# Novel Morphological Phenotypes Discovery in High-Content Screens Using Underused Features

Chen Lin, Pengyu Hong
Computer Science Department
Brandeis University
Waltham, MA, 02454, U.S.A.
clin@brandeis.edu
hongpeng@brandeis.edu

Chris Bakal
Dynamical Cell Systems Team
The Institute of Cancer Research
London, UK
cbakal@receptor.med.harvard.edu

Norbert Perrimon
Department of Genetics
Harvard Medical School
Howard Hughes Medical Institute
Boston, MA, 02115, U.S.A.
perrimon@receptor.med.harvard.edu

## Abstract

Cell-based high-content screening (HCS) is a powerful high-throughput technology for studying cellular processes through the analysis of complex cellular morphology. A typical large-scale HCS screen can generate many novel morphological phenotypes. It is challenging to identify novel phenotypes due to the huge volume of data and the lack of domain knowledge. This paper presents a new strategy that discovers novel phenotypes using underused image features, which are defined as those not fully utilized by the existing phenotypes. Our approach was successfully applied to a data set generated in a genetic HCS of *Drosophila* BG-2 cells to discover novel phenotypes and make interesting predictions.

## 1. INTRODUCTION

High-content screening (HCS) is a powerful high-throughput technology for identifying and understanding the functions of genes and pathways responsible for key cellular processes [1-9]. It has also been widely used in compound screening and drug profiling [10-16]. The automatic analysis of HCS images is an important and challenging problem. Usually, a set of image features is extracted to represent the content of each image, and used in the downstream computation.

Machine learning techniques have become essential in analyzing HCS images. The Murphy Lab is a forerunner in the use of supervised machine learning techniques that train classifiers to recognize sub-cellular patterns [17-21]. Wang et al. [22] compared supervised training of the naive Bayesian classifier, linear discriminant analysis, *K*-nearest neighbors, and support vector machine classifiers [23] in recognizing morphological phenotypes of cultured *Drosophila* Kc167 cells treated with RNA interference (RNAi). Loo et al. [24] and Bakal et al. [25] respectively trained support vector machines and neural networks as a set of classifiers to recognize images of several representative treatment conditions (TCs). The classifiers were then used to derive phenotypic profiles of the rest of the TCs. Clustering analysis of the TCs using their phenotypic profiles revealed functionally similar TCs that led to a better understanding of chemical treatments and gene functions. Slack et al. [26] trained a Gaussian mixture model (GMM) to approximate the phenotypic distribution within the overall population. The GMM was then used to score the heterogeneous responses of HeLa cells to a set of drugs.

The above approaches do not focus on discovering novel phenotypes that reflect unforeseen interesting effects of TCs. In a large-scale HCS study, the number of biologically meaningful novel phenotypes can be huge, however unknown, due to the large variety of TCs. For example, a chemical compound library can contains hundreds of thousands of compounds. Moreover, many organisms have thousands of genes that various genetic perturbations can be applied to. Usually, researchers have limited knowledge about the effects of the majority of TCs on cells. It is very possible that many novel phenotypes will be left without much exploration if the analysis relies too much on a small set of predefined phenotypes or representative TCs. Hence it is important to develop a method for identifying novel phenotypes.

Yin et al. [27] fitted a GMM to the distribution of each existing phenotype, and used an improved gap statistics [28] to judge whether new images should be merged to a known phenotype or form a new phenotype. Their approach was successfully applied to image datasets of *Drosophila* embryos, Hela cells, and synthetic polygons. Nonetheless, this approach assigns equal weight to all image features. Different phenotypes can have quite different characteristics which are reflected in their differences in utilizing features (or feature weights). In this paper, we propose an approach that explores such characteristic differences to discover novel phenotypes. The basic idea is that a set of features not generally useful for defining known phenotypes can, however, encode the morphological characteristics of novel phenotypes.

## 2. METHODS

### 2.1. Novel Phenotype Discovery Framework

Our phenotype discovery framework is outlined in Figure 1. Assuming we already have defined some phenotypes (or known phenotypes). Each phenotype is defined by a set of positive images (belonging to the phenotype) and a set of negative images (not belonging to the phenotype). We first identify the underused features (UUFs) by comparing their distributions in the positive and negative image sets, as well as the control set. We then carry out extensive clustering analysis in a bootstrap fashion to create a relationship graph of all TCs. The analysis of the relationship graph yields a set of tight clusters that may represent novel phenotypes. We say that a group of images represents a novel phenotype if (a) they share some common morphological traits; (b) they form a compact and robust cluster; and (c) they are different from known phenotypes and the wild-type. Those clusters should be visually examined and filtered by biological experts. In the phenotype definition step, biologists start with the confirmed clusters and train classifiers to recognize novel phenotypes using content-based image retrieval with relevance feed-back (CBIR-RF) [29] techniques. Finally, the newly defined phenotypes will be utilized in the next round of novel phenotype discovery.

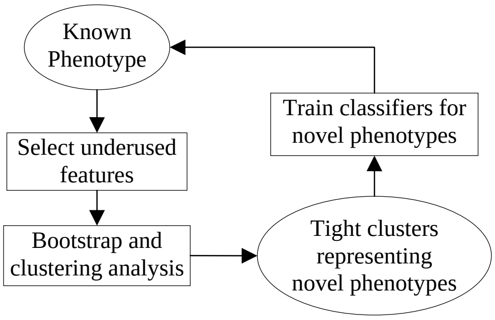
Known Phenotype
Select underused features
Bootstrap and clustering analysis
Tight clusters representing novel phenotypes
Train classifiers for novel phenotypes

**Figure 1**. The framework for novel phenotype discovery.

### 2.2. Selecting Underused Features

An UUF is defined as one that contributes little towards the recognition of known phenotypes. We can quantitatively evaluate how important a feature is in defining known phenotypes as shown below. For each known phenotype, we compare the distribution of a feature in its positive image set wtih that in its negative image set. This is done using the symmetrized form of the Kullback-Leibler divergence [30, 31], i.e., the *J*-divergence [32]. Let $p_f^c(x)$ and $q_f^c(x)$ be the distributions of a feature *f* in the positive and negative image sets of a phenotype *c*, respectively. The Kullback-Leibler divergence between $p_f^c(x)$ and $q_f^c(x)$ is:

$$KL(p_f^c \,\|\, q_f^c) = \int p_f^c(x) \log \frac{p_f^c(x)}{q_f^c(x)} dx \qquad (1)$$

The *KL* divergence is widely used to measure the difference between two distributions. However, it is asymmetric so that it is not appropriate in our application. We use its symmetrized form – the *J*-divergence – to measure the difference between $p_f^c(x)$ and $q_f^c(x)$:

$$J(p_f^c \,\|\, q_f^c) = \frac{KL(p_f^c \,\|\, q_f^c) + KL(p_f^c \,\|\, q_f^c)}{2} \qquad (2)$$

To test whether the distribution $g_f(x)$ of the feature *f* in the control condition (i.e., with the baseline treatment) is different from $p_f^c(x)$, we compute the *J*-divergence $J(p_f^c \,\|\, g_f)$ between each $p_f^c(x)$ and $g_f(x)$. Finally, we compute the overall *J*-divergence

$$\Psi_f = \sum\nolimits_c J(p_f^c \,\|\, g_f) + \sum\nolimits_c J(p_f^c \,\|\, q_f^c) \qquad (3)$$

to quantitatively indicate if the feature *f* is well-utilized by the known phenotypes. A high $\Psi_f$ value indicates that the feature *f* is useful in defining the existing phenotypes and distinguishing those phenotypes from the control. Otherwise, it is underused or a UUF.

To estimate $p_f^c(x)$, $q_f^c(x)$ and $g_f(x)$, we apply a Gaussian kernel around each data sample as proposed in [33]:

$$\tau_f(x) = \frac{1}{n\sqrt{2\pi}h}\sum_{i=1}^{n} \exp[-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2] \qquad (4)$$

where $\tau_f(x)$ represents $p_f^c(x)$, $q_f^c(x)$ or $g_f(x)$, *n* is the number of samples, and *h* is the bandwidth that can be estimated using the variance $\sigma$ of the feature *f* in those *n* samples as $h = 1.06\sigma n^{-1/5}$ [33].

### 2.3. Building a Relationship Graph

We then apply the clustering analysis using the UUFs to group TCs together. It is very likely that the TCs which are grouped together (or similar to each other) may share some novel phenotypic traits defined by UUFs. Since we do not have any prior knowledge about these novel phenotypes beyond the fact that they should be different from existing phenotypes, it is appropriate to start by using unsupervised learning techniques to discover TC clusters that may represent novel phenotypes. Clustering TCs instead of individual images will help discover novel phenotypes that are unevenly distributed across different TCs. Such phenotypes will be more interesting than those evenly distributed across many TCs, which may simply represent some common or trivial biological phenomenon instead of the effects unique to a few TCs.

To make sure our findings have enough coverage, we use 66 hierarchical clustering techniques [34]: the combination of 6 linkage analyses (average, complete, median, single, ward and weighted) and 11 distance measurements (Euclidean, Standardized Euclidean, Mahalanobis, Cityblock, Minkowski metric, cosine, Pearson Correlation, Spearman, Hamming, Jaccard and Chebychev Distance). In addition, we use the bootstrapping strategy to make sure the results are robust with respect to noise. The details of how to create a high quality relationship graph are explained below.

Let a HCS data set contain $m$ TCs. Each TC has $N_t$ images. Applying bootstrapping to each TC, we randomly sample $N_t$ images with replacement from it, which is then represented by the mean of its bootstrapped samples. Each hierarchical analysis technique is applied to the bootstrapped representations of all TCs. The GAP statistics [28] is applied to the clustering result to find the optimal number of TC clusters by comparing the change in the within-cluster dispersion with the expected result under a reference null distribution. The above bootstrapping and clustering analysis is repeated 1000 times. An *m-by-m* relationship matrix $M$ is created to store the results. In each bootstrapping and clustering analysis, we increase $M(a,b)$ by one if $\text{TC}_a$ and $\text{TC}_b$ are clustered together. By choosing a threshold, we can change the above relationship matrix into a relationship graph. Each node in the graph represents a TC. There is an edge in the graph connecting $\text{TC}_a$ and $\text{TC}_b$ if $M(a,b) > threshold$. The threshold can be chosen to be proportional to the average of the elements in $M$. Relationships discovered at a higher threshold are more robust.

### 2.4. Discovering Tight Clusters

We hypothesize that TCs sharing similar phenotypic traits should be clustered together under most circumstances. Therefore, we need to discover all possible cliques in the above relationship graph. Clique-finding is a NP-hard problem. Suboptimal solutions can be found by a simple algorithm [35] in at most $O(N^4)$ time. We define the following compact index to evaluate how likely it is that a clique represents a novel phenotype:

$$\Phi(clique) = \frac{E_{in}}{E_{total}} \tag{5}$$

where $E_{in}$ is the number of edges between the nodes of the clique and $E_{total}$ is the number of edges connected to the nodes of the clique. The higher the index of a clique, the more likely it is to represent a novel phenotype. For example, there are two cliques in Figure 2. The 1-2-3 clique has connections to three nodes (4, 5, and 9) outside of the clique while those outside nodes are sparsely connected to the clique. That is, each outside node is similar to only a small portion of nodes inside the clique. The compact index of this clique is hence as low as 3/6 = 0.5. The 6-7-8 clique only has only one connection to an outside node. Its compact index is equal to ¾. The second clique is more likely to represent a novel phenotype.

To further validate each tight cluster, we compute the $J$-divergences between the distribution of each UUF in the cluster and those in the control and in the positive samples of known phenotypes. A tight cluster will be abandoned if none of the above $J$-divergences is significantly large.

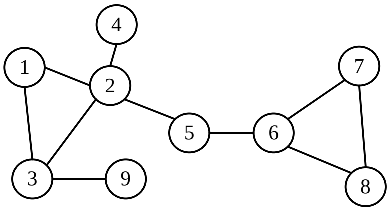

**Figure 2**. This graph shows two 3-node cliques: 1-2-3 and 6-7-8.

### 2.5. Refining Novel Phenotypes

Tight clusters are discovered in an unsupervised way. Therefore, they may not be perfectly accurate in defining these newly discovered novel phenotypes. They should be examined and refined by human experts. This can be done by using our content-based image-retrieval (CBIR) with relevance feedback (RF) software [29]. CBIR-RF techniques allow users to interactively and iteratively construct a classifier for a particular phenotype. Users usually start with a small image set of a phenotype as the query and ask the system to retrieve more images similar to the query. Users will then selectively mark some images as relevant or irrelevant. This feedback will be utilized by the system to construct a classifier for the desired phenotype. The above process can be iterated multiple times until the phenotype classifier cannot be improved further. Such a strategy allows biologists to actively apply their domain knowledge to efficiently refine any newly discovered novel phenotypes. It will also generate the positive (or relevant) image set and the negative (or irrelevant) image set of a phenotype, which can then be used to identify more novel phenotypes.

## 3. RESULTS

### 3.1. Dataset

We applied our method to an HCS image set that was generated to study the local signaling networks regulating the morphology of *Drosophila* BG-2 cells [36]. The study applied 249 TCs to *Drosophila* BG-2 cells and imaged about 12,600 individual treated cells. In each TC, a certain gene was either over-expressed or knocked down. For each cell, 145 image features were extracted to represent basic aspects of cell geometry, detailed aspects of cellular protrusions, the distribution and texture of GFP intensity within the cellular boundaries, and so on. We started with the positive samples and negative images of five phenotypes defined in [36]: large appearing wild-type, small cells with fuzzy edges, long and bipolar large cells, small round cells, and slim protrusions. Exemplar images of these phenotypes are shown in Figure 3.

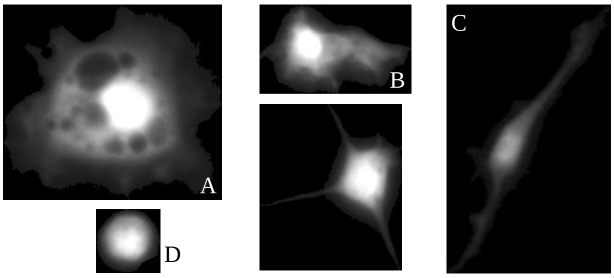

**Figure 3**. Representative images of five known phenotypes. (A) Large appear wild-type. (B) Small cells with fuzzy edges. (C) Long, bipolar large cells. (D) Small round cells. (E) Cells with slim protrusions.

### 3.2. Novel Phenotypes Discovered

We sorted all 145 features in terms of their overall *J*-divergence values which indicate their contributions in defining those phenotypes shown in Figure 3. We chose six most underused features that are listed in Table 1 (Please refer to [36] for detailed descriptions of these features).

**Table 1**. The selected UUFs

| ***Feature ID*** | ***Feature Name*** |
|---|---|
| 7 | MeanIntensity |
| 8 | StdIntensity |
| 9 | 90thPercentileIntensity |
| 37 | GFPIntensityLocationMutualInformation_8_15_24 |
| 99 | HiSmoothEllipticity |
| 132 | HiSmoothBndLargestAreaForProcessGE0.5 |

We discovered several interesting phenotypes. For example, a tight cluster contained three TCs: *Mp20* knockdown, *RacGAP50C* knockdown, and *pbl* knockdown. Visual examination revealed that a majority of cells under these three TCs have at least two nuclei (Figure 4A). This phenotype is not the focus of the original study, which makes this discovery even more interesting. The compact index of this clique is high as 0.75, indicating that it is compact and robust. The *J*-divergence values of the UUFs show that all the UUFs except for feature 99 help to separate this novel phenotype from the known phenotypes and the control (Figure 4B). These features are related to the mean and variation of all pixel intensities, the 90-percentile intensity (i.e., the brightest area), and the largest area of any process that has a maximum positive curvature >= 0.5 [36]. Visual inspection showed that the percentages of multi-nuclei cells under *Mp20, RacGAP50C, pbl* are 94.74%, 86.21% and 72%, respectively. We are thus highly confident in that this clique represents the multi-nuclei phenotype.

Literature search results strongly support this discovery. In *Drosophila*, *pbl* and *RacGAP50C* are related to cytokinesis, which is the final step in cell division and which is mediated by a complex and dynamic interplay between the microtubules of the mitotic spindle, the actomyosin cytoskeleton, and membrane fusion events [37]. *Drosophila RacGAP50C* and its homologues are essential for the formation of the central spindle and completion of cytokinesis [38-40]. Therefore, knocking down *pbl* and *RacGAP50C* could lead to unfinished cytokinesis and generate the multiple nuclei phenotype.

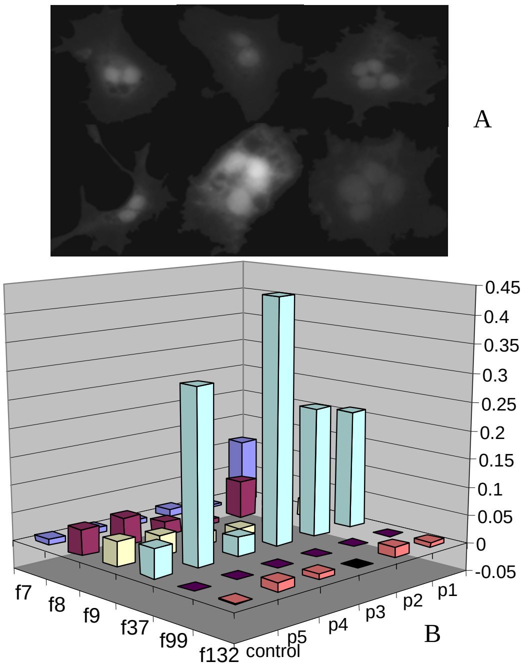

**Figure 4. (A).** Representative images of the phenotype represented by the first tight cluster (see text for details). (**B**). The *J*-divergence values of the selected UUFs (f7, f8, f9, f37, f99, and f132) with respect to the five existing phenotypes (p1-5) and the control.

Starting with the images in this cluster, we used our CBIR-RF technique [29] to train a classifier for recognizing this multi-nuclei phenotype. This newly trained classifier was used to evaluate all other TCs and identify two other TCs: *CG30158* knockdown and *Paxillin* knockdown. *Paxillin* is a focal adhesion-associated protein and is essential for completion of mammalian cytokinesis [41]. Hence, it is closely relevant to the multi-nuclei phenotype. Both *CG30158* and *Mp20* (including their orthologs in other species) have not been reported to be involved in cytokinesis. Our discovery suggests a new function of these two genes and predicts that they might contribute to the cytokinesis process.

In another discovery, we identified a tight cluster containing *mbc* knockdown, *MTL* knockdown, and *Actn* knockdown. Its compact index is 0.6. The majority of cells under these TCs are relatively small with rough lamellipodia protrusions (Figure 5A). The *J*-divergence values of the UUFs also show that this cluster is very different from the known phenotypes and the control (Figure 5B). *MTL* was reported to be involved in the actin filament bundle formation process and the lamellipodium assembly processes [42]. *Actn* was reported to be involved in the actin cytoskeleton reorganization process [43]. Hence, we extrapolate that *mbc* could be involved in the lamellipodium assembly process through controlling actin formation.

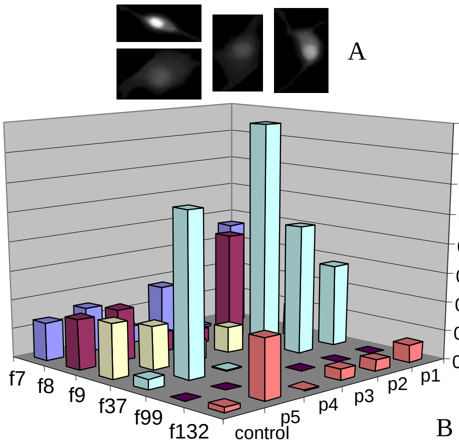

**Figure 5**. **(A).** Representative images of the phenotype represented by the first tight cluster (see text for details). (**B**). The *J*-divergence values of the selected UUFs with respect to the five known phenotypes and the control.

## 4. CONCLUSION AND DISCUSSION

Feature selection in high dimensional spaces has been an important and challenging problem in pattern recognition and machine learning research. Conventional approaches have so far focused on selecting features to boost performance on recognizing known patterns. This paper presents a new methodology that selects UUFs to mine novel phenotypes. The key idea is to utilize an obvious, and thus potentially easily ignored, observation that different phenotypes should have different characteristics which can be reflected by the differences in feature distributions. Our method directly explores such characteristic differences by selecting a set of features that is not useful for distinguishing existing phenotypes. This kind of dimensionality reduction approach effectively enhances weak signals that are keys to novel phenotypes and that can otherwise be easily overwhelmed by other image features. We show that the combination of the UUF concept with extensive unsupervised clustering analysis yields a powerful data mining tool. This is demonstrated by a successful application to analyze an HCS image dataset generated to study *Drosophila* BG-2 cells.

Our method offers the following advantages. First, the use of the UUFs effectively reduces the dimensionality of the feature space, and thus reduces the computational complexity. Second, it reduces the chance of finding a phenotype overlapping too much with the existing phenotypes because the distributions of the UUFs in the novel phenotypes are significantly different from those in the existing phenotypes. Third, unlike projective dimensionality-reduction approaches (e.g., Principal Component Analysis [44]) which transform the original feature space, it retrains the interpretability of features by using the subspace of the original feature space.

We also tried to replace the UUFs with one of the following three feature sets: all 145 features, the eigen features generated by Principal Component Analysis, and six randomly selected features. Then, we applied the same mining process: however, we failed to discover any novel phenotypes. This not only justifies the usage of the UUFs but also demonstrates the effectiveness of using these UUFs. Finally, our framework can be applied to mine novel patterns in other applications using data in high dimensional spaces.

## ACKNOWLEDGEMENT

CL and PH are supported by NIH R01 EB007042 to PH.

## 5. REFERENCES

1. Liebel, U., et al., *A microscope-based screening platform for large-scale functional protein analysis in intact cells.* FEBS Lett, 2003. **554**(3): p. 394-8.
2. Wheeler, D.B., A.E. Carpenter, and D.M. Sabatini, *Cell microarrays and RNA interference chip away at gene function.* Nat Genet, 2005. **37 Suppl**: p. S25-30.
3. Sonnichsen, B., et al., *Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans.* Nature, 2005. **434**(7032): p. 462-9.
4. Pelkmans, L., et al., *Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis.* Nature, 2005. **436**(7047): p. 78-86.
5. Muller, P., et al., *Identification of JAK/STAT signalling components by genome-wide RNA interference.* Nature, 2005. **436**(7052): p. 871-5.
6. Eggert, U.S., et al., *Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets.* PLoS Biol, 2004. **2**(12): p. e379.
7. Peng, H., *Bioimage informatics: a new area of engineering biology.* Bioinformatics, 2008. **24**(17): p. 1827-36.
8. Wollman, R. and N. Stuurman, *High throughput microscopy: from raw images to discoveries.* J Cell Sci, 2007. **120**(Pt 21): p. 3715-22.
9. Neumann, B., et al., *High-throughput RNAi screening by time-lapse imaging of live human cells.* Nat Methods, 2006. **3**(5): p. 385-90.
10. Perlman, Z.E., et al., *Multidimensional drug profiling by automated microscopy.* Science, 2004. **306**(5699): p.1194-8.
11. Pan, H., et al., *A novel small molecule regulator of guanine nucleotide exchange activity of the ADP-ribosylation factor and golgi membrane trafficking.* J Biol Chem, 2008. **283**(45): p. 31087-96.

12. Carpenter, A.E., *Image-based chemical screening.* Nat Chem Biol, 2007. **3**(8): p. 461-5.
13. Adams, C.L., et al., *Compound classification using image-based cellular phenotypes.* Methods Enzymol, 2006. **414**: p. 440-68.
14. Tanaka, M., et al., *An unbiased cell morphology-based screen for new, biologically active small molecules.* PLoS Biol, 2005. **3**(5): p. e128.
15. Young, D.W., et al., *Integrating high-content screening and ligand-target prediction to identify mechanism of action.* Nat Chem Biol, 2008. **4**(1): p. 59-68.
16. Mitchison, T.J., *Small-molecule screening and profiling by using automated microscopy.* Chembiochem, 2005. **6**(1): p. 33-9.
17. Boland, M.V. and R.F. Murphy, *A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells.* Bioinformatics, 2001. **17**(12): p. 1213-23.
18. Boland, M.V., M.K. Markey, and R.F. Murphy, *Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images.* Cytometry, 1998. **33**(3): p. 366-75.
19. Boland, M.V. and R.F. Murphy, *Automated analysis of patterns in fluorescence-microscope images.* Trends Cell Biol, 1999. **9**(5): p. 201-2.
20. Murphy, R.F., M.V. Boland, and M. Velliste, *Towards a systematics for protein subcelluar location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images.* Proc Int Conf Intell Syst Mol Biol, 2000. **8**: p. 251-9.
21. Chen, X. and R.F. Murphy, *Automated interpretation of protein subcellular location patterns.* Int Rev Cytol, 2006. **249**: p. 193-227.
22. Wang, J., et al., *Cellular phenotype recognition for high-content RNA interference genome-wide screening.* J Biomol Screen, 2008. **13**(1): p. 29-39.
23. Duda, R., P. Hart, and D. Stork, *Pattern Classification (2nd ed).* 2000, New York, NY: John Wiley and Sons, Inc.
24. Loo, L.H., L.F. Wu, and S.J. Altschuler, *Image-based multivariate profiling of drug responses from single cells.* Nat Methods, 2007. **4**(5): p. 445-53.
25. Chris Bakal, J.A., George Church, Norbert Perrimon, *Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology.* Science, 2007. **316**(5832).
26. Michael D. Slack, E.D.M., Lani F. Wu, and Steven J. Altschuler, *Characterizing heterogeneous cellular responses to perturbations.* PNAS, 2008. **105** (49): p. 6.
27. Zheng Yin, X.Z., Chris Bakal, Fuhai Li, Youxian Sun, Norbert Perrimon, Stephen TC Wong, *Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens.* BMC Bioinformatics, 2008. **9**(264).
28. Tibshirani, R., G. Walther, and T. Hastie, *Estimating the number of clusters in a dataset via the gap statistic.* J. R. Stat. Soc. Ser., 2001. **32**(2): p. 411-423.
29. Chen Lin, W.M., Pengyu Hong, Katharine Sepp, Norbert Perrimon. *Intelligent Interfaces for Mining Large-Scale RNAi-HCS Image Databases*. in *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*. 2007. Boston, MA.
30. Kullback, S. and R.A. Leibler, *On Information and Sufficiency.* The Annals of Mathematical Statistics, 1951. **22**(1): p. 79-86.
31. Basseville, M., *Distance measures for signal processing and pattern recognition.* Signal Process., 1989. **18**(4): p. 349-369.
32. Johnson, D.H. and S. Sinanovi´c, *Symmetrizing the Kullback-Leibler Distance.* IEEE Transactions on Information Theory, 2001.
33. Scott, D.W. and S.R. Sain, *Multi-dimensional Density Estimation.* Handbook of Statistics, 2004. **24**.
34. [cited; Available from: http://www.mathworks.com/access/helpdesk/help/toolbox/stats/index.html?/access/helpdesk/help/toolbox/stats/linkage.html.
35. Kim, H.-J. [cited; Available from: http://www.ibluemojo.com/school/clique_algorithm.html.
36. Chris Bakal, J.A., George Church, Norbert Perrimon. *Local Signaling Networks That Regulate Cell Morphology Defined by Quantitative Morphological Signatures (Supporting Online Material)*. 2007 [cited; Available from: http://www.sciencemag.org/cgi/content/full/sci;316/5832/1753/DC1.
37. Glotzer, M., *Animal cell cytokinesis.* Annu Rev Cell Dev Biol, 2001. **17**: p. 351-86.
38. Mishima, M., S. Kaitna, and M. Glotzer, *Central spindle assembly and cytokinesis require a kinesin-like protein/RhoGAP complex with microtubule bundling activity.* Dev Cell, 2002. **2**(1): p. 41-54.
39. Somers, W.G. and R. Saint, *A RhoGEF and Rho family GTPase-activating protein complex links the contractile ring to cortical microtubules at the onset of cytokinesis.* Dev Cell, 2003. **4**(1): p. 29-39.
40. Jantsch-Plunger, V., et al., *CYK-4: A Rho family gtpase activating protein (GAP) required for central spindle formation and cytokinesis.* J Cell Biol, 2000. **149**(7): p. 1391-404.
41. Shafikhani, S.H., K. Mostov, and J. Engel, *Focal adhesion components are essential for mammalian cell cytokinesis.* Cell Cycle, 2008. **7**(18): p. 2868-76.
42. Woolner, S., A. Jacinto, and P. Martin, *The small GTPase Rac plays multiple roles in epithelial sheet fusion--dynamic studies of Drosophila dorsal closure.* Dev Biol, 2005. **282**(1): p. 163-73.
43. Wahlstrom, G., H.L. Norokorpi, and T.I. Heino, *Drosophila alpha-actinin in ovarian follicle cells is regulated by EGFR and Dpp signalling and required for cytoskeletal remodelling.* Mech Dev, 2006. **123**(11): p. 801-18.
44. Jolliffe, I.T., *Principal Component Analysis* Springer Series in Statistics. 2002, NY: Springer.