Scalable imaging-based profiling of CRISPR perturbations with protein barcodes

Krishna Choudhary[4], Michael T. McManus[1-5]

1. Chan Zuckerberg Biohub San Francisco, San Francisco, California, 94158, United States of America
2. UCSF Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA
3. Institute for Human Genetics (IHG), University of California, San Francisco, San Francisco, CA, USA
4. UCSF Diabetes Center, University of California, San Francisco, San Francisco, CA 94143, United States of America.
5. Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94143, United States of America.

Keywords: imaging functional genomics; CRISPR interference; optical barcoding; mitotic gene network; genetic interaction mapping

Running Title: Mapping Genetic Interactions Using poolVis and cellPool

email: Krishna.Choudhary.Research@gmail.com, Michael.McManus@ucsf.edu

ABSTRACT

Imaging-based CRISPR screens enable high-content functional genomics by capturing phenotypic changes in cells after genetic perturbation. Protein barcodes provide cost-effective, easy-to-implement, and imaging-compatible barcoding for pooled perturbations, yet their scalability has been constrained by the need for arrayed cloning, lentiviral recombination between barcodes and guides, and difficulties in decoding barcodes with high confidence. Here, we introduce poolVis and cellPool, an integrated experimental and computational platform designed to address these limitations. poolVis uses Cre-lox-mediated reconfiguration to position barcode-sgRNA pairs in proximity during viral integration, which greatly reduces barcode shuffling during pooled cloning and delivery. cellPool leverages a scalable computational workflow and the unique aspects of protein barcodes to produce unpooled image galleries from multi-terabyte scale datasets. Applying this platform to single- and double-CRISPRi profiling of cell-cycle genes and chromokinesins in the MCF10A cells uncovered established and previously unrecognized phenotypes, including nuclear morphology changes and reciprocal sign epistasis in DNA damage.

INTRODUCTION

Microscopic imaging provides rich information in space and time, revealing phenotypes that DNA sequencing cannot capture. Imaging-based functional screens have propelled cell biology and drug discovery.[1,2] Recent confluence of CRISPR technologies with high-throughput, multiplexed imaging has renewed interest in pooled imaging-based profiling of genetic perturbations.[2]

Profiling thousands of perturbations in isolation is labor-intensive. Optical barcoding reduces labor by pooling perturbations. To be practical, barcodes must be robustly decodable with controlled error rates. Current approaches rely on either RNA or protein barcodes.[2]

While RNA barcoding via *in situ* sequencing offers valuable spatial information, it is costly and requires multiple amplification steps, high-temperature incubations, artisan in situ sequencing techniques, and potentially cell-line-specific spot detection methods.[3,4] Protein barcodes, in contrast, use immunostained epitope tags fused to stable proteins expressed by strong promoters, and are compatible with available automation solutions.[5,6] Their nuclear protein (e.g., H2B) fusions give bright, easy-to-detect signals even at low magnification. However, an unexpected and often underestimated burden is antibody panel optimization; failing this, debarcoding performance may suffer.

Existing lentiviral vector designs position protein barcodes and sgRNAs far apart (1 kb or more) on the backbone and are prone to recombination during viral packaging, a major source of barcode-sgRNA mispairing. In addition, libraries with identical coding sequence for an epitope tag across barcodes can exacerbate lentiviral recombination and PCR-mediated template switching during amplification, further compounding linkage errors[7–10]. Consequently, the scalability of pooled protein-barcoded libraries has remained limited.[5,11] Although one large-scale protein barcode study has been reported[6], it did not involve CRISPR screens, where the absence of gene-specific phenotypes simplified interpretation despite potential barcode shuffling. Thus, the need to scalably generate protein-barcoded libraries, read barcodes cost-effectively, and ensure performance quality despite staining variation remains unmet.

To address these challenges, we present a platform addressing these needs. It combines a wet-lab component, poolVis (Fig. 1a), with a computational pipeline, cellPool (Fig. 2a). Our platform differs from prior art by (i) use of Cre-lox reconfiguration that maintains ~50 bp protein barcode-sgRNA proximity in poolVis vectors to suppress template switching and barcode shuffling,[7,12,13] (ii) a TATAlox-U6 architecture and non-homologous barcode coding sequences to stabilize barcode and sgRNA expression with Cre induction in the transduced cells, and (iii) cellPool, a Nextflow-based workflow that performs high-quality image processing and multivariate model-based debarcoding to output unpooled single-cell galleries for quality control and interpretation.

We apply poolVis and cellPool to profile 336 single and 2,016 double CRISPR interference (CRISPRi) perturbations. To minimize antibody use and imaging time, we arrayed minipools of 84 barcoded sgRNAs, readable in three cycles. This design scales to >1,000 sgRNAs in a dozen wells, focused on specific pathways, and by extrapolation to genome-scale libraries using two 96-well plates. We recover known and previously unrecognized phenotypes, validating hits

individually. Together, these results establish a cost-effective and scalable platform for imaging-based profiling of CRISPR perturbations.

RESULTS

poolVis enables pooled CRISPR perturbations with protein barcodes

We developed poolVis to bring pooled barcode-sgRNA pairs closer than 100 bp. Our prototype considered Cas12a-CRISPR knockout because of shorter Cas12a scaffold than Cas9.[14] Here, barcodes were fused to EF1α-driven H2B-GFP, while the human U6 (hU6) promoter downstream of stop codon drove the sgRNA (Fig. 1b, top design). In flow cytometry with enAsCas12a-expressing K562s and CD81-targeting sgRNA, GFP signal and CD81 knockout were poor (Fig. 1c-d top row; Supplementary Table 1a-b; Supplementary Fig. 1a), suggesting EF1α and U6 transcriptional interference.[15]

To resolve this, we leveraged Cre-lox reconfiguration. We added lox71 between barcode and sgRNA in the oligonucleotide inserts and lox66 in the vector backbone. After viral integration, Cre inverts the floxed cassette, rearranging the barcode and sgRNA, and lox66/71 unidirectionally converts to functionally inert loxP/72. This preserves sgRNA-barcode proximity until Cre induction, after which transcriptional interference can be prevented.

To maximize barcode and sgRNA expression, we evaluated several constructs. A one-transcript design, where Cre recombination produced a single mRNA encoding the H2B-GFP barcode and the sgRNA separated by a stabilizing Triplex motif (Fig. 1b, second design), was inspired by prior mRNA-embedded guide strategies.[16] However, puromycin selection revealed poor viability, suggesting mRNA destabilization (Fig. 1c-d, second row). We therefore developed designs that reconfigure barcodes and sgRNAs into separate transcripts.

Building on these results, we tested configurations where Cre recombination repositions U6 either downstream or upstream of EF1α (Fig. 1b, third and fourth designs). Using *CD81*-targeting sgRNA, we observed the best combined CRISPR activity and barcode signal with U6 upstream of EF1α (Fig. 1c-d, bottom rows). Here, 40 bp barcode-sgRNA separation reconfigured into highly functional units.

To generalize the design, we next adapted it for Cas9 and dCas9-KRAB systems. The recombined loxP sequence is fused to the Cas12a-sgRNA in the best design above and removed by Cas12a, which Cas9 cannot do. Also, the long Cas9 scaffold should be in the vector backbone. We therefore implemented a TATAlox-mU6 promoter, embedding loxP downstream of the mouse U6 (mU6) proximal element.[17] To promote unidirectional recombination, we engineered TATAlox66/71 variants (Supplementary Table 1c). To improve lentiviral titer and eliminate significant recombination, we oriented EF1α tandem with CMV and reduced barcode-sgRNA separation to 51 bp (Fig. 1e). In HEK293Ts, these design implementations yielded efficient Cas9 *CD81* knockout, dCas9-KRAB *CD81* repression, and barcode signals (Fig. 1f-g; Supplementary Fig. 1b-c), and *AURKB* repression caused expected nuclear defects (Fig. 1h). To further improve gene knockdown, we additionally produced TATAlox-hU6 promoters (Supplementary Fig. 1d-f; Supplementary Table 1d). These results establish poolVis for pooled protein-barcoded Cas12a, Cas9, and dCas9-KRAB libraries.

Design of non-homologous coding sequences for protein barcodes

Template switching and barcode recombination can occur between identical nucleic acid sequences during lentiviral reverse transcription or PCR amplification. To reduce these artifacts, we codon-optimized epitope tag and barcode sequences to minimize homology. A brute-force pairwise comparison to minimize library-scale sequence homology is computationally impractical, e.g., V5-tag admits $>10^8$ possible coding sequences. We therefore designed a heuristic algorithm to generate 84 minimally homologous barcode coding sequences. The algorithm fragments each peptide into short motifs, enumerates codons per motif, and iteratively assembles coding sequences 5′→3′ while minimizing homology to previously sampled coding sequences. This process produced a diverse set of non-homologous barcode coding nucleic sequences resistant to template switching and recombination (Fig. 1i, Supplementary Table 1e).

cellPool enables scalable analysis of high-content imaging data

cellPool is a Nextflow[18] pipeline that transforms multiplexed imaging data into unpooled image galleries of debarcoded cells (Fig. 2a). Starting with illumination correction, imaging tiles are stitched together, nuclei are segmented, tracked across cycles, barcode intensities quantified, and debarcoded, and single-cell galleries are generated. Containerization ensures reproducibility and straightforward deployment on high-performance computing systems.

To benchmark cellPool, we transduced 19 barcodes individually in U2OS in a pre-determined plate map (Fig. 2b). We imaged and analyzed the samples without the map. cellPool recovered barcode identities with high sensitivity and specificity (Fig. 2c) attributable to our mosaicking, tracking, and debarcoding methods.

cellPool yields high-quality integration of iterative imaging data

ASHLAR is a state-of-the-art method for multiplexed imaging integration using phase correlation,[19] which estimates translational misalignments but not other transformations. It has been validated for five microscopes, predominantly pathology slide scanners. Our experiments used a well-known high-content microscope where ASHLAR frequently misaligned images (Supplementary Video 1), increasing signal-to-background overlap (Fig. 2d). Hence, cellPool stitches image tiles from each cycle, but skips pixel-to-pixel registration across cycles. This ensures robustness to instrumental variations. cellPool mosaicking is based on ASHLAR algorithm and outputs detailed quality reports and videos showing the alignments to enable human-in-the-loop fine tuning (Supplementary Fig. 2-3; Supplementary Videos 2 and 4).

cellPool unpools registered images, but internally skips pixel-to-pixel registration. This ensures robustness to unforeseen artifacts. cellPool integrates the single-cell information across imaging cycles by tracking in multiple iterations. It leverages invariance of fixed-cell nuclear orientations in the first tracking iteration (Fig. 2e-g). Although largely random, neighboring nuclei in dividing or nascently divided cells may present identical orientations causing tracking errors, which are identified based on arrangement of neighboring nuclei and corrected by positioning relative to the nuclei correctly tracked in the previous iteration. Our approach gives near perfect tracking confirmed by 100s of randomly sampled tracking mates (Supplementary Videos 3 and 5).

cellPool enables accurate decoding of protein barcodes

cellPool uses model-based clustering for debarcoding premised on the following observations. First, artifacts such as antibody aggregates may lead to multimodality, and poor signal may cause unimodality and heavy tails (Fig. 2h). Second, the multivariate barcode distributions are not elliptical, symmetric, or convex (Fig. 2i). Third, the epitope intensities depend on H2B-GFP signal and have covariance patterns that can be leveraged for debarcoding (Fig. 2j).

Debarcoding starts with a high-stringency non-parametric univariate thresholding by comparing signal-background mixture with background only populations (Fig. 2h). This reduces the downstream computational complexity by partial debarcoding without invoking distributional assumptions. Next, we group the nuclei into 5-10 sets by H2B-GFP intensity (Fig. 2i-j) and learn covariance patterns by expectation-maximization (Supplementary Fig. 4), which extends prior work on skew-elliptical distributions for flow cytometry data[20] to protein barcodes. This gives a barcode score matrix and assignments per nucleus. In benchmarking test, cellPool consistently outperformed EPICode[6] and single-cell deconvolution (SCD)[21] (Fig. 2k) as expected given explicit covariance modeling.

CRISPRi screen of cell-cycle genes in MCF10A

Cell cycle regulation ensures controlled growth and DNA replication. MCF10A breast epithelial cells are a physiologically relevant model to study this process given their stable near-diploid karyotype and intact cell-cycle checkpoints with wild-type *TP53* and *RB1*.[22] Profiling cell-cycle perturbations in MCF10As can reveal distinct functions compared to tumorigenic HeLa or A549.[23]

We applied poolVis and cellPool to profile 336 CRISPRi perturbations of cell-cycle targets (Fig. 3a; Supplementary Table 1f-g). Guides were arranged as four minipools of 84 barcoded sgRNAs each. MCF10As with genomically integrated dCas9-KRAB and ERT2-Cre-ERT2 were processed as in Fig. 1a. We used a TATAlox-mU6-driven sgRNA cassette here, which achieved 71% CD81 knockdown relative to 98% under hU6 (Supplementary Fig. 5a), indicating partial loss-of-function for majority targets as expected for the mU6 promoter.[24]

We profiled nuclei, plasma membrane, and H2B-GFP, followed by DNA damage/repair using anti-γH2AX, mitotic index using anti-phospho-histone H3 (H3S10P), and microtubules using anti-α-tubulin (Fig. 3b-c). Then, barcode profiling took three iterations, costing approximately 10-fold less per cell compared to RNA barcodes (Supplementary Table 1h). Analysis reports confirmed high-quality image processing (Supplementary Videos 4-5). With debarcoding error rate estimated at 1.4% for our confidence threshold (Fig. 2k), 454,943 debarcoded nuclei were considered for downstream analysis, excluding uninfected and multiply infected cells (median: 996; Supplementary Fig. 5b). Unpooled images revealed expected phenotypes (Fig. 3d-f).

Next, we conducted differential analysis (Supplementary Table 2). As expected for cell-cycle genes, most knockdowns caused enlarged nuclei relative to the non-targeting controls likely due to checkpoint activation and/or endoreplication[25] (Fig. 3g). The chromosomal passenger complex genes, *BIRC5*, *AURKB*, *CDCA8*, and *INCENP*, were among the top hits (Fig. 3g-h). *MAPRE2* was the only hit for smaller nuclei, which is consistent with a recent RNAi screen.[26]

The screen revealed established hits for DNA damage/repair (Fig. 3i). *TP53* knockdown increased the integrated nuclear γH2AX signal the most (Fig. 3i-j). Other established hits for increased DNA

damage were *CHEK1* and *BRCA1* (DNA repair), *NCAPH* (condensin I), *PDS5A* (cohesin regulator), and *BUB1*, *BUB1B* and *MAD2L1* (spindle assembly checkpoint). As expected, *H2AX* knockdown reduced γH2AX signal. In general, cell-cycle dysregulation increases DNA damage due to mechanisms such as replication stress, impaired DNA repair, checkpoint loss, and cytokinesis failures. However, γH2AX staining measures both DNA damage and repair, specifically, functional expression, trafficking and assembly of γH2AX at the damage sites. For multifunctional genes, γH2AX signal summarizes multiple processes.

Interestingly, more perturbations reduced γH2AX signal than enhanced. Though *UHRF1* maintains genome integrity, its knockdown reduces γH2AX assembly at damage sites[27] explaining the reduced γH2AX signal. Since microtubules help transport DNA repair proteins,[28] microtubule-associated hits such as *KIFC1*, *TACC1*, and *MAPRE2* are expected. *RPS6KA3* and *RPS6KB1* phosphorylate ribosomal protein S6 to stimulate translation. Their knockdown causes overall reduced cellular protein, plausibly including γH2AX. *NPM1*, involved in ribosome biogenesis, also facilitates DNA repair. Reduced γH2AX after *NPM1* knockdown suggests that its partial loss-of-function impacts translation more than DNA repair. Similarly, multifunctional genes *SNX9* and *SNX33*, well-known for roles in cytokinesis, reduced γH2AX (Fig. 3j). Overall, our data highlight that for multifunctional genes, perturbation dose characterization facilitates interpretations. Alternatively, more repair proteins could be profiled with our platform.

H3S10P peaks during the late G2/M phase. Perturbations that arrest cells in the late G2/M phases should increase the H3S10P signal. Since only a minority of cells is expected in the late G2/M phases, there was higher variability in mean H3S10P of nontargeting controls than for mean γH2AX (Fig. 3i). Despite this, we uncovered established hits, e.g., *GSG2* and *CENPA*, and observed correlated mean H3S10P and γH2AX. Further, we identified mitotic nuclei based on H3S10P thresholding and computed the mitotic indices (Supplementary Fig. 5c). Given good coverage, the mitotic indices for the two guides targeting the same gene were correlated (Supplementary Fig. 5d), and in the expected range (1.02%) for the non-targeting and untransduced cells. *BRCA2*, *MAPK7*, *PRKCE*, *RB1*, which regulate cell-cycle progression and mitotic entry, and *KIFC1*, which regulates spindle organization and mitotic progression, were the top hits for lowered mitotic index (FDR <0.1). No hits increased mitotic index. Similar analysis revealed *KIF2A*, a microtubule depolymerase, as the top hit for enhanced α-tubulin staining (Supplementary Table 2). Overall, these observations support that our data is high-quality.

To validate, we cloned sgRNAs targeting selected genes and nontargeting sgRNAs individually in a control vector under hU6. Then, fully arrayed samples were processed to image cycles 1-2 as in Fig. 3b and analyzed by cellPool. For confident hits identified by FDR-adjusted p-values and single-cell coverages, the arrayed minipools and fully arrayed screens largely agreed (Supplementary Fig. 6), confirming that poolVis and cellPool deliver high-quality imaging-based profiles and robust analysis.

Arrayed minipools enable low-cost double-CRISPRi screens

Chromokinesins are microtubule-based motors that ensure proper chromosome segregation.[29] Their functional epistatic interactions are of significant interest, but have not been profiled at scale by imaging. To investigate epistasis, we extended poolVis for dual-sgRNA expression by adding

another U6 promoter (Fig. 4a) and generated six backbones, each containing an hU6-driven spacer targeting both P1/P2 isoforms of *KIF15*, *KIF4A*, and *ATM*, or *KIF22* isoforms individually, or a non-targeting control. We validated each spacer individually (Supplementary Fig. 7). The four minipools in Fig. 3a were inserted for a library of 2,016 double perturbations (Fig. 4b-c). Then, we transduced MCF10A cells, induced Cre, grew cultures, and imaged four fixed plates according to Fig. 3b. cellPool debarcoded 2,159,003 nuclei (median: 751; Fig. 4d; estimated error rate: 1.4%). Technical replicates agreed closely for normalized γH2AX, H3S10P, and α-tubulin, and *H2AX* knockdown robustly reduced γH2AX (Supplementary Fig. 8).

We computed epistasis scores as differences of double- and single-knockdown effects, and a null distribution from bootstrapping non-targeting pairs.[30] The scores skewed negative relative to the null (Supplementary Fig. 9a). Since apoptosis is triggered if DNA damage accumulates beyond a threshold given the wild-type *TP53* in MCF10As, phenotypic states with DNA damage equaling the addition of single-knockdown effects may not be accessible to the perturbed cells. This means that typical methods for fitness-based epistasis analysis may not apply to DNA damage data, and new methods for imaging-based phenotypes are under active development.[2] Deep learning-based methods do not yet outperform simple linear baselines for such analysis.[31] Hence, to evaluate the biological information quality in our data, we called high-confidence epistatic interactions if double-knockdown effects were significantly negative for both guides (FDR<0.2) while single-knockdown effects were positive (known as reciprocal-sign epistasis[32] or “opposite” interactions[31]; Fig. 4e). This is a simple, biologically interpretable criterion that avoids assumptions based on prior fitness-based studies while highlighting reproducible epistatic interactions.

For *KIF4A*, interactions identified with *SMC2* and *SMC4* (condensin I subunits) and *RACGAP1* agree with roles in chromosome-axis organization[33,34] and central-spindle function,[35] respectively. We often observed cytokinesis defects for *RACGAP1* knockdown; however, manual annotation of 4,135 single-cells suggested that multinucleation alone did not explain *KIF4A-RACGAP1* interaction (Supplementary Fig. 9b-c; Additional Data 1). *ATM* interactions with *STAG2* and *CDC20* were also recovered, consistent with checkpoint and repair pathway functions.[36,37] *KIF15* showed the most hits, including *MKI67*, *BUB1*, *CDC20*, and *KIF18B*, supported by public data[38] (Supplementary Table 3). Other predicted *KIF15*-interactors were *RAD21* and *STAG2* (cohesin), *MAPK6* and *ZAK* (MAPK superfamily), *AJUBA* (spindle assembly checkpoint), *TACC1* (microtubule stabilizer), *CDK7* and *CDKN2A* (CDK regulators). *KIF15* and *KIF11* are partially redundant; their simultaneous loss-of-function shows monopolar spindles. One *KIF11* spacer narrowly missed our high-confidence threshold despite negative epistasis score, reflecting partial repression (Supplementary Table 4). No epistasis was identified for *KIF22* isoforms, suggesting mutual compensation. Single-cell galleries revealed extreme damage phenotypes even when mean γH2AX decreased, underscoring the value of unpooled visualization (Fig. 4f).

A fully arrayed experiment validated the *ATM*- and *KIF15*-centered epistasis hits where *H2AX* targeting controls showed strong γH2AX reduction indicating robust CRISPRi performance, and the epistasis scores were again strongly negative relative to the null (Supplementary Fig. 9d-e). A representative set of mitotic nuclei from the arrayed plates is shown in Fig. 4f. Overall, these results establish our platform for scalable imaging-based profiling of double perturbations.

Additional resources for screens with poolVis and cellPool

To facilitate adoption, we generated vectors for wt-Cas9, wt-Cas12a, and dCas9-KRAB engineering (Supplementary Table 1a). Common Cas vectors carry HA tags, conflicting with HA-based barcodes. Users may swap out the HA-tag in barcodes, or use our lentiviral/piggyBac vectors lacking HA or carrying alternative tags for Cas engineering. piggyBac vectors also enable one-step engineering of Cas and ERT2-Cre-ERT2 enzymes.

DISCUSSION

Protein barcodes enable scalable imaging-based profiling of genetic perturbations, yielding spatially resolved information inaccessible by sequencing. Previously, they were limited by low-throughput cloning and poor debarcoding. poolVis and cellPool address these limitations.

poolVis leverages a Cre-inducible design, a defining feature that enables temporal control of barcode and guide expression while maintaining a compact vector design. This feature will be valuable for studies requiring controlled single- or sequential double-perturbations. In addition, poolVis separates spacers and barcodes by ~50 bp during lentiviral packaging, effectively preventing recombination that complicates barcode-based pooled screening.[7,12,13] Our barcodes are read in three imaging iterations. More iterations can profile proteins, DNA, or RNA. The present library contains 84 barcodes, scalable further with more epitopes/backbones, although eventually imaging throughput becomes limiting. While in situ sequencing costs are expected to decline, immunostaining-based protein barcode detection currently reduces the per-cell reagent cost to roughly one-tenth that of RNA barcodes, and enables broader experimental replication given the specialized nature of in situ sequencing workflows.

cellPool automates the analysis of multi-terabyte high-content imaging data. It includes detailed quality control and human-in-the-loop checkpoints to verify the intermediate results. Leveraging Nextflow and Docker supports reproducible analysis on high-performance computing platforms common in imaging cores. The pipeline incorporates robust mosaicking and across-cycle integration methods and a multivariate debarcoding model tailored to the statistical properties of protein barcodes. By presenting unpooled single-cell image galleries with assigned barcodes, cellPool facilitates interpretation and follow-up.

We applied poolVis and cellPool to profile cell-cycle regulators singly and in double perturbations. We profiled nucleus, plasma membrane, microtubule, DNA damage/repair, and mitotic markers, and demonstrated low technical variability. We identified expected phenotypes and additional candidates confirmed by follow-up. Interestingly, most double perturbations lowered γH2AX more than expected, possibly because resulting early stress may enrich for damage-resistant survivors. The epistasis hits included a recently identified synthetic lethal pair (*ATM-STAG2*),[36] functionally coupled gene pairs (*KIF4A-SMC2*, *KIF4A-SMC4*, *KIF15-MKI67*, *KIF15-BUB1*, etc.), as well as novel predictions for *KIF15* interactors. These initial observations warrant further mechanistic investigation to clarify how these interactions function at the molecular level.

The current analysis has certain limitations. We did not distinguish mono- and multi-nucleated cells, the integrated nuclear intensities combined area and mean marker signal, and the epistasis analysis focused on γH2AX. Multiparametric methods considering foci number, texture, etc. could enhance statistical power. To our knowledge, we presented the first large-scale double-CRISPRi imaging data, which may provide a useful benchmark for developing such methods.

In conclusion, by combining a Cre-inducible architecture with high-quality image processing and rigorous statistical modeling of protein barcode data, poolVis and cellPool provide a robust framework for imaging-based profiling of pooled CRISPR perturbations. The platform uniquely

enables double-perturbation analysis with single-cell imaging readouts, offering a scalable and extensible foundation for dissecting genetic networks and their relevance to human disease.

METHODS

Plasmid vector design and assembly

All plasmid backbones were cloned using the Gibson Assembly method (NEBuilder Hifi DNA Assembly Master Mix, NEB E2621). DNA fragments were obtained via PCR amplification or restriction digestion of vectors available in house or sourced from Addgene, or commercial synthesis by Twist Bioscience or Integrated DNA Technologies. The CRISPR guides (with or without barcodes) were inserted into backbone vectors by Golden Gate cloning using the Esp3I restriction enzyme (NEB, R0734) and the NEBridge Ligase Master Mix (NEB, M1100; see Supplementary Table 1a-b for the insert sequences and the vectors). Plasmids were amplified by transforming the NEB Stable Competent *E. coli* (NEB, C3040). The final plasmid sequences were verified by nanopore-based whole plasmid sequencing from Plasmidsaurus or Quintara Biosciences.

Cell engineering and tissue culture

All cell lines were sourced from the McManus lab cryovial database or The Cell and Genome Engineering Core at UCSF. The cells were cultured at 37°C and 5% CO_2 in tissue culture incubators and routinely tested for mycoplasma contamination.

The K562 cells were cultured in advanced RPMI 1640 (Gibco, 12633012) supplemented with 2 mM L-glutamine, 10% FBS, 100 µg/ml streptomycin, and 100 units/ml penicillin. For testing the Cas12a focused vectors, K562 cells, originally derived from a clonal line expressing dCas9-SunTag and a scFV-sfGFP-VP64 fusion (Gilbert et al., 2014), were engineered with enAsCas12a using the pRDA_174 plasmid (Addgene, 136476) and a clone was selected for high Cas12a knockout activity (engineered line borrowed from the McManus lab cryovial database).

The U2OS cells were cultured in the McCoy's 5A (Modified) medium (Gibco, 16600082) supplemented with 10% FBS, 100 µg/ml streptomycin, and 100 units/ml penicillin.

The Hek293T cells were cultured in the Dulbecco's Modified Eagle Medium (Gibco, 11965092) supplemented with 10% FBS, 100 µg/ml streptomycin, and 100 units/ml penicillin. For testing the Cas9 and dCas9-KRAB focused vectors, Hek293T cells were engineered with the AU5-tagged Cas protein and ERT2-Cre-ERT2 expressing constructs using the piggyBac transposon system (System Biosciences, PB210PA-1; see Supplementary Table 1a for the vectors) and the jetPRIME transfection reagent (Sartorius, 101000001). The engineered cells were selected with 1000 µg/ml hygromycin B (Gibco, 10687010).

The MCF10A cells were cultured in MEGM Mammary Epithelial Cell Growth Medium (Lonza, CC-3150) supplemented with 100 ng/ml cholera toxin. They were engineered with the AU5-tagged dCas9-KRAB, hygromycin B phosphotransferase, and ERT2-Cre-ERT2 expressing cassettes using the piggyBac transposon system and electroporation using the Lonza SE 4D Nucleofector Kit (V4XC-1032), program DS-138. They were selected with 1000 µg/ml hygromycin B (Gibco, 10687010).

Lentivirus production, transduction, and Cre induction

Lentiviral particles were produced in the Hek293T cells by co-transfecting the pMD2.G, psPAX2, and the transfer plasmids using the jetPRIME transfection reagent (Sartorius, 101000001).

Supernatant was collected between 60-72 hours after transfection, cells were removed by centrifugation, and cleared supernatant was frozen at -80°C before use. All transductions of CRISPR guide RNA carrying particles were performed at MOI < 0.1. Cre recombinase was delivered to the K562 and the U2OS cells via lentiviral transduction at MOI=1. The ERT2-Cre-ERT2 engineered Hek293T and the MCF10A cells were treated with media containing 1 μM 4-hydroxytamoxifen (Sigma, SML1666) to induce Cre recombination.

Design of protein barcode coding sequences

A heuristic algorithm was developed to design 84 barcode coding sequences with minimal homology. Each epitope tag's amino acid sequence was chopped into short overlapping motifs of 5-6 amino acids and the *k*-mers of lengths 8-16 bp in the coding sequences for each motif were enumerated. The complete coding sequences for each epitope tag were generated by iteratively sampling and annealing the motif-wise coding sequences while avoiding picking a coding sequence whose *k*-mer set intersects with previously sampled coding sequences. Next, for every barcode, the RNA structure stability was assessed using RNAfold for all combinations of the coding sequences of the corresponding epitope tags. Finally, the coding sequences for all barcodes were picked such that each occurrence of an epitope tag in the library is coded by a unique DNA sequence and the overall stability of the RNA structures is maximized.

Flow cytometry

All flow cytometry data was collected using the Invitrogen™ Attune™ NxT Acoustic Focusing Cytometer. Cre recombination was induced 24 hours after transducing cells with lentivirus carrying the sgRNAs. After population-wide green fluorescence stabilized indicating effective Cre recombination, the cells were treated with media containing 2.5 μg/ml Puromycin for 60-72 hours to remove unrecombined cells. Then, flow cytometry data was collected after 21 days. Cells were detached by trypsinization (if adherent) or collected from suspension, pelleted and washed with FACS buffer (STEMCELL Technologies, 20144) twice before incubating with APC-conjugated anti-CD81 antibody (BioLegend, 349509) diluted at 1:100 in FACS buffer on ice for 20 minutes. The antibody containing solution was removed by centrifugation and the cells were resuspended in FACS buffer for flow cytometry. The resulting data was analyzed by FlowJo (10.10.0).

Pooled library design and cloning

The targeted set of genes for the single and double perturbation screens in the MCF10A cells was curated based on expert feedback. DepMap RNAi data[39] for the MCF10A cells was used to ensure that the majority of the targeted genes were not essential genes (Supplementary Fig. 5e). The gene targets were organized in four sets randomly. Then, four oligonucleotide pools were ordered from Twist Bioscience (see Supplementary Table 1g for the pooled sequences). Each pool consisted of 84 barcoded CRISPR guides. For every gene in the screen, top two CRISPR guide sequences were obtained from the hCRISPRi-v2.1 library.[40] The oligonucleotide pools were amplified using the KAPA HiFi PCR Kit (Roche, 07958935001) for 16 cycles starting from 2 ng single-stranded DNA templates per 50 μl reaction. The reaction products were purified using SPRI beads (Omega Bio-tek, M1378-00), and further by 4% TBE agarose gel electrophoresis. The double-stranded oligonucleotide pools were inserted into backbone vectors by Golden Gate cloning of mixtures containing vector to insert in ratio of 1:2 using the Esp3I restriction enzyme (NEB, R0734) and the NEBridge Ligase Master Mix (NEB, M1100). After initial incubation at 37°C

for 90 minutes, the reaction mixtures went through 9 cycles of 20 mins at 16°C followed by 20 mins at 37°C. Finally, the reaction was heat-inactivated at 60°C for 5 mins. Plasmid libraries were amplified by transforming the NEB Stable Competent *E. coli* (NEB, C3040) and growing LB liquid cultures containing 100 µg/ml ampicillin at 30°C for 24 hrs. Coverage was estimated to be >500x per vector construct by counting colonies from plating transformed bacteria on LB agar plates. Libraries were confirmed to be of good quality by nanopore-based whole plasmid sequencing and sequencing of individual colonies picked from the plated bacteria.

Single perturbation and double perturbation screens

The MCF10A cells engineered to express dCas9-KRAB, hygromycin B phosphotransferase, and ERT2-Cre-ERT2 were used in the single and double-perturbation screens in Figs. 3-4. The cells were maintained in hygromycin-free media throughout the screen (Lonza, CC-3150 with 100 ng/ml cholera toxin). Before starting the screen, the cells were maintained in hygromycin-free media for two passages. Then, the cells were passaged in twenty-eight 10 cm dishes at ~2 million cells per dish (one dish per minipool). After allowing cells to adhere for 24 hours, lentiviral minipools were added to each dish at MOI < 0.1. After another 24 hours, the media in each 10 cm dish was replaced with media containing 1 µM 4-hydroxytamoxifen to trigger nuclear translocation of ERT2-Cre-ERT2. The screening time was counted starting 36 hours after adding 4-hydroxytamoxifen. This was based on the observation of dimly green fluorescent cells in the minipool cultures, indicating successful Cre induction. This timepoint was designated t=0 days.

At t=0 days, the media was changed to media without 4-hydroxytamoxifen. At t=2.5 days, the media was changed to media containing 2.5 µg/ml puromycin. Next, at t=5.5 days, the media was changed to media without puromycin. Majority of the untransduced cells died after puromycin treatment and all cultures were observed to be sparsely confluent. Then, all cultures were maintained with passaging and media changes in 10 cm dishes until, at t=16 days, they were passaged in fibronectin-coated 96-well plates (Sigma, S5171; Revvity, 6055302) at ~10,000 cells/well. At t=18 days, cells were fixed with 4% PFA (Electron Microscopy Sciences, 15714). Then, they were processed for imaging.

Multiplexed imaging

All imaging was done using the Opera Phenix Plus high-content screening system and the Harmony 4.9 software (Revvity) using a 20X 1 NA water objective and 10% overlap between neighboring image tiles. For every tile, the different channels were imaged sequentially starting from the red fluorescent side of the spectrum to blue emission. The Hek293T images were taken in the P06-1.5H-N 6-well plates from CellVis and the U2OS and MCF10A cells were imaged after plating in the Revvity 96-well PhenoPlate (Part #6055302). In all cases, the plates were coated with fibronectin (Sigma, S5171) before seeding cells at 300,000 cells per well for 6-well plate and 10,000 cells per well in the 96-well plates. The cells were allowed to grow for 48-60 hours before fixation with 4% PFA (Electron Microscopy Sciences, 15714).

The cyclic immunofluorescence protocol was followed as described before.[41] After fixation, cells were stained with Hoechst 33342 at 1:2000 and wheat-germ agglutinin (WGA) at 2.5 µg/ml in PBS1X at room temperature for 10 minutes. Then, first imaging cycle was done after three washes with PBS1X for 5 minutes per wash. Next, the plates were treated with 140 µl/well of acidic inactivation solution containing 3% H_2O_2 and 20 mM HCl in PBS1X for 6 hours. The plates were

washed with 200 µl/well of PBS1X thrice and treated with 140 µl/well of basic inactivation solution containing 3% H2O2 and 20 mM NaOH in PBS1X for 1.5 hours. After washing with PBS1X thrice, the inactivation of GFP and Alexa Fluor dye conjugated to WGA was confirmed by imaging. Next, cells were permeabilized with ice-cold methanol, washed with PBS1X thrice, and incubated with Intercept (PBS) blocking buffer (LICORbio, 927-70001) at room temperature for 1 hour. For every subsequent imaging cycle, cells were incubated with fluorophore-conjugated antibody mixtures overnight at 4°C, imaged using the Opera Phenix system, and the fluorophores were inactivated using the basic solution. Three PBS1X washes were done between every change of staining or fluorophore inactivation solution.

Preprocessing of imaging data

The images of Hek293T cells in Fig. 1 were processed by ImageJ (version 2.3.0/1.5.3f) to enhance contrast and sharpen. All other images were processed by cellPool. First, Z-slices, if available, were combined by maximum projection. Then, vignetting correction was performed using the BaSiC tool[42] in MATLAB (R2024a) under the assumption that the vignetting pattern is identical for a tile across wells. The corrected images were stitched to create a single mosaic image per well per channel per imaging cycle. For this, an adjacency graph was created with the image tiles as nodes and edges connecting overlapping tiles based on metadata. The exact magnitude of shift was estimated by phase cross-correlation of Laplace-of-Gaussian transformed images with an upsampling factor of 10.[19] Next, every tile is positioned relative to the well center by taking a sample of paths along the graph and adding the estimated shifts along the edges. The median of all paths is used to position every tile. This approach was applied to another channel whenever one channel failed to give a reliable positional estimate and until all channels from the imaging cycle have been used. If all channels failed to give reliable positional estimate for a tile, a linear model of all confidently placed image tiles was used to predict its position. These are pasted together to create a mosaic image for each channel in each cycle. After ensuring high quality from summary quality control reports, nuclei were segmented using a custom-trained Cellpose2 model (version 2.2.3).[43] The features of interest were quantified using scikit-image.

Integrating features across imaging cycles

The nuclei were tracked across cycles in three phases using the Kuhn-Munkres algorithm in each phase. First, a low-to-moderate quality tracking is performed by using orientation of nuclei as the telling feature to set up the cost function. Next, under the assumption that neighboring nuclei have similar shift from one cycle to the next, the low confidence tracking mates are identified and returned to queue for tracking again in the next phase. Finally, cells are tracked expanding out from high confidence tracking mates. We found that the transformations not accounted for by phase correlation do not contribute significantly to distorting distances with respect to local references at order of 100s of pixels but become cumulatively significant relative to nuclear size over whole well mosaics. This allowed setting up a cost function based on a pair of high confidence tracking mates to find more high confidence tracking mates in their neighborhoods using Euclidean distances after centering on the starting pair. This process continued until all existing and new set of high confidence tracking mates were used to track their neighbors. The quality of tracking is summarized as feature comparisons of the tracking mates, and unpooled images showing random samples of tracking mates per well for human-in-the-loop verification.

Benchmarking comparison of mosaicking

For benchmarking comparison, ASHLAR was used to stitch and register the images from the barcoded U2OS samples. The mosaicked Hoechst 33342 channel from the first imaging cycle were segmented using the same custom-trained Cellpose2 model (version 2.2.3) as used for cellPool. The segmentation masks were applied to all imaging channels under the assumption of perfect pixel-to-pixel registration. Then, the barcode profiles were quantified using scikit-image in Python as before. The cycle 1 nuclear masks from cellPool and ASHLAR analyses were mapped to each other using the cellPool tracking algorithm. No mapping errors were found in 30 random samples of tracking mates from each well. Fig. 2d is a comparison of the barcode staining intensities for the set of nuclei that were tracked by cellPool across all cycles, i.e., the nuclei that washed off in later cycles were excluded. The mosaicked images from ASHLAR were unpooled by cellPool to assess the mosaicking quality.

Mixture model for debarcoding protein barcodes

The median intensities taken over a nuclear mask give the barcode profile denoted $Y_i = \{Y_{p,i}, Y_{e_1,i}, Y_{e_2,i}, \dots, Y_{e_n,i}\}$ for the corresponding cell i, where $p, e_1, \dots, e_n$ are the fluorescent barcode backbone and the n epitopes in the library, respectively. Y represents the binary logarithm of median intensity. Let $B_i = (B_{p,i}, B_{e_1,i}, B_{e_2,i}, \dots, B_{e_n,i})$ denote a barcode, where $B_{e,i}$ is a state variable. $B_{e,i} = 1$ when cell i is positive for the epitope tag e and 0 if it is negative. cellPool takes the barcode profiles as input and outputs the barcodes in two stages with the first being optional.

Population-level thresholding. This step deciphers the barcodes completely for some cells and partially for the rest by analyzing the intensity distribution for each epitope across the population. By default, we apply Li's iterative method based on minimum cross-entropy[44] to determine a threshold, t_p for GFP signal. We classify $B_{p,i}$ as 1 if $Y_{p,i}$ is greater than t_p. By default, if $Y_{p,i}$ is less than $t_p - 1$, we classify $B_i = (0, 0, \dots, 0)$. The default settings can be changed in the cellPool configuration profile to values determined by manually studying the histograms generated by cellPool. The thresholding method can be set to any method from the *filters* submodule of scikit-image. We assume that enough GFP-negative cells (order of 1000s) are available collectively on the plate. If not, cellPool skips to multivariate modeling directly.

For the GFP-positive cells, we determine high-confidence thresholds for each epitope by comparing the epitope intensity distribution in the GFP-positive cells with that in the GFP-negative cells (Fig. 2h). Say, $\hat{f}_{e,b}(y)$ is the empirical density function of the background noise for epitope tag e as quantified using the GFP-negative cells. Let w_e be the proportion of GFP-positive cells that are also positive for the epitope tag e, and $f_{e,s}(y)$ be the probability density function for the log-transformed staining intensity of e in these cells. We determine the high-confidence threshold t_e, such that for $y > t_e$, $f_{e,s}(y)/\hat{f}_{e,b}(y) \gg 1$. This happens when $\{w_e f_{e,s}(y) + (1 - w_e)\hat{f}_{e,b}(y)\}/\hat{f}_{e,b}(y) \gg 1$. Here, $\{w_e f_{e,s}(y) + (1 - w_e)\hat{f}_{e,b}(y)\}$ is the signal-background mixture in the GFP-positive cells. Hence, cellPool estimates $\{w_e f_{e,s}(y) + (1 - w_e)\hat{f}_{e,b}(y)\}/\hat{f}_{e,b}(y)$ for each epitope tag and positions the threshold such that the ratio is consistently greater than a user-specified level.

Multivariate mixture model-based debarcoding. To account for distributions with non-elliptical isodensity loci, we partition the GFP-positive cells into 5-10 sets that have relatively uniform GFP

intensity and thereby, low variation in epitope intensities as well. This permits a good fit by a mixture model based on skew elliptical distributions to the data for each set. We model the single cell intensity vectors as a mixture distribution with as many components as there are barcodes. Let $\boldsymbol{e_k}$ be an $n \times 1$ one-hot vector representing the epitope e_k, m be the number of epitopes per barcode, $\boldsymbol{E_B} = [\boldsymbol{e_{k_1}}, \dots, \boldsymbol{e_{k_m}}]$ be an $n \times m$ matrix with one-hot vectors for epitope tags in the barcode B as its columns, and $\mathbb{E}'_B$ be the set of absent epitopes, i.e., $\mathbb{E}'_B = \{k | B_{e_k} = 0\}$. Then, for barcode B and $\boldsymbol{y} = [y_{e_1,c}, \dots, y_{e_n,c}]^T$, we model the observed distribution as

$$f_B(\boldsymbol{y}) = \varphi_m(\boldsymbol{E_B^T y}; \boldsymbol{\zeta_B}, \boldsymbol{\Omega_B}, \boldsymbol{\alpha_B}) \prod_{k \in \mathbb{E}'_B} \hat{f}_{e_k,b}(\boldsymbol{e_k^T y}), \quad \boldsymbol{y} \in \mathbb{R}^n.$$

Here, φ_m is the m-dimensional multivariate skew normal distribution with $\boldsymbol{\zeta_B} \in \mathbb{R}^m$, $\boldsymbol{\Omega_B} \in \mathbb{R}^{m \times m}$, and $\boldsymbol{\alpha_B} \in \mathbb{R}^m$ as the location, scale and skewness parameters, respectively. This equation describes a multivariate skew normal distribution in the subspace of epitope tags present in barcode B and empirically estimated background distribution in the subspace of the remaining epitope tags. cellPool fits a mixture model with components defined by the above equation (Supplementary Fig. 3). It uses the expectation-maximization algorithm with the option to test multiple parameter initializations in parallel and quality control the convergence of the iterations of optimization to maximize the data likelihood,

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{B=1}^{B_{n,m}} \{p_B f_B(\boldsymbol{Y_i})\}^{z_{B,i} 1_{B \in B_{i,t}}}.$$

Here, N is the total number of nuclei, $B_{n,m}$ is the total number of barcodes, p_B is the mixing fraction of barcode B, $z_{B,i}$ is an indicator variable that is 1 if i has barcode B and is 0 otherwise, and $1_{B \in B_{i,t}}$ is an indicator variable that equals 1 if B is compatible with the epitope set $B_{i,t}$ present in i based on population-level thresholding and 0 otherwise.

cellPool initializes the model parameters as follows. The skewness parameters are set to random values, off-diagonal terms in the scale parameters are set to 0, the location parameters are set to the foreground mean based on Otsu's thresholding, and the mixing proportions are set as equal for all barcodes. Then, cellPool iteratively computes the conditional expectations for the latent variables (E-step) and updates the parameters for the multivariate skew normal components (M-step) following the equations described before,[20] except that the conditional expectations for $z_{B,i}$ and likelihood calculations use our model for protein barcode data (see $f_B(\boldsymbol{y})$ and $\mathcal{L}$ above). The EM algorithm terminates upon convergence of log-likelihood and outputs the best parameter estimates, which yield the posterior probabilities for all possible barcode assignments. cellPool assigns a barcode if the corresponding posterior probability meets the user-defined threshold.

Benchmarking comparison with other debarcoding methods

cellPool's debarcoding model was benchmarked by comparison with the Single-Cell Deconvolution (SCD)[21] and the EPICode[6] methods using custom scripts. For SCD, each epitope tag's raw signal was normalized with respect to the 99th percentile of its distribution encompassing all cells under consideration. Then, the top 3 epitope tags present in a cell were assigned as its barcode if the 3rd and 4th highest ranked epitope tags met the *barcode separation* criteria, i.e., their difference was greater than a user-defined threshold. For EPICode, each epitope tag's raw signal was converted into a percentile score based on the background distribution. Then, for each barcode, a score was calculated as the mean percentile score of its constituent epitope tags. The barcode with the highest score was assigned if the difference from the next highest score meets the barcode separation criteria. For cellPool, the difference in the posterior probabilities assigned to the top two barcodes was used to quantify confidence. The barcode separation criteria was varied from 0 to 1 to generate the curves in Fig. 2k.

Validation of guides for double perturbation screen

Top ranked guides for *ATM*, *KIF15*, *KIF22*, and *KIF4A* were validated in the MCF10A cells engineered with dCas9-KRAB and ERT2-Cre-ERT2. The targeting guides were inserted under the hU6 promoter and the nontargeting under the mU6 promoter in the same poolVis vector. After 18 days of screen, the cells were fixed in 96-well plates as described before and imaged in two cycles. In the first cycle, Hoechst 33342 and GFP signals were captured. Then, GFP was denatured using protocol described for cyclic immunofluorescence above. Then, the samples were permeabilized using ice-cold methanol and incubated with Intercept (PBS) blocking buffer (LICORbio, 927-70001) at room temperature for 1 hour. Then, the samples were incubated with primary antibodies for 24 hours at 4°C, followed by another round of incubation with secondary antibodies for 24 hours at 4°C. The samples were washed, stained with Hoechst 33342 and imaged again.

ATM, *KIF22*, and *KIF4A* are nuclear localized during interphase while *KIF15* is cytosolic. Additionally, *KIF22* and *KIF4A* are expressed primarily during the late G2/M phase. *ATM* signal distribution was quantified by the median nuclear staining intensities of its antibody. *KIF22* and *KIF4A* signal distribution was quantified as median nuclear staining intensities after thresholding for late G2/M phase nuclei using the Hoechst 33342 signal. *KIF15* signal distribution was quantified as median cytosolic staining intensities after obtaining the cytosolic mask as a strip of width 3 pixels (1.8 μm) encircling the nuclear mask.

Phenotypic feature extraction and differential analysis

Phenotypic features were extracted using scikit-image in Python. The nuclear area was computed for objects identified by Cellpose2 segmentation of images. The integrated nuclear γH2AX and H3S10P signal was obtained as the sum of fluorescent intensities over nuclear masks. Mean cytosolic α-tubulin signal was obtained as the mean fluorescence intensity in the corresponding image channel and the cytosolic mask represented by a strip of width 3 pixels (1.8 μm) encircling the nuclear mask. For single perturbation screen, log10-transformed intensities were used in differential analysis. For double perturbations, the log10-transformed intensities were further centered and scaled with respect to the median and median absolute deviation of the untransduced GFP-negative cells for every column of every imaged plate. Differential analysis was performed using the Student's t-test to compare normalized feature for a target group vs the

nontargeting control infected cells. The p-values were corrected for multiple testing by the Benjamini-Hochberg procedure. The effect of nontargeting controls was centered at 0 and standard errors were used to scale the mean effect sizes of perturbations.

For mitotic index analysis, the mitotic nuclei were identified by manually thresholding integrated nuclear γH2AX. Then, the mitotic index was measured for each targeting sgRNA as the proportion of mitotic nuclei relative to all nuclei identified with that target. The mitotic indices for well-covered guides (coverage > 1500) were compared with the mitotic index of the combined non-targeting and untransduced control population by two-sided exact binomial test. The p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. The hits were identified as the gene targets with FDR <0.1 and same sign of effect for both guides.

Genetic interaction scores were quantified as difference of double knockdown effect and the sum of single knockdown effects.[30] Interaction hits for DNA damage and repair were identified as those gene pairs where the estimates for single knockdown effects were positive for both guides targeting the Cre inducible gene perturbation and the double knockdown effect was negative with FDR-adjusted p-values <0.2. The representative images in the figures were unpooled by cellPool.

DATA AVAILABILITY STATEMENT

Data and reagent availability

The datasets generated during this study will be deposited in AWS and made publicly available upon acceptance. Plasmid constructs will be deposited on Addgene and made publicly available upon acceptance. Any additional materials or information related to the study can be obtained by reaching out to the corresponding authors with a reasonable request.

Code availability

All code used in this article is available at:: https://github.com/mcmanuslab/cellpool and https://doi.org/10.5281/zenodo.17577408.

ACKNOWLEDGEMENTS

We thank members of the McManus lab for helpful discussions, technical support, and feedback throughout this project. This work was assisted by the UCSF Parnassus Center for Advanced Technologies. A special thanks to Iain Cheeseman, Monica Bettencourt-Dias, Mariana Lince-Faria, Lee Rao, Shalin Mehta, Talon Chandler, Eduardo Hirata-Miyasaki, Soorya Pradeep for many helpful discussions. This study was supported, in part, by National Institutes of Health grants U01CA272546 and U24DK116214 to M.T.M., as well as a Sandler Program for Breakthrough Biomedical Research Award to M.T.M. Additional support was provided by the Chan Zuckerberg Biohub, Pivotal Life Sciences, and the Diabetes Center at UCSF to M.T.M.

AUTHOR CONTRIBUTIONS

K.C. and M.T.M conceived the project. K.C. and M.T.M designed the wet-lab method prototype. K.C. developed the wet-lab methods with critical input from M.T.M. K.C. developed the computational methods and the software. K.C. performed the experiments, assay optimizations, and analyzed data with feedback from M.T.M. K.C. drafted the original manuscript. K.C. and M.T.M. edited the manuscript. Both authors reviewed and approved the final manuscript.

COMPETING INTERESTS STATEMENT

We have filed a provisional patent application related to this work.

REFERENCES

1. Boyd, J., Fennell, M. & Carpenter, A. Harnessing the power of microscopy images to accelerate drug discovery: what are the possibilities? *Expert Opin Drug Discov* **15**, 639–642 (2020).

2. Walton, R. T., Singh, A. & Blainey, P. C. Pooled genetic screens with image-based profiling. *Mol Syst Biol* **18**, (2022).

3. Kudo, T. *et al.* Multiplexed, image-based pooled screens in primary cells and tissues with PerturbView. *Nat Biotechnol* **43**, 1091–1100 (2025).

4. Fandrey, C. I. *et al.* NIS-Seq enables cell-type-agnostic optical perturbation screening. *Nat Biotechnol* **43**, 1337–1347 (2025).

5. Wroblewska, A. *et al.* Protein Barcodes Enable High-Dimensional Single-Cell CRISPR Screens. *Cell* **175**, 1141-1155.e16 (2018).

6. Kudo, T., Lane, K. & Covert, M. W. A multiplexed epitope barcoding strategy that enables dynamic cellular phenotypic screens. *Cell Syst* **13**, 376-387.e8 (2022).

7. Xie, S., Cooley, A., Armendariz, D., Zhou, P. & Hon, G. C. Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS One* **13**, e0198635 (2018).

8. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e21 (2016).

9. Hanna, R. E. & Doench, J. G. A case of mistaken identity. *Nat Biotechnol* **36**, 802–804 (2018).

10. Hegde, M., Strand, C., Hanna, R. E. & Doench, J. G. Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. *PLoS One* **13**, e0197547 (2018).

11. Dhainaut, M. *et al.* Spatial CRISPR genomics identifies regulators of the tumor microenvironment. *Cell* **185**, 1223-1239.e20 (2022).

12. Replogle, J. M. *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol* **38**, 954–961 (2020).

13. Hill, A. J. *et al.* On the design of CRISPR-based single-cell molecular screens. *Nat Methods* **15**, 271–274 (2018).

14. DeWeirdt, P. C. *et al.* Optimization of AsCas12a for combinatorial genetic screens in human cells. *Nat Biotechnol* **39**, 94–104 (2021).

15. SHEARWIN, K., CALLEN, B. & EGAN, J. Transcriptional interference – a crash course. *Trends in Genetics* **21**, 339–345 (2005).

16. Campa, C. C., Weisbach, N. R., Santinha, A. J., Incarnato, D. & Platt, R. J. Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. *Nat Methods* **16**, 887–893 (2019).

17. Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proceedings of the National Academy of Sciences* **101**, 10380–10385 (2004).

18. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).

19. Muhlich, J. L. *et al.* Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613–4621 (2022).

20. Pyne, S. *et al.* Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences* **106**, 8519–8524 (2009).

21. Zunder, E. R. *et al.* Palladium-based mass tag cell barcoding with a doublet-filtering scheme and single-cell deconvolution algorithm. *Nat Protoc* **10**, 316–333 (2015).

22. Chavez, K. J., Garimella, S. V. & Lipkowitz, S. Triple negative breast cancer cell lines: One tool in the search for better treatment of triple negative breast cancer. *Breast Dis* **32**, 35–48 (2011).

23. Funk, L. *et al.* The phenotypic landscape of essential human genes. *Cell* **185**, 4634-4653.e22 (2022).

24. McDiarmid, T. A. *et al.* Diversified, miniaturized and ancestral parts for mammalian genome engineering and molecular recording. Preprint at https://doi.org/10.1101/2024.09.30.615957 (2024).

25. Fox, D. T. & Duronio, R. J. Endoreplication and polyploidy: insights into development and disease. *Development* **140**, 3–12 (2013).

26. Schibler, A. C., Jevtic, P., Pegoraro, G., Levy, D. L. & Misteli, T. Identification of epigenetic modulators as determinants of nuclear size and shape. *Elife* **12**, (2023).

27. Mistry, H. *et al.* UHRF1 is a genome caretaker that facilitates the DNA damage response to γ-irradiation. *Genome Integr* **1**, (2010).

28. Poruchynsky, M. S. *et al.* Microtubule-targeting agents augment the toxicity of DNA-damaging agents by disrupting intracellular trafficking of DNA repair proteins. *Proceedings of the National Academy of Sciences* **112**, 1571–1576 (2015).

29. Almeida, A. C. & Maiato, H. Chromokinesins. *Current Biology* **28**, R1131–R1135 (2018).

30. Axelsson, E. *et al.* Extracting quantitative genetic interaction phenotypes from matrix combinatorial RNAi. *BMC Bioinformatics* **12**, (2011).

31. Ahlmann-Eltze, C., Huber, W. & Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nat Methods* **22**, 1657–1661 (2025).

32. Poelwijk, F. J., Tănase-Nicola, S., Kiviet, D. J. & Tans, S. J. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* **272**, 141–144 (2011).

33. Takahashi, M., Wakai, T. & Hirota, T. Condensin I-mediated mitotic chromosome assembly requires association with chromokinesin KIF4A. *Genes Dev* **30**, 1931–1936 (2016).

34. Mazumdar, M., Sundareshan, S. & Misteli, T. Human chromokinesin KIF4A functions in chromosome condensation and segregation. *J Cell Biol* **166**, 613–620 (2004).

35. Capalbo, L. *et al.* The midbody interactome reveals unexpected roles for PP1 phosphatases in cytokinesis. *Nat Commun* **10**, 4513 (2019).

36. Zhou, J. *et al.* STAG2 Regulates Homologous Recombination Repair and Sensitivity to ATM Inhibition. *Advanced Science* **10**, (2023).

37. Eliezer, Y., Argaman, L., Kornowski, M., Roniger, M. & Goldberg, M. Interplay between the DNA Damage Proteins MDC1 and ATM in the Regulation of the Spindle Assembly Checkpoint. *Journal of Biological Chemistry* **289**, 8182–8193 (2014).

38. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023).

39. McFarland, J. M. *et al.* Improved estimation of cancer dependencies from large-scale RNAi screens using model-based normalization and data integration. *Nat Commun* **9**, 4610 (2018).

40. Horlbeck, M. A. *et al.* Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife* **5**, (2016).

41. Lin, J., Fallahi-Sichani, M., Chen, J. & Sorger, P. K. Cyclic Immunofluorescence (CycIF), A Highly Multiplexed Method for Single-cell Imaging. *Curr Protoc Chem Biol* **8**, 251–264 (2016).

42. Peng, T. *et al.* A BaSiC tool for background and shading correction of optical microscopy images. *Nat Commun* **8**, (2017).

43. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat Methods* **19**, 1634–1641 (2022).

44. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

FIGURE LEGENDS

Figure 1 | poolVis is a Cre inducible CRISPR platform that minimizes the randomization of protein barcode-guide RNA pairs due to recombination in pooled lentiviral libraries. a, Overview of the poolVis workflow. **b, e** Lentiviral vector designs for Cas12a- and Cas9-based screens. **c, f,** GFP signal from the design positioned in the same row to the left, scaled and centered with respect to the background green fluorescence in untransduced control cells (gray). **d, g,** *CD81* signal from the design positioned in the same row to the left, scaled and centered with respect to the non-targeting control (gray). **h,** Imaging of cells with *AURKB* knockdown shows enlarged nuclei (examples highlighted with white arrows). **i,** Distribution of pairwise homology lengths in the set of coding sequences designed for each epitope tag in the barcode library. Short homology tracts indicate minimal shared nucleotide identity between barcodes, confirming effective homology reduction by codon optimization.

Figure 2 | cellPool enables scalable computational analysis to reverse the pooling of CRISPR perturbations. a, Overview of the cellPool workflow. **b,** Plate map showing the presence or absence of each barcode element (GFP and epitope tags) in 19 barcoded U2OS samples for benchmarking cellPool (one barcode per well) and untransduced U2OS cells. **c,** cellPool debarcodes the samples with high accuracy. The colormaps are showing percentages of cells positive for each barcode element in a well. **d,** The signal-background mixture distributions estimated by ASHLAR (gray) and cellPool (black). **e-f,** cellPool tracks cells from later imaging cycles to find their “mate” cell in the previous cycles. The tracking is initialized with image coordinates after registration of image mosaics, e.g., the mate of a nucleus at position labeled with red X in cycle 2 is searched for in the cycle1 image starting from the same position with respect to the top left corner. The number of nuclei that are candidates for matching is user-defined (labeled with red numbers in cycle1 image). The nuclei with high-confidence tracking mates (example tagged with red triangle) available increase in number with every iteration of tracking. **e-g,** Clonal expansion of adherent cells leads to neighboring cells being more likely to have similar area, but orientation shows higher randomness. **h,** cellPool initializes debarcoding with univariate analysis of each epitope tag by comparing the signal-background mixture distribution with the background only distribution. **i,** The immunostaining signal for an epitope tag is dependent on other epitope tags included in the protein barcode. Example, the FLAG tag signal is much weaker in one of the shown barcodes. **j,** cellPool models the barcode profiling data as derived from a mixture of multivariate GFP-dependent skew-elliptical distributions. **k,** cellPool outperforms other existing methods for debarcoding the protein barcodes.

Figure 3 | poolVis and cellPool enable screen of cell-cycle regulating genes with multiplexed imaging readout. a, Gene ontology terms associated with the targeted genes and breakdown of the minipools. **b,** Multiplexed imaging panel showing pooled perturbations. The barcode profiles are shown as squares with nine tiles with each tile showing the median fluorescence intensity per nucleus for the epitope tags. **c,** cellPool outputs the identities of genetic targets per nucleus. **d-f,** Representative phenotypic effects of CRISPR knockdowns unpooled by cellPool. **g,** Volcano plot showing the perturbations that impact the nuclear area. **h,** Representative images of nuclei with perturbations that increase or decrease the nuclear area. **i,** Scatter plot comparing the shift in H3S10P and γH2AX signal across genetic targets. **j,**

Representative images of nuclei with perturbations that increase or decrease the nuclear γH2AX signal.

Figure 4 | poolVis and cellPool reveal epistatic interactions underlying DNA damage and repair in the MCF10A cells. a-b, Cloning design for a lentiviral vector backbone carrying a single CRISPR guide sequence and insertion of pooled guides at the poolVis library entry site. **c,** Breakdown of the 2,016 combinatorial perturbations in the current screen. **d,** Coverage distribution of the guide RNA combinations. Red dotted line is the median. **e,** Volcano plots with annotations highlighting epistatic interaction hits (gene target labels at the base of the arrows). Each arrow connects the point representing the gene target when inserted in the backbone with nontargeting guide1 (arrow base) to the point representing the same gene target in the backbone with a targeting guide1 (arrow head). For *ATM* and *KIF4A*, both guide2 spacers targeting the same interactor are shown. For *KIF15*, single guide2 spacer is shown for each interactor. **f**, Representative images of DNA damage patterns and mitotic nuclei for the epistatic interaction hits.

FIGURES

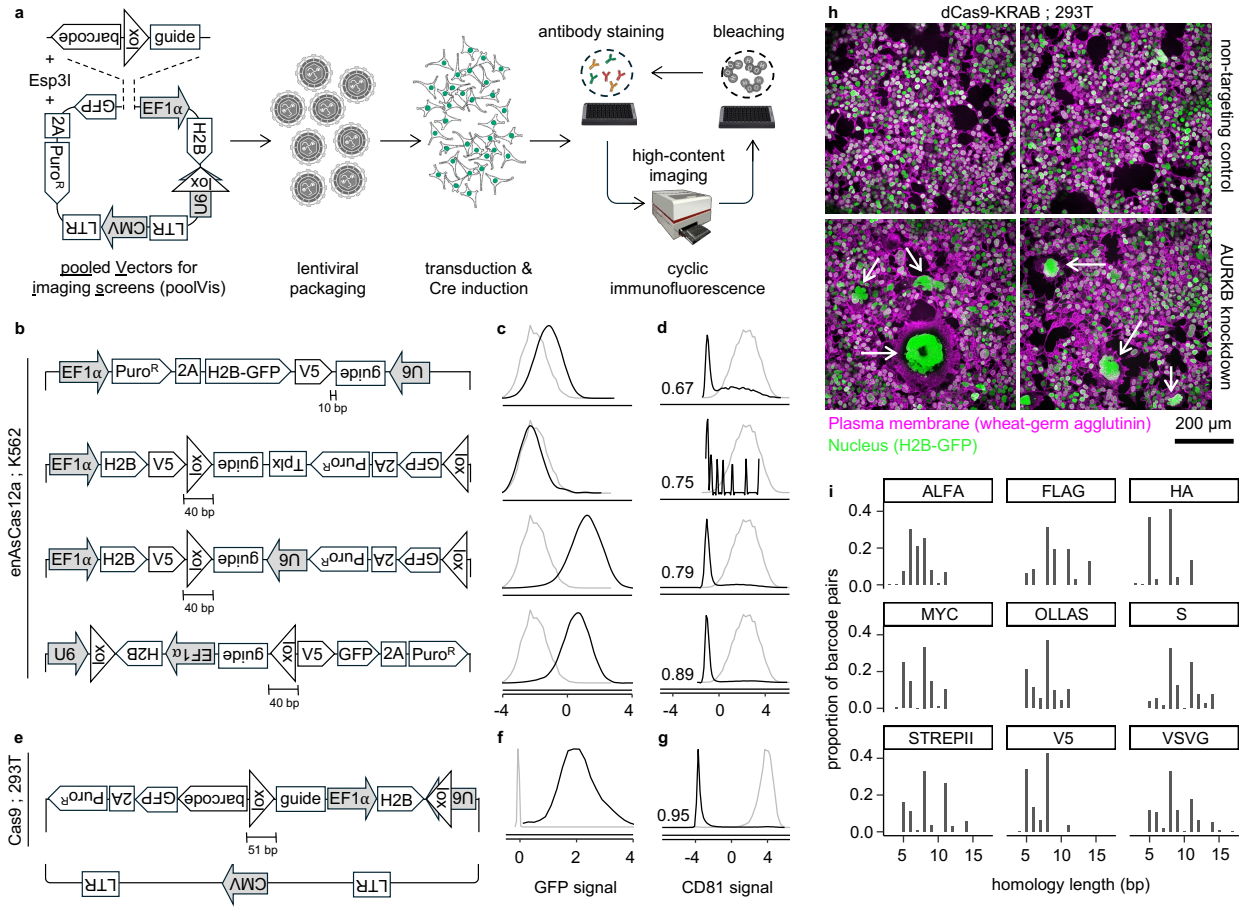

Figure 1

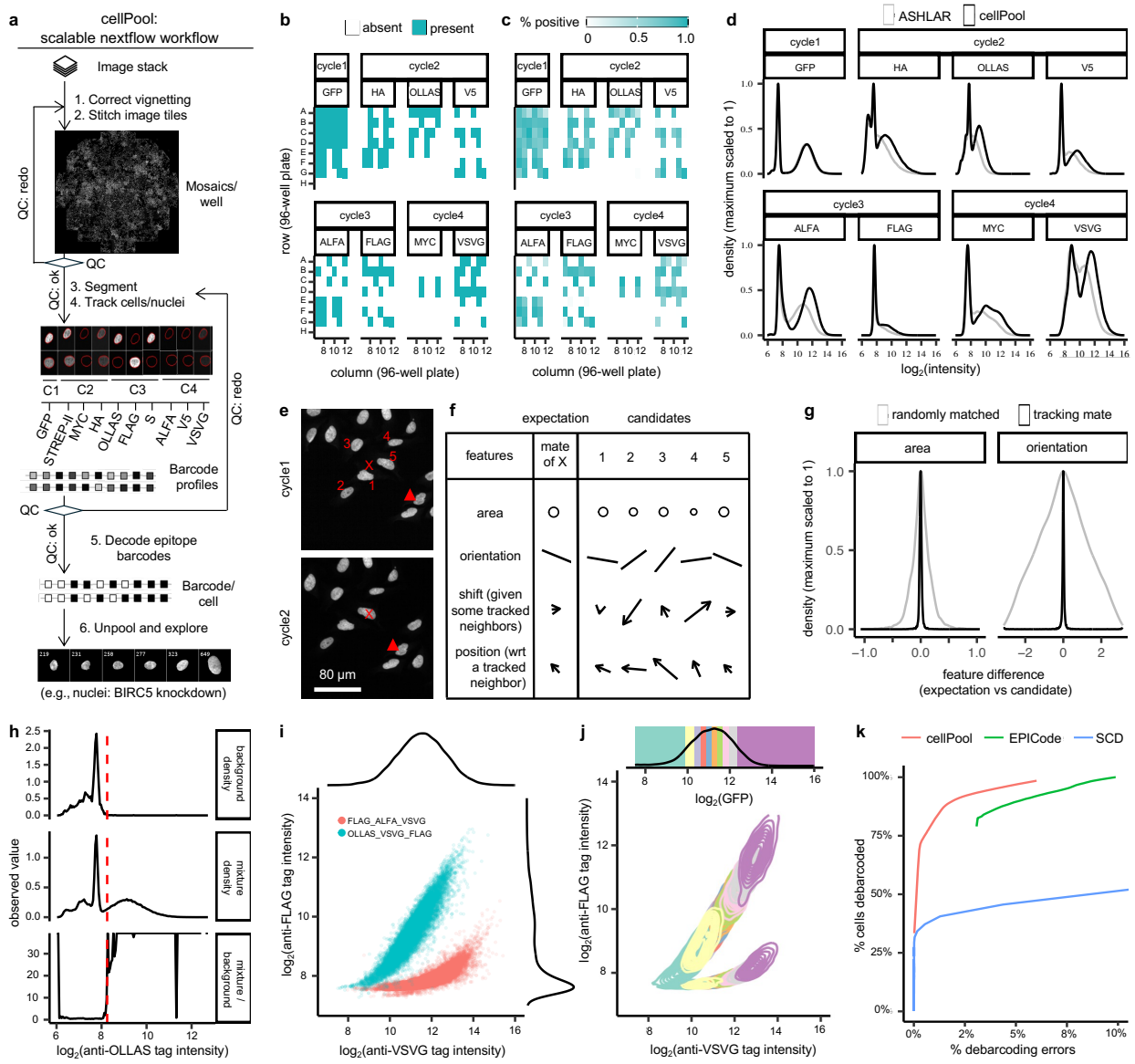
a
cellPool:
scalable nextflow workflow
Image stack
1. Correct vignetting
2. Stitch image tiles
Mosaics/
well
QC: redo
QC
QC: ok
3. Segment
4. Track cells/nuclei
C1 C2 C3 C4
GFP STREP-II MYC HA OLLAS FLAG S ALFA V5 VSVG
Barcode
profiles
QC
QC: ok
5. Decode epitope
barcodes
Barcode/
cell
6. Unpool and explore
(e.g., nuclei: BIRC5 knockdown)
b
absent
present
cycle1
cycle2
cycle3
cycle4
GFP
HA
OLLAS
V5
ALFA
FLAG
MYC
VSVG
row (96-well plate)
column (96-well plate)
c
% positive
0
0.5
1.0
d
ASHLAR
cellPool
density (maximum scaled to 1)
log2(intensity)
e
cycle1
cycle2
80 µm
f
expectation
candidates
features
mate
of X
1 2 3 4 5
area
orientation
shift (given
some tracked
neighbors)
position (wrt
a tracked
neighbor)
g
randomly matched
tracking mate
area
orientation
density (maximum scaled to 1)
feature difference
(expectation vs candidate)
h
observed value
background
density
mixture
density
mixture /
background
log2(anti-OLLAS tag intensity)
i
FLAG_ALFA_VSVG
OLLAS_VSVG_FLAG
log2(anti-FLAG tag intensity)
log2(anti-VSVG tag intensity)
j
log2(GFP)
log2(anti-FLAG tag intensity)
log2(anti-VSVG tag intensity)
k
cellPool
EPICode
SCD
% cells debarcoded
% debarcoding errors

Figure 2

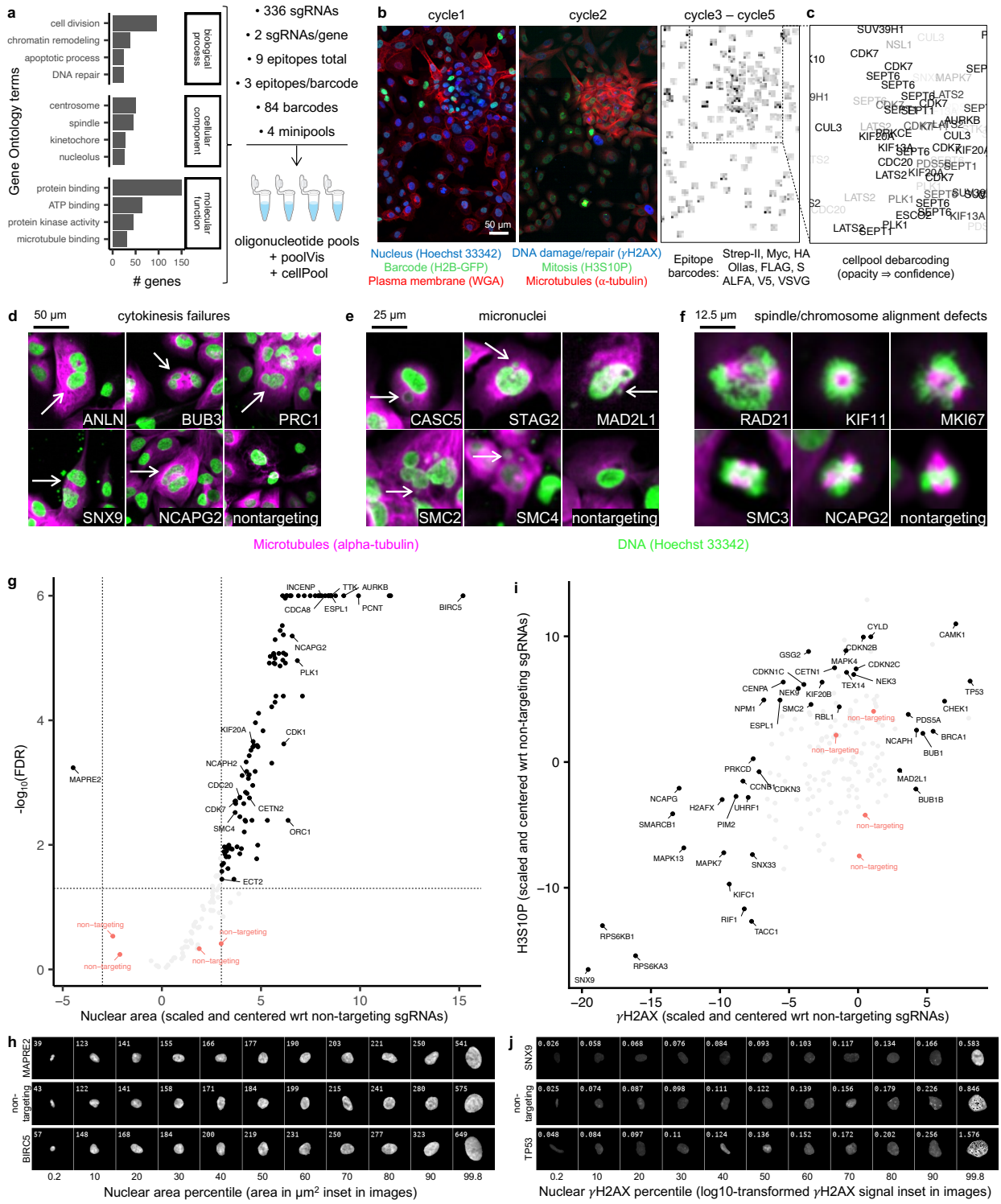
a
Gene Ontology terms
cell division
chromatin remodeling
apoptotic process
DNA repair
biological process
centrosome
spindle
kinetochore
nucleolus
cellular component
protein binding
ATP binding
protein kinase activity
microtubule binding
molecular function
genes
• 336 sgRNAs
• 2 sgRNAs/gene
• 9 epitopes total
• 3 epitopes/barcode
• 84 barcodes
• 4 minipools
oligonucleotide pools
+ poolVis
+ cellPool
b
cycle1
cycle2
cycle3 – cycle5
50 µm
Nucleus (Hoechst 33342)
Barcode (H2B-GFP)
Plasma membrane (WGA)
DNA damage/repair (γH2AX)
Mitosis (H3S10P)
Microtubules (α-tubulin)
Epitope barcodes:
Strep-II, Myc, HA
Ollas, FLAG, S
ALFA, V5, VSVG
c
cellpool debarcoding
(opacity ⇒ confidence)
d
50 µm
cytokinesis failures
ANLN
BUB3
PRC1
SNX9
NCAPG2
nontargeting
e
25 µm
micronuclei
CASC5
STAG2
MAD2L1
SMC2
SMC4
nontargeting
f
12.5 µm
spindle/chromosome alignment defects
RAD21
KIF11
MKI67
SMC3
NCAPG2
nontargeting
Microtubules (alpha-tubulin)
DNA (Hoechst 33342)
g
-log10(FDR)
Nuclear area (scaled and centered wrt non-targeting sgRNAs)
INCENP
TTK
AURKB
CDCA8
ESPL1
PCNT
BIRC5
NCAPG2
PLK1
KIF20A
CDK1
NCAPH2
MAPRE2
CDC20
CDK7
CETN2
SMC4
ORC1
ECT2
non-targeting
i
H3S10P (scaled and centered wrt non-targeting sgRNAs)
γH2AX (scaled and centered wrt non-targeting sgRNAs)
CYLD
CAMK1
GSG2
CDKN2B
MAPK4
CDKN2C
CDKN1C
CETN1
TEX14
NEK3
CENPA
NEK9
KIF20B
TP53
NPM1
SMC2
RBL1
CHEK1
ESPL1
PDS5A
NCAPH
BRCA1
BUB1
PRKCD
MAD2L1
CCNB1
CDKN3
BUB1B
NCAPG
H2AFX
UHRF1
SMARCB1
PIM2
MAPK13
MAPK7
SNX33
KIFC1
RIF1
TACC1
RPS6KB1
RPS6KA3
SNX9
non-targeting
h
MAPRE2
non-targeting
BIRC5
Nuclear area percentile (area in µm² inset in images)
j
SNX9
non-targeting
TP53
Nuclear γH2AX percentile (log10-transformed γH2AX signal inset in images)

Figure 3

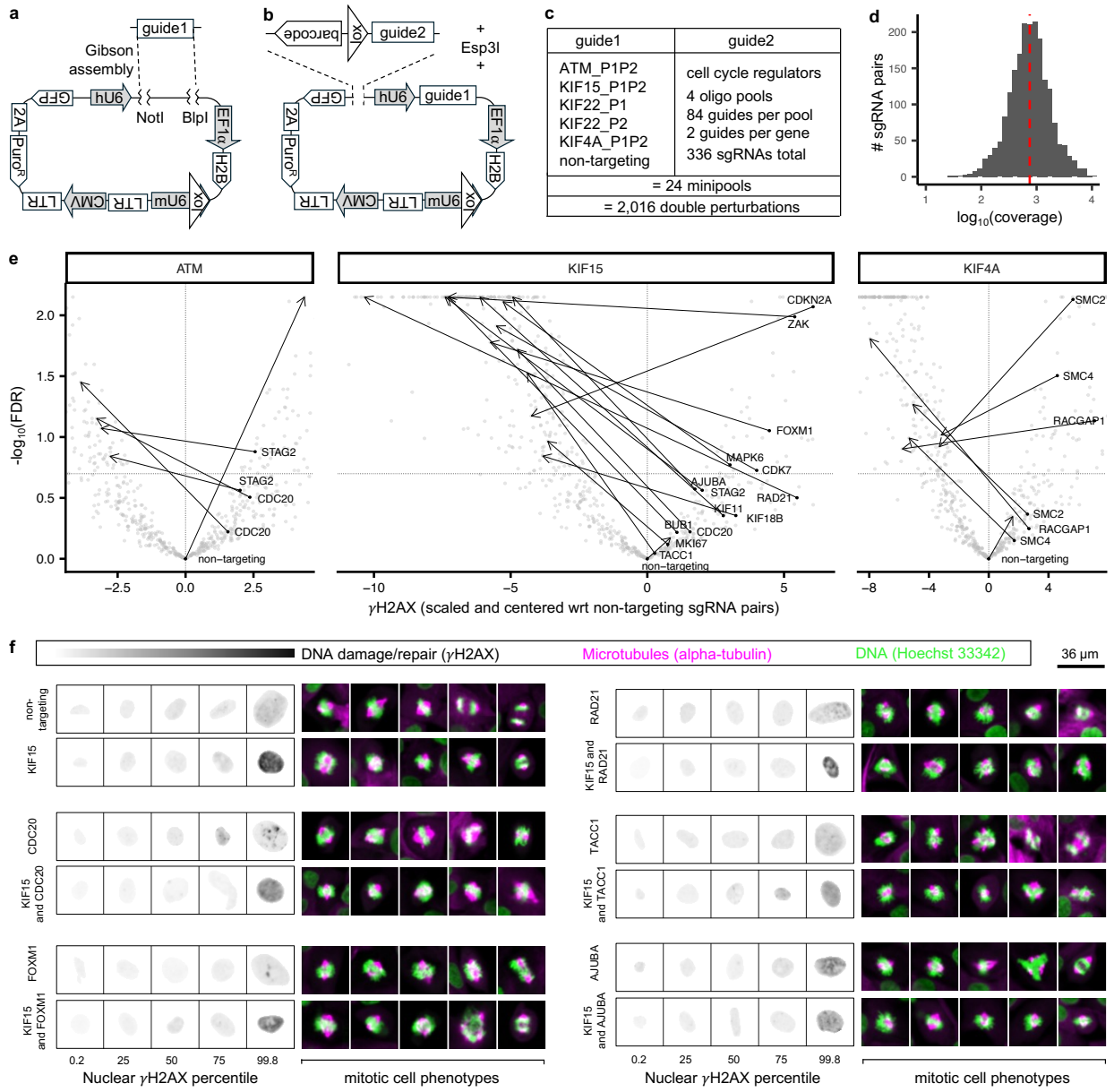
a
guide1
Gibson assembly
GFP
hU6
NotI
BlpI
EF1α
H2B
2A
PuroR
LTR
CMV
LTR
mU6
xo
b
barcode
xo
guide2
+
Esp3I
+
GFP
hU6
guide1
EF1α
H2B
2A
PuroR
LTR
CMV
LTR
mU6
xo
c
guide1
guide2
ATM_P1P2
KIF15_P1P2
KIF22_P1
KIF22_P2
KIF4A_P1P2
non-targeting
cell cycle regulators
4 oligo pools
84 guides per pool
2 guides per gene
336 sgRNAs total
= 24 minipools
= 2,016 double perturbations
d
# sgRNA pairs
log10(coverage)
e
ATM
KIF15
KIF4A
-log10(FDR)
STAG2
STAG2
CDC20
CDC20
non−targeting
CDKN2A
ZAK
FOXM1
MAPK6
CDK7
AJUBA
STAG2
RAD21
KIF11
KIF18B
BUB1
CDC20
MKI67
TACC1
non−targeting
SMC2
SMC4
RACGAP1
SMC2
RACGAP1
SMC4
non−targeting
γH2AX (scaled and centered wrt non-targeting sgRNA pairs)
f
DNA damage/repair (γH2AX)
Microtubules (alpha-tubulin)
DNA (Hoechst 33342)
36 μm
non-targeting
KIF15
CDC20
KIF15 and CDC20
FOXM1
KIF15 and FOXM1
RAD21
KIF15 and RAD21
TACC1
KIF15 and TACC1
AJUBA
KIF15 and AJUBA
0.2
25
50
75
99.8
Nuclear γH2AX percentile
mitotic cell phenotypes

Figure 4