-confidence mouse tissue interactomes

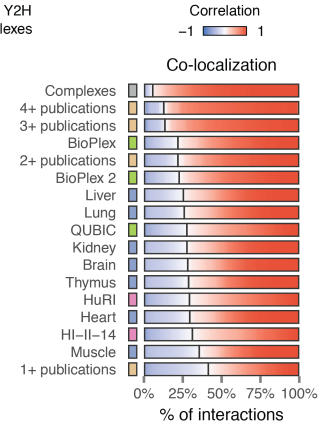
To derive interactomes from PCP-SILAM tissue proteome profiles, we developed PrInCE, a machine-learning pipeline for analysis of co-fractionation mass spectrometry data (**Figure 3.3A**) (Stacey et al., 2017). PrInCE builds on previous approaches for network inference from co-fractionation data by training a machine-learning classifier to identify interacting protein pairs based on the elution patterns of known protein complexes (Havugimana et al., 2012; Wan et al., 2015). However, in contrast to previous approaches that learn jointly from proteomic data and publicly available genomic data such as gene coexpression, phylogenetic profiles, or previously published interactions, PrInCE recovers PPIs using features derived exclusively from mass spectrometric data (Skinnider et al., 2018b). Briefly, PrInCE deconvolves each chromatogram into a mixture of Gaussian peaks, then computes six features for each protein pair that reflect the likelihood of the two proteins interacting. These include the Pearson correlation (which is calculated on both raw and imputed chromatograms) between the two proteins, and its p-value; the Euclidean distance; the number of fractions separating the maximum values of each chromatogram (the "peak location" score); and the Euclidean distance between the single closest pair of Gaussians (the "co-apex" score). These six features are subsequently provided as input to a classifier that is trained in cross-validation on a subset of known protein complexes that are known to be amenable to analysis by co-fractionation mass spectrometry (Stacey et al., 2018). This classifier subsequently calculates an interaction probability for every pair. Importantly, the identification of Gaussian peaks within the chromatogram allows PrInCE to identify interacting protein pairs that participate in multiple complexes, so long as they co-elute within a single shared peak (Stacey et al., 2017).

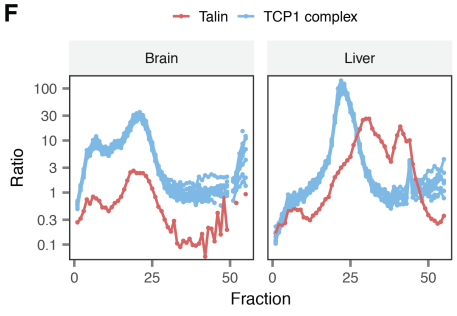

A
protein A
protein B
Abundance
Fraction
Smooth profiles
Fit Gaussians
Label known interactions
Train classifier
Raw correlation
Smoothed correlation
Euclidean distance
Peak location
Co-apex score
r = 0.69
r = 0.74
distance: 4 fractions
Protein A
Protein B
Interaction
Feature
Value
max
min
Interacting?
True
False
Precision
Interactions

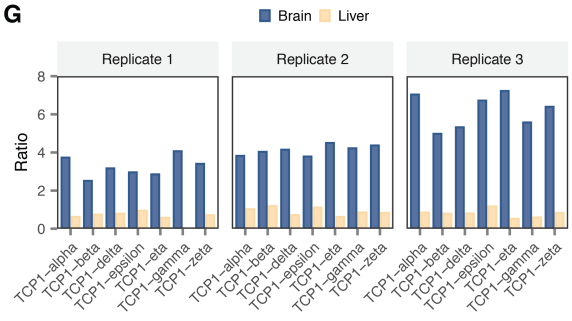

B
# of tissues
1
2
3
4
5
6
7
Interactions (thousands)
Brain
Heart
Kidney
Liver
Lung
Muscle
Thymus

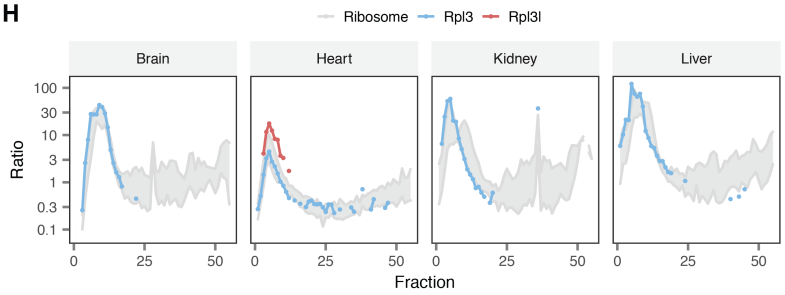

C
AUC
0.40
0.95
Functional coherence
Complexes
4+ publications
3+ publications
2+ publications
1+ publications
Thymus
BioPlex 2
Liver
QUBIC
Heart
Muscle
Lung
BioPlex
Kidney
Brain
HI–II–14
HuRI
0%
25%
50%
75%
100%
% of GO terms


D
AP–MS
PCP–SILAM
Y2H
Literature–curated
Known complexes
Co-expression
Complexes
4+ publications
3+ publications
2+ publications
QUBIC
BioPlex 2
BioPlex
Heart
Liver
Thymus
Kidney
Lung
Brain
Muscle
HI–II–14
HuRI
1+ publications
0%
25%
50%
75%
100%
% of interactions


E
Correlation
–1
1
Co-localization
Complexes
4+ publications
3+ publications
BioPlex
2+ publications
BioPlex 2
Liver
Lung
QUBIC
Kidney
Brain
Thymus
HuRI
Heart
HI–II–14
Muscle
1+ publications
0%
25%
50%
75%
100%
% of interactions


F
Talin
TCP1 complex
Brain
Liver
Ratio
Fraction


G
Brain
Liver
Replicate 1
Replicate 2
Replicate 3
Ratio
TCP1–alpha
TCP1–beta
TCP1–delta
TCP1–epsilon
TCP1–eta
TCP1–gamma
TCP1–zeta


H
Ribosome
Rpl3
Rpl3l
Brain
Heart
Kidney
Liver
Ratio
Fraction

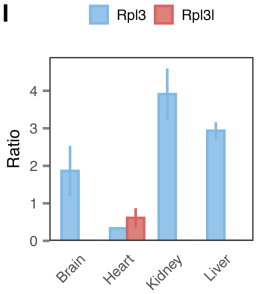

I
Rpl3
Rpl3l
Ratio
Brain
Heart
Kidney
Liver

Figure 3.3 Inference and validation of mouse tissue interactomes. (A) Schematic illustration of the computational procedure for inference of interactome networks from PCP-SILAM data (Stacey et al., 2017). (B) Number and tissue specificity of interactions detected in each mouse tissue. (C-E) Comparison of mouse tissue interactomes to high-throughput screens of the human interactome, literature-curated interactions, and known protein complexes. (C) Functional coherence of interactome networks, as quantified by the AUC of protein function prediction in cross-validation. Vertical lines indicate the proportion of GO terms with AUC less than 0.5 (random chance). (D) Co-expression of interacting protein pairs in quantitative proteomics data from 41 cancer cell lines (Lapek et al., 2017). Vertical lines indicate the proportion of negatively correlated interacting pairs (Stacey et al., 2018). (E) Co-localization of interacting protein pairs, as quantified by their correlation across cellular fractions in hyperplexed localization of organelle proteins by isotope tagging (hyperLOPIT) data (Geladaki et al., 2019). Vertical lines indicate the proportion of negatively correlated interacting pairs. (F) PCP-SILAM profiles of talin and the TCP1 complex in the mouse brain and liver. (G) Relative abundance of co-immunoprecipitated TCP1 complex subunits to talin in mouse brain and liver. (H) PCP-SILAM chromatograms for Rpl3, Rpl3l, and the 60S ribosome (median and interquartile range, ribbon) in the mouse brain, heart, kidney, and liver. (I) Relative abundance of Rpl3 and Rpl3l in purified ribosomes from the same four tissues. Error bars show standard error.

Applying PrInCE to PCP-SILAM data identified between 19,804 and 35,536 interactions in each tissue at a 5% false discovery rate (FDR), for a total of 125,696 unique interactions (**Figure 3.3B**). To evaluate the quality of the networks inferred from PCP-SILAM data, we compared each mouse tissue interactome to five recently published high-throughput human interactome screens (Hein et al., 2015; Huttlin et al., 2015, 2017; Luck et al., 2020; Rolland et al., 2014), conducted using AP-MS or Y2H, as well as literature-curated interactions reported in one, two, three, or four publications and known protein complexes. We computed three indices that reflect the concordance of each network with other large-scale genomic datasets. First, we evaluated the functional coherence of the network, defined as the degree to which the function of

any given protein can be predicted from those of its interacting partners, based on the principle of 'guilt by association' (Ballouz et al., 2017; Oliver, 2000). We used a simple neighbor-voting algorithm to predict Gene Ontology terms in cross-validation for each protein, based on the frequency of each GO term among its interacting partners, and compared the AUC across GO terms.; Second, we evaluated the tendency for interacting protein pairs to display correlated patterns of abundance in large-scale proteomic datasets (Kustatscher et al., 2019; Lapek et al., 2017). Last, we evaluated the degree to which interacting proteins localize to the same subcellular compartments by computing the Pearson correlations between interacting pairs across organellar fractions in two subcellular proteomics datasets (Geladaki et al., 2019; Orre et al., 2019). All three indices yielded broadly concordant pictures of network quality (**Figures 3.3C-E** and **3.4A-C**), generally suggesting that interactions inferred by PCP-SILAM were superior to Y2H, comparable or slightly inferior to AP-MS, and intermediate between literature-curated interactions reported in one and two publications. Thus, these findings suggest the quality of PCP-SILAM networks is comparable both to recent systematic human screens, and to interactions identified by small-scale, hypothesis-driven experiments.

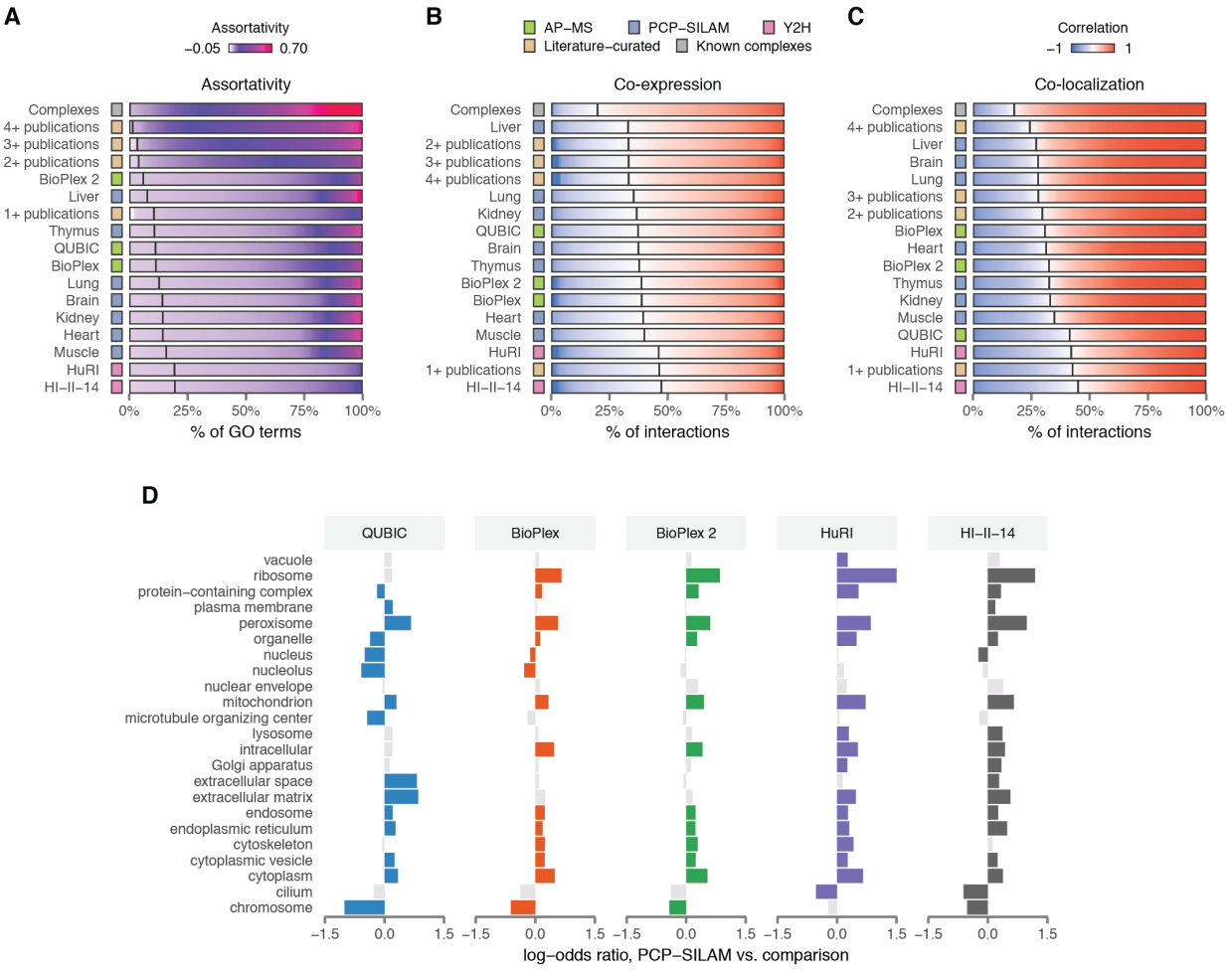


Figure 3.4 Validation of mouse tissue interactomes. (A) Assortativity of GO terms in interactome networks. Vertical lines indicate the proportion of GO terms with assortativity less than 0. (B) Co-expression of interacting protein pairs in the ProteomeHD resource (Kustatscher et al., 2019). Vertical lines indicate the proportion of negatively correlated interacting pairs. (C) Co-localization of interacting protein pairs, as quantified by their correlation across cellular fractions in subcellular proteomics data (Orre et al., 2019). Vertical lines indicate the proportion of negatively correlated interacting pairs. (D) Subcellular localizations of interacting proteins in the PCP-SILAM aggregate interactome compared to five recent human high-throughput interactome maps. Enrichments shown in light grey are not statistically significant.

We also compared the subcellular localizations of interacting proteins detected by PCP-SILAM to those detected by AP-MS or Y2H (**Figure 3.4D**). Although the PCP-SILAM interactome was slightly depleted for membrane and nuclear proteins, its subcellular distribution was broadly similar to interactomes detected with these methods.

To experimentally validate the ability of PCP-SILAM to identify novel interactions in mouse tissues, we focused on a putative novel interaction between the cytoskeletal protein talin and the subunits of the chaperonin-containing TCP1 complex, which had not previously been reported in either mouse or human (**Figure 3.3F**). Immunoprecipitation of talin confirmed the interaction with TCP1 in mouse brain (**Figure 3.3G**). To confirm the tissue specificity of the interaction, we additionally performed immunoprecipitation in mouse liver, where talin and TCP1 complex curves displayed limited correlation, and observed limited co-purification, consistent with PCP-SILAM (**Figure 3.3F-G**). We also examined an apparent heart-specific paralog switch in the ribosome, driven by the replacement of the constitutive Rpl3 by its paralog Rpl3l (**Figure 3.3H**). Rpl3l is primarily expressed in heart and skeletal muscle, and mutations of Rpl3l are implicated in atrial fibrillation (Thorolfsdottir et al., 2018). However, it has remained unclear whether Rpl3l is incorporated into ribosomes, or whether its physiological effects are mediated by extra-ribosomal functions (Chaillou et al., 2016). Isolation of purified ribosomes from mouse tissues, followed by label-free quantitation, confirmed the incorporation of Rpl3l into heart-specific ribosomes (**Figure 3.3I**). Thus, orthogonal biochemical techniques confirmed the ability of PCP-SILAM to identify dynamic rearrangements in the interactome across tissues.

3.3.3 Unbiased expansion of the mouse interactome by PCP-SILAM

The mouse is a ubiquitous model organism, yet its interactome has never been the subject of a systematic, proteome-scale mapping effort. We compared the PCP-SILAM mouse interactome to a literature-curated (LC) mouse interactome derived by assembling a total of 82,602 mouse PPIs from nine interaction databases. Strikingly, of the 125,696 unique interactions detected by PCP-SILAM, only 4,354 (2.1%) overlap with interactions detected by small-scale experiments (**Figure 3.5A**). This overlap is significantly larger than would be expected by chance ($p < 10^{-15}$, hypergeometric test), but small in magnitude, suggesting that PCP-SILAM is largely complementary to small-scale methods. Among literature-curated PPIs, interactions detected by PCP-SILAM intersect most significantly with those detected by cosedimentation, and less with PPIs detected by two-hybrid or cross-linking approaches (**Figure 3.6A**), further supporting the notion that each assay recovers characteristic subsets of the interactome. We additionally asked whether PCP-SILAM exhibits detectable bias towards stable, high-affinity PPIs, but found that PCP-SILAM interactions catalogued in the PDBbind database (Liu et al., 2015) spanned a broad range of binding affinities (**Figure 3.6B**).

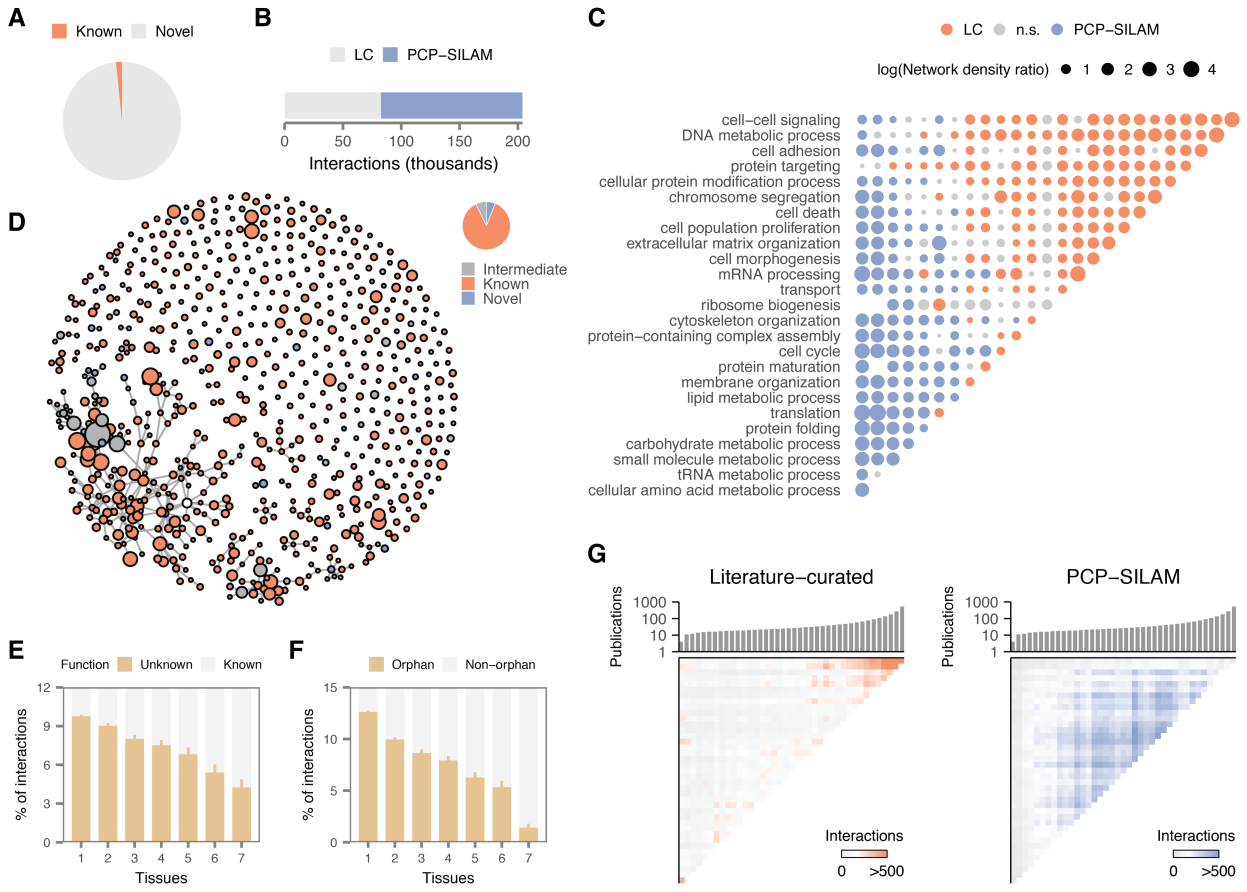


Figure 3.5 Unbiased expansion of the literature-curated mouse interactome by PCP-SILAM. (A) Proportion of previously known mouse interactions observed in PCP-SILAM tissue interactomes. (B) Size of the known mouse interactome before and after this study. LC, literature-curated. (C) Comparison of interacting protein co-annotation within and across biological processes between PCP-SILAM and literature-curated interactions. (D) Markov clustering of the global mouse interactome, and proportion of protein communities containing exclusively known interactions, exclusively novel interactions, or both. (E) Proportion of interactions involving proteins of unknown function for interactions detected in one to seven tissues. (F) Proportion of interactions involving proteins for which no interactions were previously known (interactome "orphans") for interactions detected in one to seven tissues. (G) Number of interactions between proteins binned by number of publications, and ordered along both axes. Histogram shows the median number of publications in each bin.

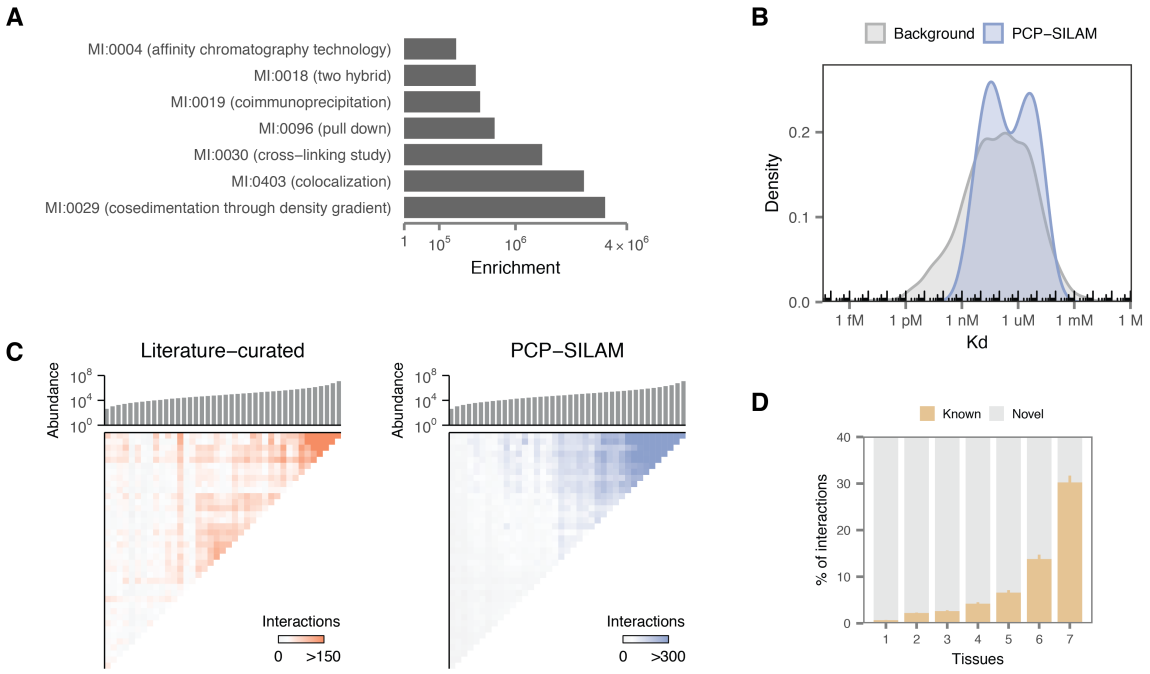


Figure 3.6 Expansion of the known mouse interactome by PCP-SILAM. (A) Enrichment for overlap with literature-curated mouse interactions detected with different methods, relative to random expectation. (B) Experimentally determined binding affinities of protein-protein interactions in the PDBbind database (Liu et al., 2015) and the subset of PDBbind recovered by PCP-SILAM. (C) Number of interactions between proteins binned by mean abundance (copies per cell) in mouse fibroblasts (Schwanhäusser et al., 2011), and ordered along both axes. Histogram shows the median abundance in each bin. (D) Proportion of previously known mouse interactions among PCP-SILAM interactions detected in one to seven tissues.

The remaining 121,342 unique mouse interactions detected in this study are novel; thus, our proteome-scale resource therefore expands the size of the known mouse interactome by a factor of almost 2.5 (**Figure 3.5B**). To functionally characterize these novel PPIs, we compared patterns of GO term co-annotation between literature-curated and PCP-SILAM interactions (**Figure 3.5C**). We calculated the number of interactions between and within GO terms in the PCP-SILAM and literature-curated networks, then evaluated the statistical significance of the

observed odds ratios. Relative to literature-curated interactions, mouse PPIs detected by PCP-SILAM were enriched for connections involving metabolism, translation, and protein folding. In contrast, PCP-SILAM PPIs were underrepresented in connections involving cell-cell signalling and proliferation. These findings are broadly consistent with the expectation that PCP-SILAM would prioritize cytosolic over nuclear or extracellular complexes.

We next asked whether PPIs detected by PCP-SILAM preferentially expanded existing regions of the global mouse interactome, or tended to form completely new subnetworks. To provide a global view of network topology, we applied unsupervised Markov clustering (Enright et al., 2002) to group the entire mouse interactome, including both literature-curated and PCP-SILAM interactions, into 696 clusters of three to 980 proteins (**Figure 3.5D**). Of the 92 clusters containing at least one PPI detected by PCP-SILAM, 50 (54%) included both literature-curated and high-throughput PPIs, while 42 were composed exclusively of PCP-SILAM interactions. Thus, while PCP-SILAM reveals several protein communities that were altogether unknown in mouse, many interactions also expand neighborhoods of the mouse interactome with foundations previously defined by small-scale experiments.

Literature-curated protein interaction datasets have been criticised on the grounds that they are biased towards a relatively small set of highly studied proteins. We organized the literature-curated mouse interactome by ranking proteins based on the number of publications in which they have been mentioned, as in a previous study in humans (Rolland et al., 2014), and found that the mouse literature-curated interaction dataset is dominated by interactions between well-studied proteins, at the expense of a “sparse zone” of poorly studied proteins, for which few PPIs are known (**Figure 3.5G**). High-throughput interactome mapping studies provide a means to define the architecture of the proteome independent of investigator biases, and in comparison

to the literature-curated dataset, PPIs detected by PCP-SILAM are distributed more homogeneously (**Figure 3.5G**). However, we anticipated that the untargeted nature of our *in vivo* approach could result in a reduced capacity to detect interactions for lowly expressed proteins, relative to targeted approaches. To test this notion, we re-organized both the literature-curated and PCP-SILAM interactomes by protein abundance (Schwanhausser et al., 2011). Consistent with this expectation, PPIs detected by PCP-SILAM display a moderate bias towards more abundant proteins, whereas interactions detected by small-scale experiments largely do not (**Figure 3.6C**).

To date, large-scale mammalian interactome mapping projects have typically been performed in yeast cells or cell lines (Snider et al., 2015). However, many biologically relevant PPIs may not occur within these systems. We therefore hypothesized that mapping the *in vivo* interactome in physiological contexts, such as in individual tissues, could preferentially reveal novel, context-specific interactions. Consistent with this hypothesis, we found that tissue-specific interactions were significantly less likely to have been catalogued in literature-curated interaction databases (**Figure 3.6D**; $p < 10^{-15}$, Kendall rank correlation). We also investigated whether mapping interactomes in mouse tissues could preferentially provide insights into interactions involving proteins whose functions are poorly understood. PCP-SILAM identified interacting partners for 218 proteins of unknown function, and these interactions were significantly more tissue-specific than the interactome average (**Figure 3.5E**; $p < 10^{-15}$, Kendall rank correlation). Similarly, we mapped interactions involving 366 proteins for which no interacting partners had previously been detected (interactome "orphans") (Kotlyar et al., 2015), and found that these interactions likewise displayed a significant trend towards increasing tissue specificity (**Figure**

3.5F; $p < 10^{-15}$, Kendall rank correlation). Thus, mapping the *in vivo* interactome of mammalian tissues can shed light on poorly studied components of the proteome.

3.3.4 Widespread interactome rewiring limits the accuracy of tissue-specific interactome prediction

In the absence of experimental tissue- or cell type-specific interactomes, computational methods have been developed to reconstruct context-specific molecular interaction networks, with a view to understanding network perturbations in tissue-specific pathologies (Greene et al., 2015; Marbach et al., 2016). The most widely used such strategy for context-specific interactome prediction proceeds from the notion that the protein products of two genes can only interact in a given context if these genes are both expressed. Thus, gene or protein expression data is overlaid onto a static interactome, and the subset of the network whose nodes are expressed above a certain threshold is extracted to generate the context-specific interactome (**Figure 3.7A**) (Bossi and Lehner, 2009; Buljan et al., 2012; de Lichtenberg et al., 2005). For instance, human tissue transcriptome data from the GTEx project (Melé et al., 2015) was overlaid onto a draft map of the human interactome by Y2H (Luck et al., 2020) to infer tissue-specific networks. An alternative strategy is to construct tissue-specific gene coexpression networks, which suggest functional association in a given tissue, if not necessarily physical interaction (Pierson et al., 2015; Saha et al., 2017). However, the degree to which predictions made by these methods capture physiologically relevant interactome rearrangements are unclear. Our PCP-SILAM dataset provides an opportunity to experimentally test the accuracy of gene expression or coexpression-based methods in context-specific interactome prediction for the first time.

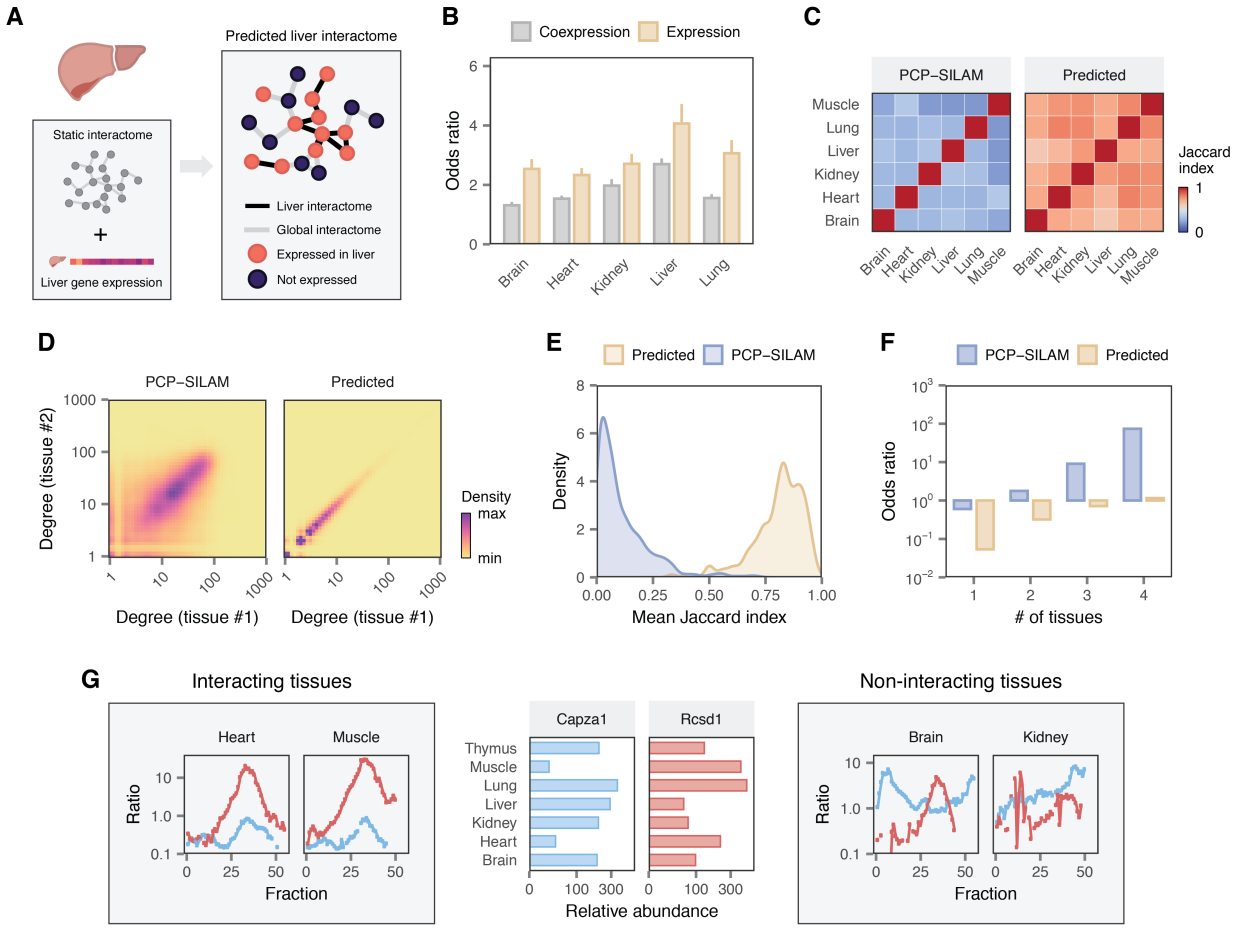


Figure 3.7 Interactome rewiring limits accuracy of tissue-specific interactome prediction. (A) Schematic illustration of tissue interactome prediction by integration of static interactome maps with tissue-specific gene expression profiles. (B) Overlap between PCP-SILAM mouse tissue interactomes and tissue-specific gene coexpression networks or tissue interactomes predicted based on gene expression, relative to rewired networks. (C) Overlap in hub proteins across tissues in predicted and PCP-SILAM tissue interactomes. (D) Variability in protein degree across tissues in predicted and PCP-SILAM tissue interactomes. (E) Protein rewiring across tissues in predicted and PCP-SILAM tissue interactomes, as quantified by the mean Jaccard index of each protein across all tissue pairs. (F) Enrichment for interactions found in one to four tissues in predicted and PCP-SILAM tissue interactomes, relative to randomized networks. (G) PCP-SILAM profiles reveal Rcsd1 and Capza2 interact in heart and muscle, but not brain or kidney, despite expression in all seven tissues. Left, PCP-SILAM chromatograms for

Rcsd1 (orange) and Capza2 (purple) in heart and muscle. Middle, summed SILAM ratios for Capza2 and Rcsd1 in each tissue. Right, PCP-SILAM chromatograms in brain and kidney.

Using our PCP-SILAM mouse tissue interactomes as a reference, we investigated the accuracy of these approaches in predicting tissue-specific interactomes. We calculated the overlap between the predicted and PCP-SILAM tissue interactomes, then compared this overlap to that observed for randomly rewired networks. Surprisingly, tissue interactomes predicted based on gene expression were only two- to four-fold enriched for experimentally detected interactions, relative to randomized networks (range, 2.3–4.1; **Figure 3.7B**), an overlap that was highly significant, but remarkably small in magnitude. Similarly modest enrichment was observed for tissue-specific coexpression networks (**Figure 3.7B**), consistent with previous findings that gene coexpression is a relatively poor predictor of physical interaction (Fortelny et al., 2017; Kühner et al., 2009). Thus, neither tissue-specific gene expression, nor coexpression, are sufficient to accurately predict tissue-specific physical PPIs.

Predicted tissue interactomes also differed markedly in their topology from experimentally determined interactomes. In interactome networks, the most highly connected ("hub") proteins are slow-evolving and physiologically indispensable (Fraser et al., 2002; Jeong et al., 2001). We identified hub proteins as the top 10% most connected proteins in each network (Batada et al., 2006), and compared the identities of hub proteins across PCP-SILAM or predicted interactomes. The hub proteins of predicted tissue interactomes were highly consistent across tissues, with 65–70% of hubs in each tissue shared between all predicted networks (**Figure 3.7C**). However, we found that hub proteins were much less consistent *in vivo*, with only 20 hubs shared across all tissues (**Figure 3.7C**). Moreover, this trend could not be attributed

to differences in proteome coverage between tissues: most hub proteins were present in all tissue interactomes, but differed specifically in their connectivity (**Figure 3.8A**). The identities of the hub proteins themselves in each tissue were also poorly predicted, with only 3.7–5.9% of hub proteins overlapping between predicted and *in vivo* interactomes (**Figure 3.8B**). More generally, the number of interactions in which any given protein participates (its degree) was considerably more variable across tissues in *in vivo* interactomes than in predicted networks (**Figures 3.7D** and **3.8C-D**). Similarly, we computed the tendency for each protein to have shared or divergent interaction partners across tissues in both the in vivo and predicted networks, and found that proteins displayed a much greater tendency to interact with different partners across tissues than was predicted by gene expression alone (**Figures 3.7E** and **3.8E**). These results can be rationalized on the basis that, in predicted tissue interactomes, a protein expressed in a given tissue retains all of its interactions with other proteins expressed in that tissue, since differences in degree can be caused solely by absence of the protein or its partners from the tissue in question. Consequently, both the degree of proteins in the network, as well as the specific identities of their interacting partners, remain artificially stable across tissues.

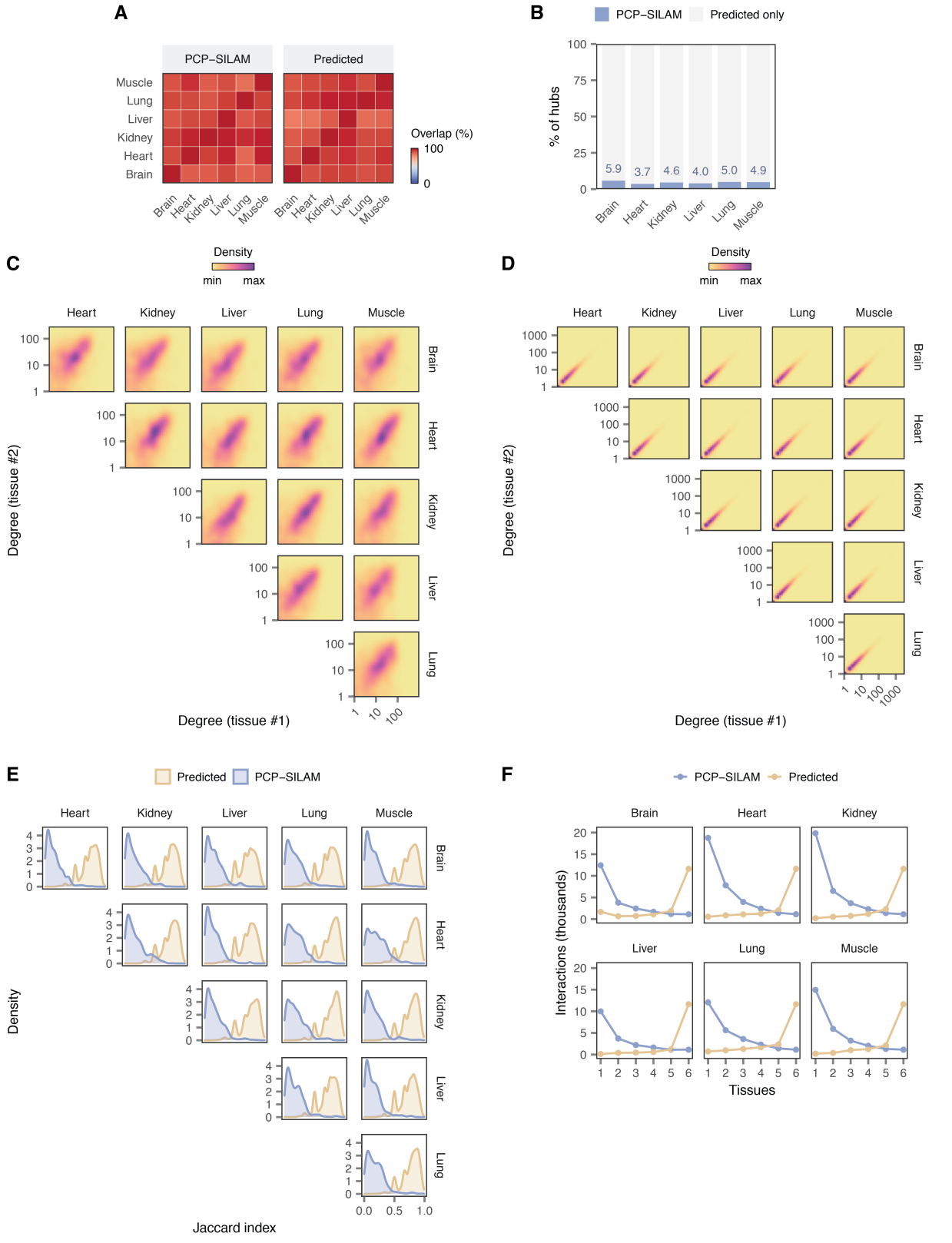

A
PCP–SILAM
Predicted
Muscle
Lung
Liver
Kidney
Heart
Brain
Overlap (%)
100
0
B
PCP–SILAM
Predicted only
% of hubs
5.9
3.7
4.6
4.0
5.0
4.9
C
Density
min
max
Degree (tissue #2)
Degree (tissue #1)
D
E
Predicted
PCP–SILAM
Density
Jaccard index
F
PCP–SILAM
Predicted
Interactions (thousands)
Tissues

Figure 3.8 Interactome rewiring limits accuracy of tissue interactome prediction. (A) Proportion of hub proteins in one tissue (x-axis) that are present in the interactome of the second tissue (y-axis). (B) Most PCP-SILAM tissue interactome hubs are not hubs in the corresponding predicted tissue interactomes. (C-D) Protein degrees across tissues in PCP-SILAM (C) and predicted (D) tissue interactomes. (E) Rewiring of protein interaction partners across each pair of tissues, as quantified by the Jaccard index, in predicted and PCP-SILAM tissue interactomes. (F) Tissue specificity of interactions across predicted and PCP-SILAM tissue interactomes.

Taken together, these analyses reveal widespread rewiring of interactome networks across mouse tissues beyond what is apparent from gene expression alone, affecting both the specific interactors of individual proteins and the global topological properties of physiological interactomes. Critically, the observation that a pair of proteins can interact in at least one context does not imply that their expression in a second context is a sufficient condition to reproduce the interaction. For example, PCP-SILAM correctly identified the known interaction between the F-actin-capping protein (CapZ) and the CapZ-interacting protein (CapZIP or Rcsd1) (Edwards et al., 2014; Eyers et al., 2005). However, the interaction was specific to heart and muscle, despite robust expression of the interacting proteins in all seven tissues (**Figure 3.7G**). This pattern of tissue specificity is consistent with the proposed function of the interaction in muscle contraction, but could not have been predicted on the basis of protein abundance alone. To evaluate whether predicted interactomes are depleted for tissue-specific interactions more systematically, we randomized interaction networks for each tissue separately, and calculated the total number of interactions found across one to seven randomized tissue interactomes. Consistent with the expectation that gene expression alone would underestimate the degree of interactome variability across tissues, predicted interactomes were significantly depleted for the most tissue-specific interactions (i.e., those found in only a single tissue), relative to PCP-

SILAM interactomes (**Figures 3.7F** and **3.8F**; $p < 10^{-15}$, Z test). Taken together, these results indicate that gene expression alone is insufficient to explain the observed rewiring of protein interactions across tissues, limiting the accuracy of existing methods to predict context-specific interactomes.

3.3.5 Evolution of interactions in mammalian tissues

Extrapolation of physical interactions detected in one organism to orthologous pairs of proteins in a different organism has been widely used to predict the interactomes of non-model organisms, or increase coverage of the human interactome (Li et al., 2017; Matthews et al., 2001; Yu et al., 2004). However, tissues may execute specialized biological functions that are less conserved between organisms than ubiquitous 'housekeeping' processes (Zhang and Li, 2004). We hypothesized that more tissue-specific interactions would show less evidence of evolutionary conservation than those occurring across many tissues. Examining literature-curated interactions for three model organisms, and five recent human high-throughput interactome maps, we found that tissue-specific interactions were less likely to be evolutionarily conserved (**Figures 3.9A** and **3.10A**; $p < 10^{-15}$ for model organisms and humans respectively, Fisher integration of Kendall rank correlations). Protein pairs that made tissue-specific interactions likewise showed less similar evolutionary rates (**Figure 3.9B**; $p < 10^{-15}$, Kendall rank correlation), and had less correlated patterns of presence and absence across eukaryotic genomes (**Figure 3.9C**; $p < 10^{-15}$), relative to universal interactions. Thus, both experimental screens and large-scale genomic evidence highlight the evolutionary novelty of tissue-specific interactions.

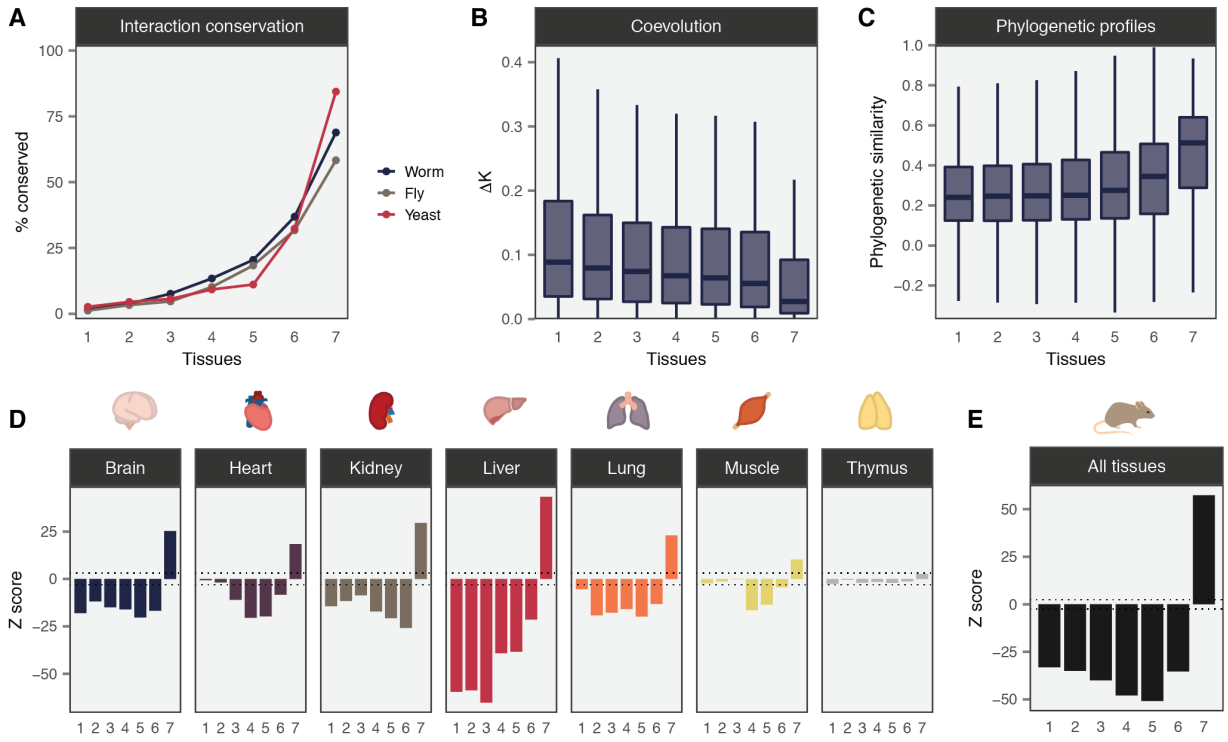


Figure 3.9 Evolution of mammalian tissue interactomes. (A) Proportion of mouse interactions conserved in worm, fly, and yeast for interactions detected in one to seven tissues. (B) Differences in evolutionary rates between interacting protein pairs detected in one to seven tissues. (C) Correlations in phylogenetic profiles between interacting protein pairs detected in one to seven tissues. (D-E) Statistical significance of interactions between housekeeping proteins and proteins quantified in one to seven tissues, relative to randomized networks in each mouse tissue interactome (D) and aggregated across mouse tissues (E).

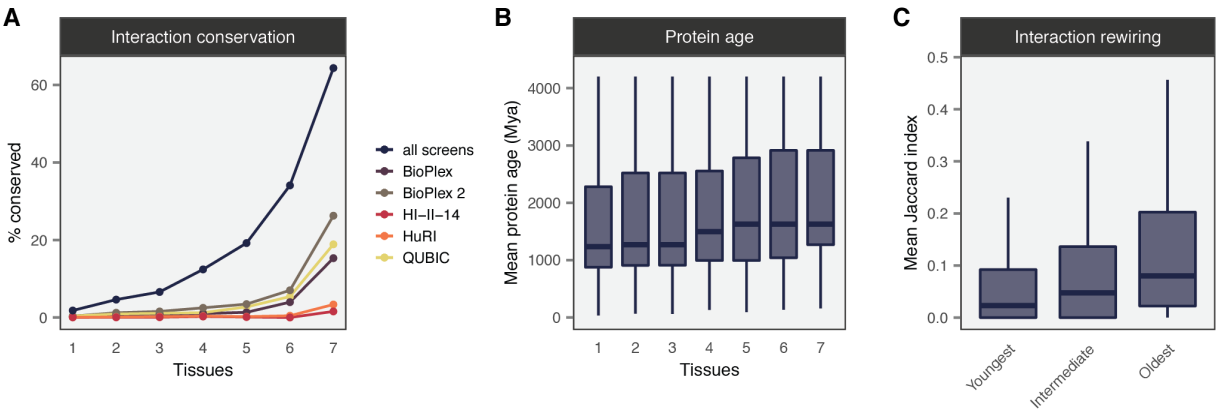


Figure 3.10 Evolution of interactions in mouse tissues. (A) Proportion of mouse interactions conserved in human, for interactions detected in between one and seven tissues. in five recent high-throughput human interactome screens. (B) Mean evolutionary age of interacting protein pairs, for interactions detected in between one and seven tissues. (C) Mean Jaccard index across all pairs of tissue interactomes for mouse proteins binned by evolutionary age.

We next asked whether tissue-specific interactions predominantly arise from tissue-specific rewiring of ancient proteins, or whether they instead involve evolutionarily young proteins. Relative to universal interactions, the average phylogenetic age of proteins involved in tissue-specific interactions was significantly younger (**Figure 3.10B**; $p < 10^{-15}$). Furthermore, ancient proteins had more conserved interaction partners across tissues, whereas younger proteins were disproportionately rewired between tissue interactomes (**Figure 3.10C**; $p < 10^{-15}$), suggesting that rewiring of ancient proteins is insufficient to explain the evolution of novel, tissue-specific interactions.

Previous analyses of predicted tissue interactomes identified extensive interactions between proteins expressed in only a subset of tissues and those expressed in all tissues (housekeeping proteins), proposing a model wherein tissue-specific functions arise by recruiting core cellular processes (Bossi and Lehner, 2009). Motivated by our observation that interactome rewiring is poorly predicted by tissue-specific gene expression, we investigated whether PCP-SILAM data supported a model of extensive cross-talk between housekeeping and tissue-specific proteins, or one in which tissue-specific functions are mediated independently of core cellular processes. To quantify the extent of cross-talk between housekeeping and tissue-specific proteins, we randomized each tissue interactome, using a degree-preserving method to control for network topology (Maslov and Sneppen, 2002), and compared the number of interactions

between proteins at each level of tissue specificity (that is, quantified in between one and seven tissues) to the number observed in randomly rewired networks. In every tissue, housekeeping proteins displayed highly significant enrichment for interactions with other housekeeping proteins, offering robust statistical support for the existence of core cellular modules within the interactome (**Figure 3.9D**). In contrast, we observed systematic depletion of interactions between tissue-specific and housekeeping proteins compared to randomized networks, particularly when aggregating results across all seven tissues (**Figure 3.9E**). Experimental tissue interactome mapping therefore reveals that evolutionarily novel tissue-specific interactions accomplish tissue-specific functions largely independent of the core modules of universal interactions.

Taken together, these analyses contrast two systems: core cellular modules present across all mouse tissues, and accessory modules that execute specialized functions within individual tissues. The former involves ancient proteins that have co-evolved over long evolutionary timeframes, and whose interaction partners are preserved across species and tissues. In contrast, tissue-specific interactions are less often conserved in other species and disproportionately involve younger proteins. Remarkably, we observe suppression of cross-talk between these two systems, with significant depletion of interactions between tissue-specific and universal proteins. Importantly, the opposite conclusion is reached when relying solely on tissue interactomes predicted based on gene expression, highlighting the importance of experimental interactome mapping to develop accurate systems-level biological models.

3.3.6 Tight regulation of tissue-specific interaction rewiring

Having established that evolutionarily ancient proteins have significantly more stable interaction partners across tissues than the interactome average, we sought to further characterize the

properties of proteins whose interactions are disproportionately rewired in a tissue-specific manner. To quantify the degree of interaction rewiring across tissues for each protein, we compared the similarity of its interaction partners across all pairs of tissues using the Jaccard index (**Figure 3.11A**). Proteins with a higher Jaccard index participate in interactions that are preserved across mouse tissues, whereas proteins with a low Jaccard index have interaction partners that are more rewired.

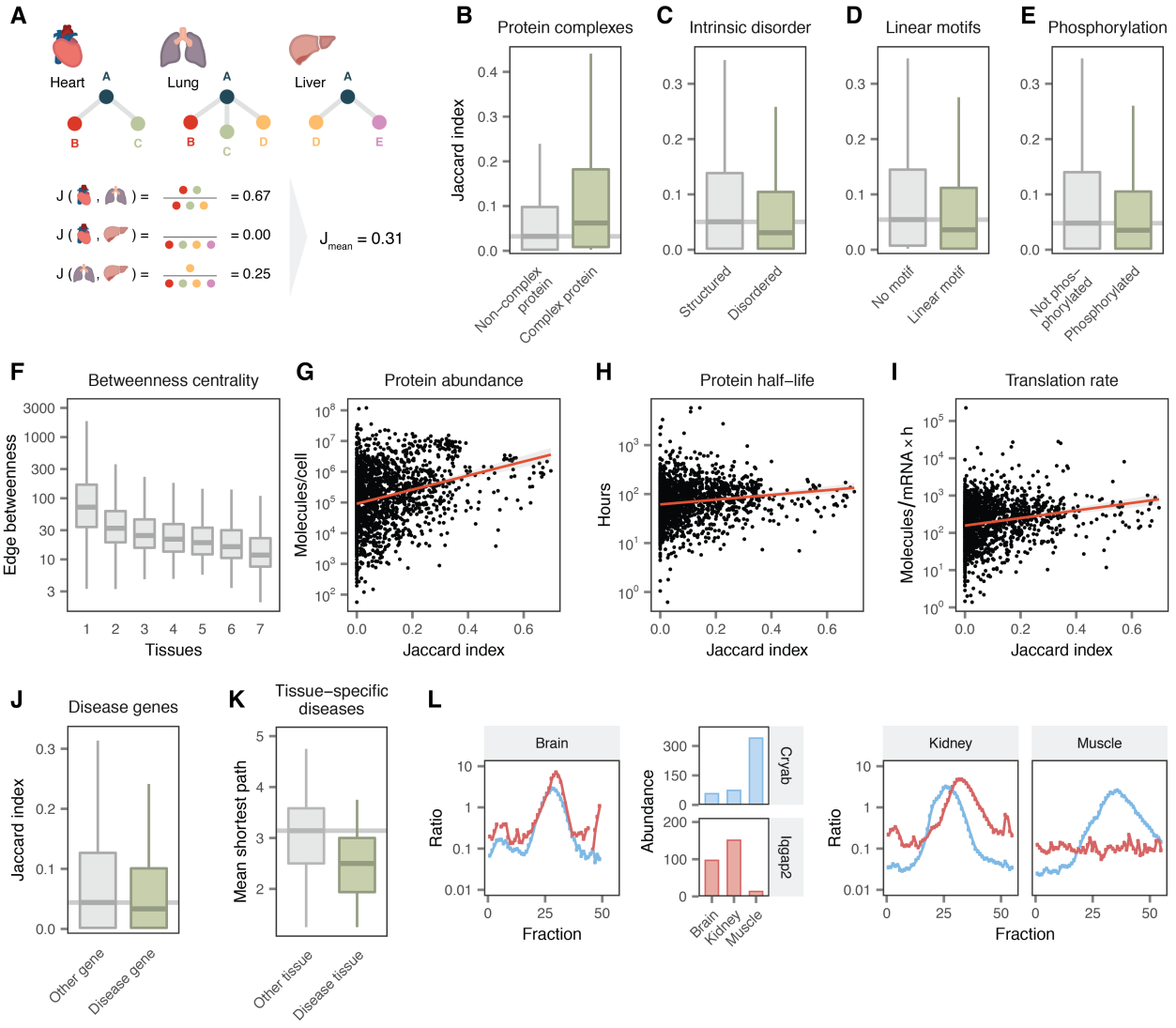


Figure 3.11 Tight regulation of interaction rewiring. (A) Schematic overview of the mean Jaccard index calculation. (B) Members of known protein complexes display a lesser degree of interaction rewiring across tissue interactomes. (C) Intrinsically disordered proteins display a greater degree of interaction rewiring across tissue interactomes. (D) Proteins containing protein-binding linear motifs display a greater degree of interaction rewiring across tissue interactomes. (E) Phosphoproteins display a greater degree of interaction rewiring across tissue interactomes. (F) Betweenness centrality of interactions detected in one to seven tissues. (G-I) Rewired proteins are characterized by low abundance (G), short half-lives (H), and slow translation rates (I). (J) Disease genes display a greater degree of interaction rewiring across tissue interactomes. (K) Disease genes associated with diseases that manifest in a specific tissue are more interconnected in the disease-associated tissue than in non-matched tissues. (L) PCP-SILAM detects a brain-specific interaction between Iqgap2 (red) and Cryab (blue), an intrinsically disordered protein and disease gene.

As expected, members of known protein complexes were significantly less rewired across tissues (**Figure 3.11B**; $p < 10^{-15}$, one-tailed Brunner-Munzel test). However, GO enrichment analysis failed to identify any functional categories overrepresented among rewired proteins. We therefore asked whether protein structural features instead would be predictive of rewiring. We hypothesized that intrinsically disordered proteins, which lack a unique structure, would possess the increased interaction surface area and conformational flexibility necessary to interact with multiple target proteins. Indeed, disordered proteins were significantly more rewired than their structured counterparts (**Figure 3.11C**; $p = 7.8 \times 10^{-7}$). Disordered protein segments often embed short peptide interaction motifs that can be bound by globular domains (Davey et al., 2012), and proteins containing such linear motifs were likewise significantly more rewired (**Figure 3.11D**; $p = 1.3 \times 10^{-6}$). Intrinsically disordered regions are also known to be hotspots for protein phosphorylation (Iakoucheva et al., 2004), raising the possibility that interaction rewiring may be coordinated by tissue-specific post-translational modification. Consistent with this possibility,

phosphoproteins were enriched among rewired proteins (**Figure 3.11E**; $p = 3.7 \times 10^{-4}$), and phosphosites on rewired proteins were significantly more tissue-specific than those found on universal proteins (**Figure 3.12A**; $p = 7.3 \times 10^{-5}$, Spearman rank correlation). Partial correlation analysis confirmed the enrichments for both linear motifs and phosphosites were independent of intrinsic disorder ($p = 1.2 \times 10^{-3}$ and 6.0×10^{-3}, respectively). Collectively, these findings suggest that binding motifs and post-translational modification sites embedded within intrinsically disordered regions facilitate the rewiring of protein interaction partners across mammalian tissues.

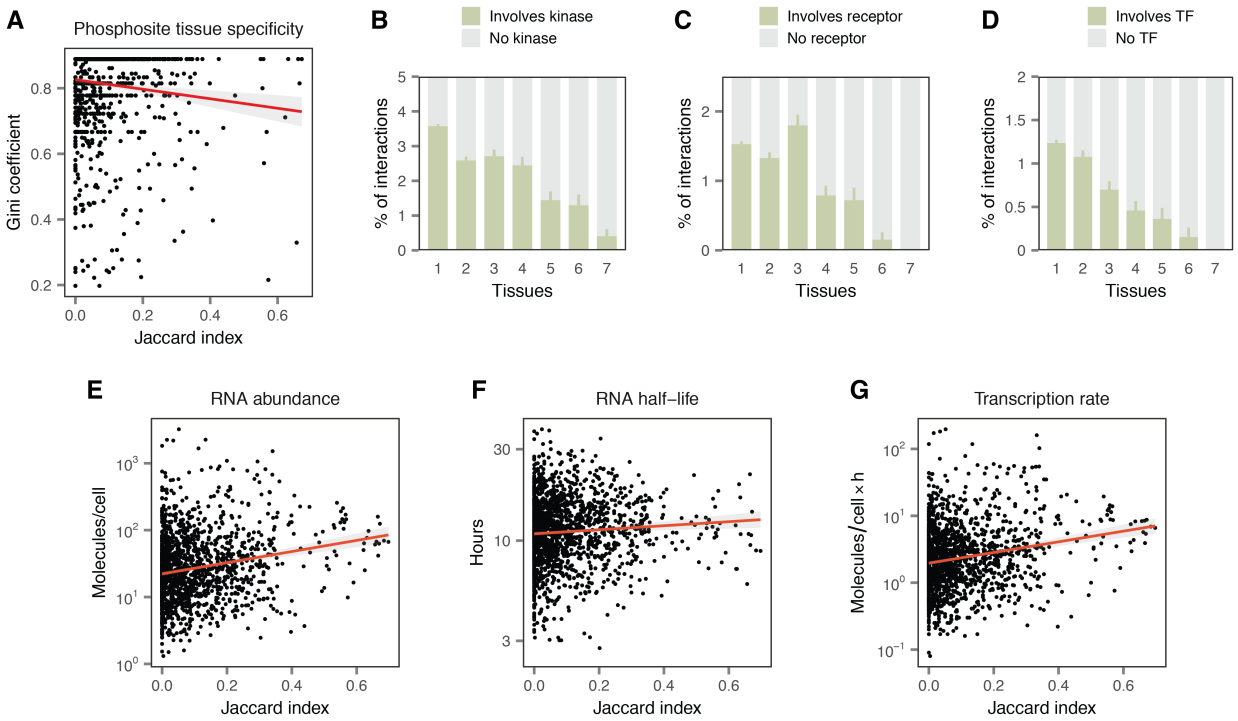


Figure 3.12 Tissue-specific interactions mediate tissue-specific biological information flow. (A) Phosphorylation sites on rewired proteins are more tissue-specific. (B) Proportion of interactions involving protein kinases at each level of tissue specificity. (C) Proportion of interactions involving transcription factors at each level of tissue specificity. (D) Proportion of interactions involving cell surface receptors at each level of tissue specificity.

(E-G) mRNAs encoding rewired proteins are characterized by low abundance (E), short half-lives (F), and slow transcription rates (G).

Disordered proteins are frequently involved in signaling pathways or mediate regulatory functions (Ward et al., 2004). We therefore hypothesized that rewiring of PPIs across tissues could facilitate tissue-specific signalling processes. Consistent with this hypothesis, tissue-specific interactions were significantly more likely to involve protein kinases, transcription factors, and cell surface protein receptors (**Figure 3.12B-D**; $p < 10^{-15}$, $p = 5.5 \times 10^{-13}$, and $p = 2.0 \times 10^{-5}$, respectively, Kendall rank correlation). In addition, we calculated the betweenness centrality of each interaction, defined as the number of shortest paths across the network that pass through each edge. Edges with a high betweenness centrality in interactome networks are associated with information flow across the network (Yan et al., 2016), and in agreement with this notion, we found tissue-specific interactions to have a significantly higher centrality than universal interactions (**Figure 3.11F**; $p < 10^{-15}$, Kendall rank correlation). Thus, both molecular and network topological perspectives highlight the key role of tissue-specific interactions in propagating biological information within tissue-specific pathways.

Within the cell, precise coordination of macromolecular interactions is required for accurate transmission of biological information. We therefore hypothesized that proteins whose interaction partners are highly variable across physiological contexts would be subject to tight regulatory mechanisms, and asked whether specific cellular strategies regulate the availability of rewired proteins. mRNAs encoding rewired proteins were expressed at lower levels, had shorter half-lives, and were transcribed at slower rates (**Figure 3.12E-G**; $p = 2.5 \times 10^{-14}$, 4.6×10^{-5}, and 1.7×10^{-13}, respectively, Spearman rank correlation) than proteins with more stable interaction

partners. Rewired proteins themselves were also less abundant (**Figure 3.11G**; $p < 10^{-15}$), and this difference in abundance was controlled both by a reduced rate of translation and increased rate of degradation (**Figure 3.11H-I**; both $p < 10^{-15}$), suggesting multiple cellular mechanisms converge to tightly regulate the abundance of proteins whose interacting partners are rewired across tissues. Disordered proteins are themselves known to be tightly regulated (Gsponer et al., 2008), but partial correlation analysis confirmed the tight regulation of rewired proteins was independent of protein disorder for all outcomes ($p \leq 5.0 \times 10^{-6}$, partial Spearman correlation).

Given this tight regulation of rewired proteins, we further asked whether proteins whose interaction partners are highly rewired between tissues were associated with deleterious phenotypes. Remarkably, we found that disease genes were significantly more rewired across tissues than the interactome average (**Figure 3.11J**; $p = 1.1 \times 10^{-2}$). Many human diseases are caused by germline mutations that are present in virtually every cell in the body, but which manifest in pathology in only one or a handful of tissues (Lage et al., 2008). We therefore drew on a resource of disease genes linked to tissue-specific pathologies (Basha et al., 2020) to ask whether the protein products of these genes are preferentially interconnected in the interactomes of the disease-associated tissues, as has long been hypothesized (Barshir et al., 2014; Hekselman and Yeger-Lotem, 2020; Kitsak et al., 2016; Magger et al., 2012). Indeed, the mean shortest paths between disease genes were significantly smaller in disease-associated tissues (**Figure 6.11K**; $p = 5.4 \times 10^{-4}$), indicating the formation of tissue-specific disease modules (Menche et al., 2015). The interaction between Cryab, a heat stock protein with an intrinsically disordered C-terminal segment implicated in a number of neurological disorders (Baldwin et al., 2012; Kuipers et al., 2017), and Iqgap2, a multifunctional signalling protein required for axon outgrowth (Wang et al., 2007), provides an example of disease gene rewiring between tissues.

PCP-SILAM detected the interaction in brain, but not kidney or muscle, despite robust expression of the interacting proteins in all three tissues (**Figure 3.11L**).

Taken together, these analyses highlight the role of protein-binding motifs and post-translational modification sites within disordered regions in mediating interaction rewiring across physiological contexts. The resulting tissue-specific interactions are associated with transmission of biological information in tissue-specific signalling pathways. Highly rewired proteins are subject to tight regulation by multiple convergent cellular mechanisms, perhaps to ensure the fidelity of biological information flow, and the elevated rate of interaction rewiring among disease genes implicates dysfunction of this regulatory cascade in disease pathophysiology. Notably, disease genes are preferentially interconnected in the interactomes of the tissues in which disease manifests, suggesting that mapping context-specific interactomes will be critical to elucidate the disease modules underlying pathobiology (Menche et al., 2015).

3.4 Discussion

Charting the complete protein-protein interactome is essential to revealing the molecular origins of cellular processes. However, earlier efforts produced static interactome maps that are fundamentally limited with respect to understanding interaction dynamics across tissue- or cell type-specific contexts. By applying PCP-SILAM to map the *in vivo* interactomes of seven mouse tissues, we provide a systematic, proteome-scale resource to understand the dynamic physiological interactome. Multiple functional genomics measures, including functional co-annotation, protein coexpression, and subcellular colocalization, indicate that our experimental and bioinformatic pipeline mapped tissue-specific interaction networks with an accuracy comparable to both small-scale experiments (Cusick et al., 2009) and the highest-quality high-

throughput screens (Hein et al., 2015; Huttlin et al., 2015, 2017; Luck et al., 2020; Rolland et al., 2014). We provide an interactive web application to facilitate exploration of the complete dataset, available at http://tissue-interactomes.msl.ubc.ca.

In the absence of high-throughput methods to define context-specific *in vivo* interactomes, computational methods have been developed to predict interactome rewiring based on gene expression (Bossi and Lehner, 2009; Buljan et al., 2012; de Lichtenberg et al., 2005). We find that widespread and physiological interactome rewiring limits the accuracy of tissue interactome predictions based on tissue-specific patterns of gene expression or coexpression. The effect of this rewiring extends beyond the mere identities of the interactors in each tissue to the global topological properties of tissue interactome networks. This finding reinforces conclusions from targeted studies, which have revealed marked dissimilarities in PPIs across cell lines (Floyd et al., 2016; Jäger et al., 2011), between cellular compartments (Markmiller et al., 2018), in response to cellular stimulation (Kerr et al., 2020; Kristensen et al., 2012), or in disease-relevant contexts (Pankow et al., 2015; Shirasaki et al., 2012). In yeast, widespread interaction rewiring has been observed in response to environmental perturbations (Celaj et al., 2017; Liu et al., 2020). Our systematic screen builds on these findings, revealing interactome rewiring at a much larger scale and across healthy mammalian tissues. As an example of an insight into interactome organization that is not apparent from gene expression-guided predictions of context-specific interactomes, we identify systematic suppression of cross-talk between the core housekeeping interactome and tissue-specific modules, whereas the opposite conclusion had been reached in analyses of predicted tissue interactomes (Bossi and Lehner, 2009).

Evolutionary analyses of mouse tissue interactomes contrast core cellular modules composed of evolutionarily ancient, housekeeping proteins that are connected via universal

interactions with evolutionarily recent proteins that interact in a more tissue-specific manner and are associated with cellular signalling. Intrinsically disordered proteins are particularly predisposed to interaction rewiring across tissues, consistent with the notion that these proteins can adopt new interacting partners over rapid evolutionary timescales (Hultqvist et al., 2017). Our findings linking proteins containing linear motifs or intrinsically disordered regions to an *in vivo* program of interactome rewiring substantiate previous bioinformatic or *in vitro* analyses suggesting that alternative splicing of disordered protein-coding exons can facilitate interactome remodeling (Buljan et al., 2012; Ellis et al., 2012; Romero et al., 2006). Proteins whose interactions are highly rewired across tissues are subject to tight cellular regulation, and are implicated in disease, suggesting dysfunction of this regulatory program is at the heart of many deleterious phenotypes. Intriguingly, topological analyses reveal the formation of tissue-specific disease modules for genes implicated in tissue-specific pathologies, potentially explaining the phenomenon whereby mutations present in every cell in the body cause dysfunction in only a subset of tissues.

The quantitative proteomic method presented in this study, PCP-SILAM, has the advantage of being an untargeted and relatively unbiased technique, apart from its moderate bias towards proteins of higher cellular abundance. As a high-throughput technique for *in vivo* interactome mapping, PCP-SILAM is uniquely suited to the simultaneous discovery of novel interactions and quantification of their dynamics across tissues. Interrogation of *in vivo* interactomes with PCP-SILAM maps new regions and functional classes within the mouse interactome, and places poorly-studied mouse proteins into tissue-specific functional contexts, suggesting PCP SILAM will be a valuable method to shed light on poorly understood components of the proteome.

The biological picture that emerges from our systematic map of seven mouse tissue interactomes is one of widespread interactome rewiring, and this rewiring may be a crucial mediator of cellular and organismal phenotype. The extent of interactome rewiring observed across healthy mammalian tissues in this study indicates that a complete understanding of the human interactome will require experimental definition of the context-specific interactome networks across cell types and tissues, and their dynamic changes in response to cellular stimulation, differentiation, and disease. Our study serves as a first step towards this goal and provides a foundation for building a systems-level understanding of the mechanistic roles interactome rewiring plays in health and disease.

Chapter 4: Conclusion

Mapping the complete network of biologically relevant protein-protein interactions has been a central goal of biomedical research since the turn of the 21st century. Over the past 20 years, classical experimental techniques such as Y2H and AP-MS have produced extensive maps of the interactome in humans and other organisms. However, the shortcomings of these approaches have prompted the development of a new generation of methods for experimental interactome mapping.

The work described in this thesis has focused on one of these second-generation methods: protein correlation profiling. From a methodological perspective, PCP has a unique set of advantages that combine the ability to monitor protein-protein interactions under native conditions at the proteome scale with the experimental resolution required to perform *de novo* interactome mapping. These advantages differentiate PCP both from classical approaches such as Y2H and AP-MS, which are labor-intensive and deliver interactome maps of questionable physiological relevance, with emerging approaches such as thermal proteome profiling (Becher et al., 2018; Dai et al., 2018; Tan et al., 2018) that enable interrogation of known interactions under physiological conditions, but not the identification of novel interactions. Accordingly, the methodological advantages of PCP position this method to deliver new biological insights about the physiological interactome as it occurs *in vivo* in the cells and tissues of the human body, rather than in immortalized cell lines or the yeast nucleus. However, this potential brings with it new analytical challenges. In this thesis, I studied the computational and statistical methods that are currently used to infer-protein interaction networks from PCP data, then applied these

methods to derive *in vivo* interactomes for seven mouse tissues. Below, I briefly review some of the key conclusions of the research presented in this thesis.

In Chapter 2, I studied the computational methods that are used to convert a raw PCP dataset, in which the abundance of each protein is quantified across an ordered series of fractions, into a network of unweighted interactions. One of the major philosophical divides in published approaches to this task is between methods that seek to infer the network exclusively from the PCP experiments themselves, and methods that incorporate external compendia of genomic data to support or discredit individual interactions. I sought to deeply understand the implications of this choice. I assembled large collections of both published PCP data, and external genomic datasets that are representative of those that have been used in the field to date. I inferred hundreds of protein-protein interaction networks, systematically varying the amount and types of input genomic datasets provided to a classifier alongside the PCP data. Superficially, incorporating external genomic datasets appeared to improve the quality of networks. I found that proteins known to have shared cellular functions were more interconnected in the resulting networks in these networks than in networks inferred from PCP data alone, long thought to be a hallmark of network quality (von Mering et al., 2002). However, this functional connectivity came at the expense of a reduced ability to discover novel interactions. I reasoned this trend could be explained if the putatively 'novel' interactions recovered without external genomic data were enriched for false positives. However, a time-split experiment showed that novel interactions recovered with or without external genomic data were equally likely to be discovered in subsequent years.

This analysis has significant implications for the analysis of PCP data. In general, the primary goal of interactome mapping projects is to discover new biology, although some have

advocated that PCP should instead be used primarily to monitor the assembly of known protein complexes (Bludau et al., 2020; Heusel et al., 2019). Collectively, the results of this analysis suggest that incorporating external genomic datasets into the machine-learning procedure used to infer networks from PCP data will unnecessarily hinder the discovery of new interactions, and that many of these interactions may connect proteins that were not previously known to have a functional link. Conversely, these findings suggest that PCP yields sufficient information to map interactomes from experimental data alone, as is typically done from Y2H or AP-MS data.

With this understanding in hand, I applied the supervised machine-learning methods that were the focus of Chapter 2 to a newly collected PCP dataset in Chapter 3. The dataset in question stemmed from the adaptation of PCP for the *in vivo* setting, an innovation that provided a basis for systematic interactome mapping across mammalian tissues for the first time. The reproducibility of this adapted technique was high and quantitatively comparable to published *in vitro* data. The application of a supervised machine-learning approach to the PCP data led to the reconstruction of seven tissue interactome networks, comprising over 125,000 unique interactions—more than doubling the size of the known mouse interactome.

I then set out to use this interactome resource to ask several questions about the physiological organization of the interactome that had previously been inaccessible. I demonstrated that tissue-specific interactions are particularly likely to be novel, and showed that these also disproportionately involve proteins of unknown function, or for which no interactions were previously known. This tissue interactome resource thus places many poorly understood proteins into a functional context. Using this resource to benchmark methods for tissue-specific interactome prediction, I systematically demonstrated the deficiencies of these approaches, showing that changes in protein abundance are insufficient to predict in vivo interactome

rewiring. This resource also allowed me to explore modes of evolution in mouse tissue interactomes, which contrast an evolutionarily ancient 'core' of the interactome present in all tissues to evolutionarily recent, 'accessory' modules present in individual tissues; moreover, I identified systematic suppression of cross-talk between these modules. I characterized the properties of proteins whose interaction partners are disproportionately rewired across tissues, identifying structural features that facilitate rewiring, and finding that these rewired proteins are subject to multiple convergent programs of tight cellular regulation. Finally, and perhaps most notably, this resource allowed me to show empirically that genes implicated in diseases that selectively impact specific tissues are more tightly interconnected in the interactome networks of those tissues. This confirms a hypothesis that has been discussed for more than a decade. Collectively, these findings underscore the power of PCP, in combination with machine-learning approaches, to probe the native physiological organization of mammalian protein-protein interaction networks.

A broad limitation affecting the work described throughout this thesis is the sensitivity of mass spectrometry. Whereas methods for profiling gene expression based on microarrays or RNA sequencing are generally able to provide quantitative estimates of abundance for essentially all of the ~20,000 genes in the human genome, typical mass spectrometry-based proteomics workflows report on at most half of this number (Aebersold and Mann, 2016). In PCP, the total number of proteins quantified is generally lower still, partly because not all cellular proteins participate in macromolecular complexes that are amenable to analysis by the various approaches to separate cellular lysates that are currently in use. Moreover, specialized workflows are required in order to monitor the presence of post-translational modifications, and the use of peptide-centric, 'bottom-up' approaches precludes the identification of complete protein

isoforms (Olsen and Mann, 2013; Tran et al., 2011). The scope of the biological questions that can be addressed using PCP, though broad, is finite. Many of the same limitations also bear on the application of AP-MS, emphasizing the need to integrate multiple complementary approaches to interactome mapping in order to develop a comprehensive view of cellular networks. More broadly, a great deal of information flow through cellular systems takes place through other types of macromolecular interactions, such as protein-DNA or protein-metabolite interactions (Marbach et al., 2016; Piazza et al., 2018). Developing approaches to systematically map these networks, and integrate them with protein-protein interaction networks, will be necessary to chart the complete cellular 'wiring diagram.'

To conclude this thesis, I highlight several of the unresolved challenges facing the field and opportunities for future work. One of the major limitations of all the work described herein is that the supervised machine learning paradigm that has come to dominate the field seeks to distinguish interacting from non-interacting protein pairs. However, PCP as an assay is not designed primarily to identify heteromeric interacting pairs. Instead, PCP should theoretically have the greatest power to resolve large and complex cellular assemblies. Consequently, a disconnect exists between the output of supervised machine-learning methods that predict binary interactions, and protein complexes as they occur in the cell. This disconnect has led many practitioners to apply graph-based clustering algorithms to the inferred interaction networks, such as Markov clustering (Enright et al., 2002) or ClusterONE (Nepusz et al., 2012). However, these methods are exquisitely sensitive to small perturbations in the network (Stacey et al., 2020). In extreme bases, randomly rewiring just 1% of the interactions in a protein-protein interaction network can produce a ~50% change in the complexes detected from that network. The same instability is observed when perturbing the raw PCP data itself, rather than the inferred

network. Because both PCP data collection and network inference are inherently noisy procedures, this observation calls into question the use of graph-based clustering to resolve protein complexes. Yet, on the other hand, there are currently few alternative approaches available to recover protein complexes from PCP data (McBride et al., 2019). The development of more robust and accurate approaches to protein complex inference from PCP data therefore represents a major challenge for the field going forward.

A second challenge in the analysis of PCP data concerns the sets of protein pairs that are used to define true positive and true negative interactions. These true positives and true negatives are used for several purposes in data analysis, the most important of which are to train the machine-learning classifier, and to assess the quality of the resulting network. With respect to the latter point, it would be particularly desirable to estimate the false-negative and false-positive rates of the inferred interaction network, in order to gauge its completeness and accuracy. In Y2H screens, these rates can be approximated by screening known positive and negative interactions (Venkatesan et al., 2009). However, an assessment of the false-negative rate in PCP data is complicated by the fact that the total set of interactions amenable to recovery by a given PCP screen is not known. Assessing the false-positive rate is in theory more tractable, but relies on the availability of a representative set of non-interacting protein pairs (Ben-Hur and Noble, 2006). In practice, most approaches use pairs of proteins from the CORUM database (Giurgiu et al., 2019) that are not known to participate in the same complex for this purpose. However, applying this definition of true negative interactions to estimate error rates in published AP-MS and Y2H networks results in the conclusion that at least 50% of these interactions are false-positives. This exceptionally high error rate raises the possibility that a significant minority of these supposedly non-interacting protein pairs are in fact *bona fide* interactors, which may be as-

of-yet undiscovered or simply unannotated in the CORUM database. Previous efforts have re-analyzed published PCP data to refine the set of true positives used as input to the classifier by removing protein complexes that are incompatible with the separation approaches used in PCP (Stacey et al., 2018). Analogous efforts to develop a PCP-specific set of true negative interactions could be similarly useful. In Chapter 3, I used several orthogonal sources of information, including protein co-expression and subcellular co-localization, to estimate the quality of the inferred networks and compare them to published screens. This approach highlights another shortcoming of the methods for genomic data integration discussed in Chapter 2: namely, once these sources of information have been used to train the classifier, they can no longer be used to evaluate the inferred network.

Collectively, the work presented in this thesis has developed new approaches to chart protein-protein interaction networks in their native, physiological context. Looking towards the future, the integration of PCP with other mass spectrometric techniques that can probe protein-protein interactions *in vivo*, such as thermal proteome profiling (Mateus et al., 2020) or cross-linking mass spectrometry (O'Reilly and Rappsilber, 2018), could further increase the coverage and accuracy of *in vivo* interactome mapping.

References

Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature *537*, 347–355.

Alanis-Lobato, G., Andrade-Navarro, M.A., and Schaefer, M.H. (2017). HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. Nucleic Acids Res. *45*, D408–D414.

Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. *33*, D418-24.

Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. Nature *426*, 570–574.

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an Arabidopsis interactome map. Science *333*, 601–607.

Azimifar, S.B., Nagaraj, N., Cox, J., and Mann, M. (2014). Cell-type-resolved quantitative proteomics of murine liver. Cell Metab. *20*, 1076–1087.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D.M., Burston, H.E., Vizeacoumar, F.J., Snider, J., Phanse, S., et al. (2012). Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature *489*, 585–589.

Baldwin, A.J., Walsh, P., Hansen, D.F., Hilton, G.R., Benesch, J.L.P., Sharpe, S., and Kay, L.E. (2012). Probing dynamic conformations of the high-molecular-weight αB-crystallin heat shock protein ensemble by NMR spectroscopy. J. Am. Chem. Soc. *134*, 15343–15350.

Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics *31*, 2123–2130.

Ballouz, S., Weber, M., Pavlidis, P., and Gillis, J. (2017). EGAD: ultra-fast functional analysis of gene networks. Bioinformatics *33*, 612–614.

Barabási, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. *5*, 101–113.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Barshir, R., Shwartz, O., Smoly, I.Y., and Yeger-Lotem, E. (2014). Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. PLoS Comput. Biol. *10*, e1003632.

Bartel, P.L., Roecklein, J.A., SenGupta, D., and Fields, S. (1996). A protein linkage map of Escherichia coli bacteriophage T7. Nat. Genet. *12*, 72–77.

Basha, O., Argov, C.M., Artzy, R., Zoabi, Y., Hekselman, I., Alfandari, L., Chalifa-Caspi, V., and Yeger-Lotem, E. (2020). Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes. Bioinformatics *36*, 2821–2828.

Bastian, F.B., Roux, J., Niknejad, A., Comte, A., Fonseca Costa, S.S., Mendes de Farias, T., Moretti, S., Parmentier, G., Rech de Laval, V., Rosikiewicz, M., et al. (2020). The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. BioRxiv.

Batada, N.N., Hurst, L.D., and Tyers, M. (2006). Evolutionary and physiological importance of hub proteins. PLoS Comput. Biol. *2*, e88.

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015).

Genomic variation. Impact of regulatory variation from RNA to protein. Science *347*, 664–667.

Becher, I., Andrés-Pons, A., Romanov, N., Stein, F., Schramm, M., Baudin, F., Helm, D., Kurzawa, N., Mateus, A., Mackmull, M.-T., et al. (2018). Pervasive Protein Thermal Stability Variation during the Cell Cycle. Cell *173*, 1495–1507.e18.

Ben-Hur, A., and Noble, W.S. (2006). Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics *7 Suppl 1*, S2.

Bludau, I., Heusel, M., Frank, M., Rosenberger, G., Hafen, R., Banaei-Esfahani, A., van Drogen, A., Collins, B.C., Gstaiger, M., and Aebersold, R. (2020). Complex-centric proteome profiling by SEC-SWATH-MS for the parallel detection of hundreds of protein complexes. Nat. Protoc. *15*, 2341–2386.

Bossi, A., and Lehner, B. (2009). Tissue specificity and the human protein interaction network. Mol. Syst. Biol. *5*, 260.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell *46*, 871–883.

Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. Nat. Methods *10*, 690–691.

Capra, J.A., Williams, A.G., and Pollard, K.S. (2012). ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. PLoS Comput. Biol. *8*, e1002567.

Carlyle, B.C., Kitchen, R.R., Kanyo, J.E., Voss, E.Z., Pletikos, M., Sousa, A.M.M., Lam, T.T., Gerstein, M.B., Sestan, N., and Nairn, A.C. (2017). A multiregional proteomic survey of the postnatal human brain. Nat. Neurosci. *20*, 1787–1795.

Celaj, A., Schlecht, U., Smith, J.D., Xu, W., Suresh, S., Miranda, M., Aparicio, A.M., Proctor,

M., Davis, R.W., Roth, F.P., et al. (2017). Quantitative analysis of protein interaction network dynamics in yeast. Mol. Syst. Biol. *13*, 934.

Chaillou, T., Zhang, X., and McCarthy, J.J. (2016). Expression of Muscle-Specific Ribosomal Protein L3-Like Impairs Myotube Growth. J. Cell Physiol. *231*, 1894–1902.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N.K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., et al. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Res. *45*, D369–D379.

Chen, B., Sheridan, R.P., Hornak, V., and Voigt, J.H. (2012). Comparison of random forest and Pipeline Pilot Naïve Bayes in prospective QSAR predictions. J. Chem. Inf. Model. *52*, 792–803.

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

Cowley, M.J., Pinese, M., Kassahn, K.S., Waddell, N., Pearson, J.V., Grimmond, S.M., Biankin, A.V., Hautaniemi, S., and Wu, J. (2012). PINA v2.0: mining interactome modules. Nucleic Acids Res. *40*, D862-5.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Crozier, T.W.M., Tinti, M., Larance, M., Lamond, A.I., and Ferguson, M.A.J. (2017). Prediction of Protein Complexes in Trypanosoma brucei by Protein Correlation Profiling Mass Spectrometry and Machine Learning. Mol. Cell Proteomics *16*, 2254–2267.

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Systems *1695*, 1–9.

Cunningham, J.M., Koytiger, G., Sorger, P.K., and AlQuraishi, M. (2020). Biophysical prediction of protein-peptide interactions and signaling networks using machine learning. Nat. Methods *17*, 175–183.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., et al. (2009). Literature-curated protein interaction datasets. Nat. Methods *6*, 39–46.

Dai, L., Zhao, T., Bisteau, X., Sun, W., Prabhu, N., Lim, Y.T., Sobota, R.M., Kaldis, P., and Nordlund, P. (2018). Modulation of Protein-Interaction States through the Cell Cycle. Cell *173*, 1481–1494.e13.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. *33*, e175.

Das, J., and Yu, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst. Biol. *6*, 92.

Davey, N.E., Van Roey, K., Weatheritt, R.J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., and Gibson, T.J. (2012). Attributes of short linear motifs. Mol. Biosyst. *8*, 268–281.

Deshmukh, A.S., Murgia, M., Nagaraj, N., Treebak, J.T., Cox, J., and Mann, M. (2015). Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. Mol. Cell Proteomics *14*, 841–853.

Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R., et al. (2012). The UniProt-GO Annotation database in 2011. Nucleic Acids Res. *40*, D565-70.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Drew, K., Lee, C., Huizar, R.L., Tu, F., Borgeson, B., McWhite, C.D., Ma, Y., Wallingford, J.B., and Marcotte, E.M. (2017). Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Mol. Syst. Biol. *13*, 932.

Edwards, M., Zwolak, A., Schafer, D.A., Sept, D., Dominguez, R., and Cooper, J.A. (2014). Capping protein regulators fine-tune actin assembly dynamics. Nat. Rev. Mol. Cell Biol. *15*, 677–689.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol. Cell *46*, 884–892.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol. *3*, 89.

Eyers, C.E., McNeill, H., Knebel, A., Morrice, N., Arthur, S.J.C., Cuenda, A., and Cohen, P. (2005). The phosphorylation of CapZ-interacting protein (CapZIP) by stress-activated protein kinases triggers its dissociation from CapZ. Biochem. J. *389*, 127–135.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. (2016). The Reactome pathway Knowledgebase. Nucleic Acids Res. *44*, D481-7.

Floyd, B.J., Wilkerson, E.M., Veling, M.T., Minogue, C.E., Xia, C., Beebe, E.T., Wrobel, R.L., Cho, H., Kremer, L.S., Alston, C.L., et al. (2016). Mitochondrial protein interaction mapping identifies regulators of respiratory chain function. Mol. Cell *63*, 621–632.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005). Protein interaction mapping: a Drosophila case study. Genome Res. *15*, 376–384.

Fortelny, N., Butler, G.S., Overall, C.M., and Pavlidis, P. (2017). Protease-Inhibitor Interaction Predictions: Lessons on the Complexity of Protein-Protein Interactions. Mol. Cell Proteomics *16*, 1038–1051.

Foster, L.J., de Hoog, C.L., Zhang, Y., Zhang, Y., Xie, X., Mootha, V.K., and Mann, M. (2006). A mammalian organelle map by protein correlation profiling. Cell *125*, 187–199.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. Science *296*, 750–752.

Fromont-Racine, M., Rain, J.C., and Legrain, P. (1997). Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat. Genet. *16*, 277–282.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Gazestani, V.H., Nikpour, N., Mehta, V., Najafabadi, H.S., Moshiri, H., Jardim, A., and Salavati, R. (2016). A Protein Complex Map of Trypanosoma brucei. PLoS Negl. Trop. Dis. *10*,

e0004533.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell Proteomics *11*, M111.014050.

Geiger, T., Velic, A., Macek, B., Lundberg, E., Kampf, C., Nagaraj, N., Uhlen, M., Cox, J., and Mann, M. (2013). Initial quantitative proteomic map of 28 mouse tissues using the SILAC mouse. Mol. Cell Proteomics *12*, 1709–1722.

Geladaki, A., Kočevar Britovšek, N., Breckels, L.M., Smith, T.S., Vennard, O.L., Mulvey, C.M., Crook, O.M., Gatto, L., and Lilley, K.S. (2019). Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. Nat. Commun. *10*, 331.

Gillis, J., Ballouz, S., and Pavlidis, P. (2014). Bias tradeoffs in the creation and analysis of protein-protein interaction networks. J. Proteomics *100*, 44–54.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727–1736.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. *47*, D559–D563.

Gorka, M., Swart, C., Siemiatkowska, B., Martínez-Jaime, S., Skirycz, A., Streb, S., and Graf, A. (2019). Protein Complex Identification and quantitative complexome by CN-PAGE. Sci. Rep. *9*, 11523.

Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular

function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576.

Grossmann, A., Benlasfer, N., Birth, P., Hegele, A., Wachsmuth, F., Apelt, L., and Stelzl, U. (2015). Phospho-tyrosine dependent protein-protein interaction network. Mol. Syst. Biol. *11*, 794.

Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365–1368.

Guo, T., Kouvonen, P., Koh, C.C., Gillet, L.C., Wolski, W.E., Röst, H.L., Rosenberger, G., Collins, B.C., Blum, L.C., Gillessen, S., et al. (2015). Rapid mass spectrometric conversion of tissue biopsy samples into permanent quantitative digital proteome maps. Nat. Med. *21*, 407–413.

Guruharsha, K.G., Rual, J.-F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al. (2011). A protein complex network of Drosophila melanogaster. Cell *147*, 690–703.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., et al. (2012). A census of human soluble protein complexes. Cell *150*, 1068–1081.

van Heesch, S., Witte, F., Schneider-Lunitz, V., Schulz, J.F., Adami, E., Faber, A.B., Kirchner, M., Maatz, H., Blachut, S., Sandmann, C.-L., et al. (2019). The translational landscape of the human heart. Cell *178*, 242–260.e29.

Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712–723.

Hekselman, I., and Yeger-Lotem, E. (2020). Mechanisms of tissue and cell-type specificity in

heritable traits and diseases. Nat. Rev. Genet. *21*, 137–150.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., et al. (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. Nat. Biotechnol. *22*, 177–183.

Heusel, M., Bludau, I., Rosenberger, G., Hafen, R., Frank, M., Banaei-Esfahani, A., van Drogen, A., Collins, B.C., Gstaiger, M., and Aebersold, R. (2019). Complex-centric proteome profiling by SEC-SWATH-MS. Mol. Syst. Biol. *15*, e8438.

Heusel, M., Frank, M., Köhler, M., Amon, S., Frommelt, F., Rosenberger, G., Bludau, I., Aulakh, S., Linder, M.I., Liu, Y., et al. (2020). A Global Screen for Assembly State Changes of the Mitotic Proteome by SEC-SWATH-MS. Cell Syst. *10*, 133–155.e6.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

Hubbell, E., Liu, W.-M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585–1592.

Hultqvist, G., Åberg, E., Camilloni, C., Sundell, G.N., Andersson, E., Dogan, J., Chi, C.N., Vendruscolo, M., and Jemth, P. (2017). Emergence and evolution of an interaction between intrinsically disordered proteins. Elife *6*.

Hurvich, C.M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. Biometrika *76*, 297–307.

Huttlin, E.L., Jedrychowski, M.P., Elias, J.E., Goswami, T., Rad, R., Beausoleil, S.A., Villén, J., Haas, W., Sowa, M.E., and Gygi, S.P. (2010). A tissue-specific atlas of mouse protein

phosphorylation and expression. Cell *143*, 1174–1189.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The bioplex network: A systematic exploration of the human interactome. Cell *162*, 425–440.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505–509.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., and Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. *32*, 1037–1049.

Ideker, T., and Krogan, N.J. (2012). Differential network biology. Mol. Syst. Biol. *8*, 565.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA *97*, 1143–1147.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Jafari, R., Almqvist, H., Axelsson, H., Ignatushchenko, M., Lundbäck, T., Nordlund, P., and Martinez Molina, D. (2014). The cellular thermal shift assay for evaluating drug target interactions in cells. Nat. Protoc. *9*, 2100–2122.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., et al. (2011). Global landscape of HIV-human protein

complexes. Nature *481*, 365–370.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science *302*, 449–453.

Jarzab, A., Kurzawa, N., Hopf, T., Moerch, M., Zecha, J., Leijten, N., Bian, Y., Musiol, E., Maschberger, M., Stoehr, G., et al. (2020). Meltome atlas-thermal proteome stability across the tree of life. Nat. Methods *17*, 495–503.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127.

Jones, A.M., Xuan, Y., Xu, M., Wang, R.-S., Ho, C.-H., Lalonde, S., You, C.H., Sardi, M.I., Parsa, S.A., Smith-Valle, E., et al. (2014). Border control--a membrane-linked interactome of Arabidopsis. Science *344*, 711–716.

Kandasamy, K., Mohan, S.S., Raju, R., Keerthikumar, S., Kumar, G.S.S., Venugopal, A.K., Telikicherla, D., Navarro, J.D., Mathivanan, S., Pecquet, C., et al. (2010). NetPath: a public resource of curated signal transduction pathways. Genome Biol. *11*, R3.

Kastritis, P.L., O'Reilly, F.J., Bock, T., Li, Y., Rogon, M.Z., Buczak, K., Romanov, N., Betts, M.J., Bui, K.H., Hagen, W.J., et al. (2017). Capturing protein communities by structural proteomics in a thermophilic eukaryote. Mol. Syst. Biol. *13*, 936.

Kerppola, T.K. (2008). Bimolecular fluorescence complementation (BiFC) analysis as a probe of protein interactions in living cells. Annu. Rev. Biophys. *37*, 465–487.

Kerr, C.H., Skinnider, M.A., Andrews, D.D.T., Madero, A.M., Chan, Q.W.T., Stacey, R.G.,

Stoynov, N., Jan, E., and Foster, L.J. (2020). Dynamic rewiring of the human interactome by interferon signaling. Genome Biol. *21*, 140.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res. *37*, D767-72.

Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. Chem. Rev. *116*, 4884–4909.

Khan, Z., Ford, M.J., Cusanovich, D.A., Mitrano, A., Pritchard, J.K., and Gilad, Y. (2013). Primate transcript and protein expression levels evolve under compensatory selection pressures. Science *342*, 1100–1104.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kim, Y., Jung, J.P., Pack, C.-G., and Huh, W.-K. (2019). Global analysis of protein homomerization in Saccharomyces cerevisiae. Genome Res. *29*, 135–145.

Kirkwood, K.J., Ahmad, Y., Larance, M., and Lamond, A.I. (2013). Characterization of native protein complexes and protein isoform variation using size-fractionation-based quantitative proteomics. Mol. Cell Proteomics *12*, 3851–3873.

Kitsak, M., Sharma, A., Menche, J., Guney, E., Ghiassian, S.D., Loscalzo, J., and Barabási, A.-L. (2016). Tissue specificity of human disease module. Sci. Rep. *6*, 35241.

Kotlyar, M., Pastrello, C., Pivetta, F., Lo Sardo, A., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., et al. (2015). In silico prediction of physical protein interactions and characterization of interactome orphans. Nat. Methods *12*, 79–84.

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res. *44*, D536-41.

Kristensen, A.R., and Foster, L.J. (2013). High throughput strategies for probing the different organizational levels of protein interaction networks. Mol. Biosyst. *9*, 2201–2212.

Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods *9*, 907–909.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Krüger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C.A., Forner, F., Schmidt, S., Zanivan, S., Fässler, R., and Mann, M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. Cell *134*, 353–364.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Kuipers, H.F., Yoon, J., van Horssen, J., Han, M.H., Bollyky, P.L., Palmer, T.D., and Steinman, L. (2017). Phosphorylation of αB-crystallin supports reactive astrogliosis in demyelination. Proc. Natl. Acad. Sci. USA *114*, E1745–E1754.

Kustatscher, G., Grabowski, P., and Rappsilber, J. (2017). Pervasive coexpression of spatially proximal genes is buffered at the protein level. Mol. Syst. Biol. *13*, 937.

Kustatscher, G., Grabowski, P., Schrader, T.A., Passmore, J.B., Schrader, M., and Rappsilber, J. (2019). Co-regulation map of the human proteome enables identification of protein functions.

Nat. Biotechnol. *37*, 1361–1371.

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E.L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S.R., Miller, R., et al. (2016). WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res. *44*, D488-94.

Lage, K., Hansen, N.T., Karlberg, E.O., Eklund, A.C., Roque, F.S., Donahoe, P.K., Szallasi, Z., Jensen, T.S., and Brunak, S. (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc. Natl. Acad. Sci. USA *105*, 20870–20875.

Lapek, J.D., Greninger, P., Morris, R., Amzallag, A., Pruteanu-Malinici, I., Benes, C.H., and Haas, W. (2017). Detection of dysregulated protein-association networks by high-throughput proteomics predicts cancer vulnerabilities. Nat. Biotechnol. *35*, 983–989.

Larance, M., Kirkwood, K.J., Tinti, M., Brenes Murillo, A., Ferguson, M.A.J., and Lamond, A.I. (2016). Global Membrane Protein Interactome Analysis using In vivo Crosslinking and Mass Spectrometry-based Protein Correlation Profiling. Mol. Cell Proteomics *15*, 2476–2490.

Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2015). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res. *43*, D321-7.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D.J., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan C. elegans. Science *303*, 540–543.

Li, T., Wernersson, R., Hansen, R.B., Horn, H., Mercer, J., Slodkowicz, G., Workman, C.T., Rigina, O., Rapacki, K., Stærfeldt, H.H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods *14*, 61–64.

Li, Y., Calvo, S.E., Gutman, R., Liu, J.S., and Mootha, V.K. (2014). Expansion of biological pathways based on evolutionary inference. Cell *158*, 213–225.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857-61.

de Lichtenberg, U., Jensen, L.J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724–727.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2015). PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics *31*, 405–412.

Liu, Z., Miller, D., Li, F., Liu, X., and Levy, S.F. (2020). A large accessory protein interactome is rewired across environments. Elife *9*.

Luck, K., Sheynkman, G.M., Zhang, I., and Vidal, M. (2017). Proteome-Scale Human Interactomics. Trends Biochem. Sci. *42*, 342–354.

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408.

Magger, O., Waldman, Y.Y., Ruppin, E., and Sharan, R. (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput. Biol. *8*, e1002690.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. *35*, D26-31.

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nat. Methods *13*, 366–370.

Markmiller, S., Soltanieh, S., Server, K.L., Mak, R., Jin, W., Fang, M.Y., Luo, E.-C., Krach, F., Yang, D., Sen, A., et al. (2018). Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. Cell *172*, 590–604.e13.

Martinez Molina, D., and Nordlund, P. (2016). The cellular thermal shift assay: A novel biophysical assay for in situ drug target engagement and mechanistic biomarker studies. Annu. Rev. Pharmacol. Toxicol. *56*, 141–161.

Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. Science *296*, 910–913.

Mateus, A., Kurzawa, N., Becher, I., Sridharan, S., Helm, D., Stein, F., Typas, A., and Savitski, M.M. (2020). Thermal proteome profiling for interrogating protein interactions. Mol. Syst. Biol. *16*, e9232.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or “interologs”. Genome Res. *11*, 2120–2126.

McBride, Z., Chen, D., Lee, Y., Aryal, U.K., Xie, J., and Szymanski, D.B. (2019). A Label-free Mass Spectrometry Method to Predict Endogenous Protein Complex Composition. Mol. Cell Proteomics *18*, 1588–1606.

McClatchy, D.B., Liao, L., Park, S.K., Venable, J.D., and Yates, J.R. (2007). Quantification of

the synaptosomal proteome of the rat cerebellum during post-natal development. Genome Res. *17*, 1378–1388.

McWhite, C.D., Papoulas, O., Drew, K., Cox, R.M., June, V., Dong, O.X., Kwon, T., Wan, C., Salmi, M.L., Roux, S.J., et al. (2020). A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies. Cell *181*, 460–474.e14.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al. (2015). The human transcriptome across tissues and individuals. Science *348*, 660–665.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. Science *347*, 1257601.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. Nature *417*, 399–403.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. *5*, e1000376.

Meyer, M.J., Beltrán, J.F., Liang, S., Fragoza, R., Rumack, A., Liang, J., Wei, X., and Yu, H. (2018). Interactome INSIDER: a structural interactome browser for genomic studies. Nat. Methods *15*, 107–114.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *42*, D374-9.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function.

Genome Biol. *9 Suppl 1*, S4.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472.

Oliver, S. (2000). Guilt-by-association goes global. Nature *403*, 601–603.

Olsen, J.V., and Mann, M. (2013). Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol. Cell Proteomics *12*, 3444–3452.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358-63.

O'Reilly, F.J., and Rappsilber, J. (2018). Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. Nat. Struct. Mol. Biol. *25*, 1000–1008.

Ori, A., Iskar, M., Buczak, K., Kastritis, P., Parca, L., Andrés-Pons, A., Singer, S., Bork, P., and Beck, M. (2016). Spatiotemporal variation of mammalian protein complex stoichiometries. Genome Biol. *17*, 47.

Orre, L.M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Fernandez Woodbridge, A., Frings, O., Fredlund, E., and Lehtiö, J. (2019). SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization. Mol. Cell *73*, 166–182.e7.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. *38*, D196-203.

Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al. (2019). The BioGRID interaction database: 2019

update. Nucleic Acids Res. *47*, D529–D541.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., et al. (2005). The MIPS mammalian protein-protein interaction database. Bioinformatics *21*, 832–834.

Pankow, S., Bamberger, C., Calzolari, D., Martínez-Bartolomé, S., Lavallée-Adam, M., Balch, W.E., and Yates, J.R. (2015). ΔF508 CFTR interactome remodelling promotes rescue of cystic fibrosis. Nature *528*, 510–516.

Parker, B.L., Calkin, A.C., Seldin, M.M., Keating, M.F., Tarling, E.J., Yang, P., Moody, S.C., Liu, Y., Zerenturk, E.J., Needham, E.J., et al. (2019). An integrative systems genetic analysis of mammalian lipid metabolism. Nature *567*, 187–193.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA *96*, 4285–4288.

Piazza, I., Kochanowski, K., Cappelletti, V., Fuhrer, T., Noor, E., Sauer, U., and Picotti, P. (2018). A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. Cell *172*, 358–372.e23.

Pierson, E., GTEx Consortium, Koller, D., Battle, A., Mostafavi, S., Ardlie, K.G., Getz, G., Wright, F.A., Kellis, M., Volpi, S., et al. (2015). Sharing and Specificity of Co-expression Networks across 35 Human Tissues. PLoS Comput. Biol. *11*, e1004220.

Pourhaghighi, R., Ash, P.E.A., Phanse, S., Goebels, F., Hu, L.Z.M., Chen, S., Zhang, Y., Wierbowski, S.D., Boudeau, S., Moutaoufik, M.T., et al. (2020). Brainmap elucidates the macromolecular connectivity landscape of mammalian brain. Cell Syst. *11*, 208.

Ramani, A.K., Li, Z., Hart, G.T., Carlson, M.W., Boutz, D.R., and Marcotte, E.M. (2008). A

map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol. Syst. Biol. *4*, 180.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. *2*, 1896–1906.

Rattray, D.G., and Foster, L.J. (2019). Dynamics of protein complex components. Curr. Opin. Chem. Biol. *48*, 81–85.

Ray, J., Pinar, A., and Seshadhri, C. (2012). Are We There Yet? When to Stop a Markov Chain while Generating Random Graphs. In Algorithms and Models for the Web Graph, A. Bonato, and J. Janssen, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 153–164.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G.C., Myers, C.L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., et al. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J. Biol. *5*, 11.

Richards, A.L., Eckhardt, M., and Krogan, N.J. Mass spectrometry-based protein-protein interaction networks for the study of human diseases. Mol. Syst. Biol. *17*, e8792.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Rogers, L.D., Fang, Y., and Foster, L.J. (2010). An integrated global strategy for cell lysis, fractionation, enrichment and mass spectrometric analysis of phosphorylated peptides. Mol. Biosyst. *6*, 822–829.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Romanov, N., Kuhn, M., Aebersold, R., Ori, A., Beck, M., and Bork, P. (2019). Disentangling genetic and environmental effects on the proteotypes of individuals. Cell *177*, 1308–1318.e10.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., et al. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. USA *103*, 8390–8395.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173–1178.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.-W. (2010). CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res. *38*, D497-501.

Saha, A., Kim, Y., Gewirtz, A.D.H., Jo, B., Gao, C., McDowell, I.C., GTEx Consortium, Engelhardt, B.E., and Battle, A. (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. *27*, 1843–1858.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647–660.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32*, D449-51.

Salwinski, L., Licata, L., Winter, A., Thorneycroft, D., Khadake, J., Ceol, A., Aryamontri, A.C., Oughtred, R., Livstone, M., Boucher, L., et al. (2009). Recurated protein interaction datasets. Nat. Methods *6*, 860–861.

Sarkans, U., Gostev, M., Athar, A., Behrangi, E., Melnichuk, O., Ali, A., Minguet, J., Rada, J.C., Snow, C., Tikhonov, A., et al. (2018). The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic Acids Res. *46*, D1266–D1270.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. Nat. Biotechnol. *18*, 1257–1261.

Scott, N.E., Brown, L.M., Kristensen, A.R., and Foster, L.J. (2015). Development of a computational framework for the analysis of protein correlation profiling and spatial proteomics experiments. J. Proteomics *118*, 112–129.

Scott, N.E., Rogers, L.D., Prudova, A., Brown, N.F., Fortelny, N., Overall, C.M., and Foster, L.J. (2017). Interactome disassembly during apoptosis occurs independent of caspase cleavage. Mol. Syst. Biol. *13*, 906.

Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., Kannaiyan, N., Manrique-Hoyos, N., Kongi, K., Cantuti, L., Hanisch, U.-K., Philips, M.-A., et al. (2015). Cell type- and brain region-resolved mouse brain proteome. Nat. Neurosci. *18*, 1819–1831.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA *104*, 4337–4341.

Sheridan, R.P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. J. Chem. Inf. Model. *53*, 783–790.

Shirasaki, D.I., Greiner, E.R., Al-Ramahi, I., Gray, M., Boontheung, P., Geschwind, D.H., Botas, J., Coppola, G., Horvath, S., Loo, J.A., et al. (2012). Network organization of the huntingtin proteomic interactome in mammalian brain. Neuron *75*, 41–57.

SIB Swiss Institute of Bioinformatics Members (2016). The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. Nucleic Acids Res. *44*, D27-37.

Simonis, N., Rual, J.-F., Carvunis, A.-R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F., et al. (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat. Methods *6*, 47–54.

Skinnider, M.A., Scott, N.E., Prudova, A., Stoynov, N., Stacey, R.G., Gsponer, J., and Foster, L.J. (2018a). An atlas of protein-protein interactions across mammalian tissues. BioRxiv.

Skinnider, M.A., Stacey, R.G., and Foster, L.J. (2018b). Genomic data integration systematically biases interactome mapping. PLoS Comput. Biol. *14*, e1006474.

Skinnider, M.A., Squair, J.W., and Foster, L.J. (2019). Evaluating measures of association for single-cell transcriptomics. Nat. Methods *16*, 381–386.

Smith, G.R., and Sternberg, M.J.E. (2002). Prediction of protein-protein interactions by docking methods. Curr. Opin. Struct. Biol. *12*, 28–35.

Smith, C.L., Blake, J.A., Kadin, J.A., Richardson, J.E., Bult, C.J., and Mouse Genome Database Group (2018). Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res. *46*, D836–D842.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I., and Stagljar, I. (2015). Fundamentals of protein interaction network mapping. Mol. Syst. Biol. *11*, 848.

Sonnhammer, E.L.L., and Östlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. *43*, D234-9.

Stacey, R.G., Skinnider, M.A., Scott, N.E., and Foster, L.J. (2017). A rapid and accurate approach for prediction of interactomes from co-elution data (PrInCE). BMC Bioinformatics *18*, 457.

Stacey, R.G., Skinnider, M.A., Chik, J.H.L., and Foster, L.J. (2018). Context-specific interactions in literature-curated protein interaction databases. BMC Genomics *19*, 758.

Stacey, R.G., Skinnider, M.A., and Foster, L.J. (2020). On the Robustness of Graph-Based Clustering to Random Network Alterations. Mol. Cell Proteomics *20*, 100002.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957–968.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368.

Tan, C.S.H., Go, K.D., Bisteau, X., Dai, L., Yong, C.H., Prabhu, N., Ozturk, M.B., Lim, Y.T., Sreekumar, L., Lengqvist, J., et al. (2018). Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. Science *359*, 1170–1177.

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., and Michnick, S.W. (2008). An in vivo map of the yeast protein interactome. Science *320*, 1465–1470.

Taşan, M., Musso, G., Hao, T., Vidal, M., MacRae, C.A., and Roth, F.P. (2015). Selecting causal

genes from genome-wide association studies via functionally coherent subnetworks. Nat. Methods *12*, 154–159.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.

Thorolfsdottir, R.B., Sveinbjornsson, G., Sulem, P., Nielsen, J.B., Jonsson, S., Halldorsson, G.H., Melsted, P., Ivarsdottir, E.V., Davidsson, O.B., Kristjansson, R.P., et al. (2018). Coding variants in RPL3L and MYZAP increase risk of atrial fibrillation. Commun. Biol. *1*, 68.

Tran, J.C., Zamdborg, L., Ahlf, D.R., Lee, J.E., Catherman, A.D., Durbin, K.R., Tipton, J.D., Vellaichamy, A., Kellie, J.F., Li, M., et al. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature *480*, 254–258.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623–627.

Uhlén, M., Fagerberg, L., Hallström, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419.

Valášek, L.S., Zeman, J., Wagner, S., Beznosková, P., Pavlíková, Z., Mohammad, M.P., Hronová, V., Herrmannová, A., Hashem, Y., and Gunišová, S. (2017). Embraced by eIF3: structural and functional insights into the roles of eIF3 across the translation cycle. Nucleic Acids Res. *45*, 10948–10968.

VanBogelen, R.A., Greis, K.D., Blumenthal, R.M., Tani, T.H., and Matthews, R.G. (1999). Mapping regulatory networks in microbial cells. Trends Microbiol. *7*, 320–328.

Vandenbon, A., Dinh, V.H., Mikami, N., Kitagawa, Y., Teraguchi, S., Ohkura, N., and

Sakaguchi, S. (2016). Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. Proc. Natl. Acad. Sci. USA *113*, E2393-402.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., et al. (2009). An empirical framework for binary interactome mapping. Nat. Methods *6*, 83–90.

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

Vizcaíno, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat. Biotechnol. *32*, 223–226.

Vizcaíno, J.A., Csordas, A., del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. *44*, D447-56.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. Science *287*, 116–122.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A., et al. (2015). Panorama of ancient metazoan macromolecular complexes. Nature *525*, 339–344.

Wang, P.I., and Marcotte, E.M. (2010). It's the machine that matters: Predicting gene function and phenotype from protein networks. J. Proteomics *73*, 2277–2289.

Wang, J., Huo, K., Ma, L., Tang, L., Li, D., Huang, X., Yuan, Y., Li, C., Wang, W., Guan, W., et al. (2011). Toward an understanding of the protein interaction network of the human liver. Mol.

Syst. Biol. *7*, 536.

Wang, S., Watanabe, T., Noritake, J., Fukata, M., Yoshimura, T., Itoh, N., Harada, T., Nakagawa, M., Matsuura, Y., Arimura, N., et al. (2007). IQGAP3, a novel effector of Rac1 and Cdc42, regulates neurite outgrowth. J. Cell Sci. *120*, 567–577.

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S.M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat. Biotechnol. *30*, 159–164.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. *337*, 635–645.

Werner, J.N., Chen, E.Y., Guberman, J.M., Zippilli, A.R., Irgon, J.J., and Gitai, Z. (2009). Quantitative genome-scale analysis of protein localization in an asymmetric bacterium. Proc. Natl. Acad. Sci. USA *106*, 7858–7863.

Wodak, S.J., Vlasblom, J., Turinsky, A.L., and Pu, S. (2013). Protein-protein interaction networks: the puzzling riches. Curr. Opin. Struct. Biol. *23*, 941–953.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M. (2013). Variation and genetic control of protein abundance in humans. Nature *499*, 79–82.

Yan, K.-K., Wang, D., Sethi, A., Muir, P., Kitchen, R., Cheng, C., and Gerstein, M. (2016). Cross-Disciplinary Network Comparison: Matchmaking Between Hairballs. Cell Syst. *2*, 147–157.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res. *14*, 1107–1118.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al. (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104–110.

Zanivan, S., Krueger, M., and Mann, M. (2012). In vivo quantitative proteomics: the SILAC mouse. Methods Mol. Biol. *757*, 435–450.

Zhang, L., and Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Biol. Evol. *21*, 236–239.