Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by

M.Sc, Francis O'Reilly
born in Enniskillen, Northern Ireland
Oral-examination: 11/09/2015

USING HIGH-THROUGHPUT METHODS TO CHART
PROTEIN-PROTEIN AND PROTEIN-LIPID INTERACTIONS
IN EUKARYOTIC CELLS

Referees: Prof. Dr. Frauke Melchoir
Dr. Kiran Raosaheb Patil

Summary

To truly understand the self-replicating eukaryotic cell we need to make significant progress unraveling the interactome, the sum of all the interactions between the proteins and the metabolites of the cell. Here, I present two projects to that end:

The first study maps protein complexes in a unicellular thermophilic eukaryote, *Chaetomium thermophilum,* using an innovative approach integrating several biological techniques. Thermophilic proteins are, by their nature, more stable than their mesophilic counterparts and *C. thermophilum* has been described as a potential model organism for structural studies.

We use a size exclusion chromatography (SEC) to separate high-molecular weight protein complexes from cell lysate. Coeleuting proteins are identified by mass spectrometry and inferred to be in a protein complex together. Chemical crosslinking combined with mass spectrometry (XL-MS) is applied to the SEC fractions to provide direct biochemical confirmation of the predicted protein-protein interactions. Together these methods have allowed us to identify protein complexes with novel subunits and also functionally related coeluting proteins not hitherto known to form protein complexes.

Additionally, negative-stain electron microscope (EM) images of protein mixtures from the SEC fractions are correlated with the elution patterns of the identified complexes to distinguish the structural signatures (Shapes) of specific complexes. This enabled manual picking of particles from cryo-EM micrographs to solve the molecular structure of *C. thermophilum* fatty acid synthase (FAS) to 4.7Å resolution directly from the SEC fractions without further purification. A novel binder of FAS was identified by EM and confirmed with XL-MS as a branching biotin dependent carboxylase, which together constitutes a potential metabolon for the production of branch chain fatty acids.

This project facilitates the use of *C. thermophilum* as a model organism for structural biology and the methods may open the way for high-throughput structural biology.

The second project used a high-content screen to discern the roles of lipid classes on the localization of proteins, protein complexes and biological processes in the cell. This approach attempts to shed light on this protein-metabolite network by genetically depleting selected lipid classes by perturbing their biosynthesis and using *in vivo* imaging to map localization changes of proteins in the cells and therefore identification of lipid dependent localizations. The database created in this project will facilitate the development of follow-up studies that will further discern the structural roles of lipids in the organization of the proteome.

Zusammenfassung

Um ein grundlegendes Verständnis von der sich selbst teilenden eukaryotischen Zelle zu bekommen, ist es nötig die Entschlüsselung des Interaktoms, die Summe aller Interaktionen zwischen Proteinen und Metaboliten der Zelle, signifikant voran zu treiben. Zu diesem Zweck stelle ich im Folgenden zwei Projekte vor.

Die erste Studie charakterisiert Proteinkomplexe in einem einzelligen thermophilen Eukaryoten, *Chaetomium thermophilum*, indem ein innovativer Ansatz benutzt wird, der mehrere biologische Techniken miteinender verbinden.

Thermophile Proteine, sind von Natur aus stabiler als ihre mesophile Gegenstücke und *C. thermophilum*, wurde als potentieler Modelorganismus für Strukturanalysen beschrieben.
Wir benutzen Größenausschluβchromatographie (SEC) um Komplexe mit hohem Molekulargewicht, die aus einem Zelllysat stammen, zu trennen. Proteine, die coeluieren werden per Massspektrometrie identifiziert und es wird angenommen, dass sie zusammen einen Komplex bilden. Chemisches Quervernetzen in Verbindung mit Massenspektrometrie (XL-MS) wird auf jede SEC Fraktion angewendet um die direkte biochemische Bestätigung der vorhergesagten Protein-Protein Interaktionen zu erhalten. In der Summe haben uns diese Methoden ermöglicht Proteinkomplexe mit neuen Untereinheiten zu identifizieren und auch funktionel verwandte coeluierende Proteine, die zuvor nicht dafür bekannt waren dass sie Proteinkomplexe bilden, zu identifizieren.

Zusätzlich wurden negativ-gefärbte elektronenmikroskopische Bilder von Proteinkomplexen aus den SEC Fraktionen mit den Elutionsmustern der identifizierten Komplexe verglichen um die strukturellen Signaturen (Shapes) spezifischer Komplexe zu unterscheiden. Dies ermöglichte manuelles Sammeln von Partikeln aus Kryo-EM Mikrofotographien um die molekulare Struktur von *C. thermophilum* Fettsäure Synthase (FAS) mit einer Auflösung von 4.7Å direkt aus den SEC Fraktionen ohne weitere Aufreinigung zu entschlüsseln. Ein neuer Bindungspartner von FAS wurde mittels EM identifiziert und mit XL-MS als verzweigende Biotin abhängige Carboxylase bestätigt, die zusammen ein potentielles Metabolon für die Herstellung von verzweigtkettigen Aminosäuren bilden.

Dieses Projekt erleichtert das Benutzen von *C. thermophilum* als Modelorganismus für die Strukturbiologie und die Methoden könnten den Weg ebnen für die Hochdurchsatz Strukturbiologie.

Das zweite Projekt verwendete einen Screen um die Rolle von Lipidklassen auf die Lokalisierung von Proteinen, Proteinkomplexen und biologischen Prozessen innerhalb der Zelle

zu erkennen. Dieser Ansatz zielt darauf Licht ins Dunkel von Protein-Metabolit Netzwerken zu bringen. Dies wird durch genetisches Vermindern von ausgewählten Lipid Klassen durch das Perturbieren ihrer Biosynthese und unter Verwendung von *in vivo* bildgebenden Verfahren um die Lokalisierungsänderungen der Proteine in der Zelle abbzubilden erzielt und dadurch die Lipid-abhängige Lokalisierung identifiziert. Die Datenbank, die innerhalb dieses Projektes generiert worden ist, wird die Weiterentwicklung von weiterführenden Studien vereinfachen und des Weiteren die strukturelle Rolle von Lipiden in der Organization des Proteomes aufzeigen.

Contents

CHAPTER 1: INTRODUCTION

1.1.1: Introduction to the field of interactomics

The long-term goal of the field of interactomics is to create a global atlas of all interactions in the cell. Details such as molecular affinities and dynamics can then be added to color this picture. This field is gradually elucidating the full complexity of the cell and enlightening us to molecular mechanisms of complex diseases (Vidal et al., 2011).

In the pre-genomic era, biomolecular interactions were discovered and studied on a case-by-case basis. While discovering the atomic detail of these interactions has largely remained labor intensive and time-consuming, simply identifying the interactors has become high-throughput through the invention of 'omics' approaches and technologies (Collura and Boissy, 2007).

The landmarks usually recognized as the beginning the 'omics' age were the first draft genomes of model organisms such as *Escherichia coli* (Blattner et al., 1997), *Saccharomyces cerevisiae* (Mewes et al., 1997), *Drosophila melanogaster* (Adams et al., 2000) and the Human Genome Project (Lander et al., 2001). These genomes were the first time that the entire assemblage of genes in an organism was catalogued and could be used as a template onto which the fields of proteomics and transcriptomics are built.

Simply generating a comprehensive 'parts list' of the RNAs and proteins expressed in cells is still an endeavor occupying biologists, especially when looking at new cell types or cellular environments (Iyer et al., 2015; Kim et al., 2014; Lappalainen et al., 2013; Mann et al., 2013). Additionally, there are also cellular components that are not directly relatable to the genome, which the field of metabolomics has emerged to address. These metabolites (sugars, lipids, etc.) have a huge variety of physicochemical properties that make them difficult to study and therefore make mapping their entire cellular repertoire an ambitious challenge (Fahy et al., 2009; Wishart et al., 2013).

These lists of cellular components do not reveal much information about the function of the cellular system but when investigated in the context of the 'interactome' these cellular constituents physically interact to produce the highly-regulated cellular phenome. Due to the number of possible interactions between these components, high-throughput approaches have been developed to measure as many of these potential interactions in parallel as possible. These are reviewed below.

1.1.2: High-throughput approaches for mapping protein-protein interactions

The meaning of 'high-throughput' molecular biology has changed over the years as new technologies make older technologies seem limited in scope. We can now map protein-protein interactions (PPIs) on scales that previously seemed unfeasible due to time and resource constraints (Wodak et al., 2013).

Some of the first studies that may be described as high-throughput were protein arrays used for studying protein-protein interactions (Michaud et al., 2003). These techniques were tedious and limited in scope as the arrays are difficult to manufacture reproducibly, though they still have specific uses such as investigating antibody-antigen interactions today.

The first proteome-wide mapping of protein-protein interactions occurred in yeast, taking advantage of its genetic tractability. Yeast two-hybrid (Y2H) studies mapped binary protein-protein interactions *in vivo* by expressing a pair of query proteins, each linked to part of a transcription factor, which caused the expression of a reporter gene upon their interaction (Uetz et al., 2000). This method is particularly suited to detecting binary interactions and even transient interactions. The early versions of Y2H screens were associated with false-positive rates of up to 50%, as some proteins when fused to the DNA-binding domain, produce spontaneous transcriptional activation and some proteins generate toxic reactions in yeast. Post-translational modifications (PTMs) on the fused proteins are also often aberrant. Many of these problems have been addressed in recent years (Vidal and Fields, 2014), e.g. split-ubiquitin methods don't require protein import to the nucleus so has allowed mapping of interactions that were previously intractable to traditional Y2H screens (e.g. full-length integral membrane proteins) (Snider et al., 2010). There are efforts still underway to map several model organisms by this technique (Fall et al., 2011; Li et al., 2004; Rajagopala et al., 2014; Rolland et al., 2014; Rozenblatt-Rosen et al., 2012).

A second large advance in high-throughput proteome mapping was the application of affinity-pull-down plus mass-spectrometry techniques (AP-MS). The first full protein complexome mapping attempts were performed in *S. cerevisiae* with TAP-tag pull-downs of endogenously-expressed proteins from (Gavin et al., 2006; Krogan et al., 2006). Identifying the interaction partners of each individual protein enabled the construction of interaction networks and, for the first time, mapping of full protein complexes and not just binary interactions. This technique is still widely used in several species (Ewing et al., 2007; Guruharsha et al., 2011; Kühner et al., 2009) and in recent years has expanded to analyzing membrane proteins (Babu et

al., 2012). The intact native complexes that can be purified with this technique can also be investigated structurally and functionally (Russell et al., 2004).

The most recent revolution in high-throughput mapping of protein-protein interactions is the use of multiple orthologous fractionation techniques; this is used to separate native protein complexes from lysate and to use quantitative mass spectrometry to identify co-fractionating proteins (Kristensen and Foster, 2013). The concept relies on the likelihood that proteins that coelute together when the lysate is separated by charge and size are part of the same complex. The most comprehensive lists of native human protein complexes to date have been generated using this method (Havugimana et al., 2012; Kristensen et al., 2012). As it is not readily possible to refine each protein complex entirely using these simple fractionation techniques, the inference of protein complexes relies upon independently generated data such as gene conservation, co-localization, co-expression and previous protein-protein interaction studies for validation. The major advantage with this technique is that genetic manipulation of cells is not needed thus it can be used to map protein complexes from any species, even those not genetically tractable.

Overall, these techniques have produced our broadest understanding of interactions in the cell. These interaction maps are not complete and will expand as we study condition-specific interactions (Kristensen et al., 2012) and use techniques such as chemical crosslinking to stabilize transient interactions for identification (Bui et al., 2013; Plaschka et al., 2015).

1.1.3: High-throughput approaches for mapping other cellular interactions

In addition to the major advances in mapping protein-protein interactions outlined, there has been considerable progress in mapping the RNA-binding proteome (Castello et al., 2012), the DNA- binding proteome (Kasinathan et al., 2014) and protein interactions with other metabolites (Li et al., 2010).

The metabolite-protein interaction network is particularly challenging to analyze due to the range of physiochemical properties of metabolites necessitating that different techniques need to be designed for each chemical sub-class (Yang et al., 2012). The current high-throughput methods for directly analyzing these interactions are: *in vitro* binding arrays (Gallego et al., 2010; Saliba et al., 2014; Vegas et al., 2008; Yu et al., 2004), *in vivo* pull-downs of metabolite-bound proteins (Li et al., 2010; Maeda et al., 2014), and *in vivo* protein-metabolite crosslinking techniques (Haberkant et al., 2013). There are also indirect techniques, such as

identifying structural changes in proteins when certain metabolites are present to infer their binding (Feng et al., 2014).

Recent advances in mass spectrometry technology have allowed identification of metabolites in a high-throughput manner and have facilitated the design of some novel approaches to map protein-metabolite interactions. However, the biggest challenge in protein-metabolite interaction networks remains that metabolites are difficult to identify reliably from mixtures. Their variety causes a huge number of fragmentation possibilities - unlike the discrete 20 amino acids of peptides. As metabolite tandem-MS databases further develop, this issue will become less pertinent and will allow for the study of metabolites that were not previously tractable (Smith et al., 2005).

1.1.4: Scope of thesis

In this thesis, I present two projects. The first, in Chapter 2, is a high-throughput protein-protein interaction mapping study in a thermophilic eukaryote using proteome fractionation and mass spectrometry. In this project, chemical crosslinking is also used to confirm PPIs and map their stoichiometries. Finally, a method was designed to solve the structures of these native complexes by electron microscopy from these complex mixtures. The methods described are designed to be generic for use in any species and do not require genetic manipulation of the organism.

The second project, in Chapter 3, is a high-content screen to discern the roles of lipid classes on the localization of proteins, protein complexes and biological processes in the cell. This approach attempts to shed light on this protein-metabolite network by genetically depleting selected lipid classes by perturbing their biosynthesis and using *in vivo* imaging to map localization changes of proteins in the cells.

Together these projects advance our knowledge of the eukaryotic interactome and provide resources to serve as the basis for many follow-up studies.

CHAPTER 2: AN INTEGRATIVE APPROACH FOR DISCOVERY AND STRUCTURAL STUDY OF LARGE PROTEIN COMPLEXES IN A EUKARYOTIC THERMOPHILE

2.1: Abstract

Protein complexes constitute most of the basic functional units of the cell. Analysis of protein complex composition and structure is thus crucial for understanding most biological functions. Structural studies of large protein complexes often utilize thermophilic archaea and bacteria as their heat adaptation implies, among other things, an increased stability of their protein-protein interactions.

The eukaryotic thermophile *Chaetomium thermophilum* has an optimal growth temperature of 55°C and represents an ideal model system for the structural study of eukaryotic protein complexes. We have systematically studied large protein complexes from *C. thermophilum* with an integrated approach that combines native protein complex extraction, quantitative mass spectrometry (MS), chemical crosslinking and electron microscopy (EM).

Protein complexes ranging from 200kDa to 5MDa from *C. thermophilum* cell lysate were separated by size exclusion chromatography (SEC). Elution profiles of individual proteins were generated by quantitative MS and correlated to predict protein complex composition. Integrating further biochemical evidence from chemical crosslinking of the SEC fractions and external database information allows for high-confidence protein complex prediction and re-annotation of proteins from this organism. Proteins not previously described as protein complex subunits are assigned to complexes and the distance restraints imposed by crosslinks also enables modeling of the topologies and stoichiometries of these complexes.

In parallel, each fraction was analyzed by negative-stain EM to search for class averages that could be correlated with protein complex abundances assigned from the MS data. This information enabled manual picking of particles from cryo-EM micrographs to solve the molecular structure of *C.thermophilum* fatty acid synthase to 4.7Å resolution directly from the SEC fractions without further purification. The solved native structure of *ct*FAS provides first-time insights into FAS function and identifies a novel binder, a branching carboxylase, which constitutes a potential metabolon for the production of branch chain fatty acids.

2.2: Contributions

I partially conceived this project and have been involved in or led all aspects of it as it has evolved over time. It is a large and highly collaborative project between the Gavin, Beck and Russell groups. Particular major contributions were;

- Panagiotis L. Kastritis performed the Electron Microscopy and Molecular Modeling
- Matt Rogon developed the machine learning approach for protein interaction prediction
- Thomas Bock performed the chemical cross-linking mass spectrometry

I established all other biochemical methods used during the project and majorly contributed to all abovementioned developments.

2.3: Introduction

2.3.1: Systematic mapping of protein complexes

Interactions between proteins are the cornerstone of many (if not most) biological processes. To understand the function of these assemblies it is important to identify them, but their diversity and varying physicochemical properties makes this a difficult task (Perutz and Raidt, 1975; Warshel et al., 2006; Yonath et al., 1988).

Many attempts have been made to produce a comprehensive list of protein-protein interactions (PPIs) as discussed in Chapter 1. The most recent advances have used fractionation approaches combined with quantitative mass spectrometry, to infer human protein complexes from coeluting native proteins (Havugimana et al., 2012; Kristensen et al., 2012). These studies have allowed for inference of PPIs without the generation of thousands of transgenic cell lines. Kristensen *et al* (2012) presented a simplified fractionation method to identify PPIs by using size exclusion chromatography (SEC) as a single dimension of separation and they showed it was sufficient to confirm known complexes from human cell lines.

These methods provide little information on topological structure or stoichiometry of the identified complexes, or the functional significance of each member.

2.3.2: *Chaetomium thermophilum* as a model organism for structural biology

To address the abovementioned issues, we study large assemblies from the thermophilic eukaryote, *C. thermophilum*. Due to their unique stability (Sterner and Liebl, 2001), thermostable proteins have yielded many landmark structures; including the first structural information for the ribosome (Yonath et al., 1984, 1988) and the proteasome (Löwe et al.,

1995). Thermophilic prokaryotes are a common source for these proteins and, obviously, cannot be used to study eukaryote-specific cellular processes. To address this limitation we turn to the thermophilic *C. thermophilum*, a fungus that thrives at temperatures up to 60°C. The availability of the *C. thermophilum* genome (Amlacher et al., 2011), transcriptome and proteome (Bock et al. 2014), and the possibility to grow this fungus under standard laboratory conditions has opened avenues to solve new eukaryotic structures, *e.g.* nucleosomal complexes (Hondele et al., 2013), nuclear transport receptors (Monecke et al., 2012) and even parts of the nuclear pore complex (Amlacher et al., 2011).

The beginning of this project coincided with the release of the genome and proteome of *C. thermophilum* from groups within EMBL-Heidelberg (Amlacher et al., 2011; Bock et al., 2014), so this species is ripe for analysis of its protein complexes in a high-throughput manner.

2.3.3: State-of-the-art methods for analyzing protein complex composition

To systematically study native protein complexes present in this species, we applied an integrative approach to the single fractionation technique described by Kristensen *et al* (2012). Protein complexes are separated from cell lysate using analytical size-exclusion chromatography and subsequently each protein is identified by quantitative mass spectrometry. We identify known protein complexes by cross-correlation of protein elution profiles and integration of external database information from orthologous complexes. Addition of chemical crosslinking combined with mass spectrometry (XL-MS) to this pipeline adds biochemical validation to these known complexes and provides additional information on stoichiometry, topology and additional novel subunits (Gingras et al., 2007). XL-MS has recently come of age as an effective tool for the study of purified large multi-subunit complexes (Plaschka et al., 2015) but has so far not been utilized in complex protein mixtures. Furthermore, XL-MS also facilitates integrative structural modeling methods for prediction of the topology of previously unknown protein complexes.

In total

2.3.4: Generating structures of protein complexes from complex mixtures

Additionally, as a method for high-throughput structural analysis, we correlate the elution profiles of the discovered protein complexes with abundances of particles from electron micrographs of the same SEC fractions.

The combination of EM and MS has been described as 'visual proteomics' in 2006 but has been utilized in few studies since then (Beck et al., 2009; Nickell et al., 2006). Electron

tomography and MS of natively derived samples has however shown heterogeneity in protein complexes not normally observed in purified recombinant samples (Han et al., 2009), demonstrating the value to investigating native material.

This approach identifies complexes that are amenable for structural study by EM and allows structure solution from complex protein mixtures. We demonstrate that the fractions from a simple one-step SEC fractionation are amenable to structural determination by this combination of techniques by solving the cryo-EM structure of native fatty acid synthase (FAS) (4.7 Å, FSC=0.143) and observing a catalytic conformation that has not been previously described. Secondly, we characterize novel additional proteins on the periphery of the complex that correspond to cellular carboxylases not previously described as FAS binders. This interaction could only have been identified by using natively derived material. Using integrative modeling, a mechanism is proposed where FAS directly binds cellular carboxylases while the acyl carrier protein (ACP) is in active proximity to facilitate acyl chain transport. In summary, our molecular model postulates a hypothesis for de novo branched-chain fatty-acid biosynthesis via a snapshot of a transient metabolon.

2.3.5: Impact of research

In summary, we were able to study the high molecular weight complexome in this organism by integrating molecular biology approaches in a novel manner. The generated dataset provides unprecedented molecular characterization for previously unresolved protein assemblies and is a useful resource for structural biologists wishing to take advantage of the thermal stability of the *C. thermophilum* proteome. Additionally, this approach is general and can form the basis for protein complex screening for any organism with an annotated genome in the future whether for whole proteome screening or investigation of the states of specific protein complexes from the native source.

2.4: Results and discussion

2.4.1: Summary of pipeline

The pipeline for mapping complexes and their structural study is summarized in Figure 2.1. During this project we developed methods for growing and lysing *C. thermophilum* cultures for the extraction of native proteins outlined in Appendix A.I. After cell lysis the lysate is enriched for high-molecular complexes by spin-filtering and this concentrated sample is separated by SEC (Figure 2.1 A) (Appendix A.II). 30 fractions were analyzed in the molecular weight range from ~200kDa – 5MDa.

Of the 4297 expressed protein in this organism (Bock et al., 2014), 1147 proteins were identified in >1 biological replicate (Appendix A.III). Using quantitative MS, the intensity of each protein in each fraction is measured and these intensities are plotted across all 30 fractions to generate elution profiles (Figure 2.1Bi). Interestingly, many well-known complexes eluted as larger than expected from the SEC column (see Appendix A.IV), this may represent cellular interactions not previously observed.

Proteins bound together as part of a complex will coelute. To identify these novel complexes a simple cross-correlation of elution profiles is used to assess the similarity of each protein's elution. However, as many protein complexes also coelute with each other, we integrated predicted protein-protein interaction interfaces and external database information from well-annotated orthologs to predict the *C. thermophilum* protein complexes. These datasets were integrated using a machine-learning approach and clustering was used to preict the protein assemblies (Figure 2.1Bii).

In parallel, SEC separated fractions are crosslinked to provide biochemical validation for the protein-protein interactions predicted (Figure 2.1C). These crosslinks also provide spatial restraints, information that enables modeling of the discovered interactions.

Finally, the elution patterns of the protein complexes predicted were correlated with EM micrographs, imaged from the same SEC fractions. This enables identification of structural signatures (shapes) of specific complexes (Figure 2.1D).

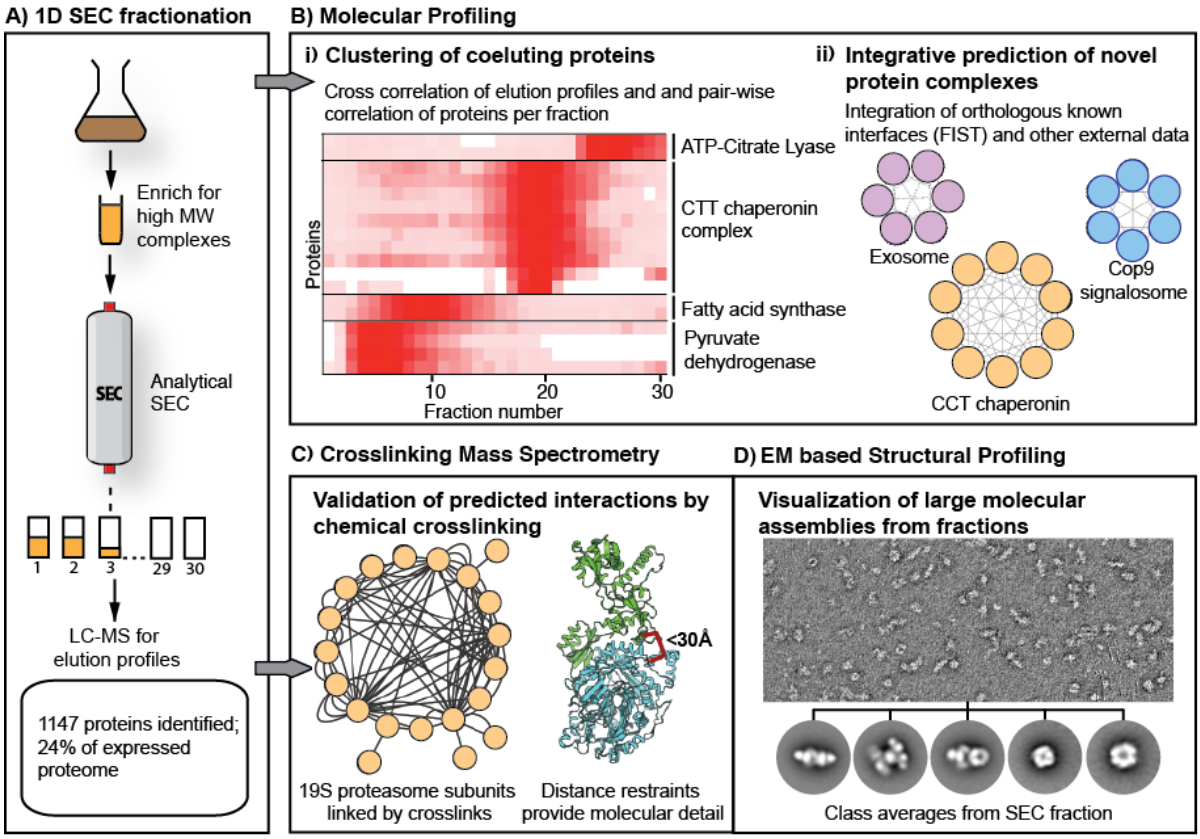

Figure 2.1: Overview of pipeline to predict and visualize protein complexes from C. thermophilum

A) 1-dimensional fractionation by SEC of lysate enriched for high molecular weight complexes. 30 fractions were collected, covering 200kDa to ~5MDa. Fractions were subjected to quantitative MS and a total of 1147 proteins were identified in >1 biological replicate. **B) i)** Each protein's elution profile is pair-wise correlated to generate a cross-correlation score which describes similarity. **ii)** Integration of external data from orthologous proteins is used to add weight to predicted interactions and clustering approaches are used to separate proteins into predicted protein complexes. **C)** Chemical crosslinking is performed on the fractions to provide chemical validation to predicted protein-protein interactions. The linking of specific lysines by the crosslinker imposes a distance restraint that allows molecular modeling of the protein-protein interaction **D)** In parallel, fractions are visualized by negative-stain EM to observe abundant structures in the fraction. These class averages can be correlated with the complexes identified by MS to allow structural study directly from non-pure native source material.

2.4.2: Predicting clusters of interacting proteins based on simple 1-dimentional SEC separation

To predict members of protein complexes we assess the similarity of elution profiles, as proteins which are in the same complex should elute together and then we integrate external information to differentiate true interactions from spurious coalitions.

Of the 1147 proteins identified in ≥2 biological replicate in the 30 SEC fractions, 974 proteins eluted across ≥4 consecutive fractions and were retained for further analysis (See Appendix 2.III and Supplementary table E.I). This filtered out very low-abundant proteins which do not provide enough elution information for cross-correlation analysis. Each of the retained elution profiles was compared in a pairwise manner and cross-correlation coefficients (CCC) were generated (see Materials and Methods).

This simple cross-correlation method does not have enough discriminatory power to deconvolute actual interacting proteins from proteins that incidentally coelute in many cases. To address this we integrate external, independently generated information from orthologous proteins in *S. cerevisiae,* which has well-annotated protein complexes*.* Orthologous proteins from *S. cerevisiae* were predicted using eggNOG (contributed by Vera van Noort) (see materials and methods) and interactions between these orthologs were extracted from the STRING interaction database (Franceschini et al., 2013). Direct physical interactions were excluded to avoid circularity issues and are retained for validation. The extracted interaction confidences are based upon co-expression, domain co-occurrence, gene neighborhoods, co-citation, functional relationship, and genetic interactions in *S. cereviaise*.

Additionally, 1553 protein-protein interaction interfaces predicted from homologous interfaces in the Protein Data were found using the FIST algorithm contributed from Prof. Rob Russell (Aloy and Russell, 2002).

We use a random-forest machine-learning technique to combine each of these three datasets (coelution, ortholog interaction information, and predicted interfaces) to produce an interaction probability for each protein pair. To weight each of these datasets, a random-forest machine-learning algorithm was then trained on a benchmark assembled from known complexes. The benchmark consisted of complexes discovered by AP-MS studies in *S. cerevisiae* and was supplemented with complexes from several other species identified by simple BLAST searches against the PDB (See Appendix A.IV and E.II).

An interaction network was built using 6146 retained high-confidence interactions (<15% FDR) as predicted by the Random forest algorithm (See Appendix A.V for a summary). This network contains 736 proteins.

In order to define protein complex membership from this protein-protein interaction network, the cluster-growth algorithm ClusterOne was used to identify highly interconnected areas containing 3 or more proteins (Nepusz et al., 2012)(see materials and methods). In total 50 clusters with were predicted with a P-value <0.05 and in total 443 proteins were assigned to clusters. Their arbitrarily assigned cluster number, 'Cluster 1' – 'Cluster 50', is used to refer to these clusters and they are summarized in Appendix E.III. Some of the predicted clusters are highly redundant as they are from highly interconnected parts of the network. These redundant clusters are merged for visualization in Figure 2.2. Of these 37 non-redundant merged clusters 19 recapitulate known complexes, some others have coeluting proteins with similar functions, such as a cluster of lipid biosynthesis proteins. There are also clusters of unannotated proteins.

Importantly, the proteins within these clusters do coelute (Figure 2.3A), showing that the prediction of interactions is driven by our experimentally derived data. In many of the known complexes interactors are identified which are not part of the canonical 'core' complex; such as, NAS6 with the proteasome regulatory particle, and Vid27 as part of the CCT chaperonin (Figure 2.3B) (Benschop et al., 2010).

Large highly interconnected clusters are also predicted, such as those containing the RNA polymerases and the ribosomes, these include not just the core complexes but also transient binders such as transcription factors and elongation factors (Figure 2.3B). External information can be used to further sub-cluster these clusters based upon known orthlogous physical known interactions (withheld from interaction prediction) to annotate the core complexes and highlight predicted novel subunits (this is yet to be implemented).

To validate some of these previously unreported interactions, we use XL-MS to provide direct biochemical validation of these predicted interactions and to map their topology.

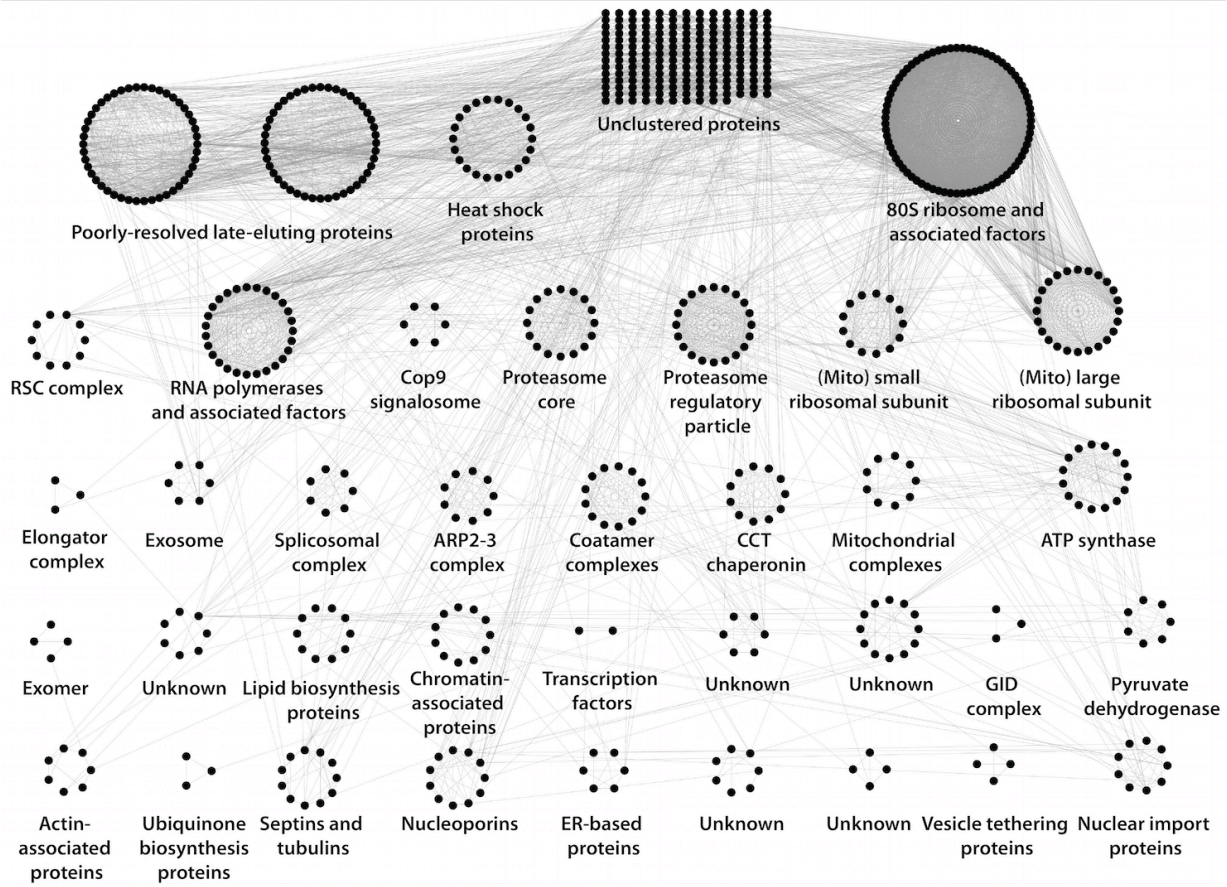

Figure 2.2: All protein clusters in the predicted interaction network.

Schematic representation of all predicted clusters and the high-confidence predicted interactions between them. Redundant clusters have been merged and clusters are named after their main constituents. Note, there are two large clusters containing poorly resolved complexes which elute at the end of the column. [Network analysis was performed in collaboration with Matt Rogon]

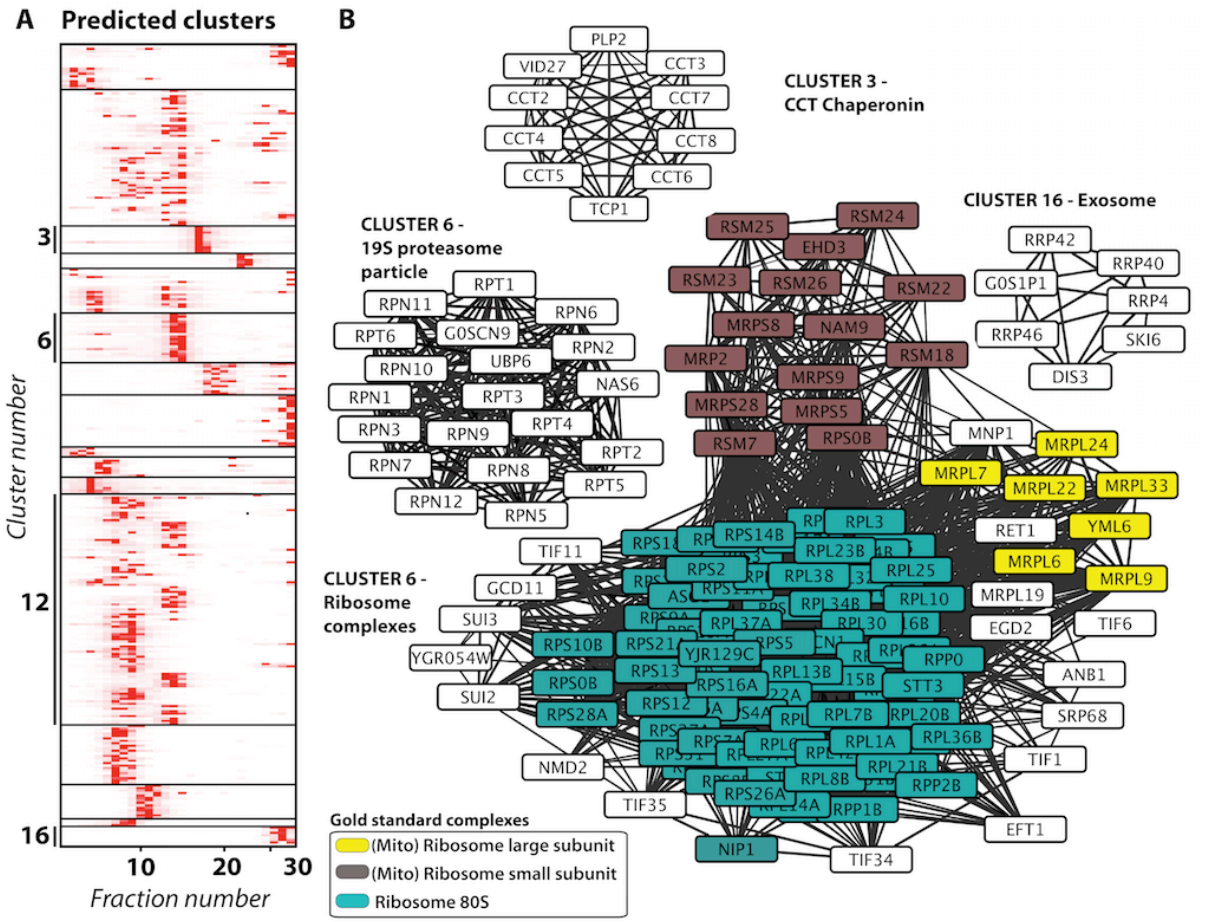

Figure 2.3: Example clusters extracted from prediction interaction networks.

A) Elution profiles of the predicted clusters 1-16, extracted from the network of predicted interactions, are shown. **B)** Some of these clusters contain very well defined complexes such as Cluster 3, 6 and 16. Others have functionally related proteins with similar elution patterns that are poorly discriminated into individual complexes such as Cluster 12. Subunits of the ribosome in cluster 12 are colored depending on their complex membership in the benchmark (See Appendix A.IV and E.II). Proteins are labeled with their *S. cerevisiae* ortholog names, if known.

2.4.3: Chemical crosslinking of complex mixtures

Traditional applications of chemical crosslinking combined with mass spectrometry (XL-MS) use purified complexes of interest. We show for the first time the application of chemical crosslinking MS to a complex mixture of protein assemblies after only one-dimensional protein fractionation.

To produce enough material for XL-MS analysis a preparative SEC column was used as it has the same separation range as the analytical column (~200kDa- >5mDa) (see materials and methods). Consecutive fractions were pooled to obtain a protein concentration of 1 µg/µl. In total, this generated 7 XL-MS pools that were subjected to mild chemical crosslinking using the chemical crosslinker disuccinimidyl suberate (DSS), which crosslinks lysine chains with a maximum distance of 30Å from Cα to Cα. Peptide mixtures were separated by gel filtration to enrich for the crosslinked peptides before MS analysis as described previously (Leitner et al., 2014).

The published methods for calculating False Discovery Rates (FDRs) of crosslinks are based upon XL-MS analysis of individual complexes (Walzthoeni et al., 2012). These methods are not applicable in this study as some of the crosslinked fractions contain >200 proteins.

The expected FDR for the crosslinks of complex mixtures was estimated on a structurally well-defined subset of proteins. In total, 31 highly conserved homomeric protein assemblies were modeled using a combination of threading/homology modeling methods (see materials and methods). These complexes covered a range of MW's and were therefore found in fractions of varying background complexity.

We used the standard software pipeline xQuest to search for crosslinks and produce an ld-score that rates the quality of the spectra used to assign the crosslink. Crosslinks can be sorted into two categories: intra-protein crosslinks (linking distinct lysines within the same protein chain) or inter-homomeric crosslinks (linking two different protein chains of the complex). In the latter category are also crosslinks identified from and to the same lysine residue in a sequence (zero-distance crosslink), and therefore must crosslink two copies of the same protein. For these protein complexes, spatial restraints of identified crosslinks (ld-score > 20) were calculated using molecular models. Spatial restraints of chemical crosslinking were considered if they satisfied the 30 Å threshold, described previously (22772729).

There were 26 crosslinks satisfied by either inter- or intra molecular contacts and thus are ambiguous. At an ld-score of >20, 30% of the crosslinks (23 crosslinks) were satisfied in both states, 17% were satisfied as only Inter-homomeric (14 crosslinks), 42% were satisfied as

Intra-protein crosslinks (34 crosslinks), and 12% were not satisfied in any of the categories. No crosslink was found violating the distance restraint with an ld-score >23.5 in the homomer subset (Figure 2.4A&B). For the rest of this study an ld-score cut-off of >23.5 is used for all reported crosslinks.

Using the chemical crosslinking we were able to provide biochemical evidence for the assembly state of the mitochondrial chaperonin-60 (GroEL-homologue). A zero-distance crosslink was discovered Lys458-Lys458. This distance restraint it can be used to distinguish between the 7mer state (~440 kDa) reported by the PDB = 1IOK and the 14mer assembly which has been reported by PDB code = 2YEY (Figure 2.4C). Further confirmation for the 14mer assembly comes from the SEC elution profile as the protein elutes as ~800kDa, close to the 848kDa of the 14mer (Figure 2.4D).

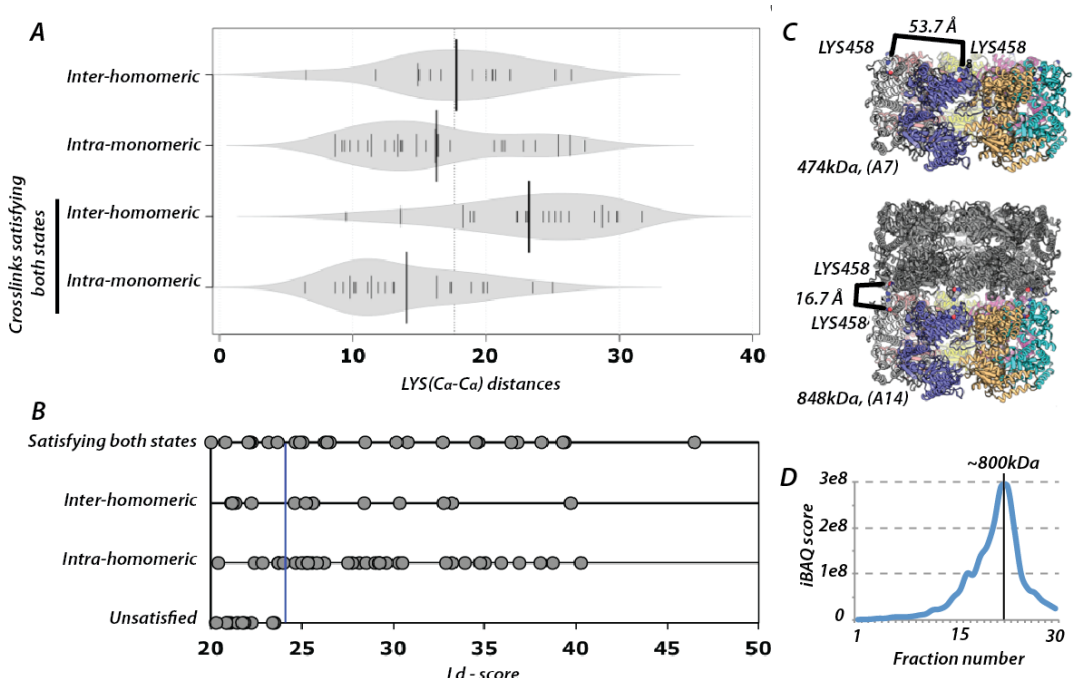

Figure 2.4: Evaluation of false-discovery rate for XL-MS in complex mixtures from modeled homomers

A) For crosslinks with ld-score >20 detected in the modeled homomers, Euclidian distances of crosslinked lysines (Cα to Cα) follows the expected normal distribution. **B)** Ld-score cut-off > 23.5 is sufficient to account for true positive crosslinks, benchmarked on 31 homomers of various oligomeric states. **C)** In Chaperonin-60 the crosslink from Lys458-Lys458 allows for discrimination between the two reported oligomeric states, A7 and A14. The distance restraint of 30Å is only satisfied in the A14 state. **D)** Elution profile of chaperonin-60-like, showing that it forms a single peak ~800 kDa. [Chemical crosslinking was performed by Thomas Bock and molecular modeling was performed by Panagiotis kastriti]

2.4.4: Integration of crosslinking data into predicted complexes validates complexes.

The 50 clusters predicted from our integrated network were searched for crosslinks for validation for some of these interactions. Not all interacting proteins will crosslink as inter-subunit lysines must be available within 30 Å (Cα to Cα) and the complexes need to be abundant enough for these crosslinked peptides to be detected.

518 inter-subunit crosslinks have been identified within the clusters, corresponding to 272 binary interactions (as two subunits can have many unique crosslinks between them). The majority of the crosslinks were found in 5 readily crosslinked clusters; the 19S regulatory proteasome particle, proteasome core, nucleosome, ribosomal subunits and Cluster 17, containing highly abundant heatshock proteins.

Table 1: summarizing number of inter-protein crosslinks found in predicted clusters

Cluster	# proteins in the cluster	# Inter-protein crosslinks	# Binary interactions
Proteasome 19S	20	98	55
Proteasome 20S	13	17	8
Nucleosome	4	12	6
Ribosome and its associated factors	94	277	164
Heat shock proteins	13	62	16
		TOTAL 466	**TOTAL 249**
OTHERS	325	57	29
		TOTAL 518	**TOTAL 273**

In cluster 12 containing the ribosome and its associated factors (see figure 2.3), 71 of the 112 proteins predicted in the cluster had a least one inter protein crosslink. If only these crosslinked proteins are plotted, three sub-clusters appear, corresponding to the 40S small ribosomal subunit, the 60S large ribosomal subunit and the mitochondrial large ribosomal subunit (Figure 2.5). Also identified are transient binders of these complexes such as the Eukaryotic initiation factors 1 & 2 and the elongation factor 2.

The crosslinking also provides information about topology in a large complex like the ribosome. The crosslinks, which link these transient subunits, provide distance restraints and therefore structural information. Subunits with multiple crosslinks can provide enough structural information to identify orientation of the binding to predict and model interfaces (work in progress).

In summary this approach validates/complements the computational prediction approaches for the identification of protein complexes. These examples show the power of XL-MS for capturing even transient interactions these complex mixtures.

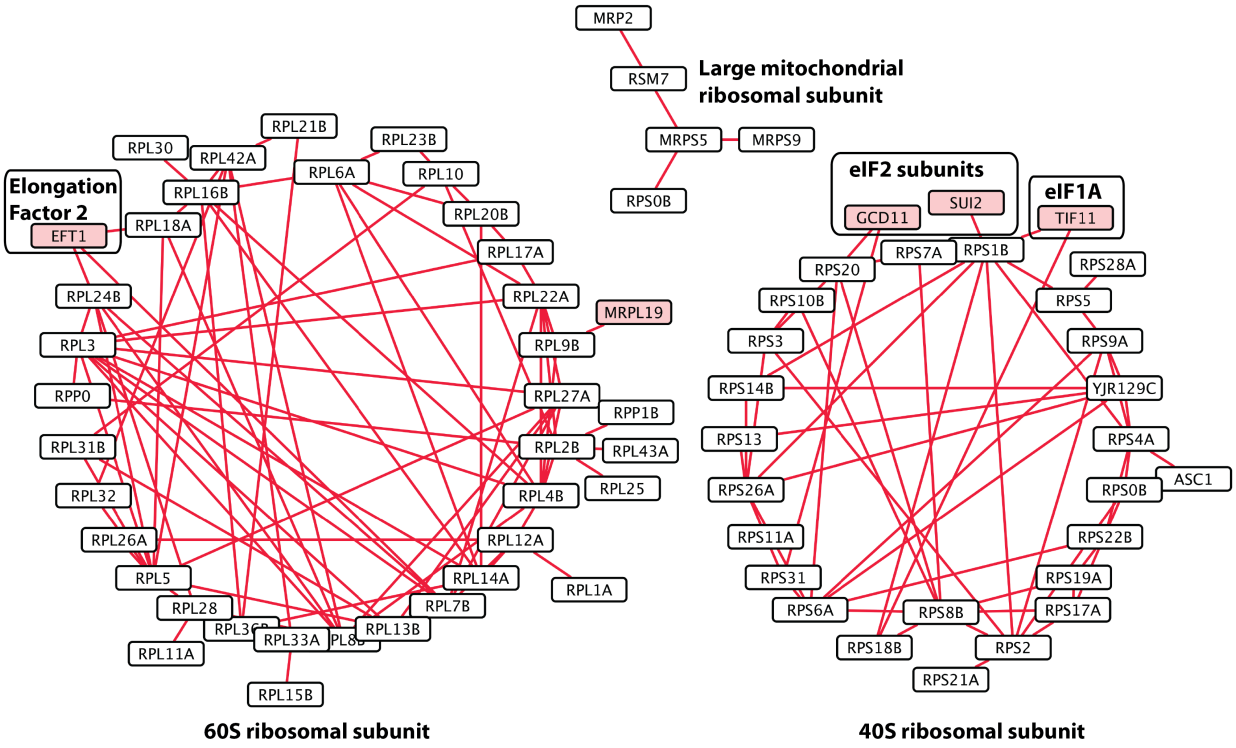

Figure 2.5: Crosslinks validate interactions predicted in the clusters

Crosslinking allows confirmation of predicted binary interactions by providing biochemical evidence. This is a simplified representation showing only confirmed crosslinked binary interactions in red as multiple unique crosslinks between two proteins are reduced to one. Crosslinks were discovered between 94 or 124 proteins predicted in cluster 12, containing the ribosomal proteins. This allows discrimination between proteins and binders of the 40S, 60S and large mitochondrial ribosomal subunits.

2.4.5: Further predicted binary interactions can be confirmed by crosslinking analysis

Additional to the 535 proteins present in the 50 predicted clusters, the network of high-confidence interactions contains an additional 144 proteins connected by 187 predicted interactions. These are represented as 'unclustered proteins' in Figure 2.2. This sparse area of the network contains binary protein interactions such as heterodimers. XL-MS is also used here to validate these interactions. Nearest neighbors of the predicted clusters in the network are

also searched as potential additional subunits of the larger complexes. The results are summarized in Table 2.

Table 2: summarizing inter-protein crosslinks found in for predicted protein-protein interactions outside of network

	Number of proteins	# Inter-protein crosslinks	# Binary interactions
Binary interactions confirmed	25	53	21

Crosslinks confirm several two subunits complexes such as the Isocitrate dehydrogenase complex, Protein phosphatase type 2A complex, Ski complex, Eukaryotic translation elongation factor 1 complex, Acetolactate complex and the Fatty acid synthase complex (FAS).

2.4.6: Crosslinking validates predicted novel subunits and novel complexes

Some predicted interactions confirmed by crosslinking have not been described before. Two examples are described:

- CTR9 (G0S4P3) and G0SDA7 (SPT6) are predicted to interact in cluster 43 and are detected with two crosslinks. These proteins bind to the RNA polymerase II C-terminal domain (CTD) on the RNA II subunit RPO21 (Qiu et al., 2012) and are critical for the accuracy of transcription and the integrity of chromatin (DeGennaro et al., 2013). They have been shown to genetically interact but a physical interaction had not previously been detected (Kaplan et al., 2005). They subunits also coelute with the RPO21, the subunit of the RNA polymerase with the CTD (Figure 2.5A).
- Cluster 42, figure 2.5B, contains all subunits of the Eukaryotic initiation factor 2B (EIF2), a subunit of the eukaryotic initiation factor 2 (EIF2), and a further protein TSR1, which binds to pre-40S ribosomes to prevent the 60S subunit joining prematurely (Strunk et al., 2011). EIF2 binds to Met-tRNA to correctly assemble the start codon and the ribosome. EIF2B with is not known to directly bind the ribosome and is the guanine nucleotide exchange factor responsible for exchanging GDP for GTP on EIF2 and is not known to interact directly with the ribosome (Gordiyenko et al., 2014; Jivotovskaya et al., 2006).

  Interestingly, two subunits of Eif2B and the GCD11 subunit of Eif2 are found crosslinked to TSR1. All of these proteins coelute with the 40S ribosome and Eif2 subunits, were in

turn, also found to be crosslinked to the 40S ribosome (Figure 2.4). The coelution of EIF2B and its interaction with a ribosomal protein is unexpected (Gordiyenko et al., 2014). This may mean that EIF2 and EIF2b are both bound to the pre-40S ribosome and make a novel subcomplex in ribosomal maturation or function. This interaction can be modeling onto the surface of the ribosome (work still ongoing).

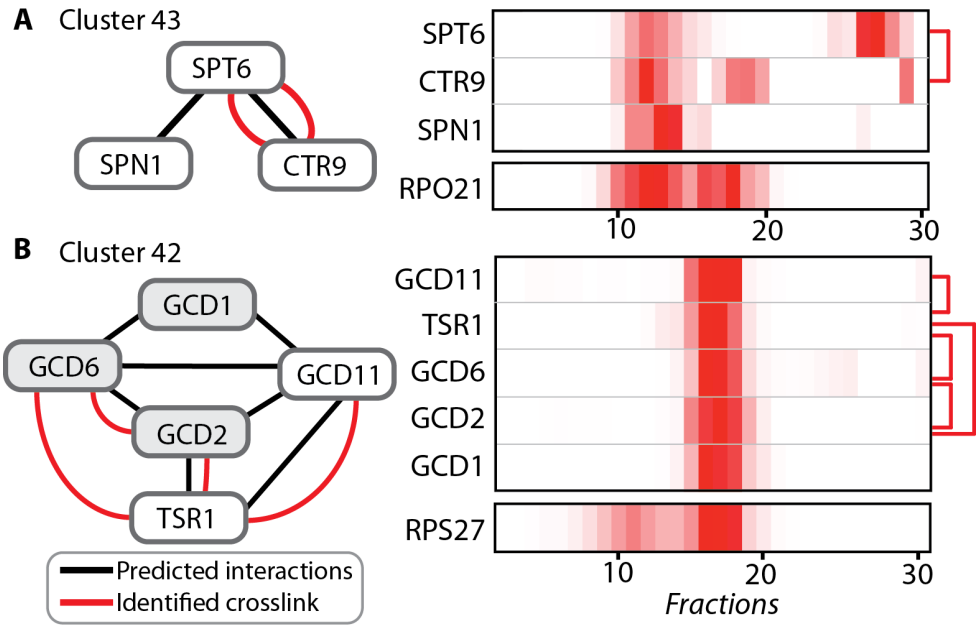

Figure 2.6: Crosslinks validate novel interactions in predicted clusters

Members of each cluster are shown connected by the predicted interactions (black) and identified unique crosslinks are mapped onto this network (red). The elution profiles of each of these proteins is also shown (right). **A)** Two unique crosslinks are identified between SPT6 and CTR9. These proteins coeute with the CTD-containing RNA pol II subunit RPO21. **B)** GCD1, GCD2 and GCD6 make up the Eif2B complex. These are found in a cluster with GCD11 of the Eif2 complex and the Pre-40S ribosomal binding protein TSR1. GCD6, GCD2 and GCD11 crosslink to TSR1 and together these coelute with the 40S ribosome (as demonstrated with the 40S protein RPS27).

2.4.7: Crosslinking analysis of most abundant proteins in each fraction reveals a network of unspecific protein-protein interactions

Additionally, to measure the overall specificity of the crosslinking analysis the most abundant proteins and protein complexes in each fraction were searched against each other to assess the number of spurious crosslinks (identifying true-negative interactions).

The relative abundance of individual complexes was calculated by averaging the intensity of each complex subunit by its known stoichiometry (from orthologous complexes with crystallographically-determined structures) i.e. the intensity for a homodimer is divided by two and those for a hetero-trimer are summed and divided by three. This is not possible for many proteins that have no known higher order assembly.

Proteins or complexes predicted to make up >2% of the MS intensity in each fraction, after normalization, were searched against each other for crosslinks. In total 1272 crosslinks were identified for 543 protein-protein interactions not present in the predicted network. No unexpected crosslinks were discovered between predicted complexes that elute together in the same fraction e.g. fatty acid synthase and the ribosome in Fraction 9, or the exomer and the Proteasome core particle in Fraction 20. This confirms that this XL-MS pipeline does not identify spurious interactions between highly abundant complexes.

Of the 1272 crosslinks detected outside the network, 85% of these involve 42 'hub' proteins; proteins that crosslink to ≥10 other proteins (see Appendix VI). These hub-proteins are predominantly heat-shock proteins and long disordered proteins, known to bind many other proteins and are common contaminates in protein-protein interaction studies.

Unexpectedly among these 'hub' proteins is Zuo1 a ribosome-associated protein not found in the predicted network, it is found crosslinked to 14 members of the 60S ribosome and the ribosome associated Elongation factor, EFT1, which is strong evidence for its membership of that complex. Also, acetyl-CoA carboxylase (Acc1) is found crosslinked to 23 other proteins, 15 of which are annotated as heat shock proteins or chaperones. Acc1 is a cellular carboxylase which produces malonyl-CoA from acetyl-CoA, the main feedstock for fatty acid synthase (discussed later). It has no known function as part of a protein complex but it has been found to bind chaperones in other protein-protein interaction studies (Brownsey et al., 2006; Gavin et al., 2002).

Table 3: Summarizing number of inter-protein crosslinks found between abundant proteins (excluding those already described from the network analysis)

	Number of proteins	# Inter-protein crosslinks	# Binary interactions
43 'hub' proteins (each binding 10 or more other proteins)	87	1137	461
Other	99	133	83
TOTAL		**1270**	**544**

Aside from crosslinks involving these 'hubs' there reamins 133 crosslinks representing 83 interactions not previously been described in the network. The majority of these add additional subunits to already predicted complexes which were missed due to their complex elution patterns.

- RPS23 is crosslinked to four 40S subunits and the elongation factor EFT1 confirming its membership of the 40S ribosome.
- G0S5N9, a protein that shares weak homology to 60S ribosomal protein L1 is crosslinked to three 60S ribosomal subunits.
- RPS35 is found extensively crosslinked to 8 members of the 60S ribosomal subunit.

All detected crosslinks are summarized in Appendix E.IV.

2.4.8: Negatively stained micrographs reproduce protein complex relative abundance and allow identification of complexes amenable to structural studies

To complement the prediction of native complexes we sought to analyze them structurally by EM in a high throughput manner directly from the SEC fractions. This, along with the XL-MS data provides information on topology of these complexes, which has not been possible previous high-throughput studies (Havugimana et al., 2012).

We developed a method to correlate MS-generated elution profiles of protein complexes with their corresponding structural signatures (shapes) as observed in negatively-stained electron micrographs using electron microscopy.

First, the 30 SEC fractions were visualized by negative-stain electron microscopy directly from the column (Figure 2.7 and 2.8A). As expected, particles are larger and much less concentrated in the early fractions and are much smaller and more concentrated in the later fractions. To investigate if there is a significant bias in which proteins can bind to the EM grid

(and can therefore be observed) the number of particles in the images was correlated with the concentration of protein as determined by BCA. Total numbers of particles in each fraction produces a good estimation of protein abundance and correlates well with protein concentration calculated by BCA assay ($R^2 = 0.84$) (Figure 2.8B).

To assess the relative abundance of particular species on the grids, 2D classification of structural signatures was employed among the fractions and clustered them by correlating their pixel intensity in an all-vs-all cross-correlation approach (see materials and methods). We were able to cluster re-occurring structural signatures in consecutive fractions to generate elution profiles. For many fractions this provided clear class averages of reoccurring structural signatures.

This correlative approach is only amenable to the more abundant particles, as there must be enough of each structural signal in consecutive fractions to generate elution profiles. As an example, in fraction 18 there were 490 proteins identified but when intensities are normalized by complex membership, only 19 proteins/complexes are predicted to comprise >2% of the particles on the grid (Figure 2.7). It is likely that the most abundant structural signatures will correspond to these most abundant particles identified by MS.

The corresponding 10 most abundant EM class averages in Figure 2.7 (consisting of >50 particles averaged from 10 images of this fraction) are the most abundant structural signatures in this fraction. The class averages of the 40S ribosome can be identified by comparing the image to known class averages from other studies (Gilbert et al., 2007) (Figure 2.7).

By correlating structural signatures across fractions with the predicted complexes, we were able to recognize other known protein complexes from other studies, including the proteasome with its top and side-views (Knispel et al., 2012), the 60S ribosome (Bradatsch et al., 2012) and the fatty acid synthase enzyme (Boehringer et al., 2013) (Figure 2.8C). Further fractionation or acquisition of larger-scale datasets will allow identification of previously unknown 2D class averages that are suitable for structural study (see Figure 2.11).

The method is complicated by certain protein complexes producing quite different structural signatures depending on how they fall on the grid as can be seen in the example of the proteasome core particle in Figure 2.8. To address this we plan to use electron tomography to match 'side views' with 'top views' of complexes and this will allow characterization of unknown protein complexes.

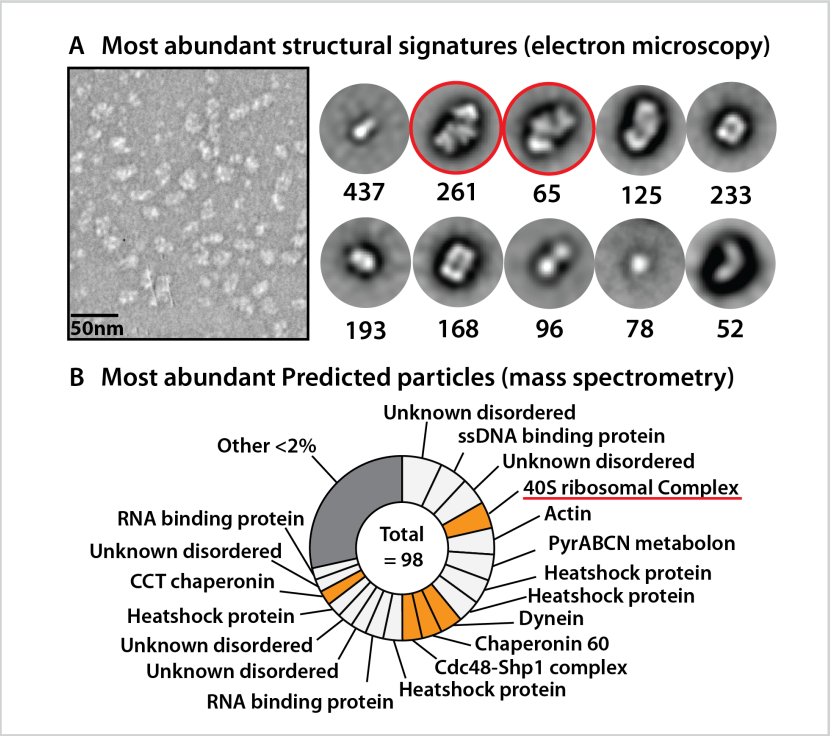

Figure 2.7: MS intensity of complexes and EM Class averages from fraction 18

A) An electron micrograph of fraction 18 is shown with the class averages comprising >50 particles from the 10 images analyzed. The 40S ribosome is one of the most abundant complexes in this fraction by MS and the classes averages belonging it are highlighted in red (recognizable from previous studies (Gilbert et al., 2007)).

B) MS intensities are normalized to identify the most abundant complexes. The relative abundance of individual complexes was calculated by averaging the intensity of each complex subunit by its known stoichiometry. Those in orange had stoichiometry information and could therefore be normalized. In total there are 98 normalized particles and 19 have >2% of the overall MS intensity. These abundant proteins/complexes are assumed to correspond to some of the structural signatures observed in (A).

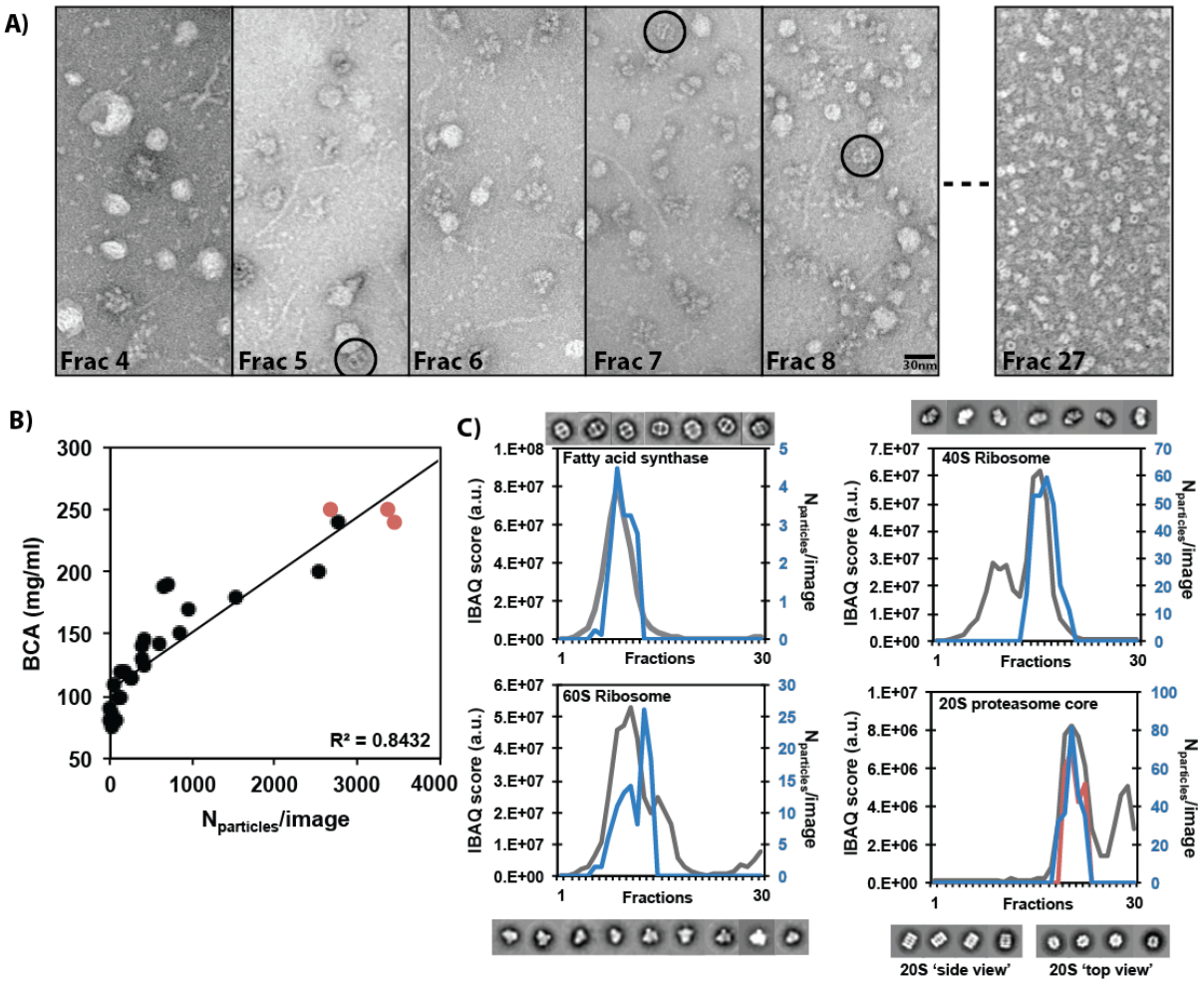

Figure 2.8: MS elution patterns correlate with structural elution patterns for known complexes

A) Negative stained micrographs were generated from each of the 30 fractions. Particles of FAS are highlighted with circles. B) Correlation of the number of particles picked per image and real protein abundance as measured by bicinchoninic acid assay (BCA). Points in red are dubious due to saturation of the assay. The R^2 is 0.84 C) Specific consecutive fractions with class averages of the proteasome (Knispel et al., 2012), ribosome (Bradatsch et al., 2012; Gilbert et al., 2007) and fatty acid synthase (Boehringer et al., 2013) are plotted with the normalized MS elution profiles in these fractions. The proteasome has two 'views' the 'side-view' and the 'top-view', elution profiles of these structural signatures these are plotted as blue and red respectively. [Samples were imaged and analyzed by Panagiotis kastritis]

2.4.9: Discovery of a structurally unknown catalytic reaction within Fatty acid Synthesis after single particle EM

To test out the hypothesis that *C. thermophilum* proteins are particularly amenable to structural study due to its thermophilicity, we decided to solve a high-resolution structure of fatty acid synthase from its SEC fractions. The class averages of FAS identified using the correlative analysis between EM and MS profiles (figure 2.8) allowed manual picking of FAS particles from cryo-EM images.

Fractions containing FAS were pooled and concentrated and negative-stain electron microscopy was performed to ensure structural integrity of the molecule (Figure 2.9A). After vitrification of the sample, particles of FAS were apparent, even at 1.2 μm of defocus, with a calculated CTF with several rings (Figure 2.9B). Gold-standard FSC (FSC=0.143) indicated that resolution of the map reaches ~4.7 Å (Figure 2.9C). The overall map of FAS is nicely homogeneous in resolution (figure 2.9D); other fungal FAS EM structures do not resolve the top part of the molecule (Boehringer et al., 2013; Gipson et al., 2010). The fit of the helices in the central wheel of the molecule is also apparent and demonstrates the quality of the density map when compared with the solved FAS crystal structure from *Thermomyces lanuginosus* (PDB code = 4V58) (Figure 2.9E).

Additional novel density is observed inside the molecule, albeit at a lower resolution. This density corresponds to the acyl carrier protein (ACP), after supervised fitting with CHIMERA. The ACP domain carries the growing fatty acid though the reaction chamber to all the active domains. This ACP conformation is orientated so that the growing acyl chain is contacting the enoyl reductase domain ERD, which catalyzes one of the intermittent steps in growing the ACP bound acyl chain (Jenni et al., 2006). This conformation of ACP was not imposed, but was emergent from the fit. Careful inspection of the fitted molecular models to the EM map, highlights a density that may correspond to a buried acyl chain, carried by ACP and found inside the catalytic site of the ERD (Figure 2.9F). Note that this catalytic intermediate of Fatty acid Synthase cycle has not been previously resolved and is the first EM structure with apparent density of the ACP domain. Even in the yeast crystal structure, where the ACP is apparent, it is interacting with the ketose synthase domain, instead.

The modeled interaction was subjected to energy calculations using HADDOCK (Kastritis et al., 2014) in order to understand determinants of the binding. Interestingly, electrostatic interactions are strong, indicating that ACP is driven by electrostatic complementarity to form an interface with the ERD (Figure 2.9F&G). Although the van der

Waals interactions are pretty weak, indicating that shape complementarity or buried surface area are not sufficient for protein-protein binding, modeling K_d based on van der Waals interaction energy, shown to linearly correlate with experimentally-measured protein-protein binding affinities (Kastritis et al., 2014), predicts this interaction to be in the low mM range (Figure 2.9H).

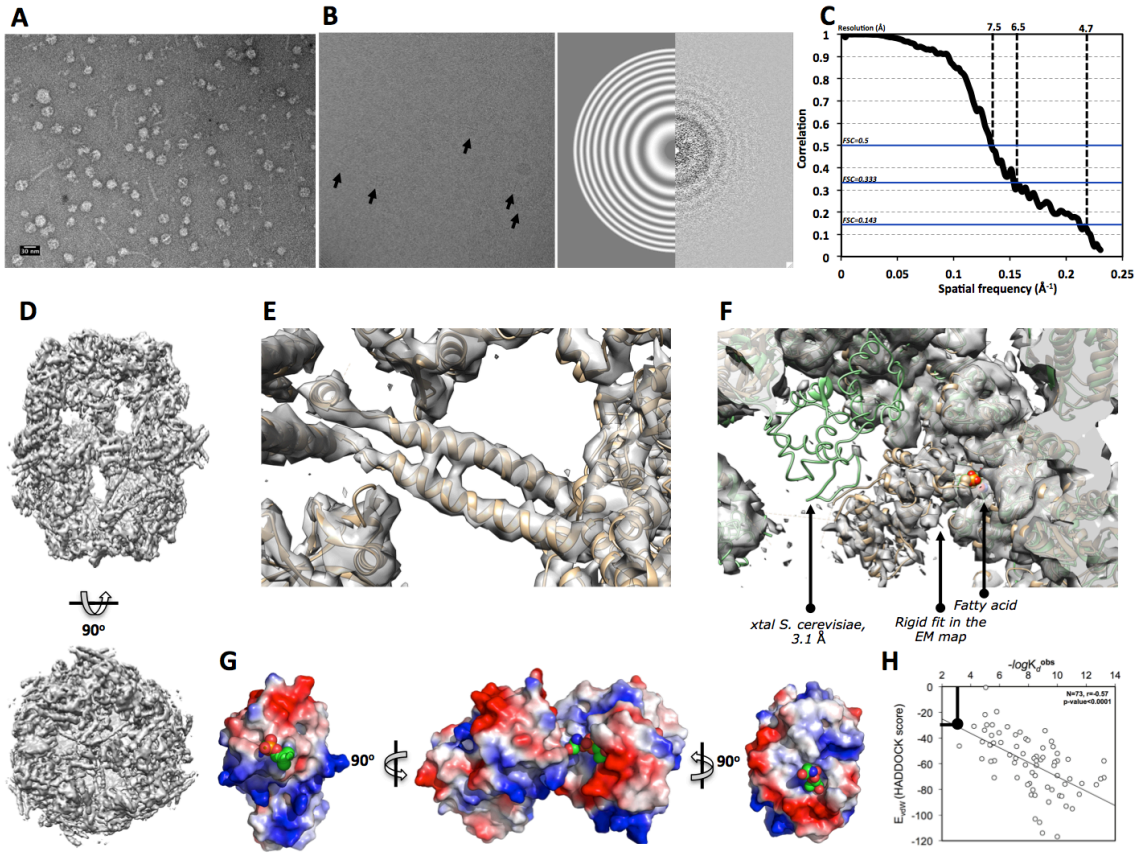

Figure 2.9: Electron microscopy of complex mixtures: A transient interaction in fatty acid synthesis as an example.

(A) Electron micrograph of pooled fractions with ~50% enrichment of FAS (B) Cryo-EM micrograph of fraction in (A), showing that FAS particles are visible, even at 1.2 um of defocus; calculated CTF with several rings show that structures can reach high resolution; (C) Gold-standard FSC (FSC=0.143) indicated that resolution of the map reaches ~4.7 Å; calculated resolutions with FSC=0.25 and 0.5 are also shown; (D) Overall electron density of the FAS complex, indicating isotropic resolution; (E) Pitch of helices in the central wheel is apparent (F) Low-resolution density is resolved for a transient intermediate in FAS catalysis (ACP protein interacting with the catalytic site of the ERD) (G) Interacting domains shown in (F); Calculated

electrostatic surface potential maps show complementarity indicating importance of electrostatic interactions in protein-protein recognition. (H) Predicted binding affinity of ACP and ER according to the model by Kastritis et al (Kastritis et al., 2014), indicating a transient interaction in the high μM to low mM range; Open circles in plot show experimental vs predicted binding affinity for 74 «near-rigid» protein-protein complexes collected from the literature (Kastritis et al., 2014). [This structure was solved and molecular analysis was performed by Panagiotis Kastritis]

2.4.10: Identifying a novel Fatty acid synthase binder

The solved FAS structure contained some unexplained density on the periphery that is not part of the known α6β6 complex (Figure 2.10A). The raw images averaged to solve the structure were manually inspected and some were found to have an additional particle bound to the side of the FAS particles (Figure 2.10B). As our sample is directly extracted from the native source this extra density likely represents a native binder of Fatty acid synthase.

These images of FAS with this extra density were picked for class average generation but structural solution was not possible due to its flexibility (figure 2.9B). FAS was not predicted to be part of any cluster in the network analysis so no potential binders were identified. It is therefore necessary to use another method to enrich for this interaction.

We hypothesized that the binder of FAS was a metabolon (temporary structural-functional complex formed between sequential enzymes of a metabolic pathway) so we searched in for related proteins in the three fractions pooled to solve the structure. Indeed Acc1, acetyl-coA carboxylase that produces the feedstock methlmanoyl-CoA for FAS (Tong, 2013), is found in these fractions with relatively low abundance but due the bulk of its elution being in a second peak receives a low cross-correlation score with FAS. We attempted to pull down and crosslink the naturally biotinylated Acc1 using a streptavadin column to enrich for this potential FAS interaction.

Unexpectedly the protein that pulled down and crosslinked to FAS was instead G0S1E6, a Biotin-dependent carboxylase subunit similar to both the β subunit of methylcrotonyl-CoA carboxlase (MCC) and the propionyl-CoA carboxylase (PCC) (also naturally biotinylated) (Figure 2.9C).

The known structures of MCC and PCC are α6β6 dodecamers that make a 1.4MDa complex using the *C. thermophilum* proteins (Huang et al., 2010, 2011). Both of these enzymes

branch short carbon chains attached to coenzyme-A (Co-A) and share a high sequence identity with each other. This carboxylase β-subunit is in turn crosslinked to G0S1X1, which is a gene-fusion of the α-subunit of the carboxylase and an acetyl-CoA hydrolase (ACH). ACH is characterized as a CoA transferase that cleaves the high-energy thioester bond between the acyl chain and CoA (Fleck and Brock, 2009; Orlandi et al., 2012).

The crosslinks from this carboxylase β-subunit to the FAS α-subunit position it at the entry tunnel to the FAS reaction chamber. Interestingly the crosslink from the carboxylase β-subunit to the FAS β-subunit is to the ACP, which binds the growing fatty acid chain and transports it to all the reaction domains within the chamber. The elution of this carboxylase β-subunit has a minor peak overlapping with the leading edge of the FAS peak, suggesting that this interaction has coelution evidence (Figure 2.9D).

Additionally Acc1 was found crosslinked to the β-subunit in the biotin pull-down experiment. To deconvolute whether all of these proteins are bound with FAS a further fractionation this experiment should be performed on the biotin pulldown sample and the fractions investigated by MS and EM.

2.4.11: Potential biological significance of FAS-carboxylase interaction

All of these enzymes crosslinked in the biotin pulldown are involved in acyl CoA metabolism and could therefore be part of a metabolon bringing together these activities. The only protein that directly crosslinked to FAS was the putative β-subunit of the MCC or the PCC. Both the MCC and the PCC are biotin-dependent carboxylases that add methyl groups to acyl-coA chains to branch them. They are currently characterized as having roles in the degradation of amino acids.

In *S. cerevisiae,* FAS is a cytosolic enzyme and both MCC and PCC are mitochondrial enzymes, though they are briefly in the cytoplasm to become biotinylated before transport to the mitochondria. Ach1 (fused to the α-subunit of the carboxylase in *C. thermophilum*) is found predominately in the mitochondria in *S. cereviaise* but also in cytoplasmic foci (it does not have an ortholog in higher eukaryotes)(Buu et al., 2003).

Biotin-dependent carboxylases have roles in the biosynthesis of fatty acids and branched-chain fatty acids (Diacovich et al., 2002; Gago et al., 2011; Kaneda and Smith, 1980) and MCC and PCC like enzymes have been implicated in the production of branched chained fatty acids in *Mycobacterium tuberculosis* (Ehebauer et al., 2015). Evidence for *de novo* production of branched-chained fatty acids in eukaryotes is sparse though two long-chain fatty

acid elongation enzymes ELO-5 and ELO-6 have been implicated in their de novo synthesis in *Caenorhabditis elegans* (Kniazeva et al., 2004).

It is therefore my hypothesis that the interaction between FAS and a branching Biotin-dependent carboxylase has a role in providing branched feedstock to the FAS to produce branched fatty acids in a novel metabolon.

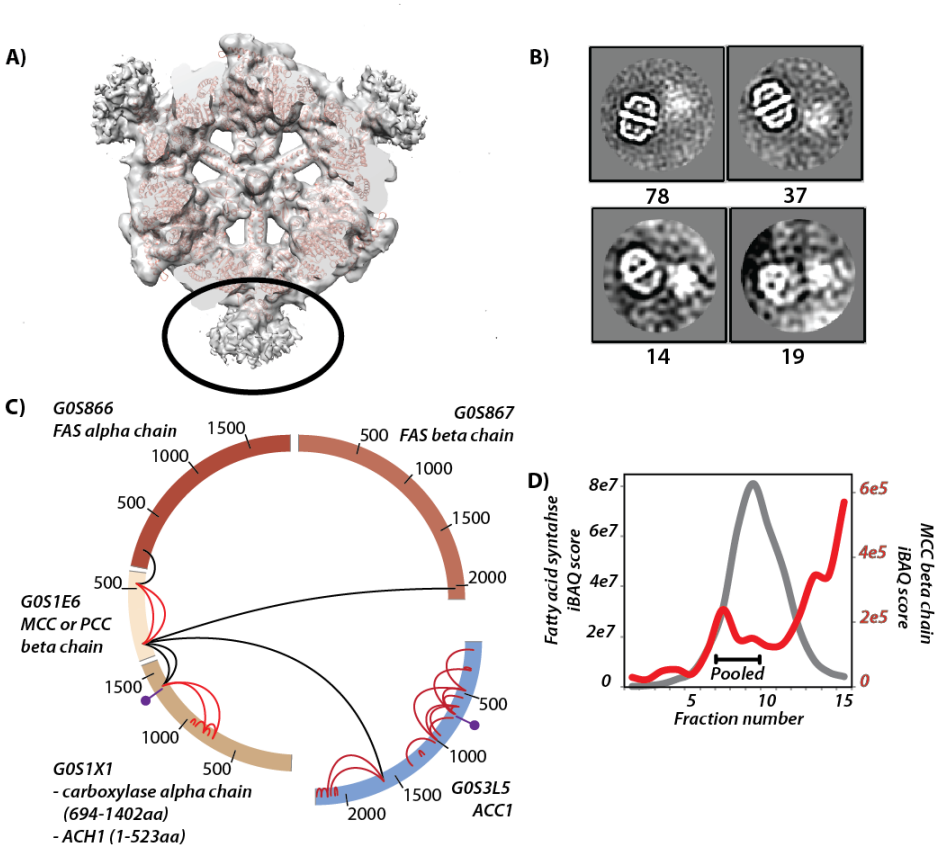

Figure 2.10: Electron microscopy of complex mixtures: A novel interaction in fatty acid synthesis as an example.

(A) Circle indicates extra poorly-resolved density on the periphery of the FAS structure. **(B)** FAS particles that have an extra binder are manually picked and reclassified. The interaction between FAS and the binder appears to be flexible and could not be satisfactorily classified. **(C)** Circular representation of the XL-MS Biotin pull-down. Interprotein crosslinks are shown in black, intraprotein crosslinks are shown in red. Purple lollipops identify the predicted sites of biotinylation. The β-subunit of the carboxylase (similar to both MCC and PCC) crosslinks to the flexible ACP domain of the α-subunit and to the FAS β-subunit. Assuming a 30 Å distance restraint, both crosslinks satisfy binding to the yet undefined density. Additionally, the β-subunit of the carboxylase crosslinks to the α-subunit of the MCC or PCC complex. A third

interprotein crosslink connects the CoA carboxylase β-subunit to the ACC1 protein, forming a potential metabolon. **(D)** Elution profiles of the FAS and the β-subunit of the MCC are shown in the first 15 fractions. (Biotin pulldown and crosslinking was performed by Thomas Bock)

2.4.12: Simplifying cell lysate for the solution of further novel protein complexes

To extend the method to identifying structural signatures to previously unknown, less abundant complexes it is necessary to simply the lysate before correlating the MS and EM, especially in the later, more complex fractions. Anion exchange was used to simplify the lysate by collecting the flow-through (i.e. proteins that did not bind to the column) (Figure 2.11A). The running buffer is not changed from the previous experiments so the complexes should be unaltered. This simplified lysate is then separated on the analystical column as before.

As a proof of principle experiment, fractions 22-28, were investigated by MS (Figure 2.11A). Only 9 proteins were detected as eluting across >1 fraction (identified by the less sensitive HCT mass spectrometer). Crosslinking studies are underway on this simplified mixture to further group these proteins that exist as complexes.

Images were collected from fraction 25 to check the complexity and to observe if any particles look susceptible to structural investigation. Indeed, there are at least 6 interesting structural signatures (Figure 2.11B). It is unclear which of these structural signatures are actually the same particles in a different orientation. We plan to deconvolute this by using electron tomography to observe these particles form different angles and create initial 3D reconstructions in a simultaneous manner.

This experiment needs to be repeated to generate true elution profiles as in Figure 2.8C. We also have preliminary data that the peak in the elution pattern in Figure 2.11A is found also when the flow-through from a cation exchange column subjected to SEC. The cation exchange flow-through may be depleted in a different group of particles allowing us to narrow down the possible combinations of proteins identified by MS and their corresponding structural signatures.

These experiments are on-going and so far show great promise that novel structures for some of the identified proteins can be solved using the same workflow as for our FAS structure. The designed method is generic and can be extended to other mixtures of similar complexity to identify complexes and to solve their structures.

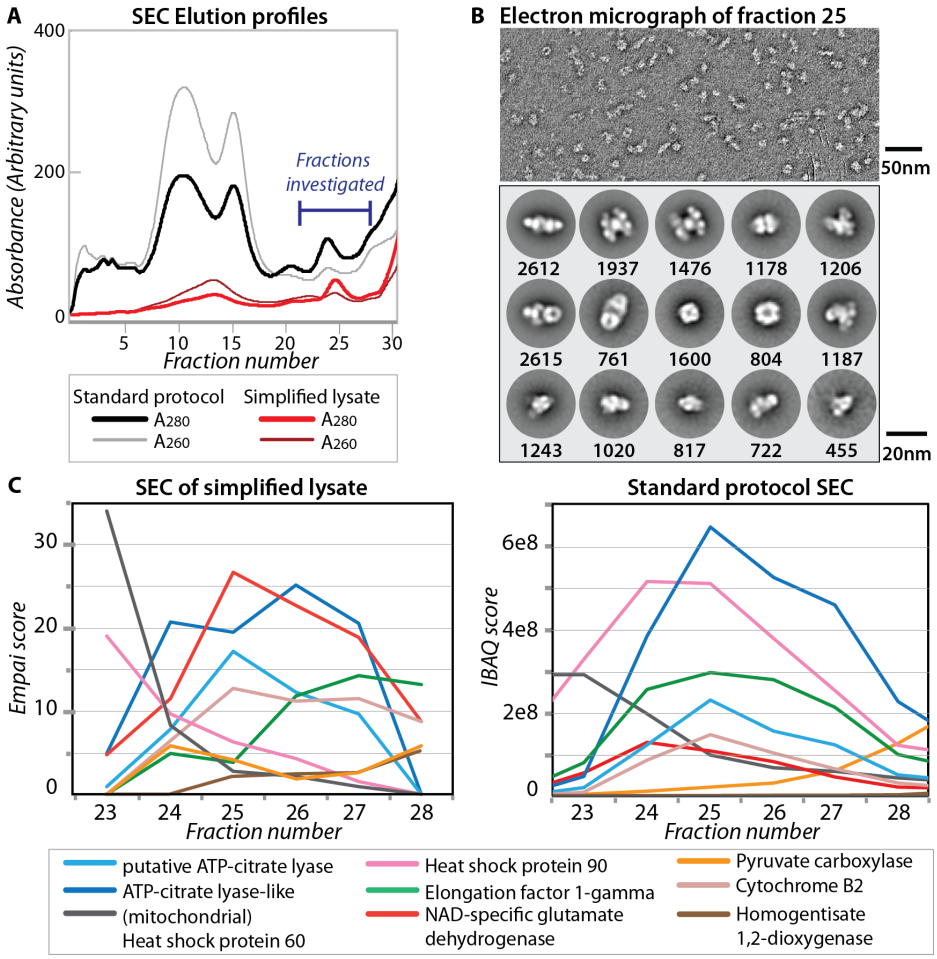

Figure 2.11: Simplifying complex mixtures allows for structural analysis of novel complexes

A) Elution profiles of the High MW proteome from SEC using the standard protocol compared with the High MW proteome depleted using anion exchange. Fractions carried on for further study are indicated. **B)** Fraction 25 is investigated using negative stain microscopy and the most abundant class averages are shown. The number of particles contributing to each class average is also indicated. **C)** MS data of fractions investigated rom the depleted lysate compared with those from the original experiment. Only proteins identified in >1 fraction in the simplified lysate experiment are shown. ATP-citrate lyase, the most abundant protein, is a cytosolic enzyme that catalyzes the production of acetyl-CoA from citrate which is then used for fatty acid biosynthesis through the fatty acid synthase (Chypre et al., 2012).

2.5: Conclusion and Perspectives

This project has demonstrated the wealth of information that can be generated by a simple one step fractionation technique with the correct combination of complementary technologies.

The assignment of many of these proteins to complexes allows for correct ortholog annotation. 130 of the 443 proteins assigned to clusters are currently annotated as 'putative uncharacterized protein' in Uniprot (www.uniprot.org). The newly mapped complex subunits will inform future structural studies using proteins from this organism, that have so far been hampered by ambiguity in their annotation.

We have shown that crosslinking-MS and electron microscopy can be applied to protein mixtures, previously thought to be too complex. This technology can facilitate future structural studies where proteins are difficult to purify due to stability issues.

The discovery of a novel binder of FAS shows the benefits of working with material from the native source as it can provide interesting biological insights and further study may identify how branched fatty acids are produced in eukaryotes for the first time. Interestingly here, EM has been used as discovery tool (serendipitous as it was).

Figure 2.11 demonstrated how many more complexes are amenable to solution by this method. Additionally Figure 2.8 showed that the simple one step purification could allow solution of the ribosome structure and potentially structural study of the 40S-TSR1-EIF2-EIF2B complex predicted by the coelution and crosslinking (discussed in section 2.4.6).

Further work can be done using these techniques on cell context dependent complexes like those described in Kristensen *et al* (2012). Additionally this technique can be used to assess the importance of post-translational modifications in the formation of complexes, some preliminary work in this direction has already been completed (Appendix A.VI).

In summary, our high-resolution EM structure of FAS and its novel interactor adds further evidence to the hypothesis that the proteins from this thermophilic eukaryote are particularly suited for structural study. When the list of complex membership is finally refined we will have produced the first list of correctly annotated complex subunits from *C. thermophilum* allowing it to become a true model organism for structural biologists.

2.6: Materials and Methods

Culture growth. Add spores to the 200ml of LB media and grow at 60C with 10% CO2 and gentle agitation 60rpm. After growing for 4 days split culture again into 4 x 400ml in 1litre flasks. And grow for two days. 4 grams of healthy fluffy culture were harvested (4 flasks).

Lysis. Equal volume lysis buffer (100mM HEPES pH8.0, 95mM NaCl, 5mM KCl, 1mM MgCl2, 1mM EDTA, 1mM DTT, 10ug/ml DNAse, pefabloc 2.5mM, E-64 40uM, Bestatin 130uM, Aprotinin 0.5uM, Leupeptin 1uM, pepstatin A 60uM, EDTA 1mM) plus additives is added to cells and mixed thoroughly. The cells are decanted into 15ml falcon tubes and lysed in a Fastprep FP120 at 4ºC (6.5mps for 40 seconds x 3 with 2 minutes rest on ice). Tubes are centrifuged for 30 seconds to pellet beads. The supernatant is untracentrifuged for 45min at 100,000g, the lipid layer is removed and the clarified cell lysate is retained concentrated by centrifugation fractionation (100,000 MWCutOff).

Fractionation. 100µl of lysate (~30mg.ml) is loaded onto a Biosep SEC-S4000 (7.8 x 600) Size exclusion column on an EttanLC (HPLC) system. Running buffer; 100mM HEPES pH7.4, 95mM NaCl, 5mM KCl, 1mM MgCl2, Flow rate 0.25ml/min, fractions 250ul. Fractions are stored at - 20ºC.

Sample preparation for MS. Proteins in fractions were denatured in 4 M urea buffer containing 0.2% (w/v) Rapigest® detergent. Carbamidomethlylation, enzymatic digest and peptide purification was performed as previously described (Ori et al., 2013). Purified peptide mixtures were lyophilized in a vacuum concentrator and stored at -20 °C until further use. Thirty fractions form one run were pooled and dimethyl labeled (heavy) and these were added to all fractions dimethyl labeled (light) in a 1:1 ratio of peptide abundance. Labeling was done in solution as described (Boersema et al., 2009).

MS of fractions and data analysis. The nano-LC system was connected to a LTQ Orbitrap Velos Pro instrument (Thermo Scientific) that was operated in data-dependent mode. The collision induced dissociation method used one survey MS scan followed by up to 20 fragmentation scans (TOP20) of the most abundant ions analysed. Peptides were separated with a BEH300 C18 (75 µm x 250 mm, 1.7 µm) nanoAcquity UPLC column (Waters) using a stepwise 45 min gradient from 3% to 85% (v/v) acetonitrile in 0.1% (v/v) formic acid at a flow rate of 300

nl/min. The LTQ-Orbitrap Velos Pro instrument was operated in data-dependent mode. Parameters for the CID based method used one survey MS scan acquired in the orbitrap followed by up to 20 fragmentation scans (TOP20) of the most abundant ions analysed in the LTQ. Only charge states of two and higher were allowed for fragmentation. Data analysis was performed with the Maxquant suite, the following filters were applied to the search results: MS1 tolerance window 20 ppm; fragment mass tolerance 0.5 Da; miss cleavages 1. Peptides were filtered for 1% FDR using target-decoy reverse database search. Protein intensities for each fraction are exported as iBAQ scores (sum of intensity from detected peptides for each protein, divided by theoretical number of measurable peptides for this protein).

Prediction of coeluting proteins. Proteins, which are found with more than one peptide, in >1 biological replicates and in >3 consecutive fractions are retained for further analysis. Data in each fraction was averaged across biological replicates and each chromatogram is correlated point by point with each other chromatogram to generate a Pearson Correlation Coefficient. A cross-correlation score was then derived as a threshold for protein co-elution discovery, calibrated by an assembled benchmark of protein complexes with known molecular structures (Table SX). Algorithm is available from the authors upon request.

Ortholog mapping. Orthologous protein groups within *C. thermophilum, C. globosum, Neurospora crassa* and some 20 related fungi (pezizomycotina) are generated by identifying triangles of best bidirectional BLAST hits between species and then connecting triangles if they share two proteins. These pezizomycotina orthologuous groups are mapped to fungal orthologous groups (fuNOG) from eggNOG (Jensen et al., 2008). If FuNOG groups have at least two members of the pezizomycotina orthologous groups in them any yeast protein in the same group is assigned as ortholog of the *C. thermophilum*.

Machine learning method and cluster prediction

A gold standard set of interactions was assembled from the PDB and AP-MS data from orthologes from *S. cerevisiae* containing 3045 TP interactions, and 57,149 TN. The Gold standard was split for the machine learning into three sets for learning, validation and testing (70/15/15 respectively). The data were subjects to a random Forest implementation was applied to weight the CCC score, STRING scores and FIST scores. The Model yields an OOB error rate of ~8% (quality scores attached in the random forest subfolder). Before learning the CCC

score is filtered for scores above 0. Scores > 0.5 are fitted to a Gaussian curve and interactions above 2sd from the mean (random forest probability >0.84) are retained (736 proteins connected by 6146 interactions). The generated network is visualized using the Cytoscape software suite and the ClusterOne cluster growth algorithm is used to predict the clusters with the following paramers: size = 3, density = 0.3, edge weights = Random forest probability, penalty = 2, haircut = 0.3, mutlipass = all nodes. 55 clusters are produced and reported.

Crosslinking of *C. thermophilum* lysate. 10g of culture was lysed and concentrated and loaded onto a Hiprep sephacryl S-300 HR SEC with 0.2ml.min and collecting 300ul fractions. Consecutive fractions are pooled to achieve 1mg and this is then concentrated to 100ul and crosslinked with a 1mM DSS lysine specific crosslinker. Crosslinked samples are digested with the above protocol and the digested peptides are separated on a superpose-6 SEC column to enrich for crosslinked peptides.

Readjustment of FDR prediction

We noted, that the standard software pipeline xQuest used in our searches, is dedicated towards searches against protein assemblies containing low numbers of protein subunits. We therefore got very high FDR predictions for our data set on a complex mixture using the standard pipeline. As chemical crosslinking on such complex mixtures has not been performed before we needed a novel method for generating error models. For FDR calculations, a subset of homomeric and heteromeric complexes from the PDB benchmark was used. From these, we can assess which crosslink scores can be trusted. In order to re-assess and adjust the FDR of our crosslinking data set, we used the fulfilment of spatial restraints for our crosslinks. We used CT molecular models of 31 highly conserved protein assemblies to readjust the FDR prediction of xQuest searches for our complex data set. Spatial restraints of identified crosslinks with an ld-score cut-off of 20 were calculated on those CT molecular models using Euclidean distances calculated by xWalk software (PMID: 10.1093/bioinformatics/btr348). A commonly accepted cut-off of 30 Å was applied. Based on those spatial restraints, all were satisfied for a minimum ld-score of 23.5.

Protein complex assignment using Protein Data Bank. Each of the 1450 proteins found in total in all 3 biological replicates were manually submitted to the NCBI BLAST server (Fernández-Recio et al., 2004). A search against the PDB was performed. A threshold of 50% of sequence

identity was assigned in the search for proceeding to visualize the biological assemblies. In case of multiple protein sequence hits, visual inspection of biological assemblies of all entries was performed. Decision on the assembly was taken after back-BLASTing the rest of the subunits of the PDB structure to the *C. thermophilum* proteome, available in Uniprot (Bairoch et al., 2005); Complexes having all subunits mapped to *C. thermophilum* high-molecular weight proteome are included in the list. Modeling of all proteins and their known complexes from the high-molecular weight fractions using I-TASSER and subsequent molecular modeling/docking. Here describe how I generated all protein models with ITASSER and how the homomers were constructed for subsequent calibration of the ld-score

Modeling of protein interfaces using crosslinking data. The HADDOCK webserver (guru access) was used (de Vries et al., 2010). Missing side-chains were properly built and the interface of the complex was optimized using the OPLS force field [10.1021/ja00214a001] and non-bonded interactions were calculated using a cut-off of 8.5 Å. Electrostatic energy (E_{elec}) was calculated by a shift function, while a switching function (between 6.5 and 8.5 Å) was used for the van der Waals energy (E_{vdw}). Desolvation energy is calculated by implementing empirical atomic solvation parameters (Fernández-Recio et al., 2004). All calculations were performed with HADDOCK version 2.1/CNS version 1.2 (Brunger, 2007). XL-MS data were implemented as interaction restraints. They were set to have an effective (and maximum) Cα-Cα distance of 35.2 Å, whereas the minimum distance was only defined by energetics. This distance was selected due to the analogy of the maximum Cα distance that the DSS crosslinker may have, when crosslinking Lys side-chains. In addition, sequence conservation was calculated using the CPORT server (de Vries and Bonvin, 2011) by default and were included as active residues in the docking procedure; surrounding residues (5 Å distance) were considered passive (de Vries et al., 2010). For docking calculations the standard HADDOCK protocol was used (Dominguez et al., 2003) with minor modifications: 10,000 structures were generated in the first iteration (it0; randomization and energy minimization step) instead of the default 1,000 to increase sampling. Standard HADDOCK scoring for it0 was applied to select the 400 top-ranking structures that are subsequently passed onto the next docking steps (semi-flexible simulated annealing and final refinement in explicit water). Clustering at 7.5 Å was finally performed.

*ct*FAS enzyme preparation and vitrification. 3 fractions derived from SEC enriched in *ct*FAS (according to quantitative MS data) were pooled and subsequently visualized for structural

integrity. *ct*FAS was ~50% enriched and overall protein concentration was determined ~40 ng/µL. Samples were then deposited on glow-discharged (60 sec) carbon-coated holey grids from Quantifoil®, type R2/1. A FEI Vitrobot® was used for plunge-freezing. In short, humidity was set to 70%, temperature to 293^{o} K, blotting and drain time to 3 and 0.5 sec, respectively. Sample volume applied was 3 µL and blot offset was set to -3 mm.

Image acquisition. The vitrified samples were recorded on a FEI Titan Krios microscope at 300 kV and automatic image acquisition was performed with FEI EPU software. Pixel size was set to 2.16 Å and a Falcon 2 GATAN camera was used in movie mode. The total number of frame groups was 7 and dose applied per frame group (e–/Å^2/sec) was set to 4/4/4/4/4/4/24, respectively. Total dose applied was summed to 48 e–/Å^2, but last frame was used only for particle picking. Magnification was set to 75000X, whereas defocus ranged between 0.6 and 3.0 µm. A total number of 13419 micrographs were acquired in 21 hours (1 frame/ 6 sec; 1 movie/42 sec).

Data processing and 3D reconstruction. Motion correction was applied to acquired micrographs leading to 1917 summed micrographs. E2BOXER was used for particle picking with a box size of 256X256 pixels. 7370 particles were selected from 1597 micrographs (4-5 particles/image). For CTF correction, CTFFIND was used. The RELION 1.2 package was then used for 2D class averaging, 3D classification and 3Dreconstruction of the density map. Briefly, 20 classes were set for 2D classification and, after visual inspection, 8 were selected for further processing that included 4898 particles. 3D classification was set to 5 classes, performed without imposing symmetry, using the cerulenin-inhibited yeast FAS as an initial model, low-pass filtered to 60 Å. 3933 particles were successfully classified in a single class and was subsequently used for reconstruction. 3D Reconstruction was performed using the same initial model but applying C3 symmetry and underwent 25 interactions, showing convergence. Dependency of the resolution according to masking was observed, but the default Gaussian mask from RELION 1.2 leads to a calculated resolution (Gold-standard FSC=0.143) of 4.7 Å.

Modeling of the ACP-enoyl reductase domain interaction. Additional density of ACP was observed close to the enoyl reductase (ER) domain of FAS; thus, coarse placement of the ACP was performed using CHIMERA and subsequently fitted to the density. Manual inspection of the fit agreed with the ACP orientation with the lipid-binding domain facing to the ER domain.

Energy calculations using the refinement webserver was performed, the domain interaction serving as input. Energy calculations were performed as previously described (Kastritis and Bonvin, 2010; Kastritis et al., 2014). Correlation of van der Waals energy with experimentally-measured equilibrium dissociation constants for known complexes is derived from Kastritis et al. 2014 (2014).

CHAPTER 3: PERTURBATIONS OF LIPID BIOSYNTHESIS PATHWAYS INDUCE SPECIFIC PROTEIN LOCALISATION CHANGES IN BUDDING YEAST

3.1: Abstract

Quantitative phenotypic profiling is a powerful tool towards linking phenotypes to genotypes to decipher novel gene functions and impacts on diverse cellular processes. High-content screens have allowed recording of phenotypes not previously identified by genetic interaction screens, which only record growth defects. In vivo screening of protein-lipid interactions is possible by genetically depleting classes of lipids, by perturbing their biosynthesis, and analyzing the localization of proteins allows for the identification of lipid dependent localizations.

Here we produced 13,507 cell lines that represent 648 GFP-tagged proteins imaged in 24 lipid perturbation backgrounds across the five main lipid biosynthesis pathways. The GFP tagged controls are compared with the proteins in the perturbed backgrounds to discover which lipid classes are structurally important in the eukaryotic cell. 243 proteins imaged were found to be miss-localized in at least one lipid perturbation background and 148 were miss-localized in >1 condition.

The dataset demonstrates the roles of lipids as structural molecules and suggests novel roles of lipids in the eukaryotic cell, which can be further examined.

3.2: Contributions

The dataset was designed and generated by Arun Kumar, a previous post-doc in the lab. I inherited it and was the main diver of the data analysis with help from Kenji Maeda in the manual hit calling and Karl Kugler and Sergej Andrejev in automated data analysis techniques.

3.3: Introduction

3.3.1: Lipids as structural molecules in the eukaryotic cell

Lipids are a major constituent of eukaryotic cells and lipidomic experiments have estimated that there are thousands of different types (Fahy et al., 2009; van Meer et al., 2008). They have three general functions; they are utilized as energy stores, their intermediates have important signaling roles as secondary messengers (Hannun and Obeid, 2008), and they have a vital

structural role forming the membranes which compartmentalize the cell and as scaffolds for protein localization.

Cellular membranes are a complex mixture of lipids but these can be broadly separated into three main classes; glycerolphospholipids, sphingolipids and sterols, with many sub-groups within these categories. There are large differences in lipid concentrations between different cellular membranes, and local concentration differences within lipid bilayers. The lipid composition of membranes provides characteristics that localize or bind certain proteins due to their thickness, fluidity, curvature and surface charge (van Meer et al., 2008).

The main constituents of eukaryotic membranes are the glycerophospholipids: phosphatiylcholine (PtdCho), phosphatiylserine (PtdSer) phosphatidylethanolamine (PtdEtn), phosphatidylinositols (PtdIns) and phosphatidic acid (PA). These have a huge repertoire of biophysical features due to their variable polar head group and the variable composition of their lipid acyl chains. PtdCho constitutes ~50% of the lipids in biological membranes with PtdEtn being the second most abundant. PtdSer, PA and PtdIns are present in much lower amounts. PtdSer is found primarily in the plasma membrane and is maintained on the cytoplasmic leaflet by flippases. PtdIns occur in comparatively low amounts but have vital roles in signaling and defining organelle identity by recruiting both soluble and membrane proteins. PtdIns(4,5)P2 is a hallmark feature of the plasma membrane (PM), phosphatidylinositol 3-phosphate (PtdIns3P) is a marker of endosomes and phosphatidylinositol 4-phosphate (PtdIns4P) is a marker of the Golgi apparatus. Various proteins with lipid binding domains (PH, PX, FYVE, or ENTH, etc.) recognize the variably phosphorylated inositol lipid head groups in order to be recruited to the membrane (Lemmon, 2008).

Sterols are the main non-polar lipids of biological membranes and in *S. cerevisiae* the principal sterol is ergosterol and not cholesterol as it is in vertebrates. Ergosterol is found in the plasma membrane and has highest intracellular concentration in the late secretary pathway and early endosomes (Hannich et al., 2011). Sphingolipids are polar lipids and longer molecules than PtdCho and so protrude from these membranes (Hannich et al., 2011). They are not distributed homogenously throughout membranes and preferentially mix with sterols to segregate into membrane rafts. This gives membranes lateral heterogeneity and allows for certain proteins to segregate into functional domains giving membranes a 'patchwork' structure (Munro, 2003; Spira et al., 2012).

3.3.2: Methods for investigating membrane protein interaction networks

Interactions between lipids and proteins are difficult to measure directly due to the insolubility of lipids and membranes, yet significant progress has been made. Methods for directly analyzing these interactions are: *in-vitro* binding assays (Gallego et al., 2010; Saliba et al., 2014; Yu et al., 2004), *in-vivo* pull-downs of lipid bound proteins (Li et al., 2010; Maeda et al., 2014), *in-vivo* protein-lipid crosslinking techniques (Haberkant et al., 2013), and lipidomics and proteomics of fractionated membranes. These techniques have all been successful in identifying lipid-binding domains but are limited in adding contextual features such as collaborative lipid protein interactions and their localizations within the cell.

Orthogonal to these biochemical techniques, depleting or increasing the concentration of certain lipids within the cell allows for the investigation of their structural role in protein localization and roles in the regulation of specific cellular processes. Many studies have done this by knocking down steps in lipid biosynthesis pathways and imaging GFP-tagged proteins on a simple case-by-case level (Fairn et al., 2011; Heese-Peck et al., 2002; Hermesh et al., 2014; Park et al., 2008). As these studies are limited in scope as they are very selective, the system-wide effects of these lipid perturbations are poorly characterized. To provide a more comprehensive account of the structural roles of lipid classes, we designed a high-content screen.

3.3.3: High-content screens provide comprehensive understanding of proteome rearrangements due to genetic or chemical perturbations

Genetic essentiality screens and genetic interaction screens simply measure growth speed of cell populations that have been genetically or chemical perturbed. This crude proxy for fitness tells us about the relative significance of these proteins but nothing of the underlying molecular biology that is occurring in the perturbed system (Costanzo et al., 2010).

High-content screens are based on fluorescence microscopy of fluorescently-tagged proteins, providing quantitative information on protein expression, dynamics and localizations in cells with perturbed genetic backgrounds (Vizeacoumar et al., 2010; Zanella et al., 2010). This adds much-needed context to genetic interactions by revealing how the proteome rearranges to cope with genetic or chemical perturbation (Collinet et al., 2010; Neumann et al., 2010; Vizeacoumar et al., 2010). These rearrangements can occur where there is no growth defect in the mutant cells, for example, a ΔMUS81 *S. cerevisiae* strain has wild type fitness but exhibits

increased numbers of DNA damage foci compared to a wild type strain (Alvaro et al., 2007).

There have been a number of these high-content screens utilizing the yeast GFP library but none so far have focused on the interactions of proteins with lipids.

3.3.4: Designing a high-content screen to investigate structural roles of lipids in *S. cerevisiae*

S. cerevisiae is an ideal model system for investigating the effects of lipid classes on protein localization *in vivo* as it is amendable to genetic manipulation and has large and stable membrane domains.

The biosynthesis of lipid species is complex. Even 'linear' pathways are complicated by their sequential intermediate steps occurring in different organelles of the cell (Natter et al., 2005). In *S. cerevisiae*, the majority of lipids are synthesized in the endoplasmic reticulum (ER), although more lipid synthesis occurs in the Golgi apparatus and in the plasma membrane. We have chosen the five main lipid biosynthesis pathways to target for creation of query strains (strains with a genetic perturbation): sphingolipid biosynthesis, glycerophospholipid biosynthesis, phoshotidylinositolphosphate biosynthesis, long-chained fatty acid and ergosterol biosynthesis. GFP-tagged strains are used as 'array' strains to be crossed with the query strains, which contain the genetic background being tested to assess localization changes.

Here, I present a high-content screen to systematically monitor translocation and/or miss-localization of the proteins by fluorescence microscopy in cells in which lipid biosynthesis has been perturbed. The biosynthesis of phospholipids, sphingolipids, ergosterol and phosphoinistol phosphates are perturbed by a mixture of knockouts, knockdowns, and drug inhibition of the enzymes involved. This provides insights into the roles for these lipids classes in the eukaryotic system.

3.3.5: Impact of research

In summary, the produced dataset is a resource that offers clues on how lipid classes are involved with global proteome regulation, specific protein binding and the regulation of cellular processes.

3.4: Results and discussion

3.4.1: Selection of pathways and perturbations for the query strains

In the five lipid biosynthesis pathways, in total, 25 enzymes were perturbed. Non-essential enzymes were knocked out and essential enzymes were targeted with temperature sensitive knockdowns, tetracycline induced knockdowns or their function perturbed with drugs (figure 3.1) (Appendix F.I).

In the main glycerophospholipids biosynthesis pathway CHO1, CHO2 and PSD1/PSD2 were knocked-out as these are non-essential genes in rich media (Figure 3.2). Cho1 and Cho2 are localized in the endoplasmic reticulum. The orthologous enzymes Psd1 and Psd2 are localized in the endomembrane system and mitochondria respectively and are therefore knocked out together in the same strain (Henry et al., 2012). In addition, a knock-out strain of the non-essential phosphatidycholine degradation enzyme Spo14, a phospholipase D, was created.

The PtnIns pathway contains tightly regulated PtnIns kinases and phosphatases that target distinct pools of lipids in distinct organelles (Figure 3.2). Stt4 and Mss4 are plasma membrane localised PtnIn kinases and are both perturbed. STT4, an essential Ptdn[4]P kinase, which causes fast cell death when knocked out, is target by two different methods; a tetracycline inducible knock-down, and a perturbation of function with the drug wortmannin (Cutler et al., 1997). MSS4, the sole Ptdn[4,5]P kinase in yeast, is also essential and is targeted by a temperature sensitive knockdown (Figure 3.1).

Sac1 and the orthologous Sjl1 and Sjl2 are non-essential PtnInps phosphatases which target different pools of PtdInPs and are knocked out (Foti et al., 2001). The non-essential PLC1, the cell's only PtnIn[4,5]P2-specific phospholipase C, is also knocked out (Flick and Thorner, 1993; Rebecchi and Pentyala, 2000; Wera et al., 2001) (Figure 3.1).

In the sphingolipid biosynthesis pathway, 10 of the 18 enzymes involved are perturbed. Two of these are essential, Lcb1 and Aur1, and are knocked-down under the control of a tetracycline inducible promoter. The other 8 were knocked out as query strains (Figure 3.1).

Ergosterol biosynthesis is a linear pathway with 10 enzymes, three of which were perturbed (Figure 3.2), the nonessential Erg2 and Erg6 were knocked out and the upstream essential Erg25 was knocked down using a tet-promoter.

Finally the long chain fatty acid synthesis enzyme Elo3 was also knocked out in a query strain (Figure 3.1).

3.4.2: Selection of arrayed proteins

To gain a comprehensive picture of the effects of these perturbations on the organization of cellular membranes, we mined membrane-associated proteins from both SGD annotations (http://www.yeastgenome.org/) and the GFP localization dataset available from Huh, et al (2003). The following localization categories from the Yeast GFP database were chosen; 'Cell periphery', 'Punctate', 'Actin', 'Endosome', 'Lipid particle', 'Bud', 'Bud neck'. Additional proteins with the following SGD Gene Ontology annotation terms were added; 'Plasma membrane', 'Actin', 'Endocytosis', 'Exocytosis', 'Eisosome', 'Bud', 'Bud tip', 'Bud neck', and 'Lipid particle'. Together this totaled 805 candidate proteins, of which, 648 proteins were available as GFP fusions strain are used in this screen. (Appendix F.II)

3.4.3: Summary of data collection

After mating, sporulation and successful strain selection steps, 13,507 double mutant strains were recovered having both GFP & lipid enzyme perturbation. Each strain had a failure rate of <5% of crosses and image collection of 9 images from each cross plus three WT controls. A random selection of 10% of the strains was used for duplication, which had similar high success rates. Including GFP control strains 14,157 strains were imaged and in total 201,464 images were taken and, of these, 35,257 were judged to be of poor quality by an automated algorithm that detected out-of-focus images (see materials and methods) (Appendix F.III).

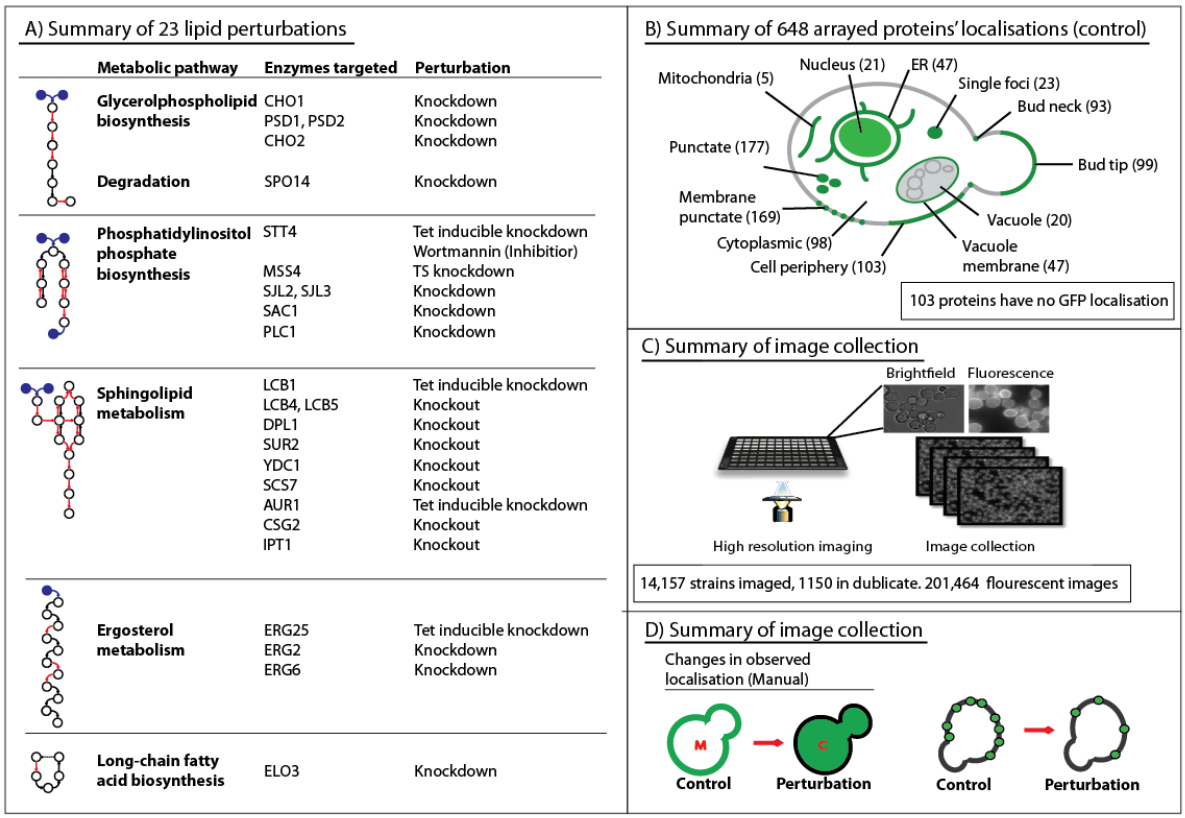

Figure 3.1: Summary of project pipeline

A) Summary of query conditions in the 5 lipid biosynthesis pathways; 25 proteins are perturbed in a total of 23 conditions. B) Summary of the manually assigned protein localizations of the 648 arrayed GFP proteins. Many proteins had more than one localization and 101 had no localization assigned. C) Each condition had 9 bright-field images and 9 corresponding fluorescence images collected. From a total of 14,157 strains were imaged including GFP controls and 201,464 images judged of 'good' quality were collected. D) Changes in localization changes were manually assigned.

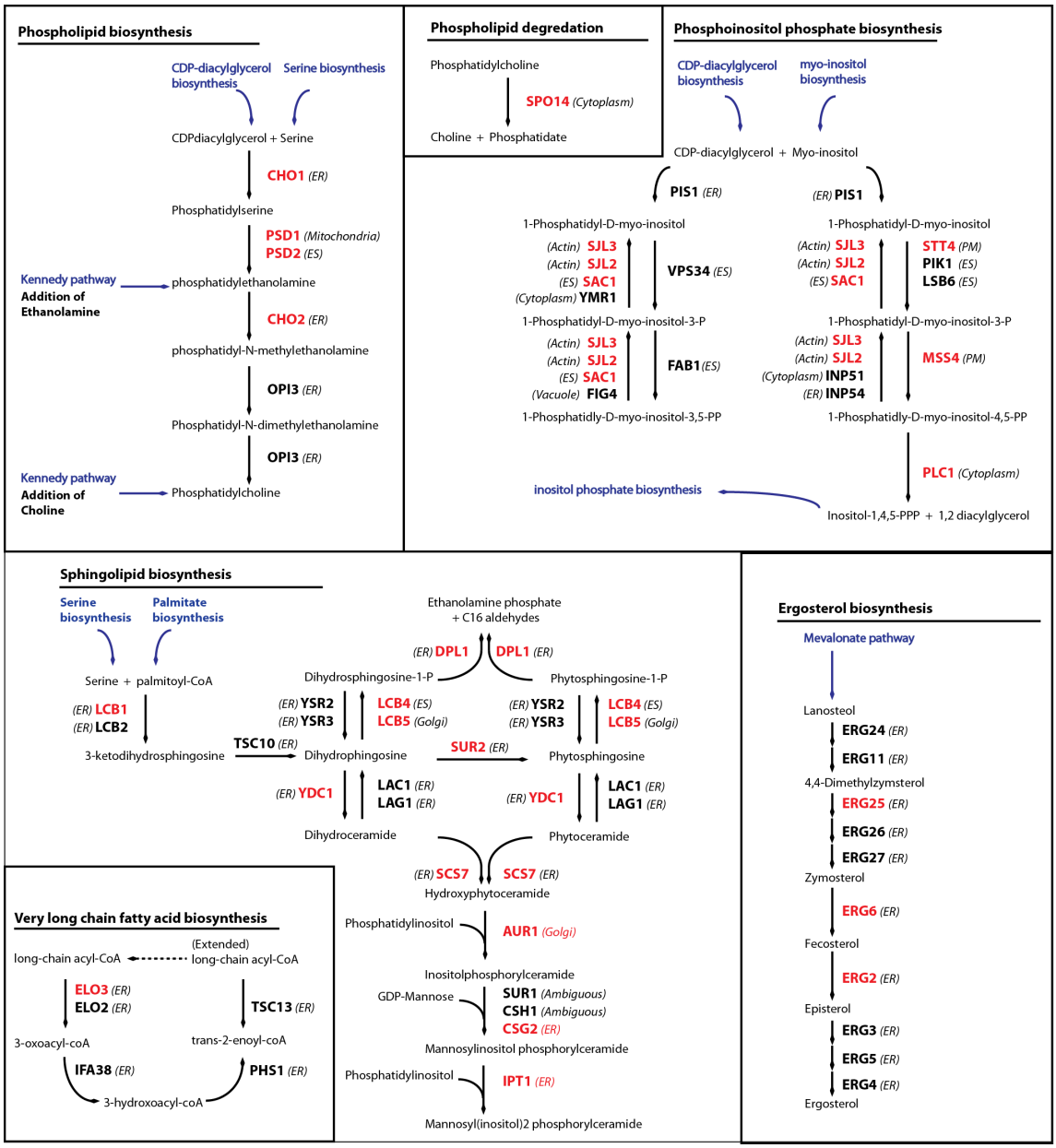

Figure 3.2: Summary of the 5 targeted lipid biosynthesis pathways.

Schematic diagrams of the lipid pathways investigated in this study. The subcellular localizations of the proteins are recorded and proteins in red are those that have been perturbed. The specific natures of the perturbations are outlined in Figure 1. ES = Endomembrane system, ER = endoplasmic reticulum, PM = Plasma membrane. Pathways and cellular localizations were from the Kegg database and several reviews [http://www.genome.jp/kegg/pathway.html](Cowart and Obeid, 2007; Daum et al., 1998; Henry et al., 2012; Ren et al., 2006).

3.4.4: Manual calling of hits

The state-of-the-art in this field is still currently manual calling of localization differences between the control and perturbed strains (Breker et al., 2013; Tkach et al., 2012). For each protein, the GFP 'wild-type' control cells were assessed for each of the following localizations; cell periphery membrane punctate, endoplasmic reticulum, mitochondria, bud neck, bud tip, punctuate, punctuate (single), vacuole interior, vacuole membrane, nucleus (Figure 3.1B). The perturbations were assigned as having an 'increase' or 'decrease' of the protein intensity at these cellular localizations. Localization changes called by two independent reviewers are retained.

103 proteins had no observable localization in the GFP control. Of the remaining 545, 243 proteins imaged were found to be miss-localized in at least one lipid perturbation background and 148 were miss-localized in >1 condition. All protein localization changes can be found in Appendix F.IV.

[Note: originally this project was designed to use image analysis and machine learning to automatically assign localization changes in perturbed strains, for a description of why this didn't work please see Appendix B]

Glycerophospholipid biosynthesis pathway biosynthesis

3.4.5: Glycerophospholipid biosynthesis; Known lipid concentration changes in phospholipid pathway perturbations

The phospholipid biosynthesis pathway is actually a complex network of tightly regulated sub-pathways and feedback mechanisms, containing redundant routes for synthesis of the resulting metabolites. Downstream glycerophospholipids can be synthesized from extracellular additives, such as ethanolamine or choline via the Kennedy pathway (Figure 3.2). ΔCHO1, ΔPSD1/ΔPSD2, ΔCHO2 and ΔSPO14 strains were grown on YPD rich media prior to imaging experiments (to provide these supplements) and then imaged with minimal media for phenotypic data collection. ΔCHO1 and ΔPSD1/ΔPSD2 showed extremely slow growth when left to grow on minimal media.

The enzymes CHO1, PSD1/PSD2 and CHO2 are well studied and their knockouts well established in many studies. The ΔCHO1 (PtdSer synthase) strain, when supplemented with choline and ethanolomine, is depleted in PtdEtn and the PtdSer concentration is reduced almost ~99% (Fairn et al., 2011). In wild-type cells PtdSer is localized in the inner leaflet of the membrane in a mesh pattern, which is excluded from endocytosis (Fairn et al., 2011; Spira et al., 2012). PtdIn and PtdCho concentrations are also increased in cho1Δ cells (Fairn et al., 2011).

The two PtdSer decarboxylases, PSD1 and PSD2, which have different cellular localizations, are knocked out together (Figure 3.2). PtdEtn concentration is reduced in these cells and PtdCho and PtdSer concentrations are moderately depleted (Schuiki et al., 2010).

ΔCHO2 cells have a severe imbalance of PtdEtn and PtdCho levels. PtdCho is depleted, PtdEtn levels increase dramatically and PtdSer and PtnIns remain relatively unperturbed (Daum et al., 1999; Summers et al., 1988; Thibault et al., 2012).

Spo14, phospholipase C, is required for the conversion of PtnCho to PA and studies have shown that its knockdown reduces levels of PA but the effect on PtnCho is unclear. Spo14 is not essential but known to be required for meiosis and spore formation (Rose et al., 1995; Waksman et al., 1996).

3.4.6: Glycerophospholipid biosynthesis; Phospholipids have widespread yet specific effects on the proteome localization

The Figures 3.3 and 3.4 summarize the proteins that are miss-localized by these perturbations and the cellular localizations affected. In total, 156 proteins were affected in the four phospholipid conditions, of the total 648 proteins screened. ΔCHO1 cells had the largest number of proteins (99) then ΔPSD1/ΔPSD2 (61), ΔCHO2 (63) and ΔSPO14 (2).

PtdEtn is depleted in both ΔCHO1 cells and ΔPSD1/ΔPSD2 cells (PtdSer is also depleted in ΔCHO1 cells), so it may be expected that proteins affected by ΔPSD1/ΔPSD2 would be a subset of those affected by the ΔCHO1 (Fairn et al., 2011). Indeed, the majority of the proteins affected in ΔPSD1/ΔPSD2 (74%) are also perturbed in ΔCHO1 cells (Figure 3.3). The cellular localizations of the affected proteins are also similar, fewer proteins localized to the bud, intracellular punctuate and membrane punctuate (Figure 3.4).

Overall, the hits in ΔCHO2 are distinctive from those in other conditions in both the proteins and the localizations affected. This may be due to difference in PtnEtn/PtnCho ratio, which has a huge effect on membrane flexibility (discussed later).

ΔSPO14 had only two hits, Gts1 and Bio5 which both have reduced membrane localization (not shown).

The large numbers of proteins affected in these glycerophospholipid conditions are expected, as phospholipids are the main constituents of cellular membranes, this is demonstrated by the wide range of cellular compartments and membrane functions affected by these perturbations. It is difficult to assign causality to these protein miss-localizations; for example, the miss-localization of plasma membrane localized transmembrane transporters may be due to the defects in intracellular transport or due to specific phospholipid binding. None of the conditions show specific enrichments of proteins affected based on Gene Ontology terms but they do confirm previously observed results in the literature and some other patterns emerge. These are discussed below.

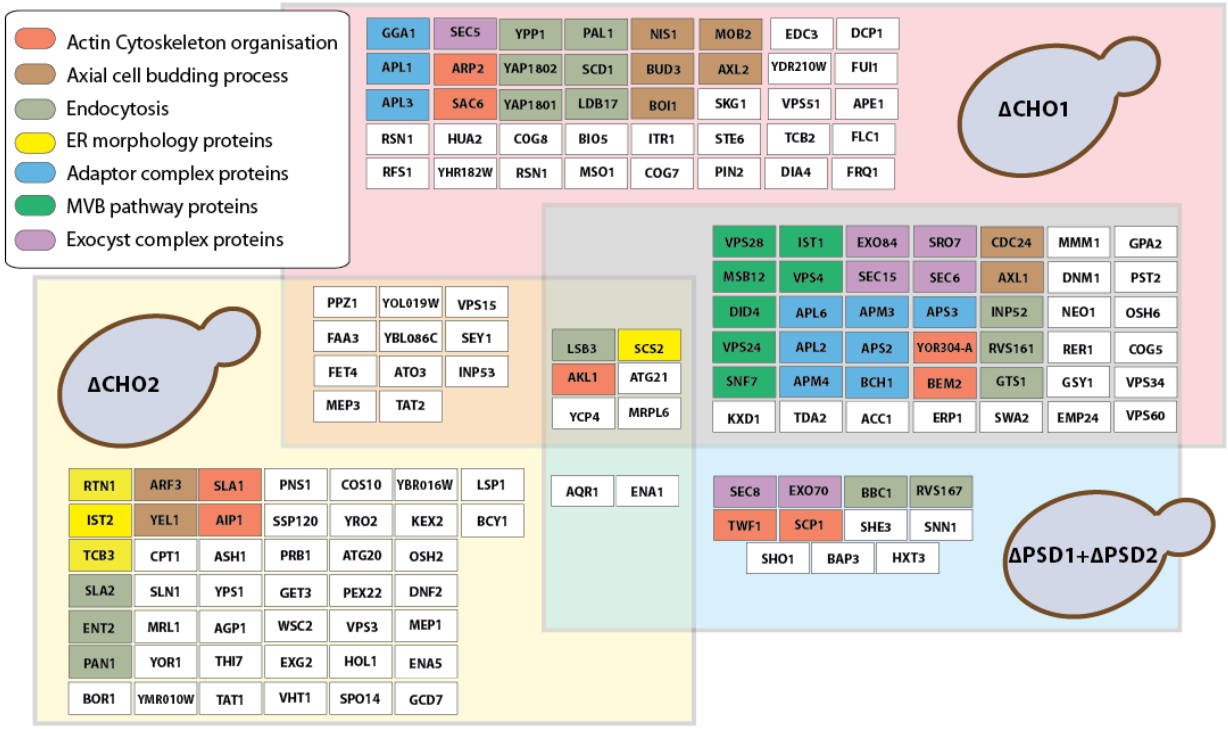

Figure 3.3: All proteins affected in the main glycerophospholipid pathway perturbations

Perturbed cells are represented; ΔCHO1, ΔPSD1/ΔPSD2, ΔCHO2. Bubbles contain proteins miss-localized in the adjoining condition. Proteins affected in more than one condition are in separate bubbles. Proteins involved in selected biological processes are indicated.

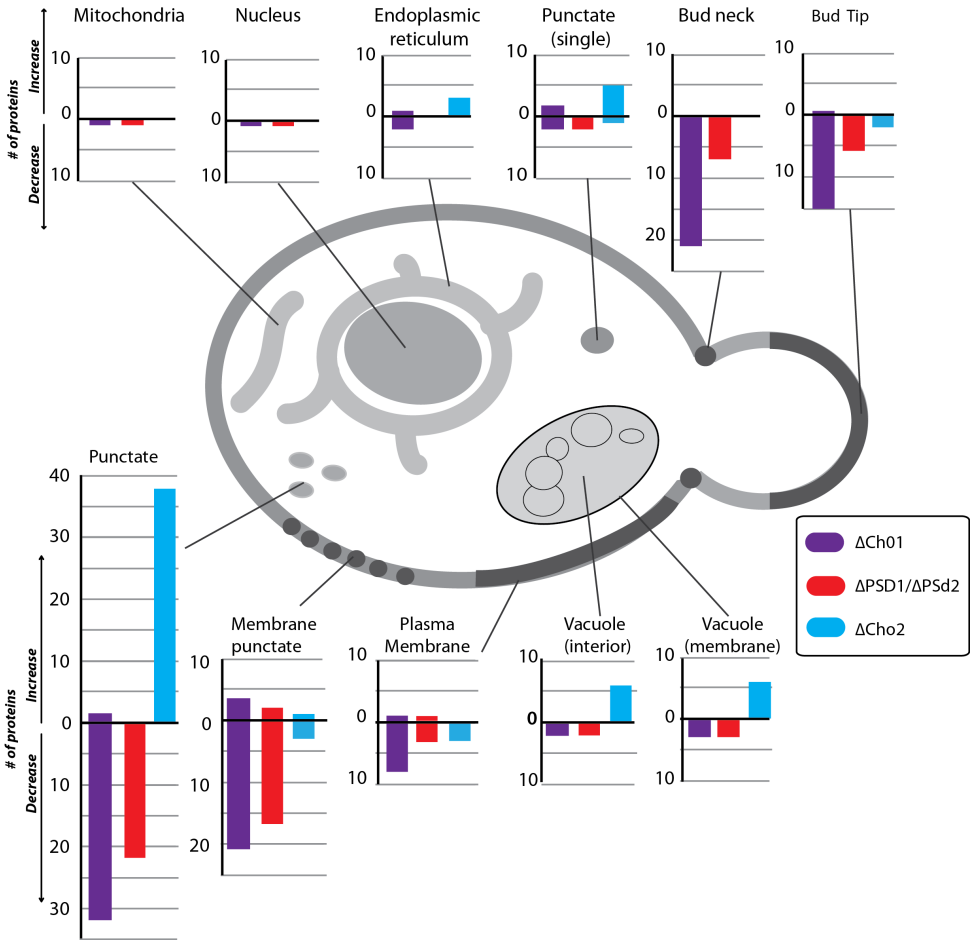

Figure 3.4: Protein localizations affected in glycerophospholipid pathway perturbations

A localization change for individual proteins is recorded as an increase or a decrease in prevalence in the 11 localizations shown. Each graph shows the number of proteins in the three conditions affected in that localization. The list is redundant as each individual protein can be affected in several cellular localizations.

3.4.7: Glycerophospholipids; CHO1 and PSD1/PSD2 deletions strongly affect intracellular protein transport

Many proteins that localize as internal punctate are have reduced intensity in ΔCHO1 and ΔPSD1/ΔPSD2 cells. Proteins that show this distinctive phenotype are involved in many intracellular transport pathways that involve the budding, transport, and merging of intracellular vesicles. These include proteins associated with the adaptor complexes, ESCRT complexes, the Exocyst, Golgi complexes, endocytosis and ER morphology. Typical phenotypes are displayed in Figure 3.5.

The common effect on the lipid profile in ΔCHO1 and ΔPSD1/ΔPSD2 cells is that there is a decrease in PtdEtn concentration. Very little is known about the roles of PtdEtn other than it is involved in the targeting of certain synthesized proteins to the plasma membrane (Opekarová et al., 2002). This role in intracellular transport may be the explanation for the observed defects in localization of transmembrane proteins to the plasma membrane.

The ratio of PtdCho and PtdEtn is important for the flexibility of membranes as PtdEtn has a conical shape compared with PtdCho's more cylindrical shape so this ratio change may effect efficiency of internal membrane budding as membranes become less flexible (Anitei and Hoflack, 2011; Li et al., 2006; Opekarová et al., 2002). PtdSer concentrations is also affected in ΔCHO1 cells but is known to be excluded from endosomes and is relatively depleted in internal membranes so is unlikely to affect these intracellular processes (Fairn et al., 2011).

3.4.8: Glycerophospholipid biosynthesis; Cho1 deletion affects axial budding initiation

ΔCHO1 cells cause miss-localization of seven proteins associated with axial budding, the preferred budding mechanism of haploid cells (Slaughter et al., 2009). AXL1, AXL2, BUD3, CDC24, NIS1, BOI1 and MOB2 lose their bud localization (Figure 3.5). This is a well-characterized process and the other components BUD4, RSR1 and BEM1 are unaffected in this screen. CDC42, the main cellular driver of symmetry breaking to allow budding, was not screened. Interestingly BEM1 and CDC42 were shown a previous study to be miss-localized in ΔCHO1 cells and the changes in axial budding were attributed to the reduction in the polarized PtdSer in this strain (Fairn et al., 2011).

Budding does occur in the ΔCHO1 cells even without the efficient localization of apical growth proteins though there is a delay in bud emergence (Fairn et al., 2011). A total 156 proteins were assigned bud neck or bud tip localizations in the wild type cells in this screen and 130 of these did not lose their bud localization in ΔCHO1 cells (Figure 3.4). The still images in

this screen do not permit assessment of budding rates or the progression of the bud site, but the strains used in this screen could be used to assess the progression of budding in cells without polarized PtdSer.

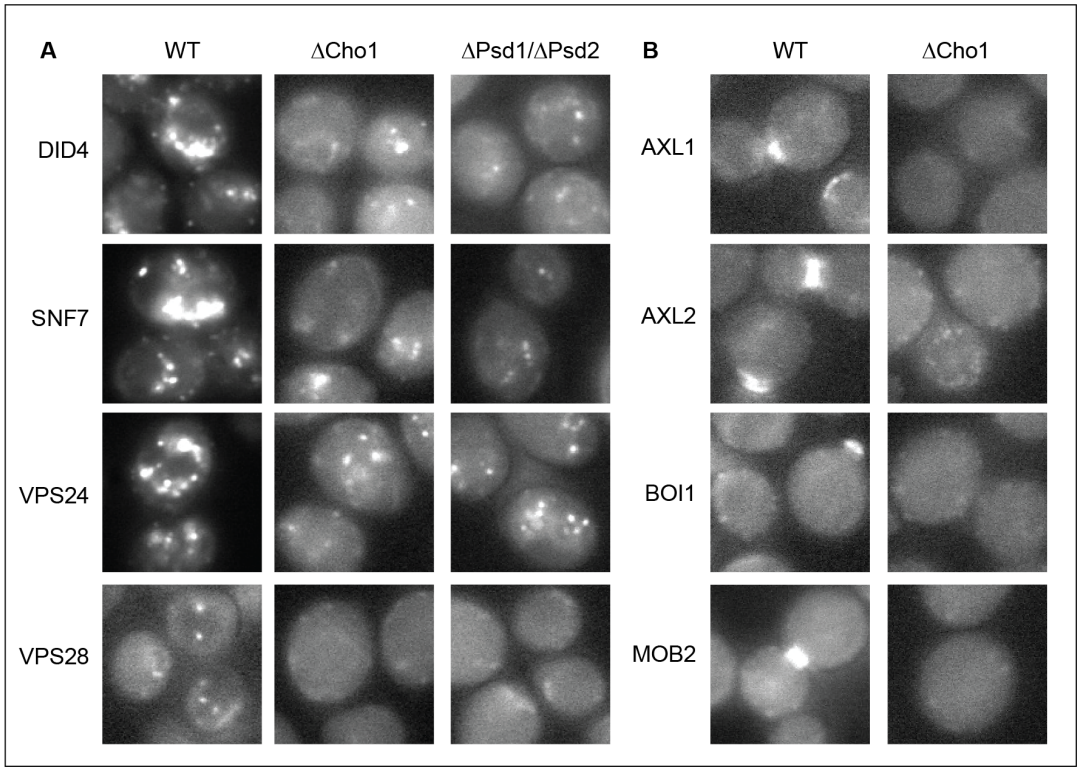

Figure 3.5: Showing intercellular transport proteins that become less punctuate in ΔCHO1 and ΔPSD1/ΔPSD2 cells and Axial budding proteins lose polarization in ΔCHO1 cells

A) Proteins involved in intercellular transport become less punctuate in ΔCHO1 and ΔPSD1/ΔPSD2 cells. The proteins displayed are part of the ESCRT complexes, representative of the phenotype of proteins involved in intracellular vesicle transport, the punctuate pattern becomes less intense in both the ΔCHO1 and ΔPSD1/ΔPSD2 cells.

B) Representative selection of proteins involved with axial budding that lose their localization/polarization in ΔCHO1 cells.

3.4.9: Glycerophospholipid biosynthesis; CHO2 deletion specifically affects ER organization and causes defects in the localization endocytotic machinery.

It has been well reported that the ER is disrupted in ΔCHO2 cells (Tavassoli et al., 2013) and indeed the ER marker proteins such as Scs2 and Rtn1 display a fragmented ER in our data (Figure 3.6). The cortical ER is also malformed, as demonstrated by the miss-localization of its anchor proteins IST2 and TCB3 (Hermesh et al., 2014) (Figure 3.6). This major structural rearrangement in the cell causes proteins such as Osh2 that are dependent on the ER for their correct localization to also become miss-localized. It is therefore difficult to interpret these new localizations in terms of specific lipid requirements, as most will simply fragment with the ER. This explains the large number of proteins which developed a punctuate localization in ΔCHO2 cells (Figure 3.4).

A functional subset of the proteins that become more punctuate in this condition are End2, Lsb3, Pan1, and Sla2, all of which are late coat proteins involved in clathrin-mediated endocytosis (Figure 3.7). These proteins normally localize in dynamic endocytotic foci at the plasma membrane and recycle back to the membrane after internalization of the endosome. However in ΔCHO2 cells these become more internally localized as punctate structures.

My hypothesis that this phenotype is due to an increase in PtnEtn concentration on these membranes and this causes a defect in the cells' ability to recycle this endocytosis machinery back to the plasma membrane. Time-lapse microscopy would be required to test this hypothesis along with co-localization studies with known marker proteins of each stage of endocytosis. This is currently planned.

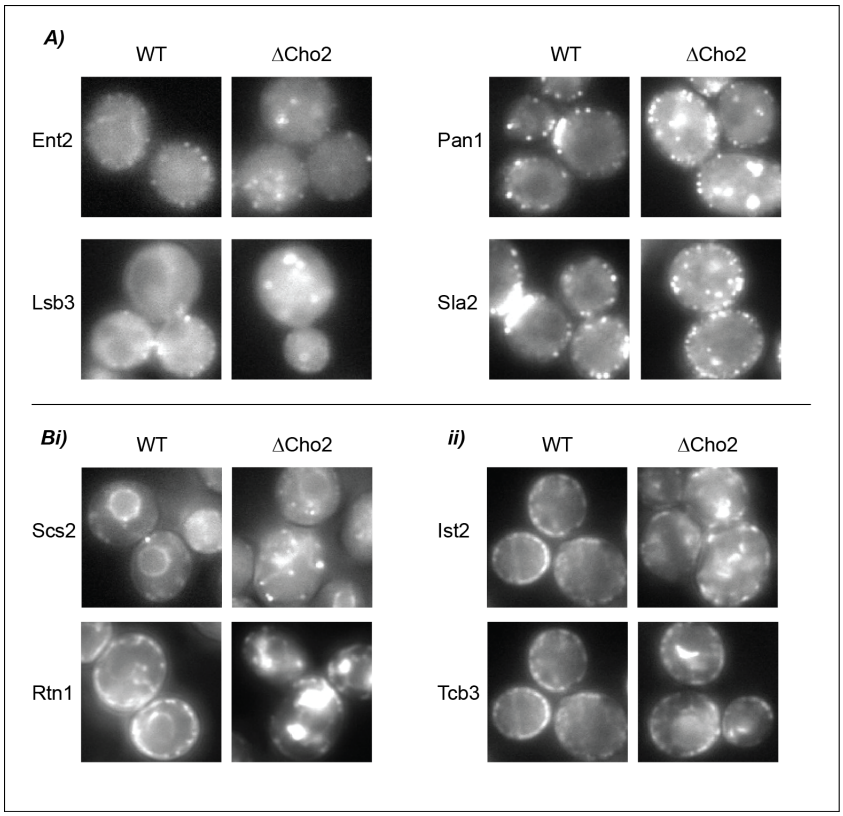

Figure 3.6: Proteins involved in endocytosis that become more punctate in ΔCHO2 cells.

A) Protein involved in the formation of the late endocytotic quote have a peripheral localization in the wild type cells but becomes internalized into punctuate in ΔCHO2 cells. Bi) ΔCHO2 cells have large defects in ER morphology as demonstrated by the ER markers Scs2 and Rtn1.

Bii) The cortical ER is also affected, as demonstrated by the cortical ER markers Ist2 and Tcb3.

Phosphatidylinositol phosphate biosynthesis pathway perturbations

3.4.10: Phosphatidylinositol phosphate biosynthesis; Known lipid concentration changes in PtnInP pathway perturbations

PtnInPs form a relatively small proportion of the cell's membranes but they regulate much of the cell's physiology as signaling molecules providing spatial and temporal downstream information (Strahl and Thorner, 2007). Perturbation of the proteins involved in the synthesis of PtnInPs is challenging. Indeed, as they have so many systemic roles that their mutations are often pleiotropic (De Camilli et al., 1996).

Stt4 is one of two essential Ptdn[4]P kinases (the other is Pik1), is localized at the plasma membrane and produces a distinct pool of Ptdn[4]P there (Figure 3.2). We have targeted Stt4 which, when depleted causes fast cell death, by two different methods, a tetracycline inducible knockdown and a perturbation of function with the drug wortmannin (both of which caused severe growth defects). Mss4 is the sole Ptdn[4,5]P kinase and is targeted by a temperature sensitive knockdown which at 37^{o}C caused cell growth to cease.

Mss4 phosphorylates both the pools of PtdIns(4)P generated from Pik1 and Stt4 (each produce roughly half of the overall cellular total). These pools are involved in different cellular processes (Audhya et al., 2000). Cells deficient in Mss4 have been reported to have ~10% of the WT amount of PtdIns(4,5)P2 (Desrivières et al., 1998).

Sac1, Sjl1 and Sjl2 are PtdInps phosphatases which are not essential and deletion of SAC1 or both SJL1 and SJL2 leads to raised levels of PtdIns(4)P (Foti et al., 2001). Plc1 is the only PtdIns[4,5]P2-specific phospholipase C in *S. cerevisiae* and is also knocked out (Flick and Thorner, 1993; Rebecchi and Pentyala, 2000; Wera et al., 2001)

3.4.11: Phosphatidylinositol phosphate biosynthesis; Overview of protein miss-localizations in PtnInP kinase perturbations

The PtnIn(4)P kinase Stt4 function was perturbed by two independent methods, the drug wortmannin and a tetracycline inducible knockdown (tet-strain). Surprisingly the overlap of proteins miss-localized in the tet-STT4 and the wortmannin conditions is small (19%). Intracellular localizations of proteins in tet-STT4 cells are difficult to interpret due to its pleiotropic affects (Audhya et al., 2000), interestingly this was not a problem in wortmannin treated cells. Therefore, localization changes called in the tet-STT4 strain usually involve the plasma membrane, as those were much easier to call.

Loss of Stt4 and Mss4 function is known to cause deregulation of the actin cytoskeleton (Strahl and Thorner, 2007). Indeed in the tsMSS4, tet-STT4 and wortmannin treated cells, proteins involved in the actin cytoskeleton and cellular processes dependent on it, such as budding and endocytosis, are miss-localized (Figure 3.8). The defects in actin cytoskeleton regulation and bud site formation explain the large number of localization changes involving the bud and membrane punctuate (Figure 3.7). Some representative images showing these phenotypes are shown in figure 3.8.

Interestingly, Slm1, a protein with a PH domain that binds PtnIn(4,5)P2 is affected in the STT4 perturbations but not in the MSS4 perturbation showing that the reduction of PtnIn(4,5)P may not be complete tsMSS4 cells. Osh2, shown in an *in vitro* screen in our lab to bind all PtnInPs (unpublished data), loses its membrane localization in all three conditions.

The membrane compartment containing Can1 (MCC), an arginine permease, is a structurally distinct compartment of the plasma membrane with distinct lipid and protein content. Changes to compartment in all lipid biosynthesis perturbations are discussed in detail later.

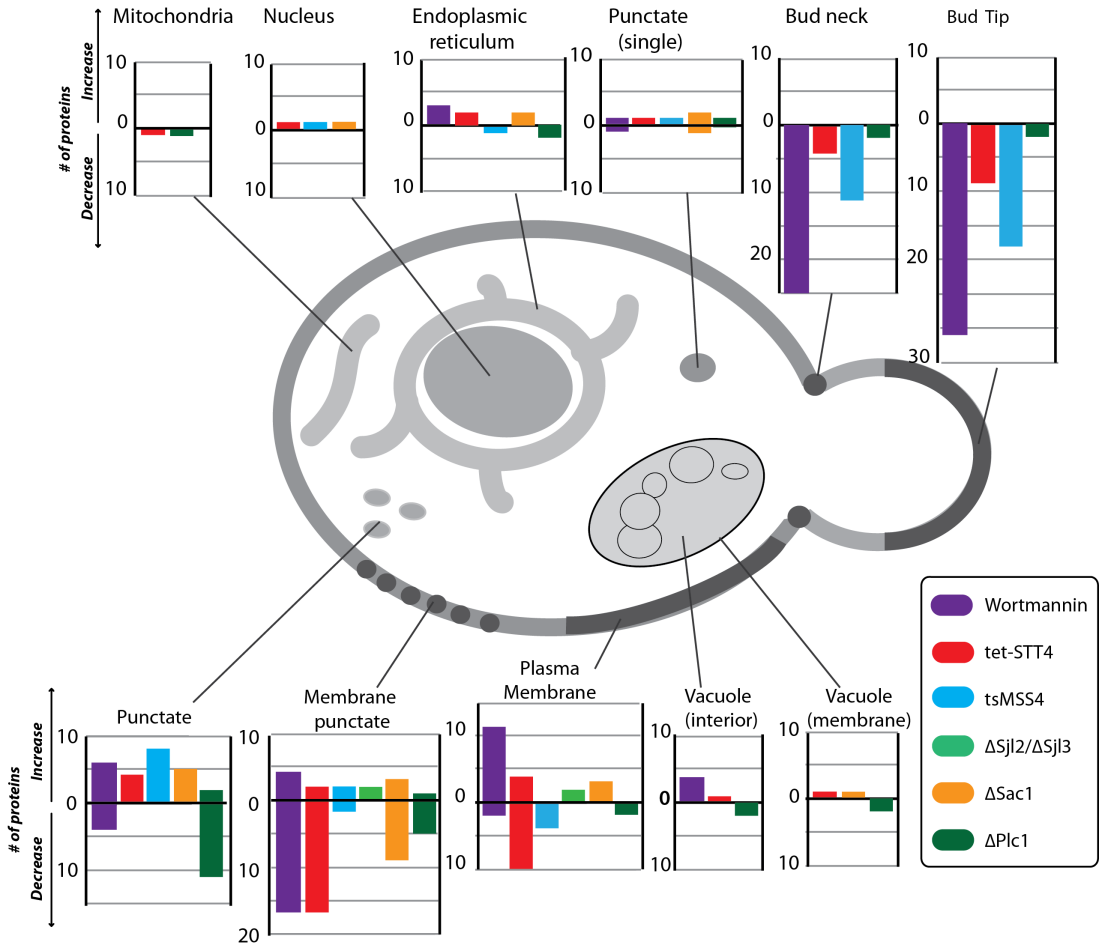

Figure 3.7: Protein localization affected in PtnInP biosynthesis perturbations

A localization change for each protein is recorded as an increase or a decrease in prevalence in the six localizations shown. Each graph shows the number of proteins in the six conditions that showed an increase or decrease in intensity in that localization. The list is redundant; proteins can be miss-localized to or from several cellular localizations in a single condition.

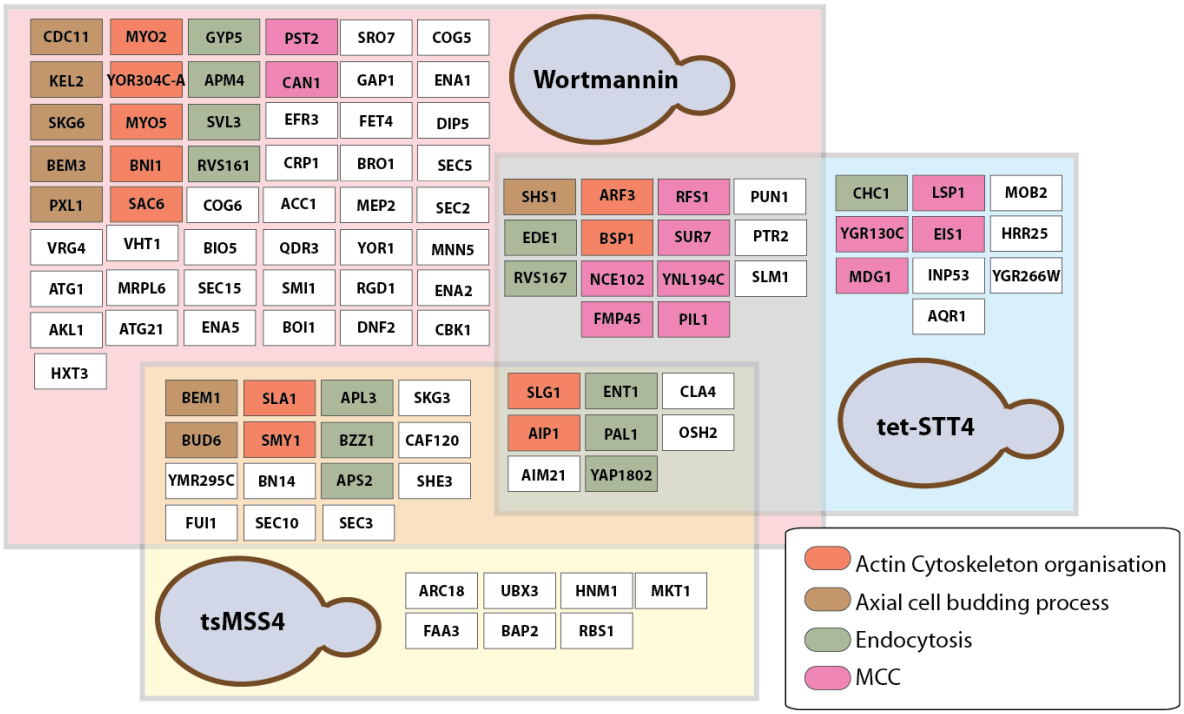

Figure 3.8: All proteins affected the PtnInP kinase perturbations

Cells with perturbed PtnInP kinases are represented; wortmannin treated, temperature sensitive MSS4 at non-permissive 37°C (tsMSS4) and tetracycline inhibited expression of STT4 (tet-STT4). Bubbles contain proteins miss-localized in the adjoining condition. Proteins affected in more than one condition are in separate bubbles. Proteins involved in selected biological processes are indicated. The MCC (membrane compartment containing Can1) is a structurally defined region in the plasma membrane.

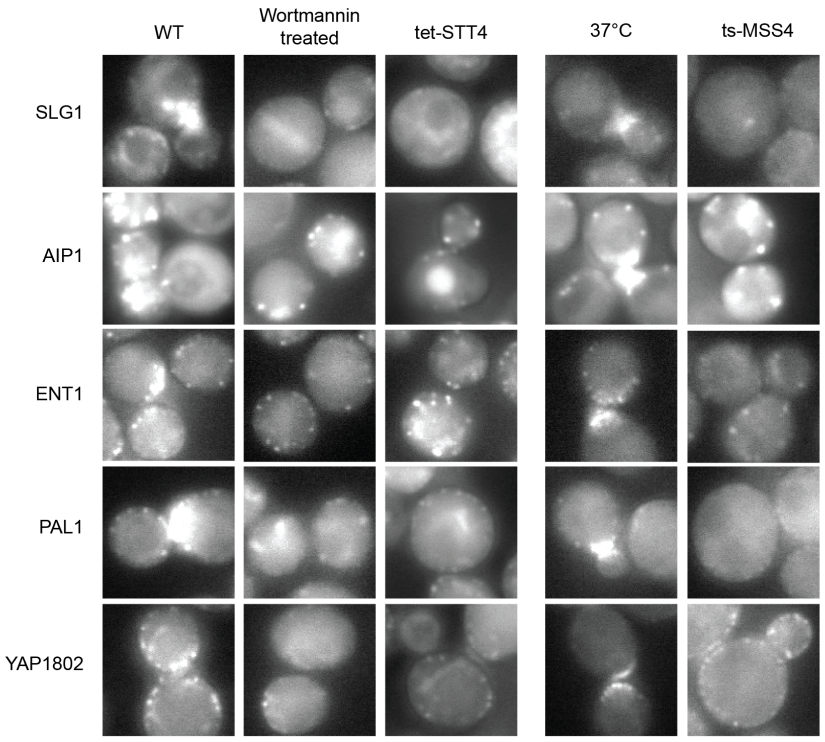

Figure 3.9: Actin cytoskeleton and endocytosis are affected in PtnInP kinase perturbations

Proteins affected in the three PtnInP kinase perturbations. The conditions are Wortmannin treated cells which have perturbed Stt4 function, cells with tetracycline-inhibited expression of STT4 (tet-STT4), and cells with temperature sensitive MSS4 at non-permissive 37^{o}C (tsMSS4) with their 37^{o}C unperturbed control. Slg1 and Aip1 are both actin cortical patch components and become internalized in PtnInP kinase perturbed cells. Ent1, Pal1 and Yap1802 are endocytosis proteins and lose polarization in PtnInP kinase perturbed cells.

3.4.12: Phosphatidylinositol phosphate biosynthesis; Protein miss-localization in PtnInP phosphatase and phospholipase C knockout strains

Sjl2, SJl3 and SAC1 are PtnInP phosphatases are active in different cellular processes and knockout strains were screened for each (Figure 3.2).

Early Golgi cisternae are depleted of PtdIns4P, relative to late cisternae, this is maintained by the Sac1 PtdIns4P phosphatase, which traffics between the ER and early Golgi compartment. ΔSAC1 cells display accumulation of PtdIns(4)P at ER and vacuolar membranes and that study also showed that late endocytosis was defective in ΔSAC1 strains (Tahirovic et al., 2005). Sac1 primarily degrades PtnIns(4)P generated by STT4 and its knockout has been described as affecting the maintenance of vacuole morphology, regulation of lipid storage, Golgi function, and actin cytoskeleton organization (Foti et al., 2001).

In this screen there were surprisingly few localization changes involving internal membranes in ΔSAC1 cells (Figure 3.7 & 3.10). Interestingly the Inp51 (Phosphatidylinositol 4,5-bisphosphate 5-phosphatase), which shares this function with Sac1 (Singer-Krüger et al., 1998) is miss-localized to membrane foci. Inp51 is an important regulator of endocytosis and a number of other endocytotic proteins are also affected. Additionally many of the proteins which form the MCC were affected (this is discussed in detail later).

The ΔSJL2/SJL3 strain is known to have defects in endocytocis, actin cytoskeleton organization, chitin deposition (Singer-Krüger et al., 1998; Strahl and Thorner, 2007), though in the screen described here very few proteins screened were miss-localized (Figure 3.10).

3.4.13: ΔPLC1 cells show defects in intracellular transport

ΔPLC1 cells have severe growth defects, impaired cell wall integrity, decreased osmotic resistance, and an inability to use carbon sources other than glucose (Flick and Thorner, 1993). Above 34^{o}C the cells cannot progress through cytokinesis and become multi-budded. Interestingly we observe the cells becoming larger and elongated at 30^{o}C) (Figure 3.11).

ΔPLC1 cells had and a reduction of the number and the intensity of the punctuate pattern of proteins involved with intracellular transport and endocytosis (Figure 3.10 & 3.11). Indeed Plc1 has been implicated with defects in endosomal sorting with respect to the cellular transport of sterols, though no mechanism has been proposed (Fei et al., 2008). The hydrolysis of Ptn(4,5)P_2 to diaglyceriol (DAG) and IP_3, which plc1 catalyzes, enables membrane fusion due to the intrinsic curvature that DAG gives to membranes due to its conical shape (Anitei and Hoflack, 2011; Asp et al., 2009; Goñi and Alonso, 1999; Kearns et al., 1997). The similarity of the

miss-localization of these proteins with those in ΔCHO1, ΔPSD1/ΔPSD2 cells, discussed in Section 3.4.9 adds weight to the hypothesis that this phenotype is due to mechanical membrane bending deficiencies and not specific protein binding.

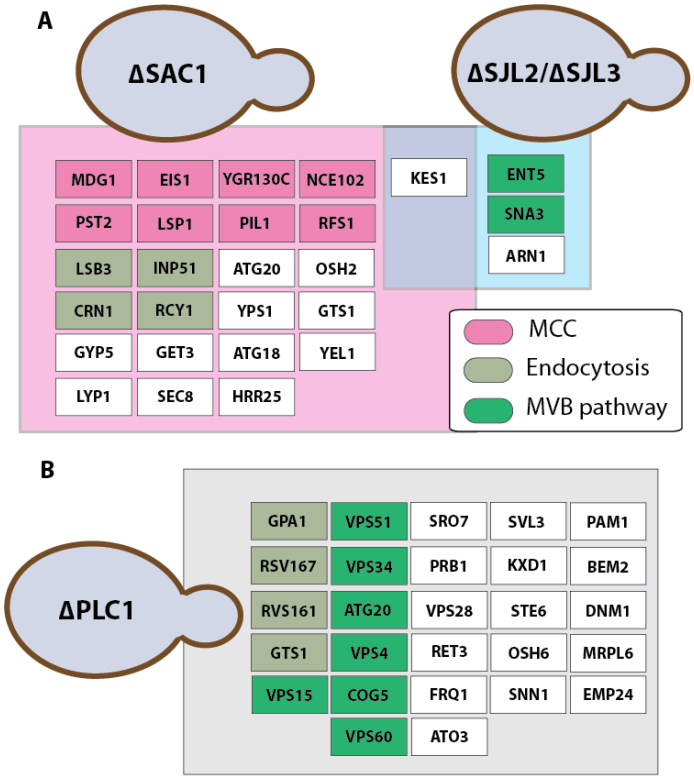

Figure 3.10: All proteins affected in the PtnInP phosphatase perturbations

Perturbed cells are represented; ΔSAC1, ΔSJL2/SJL3 and ΔPLC1. Bubbles contain proteins miss-localized in the adjoining condition. Proteins affected in more than one condition are in separate bubbles. Proteins involved in selected biological processes are indicated.

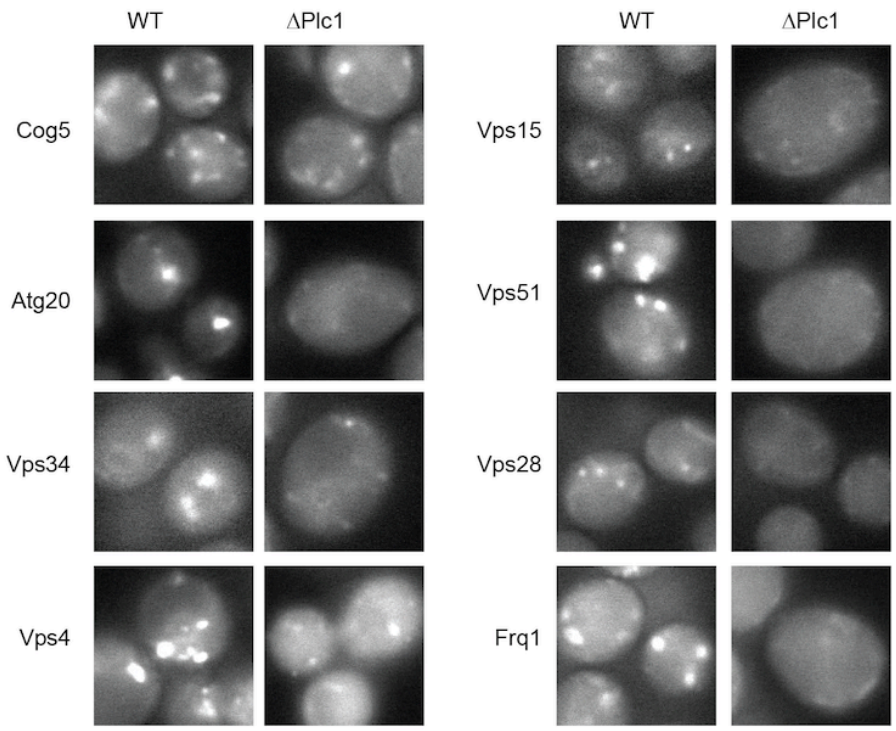

Figure 3.11: Proteins involved in intracellular transport lose punctuate pattern in ΔPLC1 cells

Proteins involved with the multivesicular body (MVB) pathway show a reduction in intracellular punctuate. Frq1 may have a role in intracellular signaling through its regulation of the phosphatidylinositol 4-kinase Pik1p.

Ergosterol biosynthesis pathway perturbations

3.4.14: Ergosterol biosynthesis; Perturbations in the ergosterol pathway confirm specific sterols are required for different cellular processes

Sterols are not major constituents of all membranes. Instead, they are known to segregate into lipid rafts with sphingolipids that act as membrane islands that sequester proteins and their functions to provide heterogeneity to the membrane.

The three enzymes perturbed are Erg25, Erg6 and Erg2 (Figure 3.2). Erg25 is essential and catalyzes the first step in the ergosterol pathway and is therefore perturbed with a tetracycline repressible promoter. Erg2 (C-8 sterol isomerase) and Erg6 (C-24 sterol methyltransferase) are knocked out as these cells can still produce sterols but without all the normal modifications; ΔERG2 cells produce ergosterol without the lack the double bond at C-7,8 and ΔERG6 cells produce ergosterol without the a side chain methyl group (Munn et al., 1999).

Overall there are 48 proteins that were affected by the sterol conditions, two in the tet-ERG25 cells, 10 in ΔERG6 cells and 39 in ΔERG2 cells (Figure 3.12). There were no specific GO terms enriched in these proteins and there is little overlap in the proteins affected.

Surprisingly for an essential protein, the inducible knockdown of ERG25 caused only two miss-localisations, Ret3 and Gts1. Care had to be taken in growing this stain and imaging them before they began to die, perhaps they were not grown for long enough for the effects of limited sterol production became apparent. This cell line also did not show a slow growth phenotype when grown with tetracycline possibly indicating a problem with the knockdown.

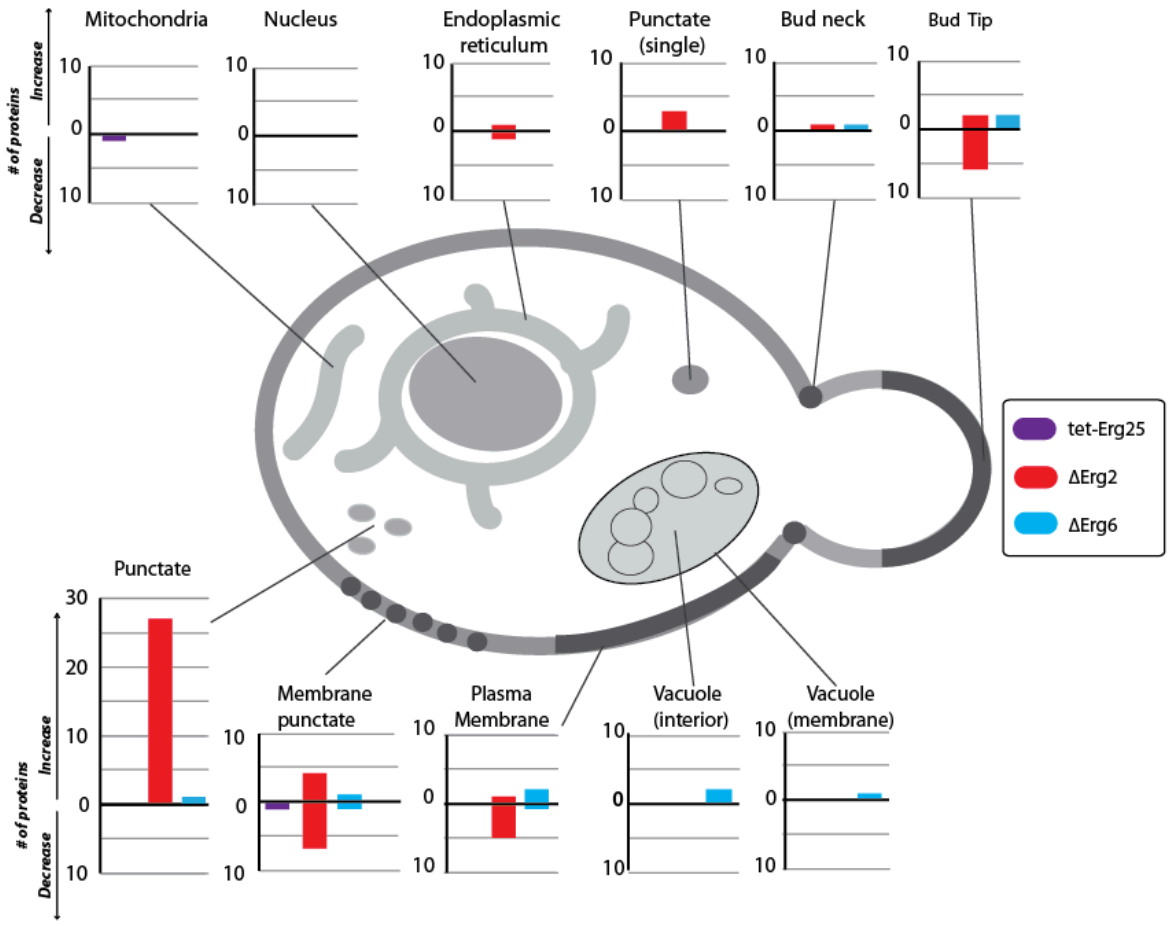

Figure 3.12: Localization of hits in ergosterol biosynthesis perturbations

Localization change for each protein is recorded as an increase or a decrease in prevalence in the three localizations shown. Each graph shows the number of proteins in the three conditions that showed an increase or decrease in intensity in that localization. The list is redundant; proteins can be miss-localized to or from several cellular localizations in a single condition.

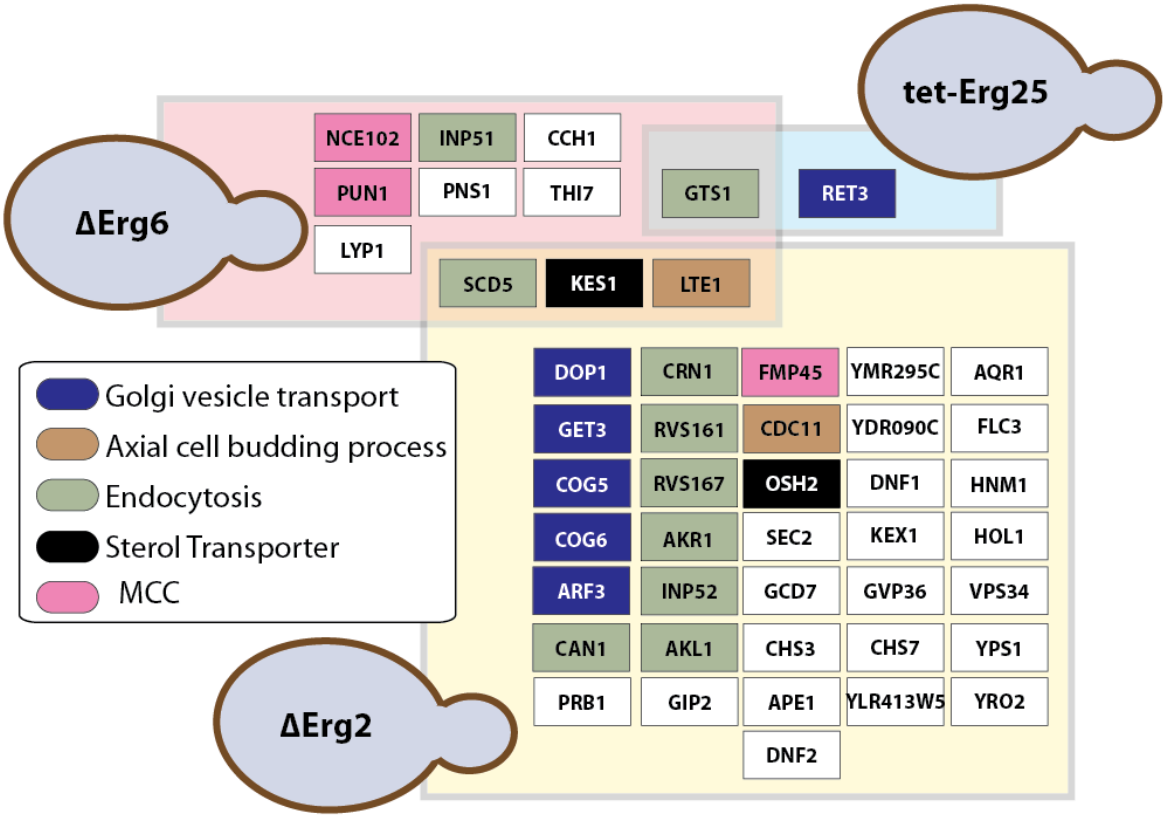

Figure 3.13: All proteins affected in the main glycerophospholipid pathway perturbations
Perturbed cells are represented; tetracycline inhibited expression of Erg25 (tet-ERG25), knockouts of ERG6 and ERG2. Bubbles contain proteins miss-localized in the adjoining condition. Proteins affected in more than one condition are in separate bubbles. Proteins involved in selected biological processes are indicated. The MCC (membrane compartment containing Can1) is a structurally defined region in the plasma membrane.

3.4.15: Ergosterol biosynthesis; ΔERG2 cells have miss-localization of endocytotic machinery

It has been previously shown that ΔERG2 cells are defective in endocytosis (Heese-Peck et al., 2002; Munn et al., 1999) and indeed we see issues with the localization of endocytotic machinery (Figure 3.14). The proteins Rvs167, Rvs161, Crn1, Lsb3 and Akl1 are all involved with scission or internalization of endosomes by interaction with the actin cytoskeleton (Weinberg and Drubin, 2011) and become more localized as internal punctate in ΔERG2 cells. It has been shown that the lipid content of the membrane is important for the binding of the RVS161 and RVS167 bar domains in a membrane raft type environment (Youn et al., 2010) and they display fewer foci in ΔERG2 cells. As with the internalization of endocytotic proteins in

ΔCHO2 cells, to dissect which stage of endocyotsis is defective and causes this phenotype, would require time-lapse analysis to investigate dynamics and co-localization studies.

3.4.16: Ergosterol biosynthesis; Lipid homeostasis proteins are affected in ergosterol mutant cells.

KES1 is a soluble protein transports ergosterol from the plasma membrane to the ER (Raychaudhuri et al., 2006) becomes miss-localized to dots in the membrane in the ERG6 condition and has a much weaker phenotype in the ERG2 mutants. Osh2 is also a sterol transporting protein located at the plasma membrane. It seems unaffected in ΔERG6 cells but becomes localized to one large punctate in ΔERG2 cells.

Additionally DNF1 and DNF2, phospholipid flippases important in maintaining the phospholipid gradients across the plasma membrane, also become miss-localized to an internal punctuate pattern (Figure 3.14).

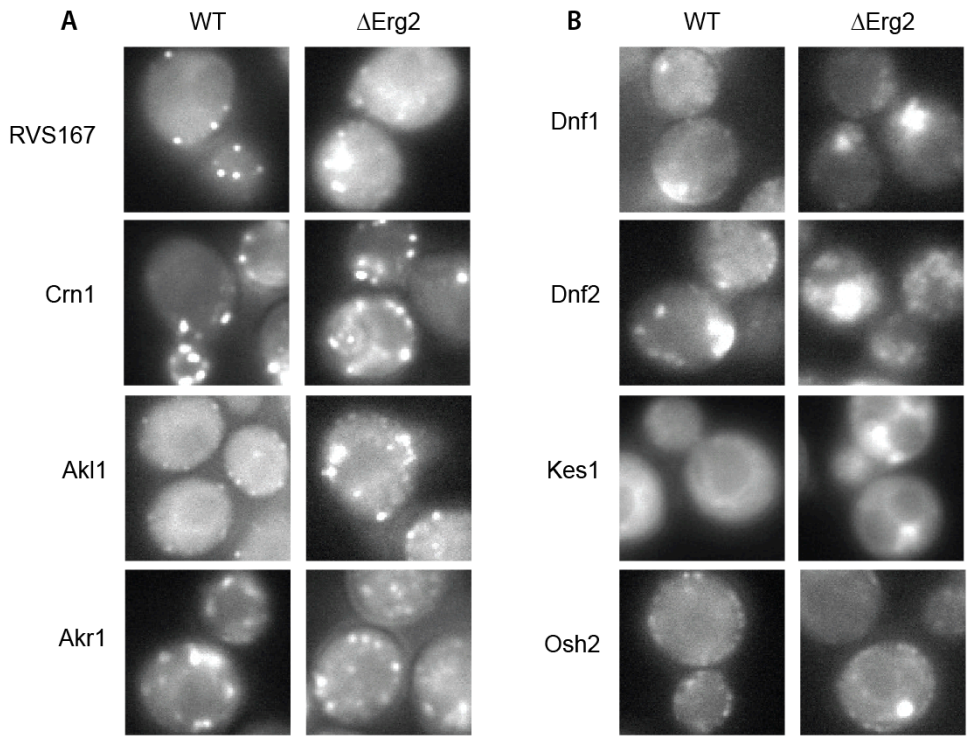

Figure 3.14: ΔERG2 cells have defects in endocytosis and lipid homeostasis proteins

A) Proteins involved with endocytosis become more internalized in ΔERG2 cells. B)The flippases Dnf1, Dnf2 and the oxysterol binding proteins Kes1 and Osh2 are miss-localized in ΔERG2 cells

Sphingolipid biosynthesis pathway perturbations

3.4.17: Sphingolipid biosynthesis; Known lipid concentration changes in PtnInP pathway perturbations

Perhaps the most interesting observation from the perturbations in the sphingolipid biosynthesis pathway is that aside from LCB1 knock-down cells, the others cause very few protein miss-localizations. LCB1 was targeted with a tetracycline inducible knockdown but this caused dramatic pleiotropic effects. As with the tet-STT4 strain, internal membranes are severely affected in this condition so the easiest changes to interpret involved the plasma membrane. Proteins are affected across most membrane involved cellular processes (Figure 3.15). These large scale proteome affects are to be expected as sphingolipids are major contributors to the lateral heterogeneity of membranes, and in particular the plasma membrane, and therefore for the organization of these processes. (Aguilera-Romero et al., 2014). These structural roles on the lateral organization of the plasma membrane are demonstrated by the effects on the MCC membrane domain (discussed in detail later). Additionally, many cellular processes use sphingolipids as signaling molecules so this can leave many miss-localizations difficult to interpret.

The enzymes downstream of Lcb1 are non-essential genes (with the exception of Aur1) so the cells can survive without the entire wild-type repertoire (Aguilera-Romero et al., 2014) (Figure 3.2). Interestingly, aside from Lcb1, there were eight other enzymes perturbed and in these only 14 proteins were miss-localized. This is unexpected considering the importance of sphingolipids in the cell. These are of particular interest as very little is known about the structural role of these lipids and no pattern has so far emerged from these protein miss-localizations (Figure 3.15).

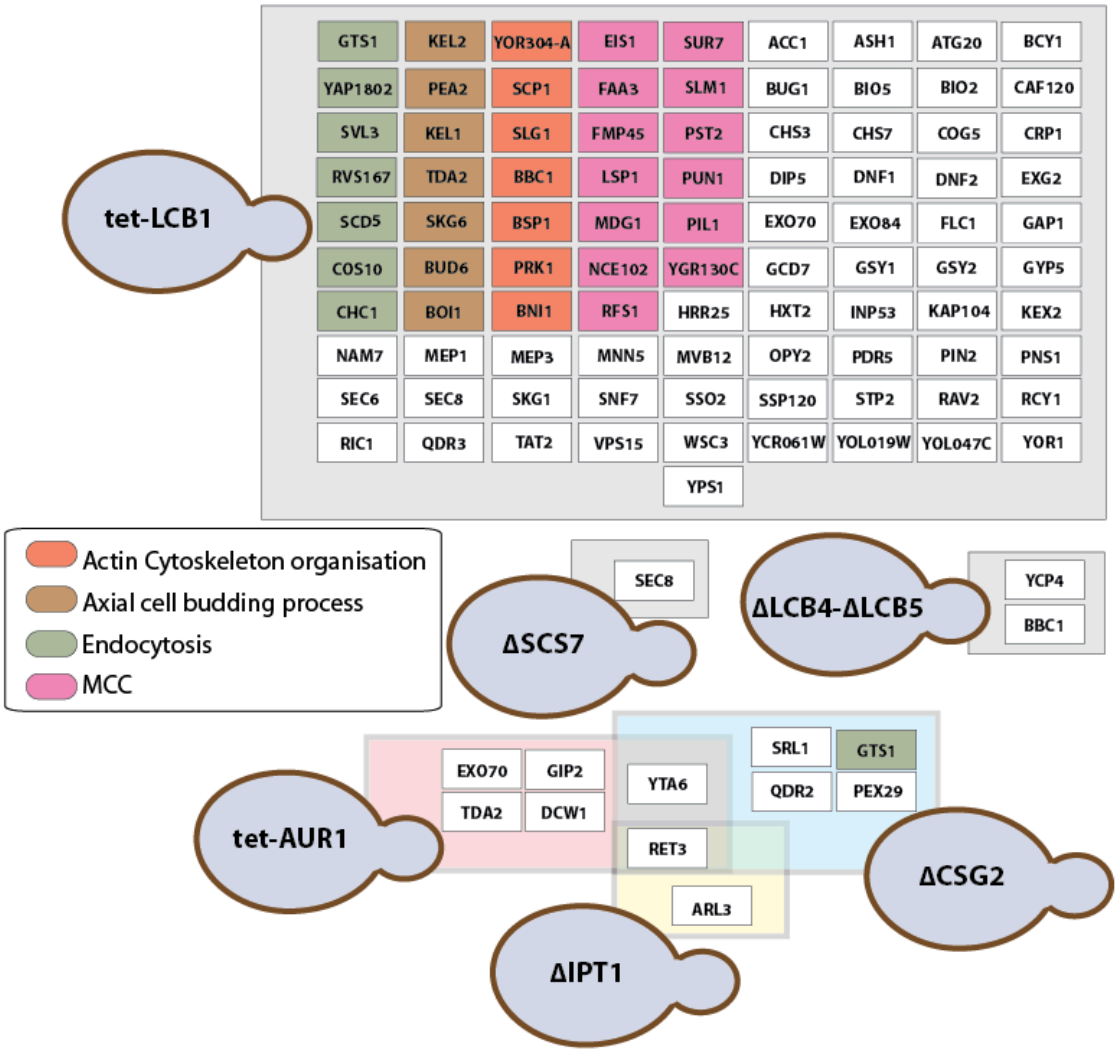

Figure 3.15: All proteins affected in the sphingolipid biosynthesis pathway perturbations

Perturbed cells are represented; tetracycline treated tet-LCB1 and tet-AUR1 strains and knockout strains of SCS7, IPT1, CSG2 and LCB4/LCB5. Bubbles contain proteins miss-localized in the adjoining condition. Proteins affected in more than one condition are in separate bubbles. Proteins involved in selected biological processes are indicated. The MCC (membrane compartment containing Can1) is a structurally defined region in the plasma membrane.

Inter-pathway features

3.4.18: The lateral heterogeneity of the plasma membrane is maintained by all lipid classes in concert

High-resolution imaging has shown that the yeast plasma membrane is segregated into at least three major domains which range from 'patchwork' to 'mesh-like' structures (Spira et al., 2012). The domain that is the best structurally described is the MCC (membrane compartment containing Can1) a distinct membrane domain that has an internally furrowed ultrastructure (Karotki et al., 2011). A complex interaction network between proteins and lipids maintains the MCC. Indeed, all of the major lipid classes have affected the localization of at least one protein from this 'organelle' in this screen..

The conditions causing the most widespread changes to the MCC are wortmannin treatment, tet-STT4, ΔSAC1 and tet-LCB1. The MCC is understood to be a lipid raft with a high concentration of sterols and sphingolipids (Karotki et al., 2011), so the affects due to LCB1 knockout are expected and known. However, there is little evidence about PtnInP roles in its MCC maintenance. It had been shown that knock outs of MSS4 affect the binding to the membrane of the MCC proteins Pil1 and Lsp1 (Karotki et al., 2011) but these proteins are not affected in this screen. Interestingly, ΔSAC1 cells have affects similar to tet-STT4 and wortmannin treated cells; a PtnInP phosphatase knockout having a similar phenotype to a PtnInp kinase knockdown with respect to the MCC.

Proteins localized to the MCC, both integral membrane proteins and peripheral proteins, either become internalized or localize to a few membrane foci in these conditions (Figure 3.16). Investigating the lipid content of these new membrane foci would be of interest to see if a new membrane domain is formed that may explain membrane domains in wild type cells.

3.5: Conclusions and perspectives

By its nature the work on this project is quite descriptive but the confirmation of known lipid-dependent protein localization changes and the discreet pathways affected these conditions shows its quality. This dataset is a resource that can also be mined for information to look for specific interactions of individual proteins. For example, the finding that Osh6 is miss-localized in ΔPSD1/ΔPSD2 cells has already has already contributed to the finding that Osh6 is in fact a PtdSer transporter (Maeda et al., 2013).

The study of the membrane interacting fraction of the proteome remains challenging due to the difficulty of working with insoluble components. The screen described here does not directly describe protein-lipid or protein-protein interactions it does inform us on which processes or groups of proteins are affected by certain lipid depletions. In many of the proteins affected if would be useful to use lipid probes to investigate if the miss-localization of the proteins correspond to the miss-allocation of specific lipids.

Almost all of these strains have been used to investigate protein localization before on the scale of a hand-full of proteins. Here, this dataset allows for a better understanding of the whole proteome affects.

The information allows the development of hypotheses and the design of further studies in this area to provide a better understanding of the structural and mechanistic roles of lipids.

3.6: Materials and methods

Media & Chemicals

Geneticin (G418, GibcoBRL) 200 mg/lt; Nourseothricin (clonNAT) Werner BioAgents, Jena, Germany; 100mg/lt, HygromycinB (Cayla, france); 300mg/lt, Canavanine (Sigma)- 50 mg /lt; 50 mg/lt , Thialysine (S-(2-aminoethyl)-L-cysteine hydrochloride, (Sigma A-2636). Sporulation medium (2% agar, 1% potassium acetate), Haploid selection plates (SD medium lacking His/Arg/Lys/leu but contains canavanine + thialysine), Final Strain selection plates (SD medium lacking His/Arg/Lys/leu but contain canavanine + thialysine + Hygromycin or Deletion antibiotic marker).

Query strain generation

Query strains were generated from BY4741 background, Y7039 strain (MATα ΔCAN1:: STE2pr-LEU2 ΔLYP1 UTA3Δ0 LEU2Δ0 HIS3Δ1 MET15Δ0 (Baryshnikova et al., 2010). A gene specific 65bp oligonucliotide was designed for construction of deletion, tet-strain and gdp-strain construction. The PCR products with antibiotic selection marker were transformed into yeast for gene disruption or promoter replacement by homologous recombination as described in (Janke et al., 2004). Gene disruptions were then confirmed by PCR using gene specific primers.

Tet-titratable promoters - The system allows expression in the absence of tetracyclines. Doxycycline was added and prevents transcription of tagged gene. Test cells with doxycycline. each strain was qualified by its growth phenotype on YPD with 10 μg/mL doxycycline in the absence of doxycycline, the Tet-promoter is fully activated. Addition of doxycycline in a titratable manner allows for down-regulation of the promoter until the gene of interest is no longer expressed at detectable levels.

Array Strain selection

A list of membrane associated proteins was generated using the Saccharomyces Genome Database (SGD) and the yeast GFP database available (Huh et al., 2003). The following localization categories from the Yeast GFP database were chosen; Cell periphery, Punctate, Actin, Endosome, Lipid particle, Bud, and Bud neck. SGD Gene ontology annotation terms Plasma membrane, Actin, Endocytosis, Exocytosis, Eisosome, Bud, Bud tip, Bud neck and lipid particle were used for selecting further protein of protein involved in Plasma membrane processes. Together 805 proteins matched these criteria, of which, 648 proteins were available

as GFP fusions strain. Genotype; MATa HIS3Δ1 LEU2Δ0 MET15Δ0 URA3Δ0 GFP::HIS3MX6) (Appendix F.II).

Strains crossing

Query strains were grown overnight in 50ml YPD media and then transferred to empty PlusPlates (Singer instruments). Query strain were pinned to YPAD plate using 384 well/plate disposable plastic pads with the help of an automated pinning robot (Singer instruments, UK) to create a 384 well format. Cells were grown for one day at 30°C. Similarly GFP strain library ((Huh et al., 2003), Invitrogen collection, USA) were revived in 96 well microtitre plates (Nunc) with the help of a Singer RoToR HDA robot using sterile liquid handling 96 long-Pin RePad. The chosen subcollection of the Plasma membrane GFP strain library sub-arraying into 96 well plates using the liquid handling robot. GFP strains were then transfer to YPD plates and arrayed into the 384 well format using 96 long-Pin RePad. For mating the GFP library with the query strains, both strain were pinned together onto fresh YPD plates and grown at 30°C incubator for one day.

Diploid selection

This resulting mated a/α diploid cells were pinned onto Diploid selection plates to prevent haploid cell growth depending on the haploid marker (eg. on synthetic dextrose medium SD medium His/Hygro plates as for GFP-HIS marker, Deletion – Hygro marker), and diploid cells array were grown for 2 days. The query strain spotted in an empty well is the diploid selection control.

Sporulation

Then diploid cells were pinned to sporulation medium plates (2% agar, 1% potassium acetate) with the robot and incubated for 7 days at 25°C.

Haploid selection

After sporulation, to select haploid cells with mating type a, sporulated cells were pinned on to haploid selection medium SD medium lacking leucine/ histidine /arginine/lysine but containing canavanine & thialysine and incubated at 30°C for 2 days. Canavanine and thialysine allow selective growth of haploid cells while leucine allows growth of only MAT A progeny due to

expression of leucine gene under mating type "a" promoter (STE2pr-LEU2). Wild type (BY4741) strains will shows no growth on leu(-) media plates as a control.

Final selection of GFP allele and deletion crosses

For double mutant strain selection having GFP allele & deletion allele, haploid strains (Mat a) were pinned onto SD medium with the selection marker (eg-GFP-HIS$^-$ marker + Deletion–Hygro marker) along with haploid selection media components (medium lacking leucine/arginine/lysine but containing canavanine & thialysine for killing any potential diploid cells and only selection of haploid progeny). Strains were then sub-arrayed into 4 x 96 well format for microscopy analysis. Several strains were randomly picked from the array and confirmed by PCR using gene specific primers for deletion & GFP Fusion proteins for their quality. For glycerol stock the strain were grown in 150 ul YPD media(96 wellplate format) overnight and 75ul glycerol((80%) were added next day and then mix and shifted to -80c freezer for storage.

Wortmannin treated cells

Wortmannin Ready Made Solution in DMSO (W3144-250µL, Sigma) was added to cells to a concentration of 1 µM, for one hour prior to, imaging experiment..

High throughput imaging live-cell imaging of Perturbed GFP Libraries

Prior to imaging Perturbed GFP library strains were grown were inoculated on SD solid medium 96 well sterile tissue culture plates (Nunc, Cat no. 167008) containing complete minimal media without tryptophan and histidine (trp(-) low fluorescence media ,his(-) for selection of GFP strains) by the liquid handling singer RoToR HDA robot. Cells were grown overnight at 30ºc.

96 well glass bottom black plates (Zell-kontakt, Cat no.5241-20, Germany) were coated with 60ul concavalin (Sigma) solution (1mg/ml) and incubated for 45 minutes at room temperature. Wells were then washed 3 times with complete minimal medium. Tet-strains were incubated with 0.1ul deoxycyline for 45mins before imaging.

Cells grown from the perturbed GFP libraries in 96 well plates were diluted to OD_{600} of 0.1 and incubated for 6 minutes on the plates. Unbound cells were washed off 3 times with minimal media, and 100 µl was added for imaging. 72 wells were used for imaging as outer wells A1-A12 or H1-H12 were avoided to avoid edge artifacts.

Cells were imaged using an Olympus fluorescence microscope system (Olympus IX81) at 30°C. 16 bit images were acquired with a 100x Oil objective with an NA of 1.45 and low-noise ORCA-R camera (Hamamastu) and MT20 illumination system, Uniblitz Electro-Programmable Shutter system. For automatic image acquisition a custom-built, in-house Scan R (Olympus) software version 2.2.0 was used. Object detection based on a gradient function and combined with an image segmentation algorithm were designed in Scan R to autofocus yeast cells in brightfield image mode. The brightfield autofocus function allows for finding and focusing cells in the right focal plane and minimizes the bleaching of fluorophores as compared to fluorescence autofocus modes. Each well was imaged at 9 different positions using both bright field and fluorescence microscopy.

To cover both low and high abundant proteins, two different fluorescence exposure time (2000ms) images were acquired. Each image was stored systematically with the well number, well position, image type (BF, FL) and exposure time information attached to each image file. High exposure time images were used for data analysis.

Two yeast GFP strains (PMA1-GFP, SLM1 –GFP) with good fluorescence signal were used as imaging controls on each plates to access any microscope fluorescence acquisition changes over different plates (Plate to plate variation). PMA1-GFP is abundant plasma membrane protein distributed throughout cell surface while SLM1-GFP form stable patch structure on membrane. These imaging controls strain were placed in duplicate at different positions within same plate (C2, C12, G1, E6 wells) to also access any within plate specific fluorescence acquisition variation & two blank well (E5) used to measure normal plate/media background

Automatic assessment of image quality

Microscopy images were processed to detect low quality images, out-of-focus, black images, overcrowded wells, saturated and unclean images using automated classification approach. As each bright field/fluorescence image pair was quantified by a set of 52 descriptors, including mean intensity, quantiles, standard deviation, saturation, spatial frequencies and granulometry features for bright field and fluorescence channels. A binary support vector machine [Boser, Guyon, Vapnik, 1992] trained on a manually annotated set of 1089 image pairs, was used to predict image quality. Prediction accuracy, estimated by 10-fold cross-validation on the training set, was about 98 %. 86.82 % of the images were of good quality and suitable for analysis.

Hit calling

The images were blinded and scored manually for localisations; cell periphery, mitochondria, bud tip, bud neck, Nucleus, punctate, single punctate, membrane punctate, vacuole, and vacuole membrane. All conditions were scored independent parallel by two scientists and the union of those localizations called is retained.

Segmentation

Next, we determined cell boundaries in the bright-field images. The method uses a matching pursuit algorithm which seeks the sparsest projection of a multidimensional signal over the complete dictionary of ring-shaped objects of different sizes. For each image, pixel intensities were linearly scaled using quantile statistics and cell edges were accentuated by global thresholding. Cell centres were detected using a variant of the matching pursuit algorithm [Mallat, Zhang, 1993], scanning for positions that locally maximise the inner product of the image with a dictionary of ring-shaped objects of 16 different sizes. Thereby, we identified 12,779,011 cells in 24481 wells (on average 522 cells per strain).

Segmentation was performed fitting different diameter circles over the bright field image. This was achieved through applying Hough circle transformation on a bright field image with radiuses 61:121:4 pixels (crown.steps) and square sliding window 20px wide (locmin.threshold.width) with a 0.2 threshold over window mean intensity. If multiple radiuses satisfy circle detection threshold bigger radius is used to fit the cell. Next we mark area in direct-segmented object proximity (8px) as Membrane area.

Membrane foci counting

We identified structures located in exact proximity of cell outer membrane. These membrane foci were detected by applying 7x7px adaptive filter with 10% threshold. From obtained structures we removed objects with area smaller than 10px (noise), uniform membranes, objects with eccentricity over 0.9 or with perimeter bigger than area (the last two steps were aimed at removing bright budding sites).

Benchmarking of the algorithm was performed by manually annotating 2681 random cells and calculating Pearson Correlation Coefficients between the manual annotation and prediction (R^2: 0.63, CI: [0.61, 0.65], p-value < 0.01). Additionally, reproducibility of membrane foci detection algorithm is supported by comparing average number of detected structures in main and duplicated screens (R^2:0.91).

Changes in the mean number of membrane foci between reference and condition was tested against the null hypothesis (the variation of MAS change between main and duplication screen was used to model noise). P-values were adjusted using Benjamini & Hochberg multiple testing correction and cutoff of 0.05 was used. Changes below this cutoff were marked as "Unchanged". Changes in conditions with not enough data and wild type were marked as "Not Defined"

CHAPTER 4: Overall conclusions and perspectives

4.1.1: Contribution of this work to the field of high-throughput biology

The two projects described in Chapter 2 and 3 show the power of high-throughput biology to give biological insights. These projects approach the field of interactomics from two different angles; one investigates protein-protein interactions in simplified in-vitro mixtures from the native source, the second investigates protein-metabolite interactions in the entire cell system perspective using imaging techniques.

The project on the *C. thermophilum* proteome integrates high-end biochemical techniques and to applies them to investigation of native biomolecular interactions. Advances in MS and electron microscopy based technologies have paved the way for this project which adds a structural dimension to large-scale protein-protein interaction mapping. This project also addressed some of the issues of solving structures from natively derived material and could be developed further for 'high-throughput structure solution'.

It has been predicted that only ~6% of the currently discovered interactions in the cell have are understood at molecular detail so high-throughput structure solution is required (Stein et al., 2011). One possible way to up-scale this 'visual proteomics' pipeline is to take advantage of the information on protein complex separation in studies such as Havugimana et al (2012). There they fractionated protein complexes from human cells using 13 different techniques taking advantage of different physiochemical properties. The complexes were detected by MS and therefore the dataset is ripe for the foundation of a visual proteomics study; it would be possible to identify the structural signatures of large protein complexes by EM in different backgrounds using these fractionation techniques. This would allow correlation with the MS. High throughput structural study of native protein complexes may then be possible.

Furthermore this technique can be further applied to mapping protein complexes from proteomes under different conditions such as cell type, developmental stage, and external stimulus. It can also be used to observe the importance of post-translational modifications on the organization of the proteome into complexes. I have done some work to that end, summarized in Appendix A.VI.

Binding pockets in proteins associated with disease have long been drugged with small molecules but protein-protein interfaces are seen as a great potential source of druggable sites

in the cell. The continued addition of molecular detail to Interactive maps by structural studies will allow a greater understanding of disease states and potentially druggable interfaces.

The project on the affect of lipids classes on the organization of the *S. cereviaise* proteome shows the power of high-throughput cell biology to draw broad systematic understanding of cellular function. The cell's membranes and the proteome associated with are usually not soluble and this has therefore proven to be a difficult area of research. The screen described here does not directly describe protein-lipid or protein-protein interactions it does inform us on which processes or groups of proteins are affected by certain lipid depletions. This information will inform the design of further studies in this area to provide a better understanding of this challenging part of the cell.

4.1.2: A note on designing high-throughput experiments

The transition in biology from the molecular level to the systems level is continuing apace and is revolutionizing our understanding of complex biological systems.

Both projects presented here involve extensive collaboration between 'wet-lab' molecular biologists and bioinformaticians. Unfortunately, often each has a poor understanding of the strength and limitations of the other's field. Communication is key to a successful project and the data analysis plan should go hand-in-hand with the experimental design of these large and expensive projects i.e. it should be there from the beginning.

Ideally a pilot project would be finished to the end of the data analysis pipeline to discover any weaknesses in the experimental design. This is made doubly important, as mistakes in high-throughput screens are by their nature more costly than their one-at-a-time counterparts. Without this due diligence projects can produce sub-optimal data for computational analysis and therefore fail to achieve all of their potential.

4.1.3: Future advances in the field of experimental systems biology

Molecular biology as a field is adept at describing, often in exquisite detail, the abundant soluble proteome, as in the project on the *C. thermophilum* proteome. Much work still needs to be done on the other parts of the cellular interactome, such as membrane-bound proteins and protein-metabolite interactions.

Major advances are being made in three converging fronts:

- The first will be to simply detect interactions that have so far evaded description by high-throughput interaction studies. The technology and chemistry of proteomics is advancing in this area (Boersema et al., 2015) and is beginning to allow the identification of metabolites from complex samples. Also, transient interactions and those interactions between non-soluble proteins are being analysed with crosslinking studies such as ours. It is also important to begin to address more the biological functionality of the enormous array of PTMs and PTM sites so far detected and how they impact the interactome. For interactions that still evade biochemical detection, theoretical and computational methods for the prediction of novel protein–protein interactions are being developed (Mosca et al., 2013; Tyagi et al., 2012; Zhang et al., 2012).

- The second will be further advances in adding molecular detail to interaction maps in a high-throughput manner. Currently many parts of the interactome are populated with simple qualitative information 'identifying X interacts with Y'. Future studies will need to discover the topologies, confirmations and molecular details of these interactions as they occur in vivo, with the molecular crowding and local cellular organization taken into account. This may occur with major advances in light microscopy (Wachsmuth et al., 2015), chemical crosslinking MS, electron tomography (Han et al., 2009) and their combinations. We may discover that much of the current understanding of protein interactions us wholly inadequate due to our often-necessary use of in-vitro systems.

- Finally there will need to be advances in our study of molecular dynamics of these interactions in a high-through-put manner. Currently much of our understanding of protein-protein interactions is qualitative but the dynamics of the interaction are often not considered except for in a few intensively studied cases. Again, here we need to consider important cellular components often stripped away in in-vitro studies such as local protein, metabolite and ion concentrations. There has been much advancement in the field of FRET in this area (Bacia et al., 2006). Mass spectrometry based technologies are also providing information on the dynamics of these interactions which was previously not possible (Feng et al., 2014; Rajabi et al., 2015).

We may never have enough information to completely model a cell in all of its complexity but with high-throughput molecular biology experiments we are generating enough information to generate models with enough detail to answer some of the most important questions such as the nature of disease and how to effectively treat it.

APPENDIX A : Supplemental information for Chapter 2

A.I Growing and lysing *C. thermophium*

For our study, we developed a lysis protocol for native protein extraction from *C. thermophilium*. The culture is grown from spores in Lysogeny broth (LB) liquid culture with gentle agitation, the culture grows as hyphae to form a ball of culture. Experiments were conducted using the exterior cells that are the youngest and healthiest (Figure 2Ai and 2B).

Three methods were tested for lysis of the culture - cryo-grinding, bead beating and sonication. Cryo-grinding followed by bead-beating produces the greatest yield of protein and complexes remain intact (read from the UV signal response in SEC) (Figure 2.2C).

As *C. thermophilum* is a thermophile, it is possible that some of its proteins are only optimized for high temperatures. We investigated lysis and SEC at room temperature, but in the ~2.5 hours that this whole process takes, all of the protein had become degraded (results not shown).

A.II Reproducible SEC for high-molecular weight proteome

The chosen column was a Biosep-SEC-S4000 as was previously used by Kristensen et al (2012). To enrich for the high molecular weight complexes, after the cell lysate was clarified by ultracentrifugation, it was then concentrated in a 100kDa cut-off spin filter, reducing the volume 25 times from ~5ml to 0.2ml. The final concentration varies between 20-30mg/ml when starting with 4g - 6g of wet culture. An estimated 2/3 of the proteins were lost through the MW cut-off filter (estimated by Bradford).

100µl of 20mg/ml clarified cell lysate was separated on the SEC column and a Comassie stained gel of the fractions showed tight peaks and good separation across all size ranges between 200kDa - 5MDa (Figure 2.3). The column began to lose separation efficiency with complexes <200kDa, as seen by Comassie gel, and the fractions became very complex. For this reason, we processed the first 30 fractions in this project covering the range 5MDa-200kDa

UV 260nm/280 nm signal was used to get an approximate estimate of protein elution over the fractions. However, the absorbance at 280nm and 260nm do not correlate very well with protein abundance as the absorbance of RNA and DNA, present in these fractions, absorb more at 260nm than at 280nm.

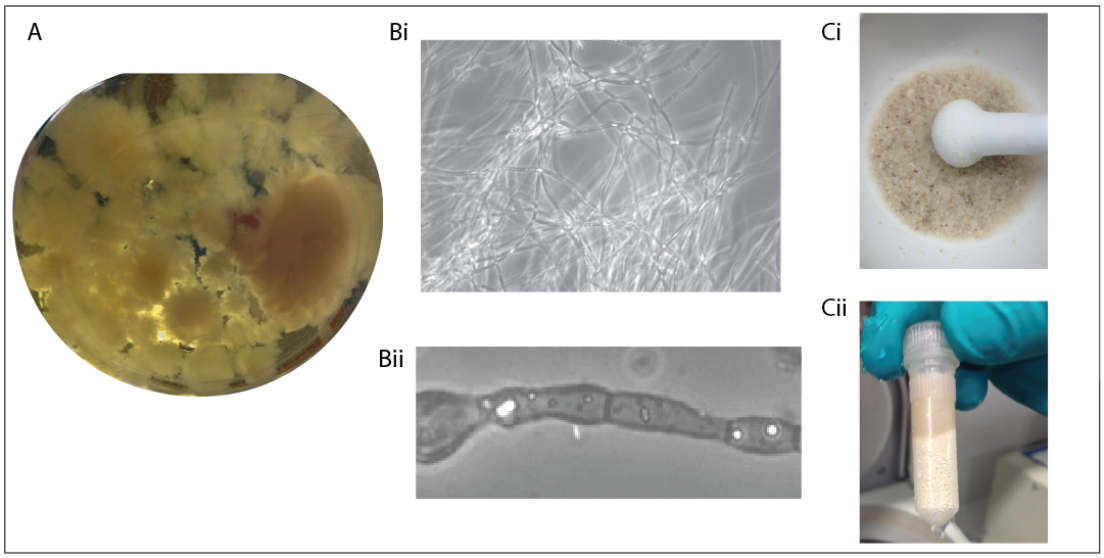

Figure A.1: C. thermophilum lysis

Figure showing growth and lysis of *C. thermophilum*. **Ai)** Optimal culture is fluffy and has a cream color. **Bi)** & **Bii)** Healthy culture under the light microscope, the culture grows in long hyphae with cells separated by septa. **C)** The optimal lysis procedure, first **i)** grinding the culture frozen in liquid nitrogen into small pebbles and **ii)** added to tubes containing silica beads and lysis buffer for lysis by bead beating (see also: Materials and methods).

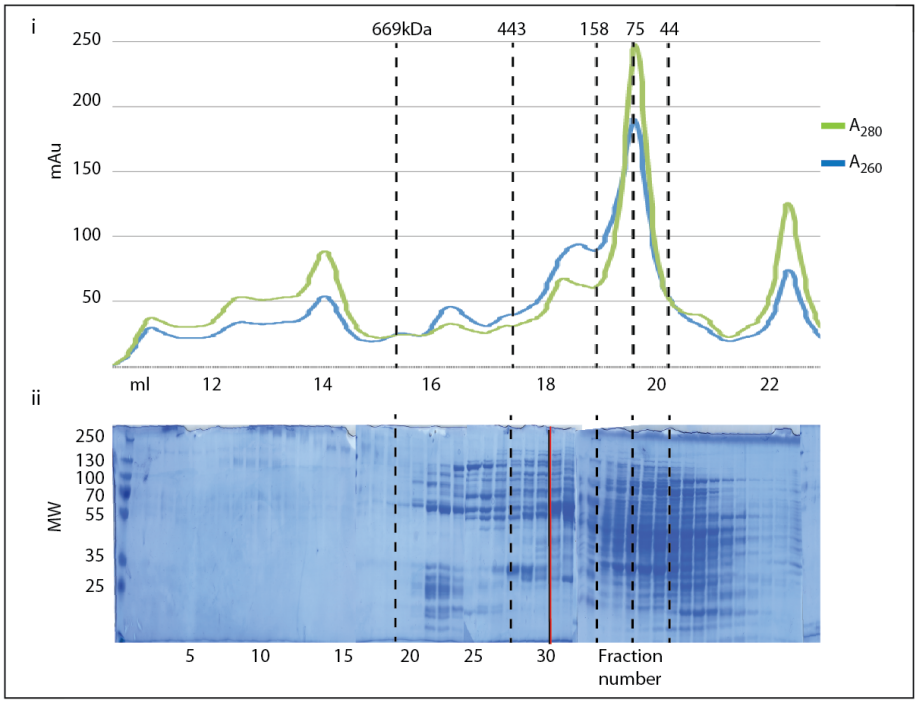

Figure A.2: Elution profile of lysate over Biosep SEC-S4000 column

Figure i) A representative elution profile from a SEC run of the CT high molecular weight proteome. Absorbance at 280nm and 260nm were recorded and with the indicated molecular weight standards (Sigma). ii) A representative gel showing each 250ul fraction. The first 30 fractions were retained for further analysis (indicated by red line).

A.III Generation of protein elution profiles using quantitative mass spectrometry

SEC was performed in biological triplicates and all fractions were analyzed by quantitative MS. To produce elution profiles the intensity of the MS ion signal of individual proteins in each fraction is normalized by converting to an iBAQ score (intensity based absolute quantitation) (21593866). The iBAQ scores are plotted per fraction to generate elution profiles for individual proteins.

Figure 2.4 shows the reproducibly plots between the three replicates 'A', 'B' and 'C', each pair having a Pearson's correlation coefficient of >0.82. The total number of proteins detected was 1281, of which 1176 were found in more than one replicate

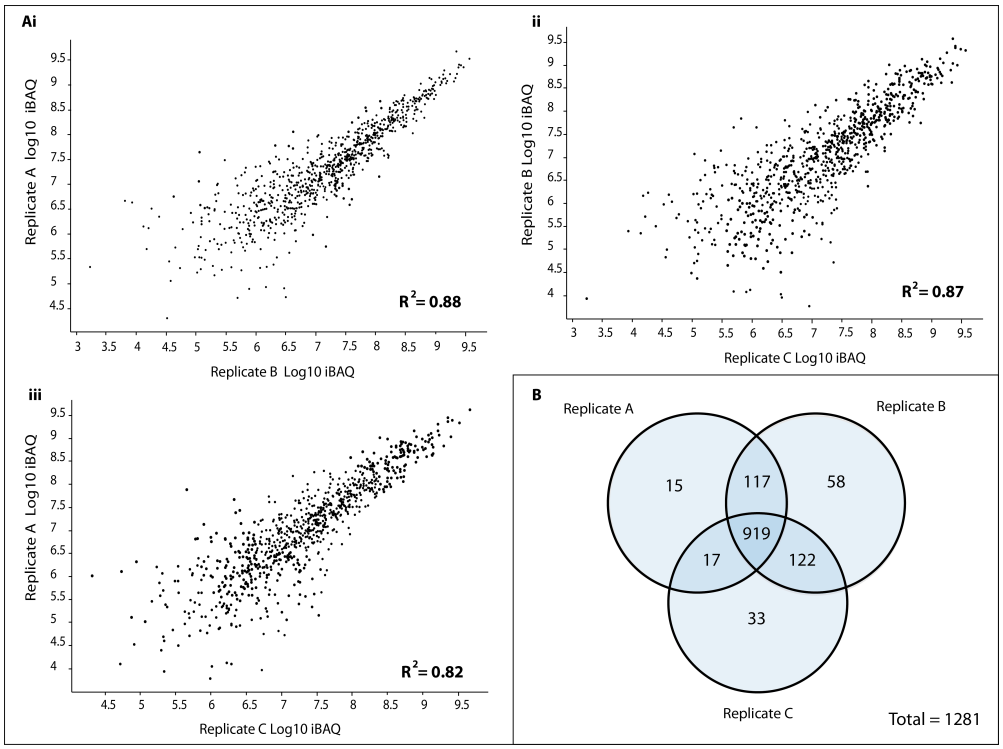

Figure A.3: Protein identification from 3 biological replicates by MS

A) Comparison of protein abundances of proteins discovered in each of the three replicates with the Person correlation coefficient (R^2 quoted). **i)** Replicate A vs B, **ii)** Replicate B vs C and **iii)** Replicate A vs C. **B)** The numbers of proteins discovered in each of the three replicates.

A.IV Assembling a benchmark of known protein complexes based on AP-MS data from *Saccharomyces cerevisiae* and PBD structures from various species

For prediction of protein complexes from the identified proteins, first a quality benchmark of true positive and true negative interactions must be produced based on conserved known orthologous protein complexes.

Coelution of proteins does not measure direct protein-protein interactions but rather can assign groups of proteins to a complex membership. As a benchmark, the most suitable external data is produced from AP-MS studies as in this technique complexes are reported without describing direct interactions. There are several studies that have reported AP-MS complexes from the well-characterized *S. cerevisiae* and several papers have been published attempting to unify this information. The study by Benschop *et al* (2010) identified common protein complex members from several *S. cerevisiae* AP-MS studies and annotated them as 'core' complexes.

To map *C. thermophilum* proteins to these complexes, all proteins were mapped to their *S. cerevisiae* orthologous proteins by mapping clusters of orthologous proteins across all annotated fungi (Amlacher et al., 2011) (see Materials and methods). This generated a list of one-to-one, one-to-many and many-to-one *C thermophilum* proteins mapped to their *S. cerevisiae* orthologs that were then mapped to the known complexes from (Benschop et al., 2010).

To be considered for the benchmark >50% of the subunits of any complex needed to be identified in >1 biological replicate and the elution profiles of the subunits were manually checked for co-elution in at least some part of the elution profile. In total, 270 proteins were assigned to 54 complexes for the benchmark (Figure 2.5).

To supplement this benchmark, and to produce a list of structural homologs useful for the structural study component of this project, each identified *C. thermophilum* protein sequence was subjected to a simple BLAST search against the Protein Data Bank. In total 427 proteins had homologs discovered with >70% sequence coverage and >30% sequence identify and can therefore be modelled during structural studies later in the project.

60 heteromeric protein complexes with >50% of the subunits identified were discovered and manually checked for co-elution in at least some part of the elution profile. Unlike proteins from AP-MS studies, these protein complexes have known stoichiometries. Together these complexes are merged into a single benchmark totaling 64 complexes containing 358 subunits (Figure A.5) (Appendix E.II). A by-product of this search of all the proteins in the PDB also allowed assignment of 137 homomeric complexes and their stoichiometries. The elution profiles of these conserved complexes reveals that many of them elute larger than is predicted (Figure A.4)

A benchmark of true positive and true negative interactions was generated from this list, to be used for the prediction of novel interactions. A matrix was produced assigning all-to-all possible within a complex as true positives, even though many of these may not be direct protein-protein interactions. For true negative interactions, all-vs-all possible interactions subunits between these benchmark complexes were used. Complex pairs excluded for generating the true negatives dataset were subunits of the ribosomes and their binders, subunits of the proteasome, Splicosomal subcomplexes, vesicle coat complexes and the RNA polymerases. In total of all the possible 1,263,376 interactions from this dataset, 3044 were assigned as true positive and 57,149 were assigned true negative.

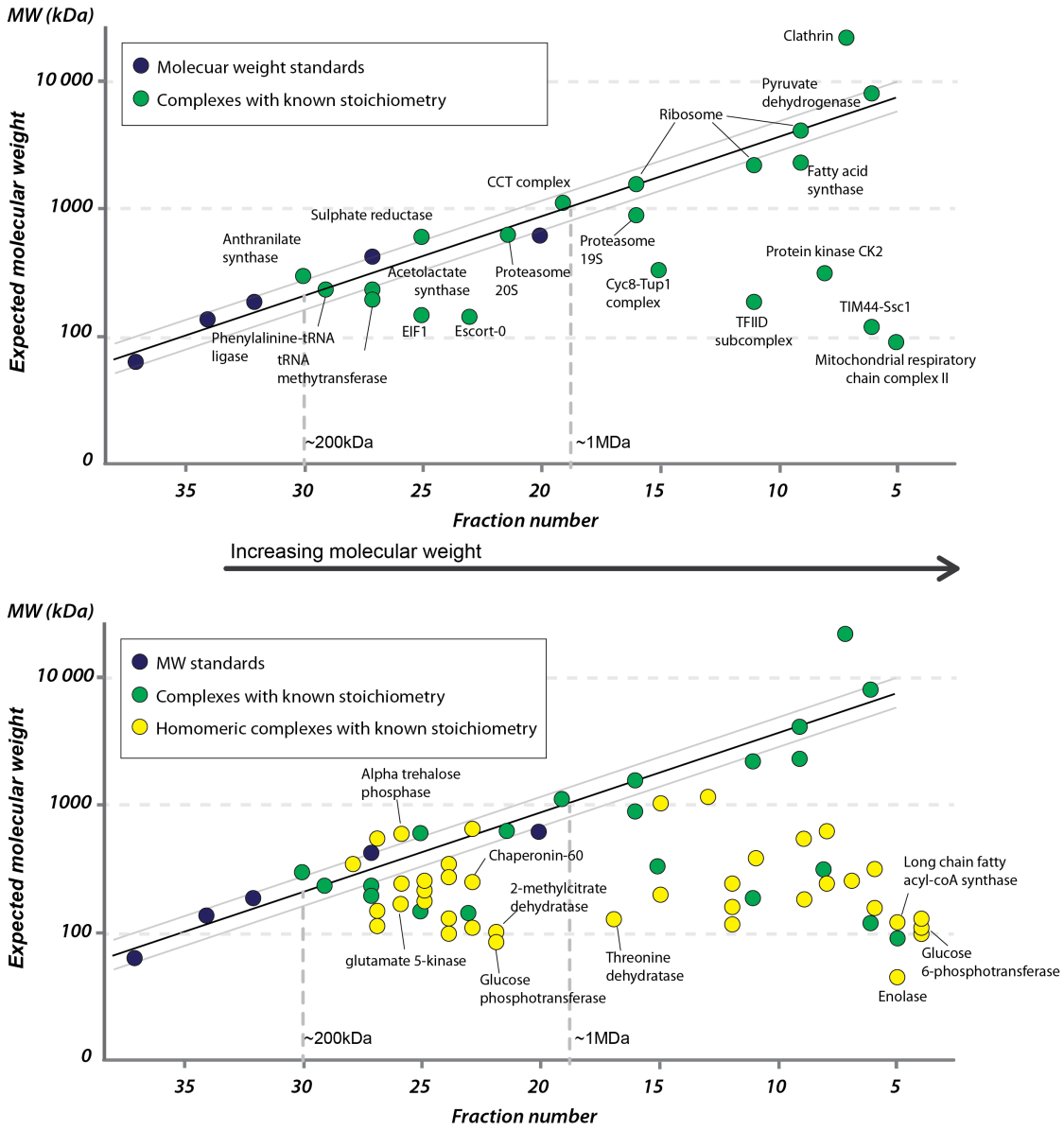

Figure A.4: Elution profile of known protein complexes shows there is much to be discovered

The SEC column is calibrated with known MW standards that have spherical shapes and thus have good MW-hydrodynamic radii correlations. MW strongly correlates with elution in the range of 200 kDa -1.5 Mda where this column is efficient. The black line is the expected elution and the grey lines represent a 10% deviation from this curve. Complexes are plotted which have PDB structures and therefore known stoichiometries from other species and elute with a single defined peak. Interestingly homomeric complexes commonly deviate from the expected elution and therefore may be subunits in heteromeric complexes.

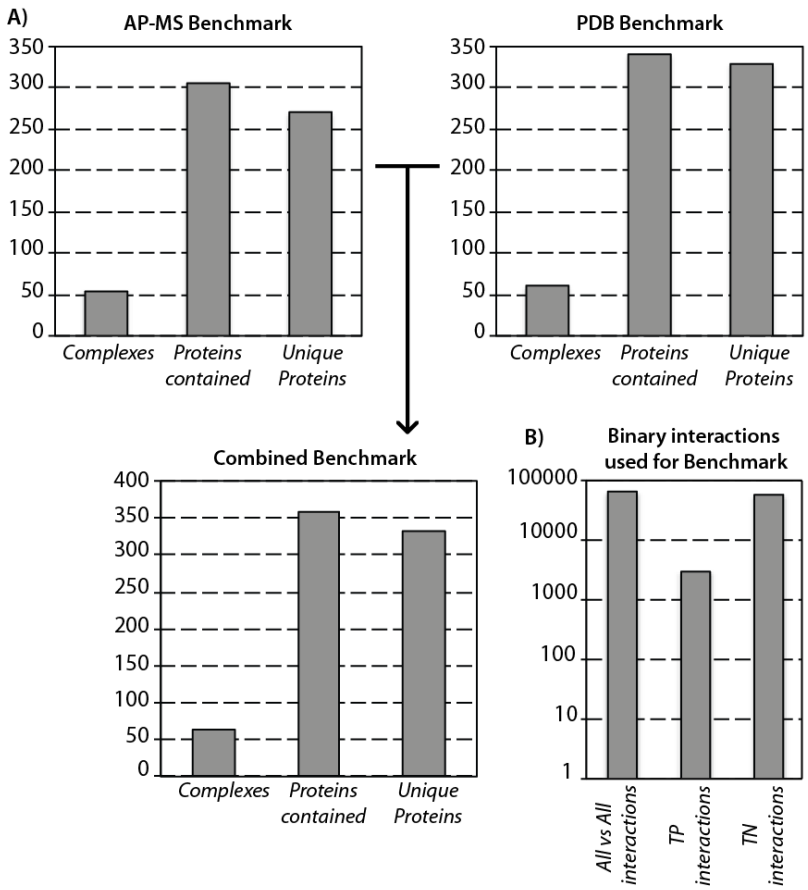

Figure A.5: Assembly of benchmark of orthologous protein complexes for complex prediction

A) 54 Orthologous *S. cerevisiae* protein complexes from (Benschop et al., 2010) with >50% of the subunits were used for the benchmark. These contained 306 proteins but some of these are shared so there are 270 unique proteins contained. The PDB benchmark contained 60 complexes containing 340 proteins and 329 are unique. The final combined benchmark contains 64 complexes containing 358 proteins and 333 are unique. **B)** For benchmark interactions all proteins inside the 64 complexes were called as True Positive interactions. Interactions of subunits between these 64 complexes were used as True Negative interactions, excluding complexes that are known to interact.

A.V Prediction of protein complexes by co-elution profiling and integrative network analysis

To discover new complexes and complex subunits we applied an integrative approach to discriminate which proteins are present in the same complex.

Contrary to published cross-correlation methodologies, which deconvolute elution profiles into their underlying peaks (Kristensen et al., 2012), a simple cross-correlation approach worked best in our hands (algorithm available upon request). We compared each of the elution profiles in a pairwise manner and generated a cross-correlation coefficient (CCC) that described how similar two elution profiles are (see Materials and Methods). Although this approach may introduce certain bias (e.g. multiple binders might be missing), a cross-correlation value > 0 is, surprisingly, calculated for more complex cases. For example, chance coelutions of complexes of the same size, or of single proteins eluting with a high MW due to binding to DNA or RNA are not necessarily missed, but rank lower compared to "cleaner" coelutions (e.g. CCT chaperonin complex proteins). Another limitation of this approach is that proteins present in many complexes will have multiple peaks and this more complicated elution pattern caused low correlation coefficients to each complexes (e.g. G0RY12 is present in both the 19S proteasome and the COP9 signalosome and therefore has a double peak), but again, such coelutions are subsequently considered in network construction.

To address the abovementioned limitations and, therefore, discriminate between spurious coelutions, an integrative method was developed based on the addition of external data and then using machine learning to decide the probability of each possible interaction. This approach takes advantage of the fact that physically interacting proteins are often conserved across species (Zhang et al., 2012).

For creating the algorithm to classify experimental coelutions into complexes, we integrate three datasets:

1- Simple cross-correlation scores from pairwise cross-correlation of elution profiles. For the simple cross-correlation procedure, the profiles from each protein eluting in >3 consecutive fractions were correlated against the others and the Pearson's product-momentum correlation coefficients were generated as a measure of similarity. Proteins with elution in <4 consecutive fractions did not have enough information to provide meaningful CCC scores. For the 974 protein species retained, 473,364 possible coelutions were calculated.

2- Predicted interfaces from homologous proteins in the Protein Data Bank. 1553 predicted interactions were found using the FIST algorithm when applied to the identified proteome (Aloy and Russell, 2002).
3- Orthologous interactions from *S. cerevisiae* were extracted from the STRING database without the physical interactions to avoid circularity issues. These interactions have been annotated by co-expression, domain co-occurance, gene neighbourhoods, co-citation, functional relationship, and genetic interactions. The STRING database gives a confidence score to each type of information (Franceschini et al., 2013).

All external data was mapped onto interactions with a CCC of >0 to avoid interactions being called that lacked coelution data from this study. We then developed a machine learning procedure using a random forest algorithm to assign weights to each interaction found using the co-elution function. In this way, usage of known interactions to train and validate the function will subsequently be used to discriminate higher confidence interactions that can be retained and used for clustering into predicted and novel complexes. The random forest algorithm was trained on the 64 Gold Standard complexes to weight each of the three variables (see materials and methods). The STRING score was assigned the most weight, followed by the FIST score and finally the CCC (Figure 2.3A), although weighting is not significantly biasing any of the variables. 30% of the TP and TN interactions were withheld from training the algorithm; the approach was highly predictive for these benchmark interactions as is seen in the Receiver Operator Curve (Figure 2.3B).

A random forest probability was assigned to each possible interaction. Interactions with a probability (P_{pred}) >0.5 were assigned 'Class 1', meaning they were potential true interactions. All the probabilities assigned to Class 1 were plotted and a Gaussian curve was fitted. A threshold of 2 standard deviations greater than the mean was applied to discard ambiguity of classification. Interactions with scores higher than 2 sd from the mean contained 736 proteins involved in 6146 interactions; each interaction had a machine learning probability of >0.84 calculated across the three datasets (Figure 2.3C).

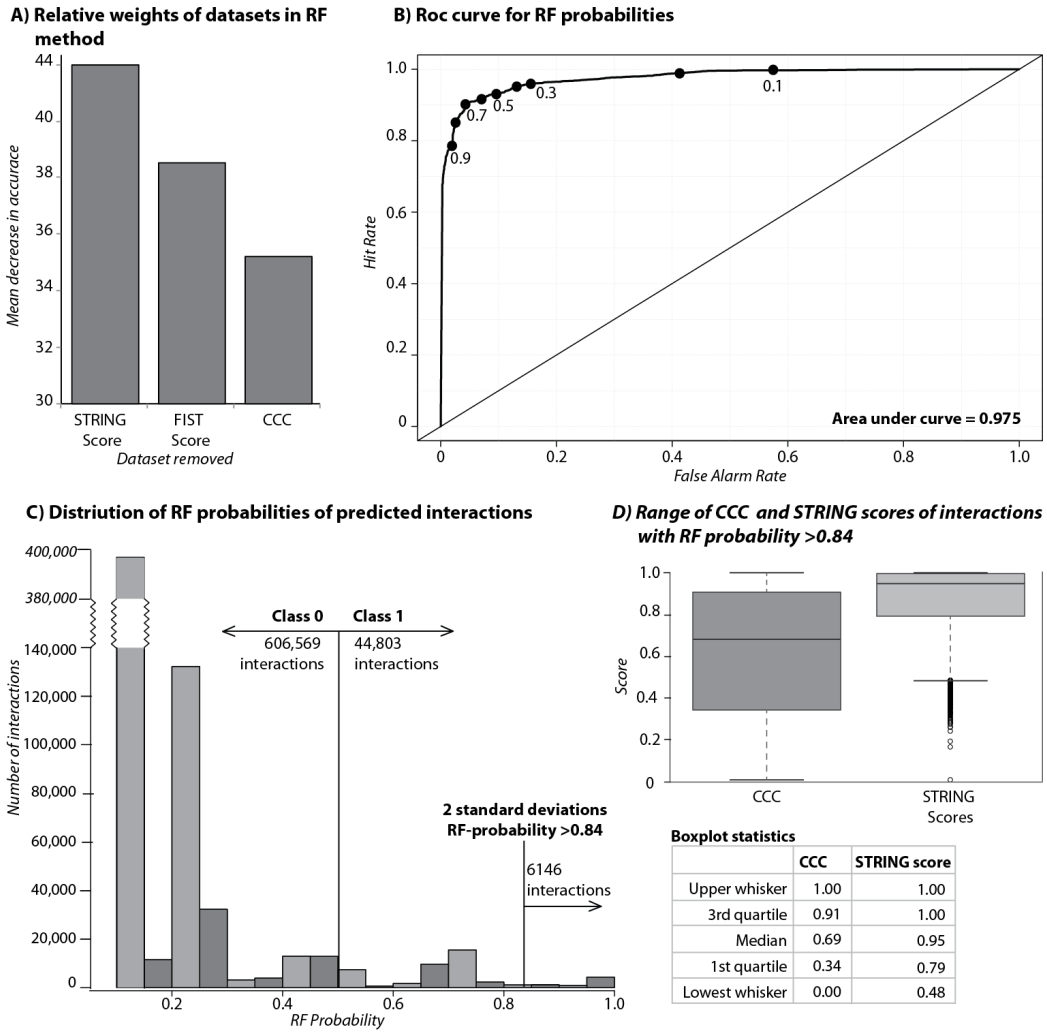

Figure A.6: Using Random Forest algorithm to assign interaction confidences.

A) Graph showing the percentage 'decrease in accuracy' of the Random Forest (RF) prediction caused by the removal of each of the three data sets in turn, i.e. the prediction is 44% poorer if the STRING data is left out. **B)** The 'Receiver-Operator Curve' showing the predictive quality of interactions based on the Gold Standard with different RF – prediction probabilities. **C)** The spread of RF probabilities assigned to each possible interaction in the dataset. The 44,803 interactions were assigned a probability >0.5. A Gaussian curve was fitted to these interactions RF prob >0.5 and if a threshold of >2 standard deviations are retained for building a network. **D)** Range of CCC and STRING scores retained in interactions with RF probabilities >0.84.

A.VI Phosphoproteomics identifies the uncharacterized protein G0SDS5 as a potential phosphorylation sink.

A scouting experiment was performed to discover the extent of phosphorylation in the high-molecular weight proteome. Studies mapping phosphorylation normally investigate bulk phosphorylation of cell lysate or on specifically purified protein assemblies but with SEC we can investigate native PTMs on proteins occurring in complexes. Post-translation modifications are known to modulate the function of proteins within the cell and many studies have shown this regulation works on the level of biomolecular interactions. However, the effect of these modifications in the context of protein complex formation has not been studied in a large-scale manner. The mapping of phosphorylation sites in the high-molecular weight proteome (as separated by SEC) allows identification of sites that can coexist on subunits when these subunits are incorporated into protein assemblies. In total 596 phosphosites were identified in they high-MW proteome (Appendix E.VI).

To check the conservation of these phosphosites in *S. Cerevisiae* we used the list of orthologues from eggNOG based on pairwise, local MAFFT L-INS-i alignments to align the proteins. When aligned, 69 of the 596 phophosites have an identical residue in *S. Cerevisiae*, however only 36 of these residues have also been discovered phosphorylated in *S. Cerevisiae***.**

26 phosphorylated proteins in the High MW fractions were from 15 of the manually curated orthologous PDB complexes used in the benchmark. A further 95 phosphopeptides were found on proteins, which map to 62 orthologues protein complexes in yeast. The large amount of phosphorylation and the number of complexes involved shows a high potential for regulation of subunit binding modulated by these sites.

Much further work is required to investigate which of these phoshosites are sites of regulation but the experimental design would need to be very careful as removal of phosphorylation may cause changes in the proteome, but these changes are likely to causes shifts in subunits binding equilibria, not all-or-nothing responses.

An interesting case from this analysis is the protein G0DS5. This protein contains 6883 residues, is predicted as completely disordered and has no detectable homologues even in the closely related *Chaetomium globosum* (mesophilic). On this protein alone we detected 152 unique phosphosites (Figure 3C) and is the most abundant protein in our experiment. This protein is of completely unknown function but its highly repetitive phosphorylated sequence may perhaps act as a sponge to soak up excess phosphorylation. Another possible explanation could be that, because ATP is highly unstable at higher temperatures, it may act as a phosphor-

donor to ADP molecules with a dedicated phosphotransferase molecule. Interestingly, this protein was also the most promiscuous crosslinked protein in the dataset, with crosslinks found to 52 other proteins indicating it may have a chaperone role.

APPENDIX B : Automated image analysis

B.I Attempted automated hit calling method

When this project was initially conceived an automatic approach to image analysis was envisaged which would use machine learning and image analysis to call perturbed cell as having quantifiable differences from the control strains. Differences would be called on an individual cell level (change in localization) and on the population level (changes in the number of budding cells etc.).

The human eye is attuned to finding differences but will miss those that it is not looking for but does have the advantage of disregarding differences that are irrelevant (dead cells or dirt etc). To teach this to a computer program which is where the difficulty lies.

Machine learning and image recognition are not trivial and require a great deal of trial and error. He workflow was as follows; After segmentation based on the bright-field images each cell was assessed with 115 features using EBI image. A support vector machine was trained on a benchmark of cells with a known localizations to allow the algorithm to learn which values of these 115 features were relevant for each localization. Each cell in the screen are then assessed for the probability that it could fall into each of these bins independently, meaning that each cell could be assigned more than one localization. The populations of localizations can then be compared between the control and perturbation (Figure B.1).

Unfortunately this approach failed overall. The support vector machine did a good job on cells with very defined single localizations (Figure B.2) but in more complex localizations up to 80% of cells were classified as ambiguous i.e. the cells could not be classified with a probability. The only classification that was called efficiently was 'Y', assigned to artifacts i.e. dead cells or segmented dirt.

The reasons for this failure are that threefold;

- Cells were often poorly segmented using center filled based on the bright-field image causing a huge source of variability. This segmented area also included the shadows on the bright-field images (Figure B.2).
- Cells with multiple localizations were particularly difficult to call as the number of variations of these combinations in terms of relative intensity is huge (Figure B.2).

- Some of the conditions caused specific artifacts that cannot easily controlled for and contribute a lot of variability. For example, ΔCHO1 and ΔPSD1/ΔPSD2 strains have a considerable amount of internal fluorescence and the tet-LCB1 and tet-STT4 strains had substantial numbers of dead or dying cells sue to their pleiotropic effects.

In the end it was decided that the time was better spent manually curating localization changes in the cells, which is currently the standard in the field. The approach showed promise and in future studies adding second marker proteins for sub-cellular or plasma membrane localizations may solve the issues with the segmentation and automatic classification of the protein localization in these cells.

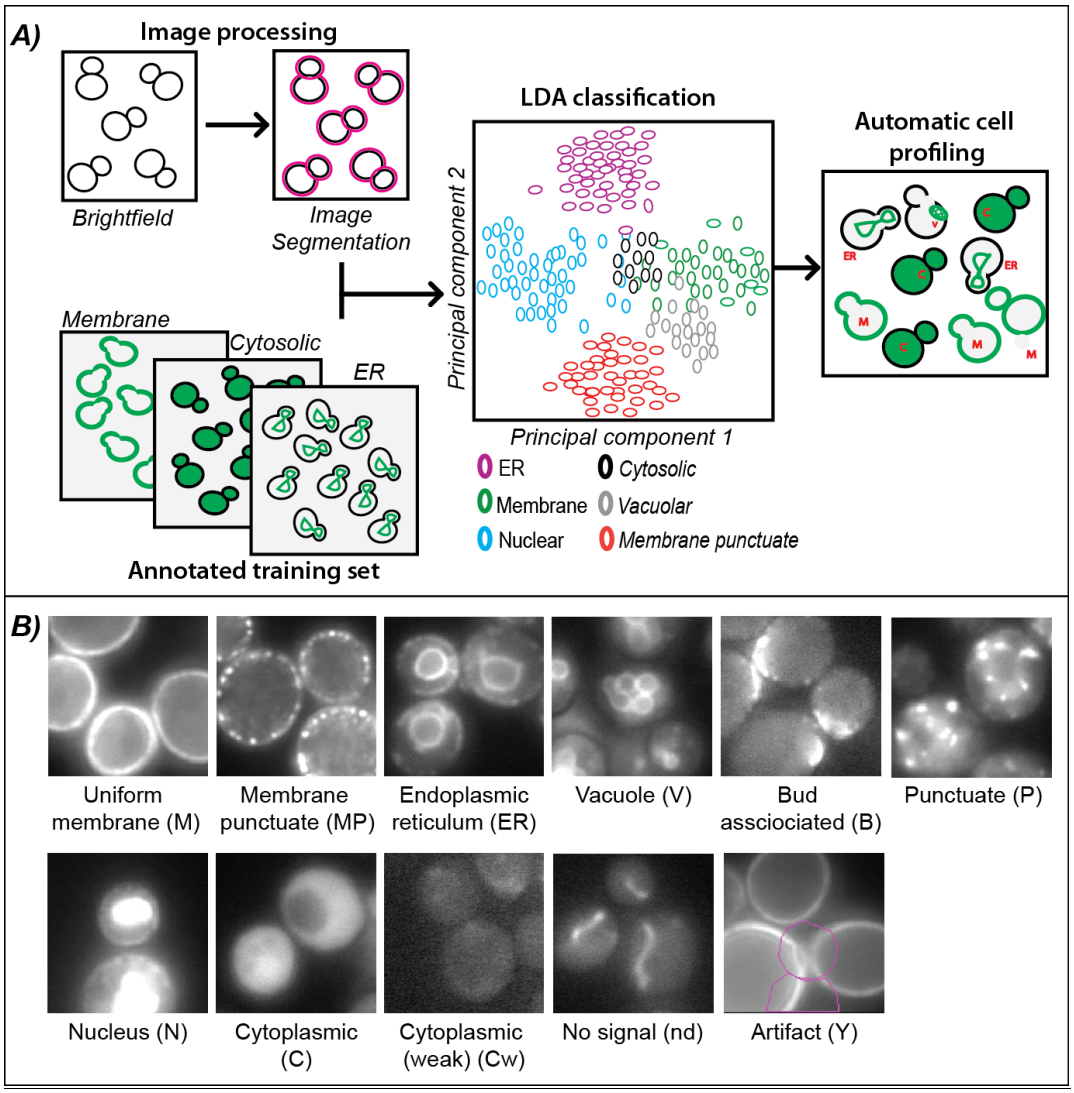

Figure B.1: Automated hit calling method

A) Fluorescent cells are segmented based upon the brightfield images. Each cell is then evaluated against 115 features from the EBImage software. A training set of 3671 cells was manually annotated with the canonical localizations (shown in **B**). Machine learning was used to plot the 115 features of the cells in multi-dimensional space. Cells with the same localization group together into a distinct 'cloud' in this plot and Linear discriminant analysis (LDA) is used to set the boundaries which assign a combination of features to a certain localization. Each cell from the main dataset is then plotted and assigned a probability of which cloud it belongs to. The populations of localizations can then be compared between the control and perturbation

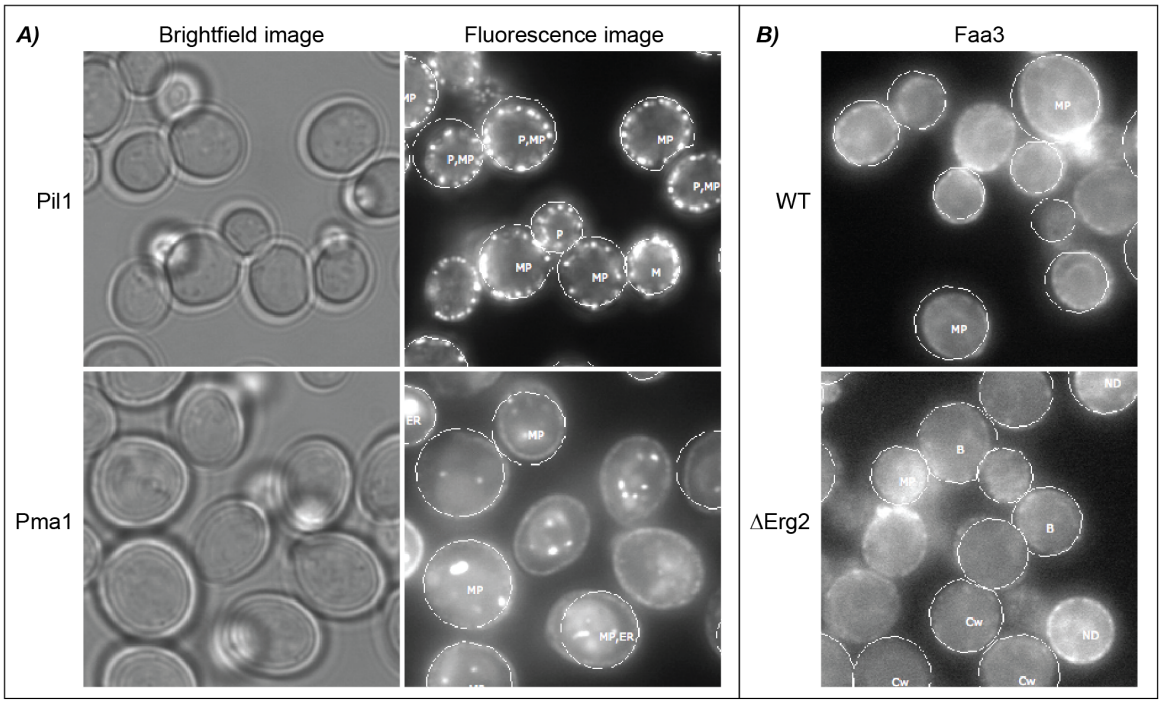

Figure B.2: Typical results from automated classification of fluorescent protein localization

A) On the left are brightfield images of the fluorescent cells on the right. Segmentation occurs on the brightfield images, which is imperfect and a source of variability. The segmented cells are displayed with white circles in the fluorescent images. Each cell is assigned a localization. Pil1 cells display a simple, single localization and are assigned 'Membrane punctuate', though due to segmentation errors many are also assigned 'Punctuate'. Pma1 has Membrane, Endoplasmic reticulum and punctuate localizations but cells are not consistently assigned 2 or 3 of these.

B) The majority of the cells with fluorescent Faa3, which localizes to the cell periphery, could not be assigned a localization. The majority of the cells that do receive a localization are assigned 'Membrane punctuate'. According to the algorithm, Faa3 in ΔERG2 cells have a significant shift to a cytoplasmic localization. Upon manual inspection, there does not appear to be a real localization change, rather the majority of cells are now assigned 'Cytoplasmic (weak)' localization. Many outcomes from like this make results from the automated hit-calling algorithm difficult to interoperate.

4.1.4: Membrane foci counting algorithm

Some changes are difficult to quantify by eye and a more quantitative method is desirable. An algorithm was developed to specifically count plasma membrane foci. These 'dots' are

important structures as they sequester specific proteins and lipids into domains and their distribution can change when the lipid ratios are disrupted.

To assess these membrane changes a 'membrane foci counting algorithm' (MFCA) was designed. This segments the cell periphery from the cellular interior and scans the membrane for peaks in intensity (Figure B.1). This approach worked nicely for peripheral proteins but it also counted many artifacts, when internal membranes or cytoplasmic foci were located towards the membrane periphery. It is possible to overcome this problem by applying the algorithm only to those cells that have plasma membrane localization in the wild type cells (bud neck, bud tip, Cell periphery or membrane punctate).

An increase or decrease in membrane foci number was called if the population average in the perturbed cells differed by greater than 20% from the GFP control population.

4.1.5: Combined approach of qualitative manual hit calling and quantitative automated dot counting

Work is being done on using the MFCA to complement the protein localization changes discovered with the manual calling. This will be a quantitative approach to better characterize for example the changes in membrane compartments described in 3.4.18

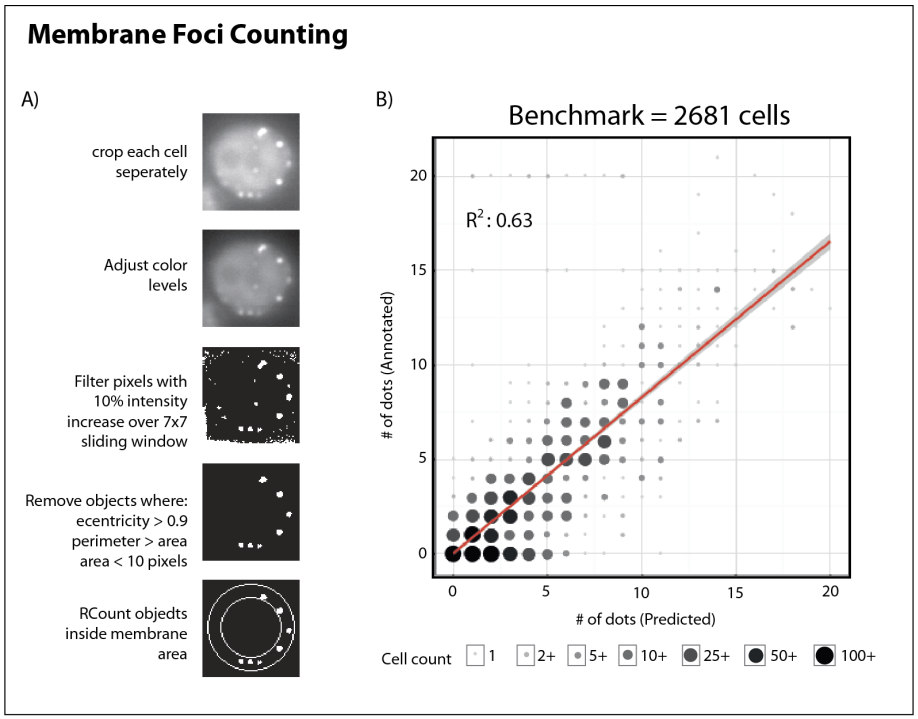

Figure B.1: Membrane foci counting algorithm (MFCA)

A) An algorithm was designed to count the number of foci present on the cell membrane. The algorithm counts the number of bright 'dot-like' objects at the cell periphery. B) The number of membrane foci was manually assigned for 2861 cells and then predicted by the MFCA. The R^2 = 0.62.

APPENDIX C : Acknowledgements

To my supervisor, Dr. Anne-Claude Gavin, thank you for your support and guidance in all stages of my graduate experiences. Your expertise, amazing collaborations and an incredible learning environment that you provided me has enhanced my graduate career in memorable ways. You have often been tough but I have learned a lot from your breadth of knowledge, and enthusiasm in work that even I had lost faith with long the way.

I would like to take this opportunity to thank members of my supervisory committee, Dr Kiran Patil, Dr. Marco Kaksonen, Dr Orsolya Barabas, Prof. Dr. Stefan Herzig, Prof. Dr. Frauke Melchior who have fostered my projects with their input, accommodating me with their flexibility and scientific advice, despite their busy schedules. Thank you for your suggestions that have been very helpful throughout the development of my project.

Over the past 3.5 years I have also benefited from incredible collaborations that have been crucial to the development of my project. In particular Panagiotis kastritis, Thomas Bock, Matt Rogon and Martin Beck, without whom, the Chaetomium thermophilum project could not have been possible. Also to Arun Kumar ,Sergaj Arvedev, Karl Kugler and Kiran Patil for the yeast high-content screen.

I would also like to thank all past and present members of the Gavin Lab, in particular the wise heads of Kenji Maeda and Marco Hennrich, Panagiotis Kastritis who have had the patience to teach me a great deal about biochemistry and mass spectrometry and computational biology.

On a personal note, I am forever grateful to my parents, family and friends for their support throughout my time at EMBL.

APPENDIX D : References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science (80-.). *287*, 2185–2195.

Aguilera-Romero, A., Gehin, C., and Riezman, H. (2014). Sphingolipid homeostasis in the web of metabolic routes. Biochim. Biophys. Acta - Mol. Cell Biol. Lipids *1841*, 647–656.

Aloy, P., and Russell, R.B. (2002). Interrogating protein interaction networks through structural biology. Proc. Natl. Acad. Sci. U. S. A. *99*, 5896–5901.

Alvaro, D., Lisby, M., and Rothstein, R. (2007). Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. PLoS Genet. *3*, 2439–2449.

Amlacher, S., Sarges, P., Flemming, D., Van Noort, V., Kunze, R., Devos, D.P., Arumugam, M., Bork, P., and Hurt, E. (2011). Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile. Cell *146*, 277–289.

Anitei, M., and Hoflack, B. (2011). Bridging membrane and cytoskeleton dynamics in the secretory and endocytic pathways. Nat. Cell Biol. *14*, 11–19.

Asp, L., Kartberg, F., Fernandez-Rodriguez, J., Smedh, M., Elsner, M., Laporte, F., Bárcena, M., Jansen, K.A., Valentijn, J.A., Koster, A.J., et al. (2009). Early stages of Golgi vesicle and tubule formation require diacylglycerol. Mol. Biol. Cell *20*, 780–790.

Audhya, A., Foti, M., and Emr, S.D. (2000). Distinct roles for the yeast phosphatidylinositol 4-kinases, Stt4p and Pik1p, in secretion, cell growth, and organelle membrane dynamics. Mol. Biol. Cell *11*, 2673–2689.

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.

Bacia, K., Kim, S.A., and Schwille, P. (2006). Fluorescence cross-correlation spectroscopy in living cells. Nat. Methods *3*, 83–89.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*.

Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B., and Boone, C. (2010). Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and schizosaccharomyces pombe. Methods Enzymol. *470*, 145–179.

Beck, M., Malmström, J.A., Lange, V., Schmidt, A., Deutsch, E.W., and Aebersold, R. (2009). Visual proteomics of the human pathogen Leptospira interrogans. Nat. Methods *6*, 817–823.

Benschop, J.J., Brabers, N., van Leenen, D., Bakker, L. V, van Deutekom, H.W.M., van Berkum, N.L., Apweiler, E., Lijnzaad, P., Holstege, F.C.P., and Kemmeren, P. (2010). A consensus of core protein complex compositions for Saccharomyces cerevisiae. Mol. Cell *38*, 916–928.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., et al. (1997). The complete genome sequence of Escherichia coli K-12. Science (80-.). *277*, 1453–1462.

Bock, T., Chen, W., Ori, A., Malik, N., Silva-martin, N., Powell, S.T., Kastritis, P.L., Smyshlyaev, G., Vonkova, I., Kirkpatrick, J., et al. (2014) An integrated approach for genome annotation of the eukaryotic thermophile Chaetomium thermophilum. 1–14.

Boehringer, D., Ban, N., and Leibundgut, M. (2013). 7.5-Å cryo-EM structure of the mycobacterial fatty acid synthase. J. Mol. Biol. *425*, 841–849.

Boersema, P.J., Raijmakers, R., Lemeer, S., Mohammed, S., and Heck, A.J.R. (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. Nat. Protoc. *4*, 484–494.

Boersema, P.J., Kahraman, A., and Picotti, P. (2015). Proteomics beyond large-scale protein expression analysis. Curr. Opin. Biotechnol. *34*, 162–170.

Bradatsch, B., Leidig, C., Granneman, S., Gnädig, M., Tollervey, D., Böttcher, B., Beckmann, R., and Hurt, E. (2012). Structure of the pre-60S ribosomal subunit with nuclear export factor Arx1 bound at the exit tunnel. Nat. Struct. Mol. Biol. *19*, 1234–1241.

Breker, M., Gymrek, M., and Schuldiner, M. (2013). A novel single-cell screening platform reveals proteome plasticity during yeast stress responses. J. Cell Biol. *200*, 839–850.

Brownsey, R.W., Boone, A.N., Elliott, J.E., Kulpa, J.E., and Lee, W.M. (2006). Regulation of acetyl-CoA carboxylase. Biochem. Soc. Trans. *34*, 223–227.

Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

Bui, K.H., Von Appen, A., Diguilio, A.L., Ori, A., Sparks, L., Mackmull, M.T., Bock, T., Hagen, W., Andrés-Pons, A., Glavy, J.S., et al. (2013). Integrated structural analysis of the human nuclear pore complex scaffold. Cell *155*, 1233–1243.

Buu, L.M., Chen, Y.C., and Lee, F.J.S. (2003). Functional characterization and localization of acetyl-CoA hydrolase, Ach1p, in Saccharomyces cerevisiae. J. Biol. Chem. *278*, 17203–17209.

De Camilli, P., Emr, S.D., McPherson, P.S., and Novick, P. (1996). Phosphoinositides as regulators in membrane traffic. Science *271*, 1533–1539.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell *149*, 1393–1406.

Chypre, M., Zaidi, N., and Smans, K. (2012). ATP-citrate lyase: A mini-review. Biochem. Biophys. Res. Commun. *422*, 1–4.

Collinet, C., Stöter, M., Bradshaw, C.R., Samusik, N., Rink, J.C., Kenski, D., Habermann, B., Buchholz, F., Henschel, R., Mueller, M.S., et al. (2010). Systems survey of endocytosis by multiparametric image analysis. Nature *464*, 243–249.

Collura, V., and Boissy, G. (2007). From protein-protein complexes to interactomics. Subcell. Biochem. *43*, 135–183.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.

Cowart, L.A., and Obeid, L.M. (2007). Yeast sphingolipids: Recent developments in understanding biosynthesis, regulation, and function. Biochim. Biophys. Acta - Mol. Cell Biol. Lipids *1771*, 421–431.

Cutler, N.S., Heitman, J., and Cardenas, M.E. (1997). STT4 is an essential phosphatidylinositol 4-kinase that is a target of wortmannin in Saccharomyces cerevisiae. J. Biol. Chem. *272*, 27671–27677.

Daum, G., Lees, N.D., Bard, M., and Dickson, R. (1998). Biochemistry, cell biology and molecular biology of lipids of Saccharomyces cerevisiae. Yeast *14*, 1471–1510.

Daum, G., Tuller, G., Nemec, T., Hrastnik, C., Balliano, G., Cattel, L., Milla, P., Rocco, F., Conzelmann, A., Vionnet, C., et al. (1999). Systematic analysis of yeast strains with possible defects in lipid metabolism. Yeast *15*, 601–614.

DeGennaro, C.M., Alver, B.H., Marguerat, S., Stepanova, E., Davis, C.P., Bähler, J., Park, P.J., and Winston, F. (2013). Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. Mol. Cell. Biol. *33*, 4779–4792.

Desrivières, S., Cooke, F.T., Parker, P.J., and Hall, M.N. (1998). MSS4, a phosphatidylinositol-4-phosphate 5-kinase required for organization of the actin cytoskeleton in Saccharomyces cerevisiae. J. Biol. Chem. *273*, 15787–15793.

Diacovich, L., Peirú, S., Kurth, D., Rodríguez, E., Podestá, F., Khosla, C., and Gramajo, H. (2002). Kinetic and structural analysis of a new group of Acyl-CoA carboxylases found in Streptomyces coelicolor A3(2). J. Biol. Chem. *277*, 31228–31236.

Dominguez, C., Boelens, R., and Bonvin, A.M.J.J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. *125*, 1731–1737.

Ehebauer, M.T., Zimmermann, M., Jakobi, A.J., Noens, E.E., Laubitz, D., Cichocki, B., Marrakchi, H., Lanéelle, M.-A., Daffé, M., Sachse, C., et al. (2015). Characterization of the Mycobacterial Acyl-

CoA Carboxylase Holo Complexes Reveals Their Functional Expansion into Amino Acid Catabolism. PLOS Pathog. *11*, e1004623.

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.

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M.J.O., and Dennis, E. a (2009). Update of the LIPID MAPS comprehensive classification system for lipids. J. Lipid Res. *50 Suppl*, S9–S14.

Fairn, G.D., Hermansson, M., Somerharju, P., and Grinstein, S. (2011). Phosphatidylserine is polarized and required for proper Cdc42 localization and for development of cell polarity. Nat. Cell Biol. *13*, 1424–1430.

Fall, S.M., Oh, S.P., Schaerer, D., Maselli, A., Blamont, J.E., Krassa, R.F., Salpeter, E.E., Garmany, C.D., and Shull, J.M. (2011). Evidence for network evolution in an Arabidopsis interactome map. Science (80-.). *334*, 1669–1675.

Fei, W., Alfaro, G., Muthusamy, B.P., Klaassen, Z., Graham, T.R., Yang, H., and Beh, C.T. (2008). Genome-wide analysis of sterol-lipid storage and trafficking in Saccharomyces cerevisiae. Eukaryot. Cell *7*, 401–414.

Feng, Y., De Franceschi, G., Kahraman, A., Soste, M., Melnik, A., Boersema, P.J., de Laureto, P.P., Nikolaev, Y., Oliveira, A.P., and Picotti, P. (2014). Global analysis of protein structural changes in complex proteomes. Nat. Biotechnol. *32*.

Fernández-Recio, J., Totrov, M., and Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. J. Mol. Biol. *335*, 843–865.

Fleck, C.B., and Brock, M. (2009). Re-characterisation of Saccharomyces cerevisiae Ach1p: Fungal CoA-transferases are involved in acetic acid detoxification. Fungal Genet. Biol. *46*, 473–485.

Flick, J.S., and Thorner, J. (1993). Genetic and biochemical characterization of a phosphatidylinositol-specific phospholipase C in Saccharomyces cerevisiae. Mol Cell Biol *13*, 5861–5876.

Foti, M., Audhya, A., and Emr, S.D. (2001). Sac1 lipid phosphatase and Stt4 phosphatidylinositol 4-kinase regulate a pool of phosphatidylinositol 4-phosphate that functions in the control of the actin cytoskeleton and vacuole morphology. Mol. Biol. Cell *12*, 2396–2411.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., et al. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*.

Gago, G., Diacovich, L., Arabolaza, A., Tsai, S.-C., and Gramajo, H. (2011). Fatty acid biosynthesis in actinomycetes. FEMS Microbiol. Rev. *35*, 475–497.

Gallego, O., Betts, M.J., Gvozdenovic-Jeremic, J., Maeda, K., Matetzki, C., Aguilar-Gurrieri, C., Beltran-Alvarez, P., Bonn, S., Fernández-Tornero, C., Jensen, L.J., et al. (2010). A systematic screen for protein-lipid interactions in Saccharomyces cerevisiae. Mol. Syst. Biol. *6*, 430.

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.

Gilbert, R.J.C., Gordiyenko, Y., von der Haar, T., Sonnen, A.F.-P., Hofmann, G., Nardelli, M., Stuart, D.I., and McCarthy, J.E.G. (2007). Reconfiguration of yeast 40S ribosomal subunit domains by the translation initiation multifactor complex. Proc. Natl. Acad. Sci. U. S. A. *104*, 5788–5793.

Gingras, A.-C., Gstaiger, M., Raught, B., and Aebersold, R. (2007). Analysis of protein complexes using mass spectrometry. Nat. Rev. Mol. Cell Biol. *8*, 645–654.

Gipson, P., Mills, D.J., Wouts, R., Grininger, M., Vonck, J., and Kühlbrandt, W. (2010). Direct structural insight into the substrate-shuttling mechanism of yeast fatty acid synthase by electron cryomicroscopy. Proc. Natl. Acad. Sci. U. S. A. *107*, 9164–9169.

Goñi, F.M., and Alonso, A. (1999). Structure and functional properties of diacylglycerols in membranes. Prog. Lipid Res. *38*, 1–48.

Gordiyenko, Y., Schmidt, C., Jennings, M.D., Matak-Vinkovic, D., Pavitt, G.D., and Robinson, C. V (2014). eIF2B is a decameric guanine nucleotide exchange factor with a γ2ε2 tetrameric core. Nat. Commun. *5*, 3902.

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.

Haberkant, P., Raijmakers, R., Wildwater, M., Sachsenheimer, T., Brügger, B., Maeda, K., Houweling, M., Gavin, A.C., Schultz, C., Van Meer, G., et al. (2013). In vivo profiling and visualization of cellular protein-lipid interactions using bifunctional fatty acids. Angew. Chemie - Int. Ed. *52*, 4033–4038.

Han, B.-G., Dong, M., Liu, H., Camp, L., Geller, J., Singer, M., Hazen, T.C., Choi, M., Witkowska, H.E., Ball, D.A., et al. (2009). Survey of large protein complexes in D. vulgaris reveals great structural diversity. Proc. Natl. Acad. Sci. U. S. A. *106*, 16580–16585.

Hannich, J.T., Umebayashi, K., and Riezman, H. (2011). Distribution and functions of sterols and sphingolipids. Cold Spring Harb. Perspect. Biol. *3*.

Hannun, Y. a, and Obeid, L.M. (2008). Principles of bioactive lipid signalling: lessons from sphingolipids. Nat. Rev. Mol. Cell Biol. *9*, 139–150.

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.

Heese-Peck, A., Pichler, H., Zanolari, B., Watanabe, R., Daum, G., and Riezman, H. (2002). Multiple functions of sterols in yeast endocytosis. Mol. Biol. Cell *13*, 2664–2680.

Henry, S. a, Kohlwein, S.D., and Carman, G.M. (2012). Metabolism and regulation of glycerolipids in the yeast Saccharomyces cerevisiae. Genetics *190*, 317–349.

Hermesh, O., Genz, C., Yofe, I., Sinzel, M., Rapaport, D., Schuldiner, M., and Jansen, R.-P. (2014). Yeast phospholipid biosynthesis is linked to mRNA localization. J. Cell Sci. 3373–3381.

Hondele, M., Stuwe, T., Hassler, M., Halbach, F., Bowman, A., Zhang, E.T., Nijmeijer, B., Kotthoff, C., Rybin, V., Amlacher, S., et al. (2013). Structural basis of histone H2A-H2B recognition by the essential chaperone FACT. Nature *499*, 111–114.

Huang, C.S., Sadre-Bazzaz, K., Shen, Y., Deng, B., Zhou, Z.H., and Tong, L. (2010). Crystal structure of the alpha(6)beta(6) holoenzyme of propionyl-coenzyme A carboxylase. Nature *466*, 1001–1005.

Huang, C.S., Ge, P., Zhou, Z.H., and Tong, L. (2011). An unanticipated architecture of the 750-kDa α6β6 holoenzyme of 3-methylcrotonyl-CoA carboxylase. Nature *481*, 219–223.

Huh, W.-K., Falvo, J. V, Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*.

Janke, C., Magiera, M.M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., et al. (2004). A versatile toolbox for PCR-based tagging of yeast genes: New fluorescent proteins, more markers and promoter substitution cassettes. Yeast *21*, 947–962.

Jenni, S., Leibundgut, M., Maier, T., and Ban, N. (2006). Architecture of a fungal fatty acid synthase at 5 A resolution. Science *311*, 1263–1267.

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2008). eggNOG: Automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. *36*.

Jivotovskaya, A. V, Valásek, L., Hinnebusch, A.G., and Nielsen, K.H. (2006). Eukaryotic translation initiation factor 3 (eIF3) and eIF2 can promote mRNA binding to 40S subunits independently of eIF4G in yeast. Mol. Cell. Biol. *26*, 1355–1372.

Kaneda, T., and Smith, E.J. (1980). Relationship of primer specificity of fatty acid de novo synthetase to fatty acid composition in 10 species of bacteria and yeasts. Can. J. Microbiol. *26*, 893–898.

Kaplan, C.D., Holland, M.J., and Winston, F. (2005). Interaction between transcription elongation factors and mRNA 3???-end formation at the Saccharomyces cerevisiae GAL10-GAL7 locus. J. Biol. Chem. *280*, 913–922.

Karotki, L., Huiskonen, J.T., Stefan, C.J., Ziolkowska, N.E., Roth, R., Surma, M. a., Krogan, N.J., Emr, S.D., Heuser, J., Grunewald, K., et al. (2011). Eisosome proteins assemble into a membrane scaffold. J. Cell Biol. *195*, 889–902.

Kasinathan, S., Orsi, G. a, Zentner, G.E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. Nat. Methods *11*, 203–209.

Kastritis, P.L., and Bonvin, A.M.J.J. (2010). Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J. Proteome Res. *9*, 2216–2225.

Kastritis, P.L., Rodrigues, J.P.G.L.M., Folkers, G.E., Boelens, R., and Bonvin, A.M.J.J. (2014). Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface. J. Mol. Biol. *426*, 2632–2652.

Kearns, B.G., McGee, T.P., Mayinger, P., Gedvilaite, A., Phillips, S.E., Kagiwada, S., and Bankaitis, V.A. (1997). Essential role for diacylglycerol in protein transport from the yeast Golgi complex. Nature *387*, 101–105.

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.

Kniazeva, M., Crawford, Q.T., Seiber, M., Wang, C.Y., and Han, M. (2004). Monomethyl branched-chain fatty acids play an essential role in Caenorhabditis elegans development. PLoS Biol. *2*.

Knispel, R.W., Kofler, C., Boicu, M., Baumeister, W., and Nickell, S. (2012). Blotting protein complexes from native gels to electron microscopy grids. Nat. Methods *9*, 182–184.

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.

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.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A.C., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

Leitner, A., Walzthoeni, T., and Aebersold, R. (2014). Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. Nat. Protoc. *9*, 120–137.

Lemmon, M. a (2008). Membrane recognition by phospholipid-binding domains. Nat. Rev. Mol. Cell Biol. *9*, 99–111.

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, X., Gianoulis, T. a, Yip, K.Y., Gerstein, M., and Snyder, M. (2010). Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses. Cell *143*, 639–650.

Li, Z., Agellon, L.B., Allen, T.M., Umeda, M., Jewell, L., Mason, A., and Vance, D.E. (2006). The ratio of phosphatidylcholine to phosphatidylethanolamine influences membrane integrity and steatohepatitis. Cell Metab. *3*, 321–331.

Löwe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W., and Huber, R. (1995). Crystal structure of the 20S proteasome from the archaeon T. acidophilum at 3.4 A resolution. Science *268*, 533–539.

Maeda, K., Anand, K., Chiapparino, A., Kumar, A., Poletto, M., Kaksonen, M., and Gavin, A.-C. (2013). Interactome map uncovers phosphatidylserine transport by oxysterol-binding proteins. Nature *501*, 257–261.

Maeda, K., Poletto, M., Chiapparino, A., and Gavin, A.-C. (2014). A generic protocol for the purification and characterization of water-soluble complexes of affinity-tagged proteins and lipids. Nat. Protoc. *9*, 2256–2266.

Mann, M., Kulak, N.A., Nagaraj, N., and Cox, J. (2013). The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. Mol. Cell *49*, 583–590.

Van Meer, G., Voelker, D.R., and Feigenson, G.W. (2008). Membrane lipids: where they are and how they behave. Nat. Rev. Mol. Cell Biol. *9*, 112–124.

Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S.G., et al. (1997). Overview of the yeast genome. Nature *387*, 7–65.

Michaud, G.A., Salcius, M., Zhou, F., Bangham, R., Bonin, J., Guo, H., Snyder, M., Predki, P.F., and Schweitzer, B.I. (2003). Analyzing antibody specificity with whole proteome microarrays. Nat. Biotechnol. *21*, 1509–1512.

Monecke, T., Haselbach, D., Voß, B., Russek, A., Neumann, P., Thomson, E., and Hurt, E. (2012). Structural basis for cooperativity of CRM1 export complex formation. PNAS *110*, 960–965.

Mosca, R., Pons, T., Céol, A., Valencia, A., and Aloy, P. (2013). Towards a detailed atlas of protein-protein interactions. Curr. Opin. Struct. Biol. *23*, 929–940.

Munn, A.L., Heese-Peck, A., Stevenson, B.J., Pichler, H., and Riezman, H. (1999). Specific sterols required for the internalization step of endocytosis in yeast. Mol. Biol. Cell *10*, 3943–3957.

Munro, S. (2003). Lipid rafts: elusive or illusive? Cell *115*, 377–388.

Natter, K., Leitner, P., Faschinger, A., Wolinski, H., McCraith, S., Fields, S., and Kohlwein, S.D. (2005). The spatial organization of lipid synthesis in the yeast Saccharomyces cerevisiae derived from large scale green fluorescent protein tagging and high resolution microscopy. Mol. Cell. Proteomics *4*, 662–672.

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

Neumann, B., Walter, T., Hériché, J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Nickell, S., Kofler, C., Leis, A.P., and Baumeister, W. (2006). A visual approach to proteomics. Nat. Rev. Mol. Cell Biol. *7*, 225–230.

Opekarová, M., Robl, I., and Tanner, W. (2002). Phosphatidyl ethanolamine is essential for targeting the arginine transporter Can1p to the plasma membrane of yeast. Biochim. Biophys. Acta - Biomembr. *1564*, 9–13.

Ori, A., Banterle, N., Iskar, M., Andrés-Pons, A., Escher, C., Khanh Bui, H., Sparks, L., Solis-Mezarino, V., Rinner, O., Bork, P., et al. (2013). Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. Mol. Syst. Biol. *9*, 648.

Orlandi, I., Casatta, N., and Vai, M. (2012). Lack of Ach1 CoA-Transferase Triggers Apoptosis and Decreases Chronological Lifespan in Yeast. Front. Oncol. *2*.

Park, W.S., Heo, W. Do, Whalen, J.H., O'Rourke, N. a, Bryan, H.M., Meyer, T., and Teruel, M.N. (2008). Comprehensive identification of PIP3-regulated PH domains from C. elegans to H. sapiens by model prediction and live imaging. Mol. Cell *30*, 381–392.

Perutz, M.F., and Raidt, H. (1975). Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. Nature *255*, 256–259.

Plaschka, C., Larivière, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E. V., Borchers, C.H., Baumeister, W., Herzog, F., et al. (2015). Architecture of the RNA polymerase II–Mediator core initiation complex. Nature *518*, 378–38.

Qiu, H., Hu, C., Gaur, N.A., and Hinnebusch, A.G. (2012). Pol II CTD kinases Bur1 and Kin28 promote Spt5 CTR-independent recruitment of Paf1 complex. EMBO J. *31*, 3494–3505.

Rajabi, K., Ashcroft, A.E., and Radford, S.E. (2015). Mass spectrometric methods to analyze the structural organization of macromolecular complexes. Methods.

Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S.B., Phanse, S., Ceol, A., et al. (2014). The binary protein-protein interaction landscape of Escherichia coli. Nat. Biotechnol. *32*, 285–290.

Raychaudhuri, S., Im, Y.J., Hurley, J.H., and Prinz, W.A. (2006). Nonvesicular sterol movement from plasma membrane to ER requires oxysterol-binding protein-related proteins and phosphoinositides. J. Cell Biol. *173*, 107–119.

Rebecchi, M.J., and Pentyala, S.N. (2000). Structure, function, and control of phosphoinositide-specific phospholipase C. Physiol. Rev. *80*, 1291–1335.

Ren, G., Vajjhala, P., Lee, J.S., Winsor, B., and Munn, A.L. (2006). The BAR domain proteins: molding membranes in fission, fusion, and phagy. Microbiol. Mol. Biol. Rev. *70*, 37–120.

Rolland, T., Tas, M., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S.D., Yang, X., et al. (2014). Resource A Proteome-Scale Map of the Human Interactome Network. Cell *159*, 1212–1226.

Rose, K., Rudge, S.A., Frohman, M.A., Morris, A.J., and Engebrecht, J. (1995). Phospholipase D signaling is essential for meiosis. Proc. Natl. Acad. Sci. U. S. A. *92*, 12151–12155.

Rozenblatt-Rosen, O., Deo, R.C., Padi, M., Adelmant, G., Calderwood, M.A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M., et al. (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature *487*, 491–495.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. (2004). A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol. *14*, 313–324.

Saliba, A.-E., Vonkova, I., Ceschia, S., Findlay, G.M., Maeda, K., Tischer, C., Deghou, S., van Noort, V., Bork, P., Pawson, T., et al. (2014). A quantitative liposome microarray to systematically characterize protein-lipid interactions. Nat. Methods *11*, 47–50.

Schuiki, I., Schnabl, M., Czabany, T., Hrastnik, C., and Daum, G. (2010). Phosphatidylethanolamine synthesized by four different pathways is supplied to the plasma membrane of the yeast Saccharomyces cerevisiae. Biochim. Biophys. Acta *1801*, 480–486.

Singer-Krüger, B., Nemoto, Y., Daniell, L., Ferro-Novick, S., and De Camilli, P. (1998). Synaptojanin family members are implicated in endocytic membrane traffic in yeast. J. Cell Sci. *111 (Pt 22)*, 3347–3356.

Slaughter, B.D., Smith, S.E., and Li, R. (2009). Symmetry breaking in the life cycle of the budding yeast. Cold Spring Harb. Perspect. Biol. *1*.

Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., and Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. Ther. Drug Monit. *27*, 747–751.

Snider, J., Kittanakom, S., Curak, J., and Stagljar, I. (2010). Split-ubiquitin based membrane yeast two-hybrid (MYTH) system: a powerful tool for identifying protein-protein interactions. J. Vis. Exp.

Spira, F., Mueller, N.S., Beck, G., von Olshausen, P., Beig, J., and Wedlich-Söldner, R. (2012). Patchwork organization of the yeast plasma membrane into numerous coexisting domains. Nat. Cell Biol. *14*, 1–11.

Stein, A., Mosca, R., and Aloy, P. (2011). Three-dimensional modeling of protein interactions and complexes is going 'omics. Curr. Opin. Struct. Biol. *21*, 200–208.

Sterner, R., and Liebl, W. (2001). Thermophilic adaptation of proteins. Crit. Rev. Biochem. Mol. Biol. *36*, 39–106.

Strahl, T., and Thorner, J. (2007). Synthesis and function of membrane phosphoinositides in budding yeast, Saccharomyces cerevisiae. Biochim. Biophys. Acta *1771*, 353–404.

Strunk, B.S., Loucks, C.R., Su, M., Vashisth, H., Cheng, S., Schilling, J., Brooks, C.L., Karbstein, K., and Skiniotis, G. (2011). Ribosome assembly factors prevent premature translation initiation by 40S assembly intermediates. Science *333*, 1449–1453.

Summers, E.F., Letts, V.A., McGraw, P., and Henry, S.A. (1988). Saccharomyces cerevisiae cho2 mutants are deficient in phospholipid methylation and cross-pathway regulation of inositol synthesis. Genetics *120*, 909–922.

Tahirovic, S., Schorr, M., and Mayinger, P. (2005). Regulation of intracellular phosphatidylinositol-4-phosphate by the Sac1 lipid phosphatase. Traffic *6*, 116–130.

Tavassoli, S., Chao, J.T., Young, B.P., Cox, R.C., Prinz, W.A., de Kroon, A.I.P.M., and Loewen, C.J.R. (2013). Plasma membrane--endoplasmic reticulum contact sites regulate phosphatidylcholine synthesis. EMBO Rep. *14*, 434–440.

Thibault, G., Shui, G., Kim, W., McAlister, G.C., Ismail, N., Gygi, S.P., Wenk, M.R., and Ng, D.T.W. (2012). The Membrane Stress Response Buffers Lethal Effects of Lipid Disequilibrium by Reprogramming the Protein Homeostasis Network. Mol. Cell *48*, 16–27.

Tkach, J.M., Yimit, A., Lee, A.Y., Riffle, M., Costanzo, M., Jaschob, D., Hendry, J. a, Ou, J., Moffat, J., Boone, C., et al. (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat. Cell Biol. *14*, 966–976.

Tong, L. (2013). Structure and function of biotin-dependent carboxylases. Cell. Mol. Life Sci. *70*, 863–891.

Tyagi, M., Hashimoto, K., Shoemaker, B.A., Wuchty, S., and Panchenko, A.R. (2012). Large-scale mapping of human protein interactome using structural complexes. EMBO Rep. *13*, 266–271.

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.

Vegas, A.J., Fuller, J.H., and Koehler, A.N. (2008). Small-molecule microarrays as tools in ligand discovery. Chem. Soc. Rev. *37*, 1385–1394.

Vidal, M., and Fields, S. (2014). The yeast two-hybrid assay : still finding connections after 25 years. Nat. Methods *11*, 1203–1206.

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

Vizeacoumar, F.J., Van Dyk, N., Vizeacoumar, F.S., Cheung, V., Li, J., Sydorskyy, Y., Case, N., Li, Z., Datti, A., Nislow, C., et al. (2010). Integrating high-throughput genetic interaction mapping and high-content screening to explore yeast spindle morphogenesis. J. Cell Biol. *188*, 69–81.

De Vries, S.J., and Bonvin, A.M.J.J. (2011). Cport: A consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS One *6*.

De Vries, S.J., van Dijk, M., and Bonvin, A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. Nat. Protoc. *5*, 883–897.

Wachsmuth, M., Conrad, C., Bulkescher, J., Koch, B., Mahen, R., Isokane, M., Pepperkok, R., and Ellenberg, J. (2015). High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. Nat. Biotechnol. *33*, 1–8.

Waksman, M., Eli, Y., Liscovitch, M., and Gerst, J.E. (1996). Identification and characterization of a gene encoding phospholipase D activity in yeast. J. Biol. Chem. *271*, 2361–2364.

Walzthoeni, T., Claassen, M., Leitner, A., Herzog, F., Bohn, S., Förster, F., Beck, M., and Aebersold, R. (2012). False discovery rate estimation for cross-linked peptides identified by mass spectrometry. Nat. Methods *9*, 901–903.

Warshel, A., Sharma, P.K., Kato, M., and Parson, W.W. (2006). Modeling electrostatic effects in proteins. Biochim. Biophys. Acta - Proteins Proteomics *1764*, 1647–1676.

Weinberg, J., and Drubin, D.G. (2011). Clathrin-mediated endocytosis in budding yeast. Trends Cell Biol. *22*, 1–13.

Wera, S., Bergsma, J.C.T., and Thevelein, J.M. (2001). Phosphoinositides in yeast: Genetically tractable signalling. FEMS Yeast Res. *1*, 9–13.

Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2013). HMDB 3.0-The Human Metabolome Database in 2013. Nucleic Acids Res. *41*.

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.

Yang, G.X., Li, X., and Snyder, M. (2012). Investigating metabolite-protein interactions: An overview of available techniques. Methods *57*, 459–466.

Yonath, A., Bartunik, H.D., Bartels, K.S., and Wittmann, H.G. (1984). Some x-ray diffraction patterns from single crystals of the large ribosomal subunit from Bacillus stearothermophilus. J. Mol. Biol. *177*, 201–206.

Yonath, A., Glotz, C., Gewitz, H.S., Bartels, K.S., von Böhlen, K., Makowski, I., and Wittmann, H.G. (1988). Characterization of crystals of small ribosomal subunits. J. Mol. Biol. *203*, 831–834.

Youn, J.-Y., Friesen, H., Kishimoto, T., Henne, W.M., Kurat, C.F., Ye, W., Ceccarelli, D.F., Sicheri, F., Kohlwein, S.D., McMahon, H.T., et al. (2010). Dissecting BAR domain function in the yeast Amphiphysins Rvs161 and Rvs167 during endocytosis. Mol. Biol. Cell *21*, 3054–3069.

Yu, J.W., Mendrola, J.M., Audhya, A., Singh, S., Keleti, D., DeWald, D.B., Murray, D., Emr, S.D., and Lemmon, M. a (2004). Genome-wide analysis of membrane targeting by S. cerevisiae pleckstrin homology domains. Mol. Cell *13*, 677–688.

Zanella, F., Lorens, J.B., and Link, W. (2010). High content screening: Seeing is believing. Trends Biotechnol. *28*, 237–245.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature *490*, 556–560.