nature cell biology

A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes

Zheng Yin[1,2,6], Amine Sadok[3,6], Heba Sailem[3], Afshan McCarthy[3], Xiaofeng Xia[1,2], Fuhai Li[1,2], Mar Arias Garcia[3], Louise Evans[3], Alexis R. Barr[3], Norbert Perrimon[4], Christopher J. Marshall[3], Stephen T. C. Wong[1,2,5,7] and Chris Bakal[3,7]

The way in which cells adopt different morphologies is not fully understood. Cell shape could be a continuous variable or restricted to a set of discrete forms. We developed quantitative methods to describe cell shape and show that *Drosophila* haemocytes in culture are a heterogeneous mixture of five discrete morphologies. In an RNAi screen of genes affecting the morphological complexity of heterogeneous cell populations, we found that most genes regulate the transition between discrete shapes rather than generating new morphologies. In particular, we identified a subset of genes, including the tumour suppressor *PTEN*, that decrease the heterogeneity of the population, leading to populations enriched in rounded or elongated forms. We show that these genes have a highly conserved function as regulators of cell shape in both mouse and human metastatic melanoma cells.

Morphological plasticity is critical to organism development—as exemplified by the reversible conversion of embryonic non-migratory epithelial cells to motile mesenchymal cells required for tissue positioning and organization[1]. The size of the shape space a cell has the potential to explore reflects its morphological plasticity[2]. Highly plastic cells explore large regions of shape space when compared with cells with stable morphologies. In adult organisms, the shape space available to most differentiated cells is relatively limited, serving to enforce tissue architecture and function. However, during the pathogenesis of diseases such as metastatic cancers, cells can re-acquire the ability to explore shape space and thus find a shape that is suitable for migration and invasion[2–6]. At present, there is little understanding of how the size and topology of cellular shape space is determined by genetic and environmental factors.

To identify how genes contribute to the size and topology of shape space we developed high-throughput imaging and computational methods to describe the morphological complexity of cellular populations and applied them to data sets generated by systematic RNA interference (RNAi) screens in *Drosophila* Kc cells. We first determined whether cells have discrete shapes or whether shape is a continuous variable. Subsequently we identified genes that contribute to the exploration of shape space in Kc cells, as well as those that regulate the topology of shape space itself. Finally we isolated a conserved gene network that regulates contractility and protrusion in *Drosophila* as well as mouse and human melanoma cells. This demonstrates that the analysis of morphological complexity provides new insights into the signalling networks regulating cell shape.

RESULTS

RNAi screening indicates that Kc cells exist in discrete shapes

We used RNAi screening in *Drosophila* Kc167 cells (Kc cells) to explore the contribution of genes to morphological complexity (Fig. 1a–f). We use the term experimental condition (EC) for cells treated with double-stranded RNA (dsRNA). Following image processing (see Methods and Supplementary Note, Fig. S1 and Tables S1–S4), we scored cells in each EC on the basis of their similarity to reference shapes[7,8]. Briefly, we used human observers (Fig. 1d) and online discovery algorithms[8] (Fig. 1e) to identify as many distinct cellular shapes (reference shapes) as possible in the data set. Most cells in the data set could be characterized as normal (N) cells, which are rounded

[1]NCI Center for Modeling Cancer Development, The Methodist Hospital Research Institute, Weill Cornell Medical College, 6670 Bertner Avenue, R6 South, Houston, Texas 77030, USA. [2]Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, 6670 Bertner Avenue, R6 South, Houston, Texas 77030, USA. [3]Institute of Cancer Research, Chester Beatty Laboratories, Division of Cancer Biology, 237 Fulham Road, London, UK. [4]Department of Genetics, Harvard Medical School and Howard Hughes Medical Institute, 77 Avenue Louis Pasteur, Boston, Massachusetts 02215, USA. [5]Department of Radiology, The Methodist Hospital, Weill Cornell Medical College, Houston, Texas 77030, USA. [6]These authors contributed equally to this work. [7]Correspondence should be addressed to S.T.C.W. or C.B. (e-mail: STWong@tmhs.org or cbakal@icr.ac.uk)

Received 4 October 2012; accepted 18 April 2013; published online 9 June 2013; DOI: 10.1038/ncb2764

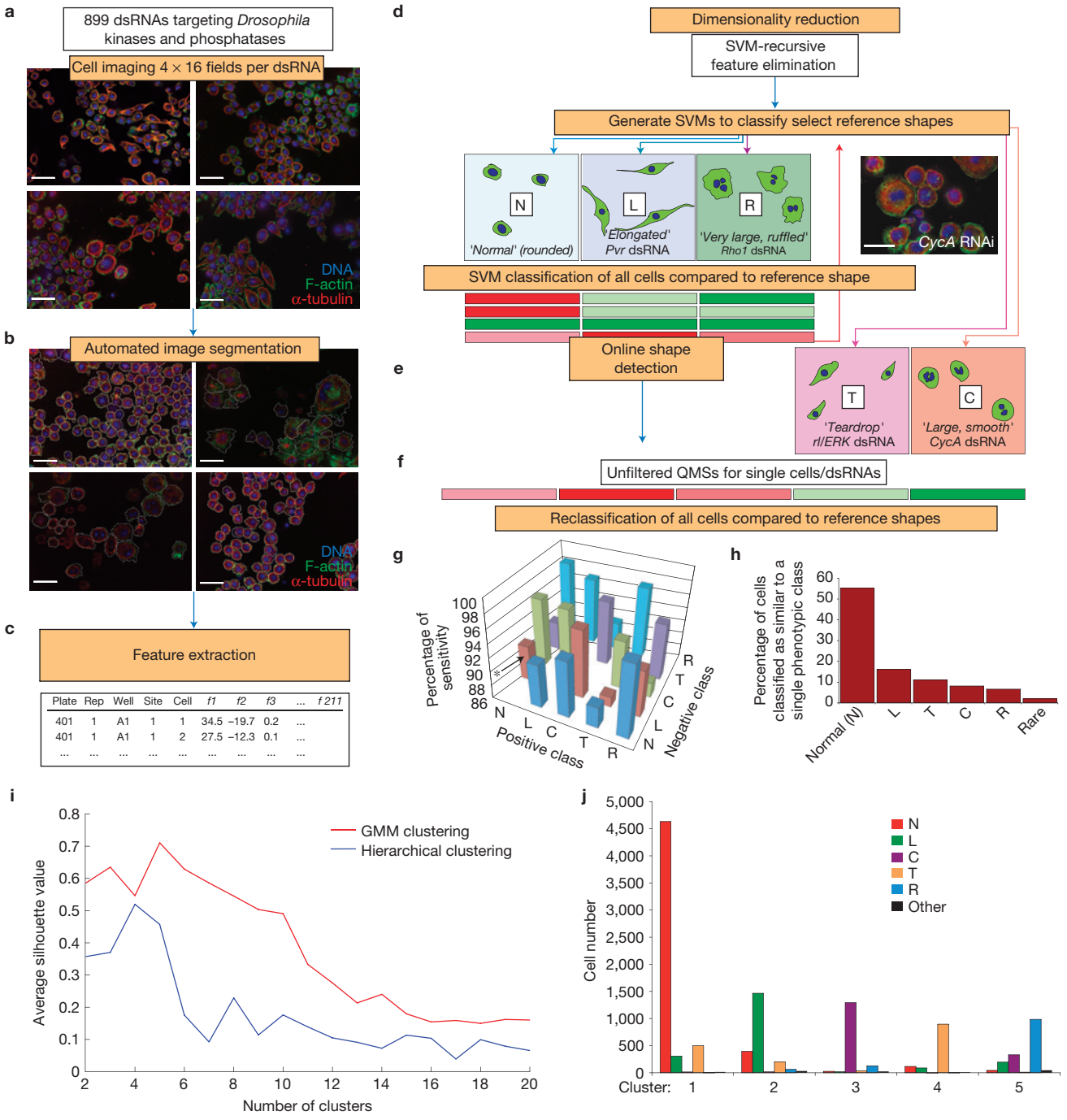

Figure 1 Automated morphological profiling. (**a**) Kc167 cells were incubated with 899 single dsRNAs targeting most *Drosophila* kinases and phosphatases. Experiments were performed in triplicate or quadruplicate in 384-well plates. Following fixation and staining using DAPI, phalloidin and an anti-α-tubulin antibody, each well was imaged at 16 sites by confocal microscopy. Scale bars, 20 μm. (**b**,**c**) Automated image segmentation and feature extraction were performed to generate feature information for 2,038,641 cell segments. Scale bars, 20 μm. (**d**) SVM-recursive feature elimination was used to reduce the dimensionality of the data, and SVM-based classifiers were generated for three initial reference shapes (N, L and R). Individual cell segments were initially classified by assigning raw QMSs based on the similarity of the segments to N, L and R shapes. Scale bar, 20 μm. (**e**) Subsequently, online phenotype-detection methods[8] were implemented to detect the presence of two other shapes, T and C. (**f**) All cells were then re-assigned QMSs based on the similarity of each cell to all five reference shapes, but the comparison to N cells is done by calculating a penetrance *Z*-score of mutant shapes before filtering (see Fig. 2). (**g**) Sensitivity as determined by cross-validation analysis to determine whether two exemplar shapes are quantitatively different from one another using a new SVM classifier. Each test was performed on 200 test cells from training classes comprised as follows: N, 2,185 cells; L, 2,053 cells, C, 2,041 cells; T, 2,002 cells; R, 2,028 cells. For example, if N is considered the positive class, and L is the negative class, the N versus L classifier correctly identifies N cells in 91% of tests (asterisk). (**h**) Cells are classified as most similar to a particular single shape. Cells not assigned to one particular shape are assigned to the rare class. (**i**) Silhouette index for different cluster numbers using hierarchical clustering (blue) or Gaussian mixture models (GMM; red) of principal component data. (**j**) Number of cells with a particular morphology (N, L, C, T, R or other) that are part of a particular cluster (1–5) using hierarchical clustering of principal components. Work-flow process steps in **a**–**f** are labeled in orange.

cells with smooth borders of cortical actin (Fig. 2a), together with another 4 reference shapes. We labelled these reference shapes as L, C, T or R, which correspond to: elongated, bipolar, spindle-shaped cells (L; Fig. 2b); large cells with smooth edges (C; Fig. 2c); small, partially polarized teardrop-shaped cells (T; Fig. 2d); and very large flat cells with ruffled edges (R; Fig. 2e). Five different support vector machine (SVM)-based classifiers were generated that could distinguish these morphological classes. We also derived specific and sensitive Gaussian SVM classifiers that distinguish between pairs of shapes versus simply one shape from all others (Fig. 1g). Every cell in the data set was then scored using each classifier to generate a multi-dimensional vector, or a quantitative morphological signature (QMS) that describes the similarity of that cell to each reference shape (Fig. 1f). Thus, unlike the use of absolute measures (for example, area, size), a QMS is a measure that describes shape relative to other reference shapes. Each cell in the data set is assigned a QMS, and a mean QMS can be calculated for any given population.

To gain a sense of whether our classification systems capture most of the morphological variance present in the data set we investigated whether most cells in the data set could be considered as similar to one of the reference shapes. When all cells in the data set are classified with respect to their similarity to a single phenotypic group, versus determining their similarity to multiple classes simultaneously, we observed that most cells could be grouped into the N, L, C, T or R classes, and that only 2.15% could be classified as other/rare shapes (Fig. 1h). We confirmed this finding using alternative unsupervised classification methods such as principal component analysis followed by hierarchical clustering, or Gaussian mixture modelling to segregate the data into distinct morphological clusters (Fig. 1i). Each of the 5 main morphological classes is populated predominantly by one of N, L, C, T or R-type cells (Fig. 1j). Thus, perhaps surprisingly, the number of different shapes present in the entire data set is low, and is well described by 5 different shapes.

As a first step towards understanding the role of different genes in the control of cell shape, we classified the effects of RNAi on the basis of the population mean of single-cell QMS scores, following the filtering out of normal cells and consolidation of replicable phenotypes. Here, the QMS is a 5-dimensional vector that describes the mean similarity of cells to L, C, T and R shapes, and a PZ score, which is the penetrance of all non-normal shapes in the population before filtering (Supplementary Table S5). Gene QMSs were organized using average linkage hierarchical clustering to describe phenoclusters (Fig. 3). However, although this analysis reveals how different genes broadly affect the morphology of different populations, it does not account for population heterogeneity.

Population of wild-type Kc cells is comprised of 5 shapes

We next sought to leverage single-cell data to determine how genes contribute to the regulation of morphological complexity. We prefer the term complexity to heterogeneity as it better describes the number of shapes that could be considered distinct, versus the total number of shapes in a population—which may represent variations on the same shape. For example, if cells in a population are mostly a single highly variable shape, the heterogeneity of the population is high but the complexity is low. After accounting for differential penetrance of different dsRNAs and identifying dsRNAs

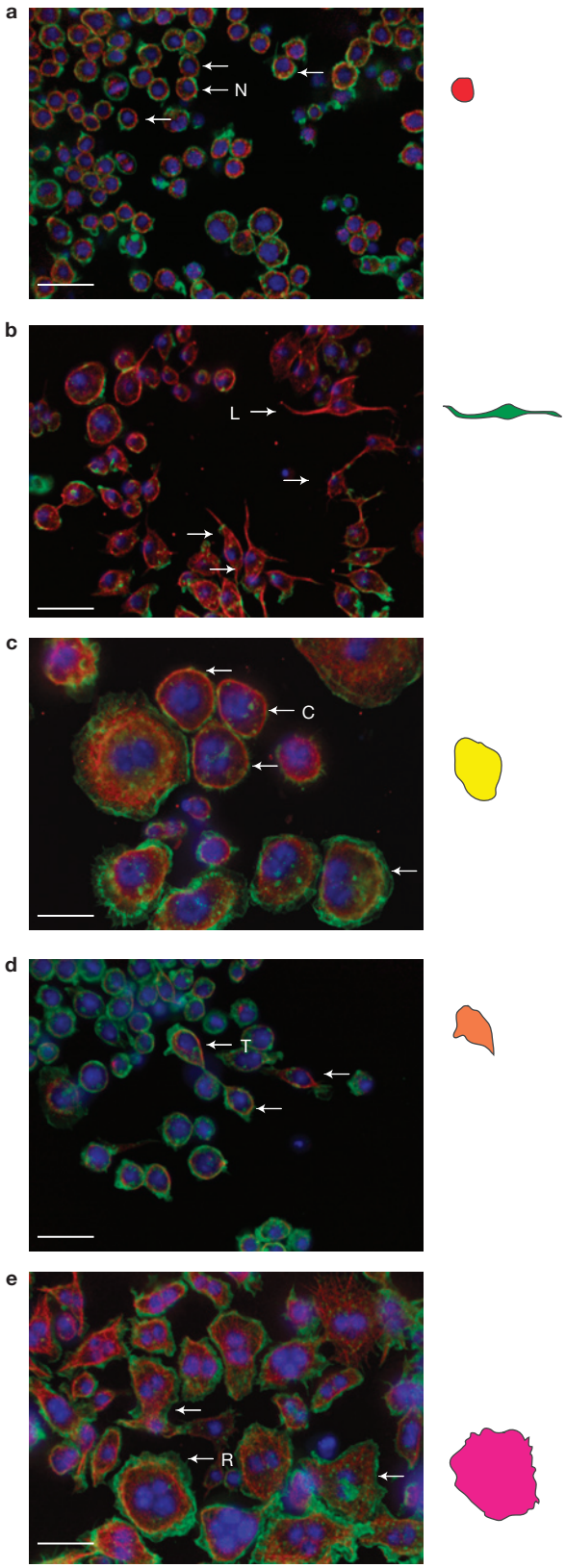

Figure 2 N, L, C, T and R cells. (**a**–**e**) Wild-type (**a**), *Pvr*- (**b**), *CycA*- (**c**), *rl/ERK*- (**d**) or *Rho1*-depleted (**e**) cells were fixed, labelled with DAPI, phalloidin and anti-α-tubulin antibody, and imaged. Arrows denote cells with representative shapes. Coloured cells on the right are traces of representative shapes. Scale bars, 20 μm.

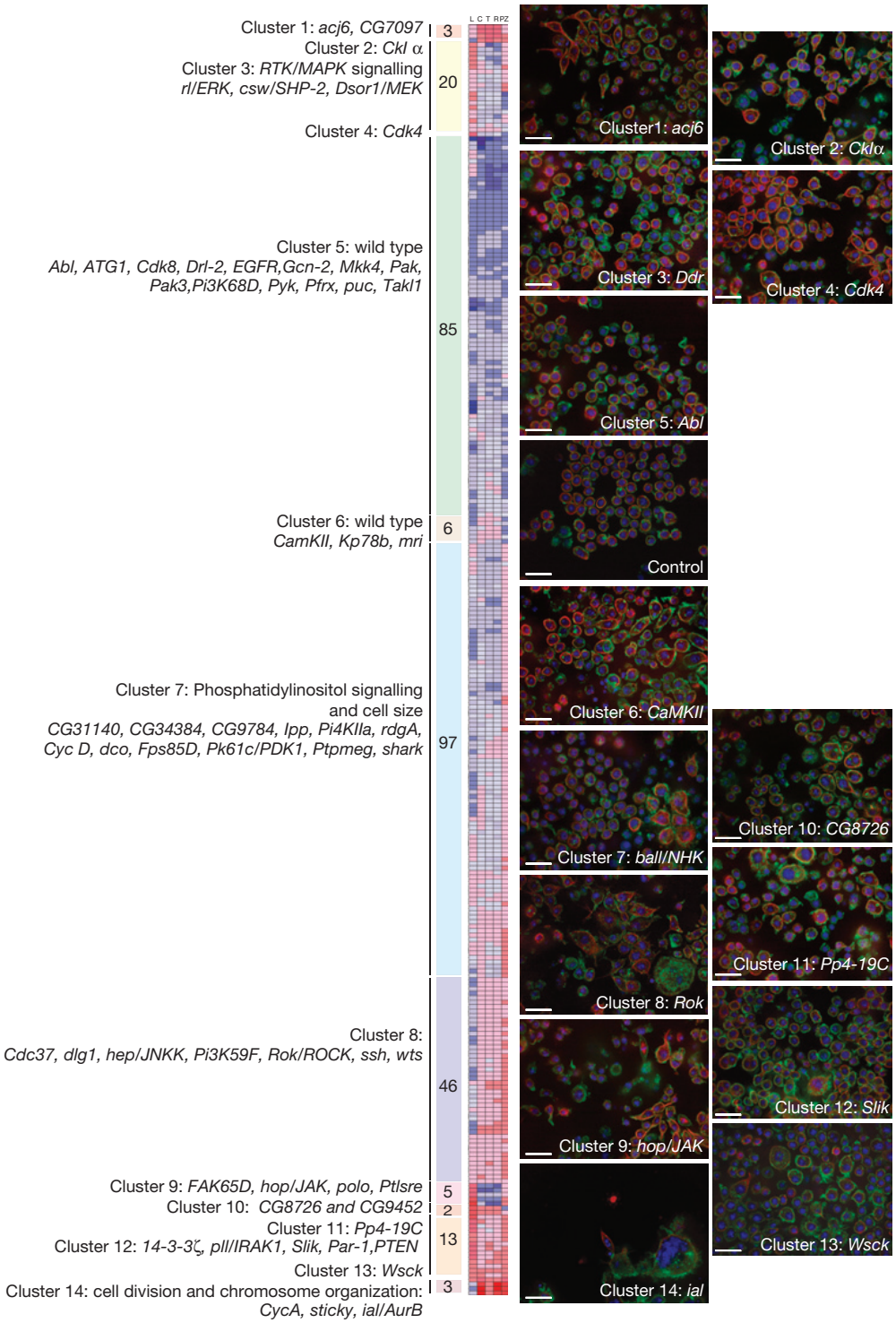

Figure 3 Hierarchical clustering of QMSs. Average linkage clustering of 284 5-feature QMSs comprising L, C, T and R SVM Z-scores as well as PZ scores. Included are experimental conditions, following the RNAi-mediated depletion of 282 genes, RNAi-mediated targeting of *lacZ*, and a signature for control wells. Genes are in the same phenocluster when clustered together at a cutoff distance (an average of uncentred Pearson correlation coefficients) greater than 0.90. At this threshold, we identified 10 different phenoclusters and 4 QMSs (*Cklα*, *Cdk4*, *Pp4-19C* and *Wsck*) that did not cluster with any other gene. The number of genes that comprise each phenocluster is shown in shaded boxes. Some genes that are members of each phenocluster are listed. The largest phenocluster, cluster 5, is composed of 85 ECs that have QMSs that are not significantly different from wild-type cells, even when RNAi penetrance is taken into account (Supplementary Fig. S1 and Supplementary Methods). The mean QMS of 6 ECs in cluster 6 is also essentially wild type. Cluster 3 is significantly enriched[19] in canonical *sevenless* receptor tyrosine kinase (RTK) components ($P = 6.21 \times 10^{-4}$), and cluster 7 is significantly enriched for genes involved in phosphatidylinositol signalling ($p = 1.19 \times 10^{-4}$) and cell size ($p = 1.07 \times 10^{-3}$). A complete list of genes in each phenocluster is included in Supplementary Table S5. Representative image fields from particular phenoclusters are shown. Nuclei are labelled with Hoechst (blue), polymerized actin is labelled with phalloidin (green) and microtubules are labelled with anti-tubulin antibody (red). Scale bars, 20 μm.

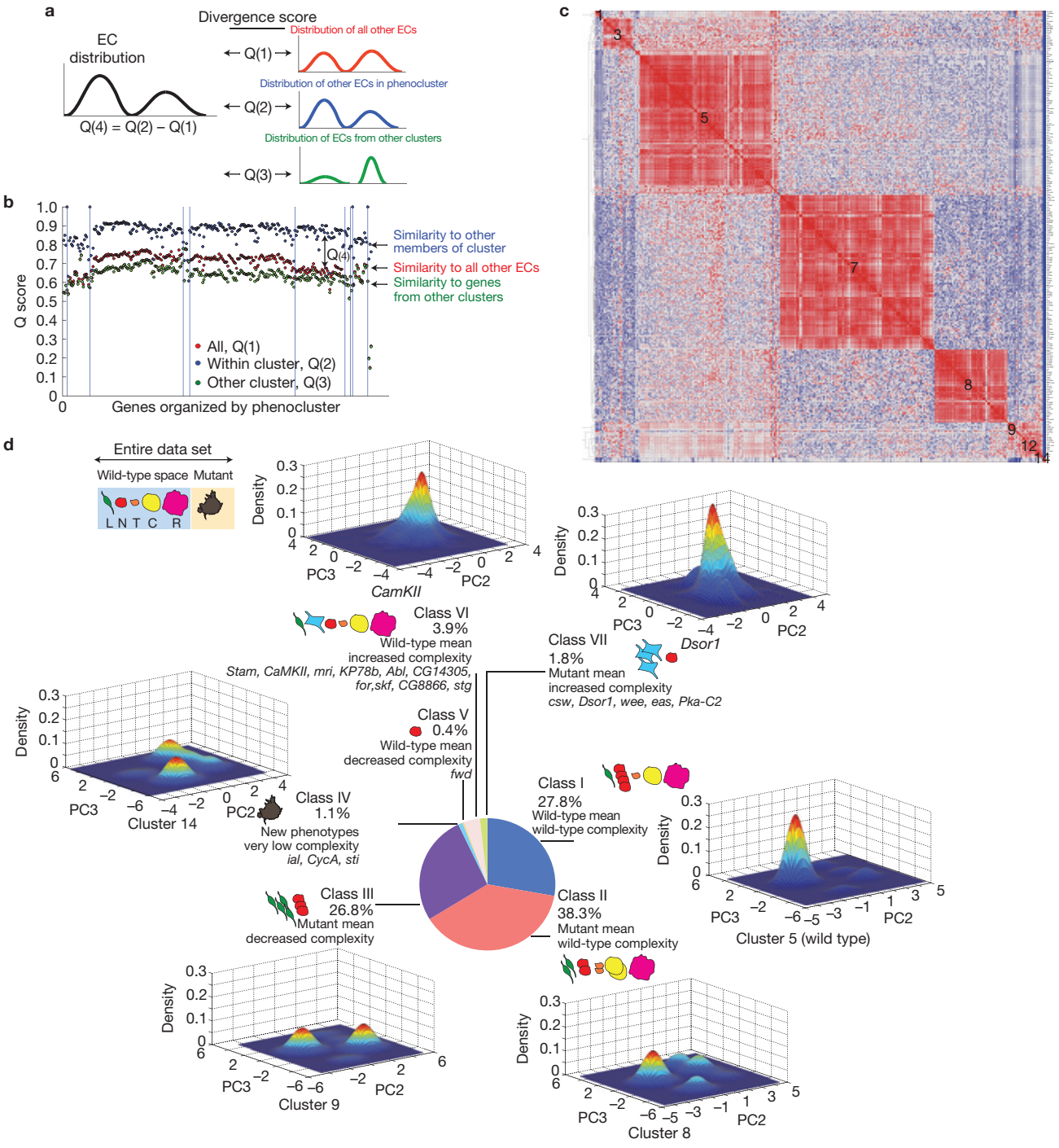

Figure 4 Morphological complexity is a phenotype that can be altered by RNAi. (**a**) Diagram explaining the generation of Q(4) scores. For any given EC, the distribution of cells in morphological space is compared with the distribution of all other ECs, all other ECs in a phenocluster (Fig. 3), and distribution of ECs from all other clusters to generate Q(1), Q(2) and Q(3) scores, respectively. A Q(4) score is the difference between a Q(2) and Q(1) score; the Q(3) score describes the uniqueness of the population. (**b**) The average similarity scores (*y* axis) between a single EC and all ECs in the data set (red), all genes in the same phenocluster (blue), or all ECs in different phenoclusters (green) are shown. Genes are organized by phenocluster (left to right) as in Fig. 3. (**c**) Q(1) similarity scores comparing all ECs to each other. Similarity scores are normalized to range between 0 and 1. Highly similar populations are coloured in red, and dissimilar populations are coloured in blue. Genes are arranged according to their phenocluster membership as described in Fig. 3 (denoted by the numbers on the graph). (**d**) We compared the Q(4) and mean QMSs of different ECs with the scores in wild-type/control EC (for example, *lacZ* RNAi) to describe how genetic inhibition can affect the exploration of cells in the pre-defined shape space. A population can belong to one of seven different categories depending on its mean QMS score and Q(4) scores. For 6 different classes we estimated the Gaussian kernel density for principal component 2 (PC2) and PC3 of populations sampled from different clusters or ECs that are representative of different classes. Each plot represents the probabilities for the occurrence of different shapes and thus describes the morphological space explored by the population. For each graph the cell numbers are as follows: cluster 5, 247,341; cluster 8, 72,196; cluster 9, 7,358; cluster 14, 4,144; *CamKII*, 3,008; *Dsor1*, 3,603.

with reproducible phenotypes (Supplementary Fig. S2), we calculated matrices describing the similarity of the space sampled by an EC to that sampled by all other ECs (Q(1) score; Fig. 4a), as well as to the space sampled by ECs in the same phenocluster (Q(2) score; Fig. 4a). From these scores, we generated a Q(4) by subtracting the Q(1) score from the Q(2) score (Fig. 4a). A Q(4) score describes the complexity of a population. Populations that sample the same morphological space, and thus have the same complexity as other populations, have low Q(4) scores, whereas homogeneous populations with low complexity have high Q(4) scores (Fig. 4b and Supplementary Table S5). By plotting the Q(1) of each EC against all others, we observed that ECs from clusters 5 and 6, or wild-type cells, are very similar to themselves, and to almost all other ECs in the data set (Fig. 4c). Moreover, control (mock-treated) and *lacZ* RNAi ECs have the fourteenth and eighteenth highest Q(1) scores in the data set, respectively, and ECs from clusters 5 and 6 include 40 of the 50 highest ranking ECs in terms of Q(1) scores. Only 3.9% of ECs have a Q(1) score significantly different from the wild type ($P < 0.05$). These data show that wild-type Kc cells have limited morphological complexity that is nearly equivalent to that of the entire data set of RNAi treatments, and are comprised of different shapes that are well represented by the 5 reference shapes. Given that the entire data set is well described by 5 shapes (Fig. 1j), this suggests that gene knockdown most often enriches for shapes that are already present at low levels in populations of wild-type cells.

RNAi most often decreases the number of shapes present in the wild-type population

To describe phenotypes on the basis of the morphological complexity of cellular populations, we classified genes by their Q(4) and mean QMS (Fig. 3) to generate seven different classes (Fig. 4d). Class (i) is comprised of ECs with a wild-type mean and wild-type complexity (unaffected cells, 27.8% of all ECs). Notably, by plotting density estimations of shape frequency in two principal components, we observe that the five different subpopulations in wild-type cells seem to exist as discrete subpopulations (Fig. 4d). Class (ii) consists of ECs that have an abnormal mean but wild-type complexity (38.3%). In this class, RNAi has altered the distribution of cells within subpopulations, but each subpopulation remains represented in the population. For example, in *Par1*-, *Rok*-, *Slik*-, *SAK*- or *trc*-depleted populations there is an enrichment of elongated shapes, resulting in a mean score that is different from wild-type cells, but the complexity of this population is the same as the wild type and the population is also enriched in other shapes. In class (iii) are ECs with decreased morphological complexity where one or more subpopulations has been enriched at significant expense of others (Q(4) Z-score > 1.0), (26.8%). Examples of these include *14-3-3ζ*-, *Pp2B-14D*-, *Pp2A-29B*-, *Dgk*-, *hop/JAK*- or *PTEN*-depleted populations where there is an enrichment of L elongated cells but a marked decrease in other shapes. Class (iv) is a small fraction (1.1%) of ECs with an abnormal mean, significantly decreased heterogeneity and morphologies that are different from those sampled by wild-type populations. Here new shapes have been generated, but the overall complexity (number of total shapes) is less than the wild type. For example, *ial/AurB*- or *sticky/CitronK*-depleted populations are very homogeneous and explore a small region of shape space not sampled by wild-type cells. Class (v) is a single EC, *fwd* RNAi, which has a wild-type mean and decreased complexity. Class (vi) (3.9% of ECs) have a wild-type mean, but are significantly more complex than the wild-type cells. Class (vii) (1.4% of ECs) have an abnormal mean and significantly increased complexity when compared with wild-type cells. Classes (vi) and (vii) include genes such as *Stam*, *CamKII* and *Abl*, which are of particular interest as morphological complexity is increased but these populations are sampling space within that explored by wild-type cells.

Thus, RNAi does not typically lead to the generation of new shapes, but rather alters the distribution of pre-existing subpopulations that exist in wild-type cells. We propose that cellular morphogenesis of Kc cells is a canalized processes[9], where cells can transition between only a limited number of stable shapes, and changes in the distribution occur following RNAi because inhibition of different signalling events prevents the ability of cells to transition from one shape to another, effectively trapping them in one or more stable shapes found in wild-type cells.

Kc cells make switch-like transitions between discrete shapes

We reasoned that cells could transition between stable shapes in one of two ways. Cells could transition between discrete shapes in a switch-like manner, where intermediate forms are highly transient and therefore rarely observed. Alternatively, cells could make continuous transitions where there are a diverse number of morphologies that appear as stable intermediates between shapes. To discriminate between these two possibilities, we calculated a RIFT score (rate of intermediate forms or transitions) for different ECs. The RIFT score quantifies the extent of misclassification by clustering that occurs when populations of cells comprised of 5 shapes from a training pool are mixed *in silico* with an equal number of cells from an EC (Fig. 5a). A high RIFT score indicates the presence of intermediate shapes in the EC, whereas a low RIFT score suggests that there are very few intermediate shapes in the EC. Deficiency of some genes results in high RIFT scores for all shape classes, and thus accumulation of intermediate forms between all shapes (Fig. 5b,c). In other cases the RIFT score is high only for a particular class, meaning that there is an accumulation of forms near a particular shape (Fig. 5b,c). We calculated the maximum RIFT score (Fig. 5d, blue bars) and the average RIFT score (Fig. 5, orange bars) of different populations, although there is typically high correlation between these values (Fig. 5d). For example, we determined the RIFT scores for ten populations with low complexity (high Q(4) score) and ten, including wild-type cells, with high complexity (low Q(4) score), and find that there are few intermediate shapes in wild-type cells. Moreover, RNAi rarely results in an increase in RIFT scores. Thus, the morphogenesis of wild-type Kc cells is both discrete and switch-like in nature. However, RNAi-mediated knockdown of *Stam* (Fig. 5e) and *CamKII* leads to high average RIFT scores (Fig. 5d), and these ECs have many shapes that can be considered continuous. This suggests that the function of these genes is essential for switch-like morphogenesis of Kc cells.

Melanoma cells exhibit discrete, switch-like morphogenesis

We next sought to determine whether our model of discrete, switch-like morphogenesis can be applied to mammalian cells. When cultured on the artificial substrate of rigid tissue culture plastic, metastatic melanoma cells, such as human WM266.4 cells, do not explore shape space in a discrete, switch-like manner akin to Kc cells, as their morphology varies continuously around a single spread morphology (Fig. 6a).

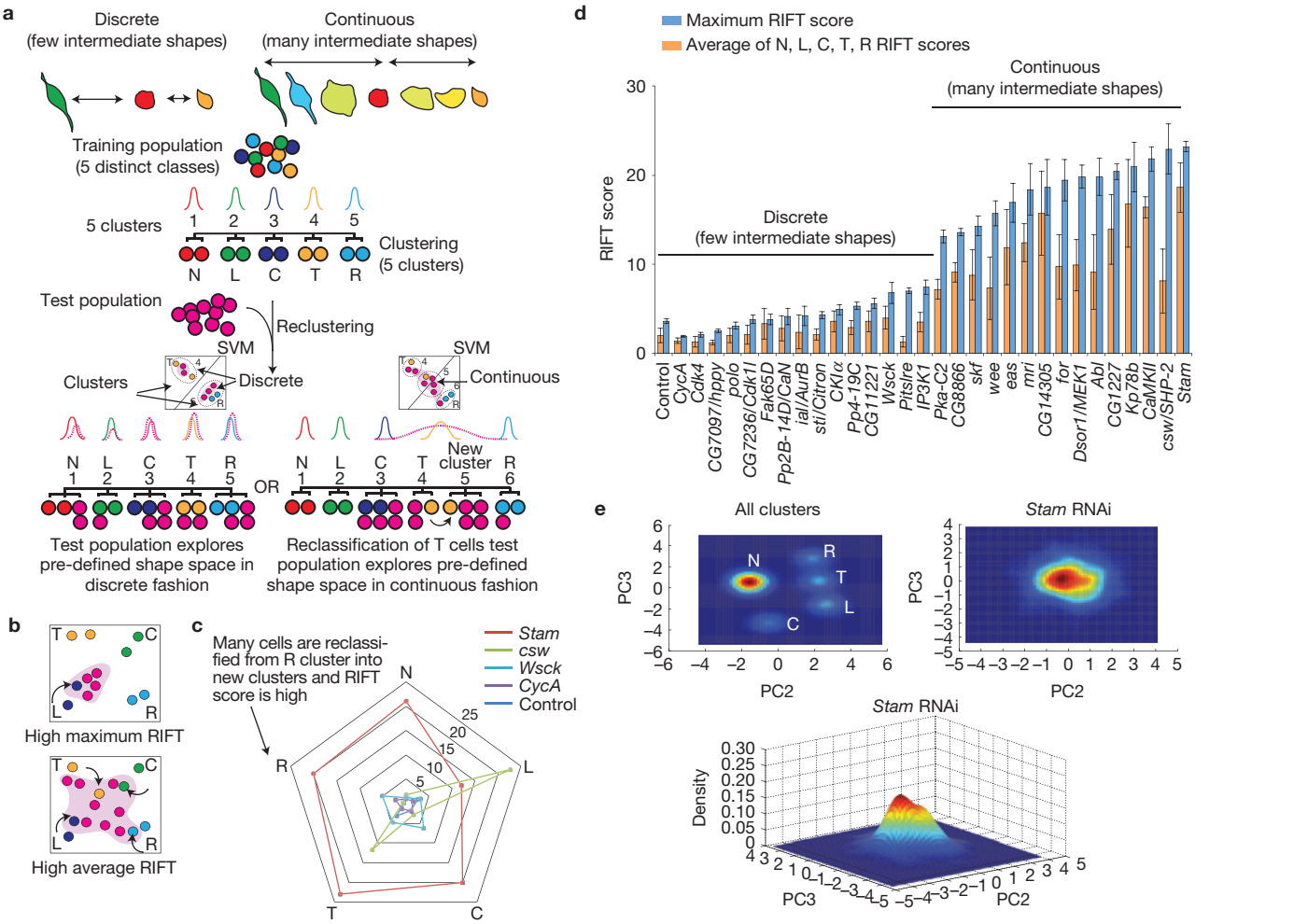

Figure 5 Kc cells exist as discrete subpopulations. (**a**) Methodology for generating the RIFT score. A training pool of 500 cells from 5 reference classes (N, L, C, T and R) is clustered, resulting in 5 different clusters. Test populations are then added to the set, and the entire population is reclustered. If the reclustering results in 5 clusters, the test population is comprised of largely discrete subpopulations and a low RIFT score. However, if after reclustering cells from the training pool are misclassified into new clusters, this indicates the test population has shapes that can be considered intermediate between the reference class, resulting in a high RIFT score. (**b**) ECs can have a high maximum RIFT score (for example, where a high percentage of L cells are assigned into new clusters) and also a high average RIFT score. (**c**) Radar-gram of RIFT scores for *Stam*, *csw/SHP-2*, *Wsck*, *CycA* and control ECs. For ECs such as *Stam*, many cells do not fall into N, L, C, T or R phenoclusters, whereas in *csw/SHP-2* populations, many cells of L or T classes are specifically reclassified. (**d**) The maximum and average RIFT score were calculated for ECs with the 10 highest and 10 lowest Q(4) scores, as well as for 10 normal populations. Error bars represent standard deviation (s.d.) following 10 recalculations (using new populations) of the RIFT score. (**e**) The top panels are a top-down view of the density estimates of randomly sampled cells from all clusters (584,452 cells), or of *Stam*-deficient cells (1,392 cells). The bottom panel is the same density estimate of *Stam*-deficient cells. RIFT scores for all ECs are listed in Supplementary Table S5. PC, principal component.

We extended these observations by quantifying the morphology of WM266.4 cells plated on plastic over time (Fig. 6b). However, when cultured on deformable collagen-I (Col-I) matrices[5,10–12] that have a stiffness comparable to the epidermis, the morphogenesis of WM266.4 cells becomes discrete. On Col-I, WM266.4 cells assume only a rounded (similar to N shape) or an elongated form (similar to L shape; Fig. 6c). Within minutes, WM266.4 cells plated on Col-I make rapid switch-like conversions between the shapes (Fig. 6d). This reveals that WM266.4 melanoma cells can explore shape space in a manner similar to Kc cells when plated on substrates that closely resemble their *in vivo* environment. Kc cells presumably can assume discrete shapes on plastic as they are only weakly adherent. We reasoned that the ability of cells to make switch-like conversions between rounded and elongated shapes could be due to dynamic regulation of protrusive and contractile forces. In support of this notion, knockdown in Kc cells of the Rho kinase *Rok* (ref. 13), a key regulator of cellular contractility, leads to an accumulation of elongated L, as well as large, flat and presumably poorly contractile C and R cells (Fig. 6e); thus, morphological transitions are inhibited in *Rok*-depleted cells. To test the role of contractility in the discrete switch-like morphogenesis of melanoma cells, we incubated WM266.4 cells plated on Col-I in increasing doses of the ROCK inhibitor H1152 and tracked their morphology over time. Inhibition of ROCK led to the accumulation of cells with a collapsed morphology that differs from both elongated forms and rounded forms (Fig. 6f) and makes only small continuous variations in shape (Fig. 6g); switch-like transitions do not occur. Thus, substrate stiffness and cellular contractility are important factors determining the extent to which cells explore shape space in a discrete versus continuous fashion.

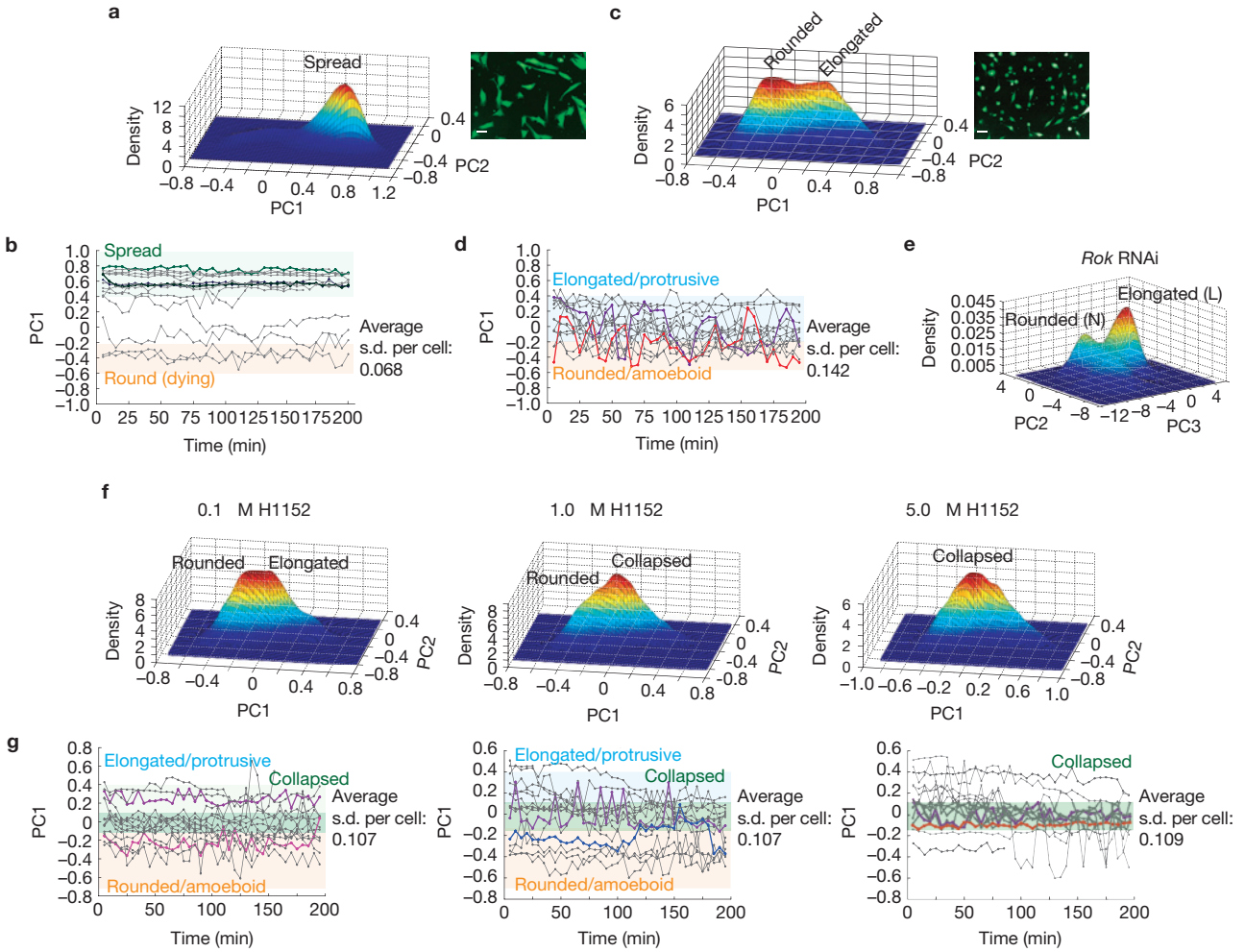

Figure 6 Melanoma cells make switch-like transitions between discrete morphologies on Col-I. (**a**–**d**) WM266.4 cells were plated either on plastic (**a**,**b**) or Col-I (**c**,**d**). The Gaussian kernel density estimate of single-cell morphology in two-dimensional principal component (PC) space is shown in **a**,**c**. In **b**,**d**, the *y* axis corresponds to the PC1 scores of single cells. Elongated cells have high PC1 scores; rounded cells have low PC1 scores and are shaded in orange. Time is described in the *x* axes. (**e**) The Gaussian kernel density estimate of *Rok*-deficient Kc cells (1,808 cells). (**f**,**g**) WM266.4 cells were plated on Col-I, exposed to increasing doses of ROCK inhibitor H1152, and morphology was quantified 6 h later at both a single time point (**f**) or over time (**g**). We calculated the magnitude of morphology fluctuations for individual cells by calculating the s.d. in PC1 scores per cell over time. (**a**,**b**) 29,476 cells, (**c**,**d**) 21,061 cells, (**f**,**g**) 0.1 μM H11552, 36,665 cells; 1.0 μM, 12,605 cells; 5.0 μM, 66,374 cells. Scale bars, 20 μm.

PTEN deficiency promotes bistable populations of rounded and elongated cells

The ratio of rounded to elongated melanoma cells is highly dependent on both environment and genetic background. For example, whereas the ratio of elongated to rounded cells can be as high as 50:50 in the case of WM266.4 cells on Col-I (Fig. 6b), melanoma cells such as A375M2 are mostly rounded[5,11]. However, the specific genes that determine the rounded/elongated ratio are largely unknown. We reasoned that we could leverage the results of our morphological screen to gain insight into the factors regulating the conversion between rounded and elongated shapes of melanoma cells on the basis of two striking observations: first, the shape of WM266.4 cells, which do not express *PTEN*, phenocopies that of *PTEN*-deficiency in *Drosophila* (high ratio of elongated to rounded; Fig. 7a) and second, *hop/JAK*-deficient Kc populations are also heavily enriched in elongated cells at the expense of other shapes, which is consistent with our recent finding that JAK1 promotes contractility in melanoma cells[12]. In fact, *PTEN* and *hop/JAK* RNAi results in the seventh and eighth highest L scores respectively in the entire Kc data set, and both ECs have high Q(4) scores demonstrating that they explore only limited regions of shape space when compared with wild-type Kc cells.

To determine whether *PTEN* status correlates with the ratio of rounded/elongated shapes in cell populations, we plated 22 melanoma cell lines (10 *PTEN* null and 12 *PTEN* wt) on Col-I gel and assessed the ratio of rounded/elongated cells. *PTEN* loss strongly correlates with an increase in the proportion of elongated to rounded cells (Fig. 7b and Supplementary Table S6). Furthermore, depletion of *PTEN* expression in PTEN wild-type mouse 4599.1 or human A375p melanoma cells (Fig. 7c and Supplementary Fig. S3a,b) by independent short hairpin shRNAs (shRNAs) increases the number of elongated cells and increases phosphorylated Akt levels (Fig. 7d and Supplementary Fig. S3a,b). We confirmed the effect of *PTEN* shRNA using quantitative readouts of morphology (Fig. 7e), and show that *PTEN* depletion increases the number of elongated cells but does not generate any other shapes. Importantly, re-expression of *PTEN* in the WM266.4 *PTEN*-null melanoma cells increases the number of rounded cells at

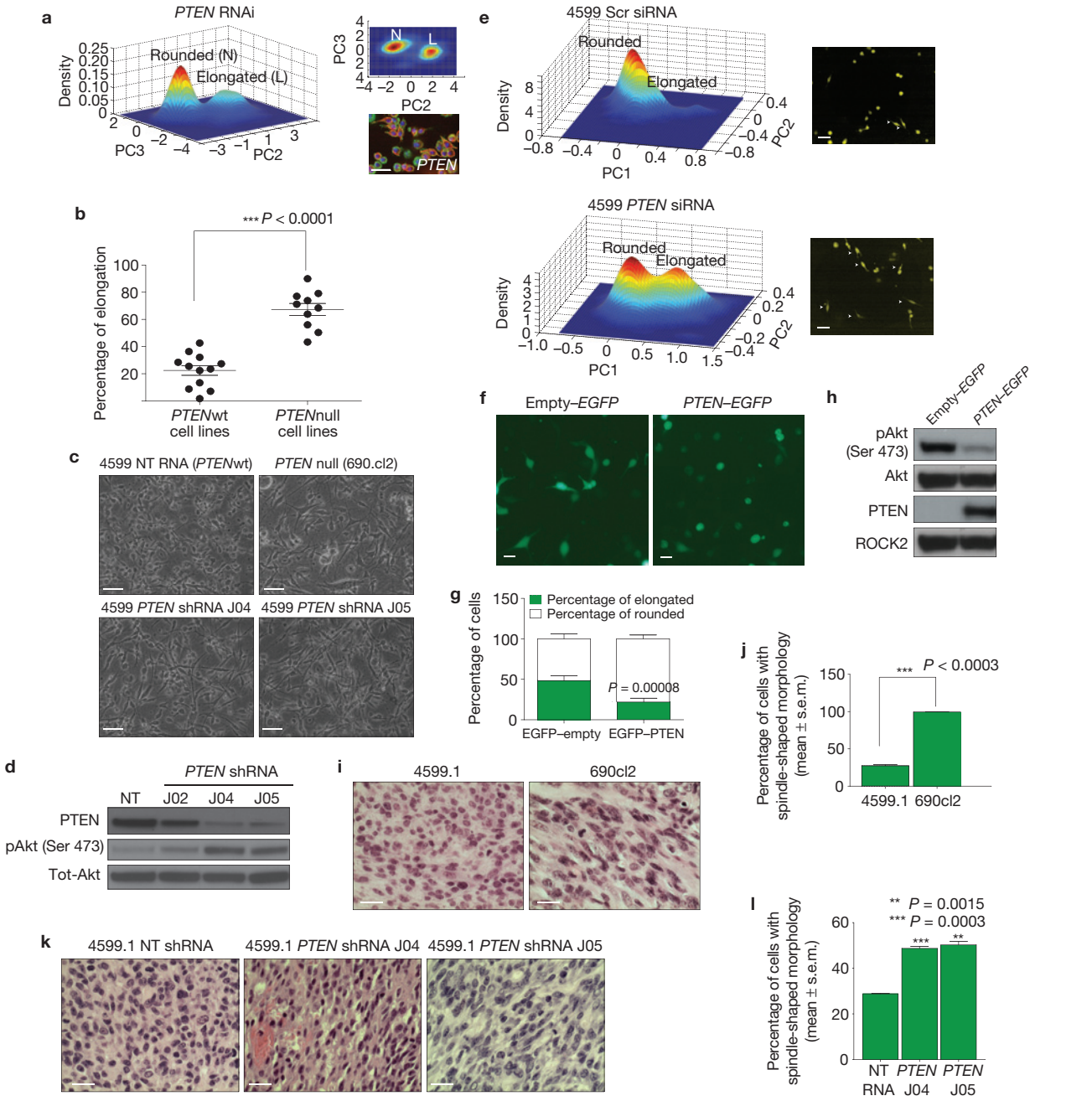

Figure 7 Loss of PTEN alters the exploration of shape space. (**a**) Density estimation of *PTEN*-deficient Kc cells (1,831 cells). The upper right panel is a side view of the density estimate shown and the lower right panel shows Kc cells stained with DAPI (blue), phalloidin (green) and anti-tubulin antibody (red) following *PTEN* RNAi. PC, principal component. Scale bar, 20 μm. (**b**) Percentage of elongated cells on the top of thick Col-I (mean ± s.e.m.); 250 cells over 5 fields of view per cell line; $n = 12$ *PTEN*wt and 10 *PTEN*null cell lines (Supplementary Table S6); Student's *t*-test was used to generate the *P* value. (**c**) Images of 4599.1 cells on thick Col-I; *PTEN* was stably depleted by two different shRNAs (J04 and J05). NT is a non-targeting shRNA. 690.cl2 cells are shown for comparison. Scale bars, 50 μm. (**d**) Representative immunoblot of pSer473 AKT, PTEN and tot-AKT in NT- and *PTEN*-shRNA-expressing 4599.1 cells. (**e**) Density estimation of 4599.1 cells treated with Scr (scrambled) siRNA (upper panel; 1,326 cells) or *PTEN* RNAi (lower panel; 305 cells) cultured on thick Col-I. The right panels show images of live cells; arrows denote elongated cells. Scale bars, 20 μm. (**f**) Representative images of WM266.4 cells transfected with Empty–*EGFP* or *PTEN*–*EGFP* on the top of thick Col-I. Scale bars, 50 μm. (**g**) Proportion of elongation/rounded cells following expression of Empty–*EGFP*- or *PTEN*–*EGFP*-expressing cells (mean ± s.d.); 200 cells per experiment, $n = 3$ experiments; Student's *t*-test was used to generate the *P* value. (**h**) Levels of PTEN, pAKT, total (Tot) AKT and ROCK2 (loading control) in Empty–*EGFP*- and *PTEN*–*EGFP*- transfected WM266.4 cells. (**i**) Representative images of 690cl2 and 4599.1 tumour sections. Scale bars, 20 μm. (**j**) The number of elongated cells in the body of either 4599.1 or 690cl2 tumours is expressed as a percentage of the total number of cells counted per tumour (mean ± s.e.m.); 200 cells per field assessed in 5 fields of view per tumour; $n = 4$ 4599.1 and 4 690cl2 tumours; statistical analysis was done using Student's *t*-test. (**k**) Representative images of tumour sections derived from NT- or *PTEN*-shRNA-expressing 4599.1 cells. Scale bars, 100 μm. (**l**) Percentage of elongated cells in the body of the tumour following control (NT) or *PTEN* RNAi. (Mean ± s.e.m.); 200 cells per field assessed in 5 fields of view over $n = 4$ NT RNAi and 4 *PTEN* RNAi tumours. Uncropped images of blots/gels are shown in Supplementary Fig. S5.

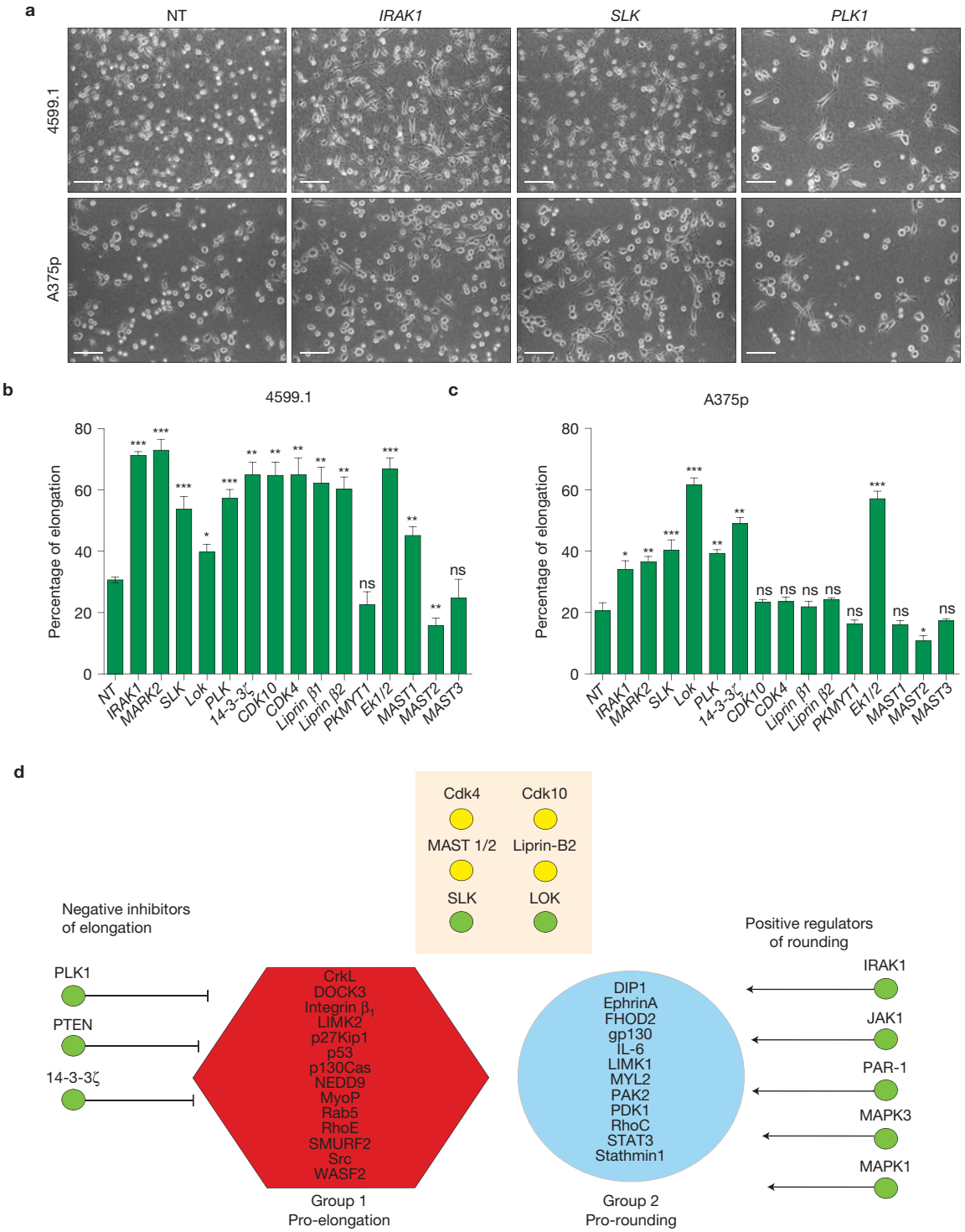

Figure 8 A conserved set of genes promotes rounding in *Drosophila*, mouse and human cells. (**a**) Mouse 4599.1 (upper panels) and human A375p (lower panels) metastatic melanoma cells plated on Col-I following RNAi-mediated gene knockdown of *IRAK1*, *SLK* and *PLK1*. NT is non-targeting RNAi. Scale bars, 50 μm. (**b**) Percentage of elongated cells following knockdown of 15 mouse genes in 4599.1 cells plated on Col-I (mean ± s.e.m.). (**c**) Percentage of elongated cells following knockdown of 15 human genes in A375p cells plated on Col-I (mean ± s.e.m.). In **b**,**c**, 250 cells over $n = 3$ experiments. Student's *t*-test was used to generate the *P* values. The asterisks denote the level of significance: $^{*}P < 0.05$, $^{**}P < 0.001$, $^{***}P < 0.0001$. (**d**) Network analysis. We calculated the proximity of different proteins identified in our screen to either pro-elongation proteins or pro-rounding proteins in the protein–protein interaction space. The length of the arrow is scaled to a *Z*-score that describes the significance of this proximity compared with random proteins, where longer arrows are less significant and thus further away in the protein–protein interaction space. As inhibition of all proteins here results in elongated shapes, we could classify different proteins as negative regulators of elongation or positive regulators of rounding. Proteins in the pale orange rectangle are not significantly close to either group. Green circles indicate that gene depletion increases the percentage of elongation in mouse and human cells; the yellow circles indicate that gene depletion increases the percentage of elongation in mouse cells.

the expense of elongated cells (Fig. 7f–h). To determine whether *PTEN* regulates the exploration of shape space *in vivo* we used orthotopic implantation of melanoma cells from 4599.1 cells (*PTEN* wt), 4599.1 cell populations stably expressing two different *PTEN* shRNAs or 690cl2 (*PTEN* null) into the dermis of NOD SCID mice and assessed the shape in haematoxylin and eosin-stained sections[5,14]. Tumour cells arising from injection of 4599.1 cells are predominantly rounded (Fig. 6i,j) whereas tumour cells arising from injection of 690cl2 cells (Fig. 7i,j), or 4599.1 cells in which *PTEN* had been knocked down, are markedly elongated (Fig. 7k,l and Supplementary Fig. S4). Thus, *PTEN* loss induces elongation cells in tissue culture and *in vivo*.

A conserved class of genes that promote cell rounding

Given that depletion of *PTEN* and *JAK* results in bistable populations of elongated and rounded cells in *Drosophila*, mouse and human cells, we sought to determine whether other genes identified in the *Drosophila* screen are conserved regulators of morphogenesis. We selected genes whose depletion in *Drosophila* cells results in a significant increase in the number of elongated cells, or the magnitude of their L scores (Supplementary Table S7). Genes were further prioritized if their inhibition resulted in low complexity (high Q(4) score), and thus were enriched in L cells at the expense of other shapes. For example, whereas *PLK1*- and *14-3-3ζ*-depleted *Drosophila* cell populations are comprised almost exclusively of rounded and elongated shapes, *Slik*-depleted populations have a high (L) score, but are also enriched in other subpopulations. We tested only genes where we could identify human homologues; in some cases this required targeting of multiple genes (for example, *MAST1*, *MAST2* and *MAST3* are homologues of *Drosophila CG6498*; Supplementary Table S7). Using short interfering RNA (siRNA) pools (Supplementary Tables S8 and S9) we depleted 15 different homologues of 11 different *Drosophila* genes in 4599.1 mouse and A375p human melanoma cells. When cultured in starving conditions on a thick Col-I matrix, both 4599.1 and A375p convert to elongated cells at a low frequency; we scored populations on the basis of whether siRNA knockdown increases the frequency of elongation (Fig. 8a). RNAi-mediated knockdown of 12/15 genes in mouse (Fig. 8b) and 7/15 genes human cells results in significant increases in elongation that phenocopy their depletion in *Drosophila* (Supplementary Table S7). For example, depletion of mouse and human *IRAK1*, *PLK1*, *PTEN*, *ERK1* and *ERK2* led to marked increases in the numbers of elongated cells (Fig. 8b). That 10/11 *Drosophila* genes whose depletion results in a high L score can be validated as regulators of cell rounding in at least one mammalian metatstatic melanoma tumour line, in addition to *PTEN* and *JAK*, highlights the ability of our RNAi screen to identify genes that have relevance to disease progression.

Classifying genes as protrusion antagonists or contractility agonists

Towards gaining systems-level mechanistic insights into how different validated genes identified in our screen regulate cell shape, we performed a network analysis to determine the proximity of different proteins in network space to regulators of protrusiveness or contractility. Proteins previously implicated in controlling cell shape were classified into either pro-elongation or pro-contractility groups[15]. We then calculated the average number of edges that separated proteins identified in our screen from proteins in either previously assigned group in protein–protein interaction networks, and judged the significance of this distance compared with that between other random proteins. Proteins such as JAK1 and IRAK1 are significantly closer in protein–protein interaction space to the pro-contractility group, whereas PTEN and $14-3-3\zeta$ are closer to the pro-elongation group (Fig. 8d). Given that depletion of all these genes results in similar elongated shapes, we conclude that JAK1 and IRAK1 promote rounding, but $14-3-3\zeta$ and PTEN negatively inhibit protrusion. This unbiased network is consistent with our previous observation that JAK1 upregulates contractility by activation of the STAT3 transcription factor[12], and that PTEN is a negative regulator of PI(3)K that acts to promote protrusion in multiple other cell types[16]. Interestingly, proteins such as ERK1/2 have not been previously associated with an upregulation of contractility. Thus, this analysis provides hypotheses for other poorly characterized genes.

DISCUSSION

By implementing methods to quantify mean morphology, complexity and presence of intermediate forms in cell populations in an RNAi screen of *Drosophila* Kc cells, we propose that cells can explore shape space in a discrete, switch-like manner. Using live-cell imaging methods in combination with morphological quantification, we demonstrate that this type of morphogenesis is not limited to *Drosophila* haemocytes, and that metastatic melanoma cells explore shape space in a similar fashion when plated on substrates that mimic their *in vivo* environment. We propose that many cell types will also exhibit discrete switch-like morphogenesis *in vivo*, and that it has been the long-standing use of rigid tissue culture plastic that has obscured this aspect of cell shape control. Although the model that cells can be constrained to specific regions of shape space is potentially counter-intuitive given the highly plastic nature of cell shape and the ability of cells to adopt radically diverse shapes, discrete morphogenesis or morphological canalization of single cells[17] is consistent with the idea that signalling networks are dynamic systems that can exist in a limited number of stable states, or attractors[18].

That systematic gene inhibition by RNAi can alter shape space and/or alter the mode of morphogenesis from switch-like to continuous (for example, *Stam* RNAi), suggests that signalling networks have evolved to couple the topology of their shape space, and how they explore it, to environmental conditions. Metastatic cancer cells may have re-engineered regulatory networks that uncouple the control of morphogenesis from environmental cues, which would otherwise dictate the number of shapes they can assume and how they convert between these shapes. In the case of *PTEN*, it is tempting to speculate that loss of *PTEN* may promote the adoption of a bistable state where rounded and elongated forms are present in high numbers. By increasing the frequency of rounded and elongated cells this would provide metastatic cells with a survival advantage that is otherwise not gained by adopting only a single shape, or being highly plastic. □

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary Information is available in the online version of the paper

ACKNOWLEDGEMENTS

We are indebted to the *Drosophila* RNAi Screening Center staff at Harvard Medical School for their invaluable assistance. We especially thank I. Flockhart for assistance with data management. We thank J. Wang, X. Zhou and P. Bradley for their initial involvement in this study. We are grateful to N. Dhomen and R. Marais for melanoma cell lines. Work was financially supported in part by NCI grants (Grants R01CA121225 and U54CA149196) to S.T.C.W. and CRUK grants to C.B. (Grant 13478) and C.J.M. (Grant C107/A10433). N.P. is an Investigator of the Howard Hughes Medical Institute. A.S. is a Marie Curie Intra-European Fellow. C.J.M. is a Gibb Life Fellow of CRUK. C.B. is a Research Career Development Fellow of the Wellcome Trust.

AUTHOR CONTRIBUTIONS

Z.Y. performed the bulk of statistical analysis of RNAi screening data and wrote the Supplementary Note. A.S. designed and performed all RNAi and cell line characterization experiments in mouse and human melanoma cells and contributed to writing of the manuscript. H.S. performed the analysis of live-cell melanoma cell imaging experiments and contributed to visualization of statistical results. A.M. performed all mouse work. X.X. and F.L. performed processing of images generated in *Drosophila* RNAi screen. M.A.G. and L.E. performed experiments describing penetrance of effects of different dsRNAs. A.R.B. contributed to writing and editing of the manuscript and the Supplementary Note. N.P. participated in the initial design of the study. S.T.C.W. coordinated image processing and statistical analysis. C.J.M. participated in design of melanoma experiments and contributed to writing the manuscript. C.B. participated in the design of experiments and statistical analysis, performed the *Drosophila* RNAi screen, performed the live-cell imaging assays, coordinated experimental and computational analysis, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at www.nature.com/doifinder/10.1038/ncb2764
Reprints and permissions information is available online at www.nature.com/reprints

1. Thiery, J.P., Acloque, H., Huang, R. Y. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139,** 871–890 (2009).
2. Keren, K. *et al*. Mechanism of shape determination in motile cells. *Nature* **453,** 475–480 (2008).
3. Mogilner, A. & Keren, K. The shape of motile cells. *Curr. Biol.* **19,** R762–R771 (2009).
4. Guarino, M., Rubino, B. & Ballabio, G. The role of epithelial-mesenchymal transition in cancer pathology. *Pathology* **39,** 305–318 (2007).
5. Sanz-Moreno, V. *et al*. Rac activation and inactivation control plasticity of tumour cell movement. *Cell* **135,** 510–523 (2008).
6. Wolf, K. *et al*. Compensation mechanism in tumour cell migration: mesenchymal-amoeboid transition after blocking of pericellular proteolysis. *J. Cell Biol.* **160,** 267–277 (2003).
7. Bakal, C., Aach, J., Church, G. & Perrimon, N. Quantitative morphological signatures define local signalling networks regulating cell morphology. *Science* **316,** 1753–1756 (2007).
8. Yin, Z. *et al*. Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens. *BMC Bioinformatics* **9,** 264 (2008).
9. Waddington, C. H. *The Strategy of Genes* (Allen Unwin, 1957).
10. Gadea, G., Sanz-Moreno, V., Self, A., Godi, A. & Marshall, C. J. DOCK10-mediated Cdc42 activation is necessary for amoeboid invasion of melanoma cells. *Curr. Biol.* **18,** 1456–1465 (2008).
11. Sahai, E. & Marshall, C. J. Differing modes of tumour cell invasion have distinct requirements for Rho/ROCK signalling and extracellular proteolysis. *Nat. Cell Biol.* **5,** 711–719 (2003).
12. Sanz-Moreno, V. *et al*. ROCK and JAK1 signalling cooperate to control actomyosin contractility in tumour cells and stroma. *Cancer Cell* **20,** 229–245 (2011).
13. Tan, C., Stronach, B. & Perrimon, N. Roles of myosin phosphatase during Drosophila development. *Development* **130,** 671–681 (2003).
14. Viros, A. *et al*. Improving melanoma classification by integrating genetic and morphologic features. *PLoS Med.* **5,** e120 (2008).
15. Sanz-Moreno, V. & Marshall, C. J. The plasticity of cytoskeletal dynamics underlying neoplastic cell migration. *Curr. Opin. Cell Biol.* **22,** 690–696 (2010).
16. Ridley, A. J. *et al*. Cell migration: integrating signals from front to back. *Science* **302,** 1704–1709 (2003).
17. Waddington, C. H. Canalization of development and genetic assimilation of acquired characters. *Nature* **183,** 1654–1655 (1959).
18. Kauffman, S. A. *The Origins of Order. Self-Organization and Selection in Evolution* (Oxford Univ. Press, 1993).
19. Dhomen, N. *et al*. Oncogenic Braf induces melanocyte senescence and melanoma in mice. *Cancer Cell* **15,** 294–303 (2009).

METHODS

Cell culture, plasmids and RNAi transfection. A375p and A375M2 cells were from R. Hynes (Howard Hughes Medical Institute, Massachusetts Institute of Technology, USA). WM266.4, LU1205, WM1361 and WM1366 cells were from R. Marais (Paterson Institute, Manchester, UK), SKMEL24 cells were from ATCC, and WM239 cells were from W. Cruz and R. Kerbel (Sunnybrook Health Science Centre, Toronto, Canada). 690cl2, 7491cl1, 690cl5, 690cl6, 4434cl2, 5537, 1840cl5, 5021cl6, 2225, 5017, A061 and 4599 cells were generated by N. Dhomen and R. Marais (Paterson Institute, Manchester, USA) either from tumours arising in the Braf V600E mouse model[20] or from tumours arising from the BrafV600E PTEN-null mouse melanoma tumour model[21]. We have generated the AM997-2 and AM993-1 lines from the BRAFV600E/*PTEN* null mouse melanoma tumour model. All of the cells were maintained in DMEM containing 10% fetal calf serum. Human GFP–PTEN was from Addgene (Plasmid #13039). Plasmid transfection was performed with Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. The On-TARGETplus siRNAs against human PTEN and the On-TARGETplus set of 4 siRNAs against mouse *PTEN* were from Dharmacon (Supplementary Table S8). For other siRNA experiments targeting genes other than *PTEN* when used On-TARGETplus pools (Dharmacon). Transfection was performed with RNAimax Lipofectamine (Invitrogen) according to the manufacturer's protocol.

Drosophila Kc167 cells were cultured in Schneider's insect media (Invitrogen), 10% fetal bovine serum (Invitrogen) and penicillin/streptomycin (Gibco). All dsRNA experiments were performed using the bathing method as described at www.flyrnai.org, and cells were fixed following five days of RNAi.

PTEN stable knockdown using shRNA. A set of four pGIPZ-mouse PTEN shRNA clones (J02, J03, J04 and J05) and a pGIPZ-non-silencing shRNA were from Open Biosystems. Lentiviral DNA was generated according to the manufacturer's instructions; 4599.1 (2×10^5 cells) mouse melanoma cells were infected with three PTEN shRNA clones and the non-silencing shRNA control for 24 h; cells were then cultured in 2 μg ml^{-1} puromycin for 2 days to enrich for the transduced cells.

Verification of mRNA depletion. To verify messenger RNA depletion in mouse and human melanoma cells, total cellular RNA was isolated from RNAi or non-targeting sequence-transfected cells using RNAeasy Mini kit (Qiagen) according to the manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) amplifications were performed using the Brilliant II SYBR Green qRT-PCR Master Mix kit (Agilent). PCR was performed in an Applied Biosystems 7900 HT Fast Real-Time PCR cycler. Fluorescence data were analysed using Applied Biosystems SDS software. The percentage of mRNA depletion was established as 100 − (the ratio of the quantity of mRNA in the RNAi condition normalized to B2microglobulin and the quantity of mRNA in the non-targeting condition normalized to B2microglobulin × 100).

Cell culture on thick layer of Col-I and time-lapse phase-contrast microscopy. Fibrillar bovine dermal Col-I was prepared at a 1.7 mg ml^{-1} dilution in DMEM according to the manufacturer's protocol (PureCol, Advanced Biomatrix), and 50 μl was placed in wells of 96-well plates, 300 μl was placed in wells of 24-well plates and 2 ml was placed in wells of 6-well plates. Cells were seeded on top of Col-I in medium containing 10% serum and allowed to adhere for 2–3 h, and medium was changed to 0% serum for 5–16 h then cells were imaged. *PTEN*-expressing WM266.4 cells were imaged after 4 h of serum starvation. A cell was considered elongated when its longest dimension was twice the shortest and when it showed at least one protrusion[5,12]. For RNAi experiments on Col-I, 48 h after transfection, cells were plated on thick Col-I in medium containing 10% serum and allowed to adhere for 2–3 h, and medium was changed to 0% serum for 16 h then cells were either imaged or lysed. WM266.4 cells were treated with the ROCK inhibitor H1152 after being transferred to Col-I.

Immunofluorescence microscopy. Following five days of incubation with individual dsRNAs, cells were fixed at room temperature in 4% UltraPure EM grade paraformaldehyde (Polysciences) in phosphate-buffered saline (PBS; Gibco) for 15 min. Cells were washed three times in PBS and then permeabilized in 0.1% Triton X-100/PBS solution for 5 min. Following three washes in PBS, cells were blocked for 1 h in 0.5% bovine serum albumin (BSA) (Sigma)/0.02% glycine/PBS solution at room temperature. Incubation with mouse anti-bovine-α-tubulin (A11126, Molecular Probes) diluted 1:1,000 was performed overnight in 20 μl 0.5%BSA/0.02% glycine/PBS at 4 °C. Cells were washed three times in PBS and then incubated with a 1:400 dilution of OregonGreen phalloidin (O7466, Molecular Probes) and a 1:500 dilution of AlexaFluor 647-labelled F(ab')2 fragment of goat anti-mouse IgG (A21237, Molecular Probes) in 20 μl of 0.5%BSA/0.02% glycine/PBS for 1 h at room temperature. Cells were washed once in PBS, incubated in 1:500 dilution of DAPI (4',6-diamidino-2-phenylindole, dihydrochloride, Molecular Probes; D1306, Molecular Probes)/PBS solution for 5 min and then washed one final time in PBS. For anti-ERK and anti-AKT staining of *Drosophila* cells, the staining procedure was identical except that the primary was either a 1:200 dilution of anti-ERK (4695, Cell Signaling Technology) or anti-Akt (4691, Cell Signaling Technology) antibody and the secondary was a 1:500 dilution of AlexaFluor 647-labelled F(ab')2 fragment of goat anti-mouse IgG (A21246, Molecular Probes).

Imaging. Imaging of *Drosophila* Kc167 was performed on the Opera QEHS (PerkinElmer) using a ×60 water-immersion objective. In addition to cells treated with different dsRNAs targeting kinases and phosphatases, we imaged 1,019 control wells with cells that had been either mock-transfected or transfected with dsRNAs targeting *lacZ*. Sixteen fields for each dsRNA were acquired in triplicate or quadruplicate. Live-cell imaging experiments of WM266.4 cells ± H1152, or 4599 cells ± *PTEN* shRNA were also performed on the Opera QEHS using a ×20 air objective. For live-cell experiments, melanoma cells were pre-labelled with CellTracker Orange CMRA (C34551, Molecular Probes) where the final concentration was 5 μM.

Immunoblotting. Whole-cell extracts from cells on thick Col-I gel were collected in Laemmli sample buffer and sonicated for 15 s before centrifugation. Lysates were fractionated by SDS–PAGE and transferred to nitrocellulose filters. Antibodies were as follows: rabbit monoclonal anti-PTEN (138G6), rabbit monoclonal anti-phospho-AKT (Ser 473; D9E), mouse monoclonal anti-AKT (pan) (40D4); all from Cell Signalling Technology. All primary antibodies were used at a dilution of 1:500. Secondary antibodies were ECL sheep anti-mouse IgG, horseradish peroxidase (NA931V, GE), or ECL donkey anti-rabbit IgG horseradish peroxidase (GE) and were used at a final dilution of 1:10,000. Detection was performed with the ECL Plus System (NA934V, GE Healthcare).

Xenografts. All animal procedures were approved by the Animal Ethics Committees of the Institute of Cancer Research in accordance with National Home Office regulations under the Animals (Scientific Procedures) Act 1986. 690.cl2 cells, 4599 cells and 4599 cells infected with PTEN shRNA (clone J04 and clone J05) or the non-targeting shRNA were injected intra-dermally into the lateral flanks of 6–8-week-old female NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) mice. Tumours were allowed to develop for a period of 24 days, after which the animals were euthanized, and tumours were excised, fixed in 4% buffered formalin overnight and embedded in paraffin. Sections (3 μm) were cut and stained with haematoxylin and eosin to enable analysis of the tumour samples. Cell shape was assessed in the body of the tumour on the haematoxylin and eosin-stained tumour samples by counting the number of round or elongated cells in 5 fields of view per tumour sample; a minimum of 200 cells were counted per field of view and for each genotype 4 individual tumours were assessed.

RNAi sequences. Sequences for all mouse and human RNAi reagents are listed in Supplementary Table S8. All *Drosophila* RNAi sequences are available at www.flybase.org.

RNAi screen data and code availability. *Drosophila* RNAi screening data has been deposited at PubChem (DRSC-P74), and is also available at flybase.org. All code is available at www.cbi-tmhs.org/GCellIQ/NCB.

20. Dankort, D. *et al.* Braf(V600E) cooperates with Pten loss to induce metastatic melanoma. *Nat. Genet.* **41,** 544–552 (2009).
21. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

DOI: 10.1038/ncb2764

nature cell biology

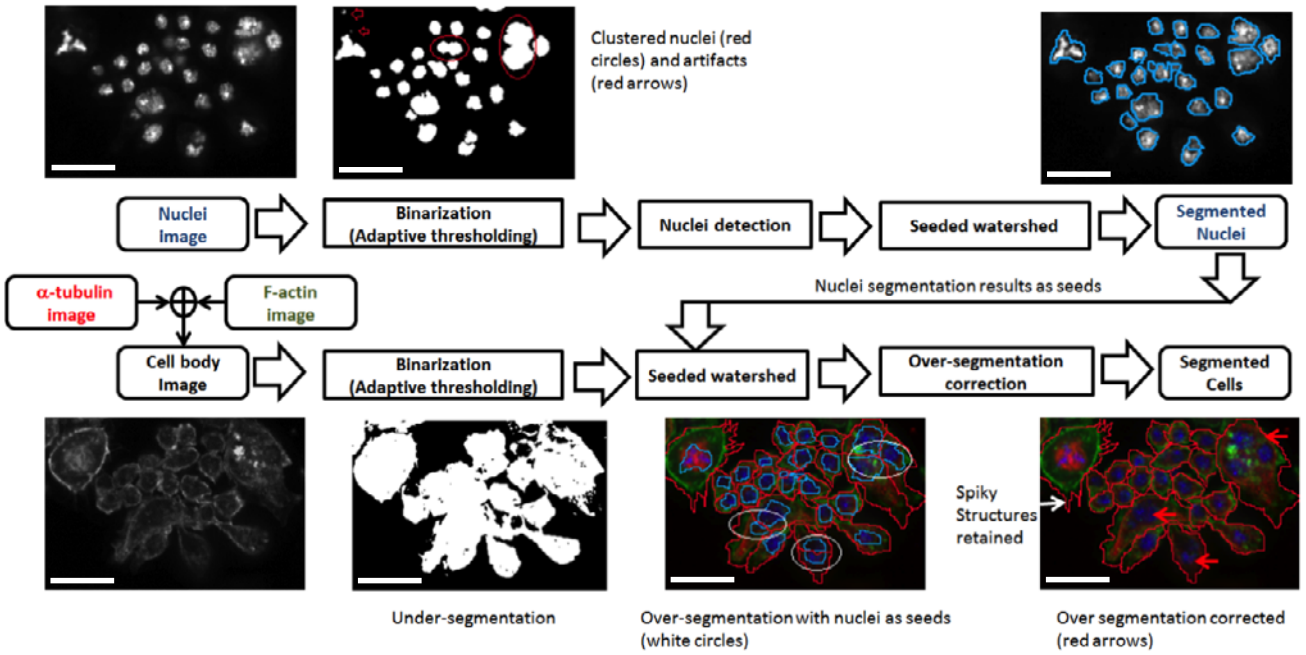

Figure S1 Workflow for image analysis of RNAi screening data. Nuclei were first segmented from the DAPI channel, and these segments were later integrated with a "cell body image" combined from the α-tubulin and F-actin channels. Scale bars, 20 μm.

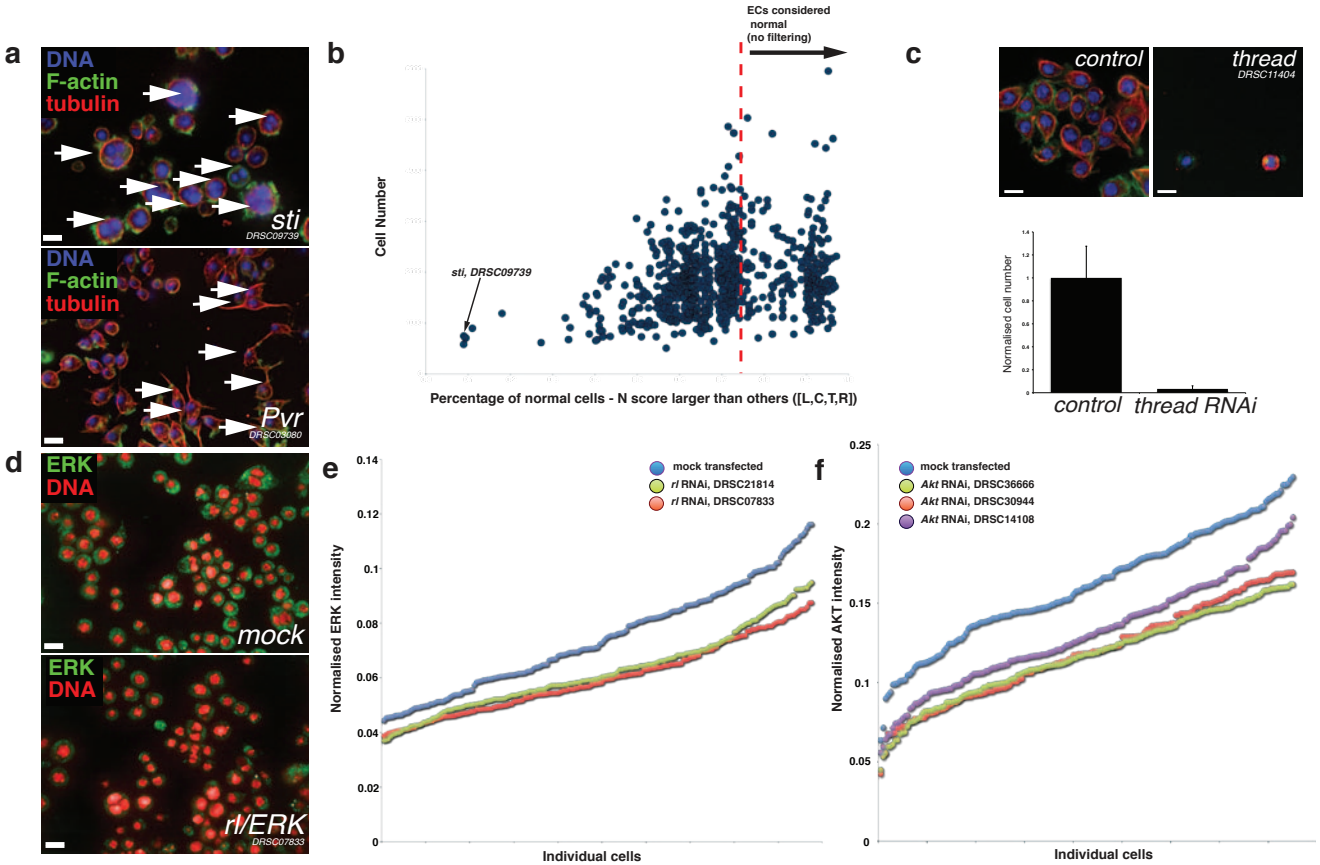

Figure S2 Accounting for differential penetrance and diverse shapes in RNAi replicates. (a) Representative image of cells treated with *sti/Citron* (DRSC09739) or *Pvr* (DRSC03080) dsRNAs. Scale bars, 20 μm. (b) The percentage of normal cells that exists in each population following treatment with a single dsRNA (x-axis) is plotted against the number of cells that were initially analysed (y-axis). ECs with >75% are considered normal. (c) Cells treated with *thread/DIAP1* RNAi (DRSC11404), show a 96.7% reduction in viability (n=4 experiments). Scale bars, 20 μm. (d) Representative images of mock-treated cells or cells treated with *rl/ERK* (DRSC07833) and stained with anti-ERK antibody (green) and DAPI (red). Scale bars, 20 μm. (e) Mean ERK intensity (normalized to DAPI intensity) from 195 individual cells randomly selected from mock-treated populations or populations treated with two different dsRNAs targetting *rl/ERK*. (f) Mean AKT intensity (normalized to DAPI intensity) from 170 individual cells that were randomly selected from mock-treated populations or populations treated with three different dsRNAs targetting *Akt*.

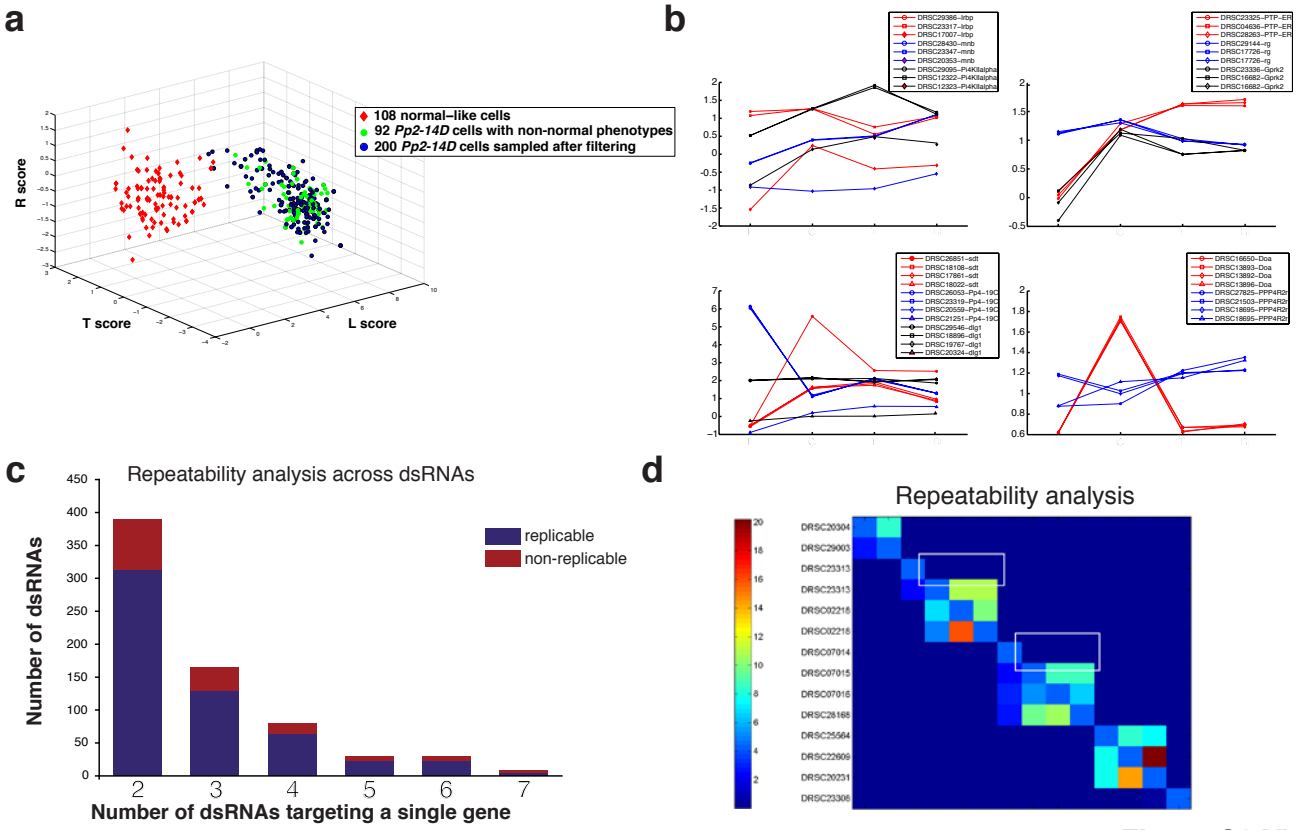

Figure S3 Repeatability analysis. (a) Example of the effects of normal cell filtering on *Pp2-14D* deficient cells. Whereas *Pp2-14D* dsRNA is 37% penetrant pre-filtering, there are almost no normal cells in the cell population post-filtering. (b) The upper panels show the similarity of the 4-dimensional QMSs (comparison to L, C, T, R shapes) generated by 3 different dsRNAs targeting the same gene. Line colour indicates dsRNAs targeting the same gene. Each point represents the mean normalised Z-score of the cell population (y-axis) describing the similarity to 4 reference shapes (x-axis). The left upper panel shows cases where dsRNAs give dissimilar QMSs, whereas the right upper panel shows cases where dsRNAs give similar QMSs. The lower panels show the similarity of the 4-dimensional QMSs generated by different 4 dsRNAs targeting the same gene. The left lower panel shows cases where dsRNAs give dissimilar QMSs, whereas the right lower panel shows cases where dsRNAs give similar QMSs. (c) The y-axis describes the number of replicable dsRNAs (blue) or non-replicable dsRNAs (red) distributed on the basis of the number of dsRNAs used to target an individual gene in the screen (x-axis). (d) Similarity matrix for dsRNAs targeting 4 genes from Clusters 1 and 2. The colour of each square represents the repeatability of each dsRNA compared with all others in the matrix. A colour towards the red end of the visible spectrum indicates increasing levels of repeatability. Squares below the diagonal depict repeatability analyses performed prior to normal cell filtering. Squares above the diagonal are analyses performed after normal cell filtering. White boxes indicate cases where normal cell filtering decreases the repeatability, meaning that the remaining shapes are dissimilar.

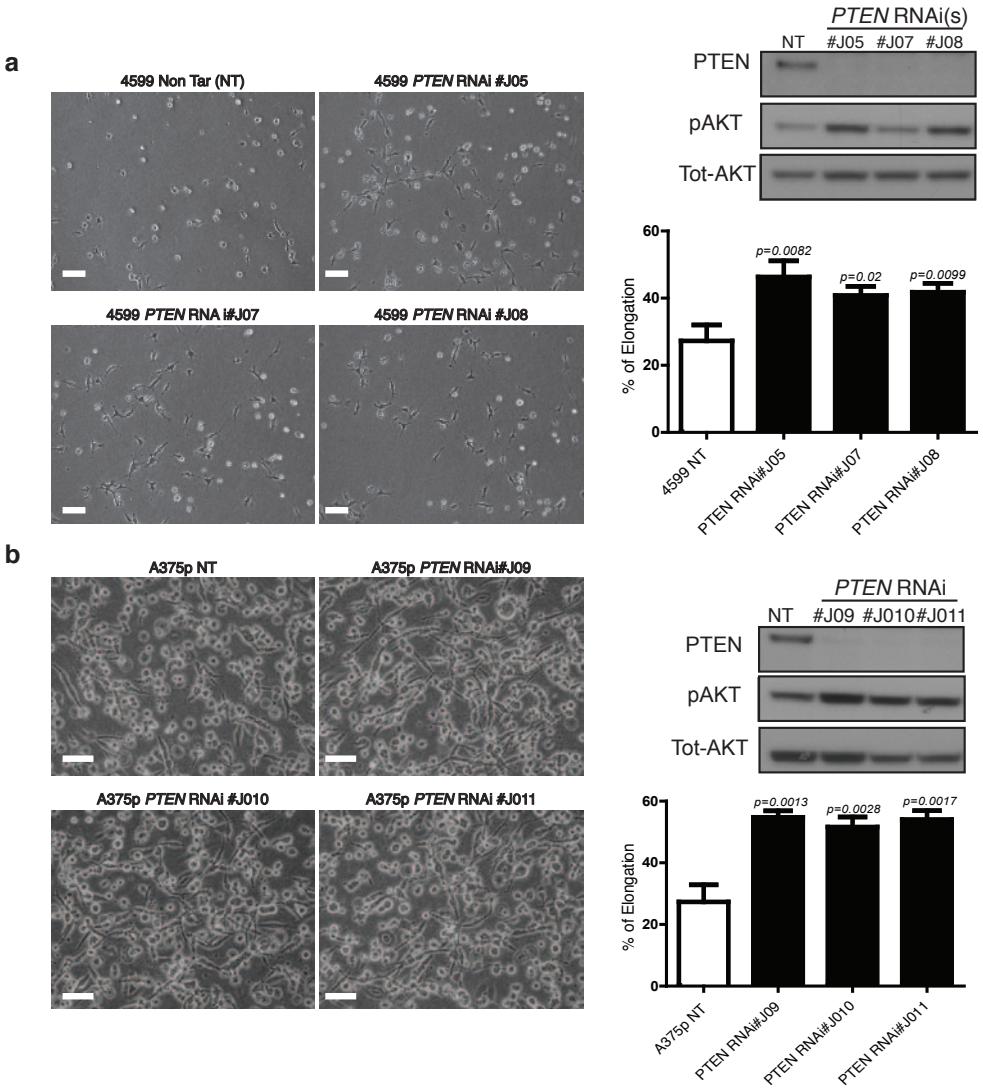

Figure S4 *PTEN* depletion by RNAi leads to increased numbers of elongated cells. 4599.1 melanoma cells (a) and A375p melanoma cells (b) were transfected with non-targetting (NT) or *PTEN* RNAi(s) and seeded on a thick layer of Col-I. After 5-16 hrs of serum starvation, cells were photographed under phase contrast. Scale bars, 50 μm. Histograms show quantification of the proportion of elongated cells (Mean±S.D.) in 4599.1 melanoma cells (a) and A375p melanoma cells (b) upon *PTEN* knockdown; 300 cells per n=3 experiments; Student's t-test was used to generate p-value. Immunoblots show the level PTEN and total (Tot) AKT in NT- and PTEN RNAi(s)-transfected 4599.1 (upper panel) and A375p (lower panel).

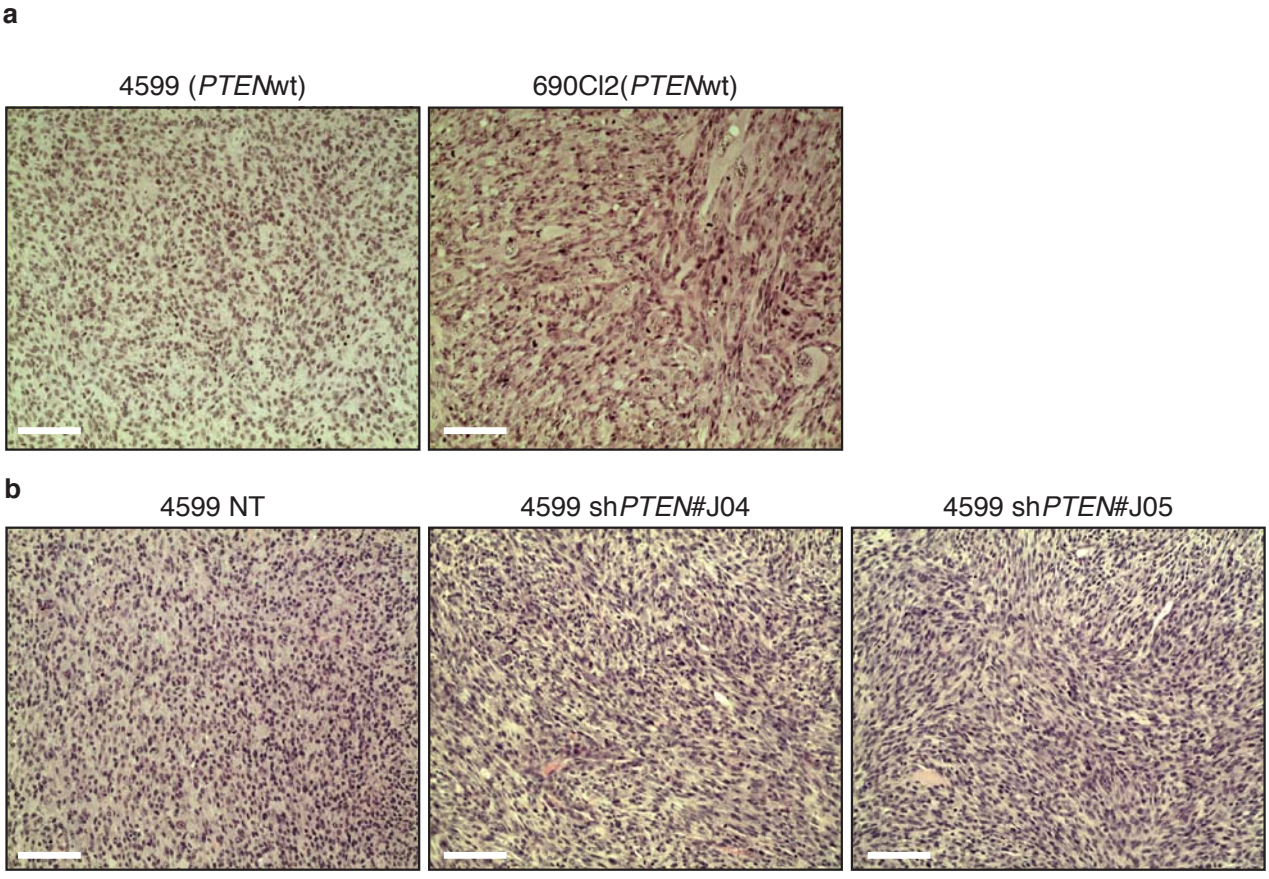

Figure S5 High magnification images of tumour sections following *PTEN* RNAi. Representative images of low magnification tumour sections derived from either non-targetting (NT), or *PTEN* shRNAs-expressing 4599.1 melanoma cells. Scale bars, 100 μm.

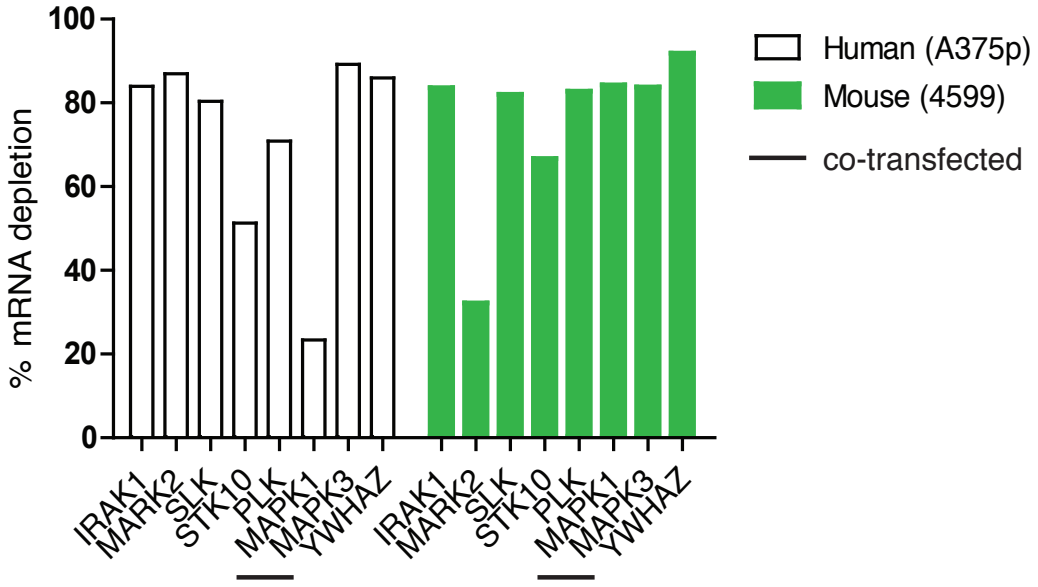

Figure S6 Levels of mRNA following siRNA-mediated knockdown in mouse and human melanoma cells.

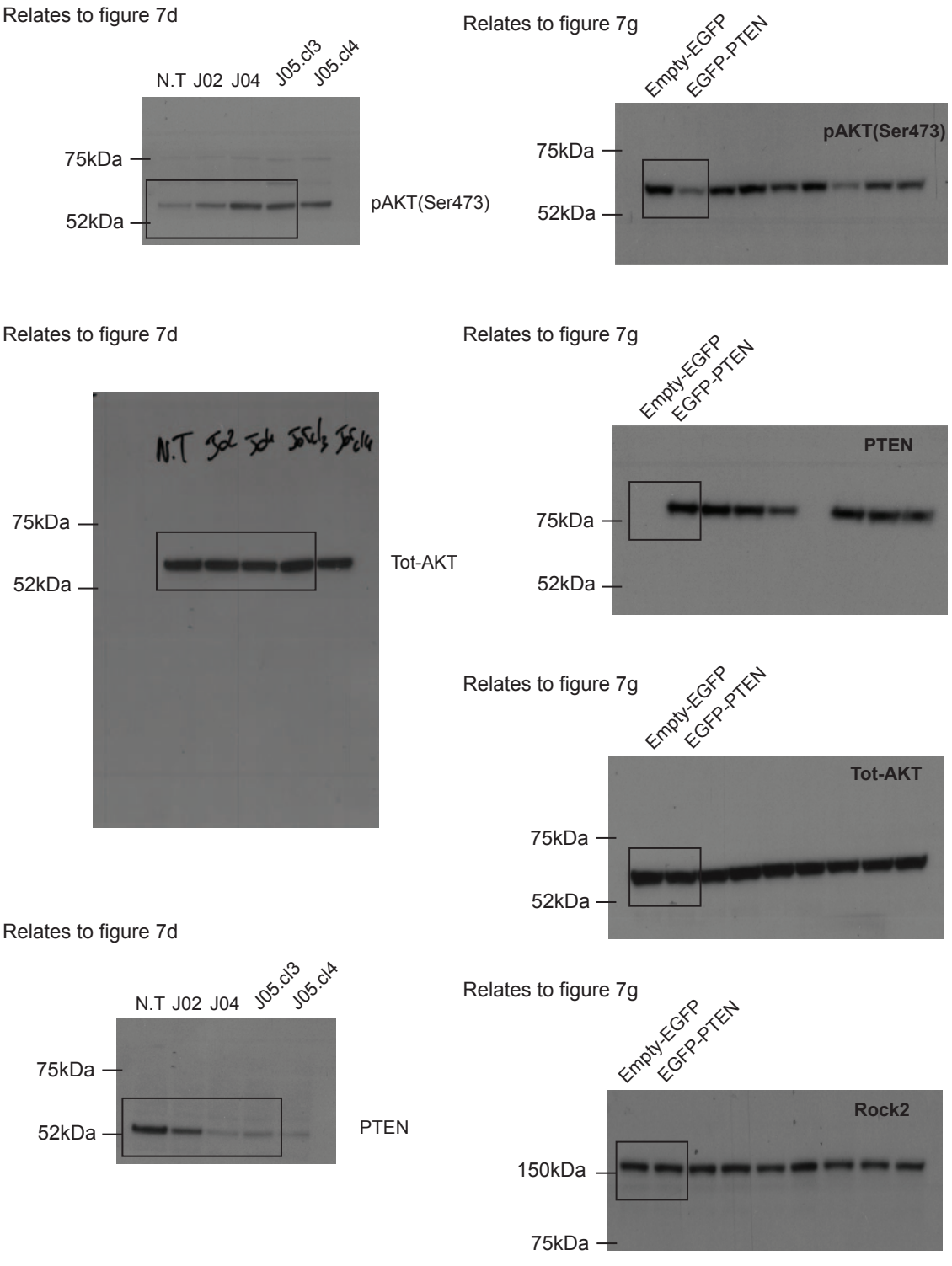

Figure S7 Uncropped Western blots.

Supplementary Table Legends

Table S1 Summary of whole-cell geometry features. Each one of 11 whole-cell geometry features is defined by a feature ID among the 211 morphology features. A brief description and the data source from where the specific feature is extracted.

Table S2 Summary of Haralick texture features extracted from the spatial-dependence matrix of each cell segment. The 14 Haralick features are divided into three groups, and the feature IDs among the 211 morphology features, as well as the feature names as defined in the original reference, are listed.

Table S3 Summary of regional geometric features. The 54 regional geometric features are divided into two groups, namely length ratios and area ratios. For each group, the feature IDs among the 211 morphology features are listed; a brief description for feature extraction are supplied; and an simple illustration for feature extraction process is shown.

Table S4 Summary of the four groups within the initial population of each GA run. The 200 individuals in each initial populations for a GA run is divided into four groups. Each group is defined based on the results of the previous SVM-RFE process, and the relationships between individuals in each group and the SVM-RFE results are defined.

Table S5 Quantitative Morphological Signature (QMS), Q(4), max RIFT, and mean RIFT scores for 287 ECs. In the first sheet we describe the number of repeatable cells that comprise each EC (column C), the number of cells from individual populations targetted by individual amplicons that contribute to total cell number. Each ECs QMS is comprised of L, T, C, and R scores and a PZ score. Cluster number is determined by hierarchical clustering (Fig. 3). In the second sheet the raw amplicon data for all tested amplicons is listed.

Table S6 Morphological comparison of *PTEN* wild-type and *PTEN*-deficient cells. Detail of data that is summarised in Fig. 7b.

Table S7 *Drosophila* genes chose for validation in mouse and human melanoma cells.

Table S8 Sequences for siRNAs and shRNAs used in mouse and human knockdown experiments.

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™, 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™.

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 $f(\mathbf{x})$ 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 $\mathbf{S_w} = [s_{\mathbf{w}}^N, s_{\mathbf{w}}^L, s_{\mathbf{w}}^C, s_{\mathbf{w}}^T, s_{\mathbf{w}}^R]$ for certain well **w**, we normalize $\mathbf{S_w}$ 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 $\mathbf{Z_w} = [z_{\mathbf{w}}^{class}, z_{\mathbf{w}}^{class}, z_{\mathbf{w}}^{class}, z_{\mathbf{w}}^{class}, z_{\mathbf{w}}^{class}]$ 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 $\mathbf{W}_D$=$\{\mathbf{w}_1,\ldots\mathbf{w}_n\}$:

1) the dispersion measurement $\mathbf{R_{W_D}} = \boldsymbol{log}(\frac{\mathbf{1}}{\mathbf{2n}}\sum_{\boldsymbol{i,j}\in\mathbf{W_D}} \boldsymbol{d}\left(\boldsymbol{Z_{w_i}}, \boldsymbol{Z_{w_j}}\right)^{\mathbf{2}})$, 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 $\mathbf{W}_D$, 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 $\mathbf{W}_D$, 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) $\mathbf{R_D^0}$, the mean value of $\mathbf{R}_{\mathbf{W}_{\boldsymbol{D}}^{(k)}}, \boldsymbol{k} = \mathbf{1}, \mathbf{2} \ldots \mathbf{1000}$, and $\mathbf{P_D^0}$, 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 $\mathbf{P_D^0}$ as the null distribution:

Null hypothesis $\mathbf{H_0}$: $\mathbf{R_{W_D}} = \mathbf{R_D^0}$;

i.e. $\mathbf{R_{W_D}}$ (dispersion for cells with a same dsRNA) is from the same distribution $\mathbf{P_D^0}$ as the cells subject to random RNAi.

Alternative hypothesis $\mathbf{H_1}$: $\mathbf{R_{W_D}} < \mathbf{R_D^0}$;

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 $\mathbf{W}_D$=$\{\mathbf{w}_1,\ldots\mathbf{w}_n\}$, the tests are repeated while members in $\mathbf{W}_D$ are iteratively removed, until:

Either null hypothesis is rejected for a subset of $\mathbf{W}_D$, 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 $\mathbf{w}_i^{(k)}$ 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: $\mathbf{Z_D} = (\sum_{i=1}^{n_{\mathbf{D}}} \frac{1}{d_{\mathbf{w}_i}} \mathbf{S}_{\mathbf{w}_i}) / (\sum_{i=1}^{n_{\mathbf{D}}} \frac{1}{d_{\mathbf{w}_i}})$.

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) $\mathbf{Z_{D_1}}$ and $\mathbf{Z_{D_2}}$, the morphological signature for D_1 and D_2, respectively;

2) $\boldsymbol{d}(\boldsymbol{Z_{D_1}}, \boldsymbol{Z_{D_2}})$, 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) d_0, the mean value for $\boldsymbol{d}(\mathbf{D}_1^{(k)}, \mathbf{D}_2^{(k)}), \boldsymbol{k = 1, 2 \ldots 1000}$; and $\mathbf{P}_0$, the estimated distribution of $\boldsymbol{d}(\mathbf{D}_1^{(k)}, \mathbf{D}_2^{(k)})$ 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 $\mathbf{H}_0$: $\boldsymbol{d}(\boldsymbol{Z}_{\mathbf{D}_1}, \boldsymbol{Z}_{\mathbf{D}_2}) = \boldsymbol{d}_0$;

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

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 $\mathbf{w}_i^{(k)}$ 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 $h = 1.06\sigma n^{-0.2}$, 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 $\mathbf{Z} = [\mathbf{z}_i], i = 1,2 \ldots n$ 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 $L_g^{class}(h) = -\sum_{i=1}^{n_g} log\mathbf{P}_{g-i}^{class}(z_i^{class}, h)$, in search of local bandwidth h_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 $\mathbf{P}(z_{\mathbf{D}}^{class}) = \frac{1}{n}\sum_{g=1}^{\mathbf{G}}\sum_{i=1}^{n_g}\frac{1}{h_g}K\left(\frac{z^{class}-z_{g,i}^{class}}{h_g}\right)$.

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 $\mathrm{J(P||Q)} = (\mathrm{KL(P||Q)} + \mathrm{KL(Q||P)})/2$ [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 $\mathbf{P}(z_{\mathbf{D_i}}^{class}), class \in \{N, L, C, T, R\}, \boldsymbol{i} = 1,2$. 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 $\mathrm{J}(\mathbf{D_1}|\mathbf{D_2}) = \sum \mathbf{J}(\mathbf{P}(z_{\mathbf{D_1}}^{class})||\mathbf{P}(z_{\mathbf{D_2}}^{class})), class \in \{N, L, C, T, R\}$. 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 $\mathbf{D}^{(k)} = \{\mathbf{D}_1^{(k)}, \mathbf{D}_2^{(k)}\}, k = 1,2 \ldots 1000$. The rule of cell sampling is the same as in 3.2.1;
3) 1000 random divergences $\mathrm{J}\left(\mathbf{D}_1^{(k)} \middle| \mathbf{D}_2^{(k)}\right), \boldsymbol{k} = \mathbf{1, 2} \ldots \mathbf{1000}$;
4) $\mathbf{J}_0$, the mean value for $\mathrm{J}\left(\mathbf{D}_1^{(k)} \middle| \mathbf{D}_2^{(k)}\right), \boldsymbol{k} = \mathbf{1, 2} \ldots \mathbf{1000}$, 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 H_0: $J(\mathbf{D_1}|\mathbf{D_2}) = \mathbf{J_0}$;

i.e. $J(\mathbf{D_1}|\mathbf{D_2})$ (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 H_1: $J(\mathbf{D_1}|\mathbf{D_2}) < \mathbf{J_0}$;

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™ 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 $\mathrm{J}_{i,j} = \mathrm{J}\left(\mathbf{D_i}\,\middle|\mathbf{D_j}\right), i, j \in \{1,2 \ldots 284\}, i \neq j$ 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™, 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™.

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.