JCB Journal of Cell Biology

TOOLS

A reference library for assigning protein subcellular localizations by image-based machine learning

Wiebke Schormann[1]*, Santosh Hariharan[1]*, and David W. Andrews[1,2,3]

Confocal micrographs of EGFP fusion proteins localized at key cell organelles in murine and human cells were acquired for use as subcellular localization landmarks. For each of the respective 789,011 and 523,319 optically validated cell images, morphology and statistical features were measured. Machine learning algorithms using these features permit automated assignment of the localization of other proteins and dyes in both cell types with very high accuracy. Automated assignment of subcellular localizations for model tail-anchored (TA) proteins with randomly mutated C-terminal targeting sequences allowed the discovery of motifs responsible for targeting to mitochondria, endoplasmic reticulum, and the late secretory pathway. Analysis of directed mutants enabled refinement of these motifs and characterization of protein distributions in within cellular subcompartments.

Introduction

Subcellular localization of proteins is a key feature of eukaryotes. Understanding subcellular localization has long been a goal for cell biologists interested in basic mechanisms of protein sorting and for understanding the generation of organelles with distinct compositions and morphologies. Moreover, there is much to be learned about disease processes, mechanisms of signal transduction, and cellular metabolism that is directly linked to subcellular localization. Traditionally, subcellular localization of a protein of interest has been assigned by visual comparison with one or more proteins of known localization (antibody based or fluorescence protein tagged) or with organelle-specific dyes (e.g., Mitotracker) in fluorescence microscope images by an experimentalist. However, human visual inspection is prone to both drift and bias. Therefore, machine learning tools have been developed to automate the analysis of subcellular localization.

Early classifiers built to distinguish subcellular structures in fluorescence micrographs in HeLa cells based on features tailored specifically for subcellular location studies functioned well with small datasets (Boland and Murphy, 2001). In addition to the newly designed region of interest (ROI) features, the well-known textural features by Haralick et al. (1973) and Zernike moments features (Zernike, 1934) were used. A new set of statistical features called threshold adjacency statistics (TAS) is faster to calculate than other commonly used statistical features (Hamilton et al., 2007), and also shows good performance (Nanni and Lumini, 2008). As a result, several supervised classification strategies to distinguish subcellular structures of the main subcellular locations (e.g., cytoplasm, nucleus, Golgi apparatus, mitochondria, and ER) have been published (Hamilton et al., 2007; Conrad et al., 2004; Li et al., 2012).

A major limitation to using automated methods of image analysis for determining subcellular localization is the paucity of high-quality images with explicit annotations. This is particularly problematic for proteins that transit between organelles or those at steady-state that are located at more than one subcellular location. For such proteins, accurate annotation based on biological experiments or images of subcellular distributions can be very difficult. One approach to dealing with this problem has been to use semi-supervised methods to assign subcellular localization from lower-quality data together with multi-label classification (Xu et al., 2016). A similar approach was used to detect mis-localization of proteins in cancer cells using images from the human protein atlas (Xu et al., 2015, 2019). However, in these cases, automated analysis was limited to detection of relatively coarse localization changes such as between the cytoplasm and nucleus or mitochondria.

More recently, deep learning approaches alone and together with crowd sourcing have been used to tackle the problem of classifying subcellular localization of proteins in yeast (Chong et al., 2015; Pärnamaa and Parts, 2017) and in human cell lines (Sullivan et al., 2018). A major advance in these studies was the

[1]Biological Sciences, Sunnybrook Research Institute, Toronto, Canada; [2]Department of Biochemistry, University of Toronto, Toronto, Canada; [3]Department of Medical Biophysics, University of Toronto, Toronto, Canada.

*W. Schormann and S. Hariharan contributed equally to this paper; Correspondence to David. W. Andrews: david.andrews@sri.utoronto.ca.

use of hundreds of thousands of images to overcome differences inherent in the data deposited in repositories such as the human protein atlas as well as the cell to cell variations inherent in normal biology. Using multiple markers in the same cell, it was possible to automatically classify a number of subcellular structures, particularly subnuclear spot types (Sullivan et al., 2018); however, automated identification of the compartments within the secretory pathway has not been achieved.

Improving automated assignment of localization requires a large dataset of high-quality images and an alternative approach to the problem of proteins having multiple subcellular localizations. Our approach was to generate a reference library of 789,011 and 523,319 optically validated landmark-based localization images in murine and human epithelia, respectively. Numerical analyses identified 160 features most useful for assignment of localization while minimizing the effects of expression levels. Rather than forcing assignment to predefined organelles, we classify images by similarity to landmarks that may themselves localize to multiple destinations. We then used this library of images to examine localization of model tail-anchor proteins. We show that automated analysis outperformed even highly trained human observers and enabled the identification multiple morphologically distinct distributions of TA proteins. Our results demonstrate the utility of the reference library of images, derived features, and machine learning to provide unbiased assignment of subcellular localizations with high accuracy in individual living cells.

Results

Distinct subcellular phenotypes identified from multidimensional descriptors of cell images of landmark proteins

Landmark proteins were generated by fusing EGFP to proteins selected from previous reports establishing localization at specific subcellular organelles (Table 1). A reference library used for automated assignment of subcellular localization in individual cells was generated by calculating multidimensional descriptors of micrographs of cells expressing these proteins. The reference library contains 789,011 and 523,319 quality-validated cell images of the EGFP-landmark fusion proteins expressed in normal murine mammary gland (NMuMG) cells and in MCF10A cells, respectively. For organelles with recognized subdomains (mitochondria, ER, and Golgi apparatus) multiple landmark proteins were used (Fig. 1 A and Table 1).

In addition to forward transit, proteins recycle between different organelles or within compartments of the same organelle (e.g., Golgi), which increases the heterogeneity of distribution. To capture the varying phenotypes caused by mobility of TA proteins, thousands of images were acquired for each landmark (Table 1). Prior to feature extraction, all images undergo automated quality control to remove out-of-focus cells and segmentation artifacts (Fig. S1 and Materials and methods).

To visualize in two dimensions the separation of these landmark proteins in 160-dimensional feature space, we used t-distributed stochastic neighbor embedding (t-SNE) projection (Van der Maaten and Hinton, 2008) to compress the data. As displaying over 500,000 multidimensional points in two dimensions is not meaningful, a subset comprising the 200 nearest-neighbor data points to the centroid of each landmark was plotted. Landmarks in the dataset are clearly separated from each other, with only a few points distributed to other areas within the t-SNE landscape (Fig. 2 A). As another way to represent the distribution of the data and demonstrate the separation of landmarks, data points for individual landmarks were over-clustered using Phenograph (Levine et al., 2015; Fig. 2 B). As a third measure of the relationships between the multidimensional classes, the Euclidean distances between landmark centroids were tabulated and plotted as a heatmap (Fig. S2).

An ensemble of Random Forest (RF) classifiers generated from a subset (less than half) of the landmark images was used to assign localization of images not used for training to generate confusion matrices (Fig. 1 B). The high accuracy of the resulting classification of individual cells (78%) demonstrated that even though many of the images for different landmarks appear visually similar, a classifier based on the image features table accurately separates and identifies fluorescent proteins targeted to different subcellular organelles for individual cells without a colocalization marker (Fig. 1). The confusion matrix indicates that even for the two most similar and visually indistinguishable landmarks, monoamine oxidase A (MAO) and cytochrome c oxidase (CCO; outer and inner mitochondrial membranes, respectively), <15% of cells were misclassified to another single landmark (Fig. 1 B). Control experiments in which we deliberately impair the classifier by training on images of the highest or lowest intensity and then classify images of different intensity levels demonstrated that the expression level (intensity) of the landmarks neither contributed to nor confounded classification (Fig. 3). Indeed, the classification accuracy drops only in the extreme cases such as when a classifier trained using images of only the lowest intensity (highest shot noise) was used to classify images from the highest-intensity quartile (Fig. 3) and the errors introduced were confined to the most closely related patterns. This effect is negligible in our other experiments as all intensity levels were used in training. Furthermore, the SD in the assignments across 20 classifiers was <1% of cells and is therefore not specified in the figures (see Materials and methods). Taken together, these results suggest that an assignment of >20% of cells to a specific landmark location is a conservative definition for significance.

At steady-state, many of the landmarks have some cells reproducibly assigned to other localizations. This type of "mis-localization" to other organelles actually reflects normal organelle and protein dynamics. Thus, assignment of a protein to a particular subcellular location is never realistically 100%.

To our surprise, different landmarks ostensibly targeted to the same organelle were efficiently discriminated by the RF classifier. For example, the images of EGFP fused to the TA sequences of the ER localized TA proteins Bcl-2 interacting killer (Bik; Germain et al., 2002) and cytochrome b5 (Cytb5) are visually similar to each other and to images of EGFP fused to Calr-KDEL, another ER landmark (Fig. 1 A). However, the cell images were well-separated from each other in multidimensional space in both cell lines (Fig. 1 B and Fig. 2 A). As a result, in NMuMG cells, only 8% of Cytb5-expressing cells were classified as Bik

Table 1. **Protein localization information used for landmarks**

Protein	Abbreviation	Number of cells	Protein sequence	UniProt	Localization	Reference
Cytochrome b5	Cytb5	Mu: 10,428	99-134	P00167	ER	D'Arrigo et al., 1993
		Mu: 4,452				
Signal sequence of calreticulin, ER retention sequence (KDEL)	Calr-KDEL	Mu: 5,367	N/A	P27797	ER	Fliegel et al., 1989; Munro and Pelham, 1987
		Hu: 24,059				
Bcl-2 interacting killer	Bik	Mu: 148,030	111-160	Q13323	ER	Germain et al., 2002
		Hu: 11,046				
Ribosome-attached membrane protein 4	RAMP4	Hu: 2,269	-1–166	Q9Y6X1	ER	Schröder et al., 1999
ER-Golgi intermediate compartment 53 kD	ERGIC53	Mu: 39,178	1–517	Q9D0F3	ERGIC	Scheel et al., 1997
		Hu: 73,059				
β1,4-galactosyl-transferase	GalT	Mu: 60,260	1–81	P15291	Trans-Golgi	Roth and Berger, 1982
		Hu: 19,797				
Golgin84	Golgin84	Mu: 147,916	674-731	Q8TBA6	Cis-Golgi	Diao et al., 2003
		Hu: 5,632				
Golgi SNAP receptor complex member 2	Membrin	Hu: 19,288	1–212	O14653	Cis-Golgi	Hong, 2005
Monoamine oxidase A	MAO	Mu: 34,521	489-527	Q49A63	OMM	de Champlain et al., 1969
		Hu: 22,935				
Cytochrome c oxidase, subunit VIII	CCO	Mu: 11,023	1–29	Q53XN1	IMM	Rizzuto et al., 1995
		Hu: 92,159				
Phosphatidylserine synthase 1	PTDSS1	Mu: 137,908	1–473	Q99LH2	MAM, ER	Stone and Vance, 2000
Ras-related protein Rab3C	Rab3	Hu: 27,545	1–227	Q96E17	Secretory vesicles	Fischer von Mollard et al., 1994
Ras-related protein Rab5A	Rab5	Mu: 18,098	1–205	A0A024R2K1	Early endosome	Chavrier et al., 1990
		Hu: 27,137				
Ras-related protein Rab7A	Rab7	Mu: 45,878	1–207	A0A158RFU6	Late endosome	Bucci et al., 2000
		Hu: 42,792				
Ras-related protein Rab11B	Rab11B	Hu: 21,961	1–218	Q15907	Recycling endosome	Schlierf et al., 2000
Vesicle-associated membrane protein 2	VAMP2	Mu: 19,411	1–116	P63027	Secretory vesicles	Grote et al., 1995; Chen and Scheller, 2001
		Hu: 12,239				
Vesicle-associated membrane protein 5	VAMP5	Mu: 43,352	1–116	O95183	Plasma membrane	Hong, 2005
		Hu: 39,946				
Neuromodulin	Neuromodulin	Hu: 17,852	1–20	P17677	Plasma membrane	Skene and Virág, 1989
Vesicle-associated membrane protein 1, TMD deleted	ΔTMD-VAMP1	Mu: 19,074	1–99	P23763	Cytoplasm, nucleus	This study
Lamin A	Lamin A	Mu: 15,197	1–690	P02545	Nuclear envelope	Scaffidi and Misteli, 2008
Peroxisome targeting signal 1	PTS-1	Mu: 18,396	SKL	N/A	Peroxisome	Gould et al., 1989
		Hu: 27,712				
Lysosomal-associated membrane protein 1	LAMP-1	Mu: 14,974	1–417	P11279	Lysosomes	Falcón-Pérez et al., 2005
		Hu: 17,162				
Emerin	Emerin	Hu: 14,277	1–254	P50402	Nuclear envelope	Pfaff et al., 2016

The common name, abbreviation, and amino acid sequence used to construct the landmarks is indicated, with the UniProt ID for each of the coding regions. The reference provided includes the data for the assignment of localization and identification of the responsible targeting sequence. The number of cells refers to the number of cell images used in the analysis here. Hu, human image library; IMM, inner mitochondrial membrane; Mu, mouse image library; OMM, outer mitochondrial membrane.

(Fig. 1 B); 5% of Bik-expressing cells were classified as Cytb5, while 9% were classified as the nearest Euclidean neighbor, phosphatidylserine synthase 1 (PTDSS1; mitochondria-associated ER membrane [MAM] and ER; Table 1); and 9% were classified as the ER–Golgi intermediate compartment (ERGIC) protein 53. Classification of Bik localization to these landmarks in a small fraction of cells is highly plausible as both compartments are related to the ER. Importantly, only 2% of cells expressing Cytb5 were classified as VAMP5, the nearest Euclidean neighbor. Thus, images that are not assigned to their own landmark are assigned according to features representing their biology, and not simply by centroid Euclidean distances. Controls (Fig. 3 and described below) demonstrated that intensity differences do not account for the different classifications. It is possible that classification differences for landmarks ostensibly located at the same organelle represent different steady-state distributions within the organelle, and different extents to which the landmarks localize to different organelles, as seen previously (Xu et al., 2016). As discussed below, different steady-state distributions within an organelle may also represent functional or physical subdomains. Because the different landmarks for the same organelle are well separated, we refer to the various subcellular locations by the landmark protein rather than the organelle name. In this way, the variability of localization for each landmark is preserved in the classification. When organelle names are used, it is to designate multiple landmarks in aggregate.

The localization of novel proteins and dyes can be identified across species using reference libraries

To validate our multiparametric definition of the subcellular landscape, we used a new set of marker proteins previously reported to target to specific organelles and classified them using

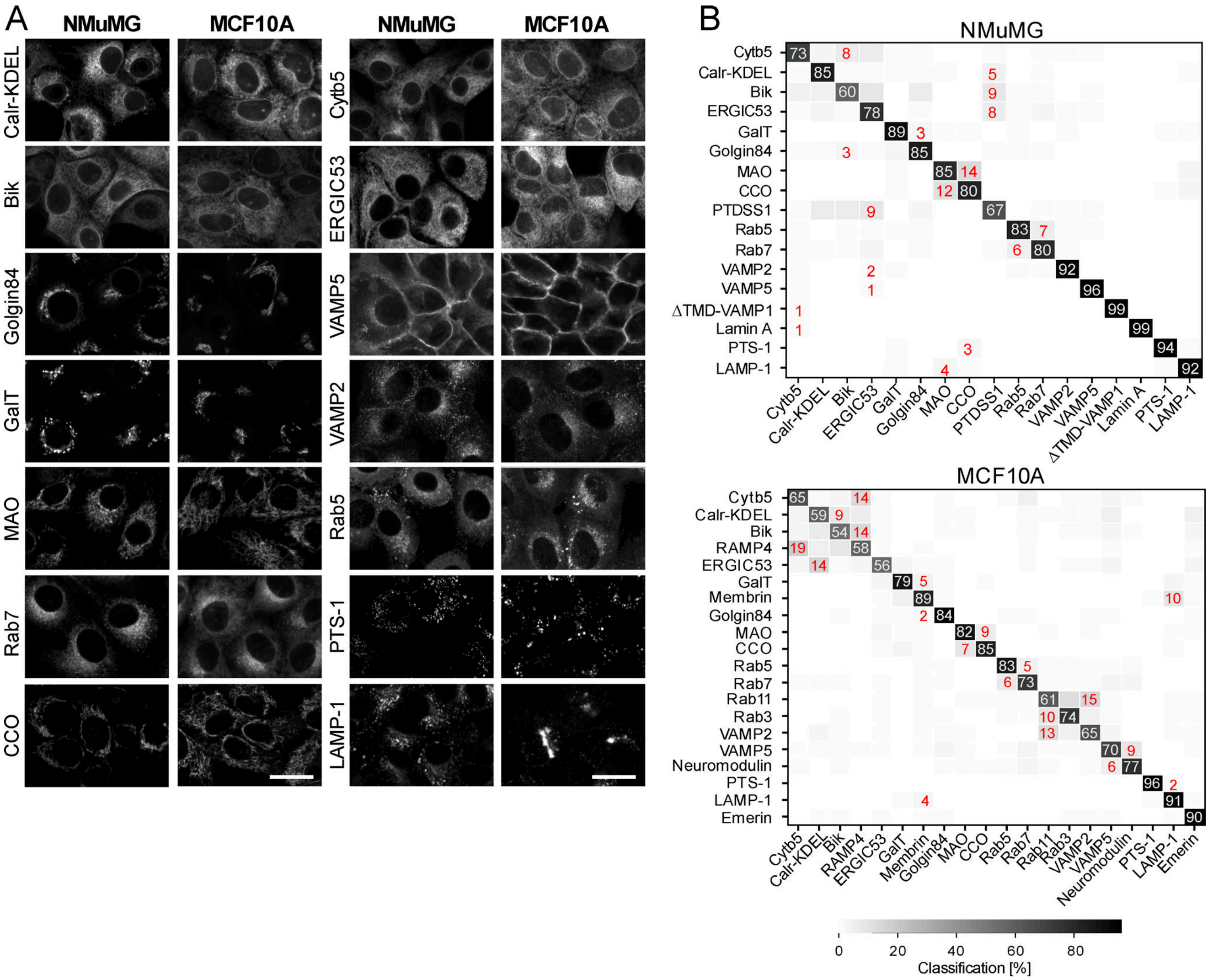


Figure 1. **Accurate assignment of subcellular localization from images of NMuMG and MCF10A cells expressing EGFP-tagged landmark proteins by RF classification. (A)** Representative microscope images of cells expressing the EGFP-tagged landmark proteins (Table 1). Scale bars, 25 µm. **(B)** Confusion matrices after RF classification of murine (NMuMG) and human (MCF10A) cells from landmark data not used in training the classifier, averaged over five independent classifications. White and red numbers show the percentage of cells assigned the highest and second highest classification landmark, respectively. SD < 1% (see Materials and methods).

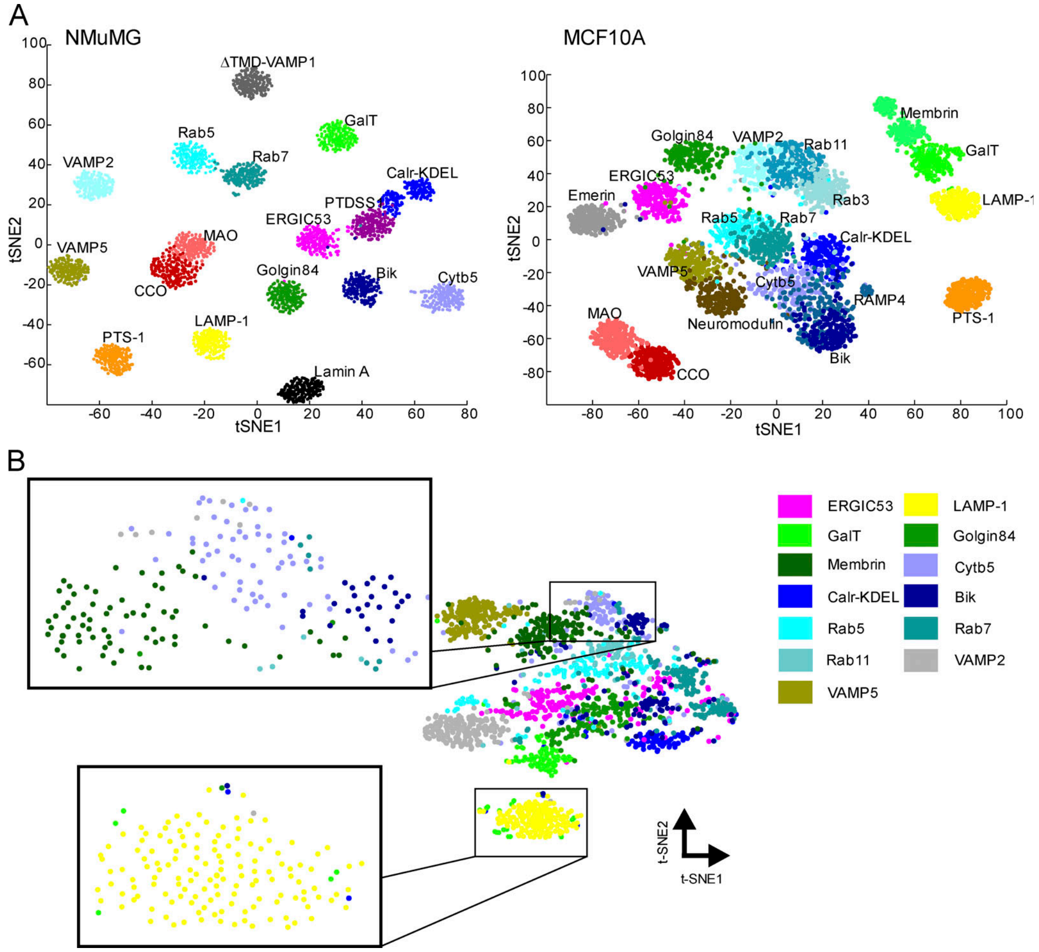


Figure 2. **Images of cell landmarks are well separated in 160-dimensional feature space. (A)** Two-dimensional visualization of 160-dimensional image data for both murine and human cells from individual landmarks using the t-SNE algorithm. Each landmark is represented by the 200 cell images (dots) closest to the centroid. **(B)** Two-dimensional t-SNE projection of landmark clusters belonging to the secretory pathway (mouse image library). To represent the 160-dimensional space encompassed by the landmarks with a number of data points practical for dimension reduction by t-SNE, individual landmarks were overclustered using Phenograph by setting the number of neighbors to five (see Materials and methods).

our existing set of NMuMG landmarks (Fig. 4 A). The ER-specific dye (BODIPY-thapsigargin) was predominantly classified with the ER marker Calr-KDEL (47%), but also with ER markers Cytb5 (15%) and Bik (20%) for a total ER classification of 82%. Similarly, the *cis*-Golgi–localized TA protein Golgi SNAP receptor complex member 2 (Membrin) was classified as mostly Golgin84 (62%) with a component assigned to the Golgi protein GalT (31%). Moreover, the mitochondrial stain Mitotracker and the mitochondrial matrix protein ornithine transcarbamylase (OTC) were both classified with the mitochondrial landmarks MAO (55% and 27%, respectively) and CCO (34% and 62%, respectively). In addition, the classifier successfully assigned outer membrane mitochondrial localization to the outer mitochondrial membrane protein Harakiri fused to the red fluorescence protein mLumin (66% outer mitochondrial MAO and 25% inner membrane CCO). The recycling endosome marker (Rab11) was classified with Rab5 (42%).

In addition to analyzing a new set of marker proteins, we also tested the performance of the NMuMG cell classifier by using human MCF10A cells expressing five of the landmarks (Rab5, ERGIC53, LAMP-1, Bik, and Golgin84). These proteins were stably expressed in MCF10A cells and imaged, and the resulting micrographs were analyzed using the classifier derived from images of NMuMG cells expressing all the landmarks. The classifier correctly assigned the localization of the landmark proteins expressed in MCF10A cells (Fig. 4 B). However, the ER protein Bik was mostly classified as the other ER-localized TA protein, Cytb5. Moreover, only 41% of Golgin84 in MCF10A was assigned to the NMuMG Golgin84 compartment, suggesting that there is a difference in the morphology of these compartments between the two cell lines. This hypothesis was verified by visual comparison of the corresponding images in the two cell lines (Fig. 4 C). This result demonstrates that subtle changes in morphology can be detected by the classifier, suggesting utility for a wide variety of genetic and chemical perturbation studies. However, for automated classification in more divergent cell types, it will be necessary to build a new library of landmark images.

We generated a second landmark library in the human cell line MCF10A (Fig. 1) that includes additional landmarks (Table 1)

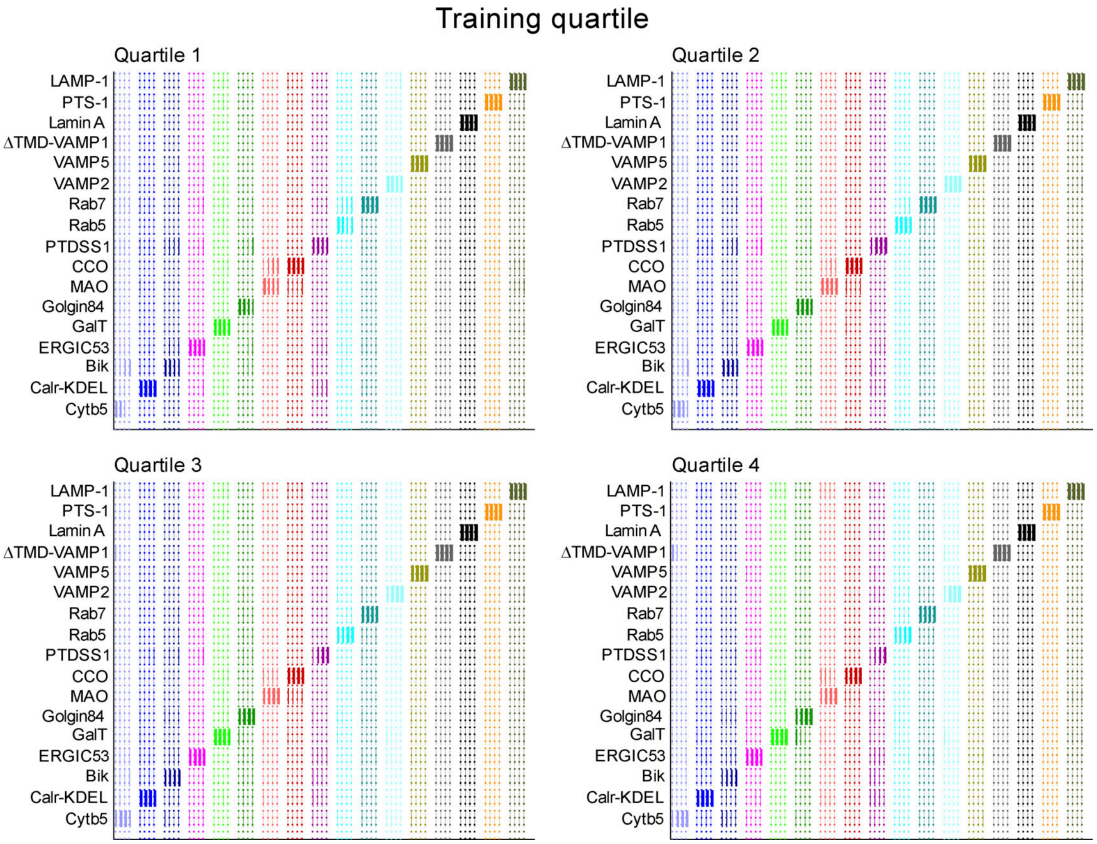


Figure 3. **Classification results for intensity quartile analysis.** Murine image data for landmarks, indicated to the left and by the color as in Fig. 2 A, were divided into quartiles based on the average pixel intensity within the cell region. RF classification was performed using randomly selected images from the quartile indicated above the panel for training and the rest of the data for testing. The thickness of the vertical bars indicates the fraction of the cells assigned to each landmark per quartile. Each vertical line corresponds to one quartile (quartile 1 to quartile 4, left to right). The dots serve as a location guide that is color-coded to facilitate identification of the row and columns.

and resulted in accurate assignment (77%) of subcellular localizations for images of individual cells not used in training (Fig. 1 B). Comparison of the five query proteins expressed in MCF10A revealed that as expected, the accuracy was somewhat (on average <20%) higher when assigned using the MCF10A instead of the NMuMG library. As expected, the improvement was dramatic for Golgin84 (41% to 84%; Fig. 4, B and C; and Fig. 1 B, respectively).

These results demonstrate that our analysis captures the subcellular landscape with sufficient resolution to accurately assign the localization of proteins and organelle specific dyes not used for training, regardless of whether they expressed RFP or GFP fusions or in human or murine epithelial cells, attesting to the robustness of the approach. Furthermore, the analysis is sensitive enough to separate landmarks ostensibly targeted to the same organelle (e.g., Cytb5, Bik, and Calr-KDEL) and thereby identify different protein distributions within organelles (Cytb5 and Bik) and/or discriminate resident versus recycling proteins (Cytb5 vs. Calr-KDEL).

Image-based analysis of the determinants of subcellular localization for TA proteins

As an example of one of the uses for the localization libraries, we systematically examined the targeting behavior of TA proteins. In multiple cell types, VAMP1 targets to the ER and then transits through the secretory pathway to the cell surface (Raptis et al., 2005; Chen and Scheller, 2001). When EGFP-TA was expressed in NMuMG cells, the images of 62,000 of 104,000 cells (~60%) were classified as VAMP5 localization. Although VAMP5 has been reported to be located primarily at the plasma membrane (Hong, 2005) due to the dynamic trafficking of proteins, images of the VAMP5 landmark in our library show protein localization throughout the secretory pathway. Thus, our results indicate that EGFP-TA has a distribution compatible with the biology of the protein the TA sequence was derived from.

We used EGFP-TA to investigate the features of the TA sequence that determine the specificity of localization within the secretory pathway and to other intracellular membranes. To analyze one region of the sequence systematically, we performed random mutagenesis on the codons for the last five amino acids of the TA sequence (IYFFT), which constitute the C-terminal sequence (CTS). Sequencing revealed 995 unique sequences, each of which was individually expressed in NMuMG cells. As expected, when classified using the reference library, many of the mutants were assigned VAMP5 localization. However, a few were localized primarily at the ER (n = 8) or mitochondria (n = 13) and were therefore selected for further analysis. Our initial focus was the mutants localized at mitochondria because the requirements for TA targeting to this organelle have been analyzed extensively, yet a clearly defined consensus has not emerged.

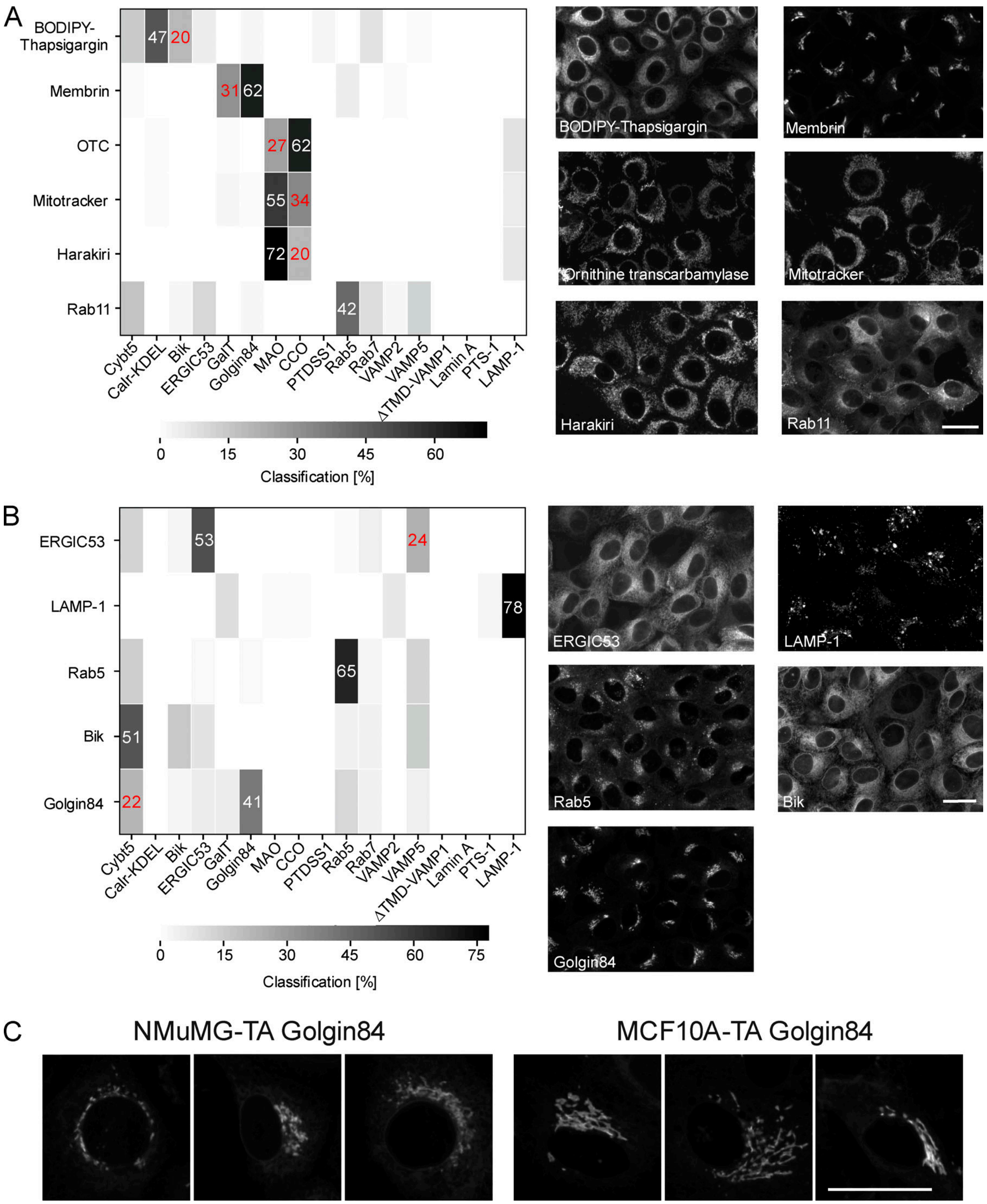


Figure 4. **Validation of the NMuMG RF classifier using images of cells expressing novel landmarks not used for training and images of MCF10A cells expressing selected landmarks. (A)** EGFP or mLumin fusions to targeting sequences from proteins with well-characterized localizations (Table 1) or organelle-specific dyes were used as queries for classification. **(B)** MCF10A cells expressing landmarks were used as queries using the NMuMG classifier. **(C)** Images of NMuMG and MCF10A cells expressing EGFP-tagged Golgin84. Numbers indicate the percentage of cells assigned to the most prevalent assigned landmark. Scale bars, 25 µm.

Targeting to the mitochondria is determined by positions of amino acid properties in the CTS

Mitochondria are a well-known targeting destination for TA proteins that play a crucial role in apoptosis (Cory and Adams, 2002), protein import (Horie et al., 2002), organelle and vesicle fission and fusion (Scott and Youle, 2010), and other functions. While several approaches have been used to decipher the sequence requirements for localization, it is still unclear what comprises a TA mitochondrial targeting sequence.

12 of the 13 CTS sequences that resulted in mitochondrial localization of EGFP-TA were five amino acids long. The one sequence with a four-amino acid CTS due to random incorporation

of a termination codon was assigned to the outer mitochondrial membrane (MAO), but not examined further. Of the five amino acid CTSs, eight were assigned only to mitochondria, all primarily to the outer mitochondrial membrane localization MAO landmark as expected for proteins anchored to mitochondria by a TA sequence. For three of these mutants, a smaller fraction was also assigned as the inner membrane landmark CCO. Of the four sequences not primarily assigned to MAO, none was primarily assigned to CCO. Even though 12 sequences is a small number, it was possible to visualize amino acid enrichment using a pLogo representation (O'Shea et al., 2013). The most overrepresented amino acid in sequences that localized EGFP-TA to mitochondria was arginine (R) at CTS position 2 (Fig. 5 A). Moreover, lysine (K) at CTS position 1 was also significantly overrepresented. There was a high occurrence of arginine at positions 3 and 5, but individually this did not reach statistical significance (Fig. 5 A). By carrying out an enrichment analysis for the number rather than position of positive charges, we observed that mitochondrial localization of mutants with at least three positively charged residues (K or R) was statistically significant (three positive charged residues, P = 5.6 e^{-5}; four positive charged residues, P = 4.9 e^{-9}, Fig. S3 A). This result is consistent with reports that positively charged residues are frequently observed in the CTS of mitochondrial TA proteins (Rapaport, 2003). However, previous publications (Marty et al., 2014; Rapaport, 2003) suggested that a net charge of the CTS region of +2 or greater is sufficient to target TA sequences to mitochondria. In contrast, for the TA protein Fis1, it has been reported that at least four positive charges are required for targeting to mitochondria (Rao et al., 2016). With our large datasets of random mutants, we can evaluate these rules systematically by examining sequences that target EGFP-TA to mitochondria and how frequently sequences that match the consensus do not target to mitochondria. When we examined the frequency of EGFP-TA mutants with five-residue CTS sequences with KR or RR as the first two amino acids, or net +2 or greater charges in the CTS targeting to mitochondria, the specificity of these simple rules was 0.86 or higher. However, the sensitivities and positive predictive values (PPV) are low (table in Fig. 5 A). Only 12 out of 133 +2 or greater charged CTS sequences (9.0%) targeted EGFP-TA to the mitochondria. Previous studies have not reported the frequency of sequences that adhere to the predicted rules but do not target as predicted, presumably due to the difficulty of manually assigning localization for large numbers of mutants, a task that is easily achievable with our automated reference library.

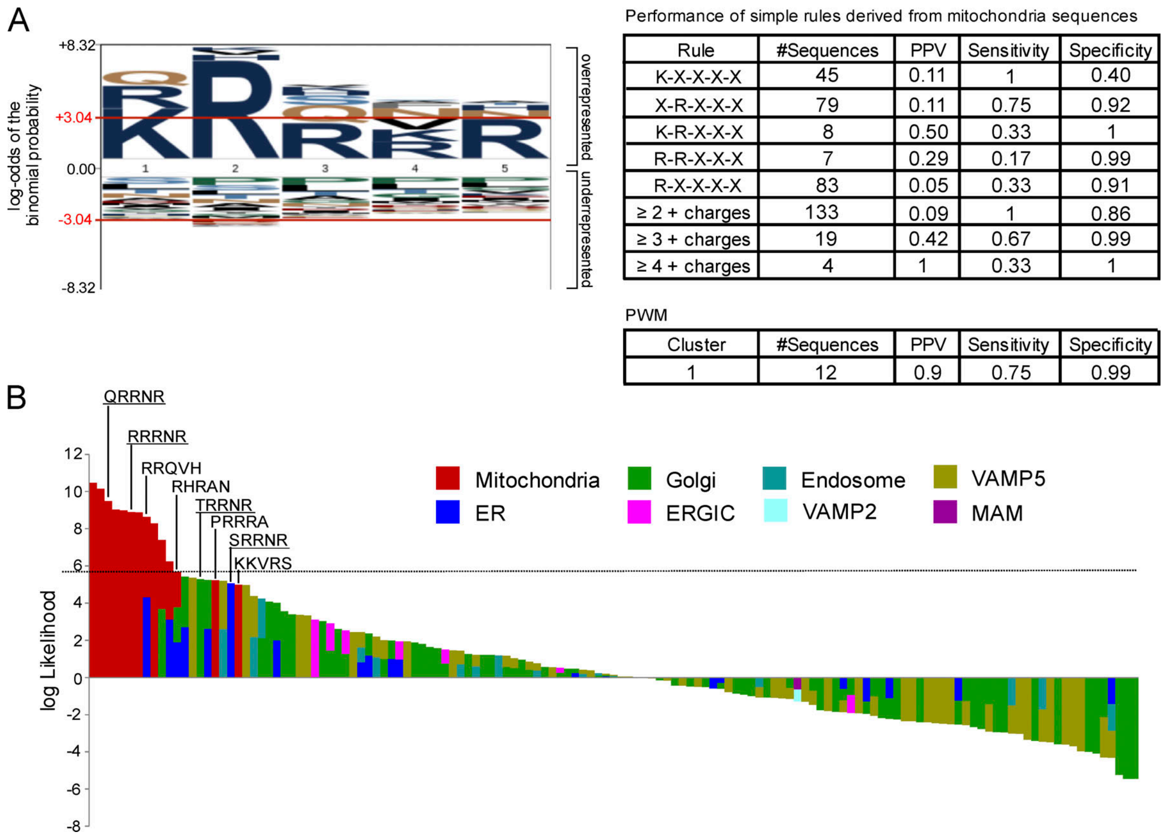
To further test the importance of positive charges, we generated mutants with a high proportion of positive amino acids (e.g., RRRNR, QRRNR, TRRNR, and SRRNR), all of which contain at least three positive charges and the over-represented R at position 2 (Fig. 5 B). The micrographs of EGFP-TA with CTS sequences RRRNR and QRRNR were assigned 89% and 80% mitochondrial localization, primarily to outer mitochondrial membrane localization (MAO, 65% and 62%), respectively. However, image-based classification assigned the mutants SRRNR and TRRNR to the ER landmark Calr-KDEL and Golgi, respectively, clearly indicating that three positive charges with an R at position 2 is not sufficient to target EGFP-TA to mitochondria.

To generate a rule for predicting localization of EGFP-TA proteins at mitochondria, we used the sequence data to generate a position weight matrix (PWM) based on amino acid properties. The PWM enabled calculating a likelihood value for localizing at mitochondria for all of the mutant sequences with at least two positive charges (Fig. 5 B). Visual inspection of this data suggests that a threshold of 5.68 (equal to that of the sequence RHRAN) optimally identifies CTSs that target EGFP-TA to mitochondria. With that threshold, the PWM correctly assigns the mutants RRQVH and RHRAN that are missed by the other rules. Moreover, it correctly assigns the designed mutants RRRNR and QRRNR to mitochondria and rejects the SRRNR and TRRNR sequences. The PPV, sensitivity, and specificity demonstrate that the PWM predictive rule clearly outperformed any of the other rules (Fig. 5, table). Thus, the locations of the biochemical properties of the amino acids contribute to the targeting signal within the CTS that specifies mitochondrial outer membrane localization.

To evaluate a CTS sequence predicted by the PWM to target to mitochondria in the context of a different protein, we examined the TA sequence of Bcl-2 fused to EGFP. Wild-type EGFP-TA-Bcl-2 targets to the ER, MAM, and mitochondria, as cell images were classified 45% Calr-KDEL, 20% PTDSS1, and 17% MAO (Fig. 6), consistent with previous reports demonstrating Bcl-2 localization at both ER and mitochondria (Zhu et al., 1996). Localization at multiple organelles also demonstrates that the TA sequence of Bcl-2 is permissive for insertion into multiple membranes.

After removing the SHK sequence that constitutes the CTS of Bcl-2 (Henderson et al., 2007), the protein Bcl-2-ΔSHK was localized almost exclusively to the ERGIC compartment, confirming that the CTS can determine localization for this protein. Replacing the SHK sequence with the CTS sequence KRRNR generated the protein EGFP-TA-Bcl-2-ΔSHK-KRRNR, which, as predicted from the PWM, was classified exclusively (96%) as MAO (outer mitochondrial membrane). Furthermore, when the CTS of EGFP-TA-Bcl-2 was replaced with the sequences FPCVN or WTNFK that localized EGFP-TA to the ER, the resulting proteins were primarily classified as targeting to the ER (~60% of the cell images). Similar to the EGFP-TA versions, most individual images were classified as Bik, but some were assigned to the other ER subdomains defined by Cytb5 and Calr-KDEL (Fig. 6). Taken together, these results indicate that for permissive TA sequences, targeting to the mitochondrial outer membrane can be achieved by CTS sequences defined by the PWM. The PWM-defined motif is the first demonstration that the position of amino acid properties within the CTS rather than sequence identity determines mitochondrial localization.

The ensemble of RF classifier correctly assigned cell images of several TA proteins that were not used in training to the outer mitochondrial rather than inner mitochondrial membrane. These included Harakiri, EGFP-TA-Bcl-2-ΔSHK-KRRNR, EGFP-TA-RRRNR, and EGFP-TA-QRRNR. This is remarkable considering a human observer would be unable to accurately determine outer from inner mitochondrial localization from the images (Fig. S4). Consistent with the classifier efficiently distinguishing localization at the outer and inner mitochondrial membrane, the



Performance of simple rules derived from mitochondria sequences

Rule	#Sequences	PPV	Sensitivity	Specificity
K-X-X-X-X	45	0.11	1	0.40
X-R-X-X-X	79	0.11	0.75	0.92
K-R-X-X-X	8	0.50	0.33	1
R-R-X-X-X	7	0.29	0.17	0.99
R-X-X-X-X	83	0.05	0.33	0.91
≥ 2 + charges	133	0.09	1	0.86
≥ 3 + charges	19	0.42	0.67	0.99
≥ 4 + charges	4	1	0.33	1

PWM

Cluster	#Sequences	PPV	Sensitivity	Specificity
1	12	0.9	0.75	0.99

Figure 5. **Targeting to the mitochondria is determined by both the position and number of positive charges in the CTS. (A)** Representation using pLogo of over- and underrepresented amino acids at each position of the CTS derived from all the sequences assigned to one or both of the classes of mitochondrial markers (MAO and CCO). The red horizontal bars correspond to P = 0.05. **(B)** Log-likelihood calculated using a PWM for all the sequences with at least two net positive charges. Bar height indicates log-likelihood, while the bar color indicates the organelle assignment by image classification. Classification threshold for the PWM (dotted line) based on the obvious breakpoint in the sequences (RHRAN). Due to the small number of EGFP-TA mutants that localized at mitochondria, mutants (underlined) were designed to test the performance of the PWM and simple rules with sequences not used for training. The table compares the performance of the different simple rules. X is any amino acid. + represents positively charged amino acids in the CTS.

confusion matrix (Fig. 1 B) indicates <15% overlap between targeting assignments for MAO and CCO. An alternative explanation is that an intensity difference between the inner and outer mitochondrial membrane protein landmarks (CCO and MAO) was sufficient to result in classification of queries as MAO based on intensity despite our selecting features less influenced by intensity. To test this hypothesis, individual cells were selected for training that expressed CCO or MAO with a similar defined intensity range (Fig. 7). This intensity filter was also applied to cell images of OTC and some of the EGFP-TA mutants that targeted to the mitochondria for use as a query set. The classification result (Fig. 7) of this image data resulted in all of the tail-anchored proteins correctly assigned as localized in the outer membrane, while OTC was correctly assigned localization to the inner membrane. Thus, the classifier accurately distinguished inner and outer mitochondrial membranes for six query proteins not used in training, independent of intensity (Fig. 7, heatmap).

ER localization and secretory pathway motifs can be identified from a sparse dataset

The 995 mutants constitute a very sparse dataset representing <0.03% of the ~3.36 million possible five–amino acid CTS sequences. The more specific the requirements are for a targeting sequence for any particular subcellular location, the fewer examples there will be in a sparse dataset. As a consequence in our dataset, even a single sequence might represent a class of related sequences that localize similarly (a motif). To test this hypothesis, we examined further the eight EGFP-TA mutants that were assigned only ER localization. These sequences already suggest that there are sequence dependencies for the localizations defined by the landmarks Bik, Cytb5, and Calr-KDEL. Of the eight sequences, three (FPCVN, WTNFK, and DPTDS) localized EGFP-TA primarily (45–59%) to the localization defined by Bik, while the other five sequences (DEPGH, PEHVS, PKWVT, PSNHQ, and RVRPG) were assigned (43–55%) to the localization identified by Cytb5. Consistent with those representing distinct distributions within or subcompartments of the ER, none of these proteins were assigned significant localization to any other locations, including the other ER landmarks (Fig. S5 B).

To assess the importance of individual amino acids in determining ER localization, we chose one mutant, FPCVN, for further study; because of the 995 sequences, it is the only one ending in CVN. Images of this mutant were assigned 61% to the

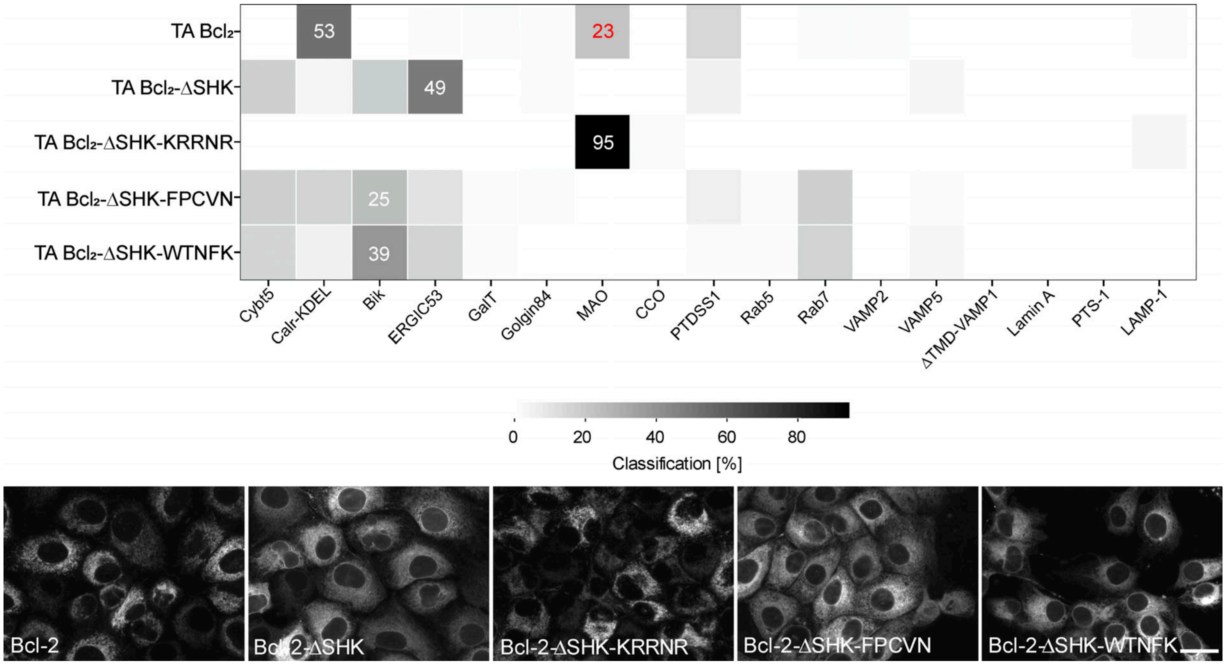


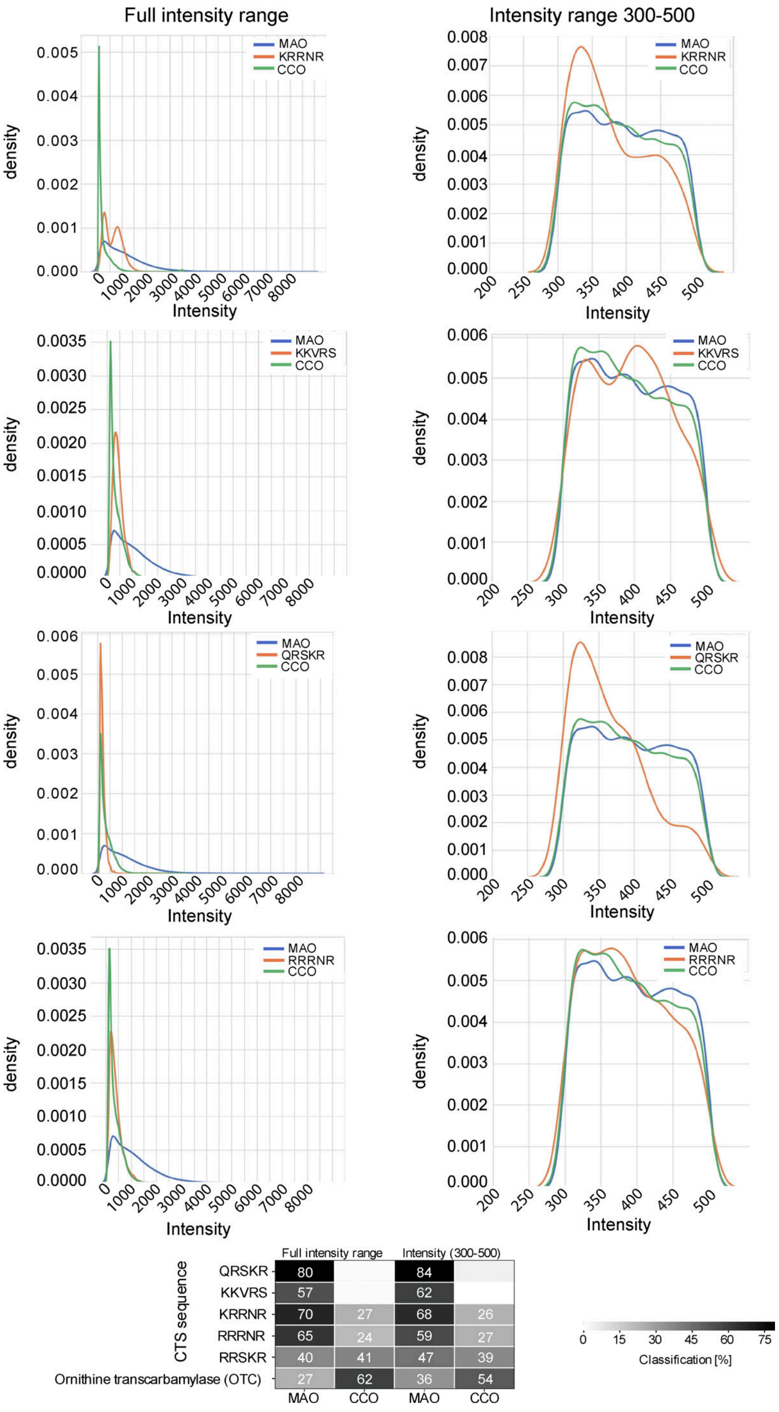
Figure 6. **CTS motifs that target the TA-Bcl-2 sequence to different landmark localizations.** Heat map: RF classifier assignments of localization according to the scale below. The most frequent assignment for each protein is indicated as a percentage within the heat map. Lower panels: Sample micrographs of the indicated EGFP-fusion proteins expressed in NMuMG cells. Scale bar, 25 µm.

ER landmarks (46% Bik localization) with <20% assigned to any other location (Fig. 8 and Fig. S5 B).

To determine if the CVN sequence is part of a motif, the first two residues of the FPCVN sequence were mutated to create a new set of CTS sequences on EGFP-TA that differ by one or two amino acids. We did not design "CVN" mutants with negative charges because mutants without negatively charged amino acids are more likely to escape the ER/post-ER compartments and localize at the plasma membrane (Fig. S3 B). Previous work on membrane proteins suggests that the length of the transmembrane domain (TMD) is a key feature that determines localization to the Golgi versus plasma membrane (Sharpe et al., 2010). To test this notion further, the changes were selected to alter the hydrophobicity at positions 1 and 2, thereby potentially extending the length of the hydrophobic region of the TA sequence by up to four residues as the C and V of the CVN sequence are also hydrophobic. Consistent with our sparse data-motif hypothesis, 14 of the 22 "X-X-C-V-N" sequences resulted in mutants that were classified as localizing primarily to the ER (Cytb5, Bik, or Calr-KDEL). This is a remarkable result considering that only 8 of 995 random mutants were localized similarly. To demonstrate that assignment to ER localization was not driven by image intensity, we repeated the analysis with selected images for training and testing with similar distributions of image intensities as shown above (Fig. 7). Restricting the distribution of intensities made no significant difference to the image classification results (Fig. S5 C).

Plotting the average hydrophobicity (Kyte and Doolittle, 1982) of the two positions (Fig. 8) demonstrated that sequences with hydrophobicity <0.6 (i.e., hydrophilic) together with CVN constitute a new motif for targeting EGFP-TA to the ER. All of the mutants with this motif were assigned significant localization at the ER with variable localization also at ERGIC and endosomes. Of the seven sequences with intermediate hydrophobicity (>0.6 and <1.7), four (FPCVN, SLCVN, GLCVN, and SVCVN) were retained in the ER. The exceptions (LRCVN, LWCVN, and LGCVN) that were assigned significant VAMP5 localization demonstrate that hydrophobicity is not required to progress through the secretory pathway. Of the seven mutants with hydrophobicity (1.7–1.8), LGCVN, WICVN, and IWCVN were all >48% assigned VAMP5, while GLCVN and SVCVN were primarily assigned to the ER with only partial localization at endosomes and no significant localization with VAMP5. Furthermore, the sequences LWCVN, WICVN, IWCVN, and CICVN that exhibited the strongest preference, >50% assigned localization to the VAMP5 compartment, ranged in CTS sequence hydrophobicity from modest (1.45) to high (3.8). Finally, some mutants suggest that rather subtle sequence features are involved in retention or export from the ER. For example, RWCVN was assigned to ER, yet switching the order of the first two amino acids to generate the CTS sequence WRCVN resulted in assigned localization throughout the secretory pathway with equivalent numbers of cell images assigned to ER (29%) and endosomes (30%). Thus, while our results confirm a weak correlation between hydrophilicity and retention at the ER, it is clear that hydrophobicity is not the only driving force for sorting to or exclusion from the plasma membrane or endosomes.

To our surprise, some of the mutants with ER-assigned localizations were classified as similar to Bik, while others were assigned to Cytb5, suggesting significant differences in their distribution in the ER. Consistent with this interpretation, several other mutants were assigned primarily to one or the other of these landmarks (Fig. S5, A and C).



	Full intensity range		Intensity (300-500)	
	MAO	CCO	MAO	CCO
QRSKR	80		84	
KKVRS	57		62	
KRRNR	70	27	68	26
RRRNR	65	24	59	27
RRSKR	40	41	47	39
Ornithine transcarbamylase (OTC)	27	62	36	54

Figure 7. **Image classification as outer or inner mitochondrial membrane does not depend on image intensity.** Intensity graphs for images randomly selected from the full intensity range or the intensity range of 300–500 arbitrary intensity units plotted as density to facilitate comparison as continuous curves. Data are for the mitochondrial landmarks (MAO, CCO) and for the query proteins including EGFP-TA mutants and images of OTC. Heat map shows classification result for images with full vs. 300–500 intensity range (see Materials and methods).

Discussion

Our development and characterization of reference datasets, each containing more than 500,000 optically validated individual cell images, enabled the identification of 160 features useful for assigning subcellular localization with an ensemble of RF classifier. These reference sets of optically validated images, computed features, classification results, and cell clones (available from Addgene) are tools that can be used to interrogate subcellular localization, for analysis of diversity in organelle morphologies, and as a reference standard for algorithm development. For example, our automated analysis of the images highlighted both morphological similarities (mitochondria) and differences (Golgi) between premalignant murine and human mammary epithelia cells grown in monolayers (Fig. 4, B and C). Furthermore, by assigning localizations to landmarks rather than specifying a specific organelle, our approach accounts for proteins having multiple localizations. For example, VAMP5 is generally referred to as a plasma membrane protein, but here the VAMP5 localization encompasses much of the secretory pathway because at steady-state, much of the protein is in transit. Because the cell lines and the derived clones have relatively stable genomes, they are expected to constitute a platform

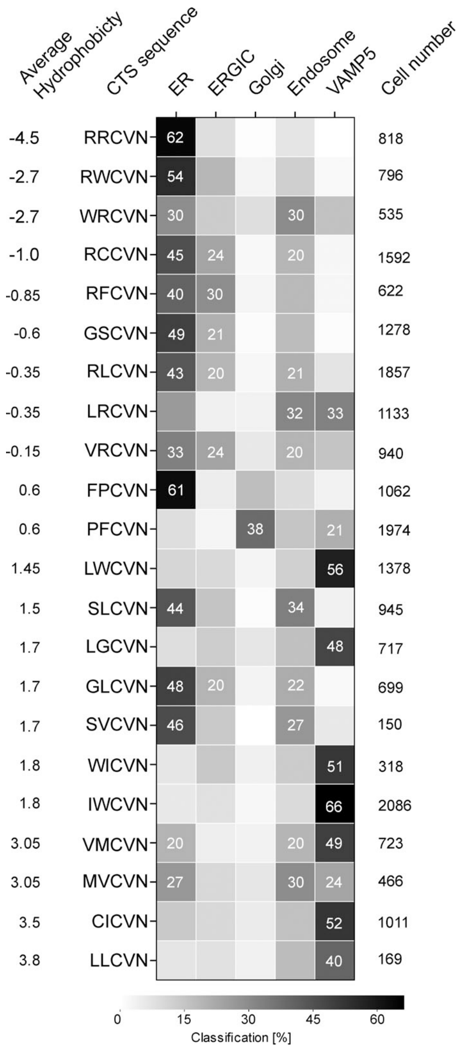


Figure 8. **The first two CTS amino acids determine localization of EGFP-TA at ER or enable transit to the plasma membrane for a CTS ending with CVN.** RF classification of localization for EGFP-TA mutants with CTS sequences derived from the ER localized mutant FPCVN arranged by the average hydrophobicity (Kyte-Doolittle) of positions 1 and 2. Assignment of the cell images (≥20%) for any one mutant to a single localization is indicated numerically and by gray shade. Localization of <20% of the cells is indicated by gray shade only. Organelles are defined as follows: ER (Cytb5, Calr-KDEL, Bik), ERGIC (ERGIC53), Golgi (Golgin84, GalT), and endosome (Rab5, Rab7). The number of cells analyzed for each mutant is indicated to the right.

for image-based analyses of genetic manipulations, cell signaling events, and quantitative measurement of cellular responses to environmental and extracellular matrix changes as well as responses to drugs and other perturbations. Here we have used the images and feature sets as tools to examine the sequence requirements for motifs that regulate TA protein subcellular localization.

Motifs for localization at mitochondria

Our dataset included only 12 mutants with a five–amino acid CTS that targeted EGFP-TA to mitochondria. However, comparison with the almost 1,000 sequences that did not result in localization at mitochondria was sufficient to identify the shared characteristics of proteins assigned mitochondrial localization as a PWM for the amino acids within the CTS. Thus, in contrast to previous results, our data directly demonstrate that the presence of positively charged amino acids is not sufficient for targeting mitochondria. Indeed, 133 EGFP-TA mutants with positively charged CTS sequences (+2 or greater) did not localize to mitochondria (Fig. 5). CTS sequences that targeted to mitochondria did not include those with a longer hydrophobic core sequence, a result consistent with other studies using smaller datasets (Costello et al., 2017). However, a shorter hydrophobic core was also not sufficient to target a TA protein to mitochondria (Fig. S3 C). Indeed, even when the ≥2 positive charges and minimal hydrophobic core were combined, there were 121 proteins targeted to nonmitochondrial locations compared with 12 assigned to the mitochondria.

Automated assignment of localization will facilitate future analyses of more sequences to determine if there are multiple independent motifs for mitochondrial localization of TA proteins. Classification based on the computed feature data clearly enables automated assignment of outer mitochondrial membrane localization even though such assignment is not possible based on visual inspection of the images (Fig. 7 and Fig. S4). Our approach will also enable generation of a more robust PWM for such motifs to fully capture the rules governing localization of TA proteins at mitochondria.

ER-localized TA proteins

Our observation that different ER-localized TA proteins have distinct distributions suggests that the ER may be composed of multiple morphologically distinct subcompartments. A related explanation is that different TA proteins have different residence times in the ER or distinct regions of the ER. Of particular interest is the observation that the TA sequences from Bik- and Cytb5-localized EGFP-TA result in different distributions of the proteins (Fig. 1 B and Fig. 2 A).

Consistent with the CTS sequence FPCVN being a representative of a bona fide localization motif for a novel Bik distribution within the ER, the sequence resulted in the same localization for EGFP-TA-Bcl-2-ΔSHK (Fig. 4). Unexpectedly, inverting the first two amino acids of FPCVN to generate PFCVN resulted in export of EGFP-TA from the ER and localization at Golgi and VAMP5 destinations. In contrast, the sequence RRCVN resulted in 48% assignment of EGFP-TA localization as most similar to Calr-KDEL with negligible protein assigned to Bik localization, indicating that the different sequences result in images that are morphologically distinct. The differences in assigned localizations are most likely due to differences in the distributions of the protein within the ER. Whether the distributions reflect differences in residence times and/or functionally distinct subdomains remains to be determined, but the differing assignments are clearly not the results of different expression levels of the proteins (Fig. S5 C).

Data from other mutants reinforce the concept that proteins ostensibly targeted to the same organelle have different distributions, potentially the result of specific ER subdomains. The sequences of two designed mutants, RWCVN and GSCVN, resulted in EGFP-TA assignment to Calr-KDEL, while the CTS sequence WTNFK resulted in assignment of both EGFP-TA and EGFP-TA-Bcl-2-ΔSHK to the Bik landmark. These results also demonstrate that despite our not identifying by random mutagenesis (i.e., within our 995 random mutants) a CTS sequence for localization of EGFP-TA to the Calr-KDEL landmark, such sequences clearly exist.

Mutants of EGFP-TA with a CVN sequence and a single R residue that lost assignment preference to a specific ER landmark provided further insight into localization requirements: those with an R in the first position (RWCVN, RCCVN, and RFCVN) were all retained in the ER (40–53%), but less than a third of the protein was assigned to any one of the ER landmarks (Fig. S3). In contrast, mutants with an R at the second position (WRCVN, LRCVN, and VRCVN) were all assigned to the Cytb5 landmark and to locations further along the secretory pathway, including endosomes and the VAMP5 compartments.

To our surprise, only mutations that resulted in a distribution similar to the Cytb5 landmark also had partial localization at endosomes. These sequences (WRCVN, RLCVN, LRCVN, VRCVN, SLCVN, SVCVN, MVCVN, and VMCVN) share no obvious similarity and vary in hydrophobicity across the entire scale. However, export of EGFP-TA to the secretory pathway is not a defining feature of the distribution characteristic of the Cytb5 landmark, as other TA CTS sequences including DEPGH, PEHVS, PKWVT, PSNHQ, and RVRPG resulted in localizations assigned as the Cytb5 landmark without significant endosome or plasma membrane localization (Fig. S5). Thus, in this case, if the localization is defined by residence times, it must reflect residences with distinct distributions within the ER. Furthermore, the extent to which the different mutant proteins were exported from the Cytb5 compartment was variable. The mutant MVCVN was equally distributed between the Cytb5, endosome, and VAMP5 compartments, while VMCVN was assigned primarily to VAMP5. In contrast, the six TA proteins assigned to either the Bik or Calr-KDEL landmarks were not exported from the ER (Fig. S5).

The bias toward export of TA proteins assigned to the Cytb5 ER landmark suggests the intriguing possibility that the distribution of this landmark represents a subcompartment that may play an important role in sorting TA proteins. Future experimentation is needed to determine if this is unique to TA proteins or if other proteins are sorted from a different ER subcompartments. For example, the Calr-KDEL distribution may represent a subcompartment involved in transport of secretory proteins since that landmark is an ER luminal protein localized to ER in part because it is efficiently recycled from the Golgi.

Conclusions

We present a database of optically validated images for both human and murine epithelia together with a set of quantitative features that can be used to study protein localization in individual living cells. It combines (1) a large number of images to provide examples of most if not all organelle morphologies including those throughout the secretory pathway that result from dynamic movement, (2) use of an unbiased classification tool to assess subcellular localization that can be readily extended to include additional landmarks, and (3) use of a PWM approach to predict and verify subcellular localization motifs. With these tools, we identified a new motif for targeting TA proteins to mitochondria and multiple morphologically and likely functionally distinct protein distributions within the ER. The sparsity of the sequence coverage provided by 995 unique mutants suggests that our dataset contains many more targeting motifs similar to the CVN sequence that can be characterized using these techniques. Ultimately, the utility of the tools described here is limited only by the creativity of the user.

Materials and methods

Plasmid construction

All coding regions were cloned into pQCXIP, a retroviral expression vector (Clontech) encoding the monomeric fluorescent reporter protein EGFP-S65T upstream of the gene of interest, unless otherwise stated. The vector pQCXIP has an incomplete retroviral 3′ long terminal repeat (LTR) to prevent replication. Therefore, the LTR was repaired by excising the missing piece from pBabe-puro and introducing it into pQCXIP. This restored the function of 3′ LTR (pQCXIP-Repaired [R]), enabling rescue of the virus as described below. To generate pQCIXP-R-EGFP-TA, EGFP from pEGFP (Clontech) was cloned into pQCIXP-R by digestion of the pEGFP vector with AgeI-HF(NEB) and BamHI-HF(NEB). The open reading frame from human VAMP1 (Open Biosystem/Dharmacon) was amplified using Phusion High-Fidelity DNA Polymerase (NEB) and forward (Fwd) and reverse (Rev) primers with the following sequences: Fwd-VAMP1: 5′-GCGTCGACATGGAGAGCAGTGCTGCCAAGCTAAA-3′, Rev-VAMP1: 5′-GCGGATCCTCAAGTAAAAAAGTAGATTACAATAAC-3′. The resulting products were subcloned into pQCXIP-R-EGFP digested with SalI-HF (NEB) and BamHI-HF (NEB). As landmarks, the following coding regions were cloned into pQCXIP: Rab5, Ras-related protein Rab7A (Rab7), and Rab11 (kindly provided by J.C. Simpson, University College Dublin, Dublin, Ireland); Golgin84 and Bik (Open Biosystem/Dharmacon); ERGIC53, VAMP2, and VAMP5 (OriGene Technologies); Bcl-2 (TA sequence residues: 213–239 [isoform 2]; Zhu et al., 1996); PTDSS1 (Stone and Vance, 2000; kindly provided by J. Vance, University of Alberta, Edmonton, Canada); cytochrome b5 (Zhu et al., 1996); harakiri, membrin, Ras-related protein Rab3C, emerin, and ribosome-attached membrane protein 4 (OriGene Technologies); and the peroxisomal targeting sequence. The sequence encoding the TA region of MAO was assembled from oligonucleotides. pBABE-puro-GFP-wt-lamin A was a gift from T. Misteli (National Cancer Institute, National Institutes of Health, Bethesda, MD; Addgene plasmid 17662; Scaffidi and Misteli, 2008), and LAMP-1-mGFP was a gift from E. Dell'Angelica (University of California, Los Angeles, Los Angeles, CA; Addgene plasmid 34831; Falcón-Pérez et al., 2005). The following coding regions were fused to the N-terminus of EGFP-S65T in the pQCXIP vector: OTC (32 amino acids of the N-terminal sequence; Horwich et al., 1986); the mitochondrial targeting sequence from subunit VIII human CCO; a sequence encoding the N-terminal 81 amino acids of GalT; and the N-terminal 20 amino acids of neuromodulin (Skene and Virág, 1989). Calr-KDEL includes the ER targeting sequence of calreticulin fused to the N-terminus of EGFP and the ER retention sequence KDEL at the C-terminus of EGFP. A plasmid encoding the fluorescent protein mLumin was kindly provided by J, Hardy (University of Massachusetts, Amherst, MA). The coding region for Harakiri was fused to the 3′ end of the mLumin coding sequence by using SalI-HF and BamHI-HF. For expression in MCF10A, the coding regions were subcloned into pLVX-EF1a-IRES-Puro (Clontech) by using NotI-HF (NEB) and AvrII (NEB).

Random mutagenesis

Generation of random mutations in the VAMP1 CTS in EGFP-TA was performed using PCR with degenerate primers. The coding sequence for the fusion protein EGFP-TA was amplified with primers (Fwd-CF: 5′-CCGCGGCCGCACCGGTCGCCACCAT-3′, Rev-CF: 5′-GCGGAATTCCGGATCCTCAMNNMNNMNNMNNMNNTACAATAACTACCACGATGATGG-3′; Integrated DNA Technologies) containing five degenerate codons NNK at the 3′ end of the coding sequence. To generate the "CVN" mutants, the following reverse primer was used: Rev-CVN: 5′-GAATTCCGGATCCTCAATTAACACAMMNMMNTACAATAACTGCCACGATGATGGC-3′ (Integrated DNA Technologies). Fwd primer is the same that was used above. The pQCXIP-R plasmid was digested with BamHI-HF (NEB) and AgeI-HF (NEB), the PCR products were ligated into the cut plasmid using the Cold Fusion technology (System Biosciences), and the DNA was transformed into *Escherichia coli* bacteria DH5α. After incubation at 37°C for 12 h to complete ligation and amplify the plasmids while minimizing the number of copies of identical plasmids, the transformants were harvested and pooled, and plasmid DNA was isolated using a Presto DNA Mini Plasmid kit (FroggaBio) and packaged into retroviral particles (Phoenix cell line). Sequence analysis after rescue from cells (see below) revealed 995 unique sequences from 1,220 clones.

Cell lines and culture

The cell lines used were selected because they are both relatively genomically stable breast epithelia. Our assumption is that genomically stable cell lines would exhibit a relatively normal distribution of localization phenotypes characteristic of the cell type. NMuMG cells (a generous gift of J. Wrana, Lunenfeld-Tanenbaum Research Institute, Toronto, Canada) were cultured in DMEM, containing 10 µg/ml bovine insulin (Sigma), 10% FBS (Gibco), and penicillin/streptomycin (Wisent). The retroviral packaging cell line (Phoenix) and HEK293T were grown in DMEM (Gibco), supplemented with 10% FBS and penicillin/streptomycin. MCF10A were cultured in DMEM/F12 (Gibco) supplemented with 5% horse serum (Gibco), 10 µg/ml insulin (Sigma), 0.02 µg/ml EGF (Preprotech), 0.5 µg/ml hydrocortisone (Sigma), and 0.25 mg/liter isoproterenol (Sigma). All cell lines were maintained in a 5% CO_2 atmosphere at 37°C. All cell lines tested mycoplasma-free using a PCR-based detection system (Hopert et al., 1993). MCF10A were analyzed by comparative genomic hybridization (Mills et al., 2015). Both Phoenix and HEK293T were genotyped at the Centre for Applied Genomics, SickKids, Toronto, Canada.

Transfection and transduction

Retrovirus was derived by transient transfection of pQCXIP-R-EGFP-TA into the Phoenix packaging cell line using Fugene HD (Promega). After 24 h, the virus-containing cell medium was filtered (0.45 µm, PALL) and transferred onto the target cell line. To increase the efficiency of transduction, 8 µg/ml polybrene (Sigma) was added. Stable colonies were selected in 10% FBS/DMEM containing 2 µg/ml puromycin (Sigma) and followed by sorting single cells (BD FACSAria II) expressing EGFP into individual wells of multiwall plates (TC plate 96-well, standard, F, Sarstedt). Once a colony formed, it was grown under puromycin selection as above until further analysis.

Rescuing of proviral DNA

To retrieve the nucleotide sequences of the mutants, cell clones individually transduced with the replication-repaired pQCXIP-R-EGFP-TA mutants were seeded in a 96-well format plate (TC plate 96-well, standard, F, Sarstedt) and transfected with the pCL-ECO packaging vector (Imgenex) using Fugene HD (Promega). Transfection with pCL-ECO enables virus production from the cells transduced with pQCXIP-R-EGFP-TA mutants. After 48 h, the supernatant was collected for viral RNA extraction (Murdoch et al., 1997). Briefly, Trizol LS (ThermoFisher) was added to the supernatant in a 3:1 ratio along with chloroform and yeast tRNA under RNase-free conditions. After centrifugation and isopropanol precipitation, viral RNA was subjected to reverse transcription by using SuperScript III Reverse transcription (ThermoFisher) to generate cDNA. Sequence analysis was performed at the Centre for Applied Genomics sequencing facility.

Lentivirus production and transduction

To express all landmarks in the human cell line MCF10A, the coding regions were subcloned into the lentiviral vector pLVX-EF1a-IRES-Puromycin (Clontech). The lentiviral DNA plasmid and both pPAX2 and pMD2.G plasmids were transfected into HEK293T cells at 1:1:0.1 ratios by means of calcium phosphate precipitation. Prior to transfection, the three plasmids were briefly mixed with sterile 1× Hepes buffered saline. 1× Hepes buffered saline is composed of Hepes (5 g/liter), NaCl (8 g/liter), dextrose (1 g/liter), KCl (3.7 g/liter), and $Na_2HPO_4 \times 7H_2O$ (0.19 g/liter). This mixture was then supplemented with 2.5 M $CaCl_2$ to a final concentration of 0.14 M and incubated for 20 min at room temperature before it was added to the HEK293T cells.

After 48 h, the virus-containing supernatant was collected from the HEK293T cells and filtered through a 0.45-µm filter unit and transferred onto MCF10A cells. To obtain stable cell lines, selection was performed by adding puromycin (2 µg/ml) 48 h after transduction.

Live cell imaging

NMuMG and MCF10A cells expressing one of the landmarks or EGFP-TA mutants were seeded in separate wells in 384-well microplates (CellCarrier-384 ultra, B128 SRI/160; Perkin Elmer) and allowed to grow for 24 h before staining with the nuclear dye DRAQ5 (5 nM; Biostatus). To account for any technical interference that may affect the classification output, cell clones expressing the landmarks or mutants were imaged on multiple days and positions within the wells. In addition, the positions within the plates where the cell clones were grown and imaged was varied. Where indicated, cells were also stained with either Mitotracker Red (500 µM; ThermoFisher) or BODIPY-thapsigargin (500 µM; Setareh Biotech) according to the manufacturer's instructions. Plates were imaged (two to eight wells, 20–30 fields of view) on two different spinning disk automated confocal microscopes of the same model (OPERA QHS; PerkinElmer) with 40× water objectives (NA = 0.9) in a defined temperature (37°C) and CO_2 (5%) environment by using Evo-Shell acquisition software. Images were collected using 3-Peltier-cooled 12-bit CCD cameras (Type sensiCam, camera resolution 1.3 megapixels; PCO.imaging) either with a binning of two or unbinned. Unbinned images were binned numerically before segmentation and feature extraction. Imaging the controls and some query lines on multiple microscopes, with and without camera binning, enabled feature counter selection to remove features characteristic of the imaging platforms. Assessment of new query cell images was processed as described above along with a set of established landmarks as controls.

Image processing

Image segmentation

To identify individual cells, as well as nuclear and cytoplasmic areas for each cell in fluorescence micrographs, image segmentation was performed using PerkinElmer Acapella 2.0 software (Nuclei Detection Algorithm A). The nuclear detection algorithm includes following parameters: threshold adjustment (1.5), individual threshold adjustment (0.4), minimum nuclei distance (15), nuclear splitting adjustment (15), minimum nuclear area (300), and minimum nuclear contrast (0.1). Briefly, the nuclear region was identified using an image of DRAQ5 staining. Using the nucleus as a seed, the cell region was identified using a watershed algorithm and the low-level cytoplasmic staining due to DRAQ5. The region of the cell that is not in the nucleus is considered to be the cytoplasm. Additionally, a ROI mask was computed by identifying all the pixels that are 1.5 times the mean pixel intensity of the cell mask in the EGFP channel. Regions with at least 30 contiguous pixels were retained as ROIs. Images of cells touching the micrograph edge were discarded.

Removing out-of-focus cells

To investigate the steady-state localization of a protein in cells, images of dividing or dying cells were removed as they have been shown to interfere with classification (Huang and Murphy, 2004). Such cells are typically out of focus and exhibit a different texture and morphology of the nucleus in the DRAQ5 channel compared with other cells. These features enabled the generation of an automated image quality control algorithm that effectively removed images of cells that were out of focus, dividing, or dead. To train a classifier to remove these images, we collected cell images of landmarks that were in-focus and 4 µm below the in-focus plane. The cells were identified by image segmentation, and image features were calculated. Since information regarding cell focus can be obtained from the nuclear staining, which should be similar for cell clones independent of the EGFP protein expressed, we used texture and morphology features from the DRAQ5 images to create a binary random forests classifier and identity features necessary to separate cell images that are in focus and out of focus. Using the new feature subset, a new random forests classifier was built to identify within fields of view images of individual cells that are in focus and out of focus. The classifier was then applied to the entire dataset to remove individual cell images that were out of focus.

Examination of the rejected images revealed that this classifier also removed many cells that, when visually inspected, appeared to be in focus, but were usually not correctly segmented, suggesting there may have been a minor problem in focus. Our approach differs from previous work on focus quality control in which an entire image field was identified as out of focus rather than images of individual cells within an image field (Bray et al., 2012). The sensitivity for individual class before and after removal of out-of-focus cells is shown in Fig. S1.

Removing cells with improper segmentation

Since segmentation was performed using images of the DRAQ5 staining, we used the same information to remove incorrectly segmented objects. We computed two parameters from our existing features: R1, a ratio of average pixel intensity inside the nuclear mask to the average pixel intensity inside the cytoplasm mask, and R2, a ratio of pixel intensity SD inside the nuclear mask to the pixel intensity SD inside the cytoplasm mask.

$$R_1 = \frac{\text{Nucleus Intensity}}{\text{Cytoplasm Intensity}}$$

$$R_2 = \frac{\text{Standard Deviation of Nucleus Intensity}}{\text{Standard Deviation of Cytoplasm Intensity}}.$$

The first parameter ensures the appropriate level of nuclear stain since DRAQ5 staining in the cytoplasm is typically far lower compared with staining in the nucleus. Similarly, the second parameter captures the changes in the staining patterns between the nucleus and the cytoplasm region. Due to differences in staining of the heterochromatin inside the nucleus, the deviation of nuclear intensity should be much higher when compared with the intensity deviation in the cytoplasm region. A threshold for both ratios was determined empirically to be 3.5. All the cells below the threshold were removed from the dataset. We then computed the ratio of the nucleus area to the cell area and removed the top and bottom 5% of the cells based on this ratio. This last step removed images of cells that were abnormally large (typically senescent or improperly segmented) or small (typically dying). The same quality control steps were also applied to the EGFP-TA CTS mutants. The 10 mutants for which <50 suitable cell images remained after the quality control steps were removed from the analysis, resulting in a total of 995 unique random mutants.

Cell image feature extraction

Features were measured for nuclear and cytoplasmic (including ROI) areas for DRAQ5 staining and EGFP expression, respectively, using a custom PerkinElmer Acapella script (code available on https://github.com/DWALab/Schormann-et-al). For each cell image, 495 morphological and statistical image features were calculated (Collins et al., 2015). Micrographs of individual cells were automatically selected for inclusion in the reference library based on numerical assessment of image and segmentation quality (R1 and R2; Fig. S1).

Feature selection

For classification of localization, a reduced set of features was selected to minimize (1) the impact of intensity variations due to changes in the intensity of the illumination or efficiency of collection, and (2) features that were sensitive to differences in imaging instrument or data collection scheme. The final feature list comprises 160 features, i.e., TAS ($n = 140$), morphology ($n = 18$; Boland and Murphy, 2001), and texture (radial moment and angular second moment; Haralick et al., 1973), that were derived exclusively from the EGFP channel. A spot detection script implemented in PerkinElmer Acapella software was used. All feature data tables and a description of feature calculation is available on GitHub (https://github.com/DWALab/Schormann-et-al).

Using the images of NMuMG cells expressing the landmarks, the following steps were taken to select features most relevant for the final classification.

We first retained only features pertaining to the EGFP channel to eliminate influence of nuclei-derived features on protein localization. Further, only texture TAS and shape features were retained to minimize the effects of protein expression on the final classification.

To minimize variation due to different microscopes, a random forests classifier was created to output features that can separate individual landmark cell images based on which microscope they were imaged on. This was done for each landmark separately. We then computed the frequency for each feature, i.e., the number of times the same feature was important in separating the microscopes across landmarks. Features that were important to separate microscopes for at least five landmarks were removed.

Correlations between all features and cell intensity were computed. Features with correlation values between −0.5 to 0.5 were retained.

A minimum set was identified from the remaining features using a "leave one feature out" test of classification accuracy. For each feature, a classifier was built in which that feature was omitted. Features were retained if the classification accuracy decreased when the feature was removed.

Intensity preprocessing

For further data processing and final classification, only features from the EGFP channel were used. Cells with <100 intensity units of EGFP expression were eliminated from the analyses. The threshold value of 100 intensity units was obtained empirically by manually thresholding 20 randomly selected EGFP channel images into foreground and background regions. All the features were scaled to have zero mean and unit SD.

Further, for each landmark and mutant dataset, cells with integrated total intensity in the top or bottom fifth percentile were removed. The remaining 789,011 and 523,319 validated landmark cell images of NMuMG and MCF10A cells, respectively, constitute the reference libraries uploaded to the Image Data Resource (idr0072; https://idr.openmicroscopy.org) and have been made available as a resource for general use.

Quantify effects of protein expression

To determine whether differences in the expression levels of the landmark proteins contributed to the localizations assigned by the random forests classifier, landmark images were divided into four different classes based on intensity quartiles (0–25, 25–50, 50–75, and 75–100 quartiles). Cross-validation was performed by training on the landmarks from a quartile bin and using the other quartile bins as queries (Fig. 3, quartile 1–quartile 4). When assessed for all quartiles, the landmarks are classified with high accuracy irrespective of the quartile fraction used for training, thus independent of protein expression levels. To further verify that assignment of localization was not due to differences in intensity, especially for proteins ostensibly targeted to the same organelle, images were randomly selected for training and classification in which the query and landmark had similar distributions of intensities (see below).

For each mutant, thousands of cells were imaged from independent experiments performed on multiple days. Further cells were deliberately imaged from different plate locations. Pooling training set cell images from independent experiments and from different automated microscopes reduces the impact of minor fluctuations between experiments. After filtering cell images for focus, segmentation, and expression artifacts, the random forests classifier created using the reference library was then used to classify images of cells not used in training to individual landmarks (Breiman, 2001).

Classification of the cell images

Random forest classifiers were used to separate the landmarks. Code scripting was performed in the MATLAB environment (R2012b, MathWorks,) including following the RF package (https://github.com/ajaiantilal/randomforest-matlab/tree/master/RF_Class_C). All codes are downloadable from GitHub (https://github.com/DWALab/Schormann-et-al). The classifier was generated setting the "mtry" variable to 12 (approximate square root of the total number of features). The "mtry" variable specifies the number of features that are available for the classifier at each split point when building decision trees. The number of trees was set to 500. Due to differences in the number of cell images for each landmark, we randomly selected 3,200 (70% of Calr-KDEL, the landmark with the smallest number of cells) cell images per landmark for training. The rest of the data were used for testing. To avoid bias due to random sampling, five different random forest classifiers were used, each with randomly sampled data from the landmarks. Every unknown cell image was classified using all five classifiers. Each classifier assigns a cell to a specific landmark. For any cell, the final class was decided using the mode of the predicted class among the five different classifiers.

To establish the extent of variation between classifications, 20 classification runs were performed on mitochondrial and CVN mutants. The result is presented as mean values (percentage classified) including SD (see Table S1). In general, the variation in percentage of cells assigned to a particular location was <1.

Classification of cell images within intensity range

For control experiments using cells with a defined range of intensities, cells were automatically selected from the image dataset based on total cellular EGFP intensity. The intensity ranges used were 300–500 and 200–1,000 arbitrary units for mitochondrial and CVN tail-anchored mutants, respectively. To display the intensity profiles as continuous lines, the cell intensities were fit to a density curve (Fig. 7). Comparison of the classification results confirmed that intensity differences did not contribute significantly (Fig. 7 and Fig. S5).

Visualization of the spatial relationships of the landmark cell images in 160-dimensional space

We used the t-SNE algorithm (Van der Maaten and Hinton, 2008) to reduce the multidimensional data into a two-dimensional plot using code obtained from the author's website (Van der Maaten and Hinton, 2008). We set the perplexity value to 10 and used 1,000 iterations to generate the two-dimensional representation. Due to the limitations of the algorithm regarding memory and computation cost, we did not use all the data for visualization. For each landmark only the 200 cells nearest to the median of the landmark were used for visualization. This resulted in a total of 3,400 (murine landmark library) and 4,000 (human landmark library) data points for visualization.

Naming system

When organelle names are used, it is to designate multiple landmarks in aggregate. The following labeling was used: ER corresponds to the landmarks Cytb5, Calr-KDEL, and Bik; Golgi apparatus corresponds to GalT and Golgin84; endosome corresponds to Rab5 and Rab7; MAM is represented by PTDSS1; mitochondria correspond to MAO and CCO; and ERGIC is represented by ERGIC53.

Motif visualization

Visualization of amino acid motifs was performed using pLogo plots (O'Shea et al., 2013). The foreground sequences were all the sequences that belonged to the cluster, while the background sequences were all the rest of the mutants generated by random mutagenesis with a CTS five amino acids long.

PWM

As an alternative approach to predict the localization of an unknown sequence of amino acids, a PWM was computed for the amino acid sequences of the CTS for mutants assigned to a specific localization. A foreground-normalized frequency matrix was first computed by calculating for each position the frequency of each amino acid at that position. If an amino acid was not present at a given position, its probability would be zero and would bring the likelihood computation to zero. To avoid this, a pseudocount value of 0.25 was added to the raw frequency for all amino acids for all positions. Thus, the frequency matrix was divided by the number of sequences in the cluster plus a value of 5 (0.25 × 20), resulting in a normalized frequency matrix. The background-normalized frequency matrix was calculated for all the sequences that were generated using random mutagenesis with a five–amino acids CTS. In this case, we did not use pseudocounts as every amino acid was represented at least once at every position. The foreground-normalized frequency matrix was then divided by the background frequency matrix. The log of the resultant gives the PWM.

The simple rules and PWMs for predicting mitochondrial localization were tested on the database of sequences having five amino acids in the CTS. The list of predicted localizations based on the simple rule or PWM was compared with the "true assignments" based on image classification. The PPV, sensitivity, and specificity were calculated from the predicted and true assignments.

Statistical enrichment

All enrichments were computed by fitting a hypergeometric distribution. For all calculations (added hydrophobic core, positive charges, and negative charges) except enrichment of actual length of the CTS, only sequences with five amino acids in the CTS were used. All of the sequences were used to calculate the enrichment for the lengths of the CTS.

For pLogo, all the sequences were fed through their website (O'Shea et al., 2013). Background sequences were computed using all the sequences of length five that were generated. Sequences classified as individual landmark were used as foreground sequences. The following equation was used to determine the height of any amino acid at a specific position:

$$\text{Residue Height }(K,N,p) = -\,log\frac{\Pr(k\ \forall k\ \geq K\mid N,p)}{\Pr(k\ \forall k\ \leq K\mid N,p)},$$

where K is the number of residues of a given type at specific position, N is total number of residues at specific position, and p is the probability of a residue at a given residue computed from background sequences. The probabilities are defined as below:

$$Pr(k\cdot\forall k\cdot\ \geq K\mid N,p)=\sum_{k=K}^{N}binomial(k,N,p)$$

$$Pr(k\cdot\forall k\cdot\ \leq K\mid N,p)=\sum_{k=0}^{K}binomial(k,N,p)$$

Online supplemental material

Fig. S1 illustrates the image processing pipeline used and sample images. Fig. S2 provides Euclidean distances between centroids of the landmarks in NMuMG cells as a colored heat map and as a table of values. Fig. S3 provides heatmaps of the amino acid enrichments associated with different localizations for EGFP-TA mutants. Fig. S4 provides sample images at high magnification demonstrating that images of different mitochondrial localized proteins correctly assigned by classification care not reliably distinguished visually. Fig. S5 provides heatmaps of the localization assignments for the XXCVN mutants to different distributions within the ER. Table S1 shows mean values (including SD) of 20 classification runs of CVN and mitochondrial mutants.

Acknowledgments

The authors thank Robert Mullen for critical reading of the manuscript and Jarkko Ylanko for technical support with the high-content imaging.

Funding for this research was provided by a foundation grant from the Canadian Institutes of Health Research (FDN 143312) to D.W. Andrews. D.W. Andrews holds the Tier 1 Canada Research Chair in Membrane Biogenesis.

The authors declare no competing financial interests.

Author contributions: All authors participated in experimental design and data interpretation. W. Schormann and S. Hariharan performed experiments, imaged the cells, and analyzed the data. W. Schormann generated all of the mutants and cell lines. S. Hariharan carried out programming for and automated image analysis and informatics analysis. D.W. Andrews supervised the project. All authors participated in writing, reviewing, and editing the manuscript.

Submitted: 16 April 2019
Revised: 30 September 2019
Accepted: 15 December 2019

References

Boland, M.V., and R.F. Murphy. 2001. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics.* 17:1213–1223. https://doi.org/10.1093/bioinformatics/17.12.1213

Bray, M.-A., A.N. Fraser, T.P. Hasaka, and A.E. Carpenter. 2012. Workflow and metrics for image quality control in large-scale high-content screens. *J. Biomol. Screen.* 17:266–274. https://doi.org/10.1177/1087057111420292

Breiman, L. 2001. Random Forests. *Mach. Learn.* 45:5–32. https://doi.org/10.1023/A:1010933404324

Bucci, C., P. Thomsen, P. Nicoziani, J. McCarthy, and B. van Deurs. 2000. Rab7: a key to lysosome biogenesis. *Mol. Biol. Cell.* 11:467–480. https://doi.org/10.1091/mbc.11.2.467

Chavrier, P., R.G. Parton, H.P. Hauri, K. Simons, and M. Zerial. 1990. Localization of low molecular weight GTP binding proteins to exocytic and endocytic compartments. *Cell.* 62:317–329. https://doi.org/10.1016/0092-8674(90)90369-P

Chen, Y.A., and R.H. Scheller. 2001. SNARE-mediated membrane fusion. *Nat. Rev. Mol. Cell Biol.* 2:98–106. https://doi.org/10.1038/35052017

Chong, Y.T., J.L.Y. Koh, H. Friesen, S.K. Duffy, M.J. Cox, A. Moses, J. Moffat, C. Boone, and B.J. Andrews. 2015. Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell.* 161:1413–1424. https://doi.org/10.1016/j.cell.2015.04.051

Collins, T.J., J. Ylanko, F. Geng, and D.W. Andrews. 2015. A Versatile Cell Death Screening Assay Using Dye-Stained Cells and Multivariate Image Analysis. *Assay Drug Dev. Technol.* 13:547–557. https://doi.org/10.1089/adt.2015.661

Conrad, C., H. Erfle, P. Warnat, N. Daigle, T. Lörch, J. Ellenberg, R. Pepperkok, and R. Eils. 2004. Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.* 14:1130–1136. https://doi.org/10.1101/gr.2383804

Cory, S., and J.M. Adams. 2002. The Bcl2 family: regulators of the cellular life-or-death switch. *Nat. Rev. Cancer.* 2:647–656. https://doi.org/10.1038/nrc883

Costello, J.L., I.G. Castro, F. Camões, T.A. Schrader, D. McNeall, J. Yang, E.-A. Giannopoulou, S. Gomes, V. Pogenberg, N.A. Bonekamp, et al. 2017. Predicting the targeting of tail-anchored proteins to subcellular compartments in mammalian cells. *J. Cell Sci.* 130:1675–1687. https://doi.org/10.1242/jcs.200204

D'Arrigo, A., E. Manera, R. Longhi, and N. Borgese. 1993. The specific subcellular localization of two isoforms of cytochrome b5 suggests novel targeting pathways. *J. Biol. Chem.* 268:2802–2808.

de Champlain, J., R.A. Mueller, and J. Axelrod. 1969. Subcellular localization of monoamine oxidase in rat tissues. *J. Pharmacol. Exp. Ther.* 166:339–345.

Diao, A., D. Rahman, D.J.C. Pappin, J. Lucocq, and M. Lowe. 2003. The coiled-coil membrane protein golgin-84 is a novel rab effector required for Golgi ribbon formation. *J. Cell Biol.* 160:201–212. https://doi.org/10.1083/jcb.200207045

Falcón-Pérez, J.M., R. Nazarian, C. Sabatti, and E.C. Dell'Angelica. 2005. Distribution and dynamics of Lamp1-containing endocytic organelles in fibroblasts deficient in BLOC-3. *J. Cell Sci.* 118:5243–5255. https://doi.org/10.1242/jcs.02633

Fischer von Mollard, G., B. Stahl, A. Khokhlatchev, T.C. Südhof, and R. Jahn. 1994. Rab3C is a synaptic vesicle protein that dissociates from synaptic vesicles after stimulation of exocytosis. *J. Biol. Chem.* 269:10971–10974.

Fliegel, L., K. Burns, D.H. MacLennan, R.A. Reithmeier, and M. Michalak. 1989. Molecular cloning of the high affinity calcium-binding protein (calreticulin) of skeletal muscle sarcoplasmic reticulum. *J. Biol. Chem.* 264:21522–21528.

Germain, M., J.P. Mathai, and G.C. Shore. 2002. BH-3-only BIK functions at the endoplasmic reticulum to stimulate cytochrome c release from mitochondria. *J. Biol. Chem.* 277:18053–18060. https://doi.org/10.1074/jbc.M201235200

Gould, S.J., G.A. Keller, N. Hosken, J. Wilkinson, and S. Subramani. 1989. A conserved tripeptide sorts proteins to peroxisomes. *J. Cell Biol.* 108:1657–1664. https://doi.org/10.1083/jcb.108.5.1657

Grote, E., J.C. Hao, M.K. Bennett, and R.B. Kelly. 1995. A Targeting Signal in VAMP Regulating Transport to Synaptic Vesicles. *Cell.* 81:581–589. https://doi.org/10.1016/0092-8674(95)90079-9

Hamilton, N.A., R.S. Pantelic, K. Hanson, and R.D. Teasdale. 2007. Fast automated cell phenotype image classification. *BMC Bioinformatics.* 8:110. https://doi.org/10.1186/1471-2105-8-110

Haralick, R., K. Shanmugam, and I. Dinstein. 1973. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* SMC-3:610–621. https://doi.org/10.1109/TSMC.1973.4309314

Henderson, M.P., Y.T. Hwang, J.M. Dyer, R.T. Mullen, and D.W. Andrews. 2007. The C-terminus of cytochrome b5 confers endoplasmic reticulum specificity by preventing spontaneous insertion into membranes. *Biochem. J.* 401:701–709. https://doi.org/10.1042/BJ20060990

Hong, W. 2005. SNAREs and traffic. *Biochim. Biophys. Acta.* 1744:120–144. https://doi.org/10.1016/j.bbamcr.2005.03.014

Hopert, A., C.C. Uphoff, M. Wirth, H. Hauser, and H.G. Drexler. 1993. Specificity and sensitivity of polymerase chain reaction (PCR) in comparison with other methods for the detection of mycoplasma contamination in cell lines. *J. Immunol. Methods.* 164:91–100. https://doi.org/10.1016/0022-1759(93)90279-G

Horie, C., H. Suzuki, M. Sakaguchi, and K. Mihara. 2002. Characterization of signal that directs C-tail-anchored proteins to mammalian mitochondrial outer membrane. *Mol. Biol. Cell.* 13:1615–1625. https://doi.org/10.1091/mbc.01-12-0570

Horwich, A.L., F. Kalousek, W.A. Fenton, R.A. Pollock, and L.E. Rosenberg. 1986. Targeting of pre-ornithine transcarbamylase to mitochondria: definition of critical regions and residues in the leader peptide. *Cell.* 44:451–459. https://doi.org/10.1016/0092-8674(86)90466-6

Huang, K., and R.F. Murphy. 2004. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics.* 5:78. https://doi.org/10.1186/1471-2105-5-78

Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132. https://doi.org/10.1016/0022-2836(82)90515-0

Levine, J.H., E.F. Simonds, S.C. Bendall, K.L. Davis, A.D. Amir, M.D. Tadmor, O. Litvin, H.G. Fienberg, A. Jager, E.R. Zunder, et al. 2015. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell.* 162:184–197. https://doi.org/10.1016/j.cell.2015.05.047

Li, J., L. Xiong, J. Schneider, and R.F. Murphy. 2012. Protein subcellular location pattern classification in cellular images using latent discriminative models. *Bioinformatics.* 28:i32–i39. https://doi.org/10.1093/bioinformatics/bts230

Marty, N.J., H.J. Teresinski, Y.T. Hwang, E.A. Clendening, S.K. Gidda, E. Sliwinska, D. Zhang, J.A. Miernyk, G.C. Brito, D.W. Andrews, et al. 2014. New insights into the targeting of a subset of tail-anchored proteins to the outer mitochondrial membrane. *Front. Plant Sci.* 5:426. https://doi.org/10.3389/fpls.2014.00426

Mills, C.E., C. Thome, D. Koff, D.W. Andrews, and D.R. Boreham. 2015. The relative biological effectiveness of low-dose mammography quality X rays in the human breast MCF-10A cell line. *Radiat. Res.* 183:42–51. https://doi.org/10.1667/RR13821.1

Munro, S., and H.R. Pelham. 1987. A C-terminal signal prevents secretion of luminal ER proteins. *Cell.* 48:899–907. https://doi.org/10.1016/0092-8674(87)90086-9

Murdoch, B., D.S. Pereira, X. Wu, J.E. Dick, and J. Ellis. 1997. A rapid screening procedure for the identification of high-titer retrovirus packaging clones. *Gene Ther.* 4:744–749. https://doi.org/10.1038/sj.gt.3300448

Nanni, L., and A. Lumini. 2008. A reliable method for cell phenotype image classification. *Artif. Intell. Med.* 43:87–97. https://doi.org/10.1016/j.artmed.2008.03.005

O'Shea, J.P., M.F. Chou, S.A. Quader, J.K. Ryan, G.M. Church, and D. Schwartz. 2013. pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods.* 10:1211–1212. https://doi.org/10.1038/nmeth.2646

Pärnamaa, T., and L. Parts. 2017. Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3 (Bethesda).* 7:1385–1392. https://doi.org/10.1534/g3.116.033654

Pfaff, J., J. Rivera Monroy, C. Jamieson, K. Rajanala, F. Vilardi, B. Schwappach, and R.H. Kehlenbach. 2016. Emery-Dreifuss muscular dystrophy mutations impair TRC40-mediated targeting of emerin to the inner nuclear membrane. *J. Cell Sci.* 129:502–516. https://doi.org/10.1242/jcs.179333

Rao, M., V. Okreglak, U.S. Chio, H. Cho, P. Walter, and S. Shan. 2016. Multiple selection filters ensure accurate tail-anchored membrane protein targeting. *eLife.* 5:e21301. https://doi.org/10.7554/eLife.21301

Rapaport, D. 2003. Finding the right organelle. Targeting signals in mitochondrial outer-membrane proteins. *EMBO Rep.* 4:948–952. https://doi.org/10.1038/sj.embor.embor937

Raptis, A., B. Torrejón-Escribano, I. Gómez de Aranda, and J. Blasi. 2005. Distribution of synaptobrevin/VAMP 1 and 2 in rat brain. *J. Chem. Neuroanat.* 30:201–211. https://doi.org/10.1016/j.jchemneu.2005.08.002

Rizzuto, R., M. Brini, P. Pizzo, M. Murgia, and T. Pozzan. 1995. Chimeric green fluorescent protein as a tool for visualizing subcellular organelles in living cells. *Curr. Biol.* 5:635–642. https://doi.org/10.1016/S0960-9822(95)00128-X

Roth, J., and E.G. Berger. 1982. Immunocytochemical localization of galactosyltransferase in HeLa cells: codistribution with thiamine pyrophosphatase in trans-Golgi cisternae. *J. Cell Biol.* 93:223–229. https://doi.org/10.1083/jcb.93.1.223

Scaffidi, P., and T. Misteli. 2008. Lamin A-dependent misregulation of adult stem cells associated with accelerated ageing. *Nat. Cell Biol.* 10:452–459. https://doi.org/10.1038/ncb1708

Scheel, J., R. Pepperkok, M. Lowe, G. Griffiths, and T.E. Kreis. 1997. Dissociation of coatomer from membranes is required for brefeldin A-induced transfer of Golgi enzymes to the endoplasmic reticulum. *J. Cell Biol.* 137:319–333. https://doi.org/10.1083/jcb.137.2.319

Schlierf, B., G.H. Fey, J. Hauber, G.M. Hocke, and O. Rosorius. 2000. Rab11b is essential for recycling of transferrin to the plasma membrane. *Exp. Cell Res.* 259:257–265. https://doi.org/10.1006/excr.2000.4947

Schröder, K., B. Martoglio, M. Hofmann, C. Hölscher, E. Hartmann, S. Prehn, T.A. Rapoport, and B. Dobberstein. 1999. Control of glycosylation of MHC class II-associated invariant chain by translocon-associated RAMP4. *EMBO J.* 18:4804–4815. https://doi.org/10.1093/emboj/18.17.4804

Scott, I., and R.J. Youle. 2010. Mitochondrial fission and fusion. *Essays Biochem.* 47:85–98. https://doi.org/10.1042/bse0470085

Sharpe, H.J., T.J. Stevens, and S. Munro. 2010. A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell.* 142:158–169. https://doi.org/10.1016/j.cell.2010.05.037

Skene, J.H., and I. Virág. 1989. Posttranslational membrane attachment and dynamic fatty acylation of a neuronal growth cone protein, GAP-43. *J. Cell Biol.* 108:613–624. https://doi.org/10.1083/jcb.108.2.613

Stone, S.J., and J.E. Vance. 2000. Phosphatidylserine synthase-1 and -2 are localized to mitochondria-associated membranes. *J. Biol. Chem.* 275: 34534–34540. https://doi.org/10.1074/jbc.M002865200

Sullivan, D.P., C.F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren, et al. 2018. Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* 36:820–828. https://doi.org/10.1038/nbt.4225

Van der Maaten, L.J.P., and G.E. Hinton. 2008. Visualizing high-dimensional data using t-sne. *J. Mach. Learn. Res.* 9:2579–2605. https://doi.org/10.1007/s10479-011-0841-3

Xu, Y.-Y., F. Yang, Y. Zhang, and H.-B. Shen. 2015. Bioimaging-based detection of mislocalized proteins in human cancers by semi-supervised learning. *Bioinformatics.* 31:1111–1119. https://doi.org/10.1093/bioinformatics/btu772

Xu, Y.-Y., F. Yang, and H.-B. Shen. 2016. Incorporating organelle correlations into semi-supervised learning for protein subcellular localization prediction. *Bioinformatics.* 32:2184–2192. https://doi.org/10.1093/bioinformatics/btw219

Xu, Y.-Y., H.-B. Shen, and R.F. Murphy. 2019. Learning complex subcellular distribution patterns of proteins via analysis of immunohistochemistry images. *Bioinformatics.*:btz844. https://doi.org/10.1093/bioinformatics/btz844

Zernike, F. 1934. Beugungstheorie des schneidenverfarhens und seiner verbesserten form, der phasenkontrastmethode. *Physica.* 1:689–704. https://doi.org/10.1016/S0031-8914(34)80259-5

Zhu, W., A. Cowie, G.W. Wasfy, L.Z. Penn, B. Leber, and D.W. Andrews. 1996. Bcl-2 mutants with restricted subcellular location reveal spatially distinct pathways for apoptosis in different cell types. *EMBO J.* 15: 4130–4141. https://doi.org/10.1002/j.1460-2075.1996.tb00788.x

Supplemental material

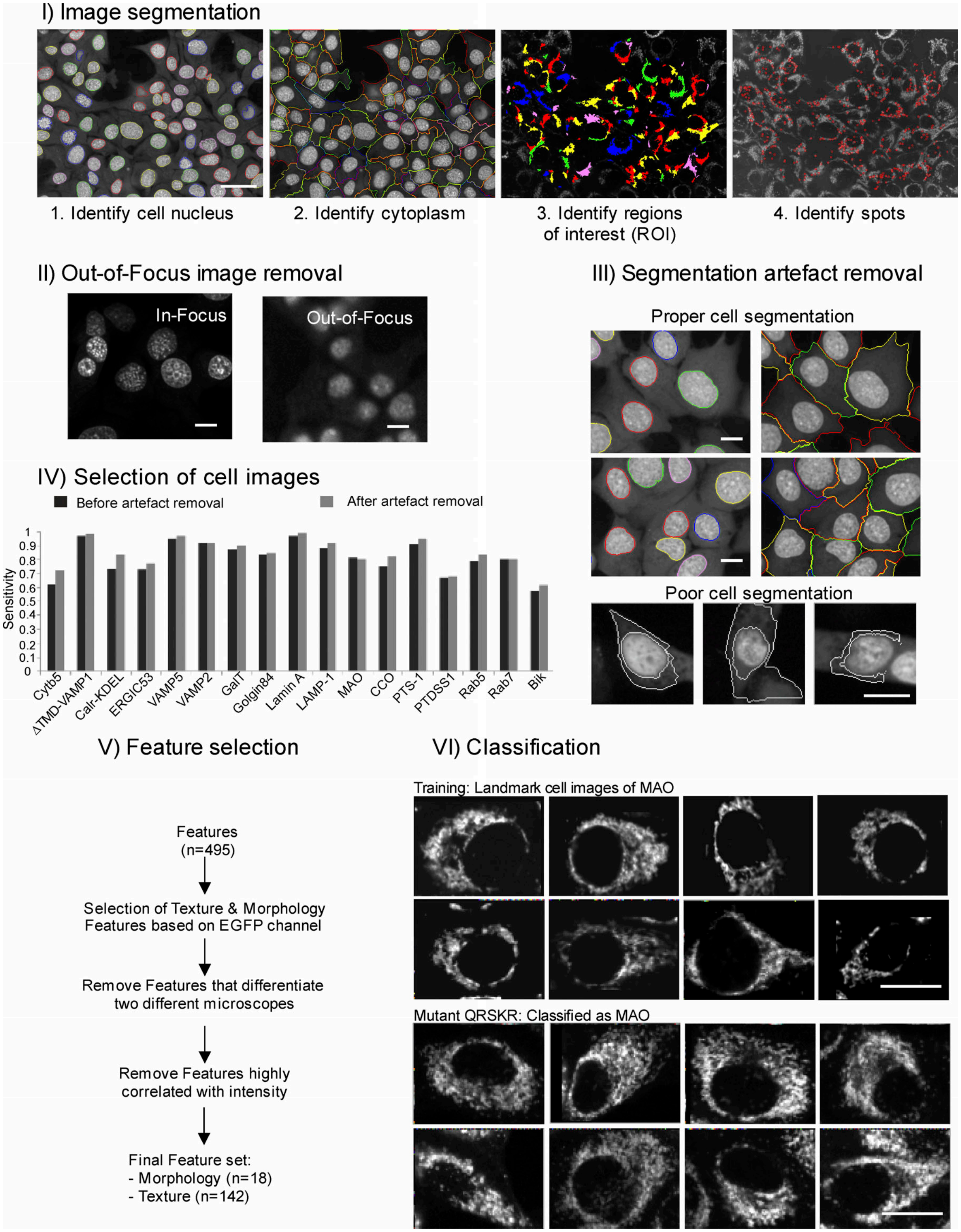
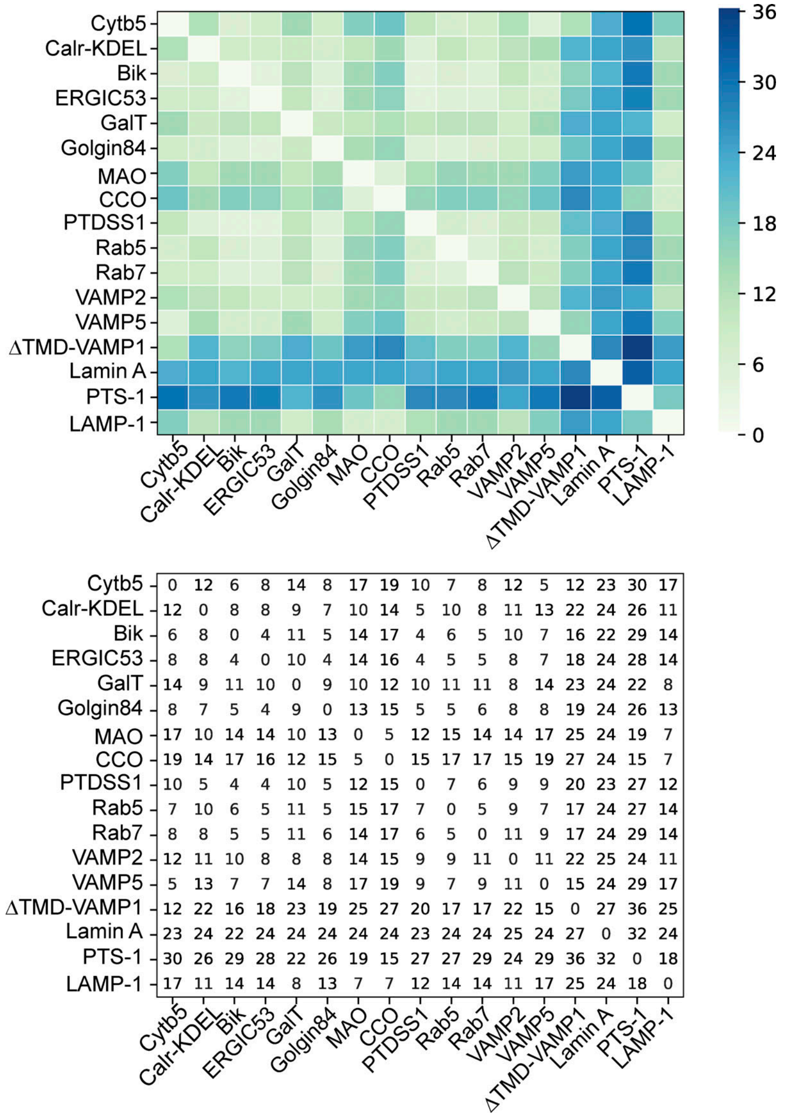


Figure S1. **Image processing pipeline. (I)** Image segmentation of cell images (scale bar, 25 µm) includes identifying the nucleus, cytoplasm, ROI, and spots. Acapella software by Perkin Elmer was used for all steps. **(II and III)** Prior to feature extraction, all images undergo a quality control process to (II) remove out-of-focus images by using a specific classifier (see Materials and methods) and (III) remove incorrectly segmented cell images by computing two parameters (see equations for R1 and R2 in Materials and methods). Representative images of proper cell segmentation (nuclei and cytoplasm) and poor cell segmentation (scale bars, 15 µm). **(IV)** Selection of optimal cell images improves classification sensitivity. **(V)** Final feature calculation included 160 morphology and texture features. For random forests classification, only features calculated from the EGFP channel were used. **(VI)** Sample training images and images of cells classified to the training landmark MAO (outer mitochondrial membrane). Scale bars, 15 µm.



	Cytb5	Calr-KDEL	Bik	ERGIC53	GalT	Golgin84	MAO	CCO	PTDSS1	Rab5	Rab7	VAMP2	VAMP5	ΔTMD-VAMP1	Lamin A	PTS-1	LAMP-1
Cytb5	0	12	6	8	14	8	17	19	10	7	8	12	5	12	23	30	17
Calr-KDEL	12	0	8	8	9	7	10	14	5	10	8	11	13	22	24	26	11
Bik	6	8	0	4	11	5	14	17	4	6	5	10	7	16	22	29	14
ERGIC53	8	8	4	0	10	4	14	16	4	5	5	8	7	18	24	28	14
GalT	14	9	11	10	0	9	10	12	10	11	11	8	14	23	24	22	8
Golgin84	8	7	5	4	9	0	13	15	5	5	6	8	8	19	24	26	13
MAO	17	10	14	14	10	13	0	5	12	15	14	14	17	25	24	19	7
CCO	19	14	17	16	12	15	5	0	15	17	17	15	19	27	24	15	7
PTDSS1	10	5	4	4	10	5	12	15	0	7	6	9	9	20	23	27	12
Rab5	7	10	6	5	11	5	15	17	7	0	5	9	7	17	24	27	14
Rab7	8	8	5	5	11	6	14	17	6	5	0	11	9	17	24	29	14
VAMP2	12	11	10	8	8	8	14	15	9	9	11	0	11	22	25	24	11
VAMP5	5	13	7	7	14	8	17	19	9	7	9	11	0	15	24	29	17
ΔTMD-VAMP1	12	22	16	18	23	19	25	27	20	17	17	22	15	0	27	36	25
Lamin A	23	24	22	24	24	24	24	24	23	24	24	25	24	27	0	32	24
PTS-1	30	26	29	28	22	26	19	15	27	27	29	24	29	36	32	0	18
LAMP-1	17	11	14	14	8	13	7	7	12	14	14	11	17	25	24	18	0

Figure S2. **Euclidean distances between centroids of the landmarks in NMuMG cells.** Top panel: Heat map of distances colored as indicated to the right. Bottom panel: Numerical data. Distances are unitless because they represent Z-scored multidimensional data but can be thought of as the number of SDs between landmarks.

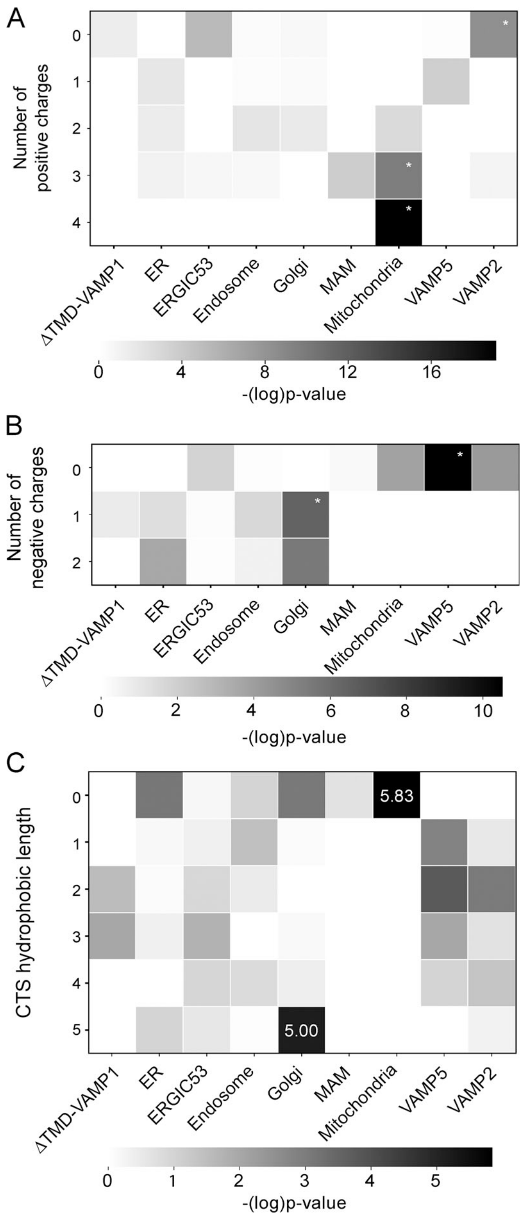


Figure S3. **Amino acid enrichment within CTS. (A)** Statistical enrichment for the number of positive charges in the CTS of EGFP-TA mutants assigned to different localizations. A hypergeometric distribution was used to calculate statistical enrichment. Asterisks highlight statistical significance at α 0.05 after Bonferroni correction. The columns indicate individual landmarks or organelles defined by merging landmarks as ΔTMD-VAMP1 (whole cell), ER (Cytb5, Calr-KDEL, Bik), ERGIC (ERGIC53), endosome (Rab5, Rab7), Golgi (Golgin84, GalT), mitochondria (MAO, CCO), MAM (PTDSS1), VAMP5, and VAMP2. **(B)** Statistical enrichment in number of negative charges within the CTS for EGFP-TA mutants assigned as localized at the indicated organelles. **(C)** Statistical enrichment for hydrophobic length added by amino acids in the CTS for EGFP-TA mutants assigned as localized at the indicated organelles. Hydrophobic length is the number of hydrophobic amino acids in the CTS before the first hydrophilic amino acid. Numbers indicate the negative log of the P value for significant enrichments (length 0, mitochondria; length 5, Golgi). Significance cut-off for negative log of P value after Bonferroni correction is 6.98.

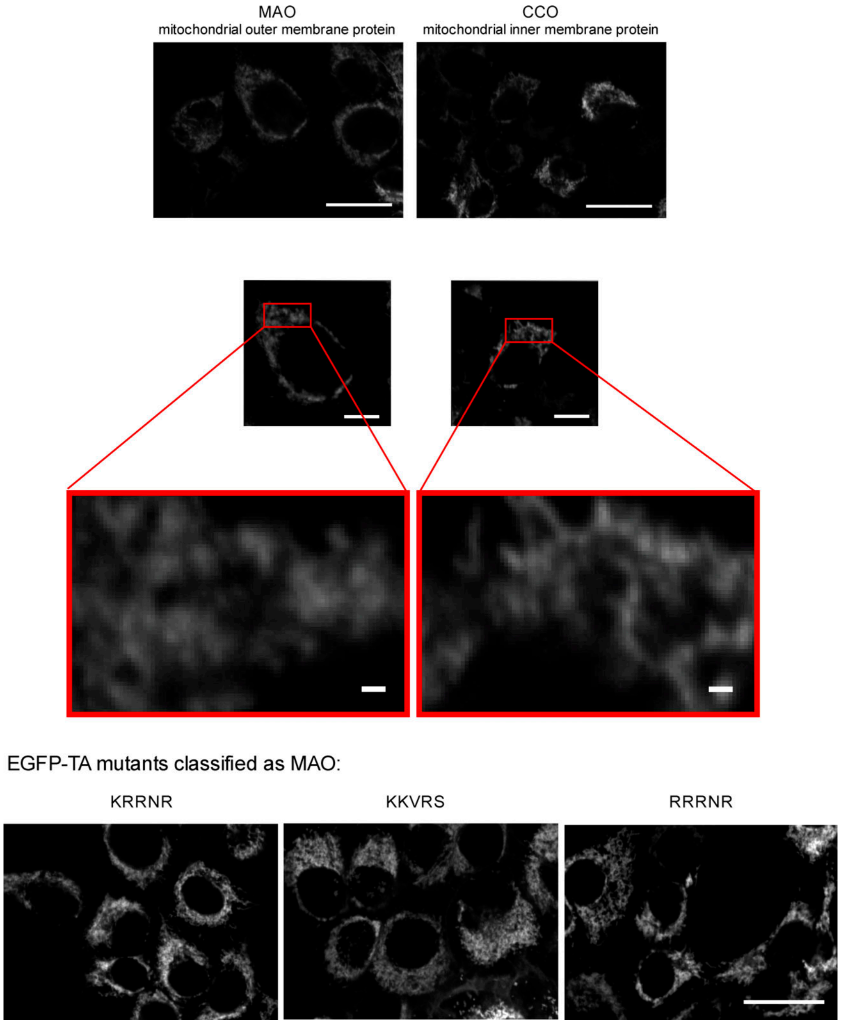

MAO
mitochondrial outer membrane protein
CCO
mitochondrial inner membrane protein
EGFP-TA mutants classified as MAO:
KRRNR
KKVRS
RRRNR

Figure S4. **Images of different mitochondrial localized proteins correctly assigned by classification are not reliably distinguishable visually.** Representative cell images (NMuMG) expressing MAO and CCO (top section, scale bars, 25 µm). Selected single-cell image of each mitochondrial landmark is magnified (middle section, scale bars, 10 µm; inset scale bars, 1 µm). Representative images of mutants classified as MAO (bottom lower section, scale bar, 25 µm).

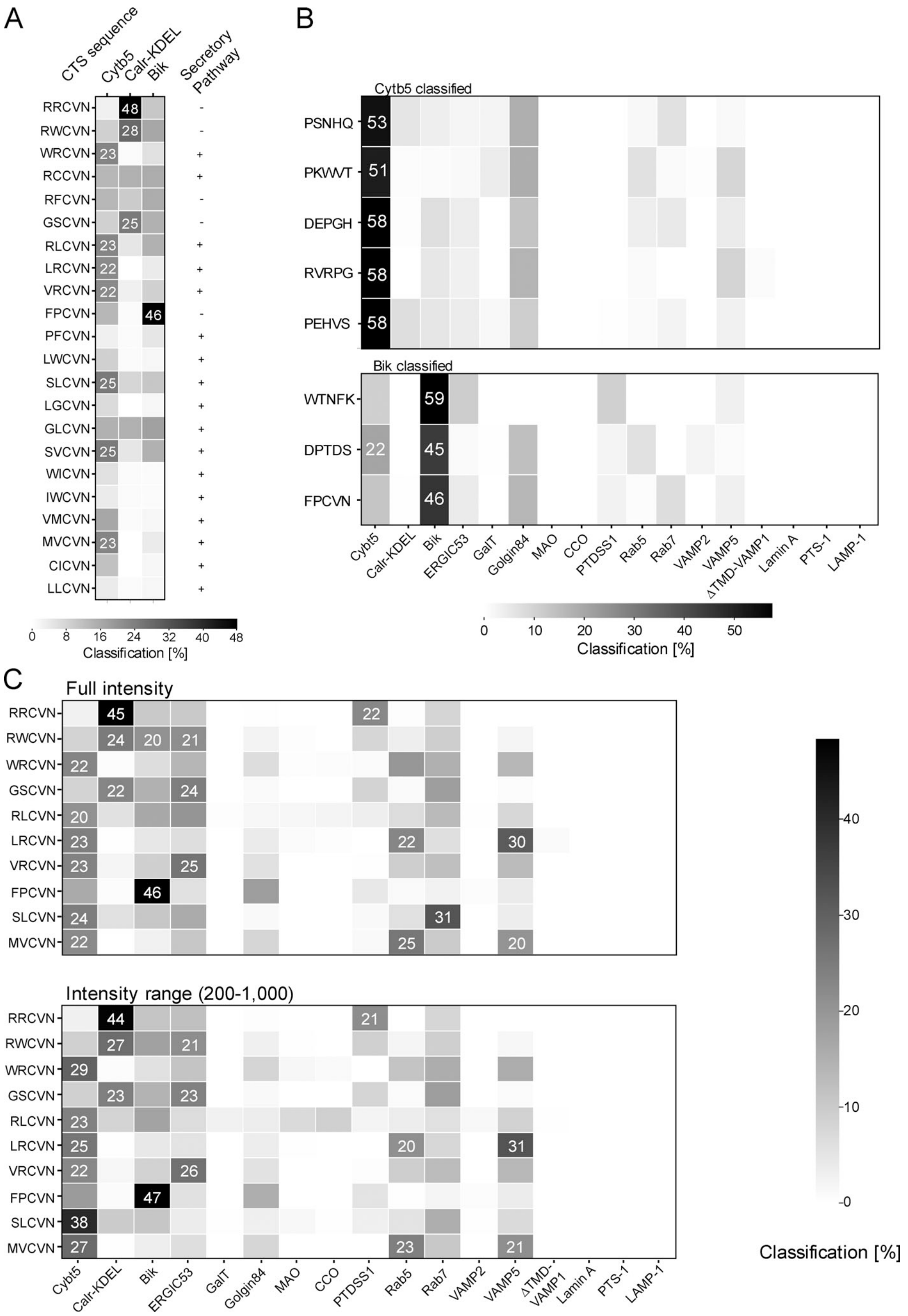


Figure S5. **Assignment of the localizations of "CVN" mutants to different distributions within the ER. (A)** Random forests classification results for "CVN" mutants to localizations defined by the landmarks Cytb5, Calr-KDEL, and Bik. Localization of at least 20% of the cell images is highlighted by numbers and gray shades. Minus (−) indicates the mutant was not localized in the post-ER secretory pathway, and plus (+) designates the mutant was also assigned localization within the post-ER secretory pathway (localization, see Fig. 8 for details). **(B)** Random forests classification results for EGFP-TA mutants that are restricted to either only Cytb5 or Bik localization. **(C)** EGFP-TA mutants classified as ER-localized independent of intensity. Random forests classification results for selected "CVN" mutants using full intensity range compared with limited range (300–500). Localization is indicated by gray shades, and numbers are provided for localizations of at least 20% of the cell images.

Provided online is one table. Table S1 shows mean values (including SD) of 20 classification runs of CVN and mitochondrial mutants.

https://doi.org/10.1083/jcb.201904090