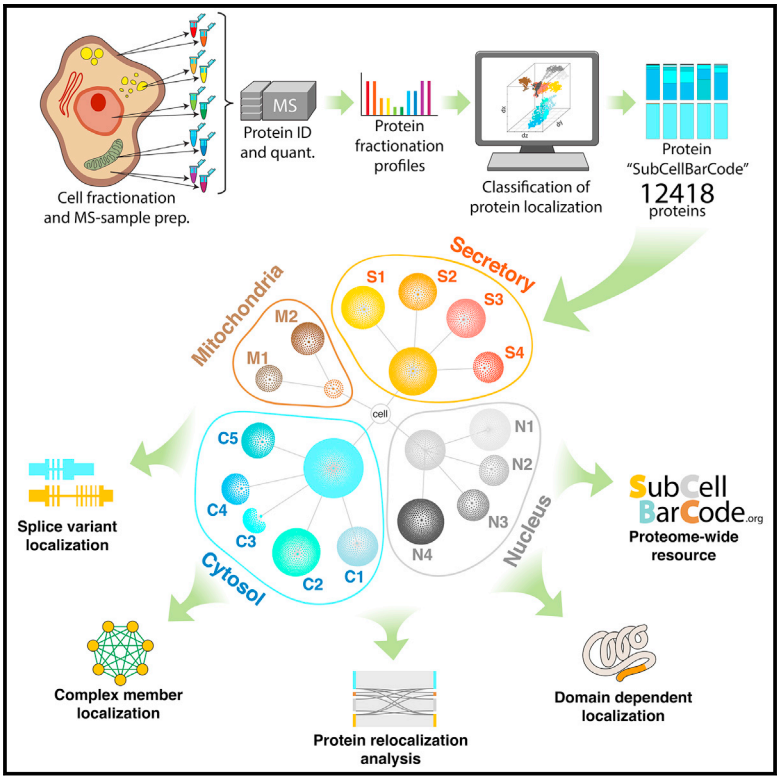
Resource

Molecular Cell

SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization

Graphical Abstract



Highlights

- Robust method for proteome-wide subcellular localization analysis
- Resource for subcellular location of 12,418 proteins in multiple cell lines
- Analysis of protein domain- and variant-associated localization
- Global analysis of protein relocalization driven by EGFR inhibition

Authors

Lukas Minus Orre, Mattias Vesterlund, Yanbo Pan, ..., Oliver Frings, Erik Fredlund, Janne Lehtiö

Correspondence

lukas.orre@ki.se (L.M.O.), janne.lehtio@ki.se (J.L.)

In Brief

Orre et al. use mass spectrometry to map the subcellular localization of more than 12,000 proteins across several cell lines and present a method for proteome-wide relocalization analysis. This study provides information about the spatial organization of the proteome and an accessible resource for protein localization, subcellbarcode.org.

Orre et al., 2019, Molecular Cell *73*, 166–182
January 3, 2019
https://doi.org/10.1016/j.molcel.2018.11.035

CellPress

CellPress

Molecular Cell
Resource

SubCellBarCode: Proteome-wide Mapping of Protein Localization and Relocalization

Lukas Minus Orre,[1,2,*] Mattias Vesterlund,[1,2] Yanbo Pan,[1,2] Taner Arslan,[1,2] Yafeng Zhu,[1] Alejandro Fernandez Woodbridge,[1] Oliver Frings,[1] Erik Fredlund,[1] and Janne Lehtiö[1,3,*]
[1]Department of Oncology and Pathology, Karolinska Institutet, Science for Life Laboratory, 17165 Solna, Sweden
[2]These authors contributed equally
[3]Lead Contact
*Correspondence: lukas.orre@ki.se (L.M.O.), janne.lehtio@ki.se (J.L.)
https://doi.org/10.1016/j.molcel.2018.11.035

SUMMARY

Subcellular localization is a main determinant of protein function; however, a global view of cellular proteome organization remains relatively unexplored. We have developed a robust mass spectrometry-based analysis pipeline to generate a proteome-wide view of subcellular localization for proteins mapping to 12,418 individual genes across five cell lines. Based on more than 83,000 unique classifications and correlation profiling, we investigate the effect of alternative splicing and protein domains on localization, complex member co-localization, cell-type-specific localization, as well as protein relocalization after growth factor inhibition. Our analysis provides information about the cellular architecture and complexity of the spatial organization of the proteome; we show that the majority of proteins have a single main subcellular location, that alternative splicing rarely affects subcellular location, and that cell types are best distinguished by expression of proteins exposed to the surrounding environment. The resource is freely accessible via www.subcellbarcode.org.

INTRODUCTION

Increasing knowledge regarding cellular signaling has revealed a highly dynamic architecture of eukaryotic cells. Protein complexes are constantly being formed and resolved, and proteins are shuttling between different subcellular localizations to execute biological processes. Historically, protein subcellular location has been determined by targeted approaches for individual proteins; e.g., cell fractionation coupled to western blot quantification. Although additional development of large-scale GFP fusion protein-based (Huh et al., 2003) and antibody-based assays (Thul et al., 2017) has greatly increased our knowledge of protein subcellular localization, these methods are labor-intensive, making expansion of data challenging. The development of mass spectrometry (MS)-based proteomics coupled with subcellular fractionation protocols has opened up new possibilities to query the spatial organization of the proteome on a larger scale; e.g., as shown for organellar proteins (Andersen et al., 2003) and soluble protein complexes (Havugimana et al., 2012). More recently, two studies have applied machine learning algorithms to assign subcellular localization based on protein quantification across multiple subcellular fractions by MS. The first study classified 2,855 proteins into 14 locations in mouse pluripotent stem cells (Christoforou et al., 2016), and the second study classified 2,423 proteins into 9 membranous organelles in HeLa cells, with an additional decision tree heuristic assigning nuclear or cytoplasmic localization for a further 3,804 proteins (Itzhak et al., 2016). These studies were performed in single-cell line systems. Thus, proteome-wide information as well as generalizability into multiple parallel cell line models are still lacking.

Here we report a comprehensive investigation of protein subcellular localization using cell fractionation combined with in-depth quantitative MS, followed by classification of localization for proteins mapping to 12,418 unique genes across five different cell lines. For each protein, classifier scores describing 15 compartments were used to generate "SubCellBarCodes" for evaluation of subcellular localization. Further, we investigate differential localization of proteins, subcellular distribution of protein classes such as kinases and E3 ligases, domain-driven localization, protein complex member co-localization, and the effect of alternative splicing on protein localization. Moreover, we demonstrate the use of our method for proteome-wide condition-dependent protein localization analysis by investigating epidermal growth factor receptor (EGFR) inhibition-dependent protein relocalization.

Collectively, our analyses provide new knowledge about the architecture of cells and the complexity of the spatial organization of the proteome. The SubCellBarCode resource (www.subcellbarcode.org) enables querying the localization of individual proteins and protein classes as well as data mining opportunities and open data sharing. Importantly, our resource can be expanded in multiple dimensions to include data representing additional cell lines, cell fractionation protocols, and chemical and genetic perturbations.

RESULTS AND DISCUSSION

Subcellular Fractionation and MS Analysis

To generate a resource of protein localization, we developed a pipeline consisting of subcellular fractionation, MS-based

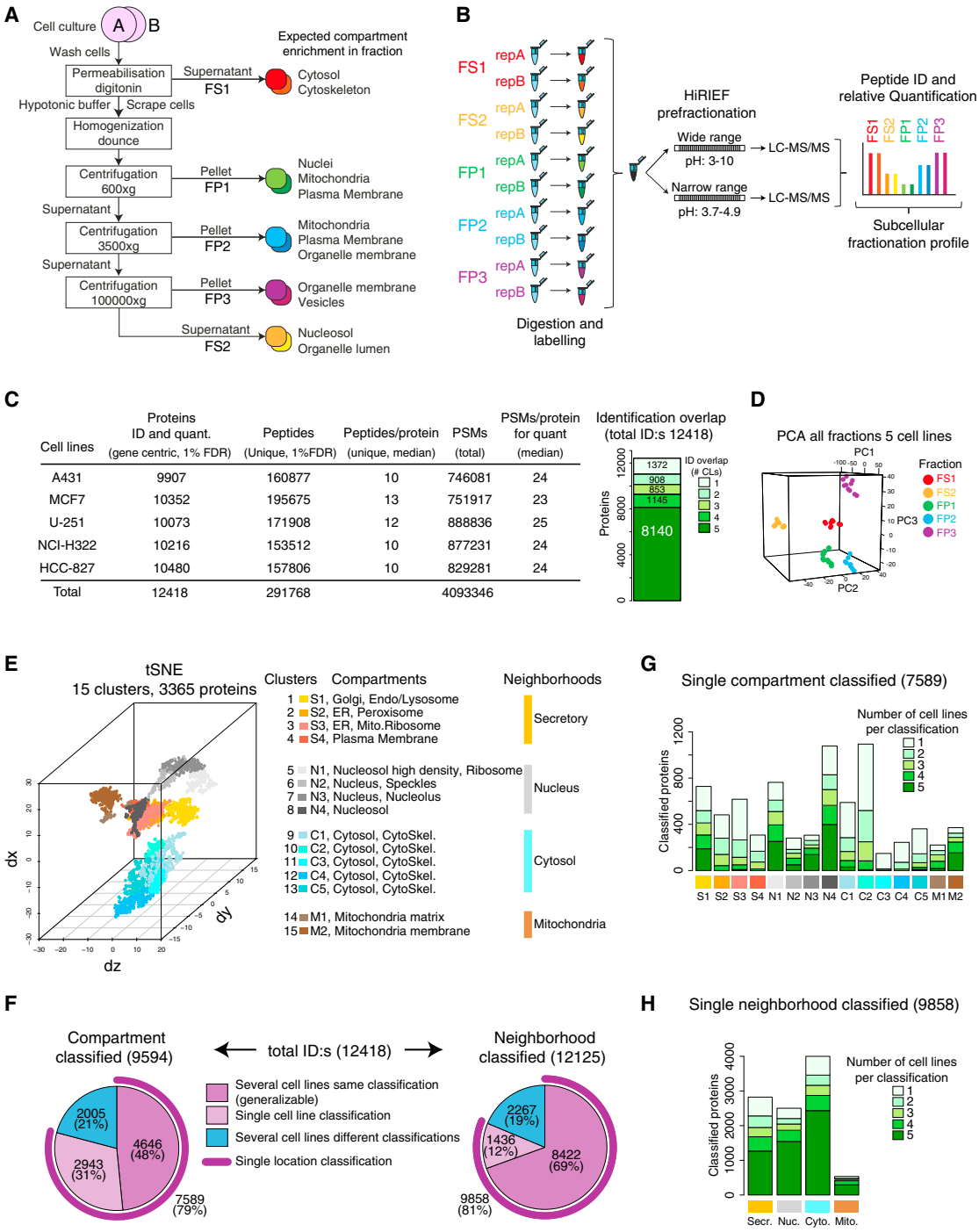


Figure 1. Subcellular Fractionation, Quantitative Mass Spectrometry, and Classification of Subcellular Localization

(A) Overview of the protocol used to generate five different subcellular fractions in duplicates for each cell line (CL).

(B) Quantitative proteomics by isobaric labeling and high resolution isoelectric focusing liquid chromatography mass spectrometry (HiRIEF-LC-MS) for generation of fractionation profiles.

(C) MS output and identification overlap across all five CLs.

(D) Principal-component analysis based on the 8,140 proteins for which fractionation profiles were generated in all five CLs.

(E) tSNE distribution of marker proteins colored based on 15 clusters generated by mClust (left). Also shown are annotated compartments and neighborhoods (right).

(legend continued on next page)

quantification of proteins across fractions, and classification of localization based on fractionation profiles. First, we established a simple and robust protocol dividing the cellular components into five different fractions to allow relative protein quantification of duplicate samples in a single 10-plex MS experiment (Figure 1A). The performance of the fractionation protocol was evaluated by western blotting using markers of different subcellular compartments (Figure S1A). Five different human cancer cell lines (epidermoid carcinoma A431, glioblastoma U251, breast cancer MCF7, lung cancer NCI-H322, and lung cancer HCC-827) were fractionated in biological duplicates, generating a total of 50 samples.

For each cell line, proteins from the 10 fractions were digested into peptides, labeled using 10-plex tandem mass tag (TMT) isobaric labels, and analyzed by high resolution isoelectric focusing liquid chromatography MS (HiRIEF-LC-MS) (Branca et al., 2014) to generate subcellular fractionation profiles for each protein (Figure 1B). This setup resulted in identification and quantification of 12,418 proteins (gene-centric; i.e., corresponding to 12,418 genes) with an overlap of 8,140 proteins across all five cell lines (Figure 1C; Table S1). High quantitative robustness was achieved, as shown by high fractionation profile correlation between duplicates (Figure S1B). Principal-component analysis (PCA; Figure 1D), unsupervised clustering (Figure S1C), and protein correlation network analysis (Figure S1D) showed distinct clustering of samples as well as proteins based on fractionation profiles, indicating that the generated data enable resolution of distinct subcellular compartments.

Relative protein abundance between the five cell lines was also determined by MS-based quantification of total protein lysates. Proteomics analysis of total lysates resulted in identification of 10,859 proteins (Figures S1E–S1G; Table S1).

Classification of the Subcellular Localization of 12,418 Proteins

The principle of the approach used here is that the fractionation profiles of proteins can be used for classification of subcellular localization. To perform a purely data-driven analysis, we postulated that any protein with a highly reproducible and robust fractionation profile across all five cell lines can be used as a marker protein for classification of subcellular localization. Accordingly, 3,365 marker proteins were selected from the 8,140 proteins identified in all cell lines based on quantitative robustness (duplicate Pearson correlation > 0.8) and cell line generalizability (between cell line Pearson corr. > 0.8 and Spearman corr. > 0.6).

Next, 15 marker protein clusters were identified by dimensionality reduction (t-distributed stochastic neighbor embedding [tSNE]) of the combined cell line fractionation profiles (50 points, 3,365 proteins) and clustering of the resulting three dimensions using mClust (Figures 1E and S2A; Table S2). The 15 clusters were then associated with specific cellular compartments by enrichment analysis against Gene Ontology (GO) and UniProt annotations covering 11 different subcellular compartments (Figure S2B), which provided cluster annotations (Figure 1E).

The enrichment analysis as well as the marker protein distribution in tSNE space and fractionation profiles (Figure S2C) indicated inter-cluster relationships and the presence of cluster "neighborhoods." Based on this and textbook knowledge of subcellular compartment relationships, we defined four neighborhoods: "secretory" (clusters 1–4), "nuclear" (clusters 5–8), "cytosol" (clusters 9–13), and "mitochondrial" (clusters 14 and 15). For readability, we will refer to the 15 clusters as compartments named according to their corresponding neighborhood (S1–S4, N1–N4, C1–C5, and M1 and M2). The final annotations of the compartments and neighborhoods are shown in Figure 1E.

For support vector machine (SVM)-based classification, the 3,365 marker proteins were divided into a training (2,362) and test (1,003) set, balanced to cover the 15 compartments. The training set proteins were used to train independent classifiers for each duplicate experiment in each cell line without using the test set proteins. Classification was performed on all proteins identified in the five cell lines analyzed, and test set proteins were used to evaluate the performance and to set thresholds for classification of proteins into compartments and neighborhoods (Table S3). Of the identified and quantified proteins in the five cell line dataset (12,418), we successfully classified 9,594 into specific compartments and 12,125 into specific neighborhoods (Figures 1F–1H).

To evaluate the robustness and accuracy of the method, an independent dataset was produced by re-performing the entire wet lab and bioinformatics workflow in three independent experiments for one of the cell lines (HCC827), with each experiment containing biological duplicates of all samples. Classification of the 9,350 proteins that were identified and quantified in all three HCC827 experiments resulted in classification of 6,190 proteins at the compartment level and 9,286 proteins at the neighborhood level in at least one experiment (Figure S2D; Table S3). Conflicting classifications between experiments were made in 3% of compartment classifications and 5% of neighborhood classifications, in line with the estimated false classification rate.

Evaluation of Determined Protein Localizations against Public Domain Data

To evaluate our results, proteins classified to a single compartment (7,589) were assessed against annotations in GO and UniProt (single location annotations, 9,016 proteins). This evaluation supports our classifications with clear enrichments of the expected annotations across our 15 compartments (Figure S3A). Further, to evaluate generalizable neighborhood classifications (i.e., the same classification in several cell lines; 8,422 proteins) against GO and/or UniProt annotations, we binned the public domain annotations into four bins matching the neighborhoods defined here (Figures 2A and S3B). For the overlapping 3,952 proteins, we noted an agreement of 84%, increasing to 89% for proteins annotated to the same location in both GO and UniProt (Figure 2B).

(F) Classification results.
(G) Single-location classifications at compartment level.
(H) Single-location classifications at neighborhood level.
See also Figures S1 and S2 and Tables S1, S2, and S3.

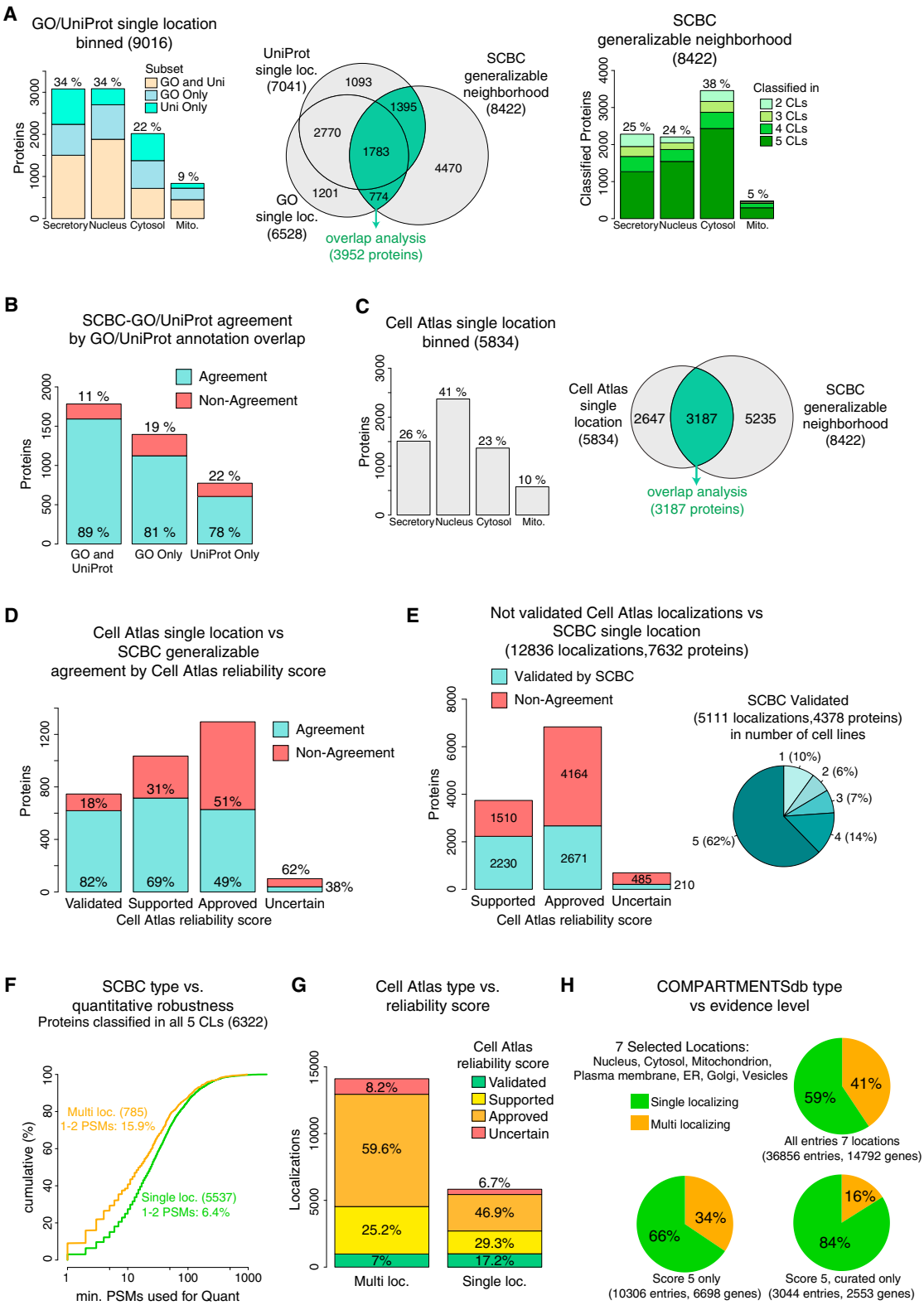


Figure 2. Evaluation of SubCellBarCode Protein Localization Classifications against GO, UniProt, and Cell Atlas

(A) Proteins annotated with a single location in GO or UniProt were binned to match the neighborhood classifications (left). Also shown is overlap between proteins annotated with a single location in GO and UniProt (center) and proteins with a generalizable neighborhood classification (right).

(legend continued on next page)

Next, we performed a similar analysis using subcellular localization information generated by antibody-based immunofluorescence microscopy in the Cell Atlas project (Thul et al., 2017). Proteins in Cell Atlas assigned to a single location (5,834) were binned to match the four neighborhoods and compared with our classifications (Figures 2C and S3C). In general, the agreement between our classifications and Cell Atlas for the overlapping 3,187 proteins was lower (63%) than the agreement with GO and/or UniProt. This result was highly dependent on the Cell Atlas-stated reliability score, with 82% agreement for proteins with the highest score (''validated,'' n = 746; Figure 2D). The agreement then dropped with decreasing Cell Atlas score from ''supported'' (69% agreement) to ''approved'' (49%) and ''uncertain'' (38%).

Cell Atlas reports close to 20,000 localizations for almost 12,000 proteins, where 10% of localizations are validated (Figure S3D). For the non-validated, almost 18,000 localizations, the data here offer orthogonal evaluation for 12,836 localizations corresponding to 7,632 proteins (Figure 2E). Of these localizations, our data validated 5,111 localizations of 4,378 proteins.

Localization of Proteins to More Than One Location

Overall, approximately 20% of proteins were classified to different locations between cell lines, dropping to 12% for the 6,322 proteins classified in all 5 cell lines. Further, quantifications of proteins classified to more than one location were based on fewer peptide spectrum matches (PSMs) than proteins classified to a single location (Figure 2F). Combined with the estimated 5%–10% false classification rate, this result suggests that the reliability of classifications made to more than one location is lower than that made to a single location.

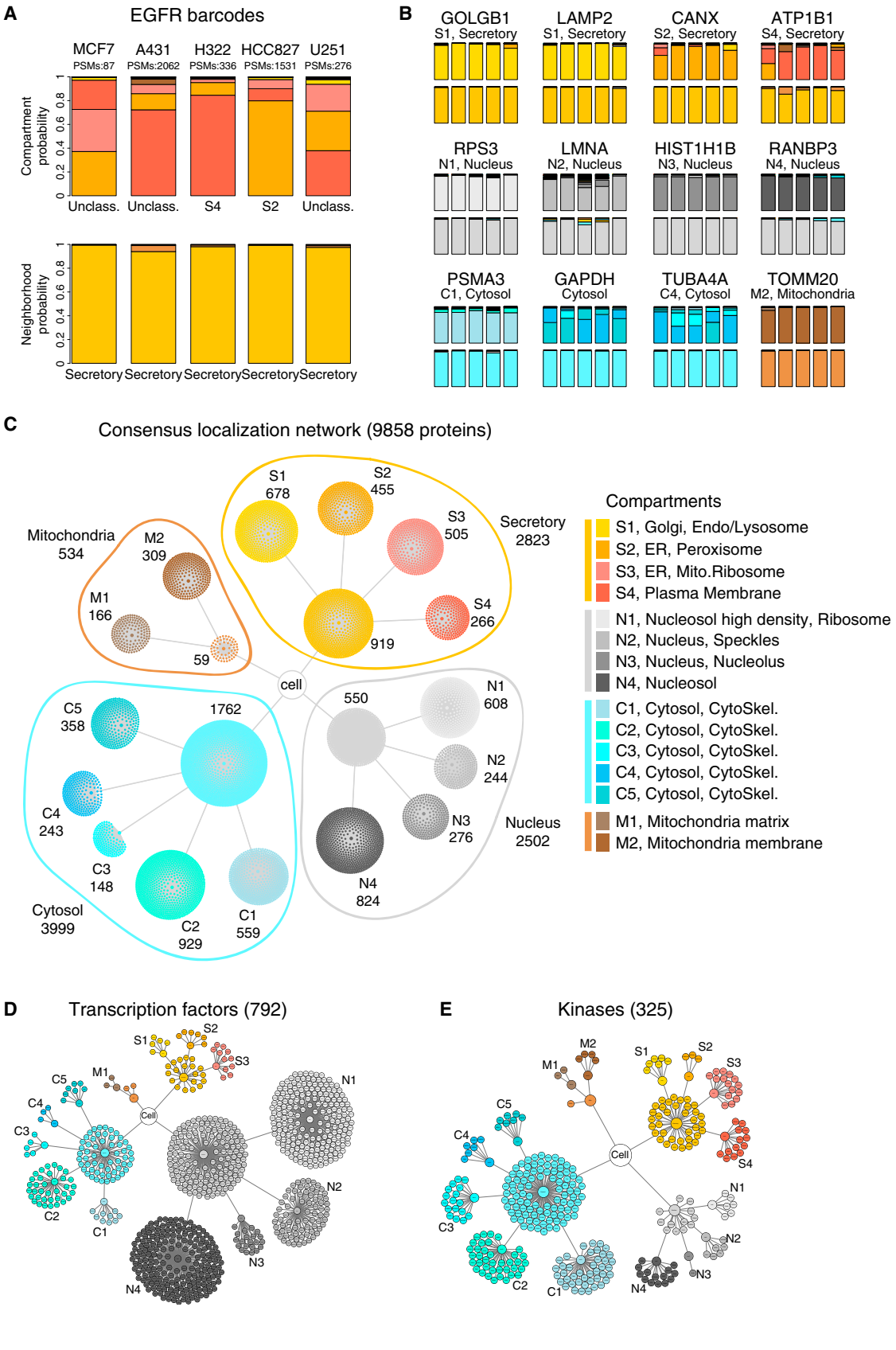
Even though our method is limited to directly determining the main localization of proteins, multi-localization of proteins can be detected indirectly. If a protein localizes to more than one location in an individual cell line, then this would result in a mixed fractionation profile, leading to reduced classification probabilities and potential loss of classification. The percentage of unclassified proteins in each cell line spans 5.8%–9.8%, indicating that mixed fractionation profiles are not a major problem in our analysis. In the triplicate HCC827 experiment, where we can assume identical localization of proteins, conflicting neighborhood classifications between experiments were still made for 5% of the proteins. Importantly, these conflicting classifications are based on less robust quantification (median, 4 PSMs for quantification) than non-conflicting classifications (median, 19 PSMs; Figure S4A). Overall, this analysis describes the effect of analytical noise on classification, resulting in misinterpretation of proteins as multi-localizing. Taking this into account, our estimation based on experimental data is that multi-localization of proteins is not a general feature and that less than 10% of proteins are multi-localizing. This is substantially less than reported previously.

The Cell Atlas resource reported that more than 50% of all proteins localize to more than one subcellular location (Thul et al., 2017). As discussed above, multi-localization of proteins can be detected in our data in different ways, either through differential classification between cell lines or through reduced classification probabilities or loss of classification in individual cell lines. We did not see any difference in our data between proteins reported in Cell Atlas to be single- or multi-localizing, neither in terms of classification output (Figure S4B) nor in classification probabilities (Figure S4C). However, proteins reported as localizing to more than one location in Cell Atlas were, in general, associated with inferior Cell Atlas reliability scores (Figure 2G), and only 369 proteins (3.1%) had the validated reliability score in more than one location. In our data, low fractionation profile correlation between peptides mapping to the same gene can indicate the presence of a protein at more than one cellular location. Such an analysis indicated slightly more low-correlating peptides in the subset of proteins assigned to multiple locations by Cell Atlas with a validated reliability score but no difference between the non-validated subset and proteins assigned to a single location (Figure S4D).

One of the most comprehensive resources of protein localization information is the COMPARTMENTS database (Binder et al., 2014), which is also used for displaying localization information in the widely used GeneCards database (https://www.genecards.org/). COMPARTMENTS combines localization information from multiple resources like GO, UniProt, and Cell Atlas with automated text mining, summarized into a score (1–5) for each localization (Figure S4E).

Using the COMPARTMENTS data, focusing on seven well-defined discrete cellular locations (nucleus, cytosol, mitochondrion, endoplasmic reticulum [ER], Golgi apparatus, plasma membrane, and vesicles), we evaluated the evidence of multi-localization of proteins. In total, 36,856 entries were available for the seven locations, corresponding to 14,792 genes (Figures S4F–S4H). Using all entries covering the seven locations, 41% of the proteins were associated with more than one of the seven locations (Figure 2H). When using only score 5 entries, multi-localizing proteins dropped to 34%. In COMPARTMENTS, 4 different types of evidence can generate a score 5 entry: ''CURATED,'' manually curated entry; ''IDA,'' inferred from direct assay (including Cell Atlas data); ''TAS,'' traceable author statement; and ''NAS,'' non-traceable author statement (Figure S4I). When using only manually curated score 5 entries, multi-localizing proteins dropped to a mere 16% (Figure 2H).

(B) Agreement between neighborhood classifications and GO and UniProt based on annotation overlap.
(C) Cell Atlas localizations were binned to match the neighborhood classifications (left). Also shown is overlap between proteins assigned to a single localization in Cell Atlas and proteins with a generalizable neighborhood classification (right).
(D) Agreement between neighborhood classifications and Cell Atlas localizations based on Cell Atlas reliability score.
(E) SubCellBarCode (SCBC) can provide orthogonal validation for Cell Atlas localizations.
(F) Cumulative plots showing the minimum number of PSMs used for quantification of proteins classified at the neighborhood level in all 5 CLs in SubCellBarCode.
(G) Cell Atlas reliability score distribution for all localizations for proteins assigned to multiple or single locations in Cell Atlas.
(H) Pie charts of COMPARTMENTS database information regarding protein localization for subsets with different evidence levels-scores.
See also Figure S3 and S4.



(legend on next page)

The COMPARTMENTS analysis illustrates two important points. First, collection-type resources risk overestimation of multi-localization of proteins because of varying quality and sensitivity of the methods used, and second, limited transparency and non-standardized annotation systems make it difficult to evaluate the quality of the data. Overall, our investigation of multi-localizing proteins show that, irrespective of method used, there is a strong enrichment of low-confidence findings in the category of multi-localizing proteins. This results in a dramatic inflation of the number of proteins assigned to more than one subcellular localization.

Navigating the Resource by SubCellBarCodes and Localization Networks

The basic output of our analysis is the protein classification probabilities across the 15 compartments or 4 neighborhoods. For easier interpretation, these probabilities can be presented as a stacked barplot, where the sum of the 15 or 4 probabilities is always 1, and each compartment and neighborhood is represented by a defined color. Hereafter, these probability barplots are referred to as SubCellBarCodes or simply bar codes. As an example, EGFR was identified in all five cell lines, resulting in five compartment bar codes and five neighborhood bar codes, visualizing the probabilities of determined locations (Figure 3A). For quantitative robustness evaluation, the number of PSMs used for quantification of each protein is indicated on top of the bar codes. EGFR was classified to the secretory neighborhood in all five cell lines and to the plasma membrane or ER at the compartment level in a cell line-dependent manner. Additional example bar codes were plotted for well-known compartment markers, showing consistent classifications across cell lines (Figure 3B).

To generate an overview of the classifications, we created a consensus localization network by extracting all proteins with single neighborhood classification (9,858 proteins). Next, we grew the network to the compartment level for proteins with a single compartment classification (Figure 3C).

Using the network described above, we created sub-networks for different sets of regulatory proteins. Genes annotated in different regulatory categories were retrieved from various sources: transcription factors (TFs; 1,569 unique gene symbols retrieved from animal transcription factor database [animalTFDB]; Zhang et al., 2015), transcription co-factors (413 genes, animalTFDB), chromatin remodeling factors (129 genes, animalTFDB), protein kinases (514 genes, 2007 update of Manning et al., 2002), ubiquitin E3 ligases (614 genes; Li et al., 2008), and protein phosphatases (189 genes; Chen et al., 2017). The analytical depth of our resource is demonstrated by the excellent identification and classification coverage in each category, ranging from 96% of all human chromatin remodeling factors to 63% of human TFs (Figure S5A). Evaluation of the localization of regulatory proteins shows, as expected, that TFs, co-factors, and chromatin remodeling factors are most commonly localized in the nucleus but also that co-factors are often found in the cytosol, suggesting alternative functions or regulation through nuclear-cytoplasmic shuttling (Figure S5B).

Next, localization networks for each regulatory category were created for single location proteins, as described above (Figures 3D, 3E, and S5C–S5F). Closer analysis of the TF network reveals that the major compartments for transcription factors are N1 (nucleosol high density) and N4 (nucleosol) (Figure 3D). Many TFs known to shuttle between the nucleus and cytoplasm were, on the contrary, classified to cytosolic compartments, as exemplified by 6 of 6 STAT, 6 of 6 SMAD, and 5 of 5 nuclear factor κB (NF-κB) family members in the network. In the kinase network, 34 of 35 receptor tyrosine kinases and 7 of 7 SRC family kinases were classified to the secretory neighborhood (Figure 3E). These membrane-associated kinases commonly transmit signals through initiation of phosphorylation cascades, such as the mitogen-activated protein kinase (MAPK) cascade. Of the 46 MAPK proteins in the network, 43 were found in cytosolic compartments, consistent with their role in transmitting signals from the plasma membrane into the cell. Interestingly, three MAPK family members were found in secretory compartments (MAP3K12, MAP3K13, and TAOK2); all three were previously associated with axon biogenesis in neurons (Chen et al., 2016). Among nuclear kinases are regulators of DNA repair (ataxia telangiectasia mutated [ATM] and ATR), mitotic regulators (AURKB and AURKC), transcriptional regulators (BRD2–BRD4 and BRDT), and regulators of mRNA splicing (CLK1–CLK4). Only 7 kinases were classified to mitochondria, all of them from two different families (PDK1–PDK3 and ADCK1, ADCK2, ADCK4, and ADCK5).

Cell-Line-Specific Localization versus Cell-Line-Specific Expression

In our total lysate analysis, the overall identification coverage of the 9,016 proteins with a single location in GO/UniProt was 57%, but only 37% for secretory neighborhood proteins, dropping to a mere 21% for plasma membrane (PM) proteins (Figure 4A). The underrepresentation of plasma membrane proteins indicates that plasma membrane is a highly specialized cellular compartment and that, compared with other compartments, a much larger proportion of plasma membrane proteins is expressed in a cell-type-dependent manner. This finding was further supported by a significantly higher variation in protein expression across the five cell lines for secretory neighborhood proteins. (Figure 4B). Next, we evaluated cell type dependency in

Figure 3. SubCellBarCodes and Localization Networks

(A) SubCellBarCodes for EGFR, showing, for each CL, the probability output from the classifiers for compartments and neighborhoods. Classifications and number of PSMs used for quantification are indicated.

(B) SubCellBarCodes for compartment markers.

(C) Consensus localization network created based on the generated classifications. The number of proteins in each location is indicated.

(D) Localization network for 792 transcription factors.

(E) Localization network for 325 protein kinases.

See also Figure S5.

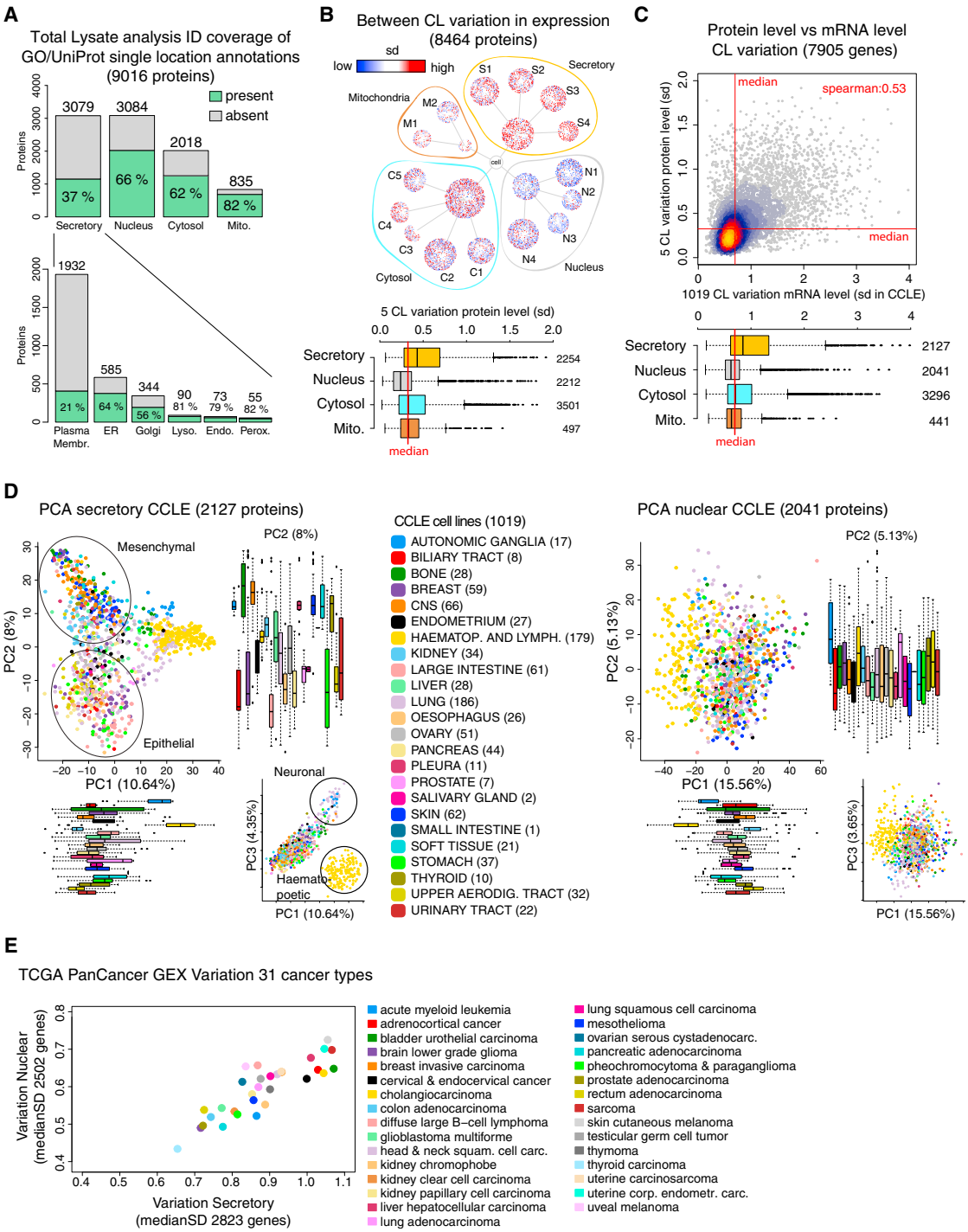


Figure 4. Cell-type-Specific Expression in Relation to Protein Localization

(A) Overlap between total lysate analysis in all five CLs and GO and/or UniProt single-location proteins by GO and/or UniProt annotation.

(B) Consensus localization network with proteins colored by variation in protein expression levels between five CLs, as determined by the total lysate analysis (top). Bottom: higher variation between CLs in secretory proteins and lower variation in nuclear proteins.

(C) Variation in protein levels between all five CLs plotted against the variation in mRNA expression between 1,019 CLs in the CCLE panel (top). Bottom: higher variation in mRNA expression for the secretory subset and lower variation in mRNA expression for the nuclear subset.

(legend continued on next page)

expression of proteins at different localizations using mRNA data from the Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012) as a proxy for protein expression because no large resource exists that contains proteomics data. A correlation analysis indicated that genes that show high variation between cell lines at the protein level in our data in general also show higher variation at the mRNA level between the 1,019 cell lines in CCLE (Figure 4C). When analyzing the genes according to our subcellular classifications, the variation in mRNA level across the CCLE cell lines was higher for the secretory neighborhood subset and lower for the nuclear subset.

Our analysis thus shows that expression of proteins in secretory compartments, and specifically at the plasma membrane, varies much more between cell lines than proteins in other compartments and that nuclear proteins are surprisingly stable in expression between different cell lines.

To investigate whether the expression variation between cell lines could be coupled to specific cell phenotypes, we performed PCA of all CCLE cell lines for secretory and nuclear subsets separately. For the secretory neighborhood subset, the first three components of the PCA clearly resolved cell lines on the 24 different origins of the cells (Figure 4D). Closer investigation of the genes with high importance to each of the components as well as the interrelationships between different cell origins indicated that cell lineage was an important determinant in the separation and, consequently, that different cell types can easily be distinguished by the expression level of genes coding for secretory neighborhood proteins. The same was not true for the PCA, based on the nuclear subset where only cells of hematopoietic origin was separated and, in addition, less well than using the secretory subset.

Finally, an analysis of the variation of mRNA expression in clinical samples from 31 different cancer types (The Cancer Genome Atlas [TCGA] and PanCancer dataset [Weinstein et al., 2013]), showed that, for each cancer type, the variation between clinical samples was much higher for the secretory neighborhood subset than for the nuclear subset (Figure 4E).

These results are supported by our recent analysis of transcription factor activity, where we show that mRNA or protein expression level analysis is insufficient to determine cell lineage and tissue type specificity of transcription factors and that activity-based assays were needed for such analysis (Wei et al., 2018). Conversely, the current analysis shows that overlaying our subcellular location information of secretory subset genes on either mRNA- or protein-level expression analysis can be used for determination of cell lineage or tissue origin. In summary, our data support that specific phenotypes of cells are driven by cell-specific expression of proteins rather than cell-specific localization.

Subcellular Distribution of Signal Peptides and Protein Domains

To evaluate and visualize the classification of proteins with domains or sequences expected to affect subcellular localization, we extracted annotations from UniProt to identify proteins with transmembrane (TM) domains, signal peptides (which target proteins to the secretory system), or transit peptides (which target proteins to mitochondria). Overlaying this information on the consensus localization network resulted in enrichment of each category in the expected locations (Figures 5A, S6A, and S6B).

The domain analysis was further extended by enrichment analysis in the single location network (Figure 3C) based on the Pfam database containing 4,624 different protein families (Finn et al., 2016). This analysis resulted in identification of 295 Pfam families significantly enriched ($q < 0.01$) more than 4-fold in specific compartments or neighborhoods (Figure 5B; Table S4). The most striking observation from this analysis was the dramatic enrichment of C2H2 zinc finger (ZF) proteins and the Krueppel-associated box (KRAB) proteins in nuclear compartment N1. ZF domains are interaction domains that bind to DNA, RNA, or other proteins, and C2H2-ZFs constitute the largest class of putative TFs (Najafabadi et al., 2015). At least one-third of human ZF proteins also include a KRAB that serves to recruit histone deacetylase complexes to regulatory regions of the genome, and, consequently, KRAB-ZF proteins function as transcriptional repressors (Huntley et al., 2006). To further evaluate the localization of different TF families, we overlaid annotation retrieved from animalTFDB on top of the TF network described above (Figure 5C). This analysis further supported the strong enrichment of C2H2-ZFs in compartment N1 and, specifically, for C2H2-ZFs also containing a KRAB domain (annotation from Huntley et al., 2006). Classification of KRAB-containing C2H2-ZF proteins into compartment N1 indicates that they are typically localized in the nucleosol as part of high-density structures or complexes. A general DNA-binding motif for C2H2-ZFs has proven difficult to determine (Jolma et al., 2013). Nevertheless, it has been demonstrated that KRAB-containing C2H2-ZF TFs show widespread binding to regulatory regions in DNA, and it has been suggested that cells must have mechanisms to overcome KRAB-based silencing of genes (Najafabadi et al., 2015). Our data suggest that KRAB-containing C2H2-ZF TFs are commonly part of large complexes and that such complex formation could be a way to regulate the effects of these transcriptional repressors on gene expression.

Protein Splice Variant-Specific Subcellular Localization

Alternative mRNA splicing is widespread, with as much as 95% of human multi-exon genes shown to generate more than one transcript variant (Pan et al., 2008). Ribosomal profiling-based analyses further indicate active translation of different variants through mRNA association with ribosomes (Weatheritt et al., 2016). However, for the vast majority of genes, the collective MS-based evidence only supports one protein variant, most commonly a translation of the longest transcript variant (Tress et al., 2017). The general lack of protein-level evidence for different variants could in part be explained by failure to detect

(D) Principal-component analysis (PCA) of 1,019 CLs in the CCLE panel based on gene expression of secretory subset (left) or nuclear subset (right) genes. CLs are colored by origin, and for each CL group, PC1 and PC2 distribution is indicated by boxplots. Also indicated for the secretory subset PCA are different cell types.

(E) Plot showing variation in gene expression for nuclear and secretory subsets in clinical cohorts of 31 different cancer types from the TCGA PanCancer dataset.

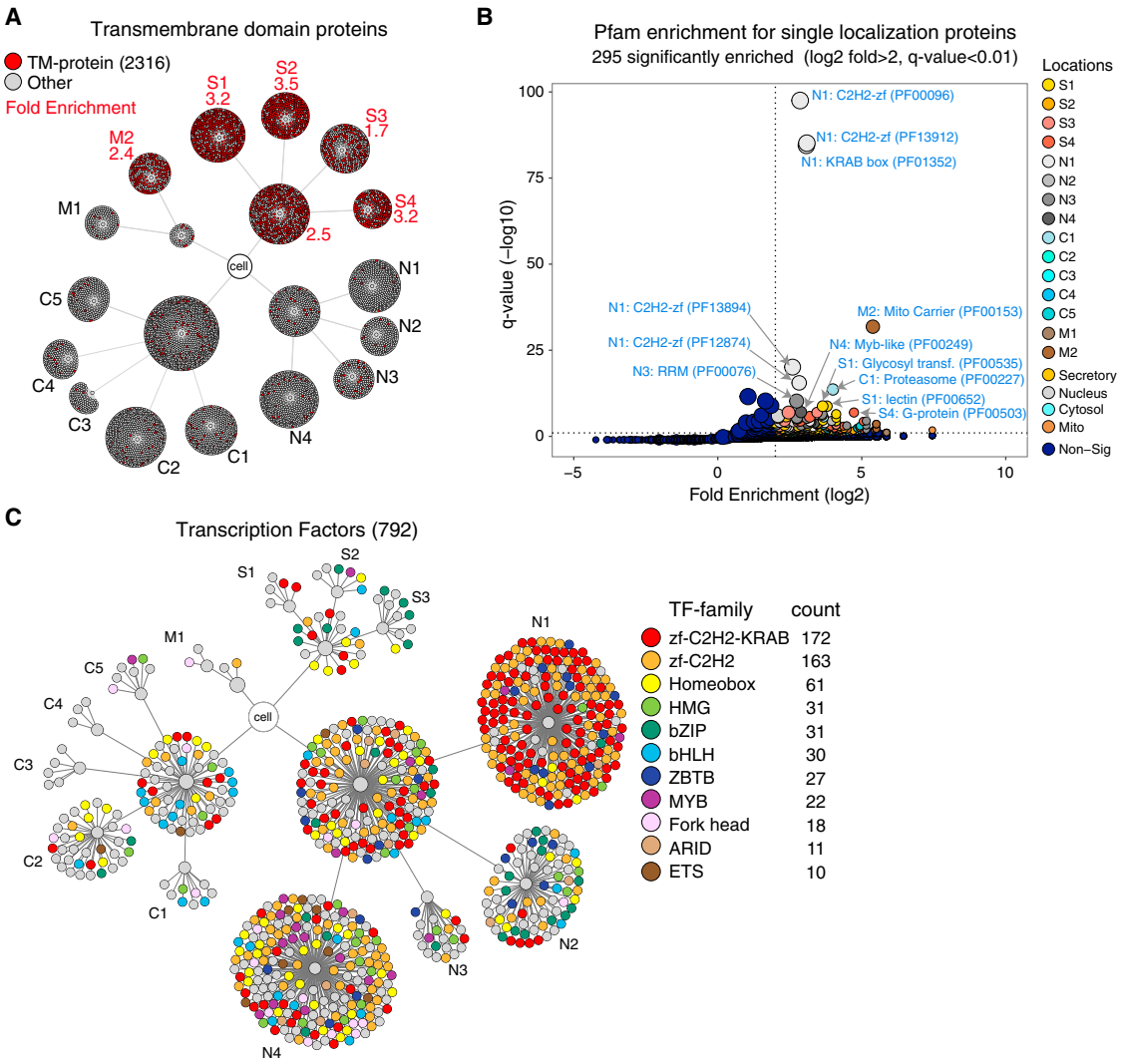


Figure 5. The Effect of Signal Peptides and Protein Domains on Subcellular Localization

(A) Localization of proteins with TM domains based on consensus network. Enriched locations ($p < 0.01$) are indicated.

(B) Volcano plot of Pfam enrichment analysis. The cutoff for fold enrichment (log2) was set to more than 2 ($q < 0.01$). The size of the circles corresponds to the number of Pfam proteins identified.

(C) Transcription factor localization network, colored by TF family.

See also Figure S6 and Table S4.

variant-specific peptides, caused by limited peptide coverage in the analysis. Nevertheless, strong evidence does exist that some genes do give rise to more than one protein. Functional differences of alternative proteins from the same gene are in most cases unknown, but it has been suggested that alternative splicing is important for localization of proteins (Kelemen et al., 2013). Based on our rich peptide-level data with subcellular resolution, we analyzed how widely splicing is used to direct protein products of the same gene to different cellular compartments.

We performed RNA sequencing (RNA-seq) of HCC827 cells and identified transcripts mapping to 10,608 genes (protein coding, TPM cutoff > 1). For genes overlapping with the localization analysis in HCC827 (8,879), the RNA-seq data were used to detect alternative splicing events. This variant analysis resulted in identification of 1,831 (20%) alternatively spliced genes (Figure 6A; Table S5). The fractionation profiles underlying the classification of subcellular localization are based on quantitative data from all peptides mapping to a specific gene. Thus, for genes with variants in different cellular compartments, a mixed fractionation profile would be generated, with a negative effect on classification. We were, however, unable to detect any differences between alternatively and non-alternatively spliced genes regarding classification frequency and probability score (Figure 6B). In addition, the alternatively spliced genes did not include more outlier peptides, as indicated by the within-gene peptide pairwise fractionation profile correlation (Figure 6B). In summary, this analysis strongly indicates that alternative splicing

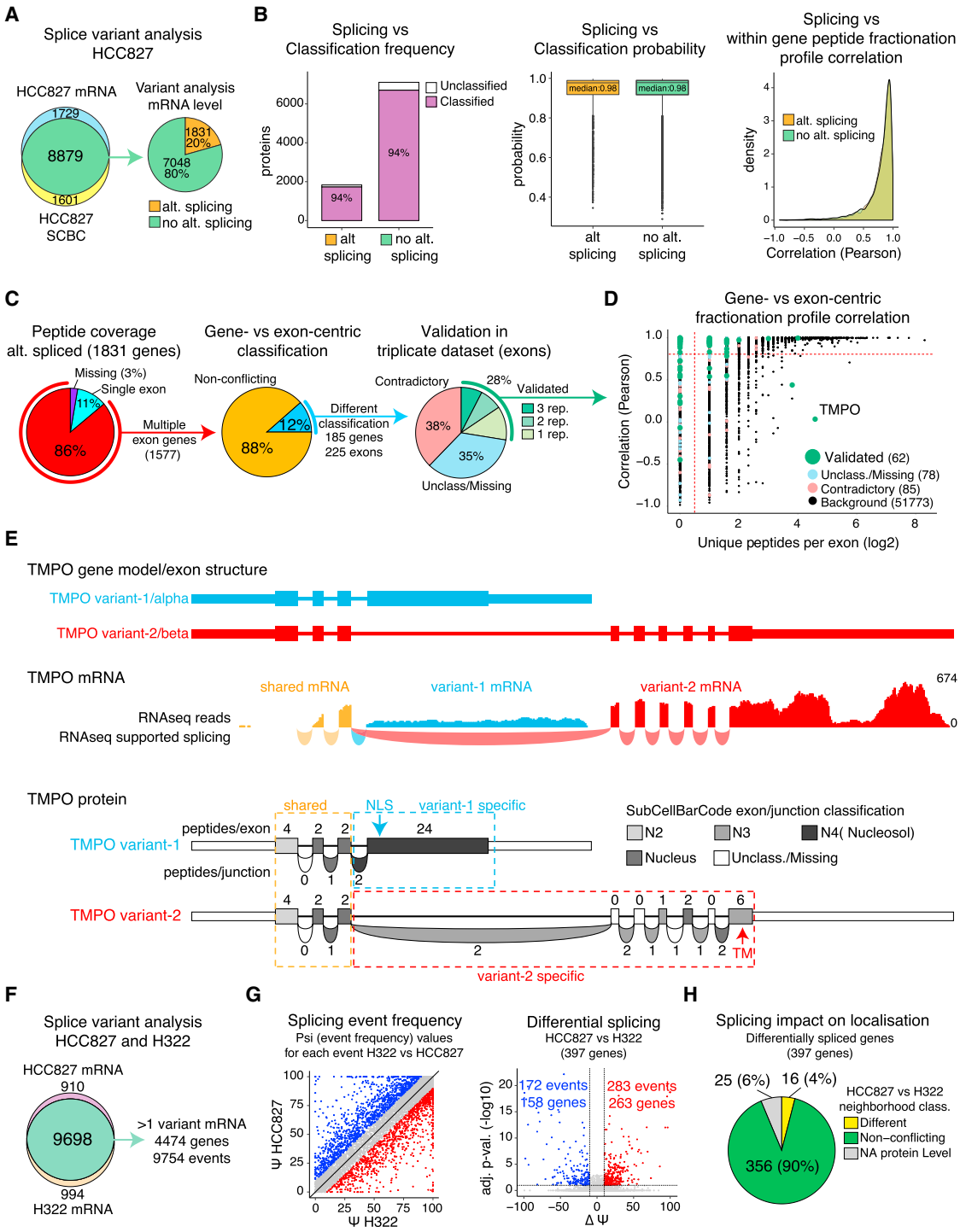


Figure 6. Variant-Specific Localization

(A) Identification overlap between RNA-seq analysis and subcellular localization in HCC827 cells (left). The pie chart (right) indicates the fraction of genes where alternative splicing was detected.

(B) No difference was found between alternatively spliced and not alternatively (alt.) spliced genes in the subcellular localization analysis, as shown for classification frequency (left), classification probability (center), and within gene peptide fractionation profile correlation (right).

(C) Pie chart of the peptide coverage of exons for alt. spliced genes in HCC827 cells (left). For genes with peptides identified from multiple exons, gene-centric was compared with exon-centric classification of localization (right).

(legend continued on next page)

is not a common mechanism for directing variants from the same gene to different subcellular locations within a specific cell.

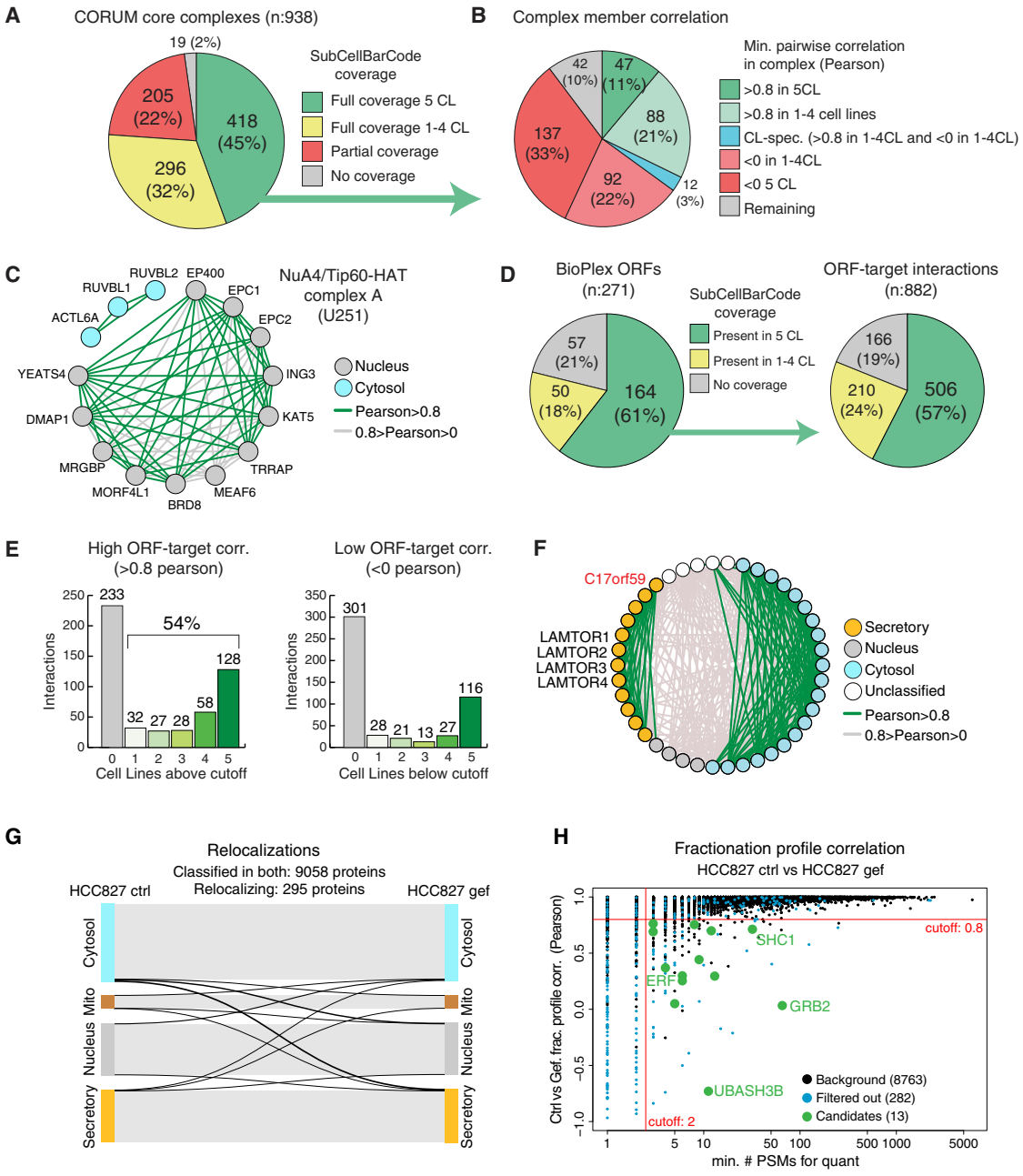
To investigate whether our method was sensitive enough to pick up individual examples of variant-specific localization, we focused on the 1,831 alternatively spliced genes. In 1,577 (86%) of these genes, peptides mapping to multiple exons were identified and quantified by MS (Figure 6C). For these genes, peptide-level data were used to calculate exon-centric fractionation profiles that were subsequently used for exon-centric classification of localization. Next, for each gene, the gene and the exon-centric classifications were compared, resulting in identification of 185 genes (11%), with exons classified to a separate location than the corresponding gene (Figure 6C; Table S5). By default, fewer PSMs for quantification are available for exons compared with genes, and, consequently, we expect a higher risk of false classification in the exon-centric analysis. Therefore, we also used the data from the HCC827 triplicate analysis and were able to validate 28% (62) of the findings on the exon level (Figure 6C). The candidates from the variant-specific analysis can also be evaluated based on the correlation between gene- and exon-centric fractionation profiles and the number of peptides per exon used for quantification (Figure 6D). Based on this analysis, the best candidate from the variant-specific analysis was the gene TMPO, with known variants containing a TM domain or a nuclear localization signal (NLS) (Dechat et al., 2000). Our analysis shows that exons and junctions mapping exclusively to the TM variant are classified to compartment N3, whereas the exons and junctions mapping to the non-TM-variant with a NLS are classified to the nucleosol (N4; Figure 6E; Table S5).

Even though our analysis shows that variant-specific localization is rare within a specific cell, differential splicing of genes could potentially explain differences in protein localization between cell lines. To evaluate this, we performed RNA-seq on H322 cells and compared the results with the analysis performed in HCC827 cells. Using the same cutoffs for H322 cells, transcripts mapping to 9,698 genes were identified in both cell lines. For the overlapping genes, variant analysis was performed, resulting in the identification of 9,754 splicing events in 4,474 genes in at least one cell line (Figure 6F; Table S5). Comparing the splicing event frequency (Ψ, 100 meaning that an event always occurs and 0 that it never occurs) between the cell lines indicated limited differences, with 397 genes differentially spliced ($|\Delta\Psi|>10$, adjusted p value [adj.p] < 0.1) between cell lines (Figure 6G). The neighborhood classification indicated that 448 proteins were differentially localized between HCC827 and H322 cells. Of these, only 16 proteins were differentially spliced (Figure 6H; Table S5). This result shows that differences in localization between cell lines are not generally explained by differences in splicing, and, conversely, that the functional effect of differential splicing between cell lines rarely relates to altered subcellular localization.

Protein Interactions and Complexes

Protein correlation profiling has been shown previously to be useful for evaluation of protein colocalization and protein complex formation (Havugimana et al., 2012). To evaluate this, we investigated the subcellular distribution of complexes from the Comprehensive Resource of Mammalian Protein Complexes (CORUM) database in our data (Ruepp et al., 2010). Of 938 investigated CORUM complexes, we observed complete coverage in at least one cell line for 76% and in all five cell lines for 45% of the complexes (Figure 7A). To evaluate colocalization of the complex members, pairwise fractionation profile correlations were calculated for all complex members (Table S6). As expected, the pairwise correlation was higher for complex members compared with a random sampling of non-complex members (Figure S7A). In the 418 complexes completely identified in all five cell lines, 32% exhibited high correlation (>0.8) between all members in at least one cell line, indicating distinct colocalization (Figures 7B and S7B). This result is in line with the output from large-scale affinity purification-MS in BioPlex (Huttlin et al., 2015). When at least two affinity-tagged CORUM complex members were used as bait to identify interacting proteins, 1 of 3 of the complexes achieved at least 90% coverage. Interestingly, we also found that a large proportion of the complexes (33%) included complex members with negative correlation to other complex members in all five cell lines (Figure 7B). For example, NuA4/Tip60 HAT complex A is responsible for histone acetylation, and the majority of its members are localized in the nucleus (Figures 7C and S7C). Three of the complex members (RUVBL1, RUVBL2, and ACTL6A) were, however, consistently poorly correlated with the remaining members of the complex in all cell lines and classified as cytosolic. Another example is the exosome complex, where DIS3 was identified as a localization outlier. In line with this, the same outliers were not identified as complex interactors in BioPlex (Figure S7C). In a few cases, the analysis also indicated cell line-specific complex member usage, as shown for the NF-κB-tumor necrosis factor α (TNF-α) complex, where TNIP2 showed high correlation with other complex members in all cell lines except for U251 (Figure S7D). It has been proposed that complexes in many cases are composed of smaller modules of more tightly interacting proteins and that these modules can be combined in different ways to regulate complex functions (Malovannaya et al., 2011). Here we present data that support this notion, as shown in the

(D) Correlation between exon- and gene-centric fractionation profiles plotted against the number of unique peptides per exon for all exons. As background are genes identified with peptides mapping to more than one exon. Candidates identified in (C) as well as the gene TMPO are indicated.
(E) Exon-intron structure for TMPO variant 1 (blue) and 2 (red), as annotated in the NCBI Reference Sequence Database (RefSeq; top). Also shown are RNA-seq reads mapping to TMPO and splicing supported by the RNA-seq analysis (center) and exon- and junction-based classification of localization for TMPO (bottom). The number of unique peptides used for quantification of each exon or junction is indicated. NLS, nuclear localization signal; TM, transmembrane.
(F) Identification overlap of RNA-seq analysis in HCC827 and H322 cells and, for overlapping genes, the number of alternative splicing events detected.
(G) Scatterplot of splicing event frequency in HCC827 and H322 cells (left). Also shown is a volcano plot describing differences in splicing between HCC827 and H322 cells (right).
(H) Pie chart of the fraction of differentially spliced genes that were classified to different subcellular localizations.
See also Table S5.



(legend on next page)

examples above, and provide a means for unbiased investigation of complex module usage.

BioPlex contributed interaction data for 271 poorly characterized open reading frames (ORFs), and to further build on this information, we evaluated the colocalization of the ORFs also found in our data and their suggested interaction partners (Table S6). Similar to CORUM complexes, ORF-target correlations were higher than a random sampling of pairwise correlations (Figure S7E). For ORFs and targets detected in all five cell lines, 54% of the ORF-target interactions showed high pairwise fractionation profile correlation in at least one cell line, with 25% in all five cell lines. In 23% of cases, ORF-target correlations were negative in all five cell lines (Figures 7D and 7E). For many ORFs with several suggested interaction partners, some were colocalizing with the ORF whereas others were not. As an example, C17orf59 has multiple reported interaction partners in BioPlex, and our analyses classify C17orf59 as a secretory neighborhood protein together with a subset of the reported interactors (Figure 7F). C17orf59 interacts with the Ragulator complex at the lysosomal membrane (Schweitzer et al., 2015). Our data thus support lysosomal localization of the complex but only some of the C17orf59 interactions reported in BioPlex.

Most protein-protein interaction studies include an initial step where cells are lysed, and, consequently, proteins in all subcellular compartments are mixed. This introduces a non-physiological setting with the potential risk of creating protein interactions that are not biologically relevant. Even though lack of colocalization is not sufficient to reject potential interactions, the SubCellBarCode data form an easily queried orthogonal information resource for refining complex analysis and can help rank candidate interactions from protein-protein interaction studies.

Condition-Dependent Relocalization Analysis

Given the reproducibility of our method, we also tested whether it can be used for proteome-wide relocalization analysis by studying the response to EGFR inhibition in the EGFR-mutated lung cancer cell line HCC827. EGFR mutations are prototypic oncogenic driver alterations in cancer, and EGFR tyrosine kinase inhibitors are commonly used in cancer therapy.

Neighborhood classifications for treated cells were first generated and compared with neighborhood classifications already generated for untreated HCC827 cells, resulting in identification of 295 candidate relocalizing proteins (Figure 7G; Table S7). The majority of these relocalizations (72%) involve the cytosol neighborhood, indicating a function of the cytosol as a reservoir for proteins that can be targeted for specific subcellular compartments under specific conditions. After filtering the candidate list based on PSMs, fractionation profile correlation and reproducibility, 13 candidate relocalizing proteins remained (Figure 7H). Among these were several proteins directly connected to EGFR signaling (Figure 7I). GRB2 and SHC1 are prototype adaptor proteins that interact with activated EGFR and help recruit other proteins needed to transmit signals (Yarden and Sliwkowski, 2001). UBASH3B also binds to active EGFR complexes and inhibits receptor ubiquitination and degradation (Kowanetz et al., 2004). All of these proteins were found to relocalize from the secretory neighborhood to the cytosol, well in line with their described functions. ERF, which relocalizes from the cytosol to the nucleus, is a transcriptional repressor that inhibits several TFs, including MYC (Mavrothalassitis and Ghysdael, 2000). When EGFR is active, the MAPK pathway protein ERK phosphorylates ERF, resulting in cytoplasmatic sequestration. Upon EGFR inhibition, ERF is dephosphorylated and can relocalize to the nucleus, where it can exert its function.

In summary, the condition-dependent relocalization analysis presented here enables unbiased investigation of temporal-spatial regulation of proteins, clearly revealing expected events as well as providing a novel hypothesis for further testing.

Perspectives

The spatial organization of the proteome is highly complex, influenced by protein expression levels, protein stability, sequence variants, post-translational modification status, and interactions with a multitude of biomolecules, all in interplay with the intra- and extracellular environment. To elucidate this biology and create a comprehensive view of the subcellular proteome, multiple techniques and resources should be used in concert because of the advantages and shortcomings inherent to each method. Much of the currently available subcellular localization data are held in the public domain by data curation consortia like the Gene Ontology and UniProt efforts. However, because of the nature of these resources, the data are a myriad of information from a variety of cell types and techniques and with a wide range of evidence levels, from mere associations to traceable experimental results. Consequently, cumbersome validation is often needed, and meta-analysis efforts can be impaired by noisy data.

Here we present data constituting the largest single-source information resource for subcellular localization of proteins, with classification of proteins mapping to more than 12,000 genes in five different cell lines. Our method relies on accurate MS-based identification of proteins and quantification of the relative abundance of each protein between subcellular fractions. Our

Figure 7. Colocalization and Relocalization Analysis

(A) Coverage of CORUM complexes in the SubCellBarCode dataset.

(B) CORUM complex member pairwise fractionation profile correlations for complexes found in all five CLs.

(C) NuA4/Tip60 HAT complex member localization classification in U251 cells.

(D) Coverage of ORFs from BioPlex in SubCellBarCode (left) and for ORFs found in all five CLs and the coverage of their reported interactions from BioPlex (right).

(E) ORF-target correlation for interaction pairs identified in all 5 CLs. Barplots show the number of cell lines with high (left) or low (right) ORF-target correlation.

(F). Neighborhood classifications and correlations in HCC827 cells of proteins reported as C17orf59-interacting in BioPlex.

(G) Sankay plot showing differences in neighborhood classifications between untreated (left) and gefitinib-treated (right) HCC827 cells.

(H) Protein fractionation profile correlations between untreated and gefitinib-treated cells plotted against the minimum number of PSMs used for quantification of each protein. Proteins that passed the filtering cutoffs are indicated in green and the proteins that failed in blue. Proteins discussed further in the text are indicated.

(I) Bar codes for selected proteins identified as relocalizing.

See also Figure S6 and Tables S6 and S7.

resource provides, for the first time, an opportunity to mine the proteome-wide subcellular localization of proteins across five different cell lines in data generated from a uniform experimental setting.

The benefit of combining data sources was recently demonstrated in a meta-analysis of subcellular proteomics data from 11 different studies (Lund-Johansen et al., 2016). Our data, in combination with Cell Atlas data as well as with UniProt and GO resources, form a solid multisource core of subcellular localization knowledge. Further, the robustness of our method allows scalable expansion of the resource.

Protein variants generated by, for instance, alternative splicing can show differential subcellular localization, but the scale of such functional regulation is not known. A prerequisite for proteome-wide variant-specific analysis is that the analytical depth of the MS not only provides quantitative information regarding a large number of proteins but also that the protein sequence coverage (i.e., the number of unique peptides identified per protein) is high. We demonstrate that the analytical depth provided by the HiRIEF-LC-MS method is sufficient to start evaluating the usage of different variants at different subcellular locations on a large scale. In extension, the sensitivity of variant-dependent analysis can be expanded by incorporating additional omics data layers to investigate the effect of post-translational modifications, somatic mutations, and germline genetic variation on subcellular protein localization.

Protein shuttling and relocalization are fundamental for cellular signaling and rapid adaptation to environmental changes. In a proof-of-principle experiment, we demonstrate the power of the method in a proteome-wide analysis of protein relocalization in response to EGFR inhibition. The possibility to perform proteome-wide investigation of condition-dependent localization will enable investigation of the molecular response to a wide range of perturbations, adding important new knowledge.

In conclusion, we present the largest single-source resource of subcellular organization of the proteome. Based on the generated data, we provide new information about the architecture of cells and the complexity of the spatial organization of the proteome. As examples, our data show that the vast majority of proteins have a single main subcellular location, that alternative splicing is rarely used to direct protein variants to specific subcellular locations, and that cells are much more different in expression of proteins exposed to the surrounding environment than in expression of intracellular proteins. Together with the proteome-wide relocation analysis, our data serve as a foundation for continued investigations into defining the key regulatory proteome. The SubCellBarCode resource is freely available at www.subcellbarcode.org.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell culture
- METHOD DETAILS
  - Subcellular fractionation
  - Western blot analysis
  - Sample preparation for MS
  - Peptide level sample fractionation through HiRIEF
  - MS-based quantitative proteomics
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Peptide and protein identification
  - Selection of Marker Proteins
  - Dimension Reduction
  - Clustering of the core proteins
  - Classification of proteins
  - Domain Enrichment Analysis
  - Annotation of marker protein clusters
  - Network visualization
  - Protein complex analysis
  - RNA sequencing and splice variant specific localization analysis
  - Relocalization analysis
  - Comparisons to additional datasets
- DATA AND SOFTWARE AVAILABILITY
- ADDITIONAL RESOURCES
  - SubCellBarCode portal

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.11.035.

ACKNOWLEDGMENTS

We are grateful to Prof. David Fenyö (Institute for Systems Genetics, NYU) for critical reading of the manuscript. We gratefully acknowledge funding from the Swedish Foundation for Strategic Research, the Swedish Cancer Society, the Swedish Research Council, the Swedish Childhood Cancer Foundation, an AstraZeneca research grant, the Cancer Research Funds of Radiumhemmet, and Stockholm's County Council (Avtal om läkarutbildning och forskning [ALF] funding). We gratefully acknowledge facility support from Swedish National Infrastructure for Biological Mass Spectrometry (BioMS) and Science for Life Laboratory (SciLifeLab).

AUTHOR CONTRIBUTIONS

L.M.O. and J.L. conceived the study. M.V. and Y.P. performed cell culture, subcellular fractionation, and MS analyses. Y.P. performed western blot and experimental reproducibility analyses. T.A., O.F., E.F., and L.M.O. performed bioinformatics related to the classification of protein localization. A.F.W. built the web resource. M.V. and L.M.O. evaluated the data in relation to other resources. T.A., E.F., and L.M.O. performed domain-driven localization analysis. Y.Z., T.A., and L.M.O. performed variant-specific localization analysis. L.M.O. performed cell-line-specific localization-expression and condition-dependent localization analysis. L.M.O., M.V., E.F., and J.L. wrote the paper. All authors contributed to finalizing the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 14, 2018
Revised: September 28, 2018
Accepted: November 27, 2018
Published: January 3, 2019

REFERENCES

Andersen, J.S., Wilkinson, C.J., Mayor, T., Mortensen, P., Nigg, E.A., and Mann, M. (2003). Proteomic characterization of the human centrosome by protein correlation profiling. Nature *426*, 570–574.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R., and Jensen, L.J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford) *2014*, bau012.

Branca, R.M., Orre, L.M., Johansson, H.J., Granholm, V., Huss, M., Pérez-Bercoff, Å., Forshed, J., Käll, L., and Lehtiö, J. (2014). HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat. Methods *11*, 59–62.

Chen, M., Geoffroy, C.G., Wong, H.N., Tress, O., Nguyen, M.T., Holzman, L.B., Jin, Y., and Zheng, B. (2016). Leucine Zipper-bearing Kinase promotes axon growth in mammalian central nervous system neurons. Sci. Rep. *6*, 31482.

Chen, M.J., Dixon, J.E., and Manning, G. (2017). Genomics and evolution of protein phosphatases. Sci. Signal. *10*, eaag1796.

Christoforou, A., Mulvey, C.M., Breckels, L.M., Geladaki, A., Hurrell, T., Hayward, P.C., Naake, T., Gatto, L., Viner, R., Martinez Arias, A., and Lilley, K.S. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. Nat. Commun. *7*, 8992.

Dechat, T., Vlcek, S., and Foisner, R. (2000). Review: lamina-associated polypeptide 2 isoforms and related proteins in cell cycle-dependent nuclear structure dynamics. J. Struct. Biol. *129*, 335–345.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44* (D1), D279–D285.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., et al. (2012). A census of human soluble protein complexes. Cell *150*, 1068–1081.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Huntley, S., Baggott, D.M., Hamilton, A.T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. *16*, 669–677.

Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex Network: a systematic exploration of the human interactome. Cell *162*, 425–440.

Itzhak, D.N., Tyanova, S., Cox, J., and Borner, G.H. (2016). Global, quantitative and dynamic mapping of protein subcellular localization. eLife *5*, e16950.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE *9*, e98679.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene *514*, 1–30.

Kowanetz, K., Crosetto, N., Haglund, K., Schmidt, M.H.H., Heldin, C.H., and Dikic, I. (2004). Suppressors of T-cell receptor signaling Sts-1 and Sts-2 bind to Cbl and inhibit endocytosis of receptor tyrosine kinases. J. Biol. Chem. *279*, 32786–32795.

Li, W., Bengtson, M.H., Ulbrich, A., Matsuda, A., Reddy, V.A., Orth, A., Chanda, S.K., Batalov, S., and Joazeiro, C.A. (2008). Genome-wide and functional annotation of human E3 ubiquitin ligases identifies MULAN, a mitochondrial E3 that regulates the organelle's dynamics and signaling. PLoS ONE *3*, e1487.

Liu, X., and Fagotto, F. (2011). A method to separate nuclear, cytosolic, and membrane-associated signaling molecules in cultured cells. Sci. Signal. *4*, pl2.

Lund-Johansen, F., de la Rosa Carrillo, D., Mehta, A., Sikorski, K., Inngjerdingen, M., Kalina, T., Røysland, K., de Souza, G.A., Bradbury, A.R., Lecrevisse, Q., and Stuchly, J. (2016). MetaMass, a tool for meta-analysis of subcellular proteomics data. Nat. Methods *13*, 837–840.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the human endogenous coregulator complexome. Cell *145*, 787–799.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912–1934.

Mavrothalassitis, G., and Ghysdael, J. (2000). Proteins of the ETS family with transcriptional repressor activity. Oncogene *19*, 6524–6532.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Romero, J.P., Muniategui, A., De Miguel, F.J., Aramburu, A., Montuenga, L., Pio, R., and Rubio, A. (2016). EventPointer: an effective identification of alternative splicing events using junction arrays. BMC Genomics *17*, 467.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Schweitzer, L.D., Comb, W.C., Bar-Peled, L., and Sabatini, D.M. (2015). Disruption of the Rag-Ragulator Complex by c17orf59 Inhibits mTORC1. Cell Rep. *12*, 1445–1455.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. *8*, 289–317.

Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*, eaal3321.

Tress, M.L., Abascal, F., and Valencia, A. (2017). Alternative splicing may not be the key to proteome complexity. Trends Biochem. Sci. *42*, 98–110.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Weatheritt, R.J., Sterne-Weiler, T., and Blencowe, B.J. (2016). The ribosome-engaged landscape of alternative splicing. Nat. Struct. Mol. Biol. *23*, 1117–1123.

Wei, B., Jolma, A., Sahu, B., Orre, L.M., Zhong, F., Zhu, F., Kivioja, T., Sur, I., Lehtiö, J., Taipale, M., and Taipale, J. (2018). A protein activity assay to measure global transcription factor activity reveals determinants of chromatin accessibility. Nat. Biotechnol. *36*, 521–529.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M.; Cancer Genome

Atlas Research Network (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Yarden, Y., and Sliwkowski, M.X. (2001). Untangling the ErbB signalling network. Nat. Rev. Mol. Cell Biol. *2*, 127–137.

Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., Jia, H., Xue, Y., and Guo, A.Y. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. *43*, D76–D81.

Zhu, Y., Hultin-Rosenberg, L., Forshed, J., Branca, R.M., Orre, L.M., and Lehtiö, J. (2014). SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. Mol. Cell. Proteomics *13*, 1552–1562.

Zhu, Y., Orre, L.M., Johansson, H.J., Huss, M., Boekel, J., Vesterlund, M., Fernandez-Woodbridge, A., Branca, R.M.M., and Lehtiö, J. (2018). Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat. Commun. *9*, 903.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-GAPDH	Santa Cruz Biotechnology	CAT#sc-69778; RRID: AB_1124759
Mouse anti-α-Tubulin	Santa Cruz Biotechnology	CAT#sc-5286; RRID: AB_628411
Mouse anti-RanBP3	Santa Cruz Biotechnology	CAT#sc-373678; RRID: AB_10918083
Mouse anti-LMNA	Cell Signaling Technology	CAT#4C11; RRID: AB_10545756
Rabbit anti-GOLGB1	Abcam	CAT#ab93281; RRID: AB_10562970
Rabbit anti-PSMA2	Cell Signaling Technology	CAT#D3A4
Rabbit anti-RPS3	Cell Signaling Technology	CAT#D50G7; RRID: AB_10839122
Rabbit anti-EGFR	Cell Signaling Technology	CAT#D38B1; RRID: AB_10828841
Rabbit anti-CANX	Cell Signaling Technology	CAT#C5C9; RRID: AB_10827903
Rabbit anti-LAMP2	Cell Signaling Technology	CAT#D5C2P
Rabbit anti-ATP1B1	Cell Signaling Technology	CAT# D8W8J
Rabbit anti-Histone H4	Santa Cruz Biotechnology	CAT#sc-8658-R; RRID: AB_2011538
Rabbit anti-TOM20	Santa Cruz Biotechnology	CAT#sc-11415; RRID: AB_2207533
Chemicals, Peptides, and Recombinant Proteins		
Gefitinib	Selleckchem	Cat# S1025; CAS: 184475-35-2
Digitonin	Sigma-Aldrich	Cat# D141; CAS: 11024-24-1
Deposited Data		
RAW-files and Galaxy-search output	This paper	ProteomeXchange: PXD006895
RNaseq raw reads (fastq)	This paper	NCBI SRA: SRP154280
Experimental Models: Cell Lines		
A431	DSMZ	CAT#ACC 91; RRID: CVCL_0037
MCF7	ATCC	Cat#HTB-22; RRID: CVCL_0031
NCI-H322	ECACC	Cat# 95111734; RRID: CVCL_1556
HCC827	ATCC	Cat#CRL-2868; RRID: CVCL_2063
U251	Laboratory of Emma Lundberg	RRID: CVCL_0021
Software and Algorithms		
Gephi 0.9.1	Jacomy et al., 2014	https://gephi.org/
SpliceVista.py	Zhu et al., 2014	https://github.com/yafeng/SpliceVista
map_peptide2genome.py	Zhu et al., 2014	https://github.com/yafeng/SpliceVista
Rtsne	van der Maaten and Hinton, 2008	https://github.com/jkrijthe/Rtsne
mClust	Scrucca et al., 2016	https://github.com/cran/mclust
Caret package		https://github.com/topepo/caret
Biomart		https://bioconductor.org/packages/release/bioc/html/biomaRt.html
networkD3	N/A	https://cran.r-project.org/web/packages/networkD3/
pyensembl	N/A	https://github.com/openvax/pyensembl
Other		
Resource website for the SubcellBarcode publication	This paper	www.subcellbarcode.org

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Janne Lehtiö (janne.lehtio@ki.se).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell culture

All cell lines were cultivated at 37°C in a 5% CO_2 humidified environment in the following growth media: Roswell Park Memorial Institute-1640 medium (NCI-H322 and HCC827; Sigma-Aldrich); Dulbecco's Modified Eagle Medium (A431 and MCF7; Sigma-Aldrich); Eagle's Minimal Essential Medium (U251; Sigma-Aldrich); All media were supplemented with 10% fetal bovine serum (FBS, Sigma-Aldrich) and 1% penicillin/streptomycin (Sigma-Aldrich). A431, MCF7 and HCC827 are female cells, while U251 and NCI-H322 are male. All cell lines were tested and found mycoplasma-free using MycoAlert mycoplasma detection kit (Lonza, Walkersville, MD, USA).

METHOD DETAILS

Subcellular fractionation

To enable true relative quantification between subcellular fractions, minimize the risk of erroneous assignment of subcellular localization of proteins and ensure maximum proteome coverage, a protocol was designed so that all five fractions derive from the same starting material (cells), without any loss of material i.e., no discarded supernatants or pellets. Cells were seeded the day before fractionation and grown to 70%–80% confluency in two (biological duplicates) 150 mm diameter plates in parallel, except A431 which was seeded in six (three plates per biological duplicate) 100 mm diameter plates. For Gefitinib treatment cells were incubated with 2.5 μM Gefitinib (Selleckchem, Houston, TX, USA) for 2h before subcellular fractionation, controls were left untreated. The fractionation procedure was adapted from Liu and Fagotto (2011). Briefly, cells were washed with ice-cold PBS and incubated with gentle rocking at 4°C with a solution containing: 42 μg/ml digitonin (Sigma-Aldrich), 2 mM dithiothretiol (DTT), 2 mM $MgCl_2$, 150 mM NaCl, 200 μM EDTA and 20 mM HEPES, pH 7.6. After 7 min of incubation the solution was collected and stored as fraction FS1. Cells were washed twice in ice cold PBS and scraped down in a hypotonic solution containing: 20 mM HEPES pH 7.6, 200 μM EDTA and 1xHalt protease inhibitor cocktail (Thermo Fischer Scientific, San Jose, CA, USA). The cell solution was homogenized on ice with 50 strokes in a dounce homogenizer (Dounce tissue grinder pestle, small clearance, Sigma-Aldrich, P1110-1EA) followed by addition of an equal volume of solution containing 300 mM NaCl, 200 mM HEPES pH 7.6, 200 μM EDTA, 1 mM DTT and 1xHalt protease inhibitor cocktail, and an additional 50 strokes in order to break up cellular organelles. The resulting suspension was centrifuged 600x*g* for 10 min at 4°C. Pellets were stored as fraction FP1. Supernatants were transferred to new tubes and centrifuged at 3500x*g* for 10 min at 4°C. Pellets were stored as fraction FP2 and supernatants were transferred to ultracentrifuge tubes and centrifuged at 100,000x*g*, 4°C for 1 h. The resulting supernatants were stored as fraction FS2 and the pellets as fraction FP3.

Western blot analysis

Subcellular fractions (25μg) from HCC827 cells were separated in SDS/PAGE gels and transferred to nitrocellulose membranes (Millipore). After blotting membranes were blocked in 5% non-fat skim milk in Tris-Buffered Saline (TBS) containing 0.1% Tween 20. Membranes were incubated with one or more of the following antibodies as specified in the figures and figure legends; mouse anti-GAPDH (Santa Cruz), mouse anti-α-Tubulin (Santa Cruz), mouse anti-RanBP3 (Santa Cruz), mouse anti-LMNA (Cell Signaling), rabbit anti-GOLGB1 (Abcam), rabbit anti-PSMA2 (Cell Signaling), rabbit anti-RPS3 (cell Signaling), rabbit anti-EGFR (Cell Signaling), rabbit anti-CANX (Cell Signaling), rabbit anti-LAMP2 (Cell Signaling), rabbit anti-ATP1B1 (Cell Signaling), rabbit anti-Histone H4 (Santa Cruz), rabbit anti-TOM20 (Santa Cruz) followed, when necessary with incubation with the appropriate HRP-conjugated secondary antibody (GE Healtcare). Membranes were visualized with the ECL western blotting detection system (Pierce) according to the manufacturer's instruction. SDS/PAGE was carried out in the Surelock XCell (Life Technologies) and blotting was performed with the same system or with the iBlot system (Life Technologies) according to the manufacturer's instructions.

Sample preparation for MS

The pellet fractions obtained from the subcellular fractionation steps were lysed by addition of lysis buffer: 4% SDS, 1 mm DTT and 25 mm HEPES pH 7.6, followed by heating to 95°C for 5 min and 1 min sonication of the lysate. The supernatants were untreated. Protein concentration was determined by Bio-Rad DCC protein assay. For each cell line, the resulting 10 fractions (5 fractions in duplicates) were digested by a modified FASP-protocol (Branca et al., 2014). 250 μg of each fractions was mixed with 1 mM DTT, 8 M urea, 25 mM HEPES, pH 7.6 and transferred to a 10-kDa cut-off centrifugation filtering unit (Pall, Nanosep®), and centrifuged at 14,000x*g* for 15 min. Proteins were alkylated by 50 mM iodoacetamide (IAA) in 8 M urea, 25 mM HEPES for 10 min. The proteins were then centrifuged at 14,000x*g* for 15 min followed by 2 more additions and centrifugations with 8 M urea, 25 mM HEPES. Proteins were digested at 37°C with gentle shaking overnight by addition of Lys-C (enzyme:protein = 1:50, Wako Pure Chemical Industries) in 500 mM Urea, 50 mM HEPES pH 7.6 followed by an additional overnight digestion with trypsin (enzyme:protein = 1:50, Thermo Fisher

Scientific) in 50 mM HEPES, pH 7.6. The filter units were centrifuged at 14,000x*g* for 15 min followed by another centrifugation with MilliQ water and the flow-through was collected. Peptide concentration was determined by the Bio-Rad DCC assay and 100 μg of peptides from each digested fraction was labeled with TMT 10-plex reagent according to the manufacturer's protocol (Thermo Scientific). Labeled samples were pooled, cleaned by strata-X-C-cartridges (Phenomenex) and dried in a Speed-Vac.

Peptide level sample fractionation through HiRIEF

The TMT labeled peptides, 300 μg, were separated by immobilized pH gradient - isoelectric focusing (IPG-IEF) on pH 3.7-4.9 and 3-10 strips using the HiRIEF method as described previously (Branca et al., 2014). Peptides were extracted from the strips by a prototype liquid handling robot, supplied by GE Healthcare Bio-Sciences AB. A plastic device with 72 wells was put onto each strip and 50 μl of MilliQ water was added to each well. After 30 min incubation, the liquid was transferred to a 96 well plate and the extraction was repeated 2 more times with 35% acetonitrile (ACN) and 35% ACN, 0.1% formic acid in MilliQ water, respectively. The extracted peptides were dried in Speed-Vac and dissolved in 3% ACN, 0.1% formic acid.

MS-based quantitative proteomics

Extracted peptide fractions were separated using an Ultimate 3000 RSLCnano system coupled to a Q Exactive or Q Exactive HF (Thermo Fischer Scientific, San Jose, CA, USA). Samples were trapped on an Acclaim PepMap nanotrap column (C18, 3 μm, 100Å, 75 μm x 20 mm, Thermo Scientific), and separated on an Acclaim PepMap RSLC column (C18, 2 μm, 100Å, 75 μm x 50 cm, Thermo Scientific). Peptides were separated using a gradient of mobile phase A (5% DMSO, 0.1% FA) and B (90% ACN, 5% DMSO, 0.1% FA), ranging from 6% to 37% B in 30-90 min (depending on IPG-IEF fraction complexity) with a flow of 0.25 μl/min. The Q Exactive was operated in a data dependent manner, selecting top 10 precursors for fragmentation by HCD. The survey scan was performed at 70,000 resolution from 400-1600 m/z, with a max injection time of 100 ms and target of 1×10^6 ions. For generation of HCD fragmentation spectra, a max ion injection time of 140 ms and AGC of 1×10^5 were used before fragmentation at 30% normalized collision energy, 35,000 resolution. Precursors were isolated with a width of 2 m/z and put on the exclusion list for 70 s. Single and unassigned charge states were rejected from precursor selection.

QUANTIFICATION AND STATISTICAL ANALYSIS

Peptide and protein identification

Peptide and protein identification was performed as described previously in Zhu et al. (2018). Briefly, Orbitrap raw MS/MS files were converted to mzML format using msConvert from the ProteoWizard tool suite. Spectra were then searched using MSGF+ (v10072) and Percolator (v2.08), where search results from 8 subsequent fraction were grouped for Percolator target/decoy analysis. All searches were done against the human protein subset of Ensembl 75 in the Galaxy platform. MSGF+ settings included precursor mass tolerance of 10 ppm, fully-tryptic peptides, maximum peptide length of 50 amino acids and a maximum charge of 6. Fixed modifications were TMT-10plex on lysines and peptide N-termini, and carbamidomethylation on cysteine residues, a variable modification was used for oxidation on methionine residues. Quantification of TMT-10plex reporter ions was done using OpenMS project's IsobaricAnalyzer (v2.0). PSMs found at 1% FDR (false discovery rate) were used to infer gene identities.

Protein quantification by TMT 10-plex reporter ions was calculated using TMT PSM ratios to the entire sample set (all 10 TMT-channels) and normalized to the sample median. The median PSM TMT reporter ratio from peptides unique to a gene symbol was used for quantification. Protein false discovery rates were calculated using the picked-FDR method using gene symbols as protein groups and limited to 1% FDR. The output from all quantitative MS experiments are available in Table S1.

Selection of Marker Proteins

A novel approach was used to identify marker proteins which are independent of previously annotated subcellular localization. Log2-transformed relative protein quantifications were used for identifying marker proteins. Three filtering steps were applied in order to select robust marker proteins with similar subcellular localization across all five cell lines. The first filtering step was identification and quantification of the protein in all five cell lines. At the second step, all proteins with pearson correlations less than 0.8 between A and B duplicates for each cell line were filtered out to select proteins with robust subcellular fractionation and MS quantification. At the final step, pairwise correlation between all cell lines for each protein (10 pairwise correlations per protein) were calculated using both pearson and spearman correlations with cut-offs set to 0.8 and 0.6 respectively, to filter out proteins with potential cell line specific subcellular localization. 4278, 615, 4160 proteins were filtered out at the first, second, and third steps, respectively. The remaining 3365 proteins were used as marker proteins. For subsequent classifications in the HCC827 triplicate dataset and the HCC827 gef cells, the subset of marker proteins that were identified were used (n:3310 and n:3359 respectively).

Dimension Reduction

Dimensionality reduction was performed using t-distributed stochastic neighborhood embedding (t-SNE) which is a non-linear, dimension-reduction, machine learning algorithm (van der Maaten and Hinton, 2008). It maps high-dimensional data into low-dimensional data by preserving the local structure of the data. A grid-search method was applied to optimize parameters, since the nature of the algorithms is stochastic and even slightly different parameters produce different mappings. t-SNE was selected instead of PCA

reduction due to the non-linear nature of the data and our interest in local similarities rather than capturing data variance. Using the Rtsne package (version 0.11) in R statistical programming language (version 3.3.1), 50 dimensions (five fractions for each cell line and five cell lines along with their biological replicates) were mapped to 3D space which are the first three dimensions of t-SNE by optimized theta value of 0.60 and perplexity value of 60 as parameters. Results of the t-SNE were visualized using the rgl package version (0.97.0).

Clustering of the core proteins

t-SNE coordinates (3D space) of all 3365 marker proteins were used for clustering of proteins using the mClust package (version 5.2) (Scrucca et al., 2016). mClust assigns the proteins to clusters with a probability by performing Expectation-Maximization algorithm on mixture of Gaussians models. Moreover, mClust fits the finite number of assumed normal distributed clusters with different cluster properties such as shape, volume, and orientation. Different cluster sizes (up to 50) were compared and 15 clusters were decided by ALMO curve generated by Bayesian Information Criteria (BIC) values and by external enrichment analysis using GO and UniProt. The identities of marker proteins as well as their cluster membership are available in Table S2.

The enrichment analysis, as well as the core protein distribution in 3D tSNE space indicated differential inter-cluster relationships and the presence of cluster 'neighborhoods'. A 3D plot of tSNE output demonstrated that several clusters were in close vicinity (e.g., clusters 1-4, clusters 5-8, clusters 9-13 and clusters 14-15). Based on this as well as enrichment analysis (see below) four neighborhoods (secretory, nuclear, cytosol, and mitochondria) were defined.

Classification of proteins

Machine learning based algorithms were performed to assign classifications to both marker as well as non-marker proteins. Support vector machine (SVM) with a Gaussian radial basis function kernel was chosen based on overall classification accuracy, kappa score of test data and concordance of classification. Basically, SVM is a decision boundary which can be linear or non-linear based classification by adding maximized margin around the boundary. Unlike other algorithms, SVM focuses more on points which are the most challenging to separate.

SVM with a Gaussian radial basis function kernel classification was performed for each biological duplicate in each cell line (10 individual classifiers were built for the 5 cell line dataset) using relative protein quantifications of core proteins. The datasets were split up into two datasets as follows; 70% for training (2362 proteins) and 30% for testing (1003 proteins) to build a classifier. Overall class distribution of the datasets was preserved on training and test data, individually. The models were trained to tune cost and sigma parameters by grid search. 10-fold cross-validation was used to avoid over-fitting on training data. The best parameters were chosen based on overall accuracy rate on training data. The cost parameter which controls the cost of misclassification rate ranges between 1 and 1000, whereas sigma which controls the bandwidth of the Gaussian ranges between 0.005 and 4, across all classifiers. The performance of the final models was evaluated by test data.

Corresponding models were applied to both marker and non-marker proteins to predict protein localization along with probabilities. The natural algorithm of the SVM does not estimate prediction probability, however, SVM outputs can be used along with the sigmoid function to map probabilities. The e1071 package (version 1.67) in the caret package was used to estimate the prediction probability for each cluster. Proteins were preliminarily assigned to the cluster with the highest prediction probability. For the neighborhood analysis, corresponding clusters probabilities were summed and proteins were preliminarily assigned to the neighborhood with the highest probability.

Probabilities which are close to 1 indicate high confidence predictions, whereas probabilities which are close to 0 show poor prediction and they are prone to lead to incorrect classifications. To increase the prediction accuracy rate and to filter out poor predictions, one criterion and two cut-offs were defined. The criterion was the consensus of preliminary predictions between biological duplicates. Proteins were kept in the analysis, if there was an agreement between biological duplicates. Then, prediction probabilities from the two duplicates of each cell line were averaged for each protein. If there was no agreement, the protein was labeled as 'unclassified'. For the cut-offs, precision (True positives / (True positives + False positives) and recall (True positives / (True positives + False Negatives)) plots were generated on test data by grid search ranging from 0 to 1. Positive and negative labels were defined as follows: If the protein of interest was correctly classified (True Positive), protein of interest was incorrectly classified (false positive), protein of non-interest was correctly classified (true negative), protein of non-interest was incorrectly classified (false negative). Due to varying performance of clusters in different cell lines, individual thresholds were defined for each cluster in each cell line. For threshold 1 (precision-based), probability cut-offs for different clusters in different cell lines were set when the precision of the test data reached 0.9. In some instances, due to well-separated clusters and very few false positives, the precision was greater than 0.9 even without applying threshold 1. To ensure the stringency of the analysis we therefore applied a second cut-off (threshold 2). For threshold 2 (recall-based), probability cut-offs for different clusters in different cell lines were set as the probability of the lowest true positive in the test data. The same workflow was applied to the neighborhood classifications but with the precision-based cut-off (threshold 1) set to 0.95. The output from compartment and neighborhood classifications as well as individual thresholds used for all classifications are available in Table S3.

Domain Enrichment Analysis

Domains were mapped to their corresponding gene id by combining the Pfam database (version 31.0) (Finn et al., 2016) and the biomart R package. Hypergeometric test and fold enrichments were used in concert to identify significantly enriched domains. Multiple testing correction for p values derived from hypergeometric tests was performed using the Benjamini-Hochberg method. 295 domains were found to be enriched in 19 locations (4 neighborhoods and 15 compartments) using a log2 fold change cutoff of 2 and a q-value threshold of 0.01 (Table S4)

For the Signal Peptides, Transit Peptides and Transmembrane Domains, Gene transfer files (GTF) for each annotation was downloaded from UniProt. Mapping of peptides and domains to gene id's for each annotation was carried out with the pyensembl package. Fold enrichments were calculated and significance testing and p values were calculated by the hypergeometric test using phyper function in R.

Annotation of marker protein clusters

For primary annotation of the 15 marker protein clusters enrichment analysis against GO and UniProt annotations for 11 different subcellular compartments were used (cytosol, nucleus, endoplasmatic reticulum, Golgi apparatus, plasma membrane, mitochondrion, cytoskeleton, endosome, lysosome, ribosome, and peroxisome). Only proteins annotated with a single compartment localization were used in the analysis as defined in a previous publication (Lund-Johansen et al., 2016). Fold enrichments were calculated and significance testing and p values were calculated by the hypergeometric test using phyper function in R. For secondary annotation of the 15 marker protein clusters a broad GO cellular component enrichment analysis of the 15 clusters were performed with GOrilla using a target-background approach (Eden et al., 2009). Output is available in Table S2. Specifically, this analysis resulted in the enrichment of *nuclear speckles* in cluster6, *nucleolus* in cluster7, *mitochondrial matrix* in cluster14 and *mitochondrial membrane* in cluster15, and these clusters were annotated accordingly. Cluster3 was enriched in mitochondrial proteins in addition to endoplasmatic reticulum, and the broad GO enrichment analysis indicated specifically mitochondrial translation elongation and termination. Of the 55 cluster3 proteins annotated as mitochondrial in GO and/or UniProt, close to 90% were related to the *mitochondrial ribosome*, and annotation was performed accordingly. In addition, the secondary annotation is based on the fractionation profiles of the cluster proteins. For the nuclear clusters (cluster 5-8) the fractionation profiles of marker proteins between clusters showed distinct differences (Figure S2E). Cluster8 is defined by proteins with the highest abundance in the FS2 fraction (soluble proteins), thus given secondary annotations as *nucleosol*. Cluster5 (also enriched for ribosome) on the other hand contain proteins with their highest abundance in the FP3 fraction, indicating that nuclear proteins in this fraction are part of high density structures or complexes. Based on this fractionation pattern, cluster 5 was given the secondary annotation *nucleosol high density*.

For the cytosol/cytoskeleton clusters (9-13) the fractionation profiles of marker proteins between clusters did not show distinct differences as all clusters contain proteins with the highest abundance in the FS1 fraction (Figure S2E). The separation of these five clusters was instead driven by gradual differences between the abundance of proteins in the FS1 fraction and the other fractions ranging from cluster 12 with the biggest difference down to cluster 9 with the smallest difference.

Network visualization

To visualize the data modularity, pearson correlations for all proteins identified in all five cell lines were calculated and correlations above 0.9 were visualized in Gephi (v. 0.9.1) the highest 10 edges per source node were used to build the network. The network was organized by Force Atlas2 and colored by modularity class (Jacomy et al., 2014).

The consensus localization network was based on proteins with a single neighborhood classification (9858 proteins). Proteins that in addition was classified with a single compartment classification was assigned with the compartment classification. Proteins with no compartment classification, or conflicting compartment classifications from the same neighborhood was assigned with the neighborhood classification. Regulatory proteome sub-networks were subsets of the consensus localization network based on previous literature (see main text). The consensus localization network as well as the regulatory proteome sub networks were visualized in Cytoscape (v. 3.5.1) and organized using AllegroLayout.

Protein complex analysis

The list of human core complexes was downloaded from CORUM (Ruepp et al., 2010) on 2017-03-11 and filtered to include complexes with 3 or more members. Pearson correlation for all pairwise interactions possible for the proteins present in individual complexes were calculated for each cell line using the entire dataset for each cell-line (Table S1) log2-transformed and pre-filtered to remove proteins with < 0.8 Pearson correlation between replicates. Similarly, Pearson correlations for the interactions reported for 271 ORFs by BioPlex (Huttlin et al., 2015) were calculated for each cell line on the same subset of proteins as above (Table S1).

RNA sequencing and splice variant specific localization analysis

Cells (HCC827 or NCI-H322) were cultured in triplicate dishes, and after harvesting of cells using trypsination, total RNA was extracted using RNeasy kit (QIAGEN, Hilden, Germany) according to manufacturer's instructions. RNA libraries were created using strand-specific TruSeq kit with poly-A selection for all RNA samples according to manufacturer's instructions. Quality control were checked by Bioanalyzer/Caliper. Sequencing (Paired-end 2 × 125 bp) was performed by HiSeq2500 (Illumina, San Diego,

CA, USA) as standard RNA-seq protocol. Library preparation as well as sequencing was performed at the sequencing facility (National Genomics Infrastructure) at SciLifeLab in Stockholm, Sweden.

RNA-sequencing reads were mapped to human reference genome (GrCh38) using STAR (Dobin et al., 2013). By setting–outFilterMultimapNmax to 1, unique reads were extracted followed by calculation of Transcript Per Million (TPM) values per sample. In order to get rid of the noise, low-expressed genes with TPM values < 1 were filtered out. Overlapped remaining genes with protein coding genes list (Ensembl 92) were used for downstream analysis.

EventPointer (Romero et al., 2016) was performed to detect alternative splicing (AE) events both for within cell line and between cell lines analysis, using bam files for each sample generated by aligner (STAR). Specified event types inlude; Casette exon, intron retention, alternative 3′ splice site, alternative 5′ splice site, alternative first, alternative last, and mutually exclusive. Event types that are not specified as above are called complex events in EventPointer. Due to exon mapping difficulties, complex events were excluded in within cell line variant specific subcellular localization analysis. The between cell lines analysis was devised at gene centric level, and here no events were excluded, i.e., all differential AE events detected in the transcriptomic data that may lead to alteration of subcellular localization were evaluated (Table S5).

Quantified 'percent spliced in' values (PSI or Ψ) were used to filter detected events. AE events with PSI values greater than 0.1 and less than 0.9 in at least one cell line were kept in the analysis. For these events, exon centric classification was performed based on MS-data from the main HCC827 experiment as well as from the triplicate HCC827 experiments.

For exon centric classification, peptide level quantification was first calculated from the PSM quant values (median). Peptides with missing values in TMT 10-plex signals were removed from the analysis. To filter the peptide level quantitative data, Pearson correlation between duplicate peptide level fractionation profiles was calculated, and peptides with a duplicate correlation < 0.8 were removed. All peptides were then mapped to transcripts annotated in Ensembl 75 using SpliceVista.py (Zhu et al., 2014). To calculate an exon-centric subcellular fractionation profiles, splice junction spanning peptides were excluded and the remaining peptides were summarized by median into exon level quantification. For a gene to be identified as a candidate for variant specific localization, the exon-centric classification must differ from the gene-centric classification, or from the classification of another exon mapping to the same gene (Table S5).

In the between cell lines analysis, the difference in PSI values ($\Delta\Psi$) were measured ($|\Delta\Psi| > 0.1$) and adopted limma framework (< 0.1 fdr) applied to detect significant AE events between cell lines (NCI-H322 and HCC827). Subsequently, proteins with different classifications (between NCI-H322 and HCC827) at the neighborhood level in the original gene centric analysis were compared with the identified differentially spliced genes (Table S5).

Relocalization analysis

For the relocalization analysis, proteins classified at the neighborhood level in both HCC827 cells from the five cell lines dataset and HCC827 cells treated with the EGFR inhibitor gefitinib (2.5 μM, 2h) were used (n:9058). Out of these 9058 proteins, different classifications were made for 295 proteins between untreated and treated cells (Table S7). Sankay plot vizualisation was performed using networkD3 R package. For stringency, cutoff for candidate relocalizing proteins were defined based on three filtering criteria; 1. > 2 PSMs for quantification; 2. Fractionation profile correlation < 0.8 between untreated and treated samples; and 3. Same classification in HCC827ctrl experiment and all three independent HCC827 experiments from the robustness analysis. After the filtering 13 candidate relocalizing proteins remained.

Comparisons to additional datasets

All evaluations of data in relation to Cell Atlas was done using data from supplementary tables associated with the Cell Atlas publication (Thul et al., 2017). For investigation of multi localization of proteins data from the COMPARTMENTS database, the human subset of the knowledge channel was downloaded from https://compartments.jensenlab.org/Downloads on 2018-06-05. For analysis of mRNA expression in CCLE cell lines, affymetrix gene expression data (CCLE_Expression_Entrez_2012-09-29) was downloaded at https://portals.broadinstitute.org/ccle/users/sign_in. For analysis of TCGA PanCancer mRNA expression, RNaseq data (HiSeqV2_PANCAN-2015-02-15) was downloaded from UCSC Cancer Browser.

DATA AND SOFTWARE AVAILABILITY

Gene symbol-centric Protein Identifications filtered at 1% FDR for all datasets: Table S1

Classification output from SVM classifier for all datasets: Table S3

The accession number for the MS proteomics data reported in this paper is ProteomeXchange: PXD006895.

The accession number for the RNA-seq data reported in this paper is NCBI SRA: SRP154280.

ADDITIONAL RESOURCES

SubCellBarCode portal

For visualization and access to the underlying datasets we have generated the SubCellBarCode portal (https://www.subcellbarcode.org). The portal uses the shiny cloud-based system (http://www.shinyapps.io/) which runs the shiny framework for R (http://shiny.

rstudio.com). The portal provides access to both the localization predictions as well as the raw fractionation and PSM data for further inspection and validation of results from the article. The portal provides access to 4 applications and download links for the RAW, Processed and Results datasets.

1 The BarCode app, provides classification output of single proteins (gene-centric) at both neighborhood level and compartment level. Mouse-over functions provide classification details for each cell line, as well as the number of PSMs used for quantification of the specified protein in the specified cell line. In addition, if no conflict in classification between cell lines, a consensus call is made at compartment level or neighborhood level. A second tab provide visualization of the fractionation profiles generated by quantitative MS.
2 The CoLocal app enables the user to input a list of genes to visualize, for each cell line separately, neighborhood classifications as well as pairwise correlations between proteins in order to assess protein co-localization. In addition, pre-defined lists are available for evaluation of corum complexes. A second tab shows a heatmap of classification predictions for the proteins under investigation.
3 The NetWork app displays the subcellular distribution of proteins in a user defined list as a network for each cell line separately. Available are also pre-defined lists for specific regulatory subsets of proteins such as kinases or transcription factors. Other tabs provide information about over- or under-representation of proteins under investigation in specific neighborhoods (tab2) or compartments (tab3). Mouse-over functions provide information about fold enrichment/depletion as well as p value (using the Hypergeometric Distribution). The 4^{th} tab provides information about the classification for the proteins under investigation in a table format.
4 The HeatMap app allows the user to perform hierarchical clustering of user defined proteins across all samples in all 5 cell lines, and to vizualise the output in a heatmap representation. Pre-defined lists of proteins are available, and the user can also evaluate number of PSMs used for quantification of proteins in each cell line separately.