*3156–3167* *Nucleic Acids Research, 2021, Vol. 49, No. 6* *Published online 2 March 2021*
*doi: 10.1093/nar/gkab077*

# Analysing the yeast complexome—the Complex Portal rising to the challenge

**Birgit H. M. Meldal [1,*], Carles Pons[2], Livia Perfetto[1], Noemi Del-Toro [1], Edith Wong [3], Patrick Aloy[2,4], Henning Hermjakob [1], Sandra Orchard [1] and Pablo Porras [1]**

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute for Science and Technology, 08028 Barcelona, Catalonia, Spain, [3]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305-5477, USA and [4]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Catalonia, Spain

Received November 03, 2020; Revised January 22, 2021; Editorial Decision January 25, 2021; Accepted January 27, 2021

## ABSTRACT

**The EMBL-EBI Complex Portal is a knowledgebase of macromolecular complexes providing persistent stable identifiers. Entries are linked to literature evidence and provide details of complex membership, function, structure and complex-specific Gene Ontology annotations. Data are freely available and downloadable in HUPO-PSI community standards and missing entries can be requested for curation. In collaboration with *Saccharomyces* Genome Database and UniProt, the yeast complexome, a compendium of all known heteromeric assemblies from the model organism *Saccharomyces cerevisiae*, was curated. This expansion of knowledge and scope has led to a 50% increase in curated complexes compared to the previously published dataset, CYC2008. The yeast complexome is used as a reference resource for the analysis of complexes from large-scale experiments. Our analysis showed that genes coding for proteins in complexes tend to have more genetic interactions, are co-expressed with more genes, are more multifunctional, localize more often in the nucleus, and are more often involved in nucleic acid-related metabolic processes and processes where large machineries are the predominant functional drivers. A comparison to genetic interactions showed that about 40% of expanded co-complex pairs also have genetic interactions, suggesting strong functional links between complex members.**

## INTRODUCTION

Many proteins exist as part of stable, macromolecular complexes that act as functional units in the cell. Identifying such complexes is crucial for a systems level understanding of biological processes. The EMBL-EBI Complex Portal (www.ebi.ac.uk/complexportal, (1,2) is a manually curated, encyclopaedic resource of macromolecular complexes from a number of key model organisms, including *Saccharomyces cerevisiae*. Entries describe assemblies of two or more macromolecules (proteins, nucleic acids, small molecules) for which there is evidence (experimental or inferred) that these molecules stably interact with each other and have a demonstrated molecular function. Judgment of what constitutes a stable complex is based on available scientific literature, experimental evidence and a consensus decision made by two curators. Homomultimers are only curated if it has been demonstrated experimentally that multimerization is required for their function or in cases where a heterodimeric complex exists and at least one of the two participants forms an orthologous homodimeric complex. Polymers are excluded and large, multi-complex machineries are reduced to their functional subcomplexes because the final assemblies are often dynamic, rather than a single instance of a functional unit existing at any moment of time. The subcomplexes of large assemblies are annotated with GO terms relating to the larger machineries, for example, all spliceosome sub-complexes are annotated with the GO term GO:0005681 (spliceosomal complex) to facilitate searching. Unlike other compendia of complexes, such as CORUM (3), the Complex Portal not only lists the protein composition of each complex but also includes nonprotein components, stoichiometry (when known) and topology (including intra-complex binary interactions) and provides both a free-text and structured description of complex function and properties (Figure 1). Each entry is linked to a range of related resources such as complex-centric Gene Ontology (GO) annotations (4,5), structure determinations deposited in the wwPDB (6) or the role of the complex in a pathway in Reactome (human-only) (7). Links to these and other resources are provided both via cross-referencing and

*To whom correspondence should be addressed. Tel: +44 1223494107; Email: bmeldal@ebi.ac.uk

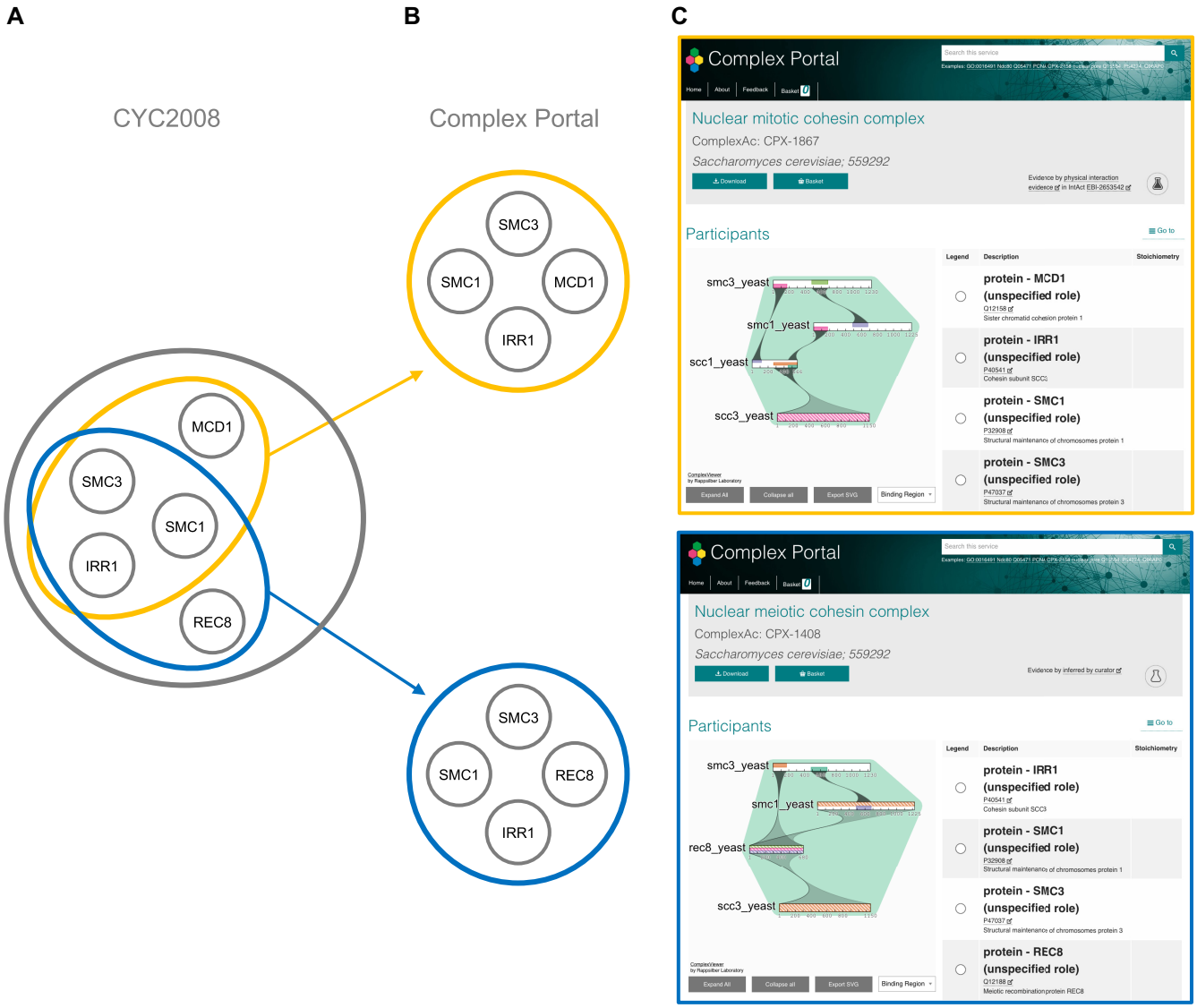


**Figure 1.** The nuclear cohesin complex is curated as one entry in CYC2008 (**A**) but represented by two, process-specific complexes in Complex Portal (**B**), one involved in mitosis and one involved in meiosis. The two complexes differ by one subunit. (**C**) Screenshots of the Details page of these complexes in Complex Portal.

the integration of widgets on the website to display Reactome pathways diagrams, structures via the PDBe LiteMol App (8) and gene expression data via the Expression Atlas widget (9). Versioning of the stable accession numbers indicates when a complex has been significantly updated, for example, by the addition or removal of a protein subunit from the list of participants. The data are freely available and downloadable in the HUPO-PSI community standard PSI-MI XML3.0 (10), MI-JSON and tab-delimited ComplexTab formats (1).

*Saccharomyces cerevisiae* (henceforth referred to as 'yeast') is an important model organism for our understanding of the biology of all eukaryotic organisms and significant effort has gone into identifying all its stable complexes. However, until recently, information about such complexes was scattered across many publications and in different databases. An early effort to concatenate these data was the, now deprecated, MIPS yeast complex database (11). Domain-specific resources such as structural data in wwPDB, functional statements and Gene Ontology annotations on the protein pages of UniProt (12) and gene pages of the *Saccharomyces* Genome Database (SGD; www.yeastgenome.org) (13) provide highly detailed, component-specific information only. It was very difficult to derive a picture of the complete yeast complexome without systematically integrating information from these and other sources. Additionally, molecular interaction databases such as those maintained by members of the IMEx Consortium (14,15) provide experimentally derived interaction data without combining evidence from multiple sources for a whole complex. Several studies in the early 2000s predicted yeast complexes based on high-throughput yeast two-hybrid (16,17) or affinity-purification methods (18–20) but only few studies included systematic validation by way of small-scale

experiments and manual curation (21). In 2009, Pu *et al.* published a comprehensive analysis of 400 highly interconnected assemblies derived from high-throughput experiments (Yeast High ThroughPut, YHTP2008) and also a compendium of 408 literature-derived, manually curated complexes based on small-scale experiments (Curated Yeast Complexes, CYC2008) (22). While both datasets contained approximately 400 entries, <20% of these were identical to each other. However, in the 12 years since this set was first published, significant advances have been made in the field of interaction biology and considerably more high-quality datasets are now available to contribute to our understanding of this field. This has allowed a re-evaluation of the data and in 2018 the first version of an updated and enhanced dataset of known yeast complexes, the 'yeast complexome', was released in the Complex Portal. Additional complexes are being added to the dataset on an ongoing basis, if and when they are experimentally verified.

In this paper, we explore the yeast complexome and compare the extent and depth of data available through the Complex Portal to other resources that contain data on yeast complexes, namely to the curated and predicted complexes from Pu *et al.* and complexes predicted based on all experimental protein–protein interactions in the IntAct molecular interaction database (23). Compared to CYC2008, the Complex Portal dataset contains almost 50% more entries (589 versus 408), covers 5% more of the yeast proteome (32% versus 27%) and includes additional detail about the complexes as described above. Finally, we compare and contrast protein complex co-membership with the global genetic interaction network (24) and found that both datasets significantly overlap.

## MATERIALS AND METHODS

### Source data for the Yeast Complexome

The data for the yeast complexome were derived from detailed literature searches and collated in collaboration with curators based at UniProt and SGD. A draft list of putative complexes was created based on the following sources: the CYC2008 dataset, UniProtKB SUBUNIT comment lines search with keywords 'found in a complex with' and a close collaboration with SGD who provided a list of identified complexes and by directed literature searches. A complex is only included in the Complex Portal dataset if there is literature evidence for its existence and functional role *in vivo*. Complexes that were identified based only on either high- or low-throughput analyses without the presence of further verification experiments or functional assays were not included. Thirteen homomers have been curated, to date, because the protein was also present in a related heteromeric complex. It should be noted that homomers have largely been omitted from manually curated datasets, because it is often challenging to demonstrate experimentally if their function requires oligomerization and their generic functions are already described in the UniProtKB database. Literature searches and the collaboration with SGD are ongoing and new complexes are being added to the dataset when they are experimentally identified.

### The datasets

The protein complex datasets analysed were the following:

- Complex Portal—589 complexes (release 228, 16 November 2019)
- CYC2008—408 manually curated complexes (22)
- YHTP2008—400 predicted complexes (22)
- IntAct-LT—332 predicted complexes derived from low-throughput experiments in IntAct (release 228, 16 November 2019)
- IntAct-HT—689 predicted complexes derived from high-throughput experiments in IntAct (release 228, November 2019)

To enable direct comparison of protein complex components represented in the Complex Portal and IntAct, gene locus IDs in CYC2008 and YHTP2008 were mapped to UniProt ACs using the UniProt Mapping service web application (UniProt Release November 2019). Ambiguous mappings, where a locus could be mapped to more than one UniProt entry with the same sequence, were expanded to include all potential mapping pairs.

Complex Portal data were exported in ComplexTab format. Where complexes are part of larger assemblies (subcomplexes) these were expanded to provide a list of unique UniProtKB identifiers. Sets of paralogous ribosomal proteins were expanded to a full list, therefore all potential UniProtKB identifiers were included in the analyses. The expansion of paralogous proteins leads to an over-inflation of the subunit count per complex for the two ribosomal subunits but is the only way to include all proteins in the comparative analysis. As stoichiometry information is only available in a limited number of Complex Portal and IntAct entries and often missing due to a lack of available evidence, it was ignored and comparisons were based on unique protein identifiers only. Nonprotein complex members such as nucleic acids and small molecules were not included as these are not provided in full by any resource other than the Complex Portal.

IntAct complexes were derived from all yeast–yeast interactions in IntAct release 228. Interactions were exported in MI-TAB2.7 format and split into those derived from papers with 100 or less interactions/paper and those with >100 interactions/paper. Complexes were predicted using the Cytoscape App ClusterONE (25) using default parameter settings, MI-score values as weights and a minimum cluster size of $n = 3$.

### Functional analyses

For the selection of genetic interactions, we used the global yeast genetic interaction network, the first comprehensive genetic interaction map in any organism (24). The network was constructed by evaluating the growth defects associated with the majority of the ∼18 million possible gene pairs in yeast, and includes ∼350 000 positive and ∼550 000 negative genetic interactions. Nonessential genes were queried by deletion alleles and essential genes by temperature-sensitive and DAmP alleles. However, we disregarded the DAmP data because few DAmP alleles had an effect on cellular fitness. For pairs of genes screened more than once (for

instance, pairs involving genes queried using different alleles), a consensus approach was implemented in which we considered a given pair to have a genetic interaction if that was the result in at least half of the screens.

Interacting protein pairs in a complex (i.e. co-complex pairs) were inferred by matrix expansion of all complexes. UniProt identifiers were mapped to ORFs in order to compare inferred physical interactions and genetic interactions as the latter are provided as ORFs. Background pairs (i.e. 'no co-complex pairs') were defined as those pairs of proteins present only in different complexes. The fractions of co-complex and background pairs with positive and negative interactions were calculated, considering only pairs of proteins whose genes were present in the genetic interaction network (52%, 51%, 55%, 58% and 64% of co-complex pairs in CP, CYC, YHTP, IntAct-LT and IntAct-HT, respectively). Statistical significance was calculated by Fisher's exact tests.

In addition to genetic interactions, we evaluated the overlap of co-complex relationships with the co-expression, co-localization and co-annotation functional standards. In all cases, only protein pairs for which functional data were available were considered. The co-expression standard was derived from the MEFIT co-expression network, which integrates data from multiple microarray datasets (26). Pairs with a MEFIT score >1.0 were considered to be co-expressed. The co-localization standard was based on a previous high-throughput study (27). Protein pairs localized in one or more shared cellular compartments were considered to be co-localized. The co-annotation standard is based on GO biological process annotations and disregards very frequently annotated GO terms as described in a previous work (24).

To obtain a comprehensive view of the differences between those proteins participating in complexes and those that do not, the following characteristics were compared: genetic interaction degree calculated on array genes and averaging estimates across the different alleles of a gene (24), co-expression degree calculated as the number of co-expression relationships per gene (see above), gene conservation in other species (28), expression variation (29), fitness of non-essential gene deletion alleles (24), PPI degree (from IntAct yeast-yeast interaction, release 234 (09 July 2020), restricted to high-throughput dataset with >100 interactions per publication as it reduces the bias from confirmatory small-scale experiments), multifunctionality of proteins based on the number of biological process annotations in GOSlim (downloaded from SGD, July 2020), fraction of disordered residues downloaded from d2p2.pro (30), being essential (31), being a gene duplicate defined as having a paralog in YeastMine (32), being a membrane protein (33) as well as subcellular localization (27) and broad functional classes (34). For each numerical feature, values were $z$-score normalized using the median and the standard deviation of the values for the background proteins. The median $z$-score value of the proteins in complexes was used for the graphical representation of the result. Statistical significance was evaluated using two-sided Mann–Whitney $U$ tests. For each binary feature, fold enrichment was calculated as the ratio of complex members with that feature divided by the ratio of noncomplex members with that feature. Statistical significance was calculated by two-sided Fisher's exact tests.

The relative difference in transcript counts, expression variance, protein abundance and protein half life was calculated for co-complex and background pairs. For every pair and measure, we calculated the maximum (MAX) and minimum (MIN) value within the pair. The relative difference was then calculated as (MAX-MIN)/MAX. The larger this score is, the larger the difference between the pair of proteins/genes. Statistical significance was calculated using two-sided Mann–Whitney $U$ tests.

Direct and indirect contacts were selected from a set of Complex Portal complexes of size 3 or larger that contained information for both types of contacts. Self interactions were ignored. Protein pairs belonging to different complexes of the selected set were defined as background. Genetic interaction profile similarity values were downloaded from http://thecellmap.org (35), considering both essential and nonessential genes, and averaging similarity values across alleles of the same gene.

A list of 12 high level GO terms (Table 2) was manually selected to best represent processes and functions related to nucleic acids as well as the component term 'nucleus'. These terms were used to build a bespoke SLIM and all annotations to yeast proteins using these terms and their children were exported on 9 October 2020. This list of GO terms was used to filter all Complex Portal complexes that were also annotated to any of these terms. This analysis was only performed on the Complex Portal dataset as there are no complex-specific GO annotations for the other datasets.

### Analysis tools

Data manipulation and visualizations were performed in R (data.table, splitstackshape, reticulate, rio, ggplot2, scales), Python and Excel. Unique versus shared sets of complexes were identified using Venny (https://bioinfogp.cnb.csic.es/tools/venny/).

## RESULTS AND DISCUSSION

### The yeast complexome in the Complex Portal

*Saccharomyces cerevisiae* complexes were captured in the Complex Portal leading to yeast being the first completed species complexome. It is the largest manually curated compendium of yeast macromolecular complexes, comprising 589 complexes, 1930 proteins and 15 863 co-complex relationships. In order to identify all known yeast complexes we gathered information from a number of sources (CYC2008 complexes, UniProt, SGD, literature publications). Some complexes that are included in these sources have not been included in the Complex Portal because they have since been identified as part of a bigger complex or they lack clear experimental evidence for their existence and their functional role *in vivo*. These putative complexes are kept in a separate list and are periodically revisited to see if more evidence has been published. Collaborations with SGD are ongoing and we update existing entries and add new ones when new evidence comes to light.

Compared to other resources, the Complex Portal provides added value through its greater scope of annotation.

**Table 1.** Basic statistics about the five complex datasets

| | Total no. proteins | Total no. complexes | Max no. of proteins/ complex | Mean no. of proteins/ complex[b] | Median no. of proteins/ complex[b] | Homomers | Dimers | No. co-complexes | Incl nonprotein components[c] | Stoichio-metry[c] | Manually curated fields[c] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Complex Portal | 1930 | 589 | 80 (73[a]) | 6.93 | 4 | Yes | Yes | 15 863 | Yes | Yes (if known) | Yes |
| CYC2008 | 1624 | 408 | 81 (44[a]) | 6.67 | 4 | No | Yes | 11 238 | No | No | Partial |
| YHTP2008 | 1911 | 400 | 181 | 8.03 | 4 | No | Yes | 28 146 | No | No | No |
| IntAct-LT | 1918 | 332 | 40 | 6.71 | 5 | No | No | 9808 | Yes | Optional | No |
| IntAct-HT | 3147 | 689 | 99 | 7.31 | 5 | No | No | 30 493 | Yes | Optional | no |

[a] max size without ribosomal subunits.
[b] for complexes size 3 or greater.
[c] in original database, e.g. IntAct PPIs.

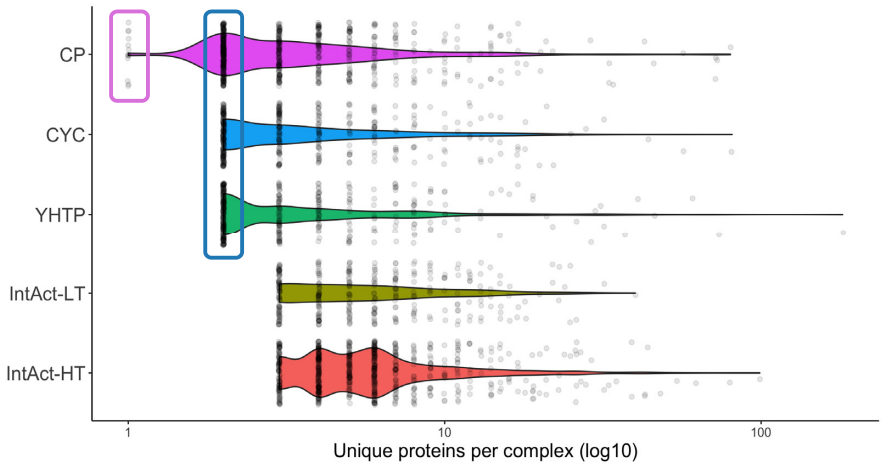


**Figure 2.** Distribution of number of unique proteins per complex. Homomers are found in the small rectangle and heterodimers in the large rectangle. Total number of complexes per dataset: CP = 589, CYC = 408, YHTP = 400, IntAct-LT = 332, IntAct-HT = 689

Each complex entry has a manually annotated description of their function and physical properties and includes stoichiometry and topological information when available. The Evidence and Conclusion Ontology (ECO) (36) is used to indicate the type of evidence we have for each entry and where interaction evidence is available in an IMEx member database, the wwPDB or EMDB (37) cross-references are provided. Each complex is annotated to GO terms specific for the complex and a selection of supporting literature references are provided. Versioning allows easy tracking of changes in complex composition. Additionally, the data are downloadable in three different community standard formats and as a live resource it gets updated every two months.

## Dataset comparisons

The yeast complex dataset published in the Complex Portal is the first manually annotated yeast complex dataset since the publication of CYC2008 by Pu *et al.* in 2009. We compare these two manually curated datasets with each other and with corresponding experimentally derived predicted complexes from YHTP2008 and IntAct release 228 (16 November 2019). The IntAct data were split into low- and high-throughput publications setting a cut-off at 100 interactions per publication. See Table 1 for a summary of the five datasets and Figure 2 for the distribution of unique proteins per complex.

The two manually curated datasets share 1543 proteins (80% and 95%, respectively): 387 proteins are unique to the Complex Portal and 81 unique to CYC2008 (Table 1, Figure 3A); overall, Complex Portal and CYC2008 complexes cover 32% and 27% of the yeast proteome, respectively. The reason for the relatively low proteome coverage may be multifaceted: both datasets have concentrated on stable, macromolecular machines whereas many proteins may be found in more transient interactions, such as signaling assemblies or enzyme–substrate interactions. The identification of protein complexes may also be limited by technological constraints and some complexes simply cannot be purified by existing methods, for example insoluble membrane components.

The Complex Portal contains 589 yeast complexes compared to 408 in the CYC2008 dataset, a 44% increase (Table 1). They share 286 identical complexes that responds to 49% of Complex Portal complexes and 70% of CYC2008 complexes (Jaccard Index = 1.0) (Figure 4A). When reducing protein identity matching to a minimum of 50% (Jaccard Index = 0.5) the overlap is over 80% for both datasets (Figure 4C). There are many more complexes in the Complex Portal than in CYC2008 because a large amount of knowledge has accumulated in the intervening 12 years and because complexes including parologous alternative proteins have often been created as a single entry in CYC2008 but split into separate, alternative entries in the Complex Portal to reflect their functional composition (Figure 1). On the

other hand, approximately 30 CYC2008 complexes were not re-curated into the Complex Portal because the available interaction evidence does not meet current curation criteria (2) or because they are now believed to be part of larger complexes. These complexes remain on a watch list and will be added if sufficient evidence becomes available. Complex Portal complexes also contain 94% of CYC2008 co-complex pairs while CYC2008 complexes only contain 66% of Complex Portal co-complex pairs (Figure 3C).

The IntAct yeast interactome contains a total of 124 918 yeast–yeast binary interactions containing 5850 unique proteins or 97% of the yeast proteome (proteome = 6049 proteins) and 18 interactions between a yeast protein and a yeast complex. A topological clustering analysis of the IntAct yeast interactome was performed using the Cytoscape App ClusterONE, restricting accepted clusters to those with three or more proteins. The resulting clusters encompassed only just over half the proteome (3280 proteins, 54%) and predicted 332 complexes from low-throughput publications (IntAct-LT) and 689 complexes from high throughput publications (IntAct-HT) (Table 1). Only a third of the proteome was present in the 400 YHTP2008 predicted complexes based on high-throughput data (1911 proteins, 32%).

Complex sizes (Figure 2) are difficult to compare as the minimum sizes are determined by the curation strategies (see Table 1 for a reference of which datasets contain homomers and dimers) and the maximum sizes determined by the handling of paralogous proteins. Where possible, Complex Portal curates separate complexes for each paralogous protein but in the case of the ribosomal subunits it creates sets for each paralogous pair. Similarly, CYC2008 often includes each paralogous gene locus in the same complex. The inclusion of paralogous proteins or loci in a complex artificially inflates its maximum (and with that the mean and median) size. Likewise, clustering algorithms tend to group paralogous proteins together. Therefore, the largest complexes are found in the predicted datasets of YHTP2008 and IntAct-HT. Excluding the ribosomal subunits that contain multiple paralogous pairs of proteins or loci, the maximum size of a complex in the Complex Portal is 73 and in CYC2008 is 44.

However, despite the issues with minimum and maximum complex sizes, the overall complex size distributions are very similar. The majority of complexes contain 10 or fewer unique proteins with a rapidly reducing tail. This is dataset-independent and demonstrates that most proteins function within a relatively small group of partners. There are a few larger complexes in the Complex Portal than in CYC2008. ClusterOne predicts no complexes >40 proteins/complex for the IntAct-LT dataset resulting in the smallest complex size distribution of all datasets. In comparison, IntAct-HT has the highest predicted complex size distribution of all datasets when ignoring the expanded ribosomal complexes. The IntAct-HT dataset includes many affinity purification experiments, which can identify large associations of co-purifying proteins which in turn result in more centralized and heavily connected areas of the underlying interactome. Such heavily connected areas in the interactome result in many overlapping clusters that have a tendency to get combined into superclusters by the ClusterOne algorithm.

We also compared the manually curated complexes with those predicted from experimental protein–protein interaction (PPI) evidence. The overlap between any curated and predicted dataset never exceeded 20% in any comparison with a Jaccard Index of 1.0 (Figure 3B). The IntAct-HT complexes contain an even smaller overlap with either of the curated complex datasets (7–8%) than the IntAct-LT or YHTP2008 complexes (13–17%). At the protein level, only 42–72% of proteins from an experimental dataset could also be found in a curated complex dataset while 68–81% of proteins in the curated datasets are also found in the experimental datasets (Figure 3A).

The low level of overlap between manually-curated and predicted complexes may be the result of a combination of factors: First, experimentally derived interactomes contain a lot more proteins than the complex datasets but incorporate fewer validated evidence than the often thoroughly and even functionally validated interaction evidence used to define curated complexes. Secondly, the need for a reductionist representation of the interactome, where multiprotein associations are reduced to binary pairs via spoke expansion methods introduces a bias in the internal topology of PPI evidence networks, potentially generating spurious associations. Finally, prediction algorithms are restricted to predicting heteromers and ClusterOne restricts clusters to size 3 and larger; therefore, any heterodimeric complexes are not included in the predicted datasets and were removed from the above comparisons for the overlap of complexes between the five datasets.

### Features of protein complexes, their proteins and genes that code for them

The properties of protein complex members were characterized using a panel of numerical and binary features (Supplementary Figure S1). Genes coding for proteins in complexes tended to have more genetic interactions and to be co-expressed with more genes. They were also more likely to be multifunctional, conserved across species and present more stable expression patterns. Additionally, they often coded for proteins with a higher percentage of disorder, higher PPI degree and were enriched for essential genes and nonessential genes with larger fitness defects. On the other hand, these genes were depleted for duplicates and were less likely to code for membrane proteins. Localization patterns changed slightly across datasets. Proteins in complexes tended to localize more often in the nucleus and the nucleolus than other proteins, while they were less likely to be found in the vacuole. To further explore this finding, complexes in the Complex Portal dataset were analyzed for annotations to nuclear and nucleic acid-related processes and functions (Table 2) taking advantage of the complex-specific GO annotations available for this dataset. More than half of complexes (304/589, 52%) are annotated to at least one of these 12 selected terms or their children (Supplementary Table S2). 65% of these complexes (197/304) are annotated to ‘GO:0005634 nucleus’ or a child term and 52% (159/304) to ‘GO:0006139 nucleobase-containing compound metabolic process’ or a child term. In all datasets, proteins found in complexes were also significantly over-represented in processes where large machineries are the predominant functional drivers such as replication, tran-

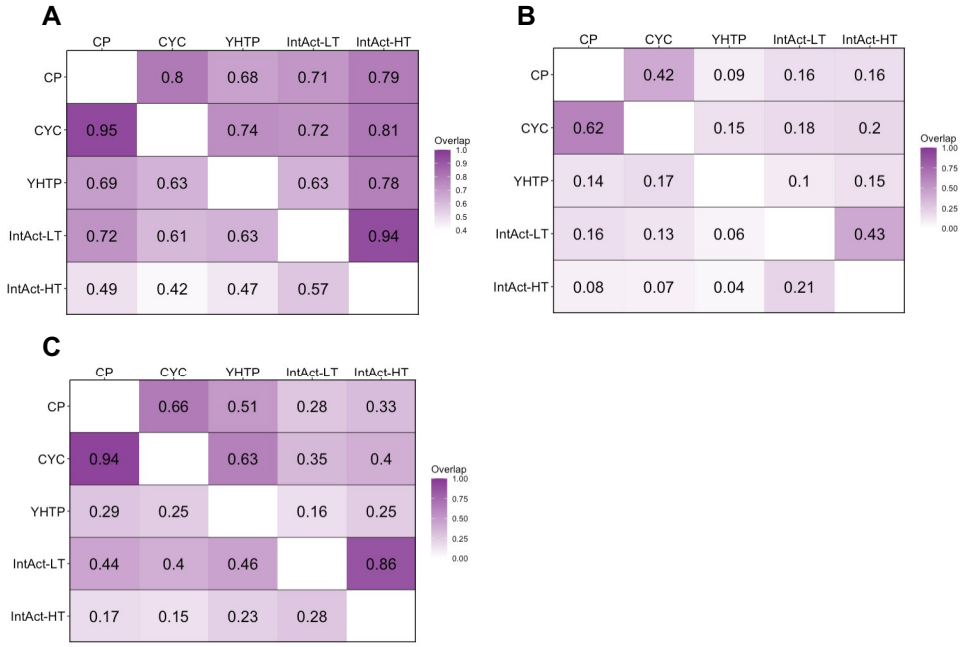


**Figure 3.** Fraction of (**A**) proteins (CP = 1930, CYC = 1624, YHTP = 1911, IntAct-LT = 1918, IntAct-HT = 3147), (**B**) complexes, based on Jaccard Index = 1.0 for complexes with a minimum of three protein participants (CP = 345, CYC = 236, YHTP = 208, IntAct-LT = 332, IntAct-HT = 689) and (**C**) co-complex pairs shared between two any datasets (CP = 15863, CYC = 11238, YHTP = 28146, IntAct-LT = 9808, IntAct-HT = 30493). Each row compares the overlap of both datasets to the total number of entities in the dataset given on the left.

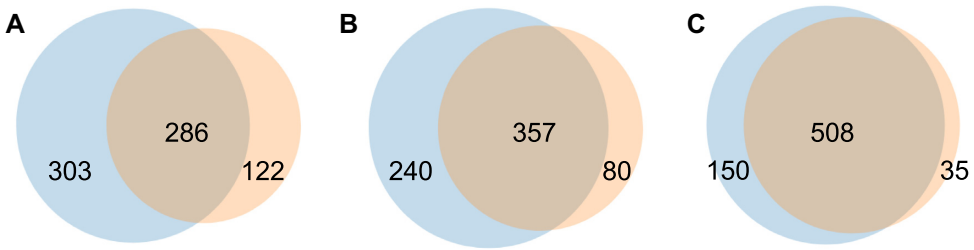


**Figure 4.** Overlap of complexes by protein identities and decreasing stringencies for complex membership between Complex Portal ($n$ = 589, left intercepting circle) and CYC ($n$ = 408, right intercepting circle). (**A**) JI = 1.0, (**B**) JI = 0.75, (**C**) JI = 0.5. JI = Jaccard Index. NB: Total numbers per dataset for JI = 0.75 and JI = 0.5 are higher than the absolute number of complexes per dataset as one complex can be broken down into more than one partial complex that matches a complex in the other dataset.

**Table 2.** Number of Complex Portal complexes annotated to nuclear and nucleic acid related GO terms

| GO term ID | GO term name | GO class | Complexes |
|---|---|---|---|
| GO:0006139 | Nucleobase-containing compound metabolic process | Biological_process | 159 |
| GO:0010467 | Gene expression | Biological_process | 101 |
| GO:0051276 | Chromosome organization | Biological_process | 73 |
| GO:0006974 | Cellular response to DNA damage stimulus | Biological_process | 38 |
| GO:0071826 | Ribonucleoprotein complex subunit organization | Biological_process | 14 |
| GO:0006997 | Nucleus organization | Biological_process | 2 |
| GO:0005634 | Nucleus | Cellular_component | 197 |
| GO:0003676 | Nucleic acid binding | Molecular function | 121 |
| GO:0140098 | Catalytic activity, acting on RNA | Molecular_function | 37 |
| GO:0140110 | Transcription regulator activity | Molecular_function | 16 |
| GO:0140097 | Catalytic activity, acting on DNA | Molecular_function | 15 |
| GO:0045182 | Translation regulator activity | Molecular_function | 12 |

scription, translation, and ER to Golgi and trans-Golgi transports. This reflects how such processes require a variety of tightly regulated multimolecular machineries whose diversity has been thoroughly explored in the literature. However, proteins found in complexes were underrepresented in many signaling, transportation and localization processes that are more often driven by single proteins. Importantly, most results were consistent across all five complex datasets.

### Multifunctionality

More than 70% of proteins in each dataset are only found in a single complex (Supplementary Figure S2), and there is no difference in this distribution between curated and predicted complexes. Only a few proteins from each dataset are found in two to five different complexes while CYC2008 contains only a few and YHTP2008 and IntAct-LT contain no proteins that occur in more than six complexes. Those proteins that were found in more than one complex were further analyzed. Most are found in complexes that carry the same components apart from the varying subunit and accordingly were deemed to be core subunits of these complexes. A check against the GO annotation of these proteins showed that many are catalytic core subunits of complexes such as cyclin-dependent kinases or ubiquitin ligases. In a recent analysis of datasets from several yeast interactome datasets, it was demonstrated that this long right-hand tail of a few proteins occurring in many complexes is almost always significantly different from a random distribution (38). The random distribution estimates that proteins should be found in a maximum of 6–9 complexes while in the real data some proteins occur in >20 complexes, matching our observations.

There are five proteins that are found in ≥4 complexes in the Complex Portal where the complexes are annotated to two or more unrelated pathways or complexes and three of these proteins are also found in more than one subcellular location when part of multiple complexes. Four of these proteins, H4 (P02309), LTV1 (P34078), SKP1 (P52286) and TAF14 (P35189), are regulatory subunits and one, PP12 (P32598), is a protein phosphatase (Supplementary Table S1). These five proteins have a relatively higher number of GO SLIM annotations compared to the rest ($P < 0.0005$, Supplementary Figure S3). All other complexes that share proteins are functionally related homologues.

### Biological assessment of complexes via omics data

Genetic interactions identify combinations of genes that yield unexpected phenotypes when simultaneously mutated. Negative genetic interactions identify cases with more severe phenotypes than expected given the individual mutant phenotypes, whereas in positive genetic interactions the resulting phenotype is healthier than expected. Both types of genetic interactions are a powerful tool for the characterization of genes and to elucidate the functional wiring of the cell (39).

Since genetic interactions identify potentially functional relationships between genes, we evaluated whether gene pairs coding for proteins within the same complex were enriched in genetic interactions using the global genetic interaction network (24). Genetic interactions have been explored for ∼52% of the co-complex pairs defined in the Complex Portal dataset. Of these, 30% and 10% of genes coding for co-complex pairs had negative and positive genetic interactions, respectively. These represent a 4.4- and 2.4-fold increase, respectively, over what was observed in background pairs, i.e. pairs of genes coding for proteins in different complexes ($P < 0.05$, Figure 5). Negative genetic interactions were particularly enriched between essential gene pairs coding for proteins in the same complex, which probably reflects the limited tolerance of the cell to sustain multiple deleterious mutations in essential complexes. On the other hand, positive genetic interactions were only enriched between nonessential genes coding for co-complex pairs. These positive interactions may identify nonessential protein complexes in which deletion of a member renders the whole complex inactive. Therefore, additional mutations on these complexes would not substantially impact fitness. The significant overlap between genetic interactions and co-complex relationships is in agreement with previous studies (24,40) and was consistent across the different complex datasets. However, the curated datasets and IntAct-LT showed a higher overlap with genetic interactions. A lower overlap of the high-throughput datasets, IntAct-HT and YHTP2008, with genetic interactions could be due to a larger fraction of indirect physical associations identified in weakly connected, large complexes in such studies. We found similar trends when comparing co-complex pairs to co-expression, co-localization and co-annotation datasets (Figure 6 and Supplementary Figure S4). In all cases, co-complex pairs had a higher overlap with these functional standards than background pairs and this overlap was particularly pertinent in the curated datasets. For instance, ∼90% of co-complex pairs in the curated datasets were co-expressed, whereas the overlap for the remaining datasets ranged from 41% to 76%. Additionally, we observed more similar transcript counts, expression variance, and protein abundance and half-life for co-complex pairs than background pairs (Figure 7 and Supplementary Figure S5), which reflects that members of the same protein complex tend to exhibit similar regulation patterns at a gene and protein level in order to act as a single coordinated biological unit.

Identifying the direct physical contacts within protein complexes can reveal sub-complex modules, improve the characterization of protein function, and help to interpret how mutations affect the phenotype. The Complex Portal is the only dataset that describes the internal connectivity of complexes, with detailed information for 237 complexes that have 3 or more participants. The functional relevance of these data were evaluated by comparing genetic interaction profiles (i.e., the set of genetic interactions of a gene) of direct and indirect contacts within protein complexes. These profiles are quantitative phenotypic signatures and revealed a higher similarity for gene pairs coding for proteins in direct contact (Figure 8; $P < 0.01$ for all pairwise comparisons). This suggests that, in protein complexes with unknown internal connectivity, the analysis of genetic inter-

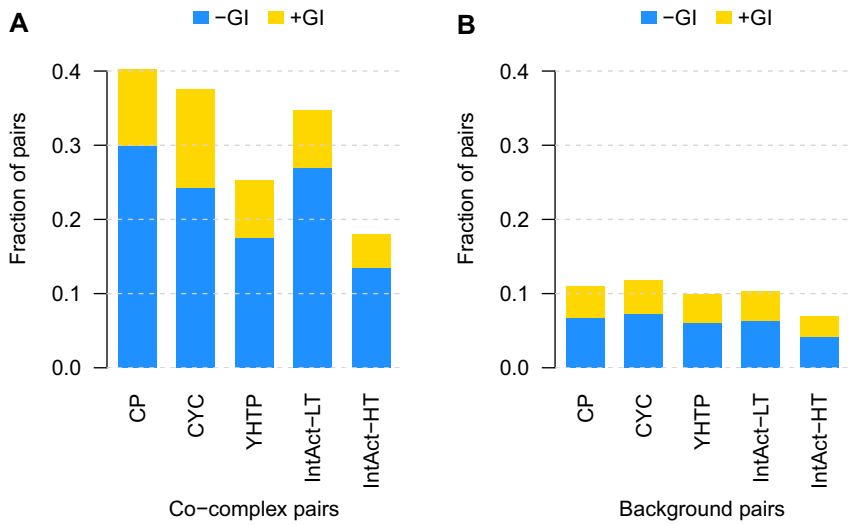


**Figure 5.** (**A**) Fraction of co-complex pairs from each complex dataset that overlaps with negative (dark bars) and positive (light bars) genetic interactions. (**B**) Fraction of protein pairs from each complex dataset that do not occur in the same complex (= background pairs) that overlaps with negative (dark bars) and positive (light bars genetic interactions.

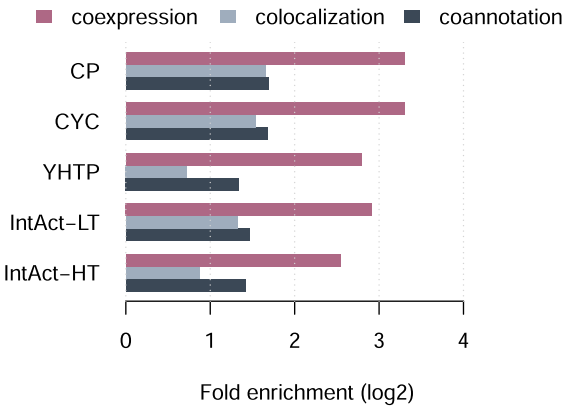


**Figure 6.** Fold enrichment of co-complex pairs compared to background pairs from all five datasets for co-expression, co-localization and co-annotation. All enrichments are statistically significant ($P < 0.05$)

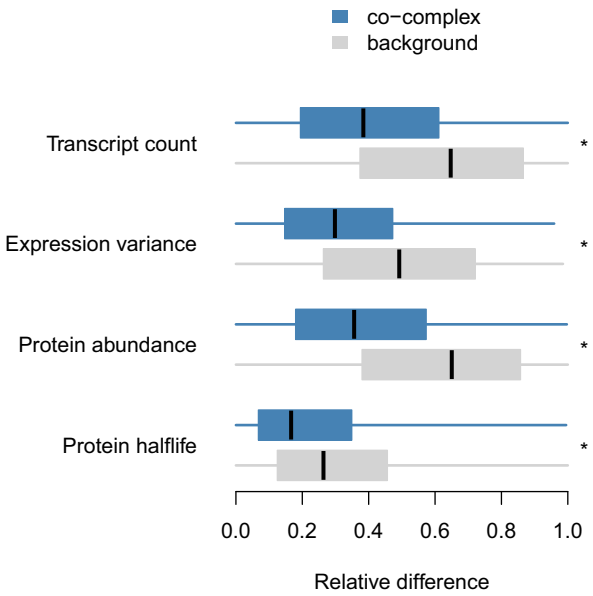


**Figure 7.** Relative difference in transcript counts, expression variance, protein abundance, and protein half life for co-complex and background pairs in the Complex Portal; * $P < 0.05$.

action profiles of the individual components may discriminate direct from indirect contacts.

## CONCLUSIONS

Our knowledge of the biology of *Saccharomyces cerevisiae* has substantially improved over the last 12 years. The Complex Portal now provides almost 50% more complexes than did the previous compendium, CYC2008 (22), and these include more protein components, details on nonprotein participants and more complex variants. The Complex Portal also provides a searchable website, a web service and three download formats.

Our set of curated yeast complexes shows a large overlap with previous curation efforts (i.e. CYC2008). However, these show a poor overlap when compared to predicted complexes. This may be due to large-scale affinity purification data producing clusters of apparently highly connected proteins as well as the presence of transient interactions in these datasets. This poor overlap also highlights that experimental protein–protein interactomes are a limited predictor for functional complexes which highlights the continuing need for a manually curated complex database.

Most proteins are found in only one complex and those found in two or more complexes tend to have the same function in multiple complexes. Only five proteins found in four or more complexes are linked to different processes showing that protein function is fairly conserved when they are part of complexes.

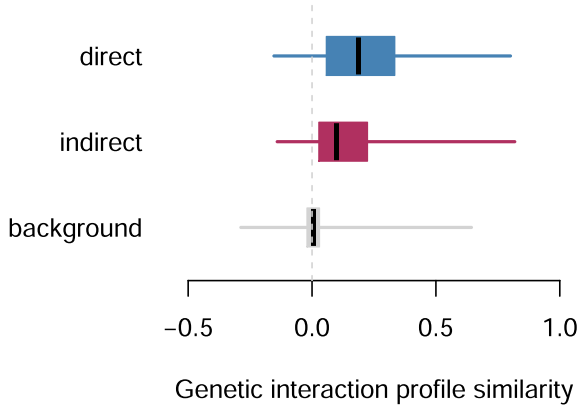


**Figure 8.** Genetic interaction profile similarities of gene pairs coding for proteins: in direct physical contact (top bar), in the same complex but not in contact (middle bar), in different complexes (bottom bar). Boxes represent second and third quartiles, whiskers first and forth. Horizontal lines in boxes represent the medians. All pairwise comparisons: $P < 0.05$

We highlight that there is a relative enrichment of multimolecular machines in the nucleus and the nucleolus. These complexes are often involved in nucleic acid-related metabolic processes like replication, transcription and translation, plus other processes where multimolecular assemblies are the predominant functional drivers such as ER to Golgi and trans-Golgi transports.

We found that the co-complex pairs overlap significantly with genetic interaction, co-expression, co-localization and co-annotation datasets, which highlights the functional relevance of co-complex membership and the potential of protein complex datasets to address questions of biological interest. Members of the same complex also tended to present more similar regulation patterns that reflect the role of the protein complex as a coordinated biological unit. Genes coding for co-complex pairs in physical contact exhibited more similar patterns of genetic interactions, illustrating that the structural organization within complexes is key to interpret the results of functional studies. Importantly, contact information within complexes is only available in Complex Portal and not in the other complex datasets.

To date, the Complex Portal yeast complexome has been used to validate complexes in several large-scale studies (38,40–43) and to define recurring patterns of complex topology (44–47). Our stable identifiers are used as annotation objects and cross-references in several other curated databases, such as IMEx consortium partners, Gene Ontology (48), Genome Properties (49), MatrixDB (50), SGD (13), Reactome, Signor (51,52) and Wikipathways (53) while other collaborations are under development, e.g. with PDBe (54). As we move to complete more complexomes, for example that of *Escherichia coli*, and continually improve our coverage of the human and mouse complexes, it will also be possible to improve our understanding of the evolution of these assemblies (55), and from there how the regulation of cellular processes has developed as organisms evolve.

We have shown how the Complex Portal yeast complexome is a key resource that significantly extends previously available datasets. Our commitment to keep it updated and freely accessible ensures the scientific community can count on a stable, high-quality reference set for the study of multimolecular machineries in yeast and other organisms.

We encourage our users to get in touch via the website if they find missing complexes or have suggestions on how to improve or extend our service.

## DATA AVAILABILITY

The complete yeast complexome is available for download from www.ebi.ac.uk/complexportal/download, the CYC2008 and YHTP2008 data from http://wodaklab.org/cyc2008/downloads and all files listing complexes and co-complexes used as input for our analyses have been deposited in Zenodo (10.5281/zenodo.4160609).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We would like to thank Eliot Ragueneau for his help with figure design and Colin Combe for his contributions to the development and maintenance of ComplexViewer.

## FUNDING

This work was supported by EMBL core funding (to B.M., H.H.); Open Targets [OTAR-044, OTAR02–048 to B.M., L.P., P.P.]; Wellcome Trust [212925/Z/18/Z to N.d-T., P.P.]; National Eye Institute (NEI) (to S.O.); National Human Genome Research Institute (NHGRI) (to S.O.); National Heart, Lung, and Blood Institute (NHLBI) (to S.O.); National Institute on Aging (NIA) (to S.O.); National Institute of Allergy and Infectious Diseases (NIAID) (to S.O.); National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (to S.O.); National Institute of General Medical Sciences (NIGMS) (to S.O.); National Cancer Institute (NCI) (to S.O.); National Institute of Mental Health (NIMH), National Institutes of Health [U24HG007822 to S.O.]; Ramon y Cajal fellowship [RYC-2017–22959 to C.P.]; National Institutes of Health (to E.W.); National Human Genome Research Institute (NHGRI) [U41HG001315, U41HG002273, U41HG02223–17S1 to E.W.]. Funding for open access charge: EMBL-EBI.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Meldal,B.H.M., Forner-Martinez,O., Costanzo,M.C., Dana,J., Demeter,J., Dumousseau,M., Dwight,S.S., Gaulton,A., Licata,L., Melidoni,A.N. *et al.* (2015) The complex portal–an encyclopaedia of macromolecular complexes. *Nucleic Acids Res.*, **43**, D479–D484.
2. Meldal,B.H.M., Bye-A-Jee,H., Gajdoš,L., Hammerová,Z., Horácková,A., Melicher,F., Perfetto,L., Pokorný,D., Lopez,M.R., Türková,A. *et al.* (2019) Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.*, **47**, D550–D558.
3. Giurgiu,M., Reinhard,J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Ruepp,A. (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.*, **47**, D559–D563.

4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
5. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
6. consortium,P.D.B. (2019) Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
7. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
8. Sehnal,D., Deshpande,M., Vařeková,R.S., Mir,S., Berka,K., Midlik,A., Pravda,L., Velankar,S. and Koča,J. (2017) LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. *Nat. Methods*, **14**, 1121–1122.
9. Papatheodorou,I., Moreno,P., Manning,J., Fuentes,A.M.-P., George,N., Fexova,S., Fonseca,N.A., Füllgrabe,A., Green,M., Huang,N. *et al.* (2020) Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*, **48**, D77–D83.
10. Sivade Dumousseau,M., Alonso-López,D., Ammari,M., Bradley,G., Campbell,N.H., Ceol,A., Cesareni,G., Combe,C., De Las Rivas,J., Del-Toro,N. *et al.* (2018) Encompassing new use cases - level 3.0 of the HUPO-PSI format for molecular interactions. *BMC Bioinformatics*, **19**, 134.
11. Mewes,H.W., Frishman,D., Mayer,K.F., Münsterkötter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. and Stümpflen,V. (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005., MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
12. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
13. Wong,E.D., Skrzypek,M.S., Weng,S., Binkley,G., Meldal,B.H.M., Perfetto,L., Orchard,S.E., Engel,S.R., Cherry,J.M. and SGD Project (2019) Integration of macromolecular complex data into the Saccharomyces Genome Database. *Database*, **2019**, baz008.
14. Orchard,S., Kerrien,S., Abbani,S., Aranda,B., Bhate,J., Bidwell,S., Bridge,A., Briganti,L., Brinkman,F.S.L., Brinkman,F. *et al.* (2012) Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods*, **9**, 345–350.
15. IMEx Consortium Curators, Del-Toro,N., Duesbury,M., Koch,M., Perfetto,L., Shrivastava,A., Ochoa,D., Wagih,O., Piñero,J., Kotlyar,M. *et al.* (2019) Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set. *Nat. Commun.*, **10**, 10.
16. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
17. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **98**, 4569–4574.
18. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
19. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dümpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
20. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
21. Aloy,P., Böttcher,B., Ceulemans,H., Leutwein,C., Mellwig,C., Fischer,S., Gavin,A.C., Bork,P., Superti-Furga,G., Serrano,L. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.
22. Pu,S., Wong,J., Turner,B., Cho,E. and Wodak,S.J. (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, **37**, 825–831.
23. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
24. Costanzo,M., VanderSluis,B., Koch,E.N., Baryshnikova,A., Pons,C., Tan,G., Wang,W., Usaj,M., Hanchard,J., Lee,S.D. *et al.* (2016) A global genetic interaction network maps a wiring diagram of cellular function. *Science*, **353**, aaf1420.
25. Nepusz,T., Yu,H. and Paccanaro,A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods*, **9**, 471–472.
26. Huttenhower,C., Hibbs,M., Myers,C. and Troyanskaya,O.G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22**, 2890–2897.
27. Huh,W.-K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
28. Koch,E.N., Costanzo,M., Bellay,J., Deshpande,R., Chatfield-Reed,K., Chua,G., D'Urso,G., Andrews,B.J., Boone,C. and Myers,C.L. (2012) Conserved rules govern genetic interaction degree across species. *Genome Biol.*, **13**, R57.
29. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
30. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztányi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L. *et al.* (2013) $D^2P^2$: database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.
31. Leeuwen,J., Pons,C., Tan,G., Wang,J.Z., Hou,J., Weile,J., Gebbia,M., Liang,W., Shuteriqi,E., Li,Z. *et al.* (2020) Systematic analysis of bypass suppression of essential genes. *Mol. Syst. Biol.*, **16**, e9828.
32. Balakrishnan,R., Park,J., Karra,K., Hitz,B.C., Binkley,G., Hong,E.L., Sullivan,J., Micklem,G. and Cherry,J.M. (2012) YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford)*, **2012**, bar062.
33. Babu,M., Vlasblom,J., Pu,S., Guo,X., Graham,C., Bean,B.D.M., Burston,H.E., Vizeacoumar,F.J., Snider,J., Phanse,S. *et al.* (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. *Nature*, **489**, 585–589.
34. van Leeuwen,J., Pons,C., Mellor,J.C., Yamaguchi,T.N., Friesen,H., Koschwanez,J., Ušaj,M.M., Pechlaner,M., Takar,M., Ušaj,M. *et al.* (2016) Exploring genetic suppression interactions on a global scale. *Science*, **354**, aag0839.
35. Usaj,M., Tan,Y., Wang,W., VanderSluis,B., Zou,A., Myers,C.L., Costanzo,M., Andrews,B. and Boone,C. (2017) TheCellMap.org: A Web-Accessible Database for Visualizing and Mining the Global Yeast Genetic Interaction Network. *G3 (Bethesda)*, **7**, 1539–1549.
36. Giglio,M., Tauber,R., Nadendla,S., Munro,J., Olley,D., Ball,S., Mitraka,E., Schriml,L.M., Gaudet,P., Hobbs,E.T. *et al.* (2019) ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Res.*, **47**, D1186–D1194.
37. Iudin,A., Korir,P.K., Salavert-Torres,J., Kleywegt,G.J. and Patwardhan,A. (2016) EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods*, **13**, 387–388.
38. Sartori,P. and Leibler,S. (2020) Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *Proc. Natl. Acad. Sci. USA*, **117**, 114–120.
39. Costanzo,M., Kuzmin,E., van Leeuwen,J., Mair,B., Moffat,J., Boone,C. and Andrews,B. (2019) Global Genetic Networks and the Genotype-to-Phenotype Relationship. *Cell*, **177**, 85–100.
40. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L.Y., Toufighi,K., Mostafavi,S. *et al.* (2010) The Genetic Landscape of a Cell. *Science*, **327**, 425–431.
41. Liebeskind,B.J., Aldrich,R.W. and Marcotte,E.M. (2019) Ancestral reconstruction of protein interaction networks. *PLoS Comput. Biol.*, **15**, e1007396.
42. Taggart,J.C. and Li,G.-W. (2018) Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst.*, **7**, 580–589.e4.
43. Michalak,W., Tsiamis,V., Schwämmle,V. and Rogowska-Wrzesińska,A. (2019) ComplexBrowser: A Tool for Identification and Quantification of Protein Complexes in Large-scale Proteomics Datasets. *Mol. Cell. Proteomics*, **18**, 2324–2334.

44. Pereira-Leal,J.B., Levy,E.D. and Teichmann,S.A. (2006) The origins and evolution of functional modules: lessons from protein complexes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **361**, 507–517.
45. Marsh,J.A. and Teichmann,S.A. (2015) Structure, dynamics, assembly, and evolution of protein complexes. *Annu. Rev. Biochem.*, **84**, 551–575.
46. Marsh,J.A., Rees,H.A., Ahnert,S.E. and Teichmann,S.A. (2015) Structural and evolutionary versatility in protein complexes with uneven stoichiometry. *Nat. Commun.*, **6**, 6394.
47. Ahnert,S.E., Marsh,J.A., Hernández,H., Robinson,C.V. and Teichmann,S.A. (2015) Principles of assembly reveal a periodic table of protein complexes. *Science*, **350**, aaa2245.
48. Kramarz,B., Roncaglia,P., Meldal,B.H.M., Huntley,R.P., Martin,M.J., Orchard,S., Parkinson,H., Brough,D., Bandopadhyay,R., Hooper,N.M. *et al.* (2018) Improving the gene ontology resource to facilitate more informative analysis and interpretation of Alzheimer's disease data. *Genes (Basel)*, **9**, 593.
49. Richardson,L.J., Rawlings,N.D., Salazar,G.A., Almeida,A., Haft,D.R., Ducq,G., Sutton,G.G. and Finn,R.D. (2019) Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. *Nucleic Acids Res.*, **47**, D564–D572.
50. Clerc,O., Deniaud,M., Vallet,S.D., Naba,A., Rivet,A., Perez,S., Thierry-Mieg,N. and Ricard-Blum,S. (2019) MatrixDB: integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.*, **47**, D376–D381.
51. Licata,L., Lo Surdo,P., Iannuccelli,M., Palma,A., Micarelli,E., Perfetto,L., Peluso,D., Calderone,A., Castagnoli,L. and Cesareni,G. (2020) SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res.*, **48**, D504–D510
52. Perfetto,L., Acencio,M.L., Bradley,G., Cesareni,G., Del Toro,N., Fazekas,D., Hermjakob,H., Korcsmaros,T., Kuiper,M., Lægreid,A. *et al.* (2019) CausalTAB: the PSI-MITAB 2.8 updated format for signalling data representation and dissemination. *Bioinformatics*, **35**, 3779–3785.
53. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K.A., Miller,R., Digles,D., Lopes,E.N., Ehrhart,F. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
54. Armstrong,D.R., Berrisford,J.M., Conroy,M.J., Gutmanas,A., Anyango,S., Choudhary,P., Clark,A.R., Dana,J.M., Deshpande,M., Dunlop,R. *et al.* (2020) PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.*, **48**, D335–D343.
55. Marchant,A., Cisneros,A.F., Dubé,A.K., Gagnon-Arsenault,I., Ascencio,D., Jain,H., Aubé,S., Eberlein,C., Evans-Yamamoto,D., Yachie,N. *et al.* (2019) The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *Elife*, **8**, e46754.