Title: Systematic bioinformatics and experimental validation of yeast complexes reduces the rate of attrition during structural investigations

Article Type: Ways and Means

Keywords: Tandem affinity tag purification, high-throughput expression, multi-protein complexes; bioinformatics

Corresponding Author: Dr. Anastassis Perrakis, PhD
Corresponding Author's Institution: Netherlands Cancer Institute
First Author: Mark A Brooks, PhD

Order of Authors: Mark A Brooks, PhD; Kamil Gewartowski, PhD; Eirini Mitsiki, PhD; Juliette Létoquart; Roland A Pache; Ysaline Billier; Michela Bertero; Margot Corrée; Mariusz Czarnocki-Cieciura; Michał Dadlez; Véronique Henriot; Noureddine Lazar, PhD; Lila Delbos; Dorothée Lebert; Jan Piwowarski; Pascal Rochaix; Bettina Böttcher, PhD; Luis Serrano, PhD; Bertrand Séraphin, PhD; Herman van Tilbeurgh, PhD; Patrick Aloy, PhD; Anastassis Perrakis, PhD; Andrzej Dziembowski, PhD

Abstract: For high-throughput structural studies of protein complexes of composition inferred from proteomics data, it is crucial that candidate complexes are selected accurately. Herein, we exemplify a procedure that combines a bioinformatics tool for complex selection with an in vivo validation, to deliver structural results in a medium-throughput manner. We have selected a set of twenty yeast complexes, which were predicted to be feasible by either an automated bioinformatics algorithm, by manual inspection of primary data, or by literature searches. These complexes were validated with two straightforward and efficient biochemical assays, and heterologous expression technologies of complex components were then used to produce the complexes to assess their feasibility experimentally. Approximately one half of the selected complexes were useful for structural studies, and we detail one particular success story. Our results underscore the importance of accurate target selection and validation in avoiding transient, unstable, or simply non-existent complexes from the outset.

Suggested Reviewers: Shoshana Wodak PhD
Scientific Director, Centre for Computational Biology, Hospital for Sick Children, Toronto
shoshana@sickkids.ca
Dr. Wodak is an expert in the field of protein-protein interaction networks.

Aled M Edwards PhD
Professor, Banting and Best Dept of Medical Research, University of Toronto
aled.edwards@utoronto.ca
Dr. Edwards is a highly knowledgeable member of the structural genomics community, and a pioneer in the field structural genomics, particularly with regards to protein complexes.

Mark B Gerstein PhD
Professor, Dept of Molecular Biophysics and Biochemistry, Yale University
Mark.Gerstein@Yale.edu
Dr. Gerstein is a renowned expert in the bioinformatic treatment of functional genomics and other information, for the purpose of predicting protein networks.

Tom S Peat PhD
Research Program Leader, Molecular & Health Technologies, CSIRO
Tom.Peat@csiro.au
Dr Peat is renowned for his expertise in structural genomics both at Los Alamos and SGX, prior to his present position.

Opposed Reviewers:
Dear Editors,

We would like to submit for your consideration our paper entitled: "Systematic bioinformatics and experimental validation of yeast complexes reduces the rate of attrition during structural investigations" for consideration as a "Ways and Means" article. This paper consolidates our experience on the use of large data sets, the yeast proteome - interactome, as the basis to for structural studies of protein complexes. Our results were obtained as part of a large European collaborative project, 3D-repertoire, which has brought together various laboratories with experience in systems biology, bioinformatics, structural biology and molecular biology, in an effort to provide new insight to the yeast proteome. A significant part of that effort was to provide new structural data of protein complexes, as the basis to promote our understanding of specific protein interactions in eukaryotic cells.

In this paper we exemplify a procedure combining bioinformatics tools for complex selection, in vivo validation and heterologous recombinant expression technologies, to deliver structural results in a medium-throughout manner. In addition, we showcase a test of twenty yeast complexes that were treated in this manner, and discuss in more detail one such complex that went all the way from identification to structural characterization.

To our knowledge this is the first time than an objective study has been done to evaluate the importance of bioinformatics analysis on pull down results to select the best possible targets for structural characterization. We believe this report is of broad interest to the molecular and structural biology communities, and that Structure is the ideal vehicle to bring our results to the attention of the broad readership that we wish to address.

The authors declare that they have no conflict of financial interest with the work presented herein.

Yours sincerely,

Anastassis Perrakis
Systematic bioinformatics and experimental validation of yeast complexes reduces the rate of attrition during structural investigations

Mark A. Brooks¹, ‡, Kamil Gewartowski², ‡, Eirini Mitsiki³, ‡, Juliette Létoquart⁴, Roland A. Pache⁵, Ysaline Billier⁴, Michela Bertero⁶, Margot Corréa⁴, Mariusz Czarnocki-Cieciura³, Michal Dadlez², Véronique Henriot⁴, Noureddine Lazar¹, Lila Delbos¹, Dorothee Lebert⁴, Jan Piwowarski², Pascal Rochaix⁴, Bettina Böttcher⁷, Luis Serrano⁶,⁸, Bertrand Séraphin⁴,⁶,⁷, Herman van Tilbeurgh¹,⁶, Patrick Aloy⁵,⁸,⁹, Anastasssis Perrakis³, ⁺, Andrzej Dziembowski², ‡

¹IBBMC-CNRS UMR8619, IFR 115, Bât. 430, Université Paris-Sud, 91405, Orsay, France.
²Department of Genetics and Biotechnology, Warsaw University & Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawińskiego 5a, 02106 Warsaw, Poland
³Department of Biochemistry, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.
⁴Equipe Labelisée La Ligue, CGM, CNRS UPR2167, Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France.
⁵Institute for Research in Biomedicine (IRB) and Barcelona Supercomputing Center (BSC). c/ Baldiri I Reixac 10-12, 08028 Barcelona, Spain.
⁶Systems Biology Laboratory, Centre for Genomic Regulation, Barcelona, Spain
⁷School of Biological Sciences, University of Edinburgh, King’s Buildings, Mayfield Road, Edinburgh EH9 3JR, United Kingdom
⁸Institució Catalana de Recerca i Estudis Avançats (ICREA)

†Present Address: Evotec (UK) Ltd., 114 Milton Park, Abingdon, Oxon, OX14 4SA, UK.
*Present address: IGBMC, 1 rue Laurent Fries, BP10142, 67404 Illkirch, France
‡These authors contributed equally to the work
Corresponding authors:

Patrick Aloy patrick.aloy@irbbarcelona.org
Andrzej Dziembowski andrzejd@ibb.waw.pl
Anastassis Perrakis a.perrakis@nki.nl
Bertrand Séraphin seraphin@igbmc.fr
Herman van Tilbeurgh Herman.Van-Tilbeurgh@u-psud.fr

Running Title:

Protein complex validation for structural biology

Keywords:

Tandem affinity tag purification, high-throughput expression, multi-protein complexes, bioinformatics
Abstract

For high-throughput structural studies of protein complexes of composition inferred from proteomics data, it is crucial that candidate complexes are selected accurately. Herein, we exemplify a procedure that combines a bioinformatics tool for complex selection with in vivo validation, to deliver structural results in a medium-throughout manner. We have selected a set of twenty yeast complexes, which were predicted to be feasible by either an automated bioinformatics algorithm, by manual inspection of primary data, or by literature searches. These complexes were validated with two straightforward and efficient biochemical assays, and heterologous expression technologies of complex components were then used to produce the complexes to assess their feasibility experimentally. Approximately one half of the selected complexes were useful for structural studies, and we detail one particular success story. Our results underscore the importance of accurate target selection and validation in avoiding transient, unstable, or simply non-existent complexes from the outset.
Introduction

Numerous large-scale proteomics initiatives in the model organism *Saccharomyces cerevisiae* have been reported over the last few years, and have provided evidence for thousands of new protein interactions and supplied a wealth of information about the composition of macromolecular complexes (Gavin et al., 2006; Ho et al., 2002; Ito et al., 2001; Krogan et al., 2006; Tarassov et al., 2008; Uetz et al., 2000). Nevertheless, the characteristics of protein interaction networks *in vivo* have not yet been rigorously untangled for any organism, let alone the faithful budding yeast. Now that such protein interaction datasets are in the public domain, a gauntlet has been thrown down to the scientific community to provide tools for assimilating these data with a view to developing algorithms and experimental methodologies for predicting the composition of complexes with high accuracy, thereby facilitating their functional and structural characterization.

However, for many predicted complexes identified in high-throughput affinity purification experiments, their subunit composition is not established with sufficient reliability to proceed to structure determination. Improvements in the confidence that can be placed in protein interaction models are therefore clearly needed, with the specific aim of identifying complexes with well-defined stoichiometry, and which are amenable to structural studies. Raising the confidence with which complex composition could be predicted would benefit enormously the field of structural biology. Ideally, it would be possible to identify stable complexes (for example ribosomes, RNA polymerases, the exosome, or the 20S proteosome) and discriminate them from more dynamic assemblies that contain transient interactors (for example spliceosomes or the 26S proteosome). It would therefore be beneficial to classify and characterize the various entities which form the central frameworks of protein-protein interaction networks (Gavin et al., 2006; Higurashi et al., 2008; Krogan et al., 2006).

Foremost among the problems encountered in complex characterization are those related to the primary data being of limited quality. For example, the heterogeneity or the extremely dilute nature of samples from proteomic experiments results in complex subunits being overlooked. Additionally, in some studies, the characterization of complex composition has been hindered by the contamination of *bona fide* complexes by so-called ‘background’ or ‘sticky’ polypeptides that interact
with other proteins in a promiscuous fashion (Shevchenko et al., 2002). One challenge is therefore to devise a computational strategy to filter through the results of many thousands of biochemical purifications which have been performed to date, and identify the complexes that will yield the optimal results during expression and purification studies (Bravo and Aloy, 2006).

The first structural genomics consortia focused on the determination of X-ray and NMR structures at the level of the single protein (Alzari et al., 2006; Graslund et al., 2008; Marsden and Orengo, 2008). More recently, the Structural Genomics Consortium (SGC) (Edwards et al., 2002), the 3D Repertoire (http://www.3drepertoire.org/) and SPINE 2 - Complexes (http://www.spine2.eu/) consortia have opted to study macromolecular complexes from a medium-throughput perspective. The expression and purification of protein complexes adds an extra level of complexity, since globular protein interfaces are often partly hydrophobic, and single partners may be insoluble. In many cases, only in the context of an assembled complex do hydrophobic interfaces become buried and the participating polypeptides can be produced as soluble entities (Dyson and Wright, 2005; Smialowski et al., 2007).

Since the inception of the European Commission-funded consortium “3D repertoire” in 2004, collaborating scientists have been addressing the problems associated with identifying complexes de novo for structural studies. Within the first step, which consisted of highly selective filtering of existing datasets for evidence of the existence of complexes in a process we term ‘complex triage’, three methods were employed. Firstly, a bioinformatics-based selection procedure, optimized using a training set composed of complexes of known three-dimensional (3D) structure, was used to screen for stable, well-folded complexes. Secondly, we examined the results from high throughput affinity purification experiments manually, focusing on the visual inspection of gels to identify complexes of which the components existed in stoichiometric quantities. Finally, a set of seven complexes was chosen on the basis of the scientific literature.

A compilation of these complexes, named the 'list of 20', were then validated by new affinity purifications of the natural complexes and their subunit compositions were confirmed using mass spectrometry. In addition, the solution sizes of these complexes were assessed by size exclusion chromatography. The subset of proteins

5
that were shown to indeed participate in macromolecular assemblies as predicted and that was also believed to be tractable for structural studies was then cloned and expressed in *E. coli*. Using various techniques, we aimed to obtain purified material suitable for structural analysis. We show the overall success in each of the steps of this procedure and present a detailed account of one example complex. The results from this test set of complexes under investigation have allowed us to evaluate the effectiveness of each of the techniques used and devise an optimal route for the production of protein complexes in structural biology pipelines.
Results

**Identification of complexes for structural studies**

**Complex triage by bioinformatics**

A system has been previously described for the ranking of the 491 complexes and the 5,488 isoforms that we had previously described from over 2,000 successful tandem affinity purifications (Gavin et al., 2006). This was based on the notion that target complexes likely amenable to structural studies should be small, compact and homogeneous. We considered biophysical, biochemical and large-scale proteomics data in the form of partial scoring functions that were normalized and combined into a final feasibility score for each complex (c.f. Methods, Supplemental Methods, and as described previously (Pache and Aloy, 2008); referred to hereafter as the Complex Feasibility (CF) algorithm). In this way, the public domain data were filtered to generate a much-reduced subset of credible complexes. To generate a list of a total of seven complexes by the CF tool, we combined four of the top ranking choices with three mid-ranking complexes (Table 1, Table S3).

**Complex triage by manual visualization of gels**

In the original genome-wide approach (Gavin et al., 2006), tandem affinity purified (TAP) assemblies were separated by denaturing gel electrophoresis and stained. The gels were then cut into 1 mm slices, digested with trypsin and analyzed by MALDI–TOF mass spectrometry (MS). However, this procedure did not take into account the relative quantities of proteins present in the TAP eluate. Complexes with apparent sub-stoichiometric components are more likely to depend on labile transient interactions and less suitable for structural studies than stable stoichiometric assemblies. We thus decided to visually inspect the original gels (Gavin et al., 2006) for bands indicative of stoichiometric complexes. The resulting assemblies were narrowed down to dimers, trimers and tetramers. Thorough inspection of about 4,000 purification experiments, we identified 64 promising complexes (Table S4; dimeric complexes, Table S5; trimeric complexes, Table S6; tetrameric complexes). Not all of the 64 chosen complexes were present in the computational selection, simply because some of these were not identified as being complexes in the original automated
annotation (Gavin et al., 2006). Notably, the best six complexes that were chosen independently by gel inspection were all in the top-50 of the CF algorithm, and two of them were in the top-10. Six complexes were finally selected by manual gel inspection (Table 1).

*The list of ‘20 complexes’*

In Table 1, we show the 20 complexes selected, with the corresponding bioinformatics and gel scores, and when possible appropriate references to the literature. Although the manual gel inspection and the bioinformatics efforts were independent, all previously identified complexes selected by manual screening had a high ranking using the CF algorithm. In contrast, not all of the complexes chosen by the algorithm could be associated with clear and conclusive gels. Notably, a top-ranked choice was associated with a gel of mediocre quality. Nonetheless, such types of selections resulted in a potentially interesting collection of complexes that would hopefully be amenable to structural studies. The selection was complemented by the choice of an additional seven complexes suggested by partners of 3D repertoire, based on specific biological interests and literature know-how, reaching the final number of 20 complexes included in this study. Interestingly, only one of the latter choices was in the top-10 bioinformatics list, and an additional two were in the top-50; the remaining four scored poorly by the CF algorithm.

*Validation of complex composition*

The twenty selected complexes were validated in a two-step TAP purification on IgG and calmodulin columns. Mass spectrometry analyses using an ESI-TRAP approach were performed using both the eluate solutions and the excised gel bands as samples. In addition, molecular weights of complexes were estimated by size exclusion chromatography of total extracts, followed by dot-blot detection of TAP-tagged proteins in eluate fractions. Finally, the molecular weights of tagged subunits and the efficiency of binding to IgG resin were verified by Western blot analyses (see Figure 1 for a schematic representation of the procedure). The conclusions regarding individual complexes are presented in Table 1 and Figure S1.

Only two of the complexes completely failed this validation stage, one for technical reasons and one could not be identified at all. Interestingly, both complexes originated from the literature additions to the list and they both scored poorly in the
bioinformatics assessment. This category, where literature knowledge was used to select complexes, gave a lower validation rate than the other strategies. Apart from the one complex for which no technically valid results were obtained, one failed, while two others showed too weak native expression to be conclusive. Another complex was highly heterogeneous and one included a very promiscuous protein as a partner and was thus inconclusive. Notably, one complex selected from the literature and validated here to be ‘excellent’, was ranked in the top-10 (20th percentile) of the bioinformatics list. The low validation rates of complexes selected from the literature, and their low bio-computing ranks stem from their specific characteristics (low abundance, specific interaction involving abundant partners flagged as promiscuous) and underline the limitation of current strategies to identify bona fide complexes. The gel-selected complexes and the bioinformatics complexes fared well in the validation, with four out of six and three out of seven, respectively, being scored as 'excellent'. From the validated complexes, eleven were chosen for heterologous expression studies and production in quantities suitable for structural studies. Analysis of the twelfth complex, Dom34:Hbs1 is described elsewhere, so was not repeated (Graille et al., 2008), but is included in Table 1.

**Recombinant production of complexes for structural studies**

For these eleven complexes, a mixture of expression strategies was employed for their evaluation: expression of the full-length individual subunits, *in vitro* complex reconstitution from subunits, and co-expression. A total of twenty-two proteins have been used in expression trials as single full-length proteins in *E. coli*, either from synthetic, codon-optimized genes (16 proteins, Figure 2, panel A) or from natural yeast genes (Figures S2, S3, S4, S5 and S6). Only three of these failed to produce soluble protein in appreciable amounts (Atg29, Psy4 and Ste11). We proceeded to reconstitute three complexes (Vps27:Hse1, Ptc2:Paa1 and Gcd10:Gcd14) from individually purified partners and succeeded in purifying them in soluble form and defined subunit composition. In parallel, we also attempted co-expression of nine complexes, and we were able to produce seven out of nine complexes by such co-expression methods, (Figure 2, panel B).
A case study of an example complex, from selection to validation

To illustrate the course of an experiment from target selection to validation, we present one particular exemplary complex. The Gcd10:Gcd14 complex was originally identified a few years ago and purified as a dimeric tRNA(1-methyladenosine) methyltransferase (Anderson et al., 1998; Anderson et al., 2000; Ozanick et al., 2007). Gavin et al. (Gavin et al., 2006) observed again this dimeric complex, which was annotated as Complex 376 in the Krogan et al. enumeration (Krogan et al., 2006). TAP purified Gcd10:Gcd14 has also been shown to be relatively homogeneous and therefore pure by electron microscopy. We selected this complex by gel analysis but it also ranked with a score of 12 by the CF algorithm.

Firstly, we re-validated the complex by repeating the TAP purification using tagged Gcd14 and the only partner that was isolated was Gcd10, with no other bands either apparent or identified by mass spectrometry Figure 3, panels A and B). Gel filtration analysis of the TAP-tag purified complex was consistent with a molecular weight of approximately 350 kDa, suggesting the formation of higher-order multimers since the expected mass of the Gcd10:Gcd14 complex with a 1:1 stoichiometry is 98.3 kDa.

The complex was reconstituted from the Ni\(^{2+}\)-NTA purified individual components and subjected to gel filtration chromatography. The resulting complex had an approximate molecular weight of around 350 kDa, in agreement with the analysis of the 'native' TAP-tagged complex (Figure 3, panel B). The purified complex was then used in a negative stain electron microscopy experiment. The sample was homogeneous and could be used for data collection (Figure 3, panel C). Image reconstructions without any imposed symmetry showed a tetrameric core with extensions at opposite surfaces, giving the entire complex two-fold, as well as quasi four-fold symmetry. Therefore, C2-symmetry was imposed for further refinement. The final reconstruction is shown in Figure 3, panel E. Projections of this reconstruction agree with class averages were determined by multivariate statistical analysis (Figure 3, panel D).
Discussion

In this work, we set out to identify an optimal strategy for the analysis of *Saccharomyces cerevisiae* complexes by combining contemporary structural biology tools with the numerous proteome-level biochemical interaction datasets. Our central tenet was that we believed such data to be essentially reliable, the use of improved bioinformatics tools, manual analysis of gels or bibliographic curation of previous data should allow the identification of complexes best suited to structural analysis.

A question that we sought to answer related to whether bioinformatics, and specifically the CF algorithm, could provide trustworthy guidance when selecting targets. Ideally, the algorithm should eliminate the need for manual inspection of data. Therefore, we first generated a target list, partly using automated tools and partly manually. The next step was to ascertain which of the selected complexes do indeed exist in a stable and stoichiometric form. Our experimental results show that the bioinformatics algorithm could select targets with a validation success rate that was very high, and comparable to visual inspection of gels.

In the final CF algorithm, the most important parameters were the yeast two-hybrid ratio and the socio-affinity index (Table S2). The usefulness of the former parameter has been obvious for some time, since yeast two-hybrid screening has been a mainstay of research into protein-protein interactions. However, the important role of the socio-affinity index in this experiment was encouraging (Gavin et al., 2006), and we believe that it is a valuable and powerful metric for the identification of protein complexes based on protein interaction datasets. Conversely, the least useful parameters were the ‘average number of problematic residues’ and the ‘co-localization ratio’; it appears that these parameters are not as useful as had been previously thought, at least in the context of this work (Pache and Aloy, 2008).

We note that some complexes identified by bibliographic analyses, which could not be validated and for which low scores were obtained with the CF algorithm, performed well using recombinant expression. These facts underline the limitation of complex analyses of low abundance complexes and/or complexes involving very abundant subunits for which it is difficult to exclude the existence of promiscuous interactions. It is possible that our complex triage procedures have been successful at
least in part, due to the clarity of primary data for which the subunits are stoichiometrically equivalent and well expressed.

The success rate of obtaining soluble subunits by heterologous recombinant expression, for the full-length proteins was high (only 3 of 22 proteins tested could not be produced in a soluble form; 86% success rate). Similarly, we were able to obtain soluble complexes corresponding to most of our validated targets using either complex reassembly or co-expression via either co-transformation of plasmids or single plasmids that contain operons encoding all of the proteins of interest (c.f. Table 1, and Supplemental Material; 9 of 11 complexes could be formed; 82% success rate). We believe that this achievement is principally due to the efficient selection criteria that we had established. It has been reported that only about 20% of full-length eukaryotic proteins are soluble when produced in a heterologous expression system (Graslund et al., 2008), but the performance of our approach is considerably superior. This is likely to be because only natively soluble proteins and complexes that are expressed at suitably high levels are detected by mass spectrometry after TAP purification, thereby biasing complex identification data towards soluble proteins.

Based on the four-year experience of a consortium of numerous structural biology groups involved in 3D repertoire, we suggest an optimal experimental strategy for the high-throughput study of protein complexes. We conclude that despite the absence of a ‘silver bullet’, much can be achieved first by triaging the targets by an efficient computational procedure, followed by simple expression and reconstitution in the first instance. For this, a LIC-based strategy to clone optimized synthetic genes in a parallel manner resulted in notable success, with 14 of 16 subunits expressed in soluble form. During complex reconstitution, we had greater success when employing co-sonication of E. coli in which each subunit had been expressed separately, compared to reconstitution using pure proteins and has become our method of choice to obtain soluble complexes (c.f. Supplemental Experimental Procedures; ‘Complex formation trials’).

We also found that producing plasmids that encode the necessary subunits as synthetic DNA, with Shine-Dalgarno sequences upstream of the successive ORFs to be a very practical and rapid method of co-expressing complexes (c.f. Supplemental
Experimental Procedures; ‘Cloning strategy used for poly-cistronic expression’). Our studies into the use of polycistronic vectors, particularly those constructed from synthetic genes (e.g. Gcd10:Gcd14 and Ssl2:Yor352 complexes, Figure S8) indicate that this is a strategy that this is a useful addition to pipelines, both because of the ease of production of plasmid constructs, and the increase in yield presented by codon-optimized genes.

In summary, we conclude that when initiating projects involving high-throughput study of protein complexes, proper triaging and validation is obligatory. Once this had been performed, it was relatively straightforward to test the association of the recombinant proteins experimentally. As we illustrate with the Gcd10:Gcd14 complex, we were able to obtain structural information during the relatively short time scale of this project. In this work, we have leveraged complementary strategies to the end of complex production for structural analysis, but we envisage the incorporation of further techniques in subsequent experiments. For example, high throughput small angle X-ray scattering studies of single proteins could be applied similarly to complexes (Hura et al., 2009), and it will be increasingly important to identify complex and sub-complex composition of samples purified directly from cells using native mass spectrometry (Hernandez et al., 2006). Accurate subunit prediction and validation methods will be beneficial to future high-throughput approaches geared towards ‘high-hanging fruit’ and increase the probability that such efforts will yield illuminating insights into macromolecular machines at work.
Acknowledgements

This work was supported in part by the European Union 6th Framework program “3D-repertoire” (LSHG-CT-2005-512028) to HvT, AP, PA, LS and to BS. An EMBO young investigator award supported AD. We thank Claire Batisse (EMBL Heidelberg) for technical assistance in electron microscopy. The authors declare that they have no conflict of financial interest with the work presented herein. LS, PA and BS conceived of the study and MB coordinated the work between the contributing laboratories. PA and RAP performed the bioinformatics analysis. KG and AD performed the in-vivo validation. EM and MC-C carried out the expression and coexpression testing using single plasmids and synthetic genes. MAB, NL, LD worked on expression and in vitro reconstitution of complexes. JL, YB, MC, VH, DL and PR worked on complex production through various co-expression methods. NL and MAB produced and characterized the Gcd10:Gcd14 and BB performed electron microscopy on it. BS, HvT, AP and AD supervised the work in their laboratories. MAB and AP and wrote the article, with contributions from MB, LS, BS, HvT, PA and AD.
Figure Legends

Figure 1: Strategy for the validation of selected complexes

A schema showing the overall pathway for the validation of complex composition and estimation of molecular weight of each complex is presented. The complexes were expressed in yeast using a C-terminal TAP-tag of the bait protein. Following cell-breakage, complexes were either subjected to TAP purification to assess the subunit composition, or to gel filtration in order to estimate the molecular weight, and thereby their stoichiometry. See Figure S1 for actual results of the validation experiments.

Figure 2: Expression and purification of yeast full-length proteins

Panel A: SDS-PAGE analysis of full-length yeast constructs produced using codon-optimized synthetic genes, Ni\(^{2+}\)-NTA-purified and visualised using Coomassie. Full-length proteins were expressed and purified as above and eluted material was analysed by SDS-PAGE. The samples are relatively pure after only one step of purification, although degradation products are sometimes present. Molecular weight markers and their sizes are indicated on both sides of the gel. Successful constructs are: Atg17 (48.7 kDa), Dug2 (98.1 kDa), Dug3 (40.2 kDa), Gcd10 (54.4 kDa), Gcd14 (43.9 kDa), Met12 (73.9 kDa), Met13 (68.6 kDa), Psy2 (98.1 kDa), Rbg2 (41 kDa), Gir2 (31 kDa), Ssl2 (95.3 kDa), Yor352w (39.3 kDa), Vps27 (71.9 kDa), Hse1 (51.1 kDa), while the unsuccessful constructs are: Atg20 (72.5 kDa) and Psy4 (50.7 kDa).

Panel B: The nine complexes successfully produced in a recombinant form. Ni\(^{2+}\)-NTA-purified samples of the results of complex formation trials were subjected to SDS-PAGE analysis and visualised using Coomassie. Co-expressed or reconstituted forms of the Gcd10:Gcd14 (54.4 and 43.9 kDa, respectively), Paa1:Ptc2 (21.9 and 50.3 kDa), Met12:Met13 (73.9 and 68.6 kDa), Dug2:Dug3 (98 and 40.2 kDa), Ssl2:Yor352w (95.2 and 40.2 kDa), Hbs1:Dom34 (68.7 and 44.1 kDa), Vps27:Hse1 (71.9 and 51.1 kDa), Gir2:Rbg2 (31 and 41 kDa), Dug2:Dug3 (98.1 and 40.2 kDa), Rps28B:Edc3 (7.6 and 61.3 kDa) complexes. Bands corresponding to the proteins of interest are arrowed. See also Figure S2 for detailed results of expression and reconstitution of complexes.

Figure 3: Validation and scale-up of an exemplary complex; Gcd10:Gcd14
A) Both Gcd10 and Gcd14 were clearly visible after purification using the TAP protocol, with little evidence of contaminating proteins, validating this complex. B) In order to estimate the size of the complexes, yeast extracts were separated with the use of size exclusion chromatography on a Superdex 200 column in low (150 mM; marked as ‘LO’) and high (500 mM; ‘HI’) concentration of NaCl. 30 fractions from this chromatography step were collected and spotted on a nitrocellulose membrane. To detect fractions containing the TAP tagged protein, western blotting using PAP antibodies was performed. See the legend to Figure S1 for further details to panels A and B. C) Micrograph of the Gcd10:Gcd14 complex which had been purified as in Figure 2, panel B, and fixed with glutaraldehyde, according to the GraFix protocols (Kästner et al., 2008) and stained with uranyl-acetate in a sandwich between two layers of carbon. The length of the scale bar equals 50 nm. D) Class averages of the data (top row) determined by multi-statistical analysis agree with projections of the 3D-map (central row). Surface presentations (bottom row) of the 3D-map are shown in the same directions as the projections above. The length of the scale bar equals 5 nm. E) Image reconstruction of the Gcd10/Gcd14 complex. C2 symmetry was imposed during the final rounds of refinement. The complex is shown along the symmetry axis (left) and perpendicular to the symmetry axis (right). The length of the scale bar equals 5 nm.
Tables

Table 1: Summary of target selection, validation and complex reconstitution results. Complexes selected by bioinformatics, gel and literature analyses respectively, are listed. The complexes were assessed according to their purity after TAP purification (column labeled “Gel quality”). The ranks according to the CF algorithm of each of the complexes (“Rank”), as well as the results of validation by tandem affinity purification (c.f. Figure S1; “TAP Validation” and Figure S2 for the results of complex production and Table S7) are shown. Results of expression, co-expression and reconstitution studies are as follows: +; successful, - ; unsuccessful, ND; not determined, NA; not applicable. *; Few of the complexes consist of 3 or more subunits. ‡The Gcd10:Gcd14 complex was not reconstituted from purified proteins, but instead cells in which the proteins had been expressed separately were combined prior to sonication. For clarity, results that were deemed to be ‘positive’ (having a ‘good’ gel quality, high ranking in the bioinformatics triage, significant expression levels or production of the relevant complex by either co-expression or by reconstitution) are shown with a green background. Similarly, ‘mediocre’ results in the TAP validation (indicating that either heterogenous or partial complexes were purified) are shown with a yellow background. Negative results, indicating either a poor gel quality, low bioinformatics rank, failed TAP validation experiment, failed expression or failed complex production, are shown in red. Expression results for the complexes not deemed to be suitable for structural analysis are shown as gray text. ‡Reconstitution of the Dom34:Hbs1 complex is described previously (Graille et al., 2008).
<table>
<thead>
<tr>
<th>Complex</th>
<th>Stage</th>
<th>Selection</th>
<th>Validation</th>
<th>Single subunit expression</th>
<th>Complex production</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gel Quality</td>
<td>Rank</td>
<td>TAP Results</td>
<td>Subunit 1</td>
</tr>
<tr>
<td>Ste11, Ste50</td>
<td>Bioinformatics Analysis</td>
<td>Good</td>
<td>1</td>
<td>Heterogeneous</td>
<td>+</td>
</tr>
<tr>
<td>Atg17, Atg20, Atg29</td>
<td>Good</td>
<td>27</td>
<td>Partial (Atg29 missing)</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Vps27, Hse1</td>
<td>Excellent</td>
<td>1</td>
<td>Excellent</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Psy2, Psy4, Pph3</td>
<td>Excellent</td>
<td>4</td>
<td>Partial (Pph3 missing)</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Nup82, Nup159, Nsp1</td>
<td>Good</td>
<td>4</td>
<td>Excellent</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Ede1, Sypl</td>
<td>Good</td>
<td>22</td>
<td>Excellent</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Dop1, Mon2</td>
<td>Excellent</td>
<td>25</td>
<td>Aggregate</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Gel Analysis</td>
<td>Gcd14, Gcd10</td>
<td>Excellent</td>
<td>12</td>
<td>Excellent</td>
<td>+</td>
</tr>
<tr>
<td>Ptc2, Paa1</td>
<td>Excellent</td>
<td>8</td>
<td>Partial</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Met12, Met13</td>
<td>Excellent</td>
<td>22</td>
<td>Excellent</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dug3, Dug2</td>
<td>Excellent</td>
<td>9</td>
<td>Excellent</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Sa2, Yor352w</td>
<td>Excellent</td>
<td>27</td>
<td>Excellent</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Spt6, Spn1</td>
<td>Excellent</td>
<td>40</td>
<td>Partial (Spn1 missing)</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Literature Analysis</td>
<td>Rad17, Mec3, Dcd1</td>
<td>Failed</td>
<td>261</td>
<td>Failed</td>
<td>ND</td>
</tr>
<tr>
<td>Orc1-6</td>
<td>Good</td>
<td>29</td>
<td>Heterogeneous</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Rho2, G12</td>
<td>Good</td>
<td>5</td>
<td>Excellent</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Dom34, Hbs1</td>
<td>No Interaction</td>
<td>364</td>
<td>No Interaction</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Rps28B, Eco3</td>
<td>No Interaction</td>
<td>364</td>
<td>No Interaction</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Sis2, Ykl088w, Vhs3</td>
<td>Partial Interaction</td>
<td>323</td>
<td>Weak expression</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Mwa1, Don1, Nid1, Nsl1</td>
<td>Partial Interaction</td>
<td>11</td>
<td>Weak expression</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>
Methods

Validation

TAP purification

TAP tagged strains of *Saccharomyces cerevisiae* were grown in 4 l of YPD medium (1% yeast extract, 1% bacto-peptone, 2% glucose) to an optical density (O.D.) of approximately 2. Yeast pellets were resuspended in 40 ml of lysis buffer (1 mM DTT, 40 mM Hepes pH 8, 250 mM NaCl) and frozen in liquid nitrogen. Cells were broken in a laboratory blender cooled with dry ice. Extracts were defrosted with protease inhibitors and spun in 35Ti rotor (Beckman) in a Beckman ultracentrifuge at 20,000 rpm for 20 minutes at 4°C. Supernatant was spun again at 32,000 rpm for 90 minutes at 4°C. Resulting extracts were dialyzed in buffer D (1 mM DTT, 40 mM Hepes pH 8, 150 mM NaCl, 1 mM PMSF) and frozen in liquid nitrogen. Extracts were then defrosted and incubated with 200 µl of IgG Sepharose 6 Fast Flow resin (GE Healthcare) in the presence of 0.1% rTX-100 for 1.5 hours at 4°C. The beads were washed twice with 10 ml IPP150 (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.1% rTX100) and twice with 10 ml TEV cleavage buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5mM EDTA, 1 mM DTT). TEV cleavage was performed for 2 hours using 20 µg of TEV protease in 300 µl of cleavage buffer at room temperature. Eluates were agitated with 300 µl of calmodulin beads suspension (Stratagene) for 0.5 hours at 4°C. The beads were washed four times with 500 µl of calmodulin wash buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 10 mM β-mercaptoethanol, 1mM CaCl₂) and the protein was eluted with 0.6 ml calmodulin elution buffer (10 mM Tris-HCl pH 8.0, 500 mM NaCl, 10 mM β-mercaptoethanol, 0.1% rTX100, 4mM EDTA). As a control, denatured elution fractions from both IgG and calmodulin beads were prepared with 250µl of 1% SDS at 60°C.

Protein precipitation and analysis by mass spectrometry

Proteins were precipitated using pyrogallol red(Aguilar et al., 1999). When salinity of buffer was higher then 200 mM of NaCl the samples were first adjusted to this concentration by dilution. Proteins were separated by electrophoresis performed on NuPAGE 4-12% gradient gels using MES buffer gel system (Invitrogen) and
stained with SimplyBlue SafeStain (Invitrogen). Mass spectrometry was performed both with IgG eluates in solution and from bands cut from gels. Samples were then processed by standard procedures with trypsin digestion and cysteine alkylation. The obtained peptide mixtures were separated on a nano-HPLC system and the column outlet was coupled to the ion source of an LTQ FTICR spectrometer.

**Western blot analyses**

After dialysis, extracts and flow-throughs after IgG Sepharose chromatography were separated by 10% SDS-PAGE and electro-blotted onto the Protran nitrocellulose membrane (Bioscience) using a Trans-Blot® system (Bio-Rad). The filters were blocked for 1 h in 5% milk powder in PBS containing 0.1% Tween-20 and then the mouse monoclonal anti-rabbit immunoglobulin–peroxidase conjugate (Sigma) diluted 3,000-fold was added. After one hour, the blots were washed three times in PBS with 0.1% Tween-20. Finally, horseradish peroxidase conjugates were visualized by enhanced chemi-luminescence system (ECL, GE Healthcare).

**Mass determination of the complexes**

In order to estimate the size of the purified complex the extract from TAP-tagged strains was separated according to size, by size exclusion chromatography on a Superdex 200 10/300 column (GE Healthcare) using an Akta Purifier FPLC. Two different salt concentrations (150 mM and 500 mM NaCl) were used for elution and fractions were collected into a 96 well plate. 60 μl of every fraction were spotted on a nitrocellulose membrane. TAP tagged subunits were detected by Dot-Blot as described for western blot analyses. The intensities of the spots were calculated with ImageQuant (GE Healthcare) and exported into chromatograms. The column was calibrated using protein markers; thyroglobulin (670 kDa), ferritin (440 kDa), catalase (232 kDa), aldolase (154 kDa), albumin (67 kDa), ovalbumin (43 kDa) and chymotrypsin (25 kDa).

**Electron Microscopy and Image Processing**

The purified, over-expressed Gcd10/Gcd14 complex was fixed on a glycerol gradient with glutaraldehyde according to the GraFix protocol(Kästner et al., 2008). Fractions of the gradient were further analyzed by dot-blot analysis using an antibody against the 6-histidine tag. The dot blot identified a single peak with a maximum at
fraction 14. Samples from the peak fractions were prepared for subsequent electron microscopy by sandwich negative stain using uranyl acetate as previously described (Ulbrich et al., 2009). Samples were imaged at room temperature in a Philips CM120 Biotwin electron microscope at 100 kV. Data was recorded on a 4kx4k Tietz-CCD camera at a nominal pixel size of 4.27 Å per pixel under low dose conditions. For further processing 10819 particle images were selected from 29 micrographs. Three-dimensional models were calculated using sinogram correlation and weighted back projection with IMAGIC 5 (van Heel et al., 1996). The process of determining initial orientations followed by calculation of a three-dimensional map was repeated several times using different class averages for starting the sinogram correlation.

Projections of the resulting three-dimensional models were compared with the initial class averages. The model that generated projections that matched most of the initial class averages, was selected for further refinement by an iterative process of projection matching followed by calculating a new 3D-map with Spider (Frank et al., 1996). After five rounds of refinement the map was stable and showed an approximately fourfold-symmetric core with extensions at opposite sides, giving the whole map a 2-fold symmetric appearance. Therefore, the map was refined for another five rounds imposing C2-symmetry. The resolution of the final map was determined by Fourier-Shell-Correlation and was 23 Å (Correlation=0.5).
References


Figure 2, Brooks et al.

A

B
## Inventory of Supplemental Information

<table>
<thead>
<tr>
<th>Name</th>
<th>Summary</th>
<th>Related element in main text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table S1</td>
<td>Bioinformatics triage: Non-redundant and manually curated set of 39 distinct yeast complexes of known 3D structure.</td>
<td>Table 1</td>
</tr>
<tr>
<td>Table S2</td>
<td>Bioinformatics triage: Final weights and effects of each parameter on the final selection.</td>
<td>Table 1</td>
</tr>
<tr>
<td>Table S3</td>
<td>Bioinformatics triage: Results of bioinformatics triage of Gavin et al complexes.</td>
<td>Table 1</td>
</tr>
<tr>
<td>Table S4</td>
<td>Manual triage: Lists of stoichiometric complexes identified by visual inspection of gels.</td>
<td>Table 1</td>
</tr>
<tr>
<td>Figure S1</td>
<td>Experimental validation of complexes.</td>
<td>Figure 1</td>
</tr>
<tr>
<td>Figure S2</td>
<td>Expression &amp; Solubility Trials</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Figure S3</td>
<td>Paa1-Ptc2 Complex Reconstitution</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Figure S4</td>
<td>Vps27-Hse1 Complex Reconstitution</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Figure S5</td>
<td>Gcd10-Gcd14 Complex Reconstitution</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Figure S6</td>
<td>Ste11-Ste50 Complex Reconstitution</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Figure S7</td>
<td>Effects of expression using synthetic genes using codon optimization relative to naturally occurring genes</td>
<td>Figure 2</td>
</tr>
<tr>
<td>Supp. Experimental Procedures</td>
<td></td>
<td>Experimental Procedures</td>
</tr>
<tr>
<td>Supp. References</td>
<td></td>
<td>References specific to the Supplementary Material</td>
</tr>
</tbody>
</table>
Supplemental Data

Supplemental Tables

**Supplemental Table 1**: Non-redundant and manually curated set of 39 distinct yeast complexes of known 3D structure.

Dotted horizontal lines indicate the thresholds corresponding to the complexes in the top 10 and top 50, after weight optimizations. Columns two and three show the accession codes of the corresponding PDB entries and the ORF IDs of the different yeast proteins in the given complex, respectively. Column 4 contains the feasibility score of the complex when ranking the 39 yeast complexes of known 3D structure together with the 491 complexes defined by Gavin *et al.*, 2006, and column 5 depicts the corresponding rank.
<table>
<thead>
<tr>
<th>Complex description</th>
<th>PDB entries</th>
<th>Yeast ORF IDs</th>
<th>Feasibility score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystal structure of the yeast kinetochore Spc24/Spc25 globular domain</td>
<td>2FTX 2FV4</td>
<td>YER018C YMR117C</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>Elongation factor complex EEF1A:EEF1BA</td>
<td>H60 1G7C 1JJE 1JF</td>
<td>YAL003W YBR118W</td>
<td>91.74</td>
<td>3</td>
</tr>
<tr>
<td>RabGDP-dissociation inhibitor in complex with prenylated YPT1 GTPase</td>
<td>1UKV 2BCG</td>
<td>YER136W YFL038C</td>
<td>91.55</td>
<td>5</td>
</tr>
<tr>
<td>Mediator MED7/MED21 subcomplex</td>
<td>1YKE</td>
<td>YDR308C YOL135C</td>
<td>91.48</td>
<td>6</td>
</tr>
<tr>
<td>Sec23/24 heterodimer</td>
<td>1M2V</td>
<td>YIL109C YPR181C</td>
<td>90.76</td>
<td>7</td>
</tr>
<tr>
<td>20S proteasome</td>
<td>1G0U 1G65 1JD2</td>
<td>YBL041W YER012W YER094C</td>
<td>89.07</td>
<td>12</td>
</tr>
<tr>
<td>Crystal structure of the yeast kinetochore Spc24/Spc25 globular domain</td>
<td>2FTX 2FV4</td>
<td>YER018C YMR117C</td>
<td>92</td>
<td>2</td>
</tr>
<tr>
<td>20S proteasome</td>
<td>1RYP 2F16 2FNY</td>
<td>YFR050C YGL011C YGR135W YGR253C YJL001W YML092C YMR314W YOL038W YOR157C YOR362C YPR103W</td>
<td>87.22</td>
<td>16</td>
</tr>
<tr>
<td>Structure of a Vps23-C:Vps28-N subcomplex</td>
<td>2F6M</td>
<td>YCL008CYPL065W</td>
<td>84.49</td>
<td>25</td>
</tr>
<tr>
<td>Mitochondrial processing peptidase</td>
<td>1HR6 1HR7 1HR8</td>
<td>YHR024C YLR163C</td>
<td>83.65</td>
<td>29</td>
</tr>
<tr>
<td>Carboxypeptidase Y inhibitor complexed with the cognate proteasome</td>
<td>1WPX</td>
<td>YLR178C YMR297W</td>
<td>83.2</td>
<td>31</td>
</tr>
<tr>
<td>AHA1/HSP90 complex</td>
<td>1USU 1USV</td>
<td>YDR214W YPL240C</td>
<td>82.23</td>
<td>33</td>
</tr>
<tr>
<td>Eukaryotic clamp loader (RFC) bound to the DNA sliding clamp (PCNA)</td>
<td>1SXJ</td>
<td>YBR087W YBR088C YJR068W YNL206W YOL094C YOR217W</td>
<td>81.28</td>
<td>36</td>
</tr>
<tr>
<td>Mms2/Ubc13 ubiquitin conjugating enzyme complex</td>
<td>1JAT</td>
<td>YDR092W YGL087C</td>
<td>81.13</td>
<td>37</td>
</tr>
<tr>
<td>MLC1P bound to IQ2 of MYO2P</td>
<td>1M45 1N2D</td>
<td>YGL106W YOR326W</td>
<td>81</td>
<td>38</td>
</tr>
<tr>
<td>Heterodimer between H48F-ySOD1 and yCCS</td>
<td>1JK9</td>
<td>YJR104CYMR038C</td>
<td>79.97</td>
<td>42</td>
</tr>
<tr>
<td>TFIIA/TBP/DNA complex</td>
<td>1NH2 1RM1 1YTF</td>
<td>YER148W YOR194C</td>
<td>79.35</td>
<td>46</td>
</tr>
<tr>
<td>A conjugating enzyme/ubiquitin thiolester complex</td>
<td>1FXT</td>
<td>YDR177W YLR167W</td>
<td>79.1</td>
<td>47</td>
</tr>
<tr>
<td>Translation initiation factor eIF4E in complex with m7GDP and eIF4GI</td>
<td>1RF8</td>
<td>YGR162W YOL139C</td>
<td>78.9</td>
<td>48</td>
</tr>
<tr>
<td>ESCRT-II endosomal trafficking complex</td>
<td>1U5T 1W7P</td>
<td>YLR417W YPL002C YJR102C</td>
<td>78.89</td>
<td>49</td>
</tr>
<tr>
<td>Nucleosome core particle</td>
<td>1ID3</td>
<td>YBL002W YBR009C YBR010W YDL225W</td>
<td>78.46</td>
<td>50</td>
</tr>
<tr>
<td>SRP receptor beta-subunit in complex with the SRX domain from the alpha-subunit</td>
<td>1NRJ</td>
<td>YDR292C YKL154W</td>
<td>78.12</td>
<td>52</td>
</tr>
<tr>
<td>Brl1/TBP/DNA ternary complex</td>
<td>1NGM</td>
<td>YER148W YGR246C</td>
<td>78.04</td>
<td>54</td>
</tr>
<tr>
<td>Sec23/Sar1 complex</td>
<td>1M2O</td>
<td>YPL218W YPR181C</td>
<td>77.34</td>
<td>61</td>
</tr>
<tr>
<td>Dsk2p UBA/ubiquitin complex</td>
<td>1WR1</td>
<td>YIL148W YMR276W</td>
<td>76.77</td>
<td>63</td>
</tr>
<tr>
<td>A peptide:N-glycanase-Rad23 complex</td>
<td>1X3W 1X3Z</td>
<td>YEL037C YPL096W</td>
<td>74.66</td>
<td>75</td>
</tr>
<tr>
<td>Complex Description</td>
<td>PDB Code(s)</td>
<td>Sequence(s)</td>
<td>Percent Similarity</td>
<td>Residues</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>------------------------------------------------------------------------------</td>
<td>--------------------</td>
<td>----------</td>
</tr>
<tr>
<td>RNA polymerase II</td>
<td>1NT9, 1I50, 1H6H</td>
<td>YBR154C YDL140C YDR404C</td>
<td>72.96</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>1NIK, 1R5U, 1R9S</td>
<td>YGL070C YHR143W-A YIL021W</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1R9T, 1SFO, 1TWA</td>
<td>YJL140W YOL005C YOR151C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1TWC, 1TWF, 1TWG</td>
<td>YOR210W YOR224C YPR187W</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1TWH, 1I3Q, 1K83</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1WCM, 1Y1W, 1Y77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2B63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exportin CSE1P in complex with its cargo (KAP60P) and RanGTP</td>
<td>1WA5</td>
<td>YGL238W YLR293C YNL189W</td>
<td>72.24</td>
<td>91</td>
</tr>
<tr>
<td>RNA polymerase II/TFIIS complex</td>
<td>1PQV, 1Y1V, 1Y1Y</td>
<td>YBR154C YDL140C YDR404C</td>
<td>71.21</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>1PQV, 1Y1V, 1Y1Y</td>
<td>YGL043W YGL070C YHR143WA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1PQV, 1Y1V, 1Y1Y</td>
<td>YIL021W YJL140W YOL005C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1PQV, 1Y1V, 1Y1Y</td>
<td>YOR151C YOR210W YOR224C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1PQV, 1Y1V, 1Y1Y</td>
<td>YPR187W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATa2/MCM1/DNA ternary transcription complex</td>
<td>1MNM</td>
<td>YCL067C YMR043W</td>
<td>69.62</td>
<td>103</td>
</tr>
<tr>
<td>CUE/ubiquitin complex</td>
<td>1OTR</td>
<td>YIL148W YKL090W</td>
<td>68.32</td>
<td>119</td>
</tr>
<tr>
<td>Cytochrome BC1 complex</td>
<td>1EZV, 1KB9, 1P84</td>
<td>Q0105 YBL045C YDR529C</td>
<td>64.61</td>
<td>151</td>
</tr>
<tr>
<td></td>
<td>1EZV, 1KB9, 1P84</td>
<td>YEL024W YFR033C YGR183C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1EZV, 1KB9, 1P84</td>
<td>YJL166W YOR065W YPR191W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MATa1/MATalpha2-3A heterodimer bound to DNA</td>
<td>1AKH, 1LE8, 1YRN</td>
<td>VCL067C YCR097W</td>
<td>64.19</td>
<td>157</td>
</tr>
<tr>
<td>Cytochrome BC1 complex with bound substrate cytochrome C</td>
<td>1KOY</td>
<td>Q0105 YBL045C YDR529C</td>
<td>64.17</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>1KOY</td>
<td>YEL024W YFR033C YGR183C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1KOY</td>
<td>YJL166W YJR048W YOR065W</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1KOY</td>
<td>YPR191W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Electron transfer Complex between cytochrome C and cytochrome C peroxidase</td>
<td>1S6V, 1U74, 2IBOZ</td>
<td>YJR048W YKR066C</td>
<td>63.26</td>
<td>167</td>
</tr>
<tr>
<td></td>
<td>1S6V, 1U74, 2IBOZ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2B10, 2B11, 2B12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2BCN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-terminal ULP1 protease domain in complex with SMT3</td>
<td>1EUV</td>
<td>YDR510W YPL020C</td>
<td>56.86</td>
<td>218</td>
</tr>
<tr>
<td>YPD1/SLN1 response regulator domain complex</td>
<td>1OXB</td>
<td>YDL235C YIL147C</td>
<td>56.71</td>
<td>222</td>
</tr>
<tr>
<td>Solution Structure of Ede1 UBA-ubiquitin complex</td>
<td>2G3Q</td>
<td>YLR167W YBL047C</td>
<td>56.35</td>
<td>225</td>
</tr>
<tr>
<td>Lif1p/Lig4p complex</td>
<td>1Z56</td>
<td>YGL090W YOR005C</td>
<td>53.77</td>
<td>244</td>
</tr>
<tr>
<td>Orc1p/Sir1p complex</td>
<td>1ZBH, 1ZHI</td>
<td>YKR101W YML065W</td>
<td>52.94</td>
<td>253</td>
</tr>
</tbody>
</table>
**Supplemental Table 2**: Final weights and effects of each parameter on the final selection. *

<table>
<thead>
<tr>
<th>Partial scoring function</th>
<th>Default weight</th>
<th>Contribution to the final score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average socio-affinity index</td>
<td>1</td>
<td>30.3 %</td>
</tr>
<tr>
<td>Maximum individual protein weight</td>
<td>0.3</td>
<td>9.1 %</td>
</tr>
<tr>
<td>Total sequence length</td>
<td>0.1</td>
<td>3.0 %</td>
</tr>
<tr>
<td>Average number of problematic residues</td>
<td>0*</td>
<td>0 %</td>
</tr>
<tr>
<td>Co-localization ratio</td>
<td>0</td>
<td>0 %</td>
</tr>
<tr>
<td>Yeast two-hybrid ratio</td>
<td>1</td>
<td>30.3 %</td>
</tr>
<tr>
<td>Complete orthologs ratio</td>
<td>0.1</td>
<td>3.0 %</td>
</tr>
<tr>
<td>Average orthologs ratio</td>
<td>0.1</td>
<td>3.0 %</td>
</tr>
<tr>
<td>Self-consistency</td>
<td>0.4</td>
<td>12.1 %</td>
</tr>
</tbody>
</table>

Proteins with trans-membrane helices were excluded from the analysis.
**Supplemental Table 3:** Results of bioinformatics triage of Gavin et al complexes.

<table>
<thead>
<tr>
<th>Complex ID</th>
<th>Subunits</th>
<th>Rank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ste11, Ste50</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Atg17, Atg29, Atg11</td>
<td>27</td>
<td>85.96</td>
</tr>
<tr>
<td>3</td>
<td>Vps27, Hse1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>Psy2, Psy4, Pph3</td>
<td>4</td>
<td>99.17</td>
</tr>
<tr>
<td>5</td>
<td>Nup82, Nup159, Nsp1</td>
<td>4</td>
<td>99.17</td>
</tr>
<tr>
<td>6</td>
<td>Ede1, Syp1</td>
<td>22</td>
<td>87.14</td>
</tr>
<tr>
<td>7</td>
<td>Dop1, Mon2</td>
<td>25</td>
<td>86.43</td>
</tr>
</tbody>
</table>
Supplemental Table 4A: List of stoichiometric dimeric complexes identified by visual inspection of gels.

<table>
<thead>
<tr>
<th>No.</th>
<th>Subunits</th>
<th>Gels (with hyperlinks)</th>
<th>No.</th>
<th>Subunits</th>
<th>Gels (with hyperlinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gsy1</td>
<td>SC-PG-235-SC23509(1-3)</td>
<td>42</td>
<td>Rnr2</td>
<td>SC-PG-213-SC2034(1-5)</td>
</tr>
<tr>
<td>3</td>
<td>Trr1</td>
<td>SC-PG-291-SC3169(1-4)</td>
<td>44</td>
<td>Gyp1</td>
<td>SC-PG-175-SC593(1-5)</td>
</tr>
<tr>
<td>4</td>
<td>Trr2</td>
<td>SC-PG-226-SC2187(1-6)</td>
<td>45</td>
<td>Trf1</td>
<td>SC-PG-235-SC2246(1-7)</td>
</tr>
<tr>
<td>5</td>
<td>Gdc14</td>
<td>SC-PG-260-SC2835(1-4)</td>
<td>46</td>
<td>Nkp1</td>
<td>SC-PG-305-SC3228(1-7)</td>
</tr>
<tr>
<td>8</td>
<td>Svl1</td>
<td>SC-PG-365-SC2732(3-3)</td>
<td>49</td>
<td>Nma1</td>
<td>SC-PG-274-SC2907(1-7)</td>
</tr>
<tr>
<td>9</td>
<td>Met12</td>
<td>SC-PG-277-SC2455(1-4)</td>
<td>50</td>
<td>Nrd1</td>
<td>SC-PG-249-SC362(1-6)</td>
</tr>
<tr>
<td>10</td>
<td>Sxm41</td>
<td>SC-PG-364-SC3020(1-9)</td>
<td>51</td>
<td>Cap2</td>
<td>SC-PG-232-SC2908(1-6)</td>
</tr>
<tr>
<td>11</td>
<td>Sxm4</td>
<td>SC-PG-459-SC5304(4-2)</td>
<td>52</td>
<td>Ptc2</td>
<td>SC-PG-248-SC1857(1-1)</td>
</tr>
<tr>
<td>12</td>
<td>Trm7</td>
<td>SC-PG-336-SC3724(1-3)</td>
<td>53</td>
<td>Cap3</td>
<td>SC-PG-241-SC3099(1-8)</td>
</tr>
<tr>
<td>13</td>
<td>Ymr25</td>
<td>SC-PG-326-SC3504(1-5)</td>
<td>54</td>
<td>Ynl119</td>
<td>SC-PG-211-SC2033(1-6)</td>
</tr>
<tr>
<td>14</td>
<td>Ymr25</td>
<td>SC-PG-447-SC3979(1-9)</td>
<td>55</td>
<td>Ynl119</td>
<td>SC-PG-402-SC1011(1-5)</td>
</tr>
<tr>
<td>15</td>
<td>9c</td>
<td>SC-PG-283-SC2908(1-1)</td>
<td>56</td>
<td>Ynl119</td>
<td>SC-PG-213-SC5408(1-3)</td>
</tr>
<tr>
<td>16</td>
<td>Ynl11</td>
<td>SC-PG-359-SC5304(2-3)</td>
<td>57</td>
<td>Ynl119</td>
<td>SC-PG-210-SC2001(1-1)</td>
</tr>
<tr>
<td>18</td>
<td>Ynl345</td>
<td>SC-PG-389-SC2079(6-4)</td>
<td>59</td>
<td>Ynl119</td>
<td>SC-PG-413-SC5018(1-3)</td>
</tr>
<tr>
<td>19</td>
<td>w</td>
<td>SC-PG-366-SC4106(1-2)</td>
<td>60</td>
<td>Ynl119</td>
<td>SC-PG-413-SC5021(1-5)</td>
</tr>
<tr>
<td>20</td>
<td>Ynl11</td>
<td>SC-PG-422-SC47511(1-5)</td>
<td>61</td>
<td>Ynl119</td>
<td>SC-PG-414-SC5032(1-2)</td>
</tr>
<tr>
<td>21</td>
<td>Ynl11</td>
<td>SC-PG-423-SC4814(1-3)</td>
<td>62</td>
<td>Ynl119</td>
<td>SC-PG-413-SC5032(1-2)</td>
</tr>
<tr>
<td>22</td>
<td>Ynl11</td>
<td>SC-PG-314-SC3454(1-4)</td>
<td>63</td>
<td>Ynl119</td>
<td>SC-PG-115-SC1438(1-3)</td>
</tr>
<tr>
<td>23</td>
<td>Ynl11</td>
<td>SC-PG-330-SC2395(3-2)</td>
<td>64</td>
<td>Ynl119</td>
<td>SC-PG-321-SC3088(1-7)</td>
</tr>
<tr>
<td>24</td>
<td>Ynl11</td>
<td>SC-PG-374-SC3088(1-7)</td>
<td>65</td>
<td>Ynl119</td>
<td>SC-PG-315-SC3602(1-9)</td>
</tr>
<tr>
<td>25</td>
<td>Ynl11</td>
<td>SC-PG-336-SC3602(1-9)</td>
<td>66</td>
<td>Ynl119</td>
<td>SC-PG-313-SC3442(1-5)</td>
</tr>
<tr>
<td>26</td>
<td>Ynl11</td>
<td>SC-PG-389-SC2079(6-4)</td>
<td>67</td>
<td>Ynl119</td>
<td>SC-PG-3020-SC2081(1-5)</td>
</tr>
<tr>
<td>27</td>
<td>Ynl11</td>
<td>SC-PG-366-SC4106(1-2)</td>
<td>68</td>
<td>Ynl119</td>
<td>SC-PG-320-SC2081(1-5)</td>
</tr>
<tr>
<td>28</td>
<td>Ynl11</td>
<td>SC-PG-341-SC3099(1-8)</td>
<td>69</td>
<td>Ynl119</td>
<td>SC-PG-322-SC47511(1-5)</td>
</tr>
<tr>
<td>29</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>70</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>30</td>
<td>Ynl11</td>
<td>SC-PG-389-SC2079(6-4)</td>
<td>71</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>31</td>
<td>Ynl11</td>
<td>SC-PG-366-SC4106(1-2)</td>
<td>72</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>32</td>
<td>Ynl11</td>
<td>SC-PG-341-SC3099(1-8)</td>
<td>73</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>33</td>
<td>Ynl11</td>
<td>SC-PG-322-SC47511(1-5)</td>
<td>74</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>34</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>75</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>35</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>76</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>36</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>77</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>37</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>78</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>38</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>79</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>40</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>80</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
<tr>
<td>41</td>
<td>Ynl11</td>
<td>SC-PG-323-SC3979(1-9)</td>
<td>81</td>
<td>Ynl119</td>
<td>SC-PG-323-SC3979(1-9)</td>
</tr>
</tbody>
</table>

Note: The table includes a list of stoichiometric dimeric complexes identified by visual inspection of gels. The complexes are identified by their subunits and the gels they are found in.
**Supplemental Table 4B:** List of stoichiometric trimeric complexes identified by visual inspection of gels.

<table>
<thead>
<tr>
<th>No.</th>
<th>Subunits</th>
<th>Gels (with hyperlinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.</td>
<td>Arc1</td>
<td>Mes1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55.</td>
<td>Tef1/Tef2</td>
<td>Cam1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56.</td>
<td>Lat1</td>
<td>Pda1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57.</td>
<td>Nup82</td>
<td>Nsp1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58.</td>
<td>Hat1</td>
<td>Hat2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59.</td>
<td>Tps1</td>
<td>Tps1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Supplemental Table 4C:** List of stoichiometric tetrameric complexes identified by visual inspection of gels.

<table>
<thead>
<tr>
<th>No.</th>
<th>Subunits</th>
<th>Gels (with hyperlinks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.</td>
<td>Cka1</td>
<td>Cka2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61.</td>
<td>Stp2</td>
<td>Ygr206w</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>62.</td>
<td>Spc2</td>
<td>Spc24</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>63.</td>
<td>Rlr1</td>
<td>Hpr1</td>
</tr>
<tr>
<td></td>
<td>Sec6</td>
<td>Sec66</td>
</tr>
<tr>
<td>64.</td>
<td></td>
<td>Sec63</td>
</tr>
</tbody>
</table>
Supplemental Figures

Supplemental Figure 1:

Bioinformatics Ste11 - 80kDa  Ste50 - 39kDa

Conclusion: Proteins interact but form heterogeneous complexes.
Supplemental Figure 1:

Bioinformatics

Atg17 - 49kDa  Atg29 - 25kDa  Atg11 - 135kDa

Conclusion:
Atg17 Atg20 dimmer

There are stable interactions between Atg17 and Atg29 but Atg11 is not present or highly substoichiometric. Apparent MW about 500 kDa suggesting stoichiometry higher than 1:1 or elongated shape.

ATG11 - missing

Atg29 - no signal from GF
Supplemental Figure 1:

**Bioinformatics** Vps27 - 71kDa  Hse1 - 51kDa

**Conclusion:**
Complex exists

Apparent MW about 350 kDa suggesting stoichiometry higher than 1:1 or elongated shape.

1. IgG SDS eluate
2. Calmodulin SDS eluate
3. IgG eluate
4. Calmodulin flow
5. Calmodulin eluate
M - Marker

**Vps27**
1. VPS27 Score: 371
2. VPS27 Score: 419
3. Hse1p Score: 237

**Hse1**
1. VPS27 Score: 419
2. VPS27 Score: 816
3. Hse1p Score: 603
4. VPS27 Score: 745
5. Hse1p Score: 228

**Vps27 tagged**

**Hse1 tagged**
Supplemental Figure 1:

**Bioinformatics**

*Psy2* - 98kDa  *Psy4* - 51kDa  *Pph3* - 35kDa

**Conclusion:**

Complex exists

*Pph3* seems to be substoichiometric. Gel filtration is not conclusive.

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M-Marker

---

**Psy2**

1- Psy2 Score: 374
   IgG eluate - Psy2 Score: 672
   Psy4 Score: 281
   PPH3 Score: 176

2- Psy4 Score: 466
   IgG eluate - missing

---

**Psy4**

1- Psy2 Score: 466
   Psy4 Score: 502
   IgG eluate

---

**Pph3**

1- Pph3 Score: 406
2- Psy4 Score: 164
3- PPH3 Score: 136
4- IgG eluate - Psy2 Score: 249
5- PPH3 Score: 151

---

**Psy2 tagged**

67kDa Thym (Exclusion volume)
440kDa Ferritin
232kDa Catalase
67kDa Albumin
43kDa Outilbin
25kDa Chymotrypsin

---

**Psy4 tagged**

67kDa Thym (Exclusion volume)
440kDa Ferritin
154kDa 42kDa
25kDa Chymotrypsin

---

**Pph3 tagged**

67kDa Thym (Exclusion volume)
440kDa Ferritin
154kDa 42kDa
25kDa Chymotrypsin
Supplemental Figure 1:

Bioinformatics

Nup82 - 82kDa  Nup159 - 159kDa  Nsp1 - 86kDa

Conclusion:
Complex exists  Apparent MW above 700kDa.

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG elute
4- Calmodulin flow
5- Calmodulin elute
M- Marker

Nup82 tagged

Nup159 tagged

Nsp1 tagged
Supplemental Figure 1:

Bioinformatics Ede1 - 151kDa, Syp1 - 96kDa

Conclusion:
Complex exists
Ede1 is partially degraded
Apparent molecular weight about 300kDa
stoichiometry 1:1

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

Syp1 missing

Ede1
1- Ede1p Score: 2154
2- Ede1p Score: 1566
Syp1p Score: 69
3- Ede1p Score: 740
IgG eluate -
Ede1p Score: 6614
Syp1p Score: 1639

- extract
- IgG flow
- M- Marker

- LQ salt
- H1 salt

- 670kDa
  - Thyrog (Exclusion volume)
- 232kDa
  - Catalase
- 67kDa
  - Albumin
- 43kDa
  - Ovoalbumin
- 25kDa
  - Chymotrypsin

Ede1 tagged
Supplemental Figure 1:

**Bioinformatics** Dop1 - 195kDa  Mon2 - 186kDa

**Conclusion:**
Proteins interact but aggregates.
There are interactions but proteins aggregate on resins making them difficult targets for structural studies. There is no peak on gel filtration. Majority is in exclusion volume and proteins smear toward lower masses.

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

1- extract
2- IgG flow
M- Marker

Dop1 tagged

1- Dop1p Score: 187
   IgG eluate - Dop1p Score: 566
2- Dop1p Score: 162
   Mon2p Score: 335
3- Dop1p Score: 89
   Mon2p Score: 74
4- Calmodulin flow
   Dop1p Score: 112
   Mon2p Score: 101
5- Calmodulin eluate

Mon2 tagged

67kDa Albumin
44kDa Ferritin
25kDa Chymotrypsin
43kDa Ovalbumin
23kDa Catalase
67kDa Albumin
154kDa Aldolase
44kDa Ferritin
67kDa Albumin
43kDa Ovalbumin
25kDa Chymotrypsin

1- extract
2- IgG flow
M- Marker
Supplemental Figure 1:

Gel analysis Gcd14 - 44kDa  Gcd10 - 54kDa

Conclusion:
Complex exists.
Apparent MW about 350 kDa suggesting stoichiometry higher than 1:1 or elongated shape.

GDC10 - missing
Supplemental Figure 1:

Gel analysis Ptc2 - 50kDa  Paa1 - 22kDa

Conclusion:
Paa1 interacts with protein phosphatases Ptc2 Ptc3 or Ptc4 forming complex which looks quite stoichiometric on the SDS PAGE. In contrast Ptc2 is mostly monomeric. Gel filtration is not conclusive.
Supplemental Figure 1:

Gel analysis Met12 - 74kDa  Met13 - 69kDa

Conclusion:
Complex exists
Apparent molecular weight about 150kDa suggesting stoichiometry 1:1

IgG SDS eluate
1- IgG eluate
2- Calmodulin flow
3- Calmodulin eluate

Met12
1- Met12 Score: 1711
2- Met13 Score: 1097
IgG eluate -
Met12 Score: 1015
Met13 Score: 1064

IgG affinity chromatography followed by GF on Superdex200 column.
Complex is not abundant but migrates as a clean peak.
Supplemental Figure 1:

Gel analysis Dug3 - 40kDa Dug2 - 98kDa

Conclusion:
Complex exists.
Purification from Dug2 TAP strain was unsuccessful. Peaks are broad suggesting heterogeneity.

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M-Marker

Dug2 - no signal
1- extract
2- IgG flow
M-Marker

Dug3 tagged
Dug2 tagged
Supplemental Figure 1:

Gel analysis Ssl2 - 95kDa Yor352w - 39kDa

Conclusion:
Complex exists  Apparent molecular weight about 250kDa

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

Ssl2
1- Ssl2 Score: 593
2- IgG eluate - Ssl Score 464
3- Yor352W Score: 682
4- Yor352W Score: 818
5- Yor352W Score: 267

Yor352w
1- extract
2- IgG flow
M- Marker

Ssl2 tagged

Yor352w tagged
Supplemental Figure 1:

Gel analysis Spt6 - 168kDa, Spn1 - 46kDa

**Conclusion:**
Complex exists
Iws1 seems to be not stoichiometric

**Spt6 missing**

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

**Spt6**

1- Spt6 Score: 4707
2- Spn1 Score: 1041

IgG eluate -
Spt6 Score: 3601
ISpn1 Score: 596
Supplemental Figure 1:

Bibliography
Rad17 - 46kDa  Mec3 - 53kDa  Dcd1 - 36kDa

Conclusion:
Purification failed
Only for Mec3 tagged subunit is detected in IgG eluate

Dcd1 - no signal
Rad17 - no signal

Mec3 tagged
Supplemental Figure 1:

Bioinformatics

Orc1 - 104kDa  Orc2 - 71kDa  Orc3 - 72kDa  Orc4 - 61kDa  Orc5 - 55kDa  Orc6 - 50kDa

1- extract
2- IgG flow
M- Marker

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

Orc3 - missing
Supplemental Figure 1:

**Bioinformatics**

- Orc1 - 104kDa
- Orc2 - 71kDa
- Orc3 - 72kDa
- Orc4 - 61kDa
- Orc5 - 55kDa
- Orc6 - 50kDa

**Conclusion:**
Proteins interact but form heterogeneous assemblies
Gel filtration is not conclusive
Supplemental Figure 1:

Bibliography Rbg2 - 41kDa Gir2 - 31kDa

Conclusion:
Complex exists
On gel looks fine but there is a broad peak on gel filtration 100 - 250kDa Rbg2 migrates aberrantly high on SDS PAGE

1- IgG SDS eluate
2- Calmodulin SDS eluate
3- IgG eluate
4- Calmodulin flow
5- Calmodulin eluate
M- Marker

1- extract
2- IgG flow

Rbg2 tagged

Gir2 tagged
Supplemental Figure 1:

Bibliography Dom34 - 44kDa  Hbs1 - 69kDa

Conclusion:
No evidence of existence of stable complex in vivo
Dom34 and Hbs1 purify as single proteins however there are data proving direct interactions in vitro.
Supplemental Figure 1:

Edc3 forms heterogeneous complexes with Dcp1 and Dcp2. No indication for interactions between Edc3 and Rps28. Edc1 migrates in gel filtration as a very broad peak suggesting existence of heterogeneous complexes.

Week signal from Rps28B - ribosomal protein
Supplemental Figure 1:

Conclusion:
Proteins interact but are very weakly expressed and detected only in IgG eluate from Ykl088w tagged. Gel filtration suggest several assemblies.

Bibliography
Sis2 - 62kDa  Ykl088w - 65kDa  Vhs3 - 74kDa
Supplemental Figure 1:

**Bibliography**
- Mtw1 - 33kDa
- Dsn1 - 66kDa
- Nnf1 - 24kDa
- Nsl1 - 25kDa

**Conclusion:**
Proteins interact but are very weakly expressed and detected only in IgG eluate.
Supplemental Figure 2: Expression & Solubility Trials

Supplemental Figure 3: Paa1-Ptc2 Purification & Complex Reconstitution Experiment
Supplemental Figure 4: Vps27-Hse1 Purification & Complex Reconstitution Experiment

Supplemental Figure 5: Gcd10-Gcd14 Purification & Complex Reconstitution Experiment
Supplemental Figure 6: Ste11-Ste50 Purification & Complex Reconstitution Experiment
Supplemental Figure 7: Effects of expression using synthetic genes using codon optimization, relative to naturally occurring genes

A

<table>
<thead>
<tr>
<th>Construct Type</th>
<th>Codon Optimized Genes</th>
<th>Molecular weight (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gcd10</td>
<td>56.4 kDa</td>
<td>97.2, 66.4</td>
</tr>
<tr>
<td>Gcd14:6His</td>
<td>43.9 kDa</td>
<td>55.6, 55.6, 42.7, 34.6</td>
</tr>
<tr>
<td>Gcd14:6His</td>
<td>43.9 kDa</td>
<td>27, 20, 14</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Construct Type</th>
<th>Codon Optimized Genes</th>
<th>Molecular weight (kDa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ssl2</td>
<td>55.3 kDa</td>
<td>97.2, 66.4</td>
</tr>
<tr>
<td>6His:Yor352w</td>
<td>40.2 kDa</td>
<td>55.6, 55.6, 42.7, 34.6</td>
</tr>
<tr>
<td>6His:Yor352w</td>
<td>40.2 kDa</td>
<td>27, 20, 14</td>
</tr>
</tbody>
</table>
Supplemental Figure Legends

**Supplemental Figure 1: Experimental validation of complexes**

Experimental validation of the complexes was based on tandem affinity purification. To test the existence of a potential complex *in vivo*, we used several yeast strains; each with expression of the TAP tagged protein from the complex. Tandem affinity purification of TAP tagged proteins was prepared in native conditions, which allows for co-purification of all other proteins forming a complex. In the top left panel of each page, the results from independent purifications of all proteins forming a potential assemble are presented. Proteins from 5 fractions obtained during TAP purification (1; IgG SDS eluate, 2; Calmodulin SDS eluate, 3; IgG eluate, 4; Calmodulin flow, 5; Calmodulin eluate) were separated on 4-12 % gradient SDS-PAGE gels. Proteins were visualized with Coomassie staining. Protein bound with numbers and proteins from the IgG eluate were analyzed with the use of mass spectrometry. Identified proteins were listed below, together with the obtained score. To conclusively validate a complex, all its components should be visible in gel, in a calmodulin eluate fraction, or at least identified by mass spectrometry in an IgG eluate. To check the expression of TAP tagged proteins, their stability and strength of binding to IgG, western blots were prepared. Total proteins from yeast extracts (1) and from the IgG flow-through (2) were separated on 10% SDS-PAGE gels and transferred to a nitrocellulose membrane. Western blots were prepared with the use of PAP antibodies. The results are presented in the top right panel of each page. In order to estimate the size of the complexes, yeast extracts were separated with the use of size exclusion chromatography on a Superdex 200 column in low (150mM) and high (500mM) concentration of NaCl. 30 fractions from each chromatography were collected and spotted on a nitrocellulose membrane. PAP antibodies were used to detect fractions containing the TAP tagged protein. Intensity of spots was calculated and visualized as curves in the bottom panel of each page. The column was calibrated with the use of protein markers. Each experiment corresponds to validation of the complexes indicated at the top of each panel.
Supplemental Figure 2: Expression & Solubility Trials

15% SDS-PAGE of total extract (T) and clarified (S) cell lysates. Each protein was expressed in BL21 Gold (DE3) (Stratagene) or Rosetta2 (DE3) (Novagen) cells at 15°C (indicated by 15), 28°C (28), and 37°C (37). 10 µL of total and soluble fraction were performed by SDS-PAGE and Coomassie blue stain was used for visualizing the gel, as was for Supplemental Figures 2-8. Molecular weight markers and their sizes are indicated at right. Each band adjudged to be of approximate molecular weight concordant with the proteins of interest is indicated by a red asterisk.

Supplemental Figure 3: Paa1-Ptc2 Purification & Complex Reconstitution

Panel A: SDS-PAGE of fractions from Ni²⁺-NTA chromatography: "FT", flow through, "0", "50", "100", "200" correspond to washes with standard buffer, including imidazole added to the concentration indicated (in mM). Each protein (Paa1, above and Ptc2, below, labelled at left with an arrow to denote the protein of interest) were expressed at both 15°C and 37°C (indicated with bars above the gel). The proteins were soluble at both of these temperatures. Molecular weight markers and their sizes are indicated to the right of all gels. Panel B: Complex Reconstitution between Paa1 and Ptc2 using a Superdex 75 10/30 (GE Healthcare) gel filtration chromatography. Control fractionations of either Paa1, injected alone (grey dotted line) or Ptc2 alone (grey solid line) eluted later than the more excluded Ptc2-Paa1 complex (solid black line). SDS-PAGE of selected fractions from a 96-well fraction collector are shown below the chromatogram (280 nm absorbance). The lanes compare the injected sample (INJ), as well as the peak fractions (D6-E7), demonstrating the presence of both proteins in the shifted peak.

Supplemental Figure 4: Vps27-Hse1 Purification & Complex Reconstitution

Panel A: SDS-PAGE of fractions from Ni²⁺-NTA chromatography: "FT", flow through, "0", "30", "100", "400" again correspond to the imidazole concentration in mM. Each protein (Vps27, above and Hse1, below, again labelled to the left with an arrow) were expressed at 37°C (indicated as the previous figure). Both proteins were soluble at both of these temperatures. Molecular weight markers and their sizes are
indicated. Panel B: Complex Reconstitution between Vps27 and Hse1 using a Superdex 200 10/30 (GE Healthcare) gel filtration chromatography. Control fractionations of either Hse1, injected alone (grey dotted line) or Vps27 alone (grey solid line) eluted later than the more excluded Vps27-Hse1 complex (solid black line). SDS-PAGE of selected fractions from a 96-well fraction collector are shown below the chromatogram (280 nm absorbance). The samples compare the injected sample (INJ), as well as the peak fractions (D6-E7), demonstrating the presence of both proteins in the shifted peak.

**Supplemental Figure 5: Gcd10-Gcd14 Purification & Complex Reconstitution**

Panel A: SDS-PAGE of fractions from Ni^{2+}-NTA chromatography: "FT", flow through, "0", "50", "100", "200" correspond to washes with standard buffer, with imidazole added to the concentration indicated (in mM). Each protein (Gcd10, above and Gcd14, below, (labelled as in Supp. Fig. 3), and both were soluble at both of these temperatures. Molecular weight markers and their sizes are indicated at right. Panel B: It was not possible to purify Gcd10 or Gcd14 in a non-aggregated form singly, so cells expressing each of the proteins alone (Gcd10 and Gcd14, as used in panel A), were combined and sonicated together after being resuspended in a buffer containing 1M NaCl and purified using Ni^{2+}-NTA chromatography as in panel A. Fractions containing Gcd10 and Gcd14 were then loaded onto a Superdex S200 10/30 column, yielding a single, symmetrical peak containing both proteins. Estimation of the molecular weight of the complex by comparison with the elution volumes of molecular weight standards ferritin (440 kDa), beta-amylase (200 kDa) and aldolase (158 kDa) (arrowed) indicates that the complex is of approximately 300 kDa.

**Supplemental Figure 6: Ste11-Ste50 Purification & Complex Reconstitution**

Panel A: SDS-PAGE of fractions from Ni^{2+}-NTA chromatography: "FT", flow through, "0", "50", "100", "200" correspond to washes of with standard buffer, with imidazole added to the concentration indicated (in mM). Each protein (Ste50, above and Ste11, below, labelled at left with an arrow to denote the protein of interest) were expressed at both 15°C and 37°C (indicated with bars above the gel). Ste11 was found to be insoluble at both of these temperatures, however. Molecular weight markers and
their sizes are indicated to the right. Stay tuned for updates on Ste11 expression. Panel B: Sonication of cells expressing Ste11 and Ste50 (as per the Gcd10:Gcd14 complex) did not yield soluble Ste11. The upper band after gel filtration (arrowed), despite having a molecular weight approximately similar to Ste11 was identified as a contaminant from *E. coli*.

**Supplemental Figure 7: Effects of expression using synthetic genes using codon optimization relative to naturally occurring genes**

SDS-PAGE of fractions from Ni\(^{2+}\)-NTA chromatography purifications of co-expressed Gcd10/Gcd14 (Panel A) and Ssl2/Yor352w (Panel B) as synthetic, codon-optimized genes (above), compared to the naturally-occurring yeast DNA sequences (below). Fractions are labelled as follows: “T”, total cells prior to sonication; “P”, pellet post-sonication, “S”; supernatant post-sonication, “FT”; flow through, “E1” & “E2” are specific elutions with buffer including 300 mM imidazole. Note the increased yield when using codon-optimized genes in both cases.
Supplemental Experimental Procedures

**Algorithms for the selection of complexes using bioinformatics**

We have already described a selection system to rank protein assemblies based on various parameters (Pache and Aloy, 2008), but recapitulate it here. The algorithm is based on the notion that promising target complexes should be small, compact and homogeneous in order to yield successful expression, purification and structure determination. To rank the complexes, biophysical, biochemical and large-scale proteomics data are incorporated in the form of partial scoring functions that we then normalized and combined into a final feasibility score for each complex.

Briefly, the first individual score refers to the average *socio-affinity* index of the complex (Gavin et al., 2006), which quantifies the tendency of two proteins to interact with each other when tagged and to co-purify when yet other proteins are tagged. The higher the average socio-affinity of a complex, the more of its proteins are predicted be in direct contact, which could be used as an indication for the compactness of the complex. We also consider the molecular weight and the total sequence length of the complex components, since larger proteins are usually more difficult to express. We penalize the presence of low complexity regions, internal repeats, coiled coils and intrinsically disordered stretches, since they often result in insoluble proteins that aggregate when over-expressed (Dale et al., 2003).

Information regarding sub-cellular localization (Huh et al., 2003) and abundance (Ghaemmaghami et al., 2003) of individual proteins is also considered, since complex components are expected to be consistent in these terms. Another criterion employed is the level of conservation of a complex across evolution, which we addressed by considering orthologous protein relationships in 83 eukaryotic species (von Mering et al., 2007). We also used binary interactions extracted from yeast two-hybrid screens in combination with the number of isoforms described for each complex to estimate its self-consistency. This is to capture the independence and homogeneity of each complex with respect to the others. For instance, if a complex contains many binary interactions between its own subunits and few with proteins from other complexes, this decreases the probability of missing components in the definition of the complex. Additionally, the fewer isoforms the more invariant the
protein cluster is. Finally, we used cumulative probabilities to normalize each score to the range \([0,1]\), and calculated a global feasibility score as the weighted average of all normalized partial scores. The final score \(S(c)\) assigned to each protein complex \(c\) is calculated by taking the weighted average of all normalized partial scores \(s_i(c)\), ignoring those which are not applicable for the respective protein complex (e.g. the 'Average abundance ratio' when the abundance of none of the proteins in the respective complex could be determined), and multiplying by 100.

Using a weighted average to combine all partial scores makes it possible to give each partial score a particular weight \(w_i\), which allows us to evaluate its importance and to control its impact on the final score:

\[
S(c) = \frac{\sum_{i} w_i P_{\sum} (s_i(c) = x)}{\sum_{i} w_i}, w_i \in [0,1]
\]

\(P_{\sum} (s_i(c) = x)\) denotes the cumulative probability that the partial score \(s_i(c)\) of the protein complex \(c\) is equal to \(x\), used for normalization to take into account the distribution of the respective partial score and calculated by taking the sum of all probabilities \(P(s_i(c) = y)\) over all values \(y \leq x\) or \(y \geq x\), depending on whether higher or lower values are better for the respective ranking criterion:
\[
P_x \sum (s_i(c) = x) = \begin{cases} 
P(s_i(c) \leq x), & \text{better scores are obtained for high } x \text{ values} \\ 
P(s_i(c) \geq x), & \text{better scores are obtained for low } x \text{ values} \\ 
\sum_{y \leq x} P(s_i(c) = y), & \text{better scores are obtained for high } x \text{ values} \\ 
\sum_{y \geq x} P(s_i(c) = y), & \text{better scores are obtained for low } x \text{ values} 
\end{cases}
\]

**Cloning into T7 promoter-based expression systems**

**Cloning strategy used for single-subunit expression**

For the expression and production of single proteins, a cloning strategy to generate C-terminally 6His-tagged proteins was employed. In order to minimize the possibility that errors could be introduced into the primers, to test for the presence of restriction sites in the gene, and to find optimal melting temperatures for primer pairs, a web interface-based script was written. The web interface is part of a basic laboratory information management system (LIMS), which additionally allows the storage of the primers in a structured query language (SQL) database. The software is open source, and freely available: http://plasmidb.sourceforge.net.

PCR reactions were performed with oligonucleotides which encoded either a 5´ Nco I or Nde I restriction site (the choice of restriction site was determined by the absence of this restriction site in the gene of interest), and a 3´ primer which contains a Not I site, followed by a sequence encoding a 6His tag and a stop codon. The oligonucleotides used in this study are listed in Supplemental Table 1. PCR using these primers and *Saccharomyces cerevisiae* S288C genomic DNA as template yielded DNA fragments of the expected size for all of the desired genes, except for Ste11, which we were unable to produce. The resulting PCR products that encoded a 5´ Nde I site were cloned into the vector pET9 (Novagen), and those containing Nco I sites were cloned into pET28 (Novagen), using standard procedures. Since it was not possible to obtain a PCR product of Ste11, despite numerous attempts with different primers and melting temperatures, the gene encoding Ste11 was ordered as a synthetic gene.

**Cloning strategy used for multi-plasmid co-expression**

The vector DNA pET-NKIB 3C/LIC (10µg) was digested with *Kpn* I (2-3h at 37°C; NEB) and purified with a QIAquick spin column (Qiagen) according to the
manufacturer's protocol. The linearised vector was treated with T4 DNA Polymerase in the presence of 25 mM dTTP to generate single-strand overhangs. The reaction was incubated at room temperature for 30 min and inactivated by incubating at 75°C for 20 min. Target genes were amplified by PCR using the Pfu Turbo polymerase (Stratagene). For the pET-NKlb 3C/LIC the 5'-end of the primers must incorporate the CAGGGACC CGGT sequence upstream the forward PCR primer and the CGAGGA GAAGCCGGTTA sequence upstream of the reverse primer (which includes a TAA stop codon). For the pET-NKlb LIC without a 6His-tag (no-tag constructs), the sequence GGGCCGGCGATG must be incorporated in the 5'-end of the primers. A web server enabling the high throughput design of PCR primers was used and is freely available at http://xtal.nki.nl/ccd. The PCR products were purified prior to T4 treatment (QI Aquick PCR purification kit by Qiagen); 0.2 pmol of purified PCR DNA was treated with T4 DNA Polymerase in the presence of 25 mM dATP to create the single-strand overhangs. The reaction was incubated at room temperature for 30 min and inactivated by incubation at 75°C for 20 minutes.

Annealing of the vector and insert was achieved by mixing 1 μl pET-NKlb 3C/LIC vector (50ng/μl) with 2 μl insert (0.02 pmol). The reactions were incubated at RT for 5 min, after which 1 μl of 25 mM EDTA was added. Typically, half of the annealing reaction (2μl) is transformed into NovaBlue competent cells (Novagen) and after overnight incubation at 37°C, the annealing and transformation efficiency can be verified. A typical transformation protocol was used (incubation on ice-20 min; heat shock- 30 sec at 42°C; incubation on ice 2 min; addition of 80 μl LB medium to each sample and incubation at 37°C for 1 hour) and transformants are plated on LB agar supplemented with kanamycin (30mg/ml for all his-tag constructs) or ampicillin (100mg/ml; for all no-tag constructs) and incubated at 37°C overnight. Plasmid DNA was extracted from single colonies using a miniprep kit (Qiagen) and restriction digestion was used to verify the presence of the insert of interest.

Cloning strategy used for poly-cistronic expression

For operon constructions, oligonucleotides were designed to amplify coding regions by PCR. Oligonucleotides contained restriction sites selected to be unique in the final plasmid, allowing the simultaneous insertion of both ORFs in the plasmid vector pBS3021. Oligonucleotides contained in addition a Shine-Dalgarno sequence
upstream of the second ORF and a sequence encoding a 6His tag fused in-frame upstream or downstream of one of the two ORFs. PCR fragments were inserted by standard cloning downstream of the T7 promoter of the pBS3021 expression vector. DNA purifications were performed on an EPmotion robot (Eppendorf) using a Macherey-Nagel mini-preparation kit followed by digestion and gel electrophoresis to ascertain the presence of the desired inserts. Inserted fragments were entirely sequenced to confirm the absence of PCR-induced mutations.

**Expression testing**

*Expression testing of individual subunits*

The initial objective in this part of the study was to define expression conditions that gave optimal yields of soluble protein for each full-length protein. Therefore, each expression vector under study was initially transformed into both Rosetta pLysS (Novagen) and Gold (Stratagene) in a 24-well block (Corning, Inc.). After incubation overnight in 5 ml per well of 2x Yeast Tryptone (2YT hereafter) medium supplemented with 30 μg/ml kanamycin at 37°C, this pre-culture was used to inoculate 10 ml per well of similar media as the expression culture, again using 2YT broth supplemented with 30 μg/ml kanamycin. The volume of the inoculum used was adapted according to its OD_{600} so as to obtain a starting optical density of 0.1 for the expression culture. This culture was incubated until the OD_{600} reached ~1, then was separated into three 1ml aliquots, one for each of the expression temperatures under study: 37°C, 28°C and 15°C. After the addition of IPTG to a final concentration of 0.5 mM, the cells were incubated either for 4 hours at 37°C, or for 16-18 hrs for the inductions at 28°C and 15°C. Cells were harvested by centrifugation of the 24-well block at 5,300 rpm for 1 hour.

The pellets in each well were re-suspended with lysis buffer (20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 5mM β-mercaptoethanol). The 24-well block containing the cell suspension was sonicated with 1 mM benzonase, 4 times for 10 minutes, and then centrifuged at 5,300 rpm for 1 hour. The crude and clarified cell lysates were analyzed using SDS-PAGE (Table 1, column labeled “Single subunit expression” and Supplemental Fig. 1). All proteins except for Ste11 could be produced in a soluble form. Expression from a construct corresponding to Ste11 did
not yield soluble protein, however. An expressed band apparent in purifications at
approximately the expected size of Ste11 was in fact identified by mass spectrometry
of tryptic digest of this band as polymyxin resistance protein arnA (UniProt accession
code; P77398), a common contaminant of purifications originating from E. coli.

Expression testing for multi-plasmid co-expression

Small-scale protein expression and solubility screening was carried out for
single constructs of full-length proteins, for constructs of individual or combinations
of domains as well as for co-expressions of partners. For the transformation of single
plasmids, plasmid DNA was transformed into E. coli Rosetta2 (DE3) T1R. Single
colonies were used for small-scale expression trials in a 24-well 'Deepwell' block
(Corning). Each well contained 3ml LB media supplemented with kanamycin
30mg/ml. For the transformation of multiple plasmids to co-express complexes,
plasmid DNA of the two partners of interest was transformed into Rosetta2 (DE3)
T1R E. coli and plated onto LB agar plates supplemented with 30 mg/ml kanamycin
and 100 mg/ml ampicillin. Single colonies were used for small-scale expression trials
in a 24-well 'Deepwell' block. Each well contained 3 ml LB media supplemented with
30mg/ml kanamycin and 100 mg/ml ampicillin.

The 24-well block was incubated in a shaking incubator at 500 rpm until an
OD600 of about 0.6-0.8 had been attained, at which point the temperature was reduced
to 16°C and the cultures were induced by the addition of IPTG to a final
concentration of 1 mM. Incubation was continued for 16-18 hrs and cells were
harvested by centrifugation of the Deepwell block at 4,000 rpm for 15 minutes. The
pellets in each well were resuspended with lysis buffer (40% sucrose, 50 mM Tris-
HCl pH 7.5, 200 mM NaCl, 5mM β-mercaptoethanol, 4mg/ml lysozyme (Novagen),
DNAse and PMSF). The 24-well block containing the cell suspension was incubated
in a temperature controlled shaking incubator at 300 rpm, for 20 minutes at 10°C and
then centrifuged at 4,000 rpm for 15 minutes. The clarified cell lysate was then mixed
with 25 μl pre-equilibrated MagneHis Ni-beads (Promega) and incubated at 4°C for
30 minutes. The Magnetight HT96 stand (Novagen) was used to pull down the
MagneHis beads. The magnetic beads were washed 3 times with 1 ml wash buffer
(lysis buffer supplemented with 20 mM imidazole). Protein elution was performed by
adding 30 $\mu$l elution buffer (wash buffer containing 400 mM imidazole) to each sample and eluted fractions were analyzed by SDS-PAGE.

**Expression testing for polycistronic co-expression**

Proteins were expressed in BL21(DE3) codon+ *E. coli* using auto-induction media (Studier, 2005). Small scale (3 ml) or large scale (100-200 ml) cultures were performed. After overnight incubation at 25°C, cells were harvested. For small-scale cultures, lysis was performed with lysozyme and benzonase. After centrifugation, the supernatant was incubated with Ni-NTA beads for 1h at 4°C. After washing the column with equilibration buffer (50 mM Tris HCl pH7.4, 20 mM imidazole, 300 mM NaCl, 2 mM $\beta$-mercaptoethanol, 0.2 % NP40, 10 % glycerol), proteins were eluted with 500 mM imidazole. For large-scale culture, pellets were washed with PBS and dissolved in 20 ml of equilibration buffer. Cells were lysed using a “Constant Cell Disruption System”. After centrifugation, the supernatant was filtered and purified by chromatography on Ni$^{2+}$-NTA (Akta system, Hitrap Ni$^{2+}$ 1 ml column volume). After washing with equilibration buffer, elution was performed with a linear gradient from 0 to 100% of elution buffer (50 mM Tris HCl pH 7.4, 500 mM imidazole, 300 mM NaCl, 2 mM $\beta$-mercaptoethanol, 0.2% NP40, 10% glycerol). All eluates were analyzed by means of SDS-PAGE.

**Complex Formation Trials**

**Reconstitution of complexes from individually purified partners**

We purified Paa1 and Ptc2 to approximate homogeneity using gel filtration and concentrated them individually to 2 mg/ml. The two proteins were combined in a 500 $\mu$l reaction mixture (250 $\mu$l of each component), and were incubated on ice at 4°C. Gel filtration chromatography illustrated that the proteins co-elute, shifting the peak of Ptc2 by 1.3 ml upon addition of the Paa1 subunit (Supplemental Fig. 3). In a similar manner to the Paa1-Ptc2 complex, both Vps27 and Hse1 could be produced in a soluble form, and the proteins were produced using expression at 37°C henceforth (Supplemental Fig. 4, Panel A). Vps27 and Hse1 proteins were purified to homogeneity using a final gel filtration step. 10nM of each of the proteins were injected separately into an S200 analytical column. Both proteins were then combined and incubated for 30 minutes at 4°C and subjected to gel filtration (Supplemental Fig.
The complex eluted at 11.03 ml, compared to the individual profiles of Vps27 (11.67 ml) and Hse1 (13.58 ml). Both proteins appear to be in the peak fractions, as judged by SDS-PAGE (Supplemental Fig. 4, lower Panel A). Gcd10 and Gcd14 posed severe problems when purified individually, even from refolded material. Despite being able to obtain both proteins in a relatively soluble form (Supplemental Fig. 5, Panel A), they had a tendency to aggregate as judged by their elution in the void volume of a Superdex S200 column. Eventually, co-sonication of the individually expressed proteins was attempted, which improved the situation considerably. However, it was only when co-sonication was performed in the presence of a high salt buffer (1M NaCl) that an acceptable elution profile was obtained, as has been previously reported (Ozanick et al., 2005) (Supplemental Fig. 5, Panel B). Formation of the complex between Ste11 and Ste50 was prevented by the lack of soluble expression of Ste11 (Supplemental Fig. 6, Panel A, lower) even when co-expressed with Ste50. Ste50 could be produced in a soluble form even when expressed alone, although what appeared to be degradation products were visible (Supplemental Fig. 6, panel A, upper).
Supplemental References


Supplemental Movies and Spreadsheets

Click here to download Supplemental Movies and Spreadsheets: SupplementaryMaterialOligonucleotides.xls