Published in final edited form as:
J Mol Biol. 2022 April 15; 434(7): 167478. doi:10.1016/j.jmb.2022.167478.

A similarity-based method for predicting enzymatic functions in yeast uncovers a new AMP hydrolase

Nir Cohen[#], **Amit Kahana**[#], **Maya Schuldiner**
Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 7610001, Israel

[#] These authors contributed equally to this work.

Abstract

Despite decades of research and the availability of the full genomic sequence of the baker's yeast *Saccharomyces cerevisiae*, still a large fraction of its genome is not functionally annotated. This hinders our ability to fully understand cellular activity and suggests that many additional processes await discovery. The recent years have shown an explosion of high-quality genomic and structural data from multiple organisms, ranging from bacteria to mammals. New computational methods now allow us to integrate these data and extract meaningful insights into the functional identity of uncharacterized proteins in yeast. Here, we created a database of sensitive sequence similarity predictions for all yeast proteins. We use this information to identify candidate enzymes for known biochemical reactions whose enzymes are unidentified, and show how this provides a powerful basis for experimental validation. Using one pathway as a test case we pair a new function for the previously uncharacterized enzyme Yhr202w, as an extra-cellular AMP hydrolase in the NAD degradation pathway. Yhr202w, which we now term Smn1 for Scavenger MonoNucleotidase 1, is a highly conserved protein that is similar to the human protein E5NT/CD73, which is associated with multiple cancers. Hence, our new methodology provides a paradigm, that can be adopted to other organisms, for uncovering new enzymatic functions of uncharacterized proteins.

Keywords

Saccharomyces cerevisiae ; Yhr202w; Smn1; AnalogYeast; Metabolomics; AMP

Introduction

One of the biggest revolutions in cell biology of eukaryotes came with the complete sequencing of the first eukaryotic genome, that of the yeast *Saccharomyces cerevisiae* (Goffeau et al., 1996). Computational analysis of the linear sequence uncovered a genome holding under 6000 genes that are mostly devoid of introns (Dujon, 2006; Ingolia et al., 2009) and lacking alternative splicing. This gives rise to a simple eukaryotic cell that can function with under 6000 protein products. Despite the simplicity of the yeast genome, a large portion of yeast proteins have not been functionally characterized. This hinders not only our understanding of cell biology but also of any biological process that relies on cells such as development and disease states.

One class of proteins that was intensively studied before the genome sequencing era is enzymes. Enzymes are molecular machines that can reduce the activation energy for metabolic reactions to occur (Förster et al., 2003) and hence to understand cellular metabolism it is essential to map enzyme substrates and products. The specific enzymatic step that enzymes performed was historically discovered either through biochemical purification step based on activity or through genetic screens coupled with biochemical follow-ups (Duntze et al., 1969; Knobling et al., 1975; Korch and Snow, 1973; Lynen, 1969; Masselot and de Robichon-Szulmajster, 1975; Schweizer et al., 1986). These approaches required a clear growth phenotype or testable biochemical activity. While this made it feasible to uncover the major players and pathways, it was not an optimal approach to identify and characterize enzymes that provide non-essential metabolites, that have isoenzymes, or other parallel pathways. The complete genome sequence of yeast provided an ability to identify the presence of new proteins and as such new enzymes, however many remained without a functional annotation.

In recent years an attempt to uncover new enzymatic functions was undertaken by utilizing untargeted metabolomics approaches. Metabolomic profiling provides a map of all metabolic changes that occur in the absence of specific genes (Fuhrer et al., 2017; Oliveira et al., 2012; Sauer, 2006; Sévin et al., 2017). Despite the ability of such methods to uncover enzymatic functions in some cases (Clasquin et al., 2011; Sévin et al., 2017), as a general rule, the immense metabolic rewiring that occurs in each strain has made it difficult to extrapolate exact gene functions from such data without a guiding hypothesis.

One way by which hypotheses could be formed for gene functions is by relying on the rich information that has been accumulating in multiple organisms at the biochemical, genomic and structural levels. To date, these data have not yet been fully tapped to create hypotheses as to protein functions and hence support the discovery of new enzyme functions.

Here, we derived sensitive sequence similarity (SSSP) predictions for all yeast proteins using the HHSearch platform (Steinegger et al., 2019). These predictions, now easily accessible through our website (https://www.weizmann.ac.il/molgen/AnalogYeast/) uncover new mammalian proteins that are similar to yeast proteins not previously found by simple sequence comparisons alone. Together with the simple sequence-based predictions, this suggests that a larger fraction of the yeast proteome is conserved to humans than previously appreciated.

Utilizing the knowledge that accumulated in other organisms, from bacteria through plants, invertebrates and vertebrates, we derive functional predictions for yeast proteins. We focused on the fraction of the proteome that has unknown function and specifically, on the likely enzymes in this group. By crossing the list of uncharacterized enzymes with a map of missing ones in known pathways, we provide a new methodology for predicting enzymatic functions of uncharacterized proteins. Using a test case of the protein product of *YHR202W*, we show how such predictions can be rapidly validated using metabolomic pipelines. These allow us to support the role of Yhr202w (which we now name, Smn1 for Scavenger MonoNucleotidase 1) as a newly-identified AMP hydrolase.

Results

Sensitive sequence similarity predictions expand the known degree of similarity between yeast and human proteomes

Ever since the establishment of the Baker's yeast as a leading model for eukaryotic cells, there has been an effort to characterize the exact extent of similarity between this simple unicellular organism and human cells. Previous efforts started even before the complete compilation of the yeast or human genome sequences (Botstein and Fink, 1988) and suggested that as little as 25% of genes in yeast will have human homologs. These efforts were simplified as whole genome sequences became available (Botstein et al., 1997) leading to a higher degree of predicted homology (31%). A compendium of many sequence-based prediction algorithms gathered by the "Quest for orthologs" database (Altenhoff et al., 2020) presented the most comprehensive comparison to date and suggested that ~ 55% of yeast gene products have homologs in humans based on sequence (Figure 1, Supplementary Table 1).

However, 1 billion years of evolution have created proteins that are similar in function and in structure but have little similarity at the sequence level. Such proteins require identification using more sensitive approaches. Here, we took advantage of such a tool, HHSearch (Steinegger et al., 2019) that was optimized to predict similarity based on amino acid sequence and secondary structure predicted from the amino acid sequence. Importantly, while most of these similarities result from divergent evolution, others might be the result of convergent evolution hence we use the term similarity and not homology.

We extracted such sensitive sequence similarity predictions (SSSPs) using HHSearch on all yeast proteins and found that over 40% of yeast proteins had predicted similar proteins in the human proteome (Figure 1) (Supplementary Table 2). Since the available databases for this method are based on the limited available protein structures and protein domains, it is not surprising that the number of similar proteins that we found based on this strategy is less than those found based on the sequence comparison alone. Importantly, the SSSP did not simply constitute a subset of the 55% of homologs based on simple sequence comparisons. In fact, we found similar proteins for an additional 10% of the yeast proteome bringing the overall predicted similarity to humans to be about two thirds of the yeast proteome (Figure 1). This stresses, again, the validity of using yeast as a representative model for understanding conserved cellular functions.

A large fraction of the yeast proteome remains uncharacterized

To truly understand cellular function, it is essential to know the activity of the constituting proteins. However, many yeast proteins have remained uncharacterized despite decades of efforts. To uncover the exact fraction of proteins for which no defined molecular function has been identified, we took two complementary approaches. First, we used text mining in the relevant fields to identify entries in the *Saccharomyces* Genome Database (SGD) (Cherry et al., 2012) that suggest a lack of known function (Figure 2A). In parallel, we manually curated all SGD descriptions of protein functions and generated a list of those that did not have sufficient support for a molecular function (Figure 2A, Supplementary Table 3). Each approach in as of itself showed that ~25% of yeast genes are yet to be characterized, and the combination of these approaches had a similar magnitude (Figure 2A). Hence, we decided to continue working with the list that we manually curated.

How have so many proteins evaded functional discovery despite decades of research? One hypothesis is that these proteins may be part of paralogous pairs or protein families. Having such similar proteins in the proteome may hinder functional annotation since backup by a paralog may reduce the phenotypes of losing a gene making the likelihood of capturing it by genetic methodologies smaller. To assay this hypothesis, we took a list of all yeast paralogous proteins (Fenech et al., 2020) and divided them into characterized and uncharacterized ones (from the curated list). We then asked what is their probability to have a paralog that is itself uncharacterized. We found that indeed proteins that are uncharacterized have a much bigger probability of having a paralog that is also not functionally characterized, supporting the back-up hypothesis (Figure 2B).

Another reason why these proteins may have been understudied is because they are yeast specific. To explore this hypothesis, we expanded our search for similar proteins to create SSSP of yeast proteins to all organisms as well as defined domains (Supplementary Table 2). To make these similarity predictions accessible to the yeast community, we organized them into an easily searchable database (https://www.weizmann.ac.il/molgen/AnalogYeast/). In this database (Using the term Analogy that includes both divergent and convergent evolution (Fitch, 1970)), for each yeast protein we present all predicted proteins based on SSSP as well as their description from UniProt (UniProt Consortium, 2021), Pfam (Mistry et al., 2021) and links to the respective databases. In addition, for relevant cases we added information regarding their involvement in human diseases (Rappaport et al., 2017) and/or their enzymatic activity (Chang et al., 2021).

Using these predictions, we found that the uncharacterized proteins are half as likely, relative to the whole proteome, to have similar proteins identified by SSSP in other model organisms (Figure 2C). This suggests that indeed conservation has both promoted research as well as, maybe, incentivized it. However, nearly 800 yeast proteins with unknown functions do have similar proteins in some species or have conserved domains. For these, advanced approaches can, and should, be used to gain a better understanding of function (Figure 2C). Importantly, 18% are conserved in humans. Hence, uncovering their function becomes critical not only for a better understanding of yeast cells (for biotechnological applications or as drug targets for antifungals) but also as a basis for better understanding conserved cellular functions.

Many uncharacterized proteins show similarity to enzymes in other organisms

Proteins can be assigned into functional categories such as structural proteins that define the building blocks of cellular architecture, regulatory and chaperoning proteins that function through binding of other biomolecules (proteins, DNA or RNA), and enzymes. The last century saw a huge burst of enzyme discovery through the fields of genetics and biochemistry, enabling us to map the enzymes carrying out the majority of the central metabolism pathways (Caspi et al., 2016; Karp et al., 2021; Lu et al., 2019). However, in recent years it is becoming clear that many additional enzymatic activities, that contribute to peripheral metabolism, signaling and stress responses, exist. Since such reactions are not essential, at least not under standard growth conditions, the identification of the enzymes carrying them out has been lagging behind.

To test whether some of the uncharacterized proteins that we defined have the potential to be enzymes we used the predictions from the SSSP algorithms to look for those proteins that show similarity to proteins with annotated enzymatic functions in any other organism. Importantly, since our analysis defines proteins as similar even if only a small part of the sequence is similar, sometimes not even in the active site of the enzyme, this type of analysis obviously gives rise to potential false positives and negatives. Taking that into account, our analysis uncovered that ~20% of the uncharacterized proteins may be enzymes (Figure 3). This is a slightly smaller number than the fraction of enzymes in characterized proteins (~28%) (Figure 3). This suggests that either proteins that are enzymes have been, historically, more likely to be identified, or rather that some enzymes are part of protein families that have not been structurally characterized in any organism.

Of the potential enzymes in our uncharacterized group, ~20% were already annotated in SGD as potential enzymes based on manual curation (Figure 3). Regardless, our analysis reveals a rough estimate of over 200 proteins that may be yeast enzymes not previously characterized (Supplementary Table 4). Moreover, these proteins show similarity to defined enzymes in other organisms, some of which have been studied and whose substrates or products are known. This suggests that we could directly use these similarities to predict the enzymatic functions of uncharacterized yeast proteins.

Metabolic gaps in known pathways can be filled by predicting enzyme functions based on similarity

Over the years there has been an effort to map metabolic pathways and the enzymes carrying out each step in multiple organisms (Karp et al., 2019, 2021). In yeast, this has led to the curation of over 572 reactions in 144 pathways consisting of 1417 enzymes. However, about 10% of the reactions (52) (Supplementary Table 5) still remain unaccounted for − meaning that it is clear that the step occurs but the enzyme carrying it out has not been identified or proven. These gaps of metabolic knowledge have been termed “pathway holes” and have awaited exploration.

We focused on these 52 pathway holes and probed whether any of our uncharacterized proteins for which we could find analogous enzymes, may account for the exact (or similar) reaction that occurs (Supplementary Table 5, Figure 4). We used Enzymatic Commission (EC) numbers, which define the exact enzyme function by a four-position hierarchical decision tree, to find proteins that can fit into the pathway “holes”. Generally, the first position of an EC number defines the general type of reaction, the second position defines a more specific reaction type, the third position defines the active enzymatic subclass and the fourth position defines the substrate (McDonald et al., 2001). We chose to focus on enzymes that are either identical in their function and substrate to the defined hole (the EC number is identical to the “hole” in all four positions) or that are largely similar in their enzymatic functions (the EC number is identical up to the third position).

We found that 15 out of the pathway holes have at least one candidate that fits in all four-positions and 42 out of the holes have candidates that fit three-positions (Figure 4). Despite the fact that a four-position match should be the identical enzymatic reaction, some holes had multiple four-position candidates. This suggests that either this step has a large enzymatic redundancy or that additional pathways with more refined substrates or distinct cellular locals should be characterized in the future.

Assigning an enzyme for a pathway hole in the periplasmic NAD degradation pathway

To test case our predictions we chose to focus on an important activity that has evaded discovery using previous approaches. The pathway is a nucleoside salvage pathway converting NAD+ to AMP and adenosine (Figure 5A). Interestingly, the enzyme carrying out the second step, converting AMP to adenosine, was suggested to reside in the periplasm (Bogan and Brenner, 2010) and its activity is described by the EC number 3.1.3.5, yet was never identified in yeast (Figure 5A). Our data suggested that two candidates are a fit for this “pathway hole” (highlighted in Figure 4) - the uncharacterized proteins Yhr202w and Ydl024c. Incidentally, both proteins have a predicted signal peptide and are soluble (Bernsel et al., 2009; Weill et al., 2019) (for a schematic of Yhr202w see Figure 5B), suggesting that they could be periplasmic. However, ydl024c was already annotated to an additional EC number (3.1.3.2) and was only matched to EC 3.1.3.5 based on similarity to a single human protein. In comparison, Yhr202w was not annotated to any other EC number and shows similarity to multiple enzymes of this function in several organisms: multiple bacterial species, Chinese cobras and humans. This made Yhr202w a more likely candidate. Moreover, *Δyhr202w* was shown to enable resistance to sodium selenite whose uptake is linked to that of phosphate, supporting a role for Yhr202w as a phosphatase as well as its extra-cellular localization. (Lazard et al., 2010; Pinson et al., 2004). We therefore first assayed whether Yhr202w is indeed a secreted enzyme by tagging the protein on its C’ with a Green Fluorescence Protein (GFP) and replacing its promotor with an inducible promotor (*GALpr*) to allow a strong inducible expression. To catch the intracellular phase by microscopy, before it is all secreted and cannot be imaged, we performed time-lapse microscopy (Supplementary Video) and could indeed capture a short phase where punctate structures, that could be secretory vesicles, appear (Figure 5C). At longer induction times very little intracellular signal was seen as would be expected from a secreted enzyme (Supplementary Video). To follow potential secretion, we performed a protein secretion assay whereby protein levels are tracked both intra-cellularly as well as in the medium (Figure 5D). As a control, we followed the levels of cytosolically expressed mCherry which should not be secreted and can serve to identify events of leakage of intracellular proteins to the medium. Indeed, Yhr202w-GFP could be found in both the media and the cellular fraction whereas mCherry could mostly be found in the cellular fraction supporting the notion that Yhr202w is a secreted enzyme.

Since Yhr202w is indeed secreted and has the correct enzymatic domains, we decided to directly assay its effect on the pathway substrates and products utilizing a metabolomics approach. We analyzed deletion and overexpression strains of Yhr202w and compared them to control strains by untargeted metabolomics on whole cell lysates using instrumentation that will allow us to focus on bases and other small polar metabolites (For a full list of metabolite changes see Supplementary Table 6). Using this we could confirm that indeed overexpression of Yhr202w causes an accumulation of adenosine in cells coupled with a reduction in AMP, ADP and ATP (Figure 5E). Conversely, the deletion mutant causes a reduction in adenosine with only a very small increase in ADP and ATP which are most likely highly regulated in their cellular levels (Figure 5E). As expected, overexpression of Yhr202w resulted in reduced NAD+ levels explained by the increased flux through the pathway (and suggesting that the limiting step of this pathway was Yhr202w) while the deletion resulted in an order of magnitude increase in NAD+ and NADH. This striking effect can be explained by the presence of a feedback loop completely blocking NAD+ degradation in the periplasm when Yhr202w is absent. We therefore decided to name Yhr202w Smn1 (Scavenger MonoNucleotidase 1).

Discussion

One of the big frontiers in modern cell biology is to map the “dark matter” of life hidden in the large percent of cellular proteins whose function is unknown. Our work shows that at least one quarter of proteins in the most highly studied eukaryotic model cell, yeast, are still uncharacterized. This highlights how fragmented our picture of cellular activity still is.

One of the places where this is clearest is yeast metabolism. While decades of biochemical studies have highlighted the central pathways and focused on core, essential metabolites, about 10% of these core pathways still have holes and more research is required to finalize their annotation. Moreover, it is now clear that yeast metabolism is more complex than previously thought. Our work suggests that ~20% of uncharacterized proteins are enzymes, highlighting at least 200 different reactions and pathways that could occur inside a yeast cell that have not yet been annotated. This may be especially relevant for metabolites that comprise single reactions or are stress induced.

Importantly, now is the time to uncover these new metabolic reactions and the enzymes mediating them. First, large metabolomic datasets showing changes in hundreds of metabolites for deletions of every cellular gene have already been published (Mülleder et al., 2016). However, by themselves these datasets can only provide limited answers to what specific enzymes are doing. This is because each perturbation causes a rewiring of cellular metabolism, resulting in myriads of changes in known, and uncharacterized, metabolites. Hence reaching a clear hypothesis as to the primary function of a single protein from the metabolic signature of its mutant has been challenging.

On the other hand, similarity predictions are now also emerging as a powerful tool due to the availability of multiple sequenced genomes, accumulation of structural data and powerful algorithms for comparing sequences and structures. However, such predictions are also, in as of themselves, often not enough. Our analysis shows this by examples of how multiple enzymes could potentially fulfill the requirement for a single specific pathway hole. Conversely, other holes have no candidate enzymes. Moreover, it is now clear that multiple enzymes have paralogs or isoenzymes that may perform an identical function but in a different compartment, slowing down correct assignment.

Luckily the last few years have brought about extensive mapping of protein localizations in yeast (Breker et al., 2014; Weill et al., 2018; Yofe et al., 2016). Hence, the true power is now to use emerging computational approaches for integrating the large amounts of available genomic data coupled with the availability of large scale metabolomic data (Ramirez-Gaona et al., 2017) and proteomic information to integrate them into predictive models (Amantonico et al., 2008; Ibáñez et al., 2013; Nobata et al., 2011; Urban et al., 2011). Indeed, many such efforts have been undertaken (from bacteria, through invertebrates to humans) to utilize the abundant genomic and proteomic data to uncover a function for uncharacterized proteins relying on computational approaches (for just a few examples see: (Kacsoh et al., 2019)(Garcia et al., 2019)(Zhang et al., 2018)).

However, any such predictions must still be tested to reach a more certain assignment of enzyme function. Our work shows that a powerful way to test these predictions is by measuring both deletion as well as over-expression versions of the enzymes. Using both mutants is important since deletions alone may not show a strong phenotype either due to rewiring of metabolism or because of buffering by enzyme redundancy. For example, Yhr202w has three similar proteins and there exist multiple other pathways to eliminate NAD+ or form AMP. Indeed, overexpression often gives a much stronger signature. However, the most powerful approach is to find metabolites that change in inverse directions in the two, inverse, mutants. While many metabolites change for each genetic manipulation, very few will display this unique characteristic signature, narrowing down the search range.

Generally, and regardless of the approach used, identifying a function for a yeast protein holds many advantages, even if its similar proteins were already studied in other organisms. First, uncovering functions for uncharacterized proteins helps to “catalog” the yeast proteome − essential for reaching at true understanding of this simple cell. In addition, yeast can serve as an evolutionary bridge between the various branches of organisms. This is especially useful when trying to uncover similarity between bacteria and mammals that have 3 billion years of evolution between them. In this case yeast can serve as a “springboard” − being only ~1 billion years distant from mammals. Finally, yeast can serve as an excellent model for functional studies as well as drug screening. For example, we identified Smn1 as a protein similar to human E5NT/CD73. While the human E5NT/CD73 was already shown to degrade AMP to adenosine (Jeffrey et al., 2020; Narravula et al., 2000) it was difficult to use for uncovering drug targets. Since it was shown to have a role in cancer progression and cyclic AMP (cAMP) signaling (Clayton et al., 2011; Gödecke, 2008; Narravula et al., 2000; Sciaraffia et al., 2014) uncovering such drugs may be beneficial. Interestingly we can also see changes in cAMP levels in the overexpression strains of Smn1 suggesting that yeast can serve as a model for E5NT/CD73 cellular activities. Now, having the yeast protein may provide a powerful platform for screening of novel inhibitors of this periplasmic protein − circumventing the need for drugs to enter cells.

More globally what is most surprising from our analysis is that yeast is much more similar to mammalian cells than previously thought. While previous assessments on similarity suggested that around half of the yeast proteome is conserved to mammals, the contribution of advanced computational approaches now suggests that it may be up to ~64% that are, in fact, conserved. This stratifies the belief that yeast is a superb model for human cellular activity and supports the need to continue and uncover the function of all of its proteins.

Materials and Methods

Computational analysis

Quest—The Quest for Orthologs database (Altenhoff et al., 2020) was downloaded on January 18th 2020, containing all homology hits between yeast and other organisms based on a combination of different homology algorithms and their benchmark criteria.

HHSearch—Protein sequences for all *saccharomyces cerevisiae* genes were obtained from SGD (Cherry et al., 2012) and rearranged to individual FASTA file formats using a homemade script. The individual FASTA files were submitted to a standalone HHSearch (from hhsuite3) (Steinegger et al., 2019) and searched against Pdb70 (Berman et al., 2000), PfamA V34 (Mistry et al., 2021), scop70-1.75 and scop40 (Andreeva et al., 2014, 2020). All proteins with the word “dubious” in their description were discarded, as well as hits with similarity score below 95 (out of 100). The result files were combined to a single.csv file using a homemade script. If the match was through a PDB structure, the host organism was added from the PDB description. Further information for each protein was added from UniProt (UniProt Consortium, 2021), including the indicated EC numbers (McDonald et al., 2001). Additionally, the involvement of each human protein with specific diseases was added based on a MalaCards search (Rappaport et al., 2017) conducted on GeneCards (Stelzer et al., 2016) version V4.13 on February 26th, 2020. Further analyses were performed on this assembled database, from here on termed, AnalogYeast, using homemade scripts (https://github.com/Maya-Schuldiner-lab/AnalogYeast). Raw prediction results can also be found on our lab webpage: https://mayaschuldiner.wixsite.com/schuldinerlab/analogyeast

Curation of genes whose proteins have unknown functions—Using the SGD database (Cherry et al., 2012), we assembled three lists containing proteins of unknown function. The first is a text mining list, containing all protein entries that included the keywords “uncharacterized”, “unknown”, “predicted”, “associates with”, “possibly”, “presumed” and “putative”, under the categories “Brief Summary” and “Function Summary”. The second list was manually curated based on SGD descriptions. The third list is a union of the two preceding lists.

Metabolic Pathway Holes—The list of pathway holes was acquired from the SGD YeastPathways database (Karp et al., 2021) on June 11 2021.

Yeast strains, strain construction and culturing conditions—All strains in this study are based on the BY4741 laboratory strain (Brachmann et al., 1998). All information on strains, plasmids and primers can be found in Supplementary Table 7.

For metabolomics analysis all yeast strains were grown in standard synthetic dextrose (SD) medium (6.7 g/L yeast nitrogen base with ammonium sulfate, 2% glucose, and all necessary amino acids) from OD_{600} of~0.1 overnight in a 30°c incubator. Then the cultures were diluted back to OD_{600} of~0.1 and incubated again overnight. Next the cultures were back diluted a last time to OD_{600} of~0.1 and left to grow until OD_{600} of~0.5 in 40ml culture breathable tubes (LIFEGENE).

Secretion assay

Yeast strains were grown in standard synthetic dextrose (SD) medium (6.7 g/L yeast nitrogen base with ammonium sulfate, 2% galactose) from OD600 of~0.1 overnight at 30°c. Next, the cultures were back diluted to OD600 of~0.1 and let to grow until OD600 of~0.5. The media fraction was separated from the cells using centrifugation (3000g for 3min) and cells were washed with DDW. Both media and cell fractions were precipitated using 10% Trichloroacetic acid (TCA) (Sigma) for 20min on ice, centrifuged for 15min at 14000g at 4°c, the supernatant was aspirated, pellet was washed in cold acetone, dried at room temperature for 30min and resuspended in urea lysis buffer (8M urea in 50mM tris pH 7.5 and oComplete Protease Inhibitor (Roche)). The cells were beaten with 100µl of glass beads (scientific industries) for 10min at 4°c. Then 0.1% SDS and 50mM DTT were added to both the cells and the media fractions and boiled at 95°c for 5min. Glass beads and cell debris were removed and the samples were resolved on 4-20% precast polyacrylamide gel (Bio-Rad), transferred to nitrocellulose membrane (PALL), and probed with a monoclonal mouse α-cherry (ab125096, Abcam) and a polyclonal rabbit α-GFP (ab290, Abcam). Secondary antibodies were alexa680 α-rabbit (ab175773, Abcam) and alexa790 α-mouse (ab186695, Abcam) that enable scanning using an Odyssey imaging system (LI-COR Biosciences).

Microscopy of yeast strains

Imaging of yeast strains was performed using a VisiScope Confocal Cell Explorer system (VisView), composed of a Yokogawa spinning disk scanning unit (CSU-W1) coupled with an inverted microscope (IX83; ×60 oil objective; Olympus) at an excitation wavelength of 488 nm for GFP. Images were taken by a connected PCO-Edge sCMOS camera (PCO) controlled by VisView software.

Metabolomics

Sample preparation—Each culture was filtered using a filtration apparatus onto a 25mm nylon membrane (GVS), washed once with 5ml of DDW and filtered again. The filter was transferred to 5ml cold 50% acetonitrile in DDW, vortexed and snapped frozen in liquid nitrogen.

Sample mass spec analysis—Sample analysis was carried out by MS-Omics using a Thermo Scientific Vanquish LC coupled to Thermo Q Exactive HF MS. An electrospray ionization interface was used as ionization source. Analysis was performed in negative and positive ionization mode. The Ultra-performance liquid chromatography was performed using a slightly modified version of a previously described protocol (Hsiao et al., 2018). Peak areas were extracted using Compound Discoverer 3.1 (Thermo Scientific). Identification of compounds was performed at four levels; Level 1: identification by retention times (compared against in-house authentic standards), accurate mass (with an accepted deviation of 3ppm), and MS/MS spectra, Level 2a: identification by retention times (compared against in-house authentic standards), accurate mass (with an accepted deviation of 3ppm). Level 2b: identification by accurate mass (with an accepted deviation of 3ppm), and MS/MS spectra, Level 3: identification by accurate mass alone (with an accepted deviation of 3ppm).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Dr. Simon Fishilevich for providing the MalaCards source data. We thank Maayan Maron and Anastasia Zarankin from the Weizmann Institute of Science IT division for creating the AnalogYeast web site. We thank Dr. Adi Millman from the laboratory of Prof. Rotem Sorek for fruitful discussions. We thank Dr. Emma Fenech for sharing the list of all yeast paralogs. We thank Hadar Meyer for technical assistance. We would like to acknowledge the professional work by MS-Omics in metabolomics data acquisition. We thank Dr. Marton Megyeri, Dr. Eden Yifrach and Dr. Einat Zalckvar for critical reading of our manuscript.

This work in this manuscript is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (EU-H2020-ERC-CoG; grant name OnTarget, grant number 864068). MS is an incumbent of the Dr. Gilbert Omenn and Martha Darling Professorial Chair in Molecular Genetics.

References

Altenhoff AM, Garrayo-Ventas J, Cosentino S, Emms D, Glover NM, Hernández-Plaza A, Nevers Y, Sundesha V, Szklarczyk D, Fernández JM, et al. The Quest for Orthologs benchmark service and consensus calls in 2020. Nucleic Acids Res. 2020; 48: W538–W545. [PubMed: 32374845]

Amantonico A, Oh JY, Sobek J, Heinemann M, Zenobi R. Mass spectrometric method for analyzing metabolites in yeast with single cell sensitivity. Angew Chem Int Ed Engl. 2008; 47: 5382–5385. [PubMed: 18543269]

Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 2014; 42: D310–4. [PubMed: 24293656]

Andreeva A, Kulesha E, Gough J, Murzin AG. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Res. 2020; 48: D376–D382. [PubMed: 31724711]

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000; 28: 235–242. [PubMed: 10592235]

Bernsel A, Viklund H, Hennerdal A, Elofsson A. TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res. 2009; 37: W465–8. [PubMed: 19429891]

Bogan KL, Brenner C. 5'-Nucleotidases and their new roles in NAD+ and phosphate metabolism. New J Chem. 2010; 34: 845.

Botstein D, Fink GR. Yeast: an experimental organism for modern biology. Science. 1988; 240: 1439–1443. [PubMed: 3287619]

Botstein D, Chervitz SA, Cherry JM. Yeast as a model organism. Science. 1997; 277: 1259–1260. [PubMed: 9297238]

Brachmann CB, Davies A, Cost GJ, Caputo E, Li J, Hieter P, Boeke JD. Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. Yeast. 1998; 14: 115–132. [PubMed: 9483801]

Breker M, Gymrek M, Moldavski O, Schuldiner M. LoQAtE--Localization and Quantitation ATlas of the yeast proteomE. A new tool for multiparametric dissection of single-protein behavior in response to biological perturbations in yeast. Nucleic Acids Res. 2014; 42: D726–30. [PubMed: 24150937]

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016; 44: D471–80. [PubMed: 26527732]

Chang A, Jeske L, Ulbrich S, Hofmann J, Koblitz J, Schomburg I, Neumann-Schaal M, Jahn D, Schomburg D. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res. 2021; 49: D498–D508. [PubMed: 33211880]

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012; 40: D700–5. [PubMed: 22110037]

Clasquin MF, Melamud E, Singer A, Gooding JR, Xu X, Dong A, Cui H, Campagna SR, Savchenko A, Yakunin AF, et al. Riboneogenesis in yeast. Cell. 2011; 145: 969–980. [PubMed: 21663798]

Clayton A, Al-Taei S, Webber J, Mason MD, Tabi Z. Cancer exosomes express CD39 and CD73, which suppress T cells through adenosine production. J Immunol. 2011; 187: 676–683. [PubMed: 21677139]

Dujon B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet. 2006; 22: 375–387. [PubMed: 16730849]

Duntze W, Neumann D, Gancedo JM, Atzpodien W, Holzer H. Studies on the regulation and localization of the glyoxylate cycle enzymes in Saccharomyces cerevisiae. Eur J Biochem. 1969; 10: 83–89. [PubMed: 5345986]

Fenech EJ, Ben-Dor S, Schuldiner M. Double the fun, double the trouble: paralogs and homologs functioning in the endoplasmic reticulum. Annu Rev Biochem. 2020; 89: 637–666. [PubMed: 32569522]

Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Zool. 1975; 19: 99–113.

Förster J, Famili L, Fu P, Palsson BØ, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 2003; 13: 244–253. DOI: 10.1007/0-306-47590-1_7 [PubMed: 12566402]

Fuhrer T, Zampieri M, Sévin DC, Sauer U, Zamboni N. Genomewide landscape of gene-metabolome associations in Escherichia coli. Mol Syst Biol. 2017; 13: 907. [PubMed: 28093455]

Garcia DC, Cheng X, Land ML, Standaert RF, Morrell-Falvey J, Doktycz MJ. Computationally Guided Discovery and Experimental Validation of Indole-3-acetic Acid Synthesis Pathways. ACS Chem Biol. 2019; 14: 2867–2875. [PubMed: 31693336]

Gödecke A. cAMP: fuel for extracellular adenosine formation? Br J Pharmacol. 2008; 153: 1087–1089. [PubMed: 18264119]

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Chabot C, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. Life with 6000 genes. Science. 1996; 274: 546. [PubMed: 8849441]

Hsiao JJ, Potter OG, Chu T-W, Yin H. Improved LC/MS Methods for the Analysis of Metal-Sensitive Analytes Using Medronic Acid as a Mobile Phase Additive. Anal Chem. 2018; 90: 9457–9464. [PubMed: 29976062]

Ibáñez AJ, Fagerer SR, Schmidt AM, Urban PL, Watson W, Jefimovs K, Geiger P, Dechant R, Heinemann M, Zenobi R. Mass spectrometry-based metabolomics of single yeast cells. Proc Natl Acad Sci USA. 2013; 110: 8790–8794. [PubMed: 23671112]

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009; 324: 218–223. [PubMed: 19213877]

Jeffrey JL, Lawson KV, Powers JP. Targeting metabolism of extracellular nucleotides via inhibition of ectonucleotidases CD73 and CD39. J Med Chem. 2020; 63: 13444–13465. [PubMed: 32786396]

Kacsoh BZ, Barton S, Jiang Y, Zhou N, Mooney SD, Friedberg I, Radivojac P, Greene CS, Bosco G. New Drosophila Long-Term Memory Genes Revealed by Assessing Computational Function Prediction Methods. G3 (Bethesda). 2019; 9: 251–267. [PubMed: 30463884]

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinformatics. 2019; 20: 1085–1093. [PubMed: 29447345]

Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong WK, Subhraveti P, Caspi R, Fulcher C, et al. Pathway Tools version 230 update: software for pathway/

genome informatics and systems biology. Brief Bioinformatics. 2021; 22: 109–126. [PubMed: 31813964]

Knobling A, Schiffmann D, Sickinger HD, Schweizer E. Malonyl and palmityl transferase-less mutants of the yeast fatty-acid-synthetase complex. Eur J Biochem. 1975; 56: 359–367. [PubMed: 1100391]

Korch CT, Snow R. Allelic complementation in the first gene for histidine biosynthesis in SACCHAROMYCES CEREVISIAE. I. characteristics of mutants and genetic mapping of alleles. Genetics. 1973; 74: 287–305. [PubMed: 17248618]

Lazard M, Blanquet S, Fisicaro P, Labarraque G, Plateau P. Uptake of selenite by Saccharomyces cerevisiae involves the high and low affinity orthophosphate transporters. J Biol Chem. 2010; 285: 32029–32037. [PubMed: 20688911]

Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marcišauskas S, Anton PM, Lappa D, Lieven C, et al. A consensus S cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun. 2019; 10 3586 [PubMed: 31395883]

Lynen, F. Lipids. Elsevier; 1969. 17–33. [3]

Masselot M, de Robichon-Szulmajster H. Methionine biosynthesis in Saccharomyces cerevisiae. Molec Gen Genet. 1975; 139: 121–132. [PubMed: 1101032]

McDonald, AG, Boyce, S, Tipton, KF. eLS John Wiley & Sons Ltd ed. Chichester, UK: John Wiley & Sons, Ltd; 2001. 1–11.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021; 49: D412–D419. [PubMed: 33125078]

Mülleder M, Calvani E, Alam MT, Wang RK, Eckerstorfer F, Zelezniak A, Ralser M. Functional metabolomics describes the yeast biosynthetic regulome. Cell. 2016; 167: 553–565. e12 [PubMed: 27693354]

Narravula S, Lennon PF, Mueller BU, Colgan SP. Regulation of endothelial CD73 by adenosine: paracrine pathway for enhanced endothelial barrier function. J Immunol. 2000; 165: 5262–5268. [PubMed: 11046060]

Nobata C, Dobson PD, Iqbal SA, Mendes P, Tsujii J, Kell DB, Ananiadou S. Mining metabolites: extracting the yeast metabolome from the literature. Metabolomics. 2011; 7: 94–101. [PubMed: 21687783]

Oliveira AP, Ludwig C, Picotti P, Kogadeeva M, Aebersold R, Sauer U. Regulation of yeast central metabolism by enzyme phosphorylation. Mol Syst Biol. 2012; 8: 623. [PubMed: 23149688]

Pinson B, Merle M, Franconi J-M, Daignan-Fornier B. Low affinity orthophosphate carriers regulate PHO gene expression independently of internal orthophosphate concentration in Saccharomyces cerevisiae. J Biol Chem. 2004; 279: 35273–35280. [PubMed: 15194704]

Ramirez-Gaona M, Marcu A, Pon A, Guo AC, Sajed T, Wishart NA, Karu N, Djoumbou Feunang Y, Arndt D, Wishart DS. YMDB 2.0: a significantly expanded version of the yeast metabolome database. Nucleic Acids Res. 2017; 45: D440–D445. [PubMed: 27899612]

Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M, Lancet D. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017; 45: D877–D887. [PubMed: 27899610]

Sauer U. Metabolic networks in motion: 13C-based flux analysis. Mol Syst Biol. 2006; 2: 62. [PubMed: 17102807]

Schweizer M, Roberts LM, Höltke HJ, Takabayashi K, Höllerer E, Hoffmann B, Müller G, Köttig H, Schweizer E. The pentafunctional FAS1 gene of yeast: its nucleotide sequence and order of the catalytic domains. Mol Gen Genet. 1986; 203: 479–486. [PubMed: 3528750]

Sciaraffia E, Riccomi A, Lindstedt R, Gesa V, Cirelli E, Patrizio M, De Magistris MT, Vendetti S. Human monocytes respond to extracellular cAMP through A2A and A2B adenosine receptors. J Leukoc Biol. 2014; 96: 113–122. [PubMed: 24652540]

Sévin DC, Fuhrer T, Zamboni N, Sauer U. Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli. Nat Methods. 2017; 14: 187–194. [PubMed: 27941785]

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019; 20: 473. [PubMed: 31521110]

Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, et al. The genecards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics. 2016; 54: 1.30.1–1.30.33. [PubMed: 27322403]

UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021; 49: D480–D489. [PubMed: 33237286]

Urban PL, Schmidt AM, Fagerer SR, Amantonico A, Ibañez A, Jefimovs K, Heinemann M, Zenobi R. Carbon-13 labelling strategy for studying the ATP metabolism in individual yeast cells by micro-arrays for mass spectrometry. Mol Biosyst. 2011; 7: 2837–2840. [PubMed: 21850345]

Weill U, Yofe I, Sass E, Stynen B, Davidi D, Natarajan J, Ben-Menachem R, Avihou Z, Goldman O, Harpaz N, et al. Genome-wide SWAp-Tag yeast libraries for proteome exploration. Nat Methods. 2018; 15: 617–622. [PubMed: 29988094]

Weill U, Cohen N, Fadel A, Ben-Dor S, Schuldiner M. Protein Topology Prediction Algorithms Systematically Investigated in the Yeast Saccharomyces cerevisiae. Bioessays. 2019; 41 e1800252 [PubMed: 31297843]

Yofe I, Weill U, Meurer M, Chuartzman S, Zalckvar E, Goldman O, Ben-Dor S, Schütze C, Wiedemann N, Knop M, et al. One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy. Nat Methods. 2016; 13: 371–378. [PubMed: 26928762]

Zhang C, Wei X, Omenn GS, Zhang Y. Structure and Protein Interaction- Based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. J Proteome Res. 2018; 17: 4186–4196. [PubMed: 30265558]

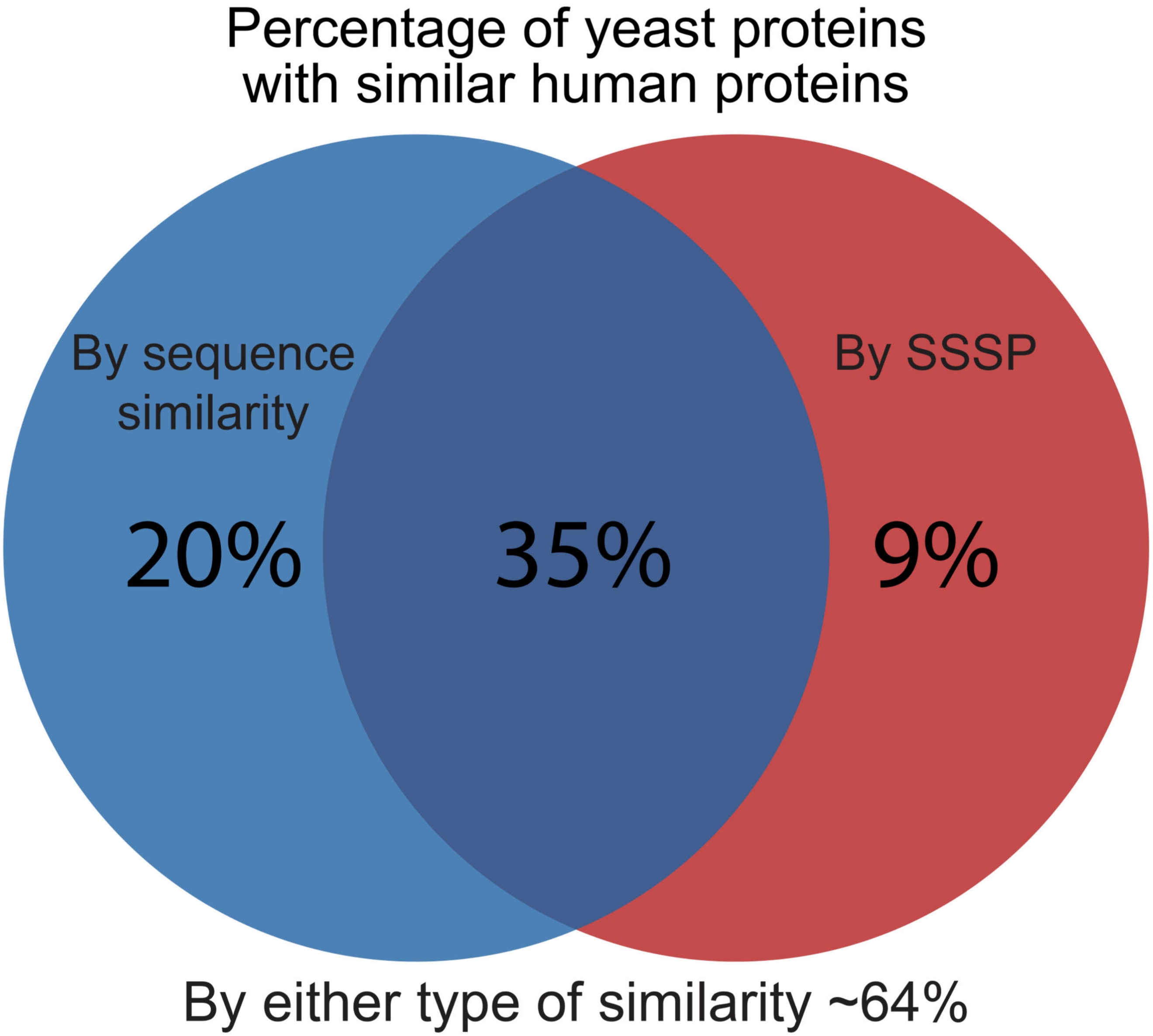

Figure 1. Sensitive sequence similarity predictions increase the percent of yeast proteins for which similar human proteins can be found.

A Venn diagram showing similarity levels between *S. cerevisiae* and human proteomes. Shown are the percentage of yeast proteins with similar human proteins uncovered by sensitive sequence similarity predictions (HHSearch), homologs uncovered by sequence alone (Quest for Orthologs), or the combination of both.

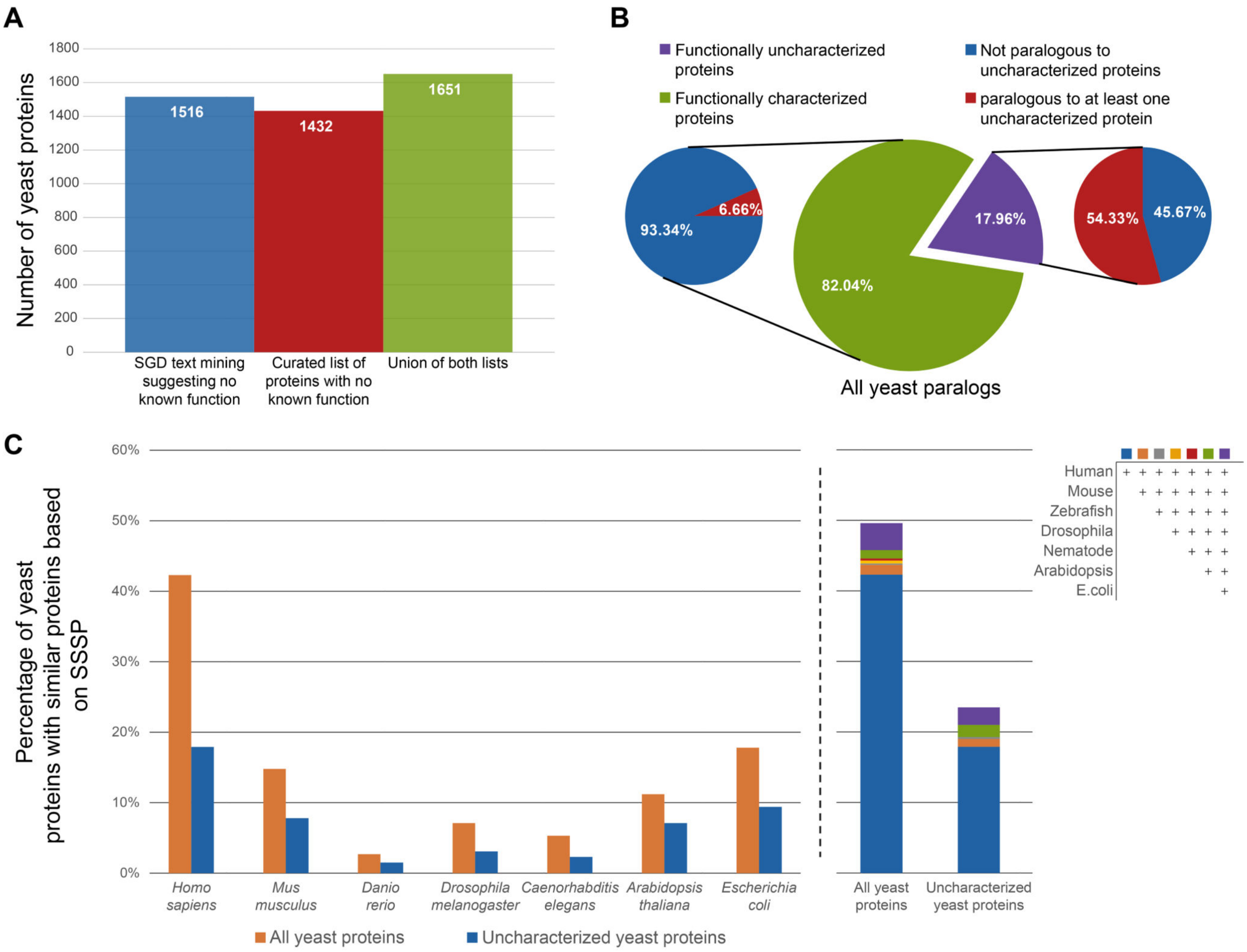

Figure 2. A large fraction of uncharacterized yeast proteins have similar proteins in other organisms

A) A bar plot of the number of yeast proteins that are still functionally uncharacterized. Shown are the numbers based on either text mining of the SGD database, or by manual curation as well as the union of both.

B) Pie chart showing the percent of functionally characterized vs uncharacterized yeast proteins out of those that have at least one paralog (center). Out of each group shown is the percentage that have at least one paralog that is also uncharacterized.

C) Bar plots presenting the percent of yeast proteins that have similar proteins in other model organisms − divided by either all proteins or only the uncharacterized ones. Left − separation to discrete model species. Right − the contribution of each model species to the overall similarity level.

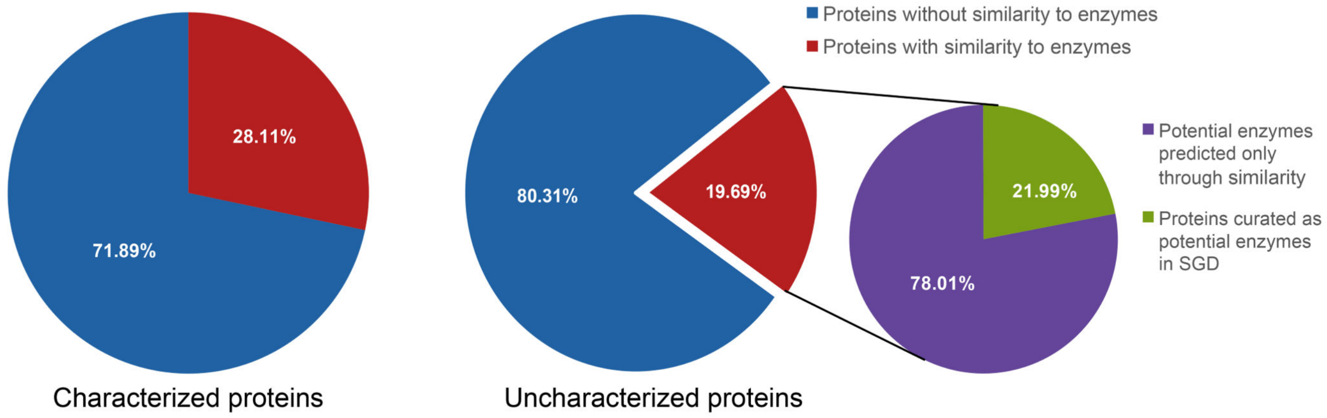

Figure 3. A substantial fraction of uncharacterized proteins have similarity to enzymes in other species.

Pie charts displaying the percent of proteins (either the functionally characterized yeast proteins or the uncharacterized ones) with predicted similarity to enzymes. From the Uncharacterized proteins an additional pie chart demonstrates the magnitude of novel detections enabled by the sensitive sequence searches.

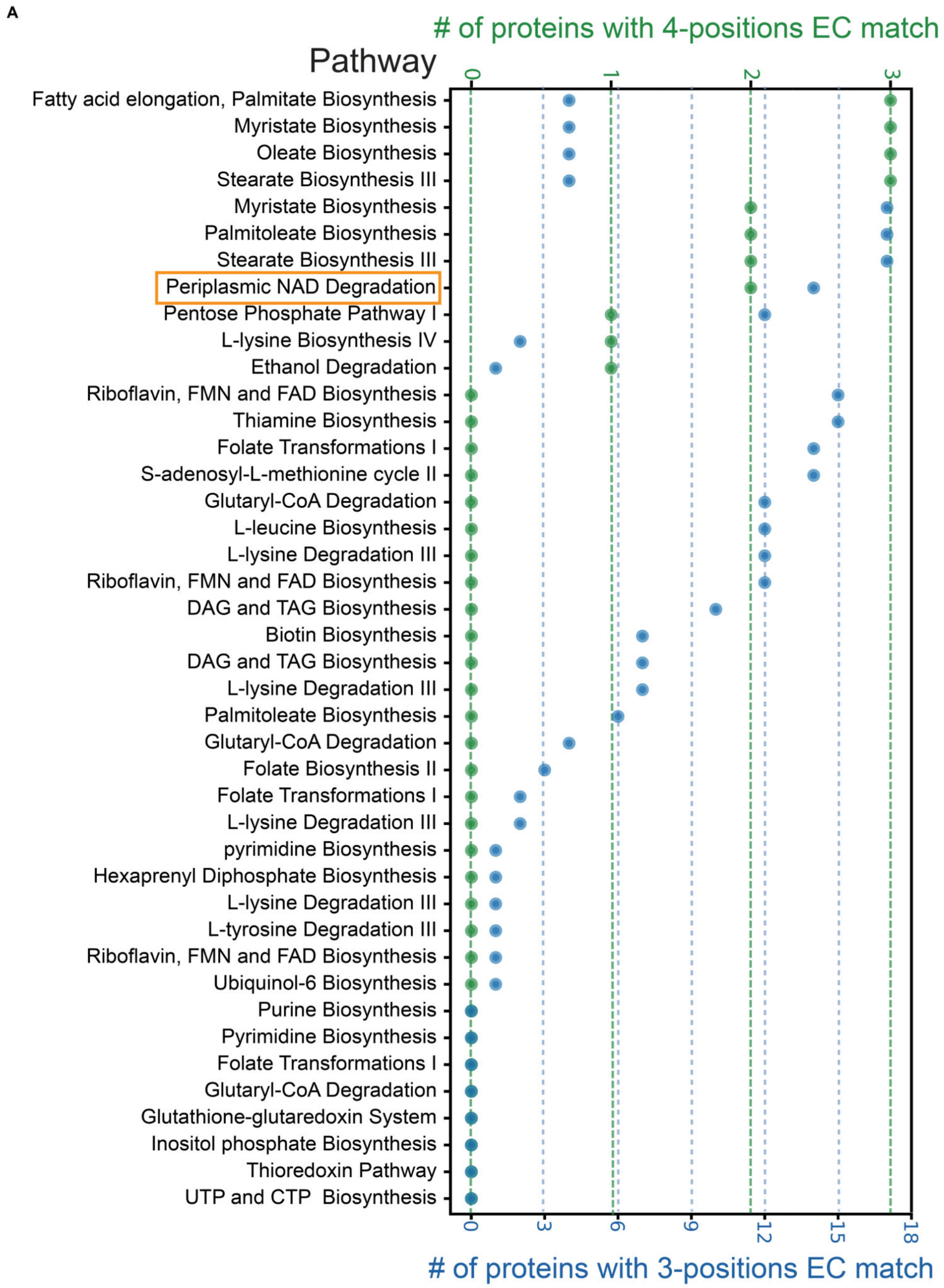

Figure 4. Matching predicted enzymes with metabolic pathway holes

A plot depicting all pathways in yeast that have an enzymatic step for which an enzyme has not yet been identified (hole). Plotted are the number of proteins with matching enzymatic-annotation (EC number) uncovered by SSSP. Green represents a full (4-positions) match. Blue represents a partial (3-positions) match. Highlighted in orange is the pathway on which we focus (Figure 5).

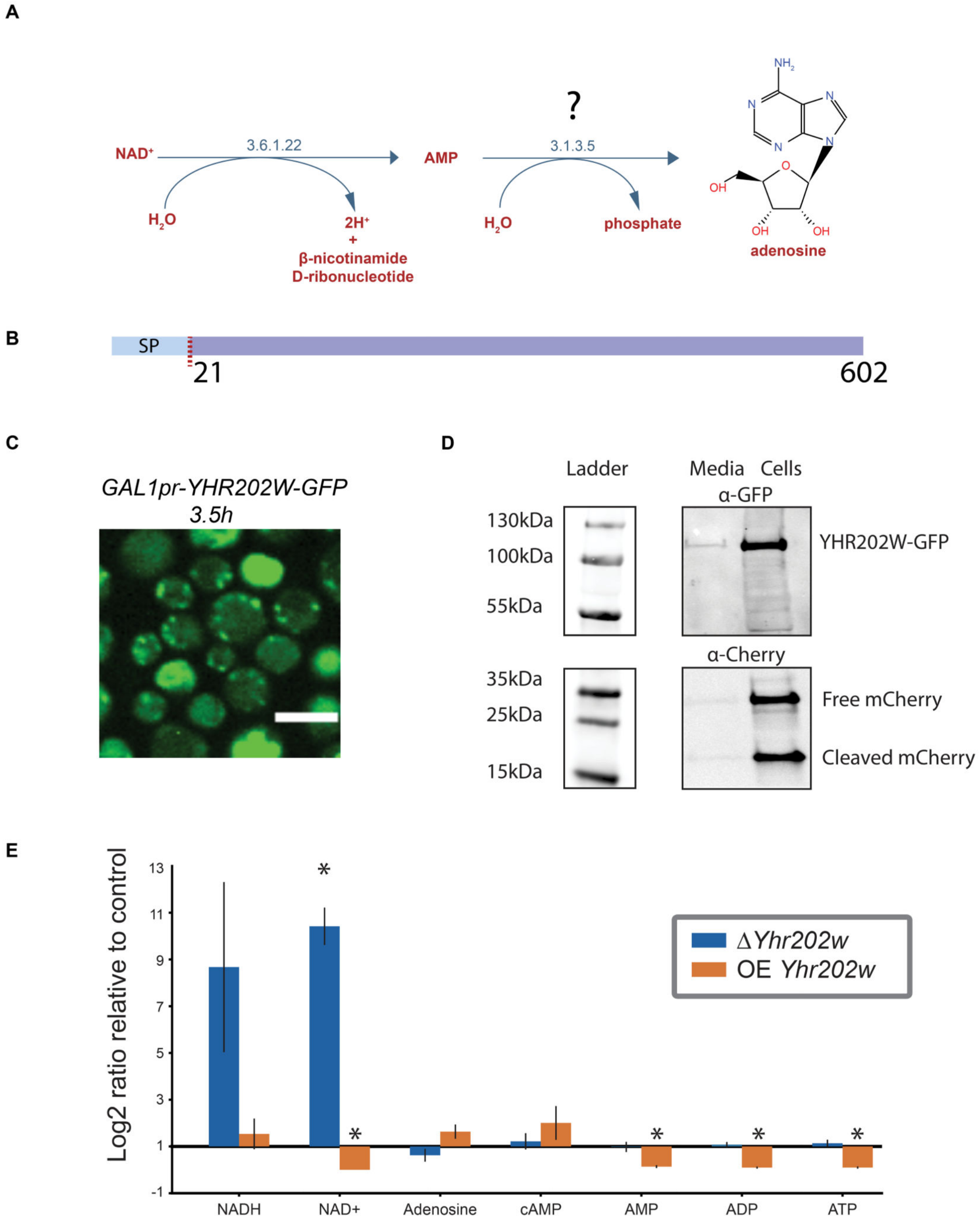

Figure 5. Assigning an enzyme for a pathway hole in the periplasmic NAD degradation pathway

A) The periplasmic NAD degradation pathway, consisting of two reactions, one with an EC number 3.6.1.22 giving rise to AMP from NAD+ and one, for which no enzyme has yet been described, giving rise to adenosine from AMP.

B) Yhr202w is a soluble protein containing a 21 amino acid signal sequence according to TOPCONS topology prediction.

C) Fluorescent microscopy image taken out of a time-lapse experiment showing the localization of Yhr202w-GFP under control of a *GALpr*. Images are shown 3.5 hours after activation of the *GAL* inducible promoter by transfer to growth in galactose containing medium. Scale bar =5μm

D) Yhr202w secretion analyzed by western blot. Yhr202w can be found in both the secreted and in the cellular fraction, while a soluble mCherry expressed under a constitutive promoter can be found mainly in the cellular fraction.

E) A bar plot showing the fold change of metabolites uncovered by metabolomics on strains in which Yhr202w/Smn1 was either deleted or overexpressed. The results depict fold change relative to genetically matched controls, with error-bars indicating the standard error of means. Fold-changes are marked with stars if they have a P-value < 0.005 in a T-test with a Benjamini-Hochberg correction.