A systematic bi-genomic split-GFP assay illuminates the mitochondrial matrix proteome and protein targeting routes

Yury S. Bykov[1,#,*], Solène Zuttion[2], Bruno Senger[2], Ofir Klein[1], Yeynit Asraf[1], Hadar Meyer[1], Hubert D. Becker[2], Róża Kucharczyk[3], and Maya Schuldiner[1,*]

[1] Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 7610001, Israel

[2] Université de Strasbourg, CNRS, Génétique Moléculaire, Génomique, Microbiologie, UMR7156, 67084 Strasbourg Cedex, France

[3] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, 02-106 Warszawa, Poland

[#] Current address: Quantitative Cell Biology, Rheinland-Pfälzische Technische Universität (RPTU), 67663 Kaiserslautern, Germany

* Correspondence to: yury.bykov@rptu.de; maya.schuldiner@weizmann.ac.il

Abstract

The majority of mitochondrial proteins are encoded in the nuclear genome and often lack clear targeting signals. Therefore, what constitutes the entire mitochondrial proteome is still unclear. We here build on our previously developed bi-genomic (BiG) split-GFP assay (Bader et al. 2020) to solidify the list of matrix and inner membrane mitochondrial proteins. The assay relies on one fragment (GFP_{1-10}) encoded in the mitochondrial DNA enabling specific visualization of only the proteins tagged with a smaller fragment, GFP_{11}, and localized to the mitochondrial matrix or the inner membrane. We used the SWAp-Tag (SWAT) strategy to tag every protein with GFP_{11} and mated them with the BiG GFP strain. Imaging the collection in six different conditions allowed us to visualize almost 400 mitochondrial proteins, 50 of which were never visualized in mitochondria before, and many are poorly studied dually localized proteins. We also show how this data can be applied to study mitochondrial inner membrane protein topology and sorting. This work brings us closer to finalizing the mitochondrial proteome and the freely distributed library of GFP_{11}-tagged strains will be a useful resource to study protein localization, biogenesis and interactions.

Keywords

Mitochondrial proteome, mitochondrial matrix, mitochondrial protein import, dual localization, eclipsed proteins, membrane protein topology, protein sorting, membrane protein insertion, stop-transfer mechanism, lateral sorting, yeast genetics, automated microscopy, split-GFP, SWAp-Tag (SWAT)

Introduction

The cell is an intricate molecular device. Full understanding of its functions requires a complete inventory of all the components tracing their positions and connections to each other. Of all eukaryotic organisms, *Saccharomyces cerevisiae*, or budding yeast, comes closest to having such an understanding thanks to its simple organization and compact genome allowing for versatile manipulations (Michaelis *et al*, 2023). A systematic collection of strains where every gene in the genome is fused to a green fluorescent protein (GFP) at the C-terminus was one of the first genetic instruments to offer a glimpse into the global distribution of yeast proteins across all organelles and its plasticity in different conditions (Huh *et al*, 2003; Breker *et al*, 2013). A new generation of SWAp-Tag (SWAT) libraries that allow quick substitution of the N-terminal or C-terminal tag with any DNA sequence of choice greatly expanded protein visualization possibilities (Yofe *et al*, 2016). The use of brighter fluorophores, and N-terminal tags combined with stronger promoters helped to gain more insight into intracellular protein distribution (Meurer *et al*, 2018; Weill *et al*, 2018; Dubreuil *et al*, 2019). Despite all these advances and systematic studies, many proteins are not yet fully characterized in terms of their subcellular and sub-organellar distribution.

The mitochondrial proteome is one of the most challenging to precisely map. First, many of Its ~800 proteins are dually localized and also have a certain fraction in an additional compartment (so called “eclipsed” distribution (Regev-Rudzki & Pines, 2007)). Such eclipsed proteins are often not clearly distinguished as mitochondrial especially if only a minority of the protein is in mitochondria and its signal is masked by high cytosolic levels. In addition, mitochondrial proteins are distributed to different sub-locations inside the organelle: outer membrane (OM), intermembrane space (IMS), inner membrane (IM), and matrix. Most of the proteins are produced in the cytosol and are imported based on the targeting signals encoded in their amino acid sequence (Herrmann & Bykov, 2023). On the OM these signals are recognized and threaded through by the translocase of the outer membrane (TOM) complex. To cross the IM, the imported proteins bind the translocase of the inner membrane (TIM) complex. Importing soluble matrix proteins requires engagement of the import motor that binds the TIM complex from the matrix side and uses the energy of ATP hydrolysis to aid translocation. The IM is a major protein delivery destination in mitochondria. Transmembrane domains (TMDs) can be either laterally sorted into the IM after translocation arrest in the TIM complex (stop-transfer mechanism), or completely imported into the matrix and inserted by the Oxa1 translocase via the so-called conservative sorting pathway (Stiller *et al*, 2016). Both, stop-transfer and conservative sorting pathways can act on different TMDs within one protein (Bohnert *et al*, 2010; Park *et al*, 2013). The mechanisms and machineries of mitochondrial protein import and

membrane insertion are well-studied thanks to experiments on isolated mitochondria; however the substrate repertoire of each pathway is still not fully mapped.

Since only a few artificially synthesized model substrates were used in *in vitro* sorting experiments, there is little understanding of how the mitochondrial protein import system handles hundreds of different proteins that are simultaneously imported in living cells (Bykov *et al*, 2020). For example, we have recently discovered that many mitochondrial proteins do not have conventional, well-predicted, targeting signals, but nevertheless are efficiently translocated into mitochondria (Bykov *et al*, 2022). Conversely, there are a lot of cytosolic proteins with high mitochondrial targeting signal prediction scores, that are not imported (Woellhaf *et al*, 2014; Mark *et al*, 2023). It is also not clear how the cell precisely directs the proteins to sub-mitochondrial locations. To study these mechanisms *in vivo*, it is essential to first obtain a detailed inventory of the mitochondrial proteome, information on the sub-mitochondrial localization and the topology of each protein.

High-throughput microscopy of GFP-tagged yeast strain collections enabled a great basis for mapping the mitochondrial proteome in living cells. However, this technique has been limited by its inability to show sub-mitochondrial location and topology. Another limitation is that proteins with eclipsed distribution and only a small mitochondria-localized fraction (echoform) are often missed while the proportion of the fraction might vary with growth conditions (Regev-Rudzki & Pines, 2007). Biochemical fractionation of the organelles and mitochondrial sub-compartments followed by mass-spectrometry-based proteomics can overcome these limitations. The most comprehensive mitochondrial protein distribution and topology inventories were produced using these approaches (Vögtle *et al*, 2017; Morgenstern *et al*, 2017; Di Bartolomeo *et al*, 2020). However, they are limited by the inability to truly purify the organelle or its sub-compartments leading to false positives and negatives.

Split-reporters are another powerful tool to detect protein localization in living cells with high sensitivity. A recent study that used a nuclear-encoded split *β*-galactosidase reporter indeed uncovered many new mitochondrial proteins (Mark *et al*, 2023).

In this work we aimed to combine the advantages of high-throughput microscopy that offers quick and sensitive protein visualization in living cells and the precision of split reporters to highlight individual mitochondrial sub-compartments. We made use of a recently developed Bi-genomic split-GFP assay (BiG Mito-Split-GFP, hereafter shortened as BiG Mito-Split) where a larger fragment of GFP, GFP_{1-10} is encoded in the mitochondrial DNA, and the small fragment GFP_{11} is fused with a nuclear-encoded protein (Bader *et al*, 2020). In this way, only the matrix-localized fraction of the GFP_{11}-tagged protein produces a fluorescence signal. The advantage of the BiG Mito-Split assay over the previously used split reporters is that matrix-limited

expression of GFP_{1-10} completely excludes the possibility of cytosolic interaction of GFP_{1-10} and GFP_{11} before import, an artifact that is hard to completely avoid when both fragments are encoded in the nuclear genome. To take this approach to the whole-proteome level we used the SWAT method and a C' SWAT acceptor library (Yofe *et al*, 2016; Meurer *et al*, 2018) to produce a whole-genomic library of C' tagged GFP_{11} proteins and after mating them with a strain expressing the mitochondrial GFP_{1-10}, tested each of them for the presence of a matrix localized fraction with high-throughput microscopy. This approach allowed us to directly visualize many mitochondrial proteins that were not accessible by microscopy before. By combining our data with targeting signal predictions, we show that a very small fraction of cytosolic proteins with high mitochondrial targeting sequences (MTS) prediction scores can actually be imported into the matrix, highlighting a strict selectivity filter for soluble proteins. We also show an unexpectedly low selectivity of TMD-containing protein sorting at the IM. Overall, our studies uncover new mitochondrial proteins and, beyond that, the collection (which will be freely distributed) and dataset will be a useful resource to enable *in vivo* studies of mitochondrial protein targeting, translocation and intra-mitochondrial sorting as well as more generally protein localization and interactions.

Results

Creation of a whole proteome GFP_{11} library enables accurate and systematic visualization of the mitochondrial matrix and inner membrane proteome

To perform the BiG Mito-Split assay on a whole proteome level we first used the SWAT method to create a haploid collection of strains (library) where in each strain one gene is fused with $3\times GFP_{11}$. We chose to use the C-SWAT library that allows C-terminal tagging (Meurer *et al*, 2018) since most mitochondrial proteins contain targeting signals at their N-termini. The original C-SWAT library contains ~5500 strains where each gene is fused to an acceptor cassette containing the *URA3* gene. Using automated mating and selection techniques we introduced a donor cassette expressing the $3\times GFP_{11}$ tag and MTS_{Su9}-mCherry as a mitochondrial marker. Recombination and negative selection for the loss of the *URA3* marker yielded a new haploid C'-$3\times GFP_{11}$ collection (Fig. 1).

To introduce mtDNA-encoded GFP_{1-10} we depleted this collection of its own mtDNA and immediately mated it with the BiG Mito-Split strain, thereby producing diploids with restored respiratory capacity and both fragments of split-GFP encoded by nuclear and mitochondrial genomes (Fig. 1).

The diploid BiG Mito-Split collection was imaged in six conditions representing various carbon sources and a diversity of stressors the cells can adapt to: logarithmic growth on glucose, post-

diauxic (stationary) phase after growth on glucose, growth on glucose in the presence of 1 mM dithiothreitol (DTT), inorganic phosphate (P_i) depletion, nitrogen starvation, and growth on oleic-acid-containing media. The imaging and initial visual analysis revealed 543 strains that showed a putative signal in at least one condition (Fig. 2A).

We selected these strains for a more detailed study and fluorescence signal quantification to verify the initial result of the visual analysis. The experiment was performed in two 384-well plates. We had the possibility to additionally include in the quantitation 123 non-fluorescent control strains. We also added 102 strains of proteins that were previously reported to be localized to the matrix or IM in proteomics studies (Vögtle *et al*, 2017; Morgenstern *et al*, 2017) and that we may have potentially overlooked in the visual analysis of whole-genomic screens (see Table S1 legend). To compare raw fluorescent measurements across different plates and conditions we normalized them to the average and standard deviation observed for multiple non-fluorescent controls present on each plate (Fig. S1A, see Materials and Methods) (The full dataset is provided in Table S1). Of note, out of 102 extra strains that were picked based on the previous annotations only three showed some signal. This means that our initial imaging had a very low false negative rate and that our two-step screening was an effective strategy to identify matrix and inner membrane proteins that can be visualized with the BiG Mito-Split collection.

Despite the highly diverse conditions that we picked – constituting both stressed and metabolic changes, the normalized fluorescence measurements in different conditions correlated very strongly with each other (Fig. 2A). The differences were most likely defined by general upregulation or downregulation of mitochondrial gene expression reflected in the linear regression slope. For example, the intensities are overall higher in stationary phase compared to logarithmic growth on glucose as expected since the former condition supposedly displays higher respiration. Our dataset did not reveal large groups of proteins that conditionally become imported into mitochondria under certain metabolic states or stresses. Proteins that were only detected in few conditions were mostly of low abundance (Table S1), so we could not conclude whether their expression or import are differentially regulated, or rather that we do not detect them in some conditions due to lower signal-to-noise ratios.

To compare our findings with published data, we created a unified list of 395 proteins that are observed with high confidence using our assay (Fig. S1B-D, Table S1). The majority of proteins that we identified are well-studied, exclusive mitochondrial matrix or IM proteins that were already reported by previous microscopy studies. Such proteins appear the same when tagged with a full-length fluorescent protein (FP) and when visualized with BiG Mito-Split showing strong mitochondrial signal and no signal in the cytosol (Ilv6, Fig. 2B). We also found known dually localized proteins like Fum1 who's mitochondrial fraction is clearly visualized by full-length FP fusion while the other localization (cytosol) is also clearly visible with the full-length

FP (Fum1, Fig. 2B). Finally, for some proteins mitochondrial localization was impossible to assign based on previously studied full-length FP fusions and they were for the first time visualized as mitochondrial in our dataset (Arc1, Fig. 2B). Overall, compared to all the previous microscopy studies using yeast collections (Huh *et al*, 2003; Breker *et al*, 2013; Weill *et al*, 2018; Meurer *et al*, 2018; Dubreuil *et al*, 2019) we visualized 56 new proteins (Table S1). Out of those, 39 were visualized in other locations meaning that in this study we managed to highlight eclipsed mitochondrial echoforms. There were also 17 proteins never before visualized by microscopy in high-throughput studies (Fig. 2C). Thus, our work significantly expands the microscopic toolkit for mitochondrial protein visualization.

BiG Mito-Split sensitivity and the definition of mitochondrial proteome

Next, we aimed to determine whether we can add new proteins to the mitochondrial proteome. For this, we compared our findings with data obtained by multiple alternate methods. As the most reliable sources we used manual annotations from the *Saccharomyces* Genome Database (SGD, yeastgenome.org (Wong *et al*, 2023)) that extensively cover individual protein localization deduced using such methods as fractionation, western blotting or manual fluorescence microscopy. From high-throughput annotations we used the two most confident proteomics datasets (Vögtle et al, 2017; Morgenstern et al, 2017). Compared to this unified list we still found eight proteins never before reported to be mitochondrial (Fig. 3A, Table S1).

It is harder to estimate how many proteins were missed in our assay. We expect to visualize all soluble matrix proteins and the IM proteins with their C-termini facing the matrix. There is no reliable annotated dataset on sub-mitochondrial protein distribution except the two mentioned proteomic studies, and only one accessible IM protein topology annotation (Morgenstern *et al*, 2017). When compared to the list of 194 soluble matrix proteins (Vögtle *et al*, 2017) we identify 136 (70%) which is superior to using a split-enzymatic system (Fig. S2A) (Mark *et al*, 2023). To check if the low abundance proteins were more likely to be overlooked by the BiG Mito-Split assay we compared the abundances of detected and undetected proteins and found that many low abundance proteins still can be visualized (Fig. 3B). Comparison with the high confidence mitochondrial proteome quantification also revealed that some groups of proteins might have preferred interaction with newly synthesized mtGFP$_{1\text{-}10}$ but did not show any specific protein groups that could not be detected by our method (Fig. S2B). Based on the presence of many low abundance proteins in our dataset, we conclude that the BiG Mito-Split assay can detect relatively low amounts of proteins translocated to the matrix, but some proteins can be missed regardless of their abundance. Hence, we only draw conclusions from the presence of signal and not from its absence.

We find eight proteins that were never connected to mitochondria before. They comprise five components of the cytosolic ribosome (r-proteins), one dubious open reading frame, the proteasome regulator Blm10, and a putative quinone oxidoreductase Ycr102c (Table S1). The high proportion of r-proteins that are very unlikely to have a function in mitochondria made us wonder if the mitochondrial proteome determined before is essentially complete and these eight proteins represent protein import noise caused by very high similarity to mitochondrial targeting sequences. To visualize how many proteins in the cytosol may be mistaken for mitochondrial ones we split all the non-mitochondrial proteome ranked by abundance into having high targeting signal prediction score and low score (Fig. 3C). The eight newly found proteins belong to both groups, but the ones with high prediction scores are highly abundant. Additional experiments with heterologous protein expression and *in vitro* import will be required to confirm the mitochondrial import and targeting mechanisms of these eight non-mitochondrial proteins. The data highlights that out of hundreds of very abundant proteins with high prediction scores only few are actually imported. Our dataset confirms that the selectivity filter recognizing targeting signals at the OM is very strict and BiG Mito-Split collection can be a valuable tool to study the mechanisms maintaining this selectivity.

Visualization of poorly studied dually localized proteins

We next aimed to determine if our assay helped to confirm localizations of poorly studied proteins that might have potential new functions in mitochondria. We compared our 56 newly visualized proteins (Fig. 2C, Table S1) with manual annotations and proteomic data and noticed that there are 20 proteins that were only found to be mitochondrial in high-throughput studies and not confirmed by other methods. We also uncovered a second localization for the abundant cytosolic glycerol-3-phosphate phosphatase Gpp1 that is in a reverse situation. It escaped detection in all proteomic studies and is not annotated to be mitochondrial in SGD. However a work that aimed to find new mitochondrial proteins arising from alternative translation starts confirmed that Gpp1-Protein A fusion has a mitochondrial fraction (Monteuuis *et al*, 2019). A full-length FP fusion of Gpp1 gives only cytosolic signal (Dubreuil *et al*, 2019). Thus, we confirm that Gpp1 is a dually localized protein with eclipsed distribution.

We wondered if the 20 poorly studied proteins are also dually localized. The full-length FP fusion data is not available for them, and we relied on the whole-cell to mitochondria abundance ratio previously calculated (Vögtle et al, 2017). The log10 of this ratio is less than zero if the protein is predominantly mitochondrial and more than zero if the protein has significant fraction localized elsewhere (Vögtle et al, 2017). Most of our newly visualized proteins are in the dual-localized protein region with the log10 of the ratio more than zero (Fig.

4A). This confirms the utility of the BiG Mito-Split collection to uncover eclipsed proteins and to aid in deciphering their targeting pathways and functions.

It is important to study such eclipsed proteins because they can use diverse mechanisms to ensure the dual localization. Understanding those can help to shed light on the factors maintaining protein import fidelity. We aimed to get an idea of the mechanisms the newly visualized dually localized proteins can use. For this, we checked in the published dataset (Monteuuis *et al*, 2019) if any of those can have an alternative translation start site that gives rise to a targeting sequence. The 20 proteins we found and Gpp1 were split into three categories: no alternative start sites but consistently high targeting signal prediction; no alternative start with low prediction; and alternative start giving rise to an echoform with a strong signal prediction (Fig 4B).

We wondered if Arc1 that has a low MTS prediction score and no alternative start to produce an echoform with a targeting signal indeed contains a targeting signal within its annotated amino-acid sequence. We expressed Arc1-GFP_{11} and a well-studied mitochondrial protein Cha1-GFP_{11} from a plasmid with a heterologous promotor and their canonical start codons in a haploid BiG Mito-Split strain (Fig. 4C). Both proteins revealed a GFP signal when observed by confocal fluorescence microscopy (Fig. 4D). This means that Arc1 harbors some poorly predicted mitochondrial targeting signal for which its exact location and recognition mechanism is still to be uncovered.

Then we turned to confirm that Gpp1 indeed contains an alternative start giving rise to an MTS. Gpp1 fell in the third category and had one of the highest signals in our dataset being consistently present in the matrix in all studied conditions. In the cytosol, Gpp1 dephosphorylates glycerol-3-phosphate and produces glycerol. This reaction is important under anaerobic conditions when mass glycerol production from glyceraldehyde-phosphate is used as a sink of reducing equivalents (Påhlman et al, 2001). Gpp1 has a paralog Gpp2 that arose from the whole genome duplication. Gpp2 plays an overlapping role and is expressed at lower levels unless upregulated by osmotic stress. Both proteins were suggested to have alternative start sites (Monteuuis *et al*, 2019) that give rise to mitochondrial echoforms (Fig. 4E). To prove this, we tagged Gpp1 and Gpp2 with 3×GFP_{11} in a haploid BiG Mito-Split strain. Interestingly, we find that Gpp1-3×GFP_{11} has a strong mitochondrial signal, but Gpp2-3×GFP_{11} does not (Fig. 4F). Then we eliminated the alternative start codons upstream of the canonical ATG by introducing either a *TEF2* promoter (*TEF2pr*) alone, *TEF2pr* followed by 3×HA tag or a *TEF2pr* followed by MTS_{Su9}-3×HA (Fig. 4F). For both proteins, introduction of *TEF2pr* or *TEF2pr-3×HA* abolished mitochondrial localization while *TEF2pr-* MTS_{Su9}-3×HA led to strong mitochondrial fluorescence (Fig. 4F). To prove that the upstream sequence in *GPP1* promoter indeed encodes a mitochondrial targeting signal, we selected two alternative start sites out of all the

predicted non-canonical ones (Monteuuis *et al*, 2019). The ATC codon at position -39 is the most upstream possible start site and the ATG codon is in the middle of the upstream region (Fig. 4G, left). When we inserted the strong *TEF2pr* one codon before each of the start sites, we observed increased mitochondrial fluorescence of the C-terminally tagged Gpp1-3×GFP_{11} (Fig. 4G, right). Interestingly, both Gpp1, Gpp2 have potential alternative start codons giving rise to an N-terminal extension with high targeting signal prediction, but only Gpp1 has a mitochondrial echoform that we could visualize. This might be due to differential activity of alternative start codons, different strength of targeting signals, or lower expression levels or more efficient degradation of Gpp2.

Visualization of protein sorting at the IM

Studying membrane protein biogenesis requires an accurate way to determine topology *in vivo*. The mitochondrial IM is one of the most protein-rich membranes in the cell supporting a wide variety of TMD topologies. We aimed to find out if our BiG Mito-Split collection can accurately visualize the localization of membrane protein C-termini protruding into the matrix.

The most reliable information on protein topology comes from structural studies. To benchmark our data we selected the structure of the supercomplex of Respiratory Complexes III and IV (Berndtsson *et al*, 2020). In total, the complexes contain six nuclear-encoded subunits whose C-termini face the matrix and all of them are visualized in the BiG Mito-Split collection (Fig. 5A, shaded green with C-terminus residue number highlighted in black). Interestingly, out of 12 nuclear-encoded subunits with IMS-facing C-termini we additionally visualize five in disagreement with their expected topology (Fig. 5A, shaded green with C-terminus residue number highlighted in red). One of them is Rip1 that is well known to be completely imported into the matrix, folded, and only then re-inserted into the IM with the help of the dedicated insertase, Bcs1 (Wagener *et al*, 2011). This suggested that our assay may not visualize only steady-state protein topology but also the biogenesis route of Rip1 as it is exposed to the matrix during the import and assembly process. This explanation is not clearly transferable to the other four proteins Cox26, Qcr8, Qcr9, and Qcr10 that are short and have a single TMD. These proteins were suggested to use a stop-transfer mechanism and therefore are not expected to be imported into the matrix completely (Park *et al*, 2013). It remains to be determined if matrix translocation can be a part of their normal biogenesis process followed by the insertion via the conservative pathway. An alternative is that we observe a failure of the stop-transfer mechanism followed by a dead-end or degradation in the matrix. More generally, to date, what other proteins use Bcs1 to insert has remained a mystery (Wagener & Neupert, 2012) and it is

highly unlikely that this conserved machine evolved only to translocate Rip1. Therefore, our assay may highlight potential new substrates.

We next expanded our analysis and selected all the membrane proteins in our dataset and analyzed their known properties from either the literature (Table S2) or predictions (Weill *et al*, 2019). We compared our observations with the expected steady-state topology (Fig. 5B. Fig. S3A). There were 23 well-studied inner membrane proteins (mono- to hexa-TMDs) in our dataset, 16 of which were expected to have their C-termini facing the matrix and detected in agreement with their steady-state topology (Fig. 5B, green labels). The other seven were expected to have IMS-facing C-terminus (Fig. 5B, red labels). Out of those there were six mono-TMD proteins and one tri-TMD protein Sdh4. The latter was reported to employ a mix of conservative and stop-transfer mechanisms for its biogenesis: the first two TMDs are imported and inserted via Oxa1 while the last TMD is sorted via stop-transfer (Park *et al*, 2013). Two proteins that we uncovered, Spg1 and Nat2, are poorly studied and are possibly IM components (Fig. 5C, grey labels). Based on the proteomics data (Morgenstern *et al*, 2017) and TMD predictions we suggest that Spg1 is a mono-TMD protein with its C-terminus facing the IMS and is detected in our experiment like the other short proteins with similar topology. Nat2 most likely is a bi-TMD protein, and we detect its steady-state topology (Fig. 5C, Fig. S3B).

Interestingly, comparison of our dataset with the expected sub-mitochondrial protein distribution revealed three proteins that were not expected to enter the matrix at all (Fig. 5D). Mcr1 has a TMD and was described to be alternatively sorted to the outer or the inner membrane (Hahne *et al*, 1994). Osm1 is imported into the IMS and follows the bipartite presequence route (Neal *et al*, 2017). Both proteins require a lateral transfer from the TIM23 complex as a biogenesis step suggesting that we observe occasional mis-sorting of these proteins to the matrix. The third protein is Mix17, CX_9C-motif containing soluble protein annotated to reside in the IMS (Longen *et al*, 2009). Its matrix localization can have physiological significance because it is a homolog of human CHCHD10 that was recently described along with several other similar proteins to contain both a CX_9C-motif and a matrix targeting signal (Peker *et al*, 2023). Such proteins are first imported into the IMS, oxidized by Mia40 and then translocated into the matrix where their oxidative state can regulate macromolecular complex assembly. The only other yeast protein known to follow a similar pathway is Mrp10 (Longen *et al*, 2014). We conclude that the BiG Mito-Split collection is useful to follow protein biogenesis and sorting at the mitochondrial IM as well as unconventional import pathways where a protein can be found in different sub-mitochondrial locations. We suggest that compared to the possible import of cytosolic proteins into the matrix that is prevented by a strict selectivity filter, the selectivity filter at the IM is relatively relaxed.

Discussion

In this work we developed a whole-proteome live-cell assay to specifically visualize proteins imported into the mitochondrial matrix or IM. We imaged ~50 proteins not observed as mitochondrial by microscopy before. Some of them were poorly studied eclipsed proteins that were localized in the cytosol or nucleus by previous studies (Table S1). Finding these proteins using the BiG Mito-Split assay was expected. However, we managed to image some proteins that were never imaged before using traditional full-length FP fusions. We argue that our collection of 3×GFP_{11}-tagged proteins can offer an advantage over the previous ones by using a smaller unfolded tag that may interfere less with protein targeting. It might be particularly useful for mitochondrial proteins for which it is long known that a tightly folded domain can inhibit their translocation (Wienhues *et al*, 1991). Thus, full-length FPs that are optimized for folding and maturation speed can be particularly problematic tags for mitochondrial proteins (Chudakov *et al*, 2010).

How sensitive is the BiG Mito-Split assay? Since the annotation of sub-mitochondrial protein localization and topology is incomplete, it is hard to make a good estimate. Here we find 70 % of proteins that were reported to be soluble in the matrix by fractionation and proteomics (Fig. S2A)(Vögtle *et al*, 2017). While low abundance proteins are probably more likely to be missed, we still find a lot of them (Fig. 3B) meaning that there is no strict sensitivity threshold for the BiG Mito-Split assay. We suggest that the major factor contributing to the assay sensitivity is the interaction probability of the two split-GFP fragments. The GFP_{1-10}/GFP_{11} fragments do not have high affinity to each other and are commonly used to study protein-protein interactions (Romei & Boxer, 2019). Indeed, in our dataset some matrix proteins related to mitochondrial translation have higher fluorescence signal than expected from their abundance (Fig. S2B) probably owing to higher probability of interacting with the newly synthesized GFP_{1-10}. Using split-reporters with higher affinity between the fragments can make the BiG Mito-Split more sensitive. The other possible reasons for not observing particular proteins with our assay are inherent to the incompleteness of yeast strain collections and interference of tags with protein expression, stability or targeting. In this version of the assay the correct tagging and expression of each protein needs to be confirmed by PCR and western blots. Developing collections where protein expression can be independently controlled with fluorescence in another channel can help to get better estimates of sensitivity and to interpret the absence of signal.

Based on the comparison of our dataset with the reported protein abundances we concluded that the detection of low abundance protein fractions is possible. We wondered if at any of the tested conditions some cytosolic proteins that we found could be imported by chance or mistake. We suggest that cytosolic r-proteins are among those. They only were detected in a

few conditions at low levels. Most of them were found on logarithmic growth phase on glucose and almost none in stationary phase (Table S1) despite many other proteins that have higher signals (Fig. 2A). We suggest that this is due to high synthesis rate of new ribosomal proteins during quick growth on glucose and due to lower rate of their synthesis and upregulation of mitochondrial quality control in stationary phase. There are hundreds of proteins that similarly to r-proteins contain regions that are predicted as mitochondrial targeting signals but only few are imported (Fig. 3C). Moreover, proteotoxic stress was suggested to trigger unspecific uptake of aggregation-prone cytosolic proteins for their subsequent degradation (Ruan *et al*, 2017). We imaged the BiG Mito-Split collection in several conditions including DTT-induced ER stress but did not detect any large group of proteins that becomes conditionally imported (Fig. 2A, Table S1). We conclude that under the conditions we investigated mitochondria maintain a strict selectivity filter at the OM. However, the specific function for r-proteins in the mitochondria cannot be completely ruled out and requires more detailed investigation. We suggest that the whole-genomic assay that we developed will be useful to study the mechanisms of the mitochondrial import specificity and to clarify what are the conditions when the selectivity is compromised.

Using the BiG Mito-Split assay we were also able to visualize IM proteins with their C-termini exposed to the matrix. Interestingly, we revealed several proteins that were suggested to have a different topology. This includes several single-TMD proteins and one 3-TMD protein (Sdh4) that were reported to use stop-transfer mechanism and to be arrested in the TIM23 complex and laterally released into the membrane. We also found one protein that uses this mechanism for import into the IMS (Osm1). One possibility is that we observe productive import intermediates that can be later reinserted into the IM by the Oxa1 translocase. Only few nuclear-encoded substrates were suggested to rely on the Oxa1 translocase for reinsertion into the IM via the conservative pathway (Bohnert *et al*, 2010; Park *et al*, 2013; Stiller *et al*, 2016). Now it is becoming clear that the substrate range of Oxa1 might be much broader (Stiller *et al*, 2016). The mechanisms that Oxa1 uses for protein insertion are similar to the ER Membrane Complex (EMC), its distant homolog (Kizmaz & Herrmann, 2023). *In vivo* fluorescence assays were among the most important tools that helped to gain considerable insight into how EMC selects its substrates and ensures correct TMD orientation within the membrane (Wu *et al*, 2023; Pleiner *et al*, 2023; Fenech *et al*, 2023). We suggest that BiG Mito-Split can be an excellent tool to address these mechanisms for Oxa1. The observation of Osm1 translocation into the matrix raises another interesting possibility that TIM23 complex is not very selective in discriminating between the substrates that need to be arrested and laterally sorted, and those that need to pass through. In this case, our assay shows considerable flexibility of the protein sorting at the IM as opposed to the strict selectivity filter at the OM. The BiG Mito-Split

collection will be helpful to study whether mis-sorting of IM proteins to the matrix is deleterious to the cells and which quality control mechanisms are responsible for managing it.

To conclude, using a new generation of yeast collections (Meurer *et al*, 2018; Yofe *et al*, 2016) we brought a fluorescence complementation assay (Bader *et al*, 2020) to a whole-genomic level. In this work we applied it to systematic visualization of mitochondrial matrix proteome by employing a matrix localized GFP_{1-10}. We used this tool to uncover new mitochondrial localized proteins and create a dataset and a resource for the mitochondrial protein field to study protein import and topogenesis *in vivo*. More broadly, our new SWAT GFP_{11} library can be utilized for many other purposes. Beyond the BiG Mito-Split assay, by targeting the GFP_{1-10} to other cellular destinations, the collection can be used to visualize targeting processes in additional organelles. In addition, it can be utilized to perform systematic protein-protein interaction studies. As such our library provides a powerful tool for systematic probing of cellular architecture.

Materials and Methods

Yeast strains and plasmids

The yeast strains used in this study are listed in Table S3, Plasmids are listed in Table S4, and primers used for yeast strain construction are listed in Table S5.

To make the BiG Mito-Split-GFP assay (Bader *et al*, 2020) compatible with the library genetic background a new strain RKY250 encoding GFP_{1-10} in the mtDNA was constructed. For this, the RKY176 (ρ^+ *atp6::*GFP_{1-10} *5`*$UTR_{COX2}$ *ATP6 3`*UTR_{COX2}) was crossed to BY4741 [ρ^o]. The diploid cells were selected on synthetic complete medium supplemented with leucine and uracil and sporulated in liquid 1% potassium acetate for one week. Tetrads were dissected and analyzed for their genotype. The presence of the mtDNA ρ^+ *atp6::*GFP_{1-10} *5`*$UTR_{COX2}$ *ATP6 3`*UTR_{COX2} was verified in each spore by checking the GFP fluorescence signal with *ATP4-*GFP_{11} expressed from the plasmid pAG416pGPD-ATP4-GFP_{11}. This plasmid was obtained by transferring *ATP4-*GFP_{11} from the original pAG414 *pGPD-ATP4-*GFP_{11} (Bader *et al*, 2020) to the pAG416 *pGPD-ccdB* with the Gateway® assembly cloning method (Alberti *et al*, 2007).

The SWAT donor strain used to make the new *C'-3×*GFP_{11} SWAT library was constructed on the basis of the published *GAL1pr-SceI* strain (Yofe *et al*, 2016). To visualize mitochondria, we inserted MTS_{Su9}-mCherry from the plasmid TEFpr-MTS-mCherry-MET into the *HO* locus. The plasmid was made by replacing mTagBFP2 in plasmid #44899 from AddGene (Lee *et al*, 2013) with TEF2pr-Su9MTS-mCherry and the KanMX with MetR from pSD-N21 (Weill *et al*, 2018). A

NAT resistance cassette was inserted into the *URA3* locus to aid diploid cell selection on rich respiratory media (see below).

The plasmid bearing a 3×GFP_{11} cassette for the SWAT library construction (3×GFP_{11} C-SWAT type I, no additional terminator and selection marker after the tag) was created by amplifying 3×GFP_{11} from the GPDpr-PGK1-3×GFP_{11} plasmid (Bader *et al*, 2020) with added BamHI and BcuI(SpeI) restriction sites, and performing digest and ligation into the C-SWAT type I donor plasmid (Meurer *et al*, 2018). The plasmid 3×GFP_{11}-ADHter-KanMX used for genomic tagging with 3×GFP_{11} was created by inserting the same amplified 3×GFP_{11} fragment into the C-SWAT type III plasmid (contains *ADH1* terminator and KanMX gene following the inserted tag) using the same restriction enzymes.

The strains to investigate the targeting of Gpp1 and Gpp2 were created by genomically tagging *GPP1* and *GPP2* with 3×GFP_{11}-*ADH*ter-KanMX using this type III plasmid. Since this plasmid contains CEN/ARS, the standard PCR-mediated transformation protocol was modified with an additional DpnI digest (37˚C, 1.5h, directly in the PCR buffer) of the amplified cassette to destroy the original plasmid and only allow transformation by genomic integration, and not by plasmid propagation. Transformants were verified by PCR and microscopy. To modify the promotor region of *GPP1-3×GFP_{11}* and *GPP2-3×GFP_{11}* strains a regular promotor swap was performed using pYM N-19 *NAT-TEF2pr, pYM N-20 NAT-TEF2pr-3HA*, or *N-SWAT NAT-MTS(Su9)-3HA* plasmid. Both the change of promotor and the elimination of the gene with native promotor was confirmed by PCR screening to avoid gene duplication events.

For visualization of Arc1 and Cha1, *3×GFP_{11}* epitope-tagging, transformation, growth of the transformed RKY176 strain and verification of both the tagged protein and GFP_{1-10} expression were performed as described in (Hemmerle *et al*, 2022). The strains where Ilv6, Fum1, and Arc1 are tagged with mNeonGreen were picked from the C-SWAT library prepared exactly like described (Meurer *et al*, 2018).

All strain modifications were performed using the standard PCR-mediated methods of genomic editing and LiAc-based yeast transformation (Longtine *et al*, 1998; Janke *et al*, 2004).

Creation of C-SWAT 3×GFP_{11} haploid collection and diploid BiG-Split collection

To create the new C-SWAT collection where every gene is tagged with 3×GFP_{11} we transformed the MTS_{Su9}-mCherry donor strain with the 3×GFP_{11} plasmid selecting the transformants on G418-containing media (1% yeast extract, 2% peptone, 3% glucose, 500 µg/l geneticin, or G418 (Formedium)). Library creation was performed using this strain essentially as described before (Meurer *et al*, 2018; Weill *et al*, 2018) using automated mating and

selection procedures (Tong & Boone, 2007; Cohen & Schuldiner, 2011) and robotic arrayer Rotor HDA (Singer Instruments). In brief, the donor strain was mated with the C-SWAT acceptor library (Meurer *et al*, 2018). The resulting diploids were selected, and sporulated. After sporulation haploids of mating type alpha were selected that were carrying the 3×GFP_{11} SWAT plasmid. The tag swapping was induced by plating the libraries on galactose-containing medium. The strains where the acceptor SWAT cassette was successfully swapped with 3×GFP_{11} were negatively selected on 5-FOA-containing media (Formedium) to kill the cells still containing the *URA3* marker in the acceptor cassette. Recombination efficiency was verified by PCR in a random set of strains and was found to be more than 90%. The resulting haploid library had every gene genomically tagged with 3×GFP_{11} followed by the native terminator without any additional selection markers (seamless tagging).

To obtain the diploid library that had both genomic 3×GFP_{11} tags and mtDNA-encoded GFP_{1-10}, the haploid library was first depleted of its own mtDNA by plating three times on 3% agar plates synthetic dextrose media (0.67% yeast nitrogen base with ammonium sulfate, 2% glucose, complete amino acid mix optimized for BY4741 background (Hanscho *et al*, 2012)) supplemented 25 µg/ml ethidium bromide. To check that the depletion was successful, the library was plated on rich respiratory media (YPGly) to confirm there is no growth in these conditions (1% yeast extract, 2% peptone, 3% glycerol). After mtDNA depletion, the haploids were immediately mated with the RKY250 BiG-Split strain. The diploids were selected and maintained on YPGly supplied with 200 µg/l nourseothricin (NAT, Jena Bioscience). Several diploids isogenic to the library and containing mtDNA-encoded GFP_{1-10} but no GFP_{11} were added to each plate as controls. The resulting diploid library was used for fluorescent microscopy screening.

Whole-genomic (primary) screening

The whole diploid library was imaged in six different conditions. All the sample preparation was performed with a robotic liquid handler Freedom EVO (Tecan) equipped with an incubator and a centrifuge. Before each experiment, the cells were transferred from agar plates to liquid synthetic dextrose (SD) medium without methionine supplied with NAT and grown overnight at 30˚C. The overnight culture was then subjected to different conditions. For imaging of logarithmically growing cells, the overnight culture was diluted 1:20 in fresh SD media supplied with complete amino acid mix and grown for 4 h at 30˚C. The resulting culture was applied onto glass-bottom microscopy plates, cells were let to settle for 25 minutes and then washed one time: the media with unattached cells was aspirated, and fresh media was added. For imaging in stationary phase, the overnight culture was directly applied onto microscopy plates (Matrical

Biosciences) and washed three times with synthetic ethanol media (0.67% yeast nitrogen base, 2% ethanol). For imaging in low phosphate conditions, the overnight culture was centrifuged at 3000 g for 3 minutes, supernatant was aspirated and the cells resuspended in distilled water, centrifugation and resuspension was repeated two more times. The resulting cell suspension was diluted 1:100 in synthetic dextrose medium without phosphate (SD-P_i, 0.67% yeast nitrogen base without inorganic phosphate, 2% glucose) and incubated for 16 h at 30˚C. Then the cells were transferred to microscopy plates as described above and washed one time with SD-P_i. For imaging in DTT-containing media the cells were prepared the same way as for imaging in SD media, except that the media contained 1 mM DTT, and the growth before imaging was limited to 3 h. For imaging in nitrogen starvation media (SD-N, 0.67% yeast nitrogen base without ammonium sulfate, 2% glucose) the overnight culture was first diluted 1:20 in SD media and grown for 4 h at 30˚C. Then the cells were pelleted by centrifugation, the media removed and substituted with SD-N. Centrifugation and media exchange was repeated one more time and after this the cells were incubated for 16 h at 30˚C, applied onto microscopy plates, washed with SD-N and imaged. All the imaging was performed using Olympus SpinSR system equipped with Hamamatsu flash Orca 4.0 camera and a CSUW1-T2SSR SD Yokogawa spinning disk unit with a 50 µm pinhole disk, and a 60× air lens with NA 0.9. Images were recorded with 488 nm laser illumination for GFP channel, 561 nm laser illumination for mCherry. The microscope was operated by ScanR Acquisition software (version 3.2). Micrographs were visualized using Fiji (Schindelin *et al*, 2012). The image of each strain was visually compared with a micrograph of the control strain from the same plate. The strains with increased signal in at least one condition, or the strains where the signal increase was ambiguous were selected for a secondary screen to measure fluorescence quantitatively (total 543 strains).

Fluorescence quantitation (secondary screen)

The 543 strains selected in the primary screen were arrayed on two 384-format plates. To these strains we added 123 non-fluorescent control strains positioned evenly across each plate and 102 additional strains that were not selected in the primary screens but were reported to reside in the matrix or the IM, based on the two most recent and reliable proteomic studies (Morgenstern *et al*, 2017; Vögtle *et al*, 2017). The strains were prepared for imaging exactly as described in the previous section. The imaging was performed with the same automated SpinSR system but using a 100× air lens with NA 1.3 equipped with an oil pump. The data was analyzed using CellProfiler v. 4.2.1+ (Stirling *et al*, 2021). The image analysis pipeline was constructed to perform background subtraction in both GFP and mCherry channels, enhance small objects in mCherry channel, segment mitochondrial objects using the mCherry channel

and measure the median GFP signal intensity within each mitochondrial object. The final dataset contained fluorescence signal measured for each mitochondrion in each strain and condition. This data was further analyzed in MATLAB (Mathworks). Median intensity of all the segmented mitochondria was used as a raw signal measurement for each strain. Since the variability of the signal was high even among the control strains without GFP_{11}, we designed a score to represent how much median mitochondrial GFP signal in GFP_{11}-containing strains deviates from non-fluorescent controls. For this we subtracted the average control fluorescence from all the values and divided them by standard deviation of all controls from this average. Less that 1% of control strain measurements were above 3 standard deviations (Fig. S1A). We designated every GFP_{11}-containing strain that has a signal exceeding control average by 5 standard deviations as a confident hit, and by more than 3 and less than 5 standard deviations as an ambiguous hit (Table S1). Most ambiguous hits were still visually well distinguishable from controls (Fig. S1B). Most of the strains were confident hits in all the conditions (Fig. S1C). The strains that were hits only in few conditions were most likely missed in the others due to low signal intensity and not due to strong differential regulation of protein expression or import (Fig. 2B). Some conditions had less hits most probably due to high media autofluorescence (Fig. S1D). Several strains were imaged one more time in all conditions to confirm the observation (Table S1). Thus, we compiled a unified list of proteins found in this study (Table S1, marked as 'OBSERVED').

Fluorescence microscopy

All microscopy of individual strains with genomically integrated tags was performed as described above for fluorescence quantitation, except each strain was grown individually in appropriate media. To visualize the potential mitochondrial import of Cha1 and Arc1, we used the Zeiss LSM 800 with Airyscan of the imaging platform of the Research Center in Biomedicine of Strasbourg (PIC-STRA, CRBS). A 63× apochromatic plane objective (1.40 digital aperture) with oil immersion, LED illumination (Colibri, Zeiss) and GFP and DsRED filter cubes were used for direct observation using eyepieces. In confocal mode, 488 nm and 640 nm lasers replaced the LEDs and high sensitivity photomultipliers (GaAsP, Phospho Arsenide de Gallium) for detection of the emitted fluorescence. The images are acquired using the Zeiss Zen Blue software, with a laser residence time per pixel (pixel dwell time) of 11.8 µs. For visualization, the image contrast was linearly adjusted in Fiji and the images were converted from 16-bit to 8-bit (Schindelin *et al*, 2012).

Protein extraction and western blot

Protein extraction, SDS-PAGE and western blotting were performed as described before with the same antibodies against GFP_{11} and GFP_{1-10} (Bader *et al*, 2020). Adh1 was probed with rabbit polyclonal primary antibody which was a gift from Claudio De Virgilio's lab (Université de Fribourg, Switzerland) used with a dilution of 1:50,000 (Calbiochem ref 126745). The secondary antibody was Goat-anti-rabbit-HRP conjugate used with a dilution of 1:5000.

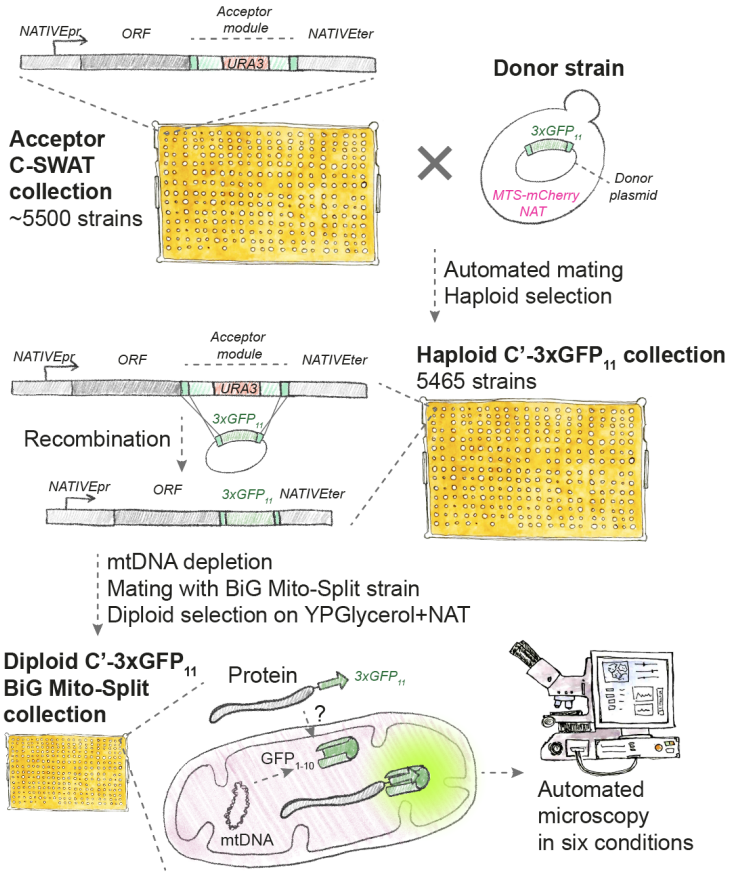
Additional data analysis

Targeting signal predictions for each protein and its alternatively translated form were taken from (Monteuuis *et al*, 2019), we always chose the highest score from the three different programs that were used in this work. Internal targeting sequences were predicted using the i-MLP web server (imlp.bio.uni-kl.de) (Boos *et al*, 2018). Unified data on protein abundance for Fig. S2B was taken from (Ho *et al*, 2018) or (Morgenstern *et al*, 2017). All plots were produced using MATLAB (Mathworks).

Acknowledgements

We would like to thank Tamara Flohr for critical reading of the manuscript and Dr. S. Friant (CNRS, UMR 7156) for her help in confocal microscopy imaging. We are grateful to Claudio De Virgilio's (Université de Fribourg, Switzerland) for kindly sharing antibodies. Work in the Schuldiner lab is supported by the German Research Foundation (DFG) collaborative grant # 1028/11-1 and an ERC CoG OnTarget (864068). Y.B. was supported by the EMBO Long-term postdoctoral fellowship (ALTF 480-2019). M.S. is an incumbent of the Dr. Gilbert Omenn and Martha Darling Professorial Chair in Molecular Genetics. The robotic system of the Schuldiner lab was purchased through the kind support of the Blythe Brenden-Mann Foundation. Work in the Kucharczyk lab was supported by the NSC grant nr 2018/31/B/NZ3/01117. The work in the Becker lab was supported by the Integrative Molecular and Cellular Biology (IMCBio), as part of the Interdisciplinary Thematic Institutes (ITI) 2021-to-2028 program of the University of Strasbourg, CNRS, and INSERM, supported by IdEx Unistra (ANR-10-IDEX-0002) and EUR IMCBio (ANR-17-EURE-0023) under the framework of the French Investments for the Future Program (to S.Z, H.D.B, B.S); by the University of Strasbourg (to S.Z., H.D.B, B.S), by the CNRS (to S.Z, H.D.B, B.S).

1 **Figures**



2

3 **Figure 1. Schematic of the BiG Mito-Split collection creation and microscopy screening.**
4 To create a whole-proteome fusion library, the acceptor C-SWAT library where each open
5 reading frame (ORF) under its native promoter (NATIVEpr) is tagged with a cassette (acceptor
6 module) that contains *URA3* marker and homology linkers (dark green) was crossed with a
7 donor strain that carries a plasmid with the desired tag (*3×GFP$_{11}$*) and genomically integrated
8 mitochondrial marker MTS$_{Su9}$-mCherry and a NAT selection marker. Following sporulation and
9 haploid selection, recombination is initiated, and the acceptor module is swapped with the
10 donor tag, following negative selection for the loss of *URA3* marker. To assay mitochondrial
11 localization, the resulting haploid C'- collection is deprived of its mtDNA on ethidium bromide
12 and crossed with the BiG Mito-Split strain carrying *GFP$_{1-10}$* in its mtDNA, diploids selected on
13 respiratory media supplemented with NAT. The diploid collection is imaged using an automated
14 fluorescence microscope and only proteins with their C-termini localized in the matrix
15 complement split-GFP and can be detected.

16

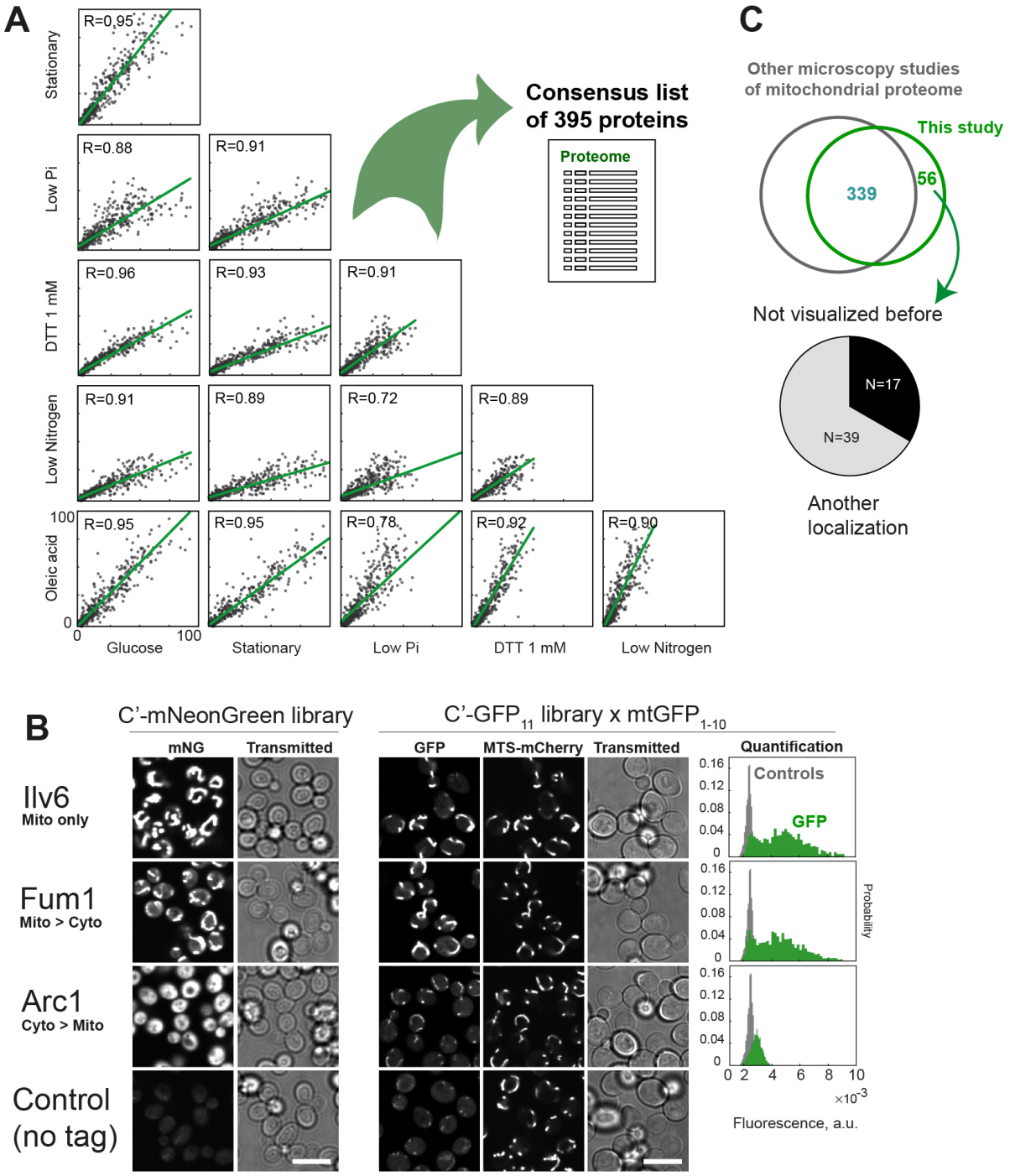


Figure 2. Visualization of mitochondrial proteins with the BiG Mito-Split collection. (A) Scatter plots of correlation between normalized fluorescence intensities of each strain imaged in different conditions, the line and value of linear regression are shown. (B) Examples of fluorescent micrographs of strictly mitochondrial and dually localized proteins visualized by full-length fluorescent protein tags (C-SWAT mNeonGreen (mNG) library, left) or using the BiG Mito-Split collection (right). Quantification of the GFP signal in mitochondria marked by MTS_{Su9}-mCherry in BiG Mito-Split strains relative to pooled controls without GFP_{11} is shown in the far right. (C) Comparison of the proteins visualized in this study with the previous microscopy studies, and the breakdown of newly visualized proteins into the ones previously found in another location, or never studied with high-throughput microscopy before. Scale bars 10 µm.

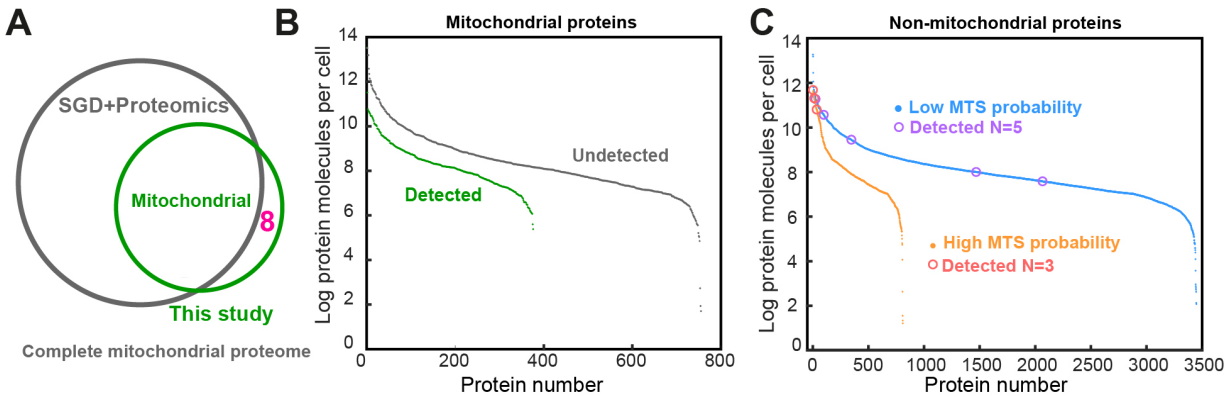


Figure 3. Comparison of the proteins visualized in the BiG Mito-Split collection with previous studies of the mitochondrial proteome. (A) A Venn diagram showing a comparison of proteins defined from this work with manual annotations in the SGD and high-throughput proteomics reveals eight proteins previously not associated with mitochondria. (B) Mitochondrial proteins visualized in our screen (green) compared to all other mitochondrial proteins (grey, list compiled as in panel A) ranked by unified protein abundance (Ho et. al 2018). (C) Non-mitochondrial proteins ranked by unified protein abundance and split into a group with a high mitochondrial targeting signal prediction score (>0.7) or low score (according to (Monteuuis *et al*, 2019)) with the eight non-mitochondrial proteins visualized in this work (panel A) marked with circles. Only proteins with known unified abundance were analyzed (Ho et. al 2018).

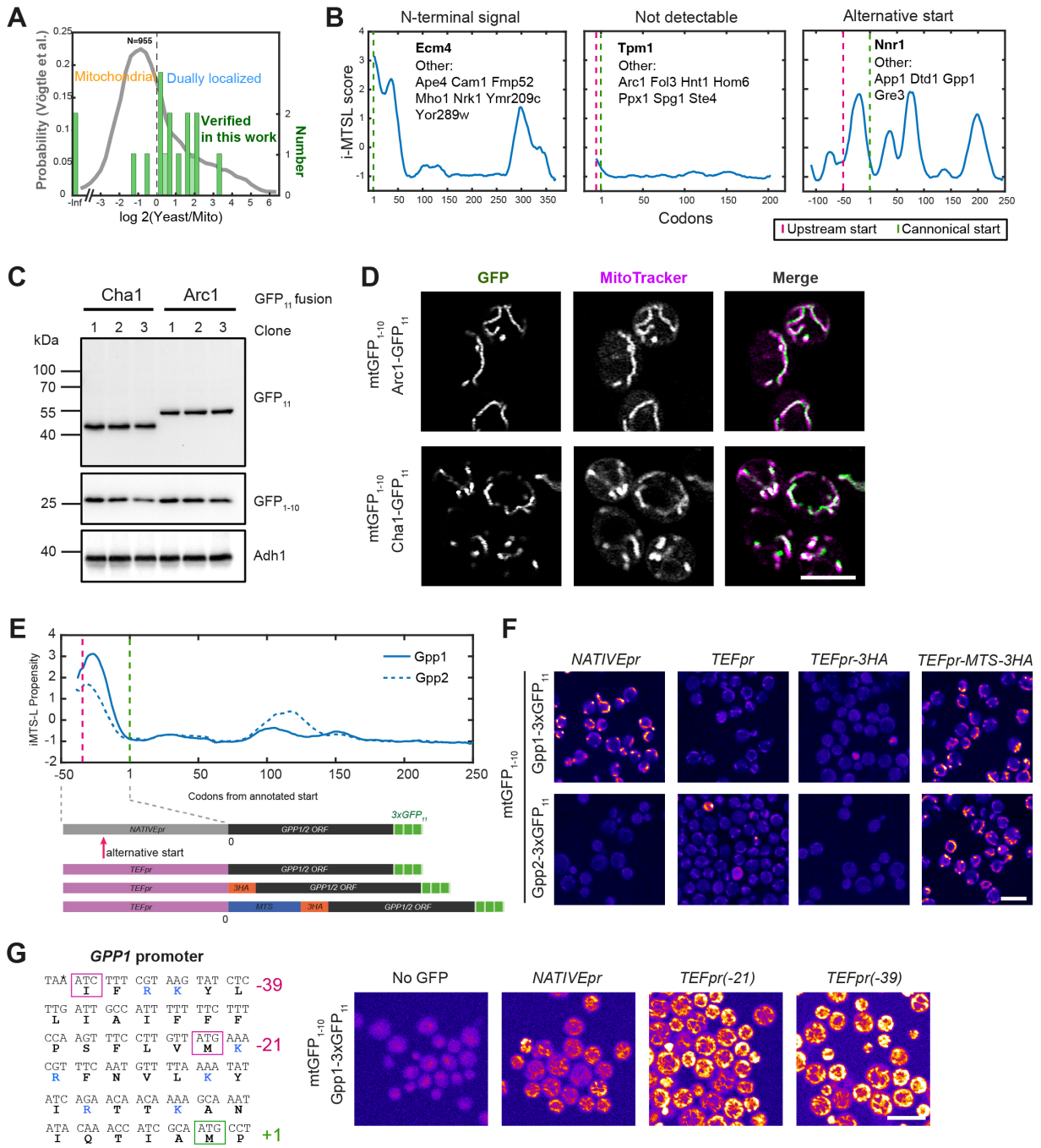


Figure 4. Dually localized proteins and their targeting signals. (A) Proteins visualized in this work and only found by high-throughput proteomics before tend to have high whole cell to mitochondria ratio (Vögtle *et al*, 2017) indicating that they are dually localized; only the proteins for which the ratio is known are plotted. (B) Three different types of potential dually localized proteins can be found: those predicted to have an N-terminal targeting signal and no alternative start codon, those with no targeting signal or alternative start prediction, those where an alternative start generates an echoform with high-scoring prediction; for each type one example is shown with its graph of the i-MTSL-score, canonical start codon (green dashed line), alternative start codon (magenta dashed line), and other similar proteins listed on the plot. (C) Western blot verifying the expression of GFP_{11}-tagged Cha1 and Arc1 cloned into an expression plasmid with their canonical start codons under a heterologous promoter and transformed into a haploid BiG Mito-Split strain; three different clones from each transformation are shown; primary antibody used for decoration is shown on the right. (D) Confocal

fluorescence microscopy of haploid BiG Mito-Split of the strains analyzed in (C), one clone is shown for each. (E) Gpp1 and Gpp2 i-MTSL start codons prediction (top), and schematics of generated constructs with mutated promotors and N-termini (bottom). (F) Fluorescence microscopy of the strains with mtDNA-encoded GFP_{1-10} where Gpp1 and Gpp2 were tagged with 3×GFP_{11} at the C-terminus (*NATIVEpr*) and then native promoter was substituted to *TEF2pr*, or *TEF2pr* followed by 3xHA tag, or MTS_{Su9}-3×HA. (G) Left: sequence of the *GPP1* promoter region with upstream on the canonical start codon (ATG, +1) showing the most upstream non-canonical start (ATC, -39) and an additional ATG (-21); right: fluorescence micrographs of the cells without GFP and of a haploid BiG Mito-Split strain where genomically-tagged *GPP1-3×GFP_{11}* has a native promoter or a *TEF2pr* integrated before the two non-canonical starts shown to the left. Scale bars are 10 μm (F,G) and 5 μm (D).

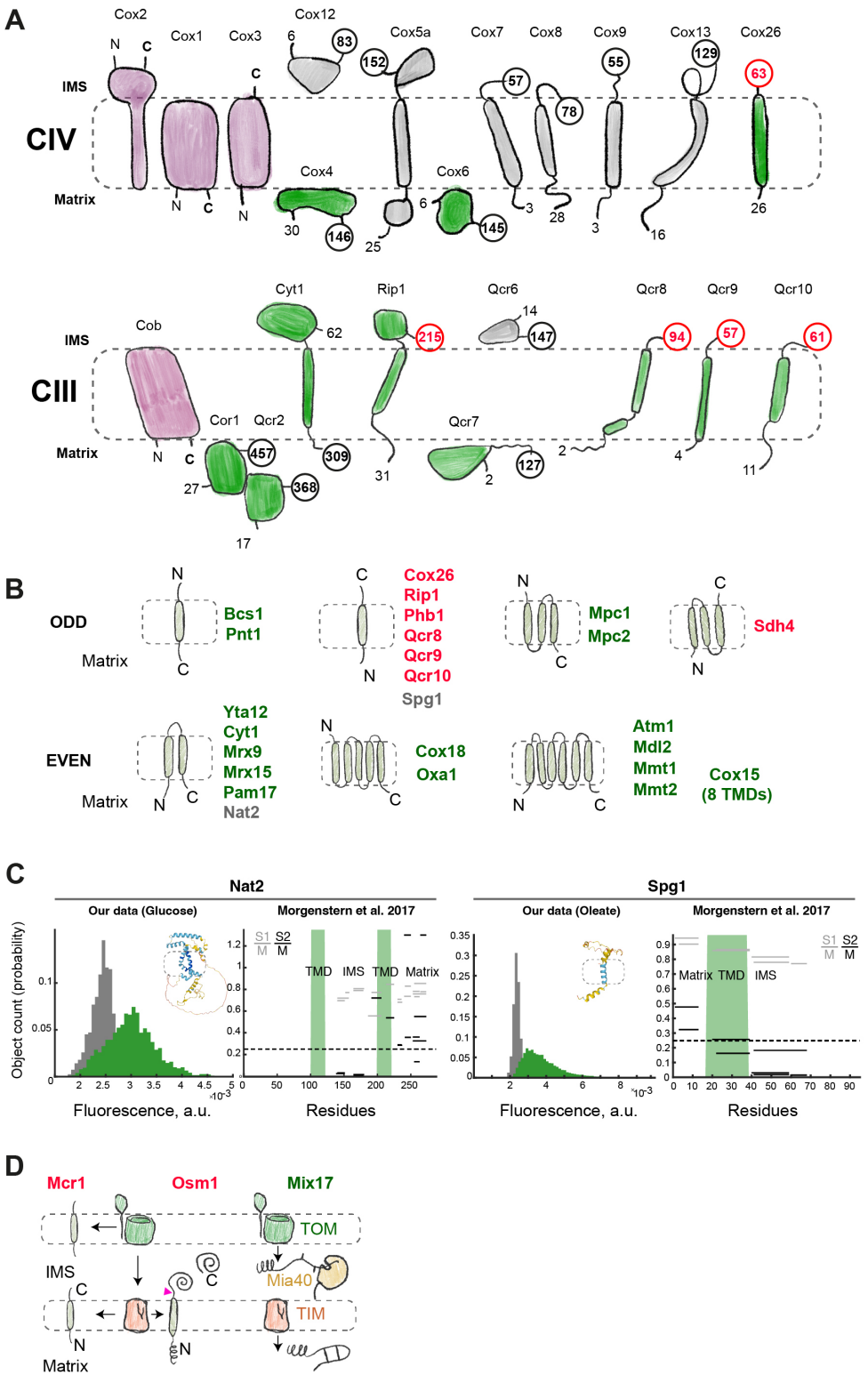


Figure 5. Protein topology visualized by BiG Mito-Split. (A) Schematic drawing of the topology and approximate sizes of Complex IV and Complex III subunits as seen in a supercomplex structure (PDB:6ymx), mtDNA-encoded subunits are shown in purple, for all other subunits the first and the last amino acid number and position occurring in the structure are shown, C-terminal label is highlighted in bold and enclosed in a circle, proteins that we observe in the dataset are colored green, C-terminal label is colored red if the observation does not agree with the known topology. (B) All membrane proteins found in our dataset sorted by the number of TMDs and topology. Those where topology does not agree with our data are

marked in red. Those with agreeing topology are marked in green. The two poorly studied proteins are marked in grey. (C) Mitochondrial GFP fluorescence intensity for Nat2-3×GFP_{11} and Spg1-3×GFP_{11} shown besides the proteomic data where peptide position and fractionation ratios S1/M and S2/M are shown in grey and black (high S1/M and S2/M mean the peptide is in the matrix; high S1/M and low S2/M mean that the peptide is in the IMS), TMD prediction is shown in green, AlphaFold2-predicted structures with TMD regions highlighted by dashed lines are shown. (D) Other proteins found in the matrix: Mcr1 is a transmembrane protein alternatively sorted into the OM and the IM, Osm1 is a soluble IMS protein sorted via a stop-transfer mechanism, Mix17 is a possible substrate of Mia40 that is subsequently imported into the matrix.

Supplementary materials

Table S1. The complete dataset of normalized fluorescence in each condition and comparisons with other data. Full legend inside file.

Table S2. Comparison of known inner membrane protein topology with the observations made using BiG Mito-Split collection. Full legend inside file.

Table S3. Yeast strains used in this work.

Table S4. Plasmids used in this work.

Table S5. Primers used for yeast strain construction.

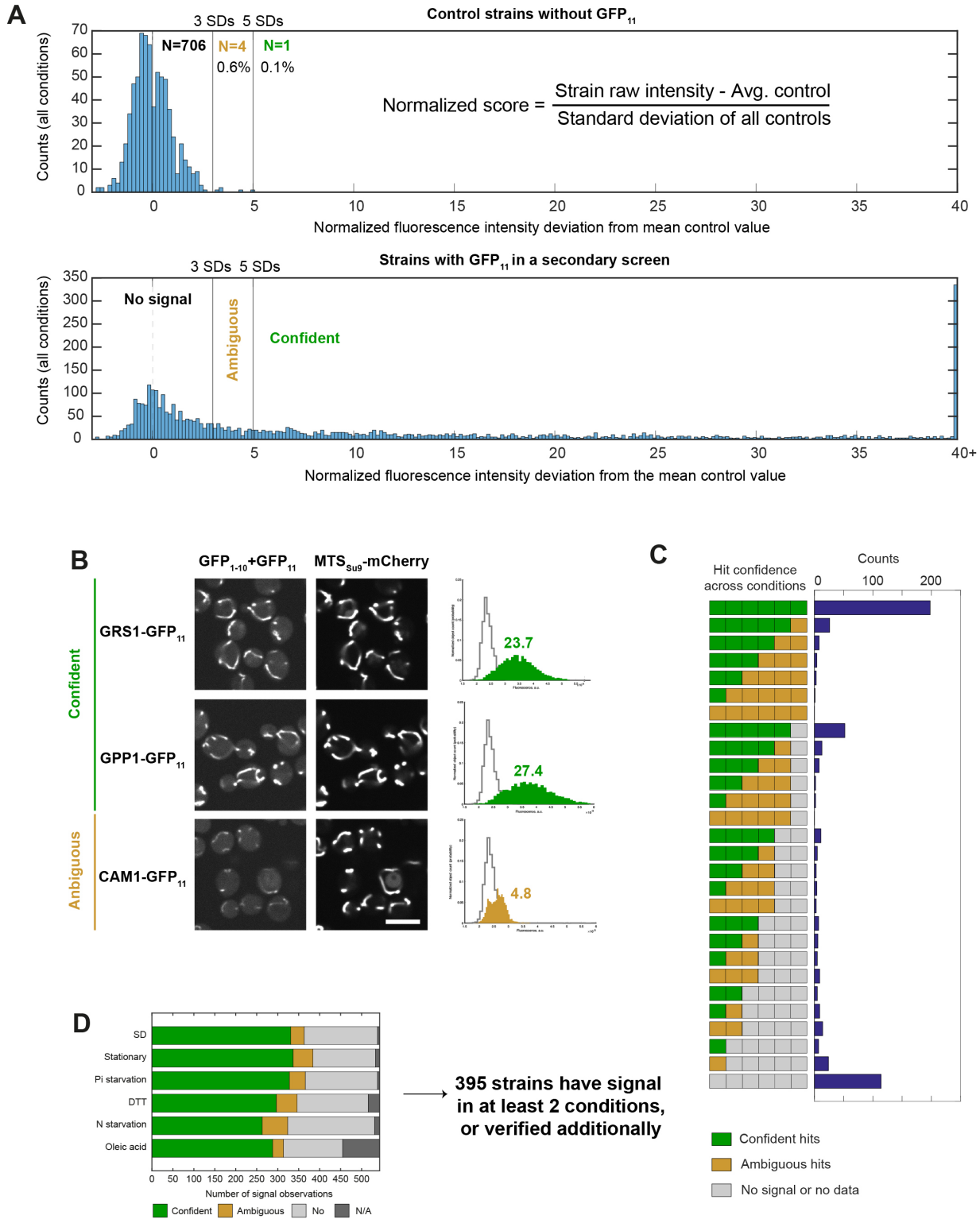


Figure S1. Quantitative analysis and normalization of GFP signal in different growth conditions. (A) Median fluorescence intensity in mitochondrial regions was measured for each control strain, then all values were normalized by subtracting the average of all measurements (N=706) and divided by standard deviation (SD), yielding a normal distribution (top) where less than 1% of measurements deviated more than 3SDs from an average control; the same normalization was applied to all the other imaged strains using the average and SD of control measurements and producing normalized fluorescence score; if the score was between 3 and 5 it was designated as ambiguous and above 5 as confident. (B) Examples of images of two

strains with confident and ambiguous scores along with histograms representing the distribution of raw intensities measured for each mitochondrial object in these strains (green or yellow shaded region) compared to the measurements in pooled controls (grey line). (C) Distribution of strains by the number of confident and ambiguous scores in six conditions: the majority of strains had a confident score in all conditions, second largest group had none, relatively few strains had different combinations of confident and ambiguous score numbers; to be added in the list of observed proteins for comparison with the other datasets, the strain had to demonstrate at least one confident and one ambiguous score or to be imaged one more time to confirm an ambiguous score. (D) The number of observed proteins differed between conditions possibly reflecting different fluorescence background and mitochondrial activity and to a lesser extent, specific regulation of individual protein expression and import. Scale bar 10 μm.

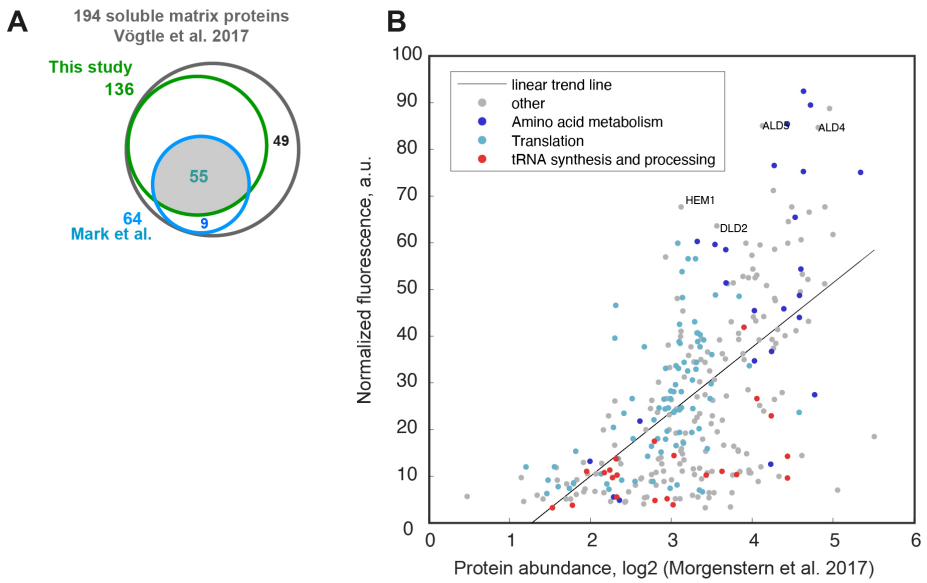
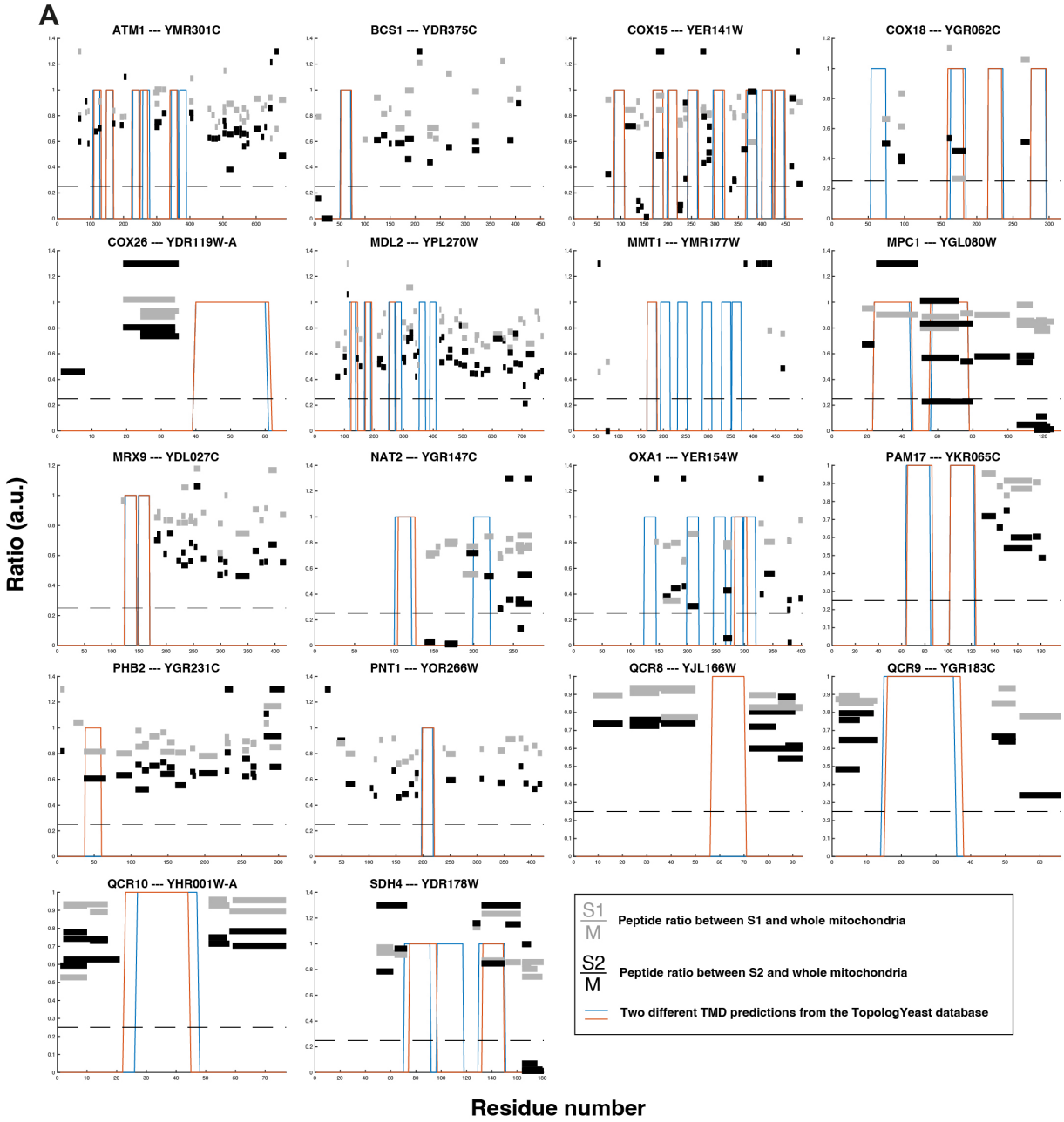


Figure S2. Proteins observed in this study compared to other works. (A) The list of soluble matrix proteins was extracted from Vögtle et al. 2017 and numbers of proteins from that list observed in our study and in Mark et al. 2023 was depicted. (B) For each protein observed in our study its normalized fluorescence score in glucose is plotted against log2 protein abundance measured by proteomics in purified mitochondria (cells grown on glucose), certain functional groups that might have disproportionally higher or lower fluorescence signal are highlighted.



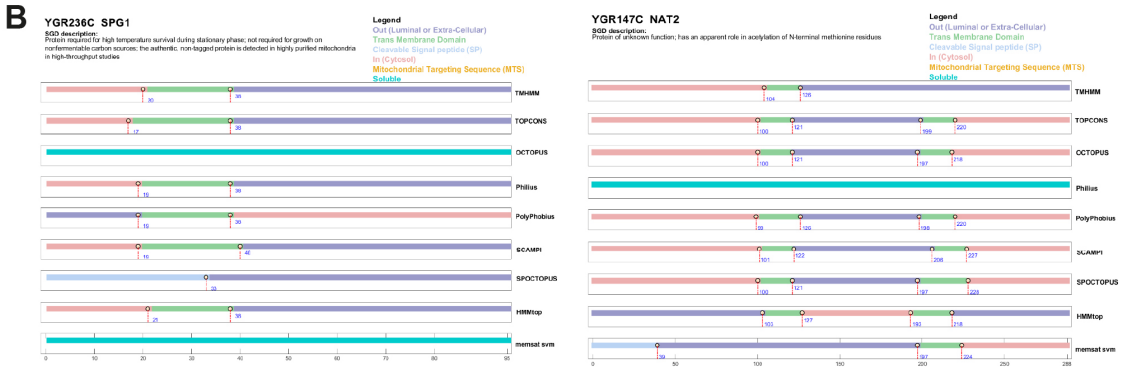


Figure S3. Comparison of observed inner membrane protein topology with other works and predictions. (A) The available proteomic data from (Morgenstern et al. 2017) for the proteins depicted in Fig. 5B showing S1/M and S2/M rations for each detected peptide (high

S1/M + low S2/M indicates IMS location, both ratios high indicate matrix location), and two TMD predictions from TopologYeast web server (which compiles TMD predictions by seven different algorithms), the X-axis corresponds to amino acid number. (B) An output for Nat2 and Spg1 proteins from TopologYeast web server.

References

Alberti S, Gitler AD & Lindquist S (2007) A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast* 24: 913–919

Bader G, Enkler L, Araiso Y, Hemmerle M, Binko K, Baranowska E, De Craene J-O, Ruer-Laventie J, Pieters J, Tribouillard-Tanvier D, *et al* (2020) Assigning mitochondrial localization of dual localized proteins using a yeast Bi-Genomic Mitochondrial-Split-GFP. *eLife* 9: e56649

Berndtsson J, Kohler A, Rathore S, Marin-Buera L, Dawitz H, Diessl J, Kohler V, Barrientos A, Büttner S, Fontanesi F, *et al* (2020) Respiratory supercomplexes enhance electron transport by decreasing cytochrome c diffusion distance. *EMBO reports* 21: e51015

Bohnert M, Rehling P, Guiard B, Herrmann JM, Pfanner N & van der Laan M (2010) Cooperation of Stop-Transfer and Conservative Sorting Mechanisms in Mitochondrial Protein Transport. *Current Biology* 20: 1227–1232

Boos F, Mühlhaus T & Herrmann JM (2018) Detection of Internal Matrix Targeting Signal-like Sequences (iMTS-Ls) in Mitochondrial Precursor Proteins Using the TargetP Prediction Tool. *Bio-protocol* 8: e2474–e2474

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

Bykov YS, Flohr T, Boos F, Zung N, Herrmann JM & Schuldiner M (2022) Widespread use of unconventional targeting signals in mitochondrial ribosome proteins. *EMBO J* 41: e109519

Bykov YS, Rapaport D, Herrmann JM & Schuldiner M (2020) Cytosolic Events in the Biogenesis of Mitochondrial Proteins. *Trends in Biochemical Sciences* 45: 650–667

Chudakov DM, Matz MV, Lukyanov S & Lukyanov KA (2010) Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. *Physiological Reviews* 90: 1103–1163

Cohen Y & Schuldiner M (2011) Advanced methods for high-throughput microscopy screening of genetically modified yeast libraries. *Methods Mol Biol* 781: 127–159

Di Bartolomeo F, Malina C, Campbell K, Mormino M, Fuchs J, Vorontsov E, Gustafsson CM & Nielsen J (2020) Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. *Proceedings of the National Academy of Sciences*: 201918216

Dubreuil B, Sass E, Nadav Y, Heidenreich M, Georgeson JM, Weill U, Duan Y, Meurer M, Schuldiner M, Knop M, *et al* (2019) YeastRGB: comparing the abundance and localization of yeast proteins across cells and libraries. *Nucleic Acids Res* 47: D1245–D1249

Fenech EJ, Cohen N, Kupervaser M, Gazi Z & Schuldiner M (2023) A toolbox for systematic discovery of stable and transient protein interactors in baker's yeast. *Molecular Systems Biology* 19: e11084

Hahne K, Haucke V, Ramage L & Schatz G (1994) Incomplete arrest in the outer membrane sorts NADH-cytochrome b5 reductase to two different submitochondrial compartments. *Cell* 79: 829–839

Hanscho M, Ruckerbauer DE, Chauhan N, Hofbauer HF, Krahulec S, Nidetzky B, Kohlwein SD, Zanghellini J & Natter K (2012) Nutritional requirements of the BY series of Saccharomyces cerevisiae strains for optimum growth. *FEMS Yeast Research* 12: 796–808

Hemmerle M, Senger B, di Rago J-P, Kucharczyk R & Becker HD (2022) Visualizing Mitochondrial Importability of a Protein Using the Yeast Bi-Genomic Mitochondrial-Split-GFP Strain and an

Ordinary Fluorescence Microscope. In *Mitochondria: Methods and Protocols*, Tomar N (ed) pp 255–267. New York, NY: Springer US

Herrmann JM & Bykov Y (2023) Protein translocation in mitochondria: Sorting out the Toms, Tims, Pams, Sams and Mia. *FEBS Letters* 597: 1553–1554

Ho B, Baryshnikova A & Brown GW (2018) Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Systems* 6: 192-205.e3

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

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

Kizmaz B & Herrmann JM (2023) Membrane insertases at a glance. *Journal of Cell Science* 136: jcs261219

Lee S, Lim WA & Thorn KS (2013) Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PLOS ONE* 8: e67902

Longen S, Bien M, Bihlmaier K, Kloeppel C, Kauff F, Hammermeister M, Westermann B, Herrmann JM & Riemer J (2009) Systematic Analysis of the Twin Cx9C Protein Family. *Journal of Molecular Biology* 393: 356–368

Longen S, Woellhaf MW, Petrungaro C, Riemer J & Herrmann JM (2014) The Disulfide Relay of the Intermembrane Space Oxidizes the Ribosomal Subunit Mrp10 on Its Transit into the Mitochondrial Matrix. *Developmental Cell* 28: 30–42

Longtine MS, Iii AM, Demarini DJ, Shah NG, Wach A, Brachat A, Philippsen P & Pringle JR (1998) Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast* 14: 953–961

Mark M, Klein O, Zhang Y, Das K, Elbaz A, Hazan RN, Lichtenstein M, Lehming N, Schuldiner M & Pines O (2023) Systematic Approaches to Study Eclipsed Targeting of Proteins Uncover a New Family of Mitochondrial Proteins. *Cells* 12: 1550

Meurer M, Duan Y, Sass E, Kats I, Herbst K, Buchmuller BC, Dederer V, Huber F, Kirrmaier D, Štefl M, *et al* (2018) Genome-wide C-SWAT library for high-throughput yeast genome tagging. *Nature Methods* 15: 598–600

Michaelis AC, Brunner A-D, Zwiebel M, Meier F, Strauss MT, Bludau I & Mann M (2023) The social and structural architecture of the yeast protein interactome. *Nature* 624: 192–200

Monteuuis G, Miścicka A, Świrski M, Zenad L, Niemitalo O, Wrobel L, Alam J, Chacinska A, Kastaniotis AJ & Kufel J (2019) Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Research* 47: 5777–5791

Morgenstern M, Stiller SB, Lübbert P, Peikert CD, Dannenmaier S, Drepper F, Weill U, Höß P, Feuerstein R, Gebert M, *et al* (2017) Definition of a High-Confidence Mitochondrial Proteome at Quantitative Scale. *Cell Reports* 19: 2836–2852

Neal SE, Dabir DV, Wijaya J, Boon C & Koehler CM (2017) Osm1 facilitates the transfer of electrons from Erv1 to fumarate in the redox-regulated import pathway in the mitochondrial intermembrane space. *MBoC* 28: 2773–2785

Påhlman A-K, Granath K, Ansell R, Hohmann S & Adler L (2001) The Yeast Glycerol 3-Phosphatases Gpp1p and Gpp2p Are Required for Glycerol Biosynthesis and Differentially Involved in the

Cellular Responses to Osmotic, Anaerobic, and Oxidative Stress*. *Journal of Biological Chemistry* 276: 3555–3563

Park K, Botelho SC, Hong J, Österberg M & Kim H (2013) Dissecting Stop Transfer versus Conservative Sorting Pathways for Mitochondrial Inner Membrane Proteins in Vivo. *J Biol Chem* 288: 1521–1532

Peker E, Weiss K, Song J, Zarges C, Gerlich S, Boehm V, Trifunovic A, Langer T, Gehring NH, Becker T, *et al* (2023) A two-step mitochondrial import pathway couples the disulfide relay with matrix complex I biogenesis. *Journal of Cell Biology* 222: e202210019

Pleiner T, Hazu M, Pinton Tomaleri G, Nguyen VN, Januszyk K & Voorhees RM (2023) A selectivity filter in the ER membrane protein complex limits protein misinsertion at the ER. *Journal of Cell Biology* 222: e202212007

Regev-Rudzki N & Pines O (2007) Eclipsed distribution: A phenomenon of dual targeting of protein and its significance. *BioEssays* 29: 772–782

Romei MG & Boxer SG (2019) Split Green Fluorescent Proteins: Scope, Limitations, and Outlook. *Annual review of biophysics* 48: 19

Ruan L, Zhou C, Jin E, Kucharavy A, Zhang Y, Wen Z, Florens L & Li R (2017) Cytosolic Proteostasis via Importing of Misfolded Proteins into Mitochondria. *Nature* 543: 443–446

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, *et al* (2012) Fiji: an open-source platform for biological-image analysis. *Nature Methods* 9: 676–682

Stiller SB, Höpker J, Oeljeklaus S, Schütze C, Schrempp SG, Vent-Schmidt J, Horvath SE, Frazier AE, Gebert N, van der Laan M, *et al* (2016) Mitochondrial OXA Translocase Plays a Major Role in Biogenesis of Inner-Membrane Proteins. *Cell Metabolism* 23: 901–908

Stirling DR, Swain-Bowden MJ, Lucas AM, Carpenter AE, Cimini BA & Goodman A (2021) CellProfiler 4: improvements in speed, utility and usability. *BMC Bioinformatics* 22: 433

Tong AHY & Boone C (2007) 16 High-Throughput Strain Construction and Systematic Synthetic Lethal Screening in Saccharomycescerevisiae. In *Methods in Microbiology*, Stansfield I & Stark MJ (eds) pp 369–707. Academic Press

Vögtle F-N, Burkhart JM, Gonczarowska-Jorge H, Kücükköse C, Taskin AA, Kopczynski D, Ahrends R, Mossmann D, Sickmann A, Zahedi RP, *et al* (2017) Landscape of submitochondrial protein distribution. *Nature Communications* 8: 290

Wagener N, Ackermann M, Funes S & Neupert W (2011) A Pathway of Protein Translocation in Mitochondria Mediated by the AAA-ATPase Bcs1. *Molecular Cell* 44: 191–202

Wagener N & Neupert W (2012) Bcs1, a AAA protein of the mitochondria with a role in the biogenesis of the respiratory chain. *J Struct Biol* 179: 121–125

Weill U, Cohen N, Fadel A, Ben-Dor S & Schuldiner M (2019) Protein topology prediction algorithms systematically investigated in the yeast Saccharomyces cerevisiae. *Bioessays* 41: e1800252

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

Wienhues U, Becker K, Schleyer M, Guiard B, Tropschug M, Horwich AL, Pfanner N & Neupert W (1991) Protein folding causes an arrest of preprotein translocation into mitochondria in vivo. *Journal of Cell Biology* 115: 1601–1609

Woellhaf MW, Hansen KG, Garth C & Herrmann JM (2014) Import of ribosomal proteins into yeast mitochondria. *Biochemistry and Cell Biology* 92: 489–498

Wong ED, Miyasato SR, Aleksander S, Karra K, Nash RS, Skrzypek MS, Weng S, Engel SR & Cherry JM (2023) Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources. *Genetics* 224: iyac191

Wu H, Smalinskaitė L & Hegde RS (2023) EMC rectifies the topology of multipass membrane proteins. *Nat Struct Mol Biol*: 1–10

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