Supplementary Materials

1. Quantitative Multi-Parametric Image Analysis (QMPIA)

a. Overview

The 3 channels (corresponding to Tfn, EGF and DAPI/SYTO 42 blue) of the fluorescent images were first corrected for microscope misalignment and uneven illumination. QMPIA was then performed in two sequential rounds of calculations. In the first round, aiming at the identification of fluorescent vesicles and nuclei, the image intensity was fitted by a sum of powered Lorenzian functions[1], the coefficients of which were used to describe the features of individual objects (e.g. intracellular position, size, fluorescence intensity, fluorescence integral intensity, elongation). In the second round, a set of statistics was extracted from the distributions of the endosome parameters measured in the first round (e.g. intracellular position relative to the nucleus, size, intensity etc.). This set (vector) of values defined the phenotypic profile of every image. Once all individual profiles were collected, data filtering and processing was performed. First, images without cells along with images with over-confluent cells and images with out-of-focus objects were removed on the basis of 4 quality control parameters; second, a first round of normalization was performed relative to internal plate controls to suppress plate-to-plate parameter variations; third, a second round of normalization was performed relative to the total set of all measurements, in order to set an “un-biased” zero-line and to weight the impact of individual parameters according to their reliability; fourth, the profiles of images belonging to the same biological conditions (i.e. the same si/esiRNA) were combined by distribution-mode searching procedures[2].

Following this analysis, we could assign a phenotype profile to each si/esiRNA screened. This data were then used to calculate the gene phenotypic profiles.

b. Microscope misalignment correction

Imaging all plates of the screen required an extended period of ~12 months. Frequent maintenance of the microscope (OPERA, Evotec Technologies GmbH, PerkinElmer) was required during the period of screening causing: 1) frequent variations in the unevenly illuminated field of view of the cameras, 2) frequent variations in the linear misalignments between the 3 image channels. To overcome these problems reference images for software-assisted image corrections were generated daily. Uneven illuminations of the fields of view were corrected on the basis of the pixel intensity distributions in images of 3 fluorescent dyes (Opera adjustment plate, Evotec Technologies GmbH, PerkinElmer). Linear misalignments between the 3 image channels and non-linear image distortions caused by small tilts of the cameras were corrected on the basis of images containing 2.5µm multicolour-beads randomly distributed in the field of view (Opera adjustment plate, Evotec Technologies GmbH, PerkinElmer). Correction parameters were calculated from 48 images with a number of beads in the range of 50-100 per image. Beads were identified independently for all 3 channels and a shift-rotate matrix with a B-spline approximation of squeezing/stretching field was calculated to minimize beads images misalignments.

c. Low quality and empty images filtering

Images without cells, with over-confluent cells and of insufficient quality (i.e. with out-of-focus objects) were removed according to the following procedure: first, all images with less than 5 or more than 65 cells were removed; then the lower 10% of the remaining images were removed according to a parameter that measures image focusing. Four quality control parameters were used for image filtering: number of nuclei and mask area (area of the view-field covered with cells) were used for the first step of filtering; nuclei intensity and nuclei contrast were used for the second. The *nuclei contrast* parameter measures the sharpness of the boundary of nuclei and *nuclei intensity* measures the nuclei fluorescence. Images with out-

of-focus objects will present low intensity of the nuclei fluorescence and smooth nuclei boundaries.

d. Fluorescent object identification

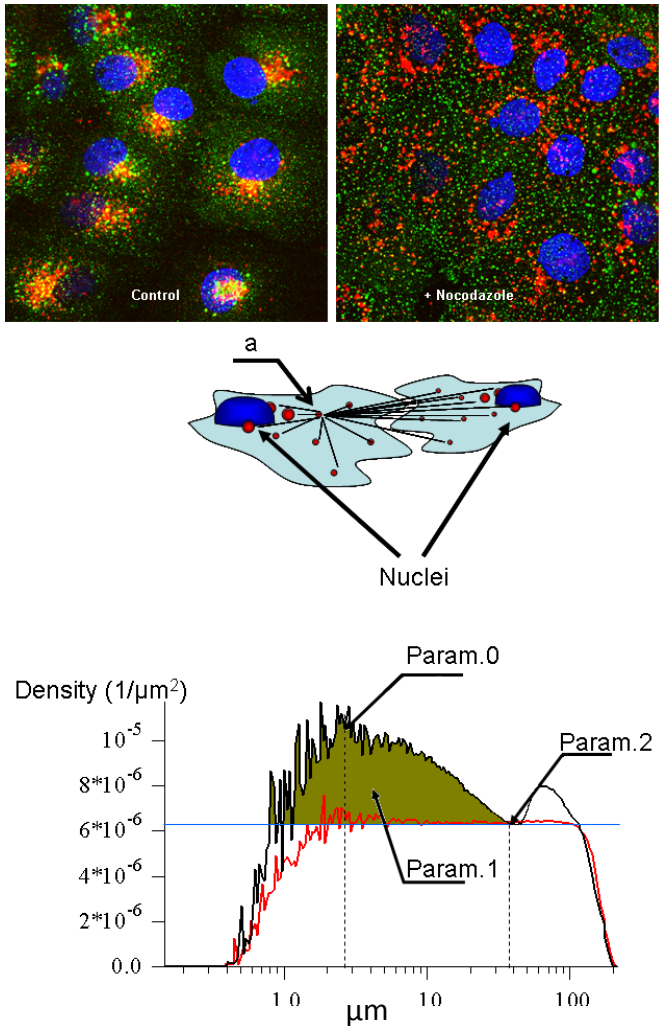
The method of fitting individual objects by a base function (i.e. point spread function of the microscope[3] as well as its approximation by computationally more tractable formulas) is common in many single object tracking algorithms[4,5,6]. Fitting single vesicles by the sum of base functions overcomes the limits of "sub-resolution" size and allows for the description of "free-shape" objects. To our knowledge, this is the first time that such an intensity fit approach has been applied to multiple (hundreds-to-thousands) vesicles using the MotionTracking/Kalaimoscope (www.kalaimoscope.com) software[1,7].

For the QMPIA, the object search was performed using the algorithm implemented in MotionTracking. This algorithm fits the image intensity by powered Lorenzian as previously described [1] in the presence of Poisson noise. The fitting procedure required solving non-linear optimization problems for fitting individual objects. Following optimizations, this procedure required ~30-100 ms on a 3GHz Intel Xeon processor per object. Given both the large number of objects and the 3 channels per image, the time required for the calculation of each image increased to minutes. Thus, vesicle search became a rate limiting step for data processing. To overcome this limitation, a Pluk-based automatic task distribution system[8] was developed to share the calculations in a heterogeneous computer environment, including Windows based PCs (31 CPUs), a Linux-based in-house cluster (60 CPUs) and the PC-Farm of the High Performance Computing department of the University of Technology in Dresden (2,584 CPUs). The server component of MotionTracking was ported by the authors to Linux, and the sharing procedure was modified to work with the SSH (Secure SHell) protocol. The performance of the task distribution system grew up almost linearly until 1,000 processors. As a result, the shared calculation provided the ability to process images within the time frame required for their acquisition by the automated microscope.

e. Phenotype parameters (statistics)

The measurement of individual endosome parameters often resulted in complicated non-Gaussian distributions that could not be used directly as part of the phenotypic profiles since they would result in non-tractable long profile vectors (where every feature, i.e. fluorescence intensity, size, intracellular positioning, etc, would have to be presented by a long set of binned data). At the same time, simple statistics such as mean values are not sufficient to describe all information encoded in the parameter distributions. For example, the addition of few large endosomes to several hundreds small endosomes changes the endosome size distribution while having little impact on the mean endosome size. To increase the sensitivity toward changes in non-Gaussian distributions, we considered two additional mean values, one weighted by the integral intensity of the endosomes and the other weighted by their mean. For a given parameter, the mean weighted by the integral intensity biased towards the vesicles containing most of the cargo, i.e. when a large fraction of cargo accumulates in few endosomes, these endosomes will be the main contributors to the mean weighted by the integral intensity but will be of limited influence on the simple, non-weighted mean value. Parameters weighted by intensity are biased towards the brightest endosomes (where cargo is concentrated) and therefore are different from the first two mean values. Since we included the weighted statistics to take into account the long tails of the measured distributions, the non-weighted means were replaced by median values to suppress the influence of outliers (see Suppl. Table I). The use of the 3 statistics described above for each parameter provided a more detailed description of the phenotypes.

Some of the parameters, such as the number of endosomes and total amount of internalized cargo, are meaningful only when normalized relatively to the cytoplasm area. The parameters measuring the distance between endosomes and nucleus were normalized by the effective radius of the nucleus (calculated as the radius of a circle with an area equal to the area of the nucleus) to suppress dependency of these parameters on cell size.



Suppl. Fig. 11. Endosome peer-to-peer distance distribution (distribution of the relative endosome density as a function of the distance of any given endosome from all others, averaged over the whole endosomal population). The set of distributions was calculated around every single endosome within the interval 0.5÷30 µm. This set was averaged to build the final distribution. The lower boundary of interval was chosen on the basis of both endosome size and microscope resolution limit. The upper boundary corresponds to the maximal cell size. The precise values of boundary do not significantly influence the final scoring. **Upper panel**: Endosome subcellular distribution. Nuclei are stained with DAPI and pseudo-coloured in blue, early endosomes are stained with an EEA1 rabbit polyclonal antibody and pseudo-coloured in green, LDL positive endosomes labelled by a 40 min internalization of LDL-DiD are pseudo-coloured in red. **Middle panel**: Illustration of the peer-to-peer distance distribution for one marked endosome "a". The total distribution is the average over all endosome distributions. **Bottom panel:** peer-to-peer distribution in control condition (black); peer-to-peer distribution after treatment of the cells with nocodazole, a microtubule depolymerising agent that scatters the endosomes randomly, (red) and theoretical uniform distribution (blue). Param.0 is the peak position of the approximated (see text) distribution. Param.1 is the integral of the area above the uniform distribution (area is coloured in brown). Param.2 is the distance to the point where the distribution crosses the uniform distribution level. It is worth to mention that the second peak on the right corresponds to endosomes which belong to different cells (corresponding to a distance > 30 µm). The peer-to-peer distance distribution of the nocodazole treated sample mimics the uniform distribution, demonstrating the efficacy of this distribution in measuring endosomes clustering.

The degree of endosome clustering was measured using 3 parameters (see below) calculated from the endosome peer-to-peer distance distribution, a distribution of the probability to find another endosome on unit area at given distance from the given endosome. From the distribution we considered 1) the position of the maximum (corresponding to the most probable distance between endosomes in a cluster, Param.0), 2) the integral of the part of the distribution that is above the uniform probability distribution level (corresponding to the "participation in the cluster", Param.1; see Suppl. Fig. 11) and 3) the peer-to-peer distance where the distribution crosses the uniform distribution level (corresponding to the size of cluster, Param.2; see Suppl. Fig. 11). Given that the distribution is noisy, the peak position (Param.0) was determined by quadratic (parabolic) approximation of the distribution in the least-square sense. The approximation was done in logarithmic scale on the abscissa axis.

Co-localization between markers was calculated on the basis of individual object analysis. Object A is considered to colocalize with object B if object B covers at least 50% of the area of object A. To correct for apparent (random) colocalization, the objects were randomly permutated while maintaining the local density distribution. The random colocalization, calculated as average of 5 permutation calculations, was subtracted from the measured colocalization using the following formula:

$$C_{corrected} = \frac{C_{measured} - C_{random}}{1 - C_{random}}$$

f. Phenotype profiles

All measured parameters were normalized relative to the internal plate control (MOCK transfected cells) and combined into multi-parametric "profiles" describing the phenotype of the gene knockdown. Every parameter is subjected to both the noise of the measurement and of the biological system. The quality of parameters, meaning their robustness to noise, is dependent on both the procedure for the calculation of the parameter and the particular assay. For example, we found that the total intensity of the Tfn and EGF channels (computationally the same parameter but applied to different endocytic markers; see below), showed very different degrees of robustness.

f.1 Parameter normalization

We calculated 62 parameters to profile the endocytic phenotypes in the microscopic images. Of these, four parameters (number of cells, nucleus contrast, area of image covered by cells and mean nucleus intensity) were used as quality control parameters, designed to reject low-quality images i.e. images without cells, over-confluent cells, out-of-focus images, etc. The remaining 58 parameters were normalized in two sequential rounds.

The first round of normalization eliminated plate-to-plate variability with respect to e.g. changes in laser intensity, fluorescent cargo, staining quality and other sources of experimental variation throughout the screen. For this purpose, any sample which is far away from saturation (i.e. an extreme phenotype) would be suitable. Since normalization is a potential source of error, the more images are available, the more accurate is the measurement of the "reference" phenotype and the lower is the normalization error. The mock sample (Mock-transfected cells, i.e. treated with the transfection reagent but no si/esiRNA) satisfied those criteria and was chosen as the reference condition. In each plate a large number of Mock images (~180 images) were acquired (see plate layout in Suppl. Fig. 5). Thus, parameters were replaced by moderated z-scores, which were calculated relative to the Mock control condition, using the formula:

$$z_i = \frac{p_i - E[p_{MOCK}]}{SD_{MOCK}} \qquad (1)$$

where $E[p_{MOCK}]$- is the mean of each parameter in Mock-treated wells and SD_{MOCK} - the standard deviation (SD) for each parameter in Mock-treated wells calculated per plate.

The second round of normalization was devoted to set to zero the general, non-specific RNAi phenotype, or in other words, nullify the "negative control" phenotype. However, "negative" control siRNAs (e.g. luciferase or scrambled siRNAs) are also not immune to off-target effects (data not shown). To circumvent this problem, in this study we used the **mode of the total set** of all siRNAs (over 100,000) as "negative control" for normalization. With the mode[2] of the total set the impact of any particular off-target effect is averaged out due to the large number of si/esiRNAs considered. Hence the second procedure sets the mode value of the

total set as a "zero-line", i.e. data from all the plates were merged and secondary moderate z-scores

$$\hat{z}_i = \frac{z_i - \mathrm{mod}\,e(\{z_i\})}{mean(SD_{z_i:equal\_condition})} \qquad (2)$$

were calculated based on the total set. Assuming that the total set of si/esiRNAs in the genome-wide library is unbiased, the mode of the total set was considered as "negative control" ("unspecific-phenotype").

The parameters were scaled (normalized) by inverse modes of the respective SDs (i.e. the most probable amplitude of the SD of any given parameter within the fixed condition). For this, the SD was first calculated for every single condition (untreated, mock, si/esiRNA, etc.) and then the mode of the SD distribution ($mean(SD_{z_i:equal\_condition})$) was used for normalization. We would like to stress that the SD mode is independent (at first approximation) of the variance of any given parameter between the different si/esiRNA and dependent mainly on the robustness of the parameter measurement for every given condition. This normalization allows us weighting the individual parameters according to their respective robustness before they can be used for phenotypic scoring. As a result, the mode of $\hat{z}$ distribution is zero, and the SD of the measurement noise equals the unity.

f.2 Parameter robustness

The parameters were tested for their robustness relative to the reproducibility of the assay. This test was performed on independent replicates of the screen of all human kinases and phosphatases (1,459 genes, 9,330 siRNA). The stability of each parameter (Pearson correlation between replicates) was calculated using the formula:

$$c_i = \frac{\sum_{j=1}^{J}\left(\hat{z}_{i,j}^{(1)} - E\left[\hat{z}_{i,j}^{(1)}\right]\right)\left(\hat{z}_{i,j}^{(2)} - E\left[\hat{z}_{i,j}^{(2)}\right]\right)}{(J-1)SD_{\hat{z}_i^{(1)}}SD_{\hat{z}_i^{(2)}}} \qquad (3)$$

Here, the upper index denotes the screen replica, the first index (*i*) the parameter, the second index (j) the si/esiRNA. Results of these calculations are presented in Suppl. Table I. Considering that the measured value $\chi_{i,j}$ of each parameter (*i*) for a given si/esiRNA consists of two

components: $\chi_{i,j} = \eta_i + \varepsilon_i^{(j)}$, where η_i - is the phenotype response to a given si/esiRNA, $\varepsilon_i^{(j)}$ - is the noise of the measurement and the biological variability of the assay in the *j*-th experiment.

Then, the Pearson correlation between two experiments for each parameter *i* is:

$$c_i = \frac{E\left[\chi_i^{(1)} \cdot \chi_i^{(2)}\right]}{\sqrt{E\left[\left(\chi_i^{(1)}\right)^2\right] \cdot E\left[\left(\chi_i^{(2)}\right)^2\right]}} = \frac{E\left[\eta_i^2 + \eta_i \varepsilon_i^{(1)} + \eta_i \varepsilon_i^{(2)} + \varepsilon_i^{(1)} \varepsilon_i^{(2)}\right]}{\sqrt{E\left[\eta_i^2 + 2\eta_i \varepsilon_i^{(1)} + \left(\varepsilon_i^{(1)}\right)^2\right] \cdot E\left[\eta_i^2 + 2\eta_i \varepsilon_i^{(2)} + \left(\varepsilon_i^{(2)}\right)^2\right]}} \quad (4)$$

where the low index (*i*) is the index of individual parameters, the upper index (1 or 2) is the index of each independent experiment and averaging is performed over the whole si/esiRNA set. It is worth mentioning that, technically, the normalization was done before the correlation analysis and was described according to the data processing flow. However, it could be done also before the second round of normalization.

Given that the noise is not correlated between independent experiments, and that the mean of the noise value is $E(\varepsilon) = 0$ with a variance $D(\varepsilon_i) = \sigma_i^2$, we have:

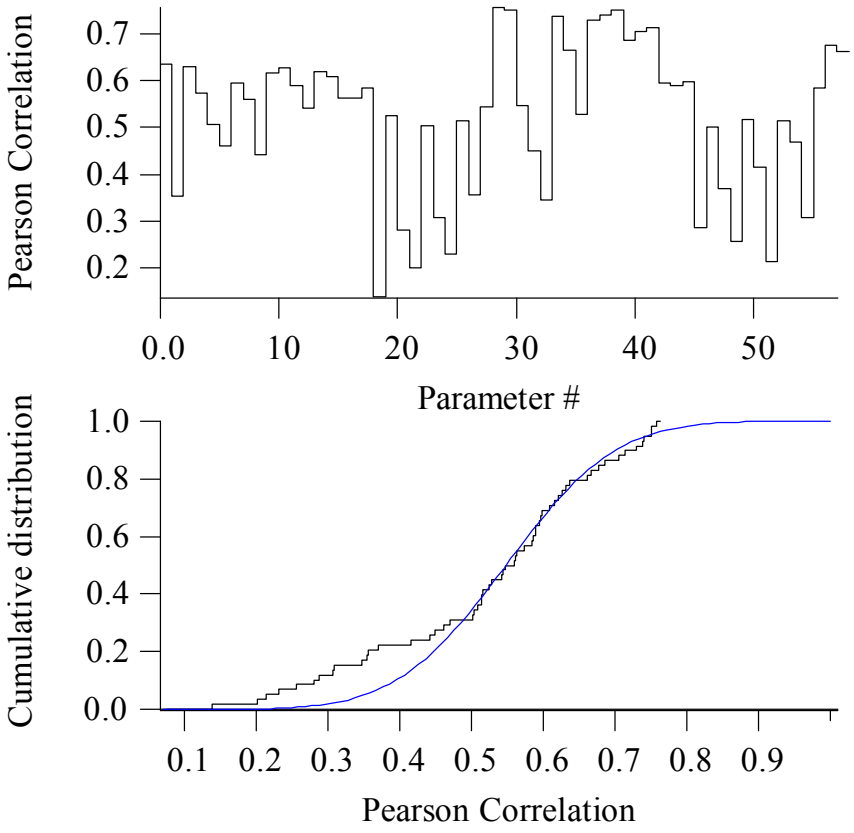
$$c_i = \frac{E\left[\eta_i^2\right]}{E\left[\eta_i^2\right] + \sigma_i^2} = \frac{SNR_i^2}{SNR_i^2 + 1} \quad (5)$$

where $SNR_i = \sqrt{\frac{E\left[\eta_i^2\right]}{\sigma_i^2}}$ denotes the "signal-to-noise" ratio for parameter *i*.

In other words, the Pearson correlation coefficient shows the robustness of the parameters with respect to measurement errors and instability of the biological assay. From Suppl. Table I, one can see that computationally equal parameters, i.e. the same parameter measured on different channels, shows essentially different Pearson correlation values in some cases. This results from the different specific-to-random (noise) signal ratios of different parameters, which is proportional to the number of genes involved in the parameter modulation and reversely proportional to the square root of the biological noise (variance), between the two different markers. Generally, the Pearson correlation coefficients provide an objective basis to select the most "specific" (reliable) parameters for phenotypic profiling.

f.3 Determination the parameters to use for the phenotype profile

The parameters used to calculate the phenotypic profiles were selected based on a correlation threshold of 0.4 (see Suppl. Table 1 and Suppl. Fig. 12). Parameters marked by "-" were excluded from subsequent analysis. Parameters marked by "-+" were kept for further analysis even though their value was below this threshold, since their counterpart in the other channel was significantly above the threshold. Parameters marked by "+-" were excluded since they were not significantly above the threshold in both channels. In total, 12 parameters were considered to be insufficiently reliable and were excluded from subsequent analysis. The RNAi profile was, thus, described by 46 parameters.



Suppl. Fig. 12. The upper panel shows the correlation of the parameters enumerated in Table 1. The bottom panel presents the cumulative distribution of the parameter Pearson correlation coefficients (black) and the fitted normal distribution (blue). The latter was used to determine the threshold for parameter discrimination (t = 0.4).

Notably, singular value decomposition (SVD) analysis showed that the parameters are partially correlated but non-redundant (data not shown).

2. Gene profile estimation and scoring

a. Estimation of *gene profiles*

The detailed description of the mathematical procedures employed for the estimation of the phenotypic *gene profiles* and the scoring will be detailed in a separate manuscript. Here we provide a short overview of the approach. Since virtually every RNAi phenotype profile is contaminated by off-target effects (manuscript in preparation), the assessment of the "on-target" component of the gene silencing phenotype (*gene profile*) requires filtering the off-target effects from the profiles of each individual si/esiRNAs. For this, we built a Bayesian probabilistic model of *gene profiles*, based on the following five assumptions:

1) Quantitative variations in target protein degradation affect the phenotype quantitatively but not qualitatively, an assumption supported by the knockdown time-course experiments (data not shown);

2) On- and off-target components of a profile are additive since it is reasonable to assume that a) off-targets are randomly distributed among the genome and b) knockdown of most genes produce mild phenotypes;

3) Off-target gene silencing of different si/esiRNAs targeting the same gene are independent (at least with respect to multi-parametric profiles)

4) Spread of directions of the phenotypic vectors is described by von Mises-Fisher distribution, analogue of Gauss distribution on unit hypersphere, i.e. is a limit distribution for the sum of random equally distributed components

5) All si/esiRNAs targeting a given gene have the same (equal) probability to be ineffective (i.e. there is no bias in the si/esiRNA design).

Based on these assumptions, the most probable phenotype profile of every gene, given the profiles of all targeting siRNAs, was estimated by a nested two-stage (for direction and for amplitude) Expectation-Maximization (EM) algorithm:

First, in order to weight the impact of the individual parameters on the *gene profile* according to their respective experimental uncertainty, all parameters were normalized to the level of measurement uncertainty as described above (f.1). Then, the multi-parametric profiles of the individual si/esiRNA were considered as vectors in the multi-dimensional parameter space whose direction and amplitude were defined by the qualitative and quantitative features of the phenotype, respectively. This model therefore considers two components: 1) the normalized phenotypic profile (the vector projection on the unit hypersphere, which corresponds to the nature of the phenotype and is considered independent of its strength) and 2) the phenotypic amplitude (length of the vector). The normalized phenotypic profile of every RNAi (directional part) obeys the von Mises-Fisher distribution with an unknown concentration coefficient k :

$$p(\vec{x} \mid \vec{\mu}, k) = \frac{k^{\frac{n}{2}-1}}{(2\pi)^{\frac{n}{2}} \cdot I_{\frac{n}{2}-1}(k)} \cdot \exp\left(k \cdot \vec{\mu}^T \vec{x}\right) \qquad (12)$$

where $\vec{x}$ is the vector of a si/esiRNA phenotype profile normalized to unity, $\vec{\mu}$ is a vector of a gene phenotype profile normalized to unity, k is the concentration parameter, n is the phenotype vector dimension (number of parameters), $I_{\frac{n}{2}-1}(k)$ is a modified Bessel function. To take into account the biological correlation between the different parameters, the angular distance was calculated as $\vec{\mu}^T C^{-1} \vec{x}$, where C is the parameter correlation matrix. The final equation includes the gene phenotype amplitude by the mean of the gene concentration coefficient (*k*).

After the estimation of the *gene profile* vectors direction their amplitude *A* was calculated by the maximization of probability of their projection on the corresponding si/esiRNA profiles. This calculation involves as additional parameter the probability of every single RNAi to be impotent (β). The cycle starts with a β = 0.1 and *k* estimated based on the amplitude of the average of all si/esiRNA profiles.

After the first estimation of the *gene profiles* direction and their amplitude, the values of *k* and β are recalculated and the loop is iterated until the calculation converges.

It is worth to mention, that the gene profiles which are found in this way are unique. They are found on the basis of all confirmatory profiles (based on profile correlations) of individual siRNAs/esiRNAs, no profile flexibility/changeability was allowed at all. To estimate the reproducibility of the gene profiles, we re-screened twice the kinome-phosphatome (9,330 siRNAs, 1,459 genes), including a large fraction of the gene hits (235). Reproducibility of gene profiles was estimated on the basis of confirmatory profiles (i.e. correlation coefficient higher than random with $p_{\text{-value}} \leq 0.05$) and found to be 84% for the genes with strong phenotype (χ^2 $p_{\text{-value}} \leq 0.05$).

b. Scoring of *Gene profiles*

The first scoring criteria of the gene phenotype is based on the phenotype strength, which is defined by the p-value of χ^2 distribution. To reduce the "over-scoring" of correlated parameters the p-value was calculated as $\chi^2 = A^2 \cdot \vec{\mu}^T \Sigma^{-1} \vec{\mu}$, where Σ – is the co-variation matrix, $\vec{\mu}$ - gene profile. In addition to the χ^2 p-values we also considered a novel *"pheno-score"* value to characterize the effect of gene knock-down. The *pheno-score* assesses the probability that the knockdown produces not just a significant, but also a specific phenotype. The model behind the calculation of this value is similar to that used to compute gene profiles, and assumes that the observed knockdown phenotype profile $\vec{r}$ is the sum of two vectors: the vector of the "pure" phenotype, defined by a discrete system state (Cluster-group profile; see below) resulting from hitting certain cellular pathway, and a random noise vector $\vec{n}$. The "pure" phenotype vector has fixed direction $\vec{p}$ and has an unknown length of l distributed as $\sqrt{\chi^2}$ with 1 degree of freedom. The noise vector has a random length drawn from the $\sqrt{\chi^2}$ distribution with d (number of dimensions i.e. number of parameters) degrees of freedom and a random direction. The cosine of the angle to the ideal phenotype vector is distributed as von Mises-Fisher with $k = 0$ marginalized over (n-1) azimuthal angles. The joint probability distribution $P(l, \vec{p}, \vec{r})dl$ is as a product of probabilities of the length of the noise vector, the angle between the noise vector and the "pure" phenotype vector, and the prior probability of l

yielding observed vector $\vec{r}$. Given $\vec{r}$ and $\vec{p}$, this probability depends only on l . The score is the integral over probability of l being more than the average noise length. The last one equals to $\sqrt{d}$.

Importantly, we use the results of mean-shift clustering of the gene profiles (see below) to implement the *pheno-score* method. For each gene profile $\vec{r}$ the mode of its respective cluster is taken as an estimate of the 'pure' phenotype vector $\vec{p}$. Moreover, we exclude from scoring the cluster groups 1 and 3, as enriched in metabolic or toxic, phenotypes.

3. Phenotype clustering by a mean-shift algorithm.

To perform clustering of the gene phenotypic profiles we developed a modified version of the mean-shift clustering algorithm. Mean-shift is a non-parametric clustering algorithm for n-dimensional vector datasets[9]. The procedure iteratively shifts every vector x of the dataset to m(x), an average of the vectors of the dataset weighted by the kernel function, which is a function of distance between the vectors. The vector m(x)–x is called the mean-shift vector, and it is a non-parametric estimate of the gradient of point density around x. By iteratively shifting the data vectors to the corresponding 'means' the algorithm migrates them towards the gradient of density until all points converge to local maxima (modes) of density, forming clusters. A variety of kernel functions may be used in the algorithm, given several restrictions: the kernel is a non-negative monotonously decreasing function of the distance between vectors with limited integral over the multi-dimensional parameter space.

We have designed a version of the mean-shift algorithm, which provides several distinct features.

We treated phenotypic profiles as directional data (vectors in the n-dimensional parameter space), in which the direction of the vector corresponded to the nature of the phenotype, whilst the length was a measure of the strength (quantity) of the phenotype. Hence, a natural measure of profile similarity was the cosine of the angle between the respective vectors.

The von Mises – Fisher distribution[10], was used as a kernel function. The distribution concentration parameter *k* defines the width of the kernel estimate of the gradient density, where the higher *k* corresponds to a narrower kernel and thus provides a finer resolution. The m(x) is the sum of all vectors in the dataset weighted by the kernel function and by their lengths, assuming that directions of stronger phenotypes are more precisely measured and, therefore, are more valuable for the estimation of the density gradient. Our algorithm is a blurring mean-shift, i.e., every point of the dataset is shifted by the corresponding m(x) value after each clustering iteration. This accelerates cluster convergence, but does not affect the quality of segmentation[11].

The disadvantage of the blurring mean-shift algorithm is that groups of vectors (*gene profiles*) formed at early iterations continue to migrate towards each other at later iterations. This causes a shift of the cluster modes in comparison to the modes of the initial dataset. In order to resolve this problem, at the end of the clustering procedure the vectors around which the clusters are formed are considered to be preliminary estimates of the modes and are included into another run of non-blurring mean-shift on the initial dataset. These vectors migrate until convergence to the real density modes, thus yielding accurate mode estimates.

The total dataset of 20386 gene profiles was clustered at a series of resolution levels defined by the values of the kernel spread parameter *k* ranged from 5 to 100 with step 5. The comparison of clusters formed at various resolution levels shows that all the major clusters (more than 100 profiles) are obtained from clustering with *k* up to 55, and clustering with higher resolution mostly increases of the amount of very small clusters (less than 10 profiles), which prompted us to select the cluster set obtained with *k* of 55 for the following analysis.

The robustness of clusters produced by mean-shift was assessed by clustering of 100 random subsets of 4000 gene profiles each are sampled from the total dataset of 20386. The clusters larger then 100 gene profiles were reproduced in 86% of cases, as measured by cluster mode similarity at the threshold of 0.95. The clusters matching between two subsets shared on average 76% of their gene profiles.

4. Bioinformatics analysis

a. Functional Annotation

Functional Annotations were obtained from several sources. NCBI RefSeq (Release 28) was used for gene sequences and descriptions, functional domains were annotated using *InterProScan*[12] from EBI, molecular function and biological process data were retrieved from both the Panther Ontology (via *iprscan*[13]) and Gene Ontology[14] (via RefSeq / NCBI Gene annotations), pathway data was retrieved from the KEGG database[15] and further functional annotation along with protein-protein interaction data was retrieved from the HPRD database[16]. Furthermore, relevant literature, human disease information (OMIM[17]) and previous characterization of the hits was obtained from the NCBI Entrez system by automated data mining, as well as manual searches.

b. Detection of Overrepresented Functional Groups in the dataset

In order to determine the over-represented functional groups (Panther Biological Process, Panther Molecular Function, InterPro Domains, Gene Ontology, KEGG Pathways) in the dataset we calculated p-values of the occurrence of all annotation terms and functional domains within the positive hits in our screen relative to a background database (the screening library mapped to RefSeq release 28). P-values of were calculated based on the hypergeometric distribution and were then ranked to determine which terms were over-represented.

References

[1] Rink, J., Ghigo, E., Kalaidzidis, Y. & Zerial, M. Rab Conversion as a Mechanism of Progression from Early to Late Endosomes. *Cell* **122**, (2005)

[2] Sivia, D.S. Dealing with Duff Data. *Proceedings of the Maximum Entropy Conference*, 131-137 (1996).

[3] Born, M. & Wolf, E. Principles of Optics, Pergamon Press, Oxford-London-Edinburgh-New York-Paris-Frankfurt. (1968)

[4] Anderson, C.M., Georgiou, G.N., Morrison, I.E.G., Stevenson, G.V.W. & Cherry, R.J. Tracking of cell surface receptors by fluorescence digital imaging microscopy using a charge-coupled device camera Low-density lipoprotein and influenza virus receptor mobility at 4°C *J. Cell Sci.* **101**, (1992)

[5] Cheezum, M.K., Walker, W.F. & Guilford, W.H. , Quantitative Comparison of Algorithms for Tracking Single *Fluorescent Particles, Biophys. J.* 81, (2001)

[6] Bonneau, S., Dahan, M. & Cohen, L.D. Tracking single quantum dots in live cell with minimal path, Proc. IEEE Comp. Soc. Conf. on Comp. Vis. and Patt. Recogn. (CVPR'05), 141-148. (2005)

[7] Helenius, J., Brouhard, G., Kalaidzidis, Y.L., Diez, S. & Howard, J., The depolymerizing kinesin MCAK uses lattice diffusion to rapidly target microtubule ends. *Nature* **441**, (2006)

[8] Kalaidzidis, Y. L., Gavrilov, A. V., Zaitsev, P. V., Kalaidzidis, A. L. & Korolev, E. V. PLUK—an environment for software development. *Program. Comput. Softw.* **23**, (1997).

[9] Cheng Y. Mean Shift, Mode Seeking, ad Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, (1995)

[10] Banerjee A., Dhillon I., Ghosh J. & Sra S. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. The Journal of Machine Learning Research, 6, (2005)

[11] Carreira-Perpiñán M. Fast nonparametric clustering with Gaussian blurring mean-shift. *ACM International Conference Proceeding Series* 148, (2006)

[12] The InterPro Consortium (Apweiler, R., et al.) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* 29, (2001)

[13] Zdobnov, E.M. & Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, (2001)

[14] The Gene Ontology Consortium (Ashburner, M. et al.) - Gene Ontology: tool for the unification of biology. *Nature Genet*. **25**, (2000)

[15] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, (1999)

[16] Peri, S. et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*. 13, (2003)

[17] McKusick, V.A. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders *Baltimore: Johns Hopkins University Press*, (1998)