Université de Montréal

Searching for novel gene functions in yeast: Identification of thousands of novel molecular interactions by protein-fragment complementation assay followed by automated gene function prediction and high-throughput lipidomics

par

Kirill Tarasov

Département de biochimie
Programme de Biochimie
Faculté de médecine
Université de Montréal

Thèse présentée à la faculté des études supérieures
en vue de l'obtention du grade *Philosophiae Doctor (Ph.D)*
en biochimie

September, 2014

Université de Montréal

Faculté des études supérieures

Cette thèse s'intitule:

Searching for novel gene functions in yeast: Identification of thousands of novel molecular interactions by protein-fragment complementation assay followed by automated gene function prediction and high-throughput lipidomics

Présentée par:

Kirill Tarasov

a été évaluée par un jury composé des personnes suivantes:

Martine Raymond	Président-rapporteur
Muriel Aubry	Membre du jury (en remplacement du directeur de recherche)
François Major	Membre du jury
Vladimir Titorenko	Examinateur externe
Lucie Parent	Représentant du doyen de la FES

Résumé

La compréhension de processus biologiques complexes requiert des approches expérimentales et informatiques sophistiquées. Les récents progrès dans le domaine des stratégies génomiques fonctionnelles mettent dorénavant à notre disposition de puissants outils de collecte de données sur l'interconnectivité des gènes, des protéines et des petites molécules, dans le but d'étudier les principes organisationnels de leurs réseaux cellulaires. L'intégration de ces connaissances au sein d'un cadre de référence en biologie systémique permettrait la prédiction de nouvelles fonctions de gènes qui demeurent non caractérisées à ce jour. Afin de réaliser de telles prédictions à l'échelle génomique chez la levure Saccharomyces cerevisiae, nous avons développé une stratégie innovatrice qui combine le criblage interactomique à haut débit des interactions protéines-protéines, la prédiction de la fonction des gènes *in silico* ainsi que la validation de ces prédictions avec la lipidomique à haut débit. D'abord, nous avons exécuté un dépistage à grande échelle des interactions protéines-protéines à l'aide de la complémentation de fragments protéiques. Cette méthode a permis de déceler des interactions *in vivo* entre les protéines exprimées par leurs promoteurs naturels. De plus, aucun biais lié aux interactions des membranes n'a pu être mis en évidence avec cette méthode, comparativement aux autres techniques existantes qui décèlent les interactions protéines-protéines. Conséquemment, nous avons découvert plusieurs nouvelles interactions et nous avons augmenté la couverture d'un interactome d'homéostasie lipidique dont la compréhension demeure encore incomplète à ce jour. Par la suite, nous avons appliqué un algorithme d'apprentissage afin d'identifier huit gènes non caractérisés ayant un rôle potentiel dans le métabolisme des lipides. Finalement, nous avons étudié si ces gènes et un groupe de régulateurs transcriptionnels distincts, non préalablement impliqués avec les lipides, avaient un rôle dans l'homéostasie des lipides. Dans ce but, nous avons analysé les lipidomes des délétions mutantes de gènes sélectionnés. Afin d'examiner une grande quantité de souches, nous avons développé une plateforme à haut débit pour le criblage lipidomique à contenu élevé des bibliothèques de levures mutantes. Cette plateforme consiste en la spectrométrie de masse à haute resolution Orbitrap et en un cadre de traitement des données dédié et supportant le phénotypage des lipides de centaines de mutations de Saccharomyces cerevisiae. Les méthodes expérimentales en lipidomiques ont confirmé les prédictions fonctionnelles en démontrant certaines différences au sein des phénotypes métaboliques lipidiques des délétions mutantes ayant une absence des gènes YBR141C et YJR015W, connus pour leur implication dans le métabolisme des lipides. Une altération du phénotype lipidique a également été observé pour une délétion mutante du facteur de transcription KAR4 qui n'avait pas été auparavant lié au métabolisme lipidique. Tous ces résultats démontrent qu'un processus qui intègre l'acquisition de nouvelles interactions moléculaires, la prédiction informatique des fonctions des gènes et une plateforme lipidomique innovatrice à haut débit , constitue un ajout important aux méthodologies existantes en biologie systémique. Les développements en méthodologies génomiques fonctionnelles et en technologies lipidomiques fournissent donc de nouveaux moyens pour étudier les réseaux biologiques des eucaryotes supérieurs, incluant les mammifères. Par conséquent, le stratégie présenté ici détient un potentiel d'application au sein d'organismes plus complexes.

Mots-clés: Interaction protéine-protéine, complémentation de fragments protéiques, protéine membranaire, métabolisme des lipides high-throughput screen, lipidomics, apprentissage automatique, prédiction de la fonction d'un gene, visualisation analytique, criblage à haut débit

Abstract

Understanding complex biological processes requires sophisticated experimental and computational approaches. The advances in functional genomics strategies provide powerful tools for collecting diverse types of information on interconnectivity of genes, proteins and small molecules for studying organizational principles of cellular networks. Integration of that knowledge into a systems biology framework enables prediction of novel functions of uncharacterized genes. For performing such predictions on a genome-wide scale in the yeast *Saccharomyces cerevisiae*, we have developed a novel strategy that combines high-throughput interactomics screen for protein-protein interactions, *in silico* gene function prediction, and validation of predictions with high-throughput lipidomics. We started by performing a large-scale screen for protein-protein interactions using a protein-fragment complementation assay. The method allowed to monitor interactions *in vivo* between proteins expressed from their natural promoters. Furthermore, the method did not suffer from bias against membrane interactions comparing to established genome-wide techniques for detecting protein interactions. As a result, we detected many novel interactions and increased coverage of an interactome of lipid homeostasis that has not been yet comprehensively explored. Next, we applied a machine learning algorithm to identify eight previously uncharacterized genes with a potential role in lipid metabolism. Finally, we investigated whether these genes and a set of distinct transcriptional regulators, not implicated previously with lipids, have a role in lipid homeostasis. For that purpose, we analyzed lipidome of deletion mutants of the selected genes. In order to probe a large number of strains, we have developed a high-throughput platform for high-content lipidomic screening of yeast mutant libraries that consists of high-resolution Orbitrap mass spectrometry and a dedicated data processing framework to support lipid phenotyping across hundreds of *Saccharomyces cerevisiae* mutants. Lipidomics experiments confirmed functional predictions by demonstrating differences of the lipid metabolic phenotypes of deletion mutants lacking *YBR141C* and *YJR015W* genes predicted to be involved in lipid metabolism. An altered lipid phenotype was also observed for a deletion mutant of the transcription factor *KAR4* that has not been linked previously with lipid metabolism. These results demonstrate that a workflow that integrates the acquisition of novel molecular interactions, computational gene function prediction and novel high-throughput shotgun lipidomics platform is a valuable contribution to an arsenal of methods for systems biology. The developments of functional genomic methods and lipidomics technologies provide means to study biological networks of higher eukaryotes, including mammals. Therefore, the presented workflow has a potential to find its applications in more complex organisms.

Keywords: Protein-protein interactions, protein-fragment complementation assays, high-throughput screen, membrane proteins, lipid metabolism, lipidomics, machine learning, gene function prediction, visual analytics.

Table of contents

List of Figures

Dedication

To my family. Thank you for your love and support.

Acknowledgements

I would like to thank all the colleagues that worked with me during these years: the fantastic lab in Montreal, Canada (Michnick lab), the young lipidomics group in Odense, Denmark (Ejsing lab) and the lab of Danila Perumov and the whole Molecular and Radiation Biophysics Department of the Nuclear Physics Institute in St. Petersburg, Russia, where my scientific career has started, teachers from school N211 St.Petersrurg. I am grateful to Valentin Ribchin for the introduction to biophysics, Embek Nikolaevich Kazbekov for opening the lab to me, Danila Perumov for all his teaching and my first car, Rimma Kreneva for being a big part of my life, Sergey Steinberg for introducing me to demanding calculations and the science in 3D, my Finnish colleagues. I would like to thank Steven Michnick for the opportunity to work on the large-scale protein interaction project, the freedom to learn every aspect of it and the tough independence lessons. I'm grateful for the support to Jean-Francois Paradis and Sylvie Beauchemin. I thank Muriel, Colm, Edith and Zoya for their comments on the thesis manuscript.

Many thanks to friends around the globe. Special thanks to those who did not forget me while I was doing science overseas. Christophe, Luc, Alex, Philippe, Dimitri, Eugene, Igor for frequently reminding me that life is not only about science. My neighbors in Lachine (Marin, Lena, Anuta) for hospitality.

Finally, I thank my family for being there all the time. Without them, this work would not have been possible. Grandpa, thank you for believing in me!

Abbreviations

ALEX	Analysis of lipid experiments
BP	Biological process
cAMP	Cyclic adenosine monophosphate
CC	Cellular compartment
Cer	Ceramide
CV	Coefficient of variation
CYGD	Comprehensive yeast genome database
DAG	Diacylglycerol
DHFR	Dihydrofolate reductase
DNA	Deoxyribonucleic acid
dTMP	Thymidine monophosphate
ER	Endoplasmic reticulum
ESI	Electrospray ionization
FT MS	Fourier transform mass spectrometry
GDP	Guanosine diphosphate
GFP	Green fluorescent protein
GO	Gene ontology
GPI	Glycosylphosphatidylinositol
GTP	Guanosine triphosphate

IPC	Inositol-phosphoceramide
KEGG	Kyoto encyclopedia of genes and genomes
LPA	Lysophosphatidic acid
LPC	Lysophosphatidylcholine
LPE	Lysophosphatidylethanolamine
LPI	Lysophosphatidylinositol
LPS	Lysophosphatidylserine
mDHFR	Murine dihydrofolate reductase
MF	Molecular function
MIPS	Munich information center for protein sequences
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NMR	Nuclear magnetic resonance
OD	Optical density
ORF	Open reading frame
PA	Phosphatidic acid
PC	Phosphatidylcholine
PC O-	Ether-linked phosphatidylcholine

PCA[1]	Protein-fragment complementation assays
PCR	Polymerase chain reaction
PE	Phosphatidylethanolamine
PE O-	Ether-linked phosphatidylethanolamine
PI	Phosphatidylinositol
PPV	Positive predicted value
PS	Phosphatidylserine
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
SC	Synthetic complete
SD	Synthetic defined
SE	Sterol ester
SGD	Saccharomyces Genome Database
SHexCer	Sulfatide
SoamD	Sum of absolute mol% difference
TAG	Triacylglycerol
TAP-MS	Tandem affinity purification coupled to mass spectrometry
Y2H	Yeast two-hybrid

[1] An abbreviation for an experimental protein-fragment complementation assays (PCA) is the same as for a mathematical method, called principal component analysis. In order to avoid confusion, abbreviation PCA in this thesis is only used in relation to protein-fragment complementation assay. The name of the mathematical method is always spelled out in the text as principal component analysis.

Chapter 1 : Introduction

Biological activity of the cell is orchestrated through the coordinated action of thousands of molecules. The coordination implies the assembly of large complexes consisting of smaller components, passing signals from one component to another, synchronization of events, and adaptation of cellular systems to perturbations. Recent developments in experimental technologies provide unprecedented opportunities to monitor the state of thousands of molecules and interactions between them at different time and conditions. Data generated by such large-scale screens stimulate development of computational methods for data integration and interpretation. Every new dataset provides observations for improving models of cellular processes and discovering new components and functions. The understanding of biological principles based on these data is a cyclic process. New data help to generate novel insights and hypotheses that can be validated and refined based on results of further experiments. The present thesis describes one of such cycles. First, we developed a novel methodology for screening interactions between proteins. Next, using computational analyses, we identified connections between proteins that suggested novel functions related to lipid metabolism for uncharacterized proteins. Finally, we conducted the second experimental screen to monitor changes in lipid composition due to inactivation of the predicted proteins to test proposed hypotheses about involvement in lipid metabolism and provided further details related to the discovered functions. In the introduction chapter, we present our model system, i.e. yeast. We describe experimental techniques for elucidating complex cellular networks and highlight advantages of our method for detecting protein-protein interactions. Further, we introduce visualization and computational methods that are commonly used for network analyses, and we emphasize developments related to the function prediction. Finally, recent advances in metabolomics, lipidomics and its applications in yeast are reviewed. We conclude the introduction with setting the goal of the study presented in the thesis.

1.1. Yeast as a model system

To understand a complex system, one needs to study its smaller components. The cell is the smallest unit of life that is able to reproduce independently. In this thesis, we use yeast to study cellular networks. The benefit of yeast is not limited to baking, brewing and wine making. Yeast is an excellent model system to study cellular processes because of the following factors: fast growth (doubling time is about 90 minutes, in comparison, human HeLa cells double in about 24 hours); ease of handling in the lab; availability of powerful strategies for genetic manipulation; and finally, complex cellular organization similar to cells of higher eukaryotes, such as humans. The most popular species of yeast, *Saccharomyces cerevisiae*, also known as baker's or budding yeast, was the source for the first sequenced eukaryotic genome completed in 1996 [1] (strain S288C). The study identified 5885 potential protein-encoding genes organized in 16 chromosomes. Because only 4% of the genes in *Saccharomyces cerevisiae* contain introns, the effect of alternative splicing is negligible. Thus, the number of proteins in the budding yeast is about the same as the number of the protein-encoding genes. About 31% of the yeast genes were found to have homologs in mammals [2] and for a number of disease-causing genes in humans corresponding yeast counterparts were identified [3].

1.2. Genetic manipulation of yeast

Success of genetic manipulation strategies is based on efficient transformation methods for introducing exogenous DNA into the yeast cells, the integration of a foreign DNA into specific regions of the genome, and the availability of the selectable markers for detecting the integration events.

The first reports of genetic transformation of *Saccharomyces cerevisiae* appeared in 1960 and currently standardized transformation protocols have been developed based on spheroplasting, electroporation, agitation with glass beads, or addition of lithium acetate, single-stranded carrier DNA and polyethylene glycol [4]. The latter method is the most commonly used because of the highest efficiency of the transformation [5]. Details about mechanisms causing the uptake of a foreign DNA are still poorly understood, and conditions of transformation protocols have been optimized empirically. The yeast cell can be transformed if incubated only with polyethylene glycol and DNA. It has been suggested that polyethylene glycol increases membrane permeability. The transformation rate is increased by heat shock and lithium acetate in intact cells, but not in cells with a disrupted cell wall. Therefore, it is likely that heat shock and lithium acetate help DNA to pass through the cell wall [6]. Double and single stranded DNA sticks to the cell wall. The addition of carrier DNA could saturate the DNA binding cites allowing the vector DNA that carry genetic material for manipulation to pass through the cell wall. It has been proposed that single stranded DNA further increases the transformation rate because it does not compete with the double-stranded vector DNA and binds more efficiently to the cell wall [4].

The transformations are performed with vectors that carry an engineered the yeast DNA construct with modified gene sequence and sequences coding for selectable markers. The transformation with vector DNA leads to incorporation of the construct sequence into a specific locus of the yeast genome [7]. The process relies on the yeast mechanisms for homologous recombination, which is a conserved process with a particular importance in repairing DNA double-strand breaks [8]. The insertion of an exogenous DNA sequence is driven by short sequences (30-50 bases) flanking the construct sequence that are homologous to the targeted region. The endogenous genomic sequence between the homologous regions is replaced with the sequence of the construct. Thus, the method is suited for gene modification by inserting particular sequences as well as for gene disruption by partly or completely deleting gene sequences [9].

Clones in which targeted DNA was successfully incorporated into the chromosomal DNA are selected based on the activity of a marker gene introduced with the construct. The functioning selectable marker allows cells to grow in the presence of an antibiotic (antibiotic resistance markers) [10] or in the absence of an indispensable compound, such as an amino acid, that cannot be synthetized by the original strain (auxotrophic markers) [11].

Yeast strains that were used in this thesis are based on the popular laboratory strain S288C that was sequenced in 1996 [1]. BY4741 strain is a MATa mating type, and BY4742 is a MATα mating type. Both strains are auxotrophic for histidine, leucine and uracil. In addition, BY4741 is auxotrophic for methionine, and BY4742 is auxotrophic for lysine, which allows mating type specific selection [11].

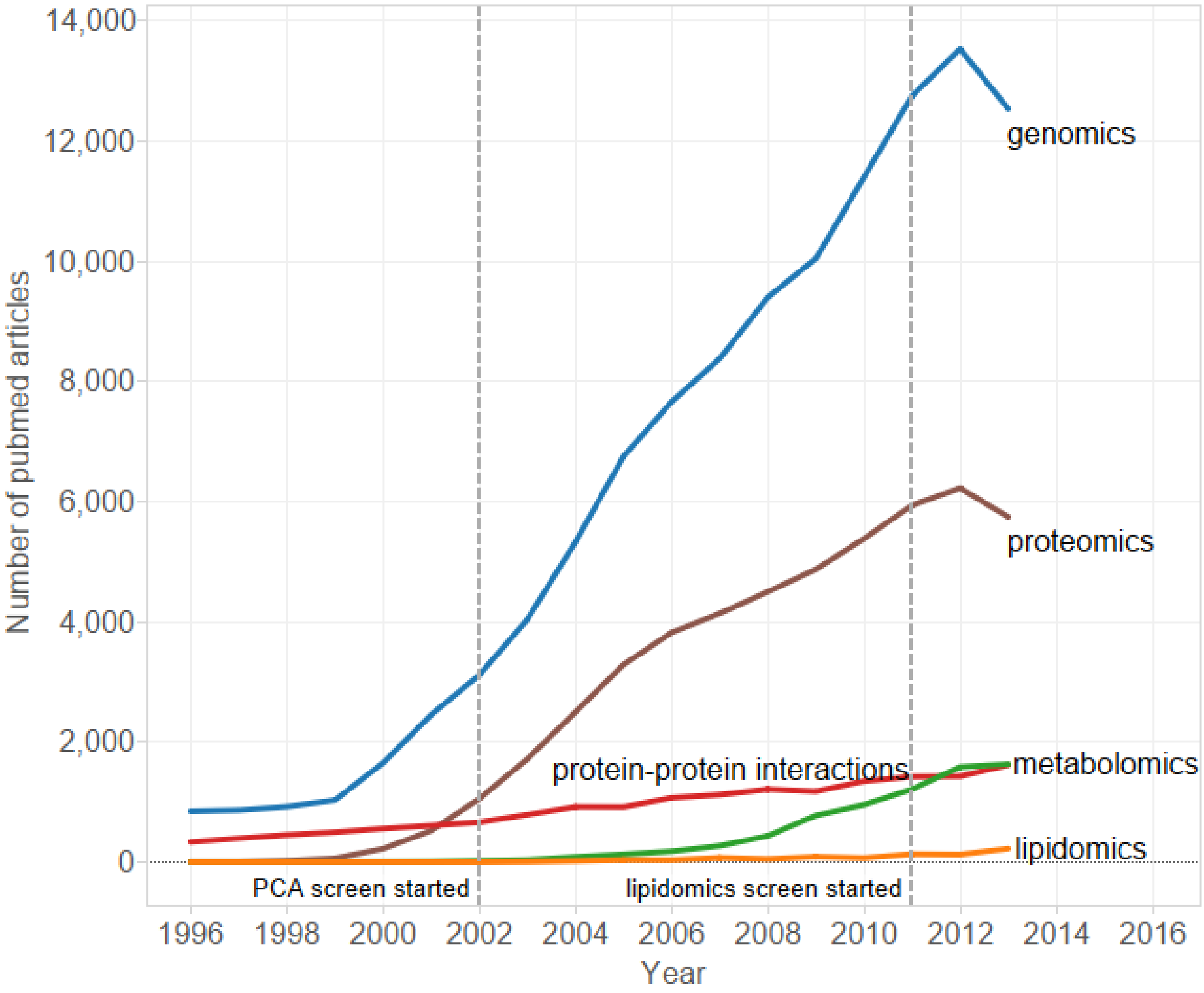


Figure 1-1. Number of PubMed citations of "omics" technologies per year.
Numbers of citations for each term per year is extracted with PubMed trend available at: http://dan.corlan.net/medline-trend.html

1.3. Functional genomics studies in yeast

Functional genomics is a discipline that attempts to study genes, proteins, small molecules and the interactions between them on a large scale. It combines the experimental high-throughput technologies and bioinformatics methods for data integration. Different types of function genomics studies performed in yeast are discussed below. Data presented in this thesis are generated using protein-protein interaction screen and lipidomics. Therefore, these technologies are discussed in detail. The level of development of each particular functional genomics field can be compared based on analysis of a number of PubMed citations per year. The trends plotted in Figure 1-1 show that protein-protein interactions and lipidomics fields are relatively new comparing to genomics and proteomics.

1.3.1. Gene expression and protein abundance

The earliest studies that took advantage of the availability of the whole genome sequence for conducting experiments on a genome-wide scale measured gene expression levels. The pioneering work that studied gene expression in the exponential growth phase and the diauxic shift defined the term “transcriptome” as a collection of gene identifiers and their expression levels in a population of cells [12]. Shortly after, changes of gene expression levels were studied during the cell cycle [13], sporulation [14] and stress response [15]. These are just a few examples of early studies that stimulated the development of technologies that were further applied to investigate transcriptome in other conditions in various organisms. Previous vast knowledge of yeast biochemistry made yeast an important validation test bed for development of novel experimental high-throughput methods, such as DNA microarrays [16], and statistical methods for data analysis, such as hierarchical clustering and heat map visualizations [17]. Collectively, these efforts provided data on transcriptional regulation of the yeast genome in thousands of conditions. A dedicated resource for storing and retrieving yeast expression data combined results from over 2400 experimental conditions, which can be mined by an efficient data search algorithm for visualization of expression patterns of genes of interest and retrieving potentially related genes based on similarity of the expression profiles [18].

Gene transcription is not the only process that defines cellular abundance of proteins. Regulation of translation, post-translational protein modifications and degradation of proteins need to be considered for making accurate assumptions regarding protein levels at a given time in the cell. Therefore, dedicated methods for probing protein abundance are needed. Protein quantification in yeast has been performed by two-dimensional gel electrophoresis [19], quantitative western blot analyses of high-affinity epitopes [20] and flow cytometry of green fluorescent protein [21] fused with a collection of the yeast proteins, and more recently, mass spectrometry based analyses [22][23]. Comparison of the results of proteomics and transcriptomics studies have demonstrated that there is a significant correlation between mRNA and protein concentrations, but the estimated correspondence of mRNA expression and protein abundance in yeast is in the range between 30% and 90% [24]. Thus, proteomics measurements are important for understanding processes related to protein homeostasis that follow gene transcription. Measuring protein abundances is experimentally more challenging. Widely available gene expression data are commonly used as an estimate of cellular quantities of corresponding proteins. Despite the differences in protein and gene expression, the latter provides highly accurate indication of whether a protein is present in the cell [25].

1.3.2. Collection of single gene deletion mutants

The genome-wide experimental strategies that followed early studies of gene expression used the sequence information and the genome transformation methods to systematically perturb or alter the yeast genes. The first study of that kind investigated phenotypes produced by a nearly complete collection of single gene deletion mutants [26,27]. A surprising outcome of the studies was the observation that most of deletion mutants did not show any growth defects. Only 18% of genes were found to be essential for growth on rich glucose medium. These data provided first hints about robustness and adaptation of an organism to perturbations of its individual components. A logical conclusion regarding the non-essential genes was that their role might be more important for surviving in specific conditions and resistance to particular perturbations. The essentiality of genes was further tested in numerous studies that investigated fitness of the gene deletion mutants under different growth conditions and in the presence of various drugs (comprehensively reviewed in [28]). The strategy helped to identify gene functions based on its response to specific perturbations. For example, screens with DNA damaging agents allowed to find genes required for maintaining the integrity of the genome [29][30]; modulation of the phosphatidylinositol metabolic pathway with wortmannin helped to establish novel pathway functions [31]. On the other hand, mechanism of action of drugs can be proposed based on the types of gene deletion mutants that are sensitive to the chemical compounds [32]. Using optical microscopy, additional cellular morphological phenotypes can be identified, such as shape, size and aggregation tendency [27], which provide further data for inferring functions of the deleted genes.

1.3.3. Protein localization studies

The important feature of the genetic modification of yeasts by homologous recombination is the ability to insert sequences at desired locations. This strategy was employed to study cellular localization of the yeast proteins whose genes were tagged with a sequence of green fluorescent protein at C terminus. 4156 tagged strains expressed fusions with green fluorescent protein that could be observed by fluorescence microscopy [33]. Proteins were classified into 22 distinct subcellular localization categories, such as cytoplasm, nucleus and mitochondrion. 70% of proteins covered by the study did not have previously any localization information. Another study employed confocal laser scanning microscopy to investigate with high-resolution localization of lipid metabolic proteins. Increased resolution of the method allowed to gain insight into suborganellar organization of lipid biosynthetic pathways and visualize proteins localized to endoplasmic reticulum membrane, membrane extensions from the nuclear envelope and lipid droplets [34]. It will be discussed below that these observations are particularly important in the context of the current thesis, because they link lipid metabolism with membrane proteins. Next level of resolution for localization studies is achieved by employing isolation of cellular suborganelles and analysis of the protein content by mass-spectrometry. Investigation of protein content of lipid droplets is one example of such analyses [35]. Lipid droplets are particles that store cellular reservoirs of non-polar lipids used as energy sources and as membrane building blocks. Mass spectrometry analysis of protein content of the particles revealed that they contain enzymes involved in fatty acid and ergosterol metabolism demonstrating that lipid droplets are essential for the lipid synthesis in addition to the role in lipid storage. Important extension to the mass spectrometry analysis of protein content of cellular components was the development of methodology for isolating protein complexes. As described below, applications using this strategy provided the most information on protein-protein interactions in yeast.

1.3.4. Studies of molecular interactions

Functional genomics strategies described above investigate a collection of states of individual components of the yeast cells, such as location or concentration of a given protein or phenotype produced by deletion of a gene. An important aspect of regulation of cellular processes is the coordinated communication between the cellular components that can be viewed as networks of molecular interactions. Biological networks are collections of associations between molecules that can be represented as graphs for computational data analysis and visualizations. The nodes of such networks are molecules (proteins, genes, metabolites) and edges are relationships between the nodes (protein-protein interactions, phosphorylation events, correlation between concentrations). Different types of experimental methodologies for capturing biological networks are described below.

Regulation of gene expression is mediated by transcription factors, proteins that bind to DNA sequence of their target genes and activate or repress the transcription. Large-scale identification of binding sites of transcription factors is enabled by chromatin immunoprecipitation followed by DNA sequencing or experiments with genomic microarrays

[36–38]. Experimentally defined binding sites of transcription factors can be used for predicting target genes based on determined binding sites specificities and promoter sequence analysis. Combination of knowledge about transcription factor bindings and data from gene expression analysis allow to identify factors that cause particular gene expression patterns [39].

Functional activity of a protein can be efficiently regulated by phosphorylation events, which lead to activation/deactivation of a protein function. This type of regulation is particularly important when a quick response to stimuli is needed. Phosphorylation events can be captured by variety of methods including mass spectrometry, kinase activity assays and western blot. Up to date, over 20 000 phosphorylation sites in yeast have been experimentally verified by high throughput mass spectrometry proteomics studies and small-scale experiments [40].

Analysis of gene essentiality in yeast demonstrated that individual deletion of most yeast genes does not result in growth defects [27]. However, when a pair of deletions is introduced it can enhance or reduce growth. Pair-wise relationships between deletion mutants are studied by accurate quantification of colony sizes or fitness of the mutants [41,42]. These data are assembled into gene interaction networks. Similarity of genetic interaction patterns between genes suggests that they might be involved in a common cellular process.

Recently, a novel large-scale approach has been developed for identification of protein-lipid interactions [43]. Authors have examined interactions between almost 200 proteins that have known lipid-binding domains and lipid enzymatic activities with lipid molecules from major lipid classes. Over 500 detected interactions provide another level of information for studying molecular networks.

1.4. Strategies for mapping protein-protein interaction networks

The focus of our study are networks of protein-protein interactions. An association of proteins into complexes of various sizes and interactions with other proteins for passing cellular signals define their functions. A collection of all protein-protein interactions forms an interactome of the cell that has been intensively studied for the last 15 years. Analysis of the interaction network can reveal functions of novel genes based on the connections with other proteins with established functional roles. The success of the protein function prediction is dependent on the coverage and accuracy of the interactome. At the time when we set our goal to contribute to the protein-protein interaction network mapping, reported experimental strategies for genome-wide analysis of the yeast interactome relied on two methods: yeast two-hybrid (Y2H) and tandem affinity purification coupled to mass spectrometry (TAP-MS). These methods and the method that we chose for novel genome-wide interaction screen, i.e. protein-fragment complementation assays (PCA)[1], are described below, followed by an overview of large-scale documented applications of the techniques.

1.4.1. Yeast two-hybrid

The yeast two-hybrid method was developed as a genetic system for detecting protein-protein interactions *in vivo*. In contrast to traditional biochemical methods such as crosslinking,

[1] An abbreviation for an experimental protein-fragment complementation assays (PCA) is the same as for a mathematical method, called principal component analysis. In order to avoid confusion, abbreviation PCA in this thesis is only used in relation to protein-fragment complementation assay. The name of the mathematical method is always spelled out in the text as principal component analysis.

co-immunoprecipitation and co-fractionation by chromatography, Y2H does not require any protein purification or isolation [44]. Therefore, Y2H can be efficiently adopted for conducting large-scale experiments [45]. Y2H is based on the activity of the transcriptional factor GAL4 (the most common option, however, applications with other transcription factors also exist) that consists of a DNA binding domain and an activating domain. Two hybrid proteins are then constructed: a protein "X" with the binding domain and a protein "Y" with the activating domain. If two proteins "X" and "Y" interact in the nucleus, they bring into proximity the GAL4 domains and enable transcription of GAL4 controlled genes, which products are used as reporters for the interaction. The advantage of Y2H is the ability to detect potentially direct protein-protein interactions. It should be noted that *in vitro* experiments with purified proteins are needed for confirming that an interaction is truly direct. Interactions that are detected by cell-based assays could be mediated by other protein, DNA and RNA molecules. Nevertheless, interactions detected by Y2H are commonly described as binary. This term implies that the method used for detecting an interaction tested a pair of proteins as opposed to a protein complex co-membership minimizing the chance for presence of intermediates. Another attractive feature of the method is the ability to use cell-based survival assays for detecting interactions. Thus, protein purification, extraction and identification steps are avoided. These factors define the cost-effectiveness and scalability of Y2H and explain why it is the most widely used method for both small and large-scale studies [46]. Early genome-wide screens for protein interactions in yeast have been suspected to contain a large number of false positive interactions [47]. However, recent developments of the method that control artifacts related to spontaneous auto-activation of the reporters, contamination by several plasmids, genetic mutations and effects of overexpression of tested proteins significantly improved the reliability of the technique [48]. The main limitation of Y2H is the fact that interactions are detected in the nucleus of *Saccharomyces cerevisiae*. Thus, non-nuclear proteins are not tested in their natural context that may lead to false positive and false negative results. Proteins are directed into nucleus by nuclear localization signal attached to binding and activating domains of the GAL4. Regardless of the presence of the nuclear localization signal, for some proteins, such as transmembrane proteins, it would be difficult or impossible to enter the nucleus. Therefore, Y2H results are biased against interactions involving this type of proteins. Finally, the fusion of proteins with GAL4 domains are expressed from plasmids, which disconnects interaction events from the physiological regulatory mechanisms.

1.4.2. Tandem affinity purification coupled to mass spectrometry

TAP-MS (tandem affinity purification coupled to mass spectrometry) is an alternative strategy that is suited for monitoring protein-protein associations under near-physiological conditions. The strategy is based on the extraction and purification of proteins that are physically associated with a tagged bait protein followed by the identification of purified proteins by mass spectrometry [49][50]. Because of the efficient methods for tagging bait proteins, scalable purification procedures and increased availability of mass spectrometers, TAP-MS method rapidly gained popularity for studying protein complexes. The advantage of the method is the untargeted identification of protein complex composition. Binary methods for detecting interactions require preparation of corresponding protein fusions for all pairs of proteins that need to be tested. In the case of TAP-MS, experiments with a few tagged baits can reveal composition of large complexes. In yeast tag can be introduced by homologous recombination, so tagged proteins remain under the regulation of the native promoters, which is a clear advantage for investigating dynamics of protein associations under different conditions in the nearly intact physiological context.

Strictly speaking, the method does not identify protein interactions. It detects collections of associated proteins that are assigned to protein complexes based on co-purification frequency and clustering analyses [51]. In the context of TAP-MS method, an interaction between two proteins means that these proteins belong to the same complex. Thus, for a pair of proteins from a macromolecular complex the method does not provide an answer of whether there is a physical contact between the two. TAP-MS experiments are associated with a technical challenge related to nonspecific binding of proteins to the tag or solid matrix used for immobilization of a ligand that require careful control experiments [52]. Furthermore, TAP-MS experiments are repeated with several baits for retrieving components of the same complexes multiple times. Processing of co-purification data with sophisticated statistical algorithms minimizes the influence of the contaminants on the determination of complex composition [51].

1.4.3. Protein-fragment complementation assays

To increase the coverage of the yeast interactome, we have conducted a genome-wide screen utilizing a technique that combines strengths of Y2H and TAP-MS approaches. Protein-fragment complementation assays (PCA) can be conducted *in vivo* as a survival cell-based assay like Y2H screens, without the need for mass spectrometers. Similarly to TAP-MS, PCA is based on protein tagging that can be performed in yeast by homologous recombination to introduce desired sequence to the targeted genes. Thus, tested gene products remain under control of their natural promoters. Tagging from the C-termini keeps the N-termini localization signal intact. Therefore, tagged proteins can follow their physiological localization path. The PCA strategy is based on a simple idea that a reporter protein can be dissected into two fragments. These fragments can be fused with proteins which ability to interact is being tested [53]. When two proteins interact, they bring into proximity the reporter fragments and allow them to fold into a normal three-dimensional structure and regain the reporter function. PCA strategy can be employed with variety of functional reporters, such as fluorescent proteins for microscopy assays and reporter enzymes that can be used for survival selection assays [54]. PCA detects direct or near-direct binary interactions between proteins. The detected interactions may be mediated by other molecules. However, the length of linkers that connect interacting proteins with the reported fragments defines how far two proteins can be apart from each other to allow fragment refolding. The large-scale screen described in Chapter 2 [55], relied on a survival selection assay based on a dihydrofolate reductase (DHFR) PCA [56]. The screening strategy employs yeast strain with inhibited endogenous DHFR activity that is not able to grow in the presence of methotrexate, a drug that inactivates cellular proliferation. Methotrexate-resistant DHFR mutant is used as a reporter. Yeast strain that expresses a pair of interacting proteins fused with fragments of methotrexate-resistant DHFR can proliferate in the presence of methotrexate when an interaction leads to refolding of the reporter fragments into a functional methotrexate-resistant DHFR reporter.

1.4.4. Large-scale studies of protein-protein interactions in yeast

We started setting up the genome-wide protein-protein interaction screen by DHFR PCA when the yeast interactome coverage was limited with few published reports describing technological developments for conducting large-scale protein interactions screens. Two studies presented large-scale Y2H screens in 2000 and 2001 and two studies published results of TAP-MS screens in 2002. The first Y2H study [57] reported 957 interactions between 1004 proteins and the second study [58] reported in total 4549 interactions between 3278 proteins. However, a smaller number of interactions from the second study was confirmed two (1533 interactions) or three times (841 interactions). The later set of interactions (core data) is considered more reliable and interactions that were not confirmed multiple times are commonly excluded from data analyses. The two TAP-MS studies reported data on complex membership of about 25% of the yeast proteins each [59,60]. One of the most intriguing observations that came from comparative analysis of the results was that the majority of interactions reported by each study were novel [47,61]. It has been proposed that the poor overlap between the studies is due to the large number of potential false-positive interactions in the early screens and method specific biases or preferences for certain types of interactions. Based on the results of these studies it was estimated that the yeast interactome contains from 16 000 to 26 000 interactions [62]. Therefore, it was evident that none of the screens reached saturation in covering the interactome. An alternative screen with a novel technique would significantly contribute to the interaction network mapping. Prior to publication of results of our DHFR PCA screen, several large-scale interaction studies were published [63–65] that were extensively compared with our data as described in Chapter 2. For consistency, it should be noted that after the publication of our results two more studies have been published presenting a comprehensive map of binary Y2H [66] and membrane interactions [67]. However, these recent results were not part of our computational analyses.

1.5. Yeast metabolomics and lipidomics

1.5.1. Metabolomics

In the previous sections, methods for elucidating biological networks of molecular interactions were discussed. The investigation of networks followed the path starting from genes (carriers of inherited information), to proteins (performers of particular cellular functions), to cellular phenotypes (collections of observable characteristics defined by combination of cellular functions). Metabolic state of a system is a novel type of information that recently received substantial attention. Small molecule metabolites are important cellular constituents that are metabolized by proteins and contribute to the cellular homeostasis. The amount of currently available information about genes and proteins on a genome-wide scale is wider than what we know about metabolites. This is mainly related to analytical challenges associated with monitoring abundances of small models. Moreover, the absence of a blue print for metabolites, such as genome sequence for proteins, makes the discovery of new bioactive molecules a much slower process than modern genome sequencing. The advances in chromatographic techniques, such as gas and liquid chromatography, nuclear magnetic resonance spectroscopy and mass spectrometry have been instrumental for commencing investigation of the cellular metabolome

- the full collection of metabolites of the cell [68]. Metabolomics is a corresponding field of research that focuses on identification and quantification of all cellular metabolites that can be classified into the following major classes: amino acids, nucleotides, toxins, vitamins, sugars and lipids. Comparing to 4 bases that define genome and 20 amino acids that are used to build proteins, metabolome is more complex in terms of diversity of chemical structures it comprises. This leads to emergence of metabolomics sub-disciplines, such as lipidomics, glycomics and peptidomics, that study particular types of metabolites [69]. The division is associated with physiochemical properties of molecules of particular types, as for example, lipidomics focused on lipids, which are generally hydrophobic compounds soluble in organic solvents.

In contrast to many functional genomics strategies discussed above that were first developed and tested in yeast before finding their applications in other organisms, the development of metabolomics is not originating from yeast research. In fact, there has been a tremendous development of metabolomics applications in plants because of the rich source of metabolites and potential scientific and applied applications [70]. Because of implications of metabolites into various diseases [71] and the promising potential of metabolic biomarkers in early disease diagnostics [72], metabolomics studies of human samples greatly outnumber attempts to scrutiny metabolome of simple eukaryotic model systems. Nevertheless, a number of elegant studies were performed in *Saccharomyces cerevisiae* that paved the way for integration of metabolomics into the yeast systems biology framework.

The first proof of concept metabolomics study in yeast investigated metabolic phenotypes of deletion mutants that did not show any growth defect in glucose-limited aerobic and anaerobic conditions [73]. Authors proposed that despite the absence of an observable growth phenotype, destitution of metabolites in the mutants could be adjusted to compensate for the effect of the mutation and maintain the normal growth rate. Measurements of metabolites by enzymatic assays and high-resolution ^{1}H-NMR spectroscopy demonstrated that deletion of *PFK26* and *PFK27* genes that encode 6-phosphofructo-2-kinase (6PF-2-K; EC 2.7.1.105), which catalyzes the conversion of fructose-6-phosphate into fructose-2,6-bisphosphate, resulted in distinct metabolic phenotypes comparing to the reference strain. Moreover, if the function of only one of the genes were known, the function of the second gene would be suggested based on the observed similarity of metabolic signatures of the two mutants. A follow-up study performed by the same group further developed the technology to enable high-throughput metabolomics screens [74]. Authors performed metabolic analysis of extracellular metabolites of about 20 gene deletion mutants from a broad range of metabolic categories. Measurement of extracellular metabolites (metabolic footprinting) was performed as alternative to determination of intracellular levels (metabolic fingerprinting) because it allows to avoid complications associated with a rapid turnover, quenching and extraction of metabolites. Unlike, quantitative metabolic profiling that focuses on quantification of all measured metabolites, the method compared raw mass spectra to identify differences between strains and experimental conditions that can be further investigated with higher accuracy. The method was not sufficiently sensitive to identify unknown peaks in the metabolic footprint. However, it could detect differences between the mutants and group together deletion strains of genes with common function, e.g. amino acid metabolism. The short running time of 2 minutes per sample makes the method attractive for a rapid systematic search for mutants with perturbed metabolism. Alternative metabolic profiling methods with higher resolution have been developed for relative or absolute quantification of metabolites by ^{1}H-NMR spectroscopy [75,76] and mass spectrometry [77–80]. These early works centered on the development of technologies for quantification of 50-100 molecules from various metabolic classes and demonstrated their performance on metabolic differences induced by selected growth conditions or few mutations. Data generated by these efforts stimulated development of computational methods for integrating metabolomics with transcriptomics [81,82] and the genome-scale yeast metabolic models [83].

The first genome-wide metabolomics study in *Saccharomyces cerevisiae* assessed amino acid levels in 5000 single-gene deletion mutants [84]. Researchers who performed the study, argued that metabolomics methods based on chromatography, mass spectrometry and nuclear magnetic resonance studies are time intensive limiting their accessibility for screening for thousands of samples. Alternatively, amino acids were analyzed starting from fluorescent derivatization of cell extracts, separation by capillary electrophoresis and detection by laser-induced fluorescence. Analysis of one sample took about 8 minutes allowing to conduct the whole screen in 2 months. Around 700 gene deletion mutants showed at least eightfold change comparing to the reference strain in at least one amino acid. The findings suggest that various factors influence amino acid levels, such as vacuolar structure and mitochondrial activity. In line with previous studies, authors demonstrated that similarity of metabolic profile help to propose gene functions to previously uncharacterized genes. The yeast metabolome has not been covered yet at this scale by mass spectrometry methods. However, recent studies demonstrate increasing feasibility of experiments employing mass spectrometry by providing data for hundreds of samples related to metabolome dynamics in various conditions [85] and genetic factors of metabolome variability [86].

1.5.2. Lipidomics

Metabolomics approaches measure a range of diverse metabolites, such as amino acids, nucleic bases, vitamins, sugars and metabolic precursors. As discussed earlier, in-depth analysis of specific metabolites requires dedicated analytical platforms. Lipids are the focus of lipidomics. The diversity of lipids define their involvement in many key biological processes, such as membrane homeostasis, energy storage and signaling [87]. That is the reason why eukaryotes have dedicated hundreds of genes to maintenance of lipid homeostasis. This makes lipidomics an attractive tool for functional genomics studies of a broad range of cellular processes. Current advances in the yeast lipidomics are summarized below and a novel lipidomics approach for discovery of lipid related genes is presented in Chapter 3.

Similarly to other "omics" technologies, the advancements in lipidomics can be viewed as a two-step process. First, analytical methods are developed and optimized to increase speed, coverage and quantification accuracy of lipid species detection. Next, the established platforms are applied for answering biological questions. The classical methods for studying lipids relied on radioactive and fluorescent labeling of lipids and separation by high performance liquid chromatography and thin-layer chromatography. Gas chromatography followed by mass spectrometry is a common method for analysis of fatty acid content of chromatographically separated lipid classes. However, these methods can be tedious and time consuming, and they are not sensitive enough to distinguish between various lipid species with a similar molecular mass. A breakthrough in lipidomics is associated with a new generation of methods that take advantage of increased sensitivity and resolution of mass spectrometry combined with the development in tandem mass spectrometry, soft ionization techniques that don't cause lipid fragmentation (matrix-assisted laser desorption/ionization and electrospray ionization (ESI)) and faster liquid chromatography methods that require lower sample volumes [88,89]. NMR spectroscopy has been also applied for determining structures of purified lipids and investigation of the structure and dynamics of lipid membranes. However, higher sensitivity of mass spectrometry-based methods make them much more common in lipidomic applications.

One of the early studies that applied mass spectrometry lipidomics in the yeast *Saccharomyces cerevisiae* utilized nanoelectrospray ionization tandem mass spectrometry to investigate membrane phospholipid composition of distinct cellular compartments [90]. It has been known that distribution of lipid classes is not uniform among the cellular compartments with examples of membrane specific classes, such as cardiolipin for the inner mitochondrial membrane, and sterol and sphingolipids for the plasma membrane. By utilizing tandem mass spectrometry lipid molecular species, i.e. lipid head group that defines the lipid class and the precise acyl chain substituent of a lipid, could be detected. A clear difference between acyl chain composition within phospholipid classes was observed in different membranes providing evidence that membrane lipid composition is regulated at the molecular species level. A later study demonstrated for the first time functional differences between two pathways of phosphatidylcholine synthesis [91]. This major lipid class of the eukaryotic membranes is synthesized either via the methylation of phosphatidylethanolamine or via the CDP-choline route. By blocking one of the pathways, phosphatidylcholine synthesis was forced to go through one of the routes. As a result, distinct profiles of molecular lipid species were observed with a greater molecular diversity of phospholipids attributed to the CDP-choline route. These pioneering studies highlighted a new level of complexity of lipid regulation that can be investigated by the modern mass spectrometry methods. The technology was rapidly adopted by several research groups that collected lipidomics data for refining the current knowledge of lipid metabolism in yeast (for a comprehensive review see [92,93]). The technology was further improved for analyzing diverse lipid classes in a single experiment. Several methods for a rapid and comprehensive coverage of the yeast lipidome by mass spectrometry have been recently published [94–97]. Electron spray ionization followed by high resolution mass spectrometry was utilized as a rapid method for detecting lipid species from the major yeast lipid classes (glycerophospholipids, free fatty acids, triacylglycerides and sphingolipids) [94,95]. In this approach, that is termed as shotgun lipidomics, lipid extracts are directly infused into a mass spectrometer avoiding the time consuming chromatographic separation, which shortens the running time of the MS analysis to 5 to 10 minutes. An alternative protocol that optimized lipid extraction procedures and solvent composition, and employed tandem mass spectrometry (MS/MS) experiments quantified 250 lipid species from 21 lipid classes [96]. It was also demonstrated that an increased coverage of lipid classes and minor lipid species could be achieved by coupling liquid chromatography with ESI-MS. An introduction of a 30 minutes separation step allowed to profile simultaneously in a simple MS experiment glycerophospholipids, sphingolipids, waxes, sterols and mono-, di- as well as triacylglycerides [97]. The availability of tools for a fast and comprehensive analysis of the yeast lipidome empowered lipidomics studies that could investigate a greater number of lipids in more conditions and mutants. The recent applications of the methods include investigation of the lipid composition of lipid rafts [98], influence of variety of growth conditions on dynamic properties of the lipidome [99], and relation of mitochondrial membrane lipidome to the yeast longevity [100]. However, there were no previous reports of attempts to analyze lipidomes of hundreds of yeast samples similarly to the above-mentioned functional genomics studies. Thus, to the best of our knowledge, the lipidomics screen described in Chapter 3 is the first step towards a large-scale identification of the lipidomic phenotypes in yeast.

1.6. Informatics and mathematics concepts related to yeast functional genomics studies

1.6.1. Hypothesis testing and statistical significance

Generation of new hypotheses based on available data and conducting experiments for confirming them is a crucial process in biological sciences. A common hypothesis tested by biologists is whether there are differences between certain properties of biological systems, such as phenotypic characteristics between different types of cells and changes in gene expression, protein and metabolite levels due to perturbation. Statistical tests evaluate whether there is a difference between observations that is unlikely due to chance or experimental error. We briefly summarize below conceptual basics of statistical testing. Details about calculations and additional methods can be found, for example, in the following references: [101,102]. In case, when a number of samples is low, graphical plots of measurements with error bars are employed to evaluate uncertainty associated with a measurement. In Chapter 3 of the thesis, we use such representation to display an average of two values with an error bar displaying two originally measured values (Figures 3-6 – 3-8). Alternatively, error bars can display standard deviation or confidence intervals. When a large number of samples is available, that is typical in high-throughput experiments, statistical significance of a result can be computationally tested. In that case, research hypothesis, i.e. a prediction made by a researcher, is reformulated with two statistical hypotheses: the null hypothesis (H_0 in mathematical notation) that states that there is no difference between observations and the alternative hypothesis (H_1) that states that there is a difference. Statistical testing is employed to calculate a test statistics. A value of a test statistics is compared to a theoretical distribution of all possible values that test statistics could have if an experiment was repeated an infinite number of times using the same number of samples. Critical values of test statistics that are unlikely to be observed if the null hypothesis is correct can be found in special statistical tables. Comparison of observed test statistics with these critical values indicate whether a result of a study is statistically significant. In scientific practice, it is common to call a result statistically significant if a likelihood of obtaining such a result by chance is not higher than 5%. A likelihood is indicated by a *p*-value, therefore for a statistically significant result the following notation is used: p <= 0.05. Methods for calculating test statistics and p-values are dependent on data distribution and can be parametric or non-parametric. Parametric methods, such as a popular Student's t-test rely on assumption that experimental outcomes are independent from one another and come from a normal distribution. If the assumptions are met, a p-value can be calculated based on a magnitude of differences between mean values of experimental outcomes detected in different conditions and a standard deviation of the measurement of the outcome. Non-parametric methods are developed for evaluation of statistical significance when data are not normally distributed. In that case, a non-parametric analogue of a t-test is Wilcoxon rank-sum test. The test statistics is calculated based on ranks rather than original values and compares medians of groups of observations instead of the means. This test is more robust to presence of outliers in the data than the parametric counterpart. However, it might yield less significant results than the t-test, i.e. has less statistical power, when applied on data that are normally distributed. Therefore, it is important to investigate whether assumptions associated with a particular test are met before performing statistical analyses. In Chapter 2, we used Wilcoxon rank-sum test for comparing protein abundances of different protein sets. The non-parametric test was selected because distribution of protein abundances was skewed and could not be approximated by normal distribution (Figure 2-4).

A special type of testing is applied when a particular property of a complex biological system is investigated. Because a property is an integral observable parameter of a system, it comes with only one value. However, one can set up a simulation experiment, in which this parameter is compared to a large number of values calculated for random systems. If a real observed value is not random, it will be rarely found in the simulated outcomes. In Chapter 2, we calculate z-score to express how different a particular value is from a mean of simulated values expressed in numbers of standard deviations (Figures 2-12, 2-13):

$$\text{z-score} = (x - \mu)/\sigma,$$

where x is an observed characteristic of a system, μ and σ are mean and standard deviation of corresponding simulated values. These scores can be converted into p-values using appropriate statistical tables.

1.6.2. Correlation methods

Large-scale functional genomics screens are often driven by analytical technologies that make it possible to observe novel types of relationships between cellular components. Therefore, instead of looking for predefined differences and their statistical significance, these results need to be explored by alternative methods developed for detection of patterns in the data when there is no *a priori* assumptions about data structure. One of such methods is a correlation analysis that assigns a score for a pair of variables. The higher the score - the higher the similarity. Same as for statistical significance tests, there are parametric (Pearson) and non-parametric (Spearman) methods for performing the correlation analysis [102]. These two methods are the most widely used. Pearson correlation measures an extent of a linear relationship between normally distributed variables. It is important to test for data normality and detect a presence of outliers in the data, because Pearson correlation is sensitive to outliers like other parametric methods. When variables are not normally distributed or when there is a deviation from linearity in the relationship between these variables, it is more appropriate to calculate calculating Spearman correlation. Spearman method is based on ranks and does not rely on normality assumption. It detects a monotonic relationship between variables that does not have to be linear.

Recently, a unified correlation approach, called maximal information coefficient, has been proposed [103]. The same method can be applied to identify different types of relationships between variables, such as linear, exponential and periodic. The maximal information coefficient is able to score similarly different types of relationships. Thus, it does not favor a particular association type. This feature is particularly valuable for exploration of datasets when there is no prior knowledge about dependencies between variables.

1.6.3. Clustering analyses

Clustering is a set of methods for multivariate data analyses that group objects in the dataset based on similarity of their properties. Unlike pair-wise correlation analyses, clustering identifies larger sets of similar objects. For example, clustering was used to identify genes that are similarly expressed under certain conditions [17] ; sets of interconnected proteins that potentially are functionally related [104]; and groups of gene deletion mutants that affect cellular metabolite concentrations in a similar way [84].

Hierarchical clustering is one of the most widely used clustering methods in biology [17]. It does not alter the original dataset but groups similar objects together. The result has the same complexity as the original dataset. The clustering procedure begins by calculating pair-wise similarity measurements between dataset objects by correlation analyses or other methods, such as Euclidian distance, which unlike correlation scores take a magnitude of differences between variables into account. This step is followed by reordering of the dataset objects moving closely related entities next to each other. The results are visualized with tree dendograms and heat maps (section 1.6.6.2 Matrix-based representation) that can be used for visual identification of modules and clusters. The method only rearranges the variables in the dataset so all relationships are kept accessible and can be visually explored.

Partitioning methods reduce dataset complexity and extract closely related modules or clusters of objects that can be represented as separate entities. This is a particularly popular method for detecting protein complexes in the protein interaction data. Various methods exist for performing such tasks [105], including algorithms that automatically cut dendogram trees produced by hierarchical clustering [106] . Rigorous performance comparison of the clustering methods [107,108] favors Marcov clustering algorithm in terms of its ability to reconstruct known protein complexes and tolerance to noise in the data. The Marcov clustering algorithm was recently applied in a number of studies aimed at consolidating the repertoire of protein complexes based on interaction data from multiple sources [109,110]. A disadvantage of the method is inability to assign proteins to multiple clusters. Recently, a ClusterOne algorithm has been reported that overcomes this limitation and has a comparable accuracy to Marcov clustering procedure [111] .

1.6.4. Filtering and quality evaluation of datasets of protein interactions

Protein-protein interactions are detected by complex experimental procedures that first generate raw experimental data (section 1.4. Strategies for mapping protein-protein interaction networks) that are processed and filtered for assigning interactions between particular proteins. Y2H studies are based on survival assays in which an interaction between two proteins results in a cell growth or altered cellular phenotype (e.g. blue colony color os strains grown with X-gal due to activation of beta-galactosidase expression). The screens were performed in an array format in which every pair of tested proteins is known [57], and in a pooled setting [57,58], in which a stain with a particular protein fused with a Y2H DNA-binding domain is mixed with many strains caring a counterpart protein fused with a Y2H activation domain. In both setting, a cell growth indicates an occurrence of an interaction. When stains are arrayed, the positions of strains corresponding to each test protein pair are known. Therefore, an interaction can be unambiguously assigned based on a cellular phenotype. In a pooled setting, the identification of interacting proteins is based on sequencing of genes associated with binding and activating domains from the selected strains. In the Y2H large-scale screens, interactions are filtered based on frequency of occurrence in replicated array experiments or frequency of identification of an interaction by sequencing. A common conclusion drawn from the evaluation of filtered results is that the interactions confirmed multiple times are much more reliable than single observations.

Frequency based filtering can be even more effective for analysis of TAP-MS results. The TAP-MS methodology detects protein composition of sets of co-purified proteins. Therefore, same protein pairs could be co-purified multiple times, which increases a confidence that they truly belong to the same protein complex. In early large-scale TAP-MS screens, the filtering was based on a simple calculation of co-occurrence frequency and exclusion of proteins that were detected in a large number of purifications [59,60]. In later studies, more sophisticated statistical methods were employed for performing data filtering. Gavin et al. presented a 'socio-affinity' index that evaluates a partnership tendency of pairs of proteins [64]. The index is based on a ratio of how often two proteins are co-purified together and their expected co-occurrence. Partnership scoring is more specific and keeps protein pairs with good socio-affinity index even if one of the proteins is observed with a high frequency. Scores based on occurrence of an individual protein would have filtered out such interactions. It was observed that interactions with a higher socio-affinity index are more reproducible. Consequently, social-affinity index cut-off value for producing a filtered dataset was selected based on the reproducibility of the interactions in the screen.

First studies aimed at the evaluation of results of large-scale protein-protein interaction screens, suggested that the important parameters of data quality evaluation are coverage and accuracy [47]. These parameters can be determined by comparing a large-scale interaction dataset with a 'gold-standard' reference set of validated interactions. Coverage is defined as a fraction of reference set interactions covered by the experiment and accuracy is a fraction of observed interactions confirmed by the reference set. Considering interactions observed in several datasets lead to the increase of data accuracy. However, such filtering decreases the coverage. Therefore, optimal filtering parameters should be selected to achieve a reasonable compromise between a comprehensive coverage and an adequate accuracy of the reported data.

In studies described above, the filtering procedures were performed with empirically selected cut-offs followed by quality evaluation of the filtered datasets. More recent TAP-MS studies integrated filtering based on interaction frequency with reference sets of validated interactions for selecting filtering parameters optimized for accuracy and coverage of the filtered datasets [51,65].

In section 2.4.3, we describe a filtering strategy of the DHFR PCA screen that combines recent developments for processing Y2H and TAP-MS data. The proposed algorithm relies on identification of protein interactions based on analysis of colony sizes of a survival assay, which is similar to Y2H processing, in combination with optimization of filtering parameters based on the data coverage and accuracy inspired by TAP-MS screens.

1.6.5. Methods for interpretation of interaction networks

General principles and concepts for studying protein-protein interaction networks are summarized below with a focus on topics related to the current thesis explored in Chapter 3, i.e. protein-protein interaction based function prediction of unknown genes and experimental validation of such predictions.

1.6.5.1. General properties of interaction networks

Large-scale methods for detecting protein-protein interaction networks as well as other methods for collecting genome-wide observations are often referred to as top-down approaches. Such approaches do not test particular hypotheses from the start, but aim at cataloging and summarizing observations. These observations collectively contribute to a network of protein or gene associations, in which nodes correspond to genes and/or proteins and edges correspond to their interactions. A network of interactions provides a basis for future studies aimed at interpretation of biological importance of observed network patterns and properties.

Similarly, computational top-down methods study general properties of networks as a whole. Learning network properties is an important step in devising general organizational principles of associations between thousands of biological molecules and constructing a network model with reduced complexity. Importance of certain repeated patterns of the network or characteristics that are different from those expected from a random network can be tested experimentally or computationally. Network properties provide a basis of comparative analysis between biological networks and knowledge transfer between extensively studied networks from model organisms to other organisms with limited available experimental data. Furthermore, a comparison of network properties helps to avoid a direct comparison of complex networks that is computationally intractable [112]. The main network properties, i.e. “degree distribution”, “network diameter”, “clustering coefficient”, and “betweenness” [113], are described below.

Degree is a property of a network node. It is equal to a number of direct connections with other nodes. Degree distribution is a characteristic of a network. It is the distribution of degrees of all nodes. In mathematical terms, degree distribution is denoted as *P(k)* and corresponds to a probability that any randomly chosen node has degree k. It has been demonstrated that different types of biological networks, such as protein-protein interactions, transcription factor binding sites, metabolic and genetic, have degree distribution significantly different from the corresponding property of a random network, reviewed in [112,114,115]. Degree distribution of these networks is scale-free that follows a “power law” distribution, $(k) \propto k^{-\gamma}$, in contrast to Poisson degree distribution characteristic to random networks. Such non-random network property of degree distribution shows that in biological networks there is a small number of highly connected nodes with high degrees termed “hubs” with the majority of nodes having a small number of connections. The possible advantage of such organization is that network is less sensitive to perturbation, thus is more robust, because inactivation of a random connection between nonhub nodes has less impact on the scale-free network structure comparing to random network. Analyses of hub nodes that are important for scale-free topology demonstrate that they tend to be more essential and more conserved than nonhub nodes [116,117]. These findings provide evidence that inferred network properties have biologically important interpretation.

“Small world” property is another non-random characteristic of biological networks. Small world networks have a small network diameter, which is defined as a maximum distance between two network nodes. A small diameter indicates that any two nodes can be connected with a relatively short path, which in the signaling networks facilitates information flow and optimizes metabolite transfer in networks of metabolic reactions [118].

Clustering coefficient is a percentage of existing connections of a node to its neighboring nodes to a total number of possible connections. In networks with high clustering coefficient if node A is connected to node B and node B is connected to C, then there is a higher probability that A and C are also connected. High clustering coefficient is another indicator of a small world network.

The importance of a node to information flow though network can be assessed with betweenness measure, which is equal to a number of shortest paths from the whole network that pass through that node or edge. Proteins with high betweenness but low connectivity a (low degree) were found to be abundant in the yeast protein-protein interaction network suggesting that they play a role of connectors between network modules. Similarly to hubs, such proteins also appeared to be essential with higher probability than proteins that did not possess high betweenness and low connectivity features [119].

It's been argued whether scale-free topology of protein-protein interaction networks could be an artifact related to incompleteness of evaluated interactome [120] and whether better estimations of network topology might exist [114,121]. Nevertheless, pioneering studies of networks and their organization principles reveal topological patterns very distinct from random networks. Understanding of general network properties will continue to mature with acquisition of more accurate and more complete data about biological networks.

1.6.5.2. Subnetworks and modules

Although the ultimate goal is to understand how cell or organism functions as a whole, system complexity requires breaking networks into smaller parts for detailed investigation of their functions, such as subnetworks and modules described below.

1.6.5.2.1. Subnetworks

A division of network can be dictated by a specific scientific question and reflect a functional subnetwork of a particular cellular process. An example of such investigation is a recent study by Gong et al, which provides an atlas of chaperone-protein interactions in *Saccharomyces cerevisiae* [122]. Authors performed investigation of protein-protein interactions of the 63 known yeast chaperons that allowed to: identify functionally promiscuous chaperones and chaperones that are functionally specific; find properties of interacting proteins that are associated with increased number of encountered chaperones (e.g. protein length and proximity to nucleus); and reveal the presence of multi-component chaperone modules. Other examples of subnetworks investigations include functional dissection of complexes involved in yeast chromosome biology [123] and a small ubiquitin-related modifier system [124]. The investigation of subnetwork brings to light details about molecular communications that are hidden in studies of global network properties. The common motives of subnetwork analysis are: a) identification of partnership between members within a particular process; b) observation of links that connect the subnetwork with other cellular pathways and processes; and c) application of various network clustering techniques to identify network modules, i.e. densely interconnected network subgraphs with limited connectivity with the rest of the network.

1.6.5.2.2. Modules

Network modules are of great interest as they provide a hint about organization of the cellular machinery from protein association data. Such modules correspond to protein complexes, i.e. proteasome and DNA/RNA polymerases, or consist of proteins that cooperate for performing a common cellular task. Analysis of early data on protein-protein interaction networks demonstrated that members of network modules tend to share the same functional annotation [104]. However, modules are rarely consisting of only known members with defined function. They also comprise members with distinct functions or proteins with unknown function. Based on an assumption that members of the same module may be important for performing the same cellular task, an unknown protein could be suspected to perform the same function as the majority of the members of the module. Therefore, studying content of the network modules and associations between genes and proteins in general has a potential for identifying novel protein functions.

1.6.5.3. Protein function prediction

1.6.5.3.1. Computational methods for function prediction

Function assignment to an uncharacterized gene or protein based on function of its interacting partner is known as guilt-by-association concept. Even in extensively studied model organism *Saccharomyces cerevisiae* hundreds of genes comprising up to one third of the yeast genome have not been associated with any cellular function or process [125]. In a complex organism such as mammals, there are more genes with unknown function than genes with some level of characterization. Therefore, an opportunity to elucidate protein function is a highly influential factor for inspiring experimental and computational studies of cellular networks.

Recently, efficient computational methods were developed for automated function prediction based on networks of protein associations and annotation ontologies of characterized proteins. Networks analyzed by these methods are not limited to protein or genetic networks and can comprise other types of data, such as co-localization or co-expression [126–128]. Generally, predictions are based on propagation of functional annotations of known network nodes to other nodes through network connections. Function assignment is rarely unambiguous, thus prediction methods generate a score value that indicates how likely the suggested function is. Two types of approaches can be distinguished in function prediction methods: module assisted and direct annotation schemes.

Module assisted approaches find first network modules with various clustering techniques, such as commonly used hierarchical clustering [104] and Markov clustering algorithm [129]. Once a module is identified, a function is assigned to all members of the module based on the function of the majority of members or based on enrichment based on hypergeometric distribution P-value can be used to identify a functional overrepresentation in the module. The distribution is based on probability that out of m nodes in the module at least k are annotated with a specific function given that there are totally n nodes in the network and f of which are annotation with that function and is calculated as:

$$p = \sum_{i=k}^{m} \frac{\binom{f}{i}\binom{n-f}{m-i}}{\binom{n}{m}} \quad ,$$

where: n is the total number of nodes, m is the total number of nodes in a module, f is the total number of nodes with a specific annotation, k is the number of nodes in a module with a specific annotation, and $\binom{a}{b}$ is a binomial coefficient.

Direct approaches assign a function to a protein based on its network connections and known function of interacting partners. In the simplest form, a protein with most of neighbors with a specific function can be suspected to have that function. Such logic is called a neighbor counting [130]. An extension of the method called naïve Bayes label propagation can use edge weights instead of counting interactions. Edge weights correspond to a likelihood that two proteins participate in the same process based on the particular evidence about their connection. Both methods only take into account direct neighborhood of the node. Network diffusion methods are designed to propagate information through network and allowing effective usage of functional knowledge about interactions that are several steps away [131,132].

1.6.5.3.2. Experimental validation of functional predictions

Accuracy of large-scale association datasets as well as performance of prediction algorithms is routinely evaluated based on the ability to recapture known biological phenomena. However, experimental confirmation of potential novel findings is required in order to translate intriguing observations into biological facts. Recently, a number of studies have been performed for defining the molecular mechanism of fatty acid chain length control by membrane-imbedded elongase complexes [133], characterization of conserved Orm proteins as regulators of sphingolipid biosynthesis [134] and linking the GDP/GTP exchange factor Rom2p to the regulation of sphingolipid metabolism [135]. These studies are examples of biochemical validation of defined novel functions based on connection patterns with known members of lipid metabolism observed in the yeast gene-gene interaction network.

The confirmation of novel hypotheses with classical genetics and biochemical tools will remain a much slower process than the generation of such hypotheses. Thus, high-throughput technologies are needed to enable a more rapid confirmation of computational predictions. Several studies have successfully combined function prediction methods with an experimental confirmation for tens of genes simultaneously. Hess et al. identified 100 proteins related to mitochondrial biogenesis in *Saccharomyces cerevisiae* [136]. Network for prediction was constructed from diverse genomic data sources and a list of 106 genes known to be involved in mitochondrial biogenesis was used as a query for retrieving genes likely to have a similar function. Top scoring genes were experimentally validated by measuring the rate of generation of cells with respiratory dysfunctional mitochondria (petite cells) [137] and growth rates in respiratory and fermentative conditions. Deletion mutants lacking genes important for mitochondrial function demonstrated higher rates of petite cells formation comparing to wild type yeast and slower growth rates under respiratory condition. In another study, whole-animal *Caenorhabditis elegans* microarray data were used to predict tissue-specific gene expression confirmed with promoter-GFP constructs [138].

1.6.6. Methods for network visualization

Visualization of results is an important aspect of data analysis that helps researchers to organize information, find unexpected patterns, predict and propose new hypotheses that could explain the observations. Communicating results visually in a clear and simple manner becomes more and more challenging because the amount of data points and heterogeneity of the data produced by “omics” experiments are rapidly increasing [115]. We review below two main visualization methods and discuss directions for further improvements of graphical tools.

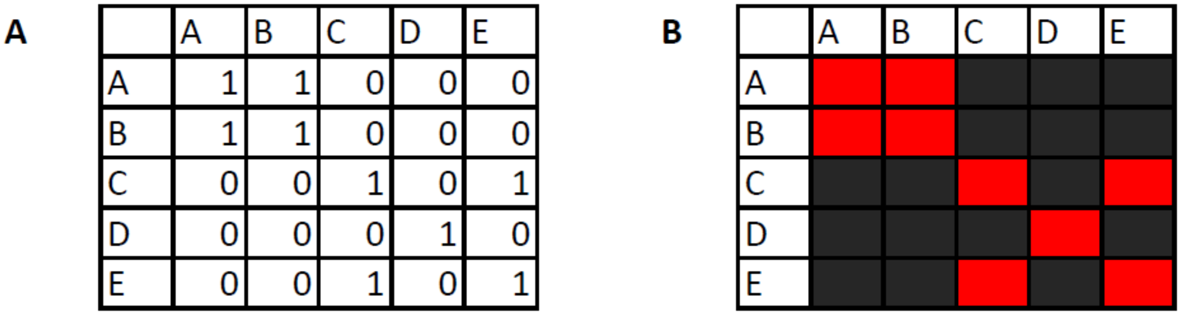
1.6.6.1. Graphs of nodes and edges

Graph of nodes connected with edges is the main method for visualizing networks of molecular interactions [139]. The main challenge of the graph representation is how to handle thousands of interactions between thousands of molecules, which is a typical scale of high-throughput experiments in yeast (in higher eukaryotes, these numbers are one to two orders of magnitude higher). Such network complexity unavoidably leads to the overlap between edges, which makes it difficult to follow connection paths between nodes and observe meaningful connectivity patterns. Dedicated network layout algorithms have been developed to address this problem and produce more informative graph visualizations by arranging network nodes in such a way that the overlap between edges is minimized [140,141]. For larger networks, clustering of interaction data can be performed before visualization to group together tightly connected components [107,142]. Graphs can be further simplified by combining in one node all clustered nodes of genes/proteins and using edges for showing connections between distinct clusters instead of visualizing all individual interactions [110]. These methods largely improve visualization clarity. However, they do not fully escape the problem of the overlap. There is still a need for further development of graphical methods especially for large networks.

Because of the popularity of the graph representation, a variety of software packages exist for representing networks as graphs, many of which are specifically dedicated to visualization of biological data, (for a comprehensive review of such software packages see reference [139]). At present, Cytoscape [143] is one of the most popular graph visualization tools in biological sciences. In addition to visualization, Cytoscape provides connectivity to major databases for extracting and annotating network data making it a handy data management tool. The modular architecture of Cytoscape that allows creation of plugins for extending a standard set of functionalities further increases popularity of the software.

1.6.6.2. Matrix-based representation

The main alternative to graphs of nodes and edges is a matrix-based representation of networks [144,145]. Matrix rows and columns correspond to genes and proteins of the network and corresponding matrix elements hold information on interaction. In a text format, such a matrix represents a table were numeric value of a matrix element indicates the relationship



	A	B	C	D	E
A	1	1	0	0	0
B	1	1	0	0	0
C	0	0	1	0	1
D	0	0	0	1	0
E	0	0	1	0	1

	A	B	C	D	E
A	🟥	🟥	⬛	⬛	⬛
B	🟥	🟥	⬛	⬛	⬛
C	⬛	⬛	🟥	⬛	🟥
D	⬛	⬛	⬛	🟥	⬛
E	⬛	⬛	🟥	⬛	🟥

Figure 1-2. Matrix-based network representation.
(A) In the text format, the relationships between rows and columns are coded by a number. **(B)** In the graphical format, the same relationships are visualized with different colors.

between a particular row and column (1 – interaction, 0 – no interaction) (Figure 1-2A). In a graphical format commonly referred to as “heat map”, the relationship between entities in rows and columns is indicated by a color of a corresponding matrix element (for example, red – interaction, black – no interaction) (Figure 1-2B). The matrix-based heat map representation is not limited to the interaction network. It has gained its popularity in biological sciences as a tool for visualizing genome wide expression profiles [17]. In this highly cited work with almost 14 000 citations recorded in Google Scholar, Eisen et al. placed genes into rows and experimental conditions into columns and used colors and color intensities to denote gene expression changes. Furthermore, the genes were grouped prior to visualization by hierarchical clustering that allowed to identify groups of coordinated genes [17]. The same workflow was used in multiple publications for displaying interaction networks and revealing modules of closely connected proteins [51,104,146]. The advantage of the heat map presentation over the graph of nodes and edges is the absence of the overlap between matrix elements that code interactions.

The heat map representation of interaction data is less common with only a few dedicated software packages available [17,147] that still lack the level of sophistication reached by the graph visualization software. An important functionality that is missing in the existing heat map software is the ability to compare several datasets in one figure. Such feature would be highly desirable because it would allow to compare visually results from different interaction screens as well as different types of molecular associations between the same nodes (e.g. physical and genetic interactions). In Chapter 4, we present an improved heat map representation method that we have developed for the matrix-based visualization of superimposed datasets. The method was implemented as a platform independent software package that was used for visualization of results presented in Chapter 2 (Figures 2-8, 2-12 and 2-13).

1.6.7. Interpretation of metabolic data

The second type of data that we analyze in the thesis, in addition to protein interactions, are metabolite abundances. In contrast to interactomics, that provides information about connections between cellular components, lipidomics and metabolomics quantifies abundances of small molecules that are products of enzymatic reactions not reflected in a genomic blueprint. Below, we discuss methods for interpretation of the metabolomics data.

1.6.7.1. Comparison of metabolite abundances

As described above, the first step of interactomics analysis is clustering that reveals organizational principles of interaction networks. In metabolomics, the first step is comparison of metabolite concentrations between different strains or growth conditions. Abundances of metabolites measured are evaluated based on methods described in section 1.6.1. Hypothesis testing and statistical significance. Comparison of metabolite abundances are summarized in tables that contain the information about a magnitude of differences between concentrations of metabolites and statistical significance of the differences.

1.6.7.2. Pathway analyses

Results of statistical analyses of metabolic abundances contain information on changes associated with single molecules. In order to enable investigation of how the whole metabolic system is changing in response to perturbation, a metabolic pathway map is needed. Unlike proteins that are encoded in a genome, metabolites are products of enzymatic reactions. Methods of analytical chemistry for detecting small molecules and biochemical characterization of enzymatic activities that modulate their concentrations are as important for metabolomics as genome sequencing efforts for genomics and proteomics. Results of metabolic studies are represented in pathway databases that connect metabolites through enzymatic reactions. One of the most developed resources for storing biochemical information is Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/ or http://www.kegg.jp/) [148]. KEGG is populated by biochemical pathways that are manually created by experts in the field of biochemistry and metabolism. The information is stored in a computer readable format, which supports development of bioinformatics tools that help visualizing organism specific pathway maps and projecting metabolic changes onto the pathway diagrams [149,150].

The development of genome-scale metabolic models is aimed at combining the information about different pathways into a unified computational framework [151]. Such representation allows to study an influence and interconnectivity of diverse metabolic pathways. Furthermore, the ability to incorporation kinetic and metabolite abundance information with metabolic models using dedicated computational tools facilitates mathematical modeling of metabolic systems [152].

1.6.8. Yeast databases

Sophisticated databases for storing biochemical information significantly contributed to the success of functional genomic screens and computation analyses of yeast *Saccharomyces cerevisiae*. These resources contain vast data collected for yeast and reflect tremendous efforts made by teams of bioinformaticians and biological curators for presenting the information in a format suitable for computations. The most important resources are discussed below.

Saccharomyces Genome Database (SGD) is one of the oldest sources of manually curated and high-throughput information collected from peer-reviewed literature (http://www.yeastgenome.org/) [153]. The resource contributes to standardization of yeast gene and protein name nomenclature, systematic curation of diverse sources of information, e.g. gene annotations, genomics, proteomics, interactomics, and development of computational tools for data retrieval and analysis. A unique feature of SGD is the information coverage and connectivity with other yeast related resources. This makes SGD a comprehensive solution that will guide a researcher to a dedicated resource based on the type of requested data.

The MIPS Comprehensive Yeast Genome Database (CYGD) is one of the most accessed databases for collecting manually curated information about validated protein complexes used for evaluating quality of high-throughput protein-protein interaction datasets (

http://mips.helmholtz-muenchen.de/genre/proj/yeast/) [154]. It should be noted, that despite an important role the resource had in pioneering interaction studies, its development has been discontinued. Therefore, other sources should be consulted for updated information on validated protein complexes, for example a catalogue created by Pu et al. [109].

Biological General Repository for Interaction Datasets (BIOGRID) is a database of protein-protein and gene-gene interactions (http://thebiogrid.org/) [155]. The database collects data for other organisms than yeast. The resource has a high quality of deposited information and is frequently updated. BIOGRID served as a template for a database of protein phosphorylation (PhosphoGRID http://www.phosphogrid.org/) [40].

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a well trusted resource of biochemical pathway maps that cover metabolic and signaling pathways (http://www.kegg.jp/) [148].

Gene Ontology Consortium (GO) is a source of standardized gene ontology information (http://www.geneontology.org/) [156]. GO provides consistent descriptions of gene products in different databases based on structured controlled vocabularies (ontologies). The structure of the ontology enables computational analysis of protein location, biological processes and molecular functions.

Because of the diversity of yeast data types, many more databases exist. A yearly update about resources dedicated to yeast and other organisms are published in a special database issue of Nucleic Acids Research journal (http://nar.oxfordjournals.org/) [157].

1.7. Rationale of the study

Despite the significant progress in understanding yeast biology powered by the genome sequencing project and functional genomics studies, functions of many yeast genes are still unknown. When we started our investigation, 30% of the yeast genes were not linked to any biological process or molecular function [125]. Because some genes are multifunctional [158], we might expect that alternative functionalities can be discovered for annotated genes as well. The identification of gene functions can be facilitated by analysis of genome-wide networks of molecular interactions based on associations of unknown genes with characterized partners [126]. However, the coverage of the yeast interaction network is still limited.

In this thesis, we sought to search for novel gene functions in yeast on a large scale. To achieve this, we set a goal to increase the coverage on the yeast interactome by performing a genome-wide screen for protein-protein interactions (Chapter 2). We employed PCA for discovering interactions between certain types of proteins that were not sufficiently sampled by existing methods. For mining the results of the screen, we wished to develop a visualization method for interactive comparison of large datasets with minimized overlap between the edges. We hoped that the visual network analysis would help to identify unique interactions not covered in previous studies (Chapter 4) and in combination with function prediction algorithms would link uncharacterized genes with established cellular processes. Finally, we aimed at developing a complementary screening approach for enabling validation of the predicted functions on a large scale (Chapter 3).

Chapter 2 : An in vivo map of the yeast protein interactome

2.1. Contribution to the published work

Results presented in this chapter were published in the following paper: Tarassov, K.*, Messier, V.*, Landry, C. R.*, Radinovic, S.*, Serna Molina, M. M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., Michnick, S. W. (2008). An in vivo map of the yeast protein interactome. Science (New York, N.Y.), 320(5882), 1465–70. [55] DOI:10.1126/science.1153878. * These authors contributed equally).

Conducting a project of such scale required an expertise of several researches for optimization of the screening procedures, development of robotic automation protocols, performing necessary control experiments and analyzing the results. My contribution to the study was dedicated to the development of programming, computational and data analysis methods for performing the PCA based high-throughput screen and publishing the results. R.S., S.M.M.M., S.I, M.Y. developed the experimental procedures for performing the high-throughput screen, R.S. developed the robotics and high-density cell culturing. L.C.R, M.V. and M.S.W. described biological examples of autophagy and bud neck network organization based on network clustering results performed by T.K. M.V. contributed to high-throughput experiments and conducted control experiments for spontaneous refolding, tested PCA reversibility and analyzed how mapping of RNA polymerase II complex corresponded to distance constrain introduced by DHFR PCA. L.C.R. performed the analyses for figure 3 B, C and D, 5B, S7 and S8, and created final versions of other figures (1, 2, 3A, 4) for which T.K.

performed the analyses and contributed with graphics. B.H. helped plan the research and edited the manuscript. V.J. helped with the biological examples presented in the paper and edited the manuscript. L.C.R. and M.S.W. drafted the manuscript with contributions from M.V. and T.K. M.S.S. formulated the original idea of the screen using PCA technique, contributed to data analyses and writing the manuscript.

In the published paper, to comply with the limits on number of words and figures allowed for papers in Science, computational methods were described only briefly and more details went into the Supplementary Methods. In this chapter, I present the work from the computational and data mining perspective. I briefly describe experimental methods for DHFR PCA screen (section 2.4.1 Experimental set-up) and focus in detail on data analysis methods that I have developed and applied for converting experimental results of a survival assay into an interaction network suitable for computational analyses (sections 2.4.2 to 2.5.5). In section 2.6, I discuss the novelty of the results and findings that inspired the follow-up investigation described in Chapter 3. Figures 2-1 to 2.13 presented in the chapter are from the published article reproduced with permission from the publisher.

2.2. Abstract

The cellular homeostasis is shaped by complex networks of molecular interactions and regulatory events. The understanding of the cell functions requires systems-level approaches capable of measuring thousands of cellular events at a time. For increasing the coverage of the interactome of yeast, *Saccharomyces cerevisiae*, we have performed a systematic screen for protein-protein interactions. We have developed a high-throughput version of a protein-

fragment complementation assay that allowed to detect *in vivo* interactions between endogenously expressed proteins on a genome-wide scale. We identified 2770 interactions between 1124 proteins, the majority of which were novel. The method covered particularly well interactions between membrane and lipid related proteins, which play a key role in such processes as energy storage and signaling.

2.3. Introduction

The investigation of networks of different types of molecular interactions is a promising approach for understanding cellular architecture. The availability of the yeast genome sequence led to development of powerful genome-wide strategies for studying transcript and protein abundances, protein localization and phenotypes of gene deletion mutants (section 1.3. Functional genomics studies in yeast). Recent studies, that utilized Y2H and TAP-MS methods, demonstrated that protein-protein interactions could be studied on a large-scale as well. However, these methods are associated with a number of limitations, such as detection of interactions in the unnatural environment or requirement of expensive analytical equipment (section 1.4. Strategies for mapping protein-protein interaction networks). The PCA is capable of detecting protein-protein interactions between proteins expressed from their natural promoters in the physiological environment. Furthermore, interactions could be observed *in vivo* by a scalable survival assay (section 1.4.3. Protein-fragment complementation assays). These advantages make the PCA method an attractive alternative to Y2H and TAP-MS methods for conducting genome-wide studies of protein-protein interactions. Moreover, the published works covered only a fraction of the yeast interactome. Thus, a large-scale screen by an alternative PCA method would contribute to the elucidation of protein-protein interaction network.

2.4. Results

2.4.1. Experimental set-up

The systematic genome-wide screen for protein-protein interactions in yeast was developed as a survival assay based on inhibition of dihydrofolate reductase activity (Figure 2-1). Dihydrofolate reductase (DHFR) enzyme catalyzes conversion of dihydrofolic acid to tetrahydrofolic acid, which is an important intermediate in *de novo* synthesis of nucleic and amino acids. DHFR is found in all organisms and encoded in yeast by *DFR1* gene. Inhibition of DHFR by methotrexate impairs cell growth making yeast inviable on minimal media similarly to deletion mutant of *DFR1* which is auxotrophic for dTMP, adenine, histidine and methionine [159].

To enable PCA screen in yeast, a double mutant of murine dihydrofolate reductase (mDHFR) was developed in our laboratory, in which catalytic activity was preserved, but sensitivity to methotrexate was reduced by 10000-fold relative to the native yeast protein [160]. mDHFR protein-complementation assay (DHFR PCA) uses two split fragments of the mutant mDHFR F[1,2] and F[3] that are fused to a pair of proteins. Interaction between two proteins brings the fragments into proximity causing reconstitution of the mDHFR activity allowing cells to grow in the presence of methotrexate.

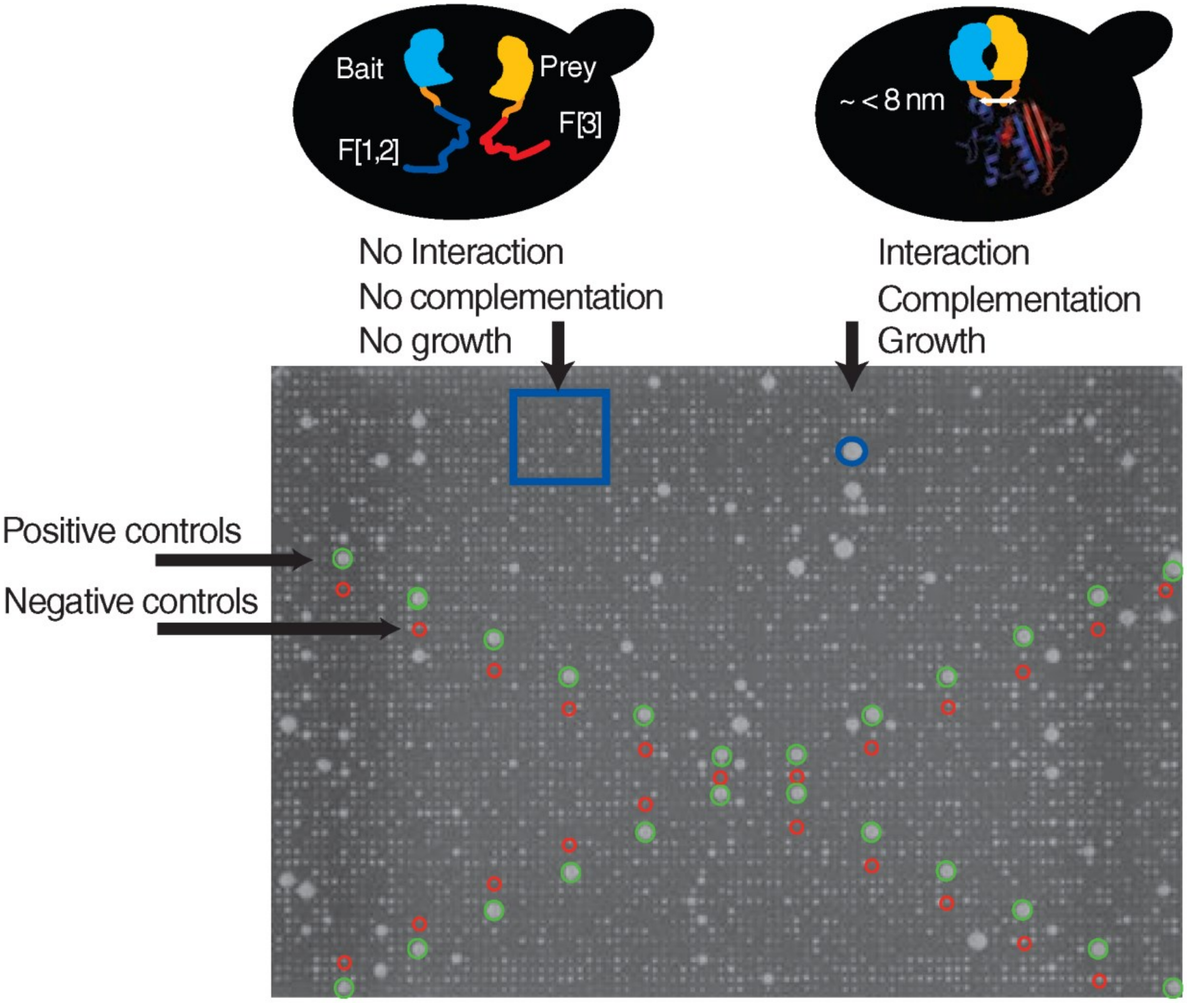


Figure 2-1. In vivo PCA screen of the yeast protein-protein interaction network.
Strategy for high-density array screening of the yeast protein-protein interaction network by DHFR PCA. Both positive [green circles (MATa/α, CDC19 fused to DHFR fragment [1,2] (CDC19-F[1,2]), and MCK1 fused to DHFR fragment. DOI:10.1126/science.1153878.

For successful protein-protein interaction screening, the assay should have high sensitivity and should be devoid of biases that can be introduced if physiological association/disassociation of interacting proteins is impaired because of trapped reconstituted fragments that are unable to unfold when interaction of linked proteins is disrupted. Experiments with DHFR PCA in mammalian and bacterial cells demonstrated a high sensitivity of the assay with potential to detect from 25 to 100 copies of complexes of interacting proteins with reconstituted DHFR fragments per cell [161,162].

Reversibility of the DHFR PCA was tested in yeast by an *in vitro* binding assay using resin-immobilized cAMP. In this assay, subunits of yeast serine/threonine complex Bcy1 and Tpk2 were fused with mDHFR fragments. After cAMP induced disruption of the Bcy1/Tpk2 complex, Bcy1 remained bound to cAMP resin while Tpk2 was found in the unbound fraction, which demonstrated that refolding of the DHFR fragments is reversible. For enabling genome-wide screening for protein-protein interactions using DHFR PCA, we developed a high-throughput platform that employs tagging of the yeast genes with the reporter fragments in MATa and MATα strains, growth of the haploid mutants expressing proteins fused with fragments, mating of the haploid strains, methotrexate selection for strains with reconstituted DHFR activity and quantification of colony intensities (Figure 2-2). We attempted to tag with both fragments each of the of the 5756 consensus yeast genes based on annotations from Saccharomyces Genome Database (SGD) [153]. We created gene specific homologous recombination cassettes with sequences of the fragments and fragment specific antibiotic resistance cassettes. Recombinant DHFR PCA cassettes with fragment F[1,2] were transformed into MATa strain, and cassettes with fragment F[3] were transformed into MATα strain.

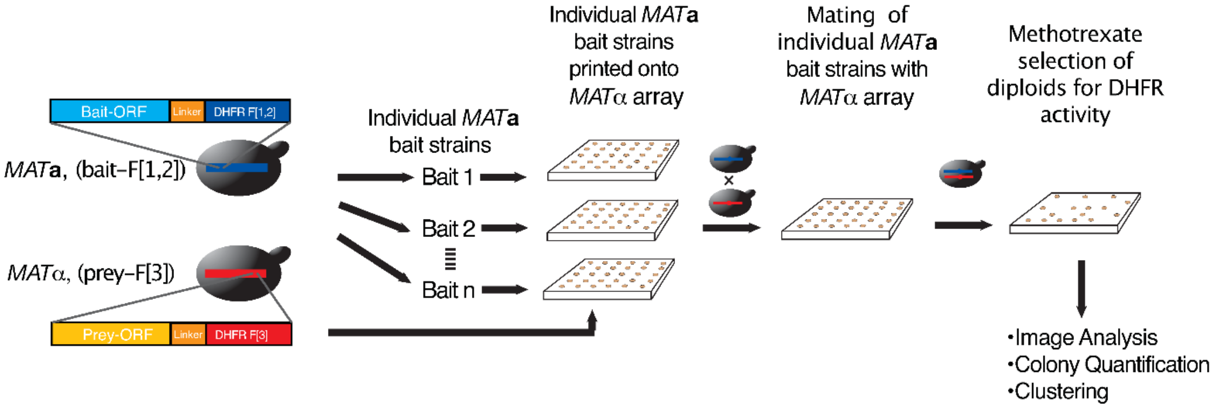

Bait-ORF
Linker
DHFR F[1,2]
MATa, (bait-F[1,2])
MATα, (prey-F[3])
Prey-ORF
Linker
DHFR F[3]
Individual MATa bait strains
Bait 1
Bait 2
Bait n
Individual MATa bait strains printed onto MATα array
×
Mating of individual MATa bait strains with MATα array
Methotrexate selection of diploids for DHFR activity
•Image Analysis
•Colony Quantification
•Clustering

Figure 2-2. Experimental set-up of a large-scale PCA screen of the yeast protein-protein interaction network.

Experimental strategy for single bait versus prey array screening of the yeast protein-protein interactions by DHFR PCA. DOI:10.1126/science.1153878.

Success of the recombination was tested by PCR using open reading frame (ORF) specific diagnostic primers from a region of 100 to 1,000 base pairs downstream of each ORF and a common primer within the terminator sequence of the antibiotic resistance marker.We verified correctness of the recombination by comparing sizes of the PCR products with expected calculated values. This analysis confirmed that 4326 (75%) ORFs with the DHFR F[1,2] fragment in MATa and 4804 (83%) ORFs with the DHFR F[3] fragment in MATα strains were transformed correctly (Table S1[*]). The screening was performed on solid-phase medium by crossing a printed array of MATα strains containing all ORF-DHFR F[3] fusions, which we called "preys", with individually grown MATa strains containing ORF-DHFR F[1,2] fusions, which we called "bait", one at a time.

We optimized mating (temperature and incubation time) and selection (amount of methotrexate) conditions by performing a screening of a subset of 380 bait and prey strains that were proteins known to interact based on manually curated information collected from CYGD MIPS database [154], a well trusted repository commonly used for benchmarking protein-protein interactions [51]. From the range of temperatures, incubation times and methotrexate concentrations we selected parameters that maximized the number of known interactions detected by DHFR PCA, i.e. true-positive results, while minimizing background growth of the colonies that leads to the false-positive results.

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

For conducting the optimized screen, the prey array was generated with very high-density using a robotic pin-tool with capacity of 6144 colonies per plate. Such density allowed us to fit the entire collection of prey strains onto a single plate and include replicates of positive and negative controls for monitoring the quality of cell transfer and the selection process. As positive controls, we used 24 replicates of MATa/ MATα diploid strains with DHFR PCA fusions with known interacting proteins Mck1p and Cdc19p. As negative controls, we used 24 replicates of fusions with a pair of proteins Cln3p and Cdc19p whose interaction was never observed; hence, such diploid is inviable in the presence of methotrexate. The control strains were positioned in cross-shaped patterns to cover areas of the whole array (Figure 2-1). The haploid MATα prey was replicated from the master array and mated with each bait MATa strain individually. The diploids were grown on solid-phase minimal medium in the presence of methotrexate.

In total, mating of 3247 bait strains was performed successfully allowing to test for interaction over 15 million protein pairs. Protein interactions were detected by novel image analysis and statistical algorithms developed for processing the photographs of the selection plates, quantification of colony sizes of the diploid strains and calculation of cut-off values that determined whether an interaction has occurred.

2.4.2. Automated image analysis

Previous high-throughput screens of protein-protein interactions in yeast relied on mass spectrometric analyses to determine which proteins belong to the same complexes of interacting subunits [64,65] or sequencing of the selected colonies [57,58]. Our survival-based assay required application of an image analysis method that quantified colony sizes of diploid strains. Similar strategy was employed in an array version of Y2H screen performed by Uetz et al. [57] and in screens for genetic interactions [32,41,163]. However, to the best of our knowledge, none of the previous screens were performed in such density as ours with maximal capacity of 6144 colonies pined onto a standard microtiter-sized plate (the array presented by Uetz et al. comprised 16 plates with 384 strains per plate [57]; arrays of Tong et al. [163] and Parsons et al. [32] consisted of 384 clones per plate in duplicate resulting in 768 colonies per plate). While such design allowed us to increase considerably the throughput and cost efficiency of the screen, it introduced processing challenges that could not be handled by existing image analysis routines available in popular image processing software, such as ImageJ [164] or NIHimage [165]. High density greatly decreased the distance between colonies, thus colony positions had to be determined with very high precision to allow correct identification of each particular colony. In addition, each plate contained a different number of colonies that survived the selection pressure and empty spots had to be assigned correctly to diploids that did not survive. Furthermore, because of limited space between the colonies, some diploids in which interaction was detected, grew over their expected boundaries and overlapped with neighboring colonies. In order to overcome these challenges, we developed image analysis routines implemented in MATLAB Image Processing Toolbox (Figure 2-3).

For the accurate determination of colony centers, we created a software interface for semi-automated grid positioning. Initial coordinates were input by a user who pointed with a computer mouse at centers of the first (top-left plate corner) and the last (bottom-right plate corner) colonies on the plate. Based on these data, the program rotated and scaled the pictures to position each row and column of colonies on a straight line, thus minimizing the effect of variation in plating or picture rotation during image acquisition. Manual input was required only for the first plate and plates whose calculated colony center coordinates after adjustment were deviating considerably from coordinates from the previous plate. In such case, the program asked again for manual definition of edge coordinates for setting new scaling and rotation parameters. In order to facilitate verification of the alignment, an image of a grid with 96 columns and 64 rows was superimposed with the plate image. Manual inspection of these images was performed to correct the alignment; reject 44 plates because colony growth was detected in the empty positions, suggesting that there was a misalignment of the grid during pinning or cross-contamination; and exclude colonies positioned too close to the plastic plate edges from further analysis. Furthermore, we implemented automatic checking of the grid alignment based on 24 positive and 24 negative control strains arranged in X-patterns across the plates. In the case of correct alignment, all positions that corresponded to positive controls contained large colonies, while all positions that corresponded to negative controls showed limited or no growth (Figure 2-1.).

Aligned images were enhanced by standard MATLAB Image Processing Toolbox functions to correct for nonuniform illumination and remove artifacts: bubbles, gel scratches, traces of the pinning tool, and plate edges. For identification of individual colonies, we applied a popular Otsu’s automatic thresholding method [166], which is well suited for discriminating between highly contrasted objects such as colonies of growing yeast and gel of the medium. The method is based on the assumption that the picture can be represented with two types of pixels: foreground, which correspond to colonies, and pixels of the background. The algorithm selects an optimal intensity threshold that best separates those two classes assigning all pixels above the threshold value to the foreground (pixel value = 1) and below threshold to the background (pixel value = 0). Such binary image was used to define a colony as a cluster of connected foreground pixels. Coordinates of the colony from the binary image were used for calculating the sum of the pixel intensities from the original image as a measure of a colony size. Next, we determined coordinates of colony centers by adopting an algorithm for spot detection based on percentiles of intensity [167] for handling high-density arrays. The algorithm is based on the observation that pixel intensities at colony centers have higher values. Algorithm processed one row and column of pixels at a time and stored values of 75th percentile of pixel intensities in an array. Next, intersection of rows and columns with highest 75th percentile values was used to determine position of centers of the arrayed colonies, which served for assigning each colony to the corresponding bait-prey pair. In the case with no overlap, these steps were sufficient to quantify colony sizes. However, algorithm improvements were necessary for deconvoluting images of fused colonies. A dedicated processing step identified such colonies by searching for single connected foreground objects that contained multiple centers within their boundaries. Colony deconovolution falls under a general category of tasks of image processing referred to as segmentation, i.e. separation of objects from one another. The watershed transform [168,169] is a common method applied for solving this task and is a default segmentation method in MATLAB Image Processing Toolbox. In topography, watersheds are ridges, like hills and mountains, that separate areas of convergence of surface waters called catchment basins. In image processing, a grey scale image can be considered as a surface, in which grey scale intensity defines the altitude of an object. In the case of the plate images from the protein-protein interaction screen, colonies were brighter and were placed higher than darker background. The watershed transform worked best with grey scale images for segmenting most of the overlapping colonies. However, pixel intensity values of colonies from overexposed images and large overlapping colonies were not sufficiently different. In such case, a strategy for segmentation of binary images was applied. Because pixels in binary image has only two values, 1 for foreground and 0 for background, there is not enough information for building a surface plot based on intensity. The surface of objects on binary images can be created by applying distance transform [170,171] that assigns for each pixels a value equal to distance (in pixels) to the nearest pixel that corresponds to the background. As a result, pixels at centers of the objects receive the highest values that descend towards the edges. The fact that colonies have circular shapes allowed to further improve the algorithm by combining distance transform with the Circular Hough transform [172], a method that detects circular shapes. The Circular Hough transform is a voting procedure that assigns higher scores to pixels at centers of objects in which a circle can be fitted. By superimposing the distance transform results with local maximum of the Circular Hough transform, we created a surface with higher altitude difference between centers and edges of the colonies. The watershed transform performed on superimposed images segmented the colonies more accurately than if applied on surfaces produced by the distance transform or the Circular Hough transform alone. At the final processing step, pixels lying on the identified boundaries between overlapping colonies were set to 0 on the original binary image and steps for quantifying non-overlapping colonies were repeated as described above. As a result, we obtained a matrix of colony intensities with rows corresponding to bait strains and columns corresponding to prey strains.

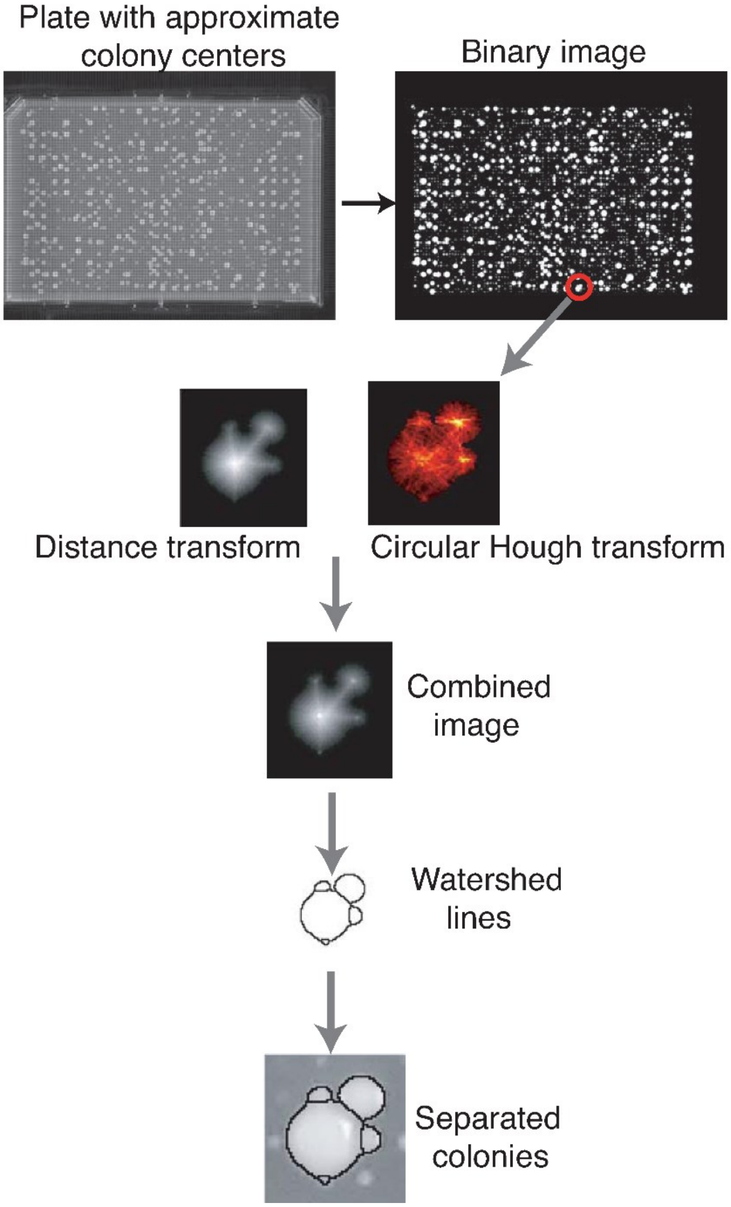


Figure 2-3. Automated extraction of colony intensities on plates.
The DHFR PCA results were inferred from the growth of diploid colonies on plates containing methotrexate. Images of the plates were taken after a 90-hour growth period with a 4.0 Mega pixel camera (Powershot A520, Canon). Plate images were saved in JPG format at a resolution of 180 dpi and a size of 2,272 × 1,704 pixels. In order to extract the intensity of the colonies, we used available image recognition routines available in the MATLAB image analysis toolbox and we modified parameters for it to be able to differentiate colonies that are in proximity to each other. The quality of the position of the grid and the recognition algorithm was examined through visual inspection of all plates. DOI:10.1126/science.1153878.

2.4.3. Data filtering

Evaluation of results of the image analysis revealed two issues that had to be considered to avoid false-positive results. The first issue is related to possibly non-specific interactions caused by experimental biases, and the second is how to discriminate colonies that grew because of an interaction from the background growth.

2.4.3.1. Detection of non-specific interactions

In our screen, certain proteins interacted with a very high number of other proteins (up to few hundreds). Importantly, some of these proteins were identified as promiscuous proteins involved in many non-specific interactions by earlier TAP-MS screens [59,64,65]. Reproducibility of such interactions and their clearly higher signal comparing to the background noise, suggested that they are not artifacts of the processing routines, but arise from biases of the experimental technique. For PCA, such patterns of interactions could be observed if certain proteins would attract complementary PCA fragment regardless of what protein is attached to it. To test for such possibility, we performed control experiments screening for all strains against fragments F[1,2] and F[3] alone or with attached peptide linker sequence not fused to any protein. We also checked for spontaneous complementation of the individual fragments. As expected, there was no interaction between fragments alone. However, we identified 344 proteins (Table S2[*]) that interacted with complementary fragments with or without linker peptides not attached to any proteins.

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

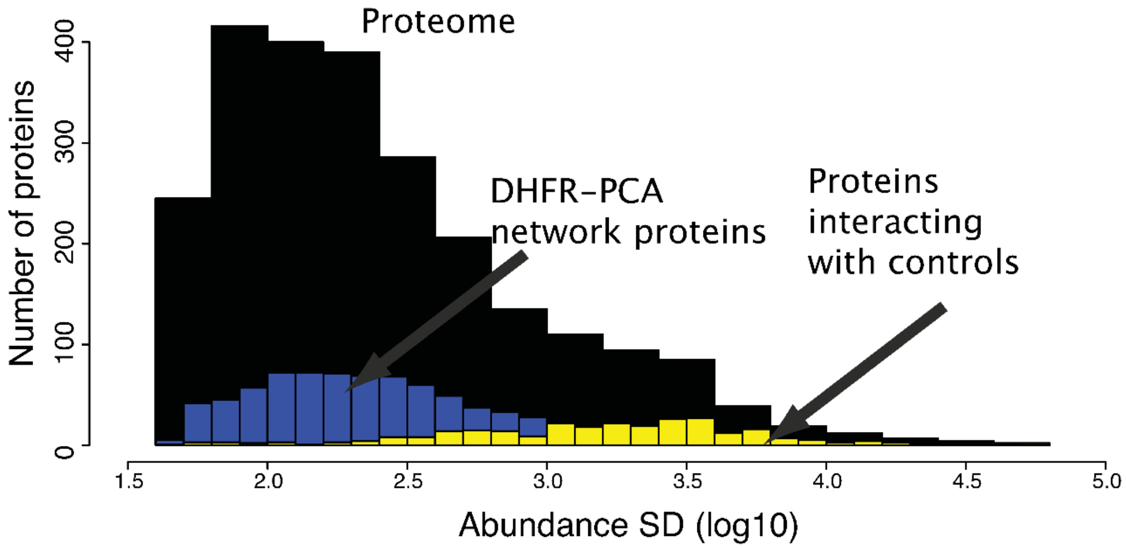


Figure 2-4. Distribution of protein abundance.
The distribution of protein abundance for cells grown on the same (SC, SD + glucose (from [21])) medium used in the DHFR PCA screen of the entire proteome (black), proteins of the DHFR PCA network (blue) and proteins interaction with the control fragments (yellow).
DOI:10.1126/science.1153878.

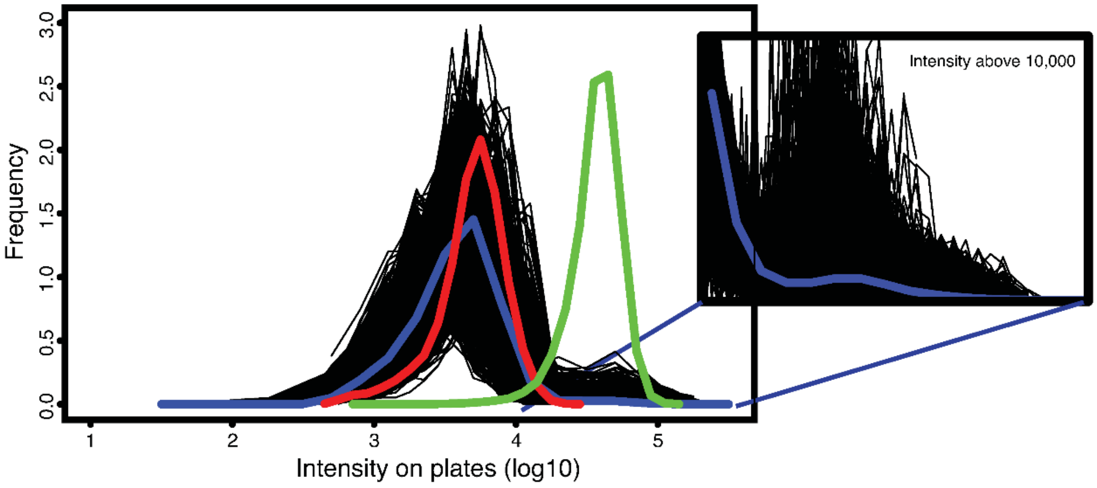


Figure 2-5. Distribution of colony intensities on plates.
Raw intensities prior to filtering. Black lines represent the intensity distributions on individual plates. The blue line represents the distribution of colony intensities across the entire experiment. The red and green lines represent the intensity distribution of the negative and positive controls respectively. The second panel shows the distribution of colony intensities above 10 000 to illustrate the growth of the methotrexate resistant diploid strains
DOI:10.1126/science.1153878.

Thus, any interaction detected in the screen involving one of such proteins could be unspecific. We excluded such interactions from the final filtered network, but reported them as supplementary data (Table S4*). Promiscuous proteins in our screen were enriched for ribosomal and ribosome associated proteins ($P < 2x10^{-65}$) and were more abundant than the proteome on average (median log10(abundance in SD medium) = 3.27 vs 2.28 for the proteome, Wilcoxon rank-sum test: $P < 2.2x10^{-16}$; Figure 2-4). Similarly, ribosomal proteins, high-abundance proteins and a few other proteins, such as Cdc19p, Eno2p, Tef2p and Tef3p, were identified as promiscuous proteins by TAP-MS screens [65].

2.4.3.2. Interaction signal vs. background noise

During the large-scale screen we used a relatively long time for growing diploid strains in the presence of methotrexate allowing cells to grow for 96 hours at 30°C before taking the plate pictures. These conditions were selected to increase the true-positive detection rate based on the control experiments with known interactions described above. Longer incubation time increased the chances of observing weak and transient interactions and interactions that involve low-abundance proteins. However, 96 hours incubation led to observable background colony growth on most of the plates. We describe below statistical procedures implemented to discriminate between the signal and the background.

First, we evaluated distribution of colony intensities in the entire screen after exclusion of signal from the strains with promiscuous proteins (Figure 2-5). The distribution was bimodal with the first distribution peak at around 4 000-5 000 counts and the second distribution peak at around 40 000 counts. Importantly, the medians of negative (5 192 counts) and positive (38 361 counts) controls overlapped with centers of the first and the second distribution respectively. This observation made us conclude that the first distribution corresponded to background colony growth and the second distribution originated from strains that survived methotrexate selection due to the interaction. A minimum overlap between the two distributions was around 20 000 counts, so we selected that value as a starting point for optimizing the selection of thresholds for inferring protein-protein interactions. With the total colony intensity threshold, we sought to define a minimum size for a colony to consider that an interaction has occurred. However, a plate-to-plate variation is a common phenomenon in high-throughput screens. Applying one threshold for processing thousands of plates is not optimal for separation of signal over noise [173]. We noticed that on some plates all colonies grew better or worse than on average in the screen. For plates with better growth, an intensity cutoff of 20 000 counts would select more false-positives, while on plates with generally poor growth, values slightly higher than 20 000 counts would correspond to the best interactions on such plates, rejection of which would lead to false-negative results.

To standardize the intensities and adjust for the plate-to-plate variation, we applied within-plate z-score transformation. For each colony, z-score was calculated as:

$$\text{z-score}_{i.j} = (\mathrm{I}_{i,j} - \mu_j))/\sigma_j,$$

where $\mathrm{I}_{i,j}$ is intensity of a colony i on plate j, μ_j is an average intensity on plate j, and σ_j is a standard deviation of intensity values on plate j.

Next, we selected optimal intensity and z-score thresholds that led to detection of maximum number of true-positive interactions, while minimized number of detected false-positive interactions (Figure 2-6.). A gold-standard positive reference set of known interactions was selected from CYGD MIPS [154], the same data source we used for optimizing experimental conditions as described above and that was used for benchmarking protein-protein interactions quality in previous large-scale experimental and *in silico* studies [51,174]. From 11 005 interactions among 1 236 proteins we selected a positive reference list of 503 interactions that could be potentially detected in our screen. We judged that an interaction could be detected if corresponding baits and preys were present in our collection, and for each protein, there was an interaction signal with at least one other protein, suggesting that the DHFR-ORFs constructs were expressed correctly. As a set of negative examples, i.e. interactions that most likely don't occur, we have selected pairs of proteins that belong to complexes present in separated cellular compartments or demonstrate anticorrelated patterns of gene expression as described by Collins et al. [51] (266 858 of interactions 6 377 of which could be potentially detected in our screen).

For optimization of the intensity and z-score thresholds, we applied an iterative process, in which for a pair of intensity (from 20 000 to 35 000 counts) and z-score (from 2 to 3.5) values a positive predictive value (PPV) was calculated as a ratio of number of true-positive interactions (interactions from the positive reference set) divided by the sum of true-positive and false-positive interactions (interactions from the negative reference set) (Figure 2-6 A). Based on the plot of intensities, z-scores and PPV values, we set the intensity threshold equal to 23 000 counts and the z-score equal to 2.4. With these thresholds, our screen achieved a PPV value of 97.7% that is comparable in precision to other large-scale as well as high-quality small-scale protein interaction studies (Figure 2-6 B, Figure 2-7). To this end, after applying intensity and z-score thresholds and filtering out interactions involving 344 promiscuous proteins, we detected 5 672 interactions at PPV score of 97.7%. However, we noticed that more than half of these interactions involved one of 83 highly connected proteins with repetitive interaction patterns similar to experimentally detected promiscuous proteins. Because of ambiguity of such connections, we decided to present these interactions as supplementary network (Table S4[*]). We marked with type 1 interactions that involved eight strains that consistently demonstrated higher growth pattern than other MATα stains. Type 2 denoted interactions involving 23 proteins with a very similar interactions pattern as the 344 promiscuous proteins. Finally, with type 3 we marked 1 830 interactions mediated by 53 proteins that showed distinct, but still very repetitive interaction pattern.

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

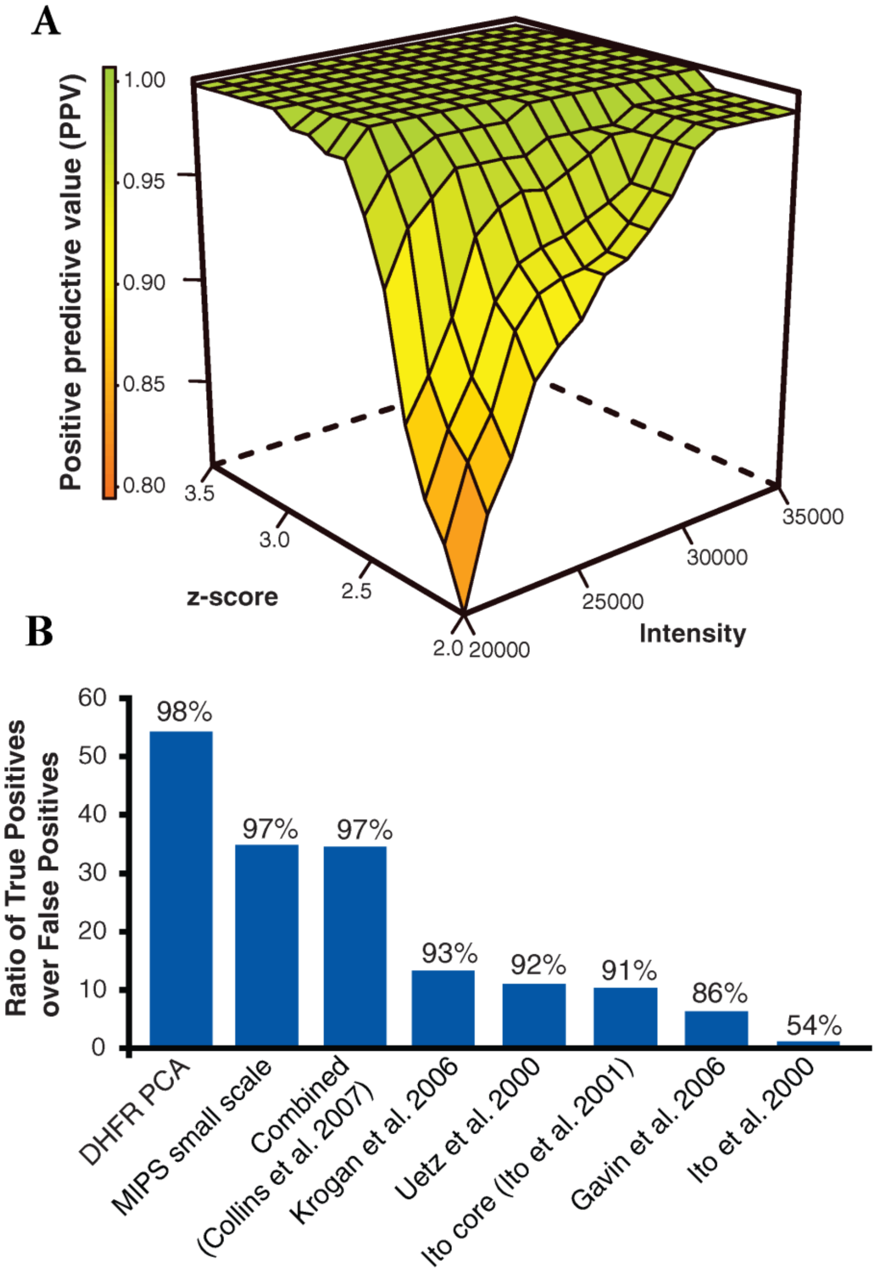


Figure 2-6. Quality assessment of DHFP- protein-protein interaction network.
(A) PPV score as a function of raw colony intensity and z-score (relative colony intensity on plates). This score represents the ratio of the number of true positive interactions over the sum of the true positive and false positive interactions predicted from the reference sets.
(B) The ratio of true positives to false positives in the DHFR PCA network compared with other large-scale data sets [51,57,58,64,65,189]. The achieved PPV is indicated above the bars.
DOI:10.1126/science.1153878.

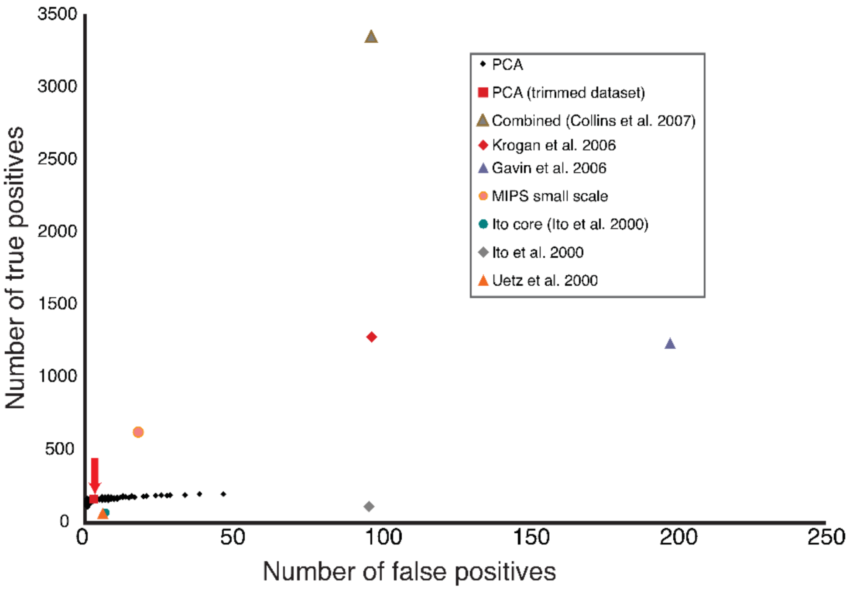


Figure 2-7. True Positives and True Negatives in PCA and other studies.
The curve represents the total number of true positive interactions and the total number of false positive interactions as a function of the score thresholds for defining protein-protein interactions in the DHFR PCA screen (ROC curve). Values for published datasets are shown as well as values of the final DHFR PCA networks. Sources for the other networks are described in the Materials and Methods Section. DOI:10.1126/science.1153878.

2.4.3.3. Final filtered network

After removal of spurious interactions with promiscuous proteins, our final filtered dataset consisted of 2770 unique interactions among 1124 endogenously expressed proteins with PPV value of 98.2% (Table S4[*]). In that dataset, we counted 3 false-positive and 163 true-positive interactions from the gold-standard dataset as compared to 2.7 true-positive and 33.8 false-positive interactions calculated for a randomly constructed network with the same number of interactions and proteins (randomization was repeated 10 000 times). Thus, it is unlikely that our PPV value is due to chance. After removal of high-abundance promiscuous proteins, the abundance of the remaining final network members was not significantly different from the average yeast proteome abundance [the median log10(protein abundance) = 2.32 versus 2.28; Wilcoxon rank-sum test, P= 0.19] (Figure 2-4), which reflects the high sensitivity of the DHFR PCA assay.

2.4.4. Hierarchical clustering of the yeast in vivo protein-protein interaction network

We identified from the literature three main graphical techniques that were utilized for visualization of results of previous large-scale studies: graph representation of nodes that correspond to individual proteins connected with edges if two proteins interact [64,66]; graphs with nodes that correspond to clusters of proteins or protein complexes connected by interaction edges [59,65,110]; and heat map representation of hierarchical clustering results [51]. For

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

presenting an overview of global network organization of the network of DHFR PCA interactions, we utilized the heat map presentation because, in our opinion, it is better suited for browsing large networks. Heat maps avoid the overlap between edges by placing proteins into rows and columns of a rectangular table marking an interaction with a colored square in the corresponding row/column position. In contrast, visualization of large networks as graphs with many nodes leads to overlap between edges, which makes it difficult to follow the connection. Dividing proteins into modules prior to building graphs helps to avoid clutter but, unlike TAP-MS methods, PCA detects pairwise interactions and not clusters of proteins. Thus, *a prio*ry we don't aim at fitting our results into complexes.

Unsupervised hierarchical clustering of 2770 DHFR PCA interactions grouped proteins into tightly interconnected clusters positioned along the diagonal of the heat map (Figure 2-8). Connections between clusters were positioned as off-diagonal elements. Careful examination of members of identified clusters revealed that many of them corresponded to known complexes with crystallographically or biochemically defined composition. Importantly, identification of known complexes in our network suggests that new modules or module members can be used with confidence to investigate novel aspects of the organization of cellular machines.

We used heat map representation for displaying enrichment/depletion with interactions between proteins annotated with specific Gene Ontology (GO) [156] terms related to Cellular Compartment (CC), Molecular Function (MF) and Biological Progress (BP) (Figures 2-12, 2-13). We further developed the heat map representation method and placed different information above and below the main heat map diagonal. Thus, for a pair of groups of proteins we could display on the same figure the number of interactions between the groups and information about how different this number was from random expectation. The method and the software that we developed for such visualization is discussed in detail in Chapter 4. Based on these heat maps, we investigated, which of the interactions we detected corresponded to known links of the yeast protein network, and which were novel connections.

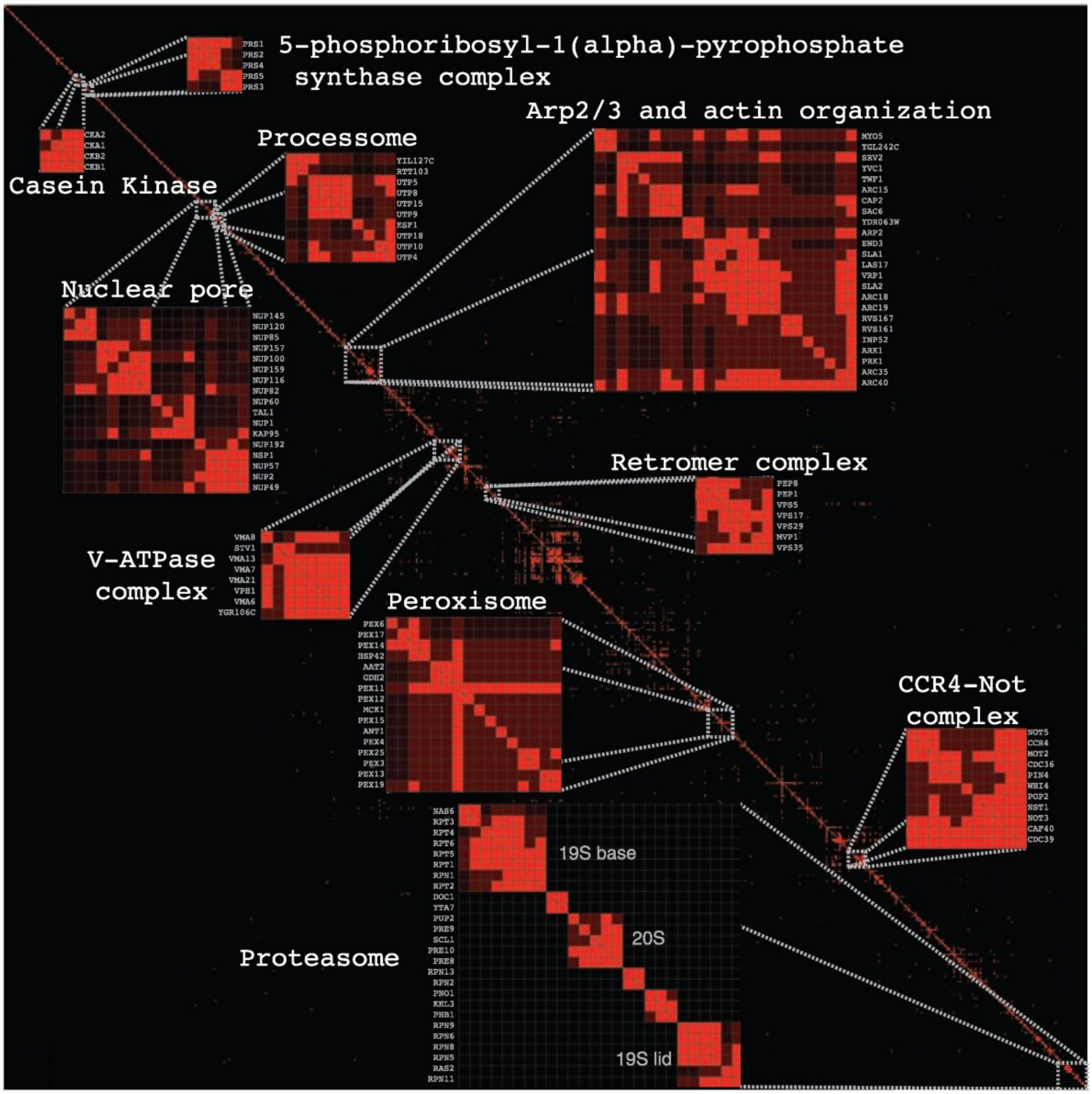


Figure 2-8. The DHFR PCA network is modular and interconnected.
Clustering of the DHFR PCA network reveals numerous known complexes, within which the substructure represents known subunits. Proteins that have interaction patterns similar to those of other proteins and that interact together are grouped along the diagonal. DOI:10.1126/science.1153878.

2.5. Discussion

In the following sections, we discuss the high-quality filtered DHFR PCA network consisting of 2770 unique interactions among 1124 proteins. We focus on comparison of the DHFR PCA interactions with previous publications and describe the novel interactome subspace covered by the method.

2.5.1. Overlap with previous studies

Evaluation of the overlap between previous large-scale protein-protein interaction studies demonstrated that only a limited number of interactions were in common between different datasets [47,175]. Such low overlap may be explained by differences between experimental technologies, biases against interactions between specific types of proteins and high error rates. While all of these reasons contribute to the discrepancies between the dataset, we noticed that the overlap and coverage are routinely calculated in a sub-optimal way leading to pessimistic results. Numbers reported previously are based on the assumption that screens reached genome-wide coverage of the interactome. However, none of the large-scale protein-protein interaction screens, including ours, probed successfully for all possible interactions, even if this was the aim at the start of the study. In our screen, we attempted to tag all undubious ORFs in yeast with DHFR fragments; however, we confirmed by PCR 75% of recombinants in MATa and 83% of MATα strains. Because none of the screens tested for all possible interactions, we decided to take into account the actual number of common interactions that could have been detected in different studies.

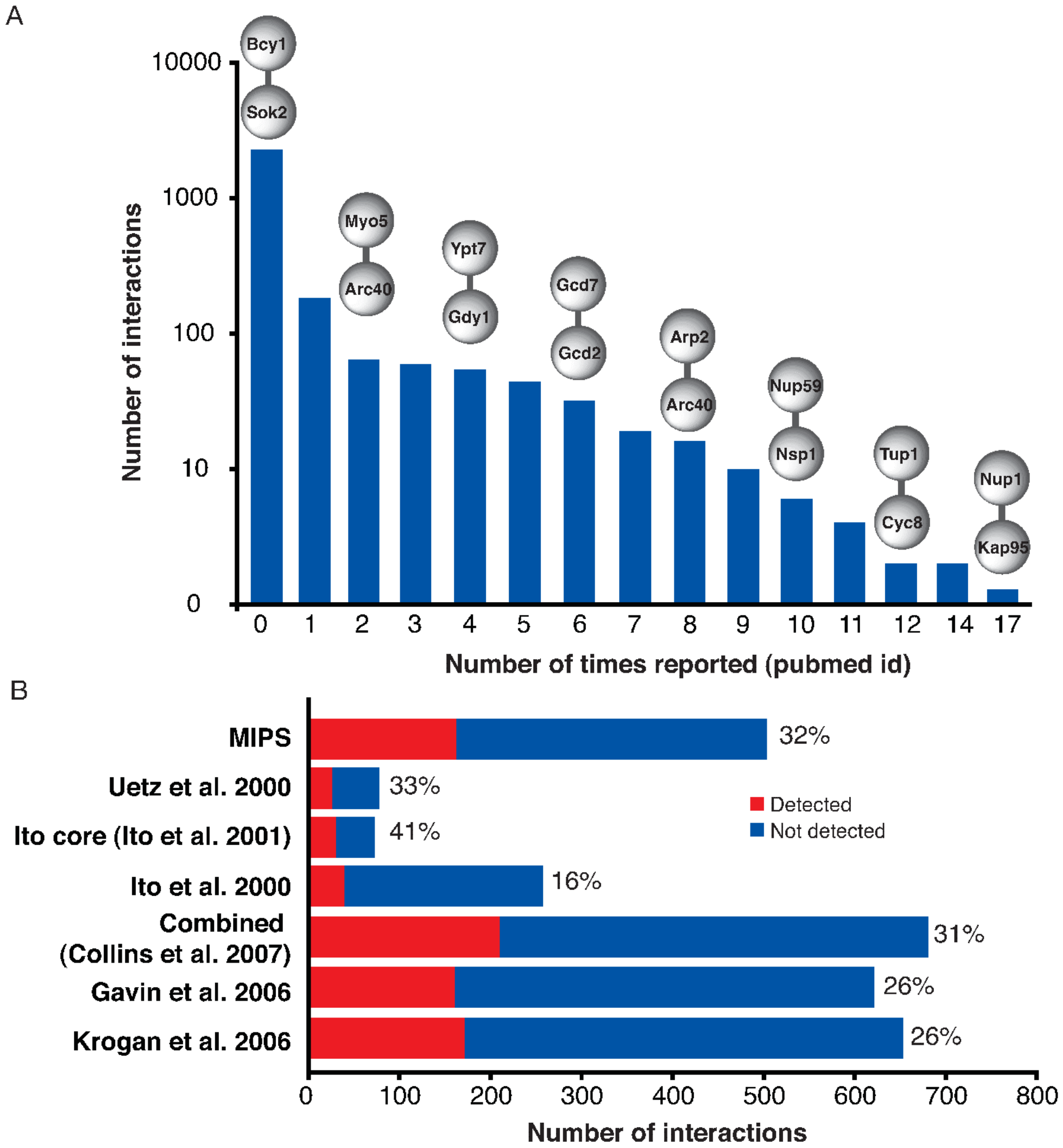


Figure 2-9. Overlap of the DHFR PCA network with other large-scale experiments. (A) Most DHFR PCA protein-protein interactions are new, since they score 0 within the distribution of the number of times a known interaction has been independently deposited in major protein-protein interaction repositories. Examples of interactions are shown above the bars. (B) The overlap of the DHFR PCA network is substantially increased when only the interactions that could be discovered are considered, i.e. only identified successful baits and preys are considered. Bars indicate the number of protein-protein interactions that could have been discovered by PCA. In red is the number of interactions that were discovered. Percentages indicate the percentage of interactions that were discovered by PCA out of the total possible. DOI:10.1126/science.1153878.

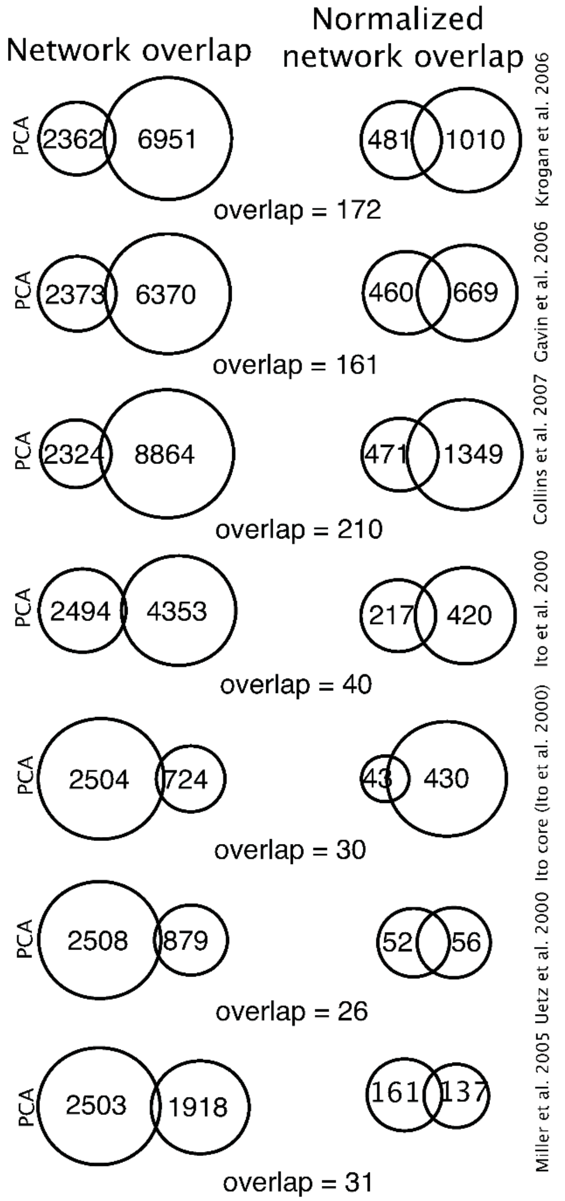
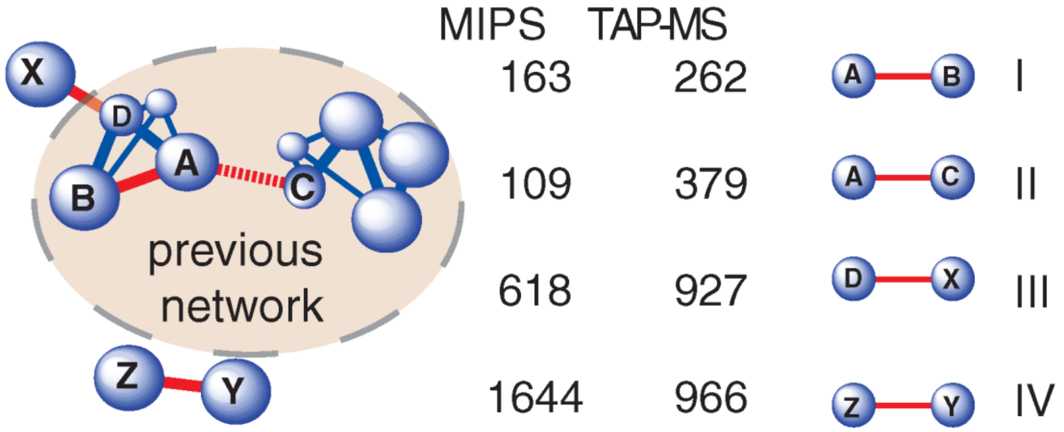


Figure 2-10. Comparison of normalized and non-normalized calculations of the overlap between datasets.

The overlap is calculated for the DHFR PCA network with other large-scale experiments. On the left is the overlap between the different networks. On the right are the same overlaps, but only for those interactions that could have been detected in both experiments; i.e. cases in which the interactions were tested for in both experiments. DOI:10.1126/science.1153878.

We note here, that while preparing our work for publication, another group independently presented a similar approach [61]. Analysis of the overlap normalized by coverage demonstrated that DHFR PCA confirmed between 16% and 41% of interactions detected by earlier screens (Figure 2-9B, Figure 2-10). These numbers are considerably higher comparing to 1%-8% of overlap calculated without this normalization. Higher overlap of DHFR PCA with Y2H studies and lower overlap with TAP-MS methods may be explained by the fact that PCA and Y2H are detecting pairwise interactions, while TAP-MS based techniques detect complexes of proteins and assign an interaction for pairs of proteins discovered in the same complexes that don't necessary interact directly. This leads to a higher number of interactions between same complex members that can be detected by TAP-MS comparing to PCA and Y2H. Pairwise interactions detected by DHFR PCA provide complementary information to complex membership about connections within and between complexes. Our analysis of DHFR PCA protein-protein interactions mapped onto protein complexes divided interactions into the following types: interactions within complexes, interactions between complexes, or interactions with one or both proteins not assigned to any complex (Figure 2-11). As a protein complex scaffold we selected a gold-standard set of CYGD [154] complexes and an assembly of complexes derived by clustering of consolidated TAP-MS studies performed prior to our screen [110]. About 10% of interactions that we detected were between members of the same complex and about 15% of interactions were connecting distinct complexes. Such intra-complex connections may contribute to coordination between biological processes mediated by distinct protein complexes.



Type I. Binary, sterically accesible interaction previously inferred to be in a complex (number of proteins: MIPS = 145; Pu et al. 2007 = 270)

Type II. Binary, sterically accesible interaction not inferred to be within a complex (MIPS = 97; Pu et al. 2007 = 276)

Type III. Binary, sterically accesible interaction between one protein being in a complex and the other not in the dataset (MIPS = 173 & 317; Pu et al. 2007 = 279 & 350)

Type IV. Binary, sterically accesible interaction between two proteins not in the datasets (MIPS = 727; Pu et al. 2007 = 445)

Figure 2-11. PCA protein-protein interactions versus protein complexes.
Comparison of the PCA network with databases of curated protein complexes (MIPS) and inferred from computational analysis of TAP-MS (15) allows classification of four types of PCA interactions: in which both proteins are found within a complex (type 1), are inferred to be in two separate complexes (type 2), one protein is in a complex and the other is not in the network (type 3), or both are absent from the network (type 4). Columns of numbers indicate the number of PCA protein-protein interactions observed for each data set and each category.
DOI:10.1126/science.1153878.

In order to verify that these connections are not due to noise in the data, we calculated semantic similarity scores of GO annotations related to Cellular Compartments (CC), Biological Processes (BP) and Molecular Functions (MF). For all categories the average semantic similarity scores calculated for pairs of interacting proteins were significantly higher than what is expected by chance (CC:3.44 versus 1.64, $P < 10^{-100}$; BP: 3.48 versus 1.51, $P < 10^{-80}$; MF: 3.53 versus 2.3, $P < 10^{-10}$). This analysis demonstrated that DHFR PCA protein-protein interactions tend to connect functionally related proteins. Importantly, semantic similarity scores calculated for TAP-MS complexes were higher than DHFR PCA protein-protein interactions scores. This again can be explained by the fact that TAP-MS methods detected complexes of proteins that collectively are closely related to specific cellular tasks. Assignment of pair-wise interactions between all co-complexed proteins could further increase the score by increasing number of pairs of proteins sharing a similar annotation. Lower semantic similarity scores of DHFR PCA protein-protein interactions shows that our technique captures remote interactions possibly responsible for coordination of a cross-talk between cellular machines. For example, we see a connection between two distinct complexes the CCR4 and the RNA-splicing complex both involved in the RNA metabolic process mediated by an interaction between Dhh1p and Lsm4p. Similar example is the connection between subunits of the serine-threonine phosphoprotein phosphatase and SNF1 complex, both involved in the regulation of metabolic carbohydrate processes mediated by an interaction between Reg1p and Snf1p (Table S7[*]).

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

2.5.2. Novel interaction

Most of the interactions from DHFR PCA network involved one or both proteins not assigned to any complex, thus covering interacting space not explored previously. The results were confirmed by contrasting DHFR PCA results against all available protein-protein interaction data from public databases, not limited to TAP-MS protein complexes, and including results of small-scale experiments (Figure 2.9A). This analysis confirmed that the majority of interactions (~80%) that we detected were novel. This is not surprising because our interactions were detected *in vivo* by a technique with different properties than Y2H and TAP-MS and executed in a different medium.

About 300 of the novel interactions involved proteins with an uncharacterized functional role. Such connections provide an opportunity to propose a potential function for uncharacterized proteins based on functional roles of their interacting partners. Function prediction based on this type of interaction will be the focus of the next chapter (Chapter 3) of this thesis.

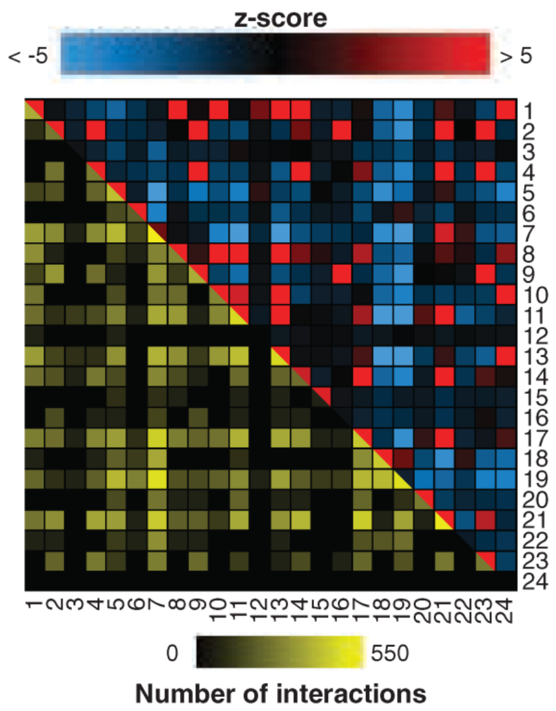
2.5.3. Enrichment of interactions with membrane proteins

In the previous section, we demonstrated that the majority of novel interactions were among proteins that are not assigned to any protein complex based on analysis of previous protein-protein interaction studies. Therefore, we suspect that DHFR PCA technique has the advantage of detecting interactions between certain types of proteins that were underrepresented in results of previous studies. To find such overrepresented proteins, we performed enrichment analysis of GO annotations for cellular compartments. We have observed that our network is highly enriched for proteins associated with membranes (corresponding GO terms: organelle membranes ($P < 10^{-12}$), endomembrane system ($P < 10^{-12}$), membrane part ($P < 10^{-9}$)) and other compartments such as proteasome regulatory particles ($P < 10^{-8}$), the nucleolus ($P < 10^{-7}$), and the cell cortex ($P < 10^{-7}$)] (Figure 2-12, Table S5[*]). In contrast, results of Y2H and TAP-MS are biased against membrane proteins [63,174]. Y2H assays could be depleted of membrane protein interactions because they require that the two proteins get into the nucleus. The aqueous nuclear environment may lead to aggregation or misfolding of the membrane proteins. Furthermore, directing a membrane protein into the nucleus could be a problem. In TAP-MS, purification steps and the requirement to isolate a protein from lipid bilayer can cause bias. We showed, for example, that in the dataset of Collins et al. proteins associated with membrane related GO terms are significantly underrepresented, i.e. they appear in the dataset less frequently than would be expected by chance (Table S6[*]). The fact that in multiple organisms about one third of genome code for proteins associated with membranes [176] illustrates the diversity and importance of cellular functions that these proteins carry out. Therefore, it is of great importance to develop methodologies capable of capturing interactions with membrane proteins and the DHFR PCA network is a significant step in that direction. Prior to our screen, two studies were published dedicated to large-scale identification of membrane protein-protein interactions: 1) Miller et al.

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

[63] using the split-ubiquitin assay, and 2) Xia et al. [174], using computational prediction. A few years after publication of our results, another large-scale experimental study of membrane interactions was published [67], but it was not considered in our analysis described below because these results were not available at the time of calculations. Because of the bias against membrane interactions, the gold-standard CYGD catalogue of protein complexes used for evaluating the quality of our network has a limited coverage of membrane interactions (only 1% of the total number of interactions). Therefore, we performed additional analyses to evaluate the quality of the membrane protein-protein interactions. The main reason for false-positive interactions could be the stabilizing effect of the membrane on protein localization causing recombination of DHFR fragments attached to a pair of proteins in the vicinity that don’t physically interact. Such bias, could be reflected in the higher number of connections (network degree) demonstrated by membrane proteins comparing to the rest of the proteome.



(1) Golgi apparatus; (2) cell cortex; (3) cell wall; (4) cellular bud; (5)unknown; (6) chromosome; (7) cytoplasm; (8) cytoplasmic membrane-bound vesicle; (9) cytoskeleton; (10) endomembrane system; (11) endoplasmic reticulum; (12) extracellular region; (13) membrane; (14) membrane fraction; (15) microtubule organizing center; (16) mitochondrial envelope; (17) mitochondrion; (18) nucleolus; (19) nucleus; (20) peroxisome; (21) plasma membrane; (22) ribosome; (23) site of polarized growth; (24) vacuole;

Figure 2-12. Interactions are enriched within Cellular Compartments Gene Ontology categories.

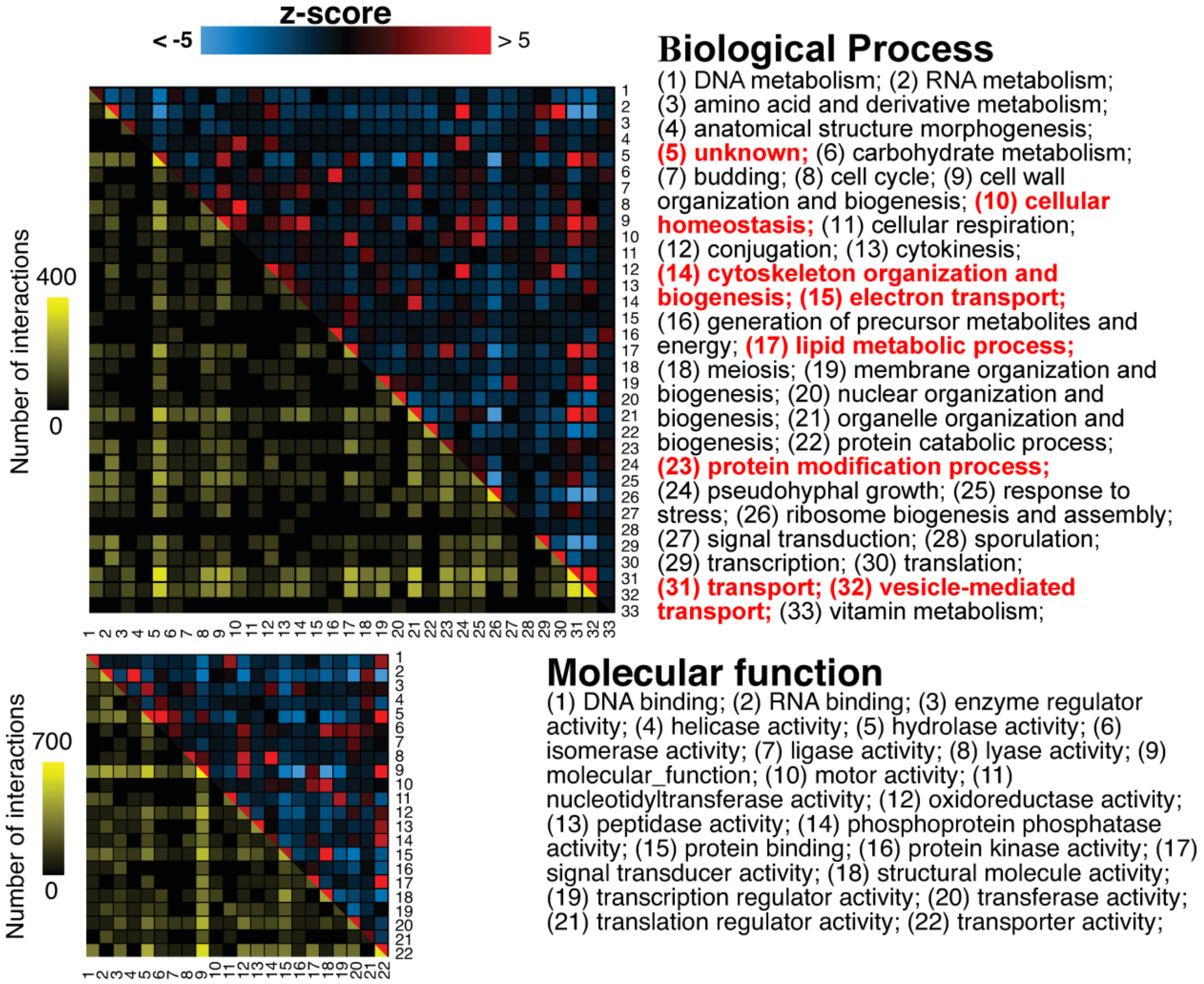
The DHFR PCA network covers several classes of protein function, location, and biological process. The colors above the diagonal represent positive and negative deviations from the expected number of interactions between two cell compartments. A positive z score indicates a larger number of interactions within or between two categories as compared with a random network. A negative z score indicates a smaller number of interactions than expected. A z score of 2 or –2 corresponds to a P value of 0.05, and a z score of 5 or –5 to a P value of 5 × 10−7. Values below–5 and above 5 were given these minimal and maximal values. Entries below the diagonal indicate the observed numbers of interactions on a log10 scale.
DOI:10.1126/science.1153878.

Consequently, we would expect a poor overlap of such unspecific interactions with previous studies of membrane protein-protein interactions. Our analysis of the DHFR PCA network demonstrated that this was not the case. The degree of membrane proteins was only slightly higher than the average network degree (2.8 vs 2.5). The overlap with previous studies was significantly higher than what would be expected by chance. From 662 interactions involving 232 putative membrane proteins in our final network, 51 were predicted by *in silico* analysis [174] versus 5 expected by chance (10 000 randomizations, $P < 10^{-94}$) and 27 interactions were confirmed by split-ubiquitin assays [63] versus 1.9 expected by chance (10000 randomizations, $P < 10^{-75}$). Finally, the semantic similarity scores for Molecular Function and Biological Process calculated for pairs of interacting membrane proteins were significantly higher than in a random network (MF: 3.89, MF random: 2.84, $P <10^{-18}$; BP: 2.86 , BP random: 2.15, $P < 10^{-15}$).

Taken together, these results demonstrated that membrane protein-protein interactions are likely to be as specific as the rest of the interactome and provide valuable information on cellular processes associated with membranes.

2.5.4. Enrichment of interactions between compartments and processes

We further investigated distribution of GO terms associated with compartments, biological processes and molecular functions in DFHR PCA interactions (Figure 2-12, Figure 2-13). Even though we observed a strong compartmentalization of interactions (preferential interaction within the group) between certain categories;



Biological Process

(1) DNA metabolism; (2) RNA metabolism; (3) amino acid and derivative metabolism; (4) anatomical structure morphogenesis; **(5) unknown;** (6) carbohydrate metabolism; (7) budding; (8) cell cycle; (9) cell wall organization and biogenesis; **(10) cellular homeostasis;** (11) cellular respiration; (12) conjugation; (13) cytokinesis; **(14) cytoskeleton organization and biogenesis; (15) electron transport;** (16) generation of precursor metabolites and energy; **(17) lipid metabolic process;** (18) meiosis; (19) membrane organization and biogenesis; (20) nuclear organization and biogenesis; (21) organelle organization and biogenesis; (22) protein catabolic process; **(23) protein modification process;** (24) pseudohyphal growth; (25) response to stress; (26) ribosome biogenesis and assembly; (27) signal transduction; (28) sporulation; (29) transcription; (30) translation; **(31) transport; (32) vesicle-mediated transport;** (33) vitamin metabolism;

Molecular function

(1) DNA binding; (2) RNA binding; (3) enzyme regulator activity; (4) helicase activity; (5) hydrolase activity; (6) isomerase activity; (7) ligase activity; (8) lyase activity; (9) molecular_function; (10) motor activity; (11) nucleotidyltransferase activity; (12) oxidoreductase activity; (13) peptidase activity; (14) phosphoprotein phosphatase activity; (15) protein binding; (16) protein kinase activity; (17) signal transducer activity; (18) structural molecule activity; (19) transcription regulator activity; (20) transferase activity; (21) translation regulator activity; (22) transporter activity;

Figure 2-13. Interactions are enriched within Biological Process and Molecular Function Gene Ontology categories.

The DHFR PCA network covers several classes of protein function, location and biological process. The colors above the diagonal represent positive and negative deviations from the expected number of interaction between two categories, Biological Process or Molecular function. A positive z-score indicates a larger number of interactions within or between two categories compared to a random network. A negative z-score indicates a smaller number of interactions than expected. A z-score of 2 or -2 corresponds to a P-value of 0.05 and a z-score of 5 or -5 to a P-value of 5×10-7. Values below -5 and above 5 were given these minimal and maximal values. z-scores were calculated by generating 10,000 random networks. Entries below the diagonal indicate the observed number of interactions on a log10 scale. DOI:10.1126/science.1153878.

for example, for nuclear and nucleolar proteins, most of DHFR PCA interactions connected distinct categories of proteins (BP: 64%, CC:56%, and MF:63%). In contrast, for TAP-MS these numbers were lower [58, 46, and 57% in the study of Krogan et al. [65] and 51, 49, and 58% in the study of Gavin et al. [64]. These results are consistent with the observation described above, that interactions from TAP-MS studies have higher semantic similarity scores. Many cross-compartment connections that we observe originate from protein categories that are better covered by DHFR PCA and reflect the natural exchange of proteins between the endoplasmic reticulum, Golgi, mitochondrial envelope, and vacuolar proteins or coordination during cell division through interactions between the bud and bud neck with the cell cortex, cytoskeleton, plasma membrane, and sites of polarized growth.

2.5.5. Enrichment of lipid related protein interactions

Similarly, we observed interactions between various processes associated with cellular compartments underrepresented in TAP-MS data. In particular, we have observed enrichment in DHFR PCA network with protein interactions involved in lipid metabolism that are frequently membrane bound [34] (about 64% of lipid metabolic proteins are annotated as membrane proteins based on GO annotation) and are consequently underrepresented in TAP-MS data (Table S6[*]). The network derived by Collins et al. [51] contained only 14 proteins (with 29 interactions) annotated with term “lipid metabolism” while 65 proteins with such annotation were expected to be present in a random network of the same size ($P < 10^{-14}$). In our network

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

we detect interactions with 85 such proteins while 45 were expected by chance ($P < 10^{-11}$). DHFR PCA interactions link lipid metabolic process with transport, protein modification processes and cellular homeostasis (Figure 2-13). Intriguingly, DHFR PCA network is enriched for interactions between proteins involved in lipid metabolism and proteins with an unknown biological function. An example of such interaction network is shown in Figure 2-14. Two proteins with unknown function Pst2p and Rfs1p were found in a middle of a large cluster of lipid related proteins. For the majority of the proteins from that cluster there is an evidence for plasma membrane localization. Eight proteins are transporters, five of which are known to transport lipids (Hnm1p, Itr1p, Osh6p, Osh7p, Vps4p). Three proteins (Tcb1p, Tcb2p, Tcb3p) regulate phosphatidylinositol-4-phosphate, which is required for recruiting effector proteins to specific membrane sites including OSH family of lipid transporters [177]. Both proteins with the unknown function Pst2p and Rfs1p belong to the same family of flavodoxin-like folded proteins. In DHFR PCA network these proteins demonstrated very similar interaction pattern and were placed, by unsupervised hierarchical clustering, next to each other. Thus, it is unlikely that cluster membership of these proteins is a random occurrence. Flavodoxins are electron transport proteins with the ability to bind to flavin mononucleotide cofactor that determines their redox activity. Although, flavodoxins are bacterial proteins [178], flavodoxins-like folded proteins have been characterized in higher eukaryotes [179]. Their function in yeast is linked to the regulation of gene expression of metabolic pathways and stress response [180]. Pst2p and Rfs1p together with another protein from the cluster Slm1p were found in lipid rafts [181], membrane microdomains enriched for signaling lipids such as sphingolipids and phosphoinositides [182] involved in trafficking, signaling, and regulation of the biosynthetic and the endocytic pathways [183]. Intriguingly, phosphatidylinositides have been implicated in similar functions [184]. The cluster from Figure 2-14 contains phosphatidylinositol-4-phosphate regulators (Tcb1-3). Rfs1p has been shown to interact with phosphatidylinositol 3,5-bisphosphate involved in signaling and trafficking [185]. Recent studies suggest that OSH proteins found in the cluster also contribute to regulation of phosphatidylinositides as well as sphingolipid pathways [186]. Observed interconnectivity of the cluster members through association with lipid rafts, lipid-protein and protein-protein interactions may reflect a complex architecture of a novel regulatory mechanism. Understanding the role of the cluster components in this mechanism requires extensive biochemical studies. However, our protein-protein interaction data contributed new potential components of the system that have to be taken into account.

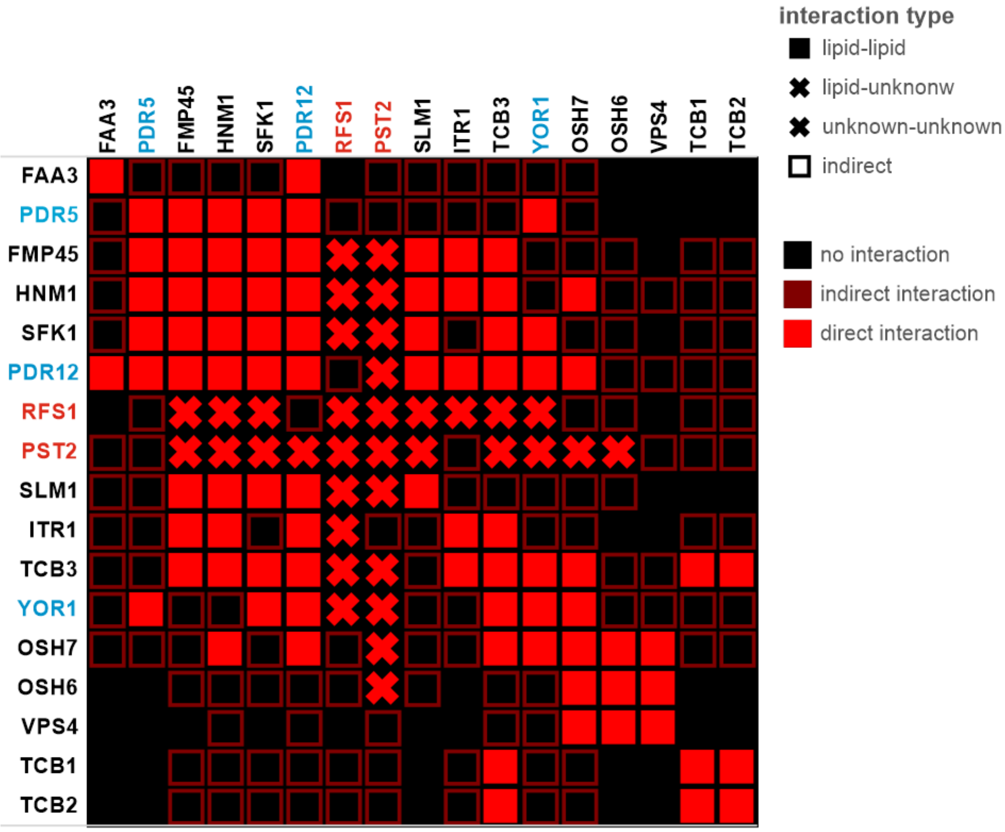


Figure 2-14. Network cluster linking proteins with unknown function with lipid related proteins. Heat map representation of connections between proteins with known functions related to lipid homeostasis and uncharacterized proteins Rfs1p and Pst2p.

2.6. Conclusions

We conducted the first large-scale screen for protein-protein interactions based on PCA strategy. To perform the screen, we developed robotic workflow that enabled to pin colonies with unprecedented density. A novel image analysis routine was developed for converting pictures of plates of the yeast colonies into digital format suitable for computation analysis of protein-protein interactions. Using statistical methods for discriminating between signal and noise, we selected a high-confidence set of interactions with quality comparable to previous large and small-scale datasets.

DHFR PCA screen was performed by different technique and in different conditions than previous protein-protein interaction studies. 2770 unique interactions were identified, the majority of which were novel. Furthermore, additional interactions were reported as supplementary data available for further evaluation. We have demonstrated that the DHFR PCA screen detects interactions between proteins with more distinct functional roles comparing to TAP-MS approaches. We observed that DHFR PCA network is particularly enriched for interactions involving membrane proteins that were depleted in previous systematic studies. Extended coverage of the new interaction space allows to gain new insights into such important processes as lipid metabolism and homeostasis. Links that we discovered with uncharacterized proteins provide data for discovering novel gene functions based on the network analysis.

Taken together, novel reported interactions contribute to the knowledge about the yeast network connectivity that combined with other available data help formulating novel biological hypotheses about cellular functions and network properties.

2.7. Materials and Methods

2.7.1. Data acquisition and image analysis

Plate images of diploid MATa/MATα array plates grown on methotrexate for 96 hours were taken with a 4.0 Mega pixel camera (Powershot A520, Canon). Plate images were saved in JPG format at a resolution of 180 dpi and a size of 2,272 x 1,704 pixels. Next, all of the 3,301 plates (3,247 plates for the 3,247 different baits, 48 repeated plates and 6 plates for the control experiments) were manually inspected during the image analysis step and positions too close to the plastic edges of the plate and therefore uninterpretable were eliminated by setting the intensities of the first and the last colonies on the first row of the array to 0. At this stage we eliminated 44 plates from the final analysis because they displayed growth of colonies at empty positions, likely resulting from grid misalignment or contamination.

Images were processed with an algorithm implemented in the Image Processing Toolbox for MATLAB (The MathWorks, Natick, Massachusetts) and consisted of the following steps:

1) Images were corrected for non-uniform illumination as described in (http://www.mathworks.se/help/images/examples/correcting-nonuniform-illumination.html)

2) Small objects that correspond to gel background, bubbles, plate edges or other anomalies were removed using the imopen function with the disk morphological structuring excluding elements of a radius smaller than 2 pixels (radius of 4 pixels was used on plate edges).

3) Images were converted into binary format using the im2bw function with a threshold calculated by the graythresh function. In this format, pixels that correspond to colonies were set

to 1 and background pixels to 0. Thus, connected components of binary images with pixel values equal to 1 (calculated with bwlabel and regionprops functions) corresponded to colonies.

4) Watershed transform was performed with watershed function. As a result of the watershed transform, pixels that lie on the border between objects were identified. These border pixels were set to zero on the original binary image, thus separating overlapping colonies and step 3 was repeated in order to analyze separated colonies.

The additional steps described below were performed if: the size of a connected region identified in step 3 was larger than expected for a single colony (on average 400 pixels); a connected region could not be matched to a position on the array due to fusion of several colonies.

5) Extraction of a rectangular subpart of an original image that fully contains a connected region that is suspected to contain fused colonies,

6) Superimposition of distance transform (bwdist function) with local maximum of Circular Hough transform (circle_hough function available at MATLAB File Exchange) for further improving the detection of circular colonies.

7) Repeat watershed transform.

8) In cases where a number of objects that were separated using steps 4 to 5 was different from the number of possible centers detected by the Circular Hough transform, we performed a watershed transform on the original grey scale image.

2.7.2. Statistical analyses

Calculations were performed in MATLAB and R software packages. Custom Java codes were used to query out protein-protein interaction data and data from public repositories and for generation of random networks.

Positive and negative datasets for optimizing DHFR PCA thresholds and benchmarking the network quality were obtained from the following sources:

A gold-standard positive reference set - CYGD catalogue of protein complexes extracted from MIPS database on 18th of May, 2006 (ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/complexcat_data_18052006).

A gold-standard negative reference set - A set of negative reference protein-protein interactions obtained from Collins et al. [51]. It contains proteins that belong to distinct protein complexes from separated cellular compartments or proteins that show anticorrelated expression patterns, which suggests that they are not present in the cell at the same time.

2.7.3. Network randomizations

To evaluate statistical significance of properties of the DHFR PCA network, we compared them to a network with randomly assigned interactions. The random network had a similar structure as the DHFR PCA network. It contained the same proteins and the same number of interactions per protein as the DHFR PCA network. To further match the random network to the structure of the experimental network it contained only interactions that could be potentially detected in the DHFR PCA screen. The statistical significance of differences between experimental network parameters and the random counterparts was calculated based on z-scores that compared these values converted to p-values.

2.7.4. Analysis of protein abundance

Data for the analysis of protein abundances in yeast were obtained from a high-throughput flow cytometry study of a library of GFP-tagged yeast strains [21]. We selected protein abundance measures collected from yeast grown in the same media as the DHFR PCA screen (synthetic complete medium).

Non-parametric wilcoxon rank-sum test was used for evaluating statistical significance of the differences between protein abundances of compared groups of proteins.

2.7.5. Analysis of Gene Ontology enrichment

GO enrichment analysis was used to evaluate whether a number of proteins of a particular type in a set of proteins was higher or lower than what is expected by random chance. Proteins were assigned to a particular type based on GO terms related to molecular function, biological process and cellular compartment.

2.7.5.1. GO enrichment of sets of proteins

To find over or under-represented types of proteins in selected sets of, GO enrichment was calculated using GOstat R library [187]. We selected conditional hypergeometric algorithm for the calculations. The algorithm takes into account hierarchical organization of GO terms structure and computes the significance of a GO term based on its neighborhood. The method is reported to perform better comparing to a simple hypergeometric test because it limits the redundancies in the results.

2.7.5.2. GO enrichment of interactions between pairs of terms

For GO enrichment analysis of the protein interaction, the following procedure was applied. The terms were extracted from GO slim map downloaded from SGD (February 17th. 2007). Enrichment was calculated for every pair of annotation terms from the biological process, cellular compartment and molecular function categories. We calculated a number of interactions that are detected between proteins associated with a specific pair of GO terms. This number was compared to the corresponding number calculated for a random network. Randomization was repeated 10 000 times. The number of interactions between proteins associated with a specific pair of terms was calculated for each randomization, and a z-score was derived by comparison of the number of interactions detected by DHFR PCA screen with 10 000 numbers calculated from random networks. The z-score for a pair of terms *i* and *j* is calculates as:

$$\text{z-score}_{i.j} = (\text{N}_{i,j} - \mu(\text{Nrand})_{i.j}))/\sigma(\text{Nrand})_{i,j},$$

where $\text{N}_{i,j}$ is the number of interactions between a pair of terms *i* and *j* in the DHFR PCA network, $\mu(\text{Nrand})_{i.j}$ is the average number of interactions between the terms calculated for the random networks and $\sigma(\text{Nrand})_{i,j}$ is the standard deviation of the interactions values for the random networks. High z-score values correspond to pairs of terms enriched for interactions comparing to random chance. Low negative z-score values correspond to pairs of terms depleted of interactions.

2.7.5.3. Semantic similarity scores

For evaluating functional relationship between pairs of proteins that do not share the same annotation, we calculated GO semantic similarity score [188]. The score searches the annotation hierarchy for parent (more general) terms that are common for a pair of proteins. If the parent is close in the hierarchy to the more specific terms characterizing the proteins, it indicates closer functional relationship, and the score is higher. To obtain the statistical significance of the results, we compared DHFR PCA semantic similarity scores with the corresponding scores calculated for random networks generated as described above.

2.7.6. Datasets compared with DHFP PCA results

We analyzed how many times each interaction from the DHFR network had been previously reported in Figure 2.9. For the calculations, we extracted physical interactions from the following databases: BIOGRID (www.thebiogrid.org/, version 2.0.29), mips-MPact (http://mips.gsf.de/genre/proj/mpact, version 18052006) and DIP (http://dip.doembi.ucla.edu/, June 2007). We separately considered the combined TAP-MS data from [51] (interaction set with Purification Enrichment score equal or above 3.19 as defined in [51]), as it overlaps considerably with what has been deposited in Biogrid by [64] and [65]. We considered only one citation for an interaction reported in Collins et al. [51], even if it had been reported in one or both original studies. We excluded interactions that were not associated with any publications by the PUBMED IDs. Each PUBMED ID we treated as an independent evidence for an interaction.

2.7.7. Overlap with previous large-scale studies

Reference datasets and criteria for normalization for the overlap calculations presented in figures 2-9 and 2-10 are described below:

1) CYGD catalogue of manually annotated complexes [154]. Downloaded from ftp://ftpmips.gsf.de/yeast/catalogues/complexcat/complexcat_data_18052006. Complexes detected by large-scale experiments were filtered out from this file and interactions were assigned between all proteins that belong to the same complex. An interaction is considered to be possible if both interacting partners are present in the CYGD catalogue.

2) Krogan et al. [65]. The core dataset was obtained from supplementary table 7, that lists successfully identified baits and preys obtained from supplementary tables 2 and 3. An interaction is considered to be possible only if one of the proteins is present in the baits list and another is in the preys list. We don't consider a co-occurrence of both proteins in the prey list since the core dataset of this study contained only bait-prey pairs.

3) Gavin et al. [64]. The network of interactions was obtained as deposited in Biogrid. This study used a statistical framework for deriving a high confidence set of interactions that makes possible interactions between two prey proteins. Therefore for normalization, we considered an interaction to be possible if for a pair of interaction partners, a bait-prey or prey-prey pair exists in the raw purification data (downloaded from http://yeast-complexes.embl.de).

4) Collins et al. [51]. We used high confidence data with a PE score cutoff of 3.19. Normalization was performed as described above for Gavin et al. using a combination of both Krogan and Gavin raw datasets.

5), Ito et al., Uetz et al. [57,58,189]. Interactions detected by yeast two-hybrid assays. Interaction is considered possible if one of the interacting partners is among proteins that showed an interaction when tagged with a binding domain and another is among proteins that showed interactions when tagged with an activation domain.

6), Miller et al. [63]. Interactions tested using the split-ubiquitin reporter. The data were extracted from supplementary Table 1. An interaction is considered possible if a corresponding pair of Cub-PLV and NubG ORF is present in the dataset.

2.7.8. Clustering of high confidence interactions

Hierarchical average linkage agglomerative clustering was performed as described in [104]. Clustering was based on the association matrix that takes into account indirect interactions between the proteins mediated by a common partner. Association values the two proteins were calculated as 1/d2, where d is the shortest path in the network between these proteins. Thus, association values ranged from 0 (non-interaction) and 1 (direct interaction). Self associations were marked as 1 for the clustering purposes regardless of whether a heteromeric interaction was observed by DHFR PCA. Clustering was performed with Cluster 3.0 software C++ libraries. (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Data were visualized with iVici software [190] described in Chapter 4. For the clarity of the visualization of the large clustered DHFR PCA network, only direct interactions are shown on a complete map (Figure 2-8). On the insets, direct interactions are bright red, while indirect interactions (2 or 3 links between two protein) are shown as two consecutively darker shades of red, respectively. Results of the clustering are available as a Supplementary File S1[*] from the original publication.

[*] http://www.sciencemag.org/content/suppl/2008/05/08/1153878.DC1/1153878s_tables.zip

Chapter 3 : High-content screening of yeast mutant libraries by shotgun lipidomics

3.1. Contribution to the published work

Chapter 3 was published as an article in Molecular Biosystems journal: Tarasov K, Stefanko A, Casanovas A, Surma M, Berzina Z, Hannibal-Bach HK, Ekroos K, Ejsing CS (2014) High-content screening of yeast mutant libraries by shotgun lipidomics. Mol Biosyst. 2014 Mar 31. [Epub ahead of print]

T.K. and E.C.S formulated the idea, designed experiments and wrote the paper. T.K. performed gene function predictions, set up a computational platform for raw data processing, preformed quality controls and analyzed the data. S.A. and H.B.H.K. performed the first round screening. S.M.A. contributed to set up of 96 format cell cultures and extractions. C.A. and H.B.H.K. performed the liquid culture experiments for the growth profile and second round screen. S.A., S.M.A. and E.K. commented on the manuscript.

High-content screening of yeast mutant libraries by shotgun lipidomics

Kirill Tarasov[a,d,*], Adam Stefanko[b,c], Albert Casanovas[b], Michal A. Surma[c], Zane Berzina[b], Hans Kristian Hannibal-Bach[b], Kim Ekroos[a] and Christer S. Ejsing[b,*]

Mol Biosyst. 2014 Mar 31. [Epub ahead of print]

Received 17th December 2013,

Accepted 20th March 2014

DOI: 10.1039/c3mb70599d

[a] Zora Biosciences Oy, Biologinkuja 1, FI-02150 Espoo, Finland

[b] Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense, Denmark

[c] Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany.

[d] Department of Biochemistry and Molecular Medicine, Université de Montréal, H3T 1J4, Montréal, Canada

[*] Corresponding authors: Christer S. Ejsing (lipidomics), Kirill Tarasov (bioinformatics).

Electronic supplementary information (ESI) available: Supplementary Fig. S1. Supplementary Tables S1–4. See DOI: 10.1039/c3mb70599d

3.2. Abstract

To identify proteins with a functional role in lipid metabolism and homeostasis we designed a high-throughput platform for high-content lipidomic screening of yeast mutant libraries. To this end, we combined culturing and lipid extraction in 96-well format, automated direct infusion nanoelectrospray ionization, high-resolution Orbitrap mass spectrometry, and a dedicated data processing framework to support lipid phenotyping across hundreds of Saccharomyces cerevisiae mutants. Our novel approach revealed that absence of genes with unknown function YBR141C and YJR015W, and the transcription factor KAR4 precipitated distinct lipid metabolic phenotypes. These results demonstrate that the high-throughput shotgun lipidomics platform is a valid and complementary proxy for high-content screening of yeast mutant libraries.

3.3. Introduction

The lipidome of eukaryotic cells consists of several hundreds to thousands of molecular lipid species that constitute membranes, store metabolic energy and function as signalling molecules[89,191]. The structural heterogeneity of lipids is defined by a metabolic network of enzymes and regulatory factors that synthesize distinct lipid species by assembling or disassembling a multitude of available hydrocarbon residues and polar head groups. Lipid species can be divided into several categories based on their chemical structures [192]. The most abundant lipid categories in eukaryotic cells include glycerophospholipids, sphingolipids, glycerolipids and sterol lipids, which mediate distinct molecular functions. Notably, several metabolic transitions interlink glycerophospholipid, sphingolipid, glycerolipid and sterol metabolism such that perturbations are prone to induce lipidome-wide ripple effects and prompt compensatory responses to sustain lipid homeostasis [96]. Compromising the lipid metabolic network is known to cause dysfunctional lipid homeostasis and cellular lipotoxicity that precipitate disorders such as obesity, atherosclerosis and neurodegeneration [193]. Importantly, the regulatory mechanisms that govern global lipid metabolism and relay physiological signals to sustain lipid homeostasis are largely unknown.

Genetic and biochemical studies using the yeast *Saccharomyces cerevisiae* have been instrumental to elucidating the blueprint of lipid metabolism and defining the physiological functions of lipids [194,195]. Early efforts have pinned lipid metabolism to the framework of global metabolism as illustrated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway resource [196]. Moreover, genetic and biochemical approaches combined with molecular biology have paved the way to cloning and functional characterization of key enzymes in the human lipid metabolic network [194]. The functional regulation of lipid metabolism depends in part on a transcriptional circuitry that sets the cellular concentration of lipid enzymes and accessory regulatory factors depending on physiological requirements. The circuitry in *S. cerevisiae* includes the transcriptional regulators Opi1p, Ino2p, Ino4p and Zap1p that control the expression level of a subset of proteins required for glycerophospholipid metabolism [197]. In addition, the circuitry also includes the regulators Mga2p and Spt23p, which are involved in controlling the expression of the fatty acid desaturase *OLE1* [198]. Notably, regulatory mechanisms controlling most of enzymes in the lipid metabolic network are still poorly understood. Intriguingly, this raises the question of how the expression levels of enzymes involved in, for example, sphingolipid and sterol lipid metabolism are controlled.

Functional genomics strategies including gene-gene [41,42] and protein-protein interactions assays [51,55,63] can be powerful approaches for identifying regulatory factors in lipid metabolism. Recently, epistatic miniarray profiling has been instrumental for defining the molecular mechanisms of how fatty acid chain length is determined by conserved membrane-imbedded elongase complexes [133], how conserved Orm proteins interact with the serine palmitoyltransferase to control sphingolipid biosynthesis [134] and linking the GDP/GTP exchange factor Rom2p to the regulation of sphingolipid metabolism [135]. These findings were prompted by high-content datasets showing particular genes interacting with known constituents of the lipid metabolic network. Notably, such interactions can also be observed in other types of publically available resource data. For example, the yeast protein interactome displays an overrepresentation of interactions between genes involved in lipid metabolism and genes with other cellular functions [55]. Thus, public repositories of gene and protein interaction data combined with function prediction algorithms can be a potential resource for shortlisting genes/proteins with a putative functional role in lipid metabolism [199]. Importantly, several studies have successfully combined function prediction methods with experimental confirmation to elucidate the molecular mechanisms of mitochondrial biogenesis in yeast [136] and tissue-specific regulation patterns in worm [138]. The success of these studies demonstrates that mining resource data and integrating lipidomic analysis can be an avenue for identifying novel lipid enzyme activities and regulators of global lipid metabolism.

Shotgun lipidomics is a relatively novel “omics” tool that affords comprehensive and quantitative profiling of cellular lipids. The efficacy of the technology has been documented in numerous studies of biological membrane organization, lipid-protein interactions and the regulation of lipid metabolism [134,200–203]. Shotgun lipidomics implies that lipid extracts of cells are directly infused into a mass spectrometer without up-front time-consuming liquid chromatographic separation thereby shortening the time required for analysis, and that identification of lipid species relies on accurately determined masses and/or tandem mass spectra acquired from corresponding lipid species [89,204]. Shotgun lipidomics enables extensive lipidome characterization by combining analyses of the same lipid extract in positive and negative ion mode, and by implementing data processing routines to merge, normalize and visualize lipidomic datasets [205]. In addition, more recent shotgun lipidomics technology based on high-resolution Orbitrap mass spectrometry and automated direct infusion nanoelectrospray ionization offer high-throughput capabilities with high sensitivity, broad dynamic quantification range and extensive lipidome coverage spanning lipid species molar abundances over 3 to 4 orders of magnitude [96,99,206]. Notably, these analytical hallmarks are ideally suited for exploratory lipidome analysis in yeast and provide a mean to screen libraries of mutant strains to identify regulatory modules in global lipid metabolism.

Here we describe a high-throughput platform for high-content lipidomic screening of yeast mutant libraries that utilizes culturing and lipid extraction in 96-well format, automated direct infusion nanoelectrospray ionization, high-mass resolution Orbitrap mass spectrometry and a dedicated data processing framework to support systematic monitoring of lipid species across hundreds of yeast strains. To catalog lipid phenotypes, we made use of ‘robust principal component analysis' and a quantitative scoring system that we term SoamD (sum of absolute mol% difference). As a test bed, we employed the platform to array a shortlist of deletion mutants of distinct transcriptional regulators and genes with unknown function predicted to play a role in lipid metabolism. Our novel approach revealed that absence of genes with previously unknown function YBR141C and YJR015W, and the transcription factor KAR4 precipitates distinct lipid metabolic phenotypes. These results show that combining functional genomics workflows and high-content lipidomic profiling can be a powerful proxy for identifying regulators of global lipid metabolism.

3.4. Materials and methods

3.4.1. Chemicals and lipid standards

Synthetic lipid standards were purchased from Avanti Polar Lipids and Larodan Fine Chemicals. Chemicals, growth media and solvents were purchased from Sigma–Aldrich, Rathburn Chemicals, MP Biomedicals and BD Biosciences.

3.4.2. Yeast strains

In this study we used *S. cerevisiae* reference strain BY4742 (MATα *his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) and the congenic deletion mutants listed in Supplementary table 1[1]. All strains were obtained from EUROSCARF. Mutant strains without genes encoding transcriptional

[1] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d2.txt

regulators were shortlisted based on gene ontology (GO) annotation in the *Saccharomyces* Genome Database (SGD) [153].

3.4.3. Prediction of uncharacterized genes with potential function in lipid metabolism.

A list of genes known to be implicated in lipid metabolism (query list of lipid-related genes) was compiled based on automated extraction of gene names with GO annotation 'lipid metabolism' or 'lipid binding' in the SGD, and thorough manual annotation based on literature (Supplementary table 2[1]). To predict uncharacterized genes with potential function in lipid metabolism we used the query list of lipid-related genes and the GeneMANIA function prediction algorithm [207]. Based on available protein-protein interaction data from BioGRID[155], the GeneMANIA function prediction algorithm was requested to output the 50 most related genes to the genes on the lipid-related query list (Supplementary table 3[2]). From the 50 top scoring genes we selected 8 genes that were annotated with "biological process unknown", and shortlisted the corresponding deletion mutants for first round lipidomic screening.

[1] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d3.txt

[2] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d4.txt

3.4.4. First round screening: 96-well plate culturing

Yeast strains were plated and cultured at 30˚C for 24 hours in 0.3 ml 96-well plates (Eppendorf AG) on an agar-based solid synthetic complete medium containing 2% glucose and supplemented with 100 µM inositol and 100 µM choline. In addition to shortlisted deletion mutants each 96-well plate contained three replicates of the control strains BY4742 and *elo2*Δ. Yeast cells were harvested by resuspension in 100 µl 155 mM ammonium acetate (average yield 1.5-2.0 OD_{600} units), transferred to 2 ml 96-well plates (Eppendorf AG) and stored at -80˚C until lipid extraction.

3.4.5. Lipid extraction in 96-well plates at 4˚C

The yeast cell suspensions (100 µl) in 2 ml 96-well plates were added glass beads (425-600 µm, Sigma–Aldrich) and subjected to cell disruption for 120 min at 1400 rpm and 4°C on a ThermoMixer (Eppendorf AG)(This high-throughput oriented cell lysis procedure was benchmarked against conventional glass bead lysis [96] by showing no differences in lipid profile (data not shown)). Cell lysates (70 µl) were transferred into a new 2 ml polytetrafluoroethylene 96-well plate (Radleys Discovery Technologies) and subjected to single-step lipid extraction in the 96-well plate. Samples were extracted by adding 250 µl chloroform/methanol (2:1, V/V) and mixing on a ThermoMixer for 120 min. The lower organic phase was collected by transferring 47 µl into two separate 150 µl 96-well plates (Eppendorf AG) that were subsequently subjected to vacuum evaporation.

3.4.6. Mass spectrometric lipid analysis and data processing for 96-well plate cultures

Lipid extracts in 96-well plates were dissolved in 20 µl 7.5 mM ammonium acetate in chloroform/methanol/propanol (1:2:4, V/V/V) for positive ion mode mass analysis, and 20 µl 0.0075% methylamine in methanol/chloroform (1:5, V/V) for negative ion mode analysis. The 96-well plates were covered with aluminum sealing tape to avoid sample evaporation. Samples were analyzed by direct infusion on a LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific) equipped with a robotic TriVersa NanoMate ion source (Advion Biosciences) as previously described [96,99]. Positive ion mode analysis was performed using multiplexed FT MS with scan ranges *m/z* 220-530 (monitoring lysophosphatidylcholine (LPC) and lysophosphatidylethanolamine (LPE) species) and *m/z* 500-1200 (monitoring phosphatidylcholine (PC), phosphatidylethanolamine (PE), diacylglycerol (DAG), triacylglycerol (TAG), sterol ester (SE) and ceramide (Cer) species). Negative ion mode analysis was performed using multiplexed FT MS with scan ranges *m/z* 200-605 (monitoring lysophosphatidic acid (LPA), lysophosphatidylserine (LPS) and lysophosphatidylinositol (LPI) species) and *m/z* 505-1400 (monitoring phosphatidylinositol (PI), phosphatidic acid (PA), phosphatidylserine (PS), and inositol-phosphoceramide (IPC) species). The total time of FT MS analysis was 3 min per polarity per sample. All FT MS spectra were acquired in profile mode using a target mass resolution of 100,000, isolation waveforms enabled, automatic gain control at 1e6, max injection time at 250 ms and acquisition of 2 µscans. Lipid species were identified, quantified and visualized using ALEX software [205], SAS software (SAS Institute Inc.) and

Tableau Desktop software (Tableau Software), respectively. Lipid species were annotated according to their sum composition [99]. Lipid species abundance was monitored by intensity profiling using the proxy intensity% (I%) calculated as the intensity of a given lipid species divided by the sum of intensities of all monitored lipid species in a given ion mode (i.e. positive or negative). A quality control procedure having two filters was implemented: i) strains with less than 70% of the average number of detected lipid species in at least one ion mode were rejected, ii) strains with less than 15% of the average total lipid intensity in at least one ion mode were rejected.

3.4.7. Classification of mutant strains into growth phase categories

The lipidomes of mutant strains surviving the quality control procedure were classified according to growth phase. To this end, the BY4742 reference strain was cultured in liquid medium as described below in the section 'Second round screening: Liquid culturing'. Samples were collected for BY4742 cells in exponential phase (0.8-3.2 OD_{600} units/ml) and stationary phase (4-4.5 OD_{600} units/ml) determined using a growth curve. These samples were subjected to two-step lipid extraction as described below, and lipidomic analysis using intensity profiling I% as described in the previous section. To classify mutant strains according to growth phase we performed average linkage agglomerative hierarchical clustering on I% values of lipid species and lipid classes that were the most different between the growth phases (i.e. sum of I% of lipid species belonging to TAG, PC, DAG, PI, PS, PE and PA lipid classes, and the sum of I% of PC, PI, PS, PE, PA and DAG species with carbon index up to 32, and equal or greater than 34). Analysis was performed in Cluster 3.0 software [208]. Results were visualized as a clustering heat map using I% standardized as Z-scores (Figure 3-3A).

3.4.8. Identification of mutants strains with perturbed lipid phenotypes

Deletion mutants with perturbed lipid phenotypes were identified using robust principal component analysis. This analysis was performed separately for mutant strains classified as either in exponential or stationary phase. I% values were used for the analysis. Missing values were substituted with zero. Calculations were performed in R software using PcaHubert function from the rrcov package for robust multivariate analysis [209,210]. Variables were scaled to have unit variance using median absolute deviation function. The number of principal components was selected based on scree-plots and was set to 4 for strains in stationary phase and to 3 for strains in exponential phase. PcaHubert function calculated orthogonal distance scores for each strains and the orthogonal distance cut-off value which was used to define the hit strains with perturbed lipid phenotypes. These strains were subjected to second round screening.

3.4.9. Second round screening: Liquid culturing

Liquid culture experiments of candidate mutant strains with perturbed lipid phenotypes and reference strains were performed at 30˚C with synthetic complete medium containing 2% glucose and supplemented with 100 µM inositol and 100 µM choline. The yeast strains were precultured for 24 hours, diluted to 0.2 OD_{600} units/ml and cultured for another 24 hours until collection of samples in the stationary phase (4.0-4.5 OD_{600} units/ml). Cells were washed in 155 mM ammonium acetate and stored at -80˚C until lipid extraction.

3.4.10. Lipid extraction at 4°C of samples obtained by liquid culturing

Samples from liquid cultures were subjected to two-step lipid extraction as previously described [96]. Briefly, yeast were resuspended in 1 ml 155 mM ammonium acetate and disrupted using glass bead lysis. Aliquots of cell lysates were diluted to 0.4 OD_{600} units in 200 µl and mixed with 17 µl of internal lipid standard mixture containing cholesterol-D7, CE 19:0, TAG 17:1/17:1/17:1, DAG 19:0/19:0, LPA O-16:0, PA 17:0/14:1, LPS 17:1, PS 17:0/20:4, LPE O-16:0, PE O-20:0/O-20:0, LPC O-17:0, PC 18:3/18:3, LPI 17:1, PI 17:0/14:1, PG 17:0/14:1, CL 14:0/14:0/14:0/14:0, Cer 18:1;2/17:0;0, IPC 18:0;2/26:0;0, MIPC 18:0;2/26:0;0 and $M(IP)_2C$ 18:0;2/26:0;0. Samples were extracted with 990 µl chloroform/methanol (15:1 V/V) for 2 hours. The lower organic 15:1-phase was collected and subjected to vacuum evaporation. The remaining aqueous phase was re-extracted with 990 µl chloroform/methanol (2:1 V/V) for 1 hour. The lower organic 2:1-phase was collected and subjected to vacuum evaporation.

3.4.11. Mass spectrometric lipid analysis and data processing for samples obtained by liquid culturing

The 15:1- and 2:1-phase lipid extracts were dissolved in 100 µl chloroform/methanol (1:2, V/V) and analysed by direct infusion on a LTQ Orbitrap XL instrument equipped with the robotic TriVersa NanoMate ion source as previously described [96,99]. Sterols were analyzed after chemical sulfation of the 15:1 phase extract [202]. The molar amount of lipid species were determined using the spiked-in internal standards and converted to mol% as previously described [96,99]. For each strain, two technical replicates of a single lipid extract were analysed.

3.4.12. SoamD calculation

A SoamD (sum of absolute mol% difference) score was calculated for each deletion mutant subjected to second round screening. SoamD was calculated as:

$$\text{SoamD}_i = \sum_{j=1}^{N} \text{abs}(\text{mol\%}_{i,j} - \text{mol\%}_{BY4742,j}),$$

where *N* is the number of lipids, mol%$_{i,j}$ is the mol% value of lipid *j* in deletion mutant *i* and mol%$_{BY4742,j}$ is the mol% value of lipid *j* in the reference strain BY4742. The score is only applicable to experiments performed with spike-in of internal standards.

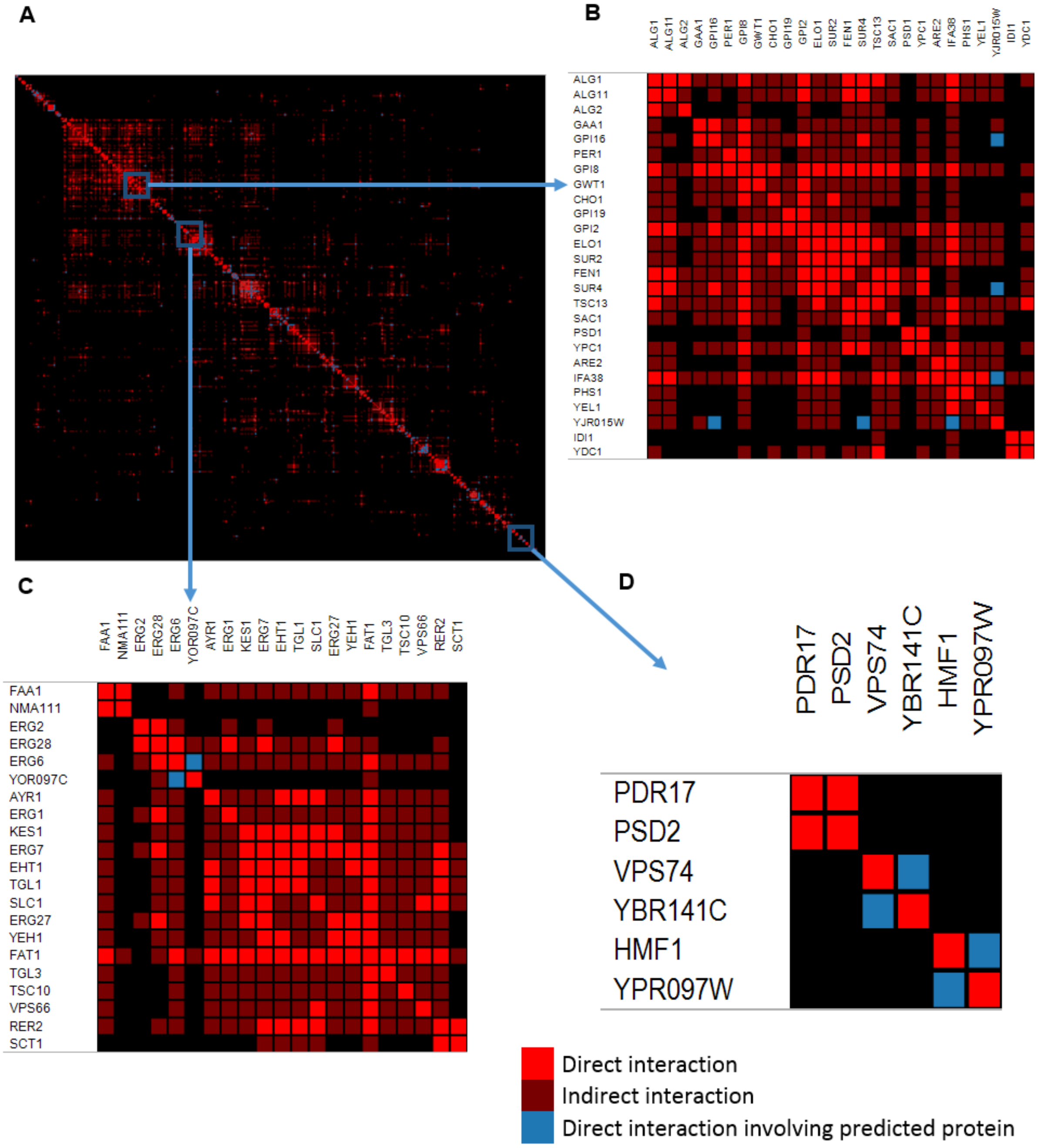


Figure 3-1. Protein-protein interaction network of lipid metabolism and function.
(A) Protein-protein interaction network of proteins known or predicted to be involved in lipid homeostasis. Interaction data was retrieved from BioGRID, predictions are made by the GeneMANIA algorithm. Interactions between known and predicted proteins are arranged using average linkage agglomerative hierarchical clustering. Interactions between known members of lipid metabolism are shown in red; interactions between proteins predicted to be involved in lipid metabolism in blue. Detailed interaction network around predicted proteins Yjr015wp (B), Yor097cp (C) and Ybr141cp (D) are shown. DOI: 10.1039/c3mb70599d.

3.5. Results and discussion

3.5.1. A high-throughput platform for lipid phenotyping

To establish a resource for identifying proteins with potential function in lipid homeostasis we developed a high-throughput lipidomics platform for quantitative high-content screening of yeast mutant libraries. First, we compiled a deletion library in 96-well format covering 178 strains divided into three groups: Group A) 168 deletion mutants of genes encoding transcriptional regulators; group B) 8 deletion mutants of genes encoding proteins with unknown function and predicted to be involved in lipid homeostasis; and group C) control strains including the BY4742 reference and the *elo2*Δ mutant with defective fatty acid elongase activity [211] (Supplementary table 1[1]). The group A strains were shortlisted based on GO annotations related to transcriptional activity. The group B strains were selected from an interaction network enriched for proteins involved in lipid metabolism and function (Figure 3-1). This network was compiled using the GeneMANIA function prediction algorithm [207] that queried available protein-protein interaction data in the BioGRID database [155] using a query list of lipid-related genes (Supplementary table 2[2]). Selected candidate proteins were all annotated as 'biological process unknown' in SGD [153] (Supplementary table 3[3]).

[1] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d2.txt

[2] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d3.txt

[3] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d4.txt

A key feature of high-throughput lipidomic screening is the ability to identify mutants with pronounced differences in lipid composition compared to reference strains or the average of the library. Differences in lipid composition are most accurately determined by absolute quantification of lipid species where the intensities of detected endogenous lipid species are normalized to the intensities and amounts of appropriate internal standards [96,98,99,201,212]. The concentration of lipid species can be expressed as the molar abundance of lipid species relative to all monitored lipid species (*i.e.* mol%). Notably, absolute quantification on a lipidome-wide level requires spiking samples with ~25 synthetic internal lipid standards, some of which are expensive or require cumbersome approaches to purify. Moreover, the workflow requires a dedicated two-step lipid extraction procedure that is difficult to execute in 96-well plate format. Thus, executing a lipidomic screening across hundreds of yeast strains utilizing absolute quantification on a lipidome-wide level is a challenging undertaking. To combat these technical and economical drawbacks we designed the screening platform to include two rounds of screening. A first round screening was designed for rapid lipid profiling across all shortlisted strains in the deletion library while the second round of screening was designed for comprehensive lipidome quantification in deletion mutants with perturbed lipid phenotypes identified in the first round screening.

For the first round screening, we devised and validated a lipidomic proxy supported by the comprehensive lipidome coverage obtained by high-resolution Orbitrap mass analysis. Lipidomic profiling in positive ion mode allows sensitive analysis of PC, LPC, PE, LPE, DAG, TAG, SE and Cer species. In addition, negative ion mode analysis allows monitoring of PI, LPI, PA, LPA, PS, LPS and IPC species. As such, the acquisition of spectral data for the same sample in both positive and negative ion mode allows monitoring of lipid species abundance by intensity profiling using the proxy ”Intensity%” (I%) calculated as the intensity of a given lipid species divided by the sum of all monitored lipid intensities monitored within a given ion mode (i.e. positive or negative).

To benchmark the proxy for lipidomic screening, we compared the abundance of lipid species monitored by I% and mol% (Figure 3-2A). To this end, we performed a comprehensive lipidomic analysis of nine yeast strains using the workflow for absolute quantification of lipid species as applied for the second round screening and in general for comparative lipidomic analysis [96,99]. The abundances of endogenous lipid species from the same strain were expressed as both mol% (using internal standard information) and I% as outlined above. Our analysis demonstrated a linear correlation between the absolute abundance of lipid species and the lipid levels monitored by I%. Importantly, we observed that lipid species with high absolute mol% values displayed high I% values, and vice versa, lipid species with low mol% values displayed low I% values. Based on this result, we concluded that I% is a valid proxy for assessing the abundance of lipid species and mapping differences in lipid composition by high-throughput lipidomic screening. We note that identified mutants with perturbed lipid composition based on I% should be further investigated using the second round screening approach as this uses the more accurate absolute quantification on a lipidome-wide level. In addition, the second round screening should be performed since the first round lipidomic screening approach does not support quantification of free sterols as these analytes require additional chemical derivatization for quantification [96,99].

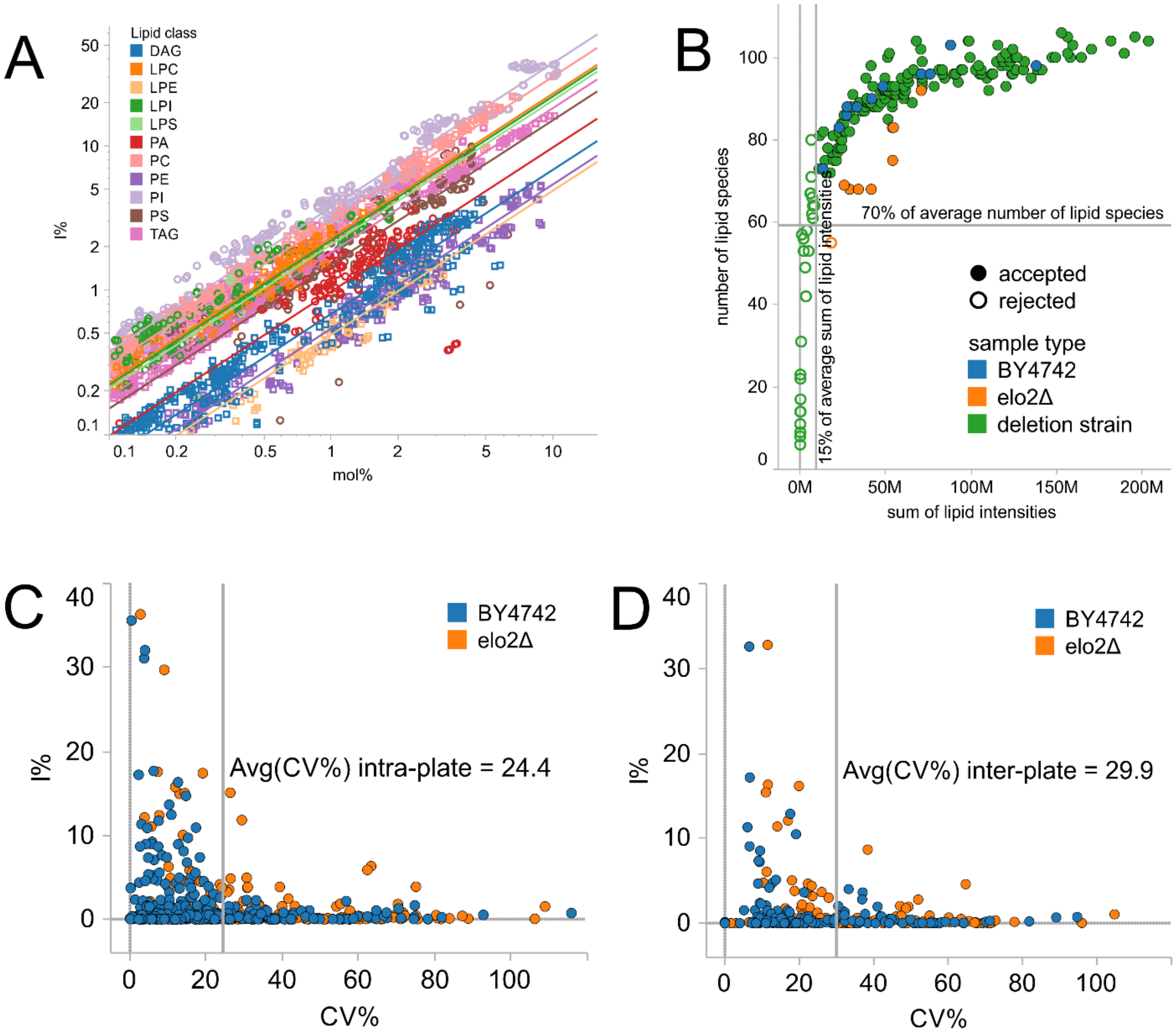


Figure 3-2. Validation of the shotgun lipidomic screening platform.
(A) Correlation between I% and mol%. A lipidomics experiment of 9 yeast strains (used for the second round screening, see Fig. 4) was executed using appropriate internal standards to allow absolute quantification of lipid species (expressed as mol%). The same dataset was used for intensity profiling (expressed as I%). Lines correspond to linear correlation between mol% and I% values for endogenous lipid species of the same class. Average R^2 value is 0.94. (B) Quality control plot with criteria for rejecting poor quality samples. Strains with less than 70% of the average number of detected lipid species and/or less than 15% of the average total intensity were excluded from subsequent analysis. Only negative mode data are shown. Positive mode data were filtered in the same way (data not shown). (C) Average intra-plate coefficient of variation. (D) Average inter-plate coefficient of variation. Inter-plate coefficient of variation reflects technical, within plate, plate-to-plate and biological variation of replicated measurements.
DOI: 10.1039/c3mb70599d.

3.5.2. Lipid profiling and data filtering for first round screening

All 178 strains in the shortlisted deletion library were subjected to the first round screening. Strains were cultured on synthetic complete solid medium in 96-well format for 24 hours in accordance with metabolomic studies [84,213]. Cells were harvested, subjected to cell lysis and single-step lipid extraction in 96-well format followed by shotgun lipidomic analysis using a robotic nanoelectrospray device and a high-resolution Orbitrap mass spectrometer [96,206,214]. The first round screening afforded comprehensive lipidomic analysis of 96 samples in approx. 12 hours. Detected lipid species were identified using ALEX software [205].

Next, we performed a quality control procedure to ensure the reliability of lipidome data across the 178 input strains. First, we introduced a filter to reject all strains featuring i) less than 70% of the average number of detected lipid species and ii) less than 15% of the average total lipid intensity (Figure 3-2B). Using this approach, we passed 128 strains having on average 120 detected lipid species. We note that rejected strains were due to technical issues (e.g. poor ion spray) and poor growth.

To assess the precision of the lipidome data we evaluated the reproducibility of I% using biological replicates of the control strains BY4742 (n=8) and *elo2*Δ (n=7) distributed over three 96-well plates, and grown and analyzed together with shortlisted deletion mutants. The coefficient of variation (CV) for each lipid species detected in these control strains was determined (Figure 3-2C,D). The average inter-plate CV of I% for lipid species was 30%. In comparison, the average intra-plate CV of I% was 24%. In addition, we observed a linear correlation between I% in biological replicates of the control strains (Pearson correlation R-square ≥ 0.973 p-value < 0.0001). We note that the CV values were determined for biological replicates of control strains grown for a fixed time rather than until the strains reached definite cell amount. Hence, one can expect variation in the growth of individual strains which contributes to the relatively high CV values. We note that the applied culturing strategy is commonly used for large-scale metabolomics screening [74,84]. Importantly, for the identification of mutant strains with altered lipid phenotypes we employed a multivariate method (described below) that differentiates mutant strains based on the composite of all lipid species I% values rather than the difference between I% values of single lipid species. Furthermore, the multivariate method differentiates lipid profiles across all surveyed strains instead of referencing only the control BY4742 strain. Using this approach minimizes the seemingly adverse impact of the relatively high CV values. Importantly, this approach successfully identified all replicates of the control mutant strain *elo2*Δ (*i.e.* no false negative identifications of *elo2*Δ) as having a perturbed lipid profile (Figure 3-4A). We note that the lipid phenotype of *elo2*Δ is only modestly different from BY4742 as compared to *elo3*Δ [96]. Moreover, the approach did not identify any of the replicates of the control BY4742 strain to display altered lipidome composition (*i.e.* no false positive identifications of BY4742). Based on these results we conclude that the first round lipidomic screening approach is a valid tool for surveying the lipid profile across hundreds of yeast strains.

3.5.3. Classification of deletion strains into growth phases

Yeast employ different lipid metabolic programs during exponential growth and stationary phase, which can result in false-positive identification of lipid phenotypes [99]. Notably, the yeast lipidome features high levels of glycerophospholipids during exponential growth which become offset by predominately TAG species and an increase in the chain length of fatty acid moieties in the stationary phase. In order to support accurate identification of mutants harbouring defects in lipid metabolism and not differences related to growth phase, we executed the first round screening using 24 hours of culturing in order to allow ample time for all shortlisted deletion strains to enter the stationary phase. Moreover, we executed a parallel lipidomic analysis of exponential and stationary BY4742 cells cultured in liquid medium. Combining these two datasets and average linkage agglomerative hierarchical clustering allowed us to group lipid profiles of mutants from the deletion library into two clusters corresponding to cells in exponential phase or stationary phase based on lipid class composition and species profile (Figure 3-3A). As expected, the majority of the deletion mutants (n=116) displayed lipid profiles corresponding to stationary phase having high levels of TAG species and glycerophospholipid species with longer chain fatty acid moieties as compared to exponential phase cells (Figure 3-3B). In comparison, only a few strains (n=12) displayed a lipidome composition similar to cells in exponential phase. Having delineated the growth-dependent effects, we subsequently surveyed each group of strains separately for altered lipid metabolic phenotypes using multivariate analysis as outlined below.

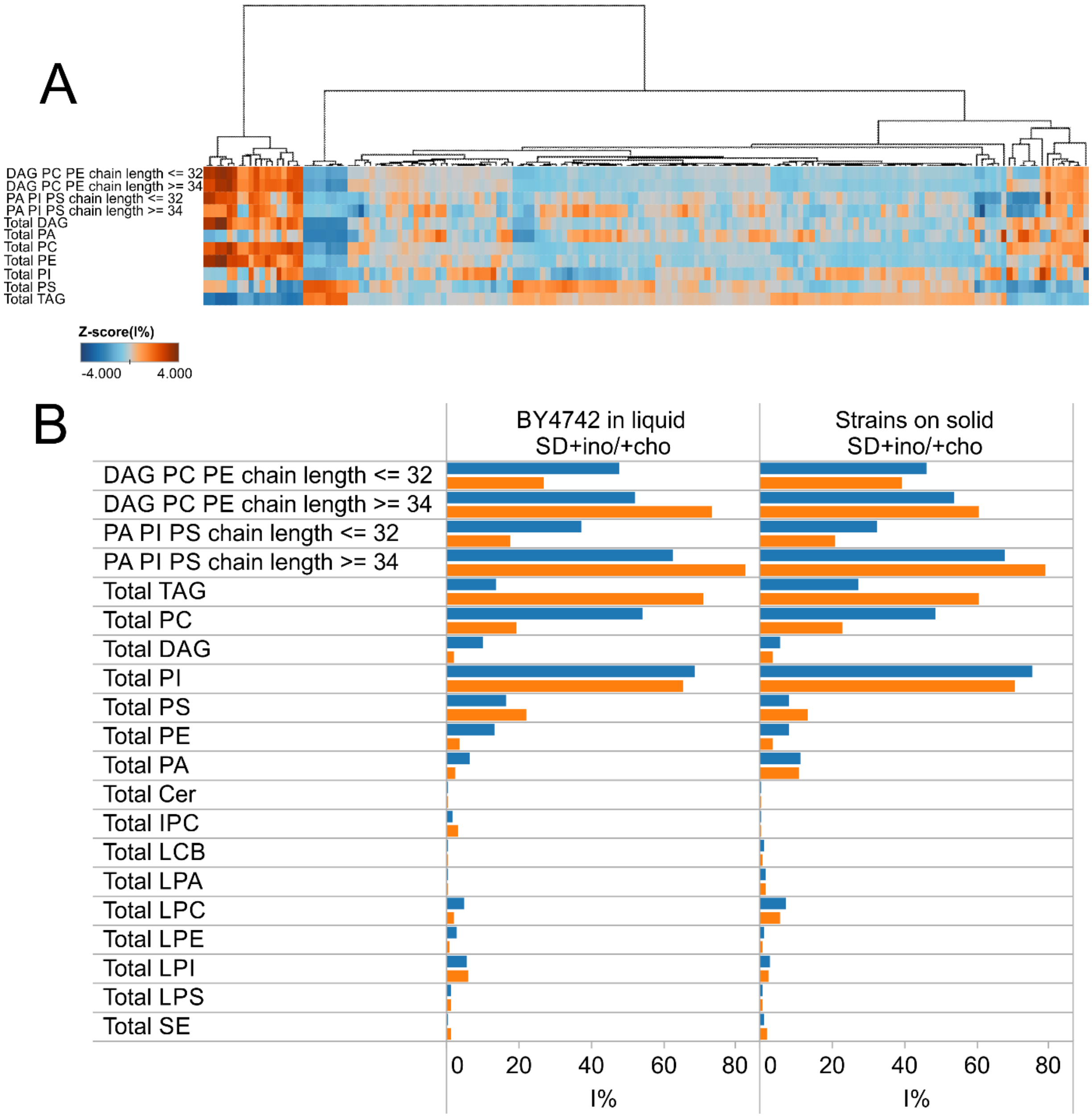


Figure 3-3. Classification of strains into groups based on growth phase.
(A) Hierarchical clustering heat map of deletion mutants classified as stationary or exponential phase based on lipid features of stationary and exponential phase in BY4742. (B) Comparison of lipid features characteristic for exponential (blue) and stationary (orange) phase. Average I% values for classified strains are shown on the right. Average I% values for BY4742 in exponential (n=3) and stationary phase (n=4) are shown on the left. The average I% values for BY4742 in exponential and stationary phase were obtained from a culture in synthetic complete liquid medium. DOI: 10.1039/c3mb70599d.

3.5.4. Identification of deletion strains with perturbed lipid phenotype

It has been demonstrated by recent genomics screening that identification of mutant strains with altered phenotypic traits can be efficiently achieved by referencing the whole collection of analyzed strains rather than comparison to a control strain [41,42,55]. The advantage of such an approach is the possibility to use a higher number of strains for better estimation of technical and biological variation. Consequently, mutants with pronounced phenotypic traits can be identified more accurately. Here we applied a similar strategy for identification of deletion mutants with perturbed lipid phenotypes that is based on "robust principal component analysis" [215]. Conventional principal component analysis can be an effective tool to identify key lipid features in multivariate lipidomic datasets [206]. However, its efficacy can easily be hampered by outlying samples and lipid species. In contrast, the robust variant overcomes the limitation of sensitivity toward outliers by replacing the covariance matrix used for conventional principal component analysis with a robust covariance estimation [216]. Consequently, the robust principal component analysis is better suited for identifying pronounced phenotypic alterations in deletion mutants rather than differences caused by technical and biological variation The robust principal component analysis reduces data dimensionality by computing principal components that explain a maximum amount of observed lipid phenotypic differences across all surveyed strains, and produces a diagnostic plot that classifies strains according to the magnitude and the similarities of lipid phenotypes (Figure 3-4A). Strains that display a common pattern of changes, but exhibit higher differences yield a higher score distance on the x-axis of the diagnostic plot. Strains that display uncommon differences that cannot be explained by the principal components receive high orthogonal distance scores displayed on y-axis of the plot. Effectively, this approach allowed us to identify 11 mutant strains having higher orthogonal distances as compared to the majority of the surveyed strains, and thus, potentially harbouring altered lipid phenotypes (Figure 3-4A). The identified strains were all part of the stationary phase group (Figure 3-3A) and represent ~9% of all shortlisted strains in the deletion library. From the 11 identified candidates, 9 strains were deletion mutants of transcriptional regulators (*cha4*Δ*, kar4*Δ*, met31*Δ*, mga2*Δ*, rrn10*Δ*, rsf2*Δ*, sir1*Δ*, sut2*Δ*, ume6*Δ) and 2 were deletion mutants of genes encoding proteins predicted to play a role in lipid homeostasis (*yjr015w*Δ, *ybr141c*Δ).

In addition, the robust principal component analysis identified all replicates of the control mutant *elo2*Δ. We note that replicates of BY4742 have relatively high score distances. This attribute can potentially be linked to the fact that the BY4742 strain is devoid of the kanamycin resistance cassette (present in deletion strains and potentially able to affect cellular fitness [217] or that BY4742 has not been subject to the same genetic selection as the mutant strains. Importantly, the orthogonal distance score for BY4742 replicates was not high and thereby illustrating no major lipidomic differences compared to the majority of the deletion mutants.

A prominent hit of the first round screening was the transcriptional regulator *MGA2* (Figure 3-4A). Mga2p is an endoplasmic reticulum membrane protein involved in the regulation of *OLE1* transcription [198]. Ole1p is the only fatty acid desaturase in *S. cerevisiae* and is therefore essential for the synthesis of monounsaturated fatty acids [218].

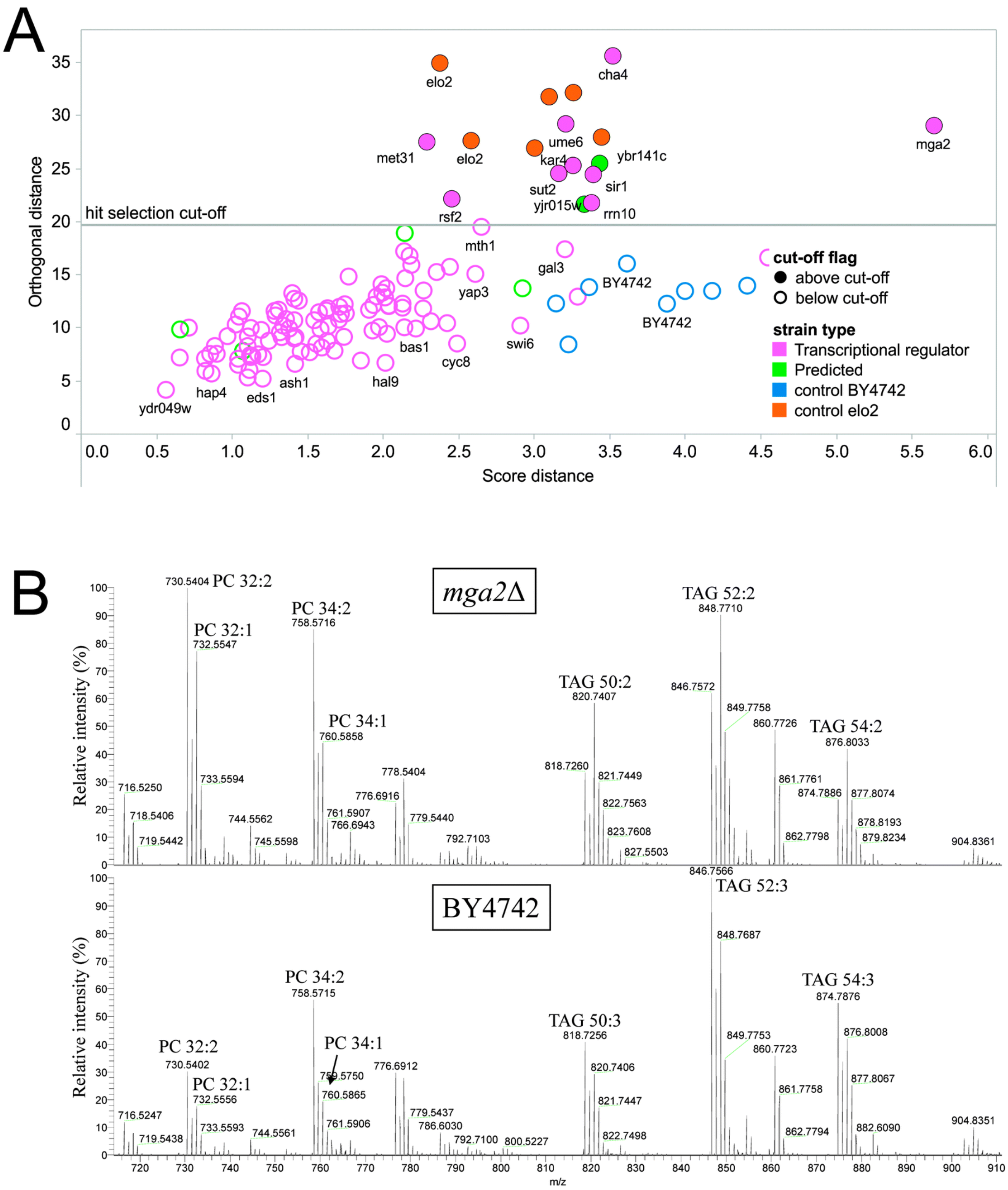


Figure 3-4. Identification of lipid phenotypes by robust principal component analysis. (A) Diagnostic plot. The score distance corresponds to the similarity of deletion mutants based on the principal component model. High score distance values correspond to scores that are different from majority of strains but demonstrate a typical pattern of changes. The orthogonal distance is a measure of how distinct a lipid phenotype is compared to the majority of strains. High orthogonal distance values indicate that a particular lipid composition cannot be explained by the model. Strains with major differences in lipid composition were identified as having orthogonal distance values above the cut-off (corresponds approximately to 97.5% quintile of the Gaussian distribution). (B) Spectral verification of the *mga2*Δ lipid phenotype. Positive ion mode FT MS spectrum of *mga2*Δ and BY4742. DOI: 10.1039/c3mb70599d.

Deletion of the *MGA2* gene reduces the expression of Ole1p which results in lower levels of the monounsaturated fatty acids C16:1 and C18:1 [219]. Consequently, the *mga2*Δ strain synthesizes elevated amounts of lipids having fewer double bonds as compared to the control strain BY4742 (Figure 3-4B). Notably, the first round screening showed that *mga2*Δ synthesizes primarily TAG species with a total of two double bonds and elevated levels of PC species having a single double bond (e.g. PC 32:1). Based on the ability to identify known constituents of the lipid metabolic network (*i.e. mga2*Δ and *elo2*Δ), we conclude that the first round lipidomic screening is a valid tool for identification of mutant strains with perturbed lipid phenotypic traits. We noted that deletion mutants for transcription factors involved in lipid metabolism Ino2p and Ino4p did not cause pronounced changes in lipid composition because the growth medium was supplemented with inositol and choline, which alleviates the phenotype of these deletions. The strain devoid of the transcriptional regulator Opi1p was excluded during the quality control procedure.

3.5.5. Second round screening of deletion strains with lipid phenotypes

In order to further substantiate the lipid phenotypes of identified strains, we executed a second round of lipidomic screening using the accurate and comprehensive workflow for absolute quantification of lipid species [96,99]. To this end, we performed an extensive lipidome analysis of 7 deletion mutants, control strains *elo2*Δ and *gup1*Δ, and the reference strain BY4742 (Supplementary table 4[1]). The strains were cultured in synthetic complete liquid medium for 24

[1] http://www.rsc.org/suppdata/mb/c3/c3mb70599d/c3mb70599d5.txt

hours to allow cells to enter the stationary phase. Quantitative lipidomic analysis was performed using spike-in of internal standards and included the quantification of ergosterol and inositol-containing sphingolipids which were not monitored in the first round screening. To rank the lipid metabolic phenotypes we made use of a scoring algorithm that calculates the sum of absolute mol% difference relative to BY4742 (SoamD). This score was applied to rank the mutant strains according to the magnitude of the differences in lipid species and lipid class composition as compared to the reference strain BY4742 (Figure 3-5). Using this approach we observed that the mutant *elo2*Δ harboured the most pronounced differences in global lipid composition followed by *ybr141c*Δ, *kar4*Δ and *yjr015w*Δ. The lipid phenotypes of *ybr141c*Δ, *kar4*Δ and *yjr015w*Δ will be discussed below.

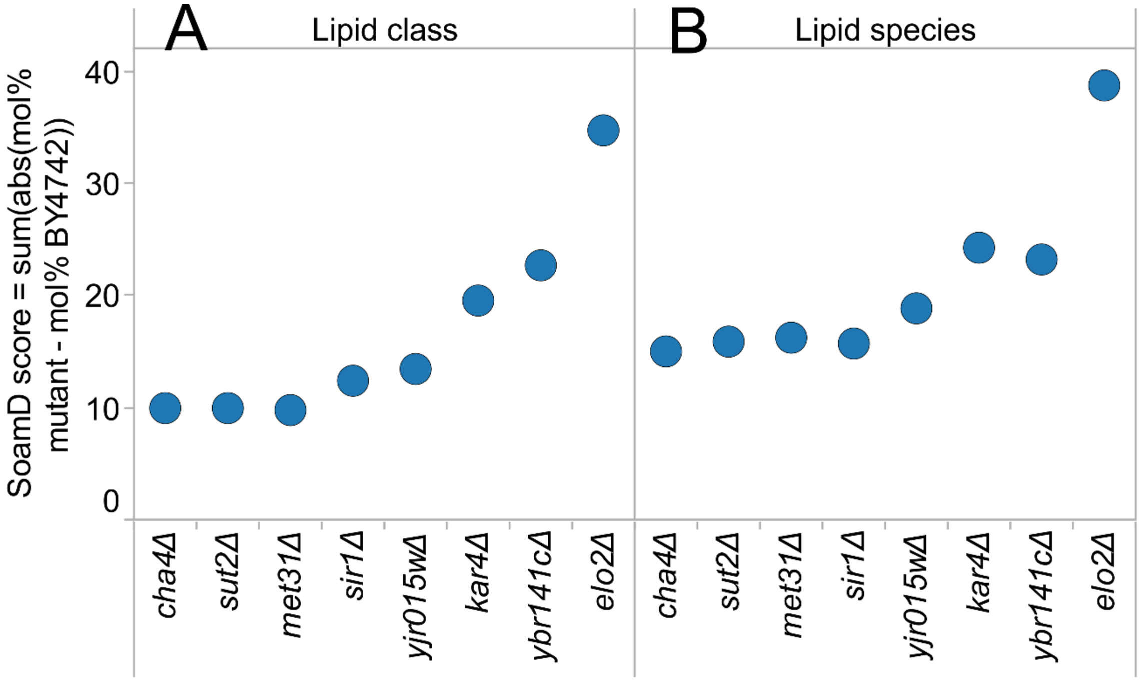


Figure 3-5. Scoring of lipid phenotypes.
Score values represent the sum of absolute differences (SoamD) in mol% lipid class (A) or species (B) as compared to BY4742. Data represent mean of two values from two separate injections of one biological replicate. DOI: 10.1039/c3mb70599d.

3.5.6. YJR015W has a plausible role in GPI-anchor synthesis

Yjr015wp is a protein of unknown function that localizes to the ER [33]. It is predicted to have 6 transmembrane domains and function as a membrane transporter [220]. The interaction landscape of *YJR015W* shows a direct interaction with key enzymes in fatty acid elongation and sphingolipid metabolism *IFA38* and *SUR4*, and two enzymes involved in glycosylphosphatidylinositol (GPI) anchor synthesis *GPI16* and *GUP1* (Figure 3-6A). All four interactors are transmembrane proteins that localize to the ER [133,221]. Ifa38p and Sur4p are components of the elongase complex that synthesizes C26:0 fatty acid for Cer synthesis and remodeling of GPI-anchors via the *O*-acyl-transferase Gup1p. Gpi16p is a subunit of the transamidase complex that adds GPI-anchors to newly synthesized proteins. Notably, GPI-anchored proteins in *S. cerevisiae* comprise either a glycerophospholipid PI species or a sphingolipid IPC species. In addition, *YJR015W* also interacts with enzymes responsible for N-linked glycosylation and machinery involved in ER to Golgi vesicle transport (Figure. 3-6A). Taken together, these interactions support the notion that Yjr015wp is potentially involved GPI-anchor synthesis. Based on this prediction we included *gup1*Δ as a control strain in the comparative lipidomic analysis.

The lipid phenotype of *yjr015w*Δ revealed a distinct set of perturbed sphingolipid features (Figure 3-6B,C). The *yjr015w*Δ lipidome showed increased levels of 46:0;4 sphingolipid species being offset by a reduction in 44:0;5 sphingolipid species. The 46:0;4 species correspond to a sphingolipid composed of a C20 phytosphingosine and an amide-linked hydroxylated C26:0 fatty acid moiety [96,222]. In comparison, the 44:0;5 species correspond to sphingolipid having a C18 phytosphingosine and a C26:0 fatty acid moiety with two hydroxyl groups. The hydroxylase that inserts the second hydroxyl group onto the C26 fatty acid moiety as well as the position on the fatty acid chain and the molecular function of the produced sphingolipid molecule are unknown [223,224]. It is unlikely that the sphingolipid phenotype of *yjr015w*Δ is attributed to a reduced activity of the fatty acid elongation complex because inactivation of the interactor Sur4p would shorten the fatty acid chain length of sphingolipids [96,211]. Instead, the perturbation of sphingolipid hydroxylation profile could be due to reduced activity of the unknown hydroxylase or selective utilization of IPC 44:0;5 species for remodeling of GPI-anchors. Interestingly, inactivation of GPI-anchor remodeling in the *gup1*Δ mutants coincide with a reduction in IPC 44:0;5 (termed IPC-D in thin-layer chromatographic analysis) and increased incorporation of a base resistant anchor lipid with chromatographic properties similar to IPC 44:0;5 [225]. In addition, it has been observed that GPI-anchor proteins in *gup1*Δ cells comprised lower levels of IPC 44:0;4. Taken together, these data support the observed reduction in all sphingolipid 44:0;5 species and elevated levels of 46:0;4 species in the *yjr015w*Δ lipidome (Figure 3-6B). Our analysis showed that the *yjr015w*Δ lipidome partially phenocopied the *gup1*Δ lipidome with respect to the top 10 decreasing lipid species (Figure 3-6B,D). In both mutant strains, the 44:0;5 sphingolipid species were among the most reduced lipid species. In contrast, the observed top 10 increased lipid species were only partially conserved in the two mutant strains. As for the lipid class phenotype we also observed a strong similarity between the *yjr015w*Δ and *gup1*Δ lipidome. Based on the distinct lipid phenotype of *yjr015w*Δ, its similarity to the *gup1*Δ lipidome and the interaction landscape of *YJR015W* we propose that Yjr015wp might play a functional role in modulation of GPI-anchor synthesis.

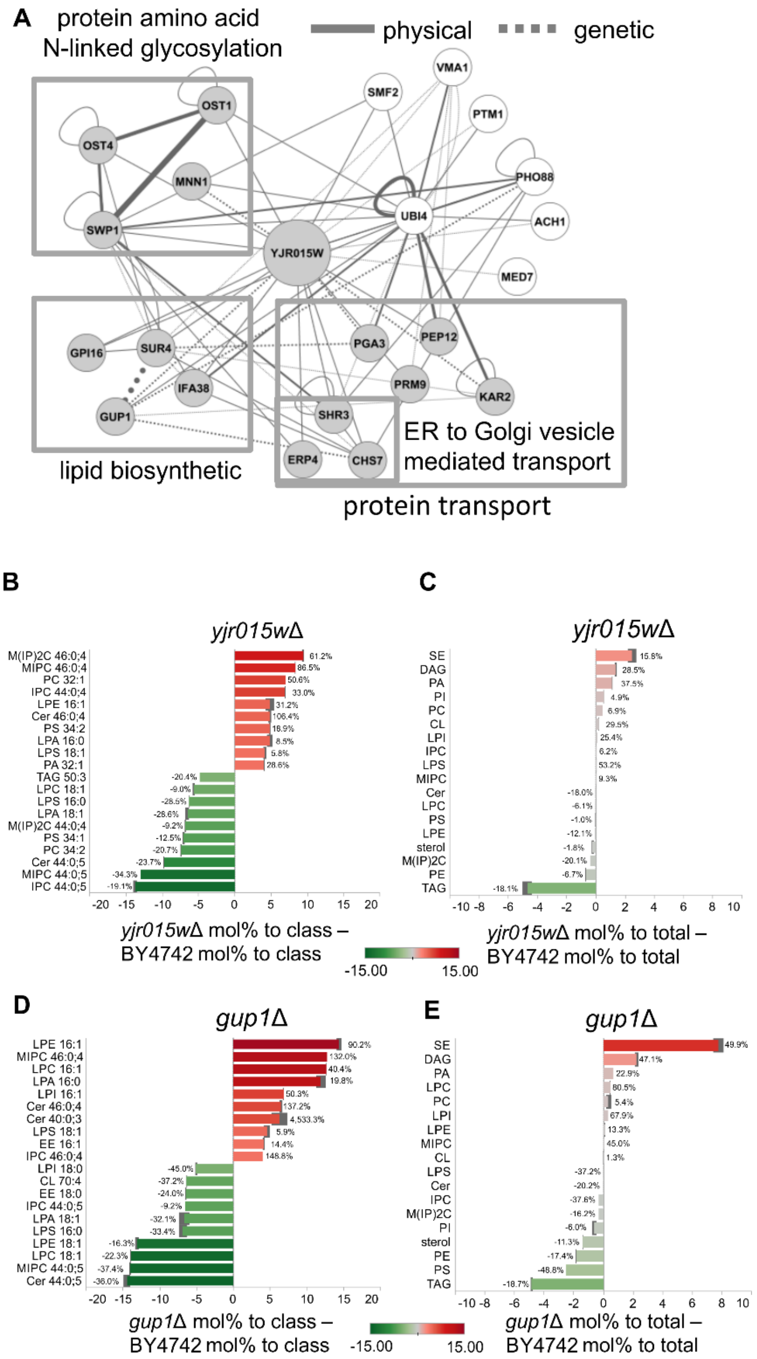


Figure 3-6. Interaction network and lipid phenotype of *yjr015w*Δ.
(A) Physical and genetic interactions with *YJR015W* from BioGIRD. Groups of genes with significantly enriched GO terms related to biological processes are highlighted (grey nodes). White nodes are genes with no significant enrichment in biological process GO terms. (B) Top 10 increased and decreased lipid species of *yjr015w*Δ compared to BY4742. (C) Top 10 increased and decreased lipid classes of *yjr015w*Δ compared to BY4742. (D) Top 10 increased and decreased lipid species of *gup1*Δ compared to BY4742. (E) Top 10 increased and decreased lipid classes of *gup1*Δ compared to BY4742. Labels correspond to percentage difference calculated as (mol% mutant – mol% BY4742) / (mol% BY4742). Data display the average of two independent analyses of a lipid extract of one biological replicate. Grey bars report the difference between the replicate data. DOI: 10.1039/c3mb70599d.

We note that in order to reveal the exact role of *YJR015W* an additional study of the *yjr015w*Δ strain and the Yjr015w protein is required.

3.5.7. Ybr141cp – a putative methyltransferase involved in sterol lipid metabolism

YBR141C encodes a putative methyltransferase that localizes to the nucleolus [33]. A recent bioinformatic study of yeast methyltransferases predicted Ybr141cp to contain a Rossmann-like catalytic domain similar to the sterol methyltransferase Erg6p that converts zymosterol to fecosterol in the ergosterol biosynthetic pathway [226]. The catalytic Rossmann-like domain spans methyltransferases with diverse substrate specificities including sterols, proteins, RNA and other small molecules. As such, Ybr141cp was predicted to use rRNA or tRNA as substrate [226], albeit this has not been experimentally verified. In addition, a number of proteins devoid of methyltransferase activity and featuring the Rossmann-like domain have been identified. These proteins include the transcription factor Kar4p and the mitochondrial RNA polymerase specificity factor Mtf1p [226,227]. The interaction network of *YBR141C* shows a link to lipid metabolism via a physical interaction with Vps74p (Figure 3-1), a phosphoinositide-binding protein involved in localizing glycosyltransferases in the Golgi [57].

The lipid phenotype of *ybr141c*Δ showed a pronounced increase of ergosterol esters offset by a reduction of ergosterol (Figure 3-7A,B). In addition, the *ybr141c*Δ lipidome showed a concomitant increase in TAG levels and a reduction in all membrane glycerophospholipids. The reason for this distinct lipid phenotype is at the present time unclear given the limited information about *YBR141C* function. Interestingly, a similar perturbation of sterol esters and

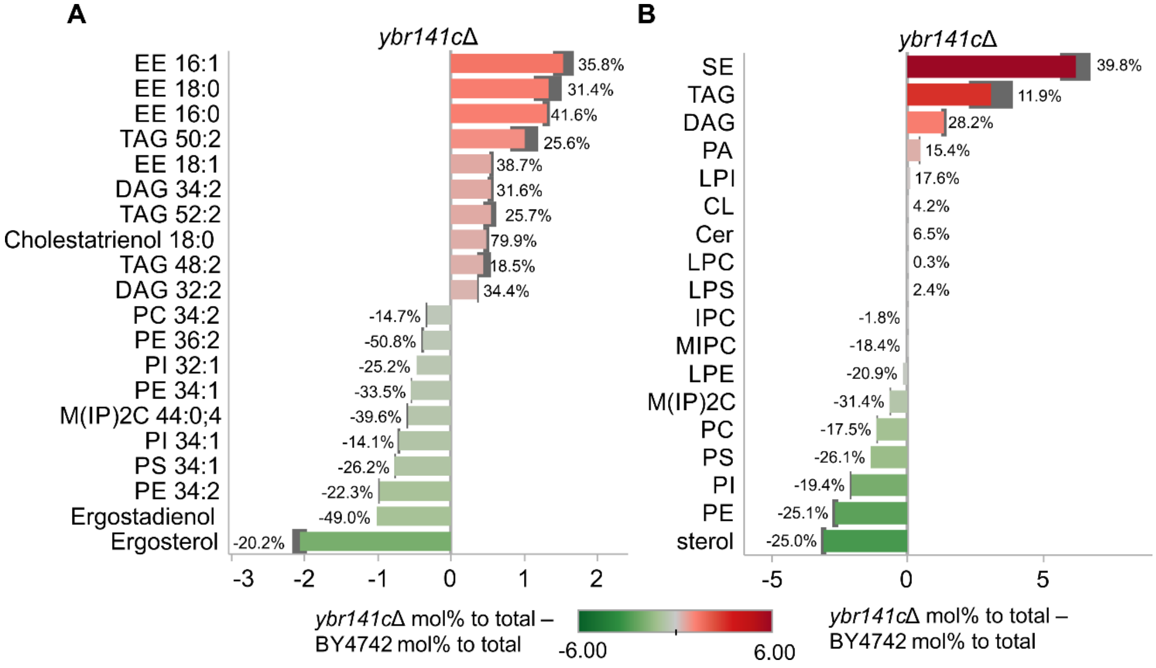


Figure 3-7. Lipid phenotype of *ybr141c*Δ.

(A) Lipid species mol% differences compared to BY4742. (B) Lipid class mol% differences compared to BY4742. Labels correspond to percentage difference calculated as (mol% mutant – mol% BY4742)/(mol% BY4742). Data display the average of two independent analyses of a lipid extract of one biological replicate. Grey bars report the difference between the replicate data. DOI: 10.1039/c3mb70599d.

free sterol levels were observed when inactivating *ERG6*[228]. Moreover, chemical genomic data shows that *YBR141C* has a co-fitness interaction with the major sterol acyl-transferase *ARE2* [229]. Although the information about *YBR141C* function is limited, our results provide a framework for testing the functional role of Ybr141cp in sterol metabolism.

3.5.8. Kar4p – a transcription factor linked to nuclear membrane dynamics

KAR4 encodes a transcription factor required for nuclear fusion during yeast mating and possibly other functions during vegetative growth [230–233]. Kar4p exists as two isoforms; a constitutive 38.5 kD protein (Kar4p-long) that predominates during vegetative growth and a 35.5 kD protein (Kar4p-short) that is induced during mating [232]. During the mating process, Kar4p-short acts together with the transcription factor Ste12p to induce the expression of *KAR3* and *CIK1* that encode a motor protein complex required for congression of nuclei prior to nuclear membrane fusion [231,233]. During vegetative growth, Kar4p-long expression is up-regulated in the G_1 phase of the cell cycle and implicated in constitutive expression of more than 50 genes [231,233]. The *kar4*Δ deletion mutant displays a slow growth phenotype attributed to a short G_1 pause during vegetative growth, and a pronounced defect in nuclear congression during mating that phenocopies the absence of *KAR3* and *CIK1* [230].

Given the functional role of Kar4p in nuclear fusion, it is plausible that Kar4p is also involved in regulating nuclear membrane dynamics during the cell cycle. Notably, perturbing lipid metabolism has previously been shown to compromise nuclear membrane growth and function [234]. Deletion of the PA phosphatase Pah1p and components of its regulatory complex Nem1p-Spo7p reduces PA to DAG conversion and causes nuclear membrane expansion [235,236]. In addition, overexpression of the nuclear/ER-localized DAG kinase Dgk1p phenocopies the Pah1p deficiency [237] indicating that regulation of the composition of DAG, PA and other glycerophospholipids is important for nuclear membrane dynamics. Conversely, deletion of integral nuclear membrane-ER proteins Brr6 and Apq12 precipitate defects in nuclear pore complex assembly, sterol metabolism and lipid droplet morphology [238].

The lipid phenotype of *kar4*Δ showed a pronounced increase in DAG and SE species being offset by a reduction in primarily PE species and ergosterol (Figure 3-8). Interestingly, this lipid phenotype is reminiscent of the lipid compositions associated with nuclear membrane defects observed in the previous studies. The elevated DAG levels and reduction in PE levels are similar to the effects of overexpressing Pah1p [239], whereas the increased levels of SE and reduced levels of ergosterol resemble effects of defective nuclear membrane growth in the *pah1*Δ deletion mutant [240]. This apparent combination of perturbed lipid features indicates that inactivation of Kar4p potentially fails to prompt an inhibition of Pah1p activity, which in turn channels PA into DAG production instead of synthesis of PE and other glycerophospholipids for membrane expansion. The accumulation of SE could be a secondary effect of *kar4*Δ cells trying to synchronize the rate of ergosterol biosynthesis and secretory vesicle flow under the reduced vegetative growth rate. We here note that the exact function of Kar4p action during vegetative growth requires further characterization of the *kar4*Δ strain and Kar4 protein.

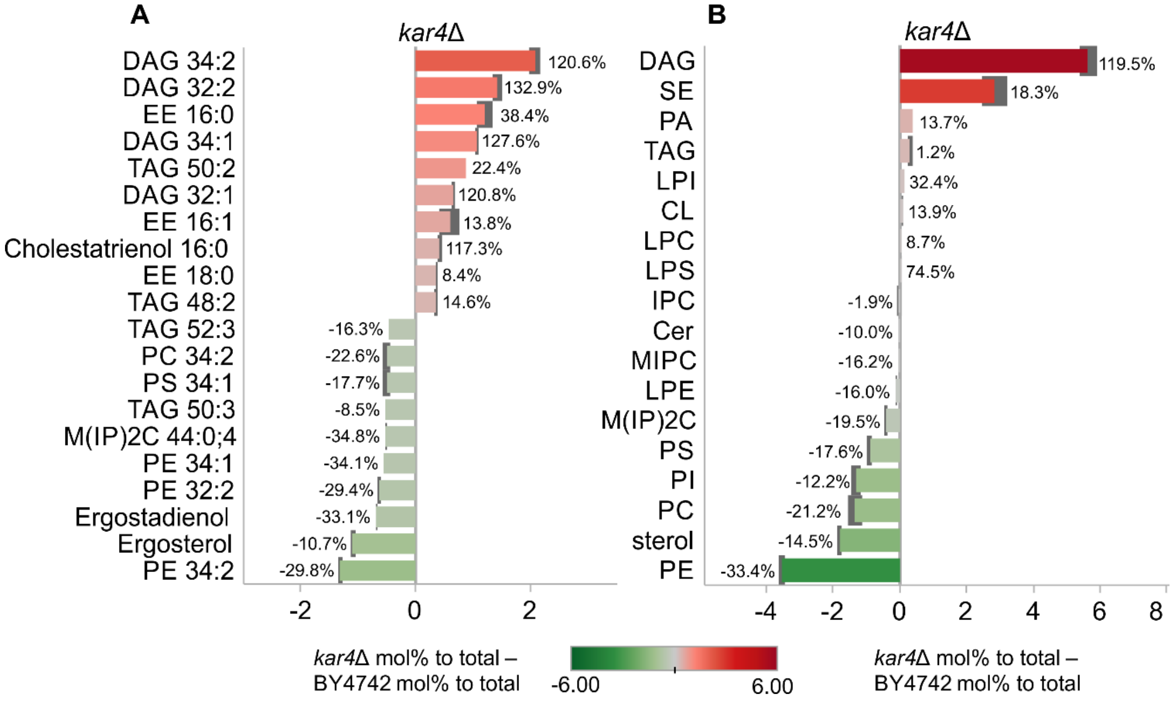


Figure 3-8. Lipid phenotype of *kar4*Δ.
(A) Lipid species mol% differences compared to BY4742. (B) Lipid class mol% differences compared to BY4741. Labels correspond to percentage difference calculated as (mol% mutant – mol% BY4742)/(mol% BY4742). Data display the average of two independent analyses of a lipid extract of one biological replicate. Grey bars report the difference between the replicate data. DOI: 10.1039/c3mb70599d.

3.6. Conclusion

Existing high throughput screening methodologies for identification of proteins with a function in lipid metabolism provide only limited phenotypic information. Epistatic miniarray profiling and protein-protein interactions assays are based on monitoring growth fitness (i.e. colony size) which allows mapping functionally related genes and proteins, respectively [41,42,55].

Alternatively, microscopy-based screening can be used for identifying mutant strains with perturbed lipid droplet dynamics based on altered lipid droplet size and morphology [241,242]. In contrast, detailed assessment of cellular lipid composition demands dedicated methods such as mass spectrometry-based lipidomics. Several groups have recently reported workflows for lipidomics analysis in yeast, but their application has so far been limited to characterisation of only few mutant strains or a reference strain grown at different conditions [96,97,99,243].

In the current study, we designed a platform for high-content lipidomic screening of yeast mutant libraries. We combined culturing and lipid extraction in 96-well format, automated direct infusion nanoelectrospray ionization, high-resolution Orbitrap mass spectrometry and a novel data processing framework to support lipid phenotyping across hundreds of *S. cerevisiae* mutants. The screening platform was designed to include two rounds of screening. A first round screening was executed for rapid lipid profiling across all shortlisted strains in the deletion library while a second round of screening was conducted for more comprehensive lipidome quantification of deletion mutants with perturbed lipid phenotypes identified in the first round screening. To our knowledge, this is the first assessment of lipidomic phenotypes across hundreds of mutants in a single screen. Notably, our platform extends the palette of analytical techniques available for functional genomics studies aimed at uncovering proteins with previously unknown function in lipid metabolism and regulators of global lipid homeostasis. The technology affords a multidimensional survey of physiological lipid parameters that helps explore uncharacterized proteins and propose valuable hypothesis for mechanistic biochemical follow-up experiments. Our case study of a library covering deletion mutants of genes with predicted function in lipid metabolism and transcriptional regulators revealed three poorly characterized genes that precipitate distinct lipid metabolic phenotypes upon deletion. In conclusion, the high-throughput lipidomic screening platform described herein is a valid and complementary tool for high-content analysis of yeast mutant libraries.

3.7. Acknowledgments

We are grateful to Nils J. Færgeman, Robin W. Klemm, Ole N. Jensen and Jens Knudsen for productive discussions and access to equipment. We thank Valentin Todorov for valuable advice on multivariate methods. We thank the reviewers for their constructive comments on the manuscript. This work was supported by the Danish Council for Independent Research | Natural Sciences (09-072484; 10-094213) and Lundbeckfonden (R44-A4342).

Chapter 4 : iVici: Interrelational Visualization and Correlation Interface

4.1. Contribution to the published work

The method presented in this chapter was published in the following paper: Tarassov K, Michnick SW, iVici: Interrelational Visualization and Correlation Interface. Genome Biology. 2005;6(13):R115. Epub 2005 Dec 30 [190]. doi:10.1186/gb-2005-6-13-r115.

M.S.W. and T.K. formulated the original idea. T.K. developed a software implementation of the method and analyzed data that illustrate the applications of the method. M.S.W and T.K. wrote the manuscript.

The paper had been published before completion of the screens presented in this thesis. Therefore, datasets from other studies were used for illustrating the advantage of the method. To correct this, I wrote an original Chapter 4 to demonstrate the application of the method on the data from the thesis.

4.2. Abstract

We have developed a novel visualization method, iVici, for interactive browsing and comparison of large datasets. The method is an extension of the heat map plot and is suitable for analysis of any types of data that can be represented as two-dimensional matrices, e.g. networks of interactions between genes and proteins and correlation networks. To demonstrate

the utility of iVici, we compared interaction networks mapped by different methods and identified interactions unique to a particular dataset.

4.2. Introduction

The functional genomic methods capture diverse molecular information in different organisms. The common question of integrative data analyses is how different dataset generated by various methods in distinct organisms correlate with each other. Powerful software applications exist for representing networks as graphs. However, because of the overlap between edges, it is difficult to navigate large networks represented as graphs. Therefore, alternative visualization methods, such as heat map representation of networks, have been developed that minimize the overlap between network edges (section 1.6.6. Methods for network visualization).

The traditional heat map tools display one data matrix at a time [17,147] using intensity of one or two colors to visualize numeric values of matrix elements. The datasets which contain only positive or only negative values are displayed with one color, whereas, datasets containing both positive and negative values are visualized with two-color schemes with a dedicated color for values above (e.g. red) and below zero (e.g. green). Below, we present two improvements to the method that we have introduced for superimposition of two datasets on the same figure: a) in the case of symmetric matrices, we display values from different datasets below and above the main matrix diagonal; and b) we extend the number of colors that can be selected for visualization.

4.3. Results

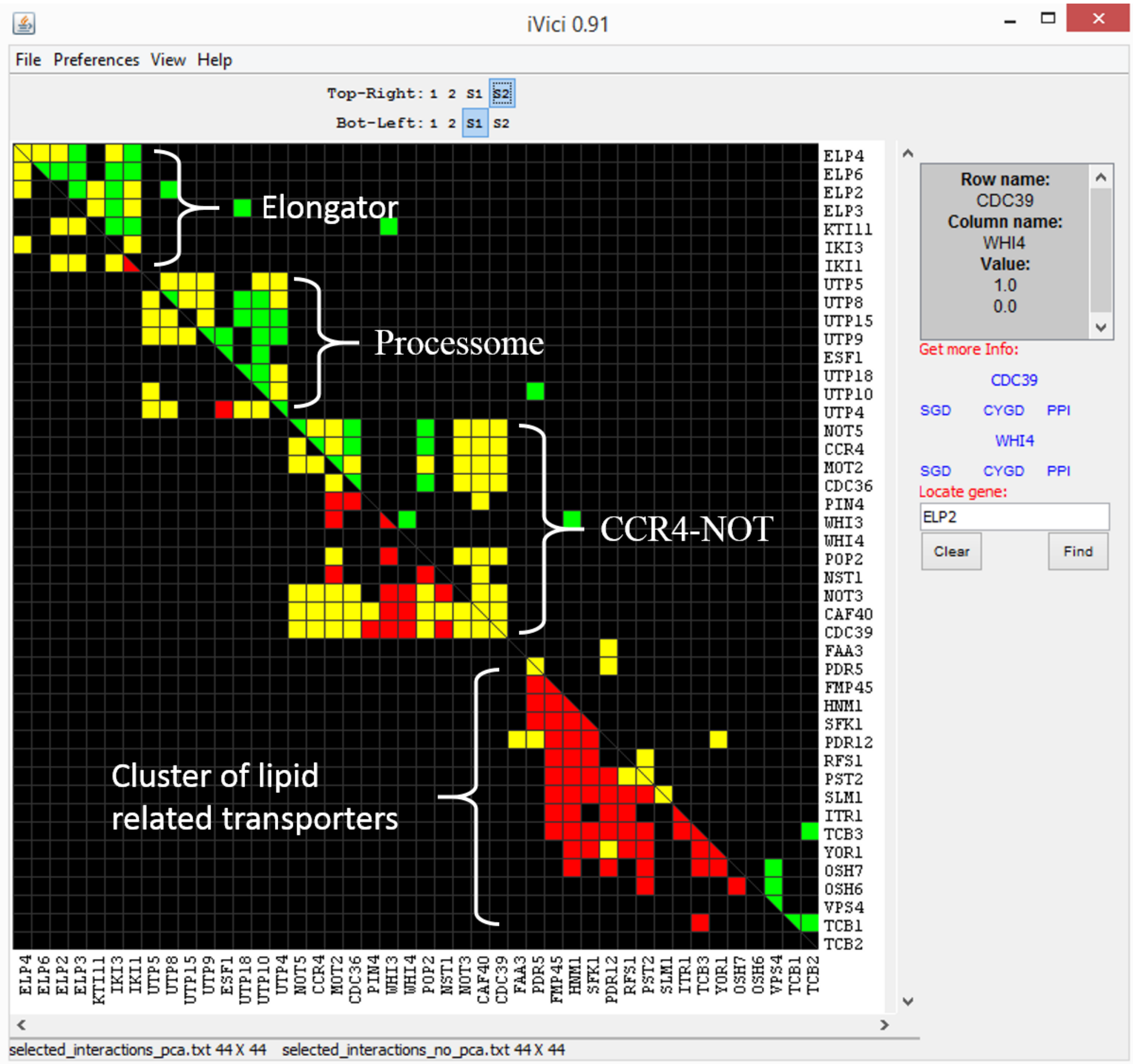
4.3.1. Two datasets within symmetric matrices

Many types of relationships can be displayed with heat maps of symmetric matrices. For example, networks of pairwise protein-protein interactions (such as the one presented in Figure 2-8 in Chapter 2) are displayed as squares with a full set of proteins from the dataset placed into rows and columns. The order of the rows and columns is kept identical by sorting protein names alphabetically or ordering by hierarchical clustering performed in two dimensions [104]. In such setting, the same interaction between proteins *i* and *j* is displayed twice: in the element in row *i* and column *j* and in the element in row *j* and column *i* (Figure 2-8). Another example of symmetric matrices is presented in Chapter 2 (sections 2.5.3 and 2.5.4) in analysis of interaction enrichment between various GO annotation terms.

We used the symmetry property of the map to introduce the first improvement to the heat maps by fitting two datasets into triangles below and above the main matrix diagonal. Thus, instead of duplicated information, elements *i,j* and *j,i* represent different aspect of a relationship between a pair of objects *i* and *j*. Our method allowed us to display on the same plot information on amount of interactions that were detected between pairs of GO terms (lower triangle) and the difference of this number from what is expected by chance (upper triangle) (Figures 2.12 and 2.13). Combining these two metrics is important for appreciating the significance of the results. In the case, where the number of detected interactions is large, it may suggest that interactions between two terms are favored in the network. However, when many proteins are annotated with these terms similarly large numbers of interactions could be discovered also in a network with randomly assigned interactions. Figure 2-12 contains an example of such phenomenon, in which large number of interactions in DHFR PCA network between proteins in cytoplasm (column 7) and nucleus (row 19) is depicted by a bright yellow color in the lower triangle. However, this number is lower than what would be expected by chance as indicated by a bright blue color (that corresponds to depletion of interactions) in row 7 column 19.

4.3.2. Extended color-schemes for comparing datasets

The second improvement that we have introduced to the heat map method was the extended number of colors that can be used. Instead of at maximally two colors available in the traditional method, we make use of up to seven colors to visualize the overlap and discrepancies between two datasets. Our implementation allows to select pairs of colors for each of the datasets that are compared. Additionally, separate colors can be selected to show overlapping values. In Figure 4-1, we present a zoomed region of the complete DHFR PCA heat map from Figure 2-8. The lower triangle bellow the main matrix diagonal shows interactions from DHFR PCA network, and the upper triangle above the main matrix diagonal depicts interactions between the same proteins detected by other studies extracted from the BIOGRID database. Interactions that are only observed by PCA are colored in red. The interactions that are not detected by PCA, but observed in other studies are colored in green. Yellow is used to show PCA interactions confirmed by other methods.



Top-Right: 1 2 S1 [✓ S2]

Bot-Left: 1 2 [✓ S1] S2

Row name: CDC39
Column name: WHI4
Value: 1.0
0.0

Locate gene: ELP2

selected_interactions_pca.txt 44 X 44 selected_interactions_no_pca.txt 44 X 44

Figure 4-1. Screenshot of iVici interface with PCA network.
Network is based on interdictions from DHFRP PCA network ordered by hierarchical clustering (lower triangle). Upper triangle shows interactions between the same proteins detected by other methods (all available data on physical protein interactions downloaded from the BIOGRID database).

4.3.3. Software implementation

We have developed a platform-independent software called iVici (iVici: Interrelational Visualization and Correlation Interface) that is capable of presenting the improved heat maps discussed above. We have programmed a stand-alone software instead of providing an extension to popular Cytoscape and R systems, because Cytoscape facilitates network visualization presented as graphs and not as heat maps, while R does not provide capabilities for development of interactive user interfaces. The software was coded in Java version 1.4. It runs on all major desktop operating systems that support Java. iVici allows to configure links to up to four different databases for getting information on network nodes. The exploration of large heat maps is facilitated by a dedicated navigation pane, which shows the full network. A movable rectangle in the navigation pane allows to choose a zoomed region for displaying in the main application window. We set up a dedicated website (http://michnick.bcm.umontreal.ca/resources.html) where iVici software, example datasets and documentation are available for download.

4.4. Discussion

We have selected four clusters to demonstrate how the heat map with superimposed colors and datasets can be used to intuitively browse and compare large networks for evaluating reproducibility and novelty of the results. Clusters in Figure 4-1 exhibit a different degree of the overlap. The interactions between members of two clusters that correspond to Elongator complex [244,245] and Processome complex [246,247] are in agreement between PCA technique and other studies. There is a good overlap between the datasets describing interactions of Ccr4-Not complex with ubiquitination and deadenylation enzymatic activities that regulates gene expression and its coordination between the nucleus and the cytoplasm [248,249]. In this example, DHFR PCA network contains new interactions (shown in red) that has not been previously reported. The novel interactions link Ccr4-Not complex with paralogous RNA-binding proteins Whi3p and Whi4p. *WHI3* has been implicated in regulation of the cell cycle progression based on observation of unusual small-cell phenotype of a *whi3*Δ deletion mutant. Deletion of *WHI4* does not cause a strong phenotype alone. However, double deletion mutation *whi3*Δ *whi4*Δ results in even smaller cells than single *whi3*Δ deletion. These results suggest that *WHI3* is partially redundant with *WHI4*, but is less important [250]. Our network supports the view that *WHI4* might have a similar function to *WHI3* because of common interaction partners detected for Whi4p and Whip3. Whi3p binds to mRNA of Cln3p [251], a key activator of the cell cycle entry, and acts as a cellular retention factor for Cdc28p [252], cyclin-dependent protein kinase that regulates the cell cycle. Cln3p acts as an activator of Cdc28p that promotes the G1 to S phase transition. Newly discovered interactions suggest association between Cln3p-Cdc28p cell cycle regulation complex, Whi3p and the Ccr4-Not complex. The mechanism by which Whi3 affects translation of Cln3 after binding to its mRNA is unknown. Based on the DFHR PCA interactions, it has been hypothesized that Whi3p bound to mRNA of Cln3p might recruit Ccr4-Not complex that would promote degradation of the Cln3p mRNA by increasing the rate of deadenylation of the poly(A) tail, which is the initial step of the mRNA turnover [253].

The last example in Figure 4.1 displays a cluster of lipid related transporters with the majority of interactions discovered only in DHFR PCA network (discussed in more details in section 2.5.5.).

Visualization of non-symmetric datasets that cannot be fit into the same square would still benefit from the extended color-scheme. In this case, only one dataset at a time is displayed; however, overlapping values would be selectively colored based on the information contained in the second dataset.

4.4. Conclusion

Here we have presented an extension of the traditional heat map visualization method for comparison of various datasets. We have provided examples on the use of the method for analysis of relationships between proteins that belong to various GO annotations terms and investigation of the overlap between interactions detected by distinct methods. Similarly, the method can be used for analysis of network dynamics by comparing interactions observed in different conditions; and comparison of heterogeneous data, such as genetic and physical interactions. Because of the absence of the overlap between connected nodes, relationships between groups of nodes can be easily explored even in very large datasets. In addition to grouping of proteins by hierarchical clustering for discovering interconnected modules, ordering can be done, for example, by cellular locations of proteins, positions of genes on chromosomes and functional annotations. To promote the application of the method, we have provided a platform-independent software implementation of the method (iVici) available at no charge.

Finally, iVici is not limited to visualization of interaction networks and can be used for comparative analysis of any type of information that can be represented in the matrix format.

Chapter 5 : Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data

5.1. Contribution to the published work

Chapter 5 was published as an article in PlosOne journal: Husen P*, Tarasov K*, Katafiasz M, Sokol E, Vogt J, Baumgart J, Nitsch R, Ekroos K, Ejsing CS (2013) Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data. PLoS One 8: e79736 (* These authors contributed equally to this work)

Chapter 5 describes software framework that was initially developed to support the high-throughput lipidomics screen presented in Chapter 3. For the publication, the framework was extended further for handling a broader range of lipidomics applications. My contribution to this work consisted of development of a software module for automated extraction of the raw data from files generated by mass spectrometer and modular design of the software implemented with visual programming and visual analytics technologies.

Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data

Peter Husen[1☯], Kirill Tarasov[2☯], Maciej Katafiasz[1], Elena Sokol[1], Johannes Vogt[3], Jan Baumgart[3], Robert Nitsch[3], Kim Ekroos[2*], Christer S. Ejsing[1*]

1 Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense, Denmark, 2 Zora Biosciences Oy, Espoo, Finland, 3 Institute for Microscopic Anatomy and Neurobiology, Johannes Gutenberg University Mainz, Mainz, Germany

Citation: Husen P, Tarasov K, Katafiasz M, Sokol E, Vogt J, et al. (2013) Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data. PLoS ONE 8(11): e79736. doi:10.1371/journal.pone.0079736

Editor: Lennart Martens, UGent / VIB, Belgium

Received July 5, 2013; **Accepted** September 24, 2013; **Published** November 7, 2013

Funding: The work was supported by Lundbeckfonden (R44-A4342) and the Danish Council for Independent Research | Natural Sciences (09-072484). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The affiliation to the company Zora Biosciences is solely based on scientific collaboration. There are no declarations to be made related to employment, consultancy, patents, products in development or marketed products etc. The affiliation to the company does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* Corresponding authors.

☯ These authors contributed equally to this work.

5.2. Abstract

Global lipidomics analysis across large sample sizes produces high-content datasets that require dedicated software tools supporting lipid identification and quantification, efficient data management and lipidome visualization. Here we present a novel software-based platform for streamlined data processing, management and visualization of shotgun lipidomics data acquired using high-resolution Orbitrap mass spectrometry. The platform features the ALEX framework designed for automated identification and export of lipid species intensity directly from proprietary mass spectral data files, and an auxiliary workflow using database exploration tools for integration of sample information, computation of lipid abundance and lipidome visualization. A key feature of the platform is the organization of lipidomics data in ”database table format” which provides the user with an unsurpassed flexibility for rapid lipidome navigation using selected features within the dataset. To demonstrate the efficacy of the platform, we present a comparative neurolipidomics study of cerebellum, hippocampus and somatosensory barrel cortex (S1BF) from wildtype and knockout mice devoid of the putative lipid phosphate phosphatase PRG-1 (plasticity related gene-1). The presented framework is generic, extendable to processing and integration of other lipidomic data structures, can be interfaced with post-processing protocols supporting statistical testing and multivariate analysis, and can serve as an avenue for disseminating lipidomics data within the scientific community. The ALEX software is available at www.msLipidomics.info.

5.3. Introduction

The lipidome of eukaryotic cells comprises hundreds to thousands of molecular lipid species that constitute and functionalize biomembranes, store metabolic energy in lipid droplets and function as signaling molecules that control cell and organism physiology [89,191,254]. A key tenet of contemporary mass spectrometry-based lipidomics methodology revolves around the identification and quantification of lipid species on a lipidome-wide scale [96,99,202,212,255]. As such, shotgun lipidomics has emerged as a powerful tool for global lipidome analysis that complements mechanistic studies of lipid metabolism, lipid homeostasis and membrane biology [134,200,201,256,257]. The efficacy of shotgun lipidomics stems from its relative technical simplicity where hundreds of lipid species in sample extracts can be quantitatively monitored at high throughput using direct infusion nanoelectrospray ionization combined with high-resolution Fourier transform mass spectrometry (FT MS) or/and tandem mass spectrometry (MS/MS) [89,204]. Notably, lipidomics analysis on a global scale generates large amounts of (spectral) data that requires software routines for automated lipid identification and quantification, and additional data management for subsequent lipidome visualization and bioinformatics analysis.

Extensive lipidome characterization by shotgun lipidomics can be achieved by executing a systematic program of mass spectrometric analyses of sample extracts in positive and negative ion mode, and by incorporating chemical derivatization procedures to specifically monitor poorly ionizing lipid molecules such as cholesterol [96,99,202,212]. Executing such an analytical program generates several mass spectral data files per sample that must be queried for lipid identification, and combined into a single dataset for lipidome quantification and visualization. Numerous software tools have been developed for the identification of lipids: LipidQA[258], LIMSA[259], FAAT[260], lipID[261], LipidSearch[262], LipidView[263], LipidInspector[264] and LipidXplorer[265]. These tools cover a broad range of applications spanning dedicated lipid identification for only certain instrumentations and specific mass analysis routines (MS and MS/MS) to cross-platform software featuring user-specified commands querying spectral data in the open-source .mzXML format. Identified lipid species are typically annotated by a shorthand nomenclature corresponding to the information detail of the mass spectrometric analysis [266,267]. The detection of lipid species by FT MS analysis or by MS/MS analysis for lipid class-specific fragment ions (e.g. m/z 184.0733 for phosphatidylcholine (PC) species) supports only "sum composition" annotation (e.g. PC 34:1). In comparison, annotation by the more detailed "molecular composition" (e.g. PC 16:0-18:1) requires MS/MS analysis and detection of structure-specific fragment ions [268]. To support the cataloging of lipid species, the LIPID MAPS Consortium recently developed the "Comprehensive Classification System for Lipids" which outlines an informatics framework for lipidomics [192,269]. Using a classification system enables the design of lipid databases where each lipid species is listed together with a range of accessory lipid features such as lipid category (e.g. glycerophospholipid, sphingolipid, glycerolipid, sterol lipid), lipid class, structural attributes (e.g. number of double bonds, fatty acid chain length), chemical formula, mono-isotopic mass and isotope information. These accessory lipid features can be incorporated into lipidomic data processing routines using database-orientated exploration tools to support computations and visualization of distinct lipidome hallmarks. Notably, none of the currently available software tools comprise streamlined processing routines that integrate lipid intensity data, the accessory lipid features and a full catalog of sample information.

Here we present a platform for processing, management and visualization of high-content shotgun lipidomics datasets acquired using high-resolution Orbitrap mass spectrometry. The platform features a novel software framework termed ALEX (Analysis of Lipid Experiments) that supports automated identification and export of lipid species intensity directly from proprietary mass spectral data files and the integration of accessory lipid features and sample information into a single output structured in “database table format”. This design supports swift data processing and lipidome visualization across large sample sizes using an auxiliary workflow powered by the database exploration tools: Orange [270] and Tableau Software. To demonstrate the efficacy of the platform, we present a comparative neurolipidomics analysis of cerebellum, hippocampus and S1BF from wild-type and knockout mice devoid of the PRG-1 gene encoding a putative lipid phosphate phosphatase [271].

5.4. Materials and Methods

5.4.1. Chemicals and lipid standards

Chemicals, solvents and synthetic lipid standards were purchased from Sigma-Aldrich, Rathburn Chemicals, Avanti Polar Lipids and Larodan Fine Chemicals AB.

5.4.2. Mouse brain tissue sampling

Animal experiments were conducted in strict accordance with German law (in congruence with 86/609/EEC) for the use of laboratory animals and approved by the local animal welfare committee at the Johannes Gutenberg University Mainz. Male C57Bl/6 wild-type and PRG-1 knockout mice [271] were euthanized by an intraperitoneal injection of ketamine at an overdose. Subsequently, the mice were perfused intracardially with 4°C 155 mM ammonium acetate, and the cerebellum, hippocampus and S1BF were dissected. The tissues were immediately frozen on dry ice and stored at -80°C until further processing.

5.4.3. Lipid extraction

Brain tissues were homogenized in 155 mM ammonium acetate and analyzed for total protein concentration using BCA Protein Assay Kit (Thermo Scientific). Aliquots of tissue homogenates corresponding to 10 µg of total protein were subjected to lipid extraction executed at 4°C as previously described [212]. Briefly, the tissue homogenates were spiked with 10 µl of internal mixture (providing a total spike of 54 pmol CE 19:0, 35 pmol TAG 17:1/17:1/17:1, 35 pmol DAG 19:0/19:0, 26 pmol LPA O-16:0, 35 pmol PA 17:0/14:1, 25 pmol LPS 17:1, 13 pmol PS 17:0/20:4, 50 pmol PE O-20:0/O-20:0, 30 pmol LPC O-17:0, 137 pmol PC 18:3/18:3, 35 pmol PI 17:0/20:4, 30 pmol PG 17:0/17:0, 55 pmol Cer 18:1;2/17:0;0, 69 pmol SM 18:1;2/17:0;0, 49 pmol HexCer 18:1;2/12:0;0, 28 pmol SHexCer 18:1;2/12:0;0) and diluted to 200 µl using 155 mM ammonium acetate. Samples were subsequently added 990 µl chloroform/methanol (10:1, v/v) and vigorously mixed for 2 h. The lower organic phase was collected (10:1-phase lipid extract). The remaining aqueous phase was re-extracted with 990 µl of chloroform/methanol (2:1, v/v) for 1 h and the lower organic phase was collected (2:1-phase lipid extract). The collected lower organic phases were vacuum evaporated.

5.4.4. Shotgun lipidomics analysis

Lipid extracts were dissolved in 60 µl of chloroform/methanol (1:2, v/v) and subjected to mass spectrometric analysis using an LTQ Orbitrap XL instrument (Thermo Fisher Scientific) equipped with a TriVersa NanoMate (Advion Biosciences) as previously described [96,212]. The 10:1-phase lipid extracts were analyzed by positive ion mode multiplexed FT MS analysis with scan ranges *m/z* 280-580 (monitoring lysophosphatidylcholine (LPC) and lysophosphatidylethanolamine (LPE) species) and *m/z* 500-1200 (monitoring sphingomyelin (SM), ceramide (Cer), diacylglycerol (DAG), PC, ether-linked PC (PC O-), phosphatidylethanolamine (PE), ether-linked phosphatidylethanolamine (PE O-) and triacylglycerol (TAG) species). The 2:1-phase lipid extracts were analyzed by negative ion mode multiplexed FT MS analysis with scan ranges *m/z* 370-660 (monitoring lysophosphatidic acid (LPA), lysophosphatidylserine (LPS) and lysophosphatidylinositol (LPI) species) and *m/z*550-1700 (monitoring phosphatidic acid (PA), phosphatidylserine (PS), phosphatidylinositol (PI), phosphatidylglycerol (PG) and sulfatide (SHexCer) species). All FT MS spectra were acquired in profile mode using a target mass resolution of 100,000 (fwhm), activation of isolation waveforms, automatic gain control at 1e6, max injection time at 250 ms and acquisition of 2 µscans.

5.4.5. Annotation of lipid species

Glycerophospholipid and glycerolipid species were annotated using sum composition: <lipid class><total number of C in the fatty acid moieties>:<total number of double bonds in the fatty acid moieties> (e.g. PI 34:1). Sphingolipid species were annotated as <lipid class><total number of C in the long-chain base and fatty acid moiety>:< total number of double bonds in the long-chain base and fatty acid moiety>;< total number of OH groups in the long-chain base and fatty acid moiety> (e.g. SM 36:1;2) [99,212].

5.4.6. ALEX software

The individual parts of the ALEX software were programmed using several programming languages, libraries and software frameworks. The ALEX lipid database is implemented using the library based SQLite database engine. The ALEX target list generator is written in C++ and uses the Qt framework for its user interface. The ALEX converter, ALEX extractor and ALEX unifier are written in Python 2.7 and make use of the Python packages PySide, NumPy and SciPy. Furthermore, the ALEX converter uses the package comtypes to interface with the MSFileReader library version 2.2 (Thermo Scientific), which must be installed. Finally, the standalone ALEX lipid calculator is written in common lisp and uses the GTK+ framework for its user interface, while the online version is written in PHP. The ALEX software is available at http://mslipidomics.info/software along with installation instructions. A sample dataset is also available for testing local installations of the software.

5.4.7. Data processing and visualization

Computation of molar abundance (fmol) of lipid species [263] was performed using open source software Orange 2.6 (www.orange.biolab.si) [270]. The Orange workflow is provided as part of the sample dataset available at http://mslipidomics.info | software. In addition, the Orange output with lipidomics data is available as Data S1[1]. Visualization and calculation of mol% values were performed using commercially available Tableau Software (www.tableausoftware.com). Lipidomic data in Tableau file format is available as Data S2[2] and can be navigated using the freeware Tableau Reader (http://www.tableausoftware.com/products/reader).

5.5. Results and Discussion

5.5.1 Input: high-resolution shotgun lipidomics data

Shotgun lipidomics platforms based on high-resolution Orbitrap mass spectrometry and automated nanoelectrospray ionization support high throughput analysis with high sensitivity, specificity and extensive lipidome coverage [96,99,206,212]. The extensive lipidome coverage is generated by combined analyses of sample extracts in negative and positive ion mode, and by implementing chemical derivatization procedures to monitor low abundant or poorly ionizing lipid species [272–274]. In order to maximize the sensitivity of Orbitrap mass analysis, we

[1] http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0079736.s001

[2] http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0079736.s002

typically record two multiplexed FT MS scans covering a low *m/z* range and high *m/z* range (e.g. +FT MS *m/z* 280-580 and +FT MS *m/z* 500-1200). Each FT MS scan is recorded in profile mode with a target mass resolution of 100,000. The rationale for this multiplexed FT MS analysis is that the instrument injects a user-specified quantum of ions into the Orbitrap for mass analysis (defined by the automated gain control). Hence, multiplexing two or more scan ranges allows separate quanta of ions within specific *m/z* ranges to be analyzed sequentially in the Orbitrap and yields better ion statistics as compared to injecting all ions at the same time when monitoring a wider *m/z* range (e.g. +FT MS *m/z* 200-1200)[275]. We note that the boundaries of the scan ranges should be chosen to cover specific lipid classes and respective internal standards. For example, the scan range +FT MS *m/z* 280-580 is used for monitoring LPC and LPE species whereas +FT MS *m/z* 500-1200 analysis is used for monitoring Cer, SM, HexCer, PC, PC O-, PE, PE O-, DAG and TAG species. The total time of analysis is typically set to 3 min. Within this time we record 25 low *m/z* and 25 high *m/z* range FT MS spectra. Likewise, negative ion mode analysis is also executed using multiplexed FT MS acquisitions to monitor negatively charged glycerophospholipid and SHexCer species (see Material and Methods). Consequently, this lipidomics approach generates four distinct mass spectral datasets per sample (two per polarity) that need to be queried for lipid identification and export of lipid species intensity. We note that this lipidomics approach is designed for high throughput-oriented studies and supports annotation of lipid species by sum composition nomenclature (e.g. PC 34:1). Characterization of molecular lipid species (e.g. PC 16:1-18:0) requires implementation of time-consuming MS/MS analysis and lipid identification by dedicated software such as LipidXplorer [265].

5.5.2. Design of the ALEX software framework

The ALEX software framework was designed for processing of shotgun lipidomics datasets obtained by multiplexed high-resolution FT MS. The rationales for the design were: i) that ALEX should support lipidomic studies with large sample sets for which a multitude of multiplexed FT MS acquisitions have been acquired; and ii) that the output format of ALEX should be compatible with an auxiliary workflow that supports robust data processing including computation of molar abundances of lipid species across numerous lipid classes, integration of sample information, implementation of data quality control procedures and rapid lipidome visualization. To this end, we designed the ALEX software framework to utilize distinct modules that identify lipid species from proprietary .RAW spectral file format, incorporate accessory lipid features stored in a lipid database and output lipidomic data in "database table format" (Figure 5-1, Data S1[1]). In this format, the lipidomic data is stored in tabulator or comma separated text files structured as database tables with a separate row for each data point. Each row separately contains fields (also termed attributes) reporting for example the originating sample (.RAW file name), the lipid species, adduct information, intensity, peak area, *m/z* values and accessory lipid features derived from the lipid database.

[1] http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0079736.s001

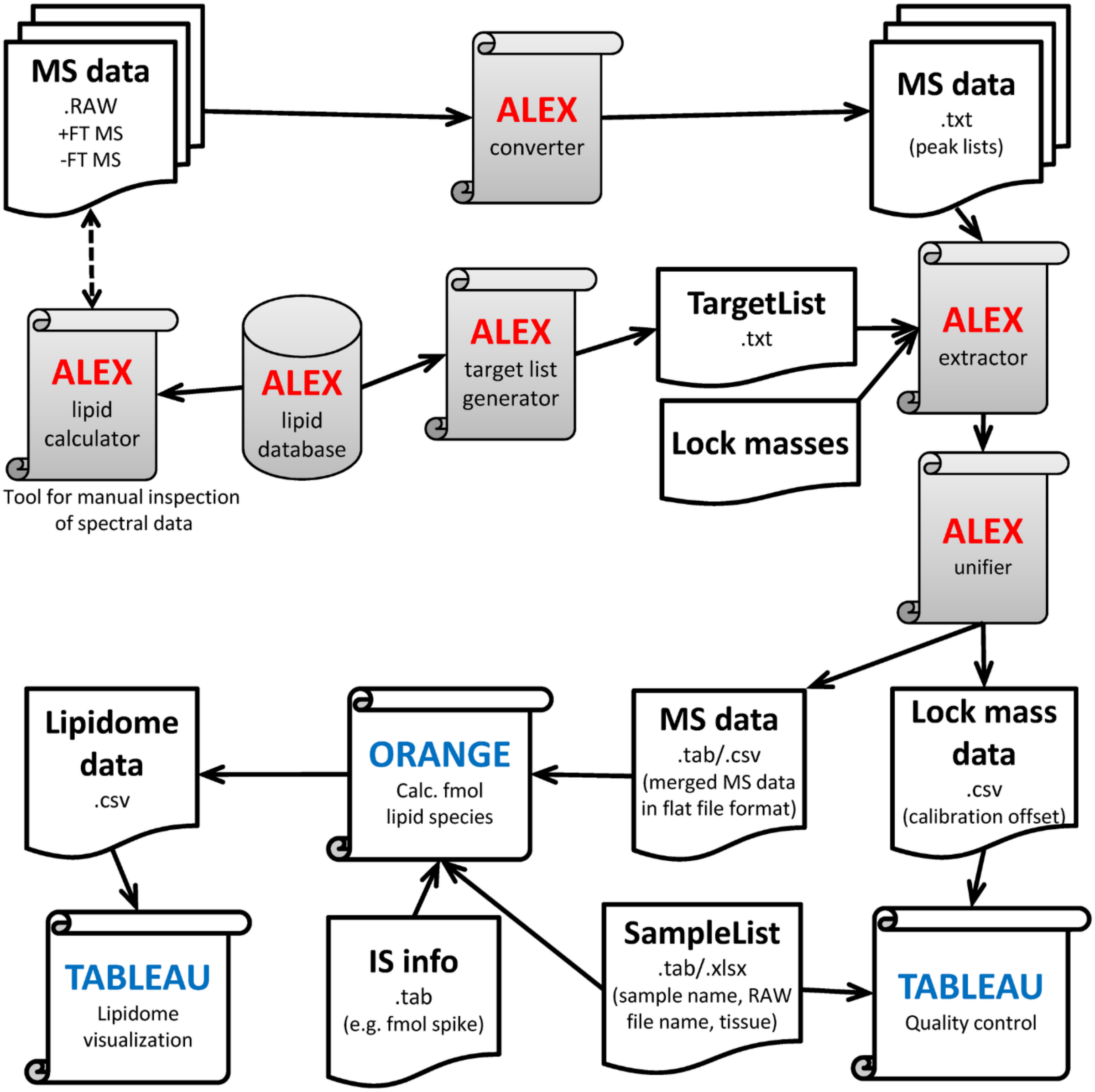


Figure 5-1. Figure 1. Overview of the ALEX software framework and auxiliary workflow.
The ALEX framework comprises six core modules (grey colored boxes). The function of each module is explained in the Results and Discussion section. The output of the ALEX framework includes a data file with identified lipid species, intensities and accessory lipid features across all processed samples and FT MS scan ranges. The ALEX output is organized in database table format that can be accessed and processed by the auxiliary workflow using Orange and Tableau software. The auxiliary workflow is designed to integrate sample information, compute lipid molar abundance, implement quality control procedures and visualize lipidome data.
doi:10.1371/journal.pone.0079736.g001

Importantly, this relational database format provides a robust way to manage large datasets programmatically and avoids the need for error-prone manual alignment of data with lists of accessory features and quantification information. We note that the output of most contemporary lipidomics software tools utilizes a “spreadsheet” format where samples/injections are arranged in the columns and lipid species in the rows of a table of with either intensities or peak areas. This “spreadsheet” format is adequate for processing and visualizing simple lipidomic datasets using tools such as Microsoft Excel, but becomes exceedingly cumbersome if sample sets include more than 10 samples and monitor more than 200 lipid species across different FT MS scan ranges.

The ALEX software framework consists of six core modules (Figure 1): i) the ALEX lipid database featuring a comprehensive range of lipid species and accessory lipid features used for lipid identification, data processing and management; ii) the ALEX lipid calculator which supports manual interpretation of mass spectra; iii) the ALEX converter which converts mass spectrometric data in proprietary .RAW format into averaged spectral peak lists in text file format; iv) the ALEX target list generator which queries the ALEX lipid database to compile target lists used for lipid species identification; v) the ALEX extractor which uses the target lists to identify lipid species and export corresponding intensities from spectral peak lists; and vi) the ALEX unifier that merges the multiple ALEX extractor outputs containing lipid species data from different FT MS scan ranges into a single data file. Further information about the function of the ALEX modules and the auxiliary workflow is outlined in the subsequent sections.

5.5.3. ALEX lipid database

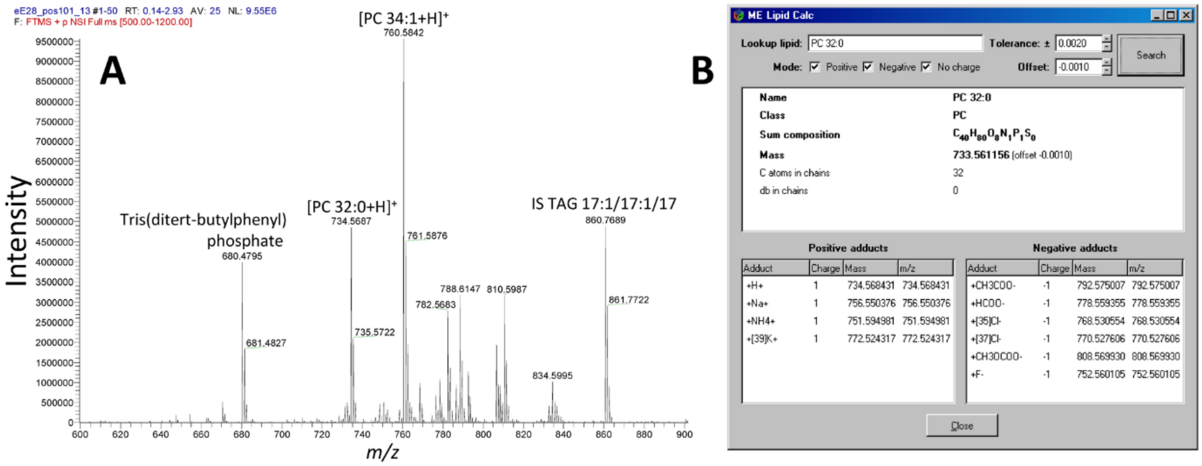
To support identification of lipid species we constructed the ALEX lipid database. Currently, the database covers more than 20,000 lipid species from more than 85 lipid classes. Each lipid species in the database is annotated by sum composition and contains a range of accessory lipid features denoting its chemical formula, mono-isotopic mass, adduction in positive and negative ion mode, lipid category, lipid class, and the total number of C atoms, double bonds and hydroxyl group in fatty acid and long chain base moieties (termed C index, db index and OH index, respectively). The ALEX lipid database is available as a collection of text files and as a binary SQLite database queried by the ALEX lipid calculator and the ALEX target list generator, respectively (outlined in the next sections). The text version also serves as the source for the binary version (compiled by a supporting Python program) and can be edited by the user. To support the editing of the database, a template Microsoft Excel file can be used to assist the enumeration and calculation of the chemical formula and mass for each species, such that catalogs of entire lipid classes can readily be added. We note that the accessory lipid features serve as important attributes in the final ALEX output that facilitate data processing and lipidome visualization.

5.5.4. ALEX lipid calculator

The ALEX lipid calculator was designed to complement manual inspection of FT MS spectra when using proprietary Xcalibur software (Figure 5-2). The calculator is available as executable program and as an online application at www.mslipidomics.info/lipid-calc. Both versions of the ALEX lipid calculator support querying *m/z* values of specific lipid species and searching the lipid database for candidate lipid species that match a measured *m/z* value within a user-specified tolerance window. Since the accuracy of lipid identification can be hampered by drifts of the FT MS calibration we also implemented an option to specify an *m/z* offset while querying the lipid database. To accurately specify the calibration offset requires that the user knows the identity of at least one well-characterized “lock mass” ion that can be used for estimating the*m/z* offset (Figure 5-2). An option to minimize potential problems with FT MS calibration drifts is to apply online lock mass calibration during sample acquisition [276]. However, the online lock mass calibration eliminates the lock mass ion(s) from the recorded FT MS spectra using waveforms that might concomitantly eliminate lipid ions having similar *m/z*. An alternative strategy is to apply an offline lock mass adjustment as implemented in the ALEX extractor (outlined below).

5.5.5. ALEX converter

A prerequisite for automated lipid identification and export of intensity is that the mass spectrometric data in proprietary .RAW format are made accessible for querying. To this end, we designed the ALEX converter to interface with the proprietary dynamic-link library MSFileReader. The MSFileReader supports export of all scan information within .RAW files including single and averaged spectral peak lists in either centroid or profile mode format. The ALEX converter was designed to export individual spectral peak lists in profile mode format, to average peak lists for specific FT MS scan ranges and to save these averaged peak lists in .txt format.



Lookup lipid: PC 32:0   Tolerance: ± 0.0020

Mode: ☑ Positive ☑ Negative ☑ No charge   Offset: -0.0010

Name	**PC 32:0**
Class	**PC**
Sum composition	**$C_{40}H_{80}O_8N_1P_1S_0$**
Mass	**733.561156** (offset -0.0010)
C atoms in chains	32
db in chains	0

Positive adducts

Adduct	Charge	Mass	m/z
+H+	1	734.568431	734.568431
+Na+	1	756.550376	756.550376
+NH4+	1	751.594981	751.594981
+[39]K+	1	772.524317	772.524317

Negative adducts

Adduct	Charge	Mass	m/z
+CH3COO-	-1	792.575007	792.575007
+HCOO-	-1	778.559355	778.559355
+[35]Cl-	-1	768.530554	768.530554
+[37]Cl-	-1	770.527606	770.527606
+CH3OCOO-	-1	808.569930	808.569930
+F-	-1	752.560105	752.560105

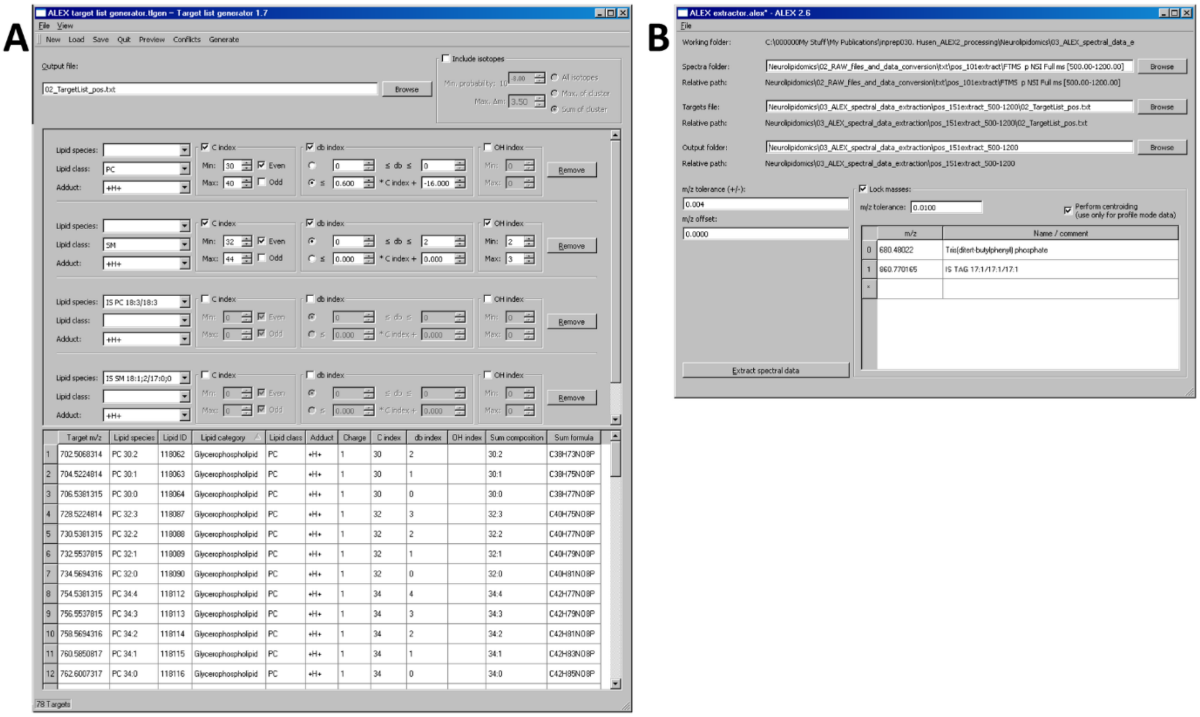
Figure 5-2. The ALEX lipid calculator.
(**A**) Representative positive ion mode FT MS spectrum of a 10:1-phase lipid extract of hippocampus from a PRG-1 knockout mouse. Note that the detection of selected lock mass ions tris(ditert-butylphenyl) phosphate (chemical background, $[M+NH_4]^+$, calculated *m/z* 680.48022, measured *m/z* 680.47945, *m/z* offset = -0.00077) and TAG 17:1/17:1/17:1 (internal standard (IS), $[M+NH_4]^+$, calculated *m/z* 860.77017, measured*m/z* 860.76889 and *m/z* offset = -0.00128). The FT MS calibration offset is estimated as the average of the *m/z* offset for both lock mass ions, i.e. the FT MS calibration offset = -0.0010. (**B**) Screenshot of the ALEX lipid calculator showing information for endogenous lipid species PC 32:0 while applying the FT MS calibration offset = -0.0010. Note that the measured *m/z* of PC 32:0 is 734.56872 and that the calculated *m/z* adjusted for the calibration offset is 734.56843 which yield a *m/z* difference of 0.00029 corresponding to a mass error of 0.4 ppm. Without applying lock mass adjustment the mass error would be 1 ppm.

doi:10.1371/journal.pone.0079736.g002

The ALEX converter output consists of a directory with separate folders named according to each FT MS scan range containing corresponding .txt files named according to the originating .RAW files (i.e. the ALEX converter does not merge overlapping FT MS scan ranges). Output files in these FT MS scan range-dependent folders are queried by the ALEX extractor (outlined below). The rationales for this design were i) that no available software supports export of multiplexed FT MS data in profile mode format; and ii) that a simple output text format reporting averaged peak list data allows users to easily review data as in contrast to accessing data in stored in the encrypted .mzXML format.

5.5.6. ALEX target list generator

Lipid identification by the ALEX software framework is based on matching intensity data in the exported peak lists with lipid species information derived from the ALEX lipid database. Important for the design of the identification routine was the ability to accurately identify lipid species, and furthermore to support integration of accessory lipid features and subsequent data management for a multitude of lipid species monitored by various FT MS scan ranges across large sample sets. To this end, the ALEX framework was designed to perform a targeted identification of lipid species and export intensity by querying spectral peak lists using target lists with lipid species and respective *m/z* values. Two modules execute this routine: (i) the ALEX target list generator (Figure 5-3A) which compiles target lists by querying the ALEX lipid database, and (ii) the ALEX extractor which uses the target lists to identify and extract lipid species intensity (outlined in the next section). We note that a distinct target list with appropriate lipid species should be manually compiled for each FT MS scan range.


A
B
ALEX target list generator.tlgen – Target list generator 1.7
File View
New Load Save Quit Preview Conflicts Generate
Browse
Remove
ALEX extractor.alex* - ALEX 2.6
File
Browse
Extract spectral data

Output file: 02_TargetList_pos.txt

□ Include isotopes
Min. probability: 10 -8.00 ○ All isotopes
Max. Δm: 3.50 ○ Max. of cluster
● Sum of cluster

Lipid species:
Lipid class: PC
Adduct: +H+
☑ C index Min: 30 ☑ Even Max: 40 □ Odd
☑ db index ○ 0 ≤ db ≤ 0 ● ≤ 0.600 * C index + -16.000
□ OH index Min: 0 Max: 0

Lipid species:
Lipid class: SM
Adduct: +H+
☑ C index Min: 32 ☑ Even Max: 44 □ Odd
☑ db index ● 0 ≤ db ≤ 2 ○ ≤ 0.000 * C index + 0.000
☑ OH index Min: 2 Max: 3

Lipid species: IS PC 18:3/18:3
Lipid class:
Adduct: +H+
□ C index Min: 0 ☑ Even Max: 0 ☑ Odd
□ db index ● 0 ≤ db ≤ 0 ○ ≤ 0.000 * C index + 0.000
□ OH index Min: 0 Max: 0

Lipid species: IS SM 18:1;2/17:0;0
Lipid class:
Adduct: +H+
□ C index Min: 0 ☑ Even Max: 0 ☑ Odd
□ db index ● 0 ≤ db ≤ 0 ○ ≤ 0.000 * C index + 0.000
□ OH index Min: 0 Max: 0

	Target m/z	Lipid species	Lipid ID	Lipid category	Lipid class	Adduct	Charge	C index	db index	OH index	Sum composition	Sum formula
1	702.5068314	PC 30:2	118062	Glycerophospholipid	PC	+H+	1	30	2		30:2	C38H73NO8P
2	704.5224814	PC 30:1	118063	Glycerophospholipid	PC	+H+	1	30	1		30:1	C38H75NO8P
3	706.5381315	PC 30:0	118064	Glycerophospholipid	PC	+H+	1	30	0		30:0	C38H77NO8P
4	728.5224814	PC 32:3	118087	Glycerophospholipid	PC	+H+	1	32	3		32:3	C40H75NO8P
5	730.5381315	PC 32:2	118088	Glycerophospholipid	PC	+H+	1	32	2		32:2	C40H77NO8P
6	732.5537815	PC 32:1	118089	Glycerophospholipid	PC	+H+	1	32	1		32:1	C40H79NO8P
7	734.5694316	PC 32:0	118090	Glycerophospholipid	PC	+H+	1	32	0		32:0	C40H81NO8P
8	754.5381315	PC 34:4	118112	Glycerophospholipid	PC	+H+	1	34	4		34:4	C42H77NO8P
9	756.5537815	PC 34:3	118113	Glycerophospholipid	PC	+H+	1	34	3		34:3	C42H79NO8P
10	758.5694316	PC 34:2	118114	Glycerophospholipid	PC	+H+	1	34	2		34:2	C42H81NO8P
11	760.5850817	PC 34:1	118115	Glycerophospholipid	PC	+H+	1	34	1		34:1	C42H83NO8P
12	762.6007317	PC 34:0	118116	Glycerophospholipid	PC	+H+	1	34	0		34:0	C42H85NO8P

78 Targets

Working folder: C:\000000My Stuff\My Publications\inprep030. Husen_ALEX2_processing\Neurolipidomics\03_ALEX_spectral_data_e

Spectra folder: Neurolipidomics\02_RAW_files_and_data_conversion\txt\pos_101extract\FTMS p NSI Full ms [500.00-1200.00]
Relative path: Neurolipidomics\02_RAW_files_and_data_conversion\txt\pos_101extract\FTMS p NSI Full ms [500.00-1200.00]

Targets file: Neurolipidomics\03_ALEX_spectral_data_extraction\pos_151extract_500-1200\02_TargetList_pos.txt
Relative path: Neurolipidomics\03_ALEX_spectral_data_extraction\pos_151extract_500-1200\02_TargetList_pos.txt

Output folder: Neurolipidomics\03_ALEX_spectral_data_extraction\pos_151extract_500-1200
Relative path: Neurolipidomics\03_ALEX_spectral_data_extraction\pos_151extract_500-1200

m/z tolerance (+/-): 0.004
m/z offset: 0.0000

☑ Lock masses:
m/z tolerance: 0.0100
☑ Perform centroiding (use only for profile mode data)

	m/z	Name / comment
0	680.48022	Tris(ditert-butylphenyl) phosphate
1	860.770165	IS TAG 17:1/17:1/17:1
*		

Figure 5-3. Screenshots of the ALEX target list generator and ALEX extractor.
(**A**) The ALEX target list generator allows users to select lipid classes and species to be identified using criteria such as lipid class, adduction, C index, db index and OH index. Individual lipid species including internal standards can also be selected. The ALEX target list generator output is a .txt file with a shortlist of selected lipid species, respective *m/z* values and accessory lipid features. The ALEX target list generator also supports inclusion of isotope information that can be used for deisotoping and isotope correction [20] by applying algorithms within the auxiliary workflow. (**B**) The ALEX extractor identifies lipid species, exports intensity data and incorporates accessory lipid features. As input the ALEX extractor requires the location of spectral peak lists generated by the ALEX converter, a target list compiled by the ALEX target list generator and a location to deposit output files. The ALEX extractor features options to specify an*m/z* tolerance window for lipid identification, to apply a constant *m/z* offset to correct lipid searches for a constant FT MS calibration offset or to apply a lock mass adjustment routine that automatically corrects lipid searches for drifts in FT MS calibration. The automated lock mass adjustment routine requires specification of well-characterized and ubiquitous lock mass ions in order to estimate the FT MS calibration offset.

doi:10.1371/journal.pone.0079736.g003

In addition to listing lipid species and respective *m/z* values, the target lists also include the accessory lipid features derived from the ALEX lipid database. Importantly, these accessory lipid features are incorporated into the final output, and used for processing and visualization by the auxiliary workflow.

5.5.7. ALEX extractor and ALEX unifier

The ALEX extractor identifies lipid species and exports intensities by querying the averaged peak lists produced by the ALEX converter (Figure 5-3B). As input the ALEX extractor requires the folder location containing averaged peak lists, an appropriate target list (i.e. specific for the FT MS scan range) to query the peak lists and a destination folder to deposit output text files. Notably, the ALEX extractor also requires an *m/z* tolerance window to identify lipid species. This *m/z* tolerance window is dependent on instrumental mass resolution and typically set to ±0.0020 amu when processing FT MS data acquired with a target resolution at 100,000. To export lipid species intensity the ALEX extractor selects the maximum intensity value within the specified tolerance window and reports the corresponding *m/z* bin value. As mentioned above, the FT MS calibration can drift during the sample analysis and depending on the time of analysis can yield either a constant calibration offset or a progressively changing offset (see Figure 5-4A). In order to monitor and adjust lipid searches for potential calibration drifts, we incorporated a novel feature in the ALEX extractor termed “lock mass adjustment”. Lock mass adjustment serves to correct the *m/z* values of targeted lipid species for calibration drifts and thereby support more accurate lipid identification. This lock mass adjustment can be specified as a constant *m/z* offset and applied across all samples being processed. Alternatively, the ALEX extractor features an in-build automatic lock mass adjustment routine that calculates *m/z* calibration offsets for each sample based on selected lock mass ions (Figure 5-4A). The calculation of calibration offset uses a three-point quadratic interpolation for estimating the centroid *m/z* values of lock mass ions. We note that the automatic lock mass adjustment requires selection of well-characterized and ubiquitous ions as lock masses. For example, by positive ion mode FT MS analysis we always detect both the chemical background ion tris(ditert-butylphenyl) phosphate and the internal standard TAG 17:1/17:1/17:1 (Figure 5-2). Using these ions as lock masses enables estimation of the FT MS calibration offset for individual samples and correcting lipid searches by adjusting the *m/z* values of targeted lipid species for each sample. As exemplified in Figure 5-2, using these lock mass ions allows identification of endogenous lipid species with a mass error of 0.4 ppm instead of 1 ppm when ignoring the FT MS calibration offset. Hence, this automated lock mass adjustment routine serves to improve the accuracy of lipid species identification despite drifts in FT MS calibration.

The ALEX extractor outputs, for each FT MS scan range, several comma-separated value (.csv) files with lipid species intensity data and calculated lock mass adjustments for all processed samples. Notably, the lipid species intensity output is organized in database table format and includes attributes that track the originating .RAW file name, lipid species intensity and peak area, measured *m/z*, calculated *m/z*, the difference between measured and calculated *m/z*, and all the accessory lipid features included on the target list. We note that these sample attributes facilitate subsequent processing and visualization, and implementation of quality control procedures. In order to merge lipid species data from different FT MS ranges, we devised the ALEX unifier to concatenate selected .csv output files into one final .csv file which contains the union of all data and an additional column with an index, “rangeID”, that tracks the FT MS scan range of the input .csv files. This output format supports further data processing including computation of lipid abundance by the auxiliary workflow using open-source Orange [270] and Tableau Software.

5.5.8. Application of the ALEX software framework

In order to demonstrate the efficacy of the ALEX software framework and describe the auxiliary workflow, we here present a neurolipidomic pilot study. Three brain tissues; cerebellum, hippocampus and S1BF from two control mice and two mice devoid of the PRG-1 gene were subjected to shotgun lipidomics analysis. Homogenates of the tissues (12 samples in total) and two blank samples were spiked with defined amounts of internal lipid standards and subjected to 2-step lipid extraction [212]. The apolar (10:1-) and polar (2:1-phase) lipid extracts were analyzed by multiplexed FT MS analysis in positive and negative ion mode, respectively. Each sample extract was analyzed twice (technical replicates). In total, this analysis produced 56 .RAW files (14 samples analyzed twice per polarity).

First, the ALEX converter was used to convert .RAW files to spectral peak lists (Figure 5-1). This processing produced a total of 112 peak list files (56 .RAW files, two FT MS scan ranges per polarity). The peak list files were automatically organized into 4 folders named according to polarity and scan range: +FTMS *m/z* 280-580, +FTMS *m/z* 500-1200, -FTMS *m/z* 370-660, and -FTMS *m/z* 550-1700 (each folder having 28 peak list files).

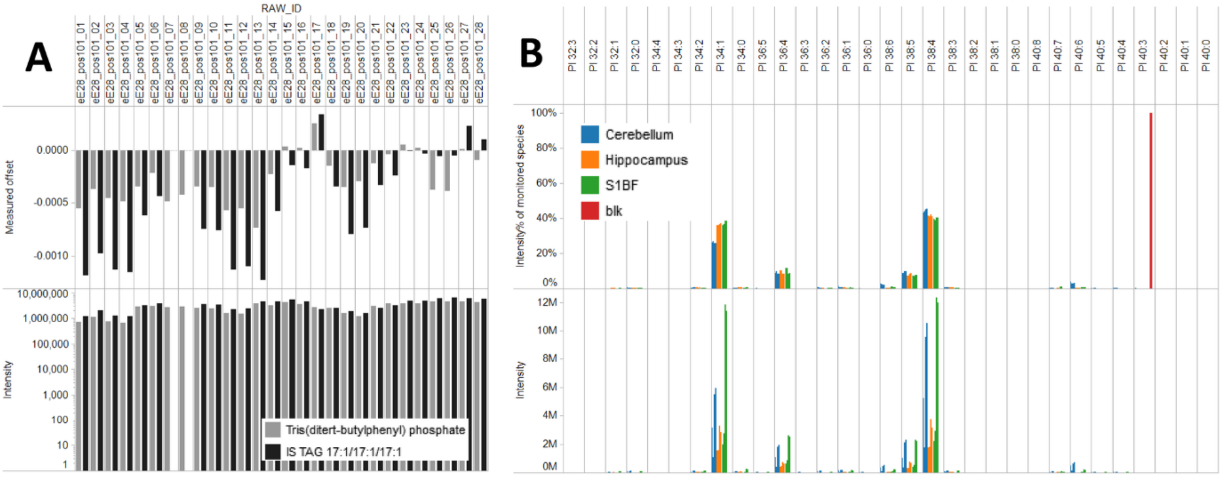


Figure 5-4. Quality control analysis.

(**A**) Monitoring of lock mass offset and lock mass ion intensity as function of sample injection. Notice that the lock mass and internal standard TAG 17:1/17:1/17:1 is not detected in injection 07 and 08. Manual inspection of FT MS spectra revealed that the particular sample had not been spiked with internal standards. (**B**) Assessing the specificity of the PI species profile and intensity across all samples from wild-type mice and the negative control blank samples. Note that in the negative control blank sample (red) a low abundant background ion is detected and falsely identified as PI 40:3. Dubious lipid species can be removed using background subtraction and filtering during subsequent processing in Orange.

doi:10.1371/journal.pone.0079736.g004

Next, target lists with lipid species for each of the four FT MS *m/z* ranges were generated using the ALEX target list generator (See Materials and Methods for details). These target lists were subsequently used by the ALEX extractor to identify lipid species and export corresponding intensities from the peak list files. In addition, the processing by the ALEX extractor was performed using lock mass ions for each FT MS *m/z* range in order to monitor and correct searches for FT MS calibration drifts (Figure 5-4A). Finally, the ALEX unifier was applied to merge the four ALEX extractor output files into single output files reporting all identified lipid species, intensities and accessory lipid features, and lock mass information and calculated FT MS calibration offsets.

As a first step in the auxiliary workflow we performed a quality control of the neurolipidomics dataset (Figure 5-1). To this end, we accessed the lock mass information using Tableau Software and displayed the estimated FT MS calibration offsets as function of sample injection. This quality control showed that the calibration offset was not constant across all samples (Figure 5-4) and thus highlighting the efficacy of the automatic lock mass adjustment. Importantly, for one sample we observed no intensity of the selected lock mass ion TAG 17:1/17:1/17:1. By manual inspection of the .RAW data, we concluded that the investigator in charge had failed to spike internal lipid standards into the particular sample. Consequently, this quality control demonstrated that the particular sample (cerebellum from a knockout mouse) could not be used for computing the molar amount of lipid species. As an additional quality control procedure we also accessed the output file with lipid species intensity data using Tableau Software, and displayed both the absolute intensity and intensity profile of monitored lipid species within lipid classes for all samples analyzed (Figure 5-4B). This analysis showed that a low abundant ion in the blank samples was falsely identified as PI 40:3. Due to the low intensity of this ion and its presence in the blank samples we concluded that this ion represents a chemical background ion. To minimize bias from falsely identified background ions one can implement a background subtraction during the subsequent computation of molar lipid abundance using the Orange software. Moreover, it is recommendable to perform additional tandem mass analysis to assess the identity of dubious identification.

5.5.9. Outlining the auxiliary workflow

In order to compute molar abundance (e.g. fmol) of lipid species we made use of the database exploration tool and open-source software Orange [270]. The molar abundance of lipid species is easily computed via a sequence of processing steps (depicted in Figure 5-5). Step (1); the lipid species intensity data generated by the ALEX framework is specified as input (icon (a)) and merged with a second input specifying sample information (e.g. tissue type, genetic information, name of mice; icon (b)). Step (2); an intensity filter is implemented in order to remove intensities below a user-specified threshold if deemed necessary. Step (3); a third input specifying the spiked amount of internal standards is incorporated (icon (c)). Step (4); a new attribute is defined by specifying the intensities of internal standards. This attribute is used in step (6) for computing the molar abundance. Step (5); internal standards and corresponding intensities are defined for the subsequent calculation of molar lipid abundances. Step (6); an equation using the attributes lipid species intensity, internal standard intensity and spiked amount of internal standard is applied for calculating the molar lipid abundance [263]. Step (7);

processed data is saved as a .csv output file (in database table format) that can be accessed for computation of mol% and lipidome visualization by Tableau Software. We here note that the Orange schema produced an output file of the neurolipidomics analysis featuring 997 targeted lipid species with accessory information across 56 sample injections producing a data matrix with a total of 1,050,504 data points (Data S1[1]). Notably, managing and processing a dataset of this magnitude is poorly suited for Microsoft Excel and highlights the benefits of managing the data using a database exploration tools.

To support rapid and efficient visualization of large lipidome datasets, we integrated Tableau Software as part of the auxiliary workflow. Tableau Software can be dynamically linked to the Orange output files such that any potential modifications within the data processing procedure can be visualized simply by updating the Orange output file and the link to Tableau. Notably, the Tableau software includes a feature that easily allows calculation and display of "mol% of lipid species" normalized to any given set of attributes and accessory lipid features in the input file. As such, the user can rapidly display the "mol% of all monitored PE species" (Figure 5-6C) or "mol% of all monitored glycerophospholipid species" (Figure 5-6D). We note, that calculation of such data formats would require implementation of additional processing steps within Orange (and equally in Microsoft Excel) in order to calculate and output such data values.

[1] http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pone.0079736.s001

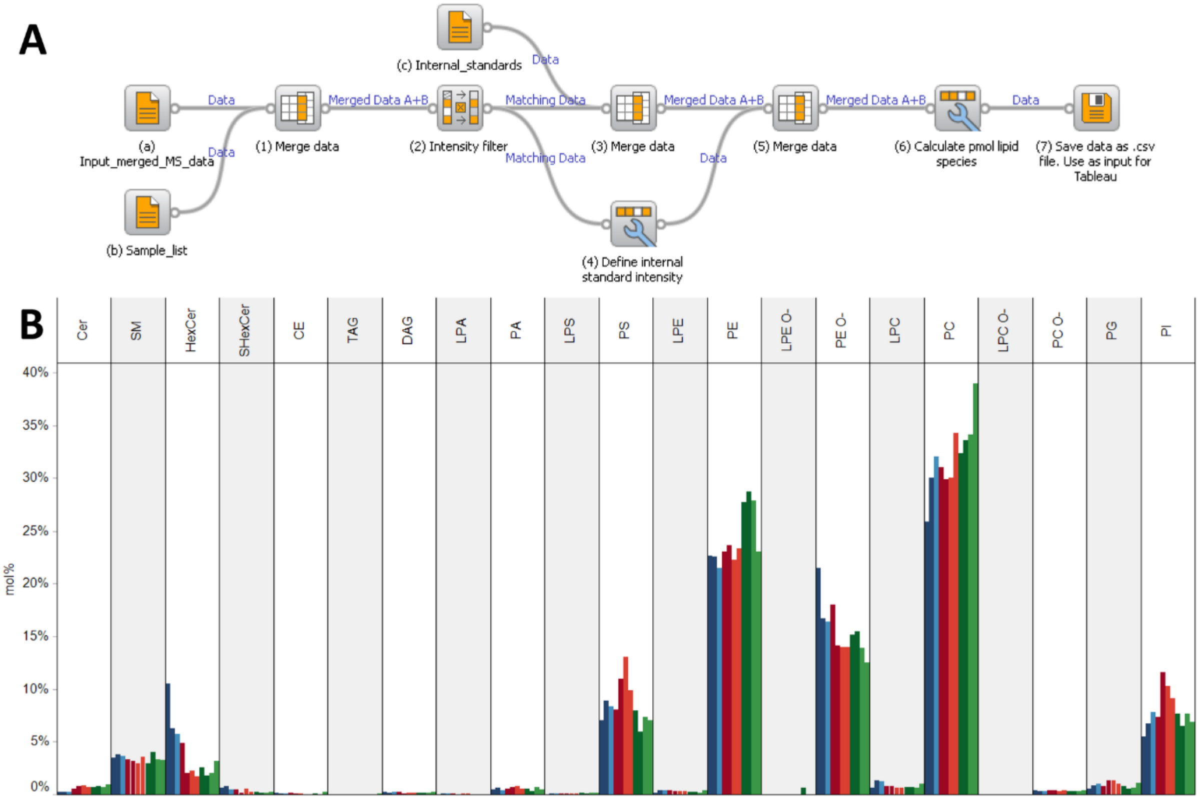


Figure 5-5. Outline of the auxiliary workflow used for lipidome data processing.
(A) A sequence of processing steps executed by the Orange software is used to compute the molar abundance of lipid species. The processing routine utilizes three input files: (a) the output file generated by the ALEX framework specifying lipid intensities, (b) a text file specifying sample information, and (c) a text file specifying internal standards and molar spike amounts. Seven processing steps are executed in order to compute the molar abundances of lipid species for all samples. The data processing generates an output file in database table format featuring the molar abundances of lipid species, originating intensity data, all accessory lipid features and all sample information. (B) Lipid class composition of cerebellum, hippocampus and S1BF from wild-type and PRG-1 knockout mice as automatically calculated and displayed using Tableau Software.
doi:10.1371/journal.pone.0079736.g005

We also note that these accessory lipid features provided by the ALEX framework, and consequently the information content in database table format, facilitates lipidome visualization by Tableau software.

5.5.10. Application of the auxiliary workflow

To illustrate the efficacy of the auxiliary workflow we here present the results of the neurolipidomic pilot study. First, we assessed the lipid class composition of the three mouse brain tissues from wild-type and PRG-1 knockout mice (Figure 5-5B). The lipid class composition of brain tissues was primarily comprised of PC, PE, PE O-, PS, PI, SM and HexCer lipids. In addition, the analysis also showed a low abundance of Cer, SHexCer, CE and TAG species (Figure 5-5B). This result corroborates previous reports on brain lipid composition [277–279]. Notably, comparing the lipid class composition of the brain tissues did not show any pronounced differences between the wild-type and PRG-1 knockout mice. Moreover, the analysis also did not show any major difference in lipid class composition between the three brain tissues. To further interrogate the lipidome data, we explored alternative display formats including lipid category composition (e.g. glycerophospholipids, sphingolipids and glycerolipids, Figure 5-6A), lipid species composition within a defined lipid class (e.g. PE species, Figure 5-6C), lipid species composition across a defined lipid category (e.g. all glycerophospholipids, Figure 5-6D) and db index within a defined lipid class (e.g. LPS, Figure 5-6B). Using the various modes of lipidome visualization, we observed that each of the three mouse brain tissues featured specific signatures of lipid species. Specifically, we observed that the composition of PE species was different between all three tissues; the hippocampus comprised a relatively high level of PE 38:4, the S1BF contained relatively high levels of PE 40:6, and the cerebellum contained a relatively high level of PE 40:6 and minor but systematically higher levels of PE 34:1, PE 36:1 and PE 36:2 as compared to the two other tissues. These distinct lipidome hallmarks could also be observed when assessing the collective lipid species composition across all glycerophospholipids (Figure 5-6D) albeit with a less pronounced differences as compared to only PE species. The specific lipid compositions of the three brain tissues were further interrogated by visualizing the distribution of double bonds within the fatty acid moieties of LPS species.

This visualization revealed that the S1BF comprised relatively high levels of LPS species with 0 double bonds (primarily attributed LPS 18:0) being offset by lower levels of LPS species having 6 double bonds (attributed LPS 22:6). In comparison, the cerebellum showed a characteristic distribution of LPS species with higher levels of species with 1 double bond as compared to the two other tissues. The LPS composition of the hippocampus comprised a double bond distribution intermediate of the S1BF and the cerebellum. We note that the lipidome visualization did not reveal any pronounced differences in lipid species composition between any of the tissues from wild-type and PRG-1 knockout mice. This observation is furthermore substantiated by results of principal component analysis (data not shown).

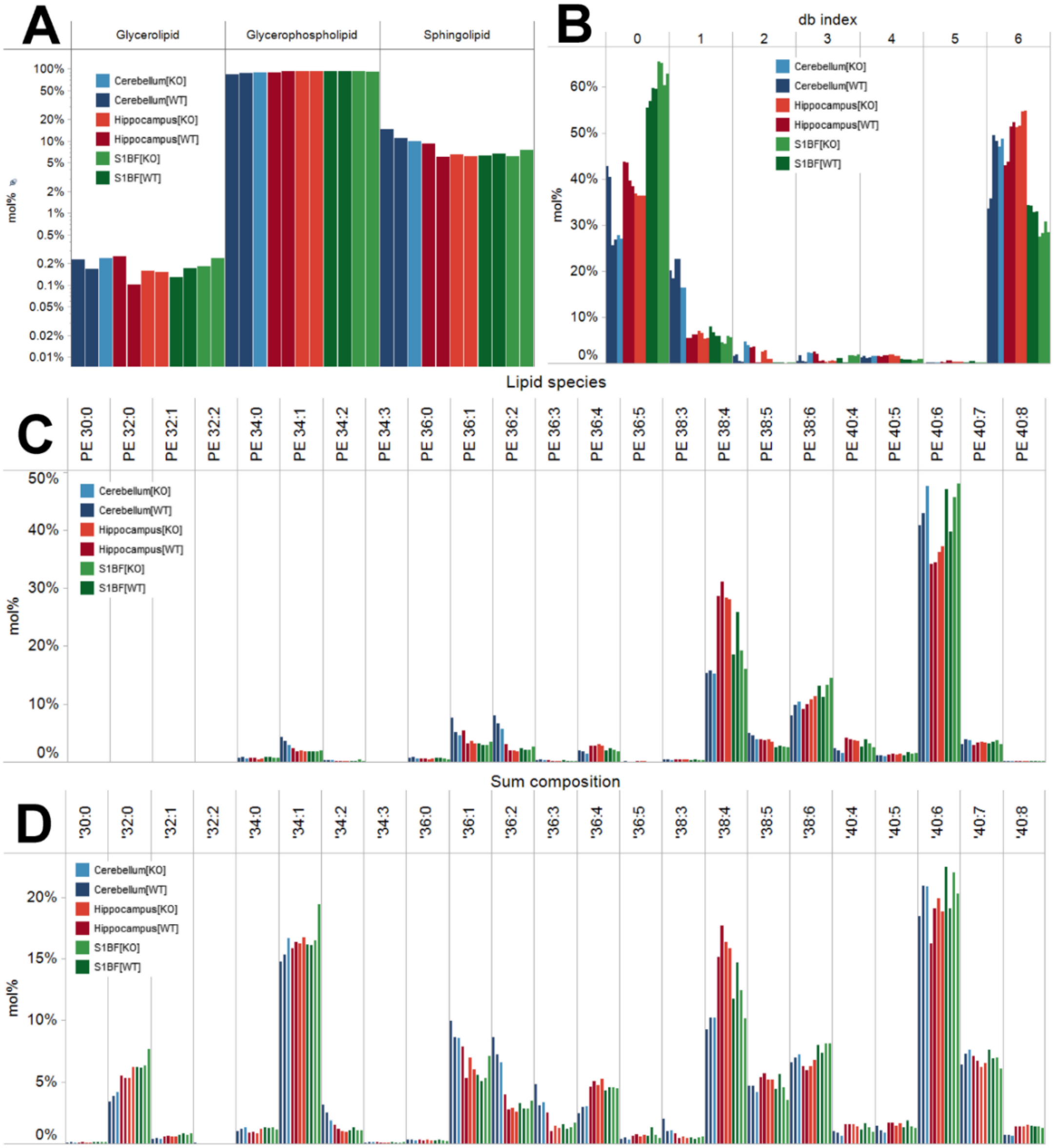


Figure 5-6. Lipidome visualization using different display formats.
(**A**) mol% lipid category. Notice that the y-axis is logarithmic. Data is displayed as the average of the two technical replicates per sample. (**B**) mol% of db index of LPS species. Note that histogram include plot for both technical replicates. (**C**) mol% of PE species. Data is displayed as the average of the two technical replicates per sample. (**D**) mol% of all GPL species. Data is displayed as the average of the two technical replicates per sample. Notice that data for only one sample of cerebellum from knockout mice is available due to lack of spiked internal standards as outline in the section "Application of the ALEX software framework". This neurolipidomics dataset is available as supporting information (Data S2).
doi:10.1371/journal.pone.0079736.g006

5.6. Conclusions

Here we presented a platform for streamlined processing, computation and visualization of high-content lipidomics datasets acquired using high-resolution Orbitrap mass spectrometry. The platform features a novel routine for querying proprietary spectral data that supports identification and quantification of lipid species, execution of quality control routines and rapid visualization of lipidome data. The platform utilizes three software modules: the ALEX framework that accesses and queries mass spectral data, the visual programming tool Orange that integrates sample information and computes molar lipid abundances, and the visual analytical tool Tableau Software for lipidome visualization. A key asset of the framework is the storage of lipidome data in “database table format” that enables using a multitude of data attributes for robust data processing and visualization. To demonstrate the efficacy of the platform, we presented a comparative neurolipidomic pilot study of mouse cerebellum, hippocampus and S1BF from wild-type and PRG-1 knockout mice [271]. The analysis demonstrated a distinct lipid species signature for each of the three brain tissues, but failed to ascertain any pronounced perturbations of lipid composition in mice devoid of the PRG-1 gene. This observation can potentially be explained by the localized expression of PRG-1 in the postsynaptic density of glutamatergic synapses. Regional differences in the lipidome composition induced by PRG-1 deficiency might potentially be concealed by the complexity of the macroscopic brain tissues investigated herein. We note that the ALEX framework can easily be adapted for processing of high-resolution shotgun lipidomics data acquired by any type of instrumentation (e.g. LTQ FT, Q Exactive, Orbitrap Fusion, solariX) provided spectral peak lists are stored in .txt file format. Moreover, the Orange processing procedure and Tableau visualization can be extended to include various filters for improved lipid identification, to calculate molar abundance of lipid species per unit of sample material (e.g. pmol lipid/µg protein), to perform statistical testing or multivariate analysis, and to integrate and support processing and visualization of lipidome data acquired by MS/MS analysis. Notably, the auxiliary workflow can also be adapted to use other database-orientated exploration tools than Orange and Tableau software. Finally, we argue that storage of lipidomics data in database table format can be a future avenue for data dissemination since it enables investigators to easily access, inspect and apply such resource data.

5.7. Acknowledgments

We are grateful to Dr. Ole N. Jensen for providing access to Triversa NanoMate. We thank Dr. Albert Casanovas and Reinaldo Almeida for critical reading of the manuscript. We thank Hans Kristian Hannibal-Bach and Nikolai Schmarowski for technical assistance.

5.8. Author Contributions

Conceived and designed the experiments: PH KT KE CSE. Performed the experiments: PH KT ES CSE. Analyzed the data: PH KT CSE. Contributed reagents/materials/analysis tools: PH KT MK JV JB RN. Wrote the manuscript: CSE.

Chapter 6 : Discussion and Conclusion

In the concluding chapter, we summarize the results of the study and propose directions for future research. We start the discussion with a focus on three main components of the work: protein-protein interaction network, high-throughput lipidomics and data visualization. Next, we discuss how the presented technologies combined together into a unified framework contribute to the functional prediction strategies. Finally, we formulate biological questions that can be addressed in future studies.

6.1. Contribution of DHFR PCA screen to the yeast interactome mapping

In a review article about pioneering genomics studies, Patrick Brown and David Botstein wrote: "Exploring the genome and the natural world with DNA microarrays Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew or expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible. We should use the unprecedented experimental opportunities that the genome sequences provide to take a fresh, comprehensive and open- minded look at every question in biology. If we succeed, we can expect that many of the new models that emerge will defy conventional wisdom."[16]

This view can be generalized to all types of genome-wide studies. Therefore, the systematic screen for protein-protein interactions was motivated by the aim to increase the coverage of the yeast interactome and expose certain types of interactions better than previous approaches. We hypothesized that apart from bringing new insights on network organization, novel data would provide the means to assign functions for uncharacterized genes.

The success of our study can be measured by the fact that DHFR PCA screen extended the repertoire of the yeast protein-protein interactions by 2770 interactions detected with high confidence, 80% of which were novel. The detected number of interactions is higher comparing to the range between 843 and 1985 filtered interactions reported by previous systematic binary interactions screens [57,58,63,66]. Combined TAP-MS protein-protein interaction network [51] contains 9 074 interactions. However, as mentioned above, TAP-MS based methods assign interactions between co-completed proteins that do not necessary interact directly leading to higher number of reported interactions [51] comparing to binary methods.

Intriguingly, most of interactions detected by each screening attempt were novel and reported error rates were low, indicating that noise is not the reason for the limited overlap between the datasets. It was argued that presented quality evaluation methods may be fitted to highlight advantages of the particular techniques challenging a uniform comparison of the error rates [66,280]. However, a general evaluation method that is based on comparison of accuracy of gene function predictions performed on each large-scale dataset confirms comparable high quality of the interaction networks [281]. Therefore, it is unlikely that limited overlap between protein-protein interaction datasets is solely due to experimental errors. Instead, some techniques cover particular network subspaces of interactions better than others do, which is caused by fundamental differences between techniques for protein-protein interaction detection techniques. It was demonstrated, for example, that Y2H and TAP-MS screens detected a higher proportion of interactions happening in the nucleus, but contained only limited information about membrane interactions [280,282].

The advantage of the DHFR PCA strategy is better coverage of interactions that involve membrane proteins comparing to systematic Y2H and TAP-MS screens. At the same time, DHFR PCA screen is different from screens developed specifically for capturing membrane binary protein interactions by Y2H [63] and membrane protein complexes by mass spectrometry [67]. DHFR PCA detects interactions between different types of proteins not limited to membrane interactions. Therefore, our network contains longer paths of connected proteins that link membranes with various cellular compartments detected in the same experiment. Analysis of such paths will be useful for modeling of signal transduction pathways that pass signals from membrane to nucleus [283][284]. The unique coverage of interactions between transmembrane proteins provides new data to study membrane related processes, such as cellular transport and lipid metabolism. Finally, over 300 interactions that were observed included uncharacterized proteins providing new data for computation function prediction.

Together with other recent screens for protein interactions, our study provided a significant contribution to mapping the yeast interactome. Combined analysis of the datasets estimated that the total number of protein interactions in yeast is higher than it was thought when we started our study. The estimation performed in 2003 reported the range of 16 000 to 26 000 interactions [62], while the analysis made in 2010 predicted that the number would be at least

36 000 [285]. Therefore, the mapping of the protein network in yeast has not been completed yet.

6.2. Normalized evaluation of the overlap between interaction datasets.

Early evaluations of the protein-protein interaction datasets raised concerns about a small overlap between them, which could be due to high false-positive and false-negative rates associated with protein-protein interaction detection by high-throughput technologies [47,286]. Previous evaluations scored datasets based on the coverage of the whole interactome. However, none of the reported methods tested all possible interactions. Each protein-protein interaction method required mutant libraries generated by genetic manipulations the efficiency of which is not constant through the genome leading to failures in tagging certain genes. Furthermore, engineered protein fusions might have an impact on the protein function and strain viability. Consequently, each method tested only a portion of all possible interactions.

To consider these factors, we devised a novel method for evaluating the overlap between interactions normalized by interactome coverage (section 2.5.1. Overlap with previous studies). When the number of potential protein-protein interactions that were actually tested by various methods was taken into account, the percentages of confirmed interactions were higher than previously reported. Results of this analysis increase credibility of the data generated by high-throughput screens.

6.3. High-throughput lipidomics

Lipidomics platform presented in Chapter 3 is one of the first attempts towards high-throughput studies of hundreds of metabolites in yeast. Comparing to the previous lipidomics studies our screen covered a larger number of mutants (128 strains passed the stringent quality control procedures). Our screen was not as comprehensive as the genome-wide profiling of amino acids [84], but measured more molecule in each strain (120 lipid species on average). For reducing the costs and the running time of the experiments, we introduced a first round screen for relative semi-quantitative lipid measurements without addition of internal standards. Hits identified in the first round were further characterized by a second absolute-quantification round. In contrast to the previous two-round procedures that compared raw mass spectra in the first round [94], our method in the first round resolved lipid identities and provided semi-quantitative values of lipid species abundances. Results of the first round screen provide the largest published collection of lipid profiles available so far in yeast.

There are some differences in growth rates of the yeast deletion mutants that has to be taken into account during experiments and data analysis. Good control over growth rates is achieved using chemostat cultivation [243]. However, the costly equipment is not easily accessible. Therefore, it is common to grow all strains for the fixed time in the metabolomics experiments for increasing the throughput. To deal with growth differences, we relied on multivariate statistics. Strains were first classified into groups corresponding to stationary and exponential growth stages by hierarchical clustering. Next, for each group of strains robust principle component analysis built a model based on common variation between the strains identifying mutants with the most unusual lipid phenotypes that are unlikely caused by growth differences or other systematic factors. To the best of our knowledge, this is the first report of application of multivariate statistical methods for minimizing the effect of growth in the yeast metabolomic studies.

6.4. Function prediction with experimental confirmation

The presented strategy for searching for novel gene functions is the first systematic attempt to find novel players in lipid homeostasis in yeast. Results of our screen for protein-protein interactions together with recent studies of membrane interactions significantly increased the coverage of the lipid related subspace of the yeast interactome. This collection of data provides unique opportunity for investigation of organizational principles of lipid homeostasis mediated by protein interactions. A few points should be considered for protein interaction network analyses. The comprehensive network of interactions is a combination of datasets obtained under different conditions by various methods. The collection of interaction is a static representation of snapshots of highly dynamic and condition dependent interactome. Thus, collectively interactomics datasets provide an evidence that an interaction can happen at a specific time, location or condition [48]. Finally, even if an interaction between certain proteins undoubtedly happens in the cell, it does not necessary carry a specific biological function [287]. For example, protein interactions could result from mega-assemblies of proteins in intracellular bodies with unclear functional role [288,289]. Furthermore, some interactions may be attributed to evolutionary noise, i.e. interactions that do exist *in vivo*, but have not been selected by evolution to carry a specific function [290]. The complexity of the network makes interpretation of the results of protein-protein interaction screens extremely challenging, even in an ideal case, in which every interaction is true and functionally meaningful.

Automated function prediction methods provide means for network mining for generation of novel hypothesis [199]. The accuracy of the prediction methods are evaluated by computational cross-validation with genes with known function. Thus, the predicted novel gene functions are probable; however, experimental validation is still required for turning a hypothesis into a biological fact. Modern algorithms for functional prediction are highly efficient and provide hundreds of potential candidate genes in a matter of minutes [207]. This gave rise to the development of experimental strategies for large-scale validation of novel functions (discussed in section 1.6.5.3.2. Experimental Validation of Functional Predictions).

We present the first application of lipidomics as a tool for systematic validation of a large number of predictions of novel lipid functions. Comparing to previously reported systematic validation strategies, lipidomics provides high-content data on abundances of hundreds of lipid species for further refining the role of the predicted genes. We note that there are some examples of application of lipidomics for defining gene functions proposed based on network associations [134]. However, in these examples, a small number of tested candidates were selected individually by reviewing results of clustering analyses. In our strategy, the predictions are made automatically, and the high-throughput version of the lipidomics platform is capable of testing hundreds of predictions in a single screen.

A large number of genes is involved in lipid homeostasis in yeast, i.e. about 200 genes are linked to lipid metabolism and another 300 genes to lipid related processes, such as transport and regulation. Moreover, a variety of important mechanisms involves lipids (building membranes, energy storage and signal transduction). Thus, our strategy can be applied for investigation of a broad range of cellular processes.

6.5. Systems for visual programming and visual analytics

The development of software packages and tools for analysis and visualization of multidimensional "omics" data is an actively growing field [139]. Despite the fact, that many powerful interactive tools exist there is still a number of challenges yet to be resolved. One of the main challenges for scientific software development is to sustain the rapid pace of evolution of experimental and bioinformatics techniques. Due to time consuming programming labor associated with development of visualization systems there is a considerable lag between acquisitions of novel experimental data and visualization of results. Moreover, biochemical labs producing novel high-throughput datasets do not necessary have programming expertise, thus have to rely on collaborations with experts form the software development domain.

Effectiveness of data analysis and visualization can be greatly improved with the help of visual programming and visual analytics systems that were employed for supporting lipidomic screen presented in Chapter 5.

There is always a substantial amount of routine processing when dealing with scientific data. The amount of routine work is even higher when a starting point of analysis is raw experimental data. Before data are ready for scientific interpretation, experimental signals (colony intensities or chromatographic peaks) should be converted into numbers (colony sizes, metabolite concentrations), data quality evaluated and filtering steps performed. Novel types of data require development of novel processing algorithms and changes at each step will have an effect on the final data. Visual programming systems combine independent data analysis tasks into executable visual workflows. Advantages of such implementation include transparency of the data processing steps, possibility to change particular steps without reprogramming other components of the system and accessibility of system modifications to researchers without programming expertise. In addition, the visual data processing workflows can be easily shared with scientific peers.

In their book, James J. Thomas and Kristin A. Cook define visual analytics as follows: “Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces. People use visual analytics tools and techniques to synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data; detect the expected and discover the unexpected; provide timely, defensible, and understandable assessments; and communicate assessment effectively for action.”[291] Visual analytics is generic field not specifically developed for biosciences. However, the description “massive, dynamic, ambiguous, and often conflicting“ fits biological data very well, so the implementation of visual analytics approaches for biosciences is highly beneficial. iVici software described in Chapter 4 contributes to the development of the visual analytics field with an interactive framework for heat map comparison. On one hand, existing software for heat map analysis that have interactive features do not support comparison of multiple datasets. On the other, graphical manipulation software that can be used for manual combination of heat maps on a single graph lack the automation and interactive features. In Chapter 5, we adopted modern software for visual analytics, such as Tableau Software and Spotfire, in which similar functionalities can be implemented. These systems produce interactive visualizations with simple drag and drop functionality without the need for programming. These visualization systems are becoming widely popular in pharmaceutical [292], internet and financial industries [293]. However, there are few examples of applications of such systems in other areas of research, including bioinformatics and biochemistry. The highly customizable visual analytics systems present information in a hierarchical manner with interactive links for navigation between layers of information starting from top level overviews to single bits of data. Diversity of available templates and graph types provide an opportunity to create appealing sets of visual representations that are best suited for particular datasets, instead of relying on preprogrammed visualization solutions with fixed functionality.

Furthermore, visualizations can be distributed as interactive graphs between scientific collaborators or as supplementary materials for scientific publications. Adding interactivity to visualizations that accompany scientific publications is particularly attractive because readers will have a chance to browse the data beyond precompiled statistical tables and figures selected by the authors.

6.6. Future directions

6.6.1. Towards dynamics of protein-protein interactions

As described above, the network of protein-protein interactions that can be extracted from public databases is a collection of static interactome snapshots studied under different physiological conditions. However, investigating the dynamics of networks related to environmental, developmental or disease related changes is important for understanding fundamental principles of complex behavior and co-operation of biological processes.

Computational modeling of network dynamics relies on superimposition of static protein-protein interaction network scaffolds with dynamic data, such as gene expression. As a result, dynamic maps of complexes and interactions between proteins that are expressed in the cell at a given time or under particular condition can be produced [294,295]. Large-scale screens for protein-protein interactions increase the interactome coverage and provide novel data for the modeling of a broader range of processes. Most importantly, the development of experimental technologies for protein-protein interaction mapping reduces the costs and increases the throughput of the screens making the experimental investigation of protein-protein interaction dynamics more accessible. Both TAP-MS and Y2H have been applied to capture protein-protein interactions in different conditions [296].

High-throughput PCA methodology extends the toolset for studying dynamics of the interactome. Arrayed collection of tagged strains built for DHFR PCA screen can be grown under different conditions. Furthermore, the collection can be effectively reengineered for performing screens with other reporter fragments. In Chapter 2, we used homologous recombination to introduce a DNA sequence coding for a linker and DHFR fragments. This process required the design of a very large number of oligonucleotide sequences specific to each particular gene, and specific diagnostic primers for verifying correct insertion of the targeted sequences. The costly effort generated a collection of the majority of the yeast open reading frames fused with the linkers and DHFR fragments at C termini. The collection can be further modified with universal primers homologous to the fusion sequence that would fit each of the strains. These technological developments empowered the recent PCA based studies of the protein-protein interaction dynamics in yeast that monitored changes of the canonical Protein Kinase A pathway caused by various stimuli [297] and the modulation of the yeast interactome in response to DNA damage [298].

Future applications, in which methods for studying dynamic protein-protein interactions could be particularly useful, include investigation of mechanisms of cellular compensatory mechanisms and understanding of the phenomena of gene multifunctionality.

Observing the reaction of a system in response to perturbation is a fundamental principle of scientific discovery. However, robustness of biological networks and compensatory mechanisms make it difficult to study certain processes, such as cellular metabolism [299]. Common strategies in yeast include investigation of mRNA expression, protein abundance, growth pattern and, recently, metabolomics analyses of deletion of over-expression mutants of genes of interest. However, deletion of only a limited number of genes leads to observable phenotypes [27]. We described above the application of complementary functional genomic tools to study effects of silent mutations (section 1.5.1. Metabolomics). We gave an example of a metabolomics study that demonstrated changes of metabolite concentrations in deletion mutants lacking certain metabolic enzymes that didn't show observable growth defects [73]. Similarly, investigation of rewiring of the protein-protein interaction network due to mutation or external perturbation may be a feasible strategy to study compensatory mechanisms.

A family of oxysterol binding proteins OSBPs is an intriguing target for studding multifunctionality. The family is present in eukaryotes from yeast to human and its members are implicated in variety of cellular processes such as metabolism of sterols and sphingomyelins, regulation of neutral lipid metabolism, signaling and transport [300]. In yeast, there are 7 members of this family and none of deletion mutants of the corresponding genes shows a significant growth defect. However, the deletion of all 7 genes is lethal [213]. Thus, any single member of the mutlifunctional family is dispensable, but together they carry an essential function. Despite the functional overlap each of OSBPs interacts with a specific set of proteins in yeast. An experiment in which protein-protein interactions will be recorded in mutants lacking each of the OSBP genes may provide an insight on the functional compensation manifested by formation of novel interactions involving remaining OSBPs as well as properties specific to particular OSBPs.

6.6.2. Database of reference lipidomic profiles

Lipidomics is an emerging technique with a growing number of applications. . However, it will take a few years before it reaches the same level of maturation as genomics and proteomics. Comparison of mutant lipid profiles performed in Chapter 3, suggests essential directions for further developments. It is crucial to gain more details about dynamics of lipid profiles related to growth and biological variation. This will increase confidence of lipidomics data interpretation and identification of lipid profile changes caused by specific mutations and not related to growth. There is a switch in lipid metabolism program from exponential to stationary phase [195]. It is mediated by Opi1p repressor that is released from nuclear/ER membrane and enters nucleus in stationary phase. Interaction of Opi1p with Ino2p attenuates the transcription from promoters of several phospholipid synthesis genes that contain an inositol-responsive *cis*-acting element, which leads to increase of phosphatidylinositol and decrease of phosphatidate. Vegetative growth is also associated with a rapid turnover of triacylglycerol, which is used to release lipid precursors for building membranes. Whereas, in stationary phase triacylglycerol is accumulated of as part of the yeast energy storage mechanism which can be quickly turned into membrane when cells start growing again [301]. However, no detailed description of growth related changes in various lipid species is available.

Analysis of the lipidomics data of deletion libraries, described in Chapter 3, demonstrated that in addition to very pronounced differences in lipid profiles between exponential and stationary growth states, there is a systematic variation between mutant strains within the same growth state. Most notably these changes affect the PC/TAG ratio and saturation degree and chain length of fatty acids. There is an indication that these changes are related to growth rate of a strain, but not to specific effect of a gene deletion on lipid metabolism. Robust principle component analysis captured over 80% of common variation between analyzed strains and identified gene deletions with lipid phenotypes that do not follow the common pattern. However, better understanding of the phenomena is important for improving our knowledge about lipid metabolism in general and increasing sensitivity of the lipidomic screens in particular. We have planned a future study of growth effect on lipidome in wild type yeast. Several hundred lipid species will be measured from 0 to 48 hours with 1 to 4 hours intervals complemented with growth curves based on optical density. The data will serve as a reference for future experiments with mutant strains, which will be analyzed in a similar way. Comparison of normal and mutant lipid growth profiles will help to identify specific lipid changes that are due to mutation. Furthermore, it will reveal mutations that have more pronounced effects in either exponential or stationary growth state, which will not be apparent by analysis of one state alone. In gene expression, normalization to levels of housekeeping genes is routinely used for quantification. Results of growth lipid profiling will be used to evaluate whether similar method can be applied in lipidomic data processing.

Furthermore, the reference profiles for deletion strains of known lipid enzymes at the high resolution of lipid species has not been systematically recorded. Creation of such reference database would be helpful for the interpretation of mutant lipid phenotypes of genes not previously implicated into lipid homeostasis. The reference database will also help to refine current metabolic pathways maps. In comprehensive metabolic databases, such as KEGG, lipid metabolism is resolved to the level of lipid classes. The mass-spectrometry based lipidomics will bring another layer of details related to lipid metabolic pathways.

6.6.3. Investigation of special organization of lipid metabolic machinery

The modular organization of networks reflects fundamental mechanisms of performance optimization and regulation of activity of cellular machines. In the context of metabolism, protein interactions between enzymes contribute to formation of metabolic channels [302].

Metabolic channels resemble assembly lines in which metabolites are passed from one enzyme to another leading to optimization of the metabolic flux. A systematic analysis of protein-protein interaction networks from *Escherichia coli* and *Saccharomyces cerevisiae* demonstrated that enzymes involved in neighboring metabolic reactions interact with each other more frequently

[303,304]. It was shown that numerous pairs of enzymes connected with a protein-protein interaction were known examples of metabolic channeling, thus suggesting that identified novel pairs of connected enzymes are potential candidates to study channeling phenomenon. Another study [305] revealed topological equivalences of protein-protein interactions and the metabolic pathway networks based on comparison of global topological network properties. These results imply possible contribution of evolved protein interactions into optimization of efficiency of metabolic processes. Moreover, this study led to an intriguing observation that in addition to interacting enzymes, non-metabolic mediator proteins may have an impact on modularity of enzymatic associations because their presence in the protein-protein interaction network shortens the distance between enzymes.

Recently, the role of non-metabolic proteins was evaluated in more detail [306]. Pérez-Bercoff and co-workers analyzed indirect connections between enzymatic pairs (i.e. connection though a common non-metabolic protein partner) in protein-protein interaction networks of *Esherichia coli*, yeast and humans and found evidence that in all three species indirect connection between enzymes are much more frequent than expected by chance. Furthermore, reactions catalyzed by enzymes connected with mediator proteins were shown to have a higher metabolic flux by computational analysis. Thus, direct interactions between enzymes and indirect associations through mediator proteins contribute to efficient organization of metabolic machinery.

Protein-protein interaction data generated in Chapter 2 in combination with methodology presented in Chapter 3 can be used to experimentally validate results from computational analyses described above. One can identify proteins that potentially contribute to channeling and investigate whether deletion mutation will affect the metabolic efficiency of a cell.

6.6.4. Correlation analysis of lipid concentrations with network constraints

Homeostasis of small molecules including lipids is dependent on coordination between metabolic, signaling and transport events that are not captured by oversimplified pathway maps. The ability to measure thousands of metabolites by MS techniques resulted in development of methods that investigate metabolite-metabolite correlations [307]. Strong correlations between certain metabolites could indicate unknown regulatory and metabolic relationships. Such correlations can be studied within wild type strains at different time points or compared with changes in correlation structure due to perturbation. Overlaying data on known metabolic pathways will reveal links that are potentially novel. These will correspond to correlated metabolites that are not connected with reactions on pathway maps. Possible hypotheses about mechanisms of such relationships can be searched in protein-protein interaction networks by investigating proteins that connect correlated metabolic processes.

6.6.5. Function prediction: from yeast to human

Experimental work of this thesis was performed in yeast. This is an excellent model organism because it can be easily grown and genetically modified to allow a multitude of trial and error experiments. On the other hand, yeast is complex enough to learn how novel strategies can be applied to study higher eukaryotes. Yeast gene function predictions efforts have a potential to reveal conserved cellular mechanisms. Moreover, genome-wide approaches are not limited to yeast. Examples of applications of functional genomics strategies include studies of obesity, diabetes and cardiovascular diseases [308,309]. At present, a wealth of experimentally determined and computationally predicted protein interactions in human are available in public databases [310]. In future, we expect a rapid growth of data coming due to such efforts as Human Interactome Project [311], which will provide a rich source of information for network modeling of disease [310,312,313]. In the meantime, lipidomics approaches are being developed for effective quantification of lipids from various human cell lines and tissue types [314]. Lipids play the major role in such diseases as Alzheimer's, diabetes and atherosclerosis. Integration of multiple layers of information that covers protein interactions maps and metabolic states is important for providing an adequate picture of complex cellular processes and disease mechanisms. Therefore, a tempting direction for continuation of the work presented in this thesis is the development of a similar strategy for studying higher eukaryotes, including humans.

6.7. Conclusion

We presented a strategy for identification of novel gene functions related to lipid metabolism by integration of interactomics and lipidomics. We began by setting up an interactomics screen for mapping protein-protein interactions by applying PCA on a large scale for the first time. The screen revealed 2770 protein-protein interactions the majority of which were novel. For performing the screen, we created high-density stain arrays that exceeded the capacity of similar arrays reported previously. To analyze the data of the survival assay performed in high density, we developed a dedicated image analysis software, capable of identifying colony positions with high precision and deconvoluting overlapping colonies. Novel interactions, that were reported, connected proteins with distinct functional roles presenting data on cross-functional communication, increased the coverage of the interactome subspace related to membrane and lipid homeostasis, and linked proteins with the unknown function. Next, we employed an automated machine learning approach that identified several genes with previously unknown function that are likely to play a role in lipid homeostasis. Finally, we developed a high-throughput lipidomics platform for measuring lipidomes of a large number of strains that we used for verifying the predictions of novel lipid functions. For building the platform, we took the advantage of batch strain culturing, lipid extraction in 96-well format, automated direct infusion nanoelectrospray ionization, high-mass resolution Orbitrap mass spectrometry and a dedicated multivariate data processing framework. Integration of interaction network analysis with lipidomics data allowed to link to lipid homeostasis two genes with unknown function *YBR141C* and *YJR015W*, and a transcription factor *KAR4* that has not been linked previously with lipid metabolism. High-content lipidomics results provided further details about the genes of interest linking them to GPI-anchor synthesis, sterol lipid metabolism and nuclear membrane dynamics correspondingly. The presented strategy is the first systematic attempt for predicting and validating novel lipid related functions. Moreover, experimental and computational methodologies that we developed for conducting each of the screens, contribute to advancing the relatively new fields of interactomics and lipidomics.

In contrast to classical biochemical approaches, our study has not begun with a specific hypothesis about a particular biological question. Instead, it has been driven by an idea that integration of new “omics” data could reveal novel biological insights, which could be further tested experimentally.

The presented study is a data driven process that relies on unbiased, model-independent exploration of experimental results. Such exploration led to identification of hints about novel gene functions suggested by the network structure of the interactome. The observations could be further refined by examining another layer of phenotypic information. Lipidomics data provided additional means for formulating specific hypotheses about functions of several genes, which could be the subject for future biochemical studies.

Chapter 7 : References

1. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. Science 274: 546, 563–567.

2. Botstein D, Chervitz SA, Cherry JM (1997) Yeast as a model organism. Science 277: 1259–1260.

3. Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli S V, White O, Botstein D, Dolinski K (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PLoS One 2: e766.

4. Gietz RD, Woods RA (2001) Genetic transformation of yeast. Biotechniques 30: 816–20, 822–6, 828 passim.

5. Gietz RD, Woods RA (2002) Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Methods Enzymol 350: 87–96.

6. Kawai S, Hashimoto W, Murata K (n.d.) Transformation of Saccharomyces cerevisiae and other fungi: methods and possible underlying mechanism. Bioeng Bugs 1: 395–403.

7. Klinner U, Schäfer B (2004) Genetic aspects of targeted insertion mutagenesis in yeasts. FEMS Microbiol Rev 28: 201–223.

8. Aylon Y, Kupiec M (2004) New insights into the mechanism of homologous recombination in yeast. Mutat Res 566: 231–248.

9. Lorenz MC, Muir RS, Lim E, McElver J, Weber SC, Heitman J (1995) Gene disruption with PCR products in Saccharomyces cerevisiae. Gene 158: 113–117.

10. Goldstein AL, McCusker JH (1999) Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast 15: 1541–1553.

11. Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast 14: 115–132.

12. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Hieter P, Vogelstein B, Kinzler KW (1997) Characterization of the yeast transcriptome. Cell 88: 243–251.

13. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9: 3273–3297.

14. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. Science 282: 699–705.

15. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11: 4241–4257.

16. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21: 33–37.

17. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.

18. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 23: 2692–2699.

19. Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI (1999) A sampling of the yeast proteome. Mol Cell Biol 19: 7357–7368.

20. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003) Global analysis of protein expression in yeast. Nature 425: 737–741.

21. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840–846.

22. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, Westphall MS, Coon JJ (2014) The one hour yeast proteome. Mol Cell Proteomics 13: 339–347.

23. Picotti P, Clément-Ziza M, Lam H, Campbell DS, Schmidt A, et al. (2013) A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis. Nature 494: 266–270.

24. De Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. Mol Biosyst 5: 1512–1526.

25. Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13: 227–232.

26. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901–906.

27. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387–391.

28. Suter B, Auerbach D, Stagljar I (2006) Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. Biotechniques 40: 625–644.

29. Chang M, Bellaoui M, Boone C, Brown GW (2002) A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. Proc Natl Acad Sci U S A 99: 16934–16939.

30. Lee W, St Onge RP, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G (2005) Genome-wide requirements for resistance to functionally distinct DNA-damaging agents. PLoS Genet 1: e24.

31. Zewail A, Xie MW, Xing Y, Lin L, Zhang PF, Zou W, Saxe JP, Huang J (2003) Novel functions of the phosphatidylinositol metabolic pathway discovered by a chemical genomics screen with wortmannin. Proc Natl Acad Sci U S A 100: 3345–3350.

32. Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C (2004) Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat Biotechnol 22: 62–69.

33. Huh W-K, Falvo J V, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK (2003) Global analysis of protein localization in budding yeast. Nature 425: 686–691.

34. Natter K, Leitner P, Faschinger A, Wolinski H, McCraith S, Fields S, Kohlwein SD (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.

35. Athenstaedt K, Zweytick D, Jandrositz A, Kohlwein SD, Daum G (1999) Identification and characterization of major lipid particle proteins of the yeast Saccharomyces cerevisiae. J Bacteriol 181: 6441–6448.

36. Horak CE, Snyder M (2002) ChIP-chip: a genomic approach for identifying transcription factor binding sites. Methods Enzym 350: 469–483.

37. Kim J, Bhinge AA, Morgan XC, Iyer VR (2005) Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. Nat Methods 2: 47–53.

38. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell 124: 207–219.

39. Kel A, Voss N, Valeev T, Stegmaier P, Kel-Margoulis O, Wingender E (2008) ExPlain: finding upstream drug targets in disease gene regulatory networks. SAR QSAR Env Res 19: 481–494.

40. Stark C, Su TC, Breitkreutz A, Lourenco P, Dahabieh M, Breitkreutz BJ, Tyers M, Sadowski I (2010) PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. Database (Oxford) 2010: bap026.

41. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell 123: 507–519.

42. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, et al. (2010) The genetic landscape of a cell. Science (80-) 327: 425–431.

43. Gallego O, Betts MJ, Gvozdenovic-Jeremic J, Maeda K, Matetzki C, Aguilar-Gurrieri C, Beltran-Alvarez P, Bonn S, Fernández-Tornero C, Jensen LJ, Kuhn M, Trott J, Rybin V, Müller CW, Bork P, Kaksonen M, Russell RB, Gavin A-C (2010) A systematic screen for protein–lipid interactions in Saccharomyces cerevisiae. Mol Syst Biol 6.

44. Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. Nature 340: 245–246.

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

46. Braun P (2012) Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. Proteomics 12: 1499–1518.

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

48. Dreze M, Monachello D, Lurin C, Cusick ME, Hill DE, Vidal M, Braun P (2010) High-quality binary interactome mapping. Methods Enzymol 470: 281–315.

49. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B (1999) A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol 17: 1030–1032.

50. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E, Wilm M, Séraphin B (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24: 218–229.

51. Collins SR, Kemmeren P, Zhao X-CC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Krogan NJ (2007) Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics 6: 439–450.

52. Dunham WH, Mullin M, Gingras A-C (2012) Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics 12: 1576–1590.

53. Michnick SW, Remy I, Campbell-Valois FX, Vallée-Bélisle A, Pelletier JN (2000) Detection of protein-protein interactions by protein fragment complementation strategies. Methods Enzymol 328: 208–230.

54. Michnick SW, Ear PH, Landry C, Malleshaiah MK, Messier V (2010) A toolkit of protein-fragment complementation assays for studying and dissecting large-scale and dynamic protein-protein interactions in living cells. Methods Enzymol 470: 335–368.

55. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW (2008) An in vivo map of the yeast protein interactome. Science 320: 1465–1470.

56. Pelletier JN, Campbell-Valois FX, Michnick SW (1998) Oligomerization domain-directed reassembly of active dihydrofolate reductase from rationally designed fragments. Proc Natl Acad Sci U S A 95: 12141–12146.

57. Uetz P, Giot L, Cagney G, Mansfield T a, Judson RS, et al. (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403: 623–627.

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

59. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.

60. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–183.

61. Gentleman R, Huber W (2007) Making the most of high-throughput protein-interaction data. Genome Biol 8: 112.

62. Grigoriev a. (2003) On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res 31: 4157–4161.

63. Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS, Fields S (2005) Large-scale identification of yeast integral membrane protein interactions. Proc Natl Acad Sci U S A 102: 12123–12128.

64. Gavin A, Aloy P, Grandi P, Krause R (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.

65. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.

66. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104–110.

67. Babu M, Vlasblom J, Pu S, Guo X, Graham C, et al. (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature 489: 585–589.

68. Weckwerth W (2003) Metabolomics in systems biology. Annu Rev Plant Biol 54: 669–689.

69. Griffiths WJ, Wang Y (2009) Mass spectrometry: from proteomics to metabolomics and lipidomics. Chem Soc Rev 38: 1882–1896.

70. Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol 169: 453–468.

71. Rabinowitz JD, Purdy JG, Vastag L, Shenk T, Koyuncu E (2011) Metabolomics in drug target discovery. Cold Spring Harb Symp Quant Biol 76: 235–246.

72. Ekroos K, Jänis M, Tarasov K, Hurme R, Laaksonen R (2010) Lipidomics: a tool for studies of atherosclerosis. Curr Atheroscler Rep 12: 273–281.

73. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff H V, van Dam K, Oliver SG (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19: 45–50.

74. Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG, Kell DB (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. Nat Biotechnol 21: 692–696.

75. Lourenço A, Ascenso J, Sá-Correia I (2011) Metabolic insights into the yeast response to propionic acid based on high resolution 1H NMR spectroscopy. Metabolomics 7: 457–468.

76. Lourenço AB, Roque FC, Teixeira MC, Ascenso JR, Sá-Correia I (2013) Quantitative (1)H-NMR-Metabolomics Reveals Extensive Metabolic Reprogramming and the Effect of the Aquaglyceroporin FPS1 in Ethanol-Stressed Yeast Cells. PLoS One 8: e55439.

77. Devantier R, Scheithauer B, Villas-Bôas SG, Pedersen S, Olsson L (2005) Metabolite profiling for analysis of yeast stress response during very high gravity ethanol fermentations. Biotechnol Bioeng 90: 703–714.

78. Villas-Bôas SG, Moxley JF, Akesson M, Stephanopoulos G, Nielsen J (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. Biochem J 388: 669–677.

79. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD (2006) Conservation of the metabolomic response to starvation across two divergent microbes. Proc Natl Acad Sci U S A 103: 19302–19307.

80. Lu W, Clasquin MF, Melamud E, Amador-Noguez D, Caudy AA, Rabinowitz JD (2010) Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer. Anal Chem 82: 3212–3221.

81. Pir P, Kirdar B, Hayes A, Onsan ZY, Ulgen KO, Oliver SG (2006) Integrative investigation of metabolic and transcriptomic data. BMC Bioinformatics 7: 203.

82. Bradley PH, Brauer MJ, Rabinowitz JD, Troyanskaya OG (2009) Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. PLoS Comput Biol 5: e1000270.

83. Cakir T, Patil KR, Onsan Z iIsen, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol 2: 50.

84. Cooper SJ, Finney GL, Brown SL, Nelson SK, Hesselberth J, MacCoss MJ, Fields S (2010) High-throughput profiling of amino acids in strains of the Saccharomyces cerevisiae deletion collection. Genome Res 20: 1288–1296.

85. Chumnanpuen P, Hansen MAE, Smedsgaard J, Nielsen J (2014) Dynamic Metabolic Footprinting Reveals the Key Components of Metabolic Network in Yeast Saccharomyces cerevisiae. Int J Genomics 2014: 894296.

86. Breunig JS, Hackett SR, Rabinowitz JD, Kruglyak L (2014) Genetic basis of metabolome variation in yeast. PLoS Genet 10: e1004142.

87. Wenk MR (2010) Lipidomics: new tools and applications. Cell 143: 888–895.

88. Wenk MR (2005) The emerging field of lipidomics. Nat Rev Drug Discov 4: 594–610.

89. Shevchenko A, Simons K (2010) Lipidomics: coming to grips with lipid diversity. Nat Rev Mol Cell Biol 11: 593–598.

90. Schneiter R, Brügger B, Sandhoff R, Zellnig G, Leber A, Lampl M, Athenstaedt K, Hrastnik C, Eder S, Daum G, Paltauf F, Wieland FT, Kohlwein SD (1999) Electrospray ionization tandem mass spectrometry (ESI-MS/MS) analysis of the lipid molecular species composition of yeast subcellular membranes reveals acyl chain-based sorting/remodeling of distinct molecular species en route to the plasma membrane. J Cell Biol 146: 741–754.

91. Boumann HA, Damen MJA, Versluis C, Heck AJR, de Kruijff B, de Kroon AIPM (2003) The two biosynthetic routes leading to phosphatidylcholine in yeast produce different sets of molecular species. Evidence for lipid remodeling. Biochemistry 42: 3054–3059.

92. Gaspar ML, Aregullin MA, Jesch SA, Nunez LR, Villa-García M, Henry SA (2007) The emergence of yeast lipidomics. Biochim Biophys Acta 1771: 241–254.

93. Santos AXS, Riezman H (2012) Yeast as a model system for studying lipid homeostasis and function. FEBS Lett 586: 2858–2867.

94. Guan XL, Wenk MR (2006) Mass spectrometry-based profiling of phospholipids and sphingolipids in extracts from Saccharomyces cerevisiae. Yeast 23: 465–477.

95. Bourque SD, Titorenko VI (2009) A quantitative assessment of the yeast lipidome using electrospray ionization mass spectrometry. J Vis Exp.

96. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, Klemm RW, Simons K, Shevchenko A (2009) Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. Proc Natl Acad Sci U S A 106: 2136–2141.

97. Shui G, Guan XL, Low CP, Chua GH, Goh JSY, Yang H, Wenk MR (2010) Toward one step analysis of cellular lipidomes using liquid chromatography coupled with mass spectrometry: application to Saccharomyces cerevisiae and Schizosaccharomyces pombe lipidomics. Mol Biosyst 6: 1008–1017.

98. Surma MA, Klose C, Klemm RW, Ejsing CS, Simons K (2011) Generic sorting of raft lipids into secretory vesicles in yeast. Traffic 12: 1139–1147.

99. Klose C, Surma MA, Gerl MJ, Meyenhofer F, Shevchenko A, Simons K (2012) Flexibility of a eukaryotic lipidome--insights from yeast lipidomics. PLoS One 7: e35063.

100. Beach A, Richard VR, Leonov A, Burstein MT, Bourque SD, Koupaki O, Juneau M, Feldman R, Iouk T, Titorenko VI (2013) Mitochondrial membrane lipidome defines yeast longevity. Aging (Albany NY) 5: 551–574.

101. Freedman D, Pisani R, Purves R (2007) Statistics. 4th ed. *W. W. Norton & Company*.

102. Sheskin DJ (2011) Handbook of Parametric and Nonparametric Statistical Procedures. 5th ed. *Chapman and Hall/CRC*.

103. Reshef DN, Reshef Y a, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC (2011) Detecting novel associations in large data sets. Science 334: 1518–1524.

104. Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci U S A 100: 1128–1133.

105. Wang J, Li M, Deng Y, Pan Y (2010) Recent advances in clustering methods for protein interaction networks. BMC Genomics 11 Suppl 3: S10.

106. Langfelder P, Zhang B, Horvath S (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics 24: 719–720.

107. Vlasblom J, Wodak SJ (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics 10: 99.

108. Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

109. Pu S, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 37: 825–831.

110. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7: 944–960.

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

112. Przulj N, Pržulj N (2011) Protein-protein interactions: making sense of networks via graph-theoretic modeling. Bioessays 33: 115–123.

113. Newman MEJ (2003) The Structure and Function of Complex Networks. SIAM Rev 45: 167–256.

114. Przulj N, Corneil DG, Jurisica I (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20: 3508–3515.

115. Zhu X, Gerstein M, Snyder M (2007) Getting connected: analysis and principles of biological networks. Genes Dev 21: 1010–1024.

116. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–42.

117. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.

118. Wagner A, Fell DA (2001) The small world inside large metabolic networks. Proc Biol Sci 268: 1803–1810.

119. Joy MP, Brock A, Ingber DE, Huang S (2005) High-betweenness proteins in the yeast protein interaction network. J Biomed Biotechnol 2005: 96–103.

120. Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 23: 839–844.

121. Friedel CC, Zimmer R (2006) Toward the complete interactome. Nat Biotechnol 24: 614–5; author reply 615.

122. Gong Y, Kakihara Y, Krogan N, Greenblatt J, Emili A, Zhang Z, Houry WA (2009) An atlas of chaperone-protein interactions in Saccharomyces cerevisiae: implications to protein folding pathways in the cell. Mol Syst Biol 5: 275.

123. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, et al. (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446: 806–810.

124. Makhnevych T, Sydorskyy Y, Xin X, Srikumar T, Vizeacoumar FJ, Jeram SM, Li Z, Bahr S, Andrews BJ, Boone C, Raught B (2009) Global map of SUMO function revealed by protein-protein interaction and genetic networks. Mol Cell 33: 124–135.

125. Hughes TR, Robinson MD, Mitsakakis N, Johnston M (2004) The promise of functional genomics: completing the encyclopedia of a cell. Curr Opin Microbiol 7: 546–554.

126. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402: 83–86.

127. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science (80-) 302: 449–453.

128. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A 100: 8348–8353.

129. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.

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

131. Ramakrishnan SR, Vogel C, Kwon T, Penalva LO, Marcotte EM, Miranker DP (2009) Mining gene functional networks to improve mass-spectrometry-based protein identification. Bioinformatics 25: 2955–2961.

132. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 9 Suppl 1: S4.

133. Denic V, Weissman JS (2007) A molecular caliper mechanism for determining very long-chain fatty acid length. Cell 130: 663–677.

134. Breslow DK, Collins SR, Bodenmiller B, Aebersold R, Simons K, Shevchenko A, Ejsing CS, Weissman JS (2010) Orm family proteins mediate sphingolipid homeostasis. Nature 463: 1048–1053.

135. Aguilar PS, Fröhlich F, Rehman M, Shales M, Ulitsky I, Olivera-Couto A, Braberg H, Shamir R, Walter P, Mann M, Ejsing CS, Krogan NJ, Walther TC, Frohlich F (2010) A plasma-membrane E-MAP reveals links of the eisosome with sphingolipid metabolism and endosomal trafficking. Nat Struct Mol Biol 17: 901–908.

136. Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, Paw J, Clore JJ, Mendoza RM, Luis BS, Nislow C, Giaever G, Costanzo M, Troyanskaya OG, Caudy AA (2009) Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLoS Genet 5: e1000407.

137. Ogur M, St John R (1956) A differential and diagnostic plating method for population studies of respiration deficiency in yeast. J Bacteriol 72: 500–504.

138. Chikina MD, Huttenhower C, Murphy CT, Troyanskaya OG (2009) Global prediction of tissue-specific gene expression and context-dependent gene networks in Caenorhabditis elegans. PLoS Comput Biol 5: e1000417.

139. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, Gavin A-CC (2010) Visualization of omics data for systems biology. Nat Methods 7: S56–68.

140. Fruchterman TMJ, Reingold EM (1991) Graph Drawing by Force-directed Placement. Softw Pr Exper 21: 1129–1164.

141. Wiese R, Eiglsperger M, Kaufmann M (2002) yFiles: Visualization and Automatic Layout of Graphs. In: Mutzel P, Jünger M, Leipert S, editors. Graph Drawing SE - 42. Lecture Notes in Computer Science. *Springer Berlin Heidelberg*, Vol. 2265. pp. 453–454.

142. Bader GD, Hogue CW V (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

143. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.

144. Bertin J (1967) Sémiologie graphique : Les diagrammes - Les réseaux - Les cartes. Paris, France: *Editions de l'Ecole des Hautes Etudes en Sciences*.

145. Ghoniem M, Fekete J-D, Castagliola P (2005) On the Readability of Graphs Using Node-link and Matrix-based Representations: A Controlled Experiment and Statistical Analysis. Inf Vis 4: 114–135.

146. Yao G, Craven M, Drinkwater N, Bradfield CA (2004) Interaction networks in yeast define and enumerate the signaling steps of the vertebrate aryl hydrocarbon receptor. PLoS Biol 2: E65.

147. Saldanha AJ (2004) Java Treeview--extensible visualization of microarray data. Bioinformatics 20: 3246–3248.

148. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109–14.

149. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36: W423–6.

150. Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. PLoS One 4: e7710.

151. Nookaew I, Olivares-Hernández R, Bhumiratana S, Nielsen J (2011) Genome-scale metabolic models of Saccharomyces cerevisiae. Methods Mol Biol 759: 445–463.

152. Garcia-Albornoz M, Thankaswamy-Kosalai S, Nilsson A, Väremo L, Nookaew I, Nielsen J (2014) BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. Nucleic Acids Res 42: W175–81.

153. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40: D700–5.

154. Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, et al. (2005) CYGD: the Comprehensive Yeast Genome Database. Nucleic Acids Res 33: D364–8.

155. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res 34: D535–9.

156. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

157. Fernández-Suárez XM, Rigden DJ, Galperin MY (2014) The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection. Nucleic Acids Res 42: D1–6.

158. Van de Peppel J, Holstege FCP (2005) Multifunctional genes. Mol Syst Biol 1: 2005.0003.

159. Huang T, Barclay BJ, Kalman TI, von Borstel RC, Hastings PJ (1992) The phenotype of a dihydrofolate reductase mutant of Saccharomyces cerevisiae. Gene 121: 167–171.

160. Remy I, Michnick SW (2006) A highly sensitive protein-protein interaction assay based on Gaussia luciferase. Nat Methods 3: 977–979.

161. Remy I, Michnick SW (1999) Clonal selection and in vivo quantitation of protein interactions with protein-fragment complementation assays. Proc Natl Acad Sci U S A 96: 5394–5399.

162. Pelletier JN, Arndt KM, Plückthun A, Michnick SW (1999) An in vivo library-versus-library selection of optimized protein-protein interactions. Nat Biotechnol 17: 683–690.

163. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Pagé N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294: 2364–2368.

164. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9: 671–675.

165. Girish V, Vijayalakshmi A (n.d.) Affordable image analysis using NIH Image/ImageJ. Indian J Cancer 41: 47.

166. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9: 62–66.

167. Shah NA, Laws RJ, Wardman B, Zhao LP, Hartman JL (2007) Accurate, precise modeling of cell proliferation kinetics from time-lapse imaging and automated image analysis of agar yeast culture arrays. BMC Syst Biol 1: 3.

168. Beucher S, Lantuejoul C (1979) Use of Watersheds in Contour Detection. International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation, Rennes, France.

169. Meyer F (1994) Topographic distance and watershed lines. Signal Processing 38: 113–125.

170. Rosenfeld A, Pfaltz JL (1966) Sequential Operations in Digital Picture Processing. J ACM 13: 471–494.

171. Rosenfeld A, Pfaltz JL (1968) Distance functions on digital pictures. Pattern Recognit 1: 33–61.

172. Duda RO, Hart PE (1972) Use of the Hough Transformation to Detect Lines and Curves in Pictures. Commun ACM 15: 11–15.

173. Brideau C, Gunter B, Pikounis B, Liaw A (2003) Improved statistical methods for hit selection in high-throughput screening. J Biomol Screen 8: 634–647.

174. Xia Y, Lu LJ, Gerstein M (2006) Integrated prediction of the helical membrane protein interactome in yeast. J Mol Biol 357: 339–349.

175. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet 18: 529–536.

176. Tan S, Tan HT, Chung MCM (2008) Membrane proteins and membrane proteomics. Proteomics 8: 3924–3932.

177. D'Angelo G, Vicinanza M, Di Campli A, De Matteis MA (2008) The multiple roles of PtdIns(4)P -- not just the precursor of PtdIns(4,5)P2. J Cell Sci 121: 1955–1963.

178. Sancho J (2006) Flavodoxins: sequence, folding, binding, function and beyond. Cell Mol Life Sci 63: 855–864.

179. Grandori R, Carey J (1994) Six new candidate members of the alpha/beta twisted open-sheet family detected by sequence similarity to flavodoxin. Protein Sci 3: 2185–2193.

180. Cardona F, Orozco H, Friant S, Aranda A, del Olmo M lí (2011) The Saccharomyces cerevisiae flavodoxin-like proteins Ycp4 and Rfs1 play a role in stress response and in the regulation of genes related to metabolism. Arch Microbiol 193: 515–525.

181. Grossmann G, Malinsky J, Stahlschmidt W, Loibl M, Weig-Meckl I, Frommer WB, Opekarová M, Tanner W (2008) Plasma membrane microdomains regulate turnover of transport proteins in yeast. J Cell Biol 183: 1075–1088.

182. Lingwood D, Simons K (2010) Lipid rafts as a membrane-organizing principle. Science 327: 46–50.

183. Simons K, Sampaio JL (2011) Membrane organization and lipid rafts. Cold Spring Harb Perspect Biol 3: a004697.

184. Martin TF (1998) Phosphoinositide lipids as signaling molecules: common themes for signal transduction, cytoskeletal regulation, and membrane trafficking. Annu Rev Cell Dev Biol 14: 231–264.

185. Michell RH, Heath VL, Lemmon MA, Dove SK (2006) Phosphatidylinositol 3,5-bisphosphate: metabolism and cellular functions. Trends Biochem Sci 31: 52–63.

186. Beh CT, McMaster CR, Kozminski KG, Menon AK (2012) A detour for yeast oxysterol binding proteins. J Biol Chem 287: 11481–11488.

187. Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. Bioinformatics 23: 257–258.

188. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275–1283.

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

190. Tarassov K, Michnick SW (2005) iVici: Interrelational Visualization and Correlation Interface. Genome Biol 6: R115.

191. Van Meer G (2005) Cellular lipidomics. EMBO J 24: 3159–3165.

192. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CRH, Shimizu T, Spener F, van Meer G, Wakelam MJO, Dennis EA (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res 50 Suppl: S9–14.

193. Wymann MP, Schneiter R (2008) Lipid signalling in disease. Nat Rev Mol Cell Biol 9: 162–176.

194. Dickson RC (2008) Thematic review series: sphingolipids. New insights into sphingolipid metabolism and function in budding yeast. J Lipid Res 49: 909–921.

195. Carman GM, Han GS (2011) Regulation of phospholipid synthesis in the yeast Saccharomyces cerevisiae. Annu Rev Biochem 80: 859–883.

196. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

197. Henry SA, Kohlwein SD, Carman GM (2012) Metabolism and regulation of glycerolipids in the yeast Saccharomyces cerevisiae. Genetics 190: 317–349.

198. Chellappa R, Kandasamy P, Oh CS, Jiang Y, Vemula M, Martin CE (2001) The membrane proteins, Spt23p and Mga2p, play distinct roles in the activation of Saccharomyces cerevisiae OLE1 gene expression. Fatty acid-mediated regulation of Mga2p activity is independent of its proteolytic processing into a soluble transcription act. J Biol Chem 276: 43548–43556.

199. Wang PI, Marcotte EM (2010) It's the machine that matters: Predicting gene function and phenotype from protein networks. J Proteomics 73: 2289–2277.

200. Brügger B, Glass B, Haberkant P, Leibrecht I, Wieland FT, Kräusslich H-G (2006) The HIV lipidome: a raft with an unusual composition. Proc Natl Acad Sci U S A 103: 2641–2646.

201. Klemm RW, Ejsing CS, Surma MA, Kaiser H-J, Gerl MJ, Sampaio JL, de Robillard Q, Ferguson C, Proszynski TJ, Shevchenko A, Simons K (2009) Segregation of sphingolipids and sterols during formation of secretory vesicles at the trans-Golgi network. J Cell Biol 185: 601–612.

202. Carvalho M, Sampaio JL, Palm W, Brankatschk M, Eaton S, Shevchenko A (2012) Effects of diet and development on the Drosophila lipidome. Mol Syst Biol 8: 600.

203. Schölz C, Parcej D, Ejsing CS, Robenek H, Urbatsch IL, Tampé R (2011) Specific lipids modulate the transporter associated with antigen processing (TAP). J Biol Chem 286: 13346–13356.

204. Han X, Yang K, Gross RW (n.d.) Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. Mass Spectrom Rev 31: 134–178.

205. Husen P, Tarasov K, Katafiasz M, Sokol E, Vogt J, Baumgart J, Nitsch R, Ekroos K, Ejsing CS (2013) Analysis of Lipid Experiments (ALEX): A Software Framework for Analysis of High-Resolution Shotgun Lipidomics Data. PLoS One 8: e79736.

206. Schwudke D, Hannich JT, Surendranath V, Grimard V, Moehring T, Burton L, Kurzchalia T, Shevchenko A (2007) Top-down lipidomic screens by multivariate analysis of high-resolution survey mass spectra. Anal Chem 79: 4083–4093.

207. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38: W214–W220.

208. De Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20: 1453–1454.

209. R Development Core Team (2013) R: A Language and Environment for Statistical Computing.

210. Todorov V, Filzmoser P (2009) An Object-Oriented Framework for Robust Multivariate Analysis. J Stat Softw 32: 1–47.

211. Oh CS, Toke DA, Mandala S, Martin CE (1997) ELO2 and ELO3, homologues of the Saccharomyces cerevisiae ELO1 gene, function in fatty acid elongation and are required for sphingolipid formation. J Biol Chem 272: 17376–17384.

212. Sampaio JL, Gerl MJ, Klose C, Ejsing CS, Beug H, Simons K, Shevchenko A (2011) Membrane lipidome of an epithelial cell line. Proc Natl Acad Sci U S A 108: 1903–1907.

213. Beh CT, Cool L, Phillips J, Rine J (2001) Overlapping functions of the yeast oxysterol-binding protein homologues. Genetics 157: 1117–1140.

214. FOLCH J, LEES M, SLOANE STANLEY GH (1957) A simple method for the isolation and purification of total lipides from animal tissues. J Biol Chem 226: 497–509.

215. Hubert M, Engelen S (2004) Robust PCA and classification in biosciences. Bioinformatics 20: 1728–1736.

216. Hubert M, Rousseeuw PJ, Vanden Branden K (2005) ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics 47: 64–79.

217. Babazadeh R, Jafari SM, Zackrisson M, Blomberg A, Hohmann S, Warringer J, Krantz M (2011) The Ashbya gossypii EF-1α promoter of the ubiquitously used MX cassettes is toxic to Saccharomyces cerevisiae. FEBS Lett 585: 3907–3913.

218. Stukey JE, McDonough VM, Martin CE (1989) Isolation and characterization of OLE1, a gene affecting fatty acid desaturation from Saccharomyces cerevisiae. J Biol Chem 264: 16537–16544.

219. Surma MA, Klose C, Peng D, Shales M, Mrejen C, Stefanko A, Braberg H, Gordon DE, Vorkel D, Ejsing CS, Farese R, Simons K, Krogan NJ, Ernst R (2013) A lipid E-MAP identifies Ubx2 as a critical regulator of lipid saturation and lipid bilayer stress. Mol Cell 51: 519–530.

220. De Hertogh B, Carvajal E, Talla E, Dujon B, Baret P, Goffeau A (2002) Phylogenetic classification of transporters and other membrane proteins from Saccharomyces cerevisiae. Funct Integr Genomics 2: 154–170.

221. Orlean P, Menon AK (2007) Thematic review series: lipid posttranslational modifications. GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. J Lipid Res 48: 993–1011.

222. Ejsing CS, Moehring T, Bahr U, Duchoslav E, Karas M, Simons K, Shevchenko A (2006) Collision-induced dissociation pathways of yeast sphingolipids and their molecular profiling in total lipid extracts: a study by quadrupole TOF and linear ion trap-orbitrap mass spectrometry. J Mass Spectrom 41: 372–389.

223. Schneiter R (1999) Brave little yeast, please guide us to thebes: sphingolipid function in S. cerevisiae. Bioessays 21: 1004–1010.

224. Beeler TJ, Fu D, Rivera J, Monaghan E, Gable K, Dunn TM (1997) SUR1 (CSG1/BCL21), a gene necessary for growth of Saccharomyces cerevisiae in the presence of high Ca2+ concentrations at 37 degrees C, is required for mannosylation of inositolphosphorylceramide. Mol Gen Genet 255: 570–579.

225. Bosson R, Jaquenoud M, Conzelmann A (2006) GUP1 of Saccharomyces cerevisiae encodes an O-acyltransferase involved in remodeling of the GPI anchor. Mol Biol Cell 17: 2636–2645.

226. Wlodarski T, Kutner J, Towpik J, Knizewski L, Rychlewski L, Kudlicki A, Rowicka M, Dziembowski A, Ginalski K (2011) Comprehensive structural and substrate specificity classification of the Saccharomyces cerevisiae methyltransferome. PLoS One 6: e23168.

227. Bujnicki JM, Feder M, Radlinska M, Blumenthal RM (2002) Structure prediction and phylogenetic analysis of a functionally diverse family of proteins homologous to the MT-A70 subunit of the human mRNA:m(6)A methyltransferase. J Mol Evol 55: 431–444.

228. McCammon MT, Hartmann M a, Bottema CD, Parks LW (1984) Sterol methylation in Saccharomyces cerevisiae. J Bacteriol 157: 475–483.

229. Hillenmeyer ME, Ericson E, Davis RW, Nislow C, Koller D, Giaever G (2010) Systematic analysis of genome-wide fitness data in yeast reveals novel gene function and drug action. Genome Biol 11: R30.

230. Kurihara LJ, Beh CT, Latterich M, Schekman R, Rose MD (1994) Nuclear congression and membrane fusion: two distinct events in the yeast karyogamy pathway. J Cell Biol 126: 911–923.

231. Kurihara LJ, Stewart BG, Gammie AE, Rose MD (1996) Kar4p, a karyogamy-specific component of the yeast pheromone response pathway. Mol Cell Biol 16: 3990–4002.

232. Gammie AE, Stewart BG, Scott CF, Rose MD (1999) The two forms of karyogamy transcription factor Kar4p are regulated by differential initiation of transcription, translation, and protein turnover. Mol Cell Biol 19: 817–825.

233. Lahav R, Gammie A, Tavazoie S, Rose MD (2007) Role of transcription factor Kar4 in regulating downstream events in the Saccharomyces cerevisiae pheromone response pathway. Mol Cell Biol 27: 818–829.

234. Siniossoglou S (2013) Phospholipid metabolism and nuclear function: roles of the lipin family of phosphatidic acid phosphatases. Biochim Biophys Acta 1831: 575–581.

235. Siniossoglou S, Santos-Rosa H, Rappsilber J, Mann M, Hurt E (1998) A novel complex of membrane proteins required for formation of a spherical nucleus. EMBO J 17: 6449–6464.

236. Santos-Rosa H, Leung J, Grimsey N, Peak-Chew S, Siniossoglou S (2005) The yeast lipin Smp2 couples phospholipid biosynthesis to nuclear membrane growth. EMBO J 24: 1931–1941.

237. Han G-S, O'Hara L, Carman GM, Siniossoglou S (2008) An unconventional diacylglycerol kinase that regulates phospholipid synthesis and nuclear membrane growth. J Biol Chem 283: 20433–20442.

238. Hodge CA, Choudhary V, Wolyniak MJ, Scarcelli JJ, Schneiter R, Cole CN (2010) Integral membrane proteins Brr6 and Apq12 link assembly of the nuclear pore complex to lipid homeostasis in the endoplasmic reticulum. J Cell Sci 123: 141–151.

239. Han G-S, Siniossoglou S, Carman GM (2007) The cellular functions of the yeast lipin homolog PAH1p are dependent on its phosphatidate phosphatase activity. J Biol Chem 282: 37026–37035.

240. Han G-S, Wu W-I, Carman GM (2006) The Saccharomyces cerevisiae Lipin homolog is a Mg2+-dependent phosphatidate phosphatase enzyme. J Biol Chem 281: 9210–9218.

241. Fei W, Shui G, Gaeta B, Du X, Kuerschner L, Li P, Brown AJ, Wenk MR, Parton RG, Yang H (2008) Fld1p, a functional homologue of human seipin, regulates the size of lipid droplets in yeast. J Cell Biol 180: 473–482.

242. Fei W, Shui G, Zhang Y, Krahmer N, Ferguson C, Kapterian TS, Lin RC, Dawes IW, Brown AJ, Li P, Huang X, Parton RG, Wenk MR, Walther TC, Yang H (2011) A role for phosphatidic acid in the formation of "supersized" lipid droplets. PLoS Genet 7: e1002201.

243. Chumnanpuen P, Zhang J, Nookaew I, Nielsen J (2012) Integrated analysis of transcriptome and lipid profiling reveals the co-influences of inositol-choline and Snf1 in controlling lipid biosynthesis in yeast. Mol Genet Genomics 287: 541–554.

244. Glatt S, Létoquart J, Faux C, Taylor NMI, Séraphin B, Müller CW (2012) The Elongator subcomplex Elp456 is a hexameric RecA-like ATPase. Nat Struct Mol Biol 19: 314–320.

245. Wittschieben BO, Otero G, de Bizemont T, Fellows J, Erdjument-Bromage H, Ohba R, Li Y, Allis CD, Tempst P, Svejstrup JQ (1999) A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme. Mol Cell 4: 123–128.

246. Gallagher JEG, Dunbar DA, Granneman S, Mitchell BM, Osheim Y, Beyer AL, Baserga SJ (2004) RNA polymerase I transcription and pre-rRNA processing are linked by specific SSU processome components. Genes Dev 18: 2506–2517.

247. Dragon F, Gallagher JEG, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, Beyer AL, Hunt DF,

Baserga SJ (2002) A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature 417: 967–970.

248. Collart MA, Panasenko OO (2012) The Ccr4--not complex. Gene 492: 42–53.

249. Collart MA, Panasenko OO, Nikolaev SI (2013) The Not3/5 subunit of the Ccr4-Not complex: a central regulator of gene expression that integrates signals between the cytoplasm and the nucleus in eukaryotic cells. Cell Signal 25: 743–751.

250. Nash RS, Volpe T, Futcher B (2001) Isolation and characterization of WHI3, a size-control gene of Saccharomyces cerevisiae. Genetics 157: 1469–1480.

251. Garí E, Volpe T, Wang H, Gallego C, Futcher B, Aldea M (2001) Whi3 binds the mRNA of the G1 cyclin CLN3 to modulate cell fate in budding yeast. Genes Dev 15: 2803–2808.

252. Wang H, Garí E, Vergés E, Gallego C, Aldea M (2004) Recruitment of Cdc28 by Whi3 restricts nuclear accumulation of the G1 cyclin-Cdk complex to late G1. EMBO J 23: 180–190.

253. Cai Y, Futcher B (2013) Effects of the yeast RNA-binding protein Whi3 on the half-life and abundance of CLN3 mRNA and other targets. PLoS One 8: e84630.

254. Yetukuri L, Ekroos K, Vidal-Puig A, Oresic M (2008) Informatics and computational strategies for the study of lipids. Mol Biosyst 4: 121–127.

255. Dennis EA, Deems RA, Harkewicz R, Quehenberger O, Brown HA, et al. (2010) A mouse macrophage lipidome. J Biol Chem 285: 39976–39985.

256. Takamori S, Holt M, Stenius K, Lemke EA, Grønborg M, et al. (2006) Molecular anatomy of a trafficking organelle. Cell 127: 831–846.

257. Zech T, Ejsing CS, Gaus K, de Wet B, Shevchenko A, Simons K, Harder T (2009) Accumulation of raft lipids in T-cell plasma membrane domains engaged in TCR signalling. EMBO J 28: 466–476.

258. Song H, Hsu F-F, Ladenson J, Turk J (2007) Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. J Am Soc Mass Spectrom 18: 1848–1858.

259. Haimi P, Uphoff A, Hermansson M, Somerharju P (2006) Software tools for analysis of mass spectrometric lipidome data. Anal Chem 78: 8324–8331.

260. Leavell MD, Leary JA (2006) Fatty acid analysis tool (FAAT): An FT-ICR MS lipid analysis algorithm. Anal Chem 78: 5497–5503.

261. Hübner G, Crone C, Lindner B (2009) lipID--a software tool for automated assignment of lipids in mass spectra. J Mass Spectrom 44: 1676–1683.

262. Houjou T, Yamatani K, Imagawa M, Shimizu T, Taguchi R (2005) A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. Rapid Commun Mass Spectrom 19: 654–666.

263. Ejsing CS, Duchoslav E, Sampaio J, Simons K, Bonner R, Thiele C, Ekroos K, Shevchenko A (2006) Automated identification and quantification of glycerophospholipid molecular species by multiple precursor ion scanning. Anal Chem 78: 6202–6214.

264. Schwudke D, Oegema J, Burton L, Entchev E, Hannich JT, Ejsing CS, Kurzchalia T, Shevchenko A (2006) Lipid profiling by multiple precursor and neutral loss scanning driven by the data-dependent acquisition. Anal Chem 78: 585–595.

265. Herzog R, Schwudke D, Schuhmann K, Sampaio JL, Bornstein SR, Schroeder M, Shevchenko A (2011) A novel informatics concept for high-throughput shotgun lipidomics based on the molecular fragmentation query language. Genome Biol 12: R8.

266. Ejsing CS, Husen P, Tarasov K (2012) Lipid Informatics: From a Mass Spectrum to Interactomics. Lipidomics. *Wiley-VCH Verlag GmbH & Co. KGaA*. pp. 147–174.

267. Liebisch G, Vizcaíno JA, Köfeler H, Trötzmüller M, Griffiths WJ, Schmitz G, Spener F, Wakelam MJO (2013) Shorthand notation for lipid structures derived from mass spectrometry. J Lipid Res 54: 1523–1530.

268. Ekroos K, Ejsing CS, Bahr U, Karas M, Simons K, Shevchenko A (2003) Charting molecular composition of phosphatidylcholines by fatty acid scanning and ion trap MS3 fragmentation. J Lipid Res 44: 2181–2192.

269. Fahy E, Subramaniam S, Brown HA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, Seyama Y, Shaw W, Shimizu T, Spener F, van Meer G, VanNieuwenhze MS, White SH, Witztum JL, Dennis E a (2005) A comprehensive classification system for lipids. J Lipid Res 46: 839–861.

270. Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B (2005) Microarray data mining with visual programming. Bioinformatics 21: 396–398.

271. Trimbuch T, Beed P, Vogt J, Schuchmann S, Maier N, et al. (2009) Synaptic PRG-1 modulates excitatory transmission via lipid phosphate-mediated signaling. Cell 138: 1222–1235.

272. Liebisch G, Binder M, Schifferer R, Langmann T, Schulz B, Schmitz G (2006) High throughput quantification of cholesterol and cholesteryl ester by electrospray ionization tandem mass spectrometry (ESI-MS/MS). Biochim Biophys Acta 1761: 121–128.

273. Sandhoff R, Brügger B, Jeckel D, Lehmann WD, Wieland FT (1999) Determination of cholesterol at the low picomole level by nano-electrospray ionization tandem mass spectrometry. J Lipid Res 40: 126–132.

274. Bilgin M, Markgraf DF, Duchoslav E, Knudsen J, Jensen ON, de Kroon AIPM, Ejsing CS (2011) Quantitative profiling of PE, MMPE, DMPE, and PC lipid species by multiple precursor ion scanning: a tool for monitoring PE metabolism. Biochim Biophys Acta 1811: 1081–1089.

275. Southam AD, Payne TG, Cooper HJ, Arvanitis TN, Viant MR (2007) Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. Anal Chem 79: 4595–4602.

276. Schuhmann K, Almeida R, Baumert M, Herzog R, Bornstein SR, Shevchenko A (2012) Shotgun lipidomics on a LTQ Orbitrap mass spectrometer by successive switching between acquisition polarity modes. J Mass Spectrom 47: 96–104.

277. Chan RB, Oliveira TG, Cortes EP, Honig LS, Duff KE, Small SA, Wenk MR, Shui G, Di Paolo G (2011) Comparative lipidomic analysis of mouse and human brain with Alzheimer disease. J Biol Chem 287: 2678–2688.

278. Rappley I, Myers DS, Milne SB, Ivanova PT, Lavoie MJ, Brown HA, Selkoe DJ (2009) Lipidomic profiling in mouse brain reveals differences between ages and genders, with smaller changes associated with alpha-synuclein genotype. J Neurochem 111: 15–25.

279. O'Brien JS, Sampson EL (1965) Lipid composition of the normal human brain: gray matter, white matter, and myelin. J Lipid Res 6: 537–544.

280. Jensen L, Bork P (2008) Not comparable, but complementary. Science (80-).

281. Mostafavi S, Morris Q (2010) Fast integration of heterogeneous data sources for predicting gene function with limited annotation. Bioinformatics 26: 1759–1765.

282. Brito GC, Andrews DW (2011) Removing bias against membrane proteins in interaction networks. BMC Syst Biol 5: 169.

283. Steffen M, Petti A, Aach J, D'haeseleer P, Church G (2002) Automated modelling of signal transduction networks. BMC Bioinformatics 3: 34.

284. Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. J Comput Biol 13: 133–144.

285. Sambourg L, Thierry-Mieg N (2010) New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. BMC Bioinformatics 11: 605.

286. Bader JS, Chaudhuri A, Rothberg JM, Chant J (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol 22: 78–85.

287. Landry CR, Levy ED, Rabbo DA, Tarassov K, Michnick SW (2013) Extracting Insight from Noisy Cellular Networks. Cell 155: 983–989.

288. O'Connell JD, Zhao A, Ellington AD, Marcotte EM (2012) Dynamic reorganization of metabolic enzymes into intracellular bodies. Annu Rev Cell Dev Biol 28: 89–111.

289. O'Connell JD, Tsechansky M, Royall A, Boutz DR, Ellington AD, Marcotte EM (2014) A proteomic survey of widespread protein aggregation in yeast. Mol Biosyst 10: 851–861.

290. Levy ED, Landry CR, Michnick SW (2009) How perfect can protein interactomes be? Sci Signal 2: pe11.

291. Thomas JJ, Cook KA (2005) Illuminating the Path: The Research and Development Agenda for Visual Analytics. Richland, Washington, USA: *National Visual Analytics Center & IEEE*.

292. Saffer JD, Burnett VL, Chen G, van der Spek P (2004) Visual analytics in the pharmaceutical industry. IEEE Comput Graph Appl 24: 10–15.

293. Economist T (2010) Special report: Managing information. Econ.

294. De Lichtenberg U, Jensen LJ, Brunak S, Bork P (2005) Dynamic complex formation during the yeast cell cycle. Science 307: 724–727.

295. Przytycka TM, Singh M, Slonim DK (2010) Toward the dynamic interactome: it's about time. Brief Bioinform 11: 15–29.

296. Diss G, Filteau M, Freschi L, Leducq J-B, Rochette S, Torres-Quiroz F, Landry CR (2013) Integrative avenues for exploring the dynamics and evolution of protein interaction networks. Curr Opin Biotechnol 24: 775–783.

297. Freschi L, Torres-Quiroz F, Dubé AK, Landry CR (2013) qPCA: a scalable assay to measure the perturbation of protein-protein interactions in living cells. Mol Biosyst 9: 36–43.

298. Rochette S, Gagnon-Arsenault I, Diss G, Landry CR (2013) Modulation of the yeast protein interactome in response to DNA damage. J Proteomics 100: 25–36.

299. Ikonen E (2008) Cellular cholesterol trafficking and compartmentalization. Nat Rev Mol Cell Biol 9: 125–138.

300. Olkkonen VM, Johansson M, Suchanek M, Yan D, Hynynen R, Ehnholm C, Jauhiainen M, Thiele C, Lehto M (2006) The OSBP-related proteins (ORPs): global sterol sensors for co-ordination of cellular lipid metabolism, membrane trafficking and signalling processes? Biochem Soc Trans 34: 389–391.

301. Kurat CF, Natter K, Petschnigg J, Wolinski H, Scheuringer K, Scholz H, Zimmermann R, Leber R, Zechner R, Kohlwein SD (2006) Obese yeast: triglyceride lipolysis is functionally conserved from mammals to yeast. J Biol Chem 281: 491–500.

302. Srere PA (1987) Complexes of sequential metabolic enzymes. Annu Rev Biochem 56: 89–124.

303. Huthmacher C, Gille C, Holzhütter H-G, Holzhutter HG (2007) Computational analysis of protein-protein interactions in metabolic networks of Escherichia coli and yeast. Genome Inform 18: 162–172.

304. Huthmache C, Gille C, Holzhutter HG, Huthmacher C, Holzhütter H-G (2008) computational analysis of protein interactions in metabolic networks reveals novel enzyme pairs potentially involved in metabolic channeling. J Theor Biol 252: 456–464.

305. Durek P, Walther D (2008) The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles. BMC Syst Biol 2: 100.

306. Pérez-Bercoff A, McLysaght A, Conant GC, Perez-Bercoff A (2011) Patterns of indirect protein interactions suggest a spatial organization to metabolism. Mol Biosyst 7: 3056–3064.

307. Steuer R, Kurths J, Fiehn O, Weckwerth W (2003) Interpreting correlations in metabolomic networks. Biochem Soc Trans 31: 1476–1478.

308. Wheelock CE, Wheelock AM, Kawashima S, Diez D, Kanehisa M, van Erk M, Kleemann R, Haeggstrom JZ, Goto S, Haeggström JZ (2009) Systems biology approaches and pathway tools for investigating cardiovascular disease. Mol Biosyst 5: 588–602.

309. Meng Q, Mäkinen V-P, Luk H, Yang X, Makinen VP (2013) Systems Biology Approaches and Applications in Obesity, Diabetes, and Cardiovascular Diseases. Curr Cardiovasc Risk Rep 7: 73–83.

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

311. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, Fan C, Svrzikapa N, Hirozane-Kishikawa T, Rietman E, Yang X, Sahalie J, Salehi-Ashtiani K, Hao T, Cusick ME, Hill DE, Roth FP, Braun P, Vidal M (2011) Next-generation sequencing to generate interactome datasets. Nat Methods 8: 478–480.

312. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. Bioinformatics 26: 1057–1063.

313. Ideker T, Sharan R (2008) Protein networks in disease. Genome Res 18: 644–652.

314. Van Meer G, de Kroon AIPM (2011) Lipid map of the mammalian cell. J Cell Sci 124: 5–8.