

Morphological map of under- and overexpression of genes in human cells

Received: 2 December 2024

Accepted: 10 June 2025

Published online: 7 August 2025

 Check for updates

Srinivas Niranj Chandrasekaran¹, Eric Alix², John Arevalo¹, Adriana Borowa³, Patrick J. Byrne¹, William G. Charles⁴, Zitong S. Chen¹, Beth A. Cimini¹, Boxiong Deng⁵, John G. Doench¹, Jessica D. Ewald¹, Briana Fritchman¹, Colin J. Fuller⁶, Jedidiah Gaetz⁷, Amy Goodale¹, Marzieh Haghighi¹, Yu Han¹, Zahra Hanifehlo⁸, Holger Hennig⁴, Desiree Hernandez¹, Christina B. Jacob⁸, Tim James⁴, Tomasz Jetka³, Alexandr A. Kalinin¹, Ben Komalo⁶, Maria Kost-Alimova¹, Tomasz Krawiec³, Brittany A. Marion¹, Glynn Martin², Nicola Jane McCarthy², Lisa Miller¹, Arne Monsees⁴, Nikita Moshkov¹, Alán F. Muñoz¹, Arnaud Ogier⁸, Magdalena Otrocka³, Krzysztof Rataj³, David E. Root¹, Francesco Rubbo⁶, Simon Scrace², Douglas W. Selinger⁷, Rebecca A. Senft¹, Peter Sommer⁸, Amandine Thibaudeau⁸, Sarah Trisorus⁶, Rahul Valiya Veetil⁴, William J. Van Trump⁶, Sui Wang⁵, Michał Warchot³, Erin Weisbart¹, Amélie Weiss⁸, Michael Wiest⁶, Agata Zaremba³, Andrei Zinovyev⁴, Shantanu Singh¹✉ & Anne E. Carpenter¹✉

Cell Painting images offer valuable insights into a cell's state and enable many biological applications, but publicly available arrayed datasets only include hundreds of genes perturbed. The JUMP Cell Painting Consortium perturbed roughly 75% of the protein-coding genome in human U-2 OS cells, generating a rich resource of single-cell images and extracted features. These profiles capture the phenotypic impacts of perturbing 15,243 human genes, including overexpressing 12,609 genes (using open reading frames) and knocking out 7,975 genes (using CRISPR–Cas9). Here we mitigated technical artifacts by rigorously evaluating data processing options and validated the dataset's robustness and biological relevance. Analysis of phenotypic profiles revealed previously undiscovered gene clusters and functional relationships, including those associated with mitochondrial function, cancer and neural processes. The JUMP Cell Painting genetic dataset is a valuable resource for exploring gene relationships and uncovering previously unknown functions.

A major biological challenge is to identify the functions of all human proteins and to understand the impact of disease on those functions. Historically, human protein functions are determined painstakingly, often through hypothesis-driven and heavily customized genetic or proteomic experiments. The pace of discovery increased with the rise of genome-scale perturbation screens that simultaneously assess all human proteins for a given function. First in arrayed format and later in a more economical pooled format, these experiments capture

readouts (reflecting phenotypes of interest) from reporter assays, molecular omics or microscopy. They can identify all genetic perturbations that affect a single or a few biological processes of interest, within limits of technical noise^{1–4}.

As technologies increased the scale of functional genomics experiments, they also increased the breadth of readouts measured from a sample. Proteomic, metabolomic and transcriptomic techniques allow measuring hundreds to thousands of individual molecular readouts.

A full list of affiliations appears at the end of the paper. ✉ e-mail: shantanu@broadinstitute.org; anne@broadinstitute.org

Profiling technologies enable rapid discovery of new protein functions in two ways. First, researchers can identify, in a single experiment, all genes whose perturbation affects each of the hundreds of individual direct readouts in the assay. Second, an entirely new mode of discovery arose: using the complex patterns among the hundreds of readouts as a 'profile' or signature to assign functions to genes based on 'guilt-by-similarity'.

Images of cells captured by fluorescence microscopy have emerged as a powerful profiling methodology alongside molecular profiling. Image analysis extracts thousands of morphological features⁵ and captures biological information on par with available high-throughput proteomic or transcriptomic methods^{6,7}, at single-cell resolution^{8–10}. Using images to generate profiles has proved useful for identifying gene functions, determining the mechanism of compounds and identifying previously unknown chemical regulators of genes, among many other applications in drug discovery and fundamental biological research¹¹.

Accordingly, information-rich imaging datasets have been publicly shared¹² that capture morphological profiles of thousands of genetic perturbations. An early Mitocheck Consortium screen decreased the expression of 21,000 human genes by RNA interference, capturing the impact on chromosome morphology in time lapse¹³. Multiple studies using deletion strains in yeast have captured the morphological impact^{14,15}. Morphological profiles of actin, DNA and α -tubulin staining were captured for 6,840 *Drosophila* genes perturbed by RNA interference, plus pairwise combinations of those genes with a 100-gene subset¹⁶.

Cell Painting has become fairly standard as an image-based profiling readout. Thus far, the largest arrayed genetic perturbation dataset using the assay is from the techbio company Recursion, using six CRISPR–Cas9 guides against 736 genes in human umbilical vein endothelial cells (with an additional 16,327 genes anonymized)¹⁷. New methods allow pooling thousands of genetic reagents in human cells while still offering image-based readouts; we helped execute three Cell Painting screens using CRISPR–Cas9 guides to knockdown each of >20,000 human genes, although an average of only 460 cells' data are available per gene (<100 per guide) and no data from overexpression perturbations are available¹⁸.

Here we aimed to create a public, large-scale dataset of Cell Painting image-based profiles from increasing and decreasing gene expression, in an arrayed format that allows thousands of cells per replicate. Over- and underexpression of proteins can both have advantages for affecting cell systems and suggesting biological functions: unlike knockdown or knockout technologies, overexpression can produce phenotypes even in (perhaps, especially in) a cell type where the gene is not expressed or when there are multiple isoforms or highly redundant protein family members. Overexpression does not suffer from off-target effects and chromosome arm effects that can affect CRISPR–Cas9 systems¹⁹ but is prone to artifactual effects: although some, such as dominant-negative effects, may still affect a biological function specifically enough to be informative²⁰. This dataset represents the genetic subset of the 136,000 chemical and genetic perturbations tested by our Joint Undertaking in Morphological Profiling (JUMP) Cell Painting Consortium, which was previously released alongside a data descriptor preprint²¹ that describes it. In this paper, we additionally validate the dataset and demonstrate how the data can be used.

Results

Genetic perturbations yield Cell Painting phenotypes

Using the Cell Painting assay v.3 (ref. 22), we tested 15,142 overexpression (open reading frame (ORF)) reagents encompassing 12,609 unique human genes, including controls from the Broad Institute lentiviral ORF library²³. We also tested 7,977 pools of CRISPR guides against 7,975 unique genes and controls from the Human Edit-R synthetic CRISPR RNA (crRNA)-Druggable Genome library from Revvity

Discovery Limited (formerly known as Horizon Discovery). In total, we tested 15,243 genes by overexpression, knockout or both. Lentiviruses containing each ORF, or pools of four CRISPR guides against a given gene, were arrayed into 384-well plates of U-2 OS human osteosarcoma cells. The cells were incubated for 48 hours to allow infection (or transfection) and expression (or transient knockout) and then fixed and stained for the Cell Painting assay, with six stains labeling eight cellular components or organelles and imaged in five fluorescent channels (Fig. 1a). The design of the experiment is shown in Fig. 1b and described in detail in Methods. We extracted 7,648 and 4,761 image-based features from the ORF and CRISPR datasets, respectively (https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/tree/v1.0.0/11.list-all-features/output). The features include metrics of size, shape, intensity, texture and correlation, measured from each of the channels and in the nucleus, cytoplasm and cell compartments. Quality control, feature selection and other preprocessing steps (Methods) yielded final profiles with 722 and 259 features for ORF and CRISPR profiles, respectively. Notably, the sphering and harmony steps used in preprocessing yield final features that cannot be readily mapped to the original feature categories, making interpretability challenging.

Overall, of the 15,243 genes tested (Fig. 2a), 68% (10,352) yielded a detectable phenotype ('phenotypic activity'), that is, an image-based profile with a signal distinct from negative controls by ORF, CRISPR or both²⁴. This includes 7,031 genes (56% of tested genes) by overexpression and 5,546 genes (70% of tested genes) by CRISPR–Cas9 knockout (Fig. 2b,c and Supplementary Table 1). Focusing on just the reagents that yielded a phenotype gave a visual overview of the dataset; many clusters represent groups of genes whose products belong to the same protein complexes (Supplementary Fig. 1).

Genes with Cell Painting phenotypes share characteristics

We performed Fisher's exact test to discover whether particular types of gene were more likely to have phenotypic activity in Cell Painting (Fig. 2d). As expected, essential genes are more likely than random genes to have a phenotype when knocked out and less likely when overexpressed; essential genes as a class are already well-expressed, so overexpression may not dramatically affect the total amount of protein. Likewise, knocking out highly expressed genes is more likely to yield a phenotypic effect than overexpressing such genes. Overexpressing disease-associated genes were more likely than nondisease-related genes to produce phenotypic changes. This has implications for drug discovery, as small molecules that reverse these phenotypic changes might be pursued as disease therapeutics.

We also tested whether particular classes of genes, as defined by the Human Protein Atlas²⁵, were more likely to yield Cell Painting phenotypes (Fig. 2e). For gene knockouts, enzymes were more likely than random genes to have phenotypic activity, while predicted secreted and membrane proteins were less likely. For overexpression, predicted secreted proteins and T cell receptors were less likely to have phenotypic activity, while predicted membrane proteins were more likely.

Cell Painting gene links are validated by existing knowledge

To assess the value of image-based profiles in representing cell states, we investigated whether genes within the same annotated group were more similar than genes from different groups (Fig. 3a–d). In both the ORF and CRISPR datasets, genes cluster based on their associated CORUM protein complex, WikiPathway and gene group. Notably, several clusters show phenotypic consistency in both ORF and CRISPR datasets (Supplementary Table 4). Genes do not cluster by disease association, possibly because diverse mechanisms can underpin a single disease.

Next, we compared gene–gene morphological connections to the Drug Repurposing Knowledge Graph (DRKG), which integrates multiple sources of existing connections among genes²⁶ (<https://github.com/gnn4dr/DRKG>). The DRKG includes protein–protein interactions, gene

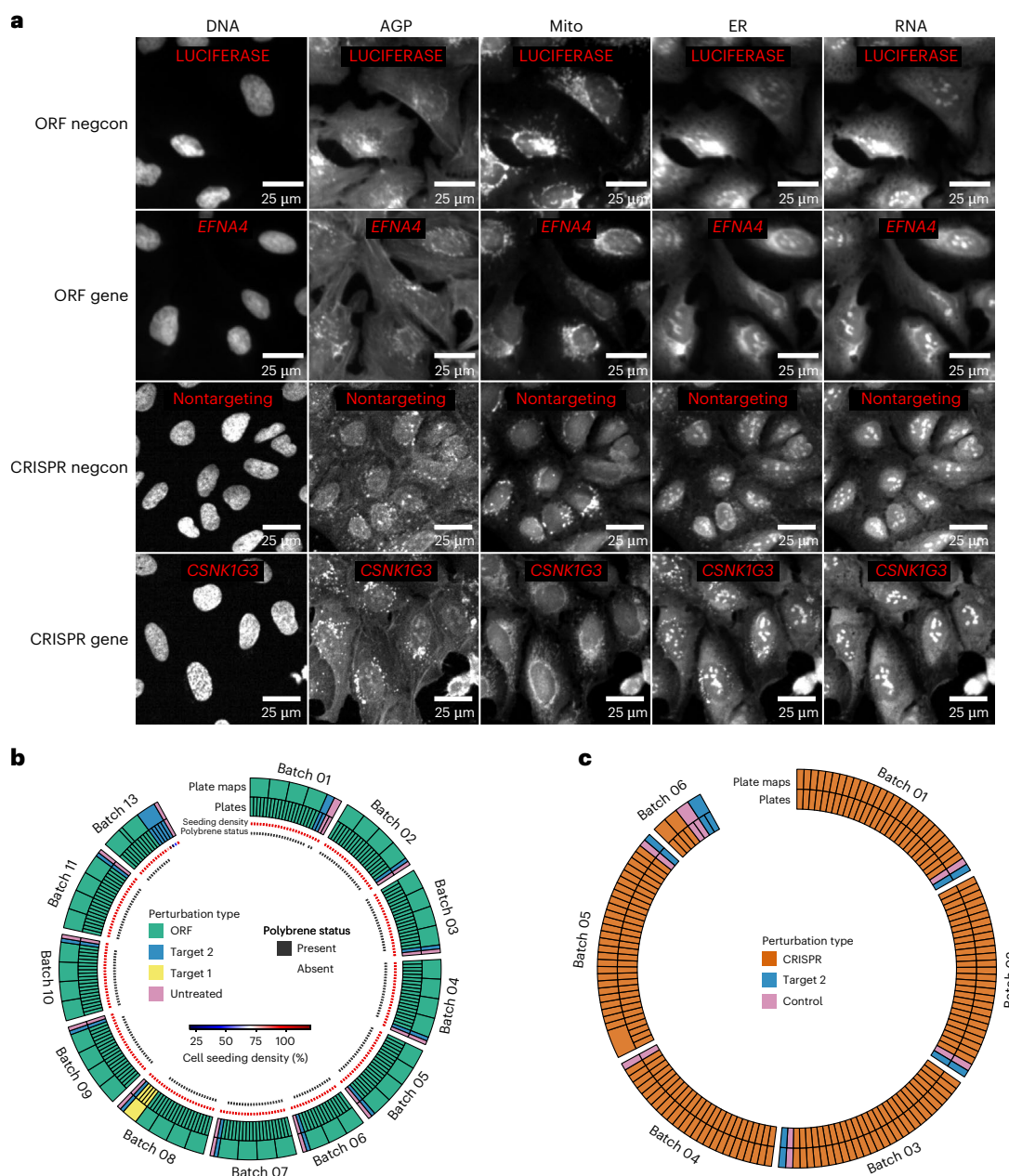


Fig. 1 | JUMP Cell Painting dataset overview. a, Example images of overexpressed genes (ORFs) and knocked-out genes (CRISPR) that are most dissimilar from negative controls (negcon) are shown. Negative control images are selected randomly from the ORF and CRISPR dataset. Channels are DNA, nucleus (Hoechst); AGP, actin cytoskeleton (phalloidin) plus the Golgi and plasma membrane (wheat germ agglutinin); Mito, mitochondria (MitoTracker); ER,

endoplasmic reticulum (concanavalin A) and RNA, nucleoli and cytoplasmic RNA (SYTO 14). **b, c**, Experimental design schematics for dataset generation. The number of plates, plate maps and batch in the ORF (**b**) and CRISPR (**c**) datasets are shown. Polybrene was added to most ORF plates, which are also shown. Detailed information regarding perturbation types per plate and other experimental conditions (for example, cell seeding density) can be found in Methods.

coexpression and other gene–gene links derived from text mining, gene ontology (GO), gene pathway, gene–disease and other annotations, but it was not built using the JUMP Cell Painting (nor indeed any image-based information). We analyzed the closeness of genes in the knowledge graph (KG) with a graph neural network (GNN) approach, where links between genes and functions in DRKG are predicted by the GNN and summarized in a GNN-based KG score (Methods). Gene–gene pairs with strong morphological similarity (or strong dissimilarity) were far more likely to have strong existing evidence of a relationship in the KG (Fig. 3e–j). Roughly half of such gene–gene pairs were supported by existing scientific evidence, as captured by the KG, and roughly half were previously unknown connections (whereas only 10% of randomly

shuffled gene pairs had supporting evidence). The JUMP dataset therefore contains reliable data and also the opportunity for discovery.

Screening for a single-cell phenotype using machine learning

Cell Painting images contain multiple channels and can reveal hundreds of morphological phenotypes of interest to biologists. To demonstrate the potential to use this data to carry out specific screens, we trained a single-cell phenotypic machine learning model to recognize rounded cells using the PhenoSorter feature in Spring Engine. Using a point-and-click interface, we provided the supervised learning tool with examples of 53 rounded cells and another 179 flat cells. The resulting classification model trained on an 80/20 training/test split dataset

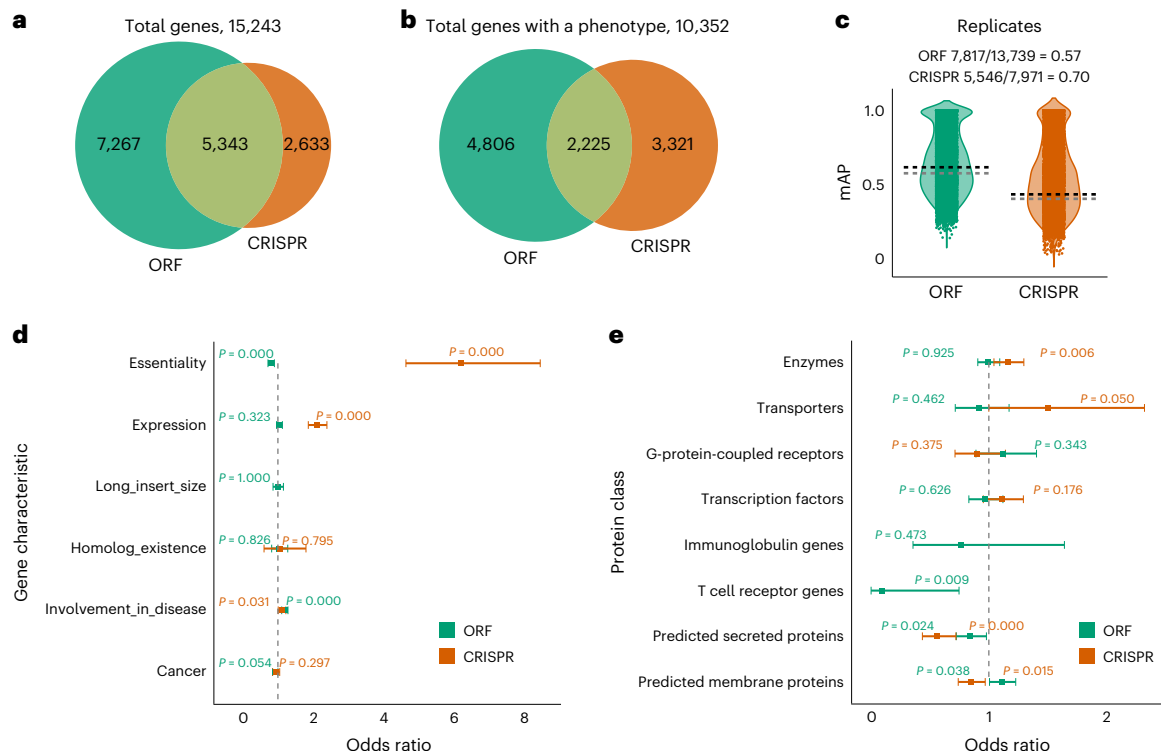


Fig. 2 | Phenotypic activity for genes and phenotypic consistency for gene groups. **a**, Number of genes overexpressed, knocked out or both. The Broad Institute lentiviral ORF library is not genome-wide because genes above 3,500 nucleotides do not package well into lentivirus, and some genes were excluded due to quality control reasons. **b**, Of the tested genes, 68% yielded a phenotype (replicates are significantly different from the negative control); 57% of tested reagents had a phenotype by ORF and 70% by CRISPR. **c**, To determine whether a reagent had a detectable phenotype ('phenotypic activity'), we computed mean average precision (mAP)²⁴ to assess how well a given replicate of a reagent retrieves other replicates of that reagent rather than a negative control. Green, ORF dataset and orange, CRISPR dataset. The fraction of reagents with

significant mAP values is shown above each panel. The gray and black dashed lines, respectively, are the 90th percentile of the reagents with a nonsignificant mAP value and the 10th percentile of reagents with a significant mAP value (the thresholds vary because there are varying numbers of replicates per gene). **d,e**, Forest plots illustrating associations between gene characteristics (**d**) and observed phenotypes (**e**) and observed phenotypes. Green represents ORF profiles and orange represents CRISPR profiles. Odds ratios (squares), the 95% confidence intervals of odds ratios (whiskers) and *P* values (calculated from a two-sided Fisher's exact test) are displayed. Multi-hypothesis correction resulted in adjusted significance thresholds ($\alpha = 0.0083$ (**a**) and $\alpha = 0.0062$ (**b**)). The number of data points in **d** and **e** are listed in Supplementary Tables 2 and 3, respectively.

yielded 100% recall and 100% precision in correctly calling rounded cells (Fig. 4a–c). Applying the model to the entire dataset as a virtual screen, we identified 187 genes in the ORF dataset with excess rounded cells, 63 of which had WikiPathways annotations. All the WikiPathways annotation groups significantly enriched in the rounded-cell gene set were related to cell death or stress (Fig. 4d).

Finding gene–gene relationships using image-based profiles

Dynein family, *FOXO3–TGF β* and *SLC39A1–ZBTB16*. Given the promising quantitative result that most relationships in the JUMP genetic dataset are supported by literature, we examined some well-known proteins, starting with some of the strongest gene clusters. In the CRISPR dataset, *FOXO3* and *TGF β 1* are negatively correlated (Fig. 5b). This interaction is well known, although the directionality of interaction is tissue dependent^{27,28}. *SLC39A1* and *ZBTB16* also strongly negatively correlate in both their ORF and CRISPR profiles (Fig. 5c,d); although lacking direct connection in the literature, the pair shares GO annotations, including appendage development, embryonic morphogenesis and skeletal system development.

In the overexpression (ORF) data, we noticed a strong cluster of *PAFAH1B1* (also known as *Lis1*) with *NDE1* (NudE) and *NDEL1* (Nudel). *HOOK1*, *HOOK2* and *SPDL1* had a strong anti-correlation with those three genes (Fig. 5e). Existing evidence links all six proteins to intracellular dynein transport²⁹; for example, the Nudel complex links *PAFAH1B1* to dynein³⁰ and *HOOK2* supports the recruitment of dynein by the Nudel complex at the nuclear envelope^{31,32}. As seen in our past

work³³, both similar and dissimilar morphological profiles can usefully point to functional relationships. 'Unexpected' directionality can be due to, for example, biological mechanisms being nonlinear, overexpression causing a dominant-negative effect (for example, by disrupting a functioning complex) or feedback loop and/or compensatory impact on the gene's function.

We then sought to identify previously unknown discoveries. Again, using the DRKG, we prioritized gene–gene connections with strong morphological similarity and/or dissimilarity but with low existing knowledge per the KG (using a cutoff of 0.5 for the KG score, Fig. 3). To distinguish whether these are false positives, reflecting technical noise or true previously undiscovered relationships between genes, we designed follow-up experiments and were able to make discoveries for several of them.

***MYT1* transcriptionally represses *RNF41*.** One of the top-15 strongest anticorrelating pairs in the CRISPR dataset was *MYT1* and *RNF41* (Fig. 6a). *MYT1* encodes a known transcriptional repressor³⁴, so we hypothesized it might directly repress transcription of *RNF41*: an obvious mechanism that could yield opposite Cell Painting phenotypes. Consistent with this, we performed CUT&RUN analysis and found that the product of *MYT1* binds the promoter of *RNF41* from adult mouse retinas (Fig. 6f).

Potential previously unknown Hippo pathway members. We previously noted strong morphological similarities between overexpression

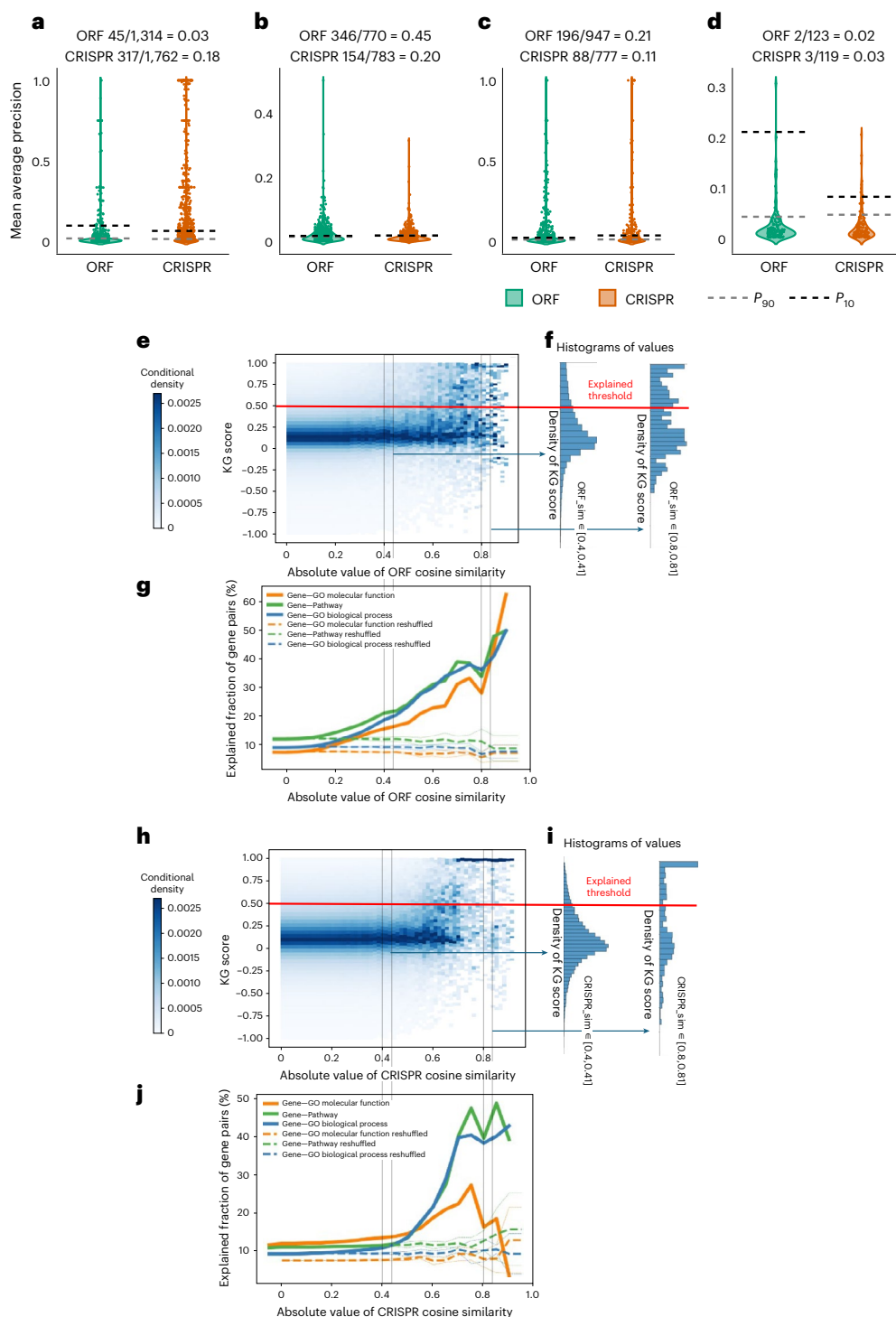


Fig. 3 | Characteristics of genes with phenotypes and KG validation.

a–d, Phenotypic consistency measured by mAP for retrieving genes sharing knowledge-based labels (CORUM Complex (**a**), WikiPathways (**b**), HGNC Gene Group (**c**) and Disease Association (**d**); Methods) versus genes with different labels. Green, ORF dataset and orange, CRISPR dataset. The fraction of gene groups with significant mAP values is shown above each panel. Gray and black dashed lines represent the 90th percentile of nonsignificant groups and the 10th percentile of significant groups, respectively (thresholds vary by group size). **e–j**, Comparison of morphological similarity in JUMP Cell Painting with evidence from the DRKG. **e,h**, Two-dimensional histograms for ORF (**e**) and CRISPR (**h**) compare morphological similarity (x axis, |cosine similarity|) versus KG score (y axis). A KG score > 0.5 indicates connection ‘explained’ via GNN. Data trends toward the upper right, showing pairs with strong morphological similarity often have higher KG scores. The large number of gene pairs is better depicted as a two-dimensional

histogram than a scatter plot. **f,i**, Example one-dimensional histograms (cuts through **e,h**) for ORF (**f**) and CRISPR (**i**) show KG score distributions for moderate (left, similarity 0.4–0.41) versus strong (right, similarity 0.8–0.81) morphological similarity. Pairs with stronger similarity (right) exhibit higher KG scores.

g,j, For ORF (**g**) and CRISPR (**j**), ~40–50% of gene pairs with strong morphological similarity (cosine similarity > 0.8) are explained by the KG (GNN prediction). This view of the data shows three variations of the KG, trained on molecular functions (gene_mf), pathways (gene_pathways) and biological processes (gene_bp). A random shuffling of gene labels shows that around 10% of gene pairs have connections if selected randomly (dashed lines show standard confidence intervals corresponding to one standard deviation for multiple shuffling of gene labels). Gene pairs with morphological similarities above ~0.4 (absolute value) show more DRKG evidence for connection than random pairs (except for the GNN trained with molecular function GO terms for CRISPR data).

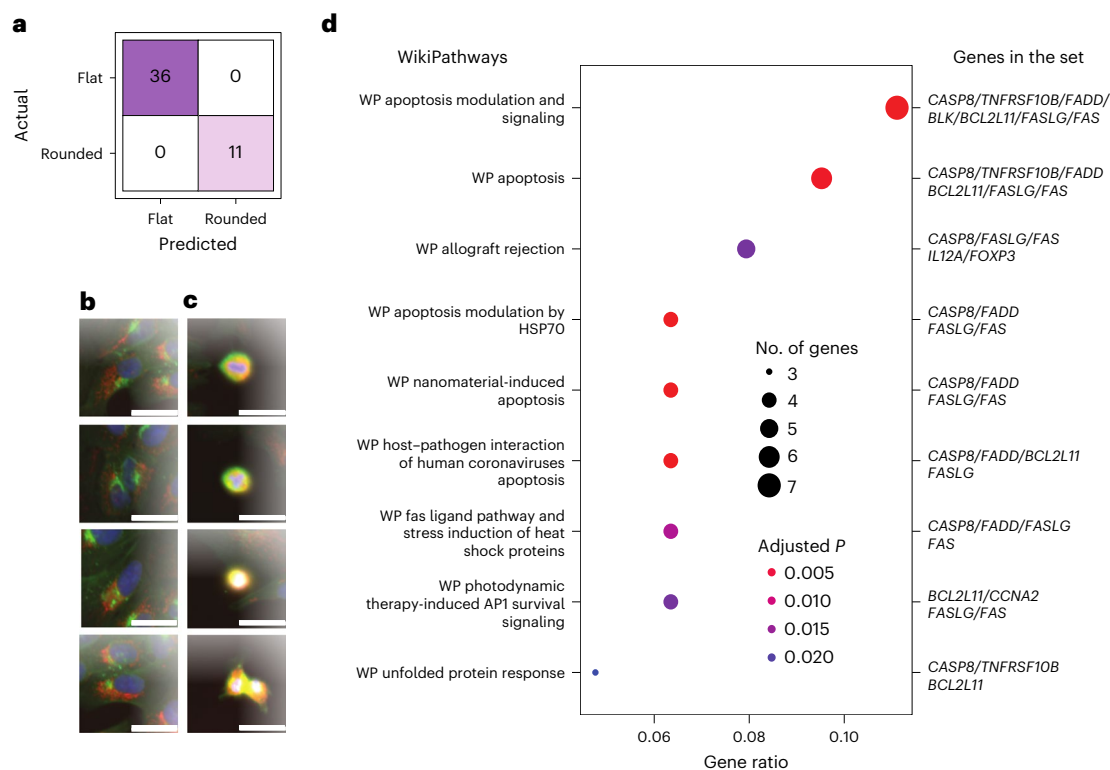


Fig. 4 | Virtual screen for genes yielding rounded cells. **a**, Confusion matrix for the single-cell classification model, showing the number of cells in the held-out test set for each combination of actual versus predicted classes. **b**, Example ‘flat’ cell image crops from the training set. **c**, Example ‘rounded’ cells. **d**, Dot plot showing significantly enriched gene annotations among genes with a high

proportion of ‘rounded’ cells. Five gene sets were related to apoptosis, and the remaining four pertain to cell stress (individual genes in each set are listed on the right). The dot size reflects the number of hit genes in the indicated gene set. The color indicates the adjusted *P* value. The reported *P* values are from a one-sided statistical test. Scale bars, 10 μ m. WP, WikiPathways.

of *YAPI* and its paralog *WWTR1* (TAZ) in an experiment with 220 genes³³; we saw the same connection here in the much larger gene set, now with other known Hippo pathway members, *STK3* (*MST2*), *VGLL4* and *PRKCE*^{35,36} (Fig. 6b). Using the Plex web application³⁷ to explore existing datasets we found that 10–12 of the 50 genes found to be most similar to *YAPI* in the ORF data are coregulated in various perturbations and/or conditions, and these genes mostly relate to actin binding and/or regulation and neuronal development^{38–43}. Furthermore, 18 of the *YAPI*-similar genes exhibit significant expression changes on *YAPI* knockout, further strengthening evidence for a valid biological relationship⁴⁴. Notably, *CEP72*, *IL2ORB* and *MTMR9* showed strong phenotypic correlation with *YAPI* despite having little KG evidence for engagement with the Hippo pathway. These genes merit investigation for involvement in the Hippo pathway and related disorders.

***INSYNI*'s potential role in cancer migration and proliferation.** In both the ORF and CRISPR data, many genes whose products are linked to the regulation of cell migration and proliferation in cancer cluster together, including the transcription factors *HOXC8* and *ZFP36L1*, the kinase-related genes *PIK3R3* and *NRBPI*, and the GTP-related genes *RAB40B* and *RAB40C* (Fig. 6c). *INSYNI*'s only known function is coordinating the postsynaptic inhibition complex⁴⁵, but it is also present in this ORF cluster (it was not tested by CRISPR). We found multiple sources of weak evidence for a link between *INSYNI* and cancer. For example, the Human Protein Atlas re-analyzed the TCGA dataset and classified *INSYNI* as a prognostic marker gene for renal cancer and as having enhanced expression in glioma cancer tissues (<https://www.proteinatlas.org/ENSG00000205363-INSYNI/pathology>). The Genetic Associations Database contains associations between uncharacterized proteins and cancer outcomes and flagged *INSYNI* (under its synonym *C15orf59*) as one of 62 new OncoORFs⁴⁶. Expression of the antisense

long noncoding RNA of *INSYNI* is associated with low-grade glioma prognosis⁴⁷ and with the overexpression of vimentin, a marker of advanced metastatic cancer⁴⁸. Collectively, this suggests *INSYNI* should be more closely investigated for its oncogenic potential.

Link between SLC and olfactory receptor superfamilies. One of the largest morphological clusters with previously unknown connections (low KG scores) contains members of two superfamilies: solute carriers (SLC) and olfactory receptors. The cluster exists in the ORF data, not the CRISPR data (mainly because none of the olfactory receptor genes showed a phenotype after CRISPR knockout). Among a tightly clustered subset of this large SLC-olfactory receptor cluster, some gene pairs are known to be connected, but many were not (Supplementary Figs. 3–5).

Cluster linked to mitochondrial function and cancer. We found that three enzymes strongly correlated in both the ORF and CRISPR profiles: *ECH1* (enoyl-CoA hydratase 1), *UQCRCF1* (ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1) and *SARS2* (seryl-tRNA synthetase 2, mitochondrial) (Fig. 6d). The connections among them are mostly unknown; *UQCRCF1* and *SARS2* are connected in the KG by inference through the GNN only, not by literature reports, and the remaining connections show weak or low KG scores (Fig. 6d). Some existing data support these new connections: *SARS2* is the most highly correlated gene with *ECH1* in terms of cell-line RNA expression; *SARS2* and *ECH1* are the fifth and tenth top matches, respectively, for *UQCRCF1* (<https://www.proteinatlas.org/ENSG00000104823-ECH1/cell+line>). Analyzing the top 52 genes most similar to this cluster in the ORF dataset using Plex, most are mitochondria-associated (GOCC_Mitochondrion), including 16 mitochondrial disease-associated genes⁴⁹. Knockdown or overexpression of *LINCO0473*, a regulator of

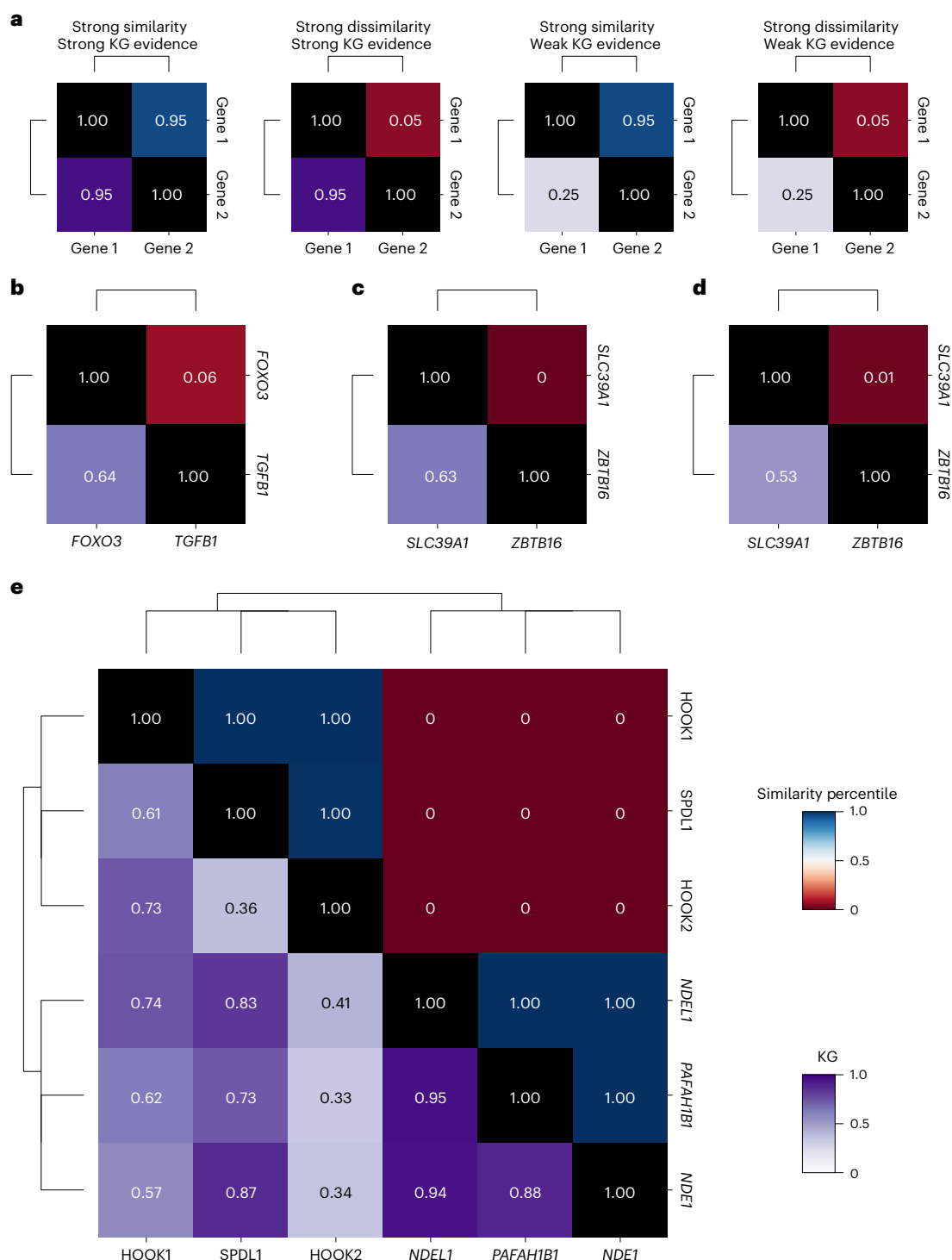


Fig. 5 | Validation of known gene-gene relationships. **a**, This schematic depicts a clustergram that integrates gene similarity with KG evidence, providing a framework for interpreting other clustergrams. The upper triangular matrix displays the percentile of cosine similarity between gene profiles (0, negative correlation and 1, positive correlation). The lower triangular matrix shows the KG

score (Methods), with scores <0.5 indicating new associations not predicted by the GNN on DRKG. **b–e**, Known gene relationships across different perturbation modalities (ORF and CRISPR). **b**, *FOXO3–TGFB1* (CRISPR). **c**, *SLC39A1–ZBTB16* (ORF). **d**, *SLC39A1–ZBTB16* (CRISPR). **e**, Dynein family cluster (ORF).

lipolysis and mitochondrial respiration, downregulates 15 of these genes⁵⁰, and knockdown of the mitochondrial chaperone encoded by *PHB2* downregulates 14 genes in the cluster⁵¹. In addition, in proteomic profiling data from cells treated with a library of 875 compounds, five of the ten profiles with the greatest overlap with this 52 gene cluster were from inhibitors of the PI3K/MTOR pathway⁵², which regulates both mitochondrial function and biogenesis⁵³. Given recent interest

in *UQCRCF1* as a mitochondrial-related oncology biomarker and drug target⁵⁴, *SARS2* and *ECHI* also merit attention.

Connection of *TSC22D1* to genes with neural function. Across both ORF and CRISPR data, we identified three strongly correlated genes: *GPR176* (*G protein-coupled receptor 176*), *CHRM4* (*cholinergic receptor muscarinic 4*), *TSC22D1* (*TSC22 domain family member 1*) and a fourth

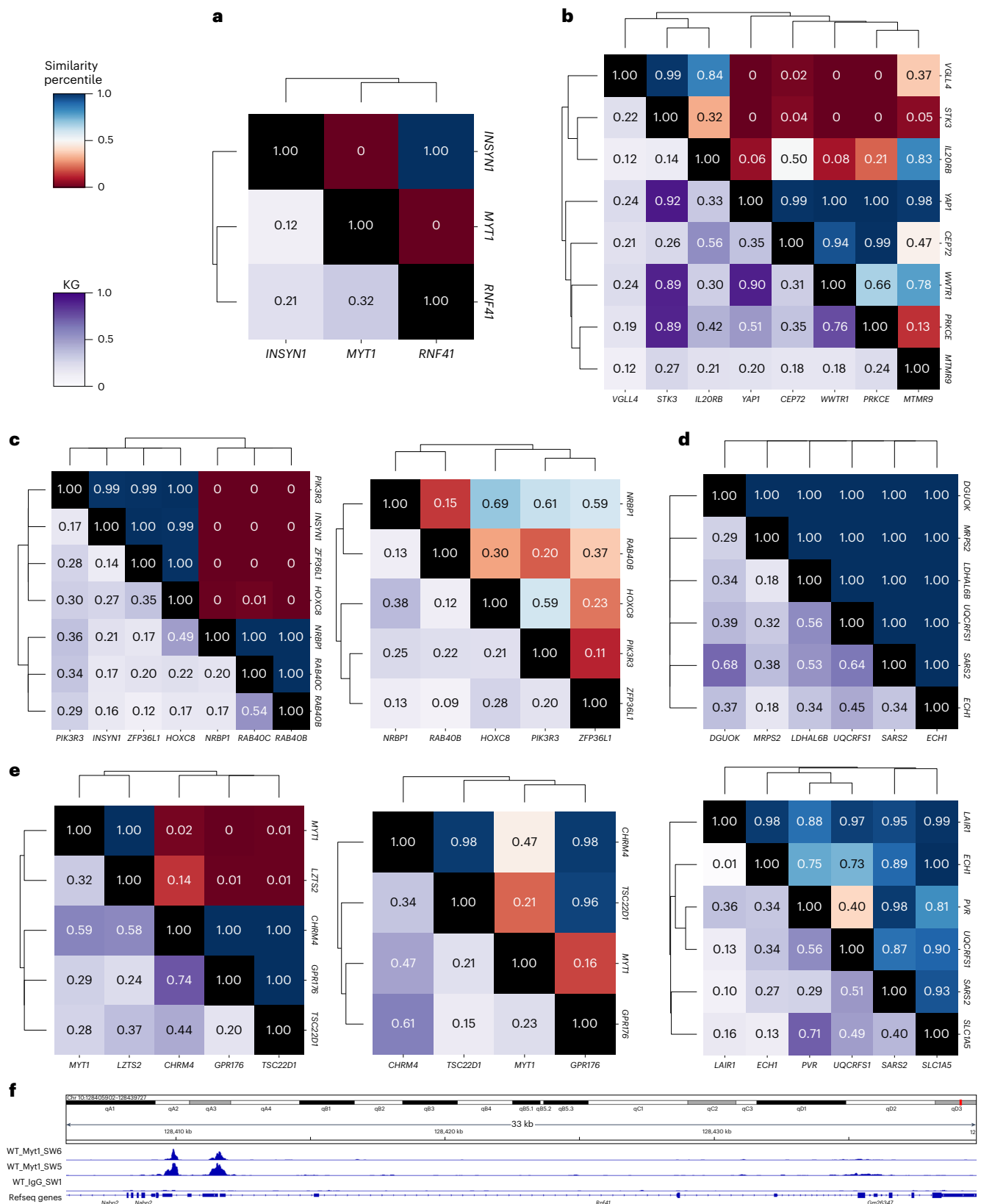


Fig. 6 | Previously unknown gene clusters with strong morphological similarity in JUMP Cell Painting. a–e, Clustergrams highlight previously unknown gene clusters with strong morphological relationships, but with little existing evidence linking the genes. Clustergrams can be interpreted using the schematic in Fig. 5a. **a,** *MYT1–RNF41* (ORF). **b,** A *YAP1*-associated cluster (ORF). **c,** Clusters of genes involved in cancer proliferation plus *INSYN1* (ORF and

CRISPR). **d,** Clusters of genes implicated in mitochondrial function and cancer (ORF and CRISPR, respectively). **e,** Clusters of genes involved in neural function (ORF and CRISPR, respectively). Note that *RNF41* (from **a**) is a near neighbor to *MYT1*, but only appears in a slightly larger version of this cluster. **f,** CUT&RUN demonstrating Myt1 binding at the promoter region of Rnf41 in the mouse retina.

gene, *MYT1* (*myelin transcription factor 1*), consistently negatively correlated to that group (Fig. 6e). The products of *GPRI76*, *CHRM4* and *MYT1* are known to be involved in neural development or upregulated in neurons^{55,56}, and their relationship is already known, according to moderate to high scores in the KG. By contrast, *TSC22D1* has little known connection to any of these and is instead annotated as involved in apoptosis, tumor suppression and cellular stress response. The morphological similarity we observed indicates that *TSC22D1* warrants further investigation for neural functions: indeed, it is most highly expressed in brain tissue according to the Human Protein Atlas (<https://www.proteinatlas.org/ENSG00000102804-TSC22D1/tissue>). Furthermore, *TSC22D1* is strongly anti-correlated to *LZTS2* (*leucine zipper tumor suppressor 2*) in our ORF dataset (*LZTS2* was not present in our CRISPR dataset), and the two genes show an inverse messenger RNA expression relationship^{55,57}. *LZTS2*'s role in the Wnt pathway⁵⁸ provides another tie to neuronal function⁵⁹. Searching public datasets in Plex, we found that knockdown of *Musashi-1* (*MSI1*), a gene that may play an essential role in nervous system development, in SU_MB002 medulloblastoma cells leads to downregulation of *GPRI76*, *CHRM4* and *TSC22D1* and upregulation of *MYT1* by RNA sequencing analysis, mirroring our observed correlations⁶⁰. In addition, *GPRI76*, *CHRM4* and *TSC22D1* are reported to be downregulated in *TREM2* (*Triggering Receptor Expressed on Myeloid Cells 2*) variants associated with Alzheimer's disease⁶¹. Together, these data indicate investigating *TSC22D1* in brain function would be worthwhile.

Public portals for exploring JUMP data

Several portals have been built to explore and analyze the JUMP genetic perturbation data. The Broad Institute's web-based portal JUMP Cell Painting Hub (<http://broad.io/jump>; unpublished communication, A. F. Muñoz) allows searching for genes with the most similar profiles to a query gene, and displays images and distinctive features (relative to negative controls), to aid in interpreting the image-based phenotypes and identifying technical artifacts. In addition, the page includes instructions for interacting with the profiles of phenotypically active ORF and CRISPR genes in the web-based data visualization and analysis software Morpheus (<https://software.broadinstitute.org/morpheus/>) using the compatible files stored at <https://doi.org/10.5281/zenodo.14025601> (ref. 62) and <https://zenodo.org/records/14165010> (ref. 63) and instructions at the JUMP Cell Painting Hub.

JUMP Consortium partner Ardigen developed a free web-based application, phenAID JUMP-CP Data Explorer (<https://phenaid.ardigen.com/jumpcpexplorer/>), that enables viewing images and metadata corresponding to each genetic perturbation and searching for and downloading up to 100 nearest neighbors. By contrast to the Broad portal, whose gene–gene similarities are based on simple cosine similarities between image profiles, phenAID tool is based on projecting the ORF and CRISPR profiles to the same two-dimensional space using the UMAP (uniform manifold approximation and projection) algorithm with Euclidean distance and color-coded by clusters as determined using the BIRCH clustering algorithm.

Discussion

This study presents a valuable resource for the research community: a large-scale dataset linking both over- and underexpression genetic perturbations to cell image phenotypes. We have demonstrated its utility in screening for particular phenotypes of interest and uncovering previously unknown connections between genes. The two modalities—overexpression by ORFs and underexpression by CRISPR–Cas9 knockout—provide complementary information rather than yielding identical gene clustering or simply producing opposing phenotypic profiles, consistent with our previous 160-gene study²⁰. Further investigation using this expanded dataset could provide more definitive answers, potentially through reprocessing ORF and CRISPR profiles

with a common pipeline to enable direct comparison rather than just at the clustering level.

This dataset has some limitations. Most notably, only 15,243 genes were tested by a reagent in the study, and only 10,352 genes had a detectable phenotype by overexpression, CRISPR or both. The Cell Painting assay captures a broad set of phenotypes, but they are only a subset of all cell morphological changes that might occur with a broader range of cellular markers. The experiment was done on a single cell type (U-2 OS) from a single demographic (white women) at a single time point after genetic perturbation, and results may not extend to other genetic backgrounds or experimental conditions. Some individual reagents may be faulty (for example, CRISPR guides' off-target effects or lack of efficacy, or unexpected mutations in overexpression reagents) and, although our processing steps attempted to mitigate them, technical artifacts affect the comparison of data collected in batches and on plate layouts that induce batch and position effects⁶⁴. For this reason, it is crucial to examine plate layouts to ensure observed connections are not artifacts (Supplementary Figs. 6–19): reagents within the same plate, row or column can exhibit spurious correlations. In addition, visual inspection is useful to identify image-based anomalies (Supplementary Figs. 20–32).

The JUMP Cell Painting Consortium is pleased to present this genetic perturbation dataset for the scientific community's exploration. All images and extracted profiles are freely available via AWS at <https://registry.opendata.aws/cellpainting-gallery/>.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-025-02753-9>.

References

- Mattiuzzi Usaj, M. et al. High-content screening for quantitative cell biology. *Trends Cell Biol.* **26**, 598–611 (2016).
- Bougen-Zhukov, N., Loh, S. Y., Lee, H. K. & Loo, L.-H. Large-scale image-based screening and profiling of cellular phenotypes. *Cytom. A* **91**, 115–125 (2017).
- Boutros, M., Heigwer, F. & Laufer, C. Microscopy-based high-content screening. *Cell* **163**, 1314–1325 (2015).
- Cheng, J. et al. Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. *Adv. Sci.* **10**, e2204484 (2023).
- Perlman, Z. E. et al. Multidimensional drug profiling by automated microscopy. *Science* **306**, 1194–1198 (2004).
- Dagher, M. et al. nELISA: a high-throughput, high-plex platform enables quantitative profiling of the secretome. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.04.17.535914> (2023).
- Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
- Moshkov, N. et al. Predicting compound activity from phenotypic profiles and chemical structures. *Nat. Commun.* **14**, 1967 (2023).
- Way, G. P. et al. Morphology and gene expression profiling provide complementary information for mapping cell state. *Cell Syst.* **13**, 911–923 (2022).
- Haghighi, M., Caicedo, J. C., Cimini, B. A., Carpenter, A. E. & Singh, S. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. *Nat. Methods* **19**, 1550–1557 (2022).
- Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
- Williams, E. et al. Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).

13. Neumann, B. et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
14. Ohya, Y. et al. High-dimensional and large-scale phenotyping of yeast mutants. *Proc. Natl Acad. Sci. USA* **102**, 19015–19020 (2005).
15. Mattiazzi Usaj, M. et al. Systematic genetics and single-cell imaging reveal widespread morphological pleiotropy and cell-to-cell variability. *Mol. Syst. Biol.* **16**, e9243 (2020).
16. Heigwer, F. et al. A global genetic interaction network by single-cell imaging and machine learning. *Cell Syst.* <https://doi.org/10.1016/j.cels.2023.03.003> (2023).
17. Fay, M. M. et al. RxRx3: phenomics map of biology. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.02.07.527350> (2023).
18. Ramezani, M. et al. A genome-wide atlas of human cell morphology. *Nat. Methods* **22**, 621–633 (2025).
19. Lazar, N. H. et al. High-resolution genome-wide mapping of chromosome-arm-scale truncations induced by CRISPR–Cas9 editing. *Nat. Genet.* **56**, 1482–1493 (2024).
20. Chandrasekaran, S. N. et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nat. Methods* **21**, 1114–1121 (2024).
21. Chandrasekaran, S. N. et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. Preprint at *bioRxiv* <https://doi.org/10.1101/2023.03.23.534023> (2023).
22. Cimini, B. A. et al. Optimizing the Cell Painting assay for image-based profiling. *Nat. Protoc.* **18**, 1981–2013 (2023).
23. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
24. Kalinin, A. A. et al. A versatile information retrieval framework for evaluating profile strength and similarity. *Nat. Commun.* **16**, 5181 (2025).
25. Sjöstedt, E. et al. An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).
26. Ioannidis, V. N. et al. Drkg - drug repurposing knowledge graph for covid-19. *GitHub* <https://github.com/gnn4dr/DRKG/> (2020).
27. Kuo, S.-J. et al. TGF- β 1 enhances FOXO3 expression in human synovial fibroblasts by inhibiting miR-92a through AMPK and p38 pathways. *Aging* **11**, 4075–4089 (2019).
28. Vivar, R. et al. Role of FoxO3a as a negative regulator of the cardiac myofibroblast conversion induced by TGF- β 1. *Biochim. Biophys. Acta, Mol. Cell.* **1867**, 118695 (2020).
29. Reck-Peterson, S. L., Redwine, W. B., Vale, R. D. & Carter, A. P. The cytoplasmic dynein transport machinery and its many cargoes. *Nat. Rev. Mol. Cell Biol.* **19**, 382–398 (2018).
30. Huang, J., Roberts, A. J., Leschziner, A. E. & Reck-Peterson, S. L. Lis1 acts as a ‘clutch’ between the ATPase and microtubule-binding domains of the dynein motor. *Cell* **150**, 975–986 (2012).
31. Kumari, A. et al. Phosphorylation and Pin1 binding to the LIC1 subunit selectively regulate mitotic dynein functions. *J. Cell Biol.* **220**, e202005184 (2021).
32. Dwivedi, D., Kumari, A., Rathi, S., Mylavarapu, S. V. S. & Sharma, M. The dynein adaptor Hook2 plays essential roles in mitotic progression and cytokinesis. *J. Cell Biol.* **218**, 871–894 (2019).
33. Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via Cell Painting. *eLife* **6**, e24060 (2017).
34. Lee, J. et al. A Myt1 family transcription factor defines neuronal fate by repressing non-neuronal genes. *eLife* **8**, e46703 (2019).
35. Fu, M. et al. The Hippo signalling pathway and its implications in human health and diseases. *Signal Transduct. Target. Ther.* **7**, 376 (2022).
36. Gong, R. et al. Opposing roles of conventional and novel PKC isoforms in Hippo-YAP pathway regulation. *Cell Res.* **25**, 985–988 (2015).
37. Selinger, D. W. et al. A framework for autonomous AI-driven drug discovery. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.12.17.629024> (2024).
38. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
39. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
40. McClintick, J. N. et al. Stress-response pathways are altered in the hippocampus of chronic alcoholics. *Alcohol* **47**, 505–515 (2013).
41. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
42. Kapeli, K. et al. Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nat. Commun.* **7**, 12143 (2016).
43. Vrenken, K. S. et al. The transcriptional repressor *SNAI2* impairs neuroblastoma differentiation and inhibits response to retinoic acid therapy. *Biochim. Biophys. Acta, Mol. Basis Dis.* **1866**, 165644 (2020).
44. Rivera-Reyes, A. et al. *YAP1* enhances NF- κ B-dependent and independent effects on clock-mediated unfolded protein responses and autophagy in sarcoma. *Cell Death Dis.* **9**, 1108 (2018).
45. Uezu, A. et al. Identification of an elaborate complex mediating postsynaptic inhibition. *Science* **353**, 1123–1129 (2016).
46. Delgado, A. P., Brandao, P., Chapado, M. J., Hamid, S. & Narayanan, R. Open reading frames associated with cancer in the dark matter of the human genome. *Cancer Genomics Proteom.* **11**, 201–213 (2014).
47. Lu, Z. & Feng, Y. Foreboding lncRNA markers of low-grade gliomas dependent on metabolism. *Medicine* **101**, e31302 (2022).
48. Usman, S. et al. Transcriptome analysis reveals vimentin-induced disruption of cell-cell associations augments breast cancer cell migration. *Cells* **11**, 4035 (2022).
49. Ochoa, D. et al. The next-generation Open Targets Platform: reimaged, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).
50. Tran, K.-V. et al. Human thermogenic adipocyte regulation by the long noncoding RNA LINC00473. *Nat. Metab.* **2**, 397–412 (2020).
51. Liu, X. et al. Regulation of mitochondrial biogenesis in erythropoiesis by mTORC1-mediated protein translation. *Nat. Cell Biol.* **19**, 626–638 (2017).
52. Mitchell, D. C. et al. A proteome-wide atlas of drug mechanism of action. *Nat. Biotechnol.* **41**, 845–857 (2023).
53. Morita, M. et al. MTOR controls mitochondrial dynamics and cell survival via MTFP1. *Mol. Cell* **67**, 922–935 (2017).
54. Sun, Q. et al. *UQCRRFS1* serves as a prognostic biomarker and promotes the progression of ovarian cancer. *Sci. Rep.* **13**, 8335 (2023).
55. Doi, M. et al. Gpr176 is a Gz-linked orphan G-protein-coupled receptor that sets the pace of circadian behaviour. *Nat. Commun.* **7**, 10583 (2016).
56. Chen, W.-Y. et al. Nerve growth factor interacts with CHRM4 and promotes neuroendocrine differentiation of prostate cancer and castration resistance. *Commun. Biol.* **4**, 22 (2021).
57. Iida, M., Anna, C. H., Gaskin, N. D., Walker, N. J. & Devereux, T. R. The putative tumor suppressor Tsc-22 is downregulated early in chemically induced hepatocarcinogenesis and may be a suppressor of Gadd45b. *Toxicol. Sci.* **99**, 43–50 (2007).

58. Liu, R., Zhou, D., Yu, B. & Zhou, Z. Phosphorylation of LZTS2 by PLK1 activates the Wnt pathway. *Cell. Signal.* **120**, 111226 (2024).
59. Tang, S.-J. Synaptic activity-regulated Wnt signaling in synaptic plasticity, glial function and chronic pain. *CNS Neurol. Disord. Drug Targets* **13**, 737–744 (2014).
60. Kameda-Smith, M. M. et al. Characterization of an RNA binding protein interactome reveals a context-specific post-transcriptional landscape of MYC-amplified medulloblastoma. *Nat. Commun.* **13**, 7506 (2022).
61. Liu, T. et al. Multi-omic comparison of Alzheimer's variants in human ESC-derived microglia reveals convergence at APOE. *J. Exp. Med.* **217**, e20200474 (2020).
62. Chandrasekaran, S. N. Phenotypically active ORF and CRISPR consensus profiles. *Zenodo* <https://doi.org/10.5281/zenodo.14025601> (2024).
63. Munoz, A. Phenotypically active ORF and CRISPR consensus profiles. *Zenodo* <https://doi.org/10.5281/zenodo.14164990> (2024).
64. Arevalo, J. et al. Evaluating batch correction methods for image-based cell profiling. *Nat. Commun.* **15**, 6516 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2025

¹Broad Institute, Cambridge, MA, USA. ²Revvity Discovery Limited (formerly known as Horizon Discovery), Cambridge, UK. ³Ardigen S.A, Kraków, Poland. ⁴Evotec SE, Hamburg, Germany. ⁵Stanford University, Stanford, CA, USA. ⁶Spring Science, San Carlos, CA, USA. ⁷Plex Research Inc., Cambridge, MA, USA. ⁸Ksilink, Strasbourg, France. ✉ e-mail: shantanu@broadinstitute.org; anne@broadinstitute.org

Methods

Genetic perturbation selection

ORF expression library. For cpg0016, we used a preexisting lentiviral ORF expression library created at the Broad Institute for gene overexpression experiments²³. We tested 15,141 overexpression reagents encompassing 12,609 unique genes, including controls, which are described in a later section. Lentiviral packaging becomes less efficient for larger genes; this library does not cover the entire genome because it excludes larger genes.

CPJUMP1 data (cpg0000) also contains ORF perturbation data, where we overexpressed 176 genes. Further details about these perturbations have been published in ref. 20.

CRISPR knockout library. We used the Revvity Human Edit-R synthetic crRNA-Druggable Genome in our CRISPR knockout experiments. This library targets 7,975 unique genes; each targeted by a pool of four pre-designed synthetic crRNAs with high specificity and functional knockout.

In the CRISPR knockout experiments in CPJUMP1, we knocked down 160 genes. Further details about these perturbations have been published in ref. 20.

Controls. We included several controls to identify and correct for different experimental artifacts. The controls can be broadly categorized as within-plate controls (those on the same plate as the treatments) and control plates (entire plates run alongside the treatment plates).

Within-plate controls. Within-plate controls include negative controls, positive controls and untreated wells.

Negative control wells are used to detect and correct plate-to-plate variations. They can also be used as a baseline for identifying perturbations with a detectable morphological signal. The following are the within-plate negative controls.

- ORF plates: ORFs of genes for blue fluorescent protein, HcRed, lacZ and Luciferase, although it should be noted that no sensible negative control exists for protein overexpression, as each of these proteins is known to induce some changes to cell state.
- CRISPR plates: nontargeting guides (CRISPR guides that do not target any gene) and no guides (Cas9 cells are transfected but did not receive any CRISPR guides). There are also wells where cells are treated with dimethylsulfoxide (DMSO) but are not used as negative control treatments.

Positive control wells are included in the plates to ensure the experiment worked as expected.

- ORF plates: we ran four positive controls on the ORF plates (Supplementary Table 5), which are a subset of the eight positive control compounds run on the CRISPR plates. Furthermore, we included eGFP (enhanced green fluorescent protein) as the ORF positive control, which showed a consistent phenotype distinct from the negative controls.
- CRISPR plates: *PLK1*, knocking it down kills cells. We also ran eight compound positive controls on each CRISPR plate (Supplementary Table 5). These compounds exhibited phenotypes that are distinct from each other and from the negative control in previous experiments²⁰.

Untreated wells received neither treatments nor controls; they contained only cells. They might also be used for normalizing samples, but they are nonideal negative controls, given that they are not mock-treated with any reagents.

Control plates. Negative control plates were run periodically (for example, one or more per batch) (Fig. 1b,c). These plates can help

disentangle staining and imaging artifacts, including batch effects and well-position effects. Examples include an entire plate of untreated cells and an entire plate of cells treated with nontargeting or no guides (for CRISPR).

JUMP-Target-2-Compound plates: one or more plates of these compounds were run in each batch to correct batch effects (unrelated plates from the same batch correlate strongly with each other compared to related plates in other batches) (Fig. 1c).

JUMP-Target-1-Compound plates: to align the production data (cpg0016) with the CPJUMP1 experiment in the pilot data (cpg0000)²⁰, we ran four JUMP-Target-1-Compound plates in one batch of cpg0016 from source_4 (Broad) (Fig. 1b,c).

JUMP-Target-2-Compound plates with polybrene: to quantify the cell morphology effects of adding polybrene, a viral transfection agent that is added to the ORF experiment, we ran two JUMP-Target-2-Compound plates with polybrene in one batch (Fig. 1b,c).

Plate layout design

ORF plates. The ORF plates were pre-designed due to their existence in a preplated library²³; each plate consists of negative control wells and untreated wells spread across the plate. Most plates contain 16 negative control wells, while some have as many as 28 wells. One replicate for each of the four compound positive controls (Supplementary Table 5) is added to wells O23, O24, P23 and P24. The remaining wells contain ORF treatments, with a single replicate of each per plate map and with five replicate plates produced per plate map.

CRISPR plates. Similar to the compound plates, the outer columns of the plate contain positive and negative controls. The outermost columns 1 and 24 contain four replicates of the eight compound positive controls. Columns 2 and 23 contain ten replicates each of wells with no guides and nontargeting guides, eight replicates of DMSO and four replicates of the CRISPR positive control, *PLK1*.

JUMP-Target-1-compound plates. The plates contain 306 compounds and DMSO; all but 14 compounds are in singlicate. There are 64 DMSO wells spread across the plate. The 14 compounds with two replicates are diverse positive controls that differ from the positive controls on the production plate. Further details about these positive controls and the criteria met by the plate's compounds have been described in ref. 20.

JUMP-Target-2-compound plates. These plates' contents are identical to those of JUMP-Target-1-Compound plates, but the layout differs. JUMP-Target-2-Compound plates meet all the criteria met by JUMP-Target-1-Compound plates²². Both layouts are provided in our GitHub repository (<https://github.com/jump-cellpainting/JUMP-Target>).

Perturbation identifier

All ORF reagents, CRISPR reagents and controls are assigned a unique JUMP identifier. These identifiers start with the code JCP2022_.

Experimental conditions

Cell Painting dyes. We optimized the concentrations of several Cell Painting dyes and chose the concentration as described in our previous publication in ref. 22.

Cell line. After comparing A549 and U-2OS cell lines in pilot experiments²⁰, we selected U-2 OS because of its performance in Cell Painting experiments and the existence of previous datasets in this cell line, allowing comparison across experiments. The U-2OS cell line was obtained directly from American Type Culture Collection (ATCC) (catalog no. HTB-96, lot no. 70016635). As the cells were sourced directly from ATCC and U2OS is not listed among commonly misidentified cell lines, further authentication was not performed. The cell cultures tested negative for mycoplasma contamination approximately 1 month after being initiated.

Time point. We compared two time points for each perturbation modality: 48 and 96 h of ORF and 96 and 144 h for CRISPR and settled on 48 h for ORF and 96 h for CRISPR, based on their performance²⁰.

Reagent vendor. Cell Painting dyes from Thermo Fisher and Revvity performed similarly in our pilot experiments²². For this dataset, we used the PhenoVue¹ Cell Painting Kit, v.2.0 (part number PING22, Revvity), which Revvity donated to the JUMP Cell Painting consortium.

Microtiter plates. After comparing different plates for their ability to minimize evaporation in the outer wells (cpg0000)²², we used Revvity Cell Carrier Ultra for data generation.

Compound concentration. The positive control compounds in the ORF and CRISPR plates, JUMP-Target-1-Compound and JUMP-Target-2-Compound plates, were assayed at 5 μ M.

ORF experiments. We followed previous protocols^{65,66}, with the following experimental parameters chosen: five replicates per virus plate, 1,525 cells per 384 wells, 30 μ l of media seeding volume per 384 wells, 1 μ l of virus per 384 wells and 4 μ g ml⁻¹ polybrene added 1 h postseeding, 30-min spin at -1,000g, media change after 24 h removing polybrene and virus adding back 40 μ l of media, no selection with blasticidin.

Creation of the Cas9 cell line. We made U2OS-311 at the Broad Institute by transducing U-2 OS cells at a multiplicity of infection <1 with lentivirus prepared from the vector pLX_311-Cas9 (Addgene plasmid 96924), which expresses blasticidin resistance from the SV40 promoter and Cas9 from the EF1 α promoter, and selected with 16 μ g ml⁻¹ blasticidin for 14 days. Briefly, a 12-well plate at 1.5×10^6 cells per well in 1.25 ml of media and 750 μ l of virus supplemented with 4 μ g ml⁻¹ polybrene were centrifuged for 2 h at 1,000g, then 2 ml of media was added per well. 24 h after infection, cells were split out of the 12-well plate and 48 h after infection, 16 μ g ml⁻¹ blasticidin was added and maintained. This is a slight modification of our published protocol⁶⁷. These lines are not available due to Broad Institute's licensing restrictions.

CRISPR experiments. Five replicates per target gene were assessed, with each replicate well containing four different sequences targeting the same gene. The cell line used for these experiments was U2OS-311 (Broad). The cells were kept for 1 week in blasticidin before plating. Blasticidin was removed one passage before plating. Before plating, 125 nl of each guide RNA pool at 10 μ M (Revvity) and 125 nl of the trans-activating crRNA (tracrRNA) at 10 μ M (Revvity) were dispensed in the 384-well plates using the ECHO (Beckman Coulter Life Science). Then 10 μ l of LipoRNAimax (Thermo Fisher) diluted in OptiMEM was added to each well using a MultiDrop Combi (Thermo Fisher) to achieve a final 0.05 μ l per well of LipoRNAimax. The plates were pulse centrifuged and then incubated for 30 min at room temperature. Then 700 cells per well were dispensed in 40 μ l using the MultiDrop Combi and incubated for 96 h in an incubator (37 C, 5% CO₂) before staining and fixation. Additional chemical treatments were included in some control wells (DMSO, C1, C2 ... C8), and some full plates of cells were treated with the Target-2 plate of compounds. These treatments were performed as described in the chemical screening and were incubated for 48 h before staining and fixing U2OS-311 cells with LipoRNAimax treatment only (no guide RNA, no tracrRNA).

ORF and CRISPR metadata. Sequences of ORF reagents and catalog number of the CRISPR reagents from the Human Edit-R synthetic crRNA-Druggable Genome library from Revvity Discovery Limited (formerly known as Horizon Discovery) are available via GitHub (https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/tree/v1.0.0/13.orf-crispr-metadata).

Sample preparation and image acquisition

The Cell Painting assay uses six fluorescent dyes, imaged across five channels, to visualize eight key cellular components: mitochondria (MitoTracker, Mito), nucleus (Hoechst, DNA), nucleoli and cytoplasmic RNA (SYTO 14, RNA), endoplasmic reticulum (concanavalin A, ER), Golgi and plasma membrane (wheat germ agglutinin) plus the actin cytoskeleton (phalloidin, alpha 1-acid glycoprotein (AGP)). We followed the optimized Cell Painting assay protocol²² (https://github.com/carpenterlab/2022_Cimini_NatureProtocols/wiki). Image acquisition for the ORF dataset was performed using the Revvity Opera Phenix microscope in the widefield mode and for the CRISPR dataset using Yokogawa CV7000 in the confocal mode. Only fluorescent channels were acquired for the CRISPR dataset, while three brightfield planes were also acquired for the ORF dataset. Nine fields of view were acquired for both the ORF and CRISPR datasets.

Image processing

We used CellProfiler bioimage analysis software (v.4.1.3 or v.4.2.1) for image processing. We corrected background illumination variations⁶⁸ and segmented nuclei and cells using classical segmentation algorithms, namely, CellProfiler's IdentifyPrimaryObjects module with Minimum Cross-Entropy thresholding and IdentifySecondaryObjects module with Otsu three-class watershed thresholding, respectively. Subsequently, we measured a suite of features for each cell in every fluorescent channel, including intensity, granularity, texture and density. We extended this analysis to brightfield planes for the ORF dataset. In addition, we quantified features at the whole-image level (refer to <http://broad.io/cellprofilermanual> for detailed methodology. Our image analysis pipeline (https://github.com/broadinstitute/imaging-platform-pipelines/tree/fc10d6acc5c1b2d9c4526a052acb1d9a196525a/JUMP_production#production-pipelines) generated up to 7,648 features (encompassing both per-cell and per-image measurements) for the ORF dataset. For the CRISPR dataset, we extracted 4,762 features per cell, as we did not measure any brightfield features.

Data exclusion

We excluded batch 12 from the ORF dataset due to incorrect dispensing of assay dyes in two plate rows. This batch was subsequently repeated as batch 13. In addition, to keep the features same across ORF, CRISPR and the compound datasets²¹, we kept only the 3,673 common features.

Overview of profile processing workflow

We mean-aggregated the single-cell features using Pycytominer⁶⁹ to generate well-level profiles. To correct for technical artifacts in the dataset, we began by correcting for systematic variations across well positions. We calculated the mean value for each feature within each well position and subtracted this mean from the corresponding feature values of individual samples. Next, recognizing that cell count can strongly contribute to data variability¹⁶, we addressed its influence on other features by regressing it out while retaining the original cell count feature for subsequent analyses. Following this, we applied a modified version of our profile processing pipeline for batch correction⁶⁴. We applied variance thresholding to remove features with minimal variation across the dataset. We then scaled the data based on its median absolute deviation, removed outlier features and performed feature selection to remove invariant and redundant features. Finally, we performed sphering transformation to decorrelate features and standardize variance, followed by Harmony batch correction⁷⁰ to mitigate technical artifacts while preserving biological variation.

CRISPR-Cas9-mediated gene knockouts, while intended to disrupt specific genes through targeted DNA cuts, can occasionally result in unintended large-scale deletions encompassing the entire remaining chromosome arm⁷¹⁻⁷³. Recent studies have identified systematic patterns in image-based and cell-line viability profiles, revealing similarities among knockouts of genes located on the same chromosome arm¹⁹.

We confirmed that this pattern is seen in our CRISPR dataset (<https://doi.org/10.5281/zenodo.13754407>, ref. 74), which had undergone different preprocessing than in the previous study, but, as expected, not in our overexpression dataset (<https://doi.org/10.5281/zenodo.13754178>, ref. 75). The proposed correction method¹⁹, involving principal components analysis followed by correction, successfully mitigated this pattern (<https://doi.org/10.5281/zenodo.13754508>, ref. 76). This correction was applied only to the CRISPR dataset.

This processing workflow can be replicated by running <https://github.com/broadinstitute/jump-profiling-recipe/tree/v0.1.0> using the appropriate config file for the CRISPR (crispr.json) and ORF datasets (orf.json). A schematic of the profile processing workflow is shown in Supplementary Fig. 33 and the number of features and samples remaining after each step in the processing workflow is shown in Supplementary Table 6.

We filtered ORF reagents based on viral infection efficiency, as determined by a CellTiter-Glo cell viability assay conducted in parallel with the Cell Painting plates. The infection efficiency values showed a bimodal distribution, with two outlier reagents (Supplementary Fig. 34). We established a threshold for low infection efficiency using Otsu thresholding. This filtration process eliminated all empty wells, as expected, due to the absence of virus in these wells. The unique, nonsymmetrical layout of empty wells in each plate map confirmed that plates were neither swapped nor rotated relative to their metadata. One exception was noted: plate map OAB41.OAC17.OAB78.79.A in batch 4, where three empty wells displayed strong infection efficiency. As the pattern did not improve with rotation and lacked a plausible explanation, we opted to exclude all five replicates of this plate map from the experiment. We also removed control wells and ORF reagents with infection efficiency below the Otsu threshold. Together, these quality control measures resulted in the removal of 2,397 ORF reagents. The infection efficiency of each ORF reagent is available via GitHub at https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/tree/v1.0.0/00.download-and-process-annotations/input/JUMP-ORF-Infection-Efficiencies.xlsx and the list of removed ORF reagents is available via GitHub at https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/tree/v1.0.0/00.download-and-process-annotations/output/orf-reagents-low-infection-efficiency-and-outliers.csv.gz.

Mean average precision (mAP) calculations for phenotypic activity and phenotypic consistency are detailed and defined²⁴. For example, mAP measures phenotypic activity when we calculate each replicate's ability to retrieve the other replicates for that gene against the background of negative control samples, using cosine similarity as the similarity metric. We assess statistical significance using permutation testing to obtain a *P* value. *P* values are adjusted for false discovery rate to yield a corrected *P* value (*q* value).

Gene annotations

Annotations were downloaded from publicly available databases as described in https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/tree/v1.0.0/00.download-and-process-annotations/README.md. To transform the predominantly continuous annotation data into categorical form for the Fisher's exact test, we established threshold values. For essentiality, we set the dependency probability threshold as 0.75, and for expression, we set the transcripts per million threshold as 75. For ORF gene insert size, we set the threshold as 2,500 nucleotides.

Single-cell PhenoSorter models

Raw images were uploaded to Spring Engine (a cloud-based software-as-a-service product provided by Spring Science). We used the PhenoSorter application to generate a single-cell-based phenotypic model that discriminated between flat and rounded cells. Briefly, this application presents individual segmented cell crops to a user, who classifies

them as rounded cells or flat cells. These cell crops are processed using ReplKNet, a pretrained large kernel convolutional neural network⁷⁷. The final layer is pooled using average pooling to yield a 128-element embedding vector for each imaging channel. Classification models are built on this embedding vector using XGBoost⁷⁸, a gradient-boosting decision tree-based framework. For this model, embeddings were concatenated from all five fluorescence channels used in the ORF dataset. The training uses an 80/20 split in which 80% of the user-classified cell images are randomly chosen to train the model, and the remaining 20% are tested to determine model performance. To accommodate for different example set sizes, the loss function is weighted so that examples from smaller classes are given greater weight, such that each class is ultimately balanced despite the different sample set sizes. Once trained, the model is applied to all single-cell image crops in the entire dataset. The data for all cells in each well are aggregated to yield a percentage value for each model class for every well, and we applied a threshold of 6.5% reduction in the percentage of living cells compared to negative controls to call hits.

Gene set enrichment analysis for the PhenoSorter model

We evaluated the results of our machine learning classification models by performing a gene set enrichment analysis on the ORF perturbations that demonstrated the greatest effect of the model. For the rounded-cell model, we compared the fraction of cells within each well that the model scored as flat for each ORF treatment against that same fraction of cells for untreated cells. ORF treatments were then ranked by the percentage reduction in flat cells. A cutoff of 6.5% increase in rounded cells was used to generate our gene subset for genes demonstrating increased levels of cell roundedness. The R programming language package clusterProfiler⁷⁹ was used for gene set enrichment analysis. The WikiPathways annotation set of the C2 subset of annotations sourced from the MSig database was used for gene annotations.

Identifying relevant datasets using Plex

Genes sets in selected clusters (that is, genes similar to *YAP1* in the ORF data, genes positively or negatively correlated with *TSC22D1* and the cluster containing *ECHI*, *UQCRFS1* and *SARS2*) were searched in the Plex Research web app (<https://plexresearch.com>)³⁷. Plex combines KGs with centrality algorithms to find convergences between search inputs and diverse, large, publicly available datasets such as Gene Expression Omnibus⁸⁰, Opentargets (<https://www.opentargets.org>) and the Broad Cancer Dependency Map Project (<https://depmap.org/portal>).

Gene-gene functional links quantified via KGs

Biomedical KGs represent information in the form of nodes (also called entities) connected by edges (also called links), where entities can have various types (for example, gene, compound, biological function, pathway, disease and so on) and the links can represent various facts connecting them (for example, a gene involved in a pathway or a disease, two genes cocited in an article, a compound is used to treat a disease). We used the DRKG²⁶, a comprehensive open-source biomedical KG relating genes, compounds, diseases, biological processes, side effects and symptoms. DRKG includes information from six existing databases, including DrugBank, Hettionet, Global Network of Biomedical Relationships, String, IntAct and DGIdb, with 4,078,154 edges (of 107 edge types) and 66,500 nodes (of 13 node types), after filtering out nonhuman genes and edges. The DRKG graph includes the results of large-scale scientific literature mining coming from the PubTator project through the Global Network of Biomedical Relationships.

We predict links between genes and functions in DRKG with a GNN approach. The analysis is limited to the 6,787 genes (ORF data) and 5,494 genes (CRISPR data) having a phenotype and existing in DRKG (most 6,787/7,031 = 97% and 5,494/5,546 = 99%). Graph representation learning methods generate vector representations for graph nodes such that the learned representations, that is, embeddings, capture

the structure and semantics of networks⁸¹. To introduce a numerical score reflecting biological functional proximity between two genes from multiple sources of evidence, we trained three predictive models using the DRKG and GNN approach, aimed at predicting the molecular function of a gene using GO Molecular Function definitions as training set, the biological process in which a gene is involved using GO Biological Process definitions as training set or the pathway in which a gene is involved (using definition of pathways from various resources, including WikiPathways and Reactome as training sets). As an architecture for GNN, we implemented a variational graph autoencoder⁸² in Pytorch Geometric⁸³ for predicting links between genes and functions, using the whole topographical structure of the KG inspired by https://colab.research.google.com/drive/1jv0GrF11jcbhiV7dK-RhxKcV_GLVvTIs?usp=sharing#scrollTo=sBWeMphti_y2. This model makes use of latent variables and is capable of learning interpretable latent representations for undirected graphs. For training, we used a standard validation strategy by retaining a part of the gene-function links from the training set as a validation set and measuring the performance of the prediction through the metrics area under the curve and Hits at k (Hits@ k)⁸⁴. Hits@ k measures the proportion of correct predictions within the top k -ranked candidates, that is, it denotes the ratio of the test triples that have been ranked among the top k triples⁸⁵.

For a final validation of the models, we applied a time machine approach⁸⁶, where computational models are trained using data published before a certain time point, and the model outputs are validated by their ability to predict links that became part of biomedical knowledge after this time point. Specifically, for validation we used those links from the latest GO^{87,88} that were established after 2020 and thus are not a part of DRKG, which was created in 2020. The prediction accuracy was high in both validation and test sets (Supplementary Table 7).

Each of the trained models provides a score $x_{g_i, f}$ for any pair of a gene g and a node f in DRKG representing a biological function (the GO Molecular Function category, GO Biological Process category or a pathway), where the known links typically score on top while other relatively large scores represent predictions of a gene function from the whole content of the KG. For a given type of biological function, the functional KG score between two genes g_i and g_j is computed as

$$SC_{KG}^F(g_i, g_j) = 2 \times \text{sigmoid}\left(\frac{\text{dot}(\mathbf{x}_{g_i}, \mathbf{x}_{g_j})}{k\sigma_F}\right) - 1,$$

where $\text{dot}()$ is the standard dot product; $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ is a sigmoid function; F is the type of the biological function node in KG, $F \in \{\text{GO_MF}, \text{GO_BP}, \text{pathway}\}$; \mathbf{x}_{g_i} is the centered vector of scores $x_{g_i, f}, f \in F$; σ_F is the standard deviation of the distribution of $\text{dot}(\mathbf{x}_{g_i}, \mathbf{x}_{g_j})$ values for all possible gene pairs g_i, g_j and k is the spreading coefficient empirically chosen at 1.5 in all experiments to avoid the saturation of sigmoid function values close to 1.0.

Therefore, the functional score between two genes is a value in $[-1; 1]$ interval, with values close to 1.0 representing two closely functionally related genes (these are usually the genes involved in the same family, with physically interacting gene products, known to be involved in the same biological function and so on) and -1.0 representing genes maximally separated in the KG. Empirically, we observed that the distribution of SC_{KG}^F scores has the maximum density for the score close to 0.2 (Fig. 3) with a relatively small number of negative values. We use 0.5 as a cutoff for whether a particular relationship is previously known.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Well-level morphological profiles, image analysis pipelines, profile generation pipelines and instructions for retrieving the cell images,

single-cell profiles and well-level profiles are publicly available via GitHub at https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap/blob/main/README.md. Plate maps and other metadata are available via GitHub at <https://github.com/jump-cellpainting/datasets/tree/main/metadata>. Cell Painting images and single-cell profiles are available via Cell Painting Gallery on the Registry of Open Data on AWS (<https://registry.opendata.aws/cellpainting-gallery/>) under accession number cp0016-jump. We have released the data with a CCO license. Several external biological databases provided critical reference information for analysis and interpretation. The DRKG was used for KG-based analyses. GO and WikiPathways annotations were obtained from the Molecular Signatures Database (Msigdb v.2023.1.Hs and v.2023.1.Hs). Human disease annotations were sourced from the Human Protein Atlas (v.23), and protein complex annotations came from the Comprehensive Resource of Mammalian protein complexes (CORUM v.4.1). Standardized gene nomenclature was referenced from Genenames.org and HGNC, while homology information was derived from HomoloGene (build 68). Finally, data regarding gene essentiality, gene expression and gene effect specifically in U2OS cells were obtained from the DepMap 23Q4 release. Source data are provided with this paper.

Code availability

The code for generating all the figures is available via GitHub at https://github.com/jump-cellpainting/2025_Chandrasekaran_NatureMethods_Morphmap. It is available under the BSD 3-clause license, a permissive open-source license.

References

- Johannessen, C. M. et al. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation. *Nature* **468**, 968–972 (2010).
- Berger, A. H. et al. High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* **30**, 214–228 (2016).
- Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
- Singh, S., Bray, M.-A., Jones, T. R. & Carpenter, A. E. Pipeline for illumination correction of images for high-throughput microscopy. *J. Microsc.* **256**, 231–236 (2014).
- Serrano, E. et al. Reproducible image-based profiling with Pycytominer. *Nat. Methods* **22**, 677–680 (2025).
- Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
- Tsuchida, C. A. et al. Mitigation of chromosome loss in clinical CRISPR–Cas9-engineered T cells. *Cell* **186**, 4567–4582 (2023).
- Nahmad, A. D. et al. Frequent aneuploidy in primary human T cells after CRISPR–Cas9 cleavage. *Nat. Biotechnol.* **40**, 1807–1813 (2022).
- Przewrocka, J., Rowan, A., Rosenthal, R., Kanu, N. & Swanton, C. Unintended on-target chromosomal instability following CRISPR/Cas9 single gene targeting. *Ann. Oncol.* **31**, 1270–1273 (2020).
- Chen, Z. & Chandrasekaran, S. N. Similarity of CRISPR genes grouped by chromosome location without chromosome arm correction. *Zenodo* <https://doi.org/10.5281/zenodo.13754407> (2024).
- Chen, Z. & Chandrasekaran, S. N. Similarity of ORF genes grouped by chromosome without chromosome arm correction. *Zenodo* <https://doi.org/10.5281/zenodo.13754178> (2024).
- Chen, Z. & Chandrasekaran, S. N. Similarity of CRISPR genes grouped by chromosome location with chromosome arm correction. *Zenodo* <https://doi.org/10.5281/zenodo.13754508> (2024).
- Ding, X. et al. Scaling up your kernels to 31×31: revisiting large kernel design in CNNs. Preprint at <https://arxiv.org/abs/2203.06717> (2022).

78. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794* (Association for Computing Machinery, 2016).
79. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
80. Clough, E. et al. NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Res.* **52**, D138–D144 (2024).
81. Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.* **6**, 1353–1369 (2022).
82. Kipf, T. N. & Welling, M. Variational graph auto-encoders. Preprint at <https://arxiv.org/abs/1611.07308> (2016).
83. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. Preprint at <https://arxiv.org/abs/1903.02428> (2019).
84. Bonner, S. et al. Understanding the performance of knowledge graph embeddings in drug discovery. *Artif. Intell. Life Sci.* **2**, 100036 (2022).
85. Ali, M. et al. Bringing light into the dark: a large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 8825–8845 (2022).
86. Ren, F. et al. A small-molecule TN1K inhibitor targets fibrosis in preclinical and clinical models. *Nat. Biotechnol.* **43**, 63–75 (2024).
87. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
88. Gene Ontology Consortium, et al. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).

Acknowledgements

We thank the scientists across the entire JUMP Cell Painting Consortium for their guidance and support throughout the project. We acknowledge a grant from the Massachusetts Life Sciences Center Bits to Bytes Capital Call program for funding the data production and catalyzing this Consortium. We appreciate funding to support data analysis and interpretation from members of the JUMP Cell Painting Consortium (Amgen, AstraZeneca, Bayer AG, Biogen, Eisai, Janssen Pharmaceutica NV, Merck KGaA, Darmstadt, Germany, Pfizer, Servier, Takeda Development Center Americas, Inc. (TDCA)), from the National Institutes of Health (NIH grant no. MIRA R35 GM122547 to A.E.C.), and from grant number 2020-225720 to B.A.C. from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. We acknowledge the Consortium's Supporting Partners for their contributions: Ardigen for their deep learning expertise and JUMP-CP Data Explorer web application (part of Ardigen's phenAID platform); Google/Verily for the compute support and configuration/optimization of Terra, which is codeveloped by the Broad Institute of MIT and Harvard, Microsoft and Verily (its use is not described in this paper); Nomic bio for their protein profiling (not described in this paper); and Revvity for the PhenoVue Cell Painting JUMP kit and the Edit-R Libraries and Edit-R tracrRNA. We also are grateful for the Amazon Web Services Registry of Open Data for hosting the public dataset. We also acknowledge the use of the Revvity Opera Phenix High-Content/High-Throughput imaging system at the Broad Institute, funded by the S10 grant no. NIH OD-026839.

Author contributions

S.N.C. was involved in the analysis, data curation, experiments and investigation, methodology, project administration, software, supervision, validation, visualization, writing the original draft and writing—review and editing. E.A. conducted experiments and investigation and provided resources. J.A. was involved in the analysis, data curation, software and validation. A.B. was involved in the methodology, supervision and writing the original draft.

P.J.B. was involved in the data curation, experiments and investigation and validation. W.G.C. was involved in the data curation, software and visualization. Z.S.C. was involved in the analysis, data curation, experiments and investigation, methodology, software, validation and writing—review and editing. B.A.C. was involved in the data curation, methodology, project administration, software, supervision, writing the original draft and writing—review and editing. B.D. was involved in the experiments and investigation. J.G.D. was involved in the project administration, resources and supervision. J.D.E. was involved in the experiments and investigation and writing the original draft. B.F. was involved in the experiments and investigation, methodology, project administration, resources, supervision and writing the original draft. C.J.F. was involved in the analysis, data curation, software, validation and writing—review and editing. J.G. was involved in the analysis and writing the original draft. A.G. was involved in the experiments and investigation, methodology, project administration, resources, supervision and writing the original draft. M.H. was involved in the analysis, experiments and investigation and software. Y.H. was involved in the analysis, software and validation. Z.H. was involved in the analysis, data curation and validation. H.H. was involved in the analysis, conceptualization, methodology, project administration, supervision, visualization and writing the original draft. D.H. was involved in the experiments and investigation and resources. C.B.J. was involved in the data curation, experiments and investigation, project administration and supervision. T. James was involved in the funding acquisition and supervision. T. Jetka was involved in the analysis, methodology, project administration, supervision, visualization and writing the original draft. A.A.K. was involved in the analysis, experiments and investigation, methodology, software and writing—review and editing. B.K. was involved in the analysis, data curation, software, validation and writing—review and editing. M.K.-A. was involved in the data curation, experiments and investigation, methodology, project administration, resources, supervision and writing—review and editing. T.K. was involved in the software and supervision. B.A.M. was involved in the experiments and investigation, methodology, project administration, resources and validation. G.M. was involved in the experiments and investigation, methodology and resources. N.J.M. was involved in the experiments and investigation, funding acquisition and supervision. L.M. was involved in the experiments and investigation, methodology, project administration, validation and writing the original draft. A.M. was involved in the analysis, data curation, methodology, software, visualization, writing the original draft and writing—review and editing. N.M. was involved in the analysis. A.F.M. was involved in the analysis, data curation, experiments and investigation, methodology, software, validation, writing the original draft and writing—review and editing. A.O. was involved in the analysis, project administration and supervision. M.O. was involved in the supervision, writing the original draft and writing—review and editing. K.R. was involved in the analysis and data curation. D.E.R. was involved in the methodology, project administration, resources and supervision. F.R. was involved in the analysis, data curation, software, validation and writing—review and editing. S.S. was involved in the experiments and investigation, project administration and resources. D.W.S. was involved in the analysis and writing the original draft. R.A.S. was involved in the analysis. P.S. was involved in the conceptualization and supervision. A.T. was involved in the experiments and investigation and validation. S.T. was involved in the analysis, data curation, software, validation and writing—review and editing. R.V.V. was involved in the analysis, data curation, methodology, software, visualization, writing the original draft and writing—review and editing. W.J.V.T. was involved in the analysis, data curation, methodology, software, supervision, visualization, writing the original draft and writing—review and editing. S.W. was involved in the analysis, data curation, experiments and investigation, validation and writing—review and editing. M. Warchot was involved in the

project administration, supervision and writing—review and editing. E.W. was involved in the analysis, data curation, experiments and investigation, software, validation and writing—review and editing. A.W. was involved in the experiments and investigation and validation. M. Wiest: analysis, data curation, software, validation and writing—review and editing. A. Zaremba was involved in the analysis and visualization. A. Zinovyev was involved in the analysis, data curation, methodology, software, supervision, visualization, writing the original draft and writing—review and editing. S.S. was involved in the analysis, conceptualization, data curation, experiments and investigation, funding acquisition, methodology, project administration, software, supervision, validation, visualization, writing the original draft and writing—review and editing. A.E.C. was involved in the analysis, conceptualization, funding acquisition, methodology, project administration, supervision, visualization, writing the original draft and writing—review and editing.

Competing interests

The authors declare the following competing interests: S.S., B.A.C. and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C., Recursion, SyzOnc, Quiver Bioscience; S.S., Waypoint Bio, Dewpoint Therapeutics, Deepcell and

B.A.C., Marble Therapeutics) and receive honoraria for occasional scientific visits to pharmaceutical and biotechnology companies. Authors with affiliations to the pharmaceutical, technology and biotechnology companies listed are or have been employees of those companies and may have real or optional ownership therein. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-025-02753-9>.

Correspondence and requests for materials should be addressed to Shantanu Singh or Anne E. Carpenter.

Peer review information *Nature Methods* thanks Susanne Muller, Sukjoon Yoon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used to collect data.

Data analysis

Image analysis and feature extraction were performed using CellProfiler. Subsequent feature processing was handled by Pycytominer, and batch effects were corrected using Harmony. Core data analysis was conducted within Jupyter notebooks, leveraging a suite of Python packages including Pandas, Numpy, Scipy, Pyarrow, copairs, and scikit-learn. Further specialized analyses included gene enrichment analysis using clusterProfiler and the identification of cells with specific phenotypes using Phenosorter. Data visualization utilized various libraries: Matplotlib, Matplotlib-venn, Seaborn, and Plotly were employed for generating standard plots, while UMAP was used for visualizing dimensionality reductions. Direct image visualization was facilitated by Tiffle and scikit-image, Pycirclize was used for creating circos plots, and final figure panels were assembled using Inkscape. Data was stored on AWS S3 and accessed programmatically via Boto3. A subset of the dataset was archived on Zenodo, and associated code was version controlled and stored on GitHub. Interactive data exploration is available through the Spring Discovery or Ardigen's phenAID portals, as well as via Morpheus or datasette.io. Interpretation of gene relationships was supported by Plex research's web application.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Well-level morphological profiles, image analysis pipelines, profile generation pipelines, and instructions for retrieving the cell images, single cell profiles and well-level profiles are publicly available online at https://github.com/jump-cellpainting/2024_Chandrasekaran_Morphmap/blob/main/README.md. Plate maps and other metadata are available online at <https://github.com/jump-cellpainting/datasets/tree/main/metadata>. Cell Painting images and single-cell profiles are available at the Cell Painting Gallery on the Registry of Open Data on AWS (<https://registry.opendata.aws/cellpainting-gallery/>) under accession number cpg0016-jump. We have released the data with a CC0 license.

Several external biological databases provided critical reference information for analysis and interpretation. The Drug Repurposing Knowledge Graph (DRKG) was utilized for knowledge graph-based analyses. Gene ontology and WikiPathway annotations were obtained from the Molecular Signatures Database (Msigdb versions 2023.1.Hs and 2023.1.Hs). Human disease annotations were sourced from the Human Protein Atlas (v23), and protein complex annotations came from the Comprehensive Resource of Mammalian protein complexes (CORUM version 4.1). Standardized gene nomenclature was referenced from Genenames.org and HGNC, while homology information was derived from HomoloGene (build 68). Finally, data regarding gene essentiality, gene expression, and gene effect specifically in U2OS cells were obtained from the DepMap 23Q4 release.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Data exclusions

Replication

Randomization

Blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)

The U-2 OS cell line was obtained directly from ATCC (Catalog #HTB-96, Lot #70016635).

Authentication

Since the cells were ordered directly from ATCC, we did not authenticate via SNP or STR profiling.

Mycoplasma contamination

We tested the parental cells ordered from ATCC for mycoplasma contamination a month into culturing them to ensure they remained negative.

Commonly misidentified lines
(See [ICLAC](#) register)

To our knowledge U-2 OS is not commonly misidentified.