SINGLE-CELL OMICS

Single-cell and multivariate approaches in genetic perturbation screens

Prisca Liberali[1], Berend Snijder[1,2] and Lucas Pelkmans[1]

Abstract | Large-scale genetic perturbation screens are a classical approach in biology and have been crucial for many discoveries. New technologies can now provide unbiased quantification of multiple molecular and phenotypic changes across tens of thousands of individual cells from large numbers of perturbed cell populations simultaneously. In this Review, we describe how these developments have enabled the discovery of new principles of intracellular and intercellular organization, novel interpretations of genetic perturbation effects and the inference of novel functional genetic interactions. These advances now allow more accurate and comprehensive analyses of gene function in cells using genetic perturbation screens.

Cell-to-cell variability
The phenomenon that individual cells in a population of genetically identical cells display variable activities and behaviours.

[1]*Institute of Molecular Life Sciences, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.*
[2]*Present address: CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, A-1090 Vienna, Austria.*
Correspondence to L.P.
e-mail: lucas.pelkmans@imls.uzh.ch
doi:10.1038/nrg3768
Published online
2 December 2014

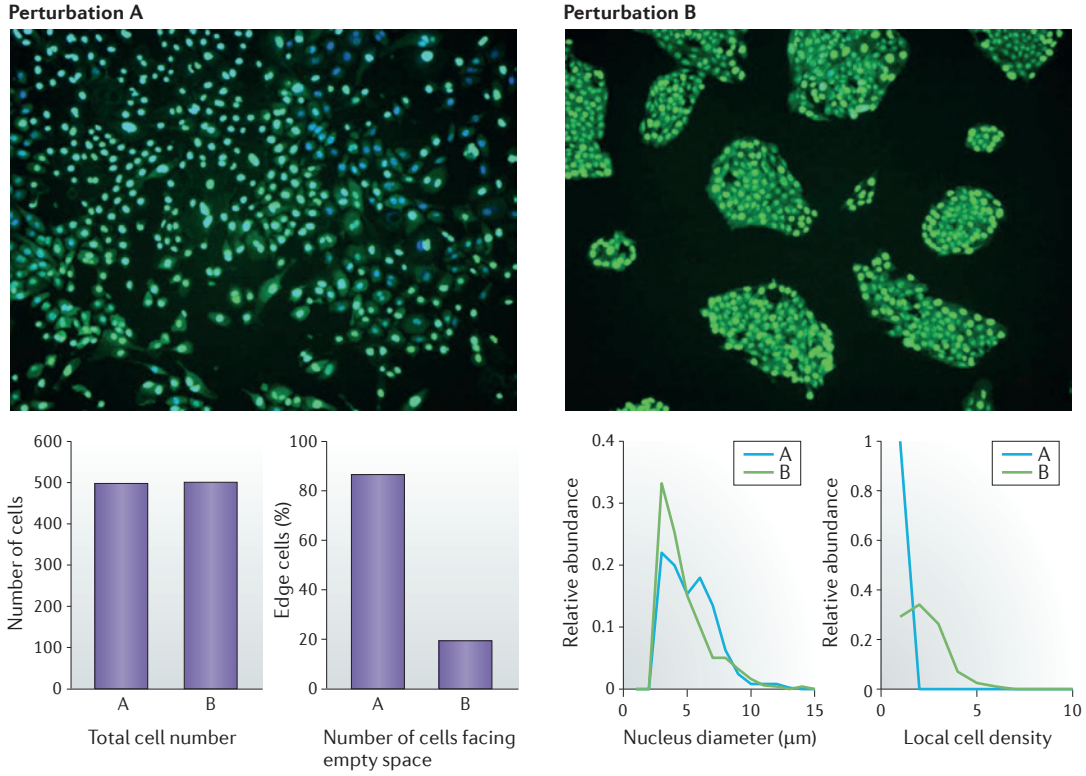
Large-scale genetic perturbation screens have been instrumental in many biological discoveries[1–3]. These screens use perturbations that act at the DNA, RNA (post-transcriptional) or protein (post-translational) level, providing a variety of different readouts. Given their fundamental importance in genetics, molecular cell biology and systems biology, these methods — as well as the various commonly applied statistical approaches to extract information from large-scale genetic perturbation screens[4–8] — have been extensively described previously[1–3,6,9,10].

In this Review, we do not elaborate on methods for genetic screens, although we provide an overview of the relevant techniques. We specifically focus on two related aspects, the importance of which only became apparent recently. First, we describe the phenomenon of cell-to-cell variability or cellular heterogeneity (BOX 1), which is a fundamental property of populations of cells, and discuss recent advances in the ability to quantify, at a large scale, multiple parameters of genetic perturbation effects in thousands of single cells[11–16]. Second, we describe the general concept of using quantitative multivariate readouts from large-scale genetic perturbation screens to infer functional interactions between phenotypic properties and between genes. Finally, we present an outlook on some of the future opportunities that the single-cell paradigm will bring to the unravelling of biological complexity from large-scale genetic perturbation screens.

Genetic perturbation screens

Traditionally, genetic perturbation approaches relied on random perturbations of the DNA of an organism or cells using chemical mutagens or random insertions[2,17], and are also termed forward genetics. These approaches create a null or mutated allele, with the latter causing either a constitutive or, sometimes, a conditional mutation such as a temperature-sensitive mutant protein. In the past decade, sequence-specific genetic perturbations (also termed reverse genetics), such as post-transcriptional gene perturbations by means of RNA interference (RNAi), have increasingly been performed, allowing large-scale targeted knockdown of specific mRNAs in *Caenorhabditis elegans* and *Drosophila melanogaster*, as well as in mammalian cells[10,18–27]. In addition, overexpression screens of either wild-type or mutated forms of genes, usually encoded as cDNAs from plasmids, have also been applied[28,29]. Chemical compound or inhibitor screens (also termed 'chemical genetics')[30], which usually rely on post-translational perturbations in which the activity or function of the protein is inhibited by a small molecule, are also now frequently applied[11,30]. Furthermore, there are multiple genome-editing approaches that target specific regions of the genome to create a null or mutated allele[31–33]. These genome-editing approaches have now become more efficient and can be applied at a large scale[34–39]. Recently, specific genome targeting approaches using single guide RNAs (sgRNAs) for the CRISPR–Cas9

Box 1 | **Cell-to-cell variability and the microenvironment**



Cell-to-cell variability refers to the phenomenon that no two genetically identical cells have identical behaviour and appearance. The extent and origins of cell-to-cell variability depend on the cellular activity that is compared between cells, but its study has revealed several common trends[88,95,141–144]. Chance, or stochasticity, has a considerable role in cellular processes that involve a small number of molecules. Small differences can lead to sizeable differences in cellular behaviour. However, in general, robustness in molecular mechanisms[145–149] buffers the intrinsic stochasticity of molecular processes, and extrinsic factors are found to explain the majority of total cell-to-cell variability. One major component of such extrinsic factors, particularly in adherent cells, is the microenvironment of individual cells. Even in environmentally controlled cell culture conditions, a growing population of adherent cells will be continuously subjected to changing microenvironments as a consequence of an increase in cell number together with increased cell adhesion and migration[79]. As the population size increases, so does the local cell density but with different rates for each single cell, and more cells will find themselves entirely surrounded by neighbouring cells. Genetic perturbations can alter the distribution of microenvironmental properties of single cells (see the figure; Perturbations A and B) to such an extent that they dominate the effect of a perturbation in genetic perturbation screens[14,16]. Although it is common practice to normalize cellular readouts obtained in high-throughput approaches for differences in the total cell number, the population context of individual cells can be substantially different even for populations with equal number of cells (see the figure). Therefore, multiparametric methods are required that correct for the influence of the population context at the single-cell level[14]. Owing to the technical challenges associated with measuring single-cell behaviour quantitatively for whole cell populations, many questions remain. There is no comprehensive understanding of which cellular processes are influenced by the population context, and few links between population context-dependent cellular processes and the *in vivo* multicellular programming of cells have been investigated so far[150,151].

Cellular heterogeneity
Similar to cell-to-cell variability. Sometimes, 'heterogeneity' is used to indicate multiple discrete phenotypes within a population, while 'variability' is used to indicate variation around a single phenotype. There is no consensus on which term to use in which occasion, and both terms are interchangeable.

Multivariate readouts
Phenotypic readouts consisting of multiple features of the cellular activity, state and microenvironment.

Functional interactions
A general term that incorporates protein–protein interactions, classical genetic interactions, regulatory interactions (such as kinase–substrate interactions) and phenotypic interactions.

(clustered regularly interspaced short palindromic repeat–CRISPR-associated protein 9) system have been repurposed to induce sequence-specific repression or activation of gene expression at a genome scale[39,40].

Different approaches have different effects. Every large-scale gene perturbation approach has its advantages and disadvantages (TABLE 1). In general, with most approaches there is a trade-off between specificity and duration in establishing a measurable perturbation[41,42]. Importantly, differences in the method of perturbation and the time required to establish the perturbation can result in differences in the observable effects. Another factor to bear in mind is that some enzymes can sufficiently perform their task in a cell even when they are reduced to a fraction of their normal concentration and, in this case, knockdown by RNAi may not result in a measurable effect, whereas a gene deletion will do so. In addition, the presence of an inhibited protein in a cell can have a different effect from that of the absence of the same protein[42]. Changes observed in a cell population after a few days of gene knockdown can also be very different from those observed in a cell population selected over the course of several weeks to harbour a specific

Table 1 | **Advantages and disadvantages of different genetic perturbation methods**

Perturbation	Level of perturbation	Advantages	Disadvantages
Haploid screens	DNA (random transposon insertion)	High specificity	Depends on selective pressure or sorting; cannot achieve single-cell resolution
Single-gene knockout	DNA	High specificity	Inefficiency in production; there can be adaptive mechanisms and off-target effects
Double-gene knockout	DNA	Experimental inference of genetic interactions	Exponential increase in experiment size
CRISPR-mediated gene knockout	DNA	High specificity and high efficiency of gene knockout	Adaptive mechanism and possible off-target effects
CRISPRi and CRISPRa	Transcription (mRNA)	High specificity and few off-target effects	–
RNAi screens (siRNA and shRNA screens)	mRNA	Experimentally accessible way to perform arrayed screens	Off-target effects and incomplete knockdown efficiency
Compound screens	Proteins	Short time of action	Poor specificity; not genome-wide

CRISPR, clustered regularly interspaced short palindromic repeat; RNAi, RNA interference; shRNA, short hairpin RNA; siRNA, small interfering RNA.

null allele and hence an adaptive phenotype. All of these factors need to be considered when designing genetic perturbation screens and analysing data produced from differing approaches.

Arrayed versus pooled screens. An important difference between genetic perturbation screens is whether the perturbation is applied all at once (pooled) or one by one (arrayed)[3,4,10,15,26,27,43–46] (FIG. 1a). In pooled screens, a mixture of single genetic perturbations — such as barcoded short hairpin RNAs (shRNAs)[47,48], sgRNAs for Cas9 (REFS 36,37) or random insertion of gene-trap cassettes[49] — or double genetic perturbations (using, for example, vectors that express 2 shRNAs[48]) is applied to one large population of cells. After positive selection of cell clones with a desired trait — such as resistance to a cytotoxic compound or pathogen, or presence of a selectable reporter — the original perturbations can be retrieved by means of sequencing, which allows the identification of the exact location of the genetic perturbation in each of the selected clones. Pooled screens are easy to carry out and therefore allow substantial up-scaling of the number of genetic perturbations tested, including the feasibility to carry out synthetic interaction screens at a large scale[48]. However, pooled screens rely on the use of a selective pressure, or the sorting of cells with a desired signal, to select relevant cells from a complex pool of perturbed cells, and on subsequent en masse identification of all selected perturbations (FIG. 1a). Therefore, perturbations that affect cell proliferation or viability are usually lost, as are perturbations that do not confer complete resistance to a selective pressure. In addition, pooled approaches currently do not have single-cell resolution and cannot obtain multivariate information from thousands of single cells subjected to the same perturbation, and they are not, or only to a limited extent, quantitative (FIG. 1b). By contrast, genetic screens in arrayed format do not have these disadvantages and can have numerous quantitative multivariate dimensions in the identification of hits and in the inference of genetic interaction (see Supplementary information S1 (table) for an overview of the multivariate dimensions currently used in arrayed screens and methods used for reducing dimensionality in single-cell experiments).

Large-scale collections of single-gene deletions. For several single-cell organisms, genome-wide collections of strains with viable single-gene deletions are available[50–52]. For mammalian organisms, such genome-wide resources do not currently exist, but there are efforts to create genome-wide collections of viable single-gene knockout mice and their cell lines[53–55]. Technologies of higher throughput are also emerging, which rely on the use of random mutational insertions in mammalian haploid cells[49,56–58] or gene-editing methods in mammalian haploid or diploid cells[34,36,37,39]. CRISPR–Cas9-mediated gene editing shows high efficiency in diploid cells and is easily targeted to a specific site in the genome by an sgRNA, and it is likely that this technology will become the method of choice for large-scale gene knockout screening in a variety of mammalian cells[36,37,39], including tissue culture cell lines, primary somatic cells and stem cells. In creating such large single-gene deletion collections, the biggest challenge will be to grow and assay each cell line in such a collection in parallel, which will require substantial efforts and more sophisticated automation and liquid-handling robotics than those currently used by most academic laboratories. Eventually, large-scale genetic perturbation screens in human cells may become most powerful when they can be applied in an arrayed format and rely on null deletions. A rapid, high-throughput one-by-one gene deletion approach in multiwell plates without the need for selection, similar to small interfering RNA (siRNA) and small compound screen methods, could provide such a solution in the future.

Synthetic interaction screens
Genetic screens in which two perturbations are combined to assess the possible synergistic and epistatic effects between the two genes perturbed.

Mammalian haploid cells
Mammalian cells that harbour only one copy of the genome.

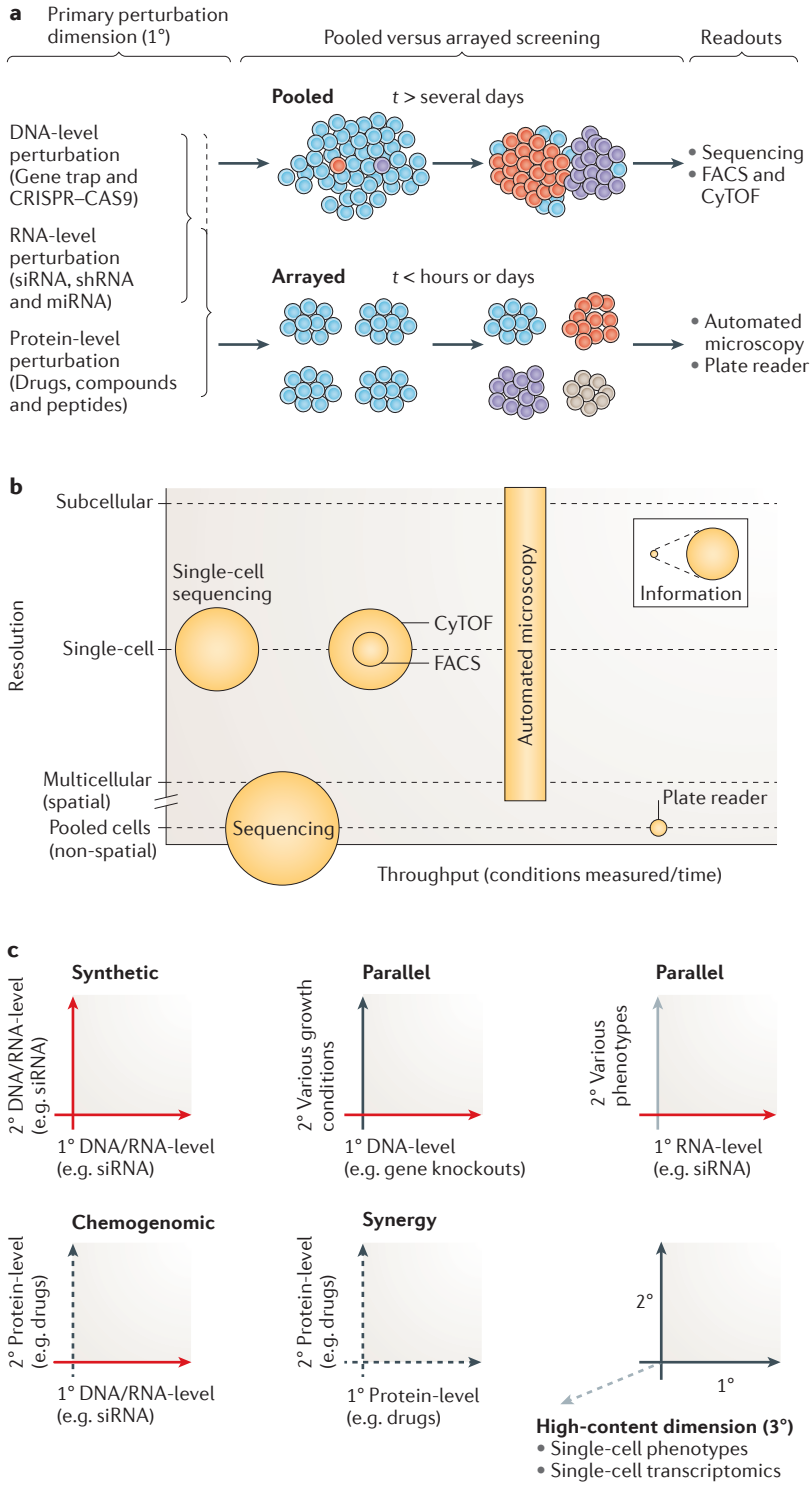


Figure 1 | **Multidimensional genetic perturbation screens.** **a** | The formats (pooled versus arrayed) and readouts routinely used in genetic perturbation screens are shown. **b** | The graph depicts the cellular resolution, throughput and quantitative information for different formats and readouts in genetic perturbation screens. **c** | Combinatorial possibilities for genetic perturbation screens and the inference of functional interactions from multidimensional data sets are shown. Depending on the two types of perturbation the system is subject to, different type of genetic interactions can be inferred. A third dimension that can be added to all previous systems is the single-cell dimension. CRISPR–Cas9, clustered regularly interspaced short palindromic repeat–CRISPR-associated protein 9; CyTOF, mass cytometry; FACS, fluorescence-activated cell sorting; miRNA, microRNA; shRNA, short hairpin RNA; siRNA, small interfering RNA.

Double-gene perturbations. In principle, any combination of gene perturbation methods can be applied to study how the combined effect of two perturbations aggravates or alleviates their respective single perturbation effects (FIG. 1c). Testing these epistatic effects between two gene perturbations is a classical approach in genetics, as it can reveal genes co-functioning in the same pathway or protein complex, but its large-scale application has traditionally been reserved to yeast[59–61]. However, approaches combining RNAi and/or chemical compound screening to evoke double-gene perturbations in mammalian cells[48,62–65], as well as the combination of large-scale insertional mutagenesis screening in mammalian haploid cells with cytotoxic drugs[66], are now increasingly being used. Clearly, pairwise double-gene perturbation screens rapidly become an enormous challenge; for example, 400 single-gene perturbations have 79,800 possible pairwise combinations. Double-gene perturbation is therefore usually either restricted to a preselected set of genes for which it would be interesting to comprehensively map their pairwise epistatic effects or applied in a pooled screening format[48,67,68] (see above).

Interpretation of genetic perturbation effects

Obviously, the interpretation of a genetic perturbation effect relies primarily on the readout. Given the complexity of cellular activities, the number of genes involved and the cell-to-cell variability that these activities can display, it is now becoming clear that measuring multiple aspects of cellular phenotype in many individual cells is necessary to avoid sampling bias or incorrect interpretation of perturbation effects. In addition, single-cell distributions of multivariate readouts allow sensitive detection of perturbation effects that occur in only a subset of cells and offer a better characterization of the effect itself.

Obtaining multivariate single-cell readouts. There are numerous ways to obtain a multivariate set of measurements from a large number of single cells, but two main approaches are practical when applied in large-scale genetic perturbation screens (FIG. 1b). These are flow cytometry and high-throughput imaging, and both methods have undergone rapid development.

In flow cytometry, the use of antibodies labelled with heavy isotope combinations and mass spectroscopy as the method of detection has allowed a marked increase in the extent of multiplexing[69]. This approach is termed mass cytometry (also known as CyTOF)[69] and has, in various studies, allowed the multiplexing of up to 35 different molecular readouts from thousands of single cells[70,71]. So far, this approach has been primarily applied to quantify the levels of signalling proteins and their phospho-specific modifications in single cells to reveal molecular heterogeneity in cancer, as well as heterogeneous adaptation of signalling during immune and drug responses[70,71].

In imaging, there have been important developments in computational methods to segment single cells within images of cells grown in culture or within their context

of an embryo or tissue, as well as in the ability to extract a large number of quantitative features from segmented cells and subcellular objects[72,73] (FIG. 2). This approach is also referred to as 'computer vision'. Although imaging lags behind flow cytometry in the number of molecular readouts that can be simultaneously measured in single cells, novel approaches[74–76] indicate that imaging may soon be able to achieve a similar or greater extent of molecular multiplexing in single cells with an unprecedented subcellular resolution (BOX 2). These technological advances in imaging can create vast amounts of data from which, besides quantifying the abundance of molecules or their activated (for example, phosphorylated) forms in single cells, much additional information can be obtained on the subcellular localization and patterning of the molecular signals[77,78]. Imaging can also provide information on the morphology and shape of single cells[12,15] and, importantly, on the relative location of a single cell with respect to other cells in a population, as well as on its microenvironment, such as bordering space without cells, the number of neighbours and local crowding[79] (FIG. 2). For the purpose of this Review, we further focus on image-based screens because this is the only technique that allows the analysis of the full spectrum of cell-to-cell variability and that has the highest spatial resolution (FIG. 1b).

Managing multivariate single-cell readouts. A crucial first step in arriving at high-quality multivariate single-cell information is eliminating technical artefacts. In imaging, this can be readily performed using approaches from artificial intelligence, such as machine learning[80,81]. Cell segmentation results can be evaluated using software tools that project segmentation outlines on the cells in images and, guided by this, classifiers can be trained to identify incorrectly segmented cells, out-of-focus cells, cells on image edges, cells with staining artefacts or, if necessary, typical cellular states that produce outlier values, such as mitotic and apoptotic cells[80,81]. It is not uncommon that this process removes up to one-third of all initially identified cells in a large-scale genetic perturbation screen[16]. This exclusion step is important because many statistical approaches using single-cell data, such as principal component analysis (PCA) or Gaussian mixture models (GMMs), are sensitive to outlier values produced by such artefacts.

The second step to consider is the dimensionality of the single-cell measurements (also known as features) (see Supplementary information S1 (table)). Many features correlate to such an extent that they basically hold identical information. Therefore, a step of data dimensionality reduction or feature elimination can often be applied to reduce the data complexity without losing information. Both linear methods (such as PCA) and nonlinear methods (such as t-distributed stochastic neighbour embedding (t-SNE)[82,83]) transform the multidimensional feature space into a space represented by fewer dimensions, which may allow an easier computational handling of the data[84,85]. As transformed dimensions can be difficult to interpret, an alternative approach is to perform iterative feature elimination without transforming the original feature space[85]. This approach can reveal features that introduce unwanted noise and thus weaken the computational identification of single-cell phenotypes, as well as uncover features that carry most of the information and thus strengthen such classifications[85].

Finally, it is often desired to perform classification of single-cell phenotypes, such as changes in cellular morphology[12,15], perturbations in cellular activities[4,11,16,43] or differences in patterns of intracellular organelles[27,46]. Typically, either unsupervised machine learning approaches (such as k-means clustering) or supervised machine learning approaches (such as support vector machine (SVM)) are used[85] (see Supplementary information S1 (table)). Supervised classification can be done by browsing through the images and recognizing, by eye, representative cells of different phenotypic classes, which are then used to train a machine learning classifier; alternatively, it can be done by using perturbations that result in a known single-cell phenotype and classifying all cells in the data set according to a set of known perturbations[4,12,15,46,86]. Although this usually results in great interpretability of the data, it has limitations. First, it assumes that each cell must belong to a predefined set of discrete classes and not part of a phenotypic continuum, which is particularly problematic for single cells that are on the boundary of two or more classes. Second, supervised approaches are not comprehensive because they overlook single-cell phenotypes that are not a priori known or expected. To address this issue, one can use unsupervised approaches that classify all single cells in a data set in an unbiased manner, such as single-cell clustering or GMM[87–91]. The challenge with these unsupervised approaches is that the single-cell clusters may not always be biologically interpretable and are less robust to 'noisy' features.

Both supervised and unsupervised classification approaches assume that the single-cell phenotypic space can be discretized. There is a natural inclination to discretize data because it greatly aids interpretability. Well-known concepts in genetics such as canalization suggest that discrete phenotypes may emerge, and bistable or multistable properties of some biological systems also support a discretized view[15,92]. However, discretization of high-quality single-cell data can also lead to a great loss of information and should, in our opinion, be used with care. Some properties of single cells may seem to be discrete in certain measurements, for example, being in the G1 or G2 phase of the cell cycle when measuring DNA content. However, the cell cycle represents a cycling phenotypic continuum in which certain transitions occur faster than others, and the cell-to-cell variability within each phase can be used as a proxy for the time spent in that phase[93]. Moreover, many single-cell measurements do not show discrete peaks within a distribution, such as single-cell endocytic activity[16]. By considering single cells on a phenotypic continuum, one opens up to the concept of valuing the whole spectrum of cell-to-cell variability as biologically meaningful. This concept is important when interpreting genetic perturbation effects (see below).

Computer vision
A field that processes, analyses and interprets images in order to produce numerical information.

Microenvironment
The local environment of a single cell within a population and their relative positioning to each other, such as the local crowding of cells, the amount of neighbours, whether cells face empty space on one site and cells on another site, and whether cells are solitary.

Cell segmentation
Automated detection and delineation of the outside of single cells and nuclei in microscope images.

Cellular states
A quantitative description of the physiological states of single cells reflected in, for instance, their sizes, shapes, cell cycle phases, senescence or other detectable readouts such as metabolic states.

Gaussian mixture models
(GMMs). Parametric probability density functions that fit the Gaussian distribution to the data set to model the presence of subpopulations within an overall population; they are represented as weighted sums of Gaussian component densities.

Support vector machine
(SVM). Supervised learning models that recognize patterns in data sets and that are used for classification and regression analyses.

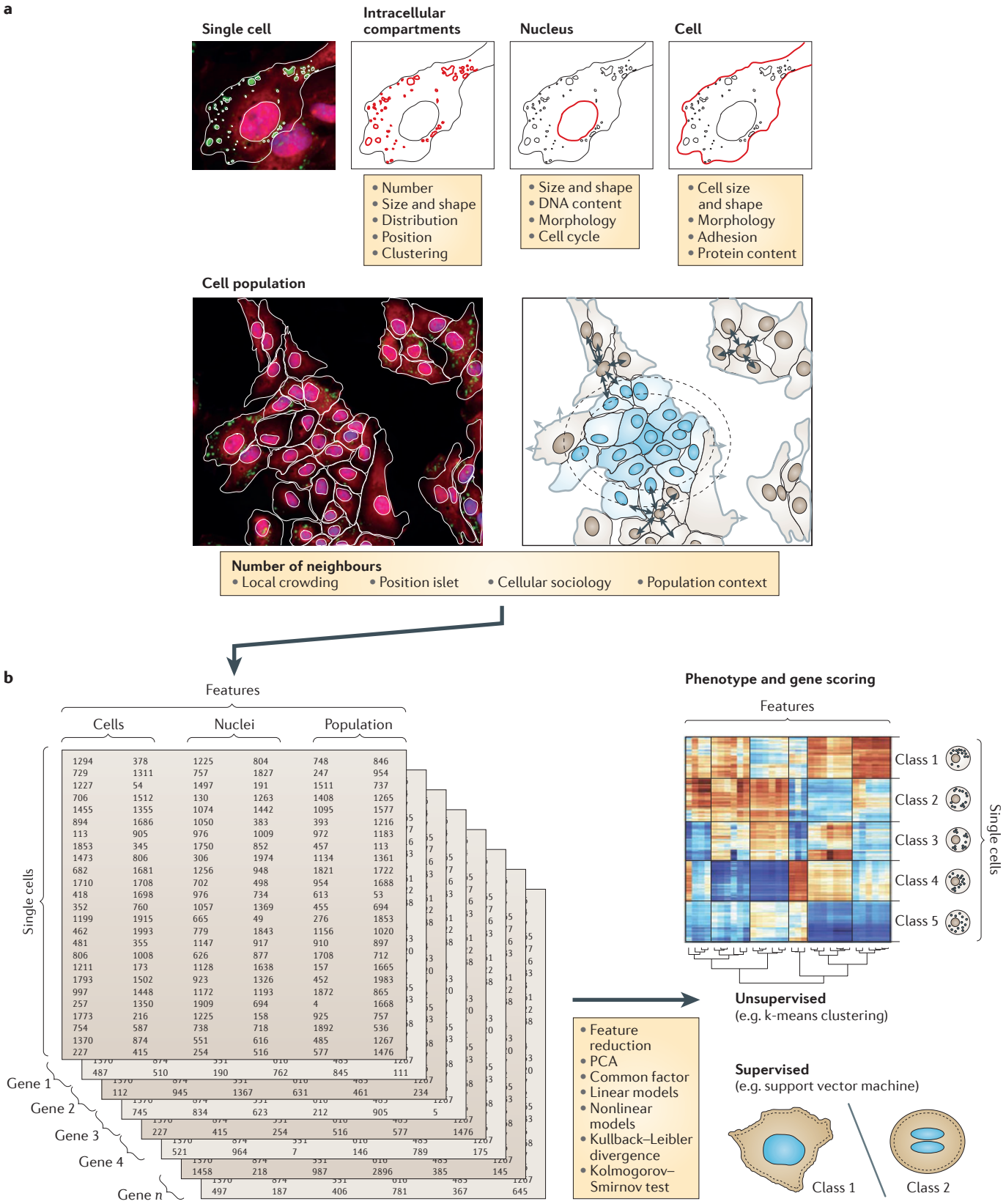


Figure 2 | **Single-cell measurements and phenotypic scoring.** **a** | The different types of measurements (features) that can be extracted from single cells and cell populations in images are illustrated. **b** | Possible dimensionality in multivariate data sets (left panel) is shown. Supervised and unsupervised methods can be used for dimensionality reduction and for scoring single-cell phenotypes and gene perturbations (right panel). Kullback–Leibler information divergence is a non-symmetrical measure that calculates the difference between two probability distributions. PCA, principal component analysis.

Box 2 | **Multiplexing molecular readouts in imaging**

In recent years, different approaches have been developed that allow the multiplexing of molecular readouts in single cells in imaging[74–76,152–154]. One approach is based on combinatorial labelling and spectral unmixing. This was applied in fluorescence *in situ* hybridization (FISH) on a complex population of multiple bacterial species, using simultaneous staining with 28 FISH probes that uniquely identified each species in the population[76]. In another approach, also using FISH, spatial barcoding of long probes with combinations of different fluorophores was applied, which could be identified using super-resolution microscopy. This approach allowed the multiplexing of 32 transcript readouts in single yeast cells[154]. It is conceivable that similar approaches can be developed for antibodies. However, it is likely that these methods of multiplexing will run into difficulties when the signals of multiple probes or antibodies overlap within the same single cells. These problems do not arise with an alternative principle for multiplexing, which relies on iterative staining and removal. In one approach, the fluorescence signal of one stain is photobleached, after which a second stain can be applied[74]. However, this becomes impractical when large numbers of single cells are imaged. Other approaches rely on antibody elution, in which both primary and secondary antibodies are removed by the use of detergent and low pH[75,155], or on the cleavage of oligonucleotides attached to antibodies[152]. Thus, it seems likely that, in the near future, these developments will allow multiplexing of up to 100 molecular readouts from thousands of single cells in an image-based approach. The information on the molecular state of single cells that is gained through this — combined with the power of quantifying numerous morphological, spatial and patterning features, both within single cells and across single cells — will empower multiple systems-biology approaches that no other method currently achieves. This would bridge the gap between omics and single-cell imaging.

The cell population context. Genetically identical cells from the same population cultured in the same medium can display a large spectrum of differences in phenotypes and activity[94–97] (FIG. 3a). In fact, the variability in single-cell properties within the same cell population can be larger than the difference of the mean of these activities between an unperturbed and a perturbed cell population[98]. Therefore, in pioneering work, the classical Kolmogorov–Smirnov test was used to compare single-cell feature distributions[11] (see Supplementary information S1 (table)). By performing such tests on multiple single-cell features separately for multiple-compound treatments across a range of different concentrations, this improved the ability to characterize drug responses and to assign mechanisms to uncharacterized drugs[11].

Although comparing full distributions instead of mean values offers a powerful approach, its greatest limitation is that it is not possible to determine which single cells are compared with each other. Cell-to-cell variability of different activities is determined by the physiological and morphological state of a cell (that is, the cellular state), which can be influenced by the microenvironment in a non-trivial manner[16,79,99–105] (FIG. 3a). As two single cells that are in different microenvironments and cellular states can both contribute to defining the same point in a one-dimensional distribution, perturbations that act in cells with different microenvironments or cellular states can have a similar effect on this distribution. We illustrate this for three theoretical single-cell features: the amount of an organelle such as the lysosome, the amount of a protein such as epithelial growth factor receptor (EGFR) and the amount of the lipid GM1 at the plasma membrane (FIG. 3a). In each scenario, we show what happens during perturbations if these features are either higher in cells that grow at high local crowding, or higher in big and spread-out cells that usually grow in regions of low local crowding. Comparing only one-dimensional single-cell distributions of activities would not distinguish between perturbations that alter the local crowding of cells or between perturbations that directly affect the specific single-cell features independently of possible effects on the cellular microenvironment[16] (FIG. 3b). To take this into account, one should compare multivariate single-cell distributions that incorporate features of the cellular state and the cellular microenvironment[13–16]. This allows one to distinguish a situation in which a single-cell activity distribution changed because the fraction of cells in certain microenvironments changed (that is, indirect perturbation) from a situation in which the single-cell activity was directly perturbed (that is, direct perturbation) (FIG. 3b). Taking patterns of cell-to-cell variability into account has been shown to be important for the interpretation of genetic perturbations of virus infection, endocytosis, membrane lipid composition and cell adhesion signalling[14,16,79]. Furthermore, it is likely to be important for studying other signal transduction pathways[94,97] and thus for studying gene transcription, protein translation, metabolic activity and possibly also intrinsically cycling cellular states such as the cell cycle[106].

In large-scale genetic perturbation screens, many perturbations can have effects on one or more aspects of the cell population context[14,16]. These effects cannot be predicted from only knowing the number of cells in a population because of the nonlinear emergence of the spectrum of single-cell microenvironments and states in a growing population of cells (BOX 1). Furthermore, some perturbations might affect different properties of the single-cell microenvironment and cellular state without affecting the number of cells, for example, by altering cell migration[14,16] (BOX 1). This possibility poses a serious problem for the comparison of genetic perturbation effects and cannot be addressed with a correction for trends in the single-cell readout as a function of cell number alone. Recently, computational approaches have been developed that allow a more direct comparison of genetic perturbation effects, provided that enough single cells are quantified for each perturbation in a large-scale genetic perturbation screen to permit statistical modelling[14,16].

Finally, cellular heterogeneity and the fact that genetic perturbations might affect different subsets of cells within one population could have important consequences for interpreting synthetic (that is, epistatic) effects between two perturbations. The complexity of this becomes very large if one takes into account that one perturbation may indirectly alter population context properties of single cells, which may allow another perturbation to exert an effect. However, such synthetic effects require a fundamentally different interpretation compared with current practice, which usually assumes that the targeted proteins are part of the same molecular complex or pathway acting in the same single cells. This population context effect may provide an explanation for some of the problematic complexity encountered in the analysis of multidimensional synthetic RNAi screens in human cells[64].

Kolmogorov–Smirnov test
A statistical non-parametric test for the comparison of continuous, one-dimensional probability distributions.

Population context
A collective term for the context in which a single cell displays its activities and behaviours, which can be determined by both local and global effects from the population to which the cell belongs. The context is determined not only by the microenvironment of a single cell but also by its physiological state that is a consequence of population effects, such as the cell size.

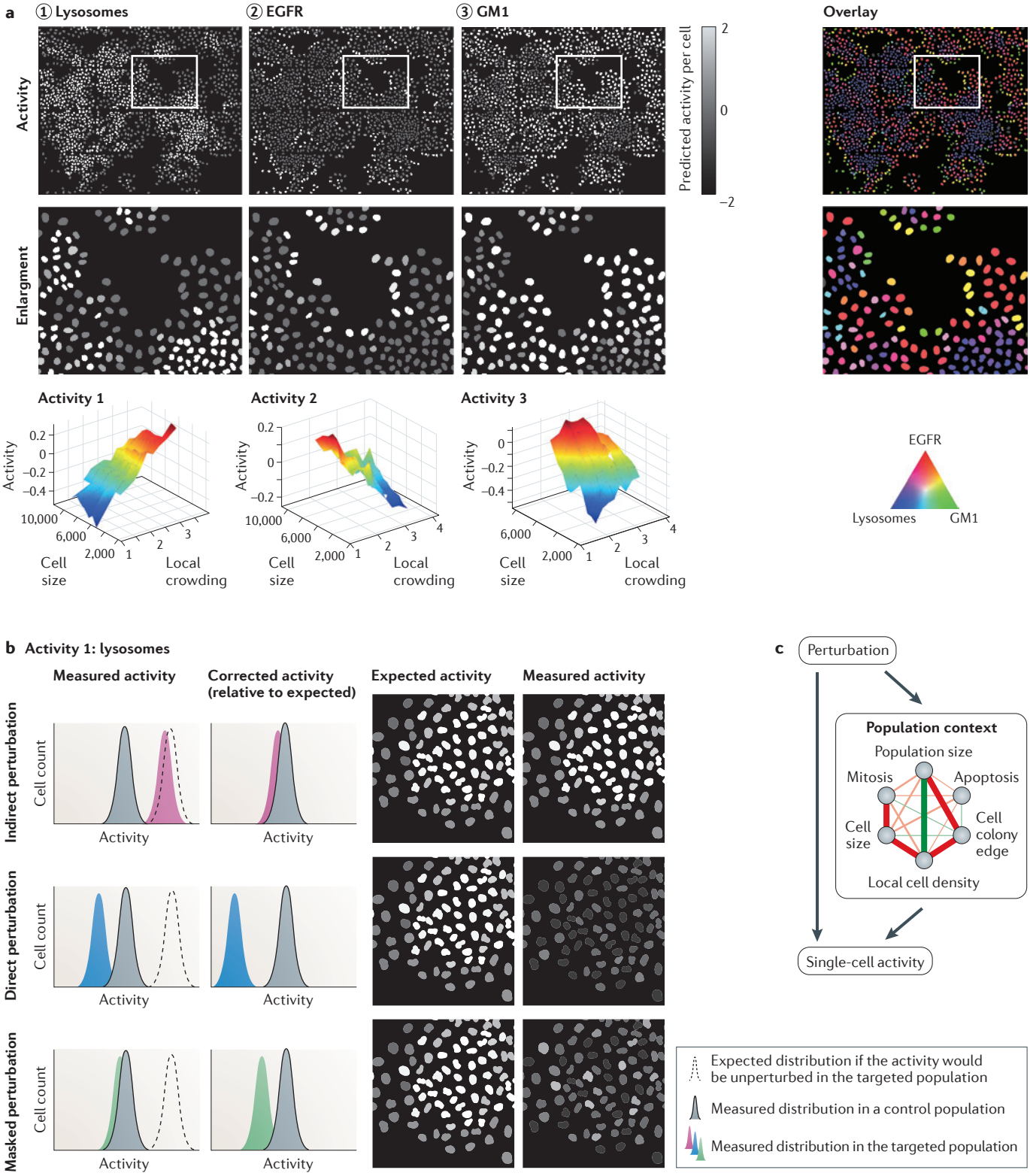


Figure 3 | **Accounting for population context in the interpretation of genetic perturbation screens.** **a** | Three theoretical single-cell features are shown: the amount of an organelle such as the lysosome, the amount of epithelial growth factor receptor (EGFR) at the plasma membrane and the amount of the lipid GM1 at the plasma membrane. Colour-coding on nuclear segmentation shows different and non-trivial patterns of cell-to-cell variability in a cell population. Three-dimensional (3D) surface plots below the images show predictors of the cell-to-cell variability patterns of the three cellular activities. **b** | Genetic perturbations can affect the population context of individual cells (for example, the crowdedness of cells in a population) and can directly or indirectly alter cellular activities. **c** | The different ways by which a genetic perturbation can affect single-cell activities are depicted. These can be mediated through indirect effects (via changes in population context) or through direct effects on the intracellular activity.

Accounting for the cell population context. The multivariate single-cell feature space of microenvironment and cellular state can often be used to predict single-cell activities in a non-parametric way[14,16,79]. For example, the amount of intracellular organelles of a single cell can be accurately predicted by the local crowding of the single cell[14,16]. Such predictive models can be learned from quantifying the cellular activity of interest in a large number of unperturbed cells plated over a wide range of cell numbers in a multiwell plate, and by creating quantile bins of cells within the multidimensional space describing microenvironment and cellular state[14,16]. For each perturbed cell population, such models predict what the expected single-cell activities, given their specific microenvironment and cellular state, are. This expected value is then subtracted from the measured single-cell activities. By correcting for this effect, it becomes possible to directly compare between two cell populations from the same cell line (and also, to a certain extent, from a different cell line), even if these cell populations show different distributions of single-cell microenvironments and cellular states (FIG. 3b). These effects have a large impact on the interpretation of results from a large-scale genetic perturbation screen[14,16]. Eliminating the effects that act through population context phenomena while maintaining single-cell autonomous effects leads to a greater enrichment of genes encoding direct regulatory and core machinery components of the cellular activity that is screened for[14,16]. In our three examples of cellular activity above (FIG. 3a), which show different patterns of cell-to-cell variability, indirect perturbation effects determined by population context could act in opposite directions on these readouts and could mask direct effects[16]. Without correction, these types of effects would dominate the interpretation of results, hiding core machinery that is potentially shared. Importantly, both types of effects are informative, and statistical approaches to separate them are merely a means to obtain more insights. Furthermore, such separations have their limitations, particularly when there is extensive and rapid feedback. However, as long as the direct and indirect effects caused by the same genetic perturbation show an additive effect and a large part of the spectrum of cellular microenvironments is present in the perturbed population, statistical approaches should be able to separate them. Finally, population context effects also occur in cell colonies of bacteria and yeast grown on agar, suggesting that such analysis may also be relevant for genetic perturbation screens in single-cell organisms[107–111].

Inferring functional interactions

To many researchers, the goal of large-scale genetic perturbation screens is to identify groups of functionally related genes and gene interactions. However, unbiased data-driven modelling on quantitative multivariate measurements from thousands of single cells at a large scale now allows many types of interactions to be identified from screens, revealing properties of the studied cellular activity at multiple levels. This is a systems-biology approach and goes beyond the simple creation of hit lists to identify gene candidates for further independent characterization. We describe below recent developments in the identification and interpretation of functional interactions between genes from perturbation screens. Finally, we discuss a different type of interaction that can be analysed from such screens — namely, between systems properties (for example, single-cell features) and perturbations.

Functional interactions between genes. The types of gene–gene interactions that can be identified from genetic perturbation screens are manifold. Here, we refer to functional interaction as a global term that incorporates both physical interactions between two protein subunits of the same complex (protein–protein interactions) and classical genetic interactions measured by epistasis between their respective loss-of-function effects (FIG. 4a). In our terminology, functional interactions also include regulatory interactions[16] such as kinase–substrate interactions and phenotypic interactions, which reflect the contribution of two genes to the emergence of a specific phenotype, without any direct physical or chemical interaction between the proteins (FIG. 4a).

The inference of functional interactions and, more specifically, classical genetic interactions has a long history. The current 'gold standard' in mapping classical genetic interactions at a large scale is by means of double-gene deletions, as these experimentally reveal epistatic effects between two gene perturbations[112,113]. The initial studies, pioneered in yeast, built a functional genetic interaction map based on the measured epistatic effects between two genes. However, it turned out that the pairwise correlation between two genes across a large set of epistatic effects with other genes is more informative for predicting functionally related genes[60,114]. Thus, statistical inference is used to derive an interaction between two genes from the similarity in their perturbation effects across a large number of genetic backgrounds (which are caused by the second gene perturbation). This concept is, in principle, not different from inferring functional genetic interactions from multiple single-gene perturbation screens performed in parallel using either multiple environmental conditions or multiple readouts[16,115]. Therefore, the large amount of information that can be extracted from single-gene perturbation screens using multivariate readouts of thousands of single cells per perturbation could allow the inference of genetic interactions of similar predictive power as inferred genetic interactions from double-gene perturbation screens[16]. In addition, as genetic interactions are readout-dependent (see below), the interaction inferred from parallel phenotypic screens may reveal novel biology that cannot be revealed by screens based on epistatic effects on colony fitness.

Genetic interactions are readout-dependent and plastic. It is important to realize that the readout used in large-scale genetic perturbation screens determines, to a large extent, the functional interactions that one finds and the novel biology that one uncovers[64,116]. Measuring colony fitness reveals interactions between genes that severely

Parallel phenotypic screens
Screens performed in parallel in the same cell line using the same perturbations but different phenotypic readouts.

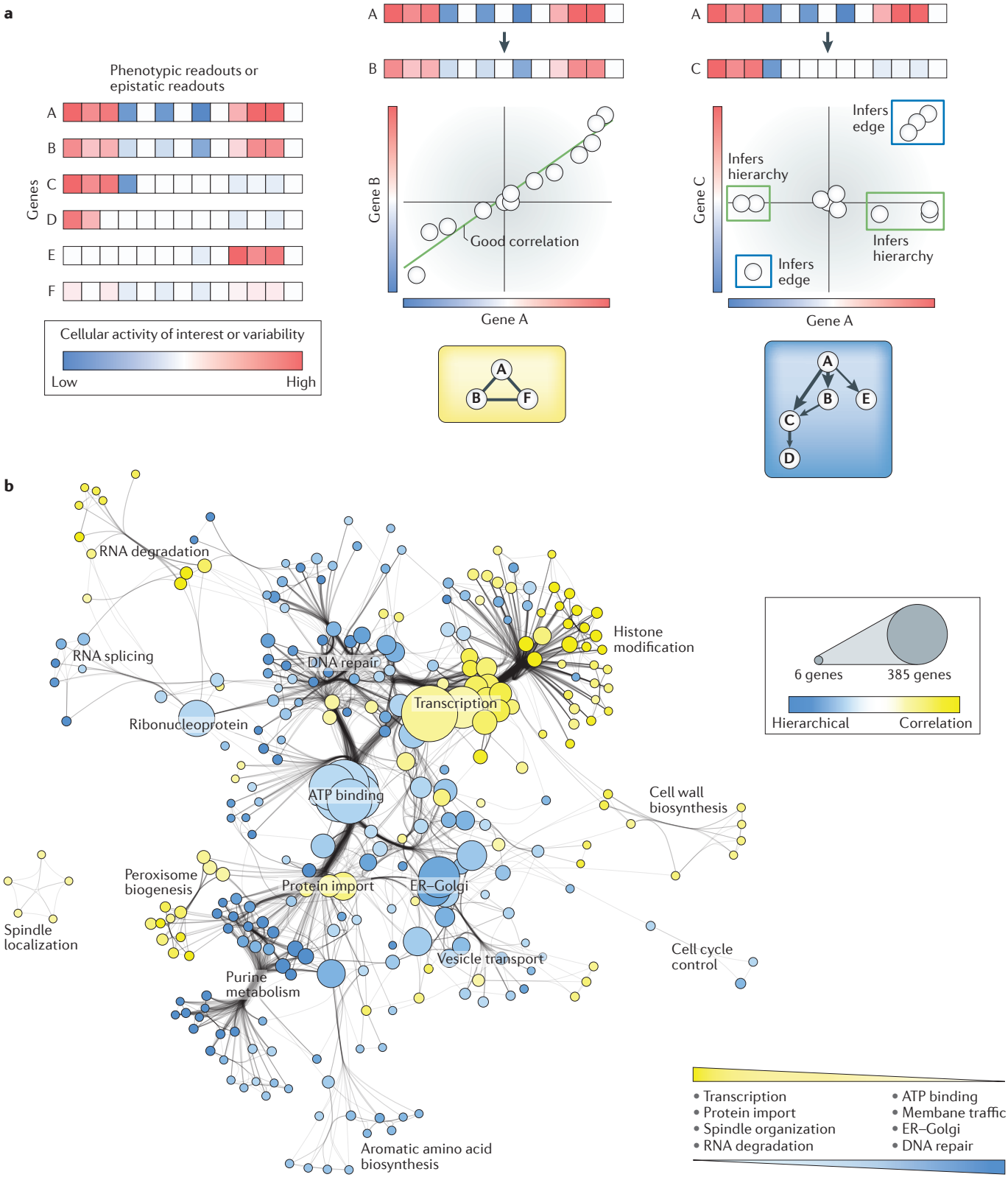


Figure 4 | **Inferring different types of genetic interactions.** **a** | Statistical inference of genetic interactions is shown. The schematic shows the mode of action of overall correlations (yellow) to infer classical genetic interactions versus subset effects in the data (for example, the Hierarchical Interaction Score (HIS); blue) to infer regulatory interactions between genes. **b** | Functional annotation enrichments in genes connected by the HIS are compared to those in genes connected by overall correlation inferred from a large double-gene perturbation screen in *Saccharomyces cerevisiae*[61]. Part **b** reprinted from *Cell*, **157**, Liberali, P., Snijder, B. & Pelkmans, L., A hierarchical map of regulatory genetic interactions in membrane trafficking, 1473–1487, Copyright (2014), with permission from Elsevier.

affect cell viability when perturbed[61], whereas measuring endocytosis will reveal interactions between genes that affect those processes when perturbed[16]. With the exception of a few protein complexes that are essential for any measurable feature of cells (such as ribosomes and proteasomes), these screens will reveal different subsets of relevant functional interactions. Furthermore, functional genetic interactions are highly plastic and can show substantial differences when cells are exposed to a single well-defined chemical perturbation[117–120]. Thus, there seems to be no single ground truth for functional genetic interaction maps, which reflects the plasticity of the underlying molecular networks.

Benchmarking functional genetic interactions. A wide variety of different omic approaches are currently being applied in numerous laboratories, leading to an explosive growth in information on the molecular networks of cells. This progress is being captured by numerous databases that contain various types of information about functional associations between genes, including co-expression from microarrays, protein–protein interactions and manually curated pathways[121–123]. As a result, there is an increasing tendency to compare results from a large-scale genetic perturbation screen with such databases, and computational approaches are developed to use such databases as a priori information in the analyses of large-scale genetic perturbation screens[124,125]. For example, iterative feature selection can be applied to compare the clustering of gene perturbations against such databases, selecting or scaling features to improve overlap[126]. However, the danger is that one could bias the results of a screen to reveal expected interactions by overfitting on the a priori data or by missing strong novel interactions arising from the data. Furthermore, these databases often generalize information and typically lack contextual and dynamic information on interactions, and are therefore currently far from being comprehensive enough to be a useful a priori source of information, especially when applied to areas of cell biology that do not immediately link to classical systematic readouts such as colony fitness or cell proliferation. However, we consider the various databases of omic information to be useful in determining general benchmarks for the predictive power of unbiased statistical methods.

Statistical inference of genetic interactions. In recent years, various statistical methods have been developed and used for the analysis of double-gene perturbation screens, and a broad set of modelling approaches have been developed and applied to such data[127–129]. This includes, for example, ordinary differential equations to model the final fitness measurements as the result of a dynamic growth process[130]. However, the vast majority of studies still rely on original clustering approaches, which are based on the pairwise correlations between two multivariate readouts of single-gene perturbations or epistatic effects. Clustering comes from the field of gene expression profiling, in which the clustering of genes with similar transcript abundance patterns in cells grown in different environmental or genetic backgrounds allows the grouping of genes with similar function[131,132]. Currently, clustering approaches using perturbation data sets are often combined with orthogonal types of large-scale data sets, For example, in an early study functional modules were identified from the combined evidence from the clustering of RNAi perturbation effects on early cell division in the *C. elegans* embryo, protein–protein interactions identified using yeast two-hybrid screens and clusters of mRNA expression profiles[133]. This approach mainly revealed well-known and tightly interconnected molecular complexes, such as the ribosome, the proteasome, the COPI coatomer, vacuolar H^+ ATPase and the anaphase-promoting complex[133].

The use of overall correlations in clustering is based on the assumption that two genes that co-function within the same molecular complex, biosynthetic pathway or regulatory mechanism must show the same perturbation effect (or synthetic effect) in many different conditions or in many different readouts. Although this is certainly true for many of the well-known interactions (such as those within the ribosome or proteasome), this is not a general rule. Many molecular complexes are known to be dynamic and show different compositions depending on the cellular activity in which they are involved, or depending on the (micro)environmental context and intrinsic physiological state of the cells in which they are acting. Furthermore, particularly in regulatory interactions, an upstream kinase often phosphorylates numerous proteins involved in different cellular activities. Therefore, perturbing this kinase might result in a broad set of effects across several different environmental or genetic backgrounds, multiple single cells in a population or multiple different readouts. Perturbing a downstream target of this kinase might only share these effects for a subset of the conditions or readouts and would thus demonstrate an overall poor correlation with the kinase[113,134]. Vice versa, it might be that a defined molecular complex is involved in multiple activities but is regulated by different upstream kinases in each of the activities. In this case, perturbing the molecular machinery would have a broader set of effects than perturbing one of the kinases. Moreover, a core set of molecular machinery components may use different subunits depending on the activity that the molecular machinery is involved in. Such combinatorial use of genes in cellular activities displays itself in large-scale genetic perturbation screens in a phenomenon called nesting, in which the effects of one perturbation are a subset within the effects of an upstream perturbation[135] (FIG. 4b). Several methods have been developed to analyse such subset or nested effects[136,137], but only recently have they been advanced to be applicable to large-scale genetic perturbation screens[16,138]. To capture these subset effects, the Hierarchical Interaction Score (HIS)[138] was developed, which connects two perturbations if they share a significant perturbation phenotype in at least one of a set of multivariate readouts (FIG. 4b). Moreover, the HIS includes directionality in the connection if one gene perturbation has a larger set of phenotypes than the other gene perturbation (with the upstream one having a larger set of phenotypes). When applied to a set of

Nesting
The phenomenon whereby the effects of a perturbation are a subset of the effects of another perturbation.

Hierarchical Interaction Score
(HIS). A statistical method that infers functional interactions between genes if they display perturbation effects in a consistent subset of readouts, or environmental or genetic backgrounds. It also infers statistical hierarchy, in which the perturbation with a broader set of effects is placed upstream of a perturbation with a narrower subset of these effects.

13 parallel RNAi screens measuring various aspects of endocytosis[16], this allowed the inference of regulatory genetic interactions between genes involved in signal transduction and membrane trafficking, and uncovered novel regulatory connections within the endosomal membrane system in human cells.

Interestingly, a comparison of functional genetic interactions inferred by the HIS and by Pearson correlation on the largest available collection of double-gene perturbations performed in yeast (*Saccharomyces cerevisiae*)[58] revealed that both methods infer complementary sets of functional interactions[16,61]. The HIS infers more interactions between genes that are enriched in vesicle transport, the endoplasmic reticulum and the Golgi, ATP binding (mainly kinases), DNA repair, aromatic amino acid biosynthesis, purine metabolism, ribonucleoprotein complexes and RNA splicing. By contrast, correlations mainly link genes that are enriched in transcription, histone modification, cell wall biosynthesis, peroxisome biogenesis, spindle localization and RNA degradation[16] (FIG. 4b). Apparently, some cellular processes rely more on combinatorial or hierarchical genetic interactions than others. As previous efforts in functional genetic interaction mapping have overlooked hierarchical interactions, the HIS provides an important additional approach to study the functional genetic landscape of cells.

Interactions between systems properties and perturbations. Besides functional interactions between genes, large-scale perturbation data sets can also be used to analyse overall trends in the data or to generalize whether certain features are co-perturbed in a multivariate feature set. Although machine learning, data dimensionality reduction and feature elimination are usually considered as data-processing steps, they do in fact provide a first overview of exactly such trends. They identify the features that correlate with each other across a large number of perturbations and show that certain combinations of feature values are more enriched in the data set than others. In addition, multiple forms of data-driven modelling — including partial correlation analysis, multivariate regression or Bayesian network learning — can be applied to multivariate single-cell measurements to reveal systems properties of a cellular activity and its response to genetic perturbations[12,14,45,79,88,139]. When data-driven modelling is applied across a large number of perturbations without a single-cell view (for example, by single-cell averaging the multiple features), this can reveal how certain properties of the cellular system under investigation are more often co-perturbed than other properties, which might indicate a causal chain of events[45]. Importantly, such systems properties can actually be revealed without the need for any perturbation by harnessing cell-to-cell variability — when enough single cells are quantified, correlations and causal interactions between properties can be inferred from the variability present within one cell population. If this cell-to-cell variability is combined with perturbations, then such correlations and causal interactions can become actual readouts in a screen, enabling the identification of genetic perturbations of systems properties. For example, this approach allowed the identification of genes perturbing the patterning of virus infection in a cell population, without necessarily changing the overall level of virus infection[14]. Without a single-cell analysis, this effect would have remained unnoticed. Such analyses are likely to become more predominant in the future, as they aid the identification of functional roles for genes in determining patterns of single-cell activities across cell populations, allowing the emergence of collective cellular behaviour, a fundamental property of all life forms.

Perspective

The ability to quantify multiple features of single cells and to identify multiple subclasses of single-cell phenotypes within a perturbed population has markedly increased the statistical power with which gene perturbations can be clustered[13,15,70,140], but there are several other implications that the single-cell paradigm might have on the inference of functional genetic interactions.

Not only is correcting for population context-determined effects important in the interpretation of single-gene perturbation effects, but it may also aid the interpretation of synthetic effects in double-gene perturbations[64,65]. The common practice of correcting for additive effects in synthetic interactions can account for the situation where two perturbations affect different subpopulation of cells, although it does not account for indirect synergies between two perturbations through population context-determined effects. One perturbation might change the microenvironment of single cells, which makes these cells sensitive to the second perturbation. Although such interactions are interesting, they are different from a direct functional interaction between two genes that are part of, for example, the same signalling pathway acting in the same single cell. The single-cell approach might also offer some unique advantages for inferring genetic interactions that have not yet been explored. If we consider all the single cells in a population to represent (slightly) different phenotypic backgrounds, then the comparison of two genetic perturbations across a multivariate single-cell space could be just as predictive of a functional interaction as comparisons across a double-gene perturbation space without a single-cell approach. This will require the quantification of a large number of single cells in each perturbed population, as well as a method that allows single cells with a similar phenotypic background to be identified and compared between two perturbed populations. The latter can be achieved, as outlined, with features of the single-cell microenvironment and state, which will become even more powerful when combined with molecular multiplexing (BOX 2).

More general regulators of a cellular activity often have a genetic perturbation effect in all cells of a population, whereas other components might only show a genetic perturbation effect in a subset of the population; in this case, statistical methods that analyse subset effects will be useful. Such population-determined subset effects may thus be derived from a varying functional

engagement of genes in a cellular activity across single cells in a population, and analyses of these patterns could reveal parallel pathways and regulatory networks depending on the single-cell microenvironment and state. Given that even classical functional genetic interactions change in the presence of a single chemical perturbation[118], it might be that different maps of functional genetic interactions will also be found within the same cell population depending on the single-cell microenvironment and state. Clearly, additional research will be required to investigate all of these aspects. With the single-cell paradigm now firmly established in biology and with rapid technological developments, it will soon be possible for research groups across disciplines to take the phenotypic spectrum of single cells into account in the interpretation of their experiments. This advance is inevitable and essential — we expect that it will improve interpretations of genetic perturbation screens and provide a rich ground for numerous novel biological discoveries for years to come.

1. Grunenfelder, B. & Winzeler, E. A. Treasures and traps in genome-wide data sets: case examples from yeast. *Nature Rev. Genet.* **3**, 653–661 (2002).
2. Adams, M. D. & Sekelsky, J. J. From sequence to phenotype: reverse genetics in *Drosophila melanogaster*. *Nature Rev. Genet.* **3**, 189–198 (2002).
3. Carpenter, A. E. & Sabatini, D. M. Systematic genome-wide screens of gene function. *Nature Rev. Genet.* **5**, 11–22 (2004).
4. Loo, L.-H., Wu, L. F. & Altschuler, S. J. Image-based multivariate profiling of drug responses from single cells. *Nature Methods* **4**, 445–453 (2007).
This paper identifies drug targets using a SVM-based method that takes into account the multivariate feature set of single cells in a population.
5. Boutros, M., Bras, L. P. & Huber, W. Analysis of cell-based RNAi screens. *Genome Biol.* **7**, R66 (2006).
6. Boutros, M. & Ahringer, J. The art and design of genetic screens: RNA interference. *Nature Rev. Genet.* **9**, 554–566 (2008).
7. Birmingham, A. *et al.* Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods* **6**, 569–575 (2009).
8. Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L. & Nolan, G. P. Computational solutions to large-scale data management and analysis. *Nature Rev. Genet.* **11**, 647–657 (2010).
9. Friedman, A. & Perrimon, N. Genome-wide high-throughput screens in functional genomics. *Curr. Opin. Genet. Dev.* **14**, 470–476 (2004).
10. Mohr, S., Bakal, C. & Perrimon, N. Genomic screening with RNAi: results and challenges. *Annu. Rev. Biochem.* **79**, 37–64 (2010).
11. Perlman, Z. E. *et al.* Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
This pioneering paper identifies drug targets by taking into account the full distribution of single cells in a population using the Kolmogorov–Smirnov test in image-based small-compound screens.
12. Bakal, C., Aach, J., Church, G. & Perrimon, N. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science* **316**, 1753–1756 (2007).
This study uses quantitative morphological features of single cells as profiles to identify genes involved in cellular morphology.
13. Loo, L.-H. *et al.* An approach for extensibly profiling the molecular states of cellular subpopulations. *Nature Methods* **6**, 759–765 (2009).
14. Snijder, B. *et al.* Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Systems Biol.* **8**, 579 (2012).
This paper shows that modelling single-cell behaviour and taking into account cell-to-cell variability strongly improve the data and comparability in siRNA screens.
15. Yin, Z. *et al.* A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. *Nature Cell Biol.* **15**, 860–871 (2013).
16. Liberali, P., Snijder, B. & Pelkmans, L. A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell* **157**, 1473–1487 (2014).
This paper infers regulatory genetic interactions from parallel siRNA screens in human cells and from double-knockout synthetic screens in yeast using the HIS.
17. Novick, P., Field, C. & Schekman, R. Identification of 23 complementation groups required for post-translational events in the yeast secretory pathway. *Cell* **21**, 205–215 (1980).
18. Gonczy, P. *et al.* Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408**, 331–336 (2000).
19. Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330 (2000).
20. Lum, L. *et al.* Identification of Hedgehog pathway components by RNAi in *Drosophila* cultured cells. *Science* **299**, 2039–2045 (2003).
21. Aza-Blanc, P. *et al.* Identification of modulators of TRAIL-induced apoptosis via RNAi-based phenotypic screening. *Mol. Cell* **12**, 627–637 (2003).
22. Brummelkamp, T. R., Nijman, S. M., Dirac, A. M. & Bernards, R. Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-κB. *Nature* **424**, 797–801 (2003).
23. Heo, W. D. & Meyer, T. Switch-of-function mutants based on morphology classification of Ras superfamily small GTPases. *Cell* **113**, 315–328 (2003).
24. Boutros, M. *et al.* Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303**, 832–835 (2004).
25. Paddison, P. J. *et al.* A resource for large-scale RNA-interference-based screens in mammals. *Nature* **428**, 427–431 (2004).
26. Neumann, B. *et al.* Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
This is the first genome-wide siRNA screen using time-lapse imaging of living cells, which extracted features from dynamic data to identify genes involved in mitosis.
27. Simpson, J. C. *et al.* Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nature Cell Biol.* **14**, 764–774 (2012).
28. Stevenson, L. F., Kennedy, B. K. & Harlow, E. A large-scale overexpression screen in *Saccharomyces cerevisiae* identifies previously uncharacterized cell cycle genes. *Proc. Natl Acad. Sci. USA* **98**, 3946–3951 (2001).
29. Pritsker, M., Ford, N. R., Jenq, H. T. & Lemischka, I. R. Genomewide gain-of-function genetic screen identifies functionally active genes in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA* **103**, 6946–6951 (2006).
30. Stockwell, B. R. Chemical genetics: ligand-based discovery of gene function. *Nature Rev. Genet.* **1**, 116–125 (2000).
31. Durai, S. *et al.* Zinc finger nucleases: custom-designed molecular scissors for genome engineering of plant and mammalian cells. *Nucleic Acids Res.* **33**, 5978–5990 (2005).
32. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nature Biotech.* **29**, 143–148 (2011).
33. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nature Rev. Genet.* **11**, 636–646 (2010).
34. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
35. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
This paper shows the first proof of principle of the CRISPR–Cas9 system in human cells.
36. Shalem, O. *et al.* Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
Using gene editing with the CRISPR–Cas9 system, this paper establishes genome-scale gene perturbation screening in a pooled format in human cancer and pluripotent stem cells.
37. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR–Cas9 system. *Science* **343**, 80–84 (2014).
Using gene editing with the CRISPR–Cas9 system, this paper establishes genome-scale gene perturbation screening in a pooled format in haploid and diploid cell lines.
38. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR–Cas9 for Genome engineering. *Cell* **157**, 1262–1278 (2014).
39. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
40. Gilbert, L. A. *et al.* Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
In references 39 and 40, the researchers repurposed the CRISPR system to induce sequence-specific repression (CRISPRi) or activation (CRISPRa) of gene expression at a genome scale.
41. Eggert, U. S., Field, C. M. & Mitchison, T. J. Small molecules in an RNAi world. *Mol. BioSystems* **2**, 93 (2006).
42. Weiss, W. A., Taylor, S. S. & Shokat, K. M. Recognizing and exploiting differences between RNAi and small-molecule inhibitors. *Nature Chem. Biol.* **3**, 739–744 (2007).
43. Bakal, C. *et al.* Phosphorylation networks regulating JNK activity in diverse genetic backgrounds. *Science* **322**, 453–456 (2008).
This paper reports a high-throughput screen that uses RNAi to systematically inhibit two genes simultaneously in 17,724 combinations to study kinase regulation.
44. Pelkmans, L. *et al.* Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature* **436**, 78–86 (2005).
This paper is the first to report parallel comparative siRNA screens.
45. Collinet, C. *et al.* Systems survey of endocytosis by multiparametric image analysis. *Nature* **464**, 243–249 (2010).
This paper reports the first multivariate genome-wide screen of endocytosis.
46. Chia, J. *et al.* RNAi screening reveals a large signaling network controlling the Golgi apparatus in human cells. *Mol. Systems Biol.* **8**, 1–33 (2012).
47. Silva, J. M. *et al.* Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617–620 (2008).
48. Bassik, M. C. *et al.* A systematic mammalian genetic interaction map reveals pathways underlying ricin susceptibility. *Cell* **152**, 909–922 (2013).
In this study, the researchers construct a double-shRNA library for pooled screens in human cells to identify genetic interaction between genes involved in ricin toxin susceptibility.
49. Carette, J. E. *et al.* Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326**, 1231–1235 (2009).
This is the first pooled screen in mammalian haploid cells using random mutational insertions.
50. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
51. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
This paper reports the construction of a collection of all viable single-gene deletion mutants of *S. cerevisiae*.

52. Kim, D. U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nature Biotech.* **28**, 617–623 (2010).
53. Tang, T. *et al.* A mouse knockout library for secreted and transmembrane proteins. *Nature Biotech.* **28**, 749–755 (2010).
54. Dolgin, E. Mouse library set to be knockout. *Nature* **474**, 262–263 (2011).
55. Skarnes, W. C. *et al.* A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* **474**, 337–342 (2011).
56. Kotecki, M., Reddy, P. S. & Cochran, B. H. Isolation and characterization of a near-haploid human cell line. *Exp. Cell Res.* **252**, 273–280 (1999).
57. Leeb, M. & Wutz, A. Derivation of haploid embryonic stem cells from mouse embryos. *Nature* **479**, 131–134 (2011).
58. Burckstummer, T. *et al.* A reversible gene trap collection empowers haploid genetics in human cells. *Nature Methods* **10**, 965–971 (2013).
59. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
In this study, the researchers develop a method for the systematic construction of a double-gene deletion library for synthetic screens in *S. cerevisiae*.
60. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005).
This paper introduces the use of pairwise correlations between two genes across a large set of epistatic effects to infer functional genetic interactions in *S. cerevisiae*.
61. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
In this study, researchers create a genome-scale genetic interaction map by examining 5.4 million gene–gene pairs for synthetic genetic interactions in *S. cerevisiae*.
62. Eggert, U. S. *et al.* Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.* **2**, e379 (2004).
63. Jiang, H., Pritchard, J. R., Williams, R. T., Lauffenburger, D. A. & Hemann, M. T. A mammalian functional-genetic approach to characterizing cancer therapeutics. *Nature Chem. Biol.* **7**, 92–100 (2011).
64. Laufer, C., Fischer, B., Billmann, M., Huber, W. & Boutros, M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nature Methods* **10**, 427–431 (2013).
This paper shows an arrayed synthetic screen in mammalian cells based on double-gene perturbation with RNAi and uses multiple readouts from single cells to infer genetic interactions.
65. Roguev, A. *et al.* Quantitative genetic-interaction mapping in mammalian cells. *Nature Methods* **10**, 432–437 (2013).
This paper shows an arrayed synthetic screen in mammalian cells based on double-gene perturbation with RNAi to infer genetic interactions, which are compared to protein–protein interaction data.
66. Reiling, J. H. *et al.* A haploid genetic screen identifies the major facilitator domain containing 2A (MFSD2A) transporter as a key mediator in the response to tunicamycin. *Proc. Natl Acad. Sci. USA* **108**, 11756–11765 (2011).
67. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic *KRAS*-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
This study uses siRNA screens to detect synthetic lethal partners of oncogenic *KRAS*.
68. Ashworth, A., Lord, C. J. & Reis, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145**, 30–38 (2011).
69. Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
70. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
71. Bodenmiller, B. *et al.* Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature Biotech.* **30**, 858–867 (2012).
72. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
73. Eliceiri, K. W. *et al.* Biological imaging software tools. *Nature Methods* **9**, 697–710 (2012).
74. Schubert, W. *et al.* Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nature Biotech.* **24**, 1270–1278 (2006).
75. Zrazhevskiy, P. & Gao, X. Quantum dot imaging platform for single-cell molecular profiling. *Nature Commun.* **4**, 1619 (2013).
76. Valm, A. M. *et al.* Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proc. Natl Acad. Sci. USA* **108**, 4152–4157 (2011).
77. Dehmelt, L. & Bastiaens, P. I. Spatial organization of intracellular communication: insights from imaging. *Nature Rev. Mol. Cell Biol.* **11**, 440–452 (2010).
78. Welch, C. M., Elliott, H., Danuser, G. & Hahn, K. M. Imaging the coordination of multiple signalling activities in living cells. *Nature Rev. Mol. Cell Biol.* **12**, 749–756 (2011).
79. Snijder, B. *et al.* Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461**, 520–523 (2009).
This paper shows that cell-to-cell variability in a population of monoclonal cells is not stochastic but can be predicted at the single-cell level by features of the cellular state and microenvironment.
80. Ramo, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009).
81. Jones, T. R. *et al.* CellProfiler Analyst: data exploration and analysis software for complex image-based screens. *BMC Bioinformatics* **9**, 482 (2008).
82. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
83. Hinton, G. & van der Maaten, L. Visualizing data using t-SNE. *J. Machine Learning Research* **9**, 2579–2605 (2008).
84. Li, L. Dimension reduction for high-dimensional data. *Methods Mol. Biol.* **620**, 417–434 (2010).
85. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification* (John Wiley, 2001).
86. Jones, T. R. *et al.* Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proc. Natl Acad. Sci. USA* **106**, 1826–1831 (2009).
87. Slack, M. D., Martinez, E. D., Wu, L. F. & Altschuler, S. J. Characterizing heterogeneous cellular responses to perturbations. *Proc. Natl Acad. Sci. USA* **105**, 19306–193011 (2008).
88. Singh, D. K. *et al.* Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. *Mol. Syst. Biol.* **6**, 639 (2010).
89. Zhong, Q., Busetto, A. G., Fededa, J. P., Buhmann, J. M. & Gerlich, D. W. Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nature Methods* **9**, 711–713 (2012).
90. Qiu, P. *et al.* Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotech.* **29**, 886–891 (2011).
91. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods* **10**, 1127–1133 (2013).
92. Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565 (1942).
93. Kafri, R. *et al.* Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. *Nature* **494**, 480–483 (2013).
94. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M. & Sorger, P. K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).
95. Altschuler, S. J. & Wu, L. F. Cellular heterogeneity: do differences make a difference? *Cell* **141**, 559–563 (2010).
96. Snijder, B. & Pelkmans, L. Origins of regulated cell-to-cell variability. *Nature Rev. Mol. Cell Biol.* **12**, 119–125 (2011).
97. Yuan, T. L., Wulf, G., Burga, L. & Cantley, L. C. Cell-to-cell variability in PI3K protein level regulates PI3K–AKT pathway activity in cell populations. *Curr. Biol.* **21**, 173–183 (2011).
98. Keren, K. *et al.* Mechanism of shape determination in motile cells. *Nature* **453**, 475–480 (2008).
99. Colman-Lerner, A. *et al.* Regulated cell-to-cell variation in a cell-fate decision system. *Nature* **437**, 699–706 (2005).
100. Castor, L. N. Flattening, movement and control of division of epithelial-like cells. *J. Cell. Physiol.* **75**, 57–64 (1970).
101. Eagle, H., Levine, E. M. & Koprowski, H. Species specificity in growth regulatory effects of cellular interaction. *Nature* **220**, 266–269 (1968).
102. Eagle, H. & Levine, E. M. Growth regulatory effects of cellular interaction. *Nature* **213**, 1102–1106 (1967).
103. Zeng, L. *et al.* Decision making at a subcellular level determines the outcome of bacteriophage infection. *Cell* **141**, 682–691 (2010).
104. St-Pierre, F. & Endy, D. Determination of cell fate selection during phage λ infection. *Proc. Natl Acad. Sci. USA* **105**, 20705–20710 (2008).
105. Robert, L. *et al.* Pre-dispositions and epigenetic inheritance in the *Escherichia coli* lactose operon bistable switch. *Mol. Syst. Biol.* **6**, 357 (2010).
106. Pelkmans, L. Cell Biology. Using cell-to-cell variability — a new era in molecular biology. *Science* **336**, 425–426 (2012).
107. Halme, A., Bumgarner, S., Styles, C. & Fink, G. R. Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast. *Cell* **116**, 405–415 (2004).
108. Avery, S. V. Microbial cell individuality and the underlying sources of heterogeneity. *Nature Rev. Microbiol.* **4**, 577–587 (2006).
109. Vlamakis, H., Aguilar, C., Losick, R. & Kolter, R. Control of cell fate by the formation of an architecturally complex bacterial community. *Genes Dev.* **22**, 945–953 (2008).
110. Parsons, B. D., Schindler, A., Evans, D. H. & Foley, E. A. Direct phenotypic comparison of siRNA pools and multiple individual duplexes in a functional assay. *PLoS ONE* **4**, e8471 (2009).
111. Cap, M., Stepanek, L., Harant, K., Vachova, L. & Palkova, Z. Cell differentiation within a yeast colony: metabolic and regulatory parallels with a tumor-affected organism. *Mol. Cell* **46**, 436–448 (2012).
112. Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B. & Boone, C. Systematic mapping of genetic interaction networks. *Annu. Rev. Genet.* **43**, 601–625 (2009).
113. Collins, S. R., Roguev, A. & Krogan, N. J. *Quantitative Genetic Interaction Mapping Using the E-MAP Approach* Ch. 9 (Elsevier, 2010).
114. Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813 (2004).
115. Nichols, R. J. *et al.* Phenotypic landscape of a bacterial cell. *Cell* **144**, 143–156 (2011).
This study combines large-scale chemical genomics with quantitative fitness measurements in hundreds of parallel conditions in *Escherichia coli*.
116. Horn, T. *et al.* Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. *Nature Methods* **8**, 341–U391 (2011).
117. Guénolé, A. *et al.* Dissection of DNA damage responses using multiconditional genetic interaction maps. *Mol. Cell* **49**, 346–358 (2013).
118. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385–1389 (2010).
119. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Systems Biol.* **8**, 565 (2012).
120. Santos, S. D. M., Verveer, P. J. & Bastiaens, P. I. H. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biol.* **9**, 324–U139 (2007).
121. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* **4**, 44–57 (2009).
122. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
123. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
124. Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Rev. Genet.* **12**, 56–68 (2011).
125. Markowetz, F. How to understand the cell by breaking it: network analysis of gene perturbation screens. *PLoS Comput. Biol.* **6**, e1000655 (2010).
126. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Machine Learn. Res.* **3**, 1157–1182 (2003).

127. Wang, L., Wang, X., Arkin, A. P. & Samoilov, M. S. Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics* **29**, 338–346 (2013).
128. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nature Rev. Genet.* **8**, 437–449 (2007).
129. Battle, A., Jonikas, M. C., Walter, P., Weissman, J. S. & Koller, D. Automated identification of pathways from quantitative genetic interaction data. *Mol. Syst. Biol.* **6**, 379 (2010).
130. Wang, L. M., Wang, X. D., Arkin, A. P. & Samoilov, M. S. Inference of gene regulatory networks from genome-wide knockout fitness data. *Bioinformatics* **29**, 338–346 (2013).
131. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
132. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
133. Gunsalus, K. C. *et al.* Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436**, 861–865 (2005).
134. Fiedler, D. *et al.* Functional organization of the *S. cerevisiae* phosphorylation network. *Cell* **136**, 952–963 (2009).
This paper analyses synthetic interactions between gene knockouts of kinases, phosphatases and their substrates in *S. cerevisiae*. It shows that kinases, phosphatases and their substrates have positive epistatic interactions between each other but no significant correlation between their epistatic effect profiles.
135. Boutros, M., Agaisse, H. & Perrimon, N. Sequential activation of signaling pathways during innate immune responses in *Drosophila*. *Dev. Cell* **3**, 711–722 (2002).
136. Markowetz, F., Bloch, J. & Spang, R. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* **21**, 4026–4032 (2005).
137. Markowetz, F., Kostka, D., Troyanskaya, O. G. & Spang, R. Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* **23**, i305–i312 (2007).
138. Snijder, B., Liberali, P., Frechin, M., Stoeger, T. & Pelkmans, L. Predicting functional gene interactions with the hierarchical interaction score. *Nature Methods* **10**, 1089–1092 (2013).
139. Young, D. W. *et al.* Integrating high-content screening and ligand–target prediction to identify mechanism of action. *Nature Chem. Biol.* **4**, 59–68 (2007).
140. Irish, J. M. *et al.* Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217–228 (2004).
141. Elowitz, M. B. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
142. Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
143. Raj, A., Rifkin, S. A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–918 (2010).
144. Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* **336**, 183–187 (2012).
145. Stelling, J. *et al.* Robustness of cellular functions. *Cell* **118**, 675–685 (2004).
146. Macarthur, B. D., Ma'ayan, A. & Lemischka, I. R. Systems biology of stem cell fate and cellular reprogramming. *Nature Rev. Mol. Cell Biol.* **10**, 672–681 (2009).
147. Barad, O. *et al.* Robust selection of sensory organ precursors by the Notch-δ pathway. *Curr. Opin. Cell Biol.* **23**, 663–667 (2011).
148. Ribrault, C., Sekimoto, K. & Triller, A. From the stochasticity of molecular processes to the variability of synaptic transmission. *Nature Rev. Neurosci.* **12**, 375–387 (2011).
149. Brandman, O. & Meyer, T. Feedback loops shape cellular signals in space and time. *Science* **322**, 390–395 (2008).
150. Connelly, J. T. *et al.* Actin and serum response factor transduce physical cues from the microenvironment to regulate epidermal stem cell fate decisions. *Nature Cell Biol.* **12**, 711–718 (2010).
151. Engler, A. J., Sen, S., Sweeney, H. L. & Discher, D. E. Matrix elasticity directs stem cell lineage specification. *Cell* **126**, 677–689 (2006).
152. Ullal, A. V. *et al.* Cancer cell profiling by barcoding allows multiplexed protein analysis in fine-needle aspirates. *Sci. Transl Med.* **6**, 219ra9 (2014).
153. Jungmann, R. *et al.* Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nature Methods* **11**, 313–318 (2014).
154. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nature Methods* **9**, 743–748 (2012).
155. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue. *Proc. Natl Acad. Sci. USA* **110**, 11982–11987 (2013).

Acknowledgements
The authors thank M. Muellner and S. Nijman for the images in the Box 1, and all members of the Pelkmans laboratory for discussions. P.L. is supported by a FEBS postdoctoral fellowship. B.S. is supported by an advanced postdoc fellowship of SNSF. Research in the Pelkmans laboratory on these topics is funded by the University of Zürich Research Priority Program (URPP) in Functional Genomics and Systems Biology, and the Swiss initiative in Systems Biology: Systemsx.ch.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
CellProfiler: http://www.cellprofiler.org/
DAVID: http://david.abcc.ncifcrf.gov/
Endocytome: http://www.endocytome.org/
GenomeRNAi: http://www.genomernai.org/
HIS: http://www.his2graph.net/
Infectome: http://www.infectome.org/
MitoCheck: http://www.mitocheck.org/
RNAi Database: http://rnai.org/
***Saccharomyces* Genome Database:** http://www.yeastgenome.org/
STRING: http://string-db.org/
Systems survey of endocytosis by multiparametric image analysis: http://endosomics.mpi-cbg.de/

SUPPLEMENTARY INFORMATION
See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF