Supplementary Note for

A Screen for Morphological Complexity Identifies Regulators of Switch-like Transitions between Discrete Cell Shapes

*To whom correspondence should be addressed
E-mail: STWong@tmhs.org (S.W.); Chris.Bakal@icr.ac.uk (C.B.); ZYin@tmhs.org (Z.Y.)

1. Image processing and cell morphology quantification

We developed G-CELLIQ (Genomic CELLular Imaging Quantitator), an integrated workflow for processing large volumes of digital images generated from high-throughput/genome-scale High-Content Screens (HCS). G-CELLIQ is freely available for academic use [1]. Our software performs both image segmentation and feature extraction as follows:

1.1 Image Segmentation

A three-stage cell segmentation method is used, consisting of nuclear segmentation, cell body segmentation, and over-segmentation correction [2-5], as shown in Supplementary Fig. S1.

Nuclear segmentation: There are three steps in this stage: binarization, nuclei detection, and seeded-watershed based nuclei segmentation [5, 6]. The **binarization** step features adaptive thresholding technology: a data-driven background correction algorithm is first used to estimate the background with cubic B-spline [7, 8]; each pixel is then classified as belonging to a nucleus or the background based on the difference between its intensity and the estimated background intensity. Because binarization usually fails to segment clustered nuclei [9], we applied further processing steps to binarization results to detect nuclei**.** First a combined image is obtained as:

$I_{\mathbf{com}} = I_{\mathbf{int}} + 0.8 * I_{\mathbf{dis}}$;

$I_{\mathbf{int}}$: the original image with intensity information;

$I_{\mathbf{dis}}$: the distance image obtained by applying the distance transform on the binary image [10].

$I_{\mathbf{com}}$ is then filtered with a Gaussian filter (with standard deviation $\sigma = 2$). In the filtered image, the noise is suppressed and the local maxima tend to correspond to the cell centers. **Nuclei detection** is then carried out in the gradient vector field (GVF) to further eliminate the possible noisy local maxima[2, 11]; here, the redundant stains in the nuclei channel are removed by the non-maxima suppression operation. Finally, given the nuclei centers defined from the combined image, marker-controlled **seeded-watershed** methods are used to delineate the nuclei shapes.

Cell body segmentation: Cell body quantification needs information from both F-actin and α-tubulin channels. The signal from these two channels are combined as $I = I_{\mathrm{F-actin}} + I_{\alpha-\mathrm{tubulin}}$. Adaptive thresholding methods [7, 8] are used to separate the cell bodies from the background. After thresholding, the nuclear segmentation results are planted onto the binary cell body image as the seed information, and the seeded-watershed method [2] uses this seed information to delineate the cell bodies. This strategy tackles the challenging cases where multiple cell bodies are touching each other.

Over-segmentation correction: Few cells are under-segmented due to the involvement of nuclei information as seed for cell body segmentations. Conversely, an over-segmentation problem arises when there are multiple nuclear regions within cells (e.g. following failed cytokinesis). We implemented a threshold based method to reduce the over-segmentation. Each cell segment is assigned a neighborhood cell segment with which it shares the longest common boundary. Then a rectangular region is defined across the common boundary of the two touching segmented patches, and the intensity variation within the rectangle is calculated. The two patches are merged if intensity variation within the rectangle is smaller than a given threshold.

Image quality control: The following procedure is implemented to select high quality cell images:

1) Before nuclei segmentation, the histogram and the calculated threshold for binarization for each image are compared to those from manually validated good quality images to exclude extremely dark or bright images.
2) Images with less than 10 candidate nuclei are discarded.
3) Cells that touch the image boundary are discarded.

1.2 Feature Extraction

To quantify the geometric and texture properties of each segmented cell, 211 morphology features were extracted [3]. The selected features include a total of 85 wavelet features (70 features from Gabor wavelet transformation [12] and 15 features from 3-level CDF97 wavelet transformation [13]), 11 whole-cell geometric features extracted from the whole cell body [3], 47 Zernike moments features with a selected order of 12 [14], 14 Haralick texture features [15], and a total of 54 regional geometric features extracted from divided parts of cell segments (36 features of ratio length of the central axis projection and 18 features of area distribution over equal sectors) [3]. All features are extracted from images generated by combining the three channels.

Wavelet features (feature No. 1~85): Two important types of discrete wavelet transformation, the Gabor wavelet [12] and the Cohen–Daubechies–Feauveau wavelet (CDF9/7) [13], were applied to extract cellular texture properties. We extracted the mean and standard deviation of Gabor texture features, as defined in [16], with 6 scales and 4 orientations. Altogether, 70 features were obtained and numbered 1~70 in the resulting feature set. Furthermore, 3-level CDF97 wavelet transformation [13] was performed to extract additional texture signatures. In each level, the minimum, maximum, mean, median of maximum distribution, and standard derivation were calculated for each transformed image. In total, we obtained 15 CDF97 wavelet features from each cell segment, and included them as feature No. 71~85.

Geometry-I: Whole-Cell Geometry features (feature No. 86~96): The 3-channel images for each cell segment were first combined into a gray level image, then 11 geometry features were extracted using the *regionprops* function in image processing toolbox of Matlab^{TM}, as defined in Supplementary Table S1.

Zernike moments features (feature No. 97~143): Based on [14], for each cell, Zernike moments of order 12 were obtained within a unit circle centered at the cell mass center. Each order generated 4 features, and with the first output in the lowest order excluded, 47 moments features in total were obtained.

Haralick texture features (feature No. 144~157): As a traditional texture signature, the Haralick Co-occurrence features, with a total of 14 attributes listed in Supplementary Table S2, were extracted from the gray-level spatial-dependence matrices for each cell segment [15].

Geometry II: Regional geometry features: Two groups of regional geometry features, "length ratios of the central axis projection" and "the area ratio over equal sectors", were extracted after further dividing each cell segment, as summarized in Supplementary Table S3[3].

We define the cell centroid (m_x, m_y) as the first order moments of the binary image $f(x,y)$ for each cell segment. A series of central radial axis are then defined as the line L_α. The central projection along L_α is quantified by the length of the cell boundary contained by two neighboring central axes. The ratio length of the central projection r_{L_α} is defined as $r_{L_\alpha} = \frac{1}{p}\int_{L_\alpha} f(r)\mathbf{d}\,r$ where $p = \int_{whole\ cell} f(r)\mathbf{d}\,r$ is the perimeter of the cell, and 36 ratio length feature sets are evenly sampled around the cell.

The entire cellular region is partitioned into 18 sectors centered at the cellular centroid with even radius angles; the ratio area is defined as the ratio between the area of the fan bin S_β to the area of entire cell segment: $r_{S_\beta} = \frac{\int_{(x,y)\in S_\beta} f(x,y)\mathbf{d}x\mathbf{d}y}{\int f(x,y)\mathbf{d}x\mathbf{d}y}$. The two shape descriptors are not invariant upon rotation of cell segments, and we sorted the calculated ratio length and ratio area in descending order to partially address this issue.

2. Phenotype modeling and cell classification

2.1 Feature selection using SVM-RFE and GA-SVM

To discriminate between different phenotypes (as judged by an expert), we need to identify a subset of relevant features to these phenotypes from our 211 morphological phenotypes. Initially, SVM-RFE (SVM-Recursive Feature Elimination) with linear kernel [17, 18] and cross validation was used to select the top 20 informative features. However, it has long been argued that such "greedy combination" of good individual features may not be the best option, and in most cases SVM-RFE tends to over-estimate the optimal feature number. Thus, we performed a secondary feature selection using Genetic Algorithm with SVM (GA-SVM) [19, 20]. Twenty initial populations, each having 200 individual features, were created. Each initial population favors one of the top 20 candidate features from SVM-RFE, and each population was divided into four groups, as detailed in Supplementary Table S4. GA optimizations were run 20 times and selected the subset giving the lowest value of target function, which is the mean error rate through 100 times 10-fold cross validations [17, 18] on the training dataset for each phenotype. In both SVM-RFE and GA-SVM stages we used SVMs with linear kernel [21-23] to assess the criteria for feature elimination and the fitness function for GA, respectively. SVMs were implemented using the **SVMTrain** and **SVMclassify** functions of Matlab^{TM}.

The implementation of GA-SVM was based on the Genetic Algorithm and Direct Search Toolbox in Matlab 7.1 (R14). Specifically, the option structure creation function **gaoptimset** used the following parameters: population size of 200, maximum generation of 100, default crossover rate of 0.7 and mutation rate of 0.1. In each generation the top 3 elite individuals with the highest fitness function were kept into the next generation.

2.2 Cell classification using SVM

A support vector machine (SVM) [21, 22] with Gaussian Radial Basis Function (RBF) kernel is used for cell classification due to its flexibility to handle the non-linear relationships between the classes. For each classifier, the continuous output from discriminate function *f*(**x**) is used directly to indicate the similarity between the specific training set and test sample. Similar to [24], the classification result for single cell is the basis of functional score for each experimental condition.

Grid search for SVM parameters: All SVMs involved were implemented using LIBSVM v3 package [23]. An SVM with Gaussian RBF kernel has two main parameters: width for Gaussian kernel γ and penalty for training error C. A two-stage search of optimal parameters was applied. Using *grid.py* in [23], the **preliminary search** employed exponentially growing sequences as $C \in \{2^t | t \in \mathrm{Z}, -8 < t < 8\}$and $\gamma \in \{2^k | k \in \mathrm{Z}, -12 < t < 3\}$. For each combination of C and γ , we carried out 10 times of 10-fold cross validation on available training sets, and made sure that all samples were used as testing sample at least once. The parameter set with the best cross-validation performance was selected as the candidate, and a **secondary search** in linear scale was carried out in the neighborhood of the candidate to determine the final parameter set for the corresponding SVM classifier.

3. Generation of Quantified Morphological Signature (QMS)

3.1 Morphological signature of each dsRNA

3.1.1 Quantifying single cell morphology using SVM classification results

SVM classifier attempts to find a hyperplane that best separates the positive and negative classes. The raw output of SVM is the distance from **x** to the discriminant hyperplane and quantifies the similarity between **x** and the given class. Thus, using five SVMs trained in 2.2, the morphology of each single cell can be quantified by five scores. Specifically for each cell x, we have:

1) $s_x^{class}, class \in \{N, L, C, T, R\}$, raw output from the SVM using *class* as the positive training set,
   s_x^N comes from SVM using **N**ormal cells as positive class; and similarly,
   L denotes elongated, bipolar, spindle shaped cells;
   C denotes very large flat cells with smooth edges;
   T denotes small, partially polarized 'teardrop' shaped cells;
   R denotes large flat ruffled cells
2) $\mathbf{s}_x = [s_x^N, s_x^L, s_x^C, s_x^T, s_x^R]$, a 5-tuple score vector quantifying the cell morphology.

3.1.2 Normal cell filtering

We observe that different dsRNAs are variably penetrant in their effects on cell shape (Supplementary Fig. S2). For example, a dsRNA targeting sticky results in a population where ~90% of cells are large and bi-nucleate (Supplementary Fig. S2a), whereas a dsRNA that depletes Pvr results in "L" cells in ~30% of the population (Supplementary Fig. S2a). By plotting the percentage of normal "N" cells in each EC we could quantify penetrance (Supplementary Fig. S2b). Differential penetrance is not due to differential uptake since dsRNAs targeting thread result in cell death in nearly 100% of cells (Supplementary Fig. S2c).

In order to gain insight into the reasons for differential penetrance we used immunofluorescence microscopy to quantify protein levels of Drosophila ERK and Akt in single cells where ERK and Akt respectively were depleted by different dsRNAs (Supplementary Fig. S2d,e). Interestingly, we see that protein levels of both ERK and AKT can vary by 2-3 fold in different wild-type populations, While dsRNAs reduce these levels overall, there is still a population of cells with ERK or AKT levels that are comparable to levels found in many wild-type cells. Thus we propose that differential penetrance is largely due to the inability of different dsRNAs to knockdown protein levels below a certain threshold.

Mahalanobis distance [25] was used to determine whether a cell has similar morphology as the predefined normal cell. Subsequently we could filter normal cells from different populations:

1) $\mathbf{N}$=[$\mathbf{s}_n$], where each row vector $\mathbf{s}_n$ quantify the morphology of a cell *n*, and cell *n* belongs to the training set for Normal phenotype;

   $\boldsymbol{\mu}_{\mathbf{N}}$, the mean vector for **N**;

   $\Sigma_{\mathbf{N}}$, the covariance matrix for **N**;

2) $\mathbf{s}_x = [s_x^N, s_x^L, s_x^C, s_x^T, s_x^R]$ for each single cell *x*;

3) $d_{\mathrm{N}}(\mathbf{s}_x, \mathbf{N}) = \sqrt{(\mathbf{s}_x - \boldsymbol{\mu}_{\mathbf{N}})\Sigma_{\mathbf{N}}^{-1}(\mathbf{s}_x - \boldsymbol{\mu}_{\mathbf{N}})^{\mathrm{T}}}$, the Mahalanobis distance between $\mathbf{s}_x$ and **N;**

4) $\mathbf{M}$=[$d_{\mathrm{N}}(\mathbf{s}_n, \mathbf{N})$], all distances between the complete dataset **N** and each row vector $\mathbf{s}_n$ within **N;**

   $\boldsymbol{\mu}_{\mathbf{M}}$, the mean of M;

   $\sigma_{\mathbf{M}}$, the standard deviation of M.

The following criteria were used for normal cell filtering:

a) Given all cells within a single well, calculate $d_{\mathrm{N}}(\mathbf{s}_x, \mathbf{N})$;

b) Calculate the mean Pearson correlation coefficients between $\mathbf{s}_x$ and every score vector in **N**;

c) If $\mathbf{s}_x$ has i) $d_{\mathrm{N}}(\mathbf{s}_x, \mathbf{N}) \leq \boldsymbol{\mu}_{\mathbf{M}} + \sigma_{\mathbf{M}}$, ii) average correlation between $\mathbf{s}_x$ and **N** is larger than 0.85 and iii) s_x^N larger than any of $\{s_x^L, s_x^C, s_x^T, s_x^R\}$, the corresponding cell *x* is considered a normal cell.

d) If less than 75% cells are considered normal in a well, remove the normal cells based on step c). The threshold of 75% is set according to the mean (0.8872) and standard deviation (0.1129) for the ratio of normal cells across all wells in control baseline (Supplementary Fig. S2b).

3.1.3 Raw morphology score for a single well

Given a single well **w**, after image quality control and normal cell filtering, the raw morphology score $\mathbf{S_w}$ is the average of single cell scores in **w**.

3.1.4 Normalization of raw well scores

899 dsRNAs were deployed into 5 different plates. Each deployment was repeated 3 times, thus any given dsRNA was repeated at least three times. Two types of negative control wells exist in each plate: “control empty” wells where no dsRNA was added and “control *LacZ*” wells where a null dsRNA targeting *LacZ* was added and was not supposed to cause any phenotype change. We have:

Control baseline $\mathbf{B} = [\mathbf{s}_b]$, all raw morphology scores of more than 200,000 cells belonging to control wells;

$\mu_{\mathbf{B}} = [\mu_B^N, \mu_B^L, \mu_B^C, \mu_B^T, \mu_B^R]$, the mean of **B**;

$\sigma_{\mathbf{B}} = [\sigma_B^N, \sigma_B^L, \sigma_B^C, \sigma_B^T, \sigma_B^R]$ the standard deviation of **B**;

Thus, given any raw score vector $\mathbf{S}_\mathbf{w} = [s_\mathbf{w}^N, s_\mathbf{w}^L, s_\mathbf{w}^C, s_\mathbf{w}^T, s_\mathbf{w}^R]$ for certain well **w**, we normalize $\mathbf{S}_\mathbf{w}$ into the Z-score of control baseline, i.e. $z_\mathbf{w}^{class} = \frac{s_\mathbf{w}^{class} - \mu_B^{class}}{\sigma_B^{class}}, class \in \{N, L, C, T, R\}$. The normalized score $\mathbf{Z}_\mathbf{w} = [z_\mathbf{w}^{class}, z_\mathbf{w}^{class}, z_\mathbf{w}^{class}, z_\mathbf{w}^{class}, z_\mathbf{w}^{class}]$ quantifies the morphological change in well **w** comparing to control baseline.

3.1.5 Repeatability test for wells using a same dsRNA

To test for repeatability, we test whether wells that are treated with dsRNA D form a compact cluster, i.e. whether cells treated with **D** have a significantly smaller dispersion than random group of cells. We denote wells with the same dsRNA ***D*** by $\mathbf{W}_{\boldsymbol{D}}=\{\mathbf{w}_1,\ldots\mathbf{w}_n\}$:

1) the dispersion measurement $\mathbf{R}_{\mathbf{W}_D} = \boldsymbol{log}(\frac{\mathbf{1}}{\mathbf{2n}}\sum_{i,j\in\mathbf{W}_D} \boldsymbol{d}\left(\boldsymbol{Z}_{w_i}, \boldsymbol{Z}_{w_j}\right)^{\mathbf{2}})$, where $d(\bullet,\bullet)$ denotes the Mahalanobis distance between two vectors;
2) 1000 randomly sampled cell groups $\mathbf{W}_{\boldsymbol{D}}^{(k)} = \{\mathbf{w}_1^{(\boldsymbol{k})}, \ldots \mathbf{w}_{\boldsymbol{n}}^{(\boldsymbol{k})}\}, k = 1,2 \ldots 1000$. Each $\mathbf{W}_{\boldsymbol{D}}^{(k)}$ has the same number of wells as $\mathbf{W}_{\boldsymbol{D}}$, and each well $\mathbf{w}_\mathrm{i}^{(\boldsymbol{k})}, \boldsymbol{i} = \mathbf{1}, \mathbf{2} \ldots \boldsymbol{n}$ consists of cells from the same plate as $\mathbf{w}_i$; *i.e.* each $\mathbf{W}_{\boldsymbol{D}}^{(k)}$ has the same cell number as $\mathbf{W}_{\boldsymbol{D}}$, while containing cells randomly sampled from the plates containing wells $\mathbf{w}_1,\ldots\mathbf{w}_n$; and thus the random sampled cells are subject to non-specific RNAi treatments;
3) 1000 random dispersion measurements $\mathbf{R}_{\mathbf{W}_{\boldsymbol{D}}^{(k)}}, \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$;
4) $\mathbf{R}_{\boldsymbol{D}}^{\mathbf{0}}$, the mean value of $\mathbf{R}_{\mathbf{W}_{\boldsymbol{D}}^{(k)}}, \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$, and $\mathbf{P}_{\boldsymbol{D}}^{\mathbf{0}}$**,** the estimated distribution of $\mathbf{R}_{\mathbf{W}_{\boldsymbol{D}}^{(k)}}$ from the non-parameterized Parzen window method [26].

A one tail permutation test is set up with $\mathbf{P}_{\boldsymbol{D}}^{\mathbf{0}}$ as the null distribution:

Null hypothesis H$_0$: $\mathbf{R}_{\mathbf{W}_D} = \mathbf{R}_{\boldsymbol{D}}^{\mathbf{0}}$;

i.e. $\mathbf{R}_{\mathbf{W}_D}$ (dispersion for cells with a same dsRNA) is from the same distribution $\mathbf{P}_{\boldsymbol{D}}^{\mathbf{0}}$ as the cells subject to random RNAi.

Alternative hypothesis H$_1$: $\mathbf{R}_{\mathbf{W}_D} < \mathbf{R}_{\boldsymbol{D}}^{\mathbf{0}}$;

i.e. when targeted by a same dsRNA, the cells show a significantly smaller dispersion than the cells undergoing random RNAi.

The null hypothesis is rejected at 5% significant level and is carried out in an iterative manner, such that when null hypothesis cannot be rejected for $\mathbf{W}_{\boldsymbol{D}}=\{\mathbf{w}_1,\ldots\mathbf{w}_n\}$, the tests are repeated while members in $\mathbf{W}_{\boldsymbol{D}}$ are iteratively removed, until:

Either null hypothesis is rejected for a subset of $\mathbf{W}_{\boldsymbol{D}}$, whose size is no smaller than *n*/2;

Or all subsets are deemed unrepeatable.

Due to the facts that a) cells for a real well $\mathbf{w}_i$ and a permuted well $\mathbf{w}_{\mathrm{i}}^{(\boldsymbol{k})}$ come from the same plate; b) each plate has different standard deviation on cellular morphology scores; and c) wells treated by a certain dsRNA can be located in a same or different plates, and each plate is repeated several times, the statistical power of this test varies when different plates are involved. When Parzen window estimation is used, 1000 permutations is always enough to detect effect size of 1.5 with 0.85 power at 5% false discovery rate (FDR).

3.1.6 Consolidation of scores through weighted average

Assume the permutation test in 3.1.5 identified a group of repeatable wells for dsRNA ***D***, the consolidated morphology signature for ***D*** is obtained, where the reciprocal of the Mahalanobis distance from one well to general control baseline serve as the weight for each well, specifically we have:

1) $n_{\mathbf{D}}$, the size for the repeatable group identified in 3.1.5;
2) $Z_{\mathbf{w}}$**,** the normalized morphology score for a (repeatable) well **w** from 3.1.4;
3) $d_{\mathbf{W}}$, Mahalanobis distance from a (repeatable) well **w** to control baseline;

The consolidated score for dsRNA D has the form of: $\mathbf{Z_D} = (\sum_{i=1}^{n_\mathbf{D}} \frac{1}{d_{\mathrm{w}_i}} \mathbf{S}_{\mathrm{w}_i}) / (\sum_{i=1}^{n_\mathbf{D}} \frac{1}{d_{\mathrm{w}_i}})$.

3.2 Morphological signature of each gene

Even after filtering (Supplementary Fig. S3a) different dsRNAs targeting a same gene can elicit very different responses (Supplementary Fig. S3b,c), similar to 3.1.5, repeatability tests were also carried out on all dsRNAs targeting a same gene.

3.2.1 Repeatability test based on Mahalanobis distance

A permutation test similar as in 3.1.5 was applied. Assume dsRNAs D_1 and D_2 are biological replicates targeting a same gene G, we have:

1) $\mathbf{Z}_{\mathbf{D}_1}$ and $\mathbf{Z}_{\mathbf{D}_2}$**,** the morphological signature for D_1 and D_2, respectively;

2) $\boldsymbol{d}(\boldsymbol{Z}_{\mathbf{D}_1}, \boldsymbol{Z}_{\mathbf{D}_2})$**,** the Mahalanobis distance between two signatures;

3) 1000 randomly sampled cell groups $\mathbf{D}^{(k)} = \{\mathbf{D}_1^{(k)},\ \mathbf{D}_2^{(k)}\}, k = 1,2 \ldots 1000$. Each sampled cell group $\mathbf{D}_i^{(k)}$ ($\boldsymbol{i} = \mathbf{1}, \mathbf{2}; \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$) has the same number of wells as real cell group D_i, and each sampled well in $\mathbf{D}_i^{(k)}$consists of cells from the same plate as the corresponding real well in D_i ; *i.e.* each $\mathbf{D}_i^{(k)}$ has the same cell number as D_i, while containing cells randomly sampled from the plates containing corresponding real wells and thus the random sampled cells are subject to non-specific RNAi treatments;

4) 1000 random Mahalanobis distances $\boldsymbol{d}(\mathbf{D}_1^{(k)},\ \mathbf{D}_2^{(k)}), \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$;

5) $\boldsymbol{d}_0$, the mean value for $\boldsymbol{d}(\mathbf{D}_1^{(\boldsymbol{k})},\ \mathbf{D}_2^{(\boldsymbol{k})}), \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$; and $\mathbf{P}_0$**,** the estimated distribution of $\boldsymbol{d}(\mathbf{D}_1^{(\boldsymbol{k})},\ \mathbf{D}_2^{(\boldsymbol{k})})$ from the non-parameterized Parzen window method [26].

A one tail permutation test is set up with $\mathbf{P}_0$ as the null distribution:

Null hypothesis H$_0$: $\boldsymbol{d}(\boldsymbol{Z}_{\mathbf{D}_1}, \boldsymbol{Z}_{\mathbf{D}_2}) = \boldsymbol{d}_0$**;**

i.e. $\boldsymbol{d}(\boldsymbol{Z}_{\mathbf{D}_1}, \boldsymbol{Z}_{\mathbf{D}_2})$ (Distance for cells with a dsRNAs targeting a same gene) is from the same distribution $\mathbf{P}_0$ as the cells subject to random RNAi;

Alternative hypothesis H$_1$: $\boldsymbol{d}(\boldsymbol{Z}_{\mathbf{D}_1}, \boldsymbol{Z}_{\mathbf{D}_2}) < \boldsymbol{d}_0$**;**

i.e. When subject to dsRNAs targeting a same gene, the morphology signatures show a significantly smaller distance than those from cells undergoing random RNAi.

The null hypothesis is rejected at 5% significant level, and unlike 3.1.5, no iterative steps are necessary because we directly work on each pair of dsRNAs. Also for each pair of dsRNAs, two p-values are calculated based on the cell populations before- and after normal cell filtering, respectively.

Due to the facts that a) cells for a real well $\mathbf{w}_i$ and a permuted well $\mathbf{w}_\mathrm{i}^{(\boldsymbol{k})}$ come from a same plate; b) each plate has different standard deviation on cellular morphology scores; and c) wells treated by a certain dsRNA can be located in a same or different plates and each plate is repeated several times, the statistical power of this test varies when different plates are involved. When Parzen window estimation is used, 1000 permutations are always enough to detect effect size of 1.5 with 0.86 power at 5% false discovery rate (FDR).

3.2.2 Repeatability test based on kernel density estimation and KL/J divergence

The test in 3.2.1 works on the average scores $\mathbf{Z_D}$ across cell populations. Next, we consider the heterogeneity within each cell population and set up a test on the similarity between two probability distributions estimated from the score matrices of two cell populations.

General Denotation: Assume that a cell population **D** contains *n* single cells, and the morphology of each cell *x* can be depicted by a 5-tuple score vector, which is normalized to Z-score of control baseline and denoted as $\mathbf{z}_x = [z_x^N, z_x^L, z_x^C, z_x^T, z_x^R]$. Thus, based on the scoring profile $\mathbf{Z} = [\mathbf{z}_i], i = 1,2 \ldots n$ for **D,** a distribution can be estimated for any $class \in \{N, L, C, T, R\}$ using the non-parameterized Parzen window method [26]. Basically, a Gaussian kernel was applied around each single score z_i^{class}as $\left(\frac{z^{class} - z_i^{class}}{h}\right) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z^{class} - z_i^{class}}{h}\right)^2\right]$, and the estimated probability distribution function (PDF) for any single score from population **D** is denoted as: $\mathbf{P}(z_\mathbf{D}^{class}) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{z^{class} - z_i^{class}}{h}\right)$. Where h is a smooth parameter referred to as bandwidth, and an acceptable guess is $\hbar = 1.06\sigma \mathrm{n}^{-0.2}$, where σ denotes the standard deviation of z_i^{class} in **D** [27].

Bandwidth Estimation: Here we use a point-wise strategy from [27] to adjust bandwidth in the distribution estimation. Given $\mathbf{Z} = [\mathbf{z}_i], i = 1,2 \dots n$ for cell population **D**, hierarchical clustering was carried out based on average linkage, unbiased Pearson correlation coefficients (PCC) between the 5-tuple vectors s_i were calculated, and those cells with PCC greater than 0.9 were assigned into a same subgroup. Thus, **G** different subgroups g=1, 2…**G** can be defined, and we use n_g to denote the number of cells in each subgroup g. A fast one-dimensional Newton optimization was used to minimize the leave-one-out cost function $L_g^{class}(\hbar) = -\sum_{i=1}^{n_g} log\mathbf{P}_{g-i}^{class}(z_i^{class}, \hbar)$, in search of local bandwidth $\hbar_g^{class}$ for each g. Finally, the estimated distribution of score f can be re-organized based on the afore-mentioned cell sub-group assignments as $\mathbf{P}(z_{\mathbf{D}}^{class}) = \frac{1}{n}\sum_{g=1}^{\mathbf{G}}\sum_{i=1}^{n_g}\frac{1}{\hbar_g}K\left(\frac{z^{class}-z_{g,i}^{class}}{\hbar_g}\right)$.

K-L divergence and J-divergence: After estimating a PDF for each score vector for each cell population, we use the J-divergence to measure the difference between two such PDFs. The Kullback–Leibler divergence [28] is a non-symmetric measure of the difference between two probability distributions P and Q, as $\mathrm{KL(P||Q)} = \int \mathrm{P}(x)\log\frac{\mathrm{P}(x)}{\mathrm{Q}(x)}\mathrm{d}x$. Further, J-divergence is developed to make it a symmetric metric in $\mathrm{J(P||Q) = (KL(P||Q) + KL(Q||P))/2}$ [29, 30].

Permutation test for repeatability: Given two cell populations $\mathbf{D_1}$ and $\mathbf{D_2}$, the distributions of five morphological scores are available as $\mathbf{P}\left(z_{\mathbf{D}_{\boldsymbol{i}}}^{class}\right), class \in \{N, L, C, T, R\}, \boldsymbol{i} = 1,2$. To avoid the computational burden in the following steps, we assume mutual independence among all scores. Thus, the overall difference between morphological profiles for $\mathbf{D_1}$ and $\mathbf{D_2}$ can be defined as $\mathrm{J}(\mathbf{D_1}\,|\mathbf{D_2}) = \sum \mathbf{J}(\mathbf{P}\left(z_{\mathbf{D_1}}^{class}\right)||\mathbf{P}\left(z_{\mathbf{D_2}}^{class}\right)), class \in \{N, L, C, T, R\}$. Let $\mathbf{D_1}$ and $\mathbf{D_2}$ denote cell populations generated by two dsRNAs targeting the same gene, similar to 3.2.1., we have:

1) $\mathrm{J}(\mathbf{D_1}\,|\mathbf{D_2})$.
2) 1000 randomly sampled cell groups $\mathbf{D}^{(k)} = \{\mathbf{D}_1^{(k)},\ \mathbf{D}_2^{(k)}\}, k = 1,2 \dots 1000$. The rule of cell sampling is the same as in 3.2.1;
3) 1000 random divergences $\mathrm{J}\left(\mathbf{D}_1^{(k)}\,\middle|\mathbf{D}_2^{(k)}\right), \boldsymbol{k} = \mathbf{1,2 \dots 1000}$;
4) $\mathbf{J}_0$, the mean value for $\mathrm{J}\left(\mathbf{D}_1^{(k)}\,\middle|\mathbf{D}_2^{(k)}\right), \boldsymbol{k} = \mathbf{1,2 \dots 1000}$, and $\mathbf{P}_0$**,** the estimated distribution of $\mathrm{J}\left(\mathbf{D}_1^{(k)}\,\middle|\mathbf{D}_2^{(k)}\right)$ from the non-parameterized Parzen window method [26].

A one tail permutation test is set up with $\mathbf{P}_0$ as the null distribution:

Null hypothesis $\mathbf{H_0}$: $\mathrm{J}(\mathbf{D_1}\ |\mathbf{D_2}) = \mathbf{J_0}$**;**

i.e. $\mathrm{J}(\mathbf{D_1}\ |\mathbf{D_2})$ (divergence for cells with dsRNAs targeting the same gene) is from the same distribution $\mathbf{P}_0$ as the cells subject to random RNAi;

Alternative hypothesis $\mathbf{H_1}$: $\mathrm{J}(\mathbf{D_1}\ |\mathbf{D_2}) < \mathbf{J_0}$**;**

i.e. when subject to dsRNAs targeting the same gene, the morphology signatures show a significantly smaller distance than those from cells undergoing random RNAi.

The null hypothesis is rejected at 5% significant level, and for each pair of dsRNAs, two p-values are calculated based on the cell populations before and after normal cell filtering, respectively. The statistical power of this test varies when different plates are involved. However, when Parzen window estimation is used, 1000 permutations is always enough to detect effect size of 1.5 with 0.90 power at 5% false discovery rate (FDR).

3.2.3 Generation of QMS (Quantitative Morphological Signature) for each gene

Phenotypic scores: Similar to 3.1.6, the consolidated score for a single gene G has the form: $\mathbf{Z_G} = (\sum_{i=1}^{n_G} \frac{1}{d_{D_i}} \mathbf{Z}_{w_i}) / (\sum_{i=1}^{n_G} \frac{1}{d_{D_i}})$**.** Where $\mathbf{Z_G}$ denotes the consolidated score for a gene, d_D is the Mahalanobis distance from the morphology signature of a repeatable dsRNA to the signature of control baseline, and n_G is the number of repeatable dsRNAs for gene G.

Penetrance Z-score: All cell populations underwent normal cell filtering, and those non-normal cells were considered to be a result of the specific RNAi treatment – or "penetrant". For each gene, we summed the number of penetrant cells in all repeatable wells, and determined the ratio of penetrant cells, which was then normalized based on the mean and standard deviation of penetrant cell ratio for control baseline wells to get a penetrance Z-score, denoted as PZ.

Finally, for each gene G, four phenotype scores were combined with PZ to form a 5-tuple QMS $[\mathbf{Z}_G^L, \mathbf{Z}_G^C, \mathbf{Z}_G^T, \mathbf{Z}_G^R, \mathbf{Z}_G^{PZ}]$.

4. Analysis of QMSs

4.1 Hierarchical clustering

In our case, 899 dsRNAs were initially used to inhibit and the majority of known predicted kinases and phosphatases and several kinase/phosphatase regulatory subunits and adapters (KP set) in Kc167 cells. 116 of these showed poor technical repeatability, 71 show poor biological repeatability, and 155 were excluded from the final analysis as *no* repeat was performed. Thus, the scores from 557 dsRNAs were used in generating the QMSs used in hierarchical clustering (Fig. 3 and Supplementary S5) or to calculate the divergence matrix (Fig. 4c of the main text). We also calculated QMS for two types of negative controls separately and incorporated them into the QMS matrix, thus generating 284 genes/conditions for hierarchical clustering. Hierarchical clustering using average linkage was performed with Cluster [31] and Java^{TM} TreeView [32] using uncentered Pearson Correlation Coefficients as the similarity metric; clusters were defined interactively by finding the highest nodes at which the distance measure became greater than 0.90. Other similarity thresholds were evaluated and this was chosen because this level of correlation resulted in coherent groups of qualitatively similar cells.

Enrichment test reveal biological relevance for resulting phenoclusters: Hierarchical clustering of 284 ECs obtained 14 phenoclusters, including 4 singular genes and 10 phenoclusters with at least two members. Fisher's exact tests were applied to identify the enrichment of biological themes, including pathways, GO terms, and other function annotation terms, using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 [33]. As input for DAVID, all gene symbols involved in hierarchical clustering were converted into Entrez ID using the Gene ID conversion tool in DAVID and the Gene and reagent lookup tool from Drosophila RNAi screening center [34]. For each obtained p-value from Fisher's test, Benjamini corrected p-value and False Discovery Rate were calculated to address for multiple tests.

4.2 Using divergence scores to address cellular heterogeneity

Cell-to-cell differences are always present to some degree in any cell population, thus the mean score of a population may not represent the behaviors of any individual cell [35]. Here, we propose to use **divergence based scores** to address the heterogeneity of phenotypic cell populations.

General Denotations: Here, the kernel density estimation and divergence calculation methods in 3.2.2 are extended to the gene level. Assuming each cell population **D** includes all n cells, and the same local bandwidth estimation has been carried out. The distribution of each feature $class \in \{N, L, C, T, R\}$ for population **D** can thus be modeled by $\mathbf{P}(z_{\mathbf{D}}^{class}) = \frac{1}{n}\sum_{g=1}^{\mathbf{G}}\sum_{i=1}^{n_g}\frac{1}{h_{g,i}}K\left(\frac{z^{class}-z_{g,i}^{class}}{h_{g,i}}\right)$. Using the symmetric divergence metric **J** between two distributions, the overall difference between two populations $\mathbf{D_1}$ and $\mathbf{D_2}$ can be obtained. Thus we had $\mathrm{J}_{i,j} = \mathrm{J}(\mathbf{D}_i \,|\mathbf{D}_j), i, j \in \{1,2 \ldots 284\}, i \neq j$ for all 284 ECs. Meanwhile, $\mathrm{J}_{i,i}$ were assigned the ceiling value of the maximum $\mathrm{J}_{i,j}, i \neq j$. Each divergence score was then normalized into a similarity measurement by $\mathrm{Q}_{i,j} = (\max(\mathrm{J}) - \mathrm{J}_{i,j})/(\max(\mathrm{J}) - \min\,(\mathrm{J}_{i,j})$. Larger values of $\mathrm{Q}_{i,j}$ indicate better similarity in phenotype composition (both morphology and ratio) between genes i and j, and **the genes in the same phenocluster should have relatively high value amongst each other.** The matrix Q is visualized as in Fig. 3c of the main text.

Quantifying the differentially sized shape space explored by cell populations: Based on $Q_{i,j}$, we defined a score $Q(4)_i$, i=1,2…284 for the analysis of gene-level cell populations. Given the matrix Q in the previous section, for each gene i, $Q(4)_i$=$Q(2)_i$-$Q(1)_i$ the difference between the mean population similarity calculated on all genes other than i, ($Q(1)_i$) and all genes within the same cluster as gene i ($Q(2)_i$). Meanwhile, $Q(3)_i$ is gene i's mean population similarity with all genes not in the same cluster as i. In an ideal case, large $Q(4)_i$ indicates that RNAi treatment of gene i results in cells exploring a unique area in phenotypic space, and the cell populations are only similar to those in the same phenocluster. When this value is larger than the mean value of the wild-type cells (0.159), the difference is greater than 1 standard deviation (S.D.) of genes in wild-type clusters, and the mean QMS of the population is also different than wild-type cells, these populations are exploring regions of morphological space that are both smaller and distinct from the space explored by the control populations. When this value is smaller than the wild-type mean by a difference greater than 1 standard deviation, the population is considered to be exploring a larger region of space than wild-type cells.

5. Network analysis for screening hits in human

5.1 Extraction of a Protein-Protein Interaction (PPI) network in human

Human PPI networks were obtained from STRING database [36] v9.0, and only those interactions with confidence score larger than 0.6 were used. As a result, 604,897 PPI entries involving 16,518 proteins were selected. Here PPIs were recorded in a directed pattern; thus, a common physical interaction between proteins A and B was recorded as two entries A->B and B->A. Meanwhile certain "directed" interaction categories like gene fusion or transcription factor binding only had one entry. According to the **graphconncomp** function from Matlab^{TM}, 16,452 out of these 16,518 proteins form a strongly connected component, meaning that given the 604,897 PPI entries, and any two proteins A and B in this component:

1) A and B can be connected by selected PPI entries;
2) Both paths A->B and B->A are available without violating the direction of each involved entry.
3) The lengths of shortest paths A->B and B->A are given by **graphallshortestpaths** function in Matlab^{TM}.

5.2 Functional Roles of human homologs for genes with high L scores

Three groups of proteins (Fig. 7d of the main text) were mapped into the connected human PPI network from 5.1:

1) Group-1– 14 proteins have been previously defined as "pro-elongation" proteins, and include:

CrkL, DOCK3, Integrin beta 1, LIMK2, p27Kip1, p53, p130Cas, NEDD9, MyoP, Rab5, RhoE, SMURF2, Src, WASF2;

2) Group-2– 12 proteins have been previously defined as “pro-contractility” proteins, and include:
3) DIP1, EphrinA, FHOD2, gp130, IL-6, LIMK1, MYL2, PAK2, PDK1, RhoC, STAT3, Stathmin1;
4) Group-3– 15 proteins, whose inhibition results in elongation in mouse or human cells (Fig. 7) and include:

PLK1, 14-3-3zeta, PTEN, IRAK1, JAK1, PAR-1, MAPK3, MAPK1, SLK, LOK, Cdk4, Cdk10, MAST1/2, Liprin-B2.

The lengths of shortest paths among any two out of these proteins was obtained in 5.1. Specifically, for each protein **p** in group-3 and any protein **x** from the connected component with 16,452 members, we have:

a) L(**p**, g1), the mean of shortest-path-lengths between protein **p** and every member in Group-1;
b) {L(**x**, g1)}, the array of mean shortest-path-lengths between any protein **x** and every member in Group-1;
   L_{g1}, the mean of {L(**x**, g1)};
   σ_{g1}, the standard deviation of {L(**x**, g1)};
c) Z(**p**, g1)=[L(**p**, g1)- L_{g1}]/ σ_{g1};
d) L(**p**, g2) and Z(**p**, g2), defined similarly between protein **p** and group-2;
e) Diff(**p,** g1, g2)=L(**p**, g1)- L(**p**, g2), **p** belongs to group-3;
   Diff(**x,** g1, g2)=L(**x**, g1)- L(**x**, g2), **x** is any one of the 16,452 connected proteins;

Z-scores Z(**p**, g1) and Z(**p**, g2) indicate whether **p** is closer/further from group-1 or -2 compared to random proteins in the connected component, with a Z-score smaller than -1.98 translating to a p-value <0.05 in standard normal distribution. Meanwhile, Diff(**p,** g1, g2) quantifies whether p is closer to group A or group B. It is worth noting that the mean of Diff(**x,** g1, g2) across 16,452 proteins is -0.08, thus a random protein tends to be closer to group-1 than group-2. However, our group-3 can be divided into 3 subsets:

i) Diff(**p,** g1, g2)>1, closer to Group-2 and projected as positive regulator of Group-2;
ii) Diff(**p,** g1, g2) between [-0.08, 1];
iii) Diff(**p,** g1, g2)<-0.08, closer to Group-1 and projected as negative regulator of Group-1.

References

1. *http://www.cbi-tmhs.org/GCellIQ/NCB*. [cited.

2. Li, F.H., X. Zhou, and S.T.C. Wong, *An automated feedback system with the hybrid model of scoring and classification for solving over-segmentation problems in RNAi high content screening.* Journal of Microscopy, 2007. **226**(2): p. 121 - 132.
3. Wang, J., et al., *Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening.* Journal of Molecular Screening, 2008. **13**(1): p. 29-39.
4. Xiong, G., et al., *Automated segmentation of Drosophila RNAi fluorescence cellular images using deformable models.* IEEE Transactions on Circuit and Sysrems, 2006. **53**: p. 2415 - 2424.
5. Yan, P., et al., *Automatic segmentation of RNAi fluorescent cellular images with interaction model.* IEEE Transactions on Information Technology in Biomedicine, 2008. **12**(1): p. 109 - 117.
6. Li, F.H., et al., *High content image analysis for human H4 neuroglioma cells exposed to CuO nanoparticles.* BMC Biotechnology, 2007. **7**: p. 66.
7. Lindblad, J., et al., *Image analysis for automatic segmentation of cytoplasms and classification of Rac1 activation.* Cytometry A, 2004. **57**(1): p. 22-33.
8. Wahlby, C., et al., *Algorithms for cytoplasm segmentation of fluorescence labelled cells.* Analytical Cellular Pathology, 2002. **24**(2-3): p. 101-11.
9. Vincent, L. and P. Soille, *Watersheds in digital spaces: an efficient algorithm based on immersion simulations.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991. **13**(6): p. 583-598.
10. Borgefors, G., *Distance transformations in digital images.* Computer Vision, Graphics, and Image Processing, 1986. **34**: p. 344-371.
11. Wang, M., et al., *Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy.* Bioinformatics, 2008. **24**(1): p. 94-101.
12. Manjunatha, B.S. and W.Y. Ma, *Texture features for browsing and retrieval of image data.* IEEE Transactions on Pattern Analysis and Machine Intelligenece, 1996. **18**: p. 837 - 842.
13. Cohen, A., I. Daubechies, and J.C. Feauveau, *Bi-orthogonal bases of compactly supported wavelets.* Communications on Pure and Applied Mathematics, 1992. **45**: p. 485 - 560.
14. Zernike, F., *Beugungstheorie des schneidencerfarhens undseiner verbesserten form, der phasenkontrastmethode.* Physica, 1934. **1**: p. 689 - 704.
15. Haralick, R.M., K. Shanmugam, and I. Dinstein, *Textural features for image classification.* IEEE Transactions on Systems, Man and Cybernetics, 1973. **6**: p. 610 - 620.
16. Daugman, J.G., *Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1988. **36**(7): p. 1169-1179.
17. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**(1-3): p. 389-422.
18. Li, G.Z., et al. *Feature selection for multi-class problems using support vector machines*. 2004: Springer-Verlag Berlin.
19. Holland, J., *Adaption in Natural and Artifical Systems*. 1975, Ann Arbor, MI: The University of Michigan Press.
20. Goldberg, D., *Genetic Algorithms in Search, Optimization and Machine Learning*. 1989, Boston, MA: Kluwer Academic Publishers.
21. Vapnik, V., *The Nature of Statistical Learning Theory*. 1995: New York, NY: Springer-Verlag.
22. Cortes, C. and V. Vapnik, *Support-vector network.* Machine Learning, 1995. **20**: p. 273-297.
23. Chang, C.-C. and C.-J. Lin, *LIBSVM: a library for support vector machines*. 2001. p. Software available at http://csie.ntu.edu.tw/cjlin/libsvm.
24. Bakal, C., et al., *Quantitative morphological signatures define local signaling networks regulating cell morphology.* Science, 2007. **316**: p. 1753 - 1756.
25. Mahalanobis, P.C., *On the generalized distance in statistics.* Proceedings of the National Institute of Science of India, 1936. **2**(1): p. 49-55.

26. Parzen, E., *On estimation of a probability density function and mode.* The Annals of Mathematics Statistics 1962. **33**: p. 1065-1076.
27. Scott, D. and S. Sain, *Multi-dimensional density estimation.* Handbook of Statistics, 2004. **24**.
28. Kullback, S. and R.A. Leibler, *On information and sufficiency.* The Annals of Mathematics Statistics, 1951. **22**: p. 76-86.
29. Lin, J., *Divergence measures based on the Shannon entropy.* Information Theory, IEEE Transactions on, 1991. **37**(1): p. 145-151.
30. Johnson, D. and S. Sinanovic, *Symmetrizing the Kullback-leibler Distance*. 2001, Rice University.
31. de Hoon, M.J.L., et al., *Open source clustering software*. 2004. p. 1453-1454.
32. Saldanha, A.J., *Java Treeview--extensible visualization of microarray data*. 2004. p. 3246-3248.
33. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat. Protocols, 2008. **4**(1): p. 44-57.
34. *http://www.flyrnai.org/cgi-bin/DRSC_gene_lookup.pl*. [cited.
35. Altschuler, S.J. and L.F. Wu, *Cellular Heterogeneity: Do Differences Make a Difference?* Cell, 2010. **141**(4): p. 559-563.
36. Jensen, L.J., et al., *STRING 8-a global view on proteins and their functional interactions in 630 organisms.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D412-D416.