Digging deep into Golgi phenotypic diversity with unsupervised machine learning

Shaista Hussain[a,†], Xavier Le Guezennec[b,†], Wang Yi[a], Huang Dong[a], Joanne Chia[b], Ke Yiping[c], Lee Kee Khoon[a], and Frédéric Bard[b,*]

[a]Institute of High Performance Computing and [b]Institute of Molecular and Cell Biology, Singapore 138673; [c]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798

ABSTRACT **The synthesis of glycans and the sorting of proteins are critical functions of the Golgi apparatus and depend on its highly complex and compartmentalized architecture. High-content image analysis coupled to RNA interference screening offers opportunities to explore this organelle organization and the gene network underlying it. To date, image-based Golgi screens have based on a single parameter or supervised analysis with predefined Golgi structural classes. Here, we report the use of multiparametric data extracted from a single marker and a computational unsupervised analysis framework to explore Golgi phenotypic diversity more extensively. In contrast with the three visually definable phenotypes, our framework reproducibly identified 10 Golgi phenotypes. They were used to quantify and stratify phenotypic similarities among genetic perturbations. The derived phenotypic network partially overlaps previously reported protein–protein interactions as well as suggesting novel functional interactions. Our workflow suggests the existence of multiple stable Golgi organizational states and provides a proof of concept for the classification of drugs and genes using fine-grained phenotypic information.**

Monitoring Editor
Thomas Sommer
Max Delbrück Center for Molecular Medicine

Received: Jun 14, 2017
Revised: Sep 8, 2017
Accepted: Oct 4, 2017

INTRODUCTION

RNA interference (RNAi) screening combined with high-throughput imaging provides a powerful experimental means of investigating the genetic regulation of subcellular structures. High-throughput imaging can acquire cell images for thousands of different treatments, requiring computationally driven image analysis. To characterize cellular phenotypes elicited by treatments, the simplest approaches rely on a dedicated, directed image analysis using one or a few image features. But obviously the phenotypes characterized are limited.

Today, image analysis can generate hundreds of numerical features for each cell image, opening up the possibility of high-content analysis and the characterization of multiple phenotypes. To convert image features into cell phenotypes, high-content analysis often relies on supervised machine learning. In this case, phenotypes are assigned to sample cells after an algorithm has been trained with sets of reference cells selected by an expert. In effect, the machine learning algorithms automate a classification scheme previously defined by a user (Conrad and Gerlich, 2010; Sommer and Gerlich, 2013). Obviously, supervised machine learning approaches are constrained by the human expert, who has to select a set of reference cell images. Although an experienced user may be able to recognize cellular phenotypes visually, it is clear that our visual system has not evolved to analyze patterns of subcellular structures in microscopic images reliably. Furthermore, visual classification cannot guarantee objectivity; it may be subject to personal bias due to prior assumptions, a problem recognized across multiple scientific disciplines (Lindblad *et al.*, 2004; Bamford *et al.*, 2009). Supervised machine learning and various high-content image–based analysis approaches

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E17-06-0379) on October 11, 2017.

The authors declare no competing financial interest.

†These authors contributed equally to this work.

Author contributions: S.H., W.Y., K.Y., H.D., and L.K.K. generated analysis codes in Python and clustering results, generated figures, analyzed data, and wrote the manuscript; X.L.G. and J.C. performed the biological experiment; X.L.G. analyzed data, generated figures, and wrote the manuscript, F.B. designed the study, analyzed the data, and wrote the manuscript.

*Address correspondence to: Frédéric Bard (fbard@imcb.a-star.edu.sg).

Abbreviations used: CDFs, cumulative distribution functions; COG, conserved oligomeric Golgi complex; CV, coefficient of variation; D-PBS, Dulbecco's phosphate-buffered saline; ER, endoplasmic reticulum; GalNAc-Ts, O-GalNAc glycosylation initiation enzyme; GMM, Gaussian mixture model; HCSU, high-content screening unit; HPL, *Helix pomatia* lectin; NQC, nuclear quality control; NT, nontargeting siRNA control; PCA, principal component analysis; Plk, Polo-like kinase; RF, random forest; RS, reproducibility score; SNARE, soluble NSF (*N*-ethyl-maleimide–sensitive factor) attachment protein receptor; STX, syntaxin; SVM, support vector machine.

have been used previously to study organelles such as the Golgi apparatus by characterizing three or four phenotypes (Farhan *et al.*, 2010; Chia *et al.*, 2012; Anitei *et al.*, 2014; Millarte *et al.*, 2015).

The Golgi apparatus is an architecturally extremely complex organelle under constant dynamic membrane flow in both the anterograde and retrograde directions. A series of four to seven cisternae, ribbon-like membranes tightly opposed to one another, form the Golgi stacks, which are flanked on each side by a network of vesicles/tubules. Several dozen to several hundred Golgi stacks are loosely linked together in a perinuclear location (Lowe, 2011). The Golgi apparatus is essential for protein posttranslational modifications, including glycan addition and protein sorting (Chia *et al.*, 2012; Goh and Bard, 2015). It also participates in the control of cell migration (Yadav *et al.*, 2009; Millarte *et al.*, 2015). Golgi organization is thought to be particularly important for the manufacture of glycans by glycosylation enzymes in an orderly distribution in specific cisternae (de Graffenried and Bertozzi, 2004; Stanley, 2011). The Golgi apparatus is thought to depend on a complex genetic network (Lowe, 2011). The precise nature of this network remains unclear, however, and is a prime target for cell-based genetic screening.

Results from a kinome/phosphatome small interfering RNA (siRNA) screen on ERGIC-53 localization were interpreted as the Golgi apparatus being either fragmented or with a tubular aspect (Farhan *et al.*, 2010). A recent restricted siRNA multiparametric screen set looking at the relation between Golgi morphology and migration used giantin as a Golgi marker and used three typical phenotypic Golgi morphological classes: small, fragmented, and big Golgi (Millarte *et al.*, 2015). Additionally, a kinome/phosphatome siRNA high-content imaging screen with a mannose-6-phosphate receptor reported some selected hits from screens to be affected by the extent of *trans*-Golgi fragmentation (Anitei *et al.*, 2014). All these studies defined Golgi phenotypes based on a chosen set of features and image training labels and were guided by human visual expertise. Similarly, our laboratory previously used a supervised machine learning approach to classify three morphologies of Golgi apparatus within single cells as diffuse, condensed, or fragmented (Chia *et al.*, 2012). However, the larger number of glycosylation profiles associated with various gene perturbations suggested that more Golgi organizations might exist.

Unsupervised machine learning methods allowing extraction of phenotypes independent from user image annotations have the potential to distinguish more phenotypes. Unsupervised methods work by identifying the underlying structure in the input data in the absence of any defined output. For example, a clustering approach discovers inherent groupings/categories based on the distribution of data points in a feature space (Sommer and Gerlich, 2013). These methods have been used, for example, to profile drugs and to classify time-lapse cellular imaging data in mitosis progression (Pelkmans, 2012; Zhong *et al.*, 2012). An unsupervised method also identified key cell heterogeneity during preadipocyte differentiation and revealed cellular subpopulations in lung cancer clones that were resistant versus sensitive to paclitaxel (Slack *et al.*, 2008; Loo *et al.*, 2009; Singh *et al.*, 2010).

Unsupervised learning can be used to characterize cellular phenotypes automatically; however, its use is challenging because of its relatively poor performance on noisy data, which is further limited by the lack of interpretation in the case of an unpredictable output (Sommer and Gerlich, 2013). To the best of our knowledge, it has not been used to date for the study of a subcellular structure such as the Golgi apparatus.

In this study, we present an unsupervised methodology for identifying Golgi phenotypes. We use a single marker to focus the phenotypic analysis as much as possible. We address three problems that commonly limit the use of unsupervised learning methods, such as clustering. The high sensitivity to noise is alleviated by extensive data quality checks. The naturally occurring phenotypic diversity is handled through a step of control morphology modeling. Finally, clustering results are validated using reproducibility as a significance metric. Using analysis at the single-cell level, with a glycosylation-dependent Golgi marker, a small library of siRNAs targeting gene transcripts important for membrane traffic and validation through reproducible associations with specific genetic perturbations, we detected more than 10 distinct Golgi phenotypes. This study demonstrates the existence of multiple stable Golgi organizations. Furthermore, we use this phenotypic information to build a phenotypic network that maps similarities between genetic perturbations. The workflow presented here can be adapted in other studies of high-content data sets to maximize the use of image-based high-dimensional data and the efficiency of phenotypic distinctions.

RESULTS

Golgi morphological RNAi screen: marker, target genes, and initial validation

To image Golgi morphology in HeLa cells, we used immunofluorescence staining with *Helix pomatia* lectin (HPL) and Hoechst to stain the nucleus as described previously (Chia *et al.*, 2012). HPL binds specifically to terminal N-acetylgalactosamine (GalNAc) residues that are added to proteins by O-GalNAc glycosylation initiation enzyme (GalNAc-Ts) located at the Golgi apparatus. GalNAc-Ts can relocate to the endoplasmic reticulum (ER) under some conditions, leading to staining with a diffuse morphology.

To define a set of genetic perturbations with a high probability of affecting Golgi morphology, we selected siRNA pools targeting SNARE (soluble NSF (*N*-ethyl-maleimide–sensitive factor) attachment protein receptor and COG complex (conserved oligomeric Golgi complex) transcripts (Supplemental Figure 1). The SNARE family of proteins mediate membrane fusion events via v-SNARE located on vesicle membranes and t-SNARE localized on target compartment membranes. The Golgi apparatus is a major center of membrane trafficking and therefore is dependent on various SNAREs (Hong, 2005; Hong and Lev, 2014; Malsam and Sollner, 2011). The COG complex plays an essential role in retrograde transport within the Golgi by providing tethering functions (Willett *et al.*, 2013). We reasoned that the disturbance in SNARE/COG targets would impact Golgi organization and reveal various Golgi phenotypic classes.

We assembled data sets for multiple screening plates and verified their experimental and biological reproducibility with several controls (details under *Reproducibility tests and replicates*). We used the Polo-like kinase (Plk) siRNA, which typically induces an over 99% decrease in cell count from mock transfected wells to verify efficient siRNA transfection in all experimental plates (Supplemental Figure 2A). Previously, we reported that depletion of the SNARE syntaxin-5 (STX5) induces a diffuse Golgi morphology with an increase in HPL staining intensity (Chia *et al.*, 2012). We used the median cell intensity of HPL staining in STX5-depleted wells to calculate a Z' for all screening plates, which was found to be between 0.36 and 0.4 (Supplemental Figure 2B and Supplemental Table 1). Analysis of the Hoechst nuclei count revealed 900–2100 nuclei/treatment depending on the siRNA treatment tested (Supplemental Figure 2C). The total nuclei counts showed a mean coefficient of variation (CV) <20% between well replicates, indicating good consistency. Overall these initial assessments indicated reasonable reproducibility and warranted further processing by the single-cell extraction pipeline.

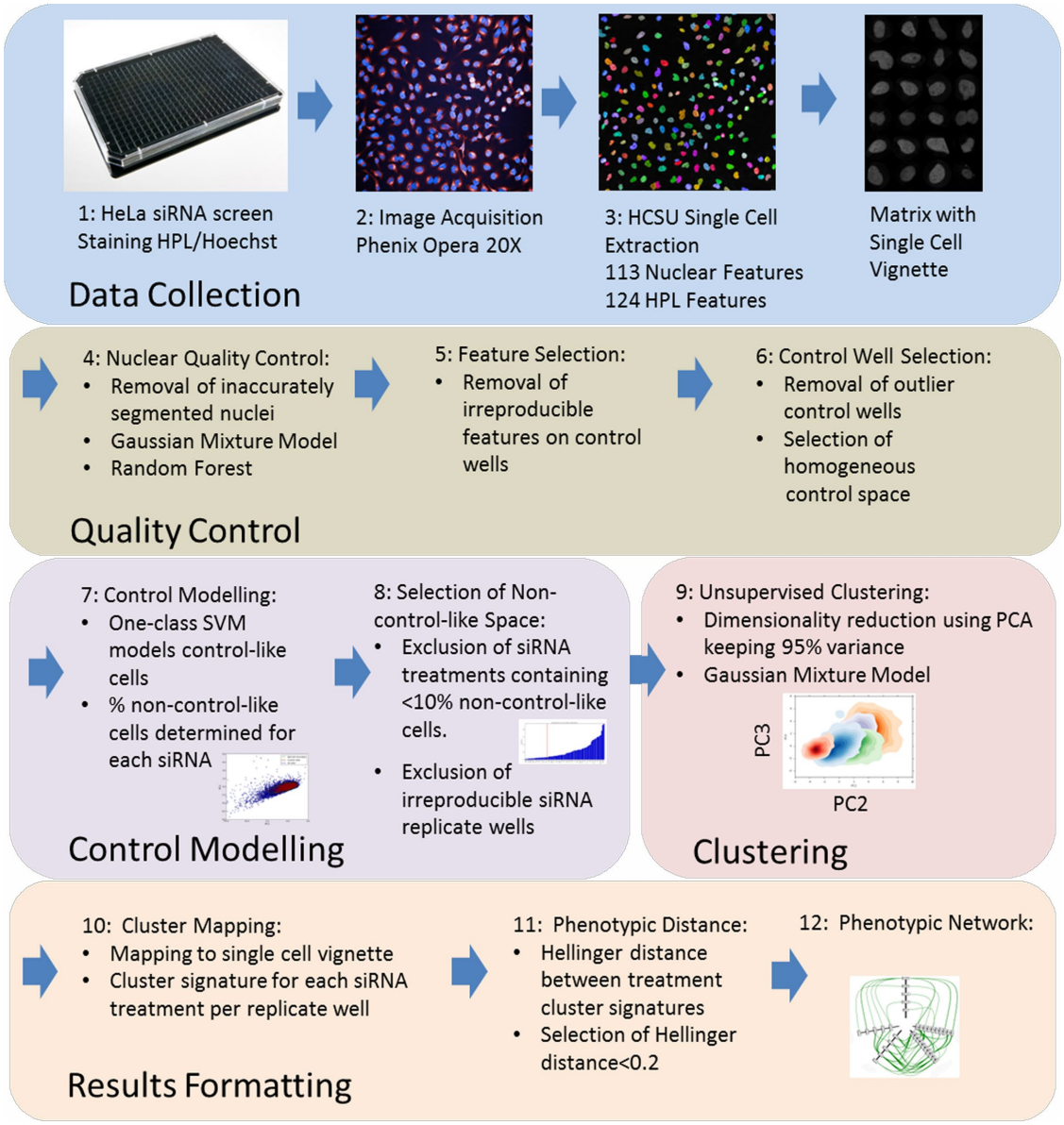

FIGURE 1: Workflow of unsupervised pipeline to uncover Golgi phenotypic classes.

Single-cell extraction pipeline: segmentation and feature extraction

We processed images using a high-content screening unit (HCSU), a dedicated automatic high-throughput workflow for the extraction of numerical features (Tjhi *et al.*, 2011; Chia *et al.*, 2012). HCSU also extracts single-cell images that can be used as cell vignettes (Figure 1). A wavelet-based segmentation and watershed algorithm for nuclear staining identifies individual nuclei and defines a cell territory from the nuclear center of mass. Images of single cells are then extracted as a vignette. Numerical features including object shape, area, and fluorescence intensity and texture were extracted for both nuclear and Golgi staining, resulting in 113 nuclear and 124 HPL features (Supplemental Table 2).

Overview of the deep phenotypic analysis workflow

Because unsupervised analysis is particularly sensitive to artifacts and noisy data, we derived a multistep approach to produce high-quality data sets and replicates at multiple levels to test for reproducibility (Figure 1). The proposed workflow for constructing image-based Golgi phenotypes consists of three major modules involving 1) a multistep machine learning–based quality control (QC) module to produce high-quality data sets, 2) control modeling to identify the non–control-like treated cells (with altered Golgi morphology content), followed by 3) unsupervised clustering analysis to identify novel types or subtypes of the known Golgi phenotypes. The first module focuses on data QC by eliminating the low-quality data at multiple levels of cells, wells, and features. The output of this module yields high-quality nuclei or cells, reproducible features, and homogeneous replicate wells. The second module identifies normal "unaltered Golgi"–looking cells in the whole data set and excludes them from subsequent unsupervised clustering. A control "unaltered Golgi" model is fitted to the high-quality control cells obtained from the previous module. The control model then predicts for all the remaining cells if they are control-like (with unaltered Golgi) or non–control-like (with altered Golgi), yielding a penetrance score, defined as the percentage of non–control-like cells for each

siRNA well or treatment. For the next step, all the replicate wells with low penetrance were excluded, retaining only siRNA treatments with a significant effect. In the third module, unsupervised clustering was performed on non–control-like cells from these high-penetrance siRNA treatment wells. The resulting Golgi phenotypic clusters formed treatment fingerprint signatures, which were compared pairwise by calculating Hellinger distance (Vajda, 1989). Finally, a Golgi phenotypic network was constructed using the Hellinger distance values between all pairs of treatments. The details of different phases and performance indicators of this workflow are given in the following sections.

Reproducibility tests and replicates

We devised different types of replicates to test for potential sources of variability and validate our results. To test the reproducibility of association of clusters with specific gene depletions, "well replicates" were acquired in each plate, with four wells for each siRNA treatment (Supplemental Figure 1). To test the reproducibility of the clustering algorithm, "technical replicates" were generated, each composed of an independent set of optical fields originating from the same well. Finally, to test the overall reproducibility of the approach, we acquired data sets from two independent "biological replicates" performed weeks apart. Hence, our data sets follow the respective nomenclature: Biological Replicate 1 with first set of optical fields (Biological Replicate 1/Technical Replicate 1) or second set of optical fields (Biological Replicate 1/Technical Replicate 2), Biological Replicate 2 with first set of optical fields (Biological Replicate 2/Technical Replicate 1) or second set of optical fields (Biological Replicate 2/Technical Replicate 2). We refer to GMMs (Gaussian mixture models) 1–4 as independent unsupervised runs of our developed pipeline with these four data sets: Biological Replicate 1/Technical Replicate 1 (GMM1), Biological Replicate 1/Technical Replicate 2 (GMM2), etc.

Nuclear quality control

Because unsupervised analysis is sensitive to artifacts and noisy data (Sommer and Gerlich, 2013), we reasoned that low-quality nuclei/cells resulting from poor segmentation, overlapping, and out-of-focus samples can affect its performance. Watershed algorithms can misidentify touching nuclei and lead to outputs with imperfectly isolated nuclei (Zhang *et al.*, 2015). Hence, we developed a nuclear quality control (NQC) step to remove these irregularly segmented nuclei (Figure 1). First, we generated nuclear clusters to select the samples for manual labelling of "high" and "low" quality nuclei. This was done by utilizing a GMM to perform unsupervised clustering on all cells and then randomly selecting a small number of cells from each cluster of nuclei for labelling (Bishop, 2006). The use of unsupervised clustering yielded different nuclear subpopulations that can capture all the nuclear image variations, and hence allowed us to generate an efficiently labeled small subset of samples.

Figure 2A shows 25 samples each from three of the total 18 nuclear clusters that represent the different nuclear subpopulations. From each of these clusters, 50 nuclear samples were randomly selected and manually labeled as high- or low-quality nuclei. Next, these labeled cell samples were used to train a random forest (RF) classifier, which then automatically removed all low-quality cells (Breiman, 2001). Figure 2B shows the samples of nuclei classified as good (well segmented, in focus) and bad (poorly segmented, overlapping). The set of manually labeled nuclear samples was divided into a training set to train the RF classifier and a test set to measure its performance. We trained the RF classifier for NQC using labels from one plate replicate, which was then used to automatically predict the good and bad cells for all the replicates tested.

To measure the performance of the RF classifier, we computed the precision and recall scores for the test data set. The precision score is defined as the fraction of high-quality cells out of the total number of correctly classified cells, and the recall score is a measure of how many high-quality cells are correctly classified. Because both precision and recall scores should be high, the training process for the RF classifier involved maximizing the area under the precision–recall curve (Mukhamedyarov *et al.*, 2016).

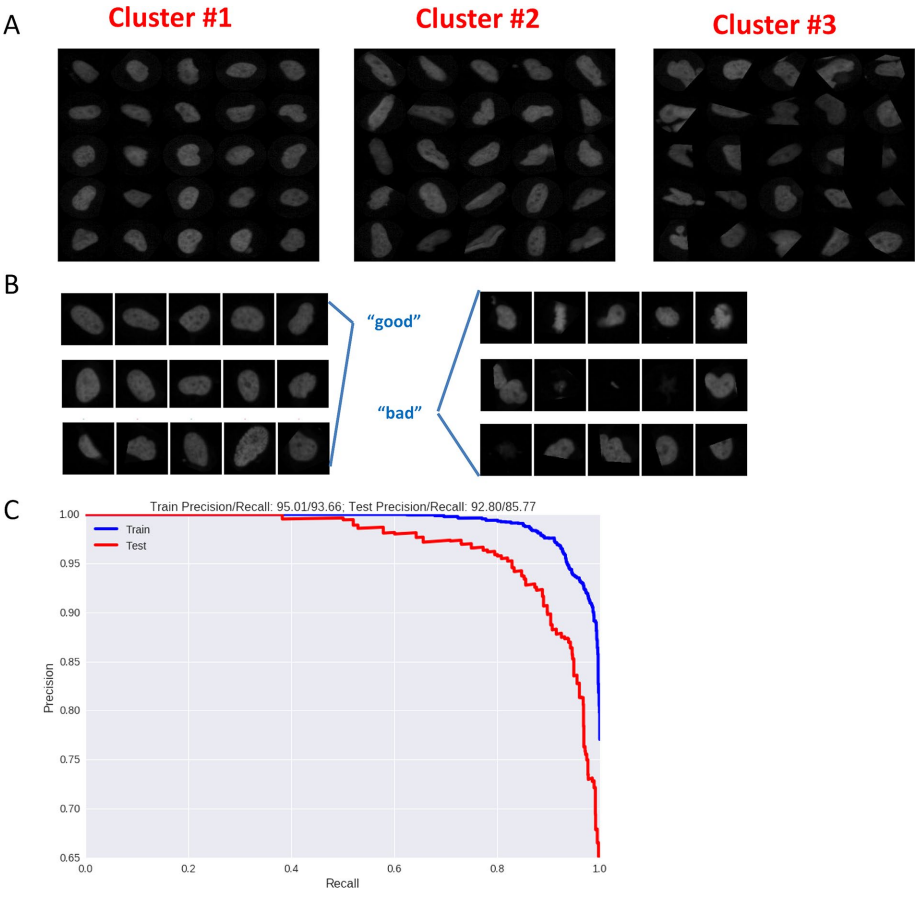

FIGURE 2: NQC with machine learning. (A) Nuclear clusters example obtained from a GMM applied on all data sets from the HCSU segmentation output. (B) Example of high- vs. low-quality nuclei predicted by the RF classifier trained on manually labeled nuclei. (C) Performance curve of RF with recall % (proportion of good nuclei identified out of total fraction of good nuclei) on horizontal axis and precision % (proportion of good nuclei assigned as correct) on vertical axis with a training data set (80% of labeled data) and a test set (20% of labeled data).

As shown in Figure 2C, area under the curve for the training data (blue) is only slightly higher than that for the test data (red), with the precision/recall scores for training data computed as 95.01/93.66 and for test data as 92.80/85.77, suggesting that the RF is robust to overfitting on the training data. Our approach yielded a mean of 700 quality nuclei/well from an input mean of 1700 nuclei/well (Supplemental Figure 3). Total nuclei counts for all plate replicates tested produced an average CV of 17%, with a maximal CV of 30.7%.

Feature quality control and control well selection

A large number of numerical features can be extracted from microscopic images, and it is not always clear in advance which ones are most useful or which are more sensitive to noise. For an unsupervised approach, eliminating noisy features is essential to avoid polluting the feature space in which unsupervised clustering is performed. Similarly, high-throughput experiments in multiwell plates can generate outlier wells, which also add noise to the clustering space.

After various trials, we decided to use a method that eliminated noisy image features and outlier control wells simultaneously (control wells refer to mock untransfected wells in Supplemental Figure 1). This combined "feature–well" selection method was motivated by the fact that an outlier well can influence the evaluation of otherwise reliable features, and reciprocally, a noisy feature can affect the evaluation of wells.

Empirical cumulative distribution functions (CDFs) for all features were generated by considering the cell population from each control well independently. The differences between these CDFs were measured to evaluate the reproducibility of a feature across multiple control wells. Figure 3A illustrates an example of how a noisy feature or noisy well is eliminated. Starting with the four features (F_1–F_4), the CDFs plotted for four wells (A_1–A_4) show that the feature F_1 is noisy across all wells, and hence it is eliminated. Further, the well A_4 (light blue) is eliminated, because it contributes to noise across all three remaining features. Hence, for this example, a set of three robust features across three homogeneous wells are selected.

The difference between feature CDFs was measured using the Kolmogorov–Smirnov statistic, which was then used to derive a reproducibility score (RS) to perform feature–well selection. This score was computed by leaving out one feature or one control well at a time and determining the number of uniform distributions for all remaining features generated across all remaining wells (Supplemental Methods). Figure 3B shows the RS corresponding to all

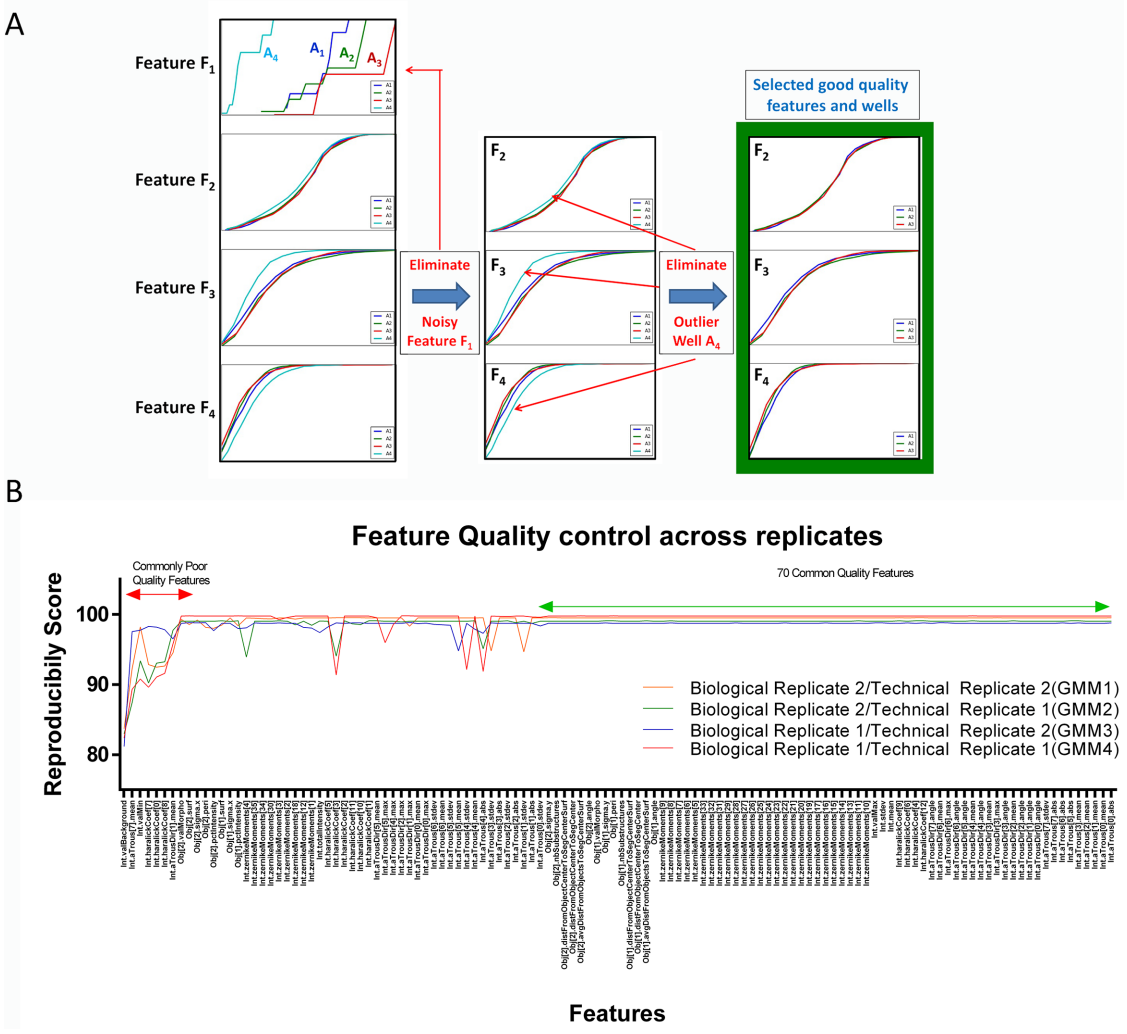

FIGURE 3: Feature–control-well QC. (A) Example of feature–well selection principle used in this workflow. (B) Reproducibility scores across all features used in this study in independent analysis with the presented workflow. Left features indicated with red arrow are commonly rejected while right features indicated with green are commonly accepted across all replicates used in this study.

features as each feature is eliminated, from left to right. The feature RS shown for all four experimental replicates increase from left to right as the poor-quality features are removed, until no further improvement in RS can be attained. The features delineated by the red arrow are eliminated across all four replicates, while those delineated by the green arrow are high-quality robust features selected for all the replicates and hence are used for further analysis. Remarkably, not all Haralick features showed similar quality levels. Haralick coefficients 9, 6, 4, and 2, respectively equivalent to Haralick difference variance, sum variance, inverse difference moment, and correlation, were highly reproducible, whereas Haralick coefficients 7, 0, and 8, respectively equivalent to Haralick sum entropy, angular second moment, and entropy, were particularly unreproducible (Murphy *et al.*, 2003).

Overall, we have established a method for selecting a large set of features and wells with high reproducibility for further processing without a priori knowledge. Given the diversity and complexity of image features and the inevitable occurrence of problematic wells in high-throughput experiments, such an automated approach should be highly valuable.

Control space modeling defines cells with altered Golgi morphology

We hypothesized that unsupervised clustering would be less efficient at defining specific Golgi phenotypes if both wild-type, unaltered Golgi (control-like cells) and altered Golgi (non–control-like cells) were used together. Indeed, wild-type Golgi morphologies are relatively diverse and principal component analysis (PCA) shows that both cell populations distribute in a continuous manner in the multiparametric phenotypic space (Figure 4A). In contrast, if only non–control-like cells are used, the clusters obtained are more likely to represent phenotypes clearly different from the wild type. In addition, removal of cells with control-like Golgi apparatus reduces the size of the data set used for unsupervised clustering, which shortens the processing time.

Therefore, to define the control Golgi phenotype, we modeled a control volume in the multiparametric space using cells from mock control wells by utilizing a one-class support vector machine (SVM) (Bishop, 2006). Figure 4A shows the decision boundary for the control space (green) learned by the control model in a two-dimensional space spanned by the first two principal components of the selected Golgi features. This region encompasses 95% of control cells, the remaining 5% of control cells being considered outliers. All sample cells were then classified using the trained one-class SVM. This sorting exercise yielded the percentage of non–control-like cells for a treatment, which we call the penetrance score.

To determine the optimal number of control cells for defining the control-like Golgi model, we followed the penetrance scores for two test treatments as a function of the number of control cells. The test treatments were STX5 siRNA, a positive control with a marked diffuse Golgi phenotype (Supplemental Figure 2B), and a nontargeting siRNA negative control (NT).

On the average, STX5 depletion had a higher penetrance score (> 30%) than that for NT controls (<8%) (Supplemental Figure 4). The penetrance score for both treatments increased significantly when the number of control cells was below ~8000 (10 wells), whereas it was stable if the number of control wells was larger (Supplemental Figure 4). This suggests that if the control group is too small, the diversity of the wild-type morphologies is not fully captured, while the penetrance value is stable once a minimum number of control cells are provided to the control model.

For replicates tested in this work, we used around 20 control wells (~16,000 cells) to define a control Golgi model. This approach resulted in the classification of between 1.8×10^5 and 2.5×10^5 control-like cells and between 1.9×10^4 and 2.5×10^4 non–control-like cells for the different replicates tested (Figure 4B).

Next, we eliminated all wells for which the penetrance score was less than 10% (red dotted line in Figure 5A). The NT siRNA produced a 5.4% (±2.8%) penetrance score. These scores were highly consistent between technical and biological replicates with a coefficient of determination, R^2, above 0.9 (Figure 5, B and C).

We used this measure of penetrance to evaluate siRNA on-targeting of the treatments. We used four deconvoluted siRNAs for each treatment and repeated the analysis of penetrance for an effect on Golgi morphology. Treatments that showed at least two single siRNA with a penetrance score >10% were considered validated.

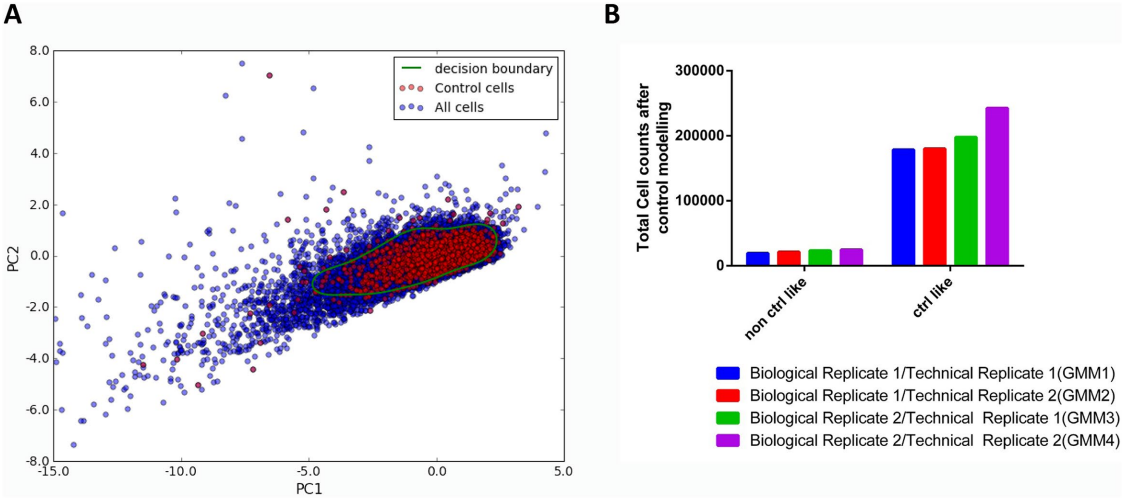

FIGURE 4: Control modeling with one-class SVM. (A) Scatterplot example of control model learned using SVM approach with all cells depicted across two major components after PCA on all features. Control cells are depicted in red and the remaining cell population is depicted in blue. Green curve represents boundary learned for defining control cells space with one-class SVM. (B) Size of total control and non–control-like spaces produced by independent SVM control modeling on replicates tested.

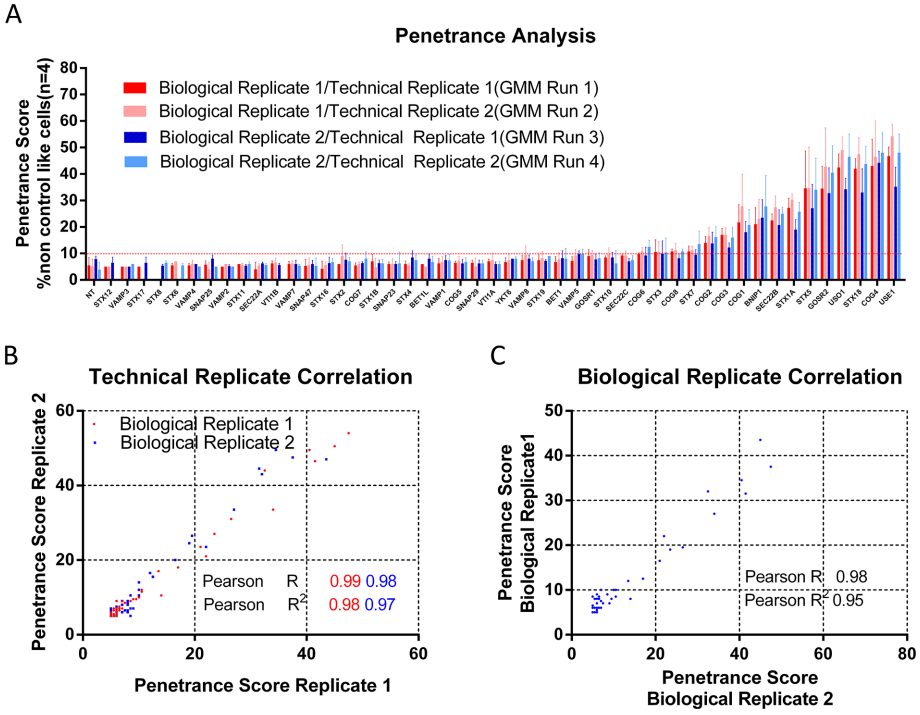

FIGURE 5: Detection of penetrance (% non–control-like cells) in biological/technical replicates. (A) Bar chart with mean and SD of % non–control-like cells for indicated siRNA treatments (from HPL–derived fluorescence channel) on horizontal axis based on four wells replicate. Control modeling was performed on various replicates. Red dotted line represents threshold for significant penetrance cutoff at 10%. (B) Correlation analysis of penetrance presented in A between sets of technical replicates for biological replicates 1 (red) and 2 (blue). Pearson correlation coefficient R and R^2 are indicated in respective replicate colors. (C) Correlation analysis of penetrance presented in A between biological replicates 1 and 2

Using this approach, we could validate ~75% of our siRNA pools, a proportion consistent with previous screens (Supplemental Figure 5).

Unsupervised clustering identifies multiple Golgi morphologies

To run the unsupervised clustering, we first performed dimensionality reduction using PCA on the image features. The use of PCA reduced data dimension from ~100 to ~30, preserving 95% of the total variance. Unsupervised clustering was then performed on the selected non–control-like population using a GMM. This resulted in splitting of non–control-like cells into 12, 10, 10, and 14 clusters in the different GMM runs (Figure 6A). A major cluster was defined as a group of non–control-like cells with an arbitrary minimal number of 100 cells/cluster. Clusters with low cell numbers were excluded. Cell count/cluster could scale up to ~2000 cells/cluster, depending on the cluster.

We plotted the five main clusters in GMM2 in the first four principal components space in a pairwise manner (Figure 6B). The density maps indicated that the centers of mass of the clusters are well separated, but that cluster boundaries tend to overlap. Images of representative cells for each cluster group were extracted from the bank of cell vignettes, and examples of typical non–control-like Golgi morphologies in each cluster group were assembled (Figure 7). Morphologies produced by unsupervised clustering agreed to some extent with our previously reported visual Golgi phenotypic classification of diffuse/condensed/fragmented, but obviously extended beyond these three visual classes (Chia *et al.*, 2012).

Cluster analysis dendrogram plots illustrate cluster relationships

To provide insight into the relationships between the phenotypic clusters, we performed hierarchical clustering on the means of the GMM components. The dendrograms plotted for the cluster means in GMM1 and GMM2 are shown in Supplemental Figure 6. The pairwise similarity between clusters was computed using Euclidean distance between means of clusters (*y*-axis) generated by a GMM. These cluster means are defined in the PCA space of the image features. For GMM1, pairs of clusters C3–C4 and C2–C5 with diffuse morphology were found to be closest to each other, with the distance between these clusters being the smallest, compared with all other pairwise distances. Moreover, clusters C9 and C12 were found to be closer to each other than to any other clusters, as shown by the first split (blue lines), and C11 was the most different from all the remaining clusters (red lines). Similarly, for GMM2, distance between the most similar clusters C2, C4 with diffuse morphology was the smallest, while C8 was found to be different from all the other clusters (first split).

Relevance of Golgi image features to the clustering structure

To identify the most relevant morphological features for the GMM clusters, we performed feature analysis by first fitting a RF classifier to the feature data to learn the cluster outputs generated by the GMM. The use of a RF consisting of several decision trees provided a direct method for measuring the importance of features by computing the Gini impurity, which is used to decide how to branch the decision trees. Hence, importance scores were generated for each feature as the average decrease in impurity from each feature. For this purpose, we considered the 70 commonly selected high-quality features across all four replicates (Figure 3B), as discussed under *Feature quality control and control well selection*. Supplemental Figure 7 shows the bar plot for these features, ranked according to decreasing importance pertaining to the clustering structure for the GMM clusters in GMM1 and GMM4, where the most important image features common to both GMMs are mean, maximum, and SD values of Golgi image intensity, some Haralick coefficients (6, 9, 12) (matching respectively Haralick difference variance, difference entropy, and Haralick info measure of correlation 2) (Murphy *et al.*, 2003), and Obj.nbSubstructures, Obj .avgDistFromObjectsToSegCenterSurf features.

Clusters are reproducibly associated with specific siRNA treatments

Next, we explored how these various phenotypic clusters relate to different siRNA treatments. A signature composed of the specific fractional number of cells present in each cluster was defined for each siRNA treatment and demonstrated using a polar plot. Representative examples are shown in Figure 8 and all signatures are available in Supplemental Table 3. Replicate well siRNA treatment signatures were overlaid on the same polar plot to visualize how

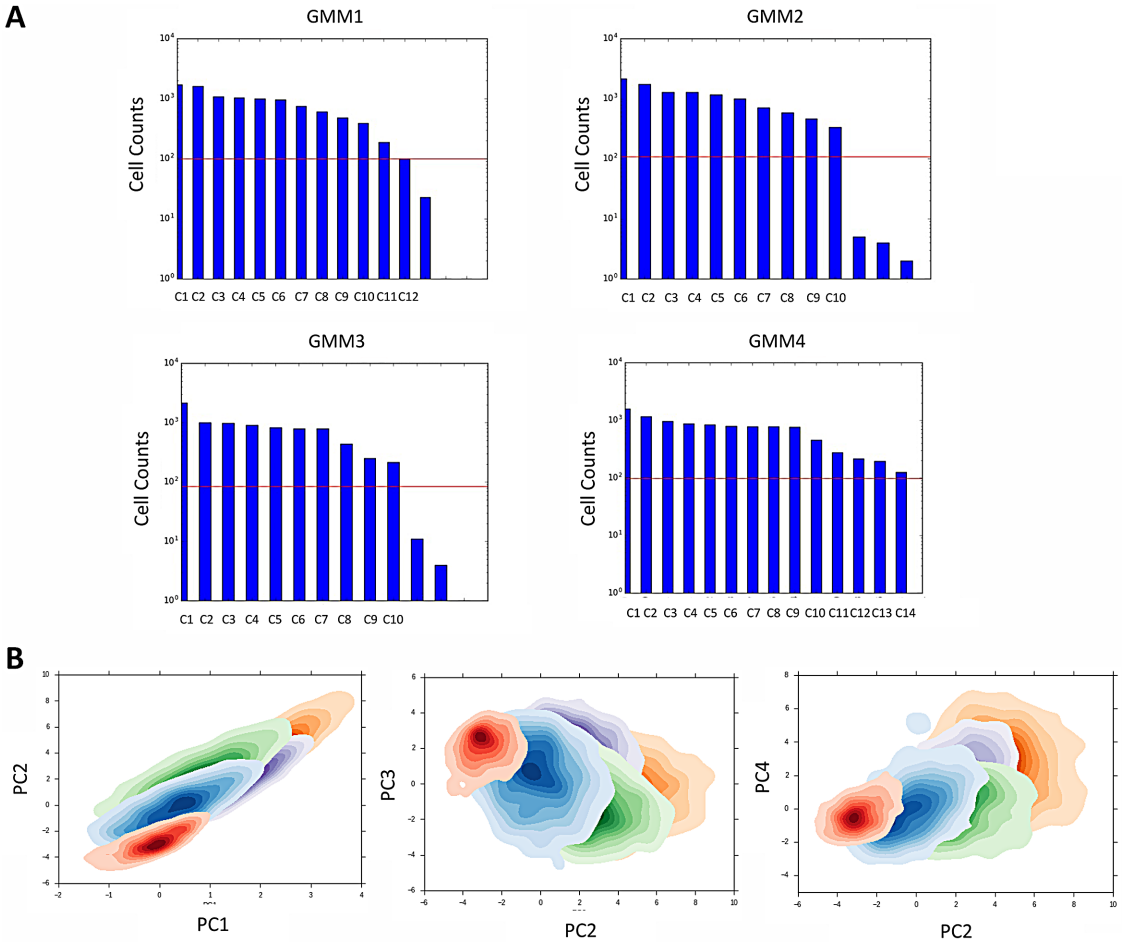

FIGURE 6: Unsupervised clustering of non–control-like cells. (A) Key cluster characteristics from unsupervised clustering performed on non–control-like cells. Bar charts depict cluster key output from GMM1 to GMM4. Cluster ID is indicated on horizontal axis; cell counts are indicated on vertical axis. Red dotted line represents threshold at 100 cells/cluster for significant cluster size. (B) Major clusters of GMM2 represented as density maps across several major components of PCA. C1, C2, C3, C4, and C5 are represented.

reproducibly treatments associate with specific clusters. As seen for replicate wells D12, D22, and K22 for STX1A, a high level of agreement was apparent, with similar fractions of cells obtained in clusters C1, C2, C3, and C6 in GMM1 (Figure 8A, Top). Therefore, our unsupervised workflow yields an almost identical phenotypic signature in wells processed completely independently and containing cells with the same perturbation. This strongly argues that the pipeline is defining a Golgi phenotypic signature unique for each treatment. Then we compared signatures between technical replicates. SXT1A signature generated independently from a technical replicate with GMM2 produced a large fraction of non–control-like cells in clusters C1, C2, and C5 (Figure 8B, Top). The number of clusters and the order changed between GMM1 and GMM2, but C1, C2, and C5 of GMM2 appeared highly related morphologically to C1, C2, C3, and C6 of GMM1 (Figure 7). This observation suggests that independent clustering performed on highly similar data sets returns very similar results.

Similarly, a biological replicate with GMM3 produced for STX1A a major fraction of non–control-like cells in clusters C1, C2, and C4 (Figure 8C, Top). The number of clusters and the order changed again from GMM1 or GMM2, but the morphology types in clusters C1, C2, and C4 from GMM3 were closely related to C1, C2, C3, and C6 in GMM1 or C1, C2, and C5 in GMM2 (Figure 7). Therefore, the results of clustering appear reproducible, with slight variations, in completely independent experiments. This reproducibility was also apparent in the similarity of signature profiles between different siRNA. For example, GMM1, GMM2, and GMM3 runs produced USE1 signatures closely related to STX1A, with a majority of non–control-like cells shared in similar clusters (Figure 8, A–C, Middle). This similarity is further explored below through the quantification of Hellinger distances. Overall, these few examples illustrate how independent clustering runs yield largely similar results.

COG4 depletion yields a unique phenotypic signature

Some treatment signatures also appeared to be very distinctive. For instance, the COG4 signature was composed of two major clusters in all the GMMs: C7/C9, C3/C8, C6/C8, and C8/C10 for GMM1–GMM4, respectively (Figure 8, A–C, for GMM1–GMM3). Visually, these Golgi morphologies presented a unique aspect: a perinuclear highly intense fragmented morphology for GMM1-C9, GMM2-C8, and GMM-C8 and a medium-intensity fragmented morphology for GMM1-C7, GMM2-C3, and GMM-C6. The aforementioned clusters C9, C8, and C8 (C10 for GMM4) were almost exclusively associated with COG4 siRNA (Figure 7 and Supplemental Tables 3 and 4). Therefore, the pipeline produced a unique phenotypic signature for COG4, a result that would have been difficult to obtain or predict

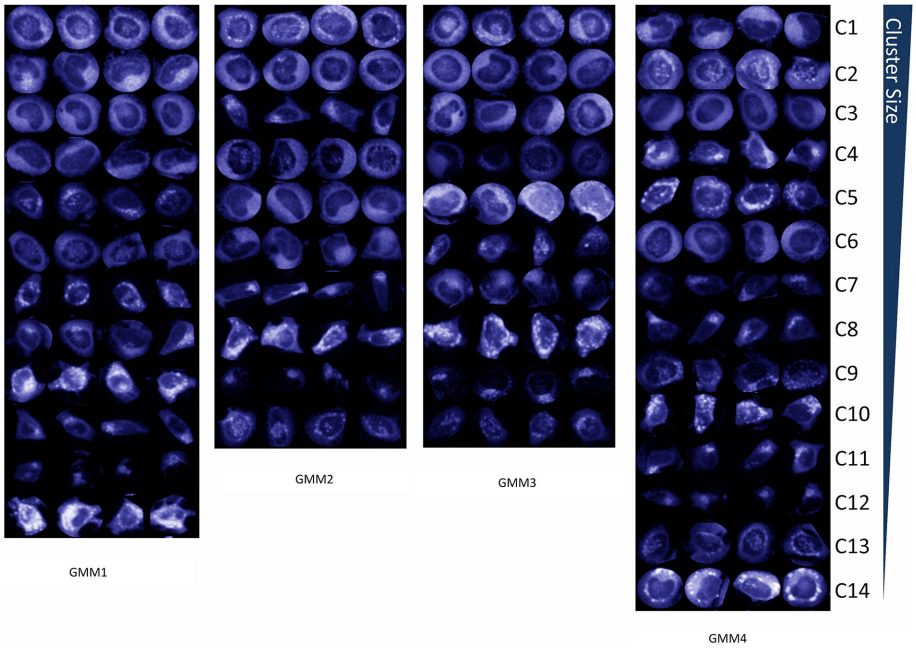

FIGURE 7: Representative phenotypic clusters for HPL Golgi stain. From left to right, output from GMM1–GMM4. Four representative non–control-like cells are shown for each cluster group. Clusters are oriented top to bottom in decreasing size order as in Figure 6. Each cell vignette is originally generated by HCSU interface from initial input of 20× Opera Phenix–acquired images with HPL Alexa647 fluorescent dye.

with a supervised approach, in the absence of specific labels for this morphology.

Comparison of phenotypic signatures using Hellinger distances

To compare cluster-based phenotypic signatures, we calculated the Hellinger distance for each siRNA treatment pair (Vajda, 1989). Distances closer to 0 reflect similarity while distances closer to 1 reflect dissimilar treatments (Figure 8). All measured Hellinger distances between treatments were compared between technical replicates (Figure 9A). The tight correlation ($R > 0.9$) indicates that the phenotypic similarities thus computed are highly reproducible between independent clustering analyses. Interestingly, the correlation between biological replicates was not much lower ($R = 0.89$), suggesting that the method is relatively robust to experimental noise (Figure 9B). Overall, the definition of phenotypic similarity appears to be highly reproducible, despite the variation in cluster numbers with different GMM modeling.

A phenotypic network of SNARE and COG subunits

We next constructed a hive plot where the nodes were genes with a significant level of penetrance. We plotted edges corresponding to Hellinger distance <0.2, which accounts for the top ~10% of distances measured (Supplemental Figure 8). An overlay of networks derived from the four GMM runs illustrates the variable degree of reproducibility in Hellinger scores (Figure 10 and Supplemental Figure 9).

We also compared the phenotypic network with a network derived from STRING predictions, adjusted for a high confidence at 0.7, based on experimental evidence (Figure 10). A network composed of STX18, STX5, GOSR2, USE1, and USO1 appeared and they were referenced as interacting in the STRING database. Interestingly, STX18, GOSR2, USE1, USO1, and STX5 are common established players in retrograde and anterograde trafficking between ER and Golgi apparatus (Xu *et al.*, 2000; Shorter *et al.*, 2002; Dilcher *et al.*, 2003).

The COG complex has been well described as a regulator of Golgi organization and is generally described as composed of two lobes, A and B. Surprisingly, we mostly recovered interactions between members of the lobe A, COG 1, 2, and 3. As previously reported, COG4 depletion produced a unique signature, while COG6 was more linked to STX7, with mostly condensed morphology–type clusters. These surprisingly different signatures suggest that depletion of these proteins has different effects on the Golgi apparatus. By extension, it is possible that different proteins of the COG complex could have additional, different effects in addition to their proposed role in the COG complex.

GOSR2 and USE1 are not reported to interact in the STRING database, but they clearly generated very similar phenotypic signatures with many diffuse cluster morphologies (Supplemental Table 3). Interestingly, a recent yeast study reported that GOSR2 and USE1 have a shared critical role during haploid nuclei fusion during yeast mating (Rogers *et al.*, 2013). The diffuse morphology of the clusters suggests a similar relocation of GalNAc-Ts from the Golgi to the ER upon GOSR2 or USE1 siRNA treatments (Dinter and Berger, 1998; Gill *et al.*, 2013).

Signature comparison between STX1A and STX5 or between STX1A and GOSR2/USE1/USO1 indicated a high level of similarity, with these signatures sharing a high fractional content of diffuse clusters also. Homologues of STX1A in *Drosophila* have been shown recently to associate with USE1, STX5, and GOSR2 in a mass spectrometry affinity approach (Guruharsha *et al.*, 2011). BNIP1 was also consistently connected to GOSR2/STX1A/STX5 from multiple GMM outputs. In line with our findings, Nakajima et al showed association between USE1 with BNIP1 and also a weak association with STX5. Furthermore, Drosophila and yeast studies also support these associations (Nakajima *et al.*, 2004; Guruharsha *et al.*, 2011; Rogers *et al.*, 2013).

DISCUSSION

Understanding the mammalian Golgi apparatus is a major scientific challenge. The range of physiological functions dependent on this organelle keeps expanding (Makowski *et al.*, 2017). Yet the debate about how cargo proteins flow through this organelle while resident proteins are retained remains ongoing after decades of publications (Farquhar, 1985; Pfeffer, 2010; Papanikou and Glick, 2014). Furthermore, how this organelle regulates the complex protein modifications that take place in its heart is still mostly unknown (Stanley, 2011; Chia *et al.*, 2012).

To understand the mammalian Golgi apparatus, the molecular machinery supporting its complex structural organization needs to be deciphered (Lowe, 2011; Stanley, 2011). It is clear that an elaborate genetic network coordinates the Golgi structural organization. More than 2000 proteins are thought to be present in this organelle (Makowski *et al.*, 2017). A set of large peripheral proteins and

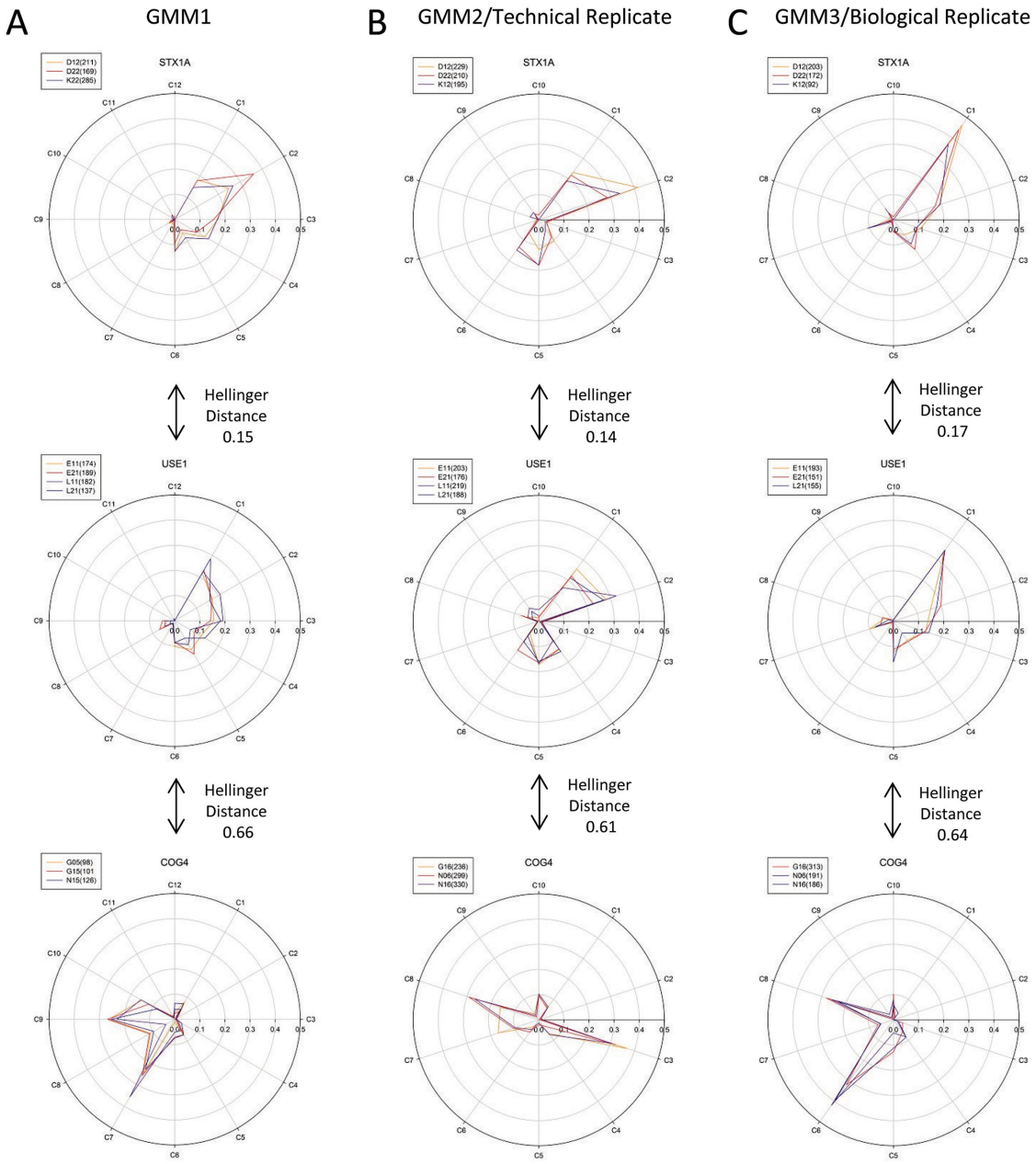

FIGURE 8: Phenotypic signature. Cluster signature composition in polar plot format for representative examples in different replicates. (A) USE1, STX1A, and COG4 siRNA treatments in GMM1 (Biological Replicate 1, Technical Replicate 1), (B) GMM2 (Biological Replicate 1, Technical Replicate 2), and (C) GMM3 (Biological Replicate 2, Technical Replicate 1). Clusters are oriented in a clockwise manner in decreasing order of size as presented in Figures 6 and 7. Radial axis indicates fraction of total non–control-like cells. Each color-coded plot corresponds to one replicate well. A replicate well reference is indicated in the top left box of each graph with total non–control-like cells number in parentheses. Hellinger distance measuring similarity of signatures is indicated for adjacent signatures.

multisubunit tethering complexes operate together with GTPase networks. Finally, a complex signaling network is coordinating and regulating this large transport machinery (Bard and Chia, 2016; Luini and Parashuraman, 2016).

Systematic approaches such as RNAi screening can help to approach this high degree of complexity. For instance, we and others previously described 160 new nodes in the form of kinase and related genes (Chia *et al.*, 2012, 2014; Galea *et al.*, 2015). However, the classification of these players and other known Golgi regulators into functional modules presents a considerable challenge.

Genetic approaches have long relied on the inference that phenotypic similarities indicate close functional relationships (Fuchs *et al.*, 2010). In the case of intracellular structures, however, visual phenotypic characterization can be limiting and challenging. Here, we presented a workflow to detect and analyze Golgi phenotypes independent of visual input using high-throughput imaging screens. The workflow presented includes several quality control steps and a module defined for exclusion of the control-like morphologies population. We showed that this workflow is reproducible and has integrated several quality control steps to handle issues such as outlier wells or irreproducible features. The main output of this workflow is

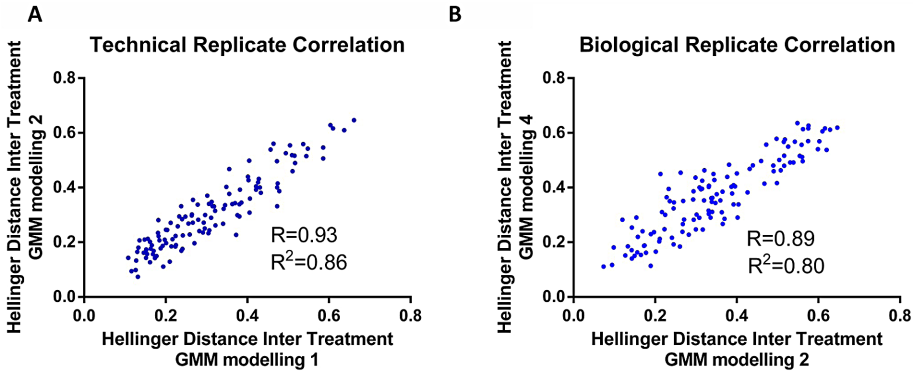

FIGURE 9: Reproducibility analysis of Hellinger distance measured between siRNA phenotypic signatures for HPL Golgi stain. (A) Treatment pair Hellinger distances from technical replicates. (B) Treatment pair Hellinger distances from biological replicates. A well-to-well reproducibility factor was set at 0.3 for all data set comparisons (Supplemental Method). Pearson correlation coefficients R and R^2 are indicated.

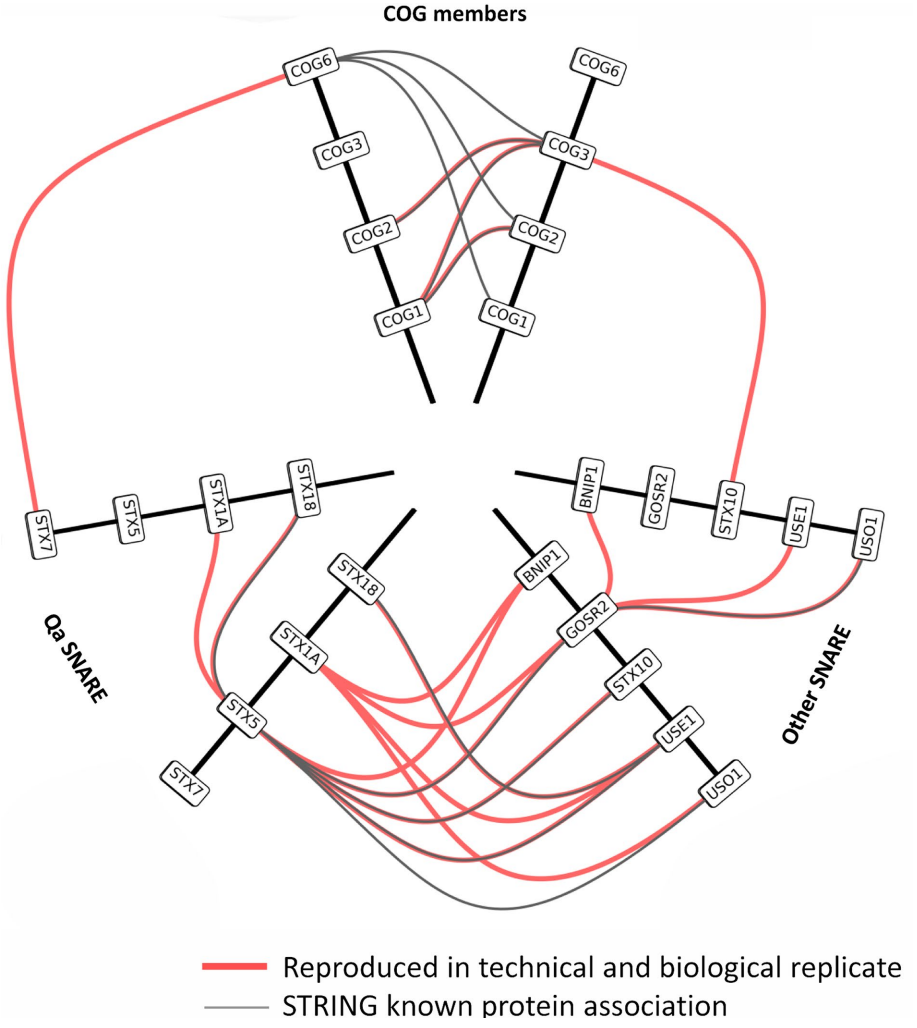

FIGURE 10: Phenotypic network: hive network plot analysis showing predicted phenotypic association in red. Each association is reproduced at least in a technical and biological replicate on the basis of Hellinger distance <0.2 for indicated paired association. A string network prediction is presented in gray (based on experimental evidence and a 0.7 threshold). A well-to-well reproducibility factor was set at 0.3 for all our Hellinger distance calculations (Supplemental Method).

the penetrance scores and the signatures defined by phenotypic clusters.

At this stage, it is not clear whether the penetrance constitutes part of a gene signature or if it is mostly a reflection of knockdown efficiency. The continuous distribution of phenotypic intensity in large-scale studies has been well documented, suggesting that penetrance reflects, at least partly, a biological parameter (Friedman and Perrimon, 2007). Consistent with this notion, the penetrance of phenotypically related perturbations such as STX18, GOSR2, USE1, USO1, and STX5 was comparable. COG1, COG2, and COG3 subunits, all implicated in transport within the Golgi, also shared similar penetrance levels.

At any rate, phenotypes defined by clusters are sufficient to obtain highly informative signatures. The strong association of clusters with specific treatments suggests that these disturbed Golgi phenotypes are not disorganized structures but rather alternative metastable states. It is currently unclear how many states Golgi can have, and it is likely that with more siRNA treatments, we would find more than 10 phenotypic clusters. The fact that different clusters are reproducibly occurring in a single condition suggests either that the Golgi can evolve along two (or more) phenotypic paths or that cells oscillate between these cooccurring states over time.

One important aspect of our workflow is the use of a single marker, in our case HPL. This lectin reveals the activity of GalNAc-Ts, which are typically localized in the Golgi apparatus. However, we have shown previously that GalNAc-Ts can traffic to the ER in an inducible manner independent of other Golgi enzymes (Gill *et al.*, 2010). The genetic network controlling the localization of GalNAc-Ts is therefore likely to be different from that for other enzymes in the Golgi. Using a different marker, for example a different enzyme, could therefore reveal a different network.

Our method would make it possible to generate different networks relatively easily, either with different markers or under different physiological conditions, which could be a significant advantage of these phenotypic networks. Indeed, as highlighted recently, different networks are highly informative for understanding biological responses (Ideker and Krogan, 2012). To be most effective, these approaches might need to be conducted at the genomewide level or at least at a large scale. This up-scaling might require some technical improvements for handling multiple plates with a larger multidimensional space.

Overall, we expect that this context-dependent, function-based approach to the discovery of genetic modules and networks will greatly enhance our capacity to obtain a system-level understanding of the regulation of Golgi organization and could be applied to many other systems within the cell.

MATERIALS AND METHODS

Cell line antibodies and reagents

HeLa cells originated from V. Malhotra (Centre for Genomic Regulation, Barcelona, Spain). HeLa cells were grown with high-glucose DMEM supplemented with 10% fetal bovine serum (FBS) at 37°C in a 10% CO_2 humidified incubator. HeLa cells from the same passage (number 23 in our lab) were exclusively used for each biological replicate tested in this study. *H. pomatia* lectin A (HPL) conjugated with 647 nm fluorophore (#L32454) and Hoechst were obtained from Invitrogen/Life technologies. On target plus siRNA pools were obtained from Dharmacon. Optimem was purchased from Invitrogen, and Hiperfect transfection reagents were from Qiagen (#301705).

siRNA transfection and imaging

A quantity of 2.5 µl of 500 nM siRNA was printed into 384 CellCarrier-Ultra Microplates (#6057308, Perkin Elmer-Cetus) with velocity 11. Reverse siRNA transfection used a defined well mixture of 0.25 µl of Hiperfect mixed with 7.25 µl of Optimem for 5 min, which was added subsequently to siRNA for complexation for 20 min. Subsequently, 40 µl of cells was added, with a content of 1000 cells/well. After 3 d of siRNA knockdown, fixing of cells was performed with 4% paraformaldehyde in Dulbecco's phosphate-buffered saline (D-PBS) for 10 min. Cells were then washed with D-PBS at pH 7.2, followed by permeabilization for 10 min with 0.2% Triton X-100. Cell staining was then performed in 2% FBS in D-PBS at pH 7.2 with HPL conjugated to Alexa Fluor 647 and Hoechst diluted in 2% FBS in PBS at pH 7.2 for 20 min on a 1 cm–span orbital shaker at 150 rpm. The plate was then washed three times with 30 µl/well D-PBS at pH 7.2 before being scanned in a high-throughput confocal imager. A multidrop combi with a small cassette was used for addition of Hiperfect mixture and cells in a 384-well plate. A standard cassette was used for fixing and washing of cells (Thermo Fisher).

Image acquisition and single-cell HCSU processing

Eight fields per well on one plan were acquired sequentially with an Opera Phenix content imager configured with CMOS cameras and a 20× NA 1.0 water objective (Perkin Elmer). Sequential measurement was performed with the pair excitation wavelength for 100 ms with Hoechst followed by Alexa647. The image data set was then used by a high-content screening unit (HCSU) to perform single-cell extraction and feature calculation (Tjhi *et al.*, 2011; Chia *et al.*, 2012). A 20× source images data set from Phenix Opera and a HCSU automatic features extraction program are available at http://dx.doi.org/10.17632/vk4yhs8h6s.1.

ACKNOWLEDGMENTS

We thank Victor Racine, William Tjhi, Emilie Chapeau, Maciek Hermanowicz, and Maja Choma for participating in the development of this work.

REFERENCES

Anitei M, Chenna R, Czupalla C, Esner M, Christ S, Lenhard S, Korn K, Meyenhofer F, Bickle M, Zerial M, Hoflack B (2014). A high-throughput siRNA screen identifies genes that regulate mannose 6-phosphate receptor trafficking. J Cell Sci 127, 5079–5092.

Bamford SP, Nichol RC, Baldry IK, Land K, Lintott CJ, Schawinski K, Slosar AE, Szalay AS, Thomas D, Torki M, *et al.* (2009). Galaxy Zoo: the dependence of morphology and colour on environment? Mon Not R Astron Soc 393, 1324–1352.

Bard F, Chia J (2016). Cracking the glycome encoder: signaling, trafficking, and glycosylation. Trends Cell Biol 26, 379–388.

Bishop CM (2006). Pattern recognition. Mach Learn 128, 1–58.

Breiman L (2001). Random forests. Mach Learn 45, 5–32.

Chia J, Goh G, Racine V, Ng S, Kumar P, Bard F (2012). RNAi screening reveals a large signaling network controlling the Golgi apparatus in human cells. Mol Syst Biol 8, 629.

Chia J, Tham KM, Gill DJ, Bard-Chapeau EA, Bard FA (2014). ERK8 is a negative regulator of O-GalNAc glycosylation and cell migration. Elife 3, e01828.

Conrad C, Gerlich DW (2010). Automated microscopy for high-content RNAi screening. J Cell Biol 188, 453–461.

de Graffenried CL, Bertozzi CR (2004). The roles of enzyme localisation and complex formation in glycan assembly within the Golgi apparatus. Curr Opin Cell Biol 16, 356–363.

Dilcher M, Veith B, Chidambaram S, Hartmann E, Schmitt HD, Fischer von Mollard G (2003). Use1p is a yeast SNARE protein required for retrograde traffic to the ER. EMBO J 22, 3664–3674.

Dinter A, Berger EG (1998). Golgi-disturbing agents. Histochem Cell Biol 109, 571–590.

Farhan H, Wendeler MW, Mitrovic S, Fava E, Silberberg Y, Sharan R, Zerial M, Hauri HP (2010). MAPK signaling to the early secretory pathway revealed by kinase/phosphatase functional screening. J Cell Biol 189, 997–1011.

Farquhar MG (1985). Progress in unraveling pathways of Golgi traffic. Annu Rev Cell Biol 1, 447–488.

Friedman A, Perrimon N (2007). Genetic screening for signal transduction in the era of network biology. Cell 128, 225–231.

Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, Steinbrink S, Horn T, Pedal A, Huber W, Boutros M (2010). Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol 6, 370.

Galea G, Bexiga MG, Panarella A, O'Neill ED, Simpson JC (2015). A high-content screening microscopy approach to dissect the role of Rab proteins in Golgi-to-ER retrograde trafficking. J Cell Sci 128, 2339–2349.

Gill DJ, Chia J, Senewiratne J, Bard F (2010). Regulation of O-glycosylation through Golgi-to-ER relocation of initiation enzymes. J Cell Biol 189, 843–858.

Gill DJ, Tham KM, Chia J, Wang SC, Steentoft C, Clausen H, Bard-Chapeau EA, Bard FA (2013). Initiation of GalNAc-type O-glycosylation in the endoplasmic reticulum promotes cancer cell invasiveness. Proc Natl Acad Sci USA 110, E3152–E3161.

Goh GY, Bard FA (2015). RNAi screens for genes involved in Golgi glycosylation. Methods Mol Biol 1270, 411–426.

Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, *et al.* (2011). A protein complex network of *Drosophila melanogaster*. Cell 147, 690–703.

Hong W (2005). SNAREs and traffic. Biochim Biophys Acta 1744, 120–144.

Hong W, Lev S (2014). Tethering the assembly of SNARE complexes. Trends Cell Biol 24, 35–43.

Ideker T, Krogan NJ (2012). Differential network biology. Mol Syst Biol 8, 565.

Lindblad J, Wahlby C, Bengtsson E, Zaltsman A (2004). Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation. Cytometry A 57, 22–33.

Loo LH, Lin HJ, Singh DK, Lyons KM, Altschuler SJ, Wu LF (2009). Heterogeneity in the physiological states and pharmacological responses of differentiating 3T3-L1 preadipocytes. J Cell Biol 187, 375–384.

Lowe M (2011). Structural organization of the Golgi apparatus. Curr Opin Cell Biol 23, 85–93.

Luini A, Parashuraman S (2016). Signaling at the Golgi: sensing and controlling the membrane fluxes. Curr Opin Cell Biol 39, 37–42.

Makowski SL, Tran TTT, Field SJ (2017). Emerging themes of regulation at the Golgi. Curr Opin Cell Biol 45, 17–23.

Malsam J, Sollner TH (2011). Organization of SNAREs within the Golgi stack. Cold Spring Harb Perspect Biol 3, a005249.

Millarte V, Boncompain G, Tillmann K, Perez F, Sztul E, Farhan H (2015). Phospholipase C gamma1 regulates early secretory trafficking and cell migration via interaction with p115. Mol Biol Cell 26, 2263–2278.

Mukhamedyarov MA, Rizvanov AA, Yakupov EZ, Zefirov AL, Kiyasov AP, Reis HJ, Teixeira AL, Vieira LB, Lima LM, Salafutdinov II, *et al.* (2016).

Transcriptional analysis of blood lymphocytes and skin fibroblasts, keratinocytes, and endothelial cells as a potential biomarker for alzheimer's disease. J Alzheimers Dis 54, 1373–1383.
Murphy RF, Velliste M, Porreca G (2003). Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. J VLSI Sig Proc Syst 35, 311–321.
Nakajima K, Hirose H, Taniguchi M, Kurashina H, Arasaki K, Nagahama M, Tani K, Yamamoto A, Tagaya M (2004). Involvement of BNIP1 in apoptosis and endoplasmic reticulum membrane fusion. EMBO J 23, 3216–3226.
Papanikou E, Glick BS (2014). Golgi compartmentation and identity. Curr Opin Cell Biol 29, 74–81.
Pelkmans L (2012). Using cell-to-cell variability—a new era in molecular biology. Science 336, 425–426.
Pfeffer SR (2010). How the Golgi works: a cisternal progenitor model. Proc Natl Acad Sci USA 107, 19614–19618.
Rogers JV, Arlow T, Inkellis ER, Koo TS, Rose MD (2013). ER-associated SNAREs and Sey1p mediate nuclear fusion at two distinct steps during yeast mating. Mol Biol Cell 24, 3896–3908.
Shorter J, Beard MB, Seemann J, Dirac-Svejstrup AB, Warren G (2002). Sequential tethering of Golgins and catalysis of SNAREpin assembly by the vesicle-tethering protein p115. J Cell Biol 157, 45–62.
Singh DK, Ku CJ, Wichaidit C, Steininger RJ 3rd, Wu LF, Altschuler SJ (2010). Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. Mol Syst Biol 6, 369.
Slack MD, Martinez ED, Wu LF, Altschuler SJ (2008). Characterizing heterogeneous cellular responses to perturbations. Proc Natl Acad Sci USA 105, 19306–19311.
Sommer C, Gerlich DW (2013). Machine learning in cell biology—teaching computers to recognize phenotypes. J Cell Sci 126, 5529–5539.
Stanley P (2011). Golgi glycosylation. Cold Spring Harb Perspect Biol 3, a005199.
Tjhi WC, Lee KK, Hung T, Tsang IW, Ong YS, Bard F, Racine V (2011). Exploratory analysis of cell-based screening data for phenotype identification in drug-siRNA study. Int J Comput Biol Drug Des 4, 194–215.
Vajda I (1989). Theory of Statistical Inference and Information, Dordrecht, Netherlands: Kluwer Academic.
Willett R, Ungar D, Lupashin V (2013). The Golgi puppet master: COG complex at center stage of membrane trafficking interactions. Histochem Cell Biol 140, 271–283.
Xu D, Joglekar AP, Williams AL, Hay JC (2000). Subunit structure of a mammalian ER/Golgi SNARE complex. J Biol Chem 275, 39631–39639.
Yadav S, Puri S, Linstedt AD (2009). A primary role for Golgi positioning in directed secretion, cell polarity, and wound healing. Mol Biol Cell 20, 1728–1736.
Zhang C, Sun C, Su R, Pham TD (2015). Clustered nuclei splitting via curvature information and gray-scale distance transform. J Microsc 259, 36–52.
Zhong Q, Busetto AG, Fededa JP, Buhmann JM, Gerlich DW (2012). Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. Nat Methods 9, 711–713.