Assessment of GFP Tag Position on Protein Localization and Growth Fitness in Yeast

Uri Weill[2,†], Gat Krieger[2,†], Zohar Avihou[2], Ron Milo[1], Maya Schuldiner[2] and Dan Davidi[1]

*1 - **Department of Plant and Environmental Sciences,** Weizmann Institute of Science, Rehovot 7610001, Israel*
*2 - **Department of Molecular Genetics,** Weizmann Institute of Science, Rehovot 7610001, Israel*

Correspondence to Maya Schuldiner and Dan Davidi: *maya.schuldiner@weizmann.ac.il, dan.davidi@weizmann.ac.il*
https://doi.org/10.1016/j.jmb.2018.12.004
Edited by M Yaniv

Abstract

While protein tags are ubiquitously utilized in molecular biology, they harbor the potential to interfere with functional traits of their fusion counterparts. Systematic evaluation of the effect of protein tags on function would promote accurate use of tags in experimental setups. Here we examine the effect of green fluorescent protein tagging at either the N or C terminus of budding yeast proteins on subcellular localization and functionality. We use a competition-based approach to decipher the relative fitness of two strains tagged on the same protein but on opposite termini and from that infer the correct, physiological localization for each protein and the optimal position for tagging. Our study provides a first of a kind systematic assessment of the effect of tags on the functionality of proteins and provides a step toward broad investigation of protein fusion libraries.

Report

Protein tags are essential for a variety of assays in biology, from affinity tags for protein purification to fluorescent tags for visualization. However, tagging proteins comes at a price: fusion proteins are different from their native form and may suffer from impaired activity, reduced stability, loss of binding partners, wrong targeting, unnatural topology, and so on [1–4]. Often, the same tag may induce different phenotypes depending on where it appears on the protein, for example, the carboxy terminus (C′) or amino terminus (N′) of the polypeptide chain. For example, a recent comparison between two whole-genome libraries in the budding yeast *Saccharomyces cerevisiae* showed that about 10% of the proteins in yeast are localized to different subcellular localizations when tagged with green fluorescent protein (GFP) at either the N′ or C′ ([5]; Fig. 1a). Furthermore, it appears that alternate tagging may introduce systematic biases, as C′ tagging seems to better support mitochondrial and nuclear localization, while N′ tagging seems to be better for supporting bud-neck targeting as well as enabling retention of proteins in the endoplasmic reticulum (ER) and later vesicular compartments rather than have them leak to the vacuole (Fig. 1a). Importantly, these differences were not a result of large expression differences as both libraries were constructed without altering the native promoters, and therefore, the expression levels of tagged proteins are similar to that in wild-type [5].

Here we set out to ask: when tagging on opposing termini gives rise to different steady-state localizations, which tagged variant represents the physiologically relevant localization? For this, we established a pairwise competition approach that relies on the assumption that there would be a growth advantage to the strain carrying the correctly localized protein form (Fig. 1b). We note that while protein function can be impaired without displaying a mis-localization, it is clear that a difference in localization affects the capacity of a protein to function properly in a cellular context. Further, while it may theoretically be that mis-localization gives rise to a growth advantage, here we assume that this is not the norm. To test our approach, 35 proteins in yeast were selected that are

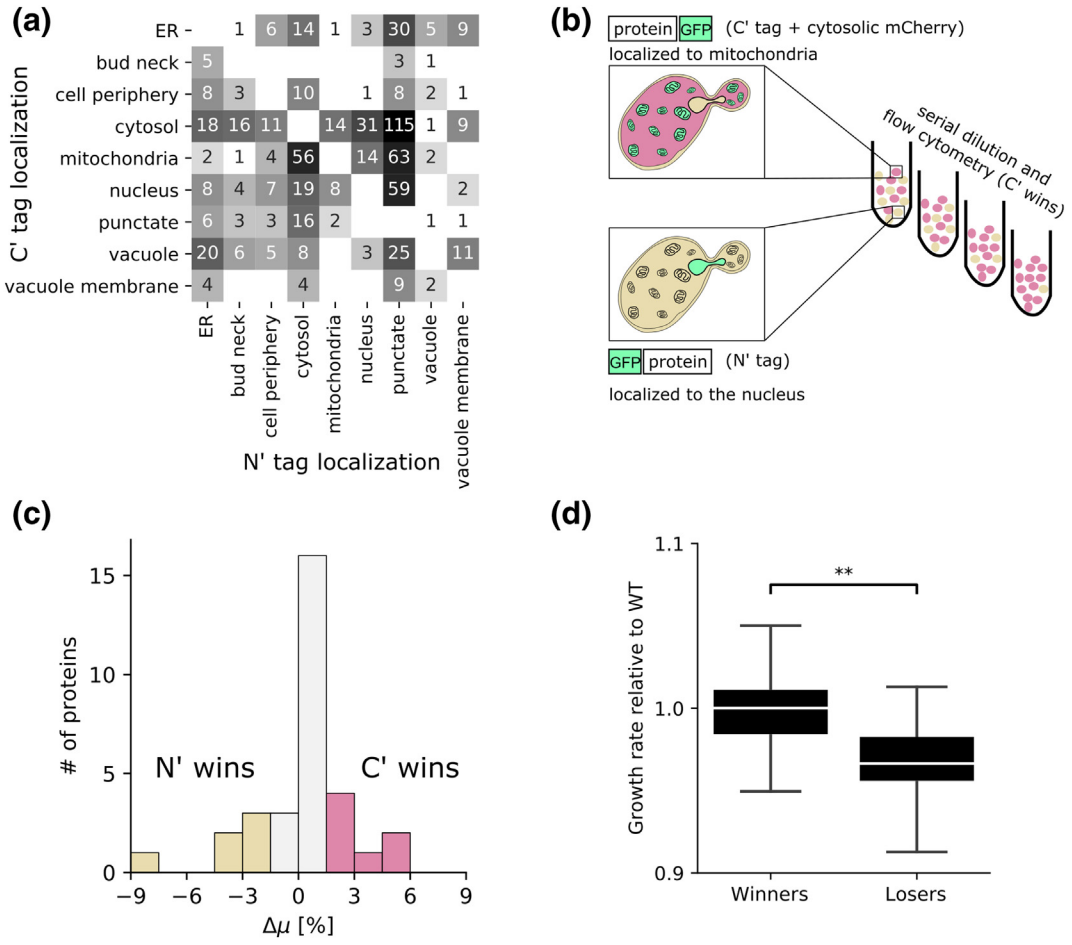

Fig. 1. Uncovering the physiological subcellular localization of proteins in yeast. (a) Comparison of localization assignments between the C′ tagged (*y*-axis [6,7]) and N′ tagged GFP genome-scale library (*x*-axis [5]). Altogether 515 proteins are differentially localized, representing about 10% all yeast proteins. Grayscale goes from white (least) to black (most) strains with altered localization (data from Ref. [5]). (b) Schematic representation of the pairwise competition approach. The C′-tagged library was genetically modified to express cytosolic mCherry, which allows for the quantitative measurement of population sizes of the two variants separately using flow cytometry on pooled mixed samples. Measurements were done every 24 h for 4 days (about 30 generations; see Methods). (c) Distribution of the relative fitness ($\Delta\mu$) for all essential proteins, which are differentially localized (35 pairwise assays); yellow bars correspond to strains with $\Delta\mu < -1.5$ and represent N′ winners, red bars are for $\Delta\mu > 1.5$ (C′ winners), and gray bars are for variants that show less than 1.5% fitness difference. (d) The effective growth rates of all "winning" or "losing" strains relative to wild type (0.35 h^{-1} in SC media), as determined from individual growth curve analyses (see Methods). The white line corresponds to the median value of each group (1.0 and 0.97 for winners and losers, respectively); black area is the interquartile range (IQR). As a group, winners exhibit a significantly faster growth rate relative to losers (p value = 0.005; independent t test).

both essential [8] and differentially localized when tagged on the opposing termini (Table S1). We further included 22 additional strains that are either non-essential and/or localized to similar subcellular localizations upon N′ or C′ tagging (a total of 57 strains; see Methods and Table S1 for the full list of proteins). We hypothesized that fitness differences could be easily monitored in essential proteins where even partial loss of the protein's function inherently leads to a growth deficiency. We then measured the fitness of the two tagged strains relative to each other in a competition experiment and tracked which has a growth advantage.

A total of 19 proteins (out of the 57) showed a significant fitness difference ($\Delta\mu$) between the two tagged forms ($|\Delta\mu| > 1.5\%$; Fig. 1c) —12 cases where the C′ tagged form was superior and 7 cases where the N′ form had an advantage. Notably, only 2 of the above 19 strains were from the control group (Sbf3, an essential protein that is non-differentially localized [see Fig. 2c) and further below, and Ysy6, a protein of unknown function that is both non-essential and non-differentially localized and showed a $\Delta\mu$ of just over 1.5% (see Table S1)], supporting the notion that the fitness difference between N′ and C′ tagging of proteins that are both essential and differentially

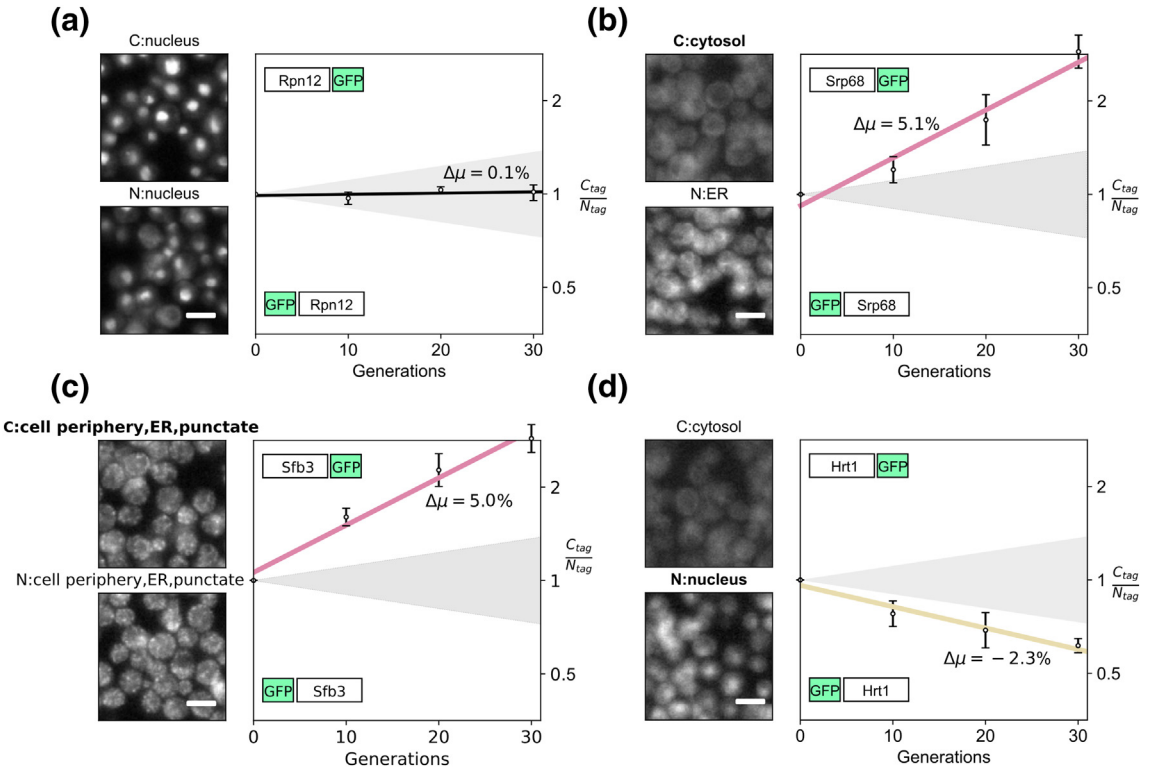

Fig. 2. Representative images showing proteins that are (a) localized to the same organelle and do not show a growth advantage with either tag, (b) localized to different places and show a growth advantage when harboring a C′ tag in comparison to an N′ tag, (c) localized to the same subcellular localizations yet show growth advantage when harboring a C′ tag, and (d) localized to different cellular locals and show a growth advantage with an N′ tag in comparison to a C′ tag. All measurements were done in triplicates; error bars represent the standard deviations; shaded triangles correspond to $|\Delta\mu| \leq 1.5$; bold font indicates the winning tag form; microscopy image scale bars are 5 μm.

localized would be larger (see Table S1 for $\Delta\mu$ values of all 57 pairs tested). To support the pairwise competition approach, we also performed growth assays to determine the effective growth rates of all our strains without competition (Methods).

Indeed, growth rate data support the competition results, since, as a group, the “winners” from the competition had faster growth rates than did the “losers” (Fig. 1d). Importantly, the winners' growth rate was similar to that of wild type (median growth rates of 0.35 and 0.34 h^{-1} for winners and losers, respectively, compared to a growth rate of 0.35 h^{-1} for wild type; growth in SC media). The fact that winners grow as well as control strains implies that the localization patterns of the winning strains indeed represent the physiological scenario. However, we show that growth rate alone could not have differentiated most pairs as well as a competition assay, as only 6 of the 19 pairs that showed significant $\Delta\mu$ values in the pairwise competition assay also showed a significant growth rate difference when measured in isolation (p value < 0.01; Table S1).

What can we learn about the cellular roles of proteins based on our analysis? Knowledge of the correct steady-state accumulation of a protein can give insights into its function. For example, Srp68 is the core component of the large ribonucleoprotein complex that constitutes the signal recognition particle (SRP), which enables co-translational targeting of proteins to the ER. The SRP complex cycles between the cytosol where it binds translating ribosomes, and the ER membrane, where it binds its receptor. The cycle of binding and release is mediated by the GTPase activity of SRP and its receptor [9]. While it is clear that the complex cycles between two cellular locals, it is not clear where the majority of the protein should reside at steady state. The fact that the cytosolic localization seen with a C′ tag supported better growth suggests that the ER localization (Fig. 2b) represents a “trapped” intermediate that cannot properly dissociate from its receptor and hence is reduced in its targeting capacity. An example of a N′ tag winner is Hrt1 (Fig. 2d) that had nuclear localization with the N′ tag and a cytosolic one with a C′ tag. Hrt1 is a RING-H2 domain core subunit of multiple ubiquitin ligase complexes, and our data suggest that its nuclear functions are the ones most affecting growth rate. Rpn12 is a proteasome subunit and is shown as a

representative of a control, where both tagged forms are localized to the nucleus as expected of the proteasome and the fitness of both strains is similar (Fig. 2a). Notably, in the set of control proteins, we observed one case, Sfb3, where the fitness of the C′ tagged variant was >5% higher than the N′ tag form, although the subcellular localization of both the N′ and C′ tags was similar. Sfb3 functions as a heterodimer of the COPII vesicle coat and has multiple phosphorylation sites at its N′ [10]. It is likely that having an N′ GFP tag interrupts this proper function of this protein on the ER.

Our work suggests a systematic approach suited for gauging the effect of a tag on global protein functionality and localization. For proof of concept, we chose to focus only on proteins that show no overlap whatsoever in the set of compartments to which they are localized when tagged on opposing termini. Proteins with multiple localizations represent a significant fraction of all proteins, and often, subcellular localizations of the C′ and N′ variants overlap but are not identical, either in the total set of subcellular localizations and/or in the percentage localization to specific compartments. It is likely that even such cases would lead to significant fitness alterations, and hence, they can also be tested using the presented methodology.

A possible confounding factor of the presented approach is if tagging a specific protein resulted in an unexpected change in the physiology of the cells that is not directly a result of the tagged protein itself. An example of this could be the disruption of an overlapping coding or regulatory region to the tagged protein. In our assay, we have only three cases in which overlapping protein coding regions exist, and only one of them, Hrt1, has a significant fitness difference of 2.3% in favor of the N′ variant (overlapping coding regions are indicated in Table S1). We are also aware that the presented approach may be more relevant for essential proteins, since for non-essential proteins, the fitness difference between the mis- and well-localized variants may be too small to detect. However, many "non-essential" proteins become essential under specific conditions (different media and/or genetic backgrounds), and hence, they could be included in tailored analyses. For example, peroxisomal biogenesis proteins become essential when cells are grown in fatty acids as a sole carbon source and mutants lacking key mitochondrial proteins become essential when yeast are forced to respire. Further, the presented approach can readily be extended to study the effect of additional tags and therefore can be used to derive multiple physiologically relevant tagged libraries. In a similar manner, one can also test the effect of a given tag on the cellular *function* of a protein by comparing the fitness of two variants that are localized to the *same* place [as in the case of Sfb3 (Fig. 2c)].

To conclude, the presented approach provides a useful tool to study the relationship between protein function and cellular fitness. Accounting for potential caveats of protein tags is essential for accurate understanding of cell biology. Such data are hence valuable for systematic, as well as for detailed, investigation of many questions in molecular biology.

Methods

A total of 57 proteins were analyzed: 35 essential and differentially localized proteins (the "study" subset) and a set of 22 control proteins that include: 9 essential and similarly localized proteins, 2 non-essential and differentially localized proteins, 9 non-essential and similarly localized proteins [5,8], and 2 different wild-type variants (Table S1). Two tagged variants were considered to be differentially localized if no overlap whatsoever was found in the set of subcellular compartments to which they are localized.

For each protein, two strains were mixed in SC media such that one strain was tagged with GFP at the C′ of the protein of interest (taken from the genome wide C′ GFP yeast collection [7]) and the second strain was tagged with GFP at the N′ (taken from the N′ genome-wide yeast collection *NATIVEpr-GFP* [5]). Importantly, both proteins were expressed under their own natural promoter. To allow for optical separation between the strains, endogenous soluble mCherry was included in the C′ library strain (TEF2pr-mCherry tag was introduced into the URA3 locus; for more details, see Ref. [7]). Cells were grown together for 24 h and diluted 32-fold, and then flow cytometry was used to monitor population sizes of the C′- and N′-tagged variants for 30 generations at 4 time points (every 24 h for 4 days).

To calculate the relative fitness difference ($\Delta\mu$), we normalized the ratio between C′ and N′ by the ratio at "day zero" to account for non-equal mixing of stains. Then, a linear regression model was fitted to the log of the ratio (*y*-axis in Fig. 2) against the number of generations (*x*-axis Fig. 2). A $\Delta\mu$ value was calculated as the slope of the fit line and was positive if the C′ strain exhibited better fitness and negative if N′ was better. We note that $\Delta\mu$ values can be treated as percent fitness advantage; for example, a $\Delta\mu$ value of −1.8 means that the N′ tagged variant is 1.8% fitter than the C′ strain. A $\Delta\mu$ value of 1.5 (in absolute values) was considered to be a significant difference as it was the maximum $\Delta\mu$ value observed across the nine pairs of non-essential proteins that are localized to similar subcellular localizations.

Flow cytometry was performed on the BD LSRII system (BD Biosciences). Fluorescent protein measurements were conducted with excitation at 488 nm

and emission at 525 ± 25 nm for GFP, excitation at 594 nm and emission at 610 ± 10 nm for mCherry. The average number of cells analyzed was 20,000. Gating of +GFP-labeled population and + GFP + mCherry-labeled population was done using a custom Matlab script; all measurements were done in triplicates. Downstream computational data processing was done using a custom Python script.

C′ and N′ GFP-tagged strain arrays were imaged using a ScanR system (Olympus) as previously described [7]. Images were acquired using a 60× air lens for GFP (excitation, 490/20 nm; emission, 535/50 nm), mCherry (excitation, 572/35 nm; emission, 632/60 nm), and brightfield channels. Images were transferred to ImageJ [1.51p Java1.8.0_144 (64-bit)] for slight, linear adjustments to contrast and brightness.

For localization assignments, we reviewed images manually. As we did not use any co-localization markers, we assigned only those localizations that could be easily discriminated by eye: ER, nuclear periphery, cytosol, cell periphery, vacuole lumen, vacuole membrane, mitochondria, nucleus, bud or bud neck, and punctate (which includes structures such as the Golgi apparatus, peroxisomes, endosomes, p-bodies, inclusions, lipid droplets, other vesicular structures and subdomain compartments).

For growth curve assays, cells were grown in 96-well plates in SC media at 30 °C until stationary phase, serial dilution of 1:75 was then applied to a volume of 150 μl. Strains were grown in a shaking incubator at 30 °C and were measured for OD600 (using infinite200 reader; Tecan Inc.) every 25 min for 2 days. Growth rates were computed by fitting a linear model to the log of the OD600 data using a custom python script.

Acknowledgments

We thank the Barkai lab for kind help with generating the competition protocol and growth curve data. We thank Naama Barkai and Einat Zalckvar for critical reading of the manuscript. This work was supported by the Azrieli Institute of Systems Biology grant to U.W. and D.D. Work in the Schuldiner lab is supported by an ERC CoG 646606 (Peroxisystem) and a VolksWagen foundation grant (93092). M.S. is an Incumbent of the Dr. Gilbert Omenn and Martha Darling Professorial Chair in Molecular Genetics.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2018.12.004.

Received 20 June 2018;
Received in revised form 5 November 2018;
Accepted 5 December 2018
Available online 12 December 2018

Keywords:
Green fluorescent protein;
Protein tag;
Genomic libraries;
Cellular fitness;
Protein localization

†These authors contributed equally to this work.

Abbreviations used:
GFP, green fluorescent protein; SRP, signal recognition particle; ER, endoplasmic reticulum.

References

[1] K. Terpe, Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems, Appl. Microbiol. Biotechnol. 60 (2003) 523–533.
[2] I. Yofe, U. Weill, M. Meurer, S. Chuartzman, E. Zalckvar, O. Goldman, S. Ben-Dor, C. Schütze, N. Wiedemann, M. Knop, A. Khmelinskii, M. Schuldiner, One library to make them all: streamlining the creation of yeast libraries via a SWAp-Tag strategy, Nat. Methods 13 (2016) 371–378.
[3] E.A. Woestenenk, M. Hammarström, S. van den Berg, T. Härd, H. Berglund, His tag effect on solubility of human proteins produced in *Escherichia coli*: a comparison between four expression vectors, J. Struct. Funct. Genom. 5 (2004) 217–229.
[4] K. Dave, H. Gelman, C.T.H. Thu, D. Guin, M. Gruebele, The effect of fluorescent protein tags on phosphoglycerate kinase stability is nonadditive, J. Phys. Chem. B 120 (2016) 2878–2885.
[5] U. Weill, I. Yofe, E. Sass, B. Stynen, D. Davidi, J. Natarajan, R. Ben-Menachem, Z. Avihou, O. Goldman, N. Harpaz, S. Chuartzman, K. Kniazev, B. Knoblach, J. Laborenz, F. Boos, J. Kowarzyk, S. Ben-Dor, E. Zalckvar, J.M. Herrmann, R.A. Rachubinski, O. Pines, D. Rapaport, S.W. Michnick, E.D. Levy, M. Schuldiner, Genome-wide SWAp-Tag yeast libraries for proteome exploration, Nat. Methods (2018), https://doi.org/10.1038/s41592-018-0044-9.
[6] W.-K. Huh, J.V. Falvo, L.C. Gerke, A.S. Carroll, R.W. Howson, J.S. Weissman, E.K. O'Shea, Global analysis of protein localization in budding yeast, Nature 425 (2003) 686–691.
[7] M. Breker, M. Gymrek, M. Schuldiner, A novel single-cell screening platform reveals proteome plasticity during yeast stress responses, J. Cell Biol. 200 (2013) 839–850.
[8] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A.P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D.J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J.H. Hegemann, S. Hempel, Z. Herman, D.F. Jaramillo, D.E. Kelly, S.L. Kelly, P. Kötter, D. Labonte, D.C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S.L. Ooi, J.L. Revuelta, C.J. Roberts, M. Rose, P. Ross-

MacDonald, B. Scherens, G. Schimmack, B. Shafer, D.D. Shoemaker, S. Sookhai-Mahadeo, R.K. Storms, J.N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-Y. Wang, T.R. Ward, J. Wilhelmy, E.A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J.D. Boeke, M. Snyder, P. Philippsen, R.W. Davis, M. Johnston, Functional profiling of the *Saccharomyces cerevisiae* genome, Nature 418 (2002) 387–391.

[9] D. Akopian, K. Shen, X. Zhang, S.-O. Shan, Signal recognition particle: an essential protein-targeting machine, Annu. Rev. Biochem. 82 (2013) 693–721.

[10] Y. Shimoni, T. Kurihara, M. Ravazzola, M. Amherdt, L. Orci, R. Schekman, Lst1p and Sec24p cooperate in sorting of the plasma membrane ATPase into COPII vesicles in *Saccharomyces cerevisiae*, J. Cell Biol. 151 (2000) 973–984.