Immunofluorescence and fluorescent-protein tagging show high correlation for protein localization in mammalian cells

Charlotte Stadler[1], Elton Rexhepaj[1], Vasanth R Singan[2,3], Robert F Murphy[4,5], Rainer Pepperkok[6], Mathias Uhlén[1], Jeremy C Simpson[2,3] & Emma Lundberg[1]

Imaging techniques such as immunofluorescence (IF) and the expression of fluorescent protein (FP) fusions are widely used to investigate the subcellular distribution of proteins. Here we report a systematic analysis of >500 human proteins comparing the localizations obtained in live versus fixed cells using FPs and IF, respectively. We identify systematic discrepancies between IF and FPs as well as between FP tagging at the N and C termini. The analysis shows that for 80% of the proteins, IF and FPs yield the same subcellular distribution, and the locations of 250 previously unlocalized proteins were determined by the overlap between the two methods. Approximately 60% of proteins localize to multiple organelles for both methods, indicating a complex subcellular protein organization. These results show that both IF and FP tagging are reliable techniques and demonstrate the usefulness of an integrative approach for a complete investigation of the subcellular human proteome.

Investigating the localization of proteins at the subcellular level is of great importance, as it leads to a better understanding of protein function, interaction networks and cellular signaling pathways. Imaging-based approaches have the advantage of providing spatial information on protein location *in situ* from single cells and can effectively address the issue of proteins that localize to multiple organelles.

A common technique is to express the target protein fused to an FP, thus enabling temporal studies of the protein's distribution in its natural environment, the living cell. However, any such engineering of a native protein carries the risk of affecting its localization. Fusion of the reporter molecule to two or more different tagging locations in each protein, such as the N and C termini, is normally needed to minimize localization artifacts[1]. In addition, FP tagging is usually carried out in cells already expressing the endogenous protein; therefore, effects arising from overexpression need to be taken into consideration.

Endogenous proteins are best visualized by the use of specific antibodies and IF in fixed cells. Although IF is widely used, there are two important complications related to this method. First, many antibodies show off-target binding due to cross-reactivity with other proteins. Second, the fixation and permeabilization of the cells can potentially cause artifacts that can affect the localization observed. In this regard, many studies have reported different results in protein distribution depending on the fixation protocol used[2–7].

To date, relatively few efforts have aimed to systematically map the subcellular localization of proteins using imaging-based techniques. In 2003, the localization of 97% of all proteins in *Saccharomyces cerevisiae* was determined through the systematic fusion of FPs to each protein[8]. A similar approach has also been applied to a fraction of the human proteome[1,9–11]. In that FP-cDNA project, mammalian cells were transiently transfected with open reading frames, and the subcellular localization of more than 1,600 proteins was determined in live cells. In the Human Protein Atlas project, IF is used to systematically map the subcellular location of the human proteome in fixed cells[12–14]. Currently, it contains subcellular localization information for 11,353 proteins obtained through the use of 12,908 unique antibodies in three human cell lines.

As IF and FP tagging are widely used experimental techniques that exhibit distinct advantages and disadvantages, we have compiled and compared subcellular localization data from more than 500 human proteins as reported by the FP-cDNA approach and the Human Protein Atlas project. We show that an integrative approach using both methods yields complementary data that together strengthen the annotation of the subcellular localization of the human proteome.

RESULTS

Localization of proteins by IF and FPs

Both N- and C-terminal tags were individually fused to 873 proteins, and the resulting FPs were expressed in Vero or HeLa cells[1].

[1]KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Stockholm, Sweden. [2]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland. [3]Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland. [4]Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [5]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [6]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany. Correspondence should be addressed to E.L. (emma.lundberg@scilifelab.se).

RECEIVED 2 DECEMBER 2012; ACCEPTED 25 JANUARY 2013; PUBLISHED ONLINE 24 FEBRUARY 2013; DOI:10.1038/NMETH.2377

Images of N- and C-terminal fusion proteins were separately annotated for subcellular localization(s) before each target protein was assigned to one main localization class: mitochondria, Golgi apparatus, plasma membrane, cytoskeleton, vesicles, nucleus, cytoplasm, endoplasmic reticulum, cytoplasm and nucleus, or Golgi and plasma membrane (**Supplementary Table 1**).

Antibodies from the Human Protein Atlas project were used in IF experiments to analyze the corresponding endogenous proteins in three human cell lines of different origin: A-431, an epidermoid carcinoma cell line; U-2 OS, an osteosarcoma cell line; and U-251MG, a glioblastoma cell line. Of the 873 proteins, 506 were studied using IF; 60 of those did not result in any staining. The remaining 446 protein localizations were characterized as 'main' (and 'additional' if more than one) on the basis of the annotations from the three cell lines (Online Methods and **Supplementary Table 2**).

IF versus FP subcellular protein localization

The localizations observed by IF and FP tagging were classified as 'identical' (one or multiple localizations observed with both methods), 'similar' (one localization observed with both methods but with additional localization(s) observed in either of the two methods) or 'dissimilar' (no common localization observed with the two methods) (Online Methods). The network in **Figure 1a** (also **Supplementary Fig. 1**) shows the distribution and overlap of protein localization obtained by the two methods, and proteins representing different cellular structures—for example, cytoplasm represented by NUDC and the endoplasmic reticulum represented by FKBP7—are shown in **Figure 1b**. For systematic comparison, the proteins were grouped into the main location classes described above, but in-depth annotations were also recorded for many proteins. For example, the RNA-binding protein PATL1 localized to cytoplasmic p-bodies, FHL2 to focal adhesion sites, EMD to the nuclear membrane and MYPN to microtubules and actin filaments (**Fig. 1b**).

The overall correlation between the FP-tagging and IF approaches was high, with 82% of all proteins sharing at least one localization. Owing to the often complex distribution of proteins to multiple locations, similar rather than identical results were seen for many proteins (**Fig. 1b**, vii). The overlap between the methods was highest for the nuclear and cytoplasmic classes, whereas the lowest correlation was seen for proteins localizing to dynamic cellular elements such as the endomembrane system and the cytoskeleton (**Fig. 1c**). There are several biological explanations for nonidentical results from IF and FP tagging in addition to the technical limitations of each method. For example, the enzyme GALNTL2 is likely to be present in different membrane components of the secretory pathway during its life cycle, and, accordingly, it localized to vesicles with IF and to the Golgi apparatus or endoplasmic reticulum and vesicles with FPs (**Fig. 1b**). A closer analysis of the proteins localized by IF in three different human cell lines revealed that only 53% were identically distributed in all three (**Supplementary Fig. 2**), which shows the extent of the possible variation in subcellular localizations. Another explanation for nonidentical results is the differential localization of protein isoforms, exemplified by the enzyme ECI2 (**Fig. 1b**): isoform 1 is mitochondrial, whereas isoform 2, which lacks the first 35 amino acids, localizes to peroxisomes. IF and C-terminal FPs detected both isoforms, but only the peroxisomal location was identified by N-terminal FP tagging, probably because the mitochondrial targeting sequence was masked[15].

Of the 446 proteins analyzed, 92 produced dissimilar results (**Fig. 1c** and **Supplementary Table 3**), and to identify systematic discrepancies between FPs and IF, we generated another network plot (**Fig. 2a** and **Supplementary Fig. 3**). Here, each protein is represented by two nodes showing the main localizations assigned by IF and FP, and these are linked by green connectors. If both methods give the same localization, the connector will be short and therefore invisible, whereas if the methods give dissimilar localizations, the green connectors will be longer. Discrepancies in assigned locations can be identified as clusters of green connectors (marked i–vi). Furthermore, when these particular nodes are grouped according to color (that is, method), this indicates a systematic bias, whereas if the colors are mixed, the bias is related to both methods.

Cluster i represents discrepancies in the localization of cytoskeletal proteins. Here, one method localized the proteins to the nucleus, whereas the other localized it to a cytoskeletal structure (**Fig. 2b**). For example, NUDCD2 localized to the cytoplasm, microtubules and centrosomes by IF as indicated by the literature, but to the nucleus and cytoplasm as indicated by both FPs.

A large proportion of the discrepancies relates to proteins of the endomembrane system, represented by clusters ii–v. Cluster ii is a mixed cluster in which proteins known to localize to vesicles or the Golgi apparatus were incorrectly localized to the nucleus by either FPs or IF (**Fig. 2b**; SEC23IP). Other clusters (iii–v) show systematic discrepancies, in which FPs localized the proteins to the endomembrane system, whereas IF shows incorrect localization to the nucleus (**Fig. 2b**; USE1 and TMEM9) or cytoplasm. These results clearly show that FPs are in general superior at localizing proteins to endomembrane structures. Unfortunately, there is also a high risk of false localization to these structures as a consequence of ectopic (over)expression of fusion proteins. Examples include DPYSL2 and the related example NDUFB8 (**Fig. 2b**), which were correctly localized to the cytoplasm and mitochondria, respectively, by IF but to the endoplasmic reticulum by FPs.

Cluster vi relates to the localization of proteins to the nucleus versus to the cytoplasm and can be subdivided into three groups: one for which IF localized the protein to the nucleus and FPs to the cytoplasm, another for which the opposite is seen and a mixed third group. The cytoplasm is usually the correct localization, whereas an observed nuclear localization is most often an artifact (**Fig. 2b**; RPS24, FSCB, PDE1A). However, as many proteins shuttle between the nucleus and cytoplasm, these discrepancies may also represent the difficulty of capturing a dynamic distribution in these static pictures.

Common among all clusters is that the most frequently incorrect localization was to the nucleus. If the target protein was not expressed, the antibody may have bound nonspecifically to a nuclear epitope in IF, as the nucleus contains a high density of proteins in proportion to the entire cellular proteome[8,12]. The nuclear localization seen in live cells might be an artifact from FP tagging of small soluble proteins that can freely diffuse through the nuclear pore and accumulate in the nucleus as a result of their true targeting sequence being masked.

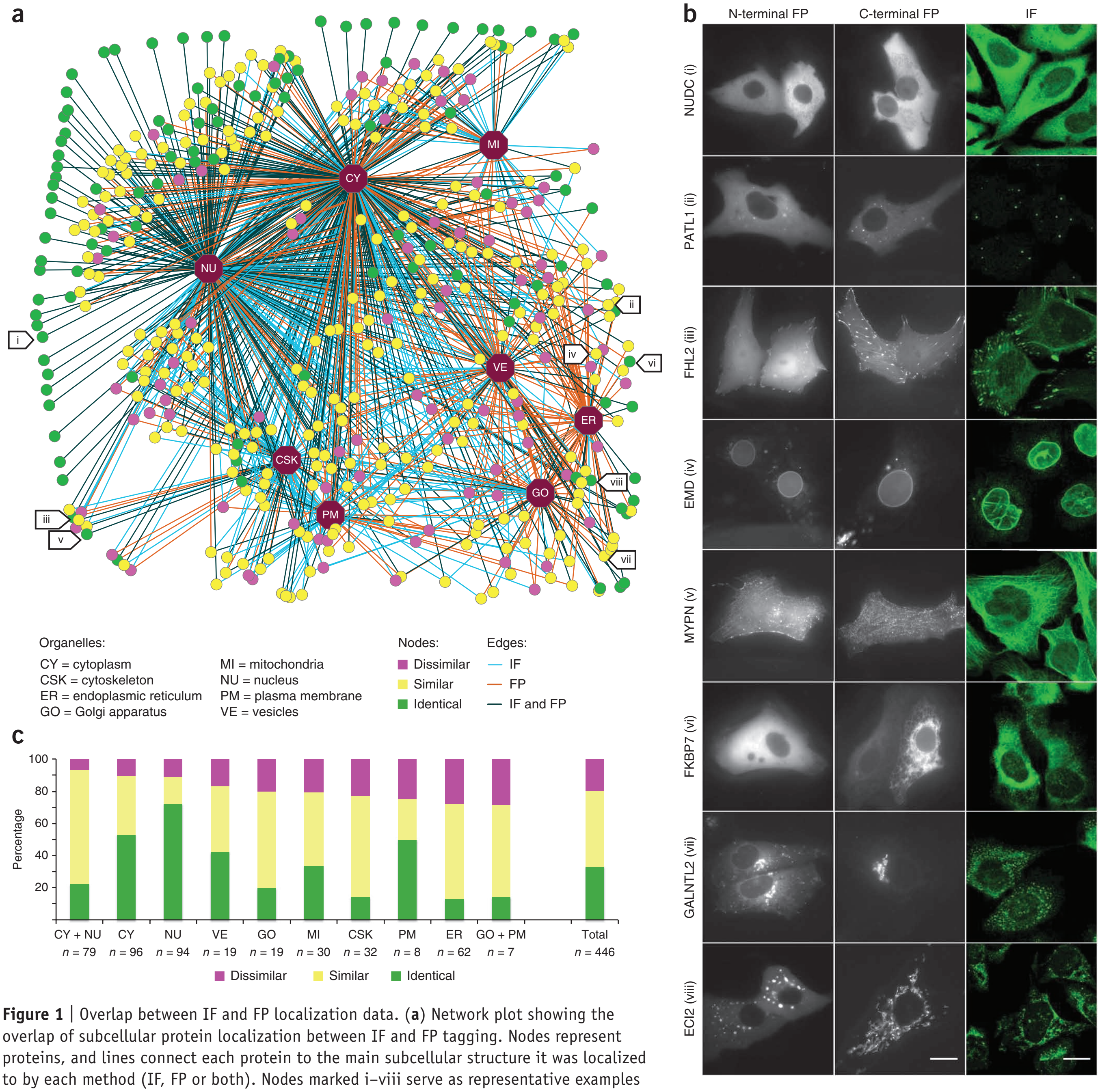


Figure 1 | Overlap between IF and FP localization data. (**a**) Network plot showing the overlap of subcellular protein localization between IF and FP tagging. Nodes represent proteins, and lines connect each protein to the main subcellular structure it was localized to by each method (IF, FP or both). Nodes marked i–viii serve as representative examples of proteins visualized in **b**. (**b**) Example fluorescence images of proteins with identical or similar localizations between IF and FPs, representing the following structures: NUCD, cytoplasm (i); PATL1, cytoplasmic p-bodies (ii); FHL2, focal adhesion sites (iii); EMD, nuclear membrane (iv); MYPN, microtubules and actin filaments (v); FKBP7, endoplasmic reticulum (ER) (vi); GALNTL2, ER, Golgi and vesicles (vii); ECI2, mitochondria and vesicles (viii). Scale bars, 10 µm. (**c**) Distribution of identical, similar and dissimilar localizations between IF and FPs for each protein main localization class. *n*, number of proteins within each class.

Multiple subcellular localizations

Discrepancies between the localizations obtained from both methods often represent a complex distribution of the corresponding protein to multiple subcellular compartments. Indeed, 60% of the proteins in this study localized to additional structures besides their main assigned location (**Fig. 2c**). This occurred with both FP tagging and IF, and it is therefore unlikely to be a result of cross-reactivity of antibodies or artifacts associated with ectopic (over)expression. Differences between the methods were seen for the proteins assigned to the cytoskeleton, for which additional locations were twice as common with IF than with FP tagging. By contrast, more multiple localizations were detected by FPs for proteins of the secretory pathway, potentially because of the increased load on the secretory machinery. For both methods, cytoplasmic or nuclear localizations were most frequently observed as the additional localization. Cytoplasmic localization is often supported by the literature, whereas the additional nuclear localizations are more likely to be an artifact seen by both methods.

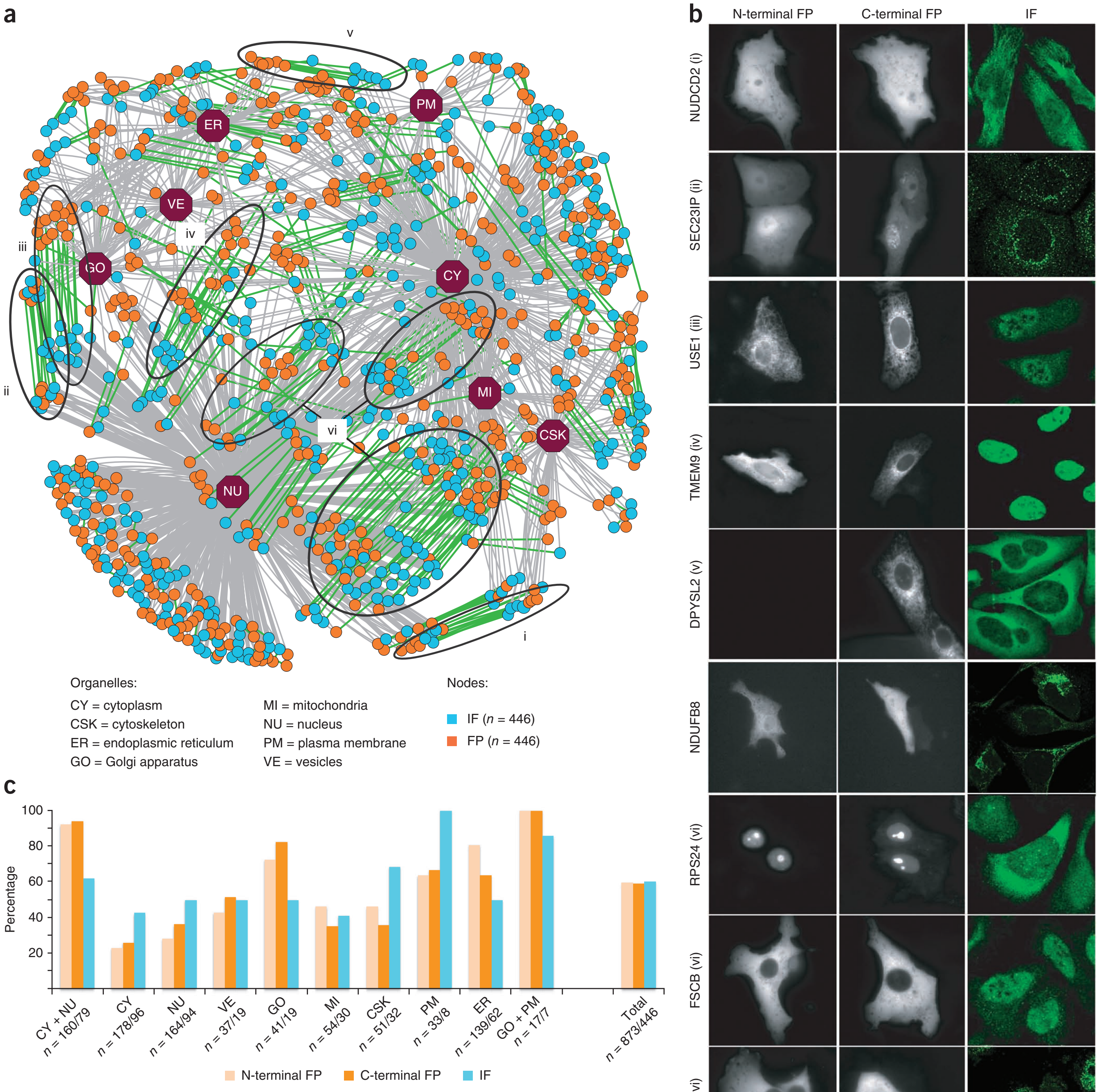


Figure 2 | Differences between IF and FP localization data. (**a**) Network plot showing the protein main localization with IF and FPs for the 446 studied proteins. Gray lines connect each protein to the subcellular structure(s) it was localized to. Green lines connect the two nodes representing the same protein to show discrepancies between the two methods. Clusters of discrepancies are circled and denoted with roman numerals. (**b**) Example fluorescence images of proteins representing the discrepancy clusters from **a**: NUDCD2 (i), SEC23IP (ii), USE1 (iii), TMEM9 (iv), DPYSL2 (v) and the related example NDUFB8, and RPS24, FSCB and PDE1A (vi). Scale bars, 10 µm. (**c**) Fraction of proteins with additional localizations in each main localization category for N- and C-terminal FP tagging and IF. *n*, number of proteins for FP and IF, respectively, in each category.

Validation of IF localization of endogenous proteins

Different approaches can be used to validate IF results (**Fig. 3**). We first used RNA sequencing to ensure expression of the target protein. This revealed that antibodies targeting 21 proteins out of the 92 with dissimilar results between the two methods are likely to be cross-reactive, as their target protein was not expressed in the cell lines used (**Fig. 3a**; FSCB and PDE1A).

RNA sequencing can also be used to support results obtained by IF across multiple cell types (**Fig. 3a**; LPCAT2). Similarly, an independent antibody that recognizes a different epitope on the target protein (**Fig. 3b**) can be used, or the target protein can be downregulated by RNA interference and the decrease in fluorescence quantified[16] (**Fig. 3c**). Colocalization experiments with IF staining of the corresponding FP-expressing cell can ensure that

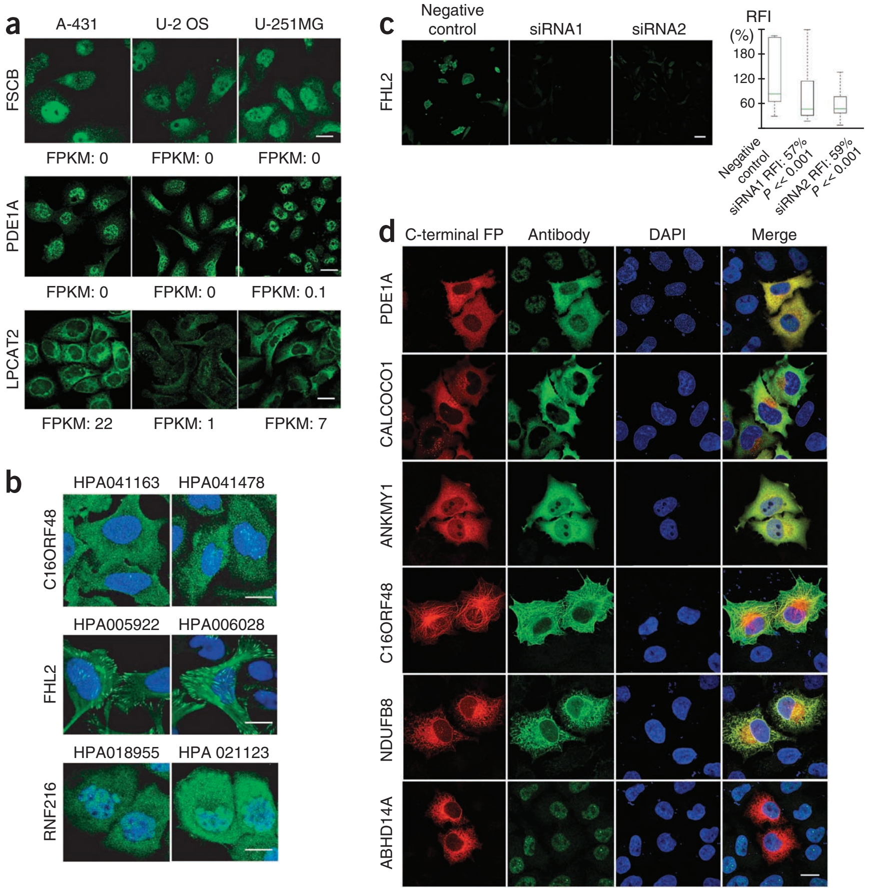


Figure 3 | Validation of antibody-based localization data. (**a**) IF images showing nuclear staining of FSCB and PDE1A and staining of LPCAT2 in the endoplasmic reticulum in A-431, U-2 OS and U-251MG cells. FPKM (fragments per kilobase of exon per million fragments mapped) serves as a measure of RNA abundance, where 0 indicates no expression. Scale bars, 10 μm. (**b**) IF images showing localization of C16ORF48, FHL2 and RNF216, as determined by the use of two independent antibodies for each target protein. Cells are counterstained with the nuclear probe DAPI (blue). Scale bars, 10 μm. (**c**) Validation of antibody specificity by knockdown ($P << 0.001$, Mann-Whitney test) of the target protein FHL2 by two different small interfering RNAs (siRNAs 1 and 2). Left, FHL2 staining after siRNA transfection. Right, relative fluorescence intensity (RFI) for the FHL2 staining for siRNA1 (283 cells) and siRNA2 (436 cells) transfected cells compared to that of the negative control (173 cells). Scale bar, 20 μm. (**d**) Colocalization experiments with IF staining (green) of the corresponding FP-expressing (red) cells for the following proteins: PDE1A, CALCOCO1, ANKMY1, C16ORF48, NDUFB8 and ABHD14A. Cells are counterstained with the nuclear probe DAPI (blue). Scale bar, 10 μm.

the antibody binds to the correct target protein and can confirm whether the expression of the FP alters its localization (**Fig. 3d**). For instance, perfect colocalization was observed for the proteins PDEA1 and ANKMY1, whereas colocalization of C16ORF48 revealed that the antibody indeed recognizes the correct protein, but IF showed additional staining of the endogenous protein in the cytoplasm. Another scenario is seen for the mitochondrial protein NDUFB8, which incorrectly localized to the endoplasmic reticulum upon FP expression, although IF also located it in the mitochondria, showing that the antibody still correctly recognized FP-NDUFB8. Some antibodies, as for ABHD14A, were also found to produce nonspecific staining patterns. All antibodies for the specific protein examples shown in **Figures 1**, **2** and **4** have been validated with western blot (**Supplementary Figs. 4–6**).

N- versus C-terminal FP tagging

For comparison of the distribution between the N- and C-terminal fusions, the overlap between the observed localizations for the 873 proteins was classified as dissimilar, similar or identical (**Fig. 4a**). In agreement with previous studies[17], this analysis revealed that 26% of the analyzed proteins showed dissimilar localization patterns. Another network plot was made to identify discrepancies in localization between N- and C-terminal FPs (**Fig. 4b** and **Supplementary Fig. 7**).

The highest N-to-C localization correlations were seen for proteins at the plasma membrane, nucleus and cytoplasm (>85%), whereas mitochondrial proteins showed the poorest overlap (17%). Clusters i and ii showed a systematic bias for mitochondrial proteins in the sense that N-terminal fusions rarely showed mitochondrial localizations but were instead found in the cytoplasm or nucleus (**Fig. 4c**; PHB, PPM1K). This strongly suggests altered membrane targeting due to masking of the N-terminal signal peptide that is required for mitochondrial import[15,18]. Nevertheless, both fusion orientations may still provide relevant information, as for MUL1 (**Fig. 4c**), and should be generated whenever possible. A systematic discrepancy was also identified for cytoskeletal proteins (cluster iii), for which fusions at the C terminus proved to be more reliable (**Fig. 4c**; SEPT3, LIMA1, SSX2IP).

Proteins localizing to secretory-pathway organelles also showed greater variations between the two fusions, probably because they often have specific targeting peptides, usually at their N terminus, that direct their transport across or into membranes[19–21]. Cluster v shows a systematic bias favoring fusions at the C terminus (**Fig. 4c**; AP1AR, CRELD1, SSR1). For cluster iv, the discrepancies are due to localizations to different membrane structures within the secretory pathway, most likely reflecting the highly dynamic nature of these proteins as already discussed. The amount of ectopically expressed protein may also influence the localization. For example, at low expression levels, the Golgi protein FLJ14495 showed a high degree of colocalization with a Golgi marker, as judged by an unbiased colocalization algorithm[22]. At higher expression levels, however, the degree of colocalization at the Golgi apparatus was reduced as the protein became trapped in the endoplasmic reticulum (**Fig. 4d**).

In summary, the correlation between IF and FP tagging was similar for N- and C-terminal fusions of the same protein. However, C-terminal fusions were more reliable for mitochondrial, cytoskeletal and endomembrane proteins, and we

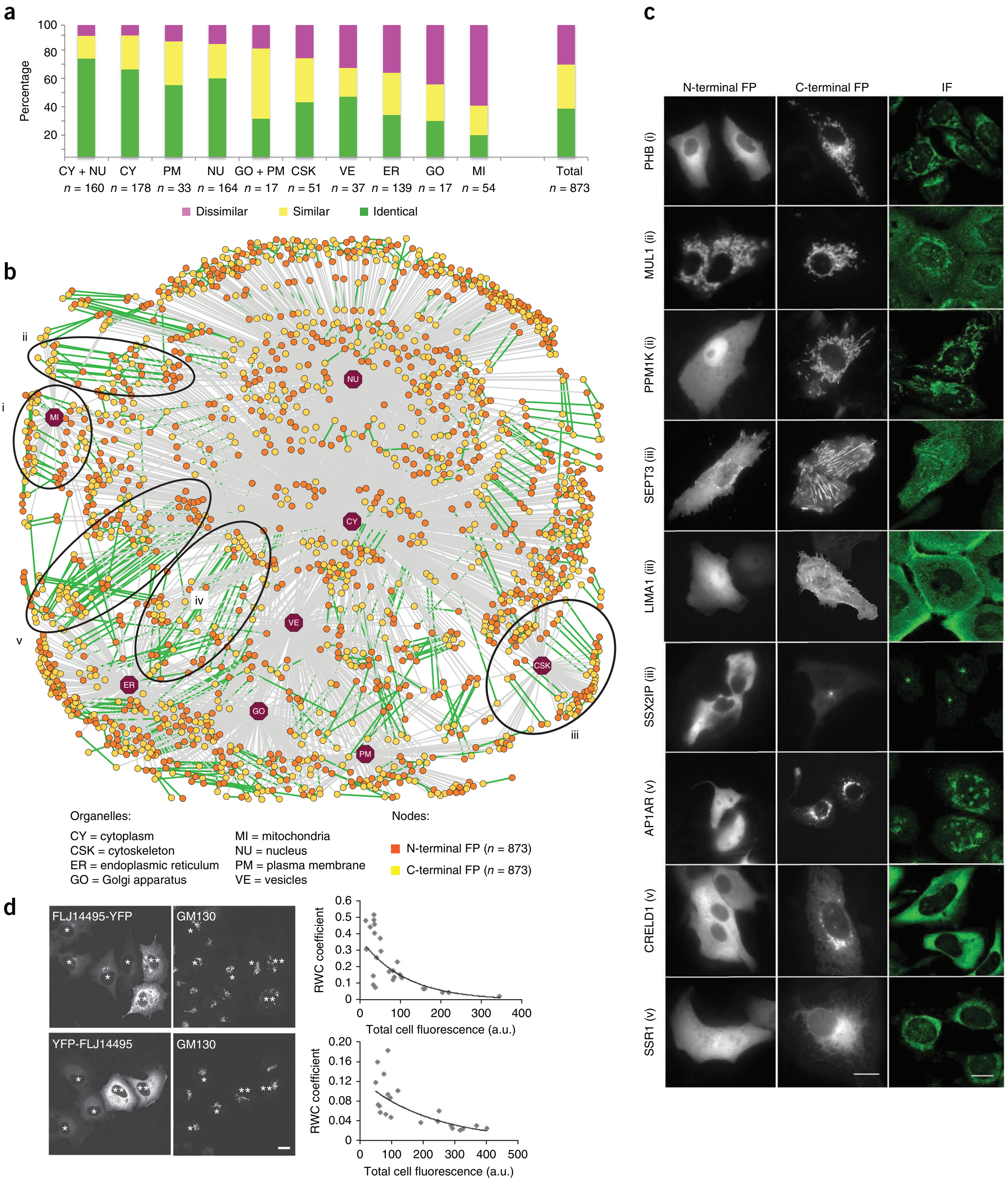


Figure 4 | Differences between N- and C-terminal FP fusions. (**a**) Distribution of identical, similar and dissimilar localizations between N- and C-terminal FPs in each main localization class. *n*, total number of proteins in each class. (**b**) Network showing the protein localization(s) by N- and C-terminal FPs. Gray lines connect each protein to the subcellular structure(s) it was localized to. Green lines connect the two nodes representing the same protein to show discrepancies between the two methods. Clusters of discrepancies are circled and denoted with roman numerals. (**c**) Example fluorescence images of proteins representing the discrepancy clusters from **a**: PHB (i), MUL1 and PPM1K (ii), SEPT3, LIMA1 and SSX2IP (iii) and AP1AR, CRELD1 and SSR1 (v). Scale bars, 10 μm. (**d**) Left, HeLa cells expressing high (**) or low (*) levels of FLJ14495-YFP (C-terminal) or YFP-FLJ14495 (N-terminal) and immunostained for the Golgi marker GM130. Right, correlation between levels of overexpression (that is, total cell fluorescence) and the rank-weighted colocalization (RWC) coefficient between FLJ14495 and GM130. Scale bar, 10 μm. a.u., arbitrary units.

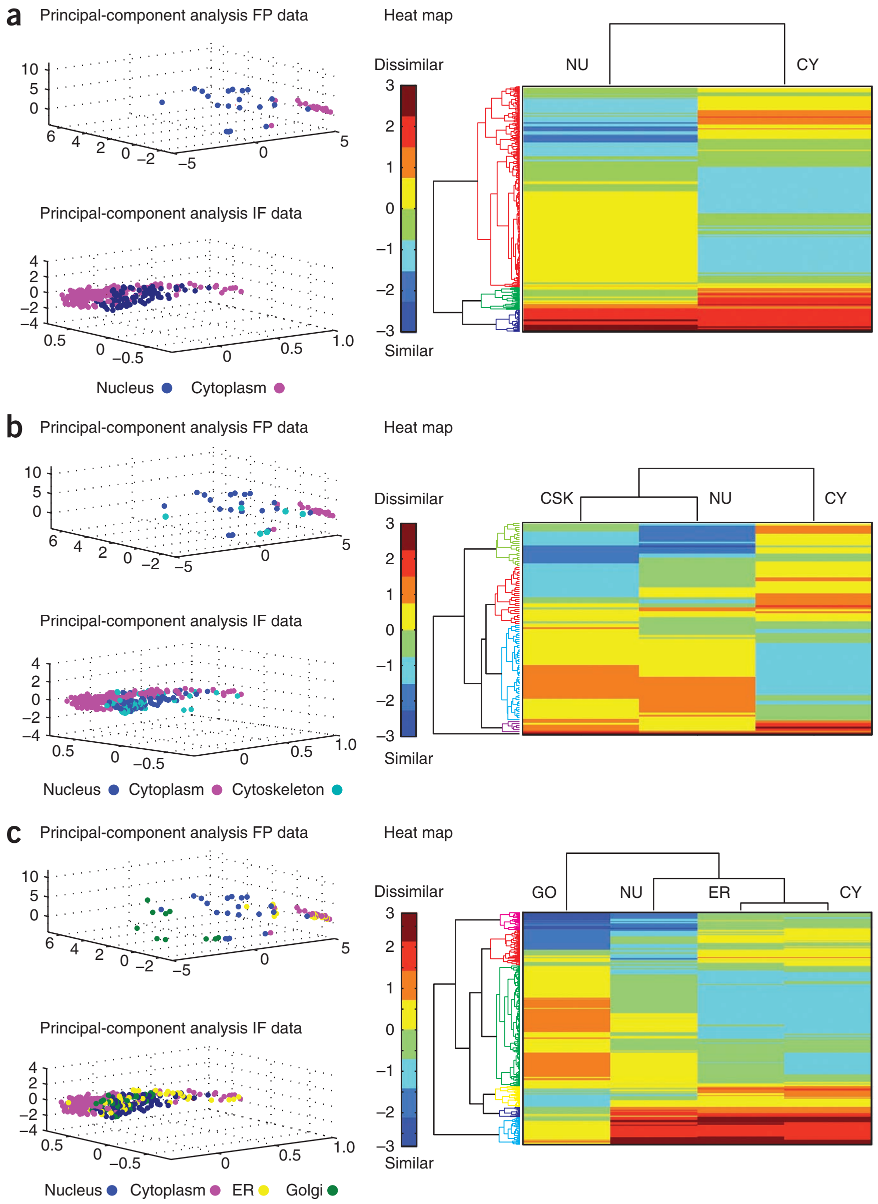


Figure 5 | Feature-based image analysis of IF and FP patterns. (**a**–**c**) Analysis of feature-based descriptions of patterns for images of proteins localized to the following main location classes: nucleus (NU) and cytoplasm (CY) (**a**); nucleus, cytoplasm and cytoskeleton (CSK) (**b**); and nucleus, cytoplasm, endoplasmic reticulum (ER) and Golgi (GO) (**c**). Left, scatter plots of pure organelle patterns based on principal-component analysis of texture and morphology features extracted from training sets of FP and IF images, respectively. These plots demonstrate how the patterns can be distinguished via analysis of image features. Right, dendrograms and associated heat maps show the feature-based similarities and differences of IF and FP images for all proteins belonging to the respective main localization class. Genes and assignment methods of the dendrograms are presented in **Supplementary Data 2**.

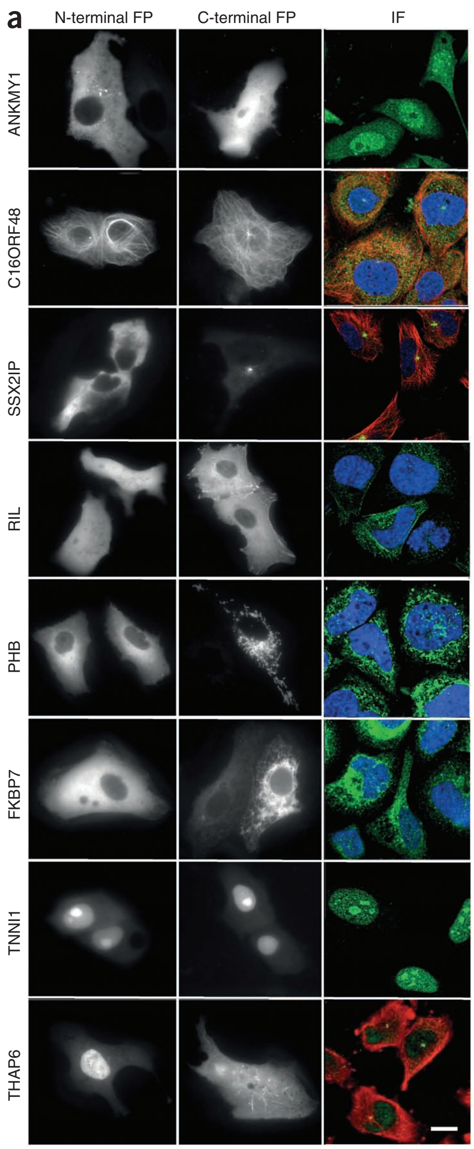
recommend that this should be the first choice when creating new FPs.

Image analysis for comparison of subcellular patterns

Automated image analysis can effectively distinguish subcellular patterns in fluorescence microscope images, in many cases better than or on par with human annotators[23–25]. We performed automated analysis of both IF and FP images, focusing on the organelles that showed the highest variation between the methods as well as a high incidence of multiple subcellular localizations: the nucleus/cytoplasm, endoplasmic reticulum/Golgi and cytoskeleton. A training set of images representing the pure organelle patterns was used to select the 50 most informative texture and morphological features for FP and IF, respectively (**Supplementary Data 1**). The measured distance to the pure pattern centers in the three-dimensional principal-component–transformed space serves as a measure of how similar the analyzed image is to this pattern: the shorter the distance, the more similar the pattern (**Fig. 5**). When comparing feature representation of proteins with the same IF and FP annotation, we saw a good correlation of average distance to the nucleus (rho = 0.341) and cytoplasm (rho = 0.349) centers. The dendrogram and the associated heat map shows the feature-based similarities of the observed IF and FP patterns for all proteins manually annotated to one or more of the organelles of interest (**Fig. 5**). The image analysis presented here clearly shows that proteins with the same annotations are grouped into subclusters with similar patterns. This approach is therefore highly suitable for obtaining a better resolution of similarities and differences among the proteins that have a complex distribution to multiple organelles.

Integrative localization of uncharacterized proteins

The localizations of 363 proteins (82%) were confirmed in this study through the overlap of IF and FPs (**Supplementary Table 4**). Of these, 263 had no prior annotation with respect to experimental data for subcellular localization in UniProtKB, and 65 of them were evident at transcript level only. Examples of such previously unlocalized proteins are SSX2IP localized to centrosomes, RIL localized to actin filaments and the plasma membrane, PHB localized to mitochondria, FKBP7 localized to the endoplasmic reticulum and TNNI1 localized to nucleoli, all together representing a variety of different subcellular structures. Furthermore, ANKMY1 localized to the nucleus and cytoplasm, C16ORF48 localized to centrosomes and microtubules, and THAP6 localized to centrosomes are examples of previously completely uncharacterized proteins evident at transcript level only (**Fig. 6a**). The possibility of applying two methods, such as FP tagging and IF, with a single aim is therefore a useful strategy for the characterization of newly identified proteins. From the results of this study, we suggest that any protein be localized initially using IF and C-terminal FPs. If the obtained results are conflicting, the IF should be validated as demonstrated in **Figure 3** and an N-terminal FP tested. We suggest that this scheme (**Fig. 6b**), combined with automated feature-based image



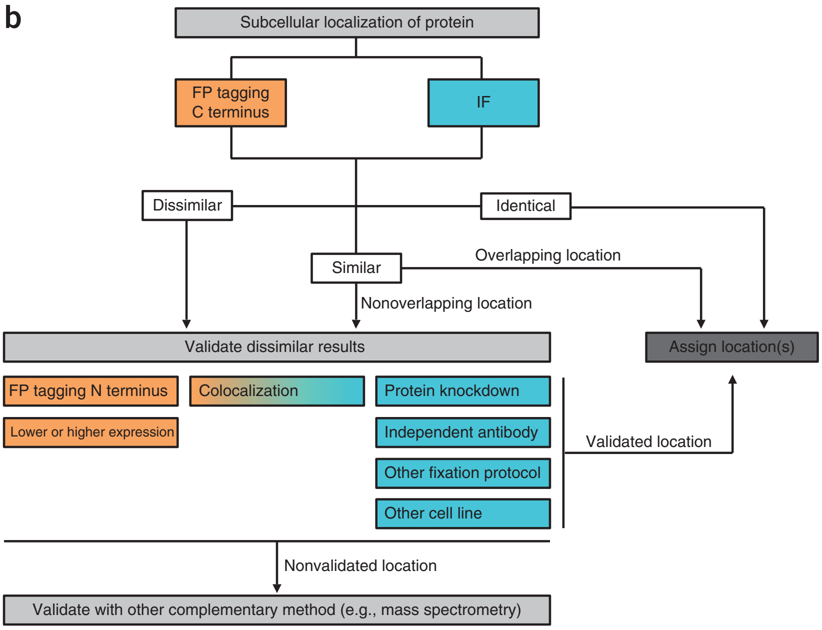


Figure 6 | Systematic localization of uncharacterized proteins. (**a**) Examples of previously uncharacterized target proteins (in UniProtKB) successfully localized according to the overlap of subcellular localizations obtained with IF and FP (N- and C-terminal fusions) for the following proteins: ANKMY1, C16ORF48, SSX2IP, RIL, PHB, FKBP7, TNNI1 and THAP6. For IF images, target proteins are shown in green, nucleus (DAPI) in blue and microtubules in red. Scale bar, 10 µm. (**b**) Schematic overview of how the subcellular localizations of proteins can be characterized and validated systematically using an integrative approach of IF in fixed cells and FP in live cells.

analysis, be used for the systematic localization of proteins and adapted by large-scale efforts aiming to characterize a mammalian proteome such as the Human Proteome Project[26].

DISCUSSION

In this study we present subcellular localization information for 506 proteins and specifically compare the methodologies of IF and FP tagging, the largest comparison of this kind to date.

We show that localization artifacts occur with a similar frequency for live-cell N- and C-terminal FP experiments and for IF experiments: hence the need for cross-validation between the two techniques. The different results obtained by the two methods often represent a complex distribution of the corresponding protein to multiple subcellular compartments and, notably, the largest discrepancies between IF and FP tagging as well as between N- and C-terminal tagging were observed for the highly dynamic cellular structures of the endomembrane system and cytoskeleton. This serves as a reminder that static images cannot capture the complete biological picture with spatiotemporal variations in protein expression resulting from dynamic cell structures, stimuli-induced translocations and cell-cycle dependency.

The nonoverlapping annotations are in most cases the result of false nuclear localizations by IF or FP tagging or of incorrect localization to the endoplasmic reticulum as a result of the ectopic (over)expression of a fusion protein, or cross-reactivity of the antibody when the target protein is not expressed. Furthermore, we show that C-terminal fusions are more reliable, in particular for mitochondrial proteins but also for endoplasmic reticulum and cytoskeletal proteins.

Although this study shows a high correlation between fixed and live-cell localization experiments, it should be noted that well-validated antibodies and the use of appropriate fixation protocols are crucial to accurately reflect the *in vivo* distribution of proteins and provide epitope accessibility[27,28]. In this study, 8% (n = 37) of the antibodies gave no staining despite the fact that RNA sequencing indicated that the target proteins were expressed. These false negative results are likely a consequence of epitope masking caused by the cross-linking fixation. A greater risk with the use of antibodies is potential false positive results due to antibody cross-reactivity when the target protein is expressed at low levels, as demonstrated in this study. For this reason, non-expressed proteins are probably a significant contributor to false localization data in systematic antibody-based studies, even when high-quality antibodies are used.

The complexity of dealing with proteins showing multiple and different localizations accentuates the difficulties in interpreting the correct localization that truly reflects the *in vivo* distribution of the protein. We have demonstrated how automated image analysis could be used to objectively group proteins with similar patterns to obtain a quantitative resolution of similarities and differences among the proteins that have a complex distribution to multiple organelles. Further analysis will be required to distinguish subpatterns of our annotation classes and to measure the fraction of protein present in each[29].

In conclusion, we believe that IF and FP tagging are indispensable techniques that are highly complementary. IF gives information about expression levels and spatial distribution of the endogenous unmodified protein, whereas FP tagging provides information on spatial protein distribution over time. On the basis of our wide experimental data set, we therefore propose a systematic and integrative strategy for the characterization of newly identified proteins. We suggest that this scheme provides the maximum potential for correctly assigning subcellular localization and, in combination with automated image analysis, should be adopted by large-scale efforts aiming to characterize a mammalian proteome.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors wish to acknowledge the entire staff of the Human Protein Atlas project, S. Wiemann and his lab (German Cancer Research Center, DKFZ) for various GFP-ORF constructs and S. Simpson for careful proofreading. The IF work within the frame of the Human Protein Atlas project was supported by grants from the Knut and Alice Wallenberg Foundation, EU Seventh Framework Programme (GA HEALTH-F4-2008-201648/PROSPECTS) and strategic grant Science for Life Laboratory. The J.C.S. lab is supported by a Principal Investigator (PI) grant (09/IN.1/B2604) from Science Foundation Ireland (SFI).

AUTHOR CONTRIBUTIONS

C.S. and E.L. provided the IF data. J.C.S. and R.P. provided the FP data. C.S. and E.R. performed the comparisons between the data sets. E.R. performed the automated image analysis. C.S. and V.R.S. performed control experiments. M.U. and R.F.M. provided intellectual input. E.L. designed and led the study. E.L., J.C.S. and C.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Simpson, J.C., Wellenreuther, R., Poustka, A., Pepperkok, R. & Wiemann, S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1**, 287–292 (2000).
2. Brock, R., Hamelers, I.H. & Jovin, T.M. Comparison of fixation protocols for adherent cultured cells applied to a GFP fusion protein of the epidermal growth factor receptor. *Cytometry* **35**, 353–362 (1999).
3. Goldenthal, K.L., Hedman, K., Chen, J.W., August, J.T. & Willingham, M.C. Postfixation detergent treatment for immunofluorescence suppresses localization of some integral membrane proteins. *J. Histochem. Cytochem.* **33**, 813–820 (1985).
4. Hoetelmans, R.W. *et al.* Effects of acetone, methanol, or paraformaldehyde on cellular structure, visualized by reflection contrast microscopy and transmission and scanning electron microscopy. *Appl. Immunohistochem. Mol. Morphol.* **9**, 346–351 (2001).
5. Stadler, C., Skogs, M., Brismar, H., Uhlen, M. & Lundberg, E. A single fixation protocol for proteome-wide immunofluorescence localization studies. *J. Proteomics* **73**, 1067–1078 (2010).
6. Shibata, T., Tanaka, T., Shimizu, K., Hayakawa, S. & Kuroda, K. Immunofluorescence imaging of the influenza virus M1 protein is dependent on the fixation method. *J. Virol. Methods* **156**, 162–165 (2009).
7. Schnell, U., Dijk, F., Sjollema, K.A. & Giepmans, B.N. Immunolabeling artifacts and the need for live-cell imaging. *Nat. Methods* **9**, 152–158 (2012).
8. Huh, W.K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
9. Starkuviene, V. *et al.* High-content screening microscopy identifies novel proteins with a putative role in secretory membrane traffic. *Genome Res.* **14**, 1948–1956 (2004).
10. Liebel, U. *et al.* A microscope-based screening platform for large-scale functional protein analysis in intact cells. *FEBS Lett.* **554**, 394–398 (2003).
11. Simpson, J.C., Neubrand, V.E., Wiemann, S. & Pepperkok, R. Illuminating the human genome. *Histochem. Cell Biol.* **115**, 23–29 (2001).
12. Fagerberg, L. *et al.* Mapping the subcellular protein distribution in three human cell lines. *J. Proteome Res.* **10**, 3766–3777 (2011).
13. Uhlén, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell Proteomics* **4**, 1920–1932 (2005).
14. Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **28**, 1248–1250 (2010).
15. Fölsch, H., Gaume, B., Brunner, M., Neupert, W. & Stuart, R.A. C- to N-terminal translocation of preproteins into mitochondria. *EMBO J.* **17**, 6508–6515 (1998).
16. Stadler, C. *et al.* Systematic validation of antibody binding and protein subcellular localization using siRNA and confocal microscopy. *J. Proteomics* **75**, 2236–2251 (2012).
17. Simpson, J.C. *et al.* Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nat. Cell Biol.* **14**, 764–774 (2012).
18. Stan, T. *et al.* Mitochondrial protein import: recognition of internal import signals of BCS1 by the TOM complex. *Mol. Cell Biol.* **23**, 2239–2250 (2003).
19. von Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* **133**, 17–21 (1983).
20. von Heijne, G. Signal sequences. The limits of variation. *J. Mol. Biol.* **184**, 99–105 (1985).
21. Janda, C.Y. *et al.* Recognition of a signal peptide by the signal recognition particle. *Nature* **465**, 507–510 (2010).
22. Singan, V.R., Jones, T.R., Curran, K.M. & Simpson, J.C. Dual channel rank-based intensity weighting for quantitative co-localization of microscopy images. *BMC Bioinformatics* **12**, 407 (2011).
23. Boland, M.V. & Murphy, R.F. 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 (2001).
24. Conrad, C. *et al.* Automatic identification of subcellular phenotypes on human cell arrays. *Genome Res.* **14**, 1130–1136 (2004).
25. Li, J., Newberg, J.Y., Uhlen, M., Lundberg, E. & Murphy, R.F. Automated analysis and reannotation of subcellular locations in confocal images from the human protein atlas. *PLoS ONE* **7**, e50514 (2012).
26. Paik, Y.K. *et al.* The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **30**, 221–223 (2012).
27. Jamur, M.C. & Oliver, C. Permeabilization of cell membranes. *Methods Mol. Biol.* **588**, 63–66 (2010).
28. Melan, M.A. & Sluder, G. Redistribution and differential extraction of soluble proteins in permeabilized cultured cells. Implications for immunofluorescence microscopy. *J. Cell Sci.* **101**, 731–743 (1992).
29. Peng, T. *et al.* Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns. *Proc. Natl. Acad. Sci. USA* **107**, 2944–2949 (2010).

ONLINE METHODS

Establishment of FP-expressing cells and live-cell imaging. Cloning and transfection for expression of the target proteins in fusion with CFP (cyan) or YFP (yellow) to the N and C termini, respectively, have previously been described in detail, including the methodology for annotating localizations[1]. Constructs were transiently transfected into Vero or HeLa cells growing in 35-mm glass-bottomed dishes (MatTek) 20 h before imaging. All images used for protein localization annotation were recorded from live cells using a Leica DMI6000B microscope equipped with a 63×/NA 1.4PL Apo objective. If more than one localization was observed, they were defined as main or additional, depending on the relative intensity between the observed organelles. All images are available at the project web site (http://gfp-cdna.embl.de/).

Antibodies and immunofluorescence. The antibodies used for immunofluorescence were rabbit polyclonal antibodies, affinity purified using the antigen as ligand, and generated and validated within the Human Protein Atlas project[13,14]. The procedures for cell cultivation, immunostaining, image acquisition and image annotation have been described elsewhere[5,12,30]. All images used for protein localization annotation were recorded from fixed cells using a Leica SP5 microscope equipped with a 63×/NA 1.4PL Apo oil-immersion objective. Within the Human Protein Atlas project, we had in-house–developed antibodies for 506 proteins out of the 873 tagged with FPs. These antibodies have been used for IF in three different human cell lines: A-431, U-2 OS and U-251MG. The localization(s) were annotated separately in each cell line, and on the basis of the results from the three cell lines together, the location(s) of the target proteins was determined. If more than one localization was observed, the most prominent localization (either by detection in several cell lines or by stronger relative staining intensity) was defined as the main localization, whereas the remaining localizations were defined as additional. All images and cell line–specific annotations are available at the Human Protein Atlas web portal (http://www.proteinatlas.org/) For a complete list of the antibodies used and the location(s) of each target protein (as decided by the union of the three cell lines), see **Supplementary Table 2**.

Image processing for article figures. All image processing for preparation of article figures (merge of channels, cropping and montage construction for figures) was performed in ImageJ version 1.410.

Overexpression and colocalization experiments. For analysis of localization impact due to overexpression, HeLa cells were transfected with constructs encoding either FLJ14495-YFP or YFP-FLJ14495 and then immunostained for the Golgi marker GM130. Transfected cells were segmented, the levels of overexpressed protein were quantified and the rank-weighted colocalization (RWC) coefficient[22] with GM130 was calculated for various cells in the population. For colocalization of the antibodies used in IF with the FP-tagged target protein, cells were fixated and immunostained using the same protocol used for all IF experiments and described in previous work[16].

Western blot. All antibodies in the study have been validated with WB within the Human Protein Project using a routine sample setup of protein lysates from a limited number of cell lines[13]. WBs for the antibodies in **Figures 1b**, **2b** and **4c** have been done on lysate from A-431, U-2 OS or U-251MG cells (**Supplementary Figs. 4–6**).

Antibody validation using siRNA. U-2 OS cells were independently transfected with two different siRNAs targeting *FHL2* (Silencer Select, s5197 and s5198, Ambion) using a solid-phase protocol. A scrambled siRNA (s229174, Silencer Select, Ambion) was used as negative control. After 72 h, cells were fixed and immunostained with the antibody HPA005922 targeting FHL2. The fluorescence intensity of FHL2 in transfected cells (two different cell populations) was compared to that of the negative scrambled control, with CellProfiler used to calculate the median of the FHL2 intensity within each population (scrambled = 173 cells, siRNA1 = 271 cells and siRNA2 = 471 cells targeting *FHL2*). The results are presented in box plots with median intensity of the FHL2 staining in the control as 100%. Significance testing was done using a Mann-Whitney ranking test. The entire procedure of transfection, staining and analysis has been described in detail in previous work[16].

RNA sequencing. RNA sequencing data of the three cell lines U-2 OS, A-431 and U-251MG were generated as part of a separate study[31]. As quantitative measurements of gene expression, FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated to normalize for both gene length and total number of reads in the measurement. FPKM values were calculated with respect to genes from Ensembl release version 63.37 using Cufflinks (v.1.0.3). The raw sequence data files were uploaded to the NCBI Sequence Read Archive with accession number SRA062599.

Definition of overlap between FP and IF. The localization results (main and additional localization) obtained for IF and FP tagging were compared, and the overlap was defined as 'identical' (one or multiple localizations observed with both methods), 'similar' (one localization observed with both methods but with additional localization(s) observed with either of the two methods) or 'dissimilar' (no common localization observed with the two methods). For nuclear proteins, the overlap was considered to be dissimilar if showing nucleoli with one method and simply the nucleus with the other. The extended nuclear annotations such as speckles or spots in IF were not taken into account for the comparisons. In the cytoskeleton category, proteins seen in different filament structures were considered as dissimilar. In cases of aggregates, no overlap was counted even if seen with both methods. In cases of no IF staining, the protein was categorized as 'negative', and if negative at either N- or C-terminal tagging, the overlap was based on the comparison with the tagged protein showing a localization.

Network analysis. Cytoscape software[32] was used for a visual exploration and mining of the complementary role of IF and FP in investigating subcellular localization of the proteins included in this study. For each protein, manual annotations were numerically categorized according to the labeling (i.e., FP or IF), tagging method (N or C terminus) and main subcellular localization (for the networks in **Figs. 1a** and **2a**) or main and additional localizations (for the network in **Fig. 4b**). Categorized annotations were then

doi:10.1038/nmeth.2377

automatically imported into Cytoscape with the node being defined by either the corresponding gene name or by the subcellular localization organelles as defined above. Edges (lines) connecting nodes were defined by the subcellular annotation. The labeling technique and tagging method were added to the network by selectively labeling the edges and the nodes with different methods (i.e., color and font). The edge-weighted spring-embedded layout algorithm[33,34], based on the force-directed drawing approach, was used with edges weight being the subcellular numerical category. Nodes were further reorganized using the layout algorithm above into clusters of nodes with similar subcellular localization patterns (subcellular organelles being the center of all node clusters).

Automated image analysis of confocal microcopy images. Raw TIFF uncompressed images were analyzed using CellProfiler[35] to find single cells using the combination of the different fluorescence channels for both IF (DAPI, endoplasmic reticulum and microtubules) and FP (N or C terminus). Independent CellProfiler rule sets where developed for the segmentation of images generated by each method (**Supplementary Notes 1** and **2**), but the same number of features (n = 292) were extracted from the protein fluorescence channel. Independent images of proteins localized to a single subcellular location were acquired for each method (n = 602 for IF, n = 223 for FP) to be used as a training set. IF and FP images of the proteins investigated in this study were used as a test set and were analyzed in the same way as the training set. Features that distinguish between the training sets were selected using the information-gain algorithm and tenfold cross-validation. All automated texture and morphological features describing IF and FP were further ranked according to the information-gain algorithm output (i.e., entropy). The 50 top-ranked features with the highest entropy were selected to describe the variability of the IF and FP subcellular patterns (**Supplementary Data 1**). Principal-component analysis (PCA) was then carried out on the selected features for the training sets of images for each method (i.e., IF and FP), and the first principal components accounting for 99.5% of the variability were automatically selected. IF and FP image data from the test sets were then visualized using the selected first three PCA components. In the PCA representation, a representative feature center is computed for each subcellular structure and method using the training data (i.e., pure patterns). A distance metric is computed for all data in the test set with regard to the training data feature centers. These distances were then clustered using a hierarchical clustering algorithm. The protein clusters and feature distances were then visualized using a dendrogram and a heat map. The labels of the dendrograms are presented in **Supplementary Data 2**. Feature selection and principal-component analysis were carried out using MATLAB R.12 (MathWorks).

30. Barbe, L. *et al.* Toward a confocal subcellular atlas of the human proteome. *Mol. Cell Proteomics* **7**, 499–508 (2008).
31. Danielsson, F. *et al.* RNA deep sequencing as a tool for selection of cell lines for systematic subcellular localization of all human proteins. *J. Proteome Res.* **12**, 299–307 (2013).
32. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
33. Fruchterman, T.M.J. & Reingold, E.M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129–1164 (1991).
34. Kamada, T. & Kawai, S. An algorithm for drawing general undirected graphs. *Inf. Process. Lett.* **31**, 7–15 (1989).
35. Carpenter, A.E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).

doi:10.1038/nmeth.2377