

Editorial Manager(tm) for Annals of Biomedical Engineering
Manuscript Draft

Manuscript Number: ABME954R1

Title: Large-Scale Automated Analysis of Location Patterns in Randomly-Tagged 3T3 Cells

Article Type: S.I.: Systems Bio & Bioinformatics

Section/Category:

Keywords: Protein Subcellular Location, Subcellular Location Trees, Subcellular Location Features, CD-tagging, Fluorescence Microscopy, Cluster Analysis, Location Proteomics

Corresponding Author: Elvira Garcia Osuna,

Corresponding Author's Institution: Carnegie Mellon University

First Author: Elvira Garcia Osuna

Order of Authors: Elvira Garcia Osuna; Juchang Hua; Nicholas W Bateman; Ting Zhao; Peter B Berget; Robert F Murphy

Manuscript Region of Origin:

1
2
3
4 **Large-Scale Automated Analysis of Location Patterns in Randomly-**
5 **Tagged 3T3 Cells**
6
7
8
9

10
11
12 Elvira García Osuna^{1,2}, Juchang Hua^{1,3,4}, Nicholas W. Bateman³, Ting Zhao^{1,2}, Peter B.
13 Berget³, Robert F. Murphy^{1,2,3,4}
14
15
16
17
18

19 Center for Bioimage Informatics¹ and Departments of Biomedical Engineering,²
20 Biological Sciences,³ and Machine Learning,⁴
21
22
23
24 Carnegie Mellon University, Pittsburgh, PA 15213
25
26
27
28

29 Large-Scale Automated Analysis of Protein Location Patterns
30
31
32
33

34 Correspondence to:

35
36 Robert F. Murphy

37
38 Carnegie Mellon University

39
40 Center for Bioimage Informatics, HHC119

41
42 5000 Forbes Ave

43
44 Pittsburgh, PA 15213

45
46 Phone: 1.412.268.3480

47
48 FAX: 1.412.268.9580

49
50 Email: murphy@cmu.edu
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 Abstract
5
6

7 Location proteomics is concerned with the systematic analysis of the subcellular location
8 of proteins. In order to perform high-resolution, high-throughput analysis of all protein
9 location patterns, automated methods are needed. Here we describe the use of such
10 methods on a large collection of images obtained by automated microscopy to perform
11 high-throughput analysis of endogenous proteins randomly-tagged with a fluorescent
12 protein in NIH 3T3 cells. Cluster analysis was performed to identify the statistically
13 significant location patterns in these images. This allowed us to assign a location pattern
14 to each tagged protein without specifying what patterns are possible. To choose the best
15 feature set for this clustering, we have used a novel method that determines which
16 features *do not* artificially discriminate between control wells on different plates and uses
17 Stepwise Discriminant Analysis (SDA) to determine which features *do* discriminate as
18 much as possible among the randomly-tagged wells. Combining this feature set with
19 consensus clustering methods resulted in 35 clusters among the first 188 clones we
20 obtained. This approach represents a powerful automated solution to the problem of
21 identifying subcellular locations on a proteome-wide basis for many different cell types.
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44

45 Key terms: Protein Subcellular Location, Subcellular Location Trees, Subcellular
46 Location Features, CD-tagging, Fluorescence Microscopy, Cluster Analysis, Location
47 Proteomics
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

I. Introduction

Current work in proteomics includes systematic analysis of protein structure, expression levels, and interactions. These projects will provide critical data for understanding and modeling cell and tissue behavior. Knowledge of the subcellular location of each protein is equally important to this task. However, this area has received far less attention.

There are two major ways of analyzing protein subcellular location: prediction and determination. A number of systems for predicting protein localization from sequence have been described.^{5,8,14,17,18} The limitation of these systems is that they can only assign new proteins to the location categories with which they have been trained. This means that proteins with previously unseen location patterns cannot be properly categorized. In addition, since they have been trained to recognize only low-resolution classes, they are typically able to predict the organelle to which a protein will be localized, but not the specific area of the organelle. Due to lack of training data, they are also unable to predict differential localization of proteins in different cell types or under different conditions.

A. Determination of Protein Location

Due to the limitations of prediction, there is a need for projects that will collect data on subcellular location for entire proteomes under a variety of conditions. These projects *determine* protein location rather than predict it. Although these projects are useful in their own right, they also serve as a way to expand the capabilities of prediction systems by providing training examples for higher-resolution and complex patterns.

1
2
3
4 Fluorescence microscopy has been widely used for determining protein subcellular
5 location, and visual examination has been the primary means of analyzing the resulting
6 images. Some large-scale projects have used fluorescence microscopy to screen hundreds
7 to thousands of proteins for particular patterns or to assign proteins to major location
8 classes.^{11,13,20,22} A particular ambitious and valuable project has been the tagging of all
9 predicted protein coding regions in the yeast *Saccharomyces cerevisiae*.¹¹
10
11
12
13
14
15
16
17
18
19
20

21 Visual examination of images is not only inefficient for high-throughput projects, but it is
22 also subjective and irreproducible. Fortunately, automated methods of analyzing protein
23 location have been described by our group^{1-3,10} and more recently by others.^{6,7,21} These
24 methods have been shown not only to perform as well as visual examination for
25 distinguishing major subcellular patterns, but also to be able to discriminate patterns that
26 a human observer cannot.¹⁶
27
28
29
30
31
32
33
34
35
36
37

38 There is not only a need for automated analysis of images, but large-scale projects also
39 require high-throughput methods for acquiring images. Automated fluorescence
40 microscopes originally developed for drug screening can meet this need.^{19,23} These
41 microscopes use multi-well plates, contain autofocus capabilities and are capable of
42 multi-color imaging as well as 3D-time-series imaging.
43
44
45
46
47
48
49
50
51
52

53 B. CD-tagging of NIH 3T3 cells

54
55 In order to perform systematic analysis of protein location by fluorescence microscopy, a
56 high-throughput means of tagging all (or most) proteins is also needed. One such method
57
58
59
60
61
62
63
64
65

1
2
3
4 is CD-tagging.¹² This method inserts a guest exon into genomic DNA. The insert consists
5
6 of an enhanced green fluorescent protein (EGFP) coding sequence flanked by splicing
7
8 signals. Therefore, when the protein with the guest exon insertion is expressed, it
9
10 contains an internal fluorescent tag. Previous studies have shown that CD-tagging has
11
12 minimal impact on protein folding, function and localization.¹³ Here, we combine CD-
13
14 tagging, automated microscopy and automated analysis to identify statistically
15
16 distinguishable location patterns NIH 3T3 cells. We present the combination of high-
17
18 throughput methods from tagging to analysis as well as fully automated methods of
19
20 imaging and analysis.
21
22
23
24
25
26
27

28 II. Methods

29 A. Production and Isolation of CD-tagged NIH 3T3 cells

30
31 The procedure described previously¹³ was followed, with some minor alterations. A CD-
32
33 tagging cassette containing the EGFP coding sequence was packaged into retrovirus
34
35 using Phoenix-GP cells. Phoenix-GP cells were seeded at a rate of 1.3×10^6 cells per
36
37 75cm^2 flask in complete Phoenix media (Dulbecco's Modified Eagle's Medium (DMEM)
38
39 containing 10% fetal bovine serum). The Phoenix-GP cells were transfected the next day
40
41 with 9 μg Stealth plasmid and 1 μg VSV-G plasmid per flask using Mirus Trans-IT-LT-1
42
43 lipofection reagent as per manufacturer's protocol. Briefly, 15 μl Trans-IT-LT-1 was
44
45 added to 500 μl serum-free media and incubated for 5 min at room temperature. The
46
47 DNA was then added to this mixture, which was then incubated for an additional 20 min
48
49 at room temperature. The resulting DNA complexes were then added to the Phoenix-GP
50
51 cells in 10 ml fresh complete media and the cells were incubated for 24 h at 37°C and 5%
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 CO₂. After 24 h, the media was replaced with 10 ml fresh media and the flasks were
5
6 incubated at 32°C and 5% CO₂ for 48 h. The resulting viral supernatant was flash frozen
7
8 in 1ml aliquots in liquid nitrogen and stored at -80°C. Viral supernatants were created
9
10 using three different versions of the Stealth plasmid, P19, P20 and P21, which encode
11
12 EGFP appropriately for class 0, class 1 and class 2 introns, respectively. A different virus
13
14 was used each week so that introns of all types could be sampled.
15
16
17
18
19
20

21 For infection, NIH 3T3 cells were plated at 2x10⁵ cells per well of a 6-well plate
22
23 containing complete media (DMEM containing 10% fetal calf serum, 100 U/ml penicillin,
24
25 and 100 µg/ml streptomycin). Six h later, the media was aspirated and viral supernatant
26
27 was added with 6 µg/ml polybrene (to neutralize the charge on the cell surface so that
28
29 viral particles will not be repelled) and incubated for 24 h at 37°C and 5% CO₂.
30
31
32
33
34
35

36 The cells were then trypsinized, expanded into a 10 cm dish and incubated for 48 h.
37
38 EGFP-expressing cells were sorted using a FACS Vantage SE using a threshold set to
39
40 include only 0.1% of untagged, control cells. Positive, singlet cells were sorted into black
41
42 polystyrene, glass-bottomed 96-well plates (Whatman) containing 200 µl of complete
43
44 medium (Dulbecco's modified Eagles medium, 10% fetal calf serum, 100 U/ml penicillin
45
46 and 100 µg/ml streptomycin). Plates were incubated for 8 d before adding 1x10⁴
47
48 untagged and positive control cells to one well each in each row (cells expressing tagged
49
50 Procollagen Type I alpha 2 were used as the positive control).
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 On days 11-15, the media was aspirated and the DNA-binding vital dye Hoechst 33342
5
6 was added at a concentration of 0.5 $\mu\text{g/ml}$ in OptiMEM (Invitrogen Corporation,
7
8 Carlsbad, CA, USA). Plates were then incubated for 45 min at 37°C and 5% CO₂ before
9
10 imaging.
11
12
13
14

15 16 B. Automated Fluorescence Imaging

17
18 Two color images (Hoechst 33342 and EGFP) were acquired using an automated
19
20 fluorescence microscope (Beckman Coulter IC-100). Images were acquired with a 40x
21
22 0.9NA objective and a Hamamatsu Orca-ERG camera at a fixed camera gain and
23
24 exposure time. 25 fields were imaged for each well using autofocus on the Hoechst
25
26 channel. Images of empty wells were discarded. The remaining images of EGFP-positive
27
28 cells were used for analysis.
29
30
31
32
33
34

35 36 C. Feature Calculation and Selection

37
38 The most common approach to describing subcellular pattern is to use features calculated
39
40 on single cell images. This requires segmenting each image into single cell regions, a
41
42 task that can be quite error prone. For the large number of images in this study, we
43
44 therefore used a new set of our Subcellular Location Features that are not sensitive to the
45
46 number of cells in an image. The starting point for this set was SLF21, which has
47
48 previously been shown to provide good performance for classifying subcellular patterns
49
50 without cell segmentation.⁹ It includes 3 morphological features, 5 edge features and 13
51
52 Haralick texture features. We augmented this set by calculating the 13 Haralick texture
53
54 features after downsampling the protein image from two to six fold and adding a new
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 feature which measures the percentage of pixels that are above threshold in the protein
5 (EGFP) image which are also above threshold in the DNA (Hoechst) image.
6
7 (Thresholding is performed as described previously.⁹) These additional 66 features gave
8
9 us a total of 87 features to describe each image. We define this set as SLF25.
10
11
12
13
14
15

16 To assess the sensitivity of a given feature to undesirable well-to-well and plate-to-plate
17 variation, t-tests were performed for all pairs of images (fields) of positive control cells.
18
19 Average p-values were calculated for all pairwise tests for a given feature, and various
20
21 thresholds on this average were used for feature elimination.
22
23
24
25
26
27

28 Step Discriminant Analysis (SDA) was then done for the remaining features on the entire
29 image dataset to select those with good discriminating power: the features that can
30
31 differentiate the patterns.
32
33
34
35
36
37

38 D. Clustering of Protein Patterns

39
40 A three-step process was used to cluster the wells that contained tagged proteins. First, *k*-
41 means clustering with a z-scored Euclidean distance function was performed on the
42 image varying *k* from 1 to 100. Akaike information content (AIC) was then calculated to
43
44 select an optimal *k* and corresponding clustering of the images. Second, each well was
45
46 allocated to that cluster which contains a plurality of the images in the well and only the
47
48 images in this cluster were kept for further analysis. If, however, the number of images
49
50 assigned to the plurality cluster was less than 1/3 of the total number of images for a
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4 given well, that well was considered not to have a unique pattern and it was removed
5
6 from the analysis.
7
8
9

10
11 Lastly, a consensus tree algorithm⁴ was applied to the remaining images. In this
12 algorithm, a hierarchical cluster tree (dendrogram), was generated from a random half of
13 images of each well. This was repeated 200 times and a consensus tree was generated in
14 which only the branches of the trees that were present in at least half of the trees were
15 kept.
16
17
18
19
20
21
22
23
24
25

26 Visual inspection was also used to cluster the tagged wells. During this process,
27 descriptive terms were assigned to each well by one of the authors (E.G.O.) after
28 carefully examining the representative images of each well (representative images were
29 chosen as described previously¹⁵). Whatever terms that were felt to accurately describe
30 the protein pattern were used, and for the consistency, the same terms were used for the
31 same patterns. Wells were then grouped into those that shared a unique combination of
32 the descriptive terms.
33
34
35
36
37
38
39
40
41
42
43
44

45 In order to measure the agreement of different clustering results, we calculated Cohen's κ
46 statistics on each pair of clustering results A and B:
47
48
49
50

$$\kappa(A,B) = \frac{\text{Observed agreement} - \text{expected agreement}}{1 - \text{expected agreement}}$$

51
52
53
54
55 where expected agreement is that expected for two random samplings from the same
56 clustering.⁴
57
58
59
60
61
62
63
64
65

E. Software and data availability

All data and Matlab code used in this paper are available at <http://murphylab.web.cmu.edu/data> and <http://murphylab.web.cmu.edu/software>, respectively.

III. Results

We have previously demonstrated the feasibility of automated clustering of randomly-tagged proteins by their location pattern using high-resolution images obtained with a spinning disk confocal microscope. This required major efforts in three areas: time and culture expense for isolating, expanding and maintaining individual clones, large reagent costs for identifying the tagged gene by RT-PCR and sequencing, and extensive time for individually plating and carrying out 3D imaging for each clone. The results provided information about the location of each protein but also about the number and type of patterns that were observed. Given the expense of this approach, we sought to evaluate a much less expensive alternative for just determining the set of possible patterns: sorting individual tagged cells directly into 96-well plates and imaging them without identifying the tagged gene. To test the feasibility of this approach, we generated and imaged ten plates per week for four weeks. After eliminating edge rows and columns (which could not be imaged due to interference by the plate skirt with the 40x objective) and the negative and positive control wells (three each per plate), we obtained images for 54 randomly-tagged wells per plate or a total of 2160 wells. Of these, 222 contained EGFP-positive cells. Examples of these images can be seen in Figure 1. After removing those images which were overcrowded or those for which valid features could not be calculated

1
2
3
4 due to low fluorescence signal, a total of 174 wells with at least 10 images remained.
5
6 These were used in clustering analysis.
7
8
9

10
11 An important issue for any image clustering approach is the nature of the features to be
12 used. Given that the images we wished to cluster were collected on different days over
13 many weeks (albeit under nominally the same conditions each day), one concern in this
14 respect is that features that are sensitive to day-to-day variations might result in clustering
15 proteins by day of acquisition (or position within a plate) rather than by protein pattern.
16
17 The presence of the positive controls wells in each plate allowed us to design a strategy to
18 minimize this concern. We sought to select features that can tell the difference between
19 the different randomly-tagged proteins but not be sensitive to the variance among the
20 positive controls from plate to plate. As described in the Methods, we did extensive t-
21 tests on each feature for each pair of images from control wells to eliminate features that
22 were significantly different between the controls. We used three thresholds (0, 0.1 and
23 0.2) on the average p-values to eliminate plate dependent features. The remaining
24 features were then subjected to Stepwise Discriminant Analysis (SDA) to eliminate
25 features that did not provide any discriminating power between the randomly-tagged
26 wells. A total of 76, 64, and 42 features were retained for thresholds of 0, 0.1 and 0.2,
27 respectively.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50

51
52
53 Using these features, we then performed k-means clustering on all images for the 174
54 clones (plus 14 positive control wells) for various values of k (the number of clusters).
55
56
57 The goodness of these different clustering runs was evaluated using the Akaike
58
59
60
61
62
63
64
65

Information Criterion (AIC), which balances tightness of the clusters against the number of clusters. These AIC values are plotted as a function of k in Figure 2. The results indicate that the optimal numbers of clusters are 41, 35 and 70 for the feature sets selected using a p-value threshold of 0, 0.1 and 0.2.

Consensus trees were then built for each feature set. These can be viewed through a web interface at <http://murphylab.web.cmu.edu/services/PSLID> that permits display of representative images for each well. The consensus tree built with a p-value threshold of 0.1 is shown in Figure 3.

Different feature sets led to different clustering results. In order to measure how much they agree with each other, the Cohen κ statistics was calculated for each pair of clustering results. Since different sets of clones were retained in each final clustering, only the common clones in both clustering results were considered in each calculation. Additionally, labels of subcellular location patterns were assigned to each well by visual inspection (shown in Figure 3), and a clustering was generated by grouping wells with the same labels. The Cohen κ statistics was also calculated between visual inspection and all three automated clustering results. The results are shown in Table 1. The agreements between visual inspection and k-means clustering results are obviously lower than those between different k-means clustering results. This indicates the consistency of automated methods of cluster analysis.

IV. Discussion

1
2
3
4 We have described a high-throughput method of analyzing randomly-tagged NIH 3T3
5 cells. This method is automated and results in clusters of protein patterns that have
6 similar distributions. This method allows us to analyze images without any previous
7 knowledge of the protein subcellular location. The work is distinguished from our prior
8 work in that we describe a higher throughput pipeline for infecting, sorting and imaging
9 tagged lines, the use of a internal control and a modified feature selection procedure to
10 minimize the effects of variability during the imaging process, and the use of a new set of
11 field level features that do not require segmentation into single cells.
12
13
14
15
16
17
18
19
20
21
22
23
24
25

26 It should be noted that in the work described here only proteins for which a consistent
27 location pattern could be found were analyzed. Future work will extend the analysis to
28 identify proteins with variable patterns, such as those that show cell cycle dependence.
29 The data collected in this study are being made publicly available to facilitate
30 development of methods for this type of analysis.
31
32
33
34
35
36
37
38
39
40

41 The current results show that many, but not all, of the positive controls were clustered
42 together. This suggests that additional effort is needed in the future to ensure consistency
43 between different runs. Incorporating a larger number of positive controls that represent
44 additional major subcellular locations would therefore appear useful. We are adopting
45 this approach in our ongoing experiments to expand our database to include thousands of
46 tagged proteins. Our goal is then to use cluster analysis as described here to determine
47 the number and types of subcellular location families that are present in NIH 3T3 cells.
48
49
50
51
52
53
54
55
56
57
58 Once the set of possible patterns is known, the methods described here can be used to
59
60
61
62
63
64
65

1
2
3
4 screen for clones with particular patterns so that the tagged gene can be sequenced. This
5
6 will be useful for identifying novel patterns and proteins that display them as well as
7
8 providing new data for training location prediction methods.
9

10 11 12 13 14 V. Acknowledgments

15
16 We would like to thank Dr. Jonathan Jarvik for helpful discussions and Yehuda Creeger
17
18 for technical assistance. This work was supported by Commonwealth of Pennsylvania
19
20 Tobacco Settlement Fund grant 017393, NIH grant GM068845-01, and NSF grant EF-
21
22 0331657.
23
24

25 26 27 28 29 VI. References

- 30
31 1. Boland M. V., M. K. Markey, and R. F. Murphy. Classification of protein localization
32 patterns obtained via fluorescence light microscopy. Proc of 19th Annual
33 International Conference of the IEEE Engineering in Medicine and Biology
34 Society. pp. 594-7, 1997.
- 35
36 2. Boland M. V., M. K. Markey, and R. F. Murphy. Automated recognition of patterns
37 characteristic of subcellular structures in fluorescence microscopy images.
38 Cytometry. 33: 366-75, 1998.
- 39
40 3. Boland M. V., and R. F. Murphy. A neural network classifier capable of recognizing
41 the patterns of all major subcellular structures in fluorescence microscope images
42 of hela cells. Bioinformatics. 17: 1213-23, 2001.
- 43
44 4. Chen X., and R. F. Murphy. Objective clustering of proteins based on subcellular
45 location patterns. J Biomed Biotechnol. 2005: 87-95, 2005.
- 46
47 5. Chou K. C., and Y. D. Cai. Prediction and classification of protein subcellular
48 location-sequence-order effect and pseudo amino acid composition. J Cell
49 Biochem. 90: 1250-60, 2003.
- 50
51 6. Conrad C., H. Erfle, P. Warnat, N. Daigle, T. Lorch, J. Ellenberg, R. Pepperkok, and
52 R. Eils. Automatic identification of subcellular phenotypes on human cell arrays.
53 Genome Research. 14: 1130-6, 2004.
- 54
55 7. Danckaert A., E. Gonzalez-Couto, L. Bollondi, N. Thompson, and B. Hayes.
56 Automated recognition of intracellular organelles in confocal microscope images.
57 Traffic. 3: 66-73, 2002.
- 58
59 8. Guda C., E. Fahy, and S. Subramaniam. Mitopred: A genome-scale method for
60 prediction of nucleus-encoded mitochondrial proteins. Bioinformatics. 20: Jul 22,
61 2004.
62
63
64
65

-
9. Huang K., and R. F. Murphy. Automated classification of subcellular patterns in multicell images without segmentation into single cells. Proc of 2004 IEEE International Symposium on Biomedical Imaging (ISBI-2004). pp. 1139-42, 2004.
 10. Huang K., and R. F. Murphy. Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. BMC Bioinformatics. 5: 78, 2004.
 11. Huh W.-K., J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Welssman, and E. K. O'Shea. Global analysis of protein localization in budding yeast. Nature. 425: 686-91, 2003.
 12. Jarvik J. W., S. A. Adler, C. A. Telmer, V. Subramaniam, and A. J. Lopez. Cd-tagging: A new approach to gene and protein discovery and analysis. BioTechniques. 20: 896-904, 1996.
 13. Jarvik J. W., G. W. Fisher, C. Shi, L. Hennen, C. Hauser, S. Adler, and P. B. Berget. In vivo functional proteomics: Mammalian genome annotation using cd-tagging. BioTechniques. 33: 852-67, 2002.
 14. Lu Z., D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics. 20: 547-56, 2004.
 15. Markey M. K., M. V. Boland, and R. F. Murphy. Towards objective selection of representative microscope images. Biophys. J. 76: 2230-7, 1999.
 16. Murphy R. F., M. Velliste, and G. Porreca. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J VLSI Sig Proc. 35: 311-21, 2003.
 17. Nakai K. Protein sorting signals and prediction of subcellular localization. Adv. Protein Chem. 54: 277-344, 2000.
 18. Park K. J., and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics. 19: 1656-63, 2003.
 19. Price J. H., A. Goodacre, K. Hahn, L. Hodgson, E. A. Hunter, S. Krajewski, R. F. Murphy, A. Rabinovich, J. C. Reed, and S. Heynen. Advances in molecular labeling, high throughput imaging and machine intelligence portend powerful functional cellular biochemistry tools. J. Cell. Biochem. Supp. 39: 194-210, 2002.
 20. Rolls M. M., P. A. Stein, S. S. Taylor, E. Ha, F. McKeon, and T. A. Rapoport. A visual screen of a gfp-fusion library identifies a new type of nuclear envelope membrane protein. J. Cell Biol. 146: 29-44, 1999.
 21. Sigal A., R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, I. Alaluf, N. Swerdlin, N. Perzov, T. Danon, Y. Liron, T. Raveh, A. E. Carpenter, G. Lahav, and U. Alon. Dynamic proteomics in individual human cells uncovers widespread cell-cycle dependence of nuclear proteins. Nat Methods. 3: 525-31, 2006.
 22. Simpson J. C., R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep. 1: 287-92, 2000.
 23. Taylor D. L., E. S. Woo, and K. A. Giuliano. Real-time molecular and cellular analysis: The new frontier of drug discovery. Curr Opin Biotechnol. 12: 75-81, 2001.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

		k-means/AIC with p-value threshold		
		0	0.1	0.2
visual inspection		0.13 (0.01)	0.06 (0.01)	0.18 (0.03)
k-means/AIC	0		0.36 (0.02)	0.75 (0.04)
	0.1			0.49 (0.04)

Table 1. Comparison of clustering results. Cohen’s κ statistic was calculated to measure the degree of agreement between each pair of clustering results. Higher values indicate better agreement. The numbers in the parentheses are the standard deviation of the statistics.

Figure Captions

Figure 1. Example images from the dataset acquired in this study. The clones varied in protein expression level, and therefore each panel was fully contrast-stretched to facilitate visualization (hence the background appears different in each panel).

Figure 2. Determination of the optimal number of clusters using AIC. Three p-value thresholds were used (solid: 0, dashed:0.1, dotted:0.2) to select a set of features and then k-means clustering was performed for various values of k. AIC was calculated to measure the goodness of each clustering. The optimal values of k are 41, 35 and 70, respectively.

Figure 3. A consensus subcellular location tree built from 126 wells of the randomly tagged 3T3 image dataset. A threshold of 0.1 was used on the average p-value of the statistic tests on control wells to select features. The first column of labels shows the well name and (positive control wells are marked with an asterisk). The second column of labels shows the locations assigned by visual inspection. In this tree, the sum of the lengths of the branches connecting two clones is proportional to the distance between them in feature space.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

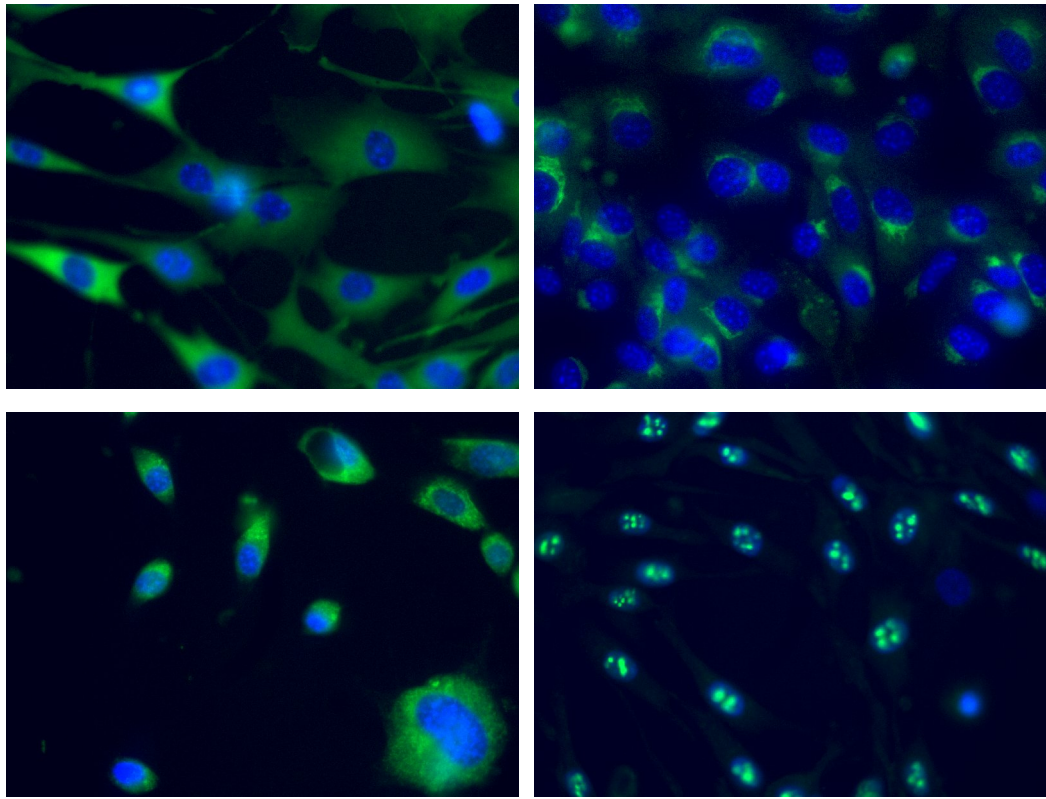


Figure 1. García Osuna, Elvira

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

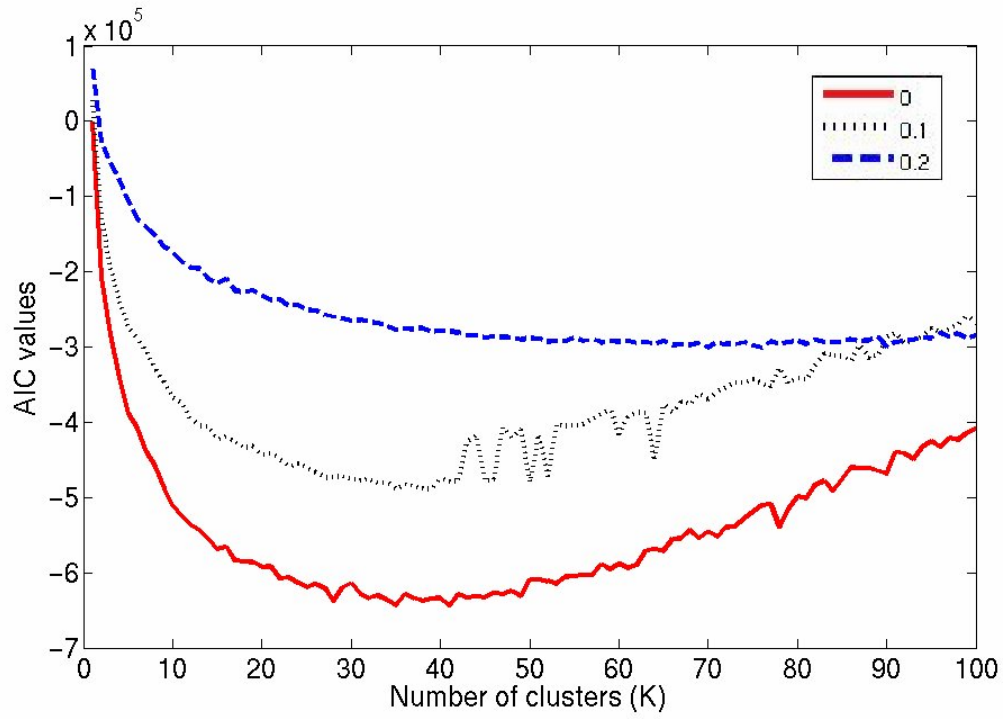


Figure 2. García Osuna, Elvira

Reviewer #1:

We see three major points raised. First, the reviewer raises a question regarding eliminating wells that exhibit variations in pattern. This point has been addressed in the manuscript in the “Discussion” section. Briefly, although we eliminate images from wells that do not exhibit the same pattern as the majority pattern in the well, we do retain those images that do demonstrate the majority pattern. Dealing with variations in pattern will be something that needs to be addressed in future work.

Secondly, the reviewer raises a very valuable point: the need for more positive controls. The results suggest that the inclusion of more positive control patterns would be useful in identifying the patterns in the cluster. We have addressed this concern in the “Discussion” section of the manuscript. At this point there is little that can be done to address the issue in dataset used for the manuscript, but it is a suggestion that can be included in future studies.

Lastly, it has been shown in previous studies that CD-tagging has minimal impact on protein localization (Jarvik, J.W., et al, 2002). However, it is entirely possible that for certain proteins, the inclusion of the EGFP encoding sequence changes the location of the protein for some insertion sites. We expect that since as the number of tagged clones grows we will have multiple instances of the same protein being tagged, a majority pattern can be discerned and presumably this will be the normal pattern.

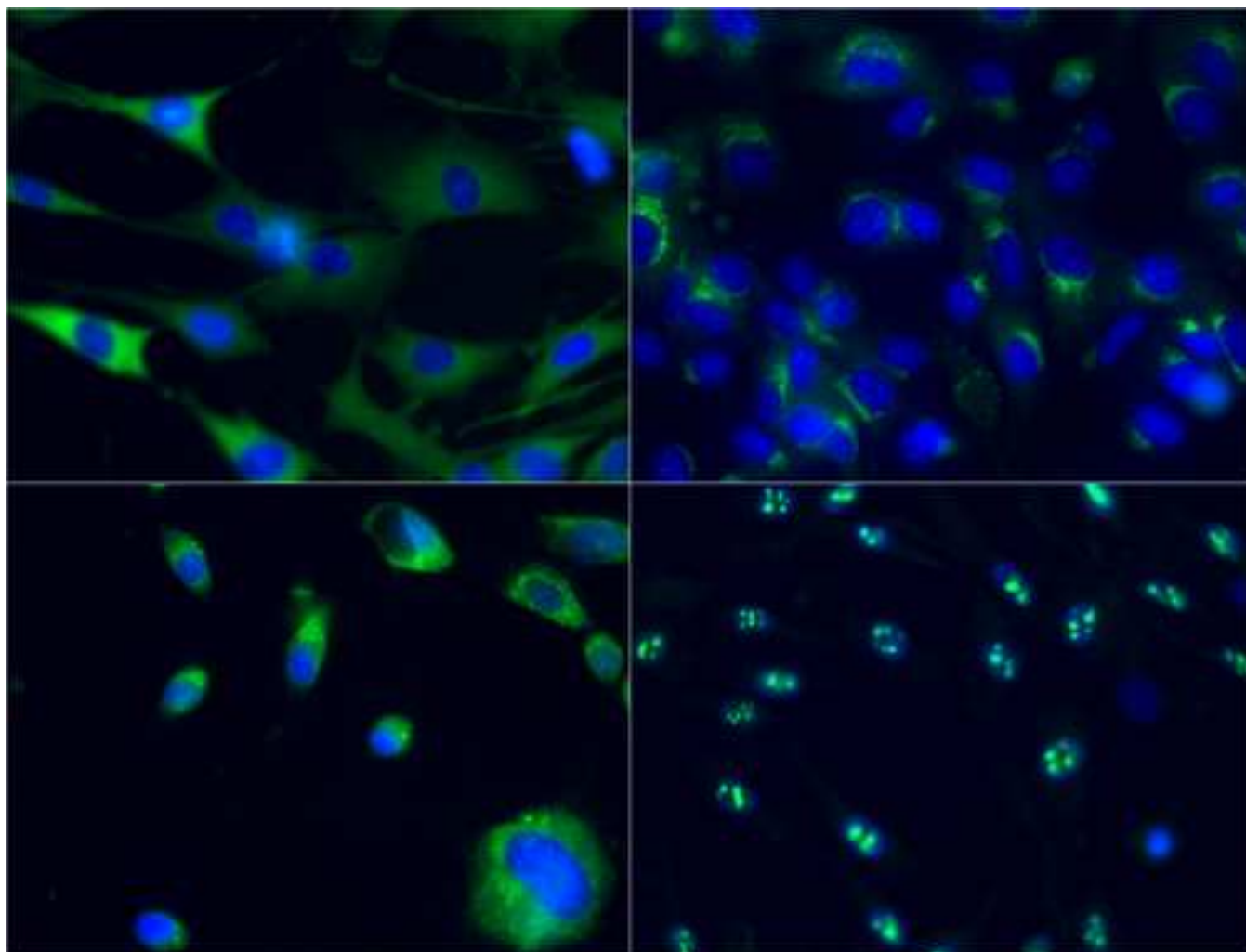
Reviewer #3:

There are three points raised by this reviewer. The first point is that there is a need to clarify the advancement of this work with respect to previous work. We discuss this now at the end of the Introduction and again in the Discussion. The new things are the higher throughput pipeline, incorporation of internal controls and new approach to feature selection, use of field features instead of single cell features.

The second point is connections to other proteomics database. We are part of the NCIBI integrated database project, and PSLID is being connected to that database so that high-resolution information can be integrated with other data sources. PSLID also can be linked to directly from other databases. We have mentioned these items briefly in the discussion.

The third point is well taken. We have been working on generative models to capture and distribute subcellular patterns for each cluster and are about to submit a manuscript describing that work. Linking directly from integrated databases (previous point) also can help with this.

Figure
[Click here to download high resolution image](#)



Figure

[Click here to download high resolution image](#)

