Supplementary Figure Legends

**Figure 1:** Plate map with siRNA treatments used

**Figure 2:** siRNA experimental quality control. (A) Efficiency of transfection of Polo Like Kinase siRNA is assessed over all 4 replicates used in this study with Hoechst Nuclei counts in Harmony software (Perkin Elmer). n refer to well amounts used for making the estimation. (B) 20X representative pictures of HPL/Hoechst staining for key morphological siRNA controls STX5 producing diffuse HPL stain, An inset with Representative single Cell is shown for each representative picture. (C) Mean Nuclei counts present in original dataset for all replicates used in this study

**Figure 3:** Number of nuclei retained after Nuclei Quality Control. All replicates used in this study are presented with mean and standard deviation as a stacked bar

**Figure 4:** Influence of Number of Controls wells on % non-control-like cells output. SVM control modelling was performed with indicated number of control wells. One control well is equivalent to ~800 cells. NT (negative siRNA) and STX5 (positive control) are used as indicator of % non-control-like cells classified by SVM in the control modelling step. Data depicts mean and standard deviation based on 10 random selections of control wells for 2 biological replicates. Variability of % non-control-like cells was more noticeable with a selection of less than 10 wells (~8000 cells).

**Figure 5:** Deconvolution analysis of siRNA pools. For each indicated siRNA treatment, each set of four duplexes was ranked based on the number of duplexes that have a penetrance score above the 10% threshold used in our control modelling analysis. A gene was validated if at least two unique siRNAs gave a penetrance score above the 10% threshold. Penetrance score for individual siRNA is indicated.

**Figure 6:** Dendrograms showing GMM cluster relationships. Means of GMM components for GMM1 (a) and GMM2 (b) were used to perform hierarchical clustering. The resulting dendrograms show that in both the replicates, 1-2 clusters (C9, C12 in GMM1 and C8 in GMM2) are very different from others. Y-axis shows the Euclidean distance between cluster means. Four sample images from each cluster are shown below the dendrograms.

**Figure 7:** Feature importance for the clustering structure generated by GMM. Random Forest was used to learn to classify the Golgi feature data according to the pre-generated GMM clusters. Bar plot shows the importance scores for the 70 commonly selected good quality features as computed by the random forest classifier.

**Figure 8:** Histogram distribution of Hellinger Distance for all replicates. Calculated Hellinger distances distribution between treatment pair for all replicates is presented in individual histogram. An arbitrary cutoff indicated with a red line was set at 0.2 to select treatment pair showing high phenotypic similarity. A well-to-well reproducibility factor was set at 0.3 for all datasets comparison (Supplementary methods).

**Figure 9:** Hive Network plot analysis showing overlay of phenotypic network analysis for all replicates (GMM modelling 1-4) in blue. Thickness and color of the blue lines from same edge and node reflect the extent of reproduced predicted links from GMM replicates. String Network Prediction is presented in red (based on experimental evidence and a 0.7 threshold). Treatment pairs with Hellinger distance <0.2 are used to draw the network. A well-to-well reproducibility factor was set at 0.3 for all datasets comparison (Supplementary methods).
Row 2 and 23 are specifically used for siRNA control wells: Polo Like Kinase (PLK), Syntaxin-5 (STX5) and non target control (NT) are used as controls. 40 SNARE and 8 COG siRNA targets were tested with 4 replicates per plate. Mock well are subjected to transfection reagents only and are used in our control modelling step.
Supplementary Figure 2

Nuclei counts derived from HCSU pipeline. Each stacked bar depicts nuclei counts in 2 biological replicates and 2 technical replicates. Each replicate is derived from the mean of 4 wells. Error Bars depict standard deviation for each replicate.
Each stacked bar depicts nuclei counts in 2 biological replicates and 2 technical replicates. Each replicate is derived from the mean of 4 wells. Error Bars depict standard deviation for each replicate.
SVM control modelling output indicator with increasing number of mock control wells used for defining control space boundaries. Y Axis indicate penetrance with fraction of non control cells from total amount of cells. One well is approximately equivalent to 800 control cells.
## Deconvolution analysis of ON-TARGETplus siRNA.

4 Single siRNA are evaluated for their individual penetrance. A penetrance $>$10% in control modelling validates the particular siRNA. siRNA pool were considered validated when a minimum of 2 single siRNA displayed a penetrance $>$10% in control modelling.

### Table: Penetration Rates

| Gene   | Column 1 | Column 2 | Column 3 | Column 4 |
|--------|----------|----------|----------|----------|
| USE1   | 29       | 12       | 30       | 14       |
| STX1B  | 25       | 48       | 17       | 11       |
| STX5   | 15       | 70       | 17       | 29       |
| STX1A  | 16       | 17       | 12       | 36       |
| SEC2B  | 13       | 23       | 41       | 60       |
| USO1   | 42       | 45       | 20       | 42       |
| BNIP1  | 19       | 15       | 12       | 0        |
| YKT6   | 14       | 16       | 19       | 0        |
| STX3   | 21       | 16       | 17       | 0        |
| COG6   | 30       | 10       | 11       | 0        |
| COG3   | 15       | 12       | 30       | 0        |
| STX19  | 21       | 10       | 14       | 0        |
| BET1L  | 19       | 12       | 17       | 0        |
| COG1   | 15       | 13       | 52       | 0        |
| GOSR2  | 17       | 26       | 23       | 0        |
| COG8   | 15       | 13       | 26       | 0        |
| COG2   | 12       | 16       | 14       | 0        |
| COG5   | 11       | 13       | 10       | 0        |
| VAMP7  | 18       | 11       | 0        | 0        |
| VAMP4  | 15       | 12       | 0        | 0        |
| COG4   | 45       | 78       | 0        | 0        |
| STX7   | 40       | 11       | 0        | 0        |
| VAMP5  | 19       | 11       | 0        | 0        |
| VAMP3  | 15       | 14       | 0        | 0        |
| COG7   | 12       | 17       | 0        | 0        |
| BET1   | 10       | 12       | 0        | 0        |
| SEC22A | 16       | 11       | 0        | 0        |
| VTI1B  | 10       | 12       | 0        | 0        |
| STX2   | 10       | 14       | 0        | 0        |
| STX1B  | 17       | 10       | 0        | 0        |
| SNAP29 | 29       | 13       | 0        | 0        |
| STX17  | 10       | 16       | 0        | 0        |
| STX10  | 13       | 19       | 0        | 0        |
| SNAP23 | 21       | 14       | 0        | 0        |
| GOSR1  | 30       | 12       | 0        | 0        |
| VTI1A  | 15       | 18       | 0        | 0        |
| STX6   | 18       | 16       | 0        | 0        |
| VAMP1  | 11       | 17       | 0        | 0        |
| SEC22C | 33       | 0        | 0        | 0        |
| STX8   | 10       | 0        | 0        | 0        |
| VAMP2  | 12       | 0        | 0        | 0        |
| STX16  | 11       | 0        | 0        | 0        |
| STX11  | 16       | 0        | 0        | 0        |
| STX4   | 24       | 0        | 0        | 0        |
| STX12  | 15       | 0        | 0        | 0        |
| SNAP47 | 10       | 0        | 0        | 0        |
| SNAP25 | 0        | 0        | 0        | 0        |
| VAMP8  | 0        | 0        | 0        | 0        |
| NT     | 0        | 0        | 0        | 0        |

Supplementary Figure 5
a) GMM1

b) GMM2

Supplementary Figure 6
Supplementary Figure 8

Histograms showing the distribution of Hellinger Distance for GMM1, GMM2, GMM3, and GMM4.
|                | Z'  | n (Mock) | n (STX5) | Mean(STX5) | SD(STX5) | Mean(Mock) | SD(Mock) |
|----------------|-----|----------|----------|------------|----------|------------|----------|
| Biological Replicate 1/Technical Replicate 1 | 0.4 | 48       | 4        | 276.1      | 11.1     | 179.7      | 7.6      |
| Biological Replicate 1/Technical Replicate 2 | 0.3 | 48       | 4        | 246.5      | 8.9      | 181.0      | 6.7      |
| Biological Replicate 2/Technical Replicate 1 | 0.42| 48       | 4        | 252.8      | 8.0      | 185.1      | 5.0      |
| Biological Replicate 2/Technical Replicate 2 | 0.32| 48       | 4        | 261.5      | 13.8     | 180.9      | 4.4      |

Supplementary Table 1: Z' analysis based on median cell Intensity HPL measurements
### List of Golgi Features:

| Feature | Description |
|---------|-------------|
| ch[1].Obj[1].nbSubstructures | |
| ch[1].Obj[1].angle | |
| ch[1].Obj[1].sigma.y | |
| ch[1].Obj[1].distFromObjectCenterToSegCenter | |
| ch[1].Obj[1].peri | |
| ch[1].Obj[1].valMorpho | |
| ch[1].Obj[1].pctIntensity | |
| ch[1].Int.aTrousDir[7].angle | |
| ch[1].Int.aTrousDir[7].max | |
| ch[1].Int.aTrousDir[7].mean | |
| ch[1].Int.aTrousDir[6].max | |
| ch[1].Int.aTrousDir[6].mean | |
| ch[1].Int.aTrousDir[5].mean | |
| ch[1].Int.aTrousDir[4].angle | |
| ch[1].Int.aTrousDir[4].mean | |
| ch[1].Int.aTrousDir[3].angle | |
| ch[1].Int.aTrousDir[3].max | |
| ch[1].Int.aTrousDir[3].mean | |
| ch[1].Int.aTrousDir[2].angle | |
| ch[1].Int.aTrousDir[2].mean | |
| ch[1].Int.aTrousDir[1].angle | |
| ch[1].Int.aTrousDir[1].max | |
| ch[1].Int.aTrousDir[1].mean | |
| ch[1].Int.aTrousDir[0].max | |
| ch[1].Int.aTrousDir[0].mean | |
| ch[1].Int.aTrous[7].abs | |
| ch[1].Int.aTrous[6].stdev | |
| ch[1].Int.aTrous[6].mean | |
| ch[1].Int.aTrous[5].stdev | |
| ch[1].Int.aTrous[5].mean | |
| ch[1].Int.aTrous[2].mean | |
| ch[1].Int.aTrous[1].stdev | |
| ch[1].Int.aTrous[1].abs | |
| ch[1].Int.aTrous[1].mean | |
| ch[1].Int.aTrous[5].mean | |
| ch[1].Int.aTrous[4].stdev | |
| ch[1].Int.aTrous[4].mean | |
| ch[1].Int.aTrous[3].stdev | |
| ch[1].Int.aTrous[3].abs | |
| ch[1].Int.aTrous[2].stdev | |
| ch[1].Int.aTrous[2].abs | |
| ch[1].Int.aTrous[0].stdev | |
| ch[1].Int.aTrous[0].abs | |
| ch[1].Int.aTrous[0].mean | |
| ch[1].Int.haralickCoef[12] | |
| ch[1].Int.haralickCoef[11] | |
| ch[1].Int.haralickCoef[10] | |
| ch[1].Int.haralickCoef[8] | |
| ch[1].Int.haralickCoef[6] | |
| ch[1].Int.haralickCoef[5] | |
| ch[1].Int.haralickCoef[4] | |
| ch[1].Int.haralickCoef[3] | |
| ch[1].Int.zernikeMoments[34] | |
| ch[1].Int.zernikeMoments[33] | |
| ch[1].Int.zernikeMoments[30] | |
| ch[1].Int.zernikeMoments[29] | |
| ch[1].Int.zernikeMoments[27] | |
| ch[1].Int.zernikeMoments[26] | |
| ch[1].Int.zernikeMoments[25] | |
| ch[1].Int.zernikeMoments[24] | |
| ch[1].Int.zernikeMoments[23] | |
| ch[1].Int.zernikeMoments[21] | |
| ch[1].Int.zernikeMoments[18] | |
| ch[1].Int.zernikeMoments[17] | |
| ch[1].Int.zernikeMoments[15] | |
| ch[1].Int.zernikeMoments[11] | |
| ch[1].Int.zernikeMoments[10] | |
| ch[1].Int.zernikeMoments[9] | |
| ch[1].Int.zernikeMoments[7] | |
| ch[1].Int.zernikeMoments[6] | |
| ch[1].Int.zernikeMoments[4] | |
| ch[1].Int.zernikeMoments[3] | |
| ch[1].Int.zernikeMoments[2] | |
| ch[1].Int.zernikeMoments[1] | |
| ch[1].Int.stdev | |
| ch[1].Int.totalIntensity | |
| ch[1].Int.mean | |
| ch[1].Int.valMin | |
| ch[1].Int.valMax | |
| ch[1].Int.valBackground | |

### List of Nuclei Features:

| Feature | Description |
|---------|-------------|
| ch[2].Obj[1].distFromObjectCenterToSegCenter | |
| ch[2].Obj[1].distFromObjectCenterToSegCenterSurf | |
| ch[2].Obj[1].peri | |
| ch[2].Obj[1].valMorpho | |
| ch[2].Obj[1].pctIntensity | |
| ch[2].Obj[2].nbSubstructures | |

Supplementary Table 2: list of calculated HCSU features
| BET1 | GMM1 | GMM2 | GMM3 | GMM4 |
|------|------|------|------|------|
| Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate |

| BNIP1 | GMM1 | GMM2 | GMM3 | GMM4 |
|------|------|------|------|------|
| Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate |

| GOSR1 | GMM1 | GMM2 | GMM3 | GMM4 |
|------|------|------|------|------|
| Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate | Did not reach Penetrance Threshold in any well replicate |
GOSR2

Did not reach Penetrance Threshold in any well replicate

SEC22B

Did not reach Penetrance Threshold in any well replicate

SEC22C

Did not reach Penetrance Threshold in any well replicate
Did not reach Penetrance Threshold so was not selected...
VAMP5

Did not reach Penetrance Threshold in any well replicate.

VAMP8

Did not reach Penetrance Threshold in any well replicate.

COG1
COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

COG2

COG3

COG4

Did not reach Penetrance Threshold in any well replicate
Did not reach Penetrance Threshold in any well replicate.
Supplementary Table 4: Major penetrant siRNA per cluster (Treatment with Cluster >10% of non control like cells)

**GMM1**

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|---|----|----|----|----|----|----|----|----|----|----|----|----|
| BNIP1 | BNIP1 | BNIP1 | BNIP1 |    |    |    |    |    |    |    |    |    |
| COG1 | COG1 | COG1 | COG1 |    |    |    |    |    |    |    |    |    |
| COG2 | COG2 | COG2 | COG2 |    |    |    |    |    |    |    |    |    |
| COG3 | COG3 | COG3 | COG3 |    |    |    |    |    |    |    |    |    |
| COG4 | COG4 | COG4 | COG4 |    |    |    |    |    |    |    |    |    |
| COG6 | COG6 | COG6 | COG6 |    |    |    |    |    |    |    |    |    |
| GOSR2 | GOSR2 | GOSR2 | GOSR2 |    |    |    |    |    |    |    |    |    |
| SEC22B | SEC22B | SEC22B | SEC22B |    |    |    |    |    |    |    |    |    |
| STX10 | STX10 | STX10 | STX10 |    |    |    |    |    |    |    |    |    |
| STX18 | STX18 | STX18 | STX18 |    |    |    |    |    |    |    |    |    |
| STX1A | STX1A | STX1A | STX1A |    |    |    |    |    |    |    |    |    |
| STX5 | STX5 | STX5 | STX5 |    |    |    |    |    |    |    |    |    |
| STX7 | STX7 | STX7 | STX7 |    |    |    |    |    |    |    |    |    |
| USE1 | USE1 | USE1 | USE1 |    |    |    |    |    |    |    |    |    |
| USO1 | USO1 | USO1 | USO1 |    |    |    |    |    |    |    |    |    |
| VAMP8 | VAMP8 | VAMP8 | VAMP8 |    |    |    |    |    |    |    |    |    |

**GMM2**

|   | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 |
|---|----|----|----|----|----|----|----|----|----|----|
| BNIP1 | BNIP1 | BNIP1 | BNIP1 |    |    |    |    |    |    |    |
| COG1 | COG1 | COG1 | COG1 |    |    |    |    |    |    |    |
| COG2 | COG2 | COG2 | COG2 |    |    |    |    |    |    |    |
| COG3 | COG3 | COG3 | COG3 |    |    |    |    |    |    |    |
| COG4 | COG4 | COG4 | COG4 |    |    |    |    |    |    |    |
| COG6 | COG6 | COG6 | COG6 |    |    |    |    |    |    |    |
| GOSR1 | GOSR1 | GOSR1 | GOSR1 |    |    |    |    |    |    |    |
| GOSR2 | GOSR2 | GOSR2 | GOSR2 |    |    |    |    |    |    |    |
| SEC22B | SEC22B | SEC22B | SEC22B |    |    |    |    |    |    |    |
| SEC22C | SEC22C | SEC22C | SEC22C |    |    |    |    |    |    |    |
| STX7 | STX7 | STX7 | STX7 |    |    |    |    |    |    |    |
| STX10 | STX10 | STX10 | STX10 |    |    |    |    |    |    |    |
| STX18 | STX18 | STX18 | STX18 |    |    |    |    |    |    |    |
| STX1A | STX1A | STX1A | STX1A |    |    |    |    |    |    |    |
| STX5 | STX5 | STX5 | STX5 |    |    |    |    |    |    |    |
| USE1 | USE1 | USE1 | USE1 |    |    |    |    |    |    |    |
| USO1 | USO1 | USO1 | USO1 |    |    |    |    |    |    |    |
| VAMP5 | VAMP5 | VAMP5 | VAMP5 |    |    |    |    |    |    |    |
| VAMP8 | VAMP8 | VAMP8 | VAMP8 |    |    |    |    |    |    |    |
|   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|----|
| GMM3 | BET1 | BET1 | BET1 | BET1 | | | | | | |
| | BNIP1 | BNIP1 | BNIP1 | BNIP1 | BNIP1 | | | | | |
| | COG1 | COG1 | COG1 | COG1 | | | | | | |
| | COG3 | COG3 | | | | | | | | |
| | | COG4 | COG4 | COG4 | | | | | | |
| | | COG6 | COG6 | COG6 | | | | | | |
| | GOSR2 | GOSR2 | GOSR2 | GOSR2 | | | | | | |
| | SEC22B | SEC22B | SEC22B | | | | | | | |
| | STX10 | STX10 | STX10 | | | | | | | |
| | STX18 | STX18 | STX18 | STX18 | | | | | | |
| | STX1A | STX1A | STX1A | | | | | | | |
| | STX5 | STX5 | STX5 | STX5 | | | | | | |
| | USE1 | USE1 | USE1 | USE1 | USE1 | | | | | |
| | USO1 | USO1 | USO1 | USO1 | USO1 | USO1 | | | | |
| | VAMP5 | VAMP5 | VAMP5 | VAMP5 | VAMP5 | VAMP5 | VAMP5 | | | |
| GMM4 | COG1 | COG1 | COG1 | COG1 | | | | | | |
| | COG2 | COG2 | COG2 | COG2 | | | | | | |
| | COG3 | COG3 | COG3 | COG3 | | | | | | |
| | | COG4 | COG4 | COG4 | | | | | | |
| | | COG6 | COG6 | COG6 | COG6 | | | | | |
| | GOSR1 | GOSR1 | GOSR1 | GOSR1 | | | | | | |
| | GOSR2 | GOSR2 | GOSR2 | GOSR2 | | | | | | |
| | STX1A | STX1A | STX1A | STX1A | | | | | | |
| | SEC22B | SEC22B | SEC22B | | | | | | | |
| | STX5 | STX5 | STX5 | STX5 | | | | | | |
| | STX7 | STX7 | STX7 | STX7 | | | | | | |
| | STX18 | STX18 | STX18 | STX18 | STX18 | | | | | |
| | USO1 | USO1 | USO1 | USO1 | USO1 | USO1 | | | | |
| | USE1 | USE1 | USE1 | USE1 | USE1 | USE1 | USE1 | | | |
| | VAMP5 | VAMP5 | VAMP5 | VAMP5 | VAMP5 | VAMP5 | | | | |
**Supplementary Methods:**

The computational framework designed for the analysis of high-content image-based data consists of quality control and unsupervised learning modules. This analysis pipeline was developed using Python 2.7 and its associated libraries available through the Anaconda Python distribution (https://www.anaconda.com/download/) and run on a Linux server with 64-core 2.6GHz processors and 256GB memory. A package consisting of the Python script for the pipeline and an example dataset has been uploaded on Mendeley data (DOI http://dx.doi.org/10.17632/pp282j4h29.1)

https://data.mendeley.com/datasets/pp282j4h29/1

1) **Nuclei Quality Control (NQC)**

It is established that watershed algorithm can misidentify overlapping nuclei and lead to outputs with imperfectly isolated nuclei (Zhang et al, 2015). To remove these poorly segmented nuclei, we developed a nuclei quality control module, which involves the following steps:

a) **Unsupervised clustering of nuclei:** First, unsupervised clustering using Gaussian Mixture Model (GMM) was performed on all nuclei to generate different nuclei subpopulations (Bishop, 2006). Before clustering, we performed data dimensionality reduction using Principal Component Analysis (PCA), which linearly transforms the high-dimensional feature data using a set of uncorrelated variables known as principal components, such that the first principal component accounts for the largest variance in the data and the second component the second largest variance and so forth (Hotelling, 1933). We performed PCA on the nuclei feature data and selected the number of components which accounted for 95% of the data variance.

The parameters of GMM were estimated using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The number of clusters (k) in the data was evaluated by computing the Bayesian Information Criterion, which attempts to maximize the log likelihood of the data samples for the given model parameters while minimizing the model complexity to avoid overfitting. Hence, the number of clusters for the GMM was chosen as the one with the smallest BIC value over the different values of k ranging from 2 to 20.

b) **Classification of nuclei:** A random selection of 100 nuclei of each subpopulation were visually inspected and labelled as “good” quality or “bad” quality cells. These labelled cell samples were used to train a random forest classifier (Breiman, 2001), which then automatically removed all bad quality cells and retained well-segmented and in-focus cell samples. We used random forest, which has the advantages of improved predictive accuracy and reduced overfitting, achieved by fitting a number of decision tree classifiers to the training data generated by randomly sampling the original data with replacement. The class of an input sample is predicted using a voting strategy in which the class with the highest mean probability estimate across decision trees is the predicted class.

We used model selection method to estimate the parameters of the RF classifier: number of trees in the random forest (n_estimators) and the number of features considered by individual trees at each split (max_features). We tested different values of these parameters, n_estimators ranging from 10 to 5000 and max_features = \{√N_f, log_2(N_f), 0.2N_f\}, where N_f is the number of nuclei features. Model selection method for estimating the performance of all the models using different
parameters and choose the best one, involved the following steps. First, we divided the data into two sets of 80% of the data samples for training set which is used for parameter estimation and 20% for the test set used to assess the generalization performance of the best model on new unknown data. To avoid overfitting of the RF parameters, we performed model selection on the training set using 10-fold cross-validation and used “Area under the Precision-Recall (P-R) curve” as the model selection criterion. Here, precision can be defined as the fraction of good quality cells out of the total number of correctly classified cells and recall is a measure of how many good quality cells can be correctly classified. In order to achieve both high precision and recall scores, our model selection method was designed to choose the model with high area under the P-R curve. Hence, the selected RF parameters corresponded to the model with the highest area under the P-R curve measured by taking the average across 10-fold cross-validation iterations.

2) Feature Quality Control and Control Well Selection

Following the NQC step, we performed quality control on the Golgi channel features to select the most reproducible features across the control population. We also eliminated the features with a constant value since these features do not provide any information pertaining to the different control and test treatments. Since, we use several control wells, the feature reproducibility depends on the feature distribution of individual control wells. Hence, noise in the feature distribution can be attributed to variations across many wells or to a few outlier wells. However, in the latter case it may be more efficient to eliminate the few outlier wells which contribute to noise in several feature distributions. To take into account these factors, we have devised a greedy method which computes a rejection score when each feature and each control well is eliminated one at a time. The rejection score is a measure of how many feature distribution functions for different control wells are different from the reference feature distributions. Hence, the proposed feature-well selection procedure recursively eliminates a feature or a control well, whichever leads to the lowest rejection score, as explained next.

Firstly, we generated the empirical cumulative distribution functions (cdf) of all Golgi features for all control wells, and used the non-parametric Kolmogorov-Smirnov test to compare the cdf’s of features. This was done by computing the KS statistic by comparing the cdf of a feature of a control well with the reference cdf of the same feature, which was generated by combining the feature values from the remaining control wells. Next, we obtained the p-value assuming that the null hypothesis, i.e. the two cdf’s are identical, is true. Since multiple KS tests are performed for all the feature cdf’s of each well, multiple comparison analysis was done using the Holm method to generate p-values corrected for multiple tests. Finally, using these corrected p-values, we computed the rejection ratio (RR) as the fraction of tests for which the null hypothesis can be rejected using a significance level of $\alpha = 0.05$. The step-wise elimination of features and control wells on the basis of RR values is described below:

1) Starting with the original set of features and all control wells, compute rejection score $\text{RR}_{\text{best}}$
2) Remove every feature, $i \in \{1, 2, ..., N_f\}$, independently and for every removed feature, compute the rejection ratio using the remaining features and all wells, $\text{RR}_{i, \text{feat}}$.
3) Remove every well, $j \in \{1, 2, ..., N_w\}$, independently and for every removed well, compute the rejection ratio using all features and the remaining wells, $\text{RR}_{i, \text{well}}$. 
4) Eliminate the feature or well resulting in the lowest rejection score, RR_{new} such that RR_{new} < RR_{best}, and set RR_{best} = RR_{new}.

5) Repeat steps 2-4 until RR_{best} doesn’t reduce further.

The reproducibility score of the \text{i}^{\text{th}} feature can be written as a function of RR_{i\text{feat}}, i.e. \text{RS}_{i} = (1 - RR_{i\text{feat}}) \times 100.

After obtaining the set of homogeneous control wells, we next performed the quality control for the test wells.

3) Test Well Quality Control

Similar to the quality control for control wells, we selected the most homogeneous replicate wells and eliminated outlier wells for each test treatment. Since we used fewer replicate wells for the test treatments as compared to the number of untreated control wells, our test well selection strategy was different from the control well selection method. For each test treatment, we performed KS tests for pairwise comparisons of the cdf’s for a particular feature of all replicate wells. This was repeated for all homogeneous features selected in the previous module. A multiple comparison analysis was then performed for the multiple KS tests for all well pairs and all features. Using the corrected p-values, for each well pair, we computed the number of features for which the null hypothesis, i.e. the feature cdf’s of the two wells are identical, can be rejected. Hence, for all well pairs, we obtained the feature rejection scores as the fraction of features for which null hypothesis can be rejected with a significance level of \( \alpha = 0.05 \). Finally, we eliminated the replicate wells with rejection scores greater than a threshold value of 0.1, when the feature cdf’s of the well are compared with the feature cdf’s of all remaining replicate wells. Hence, a treatment replicate well was removed for further analysis if it failed to pass the quality control test with any of the remaining replicate wells, for less than 90% of the features.

4) Control Modelling

In order to identify novel or subtypes of the visually detectable Golgi phenotypes, we first defined a control-like model using the control cells from the untreated homogeneous wells selected earlier. This model was constructed by fitting a one-class support vector machine (SVM) to the control cells in the multi-parametric space defined by the most reproducible features in the control space. The one-class SVM learns the decision boundary of the control space such that it encompasses about 95% of the control cells, and the remaining 5% cells were considered as outliers to allow for normal biological variability. This was done by training one-class SVM with radial basis function kernel using the parameter \( \epsilon = 0.05 \) to set an upper bound on the fraction of training outliers and lower bound on the fraction of support vectors. The kernel coefficient \( \gamma \) was varied from \( 10^{-6} \) to \( 10^{-1} \) and selected using leave-one-out cross-validation by removing one control well at a time and training the model on the data from the remaining wells. A score was computed to measure the trade-off between training error and ratio of support vectors and the \( \gamma \) value yielding the lowest cross-validation score was chosen for the best model.

After training the SVM, all the cells were classified as control-like or non-control-like. We then computed the penetrance rate (PR) for each well, defined as the percentage of non-control-like cells in that well. Finally, we eliminated the low penetrance wells by removing all the wells with PR less
than a particular penetrance threshold. We used a threshold of 10% and considered only non-control-like cells from the high penetrance wells for further unsupervised phenotypic analysis.

5) Unsupervised Clustering to Identify Golgi Morphological Phenotypes

To identify the novel or subtle Golgi phenotypes, we fitted GMM to the non-control-like cell population from the test treatment wells with PR above 10%. Similar to the unsupervised clustering of nuclei using GMM for the NQC step, we first performed dimensionality reduction using PCA and generated the low dimensional Golgi data by retaining the principal components accounting for 95% of the variance in the data. The number of components/clusters for GMM was estimated by computing the BIC criterion for different number of clusters ranging from 2 to 30, and then choosing the model with the smallest BIC. Next, we eliminated the insignificant clusters corresponding to the clusters with the number of cells less than 1% of the total number of non-control-like cells and re-numbered the remaining large clusters according to cluster sizes from 1 to N_c.

a) Measuring cluster relationships: We computed the distance between cluster means to measure the similarity between different clusters. The distance between two clusters was calculated as the Euclidean distance between the means of GMM components, which were defined in the space spanned by the principal components explaining 95% of the variance in the data. Using the pairwise cluster distances and average linkage, hierarchical clustering was performed and dendrograms were plotted to show the cluster relationships. This cluster relationship dendrogram plot can enable us to identify distinct Golgi phenotypic clusters and also clusters with some overlap or very similar clusters.

b) Measuring feature importance for clustering structure: The relevance of image features for the GMM clusters was measured by considering it as a supervised classification task. Random forest classifier was trained on the non-control-like cell data used for GMM clustering and was trained to learn the cluster output generated by the GMM. The random forest was trained using similar approach as discussed in the NQC section above. The feature importance scores were derived from the trained classifier as the Gini impurity scores associated with the random forest. This feature score indicates how often a particular feature was selected for splitting at the nodes of the decision trees comprising the random forest and hence its discriminative value for the classification problem of associating feature data with the cluster labels.

c) Measuring similarity/distance between treatments: In order to measure the similarity between different treatments and the relationship between these treatments and phenotypic clusters, the clustering results were further analyzed as discussed next. We defined cluster signatures for each treatment such that for the p^{th} treatment, signature \( S_p \) can be written as a vector:

\[
S_p = [n_p^1, n_p^2, \cdots, n_p^{N_c}]
\]

where, \( n_p^i \) is the number of cells corresponding to the \( p^{th} \) treatment in the \( i^{th} \) cluster. Further, we computed the distance between each pair of treatments using normalized signatures \( S_p^N \). We quantified the similarity between two treatments by calculating the Hellinger distance (Vajda, 1989) between their normalized signatures, which can be considered as discrete probability distributions. Hence, the distance between the \( p^{th} \) and \( q^{th} \) treatments is defined as:
\[ h(S_p^\wedge, S_q^\wedge) = \frac{1}{\sqrt{2}} \| \sqrt{S_p^\wedge} - \sqrt{S_q^\wedge} \|_2 \]

which is related to the Euclidean norm of the difference of square root of normalized signatures and the factor \( \frac{1}{\sqrt{2}} \) allows the distance to range from 0 to 1. Therefore, similar treatments with high association have distances close to 0 and very different treatments with low association result in distances close to 1.

d) Measuring signature reproducibility: In order to assess the signature reproducibility across well replicates of a particular siRNA treatment, we computed Hellinger distances between all pairs of cluster signatures computed for each well of a treatment. We then retained only reproducible well signatures by eliminating the wells whose distance from all the other treatment wells is greater than a cut-off value. This inter-well Hellinger distance cut-off value was set to 0.3 (well to well reproducibility factor). Finally, we obtained the average signature for each treatment and then computed the inter-treatment distances for all treatment pairs.

e) Constructing phenotypic network: After obtaining the Hellinger distance values between all pairs of treatments, we constructed a phenotypic network to understand how these distances reflected similar treatment signatures. We used a distance threshold value of 0.2 such that a pair of treatments with Hellinger distance less than the threshold was considered to be connected in the phenotypic network. We used hive plot to depict the network and compared it with STRING network predicted interactions (adjusted for a confidence at 0.7 based on experimental evidence) (Krzywinski et al., 2012).

Additional References:

Bishop, C.M. (2006). Pattern recognition. Machine Learning 128, 1-58.
Breiman, L. (2001). Random forests. Machine Learning 45(1), 5–32.
Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology 24, 417.
Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10(213).
Krzywinski, M., Birol, I., Jones, S.J., and Marra, M.A. (2012). Hive plots—rational approach to visualizing networks. Brief Bioinform 13, 627-644.
Vajda, I. (1989). Theory of statistical inference and information. Kluwer Academic Pub.